The specification of dynamic two-state panel data models

Tue Gørgens* Dean Hyslop†

2015-01-29

Abstract: This paper compares and contrasts dynamic panel data model and multi-spell duration model approaches to analyzing longitudinal discrete-time binary outcomes. Prototypical dynamic panel data models specify low-order Markovian state dependence and impose certain symmetry restrictions on the probabilities of transiting between states. In contrast, multi-spell duration models typically allow for state-specific elapsed duration dependence, and allow the probability of entry into and exit from a state to vary flexibly. We show that both of these approaches are special cases within a general framework. We compare specific dynamic panel data and duration models empirically using a case study of poverty transitions. In this example, both the state dependence and the symmetry restrictions imposed by the simpler dynamic panel data models are severely rejected against the more flexible duration model alternatives. Consistent with some other recent literature, we conclude that the standard dynamic panel data model is unacceptably restrictive in this context.

Keywords: Panel data, transition data, binary response, duration analysis, event history analysis, initial conditions, random effects.

JEL classification codes: C33, C35, C41, C51.

Acknowledgments: Tue Gørgens' research was supported in part by Australian Research Council Grant DP1096862. We thank Ann Huff Stevens for providing us the data and Sanghyeok Lee for excellent programming assistance.

*Address: Research School Economics, The Australian National University, Canberra 0200, Australia. E-mail: tue.gorgens@anu.edu.au.

†Address: School of Economics and Finance, Victoria University of Wellington, Wellington 6140, New Zealand. E-mail: dean.hyslop@vuw.ac.nz.

1 Introduction

Longitudinal analysis of binary outcomes is central to many studies in applied economics and other social sciences.[1] In the two-state panel data considered in this paper individuals are followed over time, and the basic outcome measure is which of two states each individual is occupying in each period.[2] The aim of the analysis is understanding the factors which influence either which state is occupied or the times of transitions between states. Often the individual dynamic processes are characterized by a degree of persistence, and an important part of the analysis is to understand whether this persistence is due either to heterogeneity across individuals or to true state dependence.[3]

There are two conceptually distinct approaches to analyzing two-state panel data.[4] The first approach focuses on modeling individuals' probabilities of occupying each state in each period. State dependence is typically modeled in terms of the effects of previous periods' state occupancy on the probability distribution for the current period's state occupancy (Markovian state dependence). Furthermore, the effects of covariates and unobserved heterogeneity on the implied probabilities of transiting into and out of a particular state are usually assumed to be symmetric. We refer to these as dynamic binary response (DBR) panel data models.

The second approach focuses on modeling the individuals' probability of moving between states between periods. State dependence is typically modeled in terms of the effects of the elapsed duration since entering the current state on the probability of a transition occurring (duration dependence). Most applications allow these effects to depend on the

[1]Typical examples include employment (Heckman, 1981a; Hyslop, 1999), unemployment (Arulampalam et al., 2000), welfare dependency (Bane and Ellwood, 1983), poverty (Capellari and Jenkins, 2004; Stevens, 1999), health (Halliday, 2008; and Contoyannis et al., 2004, using ordered rather than binary outcomes), and peace and conflict between national states (Beck et al., 2002; Beck and Katz, 1997).

[2]These are also know as discrete-time transition data (Lancaster, 1990).

[3]Heckman (1978, 1981c) first proposed the latent variable threshold model (with general intertemporal covariance matrix for the disturbances). Heckman distinguishes three sources of persistence: time-invariant unobserved heterogeneity, persistent shocks, and true state dependence. Heckman and Borjas (1980) distinguish four types of state dependence: Markovian state dependence, duration dependence, lagged duration dependence, and occurrence dependence. The focus of the models considered in this paper are generally on the first and second of these types.

[4]For example, welfare participation outcomes may be analyzed either in terms of the probability of being on-welfare (Card and Hyslop, 2005), or in terms of the probability of a spell on-welfare (or off-welfare) ending (Zabel et al., 2010).

current state (Markovian state dependence), and some applications analyze the effects of the number of previous transitions (occurrence dependence) and of completed durations of previous spells (lagged duration dependence). We refer to these as multi-spell duration (MSD) models.

The objective of this paper is to compare and contrast these two alternative approaches to modeling two-state panel data.[5] We start by outlining a general framework for representing binary outcome processes, which provides a unifying framework for the DBR and MSD models. We first show that, in general, a sequence of binary outcomes can be equivalently represented by the initial-period outcome and a sequence of transition indicators associated with subsequent periods.[6] We then show that modeling the probabilities of state occupancy or the probabilities of moving between states is nonparametrically equivalent.

The discussion of the general framework also considers issues associated with incorporating observed covariates and unobserved heterogeneity in the models, and handling left-censored spells and initial conditions at the start of the observation period. This general discussion demonstrates that the data requirements for each of the two approaches are equivalent, and they typically differ in the choice of the parameters of interest and in the way they incorporate state dependence. Furthermore, differences arise due to the typical parametric specification of DBR and MSD models used in empirical analyses. In particular, typical DBR models are comparatively parsimonious, while the MSD models are relatively flexible.

It is fair to say that DBR models are more widely used than MSD models. The use of DBR models may be motivated by specific research questions, by computational considerations, by greater familiarity with DBR models among researchers, or possibly by misconceptions about the differences between the two approaches. For example, it is

[5]To the best of our knowledge little comparative analysis has been conducted of these alternative approaches. Exceptions include Jenkins (1995) and Barmby (1998), Cappellari et al. (2007) who compare a duration and Markov model for employment transitions, and Bhuller et al. (2014) who analyze the adequacy of first-order state dependence dynamic panel data models against more general models that allow for duration and occurrence dependence.

[6]In the appendix we show that the data can also be equivalently represented by the initial outcome and a sequence of either transition times or the durations between transition times.

perhaps natural to think that DBR models should be used if one is mainly interested in modeling state occupancy and MSD models if one is interested in transitions, because these models focus directly on the parameters of interest. One might also think that DBR models are more capable of fitting and predicting state occupancy and MSD models are better at fitting and predicting transitions, since their main purpose is to model those aspects of the data. However, these conceptions are false, as we show in this paper. Another potential misconception is that if levels of covariates affect state occupancy then changes in covariates affect transitions. Conversely, if levels of covariates affect transitions then cumulative sums of covariates affect transitions. We also show this to be false in general.

The paper is organized as follows. Section 2 begins with some preliminary discussion to motivate the analysis that follows. We then outline a general framework for longitudinal binary outcome data that includes the DBR and MSD approaches as special cases. In particular, we show that, in general, the data representations of these approaches are equivalent, and the approaches are nonparametrically equivalent. In this section, we also discuss including covariates and unobserved heterogeneity, and how the approaches diverge in their respective parameterizations which affects how left-censored data are handled.

In Section 3, we use an empirical case study to compare and contrast the properties of alternative DBR and MSD models. For this purpose, we apply alternative specifications to estimation of the duration of poverty spells and poverty persistence (see Stevens, 1999). The estimation results show the MSD model dominates the more restrictive DBR models: both the DBR model's implied symmetric effects of covariates and unobserved heterogeneity on poverty entry and exit, as well as the Markovian state dependence, is rejected. Furthermore, the MSD model's within-sample predictions provides a better fit to the actual observations than the DBR model predictions.

The paper concludes with a discussion in Section 4. The appendix provides a link between the representation given in the main text and continuous-time duration analysis.

2 Modeling two-state panel data

In this section we present a general framework for handling longitudinal binary outcomes, that encompass the DBR and MSD models commonly used in analysis. The section begins with some preliminary comments, and a stylized example to help motivate the analysis. Then subsection 2.2 shows that the data representations for the DBR and MSD approaches are equivalent. Consequently, in subsection 2.3 we show that, when there is no observed or unobserved heterogeneity, the models are nonparametrically equivalent. In subsequent subsections we consider the effects of adding covariates and unobserved heterogeneity. The inclusion of covariates does not change the nonparametric equivalence results. However, equivalence is essentially lost when unobserved heterogeneity is important, because identification requires parametric assumptions and the two approaches favor different parameterizations.

2.1 Preliminaries

We begin with some preliminary comments. First, at least conceptually, some longitudinal binary outcome processes are inherently continuous time processes, such as the time to next heart attack, while others are more naturally discrete time processes, such as welfare participation where eligibility is typically determined over a period (e.g. a month or week). In this paper, we do not consider continuous time processes, although in practice such processes may be measured in discrete time.

Second, in practice, longitudinal binary outcomes may be measured either at a point in time, such as whether or not a person is employed on the 1st day of the month, or over a period, such as whether or not a person is employed (at all) during the month. The choice may be optional, as in the employment example, or context specific: e.g., measurement of welfare participation and/or income-based poverty would more naturally be period outcomes because of the welfare eligibility criteria and income measurement respectively; in contrast, wealth-based poverty may, at least conceptually, be point-in-time outcomes. Irrespective of how the outcomes are measured, we will refer to the outcomes, equivalently, as occurring at time t or in period t; and transitions will be assumed to occur between the times when the state changes.

Third, relevant measurements of a process may not be complete, and data may be missing at the beginning, during, or at the end of the process. For simplicity, the general framework we outline below explicitly allows only for left- and right-censoring. Middle-censoring can be handled using similar methods. In the following we shall refer to potential measurements, whether they are actually carried out or not, as the (data-generating) "process", and reserve the word "measurements" for the actual measurements available for analysis, the difference being missing measurements.

Before we discuss the general framework for handling longitudinal binary outcome data, we briefly discuss a stylized example to help motivate the analyses. Consider an analysis of individuals' poverty experiences over time, which has been studied by Stevens (1999). In such a study, we may have data on individuals (indexed by i), measured in years (indexed by t), and observe whether or not they are in poverty in each year (denoted by $Y_{it} = 1$ or $Y_{it} = 0$ respectively). The analysis might focus on a variety of aspects of poverty experience, such as the probability of being in poverty in one year conditional on being in poverty in the previous year (Markovian state dependence), the probability that a poverty spell ends after d years (duration dependence), the number of years and/or spells in poverty over some period, etc.

There are two common approaches taken to such analyses. The DBR approach focuses on modeling the probability that an individual is in poverty in year t, and typically conditions on relevant observable covariates (X_{it}), the poverty states experienced in the recent past (Y_{it-1}, Y_{it-2}, etc.), and time-constant unobservable factors (V_i). A typically specified DBR model assumes a low-order (e.g. first-order) Markovian state dependence in which only outcomes in the recent past affect the current state, and additionally assumes that the effects of covariates are symmetric on the probability of a transition between the states.[7] For example, the probability statement of interest for the simplest first-order

[7]Symmetric covariate effects can be relaxed, although this is usually done within a single-equation context (e.g. Card and Hyslop, 2009; Browning and Carro, 2010).

DBR model, would be:

$$\mathsf{P}(Y_{it} = 1 | Y_{it-1} = y_{it-1}, X_{it} = x_{it}, V_i = v) = G_{DBR}(\gamma y_{it-1} + x'_{it}\beta + v), \tag{1}$$

for some function G_{DBR}. (A practically useful parameterization of the influence of unobserved heterogeneity is discussed in Sections 2.5 and 2.7.) The first measurement requires special consideration, and we return to this so-called initial conditions problem later. With regards to the dynamic properties of poverty experience, the principal focus in the application of these models is the Markovian state dependence. While the focus is on the probability that an individual is poor in a given period, the estimated models can also be used to predict poverty spell durations and poverty experiences over longer periods, etc.

The MSD approach focuses on modeling the probability that an individual's poverty, or non-poverty, spell ends in year t: that is, whether or not a transition occurs between years t and $t+1$ (denoted by $C_{it+1} = 1$ or $C_{it+1} = 0$ respectively). This approach again typically conditions on relevant observable covariates (X_{it}) and time-constant unobservable factors (V_i), but focuses on the elapsed duration d in the spell at year t rather than the individual's recent poverty experience per se. For example, the probability statement of interest for a transition out of a state y spell in a simple MSD model with linear duration dependence, would be:

$$\begin{aligned}\mathsf{P}(C_{it} = 1 | Y_{it-1} = y_{it-1}, D_{it-1} = d_{it-1}, X_{it} = x_{it}, V_i = v)& \\ = G_{MSD}(\lambda_{y_{it-1}} d_{it-1} + x'_{it}\beta_{y_{it-1}} + v),& \quad (2)\end{aligned}$$

for some function G_{MSD}. (Parameterization of unobserved heterogeneity is discussed in Sections 2.5 and 2.9.) Note that this approach typically uses separate equations for the duration of poverty and non-poverty spells, to allow the transition probabilities into and out of poverty to vary. As well as the initial conditions problem, the MSD approach also needs to deal with initial left-censored spells. We return to these issues later. With regards to the dynamic properties of poverty experience, the principal focus in the application of these models is the duration dependence in the transition probabilities. This is in contrast to the DBR models. While the focus is on the probability that an individual makes a transition in a given period, like the DBR models, the estimated models can be used to predict poverty status in a given period and poverty experiences over longer periods, etc.

Stevens (1999) adopts an MSD analysis of poverty. In the empirical example below, we use this same context to compare and contrast the two approaches. But we emphasize that our intention is neither to replicate nor critique Stevens (1999) analysis. Rather, our intent is to use it as an example to illustrate and draw attention to similarities and differences in the approaches.

2.2 Equivalent data representations

The different modeling approaches require different organization of the data. In this subsection, we show that the different representations are equivalent (one-to-one) and it is possible to convert data indented for DBR modeling to MSD data and vice versa. This subsection also presents our main notation.

Suppose the data can be represented as an ordered sequence of binary measurements for each individual (or "subject") in a random sample. For convenience, we refer to the measurements as ordered with respect to time, denoted by t, and we refer to the binary values as the state occupied by the individual, labeled 0 and 1. The times are evenly spaced and may refer to points in time or periods in time. The individual are indexed by $i = 1, \ldots, N$, and the times are indexed by $t = 1, \ldots, T$. In the following, we consider a given individual and generally omit the identifier $i = 1, \ldots, N$. For simplicity, we assume the data constitute a balanced panel.

As mentioned, the data may be incomplete. If the process begins and the individual first occupies one of the states before the first measurement takes place, then the data are said to be left-censored. We assume that the time origin is not known if the data are left-censored. Although measurement ends at T, we make no assumption that the process has ended at time T. Therefore, we assume that the data are always right-censored at time T. Throughout the paper we assume that right-censoring is independent of the underlying process.

