Title
The Role of Conditional Likelihoods in Latent Variable Modeling

Journal
Psychometrika, 87(3)

ISSN
0033-3123

Authors
Skrondal, Anders
Rabe-Hesketh, Sophia

Publication Date
2022-09-01

DOI
10.1007/s11336-021-09816-8

Peer reviewed

PSYCHOMETRIKA—VOL. 87, NO. 3, 799–834
SEPTEMBER 2022
https://doi.org/10.1007/s11336-021-09816-8

THE ROLE OF CONDITIONAL LIKELIHOODS IN LATENT VARIABLE MODELING

ANDERS SKRONDAL
NORWEGIAN INSTITUTE OF PUBLIC HEALTH
UNIVERSITY OF OSLO
UNIVERSITY OF CALIFORNIA, BERKELEY

SOPHIA RABE-HESKETH
UNIVERSITY OF CALIFORNIA, BERKELEY

In psychometrics, the canonical use of conditional likelihoods is for the Rasch model in measurement. Whilst not disputing the utility of conditional likelihoods in measurement, we examine a broader class of problems in psychometrics that can be addressed via conditional likelihoods. Specifically, we consider cluster-level endogeneity where the standard assumption that observed explanatory variables are independent from latent variables is violated. Here, "cluster" refers to the entity characterized by latent variables or random effects, such as individuals in measurement models or schools in multilevel models and "unit" refers to the elementary entity such as an item in measurement. Cluster-level endogeneity problems can arise in a number of settings, including unobserved confounding of causal effects, measurement error, retrospective sampling, informative cluster sizes, missing data, and heteroskedasticity. Severely inconsistent estimation can result if these challenges are ignored.

Key words: Endogeneity, Fixed effects, Random effects, Conditional maximum likelihood, Marginal maximum likelihood, Unobserved confounding, Measurement error, Retrospective sampling, Informative cluster size, Missing data, Heteroskedasticity.

1. Introduction

As is often the case for concepts in statistics, the term "conditional likelihood" has many meanings. It has, for instance, been used to refer to likelihoods where conditioning is on (1) exogenous explanatory variables (e.g., Gourieroux & Monfort, 1995), (2) latent variables (e.g., Aigner et al., 1984), (3) the outcome variable, for instance in capture-recapture modeling of population size (e.g., Sanathanan, 1972) and ascertainment correction in biometrical genetics (e.g., Pfeiffer et al., 2001), (4) previous outcomes, for instance in autoregressive time-series models (e.g., Box & Jenkins, 1976) and peeling in phylogenetics (e.g., Felsenstein, 1981), (5) order statistics (e.g., Kalbfleisch, 1978), or (6) sufficient statistics.

In this address we follow the seminal theoretical work of Andersen (1970, 1973a) and Kalbfleisch and Sprott (1970) and let a conditional likelihood be obtained by *conditioning on sufficient statistics for incidental parameters in order to eliminate these parameters*. In the context of latent variable or mixed effects modeling, the incidental parameters are the values taken by latent variables for a set of clusters, for example individuals or organizational units.

Presidential address presented by Anders Skrondal at IMPS 2017 in Zürich, Switzerland. This article is based on joint work with Sophia Rabe-Hesketh

Correspondence should be made to Anders Skrondal, CEFH, Norwegian Institute of Public Health, P.O.Box 222 Skøyen, N-0213 Oslo, Norway. Email: anders.skrondal@fhi.no

In psychometrics, the canonical use of conditional likelihoods is in measurement relying on the Rasch model (Rasch, 1960) and its extensions. As demonstrated by Rasch, estimation of item parameters can in this case be based on a conditional likelihood where the person parameters are eliminated by conditioning on their sufficient statistics. It is often argued that Rasch models and conditional maximum likelihood (CML) estimation are advantageous in measurement (e.g., Fischer, 1995a). Indeed, Molenaar (1995) closes his excellent overview of estimation of Rasch models in the following way:

> "Unless there are clear reasons for a different decision, the present author would recommend to use CML estimates."

Conditional likelihoods have been used for a variety of problems in measurement; see Fischer (1995b; 1995c), Formann (1995), Maris and Bechger (2007), Verhelst (2019), von Davier and Rost (1995), and Zwitser and Maris (2015) for a small selected sample.

We are certainly not disputing the utility of CML estimation in measurement. However, we will argue that conditional likelihoods perhaps have a more important role to play in addressing endogeneity problems in psychometrics. Focus will be on cluster-level endogeneity, where covariates and latent variables are dependent, a problem ignored by popular methods which can therefore produce severely inconsistent estimates. Fortunately, CML estimation, an instance of what is referred to as "fixed-effects estimation" in econometrics, can rectify this problem.

Our plan is as follows. First we introduce some latent variable models and discuss the cluster-level endogeneity problem whose origins, effects and alleviation we will examine. We proceed to delineate the ideas of protective and mitigating estimation of target parameters before describing the incidental parameter problem of joint maximum likelihood (JML) estimation. Two approaches that address that problem are discussed: marginal maximum likelihood (MML) and conditional maximum likelihood (CML) estimation. We demonstrate that CML estimation, in contrast to MML estimation, handles cluster-level endogeneity, and describe an endogeneity-correcting feature of MML estimation for large clusters. The scope of CML estimation is then extended followed by a discussion of MML estimation of augmented models that can accommodate cluster-level endogeneity. Several reasons for cluster-level endogeneity are investigated (unobserved cluster-level confounding of causal effects, cluster-specific measurement error, retrospective sampling, informative cluster sizes, missing data, and heteroskedasticity) and we show how different estimators perform in these situations. Thereafter, we discuss latent variable scoring before closing the paper with some concluding remarks.

2. Clustered Data

We consider data consisting of clusters j ($j = 1, \ldots, N$) that contain units ij ($i = 1, \ldots, n_j$). Units are typically exchangeable within clusters in cross-sectional multilevel designs. An example is students ij nested in schools j, where the index i associated with the students within a school is arbitrary.

Units are non-exchangeable within clusters in two settings: (a) longitudinal designs where i is the chronological sequence number of the time-point when a subject was observed and (b) measurement designs where i is the item (or question) responded to by a subject. In the non-exchangeable case, when i corresponds to the same time-point or item across subjects j, we will refer to i as an "item." The different kinds of clustered data are illustrated in Fig. 1.

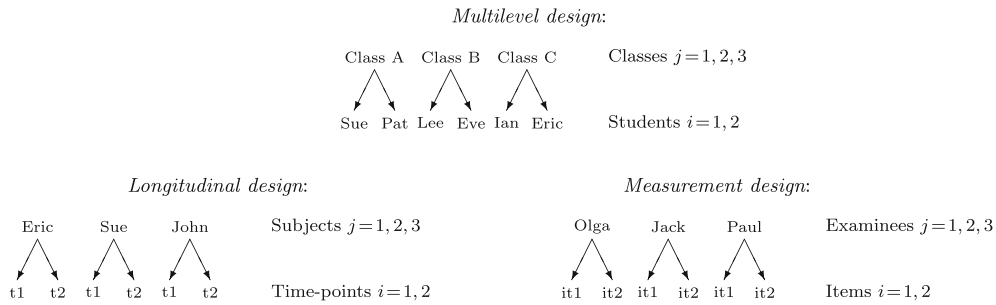

FIGURE 1.
Illustration of clustered data for $N = 3$ clusters and $n = 2$ units per cluster. Exchangeable units (upper panel) and non-exchangeable units (lower panel).

3. Latent Variable Model

We consider generalized linear mixed models (GLMMs) with canonical link functions (see, e.g., Rabe-Hesketh & Skrondal, 2009). Given the cluster-specific latent variables or random effects $\boldsymbol{\zeta}_j$, the model for an outcome y_{ij} is a generalized linear model (GLM) with three components (e.g., Nelder & Wedderburn, 1972): a linear predictor ν_{ij}, a link function $g(\mu_{ij}) = \nu_{ij}$ that links the linear predictor to the conditional expectation μ_{ij} of the outcome, and a conditional outcome distribution from the exponential family.

For a GLMM, we express the linear predictor as

$$\nu_{ij} \;=\; \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{v}'_j\boldsymbol{\gamma} + \mathbf{z}'_{ij}\boldsymbol{\zeta}_j, \tag{1}$$

where:

- $\boldsymbol{\beta}$ is a vector of parameters for the unit-specific vector $\mathbf{x}_{ij}$. For exchangeable units $\boldsymbol{\beta}$ are regression coefficients for unit-specific covariates $\mathbf{x}_{ij}$. For non-exchangeable units $\boldsymbol{\beta}$ contains a vector of item-specific intercepts and a vector of regression coefficients. Correspondingly, $\mathbf{x}_{ij}$ includes an elementary vector (where one of the elements is 1 and the other elements are 0) that picks out the intercept for item i, and item-specific and/or unit-specific covariates.
- $\boldsymbol{\gamma}$ is a vector of parameters for the cluster-specific vector $\mathbf{v}_j$. For non-exchangeable units $\boldsymbol{\gamma}$ are regression coefficients for cluster-specific covariates $\mathbf{v}_j$. For exchangeable units $\boldsymbol{\gamma}$ includes an overall intercept and regression coefficients for the cluster-specific covariates in $\mathbf{v}_j$, and $\mathbf{v}_j$ includes a 1 and cluster-specific covariates.
- $\boldsymbol{\zeta}_j$ is a vector of cluster-specific latent variables or random intercept and possibly random coefficients for the vector $\mathbf{z}_{ij}$ of item-specific and/or unit-specific covariates (that are often partly overlapping with $\mathbf{x}_{ij}$)

The conditional expectation of the outcome, given the covariates and latent variables, is

$$\mu_{ij} \equiv \mathrm{E}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{v}_j, \boldsymbol{\zeta}_j) \;=\; g^{-1}(\nu_{ij}),$$

and the conditional outcome distribution can be written as

$$p(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{v}_j, \boldsymbol{\zeta}_j) \;=\; \exp\left\{\frac{y_{ij}\nu_{ij} - b(\nu_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \tag{2}$$

where ϕ is the scale or dispersion parameter, and $b(\cdot)$ and $c(\cdot)$ are functions depending on the member of the exponential family.

We confine our treatment to three GLMs:

(i) Normal distribution (where $\phi=\sigma^2$) and identity link for continuous outcomes,
$p(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{v}_j,\boldsymbol{\zeta}_j)=(\sigma\sqrt{2\pi})^{-1}\exp\{-\frac{1}{2\sigma^2}(y_{ij}-\nu_{ij})^2\}$ and $g(\mu_{ij})=\mu_{ij}$

(ii) Bernoulli distribution and logit link for binary outcomes,
$p(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{v}_j,\boldsymbol{\zeta}_j)=\mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}$ and $g(\mu_{ij})=\log\left\{\frac{\mu_{ij}}{1-\mu_{ij}}\right\}$

(iii) Poisson distribution and log link for counts,
$p(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{v}_j,\boldsymbol{\zeta}_j)=\exp[-\exp(\nu_{ij})]\exp(\nu_{ij})^{y_{ij}}/y_{ij}!$ and $g(\mu_{ij})=\log(\mu_{ij})$

Other members of the exponential family include the gamma and inverse-Gaussian distributions.

For simplicity we concentrate on a special case of (1) with linear predictor

$$\nu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{v}_j'\boldsymbol{\gamma}+\zeta_j. \tag{3}$$

It should be emphasized that this model encompasses popular latent variable or mixed models, such as generalized linear random-intercept (multilevel or hierarchical) models, and Rasch models (Rasch, 1960) and their extensions such as "explanatory" IRT (De Boeck & Wilson, 2004).

We will in the sequel also use extensions of GLMMs such as generalized linear latent and mixed models (GLLAMMs) of Rabe-Hesketh et al. (2004) and Skrondal and Rabe-Hesketh (2004).

4. Cluster-Level Exogeneity and Endogeneity

Our focus is on the challenges that arise in estimation of latent variable models when there is cluster-level endogeneity. Before embarking on the challenges we must explicitly define what we mean by this term. Let $\mathbf{w}_j$ represent all observed covariates for cluster j. We say that there is *cluster-level exogeneity* if all covariates are independent of the cluster-specific intercepts; $\mathbf{w}_j \perp\!\!\!\perp \zeta_j$. In contrast, *cluster-level endogeneity* occurs if at least one covariate in $\mathbf{w}_j$ is not independent of ζ_j; $\mathbf{w}_j \not\perp\!\!\!\perp \zeta_j$.

The definitions of cluster-level exogeneity and cluster-level endogeneity are represented in the graphs in the left and right panels of Fig. 2, respectively. This kind of graph, which we find useful and will use throughout, resembles traditional directed acyclic graphs (DAGs) but nodes can represent vectors of random variables here. An arrow between two nodes means that the probability distribution of the node that the arrow points to depends on the value taken by the emanating node. The undirected arc between ζ_j and $\mathbf{w}_j$ in the right panel indicates that there is dependence between ζ_j and at least one element of $\mathbf{w}_j$.

When exploring reasons for cluster-level endogeneity in Section 12 we will rather informally rely on the d-separation criterion (e.g., Verma & Pearl, 1988) and the equivalent moralization criterion (Lauritzen et al., 1990) to infer cluster-level endogeneity from graphs of latent variable models, assuming "stability" or "faithfulness" to preclude dependence paths cancelling out. Sometimes we will examine likelihood contributions to show how cluster-level endogeneity arises and the consequences.

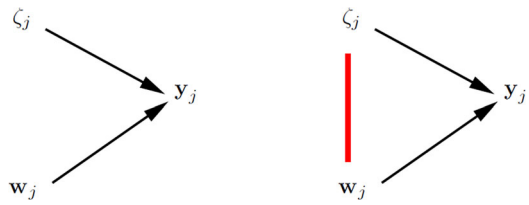

FIGURE 2.
Cluster-level exogeneity (left panel) and cluster-level endogeneity (right panel).

5. Protective and Mitigating Estimation of Target Parameters

In this address we will focus on the performance of point estimators under cluster-level endogeneity as the number of clusters N becomes large, whereas the cluster sizes n_j are fixed and could be small. A classical goal in statistical modeling is (weak) *consistency* of estimators for *all* model parameters as $N \to \infty$. However, this typically requires a correctly specified model, an assumption that we often deem to be naive.

A less ambitious but more realistic goal is *protective estimation* which is consistent for target parameters but possibly inconsistent for other parameters (Skrondal & Rabe-Hesketh, 2014). Our target parameters throughout will be the subset of coefficients in $\boldsymbol{\beta}$ corresponding to the unit-specific covariates $\mathbf{x}_{ij}$ in (3). These covariates could be time-varying variables in a longitudinal study, characteristics of units ij in a multilevel study, or attributes of items i or item-subject combinations ij in measurement.

An even less ambitious goal is what we will refer to as *mitigating estimation* where it is likely (but not guaranteed) that estimation of the target parameters $\boldsymbol{\beta}$ is less inconsistent than conventional estimation that ignores misspecification. Although mitigation in this sense cannot be formally proved, it can be made plausible by theoretical arguments and based on evidence from simulations. The hope is that "almost consistent" estimators (e.g., Laisney & Lechner, 2003) can be obtained. In reality, mitigating estimation will sometimes be the most realistic goal.

6. Incidental Parameter Problems and Their Solutions

The distinction between *structural parameters* and *incidental parameters* was introduced in a seminal paper by Neyman and Scott (1948). For linear predictor (3), the structural parameters $\boldsymbol{\vartheta}$ include $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (and σ^2 if relevant), whereas the ζ_j are incidental parameters because their number increases in tandem with the number of clusters N. In econometrics a structural parameter is usually a causal parameter and for this reason Lancaster (2000) used the term "common parameter".

