Journal of the Royal Statistical Society
SERIES C Applied Statistics
Appl. Statist. (2014)
63, *Part* 2, *pp.* 211–237

Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity

Anders Skrondal

Norwegian Institute of Public Health, Oslo, Norway

and Sophia Rabe-Hesketh

University of California, Berkeley, USA, and Institute of Education, London, UK

[Received July 2012. Final revision March 2013]

Summary. Distinguishing between longitudinal dependence due to the effects of previous responses on subsequent responses and dependence due to unobserved heterogeneity is important in many disciplines. For example, wheezing is an inflammatory reaction that may 'remodel' a child's airway structure and thereby affect the probability of future wheezing (state dependence). Alternatively, children could vary in their susceptibilities because of unobserved covariates such as genes (unobserved heterogeneity). For binary responses, distinguishing between state dependence and unobserved heterogeneity is typically accomplished by using dynamic/transition models that include both a lagged response and a random intercept. Naive maximum likelihood estimators can be severely inconsistent because of two kinds of endogeneity problem: lack of independence of the initial response and the random intercept (the initial conditions problem) and lack of independence of the covariates and the random intercept (the endogenous covariates problem). We clarify and unify previous work on handling these problems in the disconnected literatures of statistics and econometrics, suggest improved methods, investigate the asymptotic performance of competing methods and provide practical recommendations. The recommended methods are applied to longitudinal data on children's wheezing, where we investigate the extent of state dependence and unobserved heterogeneity and whether there is an effect of maternal smoking.

Keywords: Auto-regressive model; Dynamic model; Endogeneity; gllamm; Initial conditions; Panel data; State dependence; Transition model; Unobserved confounding

1. Introduction

Responses in longitudinal or panel data are invariably dependent over time, even after conditioning on observed covariates. Different kinds of statistical models have therefore been proposed to handle such longitudinal within-subject dependence. In this paper we focus on models for binary data and on the prevalent case of 'short panels' with relatively few occasions. We analyse wheezing status of children initially aged 7 years and followed up annually for 3 years.

A standard approach to handling longitudinal dependence is to use models where binary responses are regressed on previous or lagged responses (typically just the preceding response). Such models have many names; they are often called (Markov) transition models in statistics (e.g. Diggle *et al*. (2002)) and dynamic models in econometrics (e.g. Hsiao (2002)). In our

Address for correspondence: Anders Skrondal, Division of Epidemiology, Norwegian Institute of Public Health, PO Box 4404 Nydalen, N-0403 Oslo, Norway.
E-mail: anders.skrondal@fhi.no

 0035–9254/14/63211

application, the probability of a child wheezing at an occasion could depend on whether or not the child experienced wheezing at the previous occasion. This is because wheezing is an inflammatory reaction that can 'remodel' a child's airway structure. Such state dependence is also often considered in social statistics, a prominent example being unemployment, where the probability of a person being unemployed at an occasion could be higher if he was unemployed at the previous occasion than if he was employed. This might be because having experienced unemployment has changed the person in various ways and because becoming unemployed is not the same as remaining unemployed. In this case, past experience has a genuine behavioural effect in the sense that an otherwise identical person who did not experience the event would behave differently in the future from a person who experienced the event (Heckman, 1981a). A policy implication is that an intervention that changes the state at an occasion would affect future states.

Another standard approach is to use subject-specific effects to handle longitudinal dependence. In statistics, random-effects models are typically used where subject-specific random effects are (usually implicitly) assumed to be uncorrelated with the covariates (e.g. Verbeke and Molenberghs (2000)). In econometrics, fixed effects approaches that relax this assumption are more common (e.g. Wooldridge (2010)). The subject-specific effects are often interpreted as representing the combined effects of omitted time constant covariates. For example, the probability of wheezing could vary between children over and above variability explained by observed covariates (e.g. maternal smoking) and previous wheezing because of unobserved time constant covariates (e.g. genes). Regarding the unemployment example, the probability of unemployment could vary between people over and above variability explained by observed covariates (e.g. age) and previous unemployment because of omitted person-specific covariates (e.g. ability).

It is evident that effects of previous responses (or states) and unobserved heterogeneity are two competing but not mutually exclusive explanations of within-subject dependence. The two types of dependence were called 'true contagion' and 'false contagion' respectively by Bates and Neyman (1952) and 'true state dependence' and 'spurious state dependence' respectively by Heckman (1981b). There are of course other explanations for longitudinal dependence, such as serially correlated errors and dynamics in the latent responses (see, for example, Heckman (1981a)), but the focus in this paper will be on models that can distinguish between the effects of previous states and unobserved heterogeneity.

An obvious approach to distinguishing between state dependence and unobserved heterogeneity seems to be simply to include the lagged response as an additional covariate in a random-intercept model. This is frequently done in applications and is also common in statistical papers (see, for example, Albert and Follmann (2003, 2007), Sutradhar and Farrell (2007) and Song *et al.* (2011)). Unfortunately, the corresponding maximum likelihood estimators can be quite severely inconsistent owing to the initial conditions problem. The initial response at the start of the observation period is affected by the random intercept and presample responses, and ignoring this endogeneity leads to inconsistent estimation.

For continuous responses, methods for handling the initial conditions problem are standard in econometrics (e.g. Anderson and Hsiao (1982), Bhargava and Sargan (1983) and Arellano and Bond (1991)). Corresponding methods for other response types are less well developed, perhaps because the problem is more challenging and because the contributions are scattered in the largely disconnected literatures of statistics and econometrics. In this paper we discuss several versions of each of two approaches to the problem for binary responses:

(a) modelling the initial response jointly with the subsequent response (e.g. Heckman (1981a)) and
(b) conditioning on the initial response (e.g. Wooldridge (2005)).

We unify a range of models from the statistical and econometrics literatures by using a common notation and path diagrams, highlight problems with previous implementations and propose solutions. We point out, to our knowledge for the first time, that robust standard errors should be used instead of model-based standard errors because the models are known to be misspecified.

In addition to the initial conditions problem, another problem that leads to inconsistent estimators if ignored is the endogenous covariates problem where the random intercept is not independent of the covariates. Such endogeneity is due to omitted subject level (time constant) covariates that are correlated with the included covariates or, in other words, unobserved between-subject confounding. Again, methods for handling this problem for continuous responses are standard in econometrics (e.g. Mundlak (1978)), but corresponding methods for binary responses are less well known, particularly in statistics. We therefore discuss methods for handling endogenous covariates in dynamic or transition models for binary data. As far as we are aware, the endogenous covariates problem has not previously been considered in the joint modelling approach to the initial conditions problem. We also discuss, to our knowledge for the first time, how missing data can be handled. In addition to analysing the wheezing data, we investigate asymptotic performance and provide practical recommendations.

The plan of the paper is as follows. We first describe the data and research questions in Section 2. We then introduce dynamic panel models or transition models for binary responses with unobserved heterogeneity in Section 3. In Section 4 we describe the initial conditions problem and review and extend approaches to address it. In Section 5 we discuss the endogenous covariates problem and how to handle it in conjunction with the initial conditions problem. The methods proposed are used to distinguish between state dependence and unobserved heterogeneity for children's wheezing and to investigate the effect of maternal smoking on children's wheezing in Section 6. In Section 7 we investigate the asymptotic performance of different approaches. Finally, we close the paper with a brief discussion in Section 8 where we outline further extensions.

2. Children's wheezing: data and research questions

The Harvard six cities study (e.g. Ware *et al.* (1984), Ferris *et al.* (1985) and Speizer (1990)) prompted radical revisions to the US Clean Air Act. As part of this study, complete grades of elementary school children were enrolled at participating schools in each of six communities in the eastern and midwestern USA and then seen annually for a pulmonary function examination. At the time of each examination, their parents completed a respiratory illness questionnaire.

The data set that is used here was provided by Lawal (2003) and is on the history of wheezing for $N = 412$ white children who were examined annually at ages 7–10 years. The children lived in Kingston-Harriman, Tennessee, and Portage, Wisconsin: two cities chosen because they have very different ambient air quality. Kingston-Harriman is influenced by air pollution from several metropolitan and industrial areas and has high concentrations of fine particulate matter and acid aerosols whereas Portage has low concentrations.

A diagnosis of persistent wheeze required a positive response to the question

> 'Does your child's chest ever sound wheezy or whistling (1) apart from colds, or (2) most days or nights?'.

At each occasion we also have self-reported data on the number of cigarettes smoked by the mother per day.

The missingness patterns for wheezing are shown in Table 1 where we see that 65% of the

Table 1. Missingness patterns for wheeze in children's wheezing data†

Frequency	*%*	*Pattern*			
267	65	1	1	1	1
35	9	·	1	1	1
26	6	1	1	·	·
6	2	1	1	·	1
19	5	1	1	1	·
12	3	·	·	1	1
4	1	1	·	1	1
4	1	·	1	1	·
16	4	·	·	·	1
14	3	1	·	·	·
9	2		(other)		
412	100				

†Dots denote missing and 1 denotes non-missing.

children have complete data. The 373 children with patterns above the horizontal line have at least two contiguous non-missing values of wheezing and therefore contribute to the analysis if previous wheezing status is required. The patterns that are connected by braces are equivalent if previous wheezing status is required. Note that different children have their initial response at different occasions depending on the missingness pattern.

A strong association has been reported between children's respiratory infections over time (e.g. Fuhlbrigge *et al*. (2001)). One explanation of this dependence is that children have varying susceptibilities for respiratory infections. In practice we would expect that there is *unobserved heterogeneity* where unknown characteristics of the children such as their genes and environment cause wheezing (e.g. Ober and Yao (2011)). A competing explanation is *state dependence* where having experienced wheezing makes a child more prone to experience later wheezing. A possible causal mechanism is that wheezing is an inflammatory response which remodels the airway structure (e.g. An *et al*. (2007)). Our first research question concerns to what extent the within-child dependence of children's wheezing is due to state dependence and unobserved heterogeneity.

In cross-sectional studies, a strong association has also been reported between exposure to maternal smoking and wheezing in childhood (e.g. Gilliland *et al*. (2001)). However, such associations are likely to be affected by unobserved between-child confounding. Our second research question therefore concerns whether there is a within-child effect of maternal smoking on children's wheezing.

3. A dynamic/transition random-intercept model for binary panel data

We use index j for subject and i for occasion or panel wave and consider models for binary responses y_{ij} with time invariant covariates $\mathbf{z}_j$ (with first element equal to 1) and time-varying covariates $\mathbf{x}_{ij}$. The process started at occasion $S_j < 0$ for subject j and we shall later assume that the process is observed at occasions $i = 0, 1, \ldots, T-1$, apart from missing data. We further let $\mathbf{y}_j^{\text{all}} = (y_{S_j j}, \ldots, y_{0j}, \ldots, y_{T-1,j})'$ and $\mathbf{x}_j^{\text{all}} = (\mathbf{x}_{S_j j}, \ldots, \mathbf{x}_{0j}, \ldots, \mathbf{x}_{T-1,j})'$ contain all responses and

time-varying covariates respectively, from the inception of the process at occasion S_j to the end of observation at occasion $T-1$.

The following type of Markov chain is considered:

$$\Pr(y_{ij}|y_{i-1,j},\ldots,y_{S_j,j},\mathbf{z}_j,\mathbf{x}_j^{\text{all}},\zeta_j)=\Pr(y_{ij}|y_{i-1,j},\mathbf{z}_j,\mathbf{x}_{ij},\zeta_j), \tag{1}$$

where ζ_j is a random subject-specific intercept. This is a first-order Markov chain where the current response y_{ij} depends on the lag 1 response $y_{i-1,j}$ but not on $y_{i-2,j},\ldots,y_{S_j,j}$, for given $y_{i-1,j},\ldots,y_{S_j,j}$, $\mathbf{z}_j$, $\mathbf{x}_j^{\text{all}}$ and ζ_j. The Markov chain is non-stationary, since it is driven by time-varying covariates $\mathbf{x}_{ij}$.

We treat the covariates $\mathbf{z}_j$ and $\mathbf{x}_j^{\text{all}}$ as random variables because this is natural in observational studies such as the wheezing study where maternal smoking is one of the covariates. The time-varying covariates $\mathbf{x}_{ij}$ are strictly exogenous given $y_{i-1,j},\ldots,y_{S_j,j}$, $\mathbf{z}_j$ and ζ_j, meaning that only the current values $\mathbf{x}_{ij}$ of the time-varying covariates appear on the right-hand side of equation (1), although $\mathbf{x}_j^{\text{all}}$ appears in the conditioning set on the left. We assume that ζ_j is independent of the covariates, which is an assumption that will be relaxed in Section 5 on handling endogenous covariates.

We consider the following parametric dynamic/transition random-intercept model for the Markov chain in equation (1):

$$\Pr(y_{ij}=1|y_{i-1,j},\mathbf{z}_j,\mathbf{x}_{ij},\zeta_j)=h^{-1}(\mathbf{z}_j'\boldsymbol{\gamma}_z+\mathbf{x}_{ij}'\boldsymbol{\gamma}_x+\beta y_{i-1,j}+\zeta_j). \tag{2}$$

Here $\zeta_j\sim N(0,\psi)$, $\boldsymbol{\gamma}_z$ and $\boldsymbol{\gamma}_x$ are the coefficient vectors for $\mathbf{z}_j$ and $\mathbf{x}_{ij}$ respectively and β is the coefficient of the lagged response $y_{i-1,j}$. Since there is no occasion index for β, the Markov chain is time homogeneous and we hence implicitly assume that the time interval between occasions is approximately constant. $h(\cdot)$ is a link function, which is invariably taken as the logit, $h(p_{ij})=\ln\{p_{ij}/(1-p_{ij})\}$, in (bio)statistics and as the probit, $h(p_{ij})=\Phi^{-1}(p_{ij})$, in econometrics, where $p_{ij}=\Pr(y_{ij}=1|y_{i-1,j},\mathbf{z}_j,\mathbf{x}_{ij},\zeta_j)$ and $\Phi(\cdot)$ is the standard normal cumulative density function.