The data can be represented in different ways. The most obvious representation consists of indicators, Y_{it} for $t = 1, \ldots, T$ with $Y_{it} \in \{0, 1\}$, of the state occupied by individual i at time t. Another representation of the data focuses on transitions between states. Let C_{it} for $t = 2, \ldots, T$ be indicators of whether or not individual i makes a transition between times $t-1$ and t. (We subscript by t since information at time t is required to determine whether a transition has occurred or not.) Transition measurements are available only for $t = 2, \ldots, T$.

The data representations $(Y_{i1}, Y_{i2}, \ldots, Y_{iT})$ and $(Y_{i1}, C_{i2}, \ldots, C_{iT})$ are equivalent in the sense the one can be recovered from the other. Specifically, the Y_{it}s and C_{it}s are related by

$$C_{it} = 1(Y_{it-1} \neq Y_{it}), \quad t = 1, \ldots, T, \tag{3}$$

and

$$Y_{it} = \left(Y_{i1} + \sum_{k=2}^{t} C_{ik}\right) \bmod 2, \quad t = 2, \ldots, T. \tag{4}$$

Therefore, from a data perspective focusing on the sequence of states occupied or on the transitions between states is equivalent.

In duration analysis, the data are often represented as transition times or spells instead of a sequence of state indicators. (A spell is a period between consecutive transitions during which the individual stays in the same state.) We focus here on time-based representations, since they are most convenient if covariates are time-varying. However, in Appendix A, we show that spell-based and time-based representations are equivalent.

2.3 Equivalent parameterizations

In this subsection, we discuss the different parameters of interest emphasized in the DBR and MSD approaches, and compare the corresponding likelihood functions. We show that in a nonparametric framework, the two approaches are equivalent in the sense that they are simply different (one-to-one) parameterizations of the same likelihood. The number of parameters is finite and can be nonparametrically estimated. Indeed, the maximum likelihood estimates are simply the sample analogues.

Random sampling of individuals identifies the distribution of $(Y_{i1}, Y_{i2}, \ldots, Y_{iT})$ and the distribution of $(Y_{i1}, C_{i2}, \ldots, C_{iT})$. Since these distributions are discrete, they can be characterized by a finite number of probabilities, 2^T to be precise. These probabilities can be nonparametrically estimated provided $N \geq 2^T - 1$. However, the parameters of interest in most applications are not the probabilities associated with these unconditional joint distributions, but rather conditional probabilities of current outcomes given past outcomes.

The DBR approach focuses on the conditional probabilities of being in one of the states given the sequence of states previously occupied. For individual i and for $t = 1, \ldots, T$, let $\mathbf{Y}_{it}$ denote the random outcome history up to (and including) time t; that is, $\mathbf{Y}_{it} = (Y_{i1}, \ldots, Y_{it})$ for $1 \leq t \leq T$. Let $\mathbf{y}_{it} = (y_{i1}, \ldots, y_{it})$ denote the observed history. At time t, for $t = 1, \ldots, T$, the space of possible histories is $\mathbb{Y}^t = \{0, 1\}^t$. Let $\mathbf{y}_t$ with no subscript i denote a generic element of $\mathbb{Y}^t$. Then the conditional probability of being in state 1 at time t given the outcome history prior to time t is

$$\begin{aligned} &\chi = \mathsf{P}(Y_{i1} = 1), \\ &\zeta_t(\mathbf{y}_{t-1}) = \mathsf{P}(Y_{it} = 1 | \mathbf{Y}_{it-1} = \mathbf{y}_{t-1}), \quad \mathbf{y}_{t-1} \in \mathbb{Y}^{t-1}, \quad t = 2, \ldots, T. \end{aligned} \tag{5}$$

Note that there is the initial probability, and 2, 4, $\ldots$, 2^{T-1} conditional probabilities in the equations in (5), depending on the conditioning set; adding them up yields $2^T - 1$ total probabilities.

Assuming there is no left-censoring and that right-censoring is independent of outcomes, then the probabilities in (5) are fundamental parameters of interest. With left-censoring, they may or may not be, depending on whether there is interest in the effect of past outcomes which happen not to be measured. We return to this issue in Section 2.6.

In any case, treating each of the probabilities in (5) as a parameter to be estimated, the likelihood contribution for individual i is[8]

$$L_i^Y = \chi^{y_{i1}}(1-\chi)^{1-y_{i1}} \prod_{t=2}^{T} \zeta_t(\mathbf{y}_{it-1})^{y_{it}}(1-\zeta_t(\mathbf{y}_{it-1}))^{1-y_{it}}. \tag{6}$$

Combining the contributions of all N individuals yields a likelihood function which is valid for inference under the assumptions stated above. In particular, conditioning on right-censoring at T is feasible under the assumption of independent right-censoring.

In contrast to the DBR approach, the MSD approach focuses on the conditional probabilities of changing state at time t given the prior history; that is, on the hazard rates. In addition, there is the probability distribution of the initial state. The conditional probability of beginning in state 1 given the outcome history prior to time t and the hazard rates are defined as[9]

$$\begin{aligned} &\chi = \mathsf{P}(Y_{i1} = 1), \\ &\xi_t(\mathbf{y}_{t-1}) = \mathsf{P}(C_{it} = 1 | \mathbf{Y}_{it-1} = \mathbf{y}_{t-1}), \quad \mathbf{y}_{t-1} \in \mathbb{Y}^{t-1}, \quad t = 2, \ldots, T. \end{aligned} \tag{7}$$

Again, there is 1 initial probability, and 2, 4, ..., 2^{T-1} conditional probabilities in the equations in (7), giving $2^T - 1$ distinct probabilities in this representation.

Treating each of the probabilities in (7) as a parameter to be estimated, the likelihood contribution for individual i is[10]

$$L_i^C = \chi^{y_{i1}}(1-\chi)^{1-y_{i1}} \prod_{t=2}^{T} \xi_t(\mathbf{y}_{it-1})^{c_{it}}(1-\xi_t(\mathbf{y}_{it-1}))^{1-c_{it}}. \tag{8}$$

The comments following (6) apply here as well. In the absence of left-censoring, the probabilities in (7) are fundamental parameters of interest. However, the question of whether the probabilities are parameters of interest with left-censored data is complicated. We return to this issue in Section 2.8.

[8] Obviously, the likelihood can be equivalently parameterized in terms of being in state 0 rather than state 1 (i.e. $Y_{it} = 0$ versus $Y_{it} = 1$).

[9] It is possible to define $\mathbf{C}_{it} = (Y_{i1}, C_{i2}, C_{i2}, \ldots, C_{it})$ for $t \geq 1$, but since $\mathbf{Y}_{it}$ and $\mathbf{C}_{it}$ are equivalent we do not need $\mathbf{C}_{it}$.

[10] Obviously, the likelihood can be equivalently parameterized in terms of the absence rather than the presence of a transition (i.e. $C_{it} = 0$ versus $C_{it} = 1$).

To emphasize that (6) and (8) are simply reparameterizations of the same likelihood, note that

$$\begin{aligned}\zeta_t(\mathbf{y}_{t-1}) &= \xi_t(\mathbf{y}_{t-1})^{1-y_{t-1}}(1-\xi_t(\mathbf{y}_{t-1}))^{y_{t-1}} \\ &= 1-\xi_t(\mathbf{y}_{t-1})^{y_{t-1}}(1-\xi_t(\mathbf{y}_{t-1}))^{1-y_{t-1}}, \quad \mathbf{y}_{t-1}\in\mathbb{Y}^{t-1}, \quad t=2,\ldots,T,\end{aligned} \tag{9}$$

and

$$\begin{aligned}\xi_t(\mathbf{y}_{t-1}) &= \zeta_t(\mathbf{y}_{t-1})^{1-y_{t-1}}(1-\zeta_t(\mathbf{y}_{t-1}))^{y_{t-1}} \\ &= 1-\zeta_t(\mathbf{y}_{t-1})^{y_{t-1}}(1-\zeta_t(\mathbf{y}_{t-1}))^{1-y_{t-1}}, \quad \mathbf{y}_{t-1}\in\mathbb{Y}^{t-1}, \quad t=2,\ldots,T,\end{aligned} \tag{10}$$

where y_{t-1} denotes the final element of $\mathbf{y}_{t-1}$. Thus, the likelihood functions are equivalent, since both the data representations and the parameters are in one-to-one relationships.[11]

A comparison of spell-based and time-based parameterizations of the likelihood function is given in Appendix A.

2.4 Covariates

In this subsection, we discuss general issues related to including (predetermined) covariates in the analysis. In the absence of left-censoring and unobserved heterogeneity, the likelihood functions for the DBR and MSD approaches remain equivalent and the parameters of interest remain nonparametrically identified and estimable using standard smoothing techniques.

In an empirical analysis, there are two practical issues regarding the time reference for covariates. First, surveys often collect retrospective information relating to different periods, so the time of measurement may not be the same as the logical time reference for the information. (This is of course true for the outcome variable as well.) Second, some covariates which logically relate to time t are inappropriate conditioning variables

[11] Substituting (10) into (8) gives (6) since, for given t, we have

$$\begin{aligned}\xi_t(\mathbf{y}_{t-1})^{c_t}(1-\xi_t(\mathbf{y}_{t-1}))^{1-c_t} &= \zeta_t(\mathbf{y}_{t-1})^{(1-y_{t-1})c_t+y_{t-1}(1-c_t)}(1-\zeta_t(\mathbf{y}_{t-1}))^{y_{t-1}c_t+(1-y_{t-1})(1-c_t)} \\ &= \zeta_t(\mathbf{y}_{t-1})^{c_t+y_{t-1}-2c_ty_{t-1}}(1-\zeta_t(\mathbf{y}_{t-1}))^{1-c_t-y_{t-1}+2c_ty_{t-1}} = \zeta_t(\mathbf{y}_{t-1})^{y_t}(1-\zeta_t(\mathbf{y}_{t-1}))^{1-y_t}.\end{aligned}$$

The final step follows because $y_t = c_t + y_{t-1} - 2c_ty_{t-1}$. Similarly, plugging (9) into (6) gives (8).

because of simultaneity issues. For example, it is probably not interesting to condition a person's employment status in a given month on wage income earned in that month. In the discussion here, we simply assume that covariates are lagged or led so that it is sensible to condition outcomes at t on covariates with (notational) time reference t. In addition, in practice there is an issue of whether covariate effects are appropriately specified in levels or changes (or cumulative sums). We initially bypass this issue by conditioning on the entire prior covariate path. At the end of this subsection, we discuss levels and changes in terms of simple examples.

In preparation for the statement of the general likelihood functions later, we briefly present the conditional likelihood contributions for individual i given their covariate history. To keep the expressions simple and compact, the likelihood functions are stated in terms of probabilities rather than Greek-letter parameters.[12] As in the previous subsection, the likelihood functions associated with the different approaches represent equivalent parameterizations.

Let X_{it} denote a vector of covariates measured for individual i with reference to time t. For $t = 1, \ldots, T$, let $\mathbf{X}_{it}$ denote the covariate history at time t; that is, $\mathbf{X}_{it} = (X_{i1}, \ldots, X_{it})$. Let $\mathbf{x}_{it} = (x_{i1}, \ldots, x_{it})$ denote the observed history. Subscripting by t is not intended to preclude time-invariant covariates; for example, gender and ethnicity could be components of each X_{it}. Similarly, relevant information that pre-dates the time origin, such as school grades in a study of post-school labor market outcomes, can also be included as time-invariant elements of each X_{it}.

For the DBR approach, the likelihood contribution for individual i, conditional on right-censoring at T and conditional on the covariate history, is

$$L_i^Y = \mathsf{P}(Y_{i1} = y_{i1} | X_{i1} = x_{i1}) \prod_{t=2}^{T} \mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}). \tag{11}$$

As before, the total likelihood function is the product of the contributions for all the N individuals. Inference conditional on right-censoring at T is valid under the assumption of independent right-censoring.

12 For example, we write $\mathsf{P}(Y_{i1} = y_{i1})$ instead of $\chi^{y_{i1}}(1-\chi)^{1-y_{i1}}$.

Similarly, for the MSD approach, the likelihood contribution for individual i, conditional on right-censoring at T, in the time form is

$$L_i^C = \mathsf{P}(Y_{i1} = y_{i1} | X_{i1} = x_{i1}) \prod_{t=2}^{T} \mathsf{P}(C_{it} = c_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}). \tag{12}$$

The comments following (11) apply. Analogous expressions to (9) and (10) exist here, implying that (11) and (12) are reparametrizations of the same likelihood function.

If all covariates are discrete, so that X_{it} can take only a finite, say k, number of values, then there are $(2k)^{T-1}$ unknown probabilities in the likelihood contributions. Since this is also a finite number, the probabilities are in principle nonparametrically identified. If some covariates are continuous, the probabilities may be nonparametrically identified and estimable using nonparametric regression methods such as kernel regression, series estimation, or maximum penalized likelihood.

We now return to the issue of whether covariates should be included in levels or changes. A potential misconception is that if the state occupied depends on the level of a covariate, then transitions between states must depend on changes in that covariate. Conversely, if transitions depend on the level of the covariate, then the state occupied must depend on the cumulative sum of that covariate. This intuition is false. To make the relationships clearer, we discuss some simple examples in the remainder of this section.

Suppose first that changes in (but not the level of) a given covariate affects the probability of a transition. For example, suppose two cities have the same poverty rate, but moving between cities is disruptive and increases the likelihood of falling under the poverty threshold. This means the probability of a transition is lower for movers who are initially poor, and higher for those not in poverty. Let X_{it} be a scalar indicator of city. A simple model is

$$\mathsf{P}(C_{it} = 1 | Y_{it-1} = y, X_{it} = x_t, X_{it-1} = x_{t-1}) = \begin{cases} \delta_y & \text{if } x_t = x_{t-1}, \\ \delta_y + \eta_y & \text{if } x_t \neq x_{t-1}, \end{cases} \tag{13}$$

where $\delta_0 < \delta_1$ if poverty is a small risk and $\eta_1 < 0 < \eta_0$ if moving increases the risk.

The important feature of (13) is that the state-specific transition probabilities remain the same if there is no change in the covariate. Now, similarly to (9) we have

$$\mathsf{P}(Y_{it} = 1 | Y_{it-1} = y, X_{it} = x_t, X_{it-1} = x_{t-1}) \\ = \begin{cases} \delta_y^{1-y}(1-\delta_y)^y & \text{if } x_t = x_{t-1}, \\ (\delta_y + \eta_y)^{1-y}(1-\delta_y-\eta_y)^y & \text{if } x_t \neq x_{t-1}. \end{cases} \tag{14}$$

It follows immediately that if changes in a covariate affect the transition probability, then the probability of state occupancy is also affected by changes in (and not levels of) that covariate.