Let $\mathbf{y}$ and $\mathbf{w}$ denote all outcomes and covariates for the sample, respectively. Assume that the outcomes $\mathbf{y}_j$ for the clusters are conditionally independent across clusters and the outcomes y_{ij} for cluster j are conditionally independent, given the covariates and latent variable ζ_j. The joint likelihood for the structural parameters $\boldsymbol{\vartheta}$ and the latent variables $\zeta_1, \ldots, \zeta_N$ (here treated as unknown parameters) becomes

$$p(\mathbf{y}|\mathbf{w}; \boldsymbol{\vartheta}, \zeta_1, \ldots, \zeta_N) = \prod_{j=1}^{N} \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{w}_j; \boldsymbol{\vartheta}, \zeta_j). \tag{4}$$

The incidental parameter problem (Neyman & Scott, 1948; see also Lancaster, 2000) refers to the fact that *joint maximum likelihood* (JML) estimation of both structural and incidental parameters need not be consistent for the structural parameters $\boldsymbol{\vartheta}$ as $N \to \infty$ for fixed cluster sizes n_j. The problem arises because estimation of each ζ_j must often rely on a small number of units n_j in the cluster. Viewing the cluster sizes as produced by $n_j = n \times m_j$, where m_j has a mean of 1, the inconsistency in estimating $\boldsymbol{\vartheta}$ for the models considered here is of order n^{-1} (e.g., Arellano & Hahn, 2007). Note that JML estimation has also been referred to as unconditional maximum likelihood estimation (e.g., Wright & Douglas, 1977) and unconstrained maximum likelihood estimation (e.g., de Leeuw & Verhelst, 1986) in psychometrics.

There is no incidental parameter problem when the joint likelihood can be factorized into two components, one just containing structural parameters and the other just incidental parameters. Such likelihood orthogonality (e.g., Lancaster, 2000) occurs for linear predictor (3) with (a) identity link and normal conditional distribution (e.g., Chamberlain, 1980) and (b) log link and Poisson conditional distribution (e.g., Cameron & Trivedi, 1999). For these models JML estimation is consistent for $\boldsymbol{\beta}$ when $N \to \infty$ for fixed n_j.

In general, consistent JML estimation can be achieved under a double-asymptotic scheme where both the number of units per cluster increases $n \to \infty$ and the number of clusters increases $N \to \infty$. In psychometrics, a classical result is that $\widehat{\boldsymbol{\beta}}^{\text{JML}}$ is consistent for the Rasch model in this case if $\frac{N}{n} \to \infty$ (Haberman, 1977). Based on simulation evidence, Greene (2004) observed that $\widehat{\boldsymbol{\beta}}^{\text{JML}}$ appears to be consistent for many latent variable models used in econometrics under double asymptotics. However, appealing to double asymptotics is unconvincing when n is not large.

For the simple Rasch model, the inconsistency of JML estimation can be derived and corrected. When $n = 2$, Andersen (1973a) showed that $p\text{lim}\, \widehat{\boldsymbol{\beta}}^{\text{JML}} = 2\boldsymbol{\beta}$ as $N \to \infty$, so $\frac{1}{2}\widehat{\boldsymbol{\beta}}^{\text{JML}}$ is consistent. For general n, Wright and Douglas (1977) observed that the finite sample bias is approximately $\frac{1}{n-1}\boldsymbol{\beta}$ and discussed the bias correction $\frac{n-1}{n}\widehat{\boldsymbol{\beta}}^{\text{JML}}$. Andersen (1980, Theorem 6.1) stated the same result for inconsistency. For more complex models, methods that reduce inconsistency from order n^{-1} to n^{-2} are discussed in Arellano and Hahn (2007). For instance, a *modified profile likelihood* where the incidental parameters ζ_j are "profiled out" of the joint likelihood has been used for models with linear predictors such as (3) by Bellio and Sartori (2006) and Bartolucci et al. (2016). This approach can produce mitigating estimation.

An approach usually called *marginal maximum likelihood* (MML) estimation in psychometrics is the most popular for linear predictor (3). Here, ζ_j is treated as a random variable and "integrated out" of the joint likelihood, as proposed in early work by Kiefer and Wolfowitz (1956). Note that the statistical literature typically refers to this likelihood as *integrated* and that their marginal likelihood "transforms away" incidental parameters (e.g., Kalbfleisch & Sprott, 1970). The terms unconditional maximum likelihood estimation (e.g., Bock & Lieberman, 1970) and, simply, maximum likelihood estimation (e.g., Holland, 1990) have also been used in psychometrics. Under assumptions including cluster-level exogeneity, MML is consistent for all model parameters.

Alternatively, *conditional maximum likelihood* (CML) estimation can be used where ζ_j is treated as a fixed parameter and "conditioned out" of the joint likelihood. The idea of CML estimation was discussed already by Bartlett (1936, 1937a), the eminent British statistician whose name is associated with factor scores in psychometrics (e.g., Bartlett, 1937b). We will see that CML can yield protective estimation of $\boldsymbol{\beta}$ under cluster-level endogeneity.

7. Marginal Maximum Likelihood (MML) Estimation

In marginal maximum likelihood (MML) estimation the cluster-specific intercept ζ_j is treated as a *random variable* in estimation. The following assumptions are usually made:

[A.1] Cluster independence: $p(\mathbf{y}|\mathbf{w}, \zeta_1, \ldots, \zeta_N; \boldsymbol{\vartheta}) = \prod_{j=1}^{N} p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j; \boldsymbol{\vartheta})$
[A.2] Conditional unit independence: $p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j; \boldsymbol{\vartheta}) = \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{w}_j, \zeta_j; \boldsymbol{\vartheta})$
[A.3] Strict exogeneity conditional on the latent variable:
$p(y_{ij}|\mathbf{w}_j, \zeta_j; \boldsymbol{\vartheta}) = p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j; \boldsymbol{\vartheta})$; i.e., given the latent variable, the outcome for a unit only depends on covariates for that unit
[A.4] Correct conditional distribution: $p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j; \boldsymbol{\vartheta})$ follows (2) and (3)
[A.5] Cluster-level exogeneity: $p(\zeta_j|\mathbf{w}_j) = p(\zeta_j)$
[A.6] Latent variable normality: $p(\zeta_j) = \phi(\zeta_j; 0, \psi)$; a normal density with zero expectation and variance ψ

Using [A.2]-[A.6], the marginal likelihood contribution of cluster j simplifies in the following way:

$$\begin{aligned}\mathcal{L}_j^{\mathrm{MML}} \equiv p(\mathbf{y}_j|\mathbf{w}_j; \boldsymbol{\vartheta}) &= \int_{\zeta_j} p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j; \boldsymbol{\vartheta})\, p(\zeta_j|\mathbf{w}_j)\, \mathrm{d}\zeta_j \\ &= \int_{\zeta_j} \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j; \boldsymbol{\vartheta}) \right\} \phi(\zeta_j; 0, \psi)\, \mathrm{d}\zeta_j,\end{aligned}$$

where we see that ζ_j is marginalized over or integrated out of the joint likelihood.

$\phi(\zeta_j; 0, \psi)$ can be interpreted as the density of a cluster-specific disturbance in a data-generating mechanism or as a superpopulation density of clusters in survey sampling. That the ζ_j are independently and identically distributed random variables can be motivated by exchangeability of the clusters (e.g., Draper, 1995).

Using [A.1], MML estimation proceeds by maximizing the likelihood $\mathcal{L}^{\mathrm{MML}} = \prod_{j=1}^{N} \mathcal{L}_j^{\mathrm{MML}}$ w.r.t. $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and ψ (and σ^2 if relevant). If the above assumptions are satisfied, MML estimators are consistent as $N \to \infty$ for fixed n_j for *all* parameters under appropriate regularity conditions (e.g., Butler & Louis, 1997). Importantly, standard MML estimation becomes inconsistent, possibly severely, for all link functions when the exogeneity assumptions are violated. As we will see momentarily, inconsistency due to violation of [A.5] can arise because MML estimation of $\boldsymbol{\beta}$ exploits both within-cluster and between-cluster information, and the latter can be contaminated by cluster-level endogeneity.

7.1. MML Estimation for Identity Link and Normal Distribution

It is instructive to consider the linear predictor (3) with an identity link and a normal conditional distribution that can be written as

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{v}_j'\boldsymbol{\gamma} + \zeta_j + \epsilon_{ij}, \tag{5}$$

where ϵ_{ij} is an additive normally distributed unit-level error term, $p(\epsilon_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j) = \phi(\epsilon_{ij}; 0, \sigma^2)$.

For identity links the assumptions given above are stricter than necessary for consistent MML estimation. First, for identity and log links, [A.4] can be replaced by the more lenient assumption that μ_{ij} is correctly specified (and the domain of y_{ij} for the assumed exponential family distribution encompasses the domain of the correct distribution). This extends the idea of pseudo maximum likelihood (PML) estimation (Gourieroux et al., 1984) to what we may call pseudo marginal maximum likelihood estimation in the latent variable setting. For the identity link, [A.3], [A.4], and [A.5] can be replaced by (5) with a weaker set of assumptions where normality is relaxed for ζ_j and ϵ_{ij}, $\mathrm{E}(\epsilon_{ij}|\mathbf{w}_j, \zeta_j) = 0$ (a mean-independence version of "unit-level

exogeneity"), and $\mathrm{E}(\zeta_j|\mathbf{w}_j) = 0$ (e.g., Wooldridge, 2010, p. 292). Second, for the identity link, consistent estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ neither requires assumption [A.2], see Zeger et al. (1988), nor assumption [A.6], see Verbeke and Lesaffre (1997).

We now outline how MML estimation relies on both between-cluster and within-cluster information, for simplicity omitting $\mathbf{v}_j'\boldsymbol{\gamma}$ from model (5) and letting $n_j = n$. The total sum of squares of y_{ij} can then be decomposed into two contributions:

$$T_{yy} = \sum_{j=1}^{N}\sum_{i=1}^{n}(y_{ij} - \overline{y}_{..})^2 = \sum_{j=1}^{N}\sum_{i=1}^{n}(y_{ij} - \overline{y}_{.j})^2 + \sum_{j=1}^{N}\sum_{i=1}^{n}(\overline{y}_{.j} - \overline{y}_{..})^2 = W_{yy} + B_{yy}, \tag{6}$$

where W_{yy} represents the within-cluster variation and B_{yy} the between-cluster variation. We use similar decompositions of $T_{\mathbf{xx}}$ into $W_{\mathbf{xx}}$ and $B_{\mathbf{xx}}$, and $T_{\mathbf{x}y}$ into $W_{\mathbf{x}y}$ and $B_{\mathbf{x}y}$. For known variance components ψ and σ^2, the MML estimator is the generalized least squares (GLS) estimator that Maddala (1971) shows can be expressed as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{GLS}} = (W_{\mathbf{xx}} + \omega B_{\mathbf{xx}})^{-1}(W_{\mathbf{x}y} + \omega B_{\mathbf{x}y}), \tag{7}$$

where $\omega \equiv \frac{\sigma^2}{\sigma^2 + n\psi}$ is the weight given to the between-cluster variation. The GLS estimator in essence combines the between-cluster and within-cluster estimators of $\boldsymbol{\beta}$ by weighting them in inverse proportion to their respective variances. Fuller and Battese (1973) demonstrate that the GLS estimator can be obtained by using ordinary least squares (OLS) for the transformed data $\widetilde{y}_{ij} = y_{ij} - \theta\overline{y}_{.j}$ and $\widetilde{\mathbf{x}}_{ij} = \mathbf{x}_{ij} - \theta\overline{\mathbf{x}}_{.j}$, where $\theta = 1 - \sqrt{\omega}$. The probability limit of $\widehat{\boldsymbol{\beta}}^{\mathrm{GLS}}$ as $N \to \infty$ can be expressed as

$$p\lim \widehat{\boldsymbol{\beta}}^{\mathrm{GLS}} = \boldsymbol{\beta} + \omega\, \Sigma_{\widetilde{\mathbf{x}}}^{-1}\Sigma_{\mathbf{x}\zeta}, \tag{8}$$

where $\Sigma_{\widetilde{\mathbf{x}}}$ is the covariance matrix of $\widetilde{\mathbf{x}}_{ij}$, and $\Sigma_{\mathbf{x}\zeta}$ is the covariance matrix of $\mathbf{x}_{ij}$ with ζ_j. Importantly, the estimator is inconsistent if cluster-level exogeneity [A.5] is violated because this implies that $\Sigma_{\mathbf{x}\zeta} \neq \mathbf{0}$.

Analytical integration is trivial for models with conjugate latent variable densities for which $\mathcal{L}^{\mathrm{MML}}$ can be written in closed form. For a GLMM with linear predictor (3) and identity link, $\mathcal{L}^{\mathrm{MML}}$ simply takes the form of the (ordinary) likelihood of a multivariate normal regression model with $\mathrm{E}(\mathbf{y}_j \mid \mathbf{w}_j) = \mathbf{X}_j\boldsymbol{\beta} + (\mathbf{1}_{n_j} \otimes \mathbf{v}_j')\boldsymbol{\gamma}$ (where $\mathbf{x}_{1j}', \ldots, \mathbf{x}_{n_j j}'$ are the rows of $\mathbf{X}_j$) and $\mathrm{Var}(\mathbf{y}_j \mid \mathbf{w}_j) = \psi\mathbf{1}_{n_j}\mathbf{1}_{n_j}' + \sigma\mathbf{I}_{n_j}$. MML estimation of linear mixed models when variance components are unknown is discussed by Laird and Ware (1982) using the EM algorithm and by Goldstein (1986) using iterative generalized least squares (IGLS).

7.2. MML Estimation for Logit Link and Bernoulli Distribution or Log Link and Poisson Distribution

Because μ_{ij} is nonlinear in ζ_j for log and logit links, $\mathcal{L}^{\mathrm{MML}}$ cannot be expressed in closed form and maximization is usually based on numerical integration (e.g., Rabe-Hesketh et al., 2005) or Monte Carlo integration (e.g., Booth & Hobert, 1999).

Unfortunately, it is difficult to assess the normality assumption for the latent variables [A.6] for logit links. However, MML estimators for regression coefficients are almost consistent if [A.6] is violated, whereas estimators for intercepts and random effect variances can be severely inconsistent if the correct latent variable density is highly skewed (e.g., Neuhaus et al., 1992). The threats posed by violation of the cluster-level exogeneity assumption [A.5] persist.

8. Conditional Maximum Likelihood (CML) Estimation

We retain assumptions [A.1]-[A.4] stated for MML estimation but now relax assumptions [A.5] and [A.6] regarding the latent variable distribution.

Using the exponential family distribution (2) in conjunction with linear predictor (3), we obtain

$$p(\mathbf{y}_j|\mathbf{w}_j;\boldsymbol{\vartheta},\zeta_j) \propto \exp\left\{\boldsymbol{\beta}'\sum_{i=1}^{n_j}\mathbf{x}_{ij}y_{ij} + (\zeta_j+\boldsymbol{\gamma}'\mathbf{v}_j)\sum_{i=1}^{n_j}y_{ij} - \sum_{i=1}^{n_j}b(\nu_{ij})\right\}.$$

It follows from the Neyman-Fisher factorization theorem (e.g., Pawitan, 2001, Theorem 3.1) that for known $\boldsymbol{\beta}$, the cluster-specific sumscore of the outcomes, $\sum_{i=1}^{n_j} y_{ij}$, is a sufficient statistic for $\zeta_j + \boldsymbol{\gamma}'\mathbf{v}_j$ (and that $\sum_{i=1}^{n_j}\mathbf{x}_{ij}y_{ij}$ is a sufficient statistic for $\boldsymbol{\beta}$).

The conditional likelihood contribution of cluster j, given $\tau_j \equiv \sum_{i=1}^{n_j} y_{ij}$, can be expressed as

$$\mathcal{L}_j^{\text{CML}} \equiv p(\mathbf{y}_j|\tau_j,\mathbf{w}_j;\boldsymbol{\vartheta},\zeta_j) = \frac{\prod_{i=1}^{n_j}p(y_{ij}|\mathbf{x}_{ij},\mathbf{v}_j;\boldsymbol{\vartheta},\zeta_j)}{p(\sum_{i=1}^{n_j}y_{ij}|\mathbf{x}_{ij},\mathbf{v}_j;\boldsymbol{\vartheta},\zeta_j)}.$$

Importantly, the cluster-specific term $\zeta_j + \mathbf{v}_j'\boldsymbol{\gamma}$ cancels out of the numerator and denominator of $\mathcal{L}_j^{\text{CML}}$ and the latent variable assumptions [A.5] and [A.6] are therefore no longer required.