Model (2) can alternatively be written by using a latent response formulation, where a linear model for latent responses y_{ij}^*,

$$y_{ij}^*=\mathbf{z}_j'\boldsymbol{\gamma}_z+\mathbf{x}_{ij}'\boldsymbol{\gamma}_x+\beta y_{i-1,j}+\zeta_j+\varepsilon_{ij}, \tag{3}$$

is combined with a threshold model connecting the observed responses to the latent responses,

$$y_{ij}=\begin{cases}1 & \text{if } y_{ij}^*>0,\\ 0 & \text{otherwise.}\end{cases}$$

The level 1 error ε_{ij} is assumed to have zero expectation, variance θ, and to be independent of $y_{i-1,j},\ldots,y_{S_j,j}$, $\mathbf{z}_j$, $\mathbf{x}_j^{\text{all}}$ and ζ_j. Defining the 'total error' for the latent responses as $v_{ij}=\zeta_j+\varepsilon_{ij}$, model (3) implies that the total errors for different occasions have correlations $\psi/(\psi+\theta)$. For logit models, ε_{ij} has a standard logistic distribution with $\theta=\pi^2/3$ and, for probit models, ε_{ij} is standard normal with $\theta=1$.

Model (2), which is assumed to be the data-generating mechanism until Section 5, is shown graphically in Fig. 1(a). Here the circle represents an unobserved or latent variable and the squares represent observed variables. The longer arrows represent logit or probit regressions, the short arrows represent Bernoulli variability and the curved double-headed arrows represent correlations. To avoid clutter, we have not included a time constant covariate z_j in any of the diagrams in this paper (such a variable would have arrows pointing to each of the responses with coefficient γ_z and have double-headed arrows connecting it to all time-varying covariates x_{ij}). For the same reason, the subject index j is omitted from all variables. Fig. 1 shows the observed process at four occasions ($i=0,1,2,3$). The process has been on going since $S_j<0$, as indicated

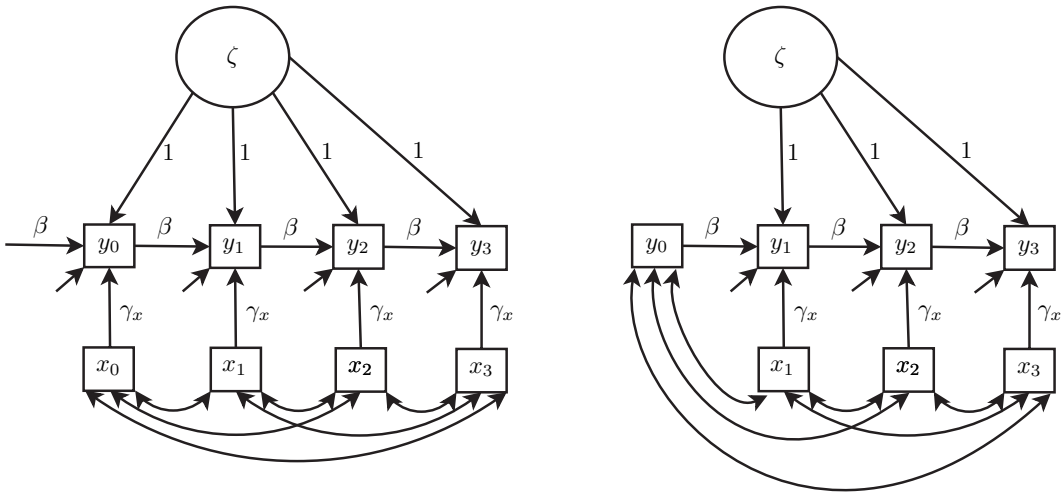

Fig. 1. (a) Data-generating mechanism and (b) naive model

by the arrow pointing to y_{0j} from the preceding (presample) response. The presample responses and their dependence on ζ_j and the presample covariates are not shown.

4. Initial conditions problem and solutions

If the process were observed from occasion S_j and there were no missing data, the likelihood contribution for subject j would be

$$\Pr(\mathbf{y}_j^{\text{all}}|\mathbf{z}_j, \mathbf{x}_j^{\text{all}}) = \int \Pr(y_{S_j j}|\mathbf{z}_j, \mathbf{x}_{S_j j}, \zeta_j) \left\{ \prod_{i=S_j+1}^{T-1} \Pr(y_{ij}|y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j) \right\} \phi(\zeta_j)\, \mathrm{d}\zeta_j,$$

where $\phi(\zeta_j)$ is a normal density with expectation 0 and variance ψ. The corresponding likelihood for a sample of N subjects becomes

$$\mathcal{L}_{\text{all}} = \prod_{j=1}^{N} \Pr(\mathbf{y}_j^{\text{all}}|\mathbf{z}_j, \mathbf{x}_j^{\text{all}}).$$

4.1. The initial conditions problem

In practice, we usually have on-going process data where observation begins at a later occasion $i=0$ than the start of the process at $i=S_j<0$ and this precludes inference based on $\mathcal{L}_{\text{all}}$. Here we let the data be observed at occasions $i=0,\ldots,T-1$, but we shall also discuss how missing data should be handled.

Estimating the dynamic/transition model with a random intercept (2) from on-going process data appears straightforward: just analyse the observed responses and include the lagged response $y_{i-1,j}$ as an additional covariate in a standard logit or probit random-intercept model. This means that only the responses $\mathbf{y}_j^{+} = (y_{1j},\ldots,y_{T-1,j})'$ after the initial response y_{0j} are modelled since the lagged response $y_{-1,j}$ is missing for y_{0j}. This naive approach assumes that ζ_j and y_{0j} are independent and the corresponding model is shown as a path diagram for four occasions in Fig. 1(b). (The correlations between the variables x_{1j}, x_{2j}, x_{3j} and y_{0j} that enter

the model only as covariates are shown because they are not assumed to be 0, but they are not explicitly modelled.)

The likelihood contribution for the naive model is

$$\Pr(\mathbf{y}_j^+|\mathbf{z}_j, \mathbf{x}_j^+, y_{0j}) = \int \left\{ \prod_{i=1}^{T-1} \Pr(y_{ij}|y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j) \right\} \phi(\zeta_j)\, \mathrm{d}\zeta_j,$$

where $\mathbf{x}_j^+ = (\mathbf{x}_{1j}, \ldots, \mathbf{x}_{T-1,j})'$. The corresponding likelihood

$$\mathcal{L}_{\text{naive}} = \prod_{j=1}^{N} \Pr(\mathbf{y}_j^+|\mathbf{z}_j, \mathbf{x}_j^+, y_{0j})$$

is then maximized to estimate β, γ_z, γ_x and ψ. Unfortunately, this estimator is inconsistent because the endogenous initial response is treated as exogenous by assuming that the conditional density of the random intercept given y_{0j} is the same as the marginal density $\phi(\zeta_j)$, giving rise to the initial conditions problem.

Endogeneity arises because the initial response y_{0j} is affected by ζ_j and the presample response $y_{-1,j}$ which is missing. For example, if wheezing at occasion 1 is associated with initial wheezing at occasion 0 because of both unobserved heterogeneity and positive state dependence, but the model allows for only the latter, then the coefficient β of the lagged response, which is assumed to be constant over time, will tend to be overestimated. Consequently, ψ will be underestimated and the coefficients of covariates that correlate with $y_{i-1,j}$, given the other covariates, will be underestimated (in absolute value) as a consequence of overcontrolling for $y_{i-1,j}$.

The initial conditions problem is not a major problem for long panel data with a large number of occasions or panel waves, since the effect of misspecification for the initial response is then swamped by the large number of responses that are correctly modelled. Indeed, consistent estimators are obtained from $\mathcal{L}_{\text{naive}}$ as $T \to \infty$ and $N \to \infty$ (e.g. Hsiao (2002)). However, naive modelling is prone to produce severely biased estimators for short panels as shown in several Monte Carlo studies (e.g. Aitkin and Alfo (1998), Fotouhi (2005), Arulampalam and Stewart (2009) and Akay (2012)).

Two principal approaches have been proposed to address the initial conditions problem in dynamic/transition random-intercept models for binary panel data: using joint models that treat y_{0j} as a response variable and using conditional models that condition on y_{0j}, taking into account the dependence between ζ_j and y_{0j}. As we shall see, the implied models are intractable in practice and approximate models must be used instead that we shall refer to as 'working models'.

Since the working models per construction are misspecified generalized linear mixed models, consistent estimation is unlikely. The aim is therefore merely to produce almost consistent estimators as $N \to \infty$ for fixed T. By 'almost consistent' we mean that the probability limit of an estimator is sufficiently close to the parameters of interest for practical purposes. Because of the misspecification, interval estimation should be based on robust standard errors from a sandwich estimator (e.g. White (1982)) instead of model-based standard errors. Rather remarkably, we are not aware of any previous work using robust standard errors in this context.

4.2. Joint working models

A joint model is specified for all observed responses $\mathbf{y}_j$, comprising the initial response y_{0j} and the subsequent responses $\mathbf{y}_j^+$, as follows:

$$\Pr(\mathbf{y}_j|\mathbf{z}_j, \mathbf{x}_j) = \int \Pr(y_{0j}|\mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j) \left\{ \prod_{i=1}^{T-1} \Pr(y_{ij}|y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j) \right\} \phi(\zeta_j)\, \mathrm{d}\zeta_j, \tag{4}$$

where $\mathbf{x}_j = (\mathbf{x}_{0j}, \ldots, \mathbf{x}_{T-1,j})'$. The corresponding likelihood function is

$$\mathcal{L}_{\text{joint}} = \prod_{j=1}^{N} \Pr(\mathbf{y}_j|\mathbf{z}_j, \mathbf{x}_j).$$

Unlike the probabilities $\Pr(y_{ij}|y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j)$ for the responses from occasion 1 onwards, the probability $\Pr(y_{0j}|\mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j)$ of the initial response is not conditional on the previous response because it is missing. The initial response probability can in principle be obtained by marginalizing the joint distribution of the initial response and all presample responses $(y_{S_j,j}, \ldots, y_{-1,j}, y_{0j})$, given the corresponding covariates and the random intercept, over all presample responses $(y_{S_j,j}, \ldots, y_{-1,j})$ and presample covariates. Such marginalization is feasible in linear models (e.g. Bollen and Curran (2004)), but only if we know exactly when the process started and the value of the initial response. If the process is believed to have started long before observation began and $|\beta| < 1$, a good approximation can be obtained by considering just a few previous occasions. However, for binary data, Heckman (1981a) noted that the required marginalization is 'somewhat computationally forbidding' even for a probit model without covariates.

4.2.1. Types of joint working models

The basic idea of Heckman (1981a) is to consider an approximation of the marginalized (or 'reduced form') distribution $\Pr(y_{0j}|\mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j)$. The link function for the initial response y_{0j} is taken to be that specified in model (2), which is an approximation since the link function for the marginalized probability is generally different from the original link. A model is then specified for the initial response that is different from that for the subsequent responses with a separate set of parameters.

4.2.1.1. The Heckman model. Following Heckman (1981a), we write his probit model in latent response form:

$$y^*_{ij} = \mathbf{z}'_j\boldsymbol{\gamma}_z + \mathbf{x}'_{ij}\boldsymbol{\gamma}_x + \beta y_{i-1,j} + \zeta_j + \varepsilon_{ij}, \qquad i = 1, \ldots, T-1,$$
$$y^*_{0j} \approx \mathbf{z}'_j\mathbf{g}_z + \mathbf{x}'_{0j}\mathbf{g}_x + \varepsilon_{0j}.$$

We see that Heckman's model for the subsequent responses is simply a probit version of model (3), with correlations $\psi/(\psi+1)$ for the total errors $v_{ij} = \zeta_j + \varepsilon_{ij}$. In the equation for the initial response, the coefficients for the covariates are allowed to differ from those for the subsequent responses. To obtain a good approximation, Heckman suggested that the model for the initial response could also contain general functions (such as polynomials) of the current covariates as well as presample time-varying covariates if they happen to be available. Importantly, Heckman allowed the error term ε_{0j} for the initial response to be 'freely correlated' with the v_{ij} $(i > 0)$, interpreted to mean that there is a free correlation ρ_{0i} between ε_{0j} and each v_{ij} (e.g. Hyslop (1999)).

A challenge of the Heckman (1981a) approach is that T-dimensional integration is required to obtain $\Pr(\mathbf{y}_j|\mathbf{x}_j, \mathbf{z}_j)$, since the probability cannot in general be simplified as in model (4). The likelihood $\mathcal{L}_{\text{joint}}$ can still be maximized by using for instance simulated maximum likelihood (e.g. Train (2009)), but the standard approach is to consider alternative models with one or two random effects that restrict the correlation structure for the latent responses (e.g. Rabe-Hesketh and Skrondal (2001)) and to perform the required lower dimensional integration in model (4) by some form of Gaussian quadrature.