Conversely, consider a simple model where the probability of being poor is different in the two cities but moving is costless,

$$\mathsf{P}(Y_{it} = 1 | Y_{it-1} = y, X_{it} = x_t, X_{it-1} = x_{t-1}) = \begin{cases} \alpha_A & \text{if } x_t = A, \\ \alpha_B & \text{if } x_t = B. \end{cases} \tag{15}$$

Similarly to (10), the corresponding transition probability is

$$\mathsf{P}(C_{it} = 1 | Y_{it-1} = y, X_{it} = x_t, X_{it-1} = x_{t-1}) = \begin{cases} \alpha_A^{1-y}(1-\alpha_A)^y & \text{if } x_t = A, \\ \alpha_B^{1-y}(1-\alpha_B)^y & \text{if } x_t = B, \end{cases} \tag{16}$$

so the conditional transition probability given Y_{it-1} does not depend on X_{it-1}. In this model, the value of X_{it-1} influences Y_{it} only indirectly through Y_{it-1}.

In sum, whether the level of a covariate or changes over time matters is an issue distinct from whether we focus on the probability of state occupancy or transitions between states. If changes in covariates matter, then both probabilities depends on the changes. If levels matter, both probabilities depends on the levels of the covariate.

2.5 Unobserved heterogeneity

The next topic which we discuss is unobserved heterogeneity. While the data may be an iid random sample, the concern is that the average probabilities may not represent the outcomes for specific individuals, even conditional on X_{it}. For example, a population may contain some people with strong immune systems and others who easily get sick. The average hazard of coming down with the flu in a particular week given that a person is not already sick may reflect a near-zero probability for the former and near-one probability for the latter group. The parameters of interest are the individual-specific probabilities of becoming sick, rather than the average probability. If characteristics of a person's immune system were measured and available in the data, they could simply be included as covariates and there would be no problem. However, if data are not available, there is important unobserved heterogeneity in the population.

It is important to be aware that unobserved heterogeneity is a "structural" concept in the sense that it precludes nonparametric identification of parameters of interest. Untestable assumptions such as parametric functional-form specifications are necessary if the data are to be used for inference. In the literature, unobserved heterogeneity is treated as equivalent to an omitted covariate. Usually it is assumed to be predetermined for each individual, independent of covariates (past and future), and independent of the measurement scheme. While these assumptions are strong and perhaps implausible in most applications, they are still not sufficient to ensure identification. We proceed in this subsection by presenting the general form of the likelihood contributions in the presence of (independent) unobserved heterogeneity, without being explicit about identifying assumptions. Specific cases are discussed in detail later.

Let V_i denote a random variable (or vector) representing unobserved heterogeneity for individual i. Let $\mathcal{V}$ denote the support of V_i, and let Ψ denote the distribution function of V_i. In the DBR framework, the likelihood contribution for individual i becomes

$$L_i^Y = \int_{\mathcal{V}} \mathsf{P}(Y_{i1} = y_{i1} | X_{i1} = x_{i1}, V_i = v) \\ \times \left(\prod_{t=2}^{T} \mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \right) d\Psi(v). \quad (17)$$

Similarly, in the MSD framework we have

$$L_i^C = \int_{\mathcal{V}} \mathsf{P}(Y_{i1} = y_{i1} | X_{i1} = x_{i1}, V_i = v) \\ \times \left(\prod_{t=2}^{T} \mathsf{P}(C_{it} = c_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \right) d\Psi(v). \quad (18)$$

An important implication is that the likelihood contributions are no longer separable across time. As we have seen, the likelihood contributions can be broken into multiplicative time-specific components when there is no unobserved heterogeneity. If unobserved heterogeneity needs to be integrated out, this is no longer the case.

In practice, there are different ways of incorporating unobserved heterogeneity in the literature. A common approach is to specify V_i as a normally distributed random variable and include V_i as a regressor with a loading similar to the covariates (e.g. Hyslop, 1999; Chay and Hyslop, 2014). Following Heckman and Singer (1984), an alternative which we adopt here is to assume that unobserved heterogeneity has a discrete distribution in a multidimensional space. The discrete distribution can be thought of either as an approximation to a true underlying continuous distribution or as a distribution of a finite number of "types". If a model has a number, say Q, of "equations" each representing a different aspect, then we assume each type is characterized by a Q-vector of constants, one for each equation. Formally, we assume that V_i is a discrete random Q-vector with support $\nu_1, \ldots, \nu_K$, where $\nu_k = (\nu_{k1}, \nu_{k2}, \ldots, \nu_{kQ}) \in \mathbb{R}^Q$ for $k = 1, \ldots, K$, and probability distribution $\pi_1, \ldots, \pi_K$ with $\sum_{k=1}^{K} \pi_k = 1$.

2.6 The DBR approach

In the DBR approach, the model focuses on low-order Markovian state dependence and assumes that the conditional probabilities of being in a given state depend only on the most recent p previous outcomes and covariates. It is also assumed that only contemporaneous covariates matter. In practice, the latter is not serious limitation since covariates which logically refer to different times (leads and lags) can be included among the contemporaneous covariates. Thus, for some fixed $p \geq 1$ it is assumed that

$$\begin{aligned} &\mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \\ &\qquad = \mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}^p_{it-1} = \mathbf{y}^p_{it-1}, X_{it} = x_{it}, V_i = v), \quad t = p+1, \ldots, T, \end{aligned} \tag{19}$$

where $\mathbf{Y}^p_{it} = (Y_{it-p+1}, \ldots, Y_{it})$ and $\mathbf{y}^p_{it} = (y_{it-p+1}, \ldots, y_{it})$. We refer to equation (19) as the DBR model's "structural equation" of interest. This equation does not restrict the probabilities for the p initial outcomes, $(Y_{i1} = y_{i1}, \ldots, Y_{ip} = y_{ip})$, referred to as the "initial conditions" of the process. In applications, there is typically less substantive interest in the probabilities associated with the initial conditions, but it is important they are dealt with unless they can be considered to be exogenous (Heckman, 1981b).

Under Assumption (19) and imposing our discrete distribution of unobserved heterogeneity, the likelihood contribution (17) can be written

$$\begin{aligned} L^Y_i = \sum_{k=1}^{K} \pi_k \Bigg[& A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, \nu_k) \\ & \times \left(\prod_{t=p+1}^{T} \mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}^p_{it-1} = \mathbf{y}^p_{it-1}, X_{it} = x_{it}, V_i = \nu_k) \right) \Bigg], \end{aligned} \tag{20}$$

where the term $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, v)$ represents the probability contribution of the initial conditions, which can be expressed as

$$A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, v) = \mathsf{P}(Y_{i1} = y_{i1}|X_{i1} = x_{i1}, V_i = v) \\ \times \left(\prod_{t=2}^{p} \mathsf{P}(Y_{it} = y_{it}|\mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \right). \quad (21)$$

If the process is ongoing prior to the observation period, so the data are left-censored, then the probabilities in (21) are simply necessary nuisance parameters. However, if the process is observed from the beginning, so the data are not left-censored, in general the initial conditions will be relevant unless all individuals have the same states for the first p periods. In this case, the probabilities in (21) may be of substantive interest in terms of the structural process.

If there is no unobserved heterogeneity, say $\mathsf{P}(V_i = \nu_1) = 1$, then the sum over K-types in (20) effectively disappears. In this case, the term $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, \nu_1)$ represents the likelihood contribution of the first p observed outcomes for individual i.[13] The remaining part of the likelihood contribution in (20) involve only observed variables, and these probabilities are nonparametrically identified. (They can be estimated by their sample analogues.) This means that the term $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, \nu_1)$ in (20) can be ignored when maximizing the likelihood or, in other words, valid inference can be obtained conditional on $\mathbf{Y}_{ip}$ and $\mathbf{X}_{ip}$.

2.7 Parametric DBR models

Adapting the ideas of Heckman (1981b), we shall model $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, v)$ using p "approximate reduced form" equations while the probabilities in the product in (20) are represented by a single common equation.[14] The p equations for $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, v)$ represent the probabilities of the first p outcomes as functions of previous outcomes, contemporaneous covariates, and unobserved heterogeneity. The final common equation is similar, except that previous outcomes are truncated at lag p. We refer to this as the DBR(p) model. The

[13]Note that $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, v)$ is defined in (21) using the most general specification for the probabilities (given covariates are assumed to be predetermined). In general $A^Y(\mathbf{y}_{ip}, \mathbf{x}_{ip}, v) \neq \mathsf{P}(\mathbf{Y}_{ip} = \mathbf{y}_{ip}|\mathbf{X}_{ip} = \mathbf{x}_{ip}, V_i = v)$.

[14]An alternative approach is to condition on the initial conditions (Wooldridge, 2005), resulting in a single-equation model.

most common DBR model used empirically adopts $p = 1$, although $p = 2$ is sometimes used in cases of either higher-frequency and/or longer-period data (e.g. Card and Hyslop, 2005, 2009). In the empirical case study in Section (3) we will consider both DBR(1) and DBR(2) models.

For the DBR(1) model, we have $Q = 2$. With Greek letters representing unknown parameters to be estimated, and G being the logistic function, the model specification is

$$\mathsf{P}(Y_{it} = 1|\mathbf{X}_{it} = \mathbf{x}_{it}, V_i = \nu_k) = G(\nu_{k1} + \beta_1' x_{it}) \equiv G_{it}^{11}(\nu_{k1}, \beta_1), \quad t = 1, \tag{22}$$

and

$$\begin{aligned}\mathsf{P}(Y_{it} = 1|\mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = \nu_k)& \\ = G(\nu_{k2} + \beta_2' x_{it} + \gamma_2 y_{it-1}) \equiv G_{it}^{12}(\nu_{k2}, \beta_2, \gamma_2),& \quad t = 2, \ldots, T.\end{aligned} \tag{23}$$

The corresponding likelihood contribution, cf. (20), for individual i is

$$\begin{aligned}&L_i^{DBR(1)}(\nu_1, \ldots, \nu_K, \pi_1, \ldots, \pi_K, \beta_1, \beta_2, \gamma_2) \\ &\quad = \sum_{k=1}^{K} \pi_k \Big[G_{it}^{11}(\nu_{k1}, \beta_1)^{y_{it}} \big(1 - G_{it}^{11}(\nu_{k1}, \beta_1)\big)^{1-y_{it}} \\ &\qquad \times \Big(\prod_{t=2}^{T} G_{it}^{12}(\nu_{k2}, \beta_2, \gamma_2)^{y_{it}} \big(1 - G_{it}^{12}(\nu_{k2}, \beta_2, \gamma_2)\big)^{1-y_{it}} \Big) \Big].\end{aligned} \tag{24}$$

The dimension of the parameters are as follows: $\beta_q \in \mathbb{R}^{\dim(x)}$ for $q = 1, 2$; and $\gamma_2 \in \mathbb{R}$.

In our DBR(2) model we relax the assumption of first-order Markovian state dependence and consider second-order Markovian state dependence. The second-order model naturally extends the first-order model above to include two equations corresponding to the first two outcomes. In the main equation, we include not only two lagged dependent variables but also an interaction term. With $Q = 3$, the model is

$$\mathsf{P}(Y_{it} = 1|\mathbf{X}_{it} = \mathbf{x}_{it}, V_i = \nu_k) = G(\nu_{k1} + \beta_1' x_{it}) \equiv G_{it}^{21}(\nu_{k1}, \beta_1), \quad t = 1, \tag{25}$$

$$\begin{aligned}
&\mathsf{P}(Y_{it} = 1 | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = \nu_k) \\
&\qquad = G(\nu_{k2} + \beta_2' x_{it} + \gamma_2 y_{i1}) \equiv G_{it}^{22}(\nu_{k2}, \beta_2, \gamma_2), \quad t = 2,
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
&\mathsf{P}(Y_{it} = 1 | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = \nu_k) \\
&\quad = G(\nu_{k3} + \beta_3' x_{it} + \gamma_{31} y_{it-1} + \gamma_{32} y_{it-2} + \gamma_{33} y_{it-1} y_{it-2}) \equiv G_{it}^{23}(\nu_{k3}, \beta_3, \gamma_3), \\
&\qquad t = 3, \ldots, T.
\end{aligned} \tag{27}$$

The corresponding likelihood contribution, c.f. (20), for individual i is

$$\begin{aligned}
&L_i^{DBR(2)}(\nu_1, \ldots, \nu_K, \pi_1, \ldots, \pi_K, \beta_1, \beta_2, \beta_3, \gamma_2, \gamma_3) \\
&\qquad = \sum_{k=1}^{K} \pi_k \Big[G_{it}^{21}(\nu_{k1}, \beta_1)^{y_{it}} \big(1 - G_{it}^{21}(\nu_{k1}, \beta_1)\big)^{1-y_{it}} \\
&\qquad \times G_{it}^{22}(\nu_{k2}, \beta_2, \gamma_2)^{y_{it}} \big(1 - G_{it}^{22}(\nu_{k2}, \beta_2, \gamma_2)\big)^{1-y_{it}} \\
&\qquad \times \Big(\prod_{t=3}^{T} G_{it}^{23}(\nu_{k3}, \beta_3, \gamma_3)^{y_{it}} \big(1 - G_{it}^{23}(\nu_{k3}, \beta_3, \gamma_3)\big)^{1-y_{it}} \Big) \Big].
\end{aligned} \tag{28}$$

The dimension of the parameters are as follows: $\beta_q \in \mathbb{R}^{\dim(x)}$ for $q = 1, 2, 3$; $\gamma_2 =\in \mathbb{R}$; and $\gamma_3 = (\gamma_{31}, \gamma_{32}, \gamma_{33}) \in \mathbb{R}^3$.

Several points are in order. First, in general G could be any known real function with range $[0, 1]$, while in practice, G is either the logistic or the standard normal distribution function.

Second, it is implicitly assumed that the parameters are the same in both states. That is, the covariate effects on the probability of being in state 1 at time t is assumed to be the same whether or not the individual is in state 0 or state 1 at time $t-1$. Several authors have pointed out the possibility of interacting covariates with the lagged dependent variables (e.g. Barmby, 1998, p263; Beck et al., 2002). However, few researchers have pursued this idea (one exception is Browning and Carro, 2010).