CML estimation proceeds by maximizing the conditional likelihood $\mathcal{L}^{\text{CML}} = \prod_{j=1}^{N}\mathcal{L}_j^{\text{CML}}$ w.r.t. $\boldsymbol{\vartheta}$, where $\boldsymbol{\vartheta}$ is a vector containing $\boldsymbol{\beta}$ (and σ^2 if relevant) here. If the above assumptions are satisfied together with appropriate regularity conditions, CML estimators are consistent as $N \to \infty$ for fixed n_j (e.g., Andersen, 1970, 1973a).

The conditional likelihood $\mathcal{L}^{\text{CML}}$ is almost invariably derived by treating the cluster-specific latent variable ζ_j as a fixed parameter, although Sartori and Severini (2004) show that the same $\mathcal{L}^{\text{CML}}$ results if ζ_j is treated as a random variable. ζ_j is usually interpreted as fixed when using CML estimation in psychometrics (e.g., Holland, 1990), whereas the "fixed effects framework" of econometrics interprets ζ_j as a random variable that can have arbitrary dependence with the covariates (e.g., Wooldridge, 2010, p. 286).

In contrast to MML estimation, CML estimation is based on solely within-cluster information. A great advantage of CML estimation is therefore that it is protective for the target parameters $\boldsymbol{\beta}$ if [A.1]-[A.4] are satisfied, regardless of the latent variable distribution and even if there is cluster-level endogeneity. It is usually not recognized that CML estimation also has a role to play under exogeneity because performance does not rely on $\mathbf{v}_j'\boldsymbol{\gamma}$ being the correct specification of the functional form for $\mathbf{v}_j$.

A cost of CML estimation is that it can be inefficient because it just exploits within-cluster information. Inefficiency is particularly acute when there is little within-cluster variation in the unit-specific covariates. Hence, CML estimation may have larger mean squared errors for estimating $\boldsymbol{\beta}$ than MML estimation, even under cluster-level endogeneity (e.g., Palta & Yao, 1991). Also, using CML estimation to remove cluster-specific slopes can lead to pronounced inefficiency.

CML estimation is primarily useful if the coefficients $\boldsymbol{\beta}$ of unit-specific covariates are the target parameters because the coefficients $\boldsymbol{\gamma}$ of cluster-specific covariates and the covariance parameters of random effects cannot be estimated. In our view, this may actually be beneficial because these parameters are inconsistently estimated by standard MML if there is cluster-level endogeneity.

Interactions between cluster-specific and unit-specific covariates become elements of $\boldsymbol{\beta}$ and can be estimated by CML for models with cluster-specific intercepts. For instance, the treatment-by-time interaction is often the target parameter in longitudinal data with time-invariant treatments. In models with cluster-specific random coefficients, the vectors $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ often include common variables and the corresponding elements of $\boldsymbol{\beta}$ are conditioned away by CML. Hence, the treatment-by-time interaction parameter cannot be estimated by CML in a model with cluster-specific slopes of time (e.g., Liang & Zeger, 2000).

In some situations covariate measurement error or misclassification problems can be more serious for CML than MML estimation (e.g., Griliches & Hausman, 1986; Frisell et al., 2012). However, CML estimation is immune to such problems for $\mathbf{v}_j$ and we will later show that cluster-specific covariate measurement error for $\mathbf{x}_{ij}$ is handled.

We now show the form taken by the conditional likelihood contribution $\mathcal{L}_j^{\text{CML}}$ for identity, logit and log links.

8.1. CML Estimation for Identity Link and Normal Distribution

Using that the sum of conditionally independent normally distributed random variables has a normal distribution, the conditional likelihood contribution for cluster j in model (5) becomes (e.g., Chamberlain, 1980)

$$\mathcal{L}_j^{\text{CML}} \;=\; n_j^{1/2}(\sqrt{2\pi}\sigma)^{-(n_j-1)}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_j}\left[(y_{ij}-\overline{y}_{\cdot j})-(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\cdot j})'\boldsymbol{\beta}\right]^2\right\}.$$

We see that the cluster-level component $\zeta_j+\mathbf{v}_j'\boldsymbol{\gamma}$ has cancelled out of $\mathcal{L}_j^{\text{CML}}$ and that a cluster does not contribute to estimation of $\boldsymbol{\beta}$ if the $\mathbf{x}_{ij}$ do not vary within the cluster.

For the identity link, assumption [A.2] is not required for consistent estimation, and [A.4] can be replaced by $\text{Cor}[(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\cdot j}),(\epsilon_{ij}-\overline{\epsilon}_{\cdot j})]=\mathbf{0}$. The effects of violating this zero-correlation assumption can in some cases be more severe for CML than MML estimation (e.g., Bound & Solon, 1999; Frisell et al., 2012). This can be addressed by using instrumental variables (IV) estimation if plausible instruments are available (e.g., Ebbes et al., 2004). Chamberlain (1984, 1985) proposes tests of [A.3] for the identity link (and the probit link) and Sjölander et al. (2016) derive the inconsistency produced by various violations of [A.2] and [A.3] for the identity link (and the logit link for some instances) in sibling designs.

For the identity link there are several alternative ways of implementing CML estimation of $\boldsymbol{\beta}$. We will briefly describe them below because they provide insight into basic features of CML estimation and its connection to other estimation methods in this particular case.

8.1.1. Cluster-Mean Centering and Maximum "Marginal" Likelihood Estimation in Statistics

Consider the cluster-mean centered or within-cluster model

$$y_{ij}-\overline{y}_{\cdot j} \;=\; (\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\cdot j})'\boldsymbol{\beta}+(\epsilon_{ij}-\overline{\epsilon}_{\cdot j}), \tag{9}$$

which is an example of a "working model" derived from the assumed data-generating mechanism. We see that ζ_j is swept out of the model and any misspecification related to ζ_j is therefore immaterial. However, $\mathbf{v}_j'\boldsymbol{\gamma}$ (and actually any cluster-specific function of $\mathbf{v}_j$) is also swept out which precludes estimation of $\boldsymbol{\gamma}$. On the other hand, MML estimation of $\boldsymbol{\gamma}$ is inconsistent if either $\mathbf{x}_{ij}$ or $\mathbf{v}_j$ are cluster-level endogenous as is estimation of both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ if the functional form $\mathbf{v}_j'\boldsymbol{\gamma}$ is incorrect.

The maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ in this model coincides with the CML estimator, which it not surprising given the resemblance of (9) with the argument of the exponential function in $\mathcal{L}_j^{\text{CML}}$. Furthermore, the CML estimator of $\boldsymbol{\beta}$ is the *within-cluster* ordinary least squares (OLS) estimator

$$\widehat{\boldsymbol{\beta}}^{\text{CML}} \;=\; W_{\mathbf{xx}}^{-1} W_{\mathbf{x}y},$$

which is the special case of (7) where $\omega = 0$ and between-cluster information is hence ignored.

Using standard terminology in statistics (e.g., Pawitan, 2001), the likelihood based on (9) can be viewed as a *marginal* likelihood that "transforms away" the nuisance parameters (here the incidental parameters ζ_j) by considering the implied model for the deviations from cluster means $y_{ij} - \overline{y}_{\cdot j}$. The canonical example of a marginal likelihood in the statistical sense is the restricted or residual likelihood of Patterson and Thompson (1971) for variance components in linear mixed models, where regression coefficients are nuisance parameters. Recall that this meaning of "marginal" is different from that in psychometrics where it refers to a likelihood where random ζ_j are "integrated out".

Goetgeluk and Vansteelandt (2008) use cluster-mean centering to consistently estimate $\boldsymbol{\beta}$ under cluster-level endogeneity by conditional generalized estimating equations (CGEE), an estimator that can also be used for log links.

8.1.2. Cluster-Specific Dummy Variables and JML Estimation Alternatively, we can use JML estimation for $\boldsymbol{\beta}$ in a working model that includes dummy or indicator variables with fixed cluster-specific coefficients ζ_j

$$y_{ij} \;=\; \mathbf{x}_{ij}'\boldsymbol{\beta} + \sum\nolimits_{r=1}^{N} d_{rj}\zeta_r + \epsilon_{ij}, \tag{10}$$

where the d_{rj} take the value 1 if $r = j$ and 0 otherwise. Note that $\mathbf{v}_j'\boldsymbol{\gamma}$ is omitted because $\mathbf{v}_j$ is collinear with the dummy variables d_{rj}. JML estimation of $\boldsymbol{\beta}$ can simply proceed by using OLS to estimate (10).

The use of a fixed parameter for each cluster explains why CML and related estimators are often referred to as *fixed-effects* estimators in econometrics. In that literature, MML and related estimators that assume cluster-level exogeneity are referred to as *random-effects* estimators, although the estimands are not the random effects.

By explicitly controlling for clusters in this way we are estimating pure within-cluster effects. The clusters are said to act as their own controls, and estimation is therefore immune to cluster-level endogeneity. As mentioned earlier, there is no incidental parameter problem in this case and $\widehat{\boldsymbol{\beta}}^{\text{JML}} = \widehat{\boldsymbol{\beta}}^{\text{CML}}$ is consistent as $N \to \infty$ for fixed n_j. Consistency of $\widehat{\zeta}_j^{\text{JML}}$ requires a double-asymptotic scheme where both $n_j \to \infty$ and $N \to \infty$.

8.1.3. Auxiliary Linear Projection and MML Estimation In the Mundlak-device (Mundlak, 1978) the cluster means $\overline{\mathbf{x}}_{\cdot j}$ of the unit-specific covariates $\mathbf{x}_{ij}$ are included in the model. This can be viewed as handling violation of [A.5] by considering an auxiliary linear projection

$$\zeta_j \;=\; \overline{\mathbf{x}}_{\cdot j}'\boldsymbol{\delta} + u_j, \tag{11}$$

where $\text{Cov}(\overline{\mathbf{x}}_{\cdot j}, u_j) = \mathbf{0}$ per construction. Substituting the linear projection in (5), we obtain

$$y_{ij} \;=\; \mathbf{x}_{ij}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{\cdot j}'\boldsymbol{\delta} + \mathbf{v}_j'\boldsymbol{\gamma} + u_j + \epsilon_{ij},$$

which can be expressed as the following model that is estimated in the "hybrid method" of Allison (2009):

$$y_{ij} \;=\; (\mathbf{x}'_{ij}-\overline{\mathbf{x}}'_{\cdot j})\boldsymbol{\beta} + \overline{\mathbf{x}}'_{\cdot j}(\boldsymbol{\beta}+\boldsymbol{\delta}) + \mathbf{v}'_j\boldsymbol{\gamma} + u_j + \epsilon_{ij}. \tag{12}$$

For the identity link, MML estimation (or ML/OLS estimation treating the composite error terms $u_j + \epsilon_{ij}$ as independent) of these working models produces the consistent CML estimator of $\boldsymbol{\beta}$, even if the linear projection does not coincide with the correct auxiliary statistical model or "data-generating mechanism". Contrary to common belief (e.g., Allison, 2009), the hybrid method is inconsistent for $\boldsymbol{\gamma}$ (and ψ) even if $\mathbf{v}_j$ is cluster-level exogenous because $\boldsymbol{\delta}$ absorbs some of the effects of the cluster-level covariates $\mathbf{v}_j$ (Castellano et al., 2014).

8.1.4. Including Deviations from Cluster Means and E-Estimation In the vector version of (12) for cluster j, the matrix of cluster mean deviations $\mathbf{X}_j - \mathbf{1}_{n_j} \otimes \overline{\mathbf{x}}'_{\cdot j}$ is orthogonal to $\mathbf{1}_{n_j} \otimes \mathbf{v}'_j$ and $\mathbf{1}_{n_j}\zeta_j$, so the CML estimator for $\boldsymbol{\beta}$ is obtained by estimating the simplified working model

$$y_{ij} \;=\; (\mathbf{x}'_{ij}-\overline{\mathbf{x}}'_{\cdot j})\boldsymbol{\beta} + \epsilon_{ij},$$

by ML/OLS.

This can be viewed as a variant of *E*-estimation of $\boldsymbol{\beta}$ (e.g., Robins et al., 1992). Here, specification of a correct model for the association between the outcome and possibly unknown cluster-level covariates in the "outcome model" is avoided by breaking the correlation between the unit-level covariates ("exposures" $\mathbf{x}_{ij}$) and cluster-level variables ζ_j and $\mathbf{v}_j$ in the "exposure model" through the inclusion of $\overline{\mathbf{x}}_{\cdot j}$ (see also Goetgeluk & Vansteelandt, 2008).

8.1.5. Using Deviations from Cluster Means as Instrumental Variables Because $\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\cdot j}$ is correlated with $\mathbf{x}_{ij}$ whereas $\mathbf{X}_j - \mathbf{1}_{n_j} \otimes \overline{\mathbf{x}}'_{\cdot j}$ is orthogonal to $\mathbf{1}_{n_j}\zeta_j$ by construction, $\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\cdot j}$ can serve as instrumental variable for $\mathbf{x}_{ij}$ in the outcome model (5). Instrumental variables (IV) estimators, such as two-stage least squares (2SLS), are then identical to the CML estimator for $\boldsymbol{\beta}$.

8.2. CML Estimation for Logit Link and Bernoulli Distribution

We now consider linear predictor (3) with a logit link and a Bernoulli conditional distribution. The conditional likelihood contribution for cluster j can be expressed as (e.g., Chamberlain, 1980)

$$\mathcal{L}_j^{\mathrm{CML}} \;=\; \frac{\prod_{i=1}^{n_j} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})^{y_{ij}}}{\sum_{\mathbf{d}\in\mathcal{B}_j} \prod_{i=1}^{n_j} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})^{d_i}}. \tag{13}$$

Here,

$$\mathcal{B}_j = \left\{\mathbf{d}=(d_1,\ldots,d_{n_j})' \,:\, d_i\in\{0,1\},\; \sum\nolimits_{i=1}^{n_j} d_i = \tau_j\right\} \tag{14}$$

is the set of all $\binom{n_j}{\tau_j}$ permutations of zeros and ones whose sum equals τ_j, the observed value of the sufficient statistic for ζ_j.

We note that $\zeta_j + \mathbf{v}'_j\boldsymbol{\gamma}$ has cancelled out of $\mathcal{L}_j^{\mathrm{CML}}$. Also, a cluster does not contribute to the conditional likelihood if its outcomes y_{ij} are all 0 or all 1, or $\mathbf{x}_{ij}$ does not vary in the cluster. CML estimation appears computationally demanding but is feasible even for large n_j by using

recursive algorithms (e.g., Howard, 1972; Gustafsson, 1980) or Markov chain Monte Carlo (e.g., Rice, 2004). For very large n_j, approximations can be based on composite conditional likelihoods (e.g., Liang, 1987) or random sampling of permutations in $\mathcal{B}_j$ (e.g., D'Haultfæuille & Iaria, 2016). For $n=2$, $\mathcal{L}_j^{\text{CML}}$ simplifies to the standard likelihood contribution of a logistic regression model with binary outcome equal to 1 if $(y_{1j}=0, y_{2j}=1)$ and equal to 0 if $(y_{1j}=1, y_{2j}=0)$, and with covariates $\mathbf{x}_{2j}-\mathbf{x}_{1j}$.

Several other models have likelihood contributions that take a similar form as (13):

(i) Case-control studies: $\mathcal{L}_j^{\text{CML}}$ corresponds to the conditional likelihood contribution for matched set j in conditional logistic regression for matched retrospective case-control designs, were the indicator y_{ij} takes the value 1 if unit i is one of a fixed number of τ_j cases and 0 if unit i is one of $n_j-\tau_j$ controls. $\mathcal{L}_j^{\text{CML}}$ then represents the conditional probability of the n_j observed covariate vectors in set j, given all potential allocations of the covariate vectors to cases and controls (e.g., Prentice & Breslow, 1978).