4.2.1.2. Factor models. One alternative to Heckman's model is a one-factor model or item response model for binary responses with occasion-specific factor loadings or discrimination parameters λ_i (e.g. Bock and Lieberman (1970)):

$$\Pr(y_{ij}=1|y_{i-1,j},\mathbf{z}_j,\mathbf{x}_{ij},\zeta_j)=h^{-1}(\mathbf{z}_j'\boldsymbol{\gamma}_z+\mathbf{x}_{ij}'\boldsymbol{\gamma}_x+\beta y_{i-1,j}+\lambda_i\zeta_j), \qquad i=1,\ldots,T-1,$$
$$\Pr(y_{0j}=1|\mathbf{z}_j,\mathbf{x}_{0j},\zeta_j)\approx h^{-1}(\mathbf{z}_j'\mathbf{g}_z+\mathbf{x}_{0j}'\mathbf{g}_x+\lambda_0\zeta_j), \tag{5}$$

where $\zeta_j\sim N(0,1)$. The corresponding latent response formulation is

$$y_{ij}^*=\mathbf{z}_j'\boldsymbol{\gamma}_z+\mathbf{x}_{ij}'\boldsymbol{\gamma}_x+\beta y_{i-1,j}+\lambda_i\zeta_j+\varepsilon_{ij}, \qquad i=1,\ldots,T-1,$$
$$y_{0j}^*\approx \mathbf{z}_j'\mathbf{g}_z+\mathbf{x}_{0j}'\mathbf{g}_x+\lambda_0\zeta_j+\varepsilon_{0j},$$

with the common factor $\zeta_j\sim N(0,1)$ assumed to be independent of the unique factors ε_{ij}, which are themselves independent across occasions $i=0,\ldots,T-1$. The total errors $\lambda_i\zeta_j+\varepsilon_{ij}$ have covariances $\lambda_i\lambda_{i'}$ and variances $\lambda_i^2+\theta$.

An advantage of this model is that the number of latent variables for subject j is reduced from T to 1, leading to dramatic computational savings when T is not small. Arulampalam and Stewart (2009) described the one-factor probit model as a simplified implementation of the Heckman (1981a) model, but the model generally implies a more restrictive covariance structure for the total errors and is not equivalent to Heckman's model unless $T=3$ (the model is not identified for $T=2$).

Since model (2) is assumed to be the data-generating mechanism, there is no need to allow for different correlations of the total errors v_{ij} and $v_{i'j}$ among subsequent responses $i,i'>0$. It is therefore natural to use a restricted version of the one-factor model (5) that has loading λ_0 for the initial response and $\lambda_i=1$ for the subsequent responses:

$$\Pr(y_{ij}=1|y_{i-1,j},\mathbf{z}_j,\mathbf{x}_{ij},\zeta_j)=h^{-1}(\mathbf{z}_j'\boldsymbol{\gamma}_z+\mathbf{x}_{ij}'\boldsymbol{\gamma}_x+\beta y_{i-1,j}+\zeta_j), \qquad i=1,\ldots,T-1,$$
$$\Pr(y_{0j}=1|\mathbf{z}_j,\mathbf{x}_{0j},\zeta_j)\approx h^{-1}(\mathbf{z}_j'\mathbf{g}_z+\mathbf{x}_{0j}'\mathbf{g}_x+\lambda_0\zeta_j), \tag{6}$$

where $\zeta_j\sim N(0,\psi)$. Such a model with a logit link was proposed by Aitkin and Alfo (2003), who also suggested the use of non-parametric maximum likelihood estimation to leave the random-intercept distribution unspecified (e.g. Laird (1978)). The structure of the restricted one-factor model (6) is shown in Fig. 2(a). The model is attractive because it is parsimonious yet seems tailor made for the desired flexibility.

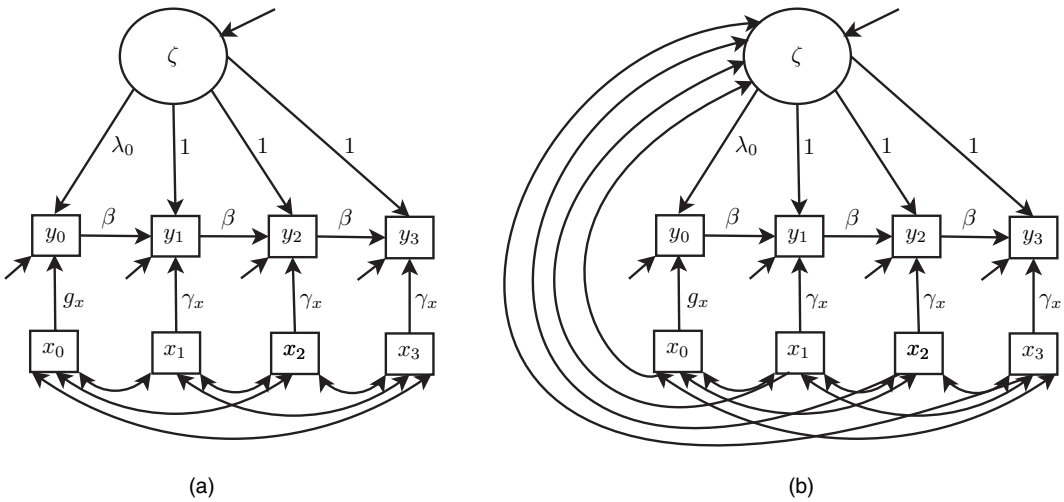

Fig. 2. Joint working model (restricted one-factor model for all responses): (a) exogenous covariate; (b) endogenous covariate

The factor working models are not generalized linear mixed models because of the model parameters λ_i, but they are special cases of generalized linear latent and mixed models (Rabe-Hesketh *et al.*, 2004a) and can be estimated by maximum likelihood with adaptive quadrature using the `gllamm` software (Rabe-Hesketh *et al.*, 2005). `gllamm` can also be used for non-parametric maximum likelihood estimation (Rabe-Hesketh *et al.*, 2003) and provides robust standard errors (Rabe-Hesketh and Skrondal, 2006). Multi-dimensional factor models could also be used for $T > 3$ to mimic the correlation structure that was used by Heckman (1981a); see the appendix of Heckman (1981c).

Arulampalam and Stewart (2009) and Akay (2012) performed simulations to investigate the finite sample performance of factor working models with probit links (referred to as 'Heckman' in both papers). Following Heckman (1981b), both Arulampalam and Stewart (2009) and Akay (2012) considered a Nerlove process (Nerlove (1971), page 367) for $\mathbf{x}_j^{\text{all}}$, and we now highlight some simulation results for this process. Arulampalam and Stewart (2009) used the unrestricted factor model (5) and found substantial bias for β when $T = 3$ and $N = 200$ (estimated as -13% with $\beta = 0.5$) but acceptable bias (less than 3%) when $T > 3$. Akay (2012) used the restricted factor model (6) and obtained better results for $T = 3$ and $N = 200$, namely an estimated bias of 5% (also with $\beta = 0.5$), that decreased to 4% for $T = 4$ and to 1% for $T > 4$. In our experience, convergence problems frequently occur when attempting to fit factor working models with $T = 3$.

4.2.1.3. Models with two correlated random intercepts. Orme (2001) and Fotouhi (2005) used models with two correlated random intercepts, ζ_{0j} for the initial response and ζ_j for all other responses:

$$\Pr(y_{ij} = 1 | y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j) = h^{-1}(\mathbf{z}_j'\boldsymbol{\gamma}_z + \mathbf{x}_{ij}'\boldsymbol{\gamma}_x + \beta y_{i-1,j} + \sigma\zeta_j), \qquad i = 1, \ldots, T-1,$$
$$\Pr(y_{0j} = 1 | \mathbf{z}_j, \mathbf{x}_{0j}, \zeta_{0j}) \approx h^{-1}(\mathbf{z}_j'\mathbf{g}_z + \mathbf{x}_{0j}'\mathbf{g}_x + \zeta_{0j}).$$

In Orme (2001), σ is a free parameter and $(\zeta_{0j}, \zeta_j)'$ have a bivariate standard normal distribution with correlation ρ_{o}. In Fotouhi (2005), $\sigma = 1$ and $(\zeta_{0j}, \zeta_j)'$ is bivariate normal with variances set equal (denoted τ^2) and with correlation ρ_{f}. Unfortunately, it turns out that both formulations imply implausible restrictions for the correlations between the total errors $v_{0j} = \zeta_{0i} + \varepsilon_{0j}$ and $v_{ij} = \sigma\zeta_j + \varepsilon_{ij}$ in the latent response formulation. With Orme's parameterization,

$$\text{corr}(v_{0j}, v_{ij}) = \rho_{\text{o}}\sqrt{\left\{\frac{\sigma^2}{2(\sigma^2+1)}\right\}} = \rho_{\text{o}}\sqrt{\left\{\frac{\text{corr}(v_{ij}, v_{i'j})}{2}\right\}}, \qquad i > 0, \quad i' \neq i.$$

It follows that, in general,

$$\text{corr}(v_{0j}, v_{ij}) < 2^{-1/2}$$

and, if $\text{corr}(v_{ij}, v_{i'j}) > \frac{1}{2}$, that

$$\text{corr}(v_{0j}, v_{ij}) < \text{corr}(v_{ij}, v_{i'j}).$$

Arulampalam and Stewart (2009) incorrectly described Orme's model as having a bivariate normal distribution for ζ_{0j} and ζ_{1j} with unstructured covariance matrix, but such a model is not identified. Fotouhi's (2005) parameterization implies that

$$\text{corr}(v_{0j}, v_{ij}) = \frac{\rho_{\text{f}}\tau^2}{\sqrt{(\tau^2+1)}} < \frac{\tau^2}{\sqrt{(\tau^2+1)}} = \text{corr}(v_{ij}, v_{i'j}),$$

which is a restriction that is implausible.

Because of the undesired restrictions that were uncovered above, we recommend against using joint working models with two correlated random intercepts. A further disadvantage, compared with using one-factor models, is that the dimension of integration is doubled from 1 to 2.

4.2.1.4. Models with a single random intercept. The most restrictive model that has been proposed includes just a common random intercept $\zeta_j \sim N(0, \psi)$, so $\lambda_0 = \lambda_i = 1$ in model (5). This model, which was used by Crouchley and Davies (2001), is easy to estimate but may provide a poor approximation.

4.2.1.5. Recommendation. We recommend use of the one-factor model (6) of Aitkin and Alfo (2003) which has one free factor loading for the initial response. The model appropriately allows for different parameters for the initial response, requires only one random effect and does not impose implausible restrictions for the correlations between the latent responses given the covariates.

When there are missing data, it is possible to analyse the data for all occasions i for which y_{ij} and $\mathbf{x}_{ij}$ are not missing for a subject. The easiest approach is to treat each occasion that follows an occasion with missing data as an 'initial' occasion and to assume that the model in the second line of expression (6) holds for all 'initial' responses. For instance, in Table 1, the initial occasions are 0 for the first pattern (1111), 1 for the second pattern (·111) and 0 and 3 for the fourth pattern (11·1). If the process started long before observation began, it is reasonable to assume that the same model, with the same parameter values, holds for all initial responses. If there are sufficient data, it is also possible to allow the parameters of the model for the initial responses to depend on the occasion number i. Another way to relax the assumption is to analyse only subjects with data at occasion 0 and to discard all data following an occasion with missing data. We expect these approaches to work well if data are missing at random.

4.3. Conditional working models

A conditional model is specified for the subsequent responses, given the initial response and the covariates $\mathbf{z}_j$ and $\mathbf{x}_j$,

$$\Pr(\mathbf{y}_j^+ | y_{0j}, \mathbf{z}_j, \mathbf{x}_j) = \int \left\{ \prod_{i=1}^{T-1} \Pr(y_{ij} | y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j) \right\} g(\zeta_j | y_{0j}, \mathbf{z}_j, \mathbf{x}_j) \, \mathrm{d}\zeta_j, \tag{7}$$

and inferences are based on the likelihood function

$$\mathcal{L}_{\mathrm{cond}} = \prod_{j=1}^{N} \Pr(\mathbf{y}_j^+ | y_{0j}, \mathbf{z}_j, \mathbf{x}_j).$$

The term $g(\zeta_j | y_{0j}, \mathbf{z}_j, \mathbf{x}_j)$ in model (7) represents the conditional distribution of the random intercept given the initial response and the covariates. The challenge in using the conditional modelling approach is that the conditional distribution $g(\zeta_j | y_{0j}, \mathbf{z}_j, \mathbf{x}_j)$ is different from the marginal random-intercept distribution $\phi(\zeta_j)$ that is used in the naive approach. The means and variances of the conditional distribution vary according to the values of the conditioning set in contrast to the marginal random-intercept distribution which has zero mean and is homoscedastic.