Third, the model with no unobserved heterogeneity is captured by $K = 1$, in which case the support points of V_i simply become constant terms. In the absence of unobserved heterogeneity the likelihood contribution is time-separable, and ν_{1Q}, β_Q and γ_Q can be consistently estimated from the last term in large parentheses.

Fourth, these specifications may also be appropriate in the absence of left-censoring. It is fair to say that the DBR literature has focused on modeling processes already in progress, and the initial conditions problem has been considered largely as an obstacle to inference (e.g. Heckman, 1981b). However, if the application has a clear time origin, there is no lagged dependent variable to condition on in the first period. Hence, even if the data are not left-censored, unless all individuals have the same initial outcomes, a different equation is still needed for the first measurement (as shown in Section 2.3). Thus, if there is no left-censoring and the specifications given in this subsection are used, then the parameters of the equations other than the last may have substantive structural interpretations as opposed to simply being flexible parameterizations of A^Y.

Fifth, it is instructive to consider the hazard rates implied by low-order DBR models. These are shown in Table 1 and Figure 1 for $p = 1, 2, 3$ in the simple case with no covariates, no unobserved heterogeneity, and without interactions between lagged outcomes. Table 1 shows the hazard rates out of state 0 and state 1 for relevant histories for each of these processes. Figure 1 graphs the time profiles of these state 0 and state 1 hazards for each case for the situation where a spell begins in period t.[15]

2.8 The MSD approach

The main interest in the MSD approach is usually the dependence of the transition probabilities on the elapsed time spent in the current state, i.e. duration dependence. For simplicity we ignore other forms of state dependence in the two MSD models discussed here. To present the models, let B_{it} denote the start-time of the spell ongoing at time t. Also, let F_i denote the time of the first observed transition and define $F_i = \infty$ if no transitions are observed for individual i.[16]

[15]Note that, in the $p = 3$ case, there are two possible hazard rates for each state shown in period $(t+1)$, depending on the state in period $(t-2)$ — i.e., whether the previous spell lasted more than one period or not.

[16]Formally, $B_{it} = \inf\{s : Y_{ik} = Y_{it} \text{ for } k = s, \dots, t\}$ and $F_i = \inf\{t : Y_{it-1} \neq Y_{it} \text{ and } 1 \leq t \leq T\}$.

The first specification is based on the assumption that outcomes prior to entering the current spell do not influence the transition probabilities. We also assume that only contemporaneous covariates matter. Formally, the assumption is that[17]

$$
\begin{aligned}
&\mathsf{P}(C_{it} = c_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \\
&\quad = \mathsf{P}(C_{it} = c_{it} | B_{it-1} = b_{it-1}, Y_{it-1} = y_{it-1}, X_{it} = x_{it}, V_i = v), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = f_i + 1, \dots, T. \quad (29)
\end{aligned}
$$

In general, equation (29) specifies separate equations for the transition probabilities out of each state, for non-left-censored "fresh spells" that are observed to start during the observation period. We refer to equation (29) as the MSD model's "structural equations" of interest. Equation (29) leaves the transition probabilities associated with the left-censored "initial spells" that are ongoing at the beginning of the observation period unrestricted (i.e. those for $t = 1, \dots, f_i$).

Under Assumption (29) and imposing our discrete distribution of unobserved heterogeneity, the likelihood contribution (18) can be written

$$
\begin{aligned}
L_i^C = \sum_{k=1}^{K} \pi_k \Bigg[& A^C(f_i, y_{i1}, \mathbf{x}_{if_i}, \nu_k) \\
& \times \left(\prod_{t=f_i+1}^{T} \mathsf{P}(C_{it} = c_{it} | B_{it-1} = b_{it-1}, Y_{it-1} = y_{it-1}, X_{it} = x_{it}, V_i = \nu_k) \right) \Bigg], \quad (30)
\end{aligned}
$$

where[18]

$$
\begin{aligned}
A^C(f_i, y_{i1}, \mathbf{x}_{if_i}, v) = {} & \mathsf{P}(Y_{i1} = y_{i1} | X_{i1} = x_{i1}, V_i = v) \\
& \times \left(\prod_{t=2}^{f_i} \mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \right). \quad (31)
\end{aligned}
$$

If $F_i = \infty$, then substitute T for f_i in (30) and (31). If the data are not left-censored,

[17] Note that $Y_{iB_{it}-1} \neq Y_{iB_{it}}$ and $Y_{iB_{it}} = Y_{ik}$ for $k = B_{it}, \dots, t$, so conditioning on $Y_{it-1} = y_{it-1}$ is here the same as conditioning on $Y_{iB_{it-1}} = y_{ib_{it-1}}$.

[18] Note that $y_{it} = y_{i1}$ for $t = 1, \dots, f_i - 1$ and $y_{if_i} \neq y_{i1}$, so it is not necessary to include $y_{i2}, \dots, y_{if_i}$ as separate arguments in A^C.

then the probabilities in (31) may be of interest; otherwise, they are simply necessary nuisance parameters.

Again, if there is no unobserved heterogeneity, the integration in (30) effectively disappears. In this case, the term $A^C(f_i, y_{i1}, \mathbf{x}_{if_i}, v)$ represents the contribution of the state occupied when first observed and the duration of the left-censored spell in progress at that time, while the probabilities in the big parentheses in (30) represent the contribution of the complete spells and the right-censored spell. The latter are obviously nonparametrically identified, and the probabilities can be estimated by sample analogues. Maximum likelihood estimation based on the non-left-censored spells is therefore consistent for those parameters.

The usefulness of Assumption (29), if satisfied, depends on the number of new spells that begin within the observation period. In applications with a lot of persistence, the number of transitions may be small relative to the number of measurements, and Assumption (29) may allow only few observations to be used for estimation. A stronger assumption combines (19) and (29). Specifically, in addition to (29) it is assumed that the effect of elapsed duration is constant after p periods in the spell so that, everything else equal, the hazard rate after p times is the same as the hazard rate at time p. To write this compactly, define $b^p_{it-1} = \max(b_{it-1}, t-p)$. Formally then, given $p \geq 1$, the assumption is that

$$\begin{aligned}
&\mathsf{P}(C_{it} = c_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \\
&\quad = \mathsf{P}(C_{it} = c_{it} | B_{it-1} \leq b^p_{it-1}, Y_{it-1} = y_{it-1}, X_{it} = x_{it}, V_i = v), \\
&\qquad\qquad\qquad\qquad\qquad\qquad t = f_i + 1, \ldots, T. \quad (32)
\end{aligned}$$

The power of Assumption (32) is that only parameters for times $1, \ldots, p$ depend on unmeasured variables. In other words, the assumption implies that all data after the first observed transition F_i or after time p, whichever is earlier, can contribute to identifying and estimating the parameters of interest. Collecting terms gives

$$L_i^C = \sum_{k=1}^{K} \pi_k \Bigg[A^D(f_i, y_{i1}, \mathbf{x}_{i\,\min(f_i,p)}, \nu_k) \Bigg(\prod_{t=\min(f_i+1,p+1)}^{T} \mathsf{P}(C_{it} = c_{it} | B_{it-1} \leq b_{it-1}^p, Y_{it-1} = y_{it-1}, X_{it} = x_{it}, V_i = \nu_k) \Bigg) \Bigg]. \quad (33)$$

where

$$A^D(f_i, y_{i1}, \mathbf{x}_{i\,\min(f_i,p)}, v) = \mathsf{P}(Y_{i1} = y_{i1} | X_{i1} = x_{i1}, V_i = v) \times \Bigg(\prod_{t=2}^{\min(f_i,p)} \mathsf{P}(Y_{it} = y_{it} | \mathbf{Y}_{it-1} = \mathbf{y}_{it-1}, \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = v) \Bigg). \quad (34)$$

Again, if the data are not left-censored, then the probabilities in (34) may be of interest; otherwise, they are simply necessary nuisance parameters.

2.9 Parametric MSD models

As discussed above, the standard MSD approach assumes that the transition probability at time t depends on the current state and the elapsed time in the current spell, but not on the history prior to entering that state. In the terminology of Heckman and Borjas (1980) this means that we allow for duration dependence, but not lagged duration dependence or occurrence dependence.[19]

We also assume that duration dependence only varies up to p periods of elapsed duration and then is constant, and refer to such models as the MSD(p) model. In particular, to capture duration dependence, we adopt a flexible specification with separate parameters for the first p-possible transition times in each state. A commonly used alternative is to specify a quadratic relationship for duration dependence (e.g. Ham and LaLonde, 1996; Beck et al., 1998, 2002).[20] In the case study in Section 3 we adopt $p = 6$. Within this framework, we consider both a five-equation model, MSD($6a$), that uses all the data

[19] By specifying separate transition probability equations for exiting each state, the MSD approach also allows for Markovian state dependence.

[20] Note that, Brown (1975) and Heckman and Borjas (1980) made the point that one can interact covariates with elapsed time in duration models, however this is rarely done.

available, and also a three-equation model, MSD($6b$), that does not model spell-hazard during the first 6 years of the sample period (details below).

Again we use Heckman's (1981b) ideas in modeling the A^D component. A fully flexible specification of the approximate reduced form for A^D would involve separate equations for each year a left-censored observation is in progress. Depending on the amount of left-censoring, this may or may not be feasible in practice. As a compromise between flexibility and feasibility, we model A^D using three equations representing the transition probabilities for the initial spells associated with each of the states, as well as the probability associated with the initial conditions. In other words, the largest model we consider has five equations: a reduced form equation for the initial state; two separate reduced form equations for modeling transitions from the initial spells; and two separate structural equations for modeling the probabilities of exiting state 1 versus state 0 spells.[21]

For the MSD($6a$) model, we have $Q = 5$. Thus each of the K types is represented by a 5-vector $\nu_k = (\nu_{k1}, \ldots, \nu_{k5}) \in \mathbb{R}^5$ for $k = 1, \ldots, K$. In this model, the A^D component in (34) is captured using three equations which represent the probability distribution of the initial state and, for each of the two states, the initial spell transition probabilities. The main probabilities in (33) are captured by two equations representing the transition probabilities out of subsequent spells, by state. In the MSD($6a$) specification, we allow the parameters representing duration dependence and the effects of covariates to vary freely across equations, and hence do not fully exploit Assumption (32). The model specification is

$$\mathsf{P}(Y_{it} = 1 | \mathbf{X}_{it} = \mathbf{x}_{it}, V_i = \nu_k) = G(\nu_{k1} + \beta_1' x_{it}) \equiv G_{it}^{11}(\nu_{k1}, \beta_1), \quad t = 1, \tag{35}$$

[21]Even with these restrictions imposed, a five-equation specification can be difficult to estimate, and in practice MSD models are rarely fully specified in this way. In particular, to our knowledge, the initial conditions problem is rarely considered in MSD models, and initial spells are often either dropped from the analysis or modeled using the same specifications as fresh spells. For example, Biewen (2006) and Devicienti (2011) ignore initial spells and condition on the initial state. Exceptions include Lacroix and Brouillette (2011), Ham and LaLonde (1996), and Eberwein et al. (1997), who model initial and fresh spells separately and have no initial conditions problem to deal with. Also, although Stevens (1999) does not model the initial conditions, she carefully considers the duration dependence associated with fresh spells, and this enables her to include initial spells within the two-equation specification for fresh spells.

$$\mathsf{P}(C_{it} = 1 | B_{it-1} \leq b^p_{it-1}, Y_{it-1} = 0, X_{it} = x_{it}, V_i = \nu_k)$$
$$= G(\nu_{k2} + \beta_2' x_{it} + \textstyle\sum_{r=2}^{p} \lambda_{2r}\, 1(t - b_{it-1} \geq r)) \equiv G^{12}_{it}(\nu_{k2}, \beta_2, \lambda_2),$$
$$t = 2, \ldots, f_i \quad (36)$$

$$\mathsf{P}(C_{it} = 1 | B_{it-1} \leq b^p_{it-1}, Y_{it-1} = 1, X_{it} = x_{it}, V_i = \nu_k)$$
$$= G(\nu_{k3} + \beta_3' x_{it} + \textstyle\sum_{r=2}^{p} \lambda_{3r}\, 1(t - b_{it-1} \geq r)) \equiv G^{13}_{it}(\nu_{k3}, \beta_3, \lambda_3),$$
$$t = 2, \ldots, f_i \quad (37)$$

$$\mathsf{P}(C_{it} = 1 | B_{it-1} \leq b^p_{it-1}, Y_{it-1} = 0, X_{it} = x_{it}, V_i = \nu_k)$$
$$= G(\nu_{k4} + \beta_4' x_{it} + \textstyle\sum_{r=2}^{p} \lambda_{4r}\, 1(t - b_{it-1} \geq r)) \equiv G^{14}_{it}(\nu_{k4}, \beta_4, \lambda_4),$$
$$t = f_i + 1, \ldots, T, \quad (38)$$

$$\mathsf{P}(C_{it} = 1 | B_{it-1} \leq b^p_{it-1}, Y_{it-1} = 1, X_{it} = x_{it}, V_i = \nu_k)$$
$$= G(\nu_{k5} + \beta_5' x_{it} + \textstyle\sum_{r=2}^{p} \lambda_{5r}\, 1(t - b_{it-1} \geq r)) \equiv G^{15}_{it}(\nu_{k5}, \beta_5, \lambda_5),$$
$$t = f_i + 1, \ldots, T. \quad (39)$$

The corresponding likelihood contribution for individual i is (c.f. (30))

$$
\begin{aligned}
&L_i^{MSD(6a)}(\nu_1,\ldots,\nu_K,\pi_1,\ldots,\pi_K,\beta_1,\ldots,\beta_5,\lambda_2,\ldots,\lambda_5)\\
&\quad=\sum_{k=1}^{K}\pi_k\Bigg[G_{i1}^{11}(\nu_{k1},\beta_1)^{y_{i1}}\big(1-G_{i1}^{11}(\nu_{k1},\beta_1)\big)^{1-y_{i1}}\\
&\quad\times\left(\prod_{t=2}^{f_i}G_{it}^{12}(\nu_{k2},\beta_2,\lambda_2)^{c_{it}}\big(1-G_{it}^{12}(\nu_{k2},\beta_2,\lambda_2)^{c_{it}}\big)^{1-c_{it}}\right)^{1-y_{it-1}}\\
&\quad\times\left(\prod_{t=2}^{f_i}G_{it}^{13}(\nu_{k3},\beta_3,\lambda_3)^{c_{it}}\big(1-G_{it}^{13}(\nu_{k3},\beta_3,\lambda_3)^{c_{it}}\big)^{1-c_{it}}\right)^{y_{it-1}}\\
&\quad\times\left(\prod_{t=f_i+1}^{T}G_{it}^{14}(\nu_{k4},\beta_4,\lambda_4)^{c_{it}}\big(1-G_{it}^{14}(\nu_{k4},\beta_4,\lambda_4)^{c_{it}}\big)^{1-c_{it}}\right)^{1-y_{it-1}}\\
&\qquad\qquad\times\left(\prod_{t=f_i+1}^{T}G_{it}^{15}(\nu_{k5},\beta_5,\lambda_5)^{c_{it}}\big(1-G_{it}^{15}(\nu_{k5},\beta_5,\lambda_5)^{c_{it}}\big)^{1-c_{it}}\right)^{y_{it-1}}\Bigg]. \qquad (40)
\end{aligned}
$$

The dimensions of the parameters are as follows: $\beta_q \in \mathbb{R}^{\dim(x)}$ and $\lambda_q = (\lambda_{q2},\ldots,\lambda_{q6}) \in \mathbb{R}^5$ for $q = 1,\ldots,5$. The MSD($6a$) model is relatively flexible and utilizes all available data, but this comes at the cost of having to estimate five equations and a large number of parameters.