(ii) Survival analysis with ties: $\mathcal{L}_j^{\text{CML}}$ corresponds to the "discrete" or "exact" partial likelihood contribution for the jth ordered survival time in a Cox proportional-hazards model with tied survival times (Cox, 1972). At the jth survival time, the indicator y_{ij} takes the value 1 if unit i is one of the τ_j units in the risk set who experienced the event and $\mathcal{B}_j$ is defined as in (14) for the n_j units in the risk set.
When there are no ties, the standard partial likelihood contribution for an event occurring at the jth survival time can be expressed as

$$\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{d\in\mathcal{R}_j}\exp(\mathbf{x}_d'\boldsymbol{\beta})}, \tag{15}$$

the conditional probability that a particular unit i experiences the event at the jth survival time, given that exactly one unit in the risk set $\mathcal{R}_j$ experiences the event. The risk sets are not disjoint because a unit in $\mathcal{R}_j$ belongs to all risk sets for earlier events and the term "partial likelihood" is used (Cox, 1975).

(iii) Discrete choice: The conditional likelihood contribution for individual j in the conditional logit model for discrete choice takes the form of the standard partial likelihood contribution (15). Here, the likelihood contribution is the conditional probability that a particular alternative i is chosen by individual j, given that exactly one alternative is chosen from the individual-specific alternative set $\mathcal{R}_j$ (e.g., McFadden, 1974).

We also note that $\mathcal{L}_j^{\text{CML}}$ is identical to the conditional likelihood contribution produced by instead conditioning on the order statistic $y_{(1)j},\ldots,y_{(n_j)j}$ (e.g., Chen, 2007).

The basic idea of standard CML estimation is extended in *exact* conditional logistic regression (e.g., Cox, 1970; Mehta et al., 1995) with linear predictor (3). Here, each element in $\boldsymbol{\beta}$ is estimated by conditioning on sufficient statistics for not just ζ_j (as in standard CML estimation) but also for the remaining elements of $\boldsymbol{\beta}$. For small or unbalanced datasets this approach can mitigate separation problems where outcomes are perfectly predicted and standard conditional likelihoods therefore do not exist (e.g., Albert & Anderson, 1984). Moreover, inferences for $\boldsymbol{\beta}$ are based on permutation distributions of the sufficient statistics that do not rely on asymptotics.

Nonparametric marginal maximum likelihood (NPMML) estimation (e.g., de Leuuw & Verhelst, 1986) leaves the latent variable distribution $p(\zeta_j)$ unspecified. NPMML estimation can be implemented by treating the latent variable as discrete and choosing the number of mass points to yield the highest likelihood. For concordant Rasch models that fit the observed sumscore distribution exactly, Lindsay et al. (1991) show that NPMML and CML estimation produce identical estimates of the item parameters $\boldsymbol{\beta}$. Rice (2004) provides conditions ensuring that marginal and

conditional likelihoods are equal. We note that standard NPMML estimation does not address the cluster-level endogeneity problem.

For a GLMM with linear predictor (1), the sufficient statistic for the latent variable vector $\boldsymbol{\zeta}_j$ is $\sum_{i=1}^{n_j} \mathbf{z}_{ij} y_{ij}$. For the logit link and the Bernoulli distribution, the conditional likelihood contribution then takes the same form as (13), with the difference that $\sum_{i=1}^{n_j} \mathbf{z}_{ij} d_i = \sum_{i=1}^{n_j} \mathbf{z}_{ij} y_{ij}$ now replaces $\sum_{i=1}^{n_j} d_i = \sum_{i=1}^{n_j} y_{ij}$ in the definition of the permutation set. In a panel data setting, Thomas (2006) considered the special case of a logit model with a cluster-specific intercept and a cluster-specific slope of time.

8.3. CML Estimation for Log Link and Poisson Distribution

Consider linear predictor (3) with a log link and a Poisson conditional distribution. Using that the sum of conditionally independent Poisson random variables has a Poisson distribution, the conditional likelihood contribution for cluster j becomes (e.g., Hausman et al., 1984)

$$\mathcal{L}_j^{\text{CML}} = \frac{(\sum_{i=1}^{n_j} y_{ij})!}{\prod_{i=1}^{n_j} y_{ij}!} \prod_{i=1}^{n_j} \left(\frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\sum_{\ell=1}^{n_j} \exp(\mathbf{x}'_{\ell j}\boldsymbol{\beta})} \right)^{y_{ij}}.$$

Again $\zeta_j + \mathbf{v}'_j\boldsymbol{\gamma}$ has cancelled out of $\mathcal{L}_j^{\text{CML}}$. We see that the product in $\mathcal{L}_j^{\text{CML}}$ that contains $\boldsymbol{\beta}$ is identical to the likelihood contribution for a unit j in a standard multinomial logit model with n_j alternatives, except that it is not required that $y_{ij} \in \{0, 1\}$ or $\sum_{i=1}^{n_j} y_{ij} = 1$ here.

Recall that there is no incidental parameter problem for the model with log link and a Poisson conditional distribution, and CML estimation and JML estimation with dummy variables for clusters produce identical estimates of $\boldsymbol{\beta}$. For the log link, assumption [A.2] is not required for consistent estimation and [A.4] can be relaxed by assuming that μ_{ij} is correctly specified (e.g., Wooldridge, 1999).

Thomas (2006) derived the CML estimator for a Poisson regression model with a cluster-specific intercept and a cluster-specific slope of time, and pointed out that there is no incidental parameter problem in this case either.

8.4. CML Estimation Beyond GLM Link Functions

It is worth noting that CML estimation can be used not just for GLMMs with continuous, binary, and count outcomes that we have investigated so far but also for other combinations of outcomes and latent variable models.

(a) Binary $y_{ij} \in \{0, 1\}$: Stratified linear odds-ratio models with cluster-specific parameters ζ_j (e.g., Storer et al., 1983)

$$p(y_{ij} = 1|\mathbf{x}_{ij}, \mathbf{v}_j; \boldsymbol{\vartheta}, \zeta_j) = \frac{\exp\{\zeta_j\}(1 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{v}'_j\boldsymbol{\gamma})}{1 + \exp\{\zeta_j\}(1 + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{v}'_j\boldsymbol{\gamma})}.$$

(b) Ordinal $y_{ij} \in \{0, \ldots, K\}$: Adjacent category logit models with cluster-specific parameters ζ_j (e.g., Heinen, 1996, p.124)

$$p(y_{ij} = k|\mathbf{x}_{ij}, \mathbf{v}_j; \boldsymbol{\vartheta}, \zeta_j) = \frac{\exp\{\alpha_{ki} + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{v}'_j\boldsymbol{\gamma} + k\zeta_j\}}{\sum_{c=0}^{K} \exp\{\alpha_{ci} + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{v}'_j\boldsymbol{\gamma} + c\zeta_j\}},$$

where we explicitly let α_{ki} denote unit and category-specific intercepts (no intercepts in $\boldsymbol{\beta}$ here), with $\alpha_{0i}=0$. Seminal special cases in psychometrics include the partial credit model (Masters, 1982) that has no covariates, and the rating scale model (Andrich, 1978) where additionally the item parameters are decomposed as $\alpha_{ki} = \alpha_i + \kappa_k$. For the exchangeable case, α_{ki} is replaced by α_k with $\alpha_0=0$.
Approximate CML estimation can be obtained via data expansion for cumulative logit models (Mukherjee et al., 2008), of which the logistic graded response model of Samejima (1969) is a special case, and for continuation-ratio logit models, as shown for sequential item response models by Tutz (1990).
Kelderman and Rijkes (1994) discuss CML estimation for a range of Rasch-type item response models for ordinal responses.

(c) Nominal $y_{ij} \in \{0, \ldots, K\}$: Multinomial logit models with cluster and category specific parameters ζ_{kj} (e.g., Chamberlain, 1980; Conaway, 1989; Lee, 2002)

$$p(y_{ij}=k|\mathbf{x}_{0ij},\ldots,\mathbf{x}_{Kij},\mathbf{x}_{ij},\mathbf{v}_j;\boldsymbol{\vartheta},\zeta_{0j},\ldots,\zeta_{Kj}) = \frac{\exp\{\alpha_{ki}+\mathbf{x}'_{kij}\boldsymbol{\beta}+\mathbf{x}'_{ij}\boldsymbol{\beta}_k+\mathbf{v}'_j\boldsymbol{\gamma}_k+\zeta_{kj}\}}{\sum_{c=0}^{K}\exp\{\alpha_{ci}+\mathbf{x}'_{cij}\boldsymbol{\beta}+\mathbf{x}'_{ij}\boldsymbol{\beta}_c+\mathbf{v}'_j\boldsymbol{\gamma}_c+\zeta_{cj}\}},$$

where α_{ki} are item and category-specific intercepts (no intercepts in $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_k$ here), with $\alpha_{0i}=0$ and $\alpha_{k1}=0$, and we let $\boldsymbol{\beta}_0=\mathbf{0}$ and $\boldsymbol{\gamma}_0=\mathbf{0}$. Special cases in psychometrics (e.g., Rasch, 1961; Andersen, 1973b) do not include covariates. For the exchangeable case, α_{ki} is replaced by α_k with $\alpha_0=0$.

(d) Survival times t: Stratified Cox-regression with cluster-specific baseline hazard function $h_j^0(t)$ (e.g., Chamberlain, 1985; Lancaster, 1990; Ridder & Tunali, 1999)

$$h_{ij}(t) = h_j^0(t)\exp\{\mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{v}'_j\boldsymbol{\gamma}\},$$

where $h_{ij}(t)$ is the continuous time hazard function for unit i in cluster j. Estimating the models by maximum partial likelihood yields fixed-effects estimators in the spirit of CML estimation.

9. MML Becomes CML for Large Clusters

Increasing the sample size usually does not improve estimation of misspecified models. However, standard MML estimation that ignores cluster-level endogeneity approaches CML estimation, and hence becomes more robust against cluster-level endogeneity, as the cluster sizes increase.

We first consider linear predictor (3) with an identity link. For known variance components, the generalized least squares (GLS) estimator is the MML estimator. Maddala (1971) showed that $\lim_{n\to\infty}\widehat{\boldsymbol{\beta}}^{\mathrm{GLS}} = \widehat{\boldsymbol{\beta}}^{\mathrm{CML}}$ for fixed N, see also (6) where $\omega \equiv \frac{\sigma^2}{\sigma^2+n\psi} \to 0$ when $n \to \infty$. Hence, GLS approaches CML estimation of $\boldsymbol{\beta}$ as the cluster sizes increase, making GLS estimation robust against cluster-level endogeneity without any ameliorating model extensions. Moreover, $\boldsymbol{\gamma}$ can in this case be consistently estimated if $\mathbf{v}_j$ is exogenous. It is clear from the IGLS formulae in Breusch (1987) that large-cluster robustness also applies to MML estimation.

Does the robustness extend beyond identity links? To shed some light on this, we performed a simulation study with exchangeable units to investigate the behaviour of MML estimation under cluster-level endogeneity for binary outcomes with a logit link and a normal latent variable distribution. For $N=1{,}000{,}000$ clusters, we gradually increased the cluster sizes n from 2 to 1,000. We

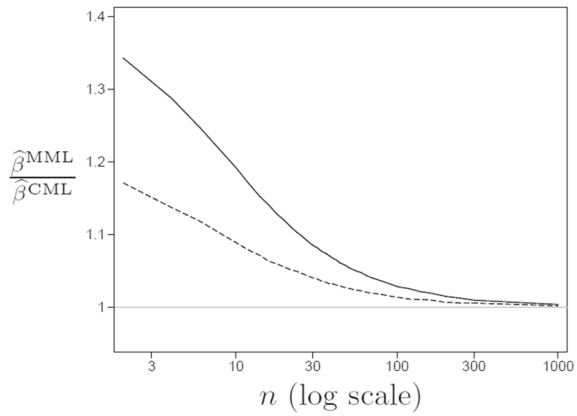

FIGURE 3.
Automatic inconsistency correction of MML estimation for logistic random-intercept model as a function of cluster size n. $\mathrm{Cor}(\zeta_j, x_{ij}) = .4$ (solid curve) and $\mathrm{Cor}(\zeta_j, x_{ij}) = .2$ (dashed curve) for $N = 1{,}000{,}000$ clusters.

simulated multivariate normal $(x_{1j}, x_{2j}, x_{3j}, x_{4j}, \zeta_j)$ with $\mathrm{Var}(x_{ij}) = 1$ and $\mathrm{Cor}(x_{ij}, x_{i'j}) = 0.2$ and parameter values $\gamma = 0$, $\beta = 1$, $\psi = 1$.

In Fig. 3 we show the similarity of the MML estimator and the consistent (as $N \to \infty$) CML estimator by plotting $\frac{\widehat{\beta}^{\mathrm{MML}}}{\widehat{\beta}^{\mathrm{CML}}}$ against n for $\mathrm{Cor}(\zeta_j, x_{ij}) = 0.4$ (solid curve) and $\mathrm{Cor}(\zeta_j, x_{ij}) = 0.2$ (dashed curve). We see that the large-cluster robustness of MML estimation extends to the logit link and that larger cluster sizes are required to approach consistency when the cluster-level endogeneity is more severe.

10. Extending the Scope of CML Estimation

For non-exchangeable data it is often plausible that the coefficients of the covariates $\mathbf{x}_{ij}$ and $\mathbf{v}_j$ are item-specific. Considering small to moderate cluster sizes, we now propose a useful extension of the model class for which CML estimation is applicable. Specifically, we generalize the GLMM in (1) by replacing $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by item-specific coefficients $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$

$$\nu_{ij} \;=\; \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \mathbf{v}'_j\boldsymbol{\gamma}_i + \mathbf{z}_{ij}\boldsymbol{\zeta}_j. \tag{16}$$

Letting $n^{\max}$ denote the maximum cluster size, the linear predictor (16) can be re-expressed as

$$\nu_{ij} \;=\; \sum_{r=1}^{n^{\max}} (d_{ri}\mathbf{x}'_{rj})\boldsymbol{\beta}_r + \sum_{r=1}^{n^{\max}} (d_{ri}\mathbf{v}'_j)\boldsymbol{\gamma}_r + \mathbf{z}_{ij}\boldsymbol{\zeta}_j,$$

where $\mathbf{x}_{rj}$ and $\mathbf{v}_j$ are now defined as in the exchangeable case. We see that this model includes solely unit-specific covariates $d_{ri}\mathbf{x}_{rj}$ and $d_{ri}\mathbf{v}_j$ and that both variable types have item-invariant coefficients. This is exactly the situation for GLMMs where CML estimation is traditionally employed, and CML estimation can therefore also be used for model (16) in a straightforward manner. Consistent estimation results for the regression coefficients in $\boldsymbol{\beta}_i$ and the differences $\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_{i'}$ for $i' \neq i$.

The validity of the traditional and rather restrictive model with invariant coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can now be investigated by contrasting it with the more general model (16), for instance by using *conditional* likelihood-ratio tests (Andersen, 1971) and fit measures based on conditional likelihoods. We refer to Maris (1998) for an insightful discussion of confidence intervals and hypothesis testing based on CML estimation.

CML estimation can also be used for models with crossed latent variables, such as panel models with both individual and year effects. For the identity link, Balazsi et al. (2017) review fixed-effects approaches, and for the logit link, Charbonneau (2017) and Kertesz (2017) use CML estimation by repeatedly conditioning on sufficient statistics to eliminate crossed latent variables one at a time. Generalized additive mixed models (GAMMs) with cluster-specific intercepts were estimated using CML by Zhang and Davidian (2004).

For models with several levels of nested latent variables, CML estimation is simply implemented by conditioning on the sufficient statistics for the latent variables at the lowest level.

11. Mimicking CML by MML Estimation of Augmented Models

How can we proceed if CML estimation is not possible for the model of interest? A subset of the parameters can in some instances be treated as known to produce a model that lends itself to CML estimation. Verhelst and Glas (1995) proposed the one parameter logistic model (OPLM) where discrimination parameters are taken to be "fixed constants supplied by hypothesis", making CML estimation feasible for $\boldsymbol{\beta}$. However, we usually prefer approaches that can provide protective or mitigating estimation in general settings without treating parameters as known.

A model closely resembling that of interest can sometimes be found. The most obvious example is use of the logit link instead of the very similar probit link which cannot be used in CML estimation. Another example is dynamic (or autoregressive) logit models with cluster-specific intercepts for binary binary outcomes. In this case Bartolucci and Nigro (2010) used a quadratic exponential model (Cox & Wermuth, 1994) to address the limitations of the CML estimator of Honoré and Kyriazidou (2000).