The required conditional density is

$$g(\zeta_j | y_{0j}, \mathbf{z}_j, \mathbf{x}_j) \propto \phi(\zeta_j) \Pr(y_{0j} | \mathbf{z}_j, \mathbf{x}_j, \zeta_j),$$

where

$$\Pr(y_{0j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j) = \sum_{y_{-1,j} \in \{0,1\}} \Pr(y_{0j}|y_{-1,j}, \mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j) \Pr(y_{-1,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j)$$

is a two-component mixture of conditional probabilities $\Pr(y_{0j}|y_{-1,j}, \mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j)$ that depend on $\mathbf{z}_j$ and $\mathbf{x}_{0j}$. If the process started at $S_j < -1$, then $\Pr(y_{-1,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j)$ is a two-component mixture of probabilities $\Pr(y_{-1,j}|y_{-2,j}, \mathbf{z}_j, \mathbf{x}_j, \zeta_j)$ with weights $\Pr(y_{-2,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j)$. However, $\Pr(y_{-1,j}|y_{-2,j}, \mathbf{z}_j, \mathbf{x}_j, \zeta_j)$ does not have a simple form because $\mathbf{x}_j$ does not include $\mathbf{x}_{-1,j}$, so we must integrate $\Pr(y_{-1,j}|y_{-2,j}, \mathbf{z}_j, \mathbf{x}_{-1,j}, \zeta_j)$ over the conditional density $f(\mathbf{x}_{-1,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j)$ and similarly for $\Pr(y_{-2,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j)$ (which is also a two-component mixture if $S_j < -2$, and so on). If the $\mathbf{x}_j^{\text{all}}$ are not independent over time, $f(\mathbf{x}_{-1,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j) \neq f(\mathbf{x}_{-1,j}|\mathbf{z}_j, \zeta_j)$, so $g(\zeta_j|y_{0j}, \mathbf{z}_j, \mathbf{x}_j)$ depends on $\mathbf{x}_j$ via the presample probabilities $\Pr(y_{-1,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j)$. In addition, $g(\zeta_j|y_{0j}, \mathbf{z}_j, \mathbf{x}_j)$ depends directly on $\mathbf{z}_j$ and $\mathbf{x}_{0j}$ via $\Pr(y_{0j}|y_{-1,j}, \mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j)$.

To derive the conditional densities and to show their dependence on $\mathbf{x}_{0j}$, we let the process start at $S_j = -1$ with $\Pr(y_{-1,j}|\mathbf{z}_j, \mathbf{x}_j, \zeta_j) = 0.5$. In this case

$$g(\zeta_j|y_{0j}, \mathbf{z}_j, \mathbf{x}_j) \propto \frac{1}{2} \sum_{y_{-1,j} \in \{0,1\}} \phi(\zeta_j) \Pr(y_{0j}|y_{-1,j}, \mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j),$$

where the normalizing constant can be obtained by integrating the above expression over ζ_j by using adaptive quadrature (Rabe-Hesketh *et al.*, 2005). Fig. 3 shows the conditional densities

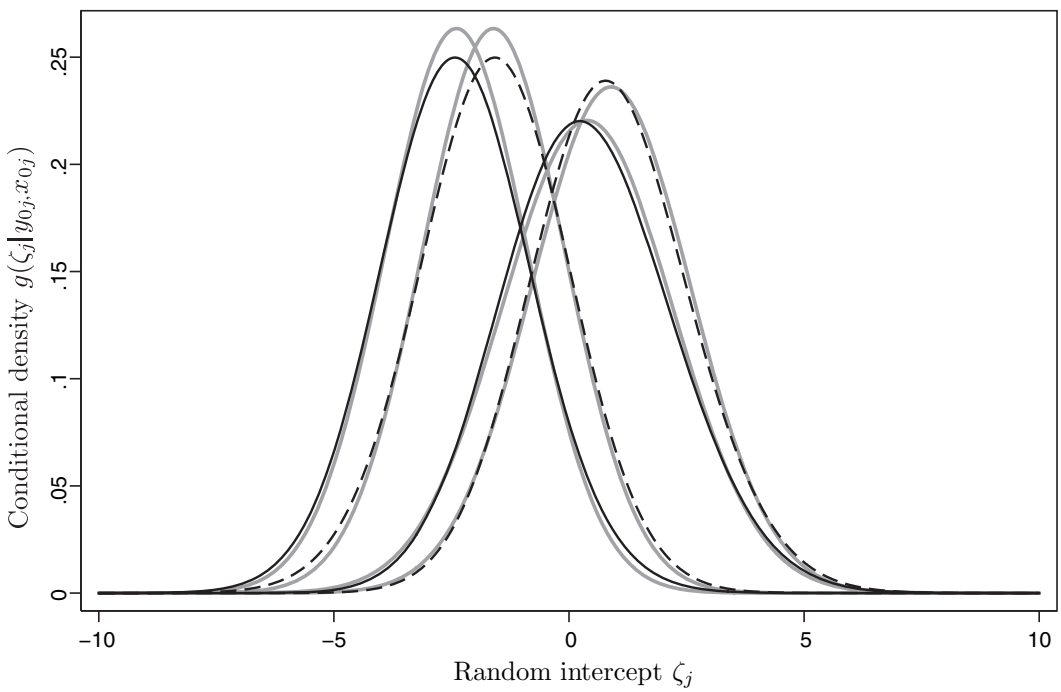

Fig. 3. Conditional densities $g(\zeta_j|y_{0j}, x_{0j})$ for a dynamic/transition random-intercept logit model with one time-varying binary covariate x_{ij} and parameters $\gamma_x = 2$, $\beta = 2$, $\psi = 4$, and no intercept, where the model for y_{0j} is the same as that for y_{ij} with $i > 0$ and $\Pr(y_{-1,j}|x_j, \zeta_j) = 0.5$ (curves on the left are for $y_{0j} = 0$ and curves on the right are for $y_{0j} = 1$): ———, true conditional density, $x_{0j} = 1$; — — —, true conditional density, $x_{0j} = 0$; ———, approximating normal densities with the same means and variances

(black curves) and normal approximations (grey curves) for a logit model with one time-varying binary covariate x_{ij} and parameters $\gamma_x = 2$, $\beta = 2$ and $\psi = 4$, and no intercept (using 50-point adaptive quadrature for the integral). The conditional densities depend on whether $x_{0j} = 1$ (full curves) or $x_{0j} = 0$ (broken curves) and on whether $y_{0j} = 0$ (curves on the left) or $y_{0j} = 1$ (curves on the right). We see that the conditional means depend strongly on x_{0j} and y_{0j} (ranging from -2.39 to 0.89), but the conditional variances are not very different (ranging from 2.29 to 2.85). The conditional densities are well approximated by the normal densities (in grey) with the same means and variances. Similar patterns were found with other sets of parameter values.

It is important to note that conditioning on the initial response y_{0j} implies that $g(\zeta_j|y_{0j}, \mathbf{z}_j, \mathbf{x}_j) \neq g(\zeta_j|y_{0j})$, even under the assumption of level 2 exogeneity, $g(\zeta_j|\mathbf{z}_j, \mathbf{x}_j) = g(\zeta_j)$.

4.3.1. Types of conditional working models

Davies and colleagues (e.g. Davies and Crouchley (1985) and Davies and Pickles (1986)), Aitkin and Alfo (1998) and Wooldridge (2005) suggested (apparently independently) approximating the conditional distribution of the random intercept given y_{0j} and the covariates by a working model.

4.3.1.1. Wooldridge's solution. Wooldridge (2005) suggested an auxiliary model for the conditional random-intercept distribution in which the mean depends on the initial response and covariates:

$$\zeta_j \approx \delta_y y_{0j} + \mathbf{z}_j'\boldsymbol{\delta}_z + \mathbf{x}_j^{+\prime}\boldsymbol{\delta}_{x+} + u_j, \tag{8}$$

where $\mathbf{x}_j^+ = (\mathbf{x}_{1j}, \ldots, \mathbf{x}_{T-1,})'$ as before. Here $u_j \sim N(0, \omega)$ is independent of y_{0j}, $\mathbf{z}_j$ and $\mathbf{x}_j^+$. Substituting approximation (8) in the latent response version of the dynamic random-intercept model (3), we obtain the following working model for $i = 1, \ldots, T-1$:

$$\Pr(y_{ij} = 1|y_{i-1,j}, y_{0j}, \mathbf{z}_j, \mathbf{x}_j, u_j) \approx h^{-1}\{\mathbf{z}_j'(\boldsymbol{\gamma}_z + \boldsymbol{\delta}_z) + \mathbf{x}_{ij}'\boldsymbol{\gamma}_x + \mathbf{x}_j^{+\prime}\boldsymbol{\delta}_{x+} + \beta y_{i-1,j} + \delta_y y_{0j} + u_j\}. \tag{9}$$

This reduced form model is a standard random-intercept model with a larger set of covariates and is therefore easy to estimate by using standard software for random-intercept logit or probit modelling.

Wooldridge (2005) showed that consistent estimators of $\boldsymbol{\gamma}_x$ and β are obtained if the auxiliary model is correct. However, the model is known to be only an approximation of the correct model, so it is assumed that estimation will be almost consistent if the auxiliary model is almost correct. Arulampalam and Stewart (2009) found that Wooldridge's solution produced smaller finite sample bias for β when $N = 200$ and $T = 3$ than the joint approach with an unrestricted factor working model. For $T > 3$, both estimators produced similar, insubstantial, bias (see also Rabe-Hesketh and Skrondal (2013)).

4.3.1.2. Constrained Wooldridge solution. Wooldridge's solution requires a separate parameter vector for the time-varying covariates at each occasion, i.e. $\mathbf{x}_{1j}$, $\mathbf{x}_{2j}$, up to $\mathbf{x}_{T-1,j}$. The model therefore becomes large if the number of occasions and/or the number of time-varying covariates is not small. Also, when there are missing data, Wooldridge's solution requires complete-case analysis (or listwise deletion) because a complete set of covariate values is required across all occasions. Perhaps for this reason, Michaud and Tatsiramos (2011) and others replace $\mathbf{x}_j^+$ by the subject means $\bar{\mathbf{x}}_{\cdot j}$ of the time-varying covariates, where the mean includes the initial values $\mathbf{x}_{0j}$ and presumably contributions from all occasions for which data are available. The auxiliary model then becomes

$$\zeta_j \approx \delta_y y_{0j} + \mathbf{z}_j'\boldsymbol{\delta}_z + \bar{\mathbf{x}}_{\cdot j}\boldsymbol{\delta}_{\bar{x}} + u_j. \tag{10}$$

Rabe-Hesketh and Skrondal (2013) show that this approach is problematic because the initial values of the time-varying covariates have an additional effect on ζ_j, after allowing for the effect of $\mathbf{x}_j$ via the presample response variable $y_{-1,j}$. It is therefore unreasonable to constrain the coefficients of $\mathbf{x}_{0j}$ to be the same as the coefficients of $\mathbf{x}_{ij}$, for $i > 0$. Akay (2012) found that the constrained Wooldridge solution in approximation (10) leads to severe finite sample bias for β when $N = 200$ for several processes for $\mathbf{x}_j^{\text{all}}$ (except the Nerlove process) unless T is large (the estimated bias for $T = 4$ was between 9% and 15% for $\beta = 0.5$, depending on the process for $\mathbf{x}_j^{\text{all}}$).

4.3.1.3. Models without covariates. Aitkin and Alfo (1998) approximated the conditional distribution $g(\zeta_j|y_{0j}, \mathbf{z}_j, \mathbf{x}_j)$ by a normal distribution:

$$\zeta_j \approx \delta_y y_{0j} + u_j,$$

where $u_j \sim N(0, \omega)$ is independent of y_{0j} (see also Mandel and Betensky (2008)). The reduced form model is a random-intercept model for binary responses that simply includes the initial response y_{0j} as an additional covariate. The model is a special case of a model considered by Fotouhi (2005) which allows the variance of the conditional random-effects distribution to depend on the initial response:

$$\zeta_j \approx \delta_y y_{0j} + u_{1j} y_{0j} + u_{0j}(1 - y_{0j}),$$

where $u_{1j} \sim N(0, \omega_1)$ is independent of $u_{0j} \sim N(0, \omega_0)$, and both u_{1j} and u_{0j} are independent of y_{0j}.

Unfortunately, it is not recognized that the conditional distribution of the random intercept also depends on the covariates $\mathbf{z}_j$ and $\mathbf{x}_j$ in these models as shown in Fig. 3 for $S_j = -1$. Davies and Crouchley (1985) and Aitkin and Alfo (1998) advocated the use of non-parametric maximum likelihood estimation for conditional working models. However, Aitkin and Alfo assumed the same non-parametric distribution for ζ_j regardless of the values taken by y_{0j} and the covariates, apart from a location shift δ_y determined by y_{0j}. Davies and Crouchley allowed for different non-parametric distributions according to the values taken by y_{0j}, but not according to the covariate values.

4.3.1.4. Recommendation. We suggest approximating the conditional distribution $g(\zeta_j|y_{0j}, \mathbf{z}_j, \mathbf{x}_j)$ with a normal distribution, using the auxiliary model

$$\zeta_j \approx \delta_y y_{0j} + \mathbf{z}_j'\boldsymbol{\delta}_z + \mathbf{x}_{0j}'\boldsymbol{\delta}_{x_0} + \bar{\mathbf{x}}_{\cdot j}'\boldsymbol{\delta}_{\bar{x}} + u_j, \tag{11}$$

where $u_j \sim N(0, \omega)$ is independent of y_{0j}, $\mathbf{z}_j$, $\mathbf{x}_{0j}$ and $\bar{\mathbf{x}}_{\cdot j}$.