In contrast, the MSD($6b$) model is more parsimonious, involving only three equations and hence fewer parameters, but at the cost that some data are ignored. For the MSD($6b$) model, we explicitly utilize Assumption (32) and assume that if $t \geq p + b_{it-1}$ then the parameters in equations (36) and (38) are the same and the parameters in equations (37) and (39) are the same. Furthermore, we ignore all data prior to period 5. This allows us to estimate a three-equation model ($Q = 3$), which represent the probability distribution of the initial state and the transition probabilities out of the state-specific spells. Thus, the likelihood contribution for individual i is

$$
\begin{aligned}
&L_i^{MSD(6a)}(\nu_1,\ldots,\nu_K,\pi_1,\ldots,\pi_K,\beta_1,\ldots,\beta_5,\lambda_2,\ldots,\lambda_5)\\
&\quad=\sum_{k=1}^{K}\pi_k\Bigg[G_{i7}^{11}(\nu_{k1},\beta_1)^{y_{i7}}\big(1-G_{i7}^{11}(\nu_{k1},\beta_1)\big)^{1-y_{i7}}\\
&\quad\times\left(\prod_{t=8}^{T}G_{it}^{14}(\nu_{k4},\beta_4,\lambda_4)^{c_{it}}\big(1-G_{it}^{14}(\nu_{k4},\beta_4,\lambda_4)^{c_{it}}\big)^{1-c_{it}}\right)^{1-y_{it-1}}\\
&\qquad\qquad\times\left(\prod_{t=8}^{T}G_{it}^{15}(\nu_{k5},\beta_5,\lambda_5)^{c_{it}}\big(1-G_{it}^{15}(\nu_{k5},\beta_5,\lambda_5)^{c_{it}}\big)^{1-c_{it}}\right)^{y_{it-1}}\Bigg],
\end{aligned}
\tag{41}
$$

where G_{it}^{11} is defined as in (35) but with $t=7$, and G_{it}^{14} and G_{it}^{15} are defined as in (38) and (39) but with $t=8,\ldots,T$.

Finally, in the absence of duration dependence, we note that the MSD models simplify. First, equations (36) and (38), and equations (37) and (39), are the same, so that the MSD(6a) and MSD(6b) models are equivalent. Second, and more importantly, the DBR(1) model is nested within the MSD model. More specifically, if the DBR(1) model is correct, then the effects of the covariates and unobserved heterogeneity in equations (38) and (39) will be symmetric (and there will be no duration terms). To see this, first note that (38) implies

$$
\mathsf{P}(Y_{it}=1|B_{it-1}\le b_{it-1}^{p},Y_{it-1}=0,X_{it}=x_{it},V_i=\nu_k)=G(\nu_{k4}+\beta_4'x_{it}),
\tag{42}
$$

and comparing this expression with the corresponding DBR(1) model probability expression given by equation (23) implies $\nu_{k4}+\beta_4=\nu_{k2}+\beta_2$. Similarly, exploiting the symmetry of G, (39) implies

$$
\mathsf{P}(Y_{it}=1|B_{it-1}\le b_{it-1}^{p},Y_{it-1}=1,X_{it}=x_{it},V_i=\nu_k)=G(-\nu_{k5}-\beta_5'x_{it}),
\tag{43}
$$

and comparing this expression with the corresponding DBR(1) model probability expression implies $-\nu_{k5}-\beta_5=\nu_{k2}+\beta_2+\gamma_2$ or that $-\nu_{k5}-\beta_5-\gamma_2=\nu_{k2}+\beta_2$. Then these two restrictions are satisfied if $\beta_4=-\beta_5$, and $\nu_{k4}=-\nu_{k5}$, allowing for an intercept difference between the MSD equations (which corresponds to the state dependence parameter γ_2).

3 Case study

We now apply each of the methods above to an empirical context, that of poverty persistence analyzed using duration models by Stevens (1999). Our focus in this example is to use this case study as an empirical setting to compare the results obtained from reasonably standard prototypical DBR and MSD models, and illustrate the differences associated with them, rather than to replicate or critique Stevens' original analysis. So, for example, we select a different analytical extract from the data provided to us than that used by Stevens (1999).

3.1 Data

Our analysis data consists of an extract of 5,248 individuals over the 20 years 1970–89 from the Panel Study of Income Dynamics (PSID).[22] Each individual's poverty status is determined by whether their family's annual income is below or above a needs threshold which depends on family size and composition, so that all individuals in a family have the same poverty status in that year.[23] The main selection criteria we apply to the sample is that all individuals experience at least one year in poverty over the extended period 1968–89 (PSID survey years), and are observed and have no missing outcome or covariate information over the analysis period 1970–89 (PSID survey years).[24] The covariates include dummy variables for age groups 0–5, 6–17, 18–24, and 55+, and dummy variables for whether the household head is female and/or black.

Tables 2 and 3 present descriptive statistics of the sample, summarized along three dimensions: first, at the person-year level; second, at the person-level, and third, at the person-spell level. The first panel of Table 2 shows the means of the covariates used in the

[22] The data we use come from the PSID survey years 1970–89, with the income and poverty measurements corresponding to calendar years 1969–88.

[23] See Stevens (1999) for more details of this and other data issues.

[24] The first criteria is used as a proxy to identify the poverty at-risk population, and follows Stevens (1999).

models for the full sample of observations and, in columns (2) and (3), for the subsamples defined by individuals' initial poverty state: i.e. in poverty (state 1) or not (state 0). This panel also shows that, on average, individuals were in poverty 35 percent of the time and had about a one-sixth (18 percent) chance of a poverty transition occurring in any year. However, if they were in poverty in the first year, they experienced poverty about 52 percent of the time, compared to 21 percent for individuals not initially in poverty.

The second panel of Table 2 shows that, on average, individuals experienced 3.35 poverty transitions over the sample period; equivalently, this implies they had 4.35 spells (poverty and non-poverty) on average. In Table 3, we briefly summarize the numbers of poverty and non-poverty spells (both separately and in total), and the average durations of these spells. The first rows show that there are more non-poverty than poverty spells and, on average, the non-poverty spells are longer (5.6 years compared to 3.4). The subsequent rows show these relative patterns hold for both initial (left-censored) and fresh (including right-censored) spells. The average observed durations of initial spells are roughly twice that of fresh spells (7.1 years compared to 3.8 years for poverty and non-poverty spells).

3.2 Estimation results

We present the estimates of the DBR and MSD models in Tables 4 and 5. Table 4 contains estimates for three DBR models: a first-order DBR model without unobserved heterogeneity, DBR(1a); and the first- and second-order DBR models with two discrete points of unobserved heterogeneity, DBR(1b) and DBR(2). Table 5 contains the estimates of the five-equation MSD model with two random effects mass points, MSD(6a), and the three-equation MSD model estimates which exploits the assumption of constant hazards after 6-years to estimate common equations for initial and fresh spells using data from 1970–89, MSD(6b).

We briefly discuss the DBR model results in Table 4. First, the estimates of the coefficients in the structural equations are consistent across the models: there is strong evidence of positive state dependence associated with poverty status; and non-prime aged (25–54) individuals, and those in female and black headed households, are more likely to be in poverty than other households. Second, although the approximate reduced form equations for the initial conditions do not have obvious interpretation, the coefficients on the covariates in these equations are generally of the same signs as those in the structural equations. (The only exceptions being some of the age-variables coefficients, but these are statistically insignificant in all cases.) Third, the results for the DBR models in Table 4 show, statistically, both unobserved heterogeneity and the more flexible second-order state dependence are important extensions in these models. (The likelihood ratio statistic for the models with versus without random effects is $LR = 2,919.8$ with 3 degrees of freedom (df), and the likelihood ratio statistic for the model with second-order versus first-order state dependence is $LR = 1,899.0$ with $9df$.)

We next discuss the MSD(6a) model results, presented in Table 5. The main (structural) estimates of interest are those in the fresh-spell equations for poverty entry and exit, and these indicate some substantive differences with the DBR models. The parameter estimates for the MSD(6b) model are broadly similar to those for the MSD(6a) estimates.[25] Our subsequent discussion of the MSD models will focus on the MSD(6a) specification.

The MSD models relax restrictions implied by the DBR model in two important respects, that we focus on here. First, one implication of the DBR model specification is that covariate coefficient magnitudes should be equal and opposite in sign in the entry and exit equations. In contrast, we find that, although the coefficients on the covariates are predominantly positive in the entry equation and negative in the exit equation (in line with the DBR estimates), there are some exceptions with the signs of the young and old age coefficients. In addition, there is also significant differences in magnitudes of the coefficients in these equations.

The second important restriction in the DBR model is that the order of state dependence implies that, for spell-durations longer than that order, the impact on the probability of a transition occurring should be zero; and furthermore, that the lower-order

[25]Eyeballing the estimates across the two specifications, the coefficients on 6+ years duration, for being aged 18–24, and for having a black head of household in the poverty entry equations, and the coefficients on being aged 0–5 in the exit equations are noticeably different. The difference in the 6+ years duration coefficients suggests the assumption of constant hazard beyond 6 years may be too strong.

elapsed-duration coefficients should be equal in the entry and exit transition equations. That the estimates of the elapsed-duration variables in both the entry and exit equations are statistically significant up to 6-plus years, strongly rejects both the first- and second-order state dependence specifications in the estimated DBR models. Also, all of the coefficients are negative, which implies that the hazard of a transition either into or out of poverty occurring is decreasing in the duration of the current spell. However, the coefficient magnitudes vary across the entry and exit equations.

Thus, perhaps not surprisingly, the MSD model provides a substantially better fit to the data than the DBR models. This is true in terms of the overall fit of the model; and also in terms of the more specific symmetry and duration-dependence restrictions implied by the DBR models.[26]

To explore the respective contributions of relaxing the DBR models' strict state dependence and symmetry effects of covariates and unobserved heterogeneity restrictions, we estimated a variety of model specifications between the DBR(1b) and MSD(6a) models. Table 6 presents a summary of these results. First, as discussed above, the DBR(1b) model is equivalent to an MSD model in which there is no duration dependence, and the effects of both the observed covariates and unobserved heterogeneity are symmetric on poverty entry and exit. This is the first model described in Table 6.

We next relax the symmetry restriction on both the covariates and unobserved heterogeneity, but maintain the constant-hazard rate restriction. This is the second model summarized in Table 6, and essentially introduces a third equation to allow separate specifications for the entry and exit transitions. The LR-statistic for the hypothesis that these symmetry restrictions are valid $(183.6,\ 7df)$ clearly rejects that hypothesis.[27]

The third model summarized in Table 6, relaxes the constant-hazard assumption but

[26]The LR-statistic of the difference between the MSD(6a) model and the nested DBR(1b) model is $3,231.4\ (43df)$. The DBR(2) model is not strictly nested within the MSD(6a) model, but the difference in the log-likelihood values, 666.2, is still large. The Vuong (1989) test statistic is 32.1 in favor of the MSD(6a) model. (The null is that both models are misspecified but fit equally well, and the statistic is asymptotically standard normally distributed.)

[27]Browning and Carro (2010) similarly reject the symmetry of effects in a first-order dynamic panel data model, although their focus is on showing heterogeneous state dependence effects that vary with the observable characteristics.

(re-)imposes the entry and exit symmetry restrictions.[28] In allowing for duration dependence, we include separate equations for the initial and fresh spells in this model. The LR-statistic for the hypothesis implied by the DBR(1b) model against this model is much greater ($1,187.4$, $19df$) than that associated with the DBR(1b) versus model 2, implying the assumption of constant hazards is a severe restriction.

The final intermediate specification we consider, relaxes the model 3 restriction of symmetric duration dependence effects on poverty entry and exit, while maintaining the symmetry of both the covariate and unobserved heterogeneity effects. Again the LR-statistic ($1,638.4$, $10df$) comparing models 3 and 4 is huge, implying that as well as duration dependence being important in poverty dynamics, the dependence on poverty entry and exit transitions are not symmetric. Finally, comparing model 4 with the full MSD(6a) model, summarized as the final model in Table 6, again shows that the hypothesis that the covariate and unobserved heterogeneity effects on poverty entry and exit are symmetric is easily rejected.

In summary, the estimation results presented here clearly show that the DBR models are too restrictive relative to the MSD alternatives. First, the Markovian state dependence implied by the DBR models is too strong against the MSD model's duration dependence alternative. Second, the implied symmetry of effects into and out of poverty transitions in the standard DBR(1) model is also strongly rejected. These conclusions are consistent with the results by Bhuller et al. (2014).

3.3 Prediction results

As well as comparing the estimates of the models, we also compare how they fit the data in the sense of their respective within-sample predictions. For example, it may be that, although the DBR models are rejected in favor of the MSD alternative, the predictive fit of the models may be substantially similar. For this purpose, we compare summary statistics of the actual data and model predictions, presented in Tables 7, 8, and 9.