In this section we consider a general approach where we mimick CML estimation by MML estimation of augmented models that can handle cluster-level endogeneity. The first variant uses an auxiliary model that specifies how the latent variable depends on the endogenous covariates and the second variant uses a joint model where the endogenous covariates are treated as outcomes. Both variants can accommodate multidimensional latent variables and, in contrast to CML, models with factor loadings or discrimination parameters and non-canonical link functions. For notational simplicity, we henceforth omit reference to parameters in all distributions.

11.1. Auxiliary Modeling of ζ_j Given $\mathbf{w}_j$

Cluster-level endogeneity can be addressed by using an auxiliary statistical model for $p(\zeta_j|\mathbf{w}_j)$. We describe two alternative methods: using a GLLAMM and using a reduced-form GLMM.

11.1.1. Using GLLAMM A GLLAMM (e.g., Rabe-Hesketh et al., 2004; Skrondal & Rabe-Hesketh, 2004) is composed of a response model for the outcomes given the covariates and latent variables and a structural model for the latent variables given the covariates. Conditional on latent variables, the response model is an extended GLM that accommodates more outcome types and different outcome types for different units. The linear predictor generalizes that of the GLMM in (1) by, for instance, allowing factor loadings or discrimination parameters for the

latent variables. Conditional on observed covariates, the structural model is a multilevel structural equation model for the latent variables with normally distributed disturbance terms.

In the present setting, we can use a simple special case of a GLLAMM where the linear predictor of the response model is specified as (3) and the structural model is specified as the chosen auxiliary model. A Mundlak-inspired auxiliary model is

$$\zeta_j \;=\; \overline{\mathbf{x}}'_{\cdot j}\boldsymbol{\delta} + u_j, \tag{17}$$

and for non-exchangeable units we can use the more flexible Chamberlain auxiliary model (Chamberlain, 1980, 1984)

$$\zeta_j \;=\; \sum\nolimits_{r=1}^{n_j} \mathbf{x}'_{rj}\boldsymbol{\delta}_r + u_j. \tag{18}$$

In (17) it is assumed that $\mathrm{E}(\zeta_j|\mathbf{w}_j) = \overline{\mathbf{x}}'_{\cdot j}\boldsymbol{\delta}$, and u_j is homoskedastic, normal and independent of $\overline{\mathbf{x}}_{\cdot j}$, and in (18) we assume that $\mathrm{E}(\zeta_j|\mathbf{w}_j) = \sum_{r=1}^{n_j} \mathbf{x}'_{rj}\boldsymbol{\delta}_r$, and that u_j is homoskedastic, normal and independent of the $\mathbf{x}_{ij}$ (Chamberlain, 1984).

Estimating the GLLAMM by MML provides consistent estimators for all parameters if the auxiliary and outcome models are correctly specified. MML estimation is mitigating for $\boldsymbol{\beta}$ if the auxiliary model is a reasonable approximation of the correct model. Note that the assumed auxiliary models can be viewed as linear projections for the identity link (see Sect. 8.1.3).

Invoking the GLLAMM framework makes it straightforward to consider useful model extensions. For non-exchangeable units, it may be plausible that the observed covariates have item-specific coefficients $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$, and that the latent variable has item-specific factor loadings or discrimination parameters λ_i. We can then use the following linear predictor for the response model:

$$\nu_{ij} \;=\; \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \mathbf{v}'_j\boldsymbol{\gamma}_i + \lambda_i\zeta_j. \tag{19}$$

The model can, for instance, be extended to include multidimensional latent variables $\boldsymbol{\zeta}_j$ and ultimately extended to the full GLLAMM response model (see e.g., Rabe-Hesketh et al., 2004).

11.1.2. Using Reduced-Form GLMM Alternatively, the auxiliary model can be substituted into (3) to yield a reduced-form working GLMM. Substituting model (17), we obtain

$$g\{\mathrm{E}(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, u_j)\} \;=\; \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{v}'_j\boldsymbol{\gamma} + \overline{\mathbf{x}}'_{\cdot j}\boldsymbol{\delta} + u_j,$$

which can be rearranged to get

$$g\{\mathrm{E}(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, u_j)\} \;=\; (\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\cdot j})'\boldsymbol{\beta} + \overline{\mathbf{x}}'_{\cdot j}(\boldsymbol{\beta}+\boldsymbol{\delta}) + \mathbf{v}'_j\boldsymbol{\gamma} + u_j. \tag{20}$$

In general, MML estimation of these models produces mitigating estimation for $\boldsymbol{\beta}$, with the exception of identity links where we have pointed out that the consistent CML estimator is obtained.

Neuhaus and McCulloch (2006) considered (20) without cluster-level covariates $\mathbf{v}_j$ and called it a "between-within model". They referred to MML estimation of (20) as the "poor-man's approximation to the conditional likelihood approach" because it is straightforward to implement in

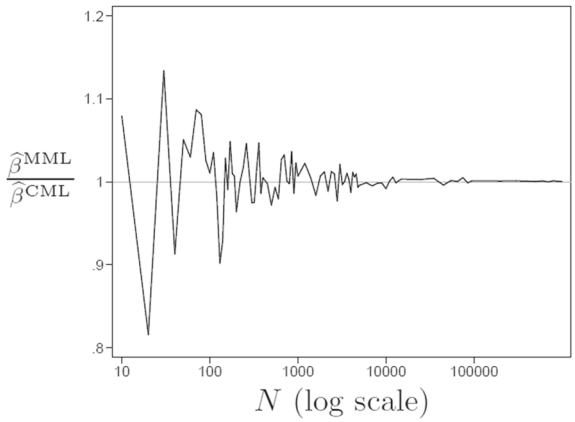

FIGURE 4.
Protective MML estimate for simulated data with correct auxiliary model for logit link as a function of N. $\mathrm{Cor}(x_{ij}, \zeta_j) = 0.4$ and $n = 4$.

practice. Brumback et al. (2010) extended the poor-man's approximation by considering nonlinear functions of the cluster-means. Note that, although $\mathbf{X}_j - \mathbf{1}_{n_j} \otimes \overline{\mathbf{x}}'_{\cdot j}$ is orthogonal to $\mathbf{1}_{n_j} \otimes \mathbf{v}'_j$, omitting $\mathbf{v}_j$ if $\boldsymbol{\gamma} \neq \mathbf{0}$ is likely to produce some additional inconsistency for models with logit links because odds ratios are not collapsible (e.g., Gail et al., 1984).

For non-exchangeable units, we can substitute model (18) in (19) with $\lambda_i = 1$ to obtain

$$g\{\mathrm{E}(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, u_j)\} = \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \mathbf{v}'_j\boldsymbol{\gamma}_i + \sum\nolimits_{r=1}^{n_j} \mathbf{x}'_{rj}\boldsymbol{\delta}_r + u_j.$$

In contrast to the GLLAMM approach, factor loadings or discrimination parameters are not accommodated.

We conducted a Monte Carlo experiment to study the performance of MML estimation for a random-intercept binary logit model with a correctly specified auxiliary model for exchangeable data. To investigate consistency as $N \to \infty$ for $n = 4$, we simulated multivariate normal $(x_{1j}, x_{2j}, x_{3j}, x_{4j}, \zeta_j)$ with $\mathrm{Var}(x_{ij}) = 1$, $\mathrm{Cor}(x_{ij}, x_{i'j}) = 0.2$ and $\mathrm{Cor}(x_{ij}, \zeta_j) = 0.4$ for all i and parameter values $\gamma = 0$, $\beta = 1$, and $\psi = 1$. The cluster mean $\overline{x}_{\cdot j}$ was used in the auxiliary model.

Figure 4 plots the ratio of estimates $\frac{\widehat{\beta}^{\mathrm{MML}}}{\widehat{\beta}^{\mathrm{CML}}}$ against N. Because this ratio seems to converge to 1 as $N \to \infty$ for fixed n and we know that $\widehat{\boldsymbol{\beta}}^{\mathrm{CML}}$ is consistent, we conclude that $\widehat{\boldsymbol{\beta}}^{\mathrm{MML}}$ also appears to be consistent. MML estimation using a correct auxiliary model is therefore protective in this case.

11.2. Joint Modeling of $\mathbf{y}_j$ *and* $\mathbf{w}_j$

We can also handle cluster-level endogeneity by specifying a joint statistical model $p(\mathbf{y}_j, \mathbf{w}_j)$ for the outcomes $\mathbf{y}_j$ and covariates $\mathbf{w}_j$. For continuous outcomes we discuss joint modeling via conventional structural equation modeling (SEM) and for other outcome types we briefly describe joint modeling using GLLAMMs.

11.2.1. Using Conventional SEM In conventional SEM with identity links and normal conditional distributions, analytic integration over the latent variables is straightforward. In this case

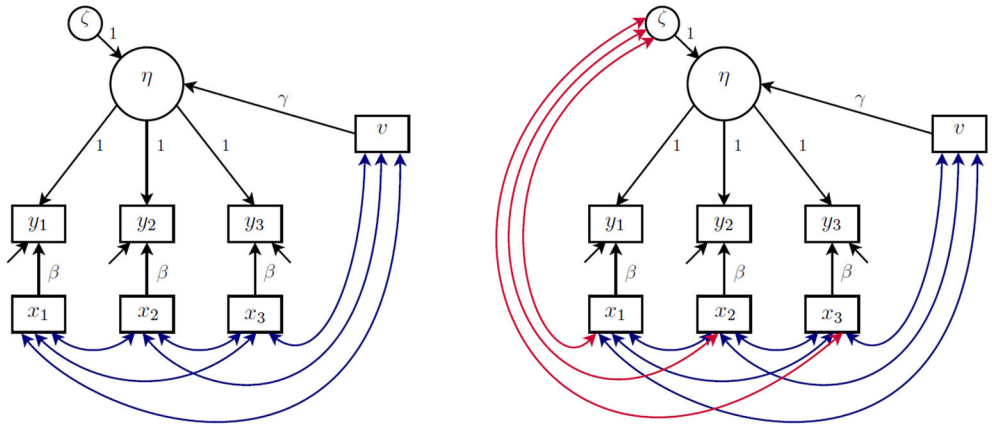

FIGURE 5.
Joint modeling using SEM for identity link and normal conditional distribution. Path diagrams ($n = 3$) for standard random-intercept model where $\text{Cor}(x_{ij}, \zeta_j) = 0$ and $\text{Cor}(v_j, \zeta_j) = 0$ (left panel) and joint SEM specifying $\text{Cor}(x_{ij}, \zeta_j) \neq 0$ and $\text{Cor}(v_j, \zeta_j) = 0$ (right panel).

joint models are usually expressed as

$$p(\mathbf{y}_j, \mathbf{w}_j) = \int_{\zeta_j} p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j) p(\mathbf{w}_j, \zeta_j)\, \mathrm{d}\zeta_j,$$

which requires specification of a model for $p(\mathbf{w}_j, \zeta_j)$. Specifically, a SEM that includes covariances between $\mathbf{x}_{ij}$ and ζ_j is specified and estimated by MML (e.g., Teachman et al., 2001; Bollen & Brand, 2010). In this case consistency requires a correctly specified covariance structure, but does not rely on normality (e.g., Browne, 1974), and the approach can be viewed as an instance of pseudo maximum (marginal) likelihood estimation (e.g., Arminger & Schoenberg, 1989).

Figure 5 shows path diagrams for a SEM representation of a standard random-intercept model with exogenous unit-specific covariate x_{ij} and exogenous cluster-specific covariate v_j (left panel) and a joint SEM for the same model but allowing the random intercept to be correlated with x_{ij} to accommodate cluster-level endogeneity (right panel).

MML estimation of the joint model in the right panel produces CML estimates of $\boldsymbol{\beta}$. In contrast to CML estimation, the joint MML approach is also consistent for $\boldsymbol{\gamma}$ when $\mathbf{v}_j$ is cluster-level exogenous (as in the figure). Because all the bells and whistles of SEMs are available, it is straightforward to include, for instance, factor loadings in the models. Note that appropriate parameter restrictions should be imposed for exchangeable data (Sim, 2019).

11.2.2. Using GLLAMM The SEM approach outlined above is not feasible beyond the identity link, and joint models in biometrics and statistics are typically formulated as (e.g., Neuhaus & McCulloch, 2006)

$$p(\mathbf{y}_j, \mathbf{w}_j) = \int_{\zeta_j} p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j) p(\mathbf{w}_j|\zeta_j) p(\zeta_j)\, \mathrm{d}\zeta_j,$$

which requires specification of a model for $p(\mathbf{w}_j|\zeta_j)$. In this case a GLLAMM response model such as (19) with different outcome types for different units is specified for both the outcome model $p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j)$ and the covariate model $p(\mathbf{w}_j|\zeta_j)$.

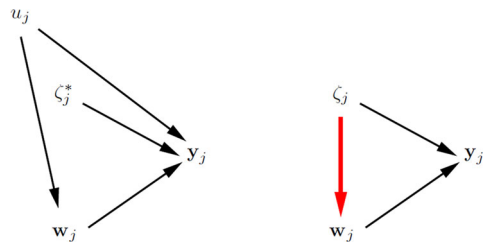

FIGURE 6.
Unobserved cluster-level confounding. Cluster-level unobserved confounder u_j (left panel) and resulting cluster-level endogeneity (right panel).

MML estimation for joint models is consistent for all model parameters if the model for $p(\mathbf{w}_j|\zeta_j)$ is correctly specified, in addition to a correct outcome model for $p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j)$ and a correct $p(\zeta_j)$.

11.3. Auxiliary or Joint Modeling?

Both auxiliary and joint modeling are useful for mimicking protective CML estimation of the target parameters $\boldsymbol{\beta}$ when there is cluster-level endogeneity. However, it seems unlikely that auxiliary models represent plausible data-generating mechanisms (e.g., Goetgeluk & Vansteelandt, 2008), whereas joint models may do so, for instance when there is unobserved cluster-level confounding (see Sect. 12.1). In the unlikely event that the entire joint model is correctly specified, MML estimation will be consistent for *all* model parameters. Neuhaus and McCulloch (2006) discuss conditions for consistent estimation of $\boldsymbol{\beta}$ using auxiliary modeling when the data-generating mechanism is a joint model. Auxiliary modeling can be implemented in standard GLMM software, and very flexible nonlinear parametric models can be used for $p(\zeta_j|\mathbf{w}_j)$, whereas modeling of $p(\mathbf{w}_j|\zeta_j)$ requires specification of an appropriate link function and conditional distribution for each covariate. Joint modeling in effect assumes a particular dependence structure for the covariates but accommodates covariates missing at random. We advocate performing sensitivity analysis by using both auxiliary and joint modeling.

Finally, it should be kept in mind that the choice between CML estimation and MML estimation of augmented models may in practice involve a trade-off between inconsistencies due to CML estimation of overly simple models (e.g., without discrimination parameters) and misspecified endogeneity models.

12. Reasons for Cluster-Level Endogeneity

Cluster-level endogeneity can arise for a variety of reasons, including unobserved cluster-level confounding, covariate measurement error, retrospective sampling, informative cluster sizes, missing data, and heteroskedasticity.

12.1. Unobserved Cluster-Level Confounding of Causal Effects

Recall that consistent MML estimation in general requires cluster-level exogeneity as shown in the left panel of Fig. 2. Consider now the case where the data-generating mechanism contains an unobserved cluster-level confounder u_j as in the left panel of Fig. 6 (where the cluster-level error term is now denoted ζ_j^*).

In a statistical model with linear predictor such as (3), the unobserved cluster-level confounder u_j becomes absorbed by the cluster-level error term $\zeta_j = \zeta_j^* + u_j$ as displayed in the right panel of Fig. 6. It is evident that unobserved cluster-level confounding leads to cluster-level endogeneity.