This model has the advantage that it takes into account dependence of the distribution on the covariates and can be used without requiring complete-case analysis when there are missing data. In contrast with the constrained Wooldridge solution (10), which implicitly constrains the coefficients of the time-varying covariates $\mathbf{x}_{ij}$ at all occasions i to be the same, our recommended model allows the coefficients of the initial values $\mathbf{x}_{0j}$ of the time-varying covariates to be different from the coefficients for subsequent occasions. This model was proposed by Rabe-Hesketh and Skrondal (2013), who showed that the substantial finite sample bias that was found by Akay (2012) for the constrained Wooldridge solution becomes negligible (and similar to the bias for Wooldridge's solution) when $\mathbf{x}_{0j}$ are included as additional covariates. The structure of the working model is shown in Fig. 4, where the short arrow pointing to ζ_j represents an additive normally distributed error. For simplicity, we have omitted the coefficients for the paths from the

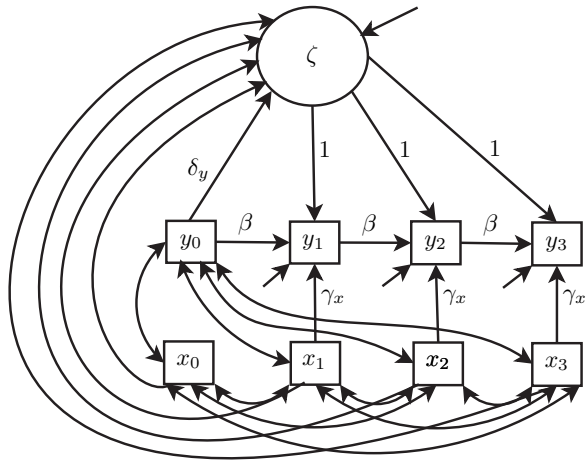

Fig. 4. Conditional working model: random-intercept model for subsequent responses, conditioning on the initial response y_{0j}, the initial value x_{0j} and mean $\bar{x}_{\cdot j}$ of the time-varying covariate x_{ij}, and the time constant covariate z_j (not shown)

x_{ij} to ζ_j. When the number of occasions is small and there are little missing data Wooldridge's solution can also be used.

We now consider approaches for handling missing data. Isolated observations (preceded and succeeded by missing data) cannot be used. However, it is possible to utilize several sequences of non-missing data for a subject (e.g. $11 \cdot 11$). In this case, the 'initial' values of the response and time-varying covariates change between sequences (for example, for $11 \cdot 11$, the initial response is y_{0j} for the first sequence and y_{3j} for the second sequence). It is possible to let the parameters of the auxiliary model differ depending on which occasion is the initial occasion. Another possibility is to analyse only those contiguous sequences of non-missing data that start at occasion 0, i.e. to discard subjects with patterns such as $\cdot\cdot 11$. In these somewhat *ad hoc* approaches, the missing values of $\mathbf{x}_{ij}$ are implicity imputed by $\bar{\mathbf{x}}_{ij}$ and the responses are assumed to be missing at random.

5. The endogenous covariates problem and solutions

In the statistical literature, the covariates $\mathbf{z}_j$ and $\mathbf{x}_j$ are usually either treated as fixed (e.g. Verbeke and Molenberghs (2000)) or implicitly assumed to be independent of the random intercept ζ_j. As pointed out by Snijders and Berkhof (2008), if treated as fixed, the random-intercept distribution is assumed not to depend on the covariates. In both statistics and econometrics, it is typically assumed that there is level 1 exogeneity in the sense that the level 1 error ε_{ij} in the latent response formulation (3) is independent of the covariates (or its distribution does not depend on the covariates if covariates are treated as fixed).

However, these assumptions may be unrealistic in observational studies. The random intercept represents the combined effect of omitted time constant covariates. If these omitted covariates, and therefore the random intercept, are associated with the covariates in the model, we have level 2 endogeneity, which we also refer to as between-subject confounding. This kind of endogeneity could be represented in Fig. 1(a) by including curves with double arrowheads connecting each of the time-varying covariates x_{ij} with the random intercept ζ_j. The level 1 errors ε_{ij} in the latent response formulation represent the combined effects of omitted time-varying covariates

and, if these are not independent of the included covariates, we have level 1 endogeneity or within-subject confounding.

We conjecture that the problem of inconsistent estimation due to level 2 endogeneity disappears as $T \to \infty$, since the 'random-effects estimator' of γ_x in this case may converge to the fixed effects estimator which is consistent as $N \to \infty$ (see, for example, Blundell and Windmeijer (1997) for the case of *linear* random-intercept models). Once again, naive modelling can produce severe inconsistency for the short panels that we focus on in this paper.

We assume in what follows that there is level 1 exogeneity, although level 1 endogeneity can in principle be addressed if credible instrumental variables are available. To relax level 2 exogeneity in standard random-intercept logit models (without lagged responses), econometricians typically use conditional maximum likelihood estimation (e.g. Chamberlain (1980)). This approach produces consistent estimators for γ_x even if there is level 2 endogeneity. No estimates are produced for γ_z and ψ, but these parameters cannot generally be consistently estimated anyway if there is endogeneity. Fixed effects approaches for dynamic logit models have been developed by Honoré and Kyriazidou (2000) and Bartolucci and Nigro (2010). The first approach has many limitations (e.g. Wooldridge (2005)) whereas the latter is more promising. Unfortunately, a conditional likelihood cannot be constructed for probit models.

In practice, level 2 endogeneity in static logit or probit random intercept models is addressed by specifying an auxiliary model where the random intercept is regressed on the time-varying covariates (e.g. Chamberlain (1980, 1984) or vice versa (e.g. Neuhaus and McCulloch (2006)). Surprisingly, there seems to be no previous reference discussing these approaches for joint working models.

5.1. Joint working models

To allow for level 2 endogeneity of the time-varying covariates $\mathbf{x}_j$ in the restricted one-factor model (6), we propose to use the auxiliary model

$$\zeta_j = \bar{\mathbf{x}}'_{\cdot j}\boldsymbol{\delta}_{\bar{x}} + u_j, \tag{12}$$

where $u_j \sim N(0, \omega)$ is independent of $\bar{\mathbf{x}}_{\cdot j}$.

For probit random-intercept models, Chamberlain (1980) suggested a variant where $\bar{\mathbf{x}}_{\cdot j}$ is replaced by $\mathbf{x}_j$ with coefficients $\boldsymbol{\delta}_x$. In linear random-intercept models, the auxiliary equation is just a device to obtain an appropriate linear predictor for consistent estimation (e.g. Mundlak (1978)). In that case, u_j need neither be normal nor homoscedastic and it is immaterial whether $\bar{\mathbf{x}}_{\cdot j}$ or $\mathbf{x}_j$ are included (Chamberlain, 1982). In contrast, for logit or probit random-intercept models, the auxiliary equation represents a proper statistical model which must be correctly specified to ensure consistency (Chamberlain, 1984). Using $\bar{\mathbf{x}}_{\cdot j}$ in place of $\mathbf{x}_j$ now restricts the correlations between the random intercept and the time-varying covariates to be constant over time, a point that appears to have been overlooked by Arulampalam and Stewart (2009). However, using the subject means $\bar{\mathbf{x}}_{\cdot j}$ is often the only viable option in practice, especially when $\mathbf{x}_{ij}$ has missing values or is not low dimensional, or T is not small. When there are missing data, we believe that the means $\bar{\mathbf{x}}_{\cdot j}$ should be based only on those occasions for which y_{ij} contributes to the analysis. The $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\cdot j}$ for occasions that contribute to the analysis are then correlated with $\mathbf{x}_{ij}$ but uncorrelated with ζ_j, and hence instrumental variables per construction.

Substituting equation (12) in equation (6), we obtain

$$\begin{aligned} \Pr(y_{ij} = 1 | y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_j, u_j) &= h^{-1}(\mathbf{z}'_j\boldsymbol{\gamma}_z + \mathbf{x}'_{ij}\boldsymbol{\gamma}_x + \bar{\mathbf{x}}'_{\cdot j}\boldsymbol{\delta}_{\bar{x}} + \beta y_{i-1,j} + u_j), \qquad i = 1, \ldots, T-1, \\ \Pr(y_{0j} = 1 | \mathbf{z}_j, \mathbf{x}_j, u_j) &\approx h^{-1}(\mathbf{z}'_j\mathbf{g}_z + \mathbf{x}'_{0j}\mathbf{g}_x + \bar{\mathbf{x}}'_{\cdot j}\lambda_0\boldsymbol{\delta}_{\bar{x}} + \lambda_0 u_j). \end{aligned} \tag{13}$$

A graphical representation of this model is given in Fig. 2(b). Note that γ_z will no longer be consistently estimated when $\mathbf{x}_j$ is endogenous and should hence not be interpreted.

Model (13) imposes the restriction that the coefficients $\lambda_0\boldsymbol{\delta}_{\bar{x}}$ of $\bar{\mathbf{x}}_{\cdot j}$ for the initial response are λ_0 times the corresponding coefficients $\boldsymbol{\delta}_{\bar{x}}$ for the subsequent responses. We can either impose this restriction in the estimation or simply ignore it by specifying a separate coefficient vector $\boldsymbol{\delta}_{\bar{x}0}$ for the initial response. Whichever approach is taken, we can then maximize the likelihood aiming to obtain almost consistent estimators of β and $\boldsymbol{\gamma}_x$.

5.2. Conditional working models

The conditional working model (11), which was proposed in Section 4.3.1 to handle the initial conditions problem, already accommodates level 2 endogeneity by including the cluster means $\bar{\mathbf{x}}_{\cdot j}$. As for the joint working models (and for the same reason), we believe that $\bar{\mathbf{x}}_{\cdot j}$ should be based only on those occasions for which y_{ij} contributes to the analysis. Hence, no extra remedies are required for conditional working models, unless one of the auxiliary models without covariates is used.

6. Analysis of children's wheezing data

6.1. Model specification

To address the research questions that were stated in Section 2, we consider a logit version of model (2) for children $j = 1, \ldots, 412$ and occasions $i = 1, 2, 3$:

$$\text{logit}\{\Pr(y_{ij} = 1 | y_{i-1,j}, \mathbf{z}_j, \mathbf{x}_{ij}, \zeta_j)\} = \gamma_{z_0} + \gamma_{z_1} z_j + \gamma_{x_1} x_{1ij} + \gamma_{x_2} x_{2i} + \beta y_{i-1,j} + \zeta_j. \tag{14}$$

Here y_{ij} is a binary response variable taking the value 1 if child j is wheezing at occasion i and 0 otherwise, z_j is an indicator variable taking the value 1 if residency was in Kingston-Harriman and 0 for Portage, x_{1ij} represents the number of cigarette packs smoked by the mother of child j at occasion i, and x_{2i} represents the age at occasion i (which is the same for all children at a given occasion).

For the joint modelling approach, we model the initial response at $i = 0$ as

$$\text{logit}\{\Pr(y_{0j} = 1 | \mathbf{z}_j, \mathbf{x}_{0j}, \zeta_j)\} = g_{z_0} + g_{z_1} z_j + g_{x_1} x_{10j} + g_{x_2} x_{20j} + \lambda_0 \zeta_j.$$

To handle level 2 endogeneity, we use the auxiliary model

$$\zeta_j = \delta_{\bar{x}_1} \bar{x}_{1\cdot j} + \delta_{\bar{x}_2} \bar{x}_{2\cdot j} + u_j.$$

All available data are analysed, specifying an identical model for all initial responses that have no observed lagged response. For this reason, age at the initial occasion, x_{20j}, is not constant and can be included in the model. The subject means $\bar{x}_{1\cdot j}$ and $\bar{x}_{2\cdot j}$ include contributions from only those occasions where the corresponding y_{ij} contributes to the analysis. The subject mean of age therefore varies between children and can be included in the auxiliary model.

For the conditional approach, we use the auxiliary model

$$\zeta_j = \delta_{y_0} y_{0j} + \delta_{x_{10}} x_{10j} + \delta_{x_{20}} x_{20j} + \delta_{\bar{x}_1} \bar{x}_{1\cdot j} + \delta_{\bar{x}_2} \bar{x}_{2\cdot j} + u_j,$$

where we note that level 2 endogeneity of x_{1ij} is accommodated. The first (and here only) contiguous sequence of at least two non-missing responses for each child is analysed. As for the joint approach, the subject means $\bar{x}_{1\cdot j}$ and $\bar{x}_{2\cdot j}$ are based on the sets of occasions for which the response variable contributes to the analysis.

6.2. Results

Table 2 gives estimates from the naive approach that ignores both the initial conditions and endogenous covariates problems, the joint working model (assuming either exogeneity or level 2 endogeneity for smoking and age) and the conditional working model.

According to the naive estimates, the longitudinal within-child dependence is wholly due to state dependence, with the odds of wheezing estimated to be as much as $\exp(2.34) = 10.38$ times as high if wheezing was experienced at the previous occasion as if it was not, and an estimated random-intercept variance of 0.00.