In Table 7, we present summaries using two-way frequency tables of the number of

[28]We also impose symmetry of the duration dependence effects in this specification.

years that a person is poor and the number of transitions that occur between poverty and non-poverty, for each of the actual data and the predictions based on the first- and second-order DBR models with unobserved heterogeneity and the five-equation MSD model.[29]

The first panel of Table 7 summarizes the actual poverty experience of individuals over the sample period. This shows that there is substantial variation in poverty experience around the 7-year average: about 5 percent of the sample had no spells of poverty,[30] 18 percent had a single year poor, while at the other extreme, 3 percent of individuals were always in poverty, and the remaining three-quarters experienced between 2 and 19 years of poverty. Similarly, there was substantial heterogeneity in poverty transitions around the 3.35 average: corresponding to those who were either never or always poor, about 8 percent experience no transitions, while over 40 percent had at least 4 transitions.

The next three panels in Table 7 present analogous summaries of the predictions from the first- and second-order DBR models, and the MSD model respectively. First, the average incidence of poverty, or equivalently the number of years poor, is predicted well by each of the models: compared to the actual average of 0.353 (7.06 years over the 20 year sample), the DBR(1b) model average prediction is 0.350 (7.00 years), while the DBR(2) model's average is 0.351 (7.03 years) and the MSD(6a) model's average is 0.350 (7.01 years). The frequency distribution of the MSD model's predicted years poor is noticeably closer to the actual distribution than those of the DBR models. However, the models all substantially overpredict the number of individuals who have no poverty experience, and underpredict the number with a single year; in addition, the DBR models also substantially underpredict the number of individuals who are always in poverty.

Second, the models also accurately predict the average number of transitions: the DBR(1b), DBR(2) and MSD(6a) model average predicted transitions are 3.36, 3.41 and

[29]In order to obtain manageable summaries and limit the extent of small cell frequencies, we group the number of years poor as 0, 1, 2–5, 6–10, 11–15, 16–19, and 20 years, and the number of transitions 0, 1, 2, 3, 4+ even (so the initial and final states are the same), and 5+ odd (so the initial and final states are different). In this two-way tabulation, some cells are necessarily null (e.g. if a person is either never or always poor, they will experience no transitions); and similarly, if a person is poor in only 1 year, they must experience either 1 or 2 transitions. The predictions from each of the models are based on 20 simulations per individual taking the covariates as given.

[30]Recall this is among those individuals were experienced some poverty between survey years 1968 and 1989.

3.36 respectively, compared to 3.35 actual transitions. However, associated with overpredicting the zero poverty incidence, the models overpredict the number of cases with zero transitions and dramatically underpredict the incidence of 1–3 transition cases. For indicative purposes, we have constructed Pearson goodness-of-fit statistics for each of the models based on the tables of actual and predicted frequencies in Table 7. The MSD model's goodness-of-fit statistic (98.9, $23df$) is substantially lower than those of the two DBR models. Thus, although this implies the MSD model does not provide an adequate statistical fit to the data using conventional significance levels, the relative magnitudes are consistent with the MSD model fitting substantially better than the two DBR models, and second-order DBR model fitting better than the first-order model.

In Table 8 we present a different summary of the actual and predicted poverty experiences from the DBR(1b) and MSD(6a) models.[31] The table shows the frequency distribution of the number of distinct spells experienced over the sample period separately by the initial state. The actual experiences include up to 15 separate spells, while the maximum number of spells predicted by the DBR(1b) and the MSD(6a) models is 17. Note that the total number of transitions given in Table 7 equals the sum of the number of spells for those initially not poor and those in poverty minus one. Table 8 shows that the tendency for the models to overpredict the frequency of single spells and underpredict the frequencies of 2 and 3 poverty spells is especially strong for those whose initial state is not-in-poverty. On the other hand, the models fit a little better for those who are poor initially; this is particularly true for the MSD(6a) model.

In Table 9 we present the average durations of the actual and predicted spells. The averages of the MSD(6a) model predictions are again closer to the actual spell average durations than those of the DBR(1b) model. Thus, these prediction results are consistent with the estimation results indicating the MSD model fits better than the DBR models.

[31] We exclude the DBR(2) model predictions here as these are comparatively similar to those of the first-order model.

4 Concluding remarks

This paper shows that the commonly used alternative dynamic panel data and spell duration approaches to modeling longitudinal binary outcomes can be viewed as special cases of a general analytical framework. Each of these approaches typically focus on and emphasize different aspects of the state dependence properties of the data. In addition, the dynamic panel data models are generally much more tightly specified than the multi-spell duration model alternatives.

The case study analysis of poverty experiences provided some clear conclusions regarding the relative efficacy of the models. The MSD model fits the data better than the first- and second-order DBR models. That this is the case in terms of the model estimates is perhaps not surprising given the MSD model is substantially more flexibly specified than the parsimonious DBR models, and especially the DBR(1) is strictly nested within the MSD model specification. However, this conclusion also holds in terms of the model predictions: i.e. the MSD model predictions were substantially better than the predictions from the DBR model.

Finally, the analysis of intermediate models between the parsimonious DBR(1) model and general MSD model implies that the DBR(1) is overly restrictive both in terms of the state dependence, and also in terms of the symmetric effects of covariates and unobserved heterogeneity on transitions into and out of poverty, that it imposes.

References

Allison, P. (1982). Discrete time methods for the analysis of event histories. In S. Leinhardt (Ed.), *Sociological Methodology 1982*, Chapter 2, pp. 61–97. San Francisco: Jossey-Bass.

Arulampalam, W., A. L. Booth, and M. P. Taylor (2000). Unemployment persistence. *Oxford Economic Papers 52*, 24–50.

Bane, M. J. and D. T. Ellwood (1983). The dynamics of dependence: The routes to self sufficiency. Technical report, The U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.

Barmby, T. (1998). The relationship between event history and discrete time duration models: An application to the analysis of personnel absenteeism. *Oxford Bulletin of Economics and Statistics 60*(2), 261–265.

Beck, N., D. Epstein, S. Jackman, and S. O'Halloran (2002). Alternative models of dynamics in binary time-series-cross-section models: The example of state failure. Paper presented at the annual meeting of the political methodology group, Emory University.

Beck, N. and J. Katz (1997). The analysis of binary time-series-cross-section data and/or the democratic peace. Paper presented at the annual meeting of the political methodology group, Ohio State University.

Beck, N., J. N. Katz, and R. Tucker (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Political Science Review 42*(4), 1260–1288.

Bhuller, M., C. N. Brinch, and S. Königs (2014). Time aggregation and state dependence in welfare receipt. Discussion paper 3, Statistics Norway Research Department.

Biewen, M. (2006). Who are the chronic poor? An econometric analysis of chronic poverty in Germany. In *Dynamics of Inequality and Poverty*, Volume 13 of *Research on Economic Inequality*. Emerald Group Publishing Limited.

Brown, C. C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics 31*, 863–872.

Browning, M. and J. M. Carro (2010). Heterogeneity in dynamic discrete choice models. *Econometrics Journal 13*, 1–39.

Capellari, L. and S. P. Jenkins (2004). Modelling low income transitions. *Journal of Applied Econometrics 19*(5), 593–610.

Cappellari, L., R. Dorsett, and G. Haile (2007). State dependence, duration dependence and unobserved heterogeneity in the employment transitions of the over-50s. Working paper 2007-16, Institute for Social and Economic Research.

Card, D. and D. R. Hyslop (2005). Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica 73*(6), 1723–1770.

Card, D. and D. R. Hyslop (2009). The dynamic effects of an earnings subsidy for longterm welfare recipients: Evidence from the self sufficiency project applicant experiment. *Journal of Econometrics 153*(1), 1–20.

Chay, K. Y. and D. R. Hyslop (2014). Identification and estimation of dynamic binary response panel data models: Empirical evidence using alternative approaches. *Research in Labor Economics 39*, 1–39.

Contoyannis, P., A. M. Jones, and N. Rice (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics 19*, 473–503.

Devicienti, F. (2011). Estimating poverty persistence in Britain. *Empirical Economics 40*, 657–686.

Eberwein, C., J. C. Ham, and R. J. Lalonde (1997). The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: Evidence from experimental data. *The Review of Economic Studies 64*(4), 655–682.

Halliday, T. J. (2008). Heterogeneity, state dependence and health. *Econometrics Journal 11*(3), 499–516.

Ham, J. C. and R. J. LaLonde (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica 64*(1), 175–205.

Heckman, J. J. (1978). Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence. *Annales de l'INSEE 30–31*, 227–269.

Heckman, J. J. (1981a). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in Labor Markets*, pp. 91–139. Chicago and London: University of Chicago Press.

Heckman, J. J. (1981b). The incidental parameter problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 179–195. Cambridge, Massachusetts: MIT Press.

Heckman, J. J. (1981c). Statistical models for discrete panel data. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Economic Applications*, pp. 114–178. Cambridge, Massachusetts: MIT Press.

Heckman, J. J. and G. J. Borjas (1980). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica 47*, 247–283.

Heckman, J. J. and B. Singer (1984). Econometric duration analysis. *Journal of Econometrics 24*, 63–132.

Honoré, B. (1993). Identification results for duration models with multiple spells. *Review of Economic Studies 60*, 241–246.

Horowitz, J. L. and S. Lee (2004). Semiparametric estimation of a panel data proportional hazards model with fixed effects. *Journal of Econometrics 119*(1), 155–198.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation behavior of married women. *Econometrica 67*(6), 1255–1294.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics 57*(1), 129–138.

Lacroix, G. and D. Brouillette (2011). Assessing the impact of a wage subsidy for single parents on social assistance. *Canadian Journal of Economics 44*(4), 1195–1221.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Econometric Society Monographs No. 17. Cambridge; New York: Cambridge University Press.

Stevens, A. H. (1999). Climbing out of poverty, falling back in: Measuring the persistence of poverty over multiple spells. *Journal of Human Resources 34*, 557–588.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics 20*, 39–54.

Zabel, J., S. Schwartz, and S. Donald (2010). The impact of the Self-Sufficiency Project on the employment behaviour of former welfare recipients. *Canadian Journal of Economics 43*(3), 883–918.

Appendix A Spell-based representations

As mentioned, in duration analysis and event history approaches, the data are often represented as transition times or spells instead of a sequence of state indicators. In this literature, the data are often considered as measurement in continuous time, in which case the spell-based representations are the only ones practical. In discrete time, spell-based representations may be computationally efficient if covariates are constant within each spell. Time-based representations may be more convenient if the covariates are time-varying rather than spell-varying.

In this appendix, we show that spell-based and time-based data representations are equivalent. The nonparametrically likelihood functions are also equivalent in the sense that there is a one-to-one relationship between the different parameter sets. The final subsection discusses covariates.

A.1 Data representation

Let J_i denote the number of transitions observed between time 1 and time T; that is,

$$J_i = \sum_{t=2}^{T} C_{it}. \tag{44}$$

Let Z_{ij} for $j = 1, \ldots, J_i$ denote the measured times of change of state (spell endings). Then $(Y_{i1}, Z_{i1}, \ldots, Z_{iJ_i}, T)$ is an equivalent complete representation of the data. To see this, note that the transition times can be defined recursively (assuming $J_i > 0$) by[32]

$$\begin{aligned} Z_{i1} &= \inf\{t \in \mathbb{N} : 1 \leq t < T,\ Y_{i1} \neq Y_{it+1}\}, \\ Z_{ij} &= \inf\{t \in \mathbb{N} : Z_{j-1} + 1 \leq t < T,\ Y_{iZ_{j-1}+1} \neq Y_{it+1}\}, \quad j = 2, \ldots, J_i. \end{aligned} \tag{45}$$

[32]Note that for $j = 1, \ldots, J_i$ we have $Z_{ij} = t \Rightarrow C_{it+1} = 1$, and similarly for $t = 2, \ldots, T$, we have $C_{it} = 1 \Rightarrow \exists j \in \mathbb{N} : 1 \leq j \leq J_i,\ Z_{ij} = t - 1$.

Conversely, the state indicators can be recovered from the transition times by

$$Y_{it} = \begin{cases} 1 - Y_{it-1} & \text{if } \exists j \in \mathbb{N} : 1 \leq j \leq J_i,\ Z_{ij} = t-1, \\ Y_{it-1} & \text{otherwise,} \end{cases} \qquad t = 2, \ldots, T, \tag{46}$$

or by

$$Y_{it} = \left(Y_{i1} + \sum_{j=1}^{J_i} 1(Z_{ij} < t) \right) \bmod 2, \quad t = 2, \ldots, T. \tag{47}$$

For simplicity, in this paper we have assume all histories are right-censored at time T. Since we do not know the state at time $T+1$, we therefore do not know whether or not there is a transition at time T.

The data can also be represented as a panel of durations. Let D_{ij} for $j = 1, \ldots, J_i$ denote the duration of the jth spell. Formally,

$$D_{ij} = Z_{ij} - Z_{ij-1}, \quad j = 2, \ldots, J_i. \tag{48}$$

If $J_i > 0$, we may also define the (possibly left-censored) duration at the beginning of the measurement period by $D_{i1} = Z_{i1}$ and the (possibly right-censored) duration at the end of the observation period by $D_{iJ_i+1} = T - Z_{iJ_i}$. If $J_i = 0$, define $D_{i1} = T$. Then $(Y_{i1}, D_{i1}, \ldots, D_{iJ_i}, D_{iJ_i+1})$ is an equivalent representation of the data.

Example Suppose $T = 4$ and the state occupancy indicators are $Y_{i1} = 0$, $Y_{i2} = 0$, $Y_{i3} = 1$, and $Y_{i4} = 1$. Then there is one transition and two spells, which can be represented in four different ways: the first representation is $(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}) = (0, 0, 1, 1)$, the second is $(Y_{i1}, C_{i2}, C_{i3}, C_{i4}) = (0, 0, 1, 0)$, the third is $(Y_{i1}, Z_{i1}, Z_{i2}) = (0, 2, 4)$, and the fourth representation is $(Y_{i1}, D_{i1}, D_{i2}) = (0, 2, 2)$.