Use of the term confounding presupposes that regression coefficients represent causal effects that can be confounded. Lancaster (2000, p. 296) points out that econometricians emphasize that some or all covariates may be "chosen" by an individual j in light of his knowledge of ζ_j (e.g., attending a training program, $x_{ij} = 1$ rather than $x_{ij} = 0$, because ability ζ_j is low). Hence, economic theory provides a presumption that ζ_j and $\mathbf{x}_{ij}$ are dependent in the population. Lancaster concludes that "This point plays absolutely no role in the statistics literature", where ζ_j is invariably, and usually implicitly, either assumed to be independent or uncorrelated with random covariates or assumed to not depend on the values taken by fixed covariates. Here, regression coefficients merely represent associations between included variables, or linear projections in the case of linear models, in which case the error terms are orthogonal to the covariates by construction. Spanos (2006) contrasts the conventional meaning of models in econometrics and statistics.

Importantly, CML estimation can be consistent for causal effects even when there is unobserved cluster-level confounding. Under assumptions [A.1]-[A.4], $\widehat{\beta}^{\text{CML}}$ for a treatment x_{ij} in (3) can be interpreted as estimating a causal effect that is homogeneous in the population. However, causal effects are usually viewed as heterogeneous and the estimand taken to be some average causal effect (ACE) in the modern literature on causal inference.

For the identity link, $\widehat{\beta}^{\text{CML}}$ represents an estimated ACE for the subpopulation of clusters where the treatment varies between the units (e.g., Imai & Kim, 2019; Petersen & Lange, 2020). Sobel (2012) and Wooldridge (2010: sect. 21.6.4) explore causal effects that can be estimated by fixed-effects methods for different treatment regimes and state assumptions required for identification. For the logit link, there is no simple interpretation of $\widehat{\beta}^{\text{CML}}$ when the causal odds ratio is heterogeneous (Sjölander et al., 2012; Petersen & Lange, 2020).

12.2. Cluster-Specific Measurement Error

Sometimes variables are fallibly measured with cluster-specific measurement errors (e.g., Wang et al., 2012). Examples include teacher-specific bias in ratings of students and laboratory tests analyzed in batches.

12.2.1. Covariate Measurement Error We now consider the following version of linear predictor (3):

$$\nu_{ij} \;=\; \beta x_{ij} + \mathbf{v}_j'\boldsymbol{\gamma} + \zeta_j, \tag{21}$$

where the unit-specific covariate x_{ij} is continuous.

If x_{ij} were observed and cluster-level exogenous, consistent estimation of all model parameters could proceed by MML estimation. The new feature is that x_{ij} is latent and fallibly measured by a continuous variable m_{ij} with additive cluster-specific covariate measurement error δ_j

$$m_{ij} \;=\; x_{ij} + \delta_j.$$

Rearranging this classical covariate measurement model as $x_{ij} = m_{ij} - \delta_j$, we substitute it in (21) to obtain a working model with linear predictor

$$\nu_{ij} \;=\; \beta m_{ij} + \mathbf{v}_j'\boldsymbol{\gamma} + \zeta_j^*,$$

where $\zeta_j^* \equiv \zeta_j - \beta\delta_j$. Having replaced the latent covariate x_{ij} by the fallibly observed covariate m_{ij}, we see that ζ_j has been replaced by a composite cluster-specific error term. Unfortunately, even if x_{ij} is cluster-level exogenous, the fallibly observed covariate m_{ij} becomes cluster-level endogenous. This is because the component $\beta\delta_j$ of ζ_j^* is not independent of m_{ij}.

MML estimation would be inconsistent for β in the working model. Joint modeling of the outcomes y_{ij} and measures m_{ij} would enable consistent MML estimation of all model parameters, but only if the entire model is correctly specified. In contrast, CML estimation is protective for β. In this case parametric assumptions are not required for the distributions of ζ_j or δ_j, and these terms could even be dependent, producing a form of differential measurement error. Moreover, x_{ij} could be cluster-level endogenous. CML estimation remains protective if several continuous unit-specific covariates are measured with covariate- and cluster-specific errors.

12.2.2. Latent-Response Measurement Error Consider now the class of GLMMs that can be expressed as latent response models

$$y_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{v}_j'\boldsymbol{\gamma} + \zeta_j + \epsilon_{ij}. \tag{22}$$

For instance, models with logit links for binary outcomes y_{ij} arise if ϵ_{ij} has a standard logistic density and the observed outcome is produced by thresholding the latent response y_{ij}^* as $y_{ij} = \mathrm{I}(y_{ij}^* > 0)$.

If y_{ij}^* is contaminated by cluster-specific additive error δ_j, we obtain $y_{ij}^\bullet = y_{ij}^* + \delta_j$ yielding observed outcomes $y_{ij} = \mathrm{I}(y_{ij}^\bullet > 0)$. Substituting the latent response model (22) we see that ζ_j is replaced by a composite cluster-specific intercept $\zeta_j^\bullet = \zeta_j + \delta_j$.

Again, CML estimation is protective for β and parametric assumptions are not required for ζ_j and δ_j. Moreover, differential measurement error, in the sense that δ_j depends on $\mathbf{x}_{ij}$, $\mathbf{v}_j$ and ζ_j, is accommodated.

12.3. Retrospective Sampling

We will discuss two kinds of retrospective sampling schemes that produce cluster-level endogeneity, the first by sampling units and the second by sampling clusters.

12.3.1. Case-Control and Choice-Based Sampling of Units Case-control sampling is very useful for rare binary outcomes $\mathbf{y}_j$ when obtaining one or more of the covariates is expensive or invasive (e.g., Breslow, 1996). Examples include drawing blood samples from individuals and conducting comprehensive psychiatric interviews with patients.

The basic idea of case-control designs is to under-sample units i with outcome $y_{ij} = 0$. This is an example of retrospective sampling because the probability of being sampled depends on the value taken by an outcome variable. Letting S_{ij} be an indicator variable for sampling unit i in cluster j, the probability of selecting the unit is then dependent on whether the unit is a case ($y_{ij} = 1$) or control ($y_{ij} = 0$):

$$p(S_{ij}=1|y_{ij}) \equiv \pi(y_{ij}).$$

We assemble the selection indicators for the units in cluster j in the selection vector $\mathbf{s}_j$. In a cumulative case-control study (e.g., Rothman et al., 2008, p. 125), the researcher samples all cases, $\pi(1) = 1$, whereas $\pi(0)$ is small in order to under-sample controls.

In choice-based sampling individuals are sampled retrospectively by stratifying on their individual choices (e.g., Manski, 1981), in which case the marginal distribution of the choices in

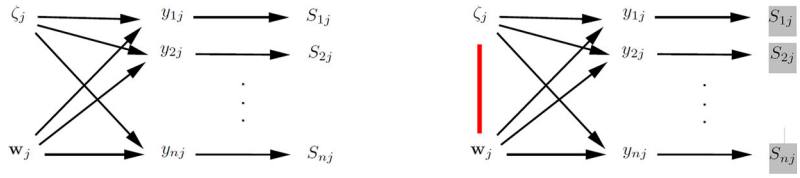

FIGURE 7.
Retrospective sampling of units. Unselected population (left panel) and selected sample (right panel).

the selected sample typically differs from the corresponding population distribution. A canonical example is choice of transport mode (such as bus, plane or train) where travelers are interviewed at their chosen mode.

The anatomy of retrospective sampling of units is depicted in the left panel of Fig. 7, where we see that the probability of selecting a unit depends on the outcome of that particular unit. Importantly, the outcomes $\mathbf{y}_j$ are colliders because they are affected by both ζ_j and $\mathbf{w}_j$. The elements of $\mathbf{s}_j$ are descendants of the colliders in $\mathbf{y}_j$ and conditioning on $\mathbf{s}_j$ (performing selection) therefore induces cluster-level endogeneity as illustrated in the right panel of Fig. 7 (conditioning is henceforth signalled by placing variables in grey background in the figures). The produced dependence between the "parents" ζ_j and $\mathbf{w}_j$ is due to "moralization" according to the terminology of Lauritzen et al. (1990).

Applications proceed with the logit link

$$\text{logit}\{p(y_{ij}=1|\mathbf{x}_{ij},\mathbf{v}_j,\zeta_j)\} = \alpha_i + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{v}_j'\boldsymbol{\gamma} + \zeta_j, \tag{23}$$

where we temporarily denote the intercept for unit i as α_i (no intercepts in $\boldsymbol{\beta}$ here).

The model is often estimated by standard MML but the marginal likelihood contribution is misspecified for the selected sample (where we denote the outcome vector as $\mathbf{y}_j^{\text{sel}}$)

$$\begin{aligned} p(\mathbf{y}_j^{\text{sel}}|\mathbf{s}_j,\mathbf{w}_j) &= \int_{\zeta_j}\left\{\prod_i p(y_{ij}^{\text{sel}}|S_{ij}=1,\mathbf{x}_{ij},\mathbf{v}_j,\zeta_j)\right\} p(\zeta_j|\mathbf{s}_j,\mathbf{w}_j)\,\mathrm{d}\zeta_j \\ &\neq \int_{\zeta_j}\left\{\prod_i p(y_{ij}^{\text{sel}}|\mathbf{x}_{ij},\mathbf{v}_j,\zeta_j)\right\} p(\zeta_j)\,\mathrm{d}\zeta_j \end{aligned}$$

We see that the correct contribution in the first line differs from the standard marginal likelihood contribution in the second line in two ways: (1) the correct conditional outcome distribution $p(y_{ij}^{\text{sel}}|S_{ij}=1,\mathbf{x}_{ij},\mathbf{v}_j,\zeta_j)$ differs from the naive one $p(y_{ij}^{\text{sel}}|\mathbf{x}_{ij},\mathbf{v}_j,\zeta_j)$ and the correct latent variable distribution $p(\zeta_j|\mathbf{s}_j,\mathbf{w}_j)$ differs from the naive one $p(\zeta_j)$. The fact that the correct latent variable distribution becomes $p(\zeta_j|\mathbf{s}_j,\mathbf{w}_j)$ corresponds to the dependence between ζ_j and $\mathbf{w}_j$ seen in the right panel of Fig. 7. As a result, standard MML estimation leads to inconsistent estimation of $\boldsymbol{\beta}$, an instance of collider-stratification bias (e.g., Greenland et al., 1999).

It is important to note that the logit link is preserved in selected samples

$$\text{logit}\{p(y_{ij}^{\text{sel}}=1|S_{ij}=1,\mathbf{x}_{ij},\mathbf{v}_j,\zeta_j)\} = \alpha_i^{\text{sel}} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{v}_j'\boldsymbol{\gamma} + \zeta_j,$$

where $\alpha_i^{\text{sel}} = \alpha_i + \log[\pi(1)/\pi(0)]$. It follows that standard CML estimation is protective. Prentice (1976) argues that protective estimation of the coefficient β for a binary x_{ij} can alternatively be

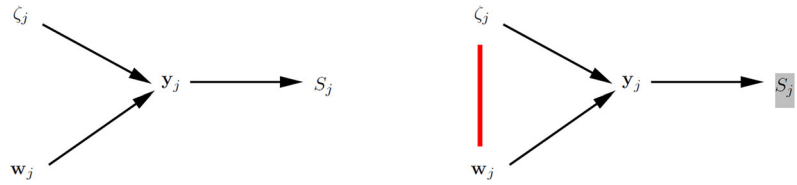

FIGURE 8.
Retrospective sampling of clusters. Unselected population (left panel) and selected sample (right panel).

achieved by CML estimation of a *retrospective* logit model with cluster-specific effects where x_{ij} is the outcome and β is now the coefficient of y_{ij}. In contrast, MML estimation with an auxiliary model produces inconsistent estimation of the intercepts but protective or mitigating estimation of the coefficients of covariates, depending on whether the auxiliary model is correct or not.

12.3.2. Sumscore-Based Sampling of Clusters For rare binary outcomes $\mathbf{y}_j$, clusters where many of the outcomes take the value zero are sometimes under-sampled. For instance, in a genetic study a family j could be more likely to be ascertained if one or more members of the family has a particular disease.

Here we consider the case where the probability of sampling a cluster j is a function of the sumscore or number of "successes" in the cluster (e.g., Neuhaus & Jewell, 1990)

$$p(S_j = 1|\mathbf{y}_j) \;=\; f\left(\sum_{i=1}^{n_j} y_{ij}\right).$$

The structure of retrospective sumscore-based sampling is shown in the left panel of Fig. 8. Importantly, we see from the right panel that conditioning on the cluster-selection indicator S_j induces cluster-level endogeneity because S_j is a descendant of the colliders in $\mathbf{y}_j$.

We proceed with a logit link as we did for retrospective sampling of units. The standard marginal likelihood contribution is now misspecified in the selected sample

$$\begin{aligned} p(\mathbf{y}_j^{\text{sel}}|S_j = 1, \mathbf{w}_j) &= \int_{\zeta_j} p(\mathbf{y}_j^{\text{sel}}|S_j = 1, \mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j) p(\zeta_j|S_j = 1, \mathbf{w}_j)\, \mathrm{d}\zeta_j \\ &\neq \int_{\zeta_j} \left\{ \prod_{i=1}^{n_j} p(y_{ij}^{\text{sel}}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j) \right\} p(\zeta_j)\, \mathrm{d}\zeta_j, \end{aligned}$$

which yields inconsistent MML estimation of $\boldsymbol{\beta}$. In contrast to retrospective sampling of units, $p(\mathbf{y}_j^{\text{sel}}|S_j = 1, \mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j)$ is no longer a product of conditionally independent logit models, but rather corresponds to the Rosner (1984) model for correlated data. MML estimation based on an auxiliary model therefore produces inconsistent estimation of $\boldsymbol{\beta}$. Fortunately, standard CML estimation is once again protective for $\boldsymbol{\beta}$ (Neuhaus & Jewell, 1990).

12.4. Informative Cluster Sizes

It is sometimes plausible that the cluster sizes n_j depend on a cluster-specific latent variable ζ_j and cluster-specific covariates $\mathbf{v}_j$. For example, for a clinical psychologist j, the patient volume n_j may depend on his latent skill ζ_j and whether he works in a public or private hospital z_j. The patient outcomes $\mathbf{y}_j$ may depend on ζ_j, z_j, and observed individual patient characteristics $\mathbf{x}_j$.

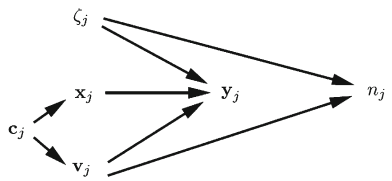

FIGURE 9.
Informative cluster-sizes.

This situation is shown in Fig. 9, where $\mathbf{c}_j$ is included to signify that $\mathbf{x}_j$ and $\mathbf{v}_j$ are typically dependent.

A joint model can be specified for the outcomes $\mathbf{y}_j$ and cluster size n_j, the latter part typically a Poisson model with log link

$$p(\mathbf{y}_j, n_j|\mathbf{x}_j, \mathbf{v}_j) = \int_{\zeta_j} p(n_j|\mathbf{v}_j, \zeta_j) \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j) \right\} p(\zeta_j)\, \mathrm{d}\zeta_j.$$

MML estimation of the joint model is consistent if the cluster-size model $p(n_j|\mathbf{v}_j, \zeta_j)$ is correctly specified in addition to the outcome model $p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j)$ and the mixing distribution $p(\zeta_j)$.

Researchers occasionally condition on n_j by including it as a covariate in the model (e.g., Seaman et al., 2014). The standard marginal likelihood contribution in this case is misspecified

$$\begin{aligned} p(\mathbf{y}_j|\mathbf{x}_j, \mathbf{v}_j, n_j) &= \int_{\zeta_j} \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j, n_j) \right\} p(\zeta_j|\mathbf{x}_j, \mathbf{v}_j, n_j)\, \mathrm{d}\zeta_j \\ &= \int_{\zeta_j} \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j, n_j) \right\} p(\zeta_j|\mathbf{v}_j, n_j)\, \mathrm{d}\zeta_j \\ &\neq \int_{\zeta_j} \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\mathbf{x}_{ij}, \mathbf{v}_j, \zeta_j, n_j) \right\} p(\zeta_j)\, \mathrm{d}\zeta_j. \end{aligned}$$

It is evident from Fig. 9 that n_j is a collider and that conditioning on n_j opens a confounding path between ζ_j and $\mathbf{x}_j$ that makes $\mathbf{x}_j$ cluster-level endogenous. However, also conditioning on $\mathbf{v}_j$ blocks this path, in which case $\mathbf{x}_j$ remains cluster-level exogenous, whereas $\mathbf{v}_j$ becomes cluster-level endogenous. The latent-variable distribution becomes $p(\zeta_j|\mathbf{v}_j, n_j)$ (shown in the second line of the likelihood contribution above), which differs from the assumed $p(\zeta_j)$ in naive MML estimation.