In contrast, the estimates for the joint working models suggest that the longitudinal dependence is due to both state dependence and unobserved heterogeneity. For the model with exogenous smoking and age, the estimated odds ratio for previous wheezing is considerably reduced to $\exp(0.89) = 2.43$, with a 95% confidence interval of (1.14, 5.17), and the estimated random-intercept variance is 3.17. The intraclass correlation of the latent responses y^*_{ij}, given the observed covariates, is estimated as $3.17/(3.17 + \pi^2/3) = 0.49$. The estimates are very similar for the joint model where level 2 endogeneity of smoking and age is accommodated. Here, testing the null hypothesis $\delta_{\bar{x}_1} = \delta_{\bar{x}_2} = 0$ can be viewed as testing the hypothesis of level 2 exogeneity. The estimates of these coefficients are minute and far from significant, suggest-

Table 2. Estimated parameters and robust standard errors for dynamic/transition random-intercept logit models for children's wheezing data: the naive model N_C that ignores initial conditions and endogenous $\mathbf{x}_{ij}$, joint working models A_J and R_J with exogenous and endogenous $\mathbf{x}_{ij}$, and the conditional working model R_C that accommodates endogenous $\mathbf{x}_{ij}$†

Parameter	*Estimate for naive model* N_C	*Estimate for joint models*		*Estimate for conditional model* R_C
		Exogenous $\mathbf{x}_{ij}$ *(model* A_J*)*	*Endogenous* $\mathbf{x}_{ij}$ *(model* R_J*)*	
Structural parameters				
β	2.34 (0.21)	0.89 (0.39)	0.86 (0.39)	0.81 (0.37)
γ_{z_0}	−1.57 (0.97)	−1.62 (1.18)	−1.66 (2.88)	−5.90 (3.96)
γ_{z_1} [City]	0.41 (0.18)	0.82 (0.33)	0.83 (0.33)	0.61 (0.30)
γ_{x_1} [Smoke]	0.01 (0.01)	0.02 (0.01)	0.01 (0.02)	−0.06 (0.03)
γ_{x_2} [Age]	−0.09 (0.10)	−0.14 (0.13)	−0.15 (0.13)	−0.21 (0.14)
ψ	0.00	3.17		
Nuisance parameters				
$g_{z_0} - \gamma_{z_0}$		−0.22 (1.89)	−0.15 (1.90)	
$g_{z_1} - \gamma_{z_1}$		−0.05 (0.42)	−0.07 (0.42)	
$g_{x_1} - \gamma_{x_1}$		0.00 (0.02)	−0.01 (0.02)	
$g_{x_2} - \gamma_{x_2}$		0.07 (0.24)	0.06 (0.25)	
$\lambda_0 - 1$		0.17 (0.46)	0.11 (0.44)	
$\delta_{y_0} - \gamma_{z_0}$				2.40 (0.56)
$\delta_{x_{10}} - \gamma_{x_1}$				−0.08 (0.04)
$\delta_{x_{20}} - \gamma_{x_2}$				−0.40 (0.50)
$\delta_{\bar{x}_1}$			0.01 (0.02)	0.16 (0.06)
$\delta_{\bar{x}_2}$			0.00 (0.33)	0.85 (0.62)
ω			3.37	2.52
Number of children	373	412	412	373
Log-likelihood	−402.5	−620.0	−619.9	−385.1

†Estimated standard errors are given in parentheses.

ing that smoking and age are level 2 exogenous. For the conditional working model, which accommodates level 2 endogeneity of smoking and age, the estimated odds ratio for previous wheezing status is similar to those for the joint working models.

We also investigated whether the extent of state dependence varies according to the covariates, but none of these interactions are significant at the 5% level. No significant relationship is found between the number of packs of cigarettes smoked by the mother and wheezing with any of the approaches that were used. This result persists when exposure to smoking is represented by a time-varying dummy variable for whether the mother smokes any cigarettes at each occasion.

Regarding the included confounders, age appears to have a negative linear relationship with the log-odds of wheezing. To assess the possibility of a non-linear relationship we also included quadratic and cubic components of age but these are not significant. As would be expected, the estimate $\hat{\gamma}_{z_1}$ from the joint model with exogenous $\mathbf{x}_{ij}$ implies that residing in highly polluted Kingston-Harriman is associated with more wheezing than residing in Portage. In the naive approach γ_{z_1} is severely underestimated because β is overestimated and z_{1j} is highly correlated with $y_{i-1,j}$ given $\mathbf{x}_{ij}$. Note that γ_{z_1} cannot be consistently estimated by the joint endogenous or conditional approaches. For parameters that can be almost consistently estimated by several approaches, such as β and γ_{x_1}, we recommend comparing the estimates as a sensitivity analysis.

Causal inferences based on these data should be made with considerable caution. We have access to only a subset of the Harvard six cities study and the study itself has many limitations. Using self-reported maternal smoking as a proxy for children's postnatal smoke exposure is problematic and using the nicotine in the child's hair as a biomarker would have been preferable (e.g. Nafstad *et al*. (1997)). *In vitro* exposure to smoking was not obtained in the study, which is unfortunate since it is more strongly associated with wheezing than postnatal exposure (e.g. Gilliland *et al*. (2001)). We also lack information on respiratory syncytial virus bronchiolitis which has been shown to be associated with wheezing (e.g. Stein *et al*. (1999)). Although we have extended previous analyses to handle unobserved between-child confounding, there could still be unobserved within-child confounding (level 1 endogeneity). We have also assumed that the missing data are missing at random. Finally, the concept of wheezing, how to measure it and its relationship to other respiratory outcomes such as asthma remain controversial in respiratory epidemiology (e.g. Dundas and McKenzie (2006)).

The children's wheezing data and a Stata 'do file' to perform the analyses presented here can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

7. Asymptotic performance

Since the aim of using working models is to produce *almost* consistent estimators, it is of interest to compare the asymptotic performance of the various estimators and to investigate how close to consistent they turn out to be in different situations.

To study asymptotic performance, we use an approach that was proposed by Rotnitzky and Wypij (1994) and also used by Heagerty and Kurland (2001). The idea is to create 'population data' in which each possible response pattern $y_{0j}, \ldots, y_{T-1,j}$ occurs with a relative frequency equal to the joint probability that is implied by the assumed model for each set of covariate values. In practice, this can be achieved by simulating data for a huge sample and making estimation feasible by using categorical covariates and collapsing the data so that there is one observation for each combination of covariate–pattern and response–pattern. The corresponding frequencies are then used to weight the log-likelihood contributions from these covariate–response patterns.

Maximum likelihood estimates for a misspecified model minimize the Kullback–Leibler

divergence, and the differences between these estimates and the parameter values represent the inconsistencies of the estimators (e.g. White (1982)). The standard errors can be viewed as asymptotic standard errors for a sample size equal to the sum N_s of the absolute frequencies used to weight the log-likelihood contributions. To obtain standard errors for a given desired sample size N_d, these standard errors are multiplied by a factor $\sqrt{(N_s/N_d)}$. The corresponding robust standard errors can be viewed as the true asymptotic sampling standard deviations which can be compared with the model-based asymptotic standard errors.

7.1. Design

Following Heckman (1981a), we consider a process that has been on going for 25 occasions $i=-25, -24, \ldots, -1$ before it is observed at occasions $i=0, \ldots, T-1$. The process starts as

$$\Pr(y_{-25,j}=1)=0.5$$

and then progresses as

$$\Pr(y_{ij}=1|z_j, x_{ij}, y_{i-1,j}, \zeta_j)=\text{logit}^{-1}(\gamma_{z_0}+\gamma_{z_1}z_j+\gamma_x x_{ij}+\beta y_{i-1,j}+\zeta_j), \qquad i=-24, \ldots, T-1.$$

The covariates z_j and x_{ij} are binary, taking values -1 and 1, with zero means, unit standard deviations, correlations $\text{corr}(x_{ij}, x_{i'j})=0.5$ for each pair of occasions $i \neq i'$ and $\text{corr}(z_j, x_{ij})=0.1$ for each occasion i. $\zeta_j \sim N(0, 4)$ and $\text{corr}(\zeta_j, x_{ij})$ is either 0 or 0.5 for each occasion i, i.e. x_{ij} is either exogenous or level 2 endogenous. The coefficients are $\gamma_{z_0}=0$, $\gamma_{z_1}=2$ and $\beta=1$.

Data were simulated for $N=10$ million subjects and $T=4, 5, 6$ observed occasions. After deleting the presample data, the remaining data were collapsed to 512 unique covariate–response patterns of z_j, x_{0j}, x_{1j}, x_{2j}, x_{3j}, y_{0j}, y_{1j}, y_{2j} and y_{3j} for $T=4$ (1024 and 2048 patterns for $T=5$ and $T=6$ respectively). The corresponding frequency weights were used to weight the log-likelihood contribution from each covariate–response pattern.

We used several versions of the joint and conditional modelling approaches to estimate the target parameters γ_x and β. For the joint modelling approach, we used our recommended model (6) when x_{ij} is assumed to be exogenous. Since λ_0 cannot be estimated in standard software, we also fitted the model with the constraint $\lambda_0=1$, as in Crouchley and Davies (2001). To allow for endogenous x_{ij}, we used model (13) which assumes that $\delta_{\bar{x}0}=\delta_{\bar{x}}\lambda_0$. We also considered three alternative versions of this model, with

(a) $\delta_{\bar{x}0}$ estimated freely (relaxing the constraint $\delta_{\bar{x}0}=\delta_{\bar{x}}\lambda_0$),
(b) $\bar{x}_{\cdot j}$ replaced by $\mathbf{x}_j$ with constraint $\boldsymbol{\delta}_{x0}=\lambda_0\boldsymbol{\delta}_x$ and
(c) $\bar{x}_{\cdot j}$ replaced by $\mathbf{x}_j$ and with $\boldsymbol{\delta}_{x0}$ estimated freely.

For the conditional modelling approaches, we used working model (11), with different replacements for x_{0j} and $\bar{x}_{\cdot j}$. To consider a case analogous to assuming exogeneity in the joint modelling approach, we used only x_{0j}, based on the assumption that the conditional distribution of ζ_j given y_{0j} and x_{0j} does not depend greatly on x_{ij} at other occasions $i>0$. As a naive alternative, we also excluded x_{0j} from the auxiliary model. We also considered four alternatives with x_{0j} and $\bar{x}_{\cdot j}$ replaced by

(a) $\mathbf{x}_j$,
(b) $\mathbf{x}_j^+=(x_{1j}, \ldots, x_{T-1,j})'$ as in Wooldridge (2005),
(c) $\bar{x}_{\cdot j}$ as in Akay (2012) and
(d) $\bar{x}_{\cdot j}^+=(x_{1j}+\ldots+x_{T-1,j})/(T-1)$.

All models were estimated by using `gllamm` (Rabe-Hesketh *et al.*, 2004b, Rabe-Hesketh and Skrondal, 2012) with 20-point adaptive quadrature.

7.2. Results

We start by discussing the results for $T=4$. When x_{ij} is exogenous, the naive approach severely overestimates the lag parameter as $\hat{\beta}=2.08$ (the true value is 1) and underestimates the random-intercept variance as $\hat{\psi}=0.86$ (the true value is 4) as expected. The coefficients of the time-varying and time invariant covariates are underestimated as $\hat{\gamma}_x=1.49$ (the true value is 2) and as $\hat{\gamma}_{z_1}=1.27$ (the true value is 2). This underestimation is likely to be due to overcontrolling for the lagged response because β is overestimated, since the lagged response is strongly associated with x_{ij} (through direct dependence on $x_{i-1,j}$ which is correlated with x_{ij}) and with z_j.

In contrast, the joint modelling approach assuming exogeneity and with a free parameter λ_0 gives estimates of these parameters within 1% of the true values when x_{ij} is exogenous. The conditional modelling approaches also produce almost consistent estimates of β and γ_x, underestimating the parameters by 3% when only x_{0j} is included in the auxiliary model (in addition to y_{0j}) and producing estimates that differ no more than 2% from the true values for the models that include both x_{0j} and $\bar{x}_{\cdot j}$ or $\mathbf{x}_j$ or $\mathbf{x}_j^+$. However, the asymptotic standard errors of $\hat{\beta}$ are up to 20% greater for the conditional approaches, whereas the standard errors for $\hat{\gamma}_x$ differ by no more than 5% between the different methods.

Graphs of the point estimates of β and γ_x for several of the estimation methods are given in Fig. 5 for $T=4$, where the horizontal lines represent the true parameter values. The x-axis labels denote the simulation conditions (exogenous and endogenous x_{ij}), and different line styles are used for the different estimation methods. In the line styles, long dashes correspond to assuming exogenous x_{ij} and short dashes correspond to allowing for endogenous x_{ij} (handled by including $\bar{x}_{\cdot j}$ in the models). The line styles with dots represent less constrained models (λ_0 and $\delta_{\bar{x}0}$ are not constrained in the joint exogenous and endogenous approaches respectively, and x_{0j} is not excluded in the conditional approaches).

Fig. 5(a) shows the estimates of β by using joint modelling approaches. Assuming that x_{ij} is exogenous (lines with long dashes) leads to severe inconsistency when this assumption is violated. In contrast, constraining $\lambda_0=1$ when exogeneity is assumed or $\delta_{\bar{x}0}=\delta_{\bar{x}}\lambda_0$ when endogeneity is allowed for by including $\bar{x}_{\cdot j}$ in the model (comparing lines without dots with lines with dots) makes little difference (the absolute difference in point estimates is at most 0.05). Fig. 5(c) shows that assuming exogeneity (lines with long dashes) leads to inconsistency also for γ_x when the assumption is violated, but much less severe as a percentage of the true value than for β. Constraining $\lambda_0=1$ (the line with long dashes and dots) makes this inconsistency considerably larger. The overall conclusion appears to be that the models allowing for endogeneity produce almost consistent estimators, regardless of whether the constraint $\delta_{\bar{x}0}=\delta_{\bar{x}}\lambda_0$ is used. Using $\mathbf{x}_j$ instead of $\bar{x}_{\cdot j}$ gives nearly identical estimates (which are not shown in Fig. 5). The standard errors are smaller when the constraints $\delta_{\bar{x}0}=\delta_{\bar{x}}\lambda_0$ or $\boldsymbol{\delta}_{x0}=\boldsymbol{\delta}_x\lambda_0$ are used but, with these constraints, whether $\mathbf{x}_j$ or $\bar{x}_{\cdot j}$ are used makes little difference to the standard errors (less than 1%).