A.2 Parameterization

The likelihood contribution in the MSD approach can also be written in a spell-based form instead of the time-based form given in (8).[33] The notation becomes slightly more involved, since the number of spells may vary across individuals. First, let $\mathbf{Z}_{ij}$ denote the random outcome history at the time of the jth transition (the end of the jth spell) for individual i; that is, define $\mathbf{Z}_{i0} = (Y_{i1})$ and $\mathbf{Z}_{ij} = (Y_{i1}, Z_{i1}, \ldots, Z_{ij})$ for $j = 1, \ldots, J_i$. Let $\mathbf{z}_{i0} = (y_{i1})$ and $\mathbf{z}_{ij} = (y_{i1}, z_{i1}, \ldots, z_{ij})$ denote the observed history. Second, let $\mathbb{Z}_{t-1}^{j-1}$ denote the space of possible prior histories when spell j is in progress at time t. If there has been no transitions before time t, we have $\mathbb{Z}_{t-1}^{0} = \{0, 1\}$ for $t = 1, \ldots, T-1$, and if there has been one previous transition, then $\mathbb{Z}_{t-1}^{1} = \{0, 1\} \times \{1, \ldots, t-1\}$ for $t = 2, \ldots, T-1$. For $j = 2, \ldots, t-1$ and $t = 2, \ldots, T-1$, the space of possible prior histories is $\mathbb{Z}_{t-1}^{j-1} = \{0, 1\} \times \{1, \ldots, t-j\} \times \cdots \times \{j-1, \ldots, t-1\}$. Let $\mathbf{z}_{j-1}$ with no subscript i denote a generic element of $\mathbb{Z}_{t-1}^{j-1}$. Then the conditional probability of beginning in state 1 given the history and the hazard rates at each time t are defined as[34]

$$\begin{aligned}
&\chi = \mathsf{P}(Y_{i1} = 1), \\
&\varphi_t(\mathbf{z}_{j-1}) = \mathsf{P}(Z_{ij} = t | Z_{ij} \geq t, \mathbf{Z}_{ij-1} = \mathbf{z}_{j-1}), \\
&\qquad\qquad\qquad\qquad \mathbf{z}_{j-1} \in \mathbb{Z}_{t-1}^{j-1}, \quad j = 1, \ldots, t, \quad t = 1, \ldots, T-1.
\end{aligned} \tag{49}$$

Of course, there are also $2^T - 1$ distinct probabilities in this representation. This can be verified as follows: given T and j with $0 \leq j \leq T-1$, there is $T-1$ choose j possible transition times; by the binomial formula there are then a total of $\sum_{j=0}^{T-1} \binom{T-1}{j} = 2^{T-1}$ possible transition times; each of which can begin in either of the two states, so 2^T possible outcomes; and one probability is determined by the adding-up constraint, so $2^T - 1$ probabilities.

[33] The time-based form is standard in the continuous-time literature, see e.g. Honoré (1993) and Horowitz and Lee (2004).

[34] Allison (1982, p92) defined the discrete-time hazard rate for repeated events, but did not provide the likelihood function.

Example For $\tau = 3$, there are 7 parameters depending on time t and the history prior to t; namely,

$$\begin{aligned}
\chi &= \mathsf{P}(Y_{i1} = 1), \\
\varphi_1(0) &= \mathsf{P}(Z_{i1} = 1 | Z_{i1} \geq 1, (Y_{i1}) = (0)) \text{ with } \mathbf{z}_0 = (0) \in \mathbb{Z}_0^0, \\
\varphi_1(1) &= \mathsf{P}(Z_{i1} = 1 | Z_{i1} \geq 1, (Y_{i1}) = (1)) \text{ with } \mathbf{z}_0 = (1) \in \mathbb{Z}_0^0, \\
\varphi_2(0) &= \mathsf{P}(Z_{i1} = 2 | Z_{i1} \geq 2, (Y_{i1}) = (0)) \text{ with } \mathbf{z}_0 = (0) \in \mathbb{Z}_1^0, \\
\varphi_2(1) &= \mathsf{P}(Z_{i1} = 2 | Z_{i1} \geq 2, (Y_{i1}) = (1)) \text{ with } \mathbf{z}_0 = (1) \in \mathbb{Z}_1^0, \\
\varphi_2(0,1) &= \mathsf{P}(Z_{i2} = 2 | Z_{i2} \geq 2, (Y_{i1}, Z_{i1}) = (0,1)) \text{ with } \mathbf{z}_1 = (0,1) \in \mathbb{Z}_1^1, \\
\varphi_2(1,1) &= \mathsf{P}(Z_{i2} = 2 | Z_{i2} \geq 2, (Y_{i1}, Z_{i1}) = (1,1)) \text{ with } \mathbf{z}_1 = (1,1) \in \mathbb{Z}_1^1.
\end{aligned}$$

For $\tau = 4$, there are additionally 8 parameters; namely,

$$\begin{aligned}
\varphi_3(0) &= \mathsf{P}(Z_{i1} = 3 | Z_{i1} \geq 3, (Y_{i1}) = (0)) \text{ with } \mathbf{z}_0 = (0) \in \mathbb{Z}_2^0, \\
\varphi_3(1) &= \mathsf{P}(Z_{i1} = 3 | Z_{i1} \geq 3, (Y_{i1}) = (1)) \text{ with } \mathbf{z}_0 = (1) \in \mathbb{Z}_2^0, \\
\varphi_3(0,1) &= \mathsf{P}(Z_{i2} = 3 | Z_{i2} \geq 3, (Y_{i1}, Z_{i1}) = (0,1)) \text{ with } \mathbf{z}_1 = (0,1) \in \mathbb{Z}_2^1, \\
\varphi_3(1,1) &= \mathsf{P}(Z_{i2} = 3 | Z_{i2} \geq 3, (Y_{i1}, Z_{i1}) = (1,1)) \text{ with } \mathbf{z}_1 = (1,1) \in \mathbb{Z}_2^1, \\
\varphi_3(0,2) &= \mathsf{P}(Z_{i2} = 3 | Z_{i2} \geq 3, (Y_{i1}, Z_{i1}) = (0,2)) \text{ with } \mathbf{z}_1 = (0,2) \in \mathbb{Z}_2^1, \\
\varphi_3(1,2) &= \mathsf{P}(Z_{i2} = 3 | Z_{i2} \geq 3, (Y_{i1}, Z_{i1}) = (1,2)) \text{ with } \mathbf{z}_1 = (1,2) \in \mathbb{Z}_2^1, \\
\varphi_3(0,1,2) &= \mathsf{P}(Z_{i3} = 3 | Z_{i3} \geq 3, (Y_{i1}, Z_{i1}, Z_{i2}) = (0,1,2)) \text{ with } \mathbf{z}_2 = (0,1,2) \in \mathbb{Z}_2^2, \\
\varphi_3(1,1,2) &= \mathsf{P}(Z_{i3} = 3 | Z_{i3} \geq 3, (Y_{i1}, Z_{i1}, Z_{i2}) = (1,1,2)) \text{ with } \mathbf{z}_2 = (1,1,2) \in \mathbb{Z}_2^2.
\end{aligned}$$

The likelihood contribution of individual i can be built up by considering the initial state and each subsequent transition separately. To simplify the notation, it is customary to state the likelihood using a "latent" variable, Z_{iJ_i+1}, which represents an $(J_i + 1)$th transition. It is imagined that this transition would have been observed, were it not for right-censoring. We then have[35]

$$L_i^Z = \chi^{y_{i1}}(1-\chi)^{1-y_{i1}}\left(\prod_{j=1}^{j_i}\mathsf{P}(Z_{ij}=z_{ij}|\mathbf{Z}_{ij-1}=\mathbf{z}_{ij-1})\right)\mathsf{P}(Z_{ij_i+1}\geq T|\mathbf{Z}_{ij_i}=\mathbf{z}_{ij_i}), \quad (50)$$

where

$$\mathsf{P}(Z_{ij}=z_{ij}|\mathbf{Z}_{ij-1}=\mathbf{z}_{ij-1}) = \varphi_{z_{ij}}(\mathbf{z}_{ij-1})\prod_{t=z_{ij-1}+1}^{z_{ij}-1}(1-\varphi_t(\mathbf{z}_{ij-1})), \quad j=1,\ldots,j_i, \quad (51)$$

and

$$\mathsf{P}(Z_{ij_i+1}\geq T|\mathbf{Z}_{ij_i}=\mathbf{z}_{ij_i}) = \prod_{t=z_{ij_i}+1}^{T-1}(1-\varphi_t(\mathbf{z}_{ij_i})). \quad (52)$$

In (50), the first term on the right-hand side is the contribution of the initial state, the term in large parentheses is the contribution of the j_i observed transitions, and the last term is the contribution of the fact that no event took place between z_{ij_i} and T.

The two representations of the MSD likelihood contributions (8) and (50) are of course equivalent. In particular, the parameters are one-to-one and the likelihood values are identical. To verify the first claim, given t and $\mathbf{y}_t \in \mathbb{Y}^t$ arguments similar to those given in Section 2.2 can be used to deduce $\mathbf{z}_{j-1} \in \mathbb{Z}_{t-1}^{j-1}$, where either $j=1$ and $0<t\leq z_1$ or $j>1$ and $z_{j-1}<t\leq z_j$. Conversely, given t and $\mathbf{z}_{j-1}\in\mathbb{Z}_{t-1}^{j-1}$, it is straightforward to deduce $\mathbf{y}_t \in \mathbb{Y}^t$. Therefore, given t and j and compatible histories $\mathbf{y}_t\in\mathbb{Y}^t$ and $\mathbf{z}_{j-1}\in\mathbb{Z}_{t-1}^{j-1}$, we have the one-to-one relationship between parameters

$$z_{j-1}<t\leq z_j \;\Rightarrow\; \varphi_t(\mathbf{z}_{j-1}) = \xi_{t+1}(\mathbf{y}_t). \quad (53)$$

Intuitively, the conditional probability of spell j ending at t given prior history is the same as the conditional probability of a transition between t and $t+1$; or in other words the hazard rates can be expressed in terms of C_{it}s or Z_{ij}s.

[35]It is possible to state the likelihood without the use of a latent variable, by noting that the probability of no events taking place between z_{ij_i} and T is the same as $\mathsf{P}(J_i=j_i|\mathbf{Z}_{ij_i}=\mathbf{z}_{ij_i}, T=T)$.

To verify that the likelihood values are identical, $L_i^C = L_i^Z$, note that (assuming $j_i > 0$)

$$\begin{aligned}
L_i^Z &= \chi^{y_{i1}}(1-\chi)^{1-y_{i1}}\left(\prod_{j=1}^{j_i}\prod_{t=z_{ij-1}+1}^{z_{ij}} \varphi_t(\mathbf{z}_{ij-1})^{1(t=z_{ij})}(1-\varphi_t(\mathbf{z}_{ij-1}))^{1(t\neq z_{ij})}\right)\\
&\qquad\qquad \times\left(\prod_{t=z_{ij_i}+1}^{T-1} 1-\varphi_t(\mathbf{z}_{ij_i})\right)\\
&= \chi^{y_{i1}}(1-\chi)^{1-y_{i1}}\\
&\qquad\qquad \times\left(\prod_{j=1}^{j_i}\prod_{t=1}^{T-1} 1(z_{ij-1} < t \leq z_{ij})\varphi_t(\mathbf{z}_{ij-1})^{1(t=z_{ij})}(1-\varphi_t(\mathbf{z}_{ij-1}))^{1(t\neq z_{ij})}\right)\\
&\qquad\qquad \times\left(\prod_{t=1}^{T-1} 1(z_{ij_i} < t \leq T-1)(1-\varphi_t(\mathbf{z}_{ij_i}))\right)\\
&= \chi^{y_{i1}}(1-\chi)^{1-y_{i1}}\left(\prod_{t=1}^{T-1}\prod_{j=1}^{j_i} 1(z_{ij-1} < t \leq z_{ij})\xi_{t+1}(\mathbf{y}_{it})^{1(t=z_{ij})}(1-\xi_{t+1}(\mathbf{y}_{it}))^{1(t\neq z_{ij})}\right)\\
&\qquad\qquad \times\left(\prod_{t=1}^{T-1} 1(z_{ij_i} < t \leq T-1)(1-\xi_{t+1}(\mathbf{y}_{it}))\right)\\
&= \chi^{y_{i1}}(1-\chi)^{1-y_{i1}}\prod_{t=1}^{T-1}\xi_{t+1}(\mathbf{y}_{it})^{c_{it+1}}(1-\xi_{t+1}(\mathbf{y}_{it}))^{1-c_{it+1}}\\
&= L_i^C. \qquad (54)
\end{aligned}$$

This shows that the likelihood contribution in the multi-spell duration approach can be expressed equivalently either in a j- or a t-dimension.

Example Suppose $Y_{i1} = 0$, $Y_{i2} = 0$, $Y_{i3} = 1$, and $Y_{i4} = 1$ with $T = 4$. Then

$$\begin{aligned}
L_i^Y &= \mathsf{P}(Y_{i1}=0)\mathsf{P}(Y_{i2}=0|\mathbf{Y}_{i1}=(0))\\
&\qquad\qquad \times \mathsf{P}(Y_{i3}=1|\mathbf{Y}_{i2}=(0,0))\mathsf{P}(Y_{i4}=1|\mathbf{Y}_{i3}=(0,0,1)), \quad (55)
\end{aligned}$$

while

$$\begin{aligned}
L_i^Z &= \mathsf{P}(Y_{i1}=0)\big(1-\mathsf{P}(Z_{i1}=1|Z_{i1}\geq 1,\mathbf{Z}_{i0}=(0))\big)\\
&\qquad \times \mathsf{P}(Z_{i1}=2|Z_{ij}\geq 2,\mathbf{Z}_{i0}=(0))\big(1-\mathsf{P}(Z_{i2}=3|Z_{i2}\geq 3,\mathbf{Z}_{i1}=(0,1))\big). \quad (56)
\end{aligned}$$

Note that

$$\mathsf{P}(Y_{i2}=0|\mathbf{Y}_{i1}=(0)) = 1-\mathsf{P}(Z_{i1}=1|Z_{i1}\geq 1,\mathbf{Z}_{i0}=(0)), \tag{57}$$

$$\mathsf{P}(Y_{i3}=1|\mathbf{Y}_{i2}=(0,0)) = \mathsf{P}(Z_{i1}=2|Z_{ij}\geq 2,\mathbf{Z}_{i0}=(0)), \tag{58}$$

$$\mathsf{P}(Y_{i4}=1|\mathbf{Y}_{i3}=(0,0,1)) = 1-\mathsf{P}(Z_{i2}=3|Z_{i2}\geq 3,\mathbf{Z}_{i1}=(0,1)). \tag{59}$$

The two nonparametric likelihood representations are therefore equivalent.