For the identity link, MML estimation is consistent for $\boldsymbol{\beta}$ under cluster-level exogeneity whatever the distribution of ζ_j (e.g., Verbeke & Lesaffre, 1997), but this is not the case for other links. Naive MML estimation that includes n_j as a covariate therefore gives protective estimation of $\boldsymbol{\beta}$ for the identity link, but otherwise inconsistent estimation, albeit mildly inconsistent according to simulations. Fortunately, standard CML estimation is protective for $\boldsymbol{\beta}$ for all canonical links because n_j is a cluster-level characteristic. Important advantages of this approach are that we neither need to know about the relevant $\mathbf{v}_j$ nor specify a correct model for the dependence on n_j.

In practice, the cluster size n_j is often ignored in the estimating model. From Fig. 9 we see that conditioning on $\mathbf{v}_j$ makes n_j and $\mathbf{x}_j$ conditionally independent. This implies that MML estimation for the identity link remains protective for $\boldsymbol{\beta}$. For the logit link, the estimator of $\boldsymbol{\beta}$ for

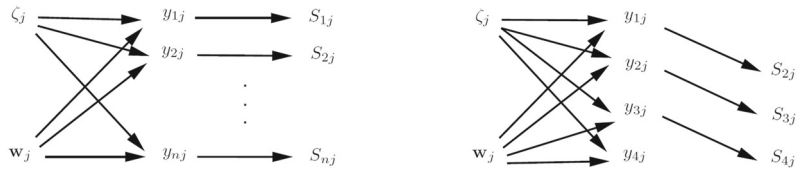

FIGURE 10.
Outcome dependent missingness. Current outcome dependent missingness (left panel) and lag(1) dependent missingness for $n = 4$ (right panel).

the model that ignores n_j has a probability limit that differs from the inconsistent MML estimator for the model where n_j is included as a covariate because odds-ratios are not collapsible (e.g., Gail et al., 1984). It is extremely unlikely that this estimator is consistent. For the log link, the inconsistency is expected to be mild. Fortunately, standard CML estimation remains protective for $\boldsymbol{\beta}$ for all canonical links.

Neuhaus and McCulloch (2011) considered a restrictive version of the model in Fig. 9 where the cluster size just depends on ζ_j. Here, there is no collider problem and the correct mixing distribution becomes $p(\zeta_j|n_j)$. The conclusions reported above for MML and CML estimation persist.

12.5. Data Missing Not at Random

12.5.1. Outcome-Dependent Missingness Missingness of outcomes that depends on the values taken by the outcomes violates the missing at random assumption (e.g., Seaman et al., 2013) and is therefore referred to as not missing at random (NMAR).

In the longitudinal non-exchangeable setting, current-outcome dependent missingness occurs if the probability that an outcome is missing at an occasion i, $S_{ij} = 1$, depends on the value taken by the outcome for that particular occasion y_{ij}. An example would be when the outcome is a disease symptom that makes it more difficult to visit a clinic for an assessment. The structure of current-outcome dependent missingness is shown in the left panel of Fig. 10, which is identical to that previously shown for retrospective sampling of units in Fig. 7.

Conditioning on the selection indicators S_{ij} leads to cluster-level endogeneity, as shown in the right panel of Fig. 7, because the S_{ij} are decendants of the colliders y_{ij}. For the non-exchangeable case, it is plausible that the missingness probabilities will differ between units i, $\pi_i(y_{ij})$.

For exchangable units, outcome-dependent missingness only makes sense if the probability that an outcome y_{ij} is missing for a unit, $S_{ij} = 1$, depends on the value taken by the outcome for that particular unit. In contrast, for longitudinal data we can also consider lag(1) outcome-dependent missingness where the probability that an outcome is missing at an occasion, $S_{ij} = 1$, depends on the previous outcome $y_{i-1,j}$. Such a process is shown in the right panel of Fig. 10 for the case of $n = 4$. This can occur if an outcome, such as a diagnosis, only affects missingness after having been relayed to the subject. Again, missingness produces cluster-level endogeneity.

Hausman and Wise (1979) and Diggle and Kenward (1994) used MML to estimate joint models where linear predictor (3) with an identity link for the outcome is combined with probit/logit models for current-outcome dependent and current plus lag(1) outcome-dependent missingness, respectively. MML estimation is consistent for all parameters under correct model specification, but this approach has been criticized for relying heavily on unverifiable distributional assumptions (e.g., Little, 1985). Standard MML estimation (ignoring the missingness) suffers from collider-stratification bias and is inconsistent for $\boldsymbol{\beta}$.

Skrondal and Rabe-Hesketh (2014) obtain several useful results for model (23). For current outcome-dependent missingness, CML estimation is protective for $\boldsymbol{\beta}$. MML estimation using an

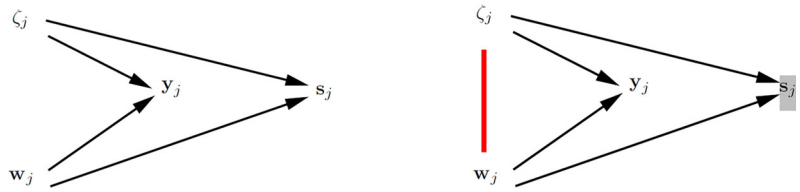

FIGURE 11.
Latent-variable and covariate dependent missingness. Unselected population (left panel) and selected sample (right panel).

auxiliary model yields protective estimation of $\boldsymbol{\beta}$ if that model is correct and mitigating estimation otherwise. For lag(1) outcome-dependent missingness, CML estimation stratified on the missingness pattern is protective. If missingness depends on both the current and all lagged outcomes (whether observed or missing), CML estimation is protective only if complete data are analyzed stratified on judiciously chosen values of the sufficient statistic. Note that none of these results hinge on specification of a parametric or nonparametric model for missingness.

12.5.2. Latent-Variable and Covariate Dependent Missingness Missingness of outcomes can depend on the latent variable in addition to the covariates and is in this case also not missing at random (NMAR). For example, the probability of visiting a clinic/answering an item could depend on the unobserved frailty/ability ζ_j of the subject as well as his observed characteristics $\mathbf{w}_j$.

The structure shown in Fig. 11 is similar to that previously discussed for informative cluster-sizes but here $\mathbf{s}_j$ could also depend on $\mathbf{x}_j$, so conditioning on $\mathbf{v}_j$ does not block the confounding path to give protective estimation of $\boldsymbol{\beta}$.

Joint modeling can be performed with "shared-parameter" models (e.g., Wu & Carroll, 1988; Ten Have et al., 1998) where the outcome and missingness processes share latent variables in addition to observed covariates. MML estimation is consistent for all parameters if the joint model is correctly specified, but note that such models rely on unverifiable distributional assumptions (e.g., Little, 1985). Standard MML estimation that ignores missingness suffers from collider-stratification bias and is inconsistent. CML estimation is protective for $\boldsymbol{\beta}$ for identity and log link functions because it can be cast as JML estimation of models with cluster-specific dummy variables.

For the logit link, Skrondal and Rabe-Hesketh (2014) prove that CML estimation is protective for $\boldsymbol{\beta}$, even when missingness also depends on missing outcomes (in addition to the latent variable and covariates). Again, this does not require specification of a parametric or nonparametric missingness model.

The case where missingness just depends on the latent variable is shown in Fig. 12. Here the shape of the latent variable distribution is changed from $p(\zeta_j)$ to $p(\zeta_j|\mathbf{s}_j)$ but there is no collider problem. For the identity link, standard MML estimation is now protective for $\boldsymbol{\beta}$. In contrast, standard MML estimation of $\boldsymbol{\beta}$ for the logit link is just mitigating, but with mild inconsistency (e.g., Neuhaus et al., 1992). CML estimation remains protective for $\boldsymbol{\beta}$.

12.6. Heteroskedastic Latent Variable

Cluster-level endogeneity occurs if the variance of the latent variable distribution depends on cluster-level covariates $\mathbf{v}_j$. For example, Heagerty (1999) considered a longitudinal study of schizophrenia where the latent variable variance depends on gender. Such heteroskedasticity is illustrated in Fig. 13 where $\mathbf{v}_j$ is a subset of $\mathbf{w}_j$.

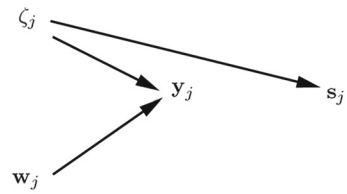

FIGURE 12.
Latent-variable dependent missingness.

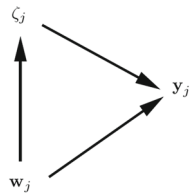

FIGURE 13.
Heteroskedastic latent variable.

For the identity link, standard MML estimation remains consistent under this misspecification. For the logit link, Heagerty and Kurland (2001) demonstrated that standard MML estimation becomes inconsistent for all parameters. If the structure of the heteroskedasticity is known, consistent MML estimation can be achieved if an appropriate heteroskedasticity model can be specified, for instance by using a GLLAMM. Fortunately, standard CML estimation is protective for $\boldsymbol{\beta}$, and does not require additional modeling or even knowing the variables that induce heteroskedasticity.

13. Latent Variable Scoring

When considering a latent variable model without an incidental parameters problem, it is straightforward to obtain estimates of ζ_j by JML estimation using (4). In general, this scoring method can work well for large cluster sizes n_j.

CML estimation is of limited value if the target of inference is the value of the latent variable ζ_j. However, we can estimate the cluster-specific component $u_j \equiv \zeta_j + h(\mathbf{v}_j)$ that is eliminated in CML estimation by maximizing the scoring likelihood

$$\prod_{i=1}^{n_j} p(y_{ij}|\mathbf{x}_{ij}; \widehat{\boldsymbol{\vartheta}}^{\mathrm{CML}}, u_j)$$

with respect to u_j. The estimated scores $\widehat{u}_j^{\mathrm{ML}}$ can then be plugged in to obtain predictions of outcomes $p(y_{ij}|\mathbf{x}_{ij}; \widehat{\boldsymbol{\vartheta}}^{\mathrm{CML}}, \widehat{u}_j^{\mathrm{ML}})$. Improvements of the ML estimator, such as variants of the weighted likelihood method of Warm (1989), can also be employed.

If CML estimation is mimicked by MML estimation of augmented models, we can use empirical Bayes (EB) prediction that performs partial pooling of information from other clusters and is therefore more precise. For instance, for joint modeling (see Sect. 11.2) the predictions can

be obtained as:

$$\widetilde{\zeta}_j = \frac{\int_{\zeta_j} \zeta_j \, p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j; \widehat{\boldsymbol{\vartheta}}^{\text{MML}}) p(\mathbf{w}_j|\zeta_j; \widehat{\boldsymbol{\vartheta}}^{\text{MML}}) \phi(\zeta_j; 0, \widehat{\psi}^{\text{MML}}) \, \mathrm{d}\zeta_j}{\int_{\zeta_j} p(\mathbf{y}_j|\mathbf{w}_j, \zeta_j; \widehat{\boldsymbol{\vartheta}}^{\text{MML}}) p(\mathbf{w}_j|\zeta_j; \widehat{\boldsymbol{\vartheta}}^{\text{MML}}) \phi(\zeta_j; 0, \widehat{\psi}^{\text{MML}}) \, \mathrm{d}\zeta_j}.$$

The performance of EB prediction in this case also relies on correct specification of $p(\mathbf{w}_j|\zeta_j; \boldsymbol{\vartheta})$. Parametric assumptions are moreover made regarding the latent variable distribution unless non-parametric marginal maximum likelihood ((NPMML) is used (Rabe-Hesketh et al., 2003). However, simulation studies suggest that violations of the distributional assumptions may have a modest impact on the mean squared error of prediction unless the assumed distribution has more limited support than the correct distribution, the latent variable variance is large, or the cluster sizes are large (e.g., McCulloch & Neuhaus, 2011a; 2011b).

We refer to Skrondal and Rabe-Hesketh (2009) for latent variable scoring and various kinds of prediction of outcomes in GLMMs and related models with multidimensional latent variables.

14. Concluding Remarks

We have demonstrated that conditional likelihoods have an important role to play in latent variable modeling that extends well beyond Rasch models for measurement. For the class of models considered here, a great advantage of CML estimation is that it can *simultaneously* handle cluster-level endogeneity problems induced by, for instance, unobserved cluster-level confounding of causal effects, cluster-specific measurement error, retrospective sampling, informative cluster sizes, missing data, and heteroskedasticity.

Although randomized experimental designs ensure that there is no confounding of treatment effects, there could be cluster-level endogeneity due to, for instance, covariate measurement error and data not missing at random. The famous Hausman (1978) specification test that compares fixed-effect estimates (e.g., from CML estimation) with random-effects estimates (e.g., from MML estimation) is routinely used in the context of model (3). Contrary to common belief, a significant test cannot be interpreted as flagging unobserved confounding because cluster-level endogeneity can arise for a variety of reasons.

In psychology, split-plot analysis of variance (ANOVA) is sometimes used for repeated measures designs. The hypothesis test for a within-subject effect is in this case robust against cluster-level endogeneity. However, the focus is traditionally solely on hypothesis testing and not estimation of parameters or effect sizes. A very similar fixed-effects approach known as difference-in-differences in economics is popular for estimating effects in natural/quasi experiments (e.g., Angrist & Pischke, 2009).

Hybrid estimation approaches can be obtained by combining CML with other estimation methods:

(a) CML and MML (and related) estimation:
In a Rasch context, Andersen and Madsen (1977) used CML to estimate item parameters $\boldsymbol{\beta}$ and subsequently MML to estimate the expectation and variance of a parametric person distribution, given $\widehat{\boldsymbol{\beta}}^{\text{CML}}$. For a linear mixed model with random slopes, Verbeke et al. (2001) used a conditional likelihood to eliminate the cluster-specific intercept ζ_j (and $\boldsymbol{\gamma}$), and MML or restricted maximum likelihood (REML) to estimate $\boldsymbol{\beta}$, σ^2 and the covariance matrix of the slopes in $\boldsymbol{\zeta}_j$. Tibaldi et al. (2007) combined CML and composite-likelihood estimation of crossed random-effects models with identity and logit links.

(b) CML and instrumental variables (IV) estimation:
For model (3) with identity link, Hausman and Taylor (1981) proposed a multi-stage estimation approach where CML is used to estimate $\boldsymbol{\beta}$ (and σ^2) followed by IV estimation of $\boldsymbol{\gamma}$ (and the variance of ζ_j) for given $\widehat{\boldsymbol{\beta}}^{\text{CML}}$. The instruments are internal in the sense that they are constructed from the covariates $\mathbf{x}_{ij}$ and $\mathbf{v}_j$ in the model. All parameters can be consistently estimated if one can correctly designate which unit- and cluster-specific covariates are cluster-level endogenous. In a panel data setting with identity link, Lee (2002) used first differencing to remove cluster-specific intercepts. Subsequently, he used IV methods to estimate $\boldsymbol{\beta}$, where various exogeneity assumptions for the ϵ_{ij} dictate if covariates at an occasion can serve as internal instruments for covariates at other occasions.

(c) CML and Bayes estimation:
Conditional likelihoods have been used in conjunction with prior distributions for model parameters in Bayesian inference. This was motivated by Diggle et al. (2000) to handle retrospective sampling and by Lancaster (2004) to handle unobserved confounding. It is worth pointing out that standard Bayesian inference that ignores cluster-level endogeneity performs similarly to standard MML estimation.

(d) Mixed-type ML estimation:
Cook and Farewell (1999) discussed the construction of mixed-type likelihoods where the likelihood contributions are of different types for different clusters. They considered a version of linear predictor (3) with a logit link, where conditional likelihoods were used for small clusters whereas joint likelihoods were used for large clusters.