In Figs 5(b) and 5(d) for the conditional modelling approaches, we see that assuming exogeneity (lines with long dashes) when it is violated leads to the greatest inconsistency. Excluding x_{0j} (lines without dots) increases the inconsistency in the model assuming exogeneity when it holds and in the model allowing for endogeneity. Note, however, that excluding x_{0j} makes little difference when $\mathbf{x}_j^+$ or $\bar{x}_{\cdot j}^+$ are included instead of $\bar{x}_{\cdot j}$ (which is not shown in Fig. 5) and gives almost consistent estimates. The reason for the poor performance of the model that includes $\bar{x}_{\cdot j}$, but not x_{0j} (short dashed lines without dots), appears to be that, in the model that includes $\mathbf{x}_j$ (equivalent to x_{0j} and $\mathbf{x}_j^+$), the estimated coefficient of x_{0j} is quite different from the coefficients of the x_{ij} for $i=1,2,3$ (when exogeneity holds, the coefficient of x_{0j} is negative and the other coefficients are 0). By including only $\bar{x}_{\cdot j}$, all coefficients are effectively set equal, giving poor estimates of the coefficients of x_{ij} for $i=1,2,3$ and therefore inadequate control

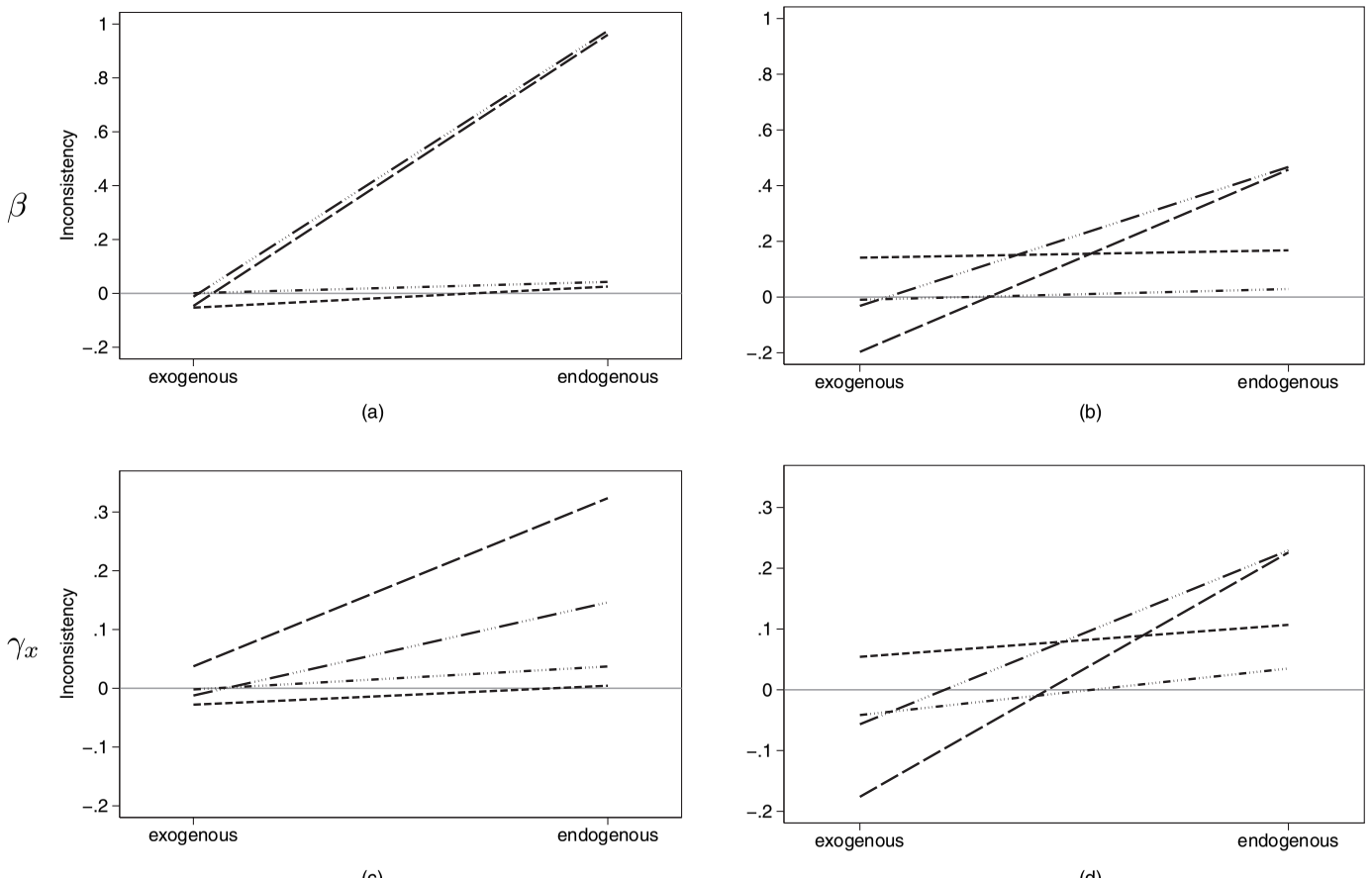

Fig. 5. Inconsistency (difference between the asymptotic estimate and the parameter value) for (a), (b) β and (c), (d) γ_x, by using (a), (c) joint approaches and (b), (d) conditional approaches with $T = 4$: (a), (c) — — —, exogenous, $\lambda_0 = 1$; ▬ ⋯ ▬, exogenous, free λ_0; ------, endogenous, with $\bar{x}_{\cdot j}$, $\delta_{\bar{x}0} = \delta_{\bar{x}}\lambda_0$; ▪ ⋯ ▪ ⋯, endogenous, with $\bar{x}_{\cdot j}$, free $\delta_{\bar{x}0}$; (b), (d) — — —, exogenous, no x_{0j}; ▬ ⋯ ▬, exogenous, with x_{0j}; ------, endogenous, with $\bar{x}_{\cdot j}$, no x_{0j}; ▪ ⋯ ▪ ⋯, endogenous, with $\bar{x}_{\cdot j}$, with x_{0j}

Table 3. 100(asymptotic estimate − parameter value) for the naive approach N_C, joint approaches and conditional approaches†

Condition	*Parameter*	N_C	*Results for the following joint approaches:*				*Results for the following conditional approaches:*				
			C_J	A_J	R_J	E_J	E_C	A_C	R_C	W_C	W_C^*
Exogenous											
$T=4$	β	108	−5	−2	−5 (0.65)	0	−3	14	−1 (0.71)	1	−1
$T=5$		63	−4	−1	−3 (0.52)	0	−3	8	−1 (0.55)	0	−1
$T=6$		38	−4	−1	−3 (0.44)	0	−3	5	−1 (0.46)	0	−1
Endogenous											
$T=4$	β	134	96	97	2 (0.75)	4	46	16	2 (0.76)	3	2
$T=5$		129	84	88	2 (0.58)	4	42	11	2 (0.59)	3	2
$T=6$		117	74	78	2 (0.49)	3	38	8	2 (0.50)	2	2
Exogenous											
$T=3$	γ_x	−51	4	−1	−3 (0.46)	0	−6	6	−4 (0.46)	−2	−4
$T=4$		−22	3	−1	−3 (0.37)	0	−6	3	−4 (0.37)	−2	−4
$T=5$		−11	2	−1	−3 (0.32)	0	−5	2	−4 (0.32)	−3	−4
Endogenous											
$T=4$	γ_x	0	32	15	0 (0.42)	4	23	11	3 (0.47)	4	3
$T=5$		8	37	24	1 (0.35)	4	25	9	3 (0.37)	4	3
$T=6$		18	39	29	2 (0.30)	4	25	7	3 (0.32)	4	3

†Joint approaches: C_J without $\bar{x}_j$ and with $\lambda_0 = 1$ (Crouchley and Davies, 2001), A_J without $\bar{x}_j$ and with free λ_0 (Aitkin and Alfo, 2003), recommended R_J with $\bar{x}_j$ and $\delta_{\bar{x}0} = \delta_{\bar{x}}\lambda_0$ and experimental E_J with $\bar{x}_j$ and free $\delta_{\bar{x}0}$. Conditional approaches: E_C conditions on x_{0j} and y_{0j}, A_C conditions on $\bar{x}_j$ and y_0 (Akay, 2012), recommended R_C conditions on $\bar{x}_j$, x_{0j} and y_0 (Rabe-Hesketh and Skrondal, 2013), W_C conditions on $\mathbf{x}_j^+$ and y_0 (Wooldridge, 2005) and W_C^* conditions on $\mathbf{x}_j$ and y_0. For the recommended approaches, robust standard errors for a sample size of 100 are given in parentheses.

for the between-subject effect of x_{ij}. The overall conclusion appears to be that models allowing for endogeneity produce almost consistent estimators, as long as the coefficient of $\bar{x}_{\cdot j}^+$ is not constrained equal to the coefficient of x_{0j}. The standard errors differ by less than 2% between the different almost consistent approaches that allow for endogeneity.

Comparing the joint and conditional approaches that include $\bar{x}_{\cdot j}$ (with constraint $\delta_{\bar{x}0} = \delta_{\bar{x}}\lambda_0$ for the joint approach and including x_{0j} in the conditional approach) when x_{ij} is endogenous, the standard errors for the conditional approach are 12% greater for γ_x, but only 1% greater for β. For both approaches, the model-based asymptotic standard errors are lower than the true asymptotic standard errors but only by at most 2%.

Table 3 reports 100 times the difference between the asymptotic estimate and the parameter value for all the methods described in Section 7.1 for $T = 4, 5, 6$. The naive estimator (denoted N_C) is severely inconsistent for β and less so for γ_x. When $\mathbf{x}_j$ is endogenous, the negative inconsistency for γ_x due to the initial conditions problem is cancelled by the positive inconsistency due to endogeneity. The joint approaches that assume exogeneity of $\mathbf{x}_j$ (C_J with $\lambda_0 = 1$ and A_J with free λ_0) are severely inconsistent when $\mathbf{x}_j$ is endogenous. The recommended joint and conditional approaches (R_J and R_C) are almost consistent regardless of the number of occasions T, as are Wooldridge's solution W_C and Wooldridge's solution with x_{0j} as an additional covariate (W_C^*). The recommended conditional approach has a slightly larger asymptotic sampling standard deviation than the recommended joint approach, but only when $\mathbf{x}_j$ is exogenous. The constrained

Wooldridge solution (A_C) is substantially inconsistent when $T=4$ with decreasing inconsistency as T increases.

8. Discussion

We have clarified, unified and extended methods for estimating dynamic/transition models for binary data with unobserved heterogeneity. Specifically, we have discussed two approaches to the initial conditions problem: joint modelling of the initial and subsequent responses and conditional modelling of subsequent responses given the initial response. Both approaches require approximate working models for which we have made recommendations. We also discussed the role of robust standard errors, how to handle missing data and extensions of joint working models to handle endogenous covariates.

To assess how close to consistent the estimators are, we have investigated their asymptotic performance. When the time-varying covariates are exogenous, or when the working models allow for level 2 endogeneity, the estimators are almost consistent. In the conditional approach, it is important not to include the means of time-varying covariates across all occasions (including the initial occasion) unless the time-varying variables at the initial occasion are also included.

What are the advantages and disadvantages of joint and conditional working models? Under level 2 exogeneity of all covariates, an advantage of the joint approach is that γ_z and ψ can be almost consistently estimated, in addition to γ_x and β. However, only γ_x and β can be estimated almost consistently if there is level 2 endogeneity. As we saw in the children's wheezing application, the joint approach allows testing for level 2 endogeneity, unlike the conditional approach. The joint approach may seem more natural, since the initial response is treated as a response and not a covariate, but both approaches are based on approximations. An advantage of joint modelling is that it is slightly more efficient, but the conditional approach is straightforward to implement in standard software for random-intercept modelling of binary data, whereas the joint approach requires more specialized software such as `gllamm`. However, estimation of robust standard errors is not implemented in standard software.

The class of model that was considered in this paper can be extended in various directions. Obvious extensions would be to relax the first-order Markov structure by considering further lags of the response or to specify antedependence models where the lagged response has time-varying coefficients β_i. Francis *et al.* (1996) and Albert and Follmann (2003) considered models where both the coefficients of the covariates and the effects of the random intercept depend on the previous state. Another obvious extension would be to censored, ordinal or nominal responses, or counts (see Wooldridge (2005) for discussion of the conditional approach for different response types).

Instead of specifying the dynamics in terms of lags in observed responses $y_{i-1,j}$, we could consider latent Markov models (Coleman, 1964) with lags in the latent response $y^*_{i-1,j}$ (e.g. Pudney (2008)). Heckman (1981c) discussed a very general dynamic model for binary responses that also includes lags for both observed and latent responses. The assumption of longitudinal independence for the level 1 errors ε_{ij} can also be relaxed (e.g. Hyslop (1999), Stewart (2006) and Hajivassiliou and Ioannides (2007)). More general specifications of unobserved heterogeneity than a random intercept can be used by considering several random coefficients or common factors (e.g. Heckman (1981c)). For instance, the auto-regressive latent trajectory model of Bollen and Curran (2004) can be extended to non-continuous responses.

The models that were considered in this paper can also be viewed as special cases of non-linear state space models for longitudinal data, such as dynamic generalized linear mixed models (e.g. Fahrmeir and Tutz (2001), section 8.4). Bayesian inference can also be performed; for instance,

Hasegawa (2009) discussed different kinds of Markov chain Monte Carlo methods for dynamic ordered probit models.

The single-equation models that were considered here could be extended to generalized structural equation models (e.g. Skrondal and Rabe-Hesketh (2004) and Rabe-Hesketh *et al*. (2004a)). For instance, a dynamic simultaneous equation model with several non-continuous response variables was used by Hajivassiliou and Ioannides (2007). Another possibility would be a structural equation model where the response and/or covariates of the dynamic/transition model are treated as latent variables measured with error.

Acknowledgements

We thank the Joint Editor, the Guest Associate Editor and two reviewers for constructive comments that helped to improve the paper. We also thank the Fulbright Foundation for supporting this work.

References

Aitkin, M. and Alfo, M. (1998) Regression models for binary longitudinal responses. *Statist. Comput.*, **8**, 289–307.
Aitkin, M. and Alfo, M. (2003) Longitudinal analysis of repeated binary data using autoregressive and random effect modelling. *Statist. Modllng*, **3**, 291–303.
Akay, A. (2012) Finite-sample comparison of alternative methods for estimating dynamic panel data models. *J. Appl. Econmetr.*, **27**, 1189–1204.
Albert, P. S. and Follmann, D. A. (2003) A random effects transition model for longitudinal binary data with informative missingness. *Statist. Neerland.*, **57**, 100–111.
Albert, P. S. and Follmann, D. A. (2007) Random effects and latent processes approaches for analyzing binary longitudinal data with missingness: a comparison of approaches using opiate clinical trial data. *Statist. Meth. Med. Res.*, **16**, 417–439.
An, S. S., Bai, T. R., Bates, J. H. T., Black, J. L., Brown, R. H., Brusasco, V., Chitano, P., Deng, L., Dowell, M., Eidelman, D. H., Fabry, B., Fairbank, N. J., Ford, L. E., Fredberg, J. J., Gerthoffer, W. T., Gilbert, S. H., Gosens, R., Gunst, S. J., Halayko, A. J., Ingram, R. H., Irvin, C. G., James, A. L., Janssen, L. J., King, G. G., Knight, D. A., Lauzon, A. M., Lakser, O. J., Ludwig, M. S., Lutchen, K. R., Maksym, G. N., Martin, J. G., Mauad, T., McParland, B. E., Mijailovich, S. M., Mitchell, H. W., Mitchell, R. W., Mitzner, W., Murphy, T. M., Par, P. D., Pellegrino, R., Sanderson, M. J., Schellenberg, R. R., Seow, C. Y., Silveira, P. S. P., Smith, P. G., Solway, J., Stephens, N. L., Sterk, P. J., Stewart, A. G., Tang, D. D., Tepper, R. S., Tran, T. and Wang, L. (2007) Airway smooth muscle dynamics: a common pathway of airway obstruction in asthma. *Eur. Resp. J.*, **29**, 834–860.
Anderson, T. W. and Hsiao, C. (1982) Formulation and estimation of dynamic models using panel data. *J. Econmetr.*, **18**, 47–82.
Arellano, M. and Bond, S. (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Stud.*, **58**, 277–297.
Arulampalam, W. and Stewart, M. B. (2009) Simplified implementation of the Heckman estimator of the dynamic probit model and a comparison with alternative estimators. *Oxf. Bull. Econ. Statist.*, **71**, 659–681.
Bartolucci, F. and Nigro, V. (2010) A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica*, **78**, 719–733.
Bates, G. E. and Neyman, J. (1952) Contributions to the theory of accident proneness II: True or false contagion. *Univ. Calif. Publ. Statist.*, **1**, 255–275.
Bhargava, A. and Sargan, J. D. (1983) Estimating dynamic random effects models from panel data covering short time periods. *Econometrica*, **51**, 1635–1659.
Blundell, R. and Windmeijer, F. (1997) Cluster effects and simultaneity in multilevel models. *Hlth Econ.*, **6**, 439–443.
Bock, R. D. and Lieberman, M. (1970) Fitting a response model for n dichotomously scored items. *Psychometrika*, **33**, 179–197.
Bollen, K. A. and Curran, P. J. (2004) Autoregressive latent trajectory (ALT) models: a synthesis of two traditions. *Sociol. Meth. Res.*, **32**, 336–383.
Chamberlain, G. (1980) Analysis of covariance with qualitative data. *Rev. Econ. Stud.*, **47**, 225–238.
Chamberlain, G. (1982) Multivariate regression models for panel data. *J. Econmetr.*, **18**, 5–46.
Chamberlain, G. (1984) Panel data. In *Handbook of Econometrics*, vol. II (eds Z. Griliches and M. D. Intriligator), pp. 1247–1318. Amsterdam: North-Holland.

Coleman, J. S. (1964) *Models of Change and Response Uncertainty*. Englewood Cliffs: Prentice Hall.
Crouchley, R. and Davies, R. B. (2001) A comparison of GEE and random effects models for distinguishing heterogeneity, nonstationarity and state dependence in a collection of short binary event series. *Statist. Modllng*, **1**, 271–285.
Davies, R. B. and Crouchley, R. (1985) Control for omitted variables in the analysis of panel and other longitudinal data. *Geogr. Anal.*, **17**, 1–15.
Davies, R. B. and Pickles, A. (1986) Accounting for omitted variables in a discrete time panel data model of residential mobility. *Qual. Quant.*, **20**, 219–233.
Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*. Oxford: Oxford University Press.
Dundas, I. and McKenzie, S. (2006) Spirometry in the diagnosis of asthma in children. *Curr. Opin. Pulm. Med.*, **12**, 28–33.
Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn. New York: Springer.
Ferris, B. G., Ware, J. H., Berkey, C. S., Dockery, D. W., Spiro, A. and Speizer, F. E. (1985) Effects of passive smoking on health of children. *Environ. Hlth Perspect.*, **62**, 289–295.
Fotouhi, A. R. (2005) The initial conditions problem in longitudinal binary process: a simulation study. *Simuln Modllng Pract. Theor.*, **13**, 566–583.
Francis, B. J., Stott, D. N. and Davies, R. B. (1996) *SABRE: a Guide for Users, Version 3.1*. Lancaster: Lancaster University. (Available from `http://www.cas.lancs.ac.uk/software/sabre3.1/sabre.html`.)
Fuhlbrigge, A. L., Kitch, B. T., Paltiel, A. D., Kuntz, K. M., Neumann, P. J., Dockery, D. W. and Weiss, S. T. (2001) FEV1 is associated with risk of asthma attacks in a pediatric population. *J. Allgy Clin. Immun.*, **107**, 61–67.
Gilliland, F. D., Li, Y.-F. and Peters, J. M. (2001) Effects of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children. *Am. J. Resp. Crit. Care Med.*, **163**, 429–436.
Hajivassiliou, V. A. and Ioannides, Y. M. (2007) Unemployment and liquidity constraints. *J. Appl. Econmetr.*, **22**, 479–510.
Hasegawa, H. (2009) Bayesian dynamic panel-ordered probit model and its application to subjective well-being. *Communs Statist. Simuln Computn*, **38**, 1321–1347.
Heagerty, P. J. and Kurland, B. F. (2001) Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika*, **88**, 973–985.
Heckman, J. J. (1981a) The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications* (eds C. F. Manski and D. L. McFadden), pp. 179–195. Cambridge: MIT Press.
Heckman, J. J. (1981b) Heterogeneity and state dependence. In *Studies in Labor Markets* (ed. S. Rosen), pp. 91–139. Chicago: University of Chicago Press.
Heckman, J. J. (1981c) Statistical models for discrete panel data. In *Structural Analysis of Discrete Data with Econometric Applications* (eds C. F. Manski and D. L. McFadden), pp. 179–195. Cambridge: MIT Press.
Honoré, B. E. and Kyriazidou, E. (2000) Panel data discrete choice models with lagged dependent variables. *Econometrica*, **68**, 839–874.
Hsiao, C. (2002) *Analysis of Panel Data*, 2nd edn. Cambridge: Cambridge University Press.
Hyslop, D. R. (1999) State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, **67**, 1255–1294.
Laird, N. M. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Statist. Ass.*, **73**, 805–811.
Lawal, B. (2003) *Categorical Data Analysis with SAS and SPSS Applications*. Mahwah: Erlbaum.
Mandel, M. and Betensky, R. A. (2008) Estimating time-to-event from longitudinal ordinal data using random-effects Markov models: application to multiple sclerosis progression. *Biostatistics*, **9**, 750–764.
Michaud, P.-C. and Tatsiramos, K. (2011) Fertility and female employment dynamics in Europe: the effect of using alternative econometric modeling assumptions. *J. Appl. Econmetr.*, **26**, 641–668.
Mundlak, Y. (1978) On the pooling of time series and cross section data. *Econometrica*, **84**, 69–85.
Nafstad, P., Jaakkola, J. J., Hagen, J. A., Zahlsen, K. and Magnus, P. (1997) Hair nicotine concentrations in mothers and children in relation to parental smoking. *J. Expos. Anal. Environ. Epidem.*, **7**, 235–239.
Nerlove, M. (1971) Further evidence on the estimation of dynamic economic relations from a time series of cross sections. *Econometrica*, **39**, 359–382.
Neuhaus, J. M. and McCulloch, C. M. (2006) Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Statist. Soc.* B, **68**, 859–872.
Ober, C. and Yao, T.-C. (2011) The genetics of asthma and allergic disease: a 21st century perspective. *Immun. Rev.*, **242**, 10–30.
Orme, C. D. (2001) Two-step inference in dynamic non-linear panel data models. *Technical Report*. School of Economic Studies, University of Manchester, Manchester.
Pudney, S. (2008) The dynamics of perception: modelling subjective wellbeing in a short panel. *J. R. Statist. Soc.* A, **171**, 21–40.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2003) Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statist. Modllng*, **3**, 215–232.
Rabe-Hesketh, S. and Skrondal, A. (2001) Parameterization of multivariate random effects models for categorical data. *Biometrics*, **57**, 1256–1264.
Rabe-Hesketh, S. and Skrondal, A. (2006) Multilevel modelling of complex survey data. *J. R. Statist. Soc.* A, **169**, 805–827.
Rabe-Hesketh, S. and Skrondal, A. (2012) *Multilevel and Longitudinal Modeling using Stata*, 3rd edn, vol. II, *Categorical Responses, Counts, and Survival*. College Station: Stata Press.
Rabe-Hesketh, S. and Skrondal, A. (2013) Avoiding biased versions of Wooldridge's simple solution to the initial conditions problem. *Econ. Lett.*, **120**, 346–349.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004a) Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167–190.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004b) GLLAMM manual. *Technical Report 160*. Division of Biostatistics, University of California, Berkeley. (Available from `http://www.bepress.com/ucbbiostat/paper160/`.)
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econmetr.*, **128**, 301–323.
Rotnitzky, A. G. and Wypij, D. (1994) A note on the bias of estimators with missing data. *Biometrics*, **50**, 1163–1170.
Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman and Hall–CRC.
Snijders, T. A. B. and Berkhof, J. (2008) Diagnostic checks for multilevel models. In *Handbook of Multilevel Analysis* (eds J. de Leeuw and E. Meijer), pp. 141–175. New York: Springer.
Song, C., Kuo, L., Derby, C. A., Lipton, R. B. and Hall, C. B. (2011) Multi-stage transitional models with random effects and their application to the Einstein aging study. *Biometr. J.*, **53**, 938–955.
Speizer, F. E. (1990) Asthma and persistent wheeze in the Harvard Six Cities Study. *Chest*, **98**, suppl., 191S–195S.
Stein, R. T., Sherill, D., Morgan, W. J., Holberg, C. J., Halonen, M., Taussig, L. M., Wright, A. L. and Martinez, F. D. (1999) Respiratory syncytial virus in early life and risk of wheeze and allergy by age 13 years. *Lancet*, **354**, 541–545.
Stewart, M. B. (2006) Maximum simulated likelihood estimation of random-effects dynamic probit models with autocorrelated errors. *Stata. J.*, **6**, 256–272.
Sutradhar, B. C. and Farrell, P. J. (2007) On optimal lag 1 dependence estimation for dynamic binary models with application to asthma data. *Sankyha* B, **69**, 448–467.
Train, K. E. (2009) *Discrete Choice Methods with Simulation*, 2nd edn. Cambridge: Cambridge University Press.
Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer.
Ware, J. H., Dockery, D. W., Spiro, A., Speizer, F. E. and Ferris, B. G. (1984) Passive smoking, gas cooking and respiratory health in children living in six cities. *Am. Rev. Resp. Dis.*, **129**, 366–374.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.
Wooldridge, J. M. (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *J. Appl. Econmetr.*, **20**, 39–54.
Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edn. Cambridge: MIT Press.