An interesting recent development is the use of conditional likelihoods in double-robust estimation. Zetterquist et al. (2019) consider model (3) with a logit link and a binary treatment x_{ij} of interest. They argue that consistency for the corresponding β can be achieved if at least one of the following approaches is consistent for β: i) CML estimation of β in the prospective model (3) for y_{ij} or ii) CML estimation of β as the coefficient of y_{ij} in a retrospective logit model for x_{ij} with a cluster-specific intercept. One need not know which of the models is correct so there are in this sense two opportunities of getting it right.

In closing, it is evident that we have drawn on and extended results not just from psychometrics but also from other "ics", such as econometrics, biometrics and statistics, in this address. Unfortunately, progress in psychometrics has been hampered by a paucity of cross-fertilization from the other "ics" — but the opposite is certainly also true! A case in point is the extensive literature on covariate measurement error where the psychometric wheel is regularly reinvented, albeit often in a somewhat square fashion. We end with a perceptive citation from a *Psychometrika* article by the renowned econometrician Arthur Goldberger (Goldberger, 1971, p. 83):

> "Economists and psychologists have been developing their statistical techniques quite independently for many years. From time to time, a hardy soul strays across the frontier but is not met with cheers when he returns home."

Funding This work was partially supported by The Research Council of Norway through it Centres of Excellence funding scheme, project number 26270. Open access funding provided by University of Oslo (incl Oslo University Hospital).

References

Aigner, D., Hsiao, C., Kapteyn, A., & Wansbeek, T. (1984). Latent variable models in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, pp. 1321–1393). North-Holland.

Albert, A., & Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71,* 1–10.

Allison, P. (2009). *Fixed effects regression models*. Sage.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society Series B, 32,* 283–301.

Andersen, E. B. (1971). The asymptotic distribution of conditional likelihood ratio tests. *Journal of the American Statistical Association, 66,* 630–633.

Andersen, E. B. (1973a). *Conditional inference and models for measuring*. Mentalhygiejnisk Forsknings Institut.

Andersen, E. B. (1973b). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. North-Holland.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357–374.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Angrist, J., & Pischke, J. (2009). *Mostly harmless econometrics*. Princeton University Press.

Arellano, M., & Hahn, J. (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. Newey, & T. Persson (Eds.), *Advances in economics and econometrics: Ninth world congress* (pp. 381–409). Cambridge University Press.

Arminger, G., & Schoenberg, R. (1989). Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika, 54,* 409–425.

Balazsi, L., Matyas, L., & Wansbeek, T. (2017). Fixed effects models. In L. Matyas (Ed.), *The econometrics of multidimensional panels* (pp. 1–34). Springer.

Bartlett, M. (1936). The information available in small samples. *Proceedings of the Cambridge Philosophical Society, 32,* 560–566.

Bartlett, M. (1937a). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, A, 160,* 268–282.

Bartlett, M. (1937b). The statistical conception of mental factors. *British Journal of Psychology, 28,* 97–104.

Bartolucci, F., & Nigro, V. (2010). A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica, 78,* 719–733.

Bartolucci, F., Bellio, R., Salvan, A., & Sartori, N. (2016). Modified profile likelihoods for fixed-effects panel models. *Econometric Reviews, 35,* 1271–1289.

Bellio, R., & Sartori, N. (2006). Practical use of modified maximum likelihoods for stratified data. *Biometrical Journal, 48,* 876–886.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179–197.

Bollen, K., & Brand, J. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 81,* 1–34.

Booth, J., & Hobert, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B, 61,* 265–285.

Bound, J., & Solon, G. (1999). Double trouble: On the value of twins-based estimation of the return to schooling. *Economics of Education Review, 18,* 169–182.

Box, G., & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. Holden Day.

Breslow, N. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association, 91,* 14–28.

Breusch, T. (1987). Maximum likelihood estimation of random effects models. *Journal of Econometrics, 36,* 383–389.

Browne, M. (1974). Generalized least squares estimation in the analysis of covariance structures. *South African Statistical Journal, 8,* 1–24.

Brumback, B., Dailey, A., Brumback, L., Livingston, M., & He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters, 80,* 1650–1654.

Butler, S., & Louis, T. (1997). Consistency of maximum likelihood estimators in general random effects models for binary data. *Annals of Statistics, 25,* 351–377.

Cameron, C., & Trivedi, P. (1999). *Regression analysis of count data*. Cambridge: Cambridge University Press.

Castellano, K., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39,* 333–367.
Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies, 47,* 225–238.
Chamberlain, G. (1984). Panel data. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, pp. 131–1247). North-Holland.
Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 3–38). Cambridge University Press.
Charbonneau, K. (2017). Multiple fixed effects in binary response panel data models. *Econometrics Journal, 20,* S1–S13.
Chen, H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics, 63,* 413–421.
Conaway, M. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association, 84,* 53–62.
Cook, R., & Farewell, V. (1999). The utility of mixed-form likelihoods. *Biometrics, 55,* 284–288.
Cox, D. R. (1970). *The analysis of binary data*. Methuen.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, 34,* 187–202.
Cox, D. R. (1975). Partial likelihood. *Biometrika, 62,* 269–276.
Cox, D. R., & Wermuth, N. (1994). A note on the quadratic exponential binary model. *Biometrika, 81,* 403–408.
De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.
de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.
D'Haultfæuille, X., & Iaria, A. (2016). A convenient method for the estimation of the multinomial logit model with fixed effects. *Economics Letters,141*, 77–79.
Diggle, P., & Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics, 43,* 49–93.
Diggle, P., Morris, S., & Wakefield, J. (2000). Point-source modeling using matched case-control data. *Biostatistics, 1,* 89–105.
Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20,* 115–147.
Ebbes, P., Böckenholt, U., & Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica, 58,* 161–178.
Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution, 35,* 1229–1242.
Fischer, G. (1995a). Derivation of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models. foundations, recent developments, and applications* (pp. 15–38). Springer.
Fischer, G. (1995b). The linear logistic test model. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 131–155). New York: Springer.
Fischer, G. (1995c). Linear logistic models for change. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 157–180). Springer.
Formann, A. (1995). Linear logistic latent class analysis and the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 239–255). Springer.
Frisell, T., Öberg, S., Kuja-Halkola, R., & Sjölander, A. (2012). Sibling comparison designs. *Epidemiology, 23,* 713–720.
Fuller, W., & Battese, G. (1973). Transformations for estimation of linear models with nested error structure. *Journal of the American Statistical Association, 68,* 626–632.
Gail, M., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71,* 431–444.
Goetgeluk, S., & Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics, 64,* 772–780.
Goldberger, A. S. (1971). Econometrics and psychometrics: A survey of communalities. *Psychometrika, 36,* 83–107.
Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.
Gourieroux, C., & Monfort, A. (1995). *Statistics and econometrics* (Vol. 2). Cambridge University Press.
Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica, 52,* 681–700.
Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal, 7,* 98–119.
Greenland, S., Pearl, J., & Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10,* 37–48.
Griliches, Z., & Hausman, J. (1986). Errors in variables in panel data. *Journal of Econometrics, 31,* 93–118.
Gustafsson, J.-E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 377–385.
Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5,* 815–841.
Hausman, J. (1978). Specification tests in econometrics. *Econometrica, 46,* 1251–1271.
Hausman, J., & Taylor, W. (1981). Panel data and unobservable individual effects. *Econometrica, 49,* 1377–1398.
Hausman, J., & Wise, D. (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica, 47,* 455–473.
Hausman, J., Hall, B., & Griliches, Z. (1984). Econometric models for count data with an application to the Patents-R&D relationship. *Econometrica, 52,* 909–939.
Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage.

Honoré, B., & Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica, 68,* 839–874.
Heagerty, P. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics, 55,* 688–698.
Heagerty, P., & Kurland, B. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika, 88,* 973–985.
Holland, P. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.
Howard, S. (1972). Discussion on professor Cox's paper. *Journal of the Royal Statistical Society, Series B, 34*, 210–211.
Imai, K., & Kim, I. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science, 63,* 467–490.
Kalbfleisch, J. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association, 73,* 167–170.
Kalbfleisch, J., & Sprott, D. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B, 32,* 175–208.
Kelderman, H., & Rijkes, C. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149–176.
Kertesz, B. (2017). Discrete response models. In L. Matyas (Ed.), *The econometrics of multi-dimensional panels* (pp. 163–194). Springer.
Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–906.
Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963–974.
Laisney, F., & Lechner, M. (2003). Almost consistent estimation of panel probit models with 'small' fixed effects. *Econometric Reviews, 22*, 1–28.
Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge University Press.
Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics, 95,* 391–413.
Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Wiley.
Lauritzen, S., Dawid, A., Larsen, B., & Leimer, M. (1990). Independence properties of directed Martkov fields. *Networks, 20,* 491–505.
Lee, M.-J. (2002). *Panel data econometrics: Methods-of-moments and limited dependent variables*. Academic Press.
Liang, K.-Y. (1987). Extended Mantel–Haenszel estimating procedure for multivariate logistic regression models. *Biometrics, 43,* 289–299.
Liang, K.-Y., & Zeger, S. (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhya, 62,* 134–148.
Lindsay, B. G., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96–107.
Little, R. (1985). A note about models for selectivity bias. *Econometrica, 53,* 1469–1474.
Maddala, G. (1971). The use of variance components models in pooling cross section and time series data. *Econometrica, 39,* 341–358.
Manski, C. (1981). Models for discrete data: The analysis of discrete choice. *Sociological Methodology, 12,* 58–109.
Maris, E. (1998). On the sampling interpretation of confidence intervals and hypothesis tests in the context of conditional maximum likelihood estimation. *Psychometrika, 63,* 65–71.
Maris, G., & Bechger, T. (2007). Scoring open ended questions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 663–681). Elsevier.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.
McCulloch, C., & Neuhaus, J. (2011a). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science, 26,* 388–402.
McCulloch, C., & Neuhaus, J. (2011b). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics, 67,* 270–279.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic.
Mehta, C., & Patel, N. (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine, 19,* 2143–2160.
Molenaar, I. (1995). Estimation of item parameters. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 39–51). Springer.
Mukherjee, B., Ahn, J., Liu, I., Rathouz, P., & Sanchez, B. (2008). Fitting stratified proportional odds models by amalgamating conditional likelihoods. *Statistics in Medicine, 27,* 4950–4971.
Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46,* 69–85.
Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, 135,* 370–384.
Neuhaus, J., Hauck, W., & Kalbfleisch, J. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika, 79,* 755–762.
Neuhaus, J., & Jewell, N. (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics, 46,* 977–990.
Neuhaus, J., & McCulloch, C. (2006). Separating between-and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society, Series B, 68,* 859–872.
Neuhaus, J., & McCulloch, C. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika, 98,* 147–162.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16,* 1–32.
Palta, M., & Yao, T.-J. (1991). Analysis of longitudinal data with unmeasured confounders. *Biometrics, 47,* 1355–1369.
Patterson, H., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58,* 545–554.
Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford University Press.
Petersen, A., & Lange, T. (2020). What is the causal interpretation of sibling comparison designs? *Epidemiology, 31,* 75–81.
Pfeiffer, R., Gail, M., & Pee, D. (2001). Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika, 88,* 933–948.
Prentice, R. (1976). Use of the logistic model in retrospective studies. *Biometrics, 32,* 599–606.
Prentice, R., & Breslow, N. (1978). Retrospective studies and failure time models. *Biometrika, 65,* 153–158.
Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modeling, 3,* 215–232.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128,* 301–323.
Rabe-Hesketh, S., & Skrondal, A. (2009). Generalized linear mixed-effects models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 79–106). Chapman & Hall/CRC.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Pædagogiske Institut.
Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Volume 4: Contributions to Biology and Problems of Medicine* (pp. 321–333). University of California Press.
Rice, K. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association, 99,* 510–522.
Ridder, G., & Tunali, I. (1999). Stratified partial likelihood estimation. *Journal of Econometrics, 92,* 193–232.
Robins, J., Mark, S., & Newey, W. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics, 48,* 479–495.
Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics, 40,* 1025–1035.
Rothman, K., Greenland, S., & Lash, T. (2008). *Modern epidemiology* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.
Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores*. Bowling Green: Psychometrika Monograph 17, Psychometric Society.
Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics, 43,* 142–152.
Sartori, N., & Severini, T. (2004). Conditional likelihood inference in generalized linear mixed models. *Statistica Sinica, 14,* 349–360.
Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by "missing at random" ? *Statistical Science, 28*, 257–268.
Seaman, S., Pavlou, M., & Copas, A. (2014). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine, 33,* 5371–5387.
Sim, N. (2019). Beyond standard assumptions—semiparametric models, a dyadic item response theory model, and cluster-endogenous random intercept models. Ph.D. Dissertation, Berkeley: University of California.
Sjölander, A., Johansson, A., Lundholm, C., Altman, D., Almqvist, C., & Pawitan, Y. (2012). Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Statistical Science, 27,* 395–411.
Sjölander, A., Frisell, T., Kuja-Halkola, R., Öberg, S., & Zetterquist, J. (2016). Carryover effects in sibling comparison designs. *Epidemiology, 27,* 852–858.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC.
Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A, 172,* 659–687.
Skrondal, A., & Rabe-Hesketh, S. (2014). Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika, 101,* 175–188.
Sobel, M. (2012). Does marriage boost men's wages?: Identification of treatment effects from fixed effects regression models for panel data. *Journal of the American Statistical Association, 107*, 521–529.
Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. IMS Lecture Notes-Monograph Series 2nd Lehmann Symposium-Optimality, **49**, 98–119.
Storer, B., Wacholder, S., & Breslow, N. (1983). Maximum likelihood fitting of general risk models to stratified data. *Applied Statistics, 32,* 172–181.
Teachman, J., Duncan, G., Yeung, J., & Levy, D. (2001). Covariance structure models for fixed and random effects. *Sociological Methods and Research, 30,* 271–288.
Ten Have, T., Kunselman, A., Pulkstenis, E., & Landis, J. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics, 54,* 367–383.
Thomas, A. (2006). Consistent estimation of binary-choice panel data models with heterogeneous linear trends. *Econometrics Journal, 9,* 177–195.

Tibaldi, F., Verbeke, G., Molenberghs, G., Renard, D., van den Noortgate, W., & De Boeck, P. (2007). Conditional mixed models with crossed random effects. *British Journal of Mathematical and Statistical Psychology, 60,* 351–365.
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.
Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis, 23,* 541–556.
Verbeke, G., Spiessens, B., & Lesaffre, E. (2001). Conditional linear mixed models. *American Statistician, 55,* 25–34.
Verhelst, N. (2019). Exponential family models for continuous responses. In B. Veldkamp & C. Sluijter (Eds.), *Methodology of educational measurement and assessment* (pp. 135–160). Springer.
Verhelst, N., & Glas, C. (1995). The one parameter logistic model. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 215–237). Springer.
Verma, T. & Pearl, J. (1988). Causal networks: Semantics and expressiveness. In: R. Schachter, T. Levitt, L. Kanal & J. Lemmer (Eds.), *Proceedings of the 4th conference on uncertainty and artificial intelligence* (pp. 69–76). Elsevier.
von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 371–379). Springer.
Wang, M., Flanders, W., Bostick, R., & Long, Q. (2012). A conditional likelihood approach for regression analysis using biomarkers measured with batch-specific error. *Statistics in Medicine, 31,* 3896–3906.
Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.
Wooldridge, J. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics, 90,* 77–97.
Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.
Wright, B., & Douglas, G. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281–294.
Wu, M., & Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics, 44,* 175–188.
Zeger, S., Liang, K.-Y., & Albert, P. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics, 44,* 1049–1060.
Zetterqvist, J., Vermeulen, K., Vansteelandt, S., & Sjölander, A. (2019). Doubly robust conditional logistic regression. *Statistics in Medicine, 38,* 4749–4760.
Zhang, D., & Davidian, M. (2004). Likelihood and conditional likelihood inference for generalized additive mixed models for clustered data. *Journal of Multivariate Analysis, 91,* 90–106.
Zwitser, R., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika, 80,* 65–84.

Manuscript Received: 11 SEP 2021
Final Version Received: 3 OCT 2021
Published Online Date: 10 JAN 2022