A.3 Likelihood contribution with covariates

Suppose the covariates are constant within each spell and only vary between spells. For $j=0,\ldots,j_i+1$, let X^*_{ij} denote the vector of spell-constant covariates, and let $\mathbf{X}^*_{ij}$ and $\mathbf{x}^*_{ij}$ denote the random and observed covariate histories up to (and including) spell j.[36] Then the likelihood contribution for individual i in the spell-based form becomes

$$\begin{aligned} L_i^Z = \chi^{y_{i1}}(1-\chi)^{1-y_{i1}}&\left(\prod_{j=1}^{j_i}\mathsf{P}(Z_{ij}=z_{ij}|\mathbf{Z}_{ij-1}=\mathbf{z}_{ij-1},\mathbf{X}^*_{ij-1}=\mathbf{x}^*_{ij-1})\right)\\ &\times\mathsf{P}(Z_{ij_i+1}\geq T|\mathbf{Z}_{ij_i}=\mathbf{z}_{ij_i},\mathbf{X}^*_{ij_i}=\mathbf{x}^*_{ij_i}), \end{aligned} \tag{60}$$

It can be shown that this is equivalent to (11) and (12).

Conceptually, it makes no difference if the covariates are not spell-constant, but time-varying; however, the expression for the likelihood contribution is more complicated. With time-varying covariates, the likelihood contribution for individual i becomes

$$\begin{aligned} L_i^Z = \mathsf{P}(Y_{i1}=y_{i1}|X_{i1}=x_{i1})&\\ \times&\left(\prod_{j=1}^{j_i}\mathsf{P}(Z_{ij}=z_{ij}|Z_{ij}\geq z_{ij},\mathbf{Z}_{ij-1}=\mathbf{z}_{ij-1},\mathbf{X}_{iz_{ij}}=\mathbf{x}_{iz_{ij}})\right.\\ \times&\left.\prod_{t=z_{ij-1}+1}^{z_{ij}-1}\big(1-\mathsf{P}(Z_{ij}=t|Z_{ij}\geq t,\mathbf{Z}_{ij-1}=\mathbf{z}_{ij-1},\mathbf{X}_{it}=\mathbf{x}_{it})\big)\right)\\ \times&\left(\prod_{t=z_{ij_i}+1}^{T-1}\big(1-\mathsf{P}(Z_{ij_i+1}=t|Z_{ij_i+1}\geq t,\mathbf{Z}_{ij_i}=\mathbf{z}_{ij_i},\mathbf{X}_{it}=\mathbf{x}_{it})\big)\right). \end{aligned} \tag{61}$$

[36]Endogenous spell-varying covariates is beyond the scope of this paper.

In general, it is not possible to simplify further. If the covariates remain constant within each spell, then (61) simplifies to (60). In terms of computing time and memory requirements, (60) is likely to be more efficient than (61).

Table 1: Hazard rates for DBR models

Hazard out of state 0 at time t				Hazard out of state 1 at time t			
History			Hazard	History			Hazard
$t-3$	$t-2$	$t-1$	t	$t-3$	$t-2$	$t-1$	t
DBR(1): $\mathsf{P}(Y_{it}=1\vert\mathbf{Y}_{it-1}=\mathbf{y}_{it-1})=G(\gamma_0+\gamma_1 y_{it-1})$							
		0	$G(\gamma_0)$			1	$1-G(\gamma_0+\gamma_1)$
DBR(2): $\mathsf{P}(Y_{it}=1\vert\mathbf{Y}_{it-1}=\mathbf{y}_{it-1})=G(\gamma_0+\gamma_1 y_{it-1}+\gamma_2 y_{it-2})$							
	1	0	$G(\gamma_0+\gamma_2)$		0	1	$1-G(\gamma_0+\gamma_1)$
	0	0	$G(\gamma_0)$		1	1	$1-G(\gamma_0+\gamma_1+\gamma_2)$
DBR(3): $\mathsf{P}(Y_{it}=1\vert\mathbf{Y}_{it-1}=\mathbf{y}_{it-1})=G(\gamma_0+\gamma_1 y_{it-1}+\gamma_2 y_{it-2}+\gamma_2 y_{it-3})$							
1	1	0	$G(\gamma_0+\gamma_2+\gamma_3)$	0	0	1	$1-G(\gamma_0+\gamma_1)$
0	1	0	$G(\gamma_0+\gamma_2)$	1	0	1	$1-G(\gamma_0+\gamma_3)$
1	0	0	$G(\gamma_0+\gamma_3)$	0	1	1	$1-G(\gamma_0+\gamma_1+\gamma_2)$
0	0	0	$G(\gamma_0)$	1	1	1	$1-G(\gamma_0+\gamma_1+\gamma_2+\gamma_3)$

Figure 1: Hazard rates for DBR models

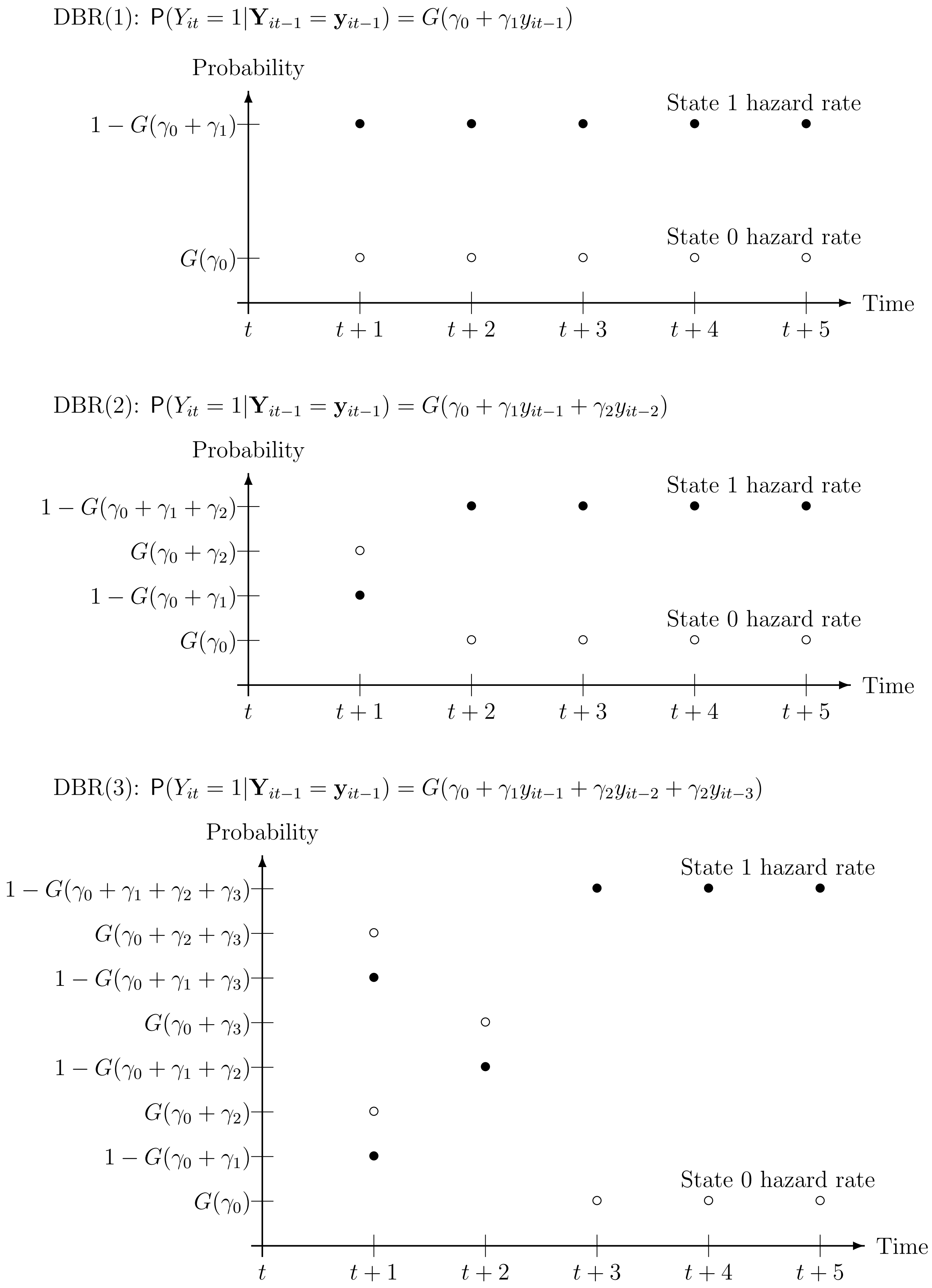

Table 2: Descriptive statistics by initial state

	Full sample	Initial state	
		Not poor	In-poverty
Person-years: means and standard deviations			
Aged 0–5	0.025	0.026	0.025
	(.0005)	(.0005)	(.0005)
Aged 6–17	0.225	0.213	0.239
	(.001)	(.002)	(.002)
Aged 18–24	0.204	0.188	0.223
	(.001)	(.002)	(.002)
Aged 25–54	0.420	0.432	0.406
	(.002)	(.002)	(.002)
Aged 55+	0.126	0.143	0.107
	(.001)	(.001)	(.001)
Female head	0.336	0.270	0.411
	(.001)	(.002)	(.002)
Black head	0.582	0.411	0.775
	(.002)	(.002)	(.002)
Poor (Y_{it})	0.353	0.208	0.517
	(.001)	(.002)	(.002)
Transition (C_{it})	0.177	0.168	0.186
	(.001)	(.002)	(.002)
No. person-years	104,960	55,700	49,260
Persons: means and standard deviations			
Transitions	3.35	3.20	3.53
	(.032)	(.043)	(.048)
No. persons	5,248	2,785	2,463

Table 3: Descriptive statistics by poverty status

	Full sample	Poverty status	
		Not poor	In-poverty
All spells: means and standard deviations			
Duration	4.59	5.63	3.44
	(.033)	(.050)	(.038)
No. spells	22,849	12,062	10,787
Initial spells: means and standard deviations			
Duration	7.11	8.23	5.86
	(.084)	(.121)	(.111)
No. spells	5,248	2,785	2,463
Fresh spells: means and standard deviations			
Duration	3.84	4.85	2.72
	(.032)	(.051)	(.032)
No. spells	17,601	9,277	8,324

No adjustments for censoring.

Table 4: Dynamic binary response model estimates

	DBR(1a)		DBR(1b)		DBR(2)		
	IC	Strl	IC	Strl	IC1	IC2	Strl
Variables							
y_{it-1}		2.708		2.191		2.426	1.859
		0.017		0.019		(.077)	(.030)
y_{it-2}							0.942
							(.031)
$y_{it-1}y_{it-2}$							0.039
							(.042)
Aged 0–5	0.183	0.357	0.253	0.549	0.218	0.163	0.328
	(.097)	(.064)	(.106)	(.072)	(.104)	(.130)	(.089)
Aged 6–17	0.463	0.395	0.601	0.560	0.558	0.165	0.449
	(.077)	(.023)	(.083)	(.029)	(.083)	(.093)	(.030)
Aged 18–24	0.220	0.070	0.397	0.186	0.349	–0.13	0.109
	(.102)	(.023)	(.110)	(.027)	(.109)	(.119)	(.027)
Aged 55+	0.020	0.314	–0.131	0.272	–0.105	–0.236	0.288
	(.150)	(.027)	(.159)	(.034)	(.158)	(.175)	(.035)
Female head	0.957	0.717	1.076	0.935	1.085	0.744	0.874
	(.068)	(.018)	(.074)	(.024)	(.074)	(.083)	(.024)
Black head	1.446	0.607	1.429	0.620	1.480	0.855	0.527
	(.064)	(.018)	(.071)	(.029)	(.070)	(.081)	(.028)
Random effects (mass points and probabilities)							
ν_1	–1.519	–2.599	–2.178	–3.186	–2.121	–2.686	–3.206
	(.075)	(.020)	(.093)	(.033)	(.096)	(.108)	(.036)
ν_2			–0.654	–1.604	–0.767	–1.607	–1.926
			(.091)	(.036)	(.095)	(.109)	(.041)
π_1				0.638			0.640
				(.012)			(.017)
Statistics							
No. persons		5248		5248			5248
No. years		20		20			20
Log L		–46520.1		–45060.2			–44110.7
LR				2,919.8			1,899.0
(df)				(3)			(11)

y_{it} indicates poverty in year t; IC: initial conditions equation; Strl: structural equation; LR: likelihood ratio statistics; df: degrees of freedom. For DBR(1b), the null hypothesis for the LR statistic is DBR(1a). For DBR(2), the null hypothesis for the LR statistic is DBR(1b).

Table 5: Multi-spell duration model estimates

	MSD(6a)					MSD(6b)		
	IC	Initial spells		Fresh spells		IC	Fresh spells	
		Entry	Exit	Entry	Exit		Entry	Exit
Variables								
$1(d_{it} \geq 2)$		–0.070	–0.425	–0.507	–0.562		–0.509	–0.544
		(.082)	(.083)	(.044)	(.044)		(.047)	(.047)
$1(d_{it} \geq 3)$		–0.490	0.234	–0.365	–0.188		–0.373	–0.163
		(.101)	(.095)	(.059)	(.060)		(.061)	(.063)
$1(d_{it} \geq 4)$		0.007	–0.046	–0.186	–0.290		–0.228	–0.321
		(.116)	(.104)	(.074)	(.078)		(.075)	(.079)
$1(d_{it} \geq 5)$		0.003	–0.028	–0.051	–0.169		–0.068	–0.258
		(.121)	(.118)	(.087)	(.098)		(.078)	(.085)
$1(d_{it} \geq 6)$		0.257	–0.051	–0.309	–0.056		0.029	–0.055
		(.093)	(.099)	(.076)	(.090)		(.059)	(.066)
Aged 0–5	0.156	0.332	–0.068	–0.340	–0.049	0.036	–0.206	–1.124
	(.103)	(.113)	(.121)	(.205)	(.187)	(.184)	(.450)	(.639)
Aged 6–17	0.512	–0.098	–0.440	0.511	–0.188	0.497	0.431	–0.272
	(.082)	(.059)	(.066)	(.051)	(.049)	(.083)	(.045)	(.047)
Aged 18–24	0.314	0.513	0.401	0.111	0.169	0.073	0.310	0.197
	(.108)	(.059)	(.068)	(.043)	(.043)	(.099)	(.038)	(.041)
Aged 55+	–0.078	0.047	–0.289	0.366	–0.287	0.150	0.327	–0.289
	(.157)	(.074)	(.108)	(.054)	(.051)	(.137)	(.047)	(.050)
Female head	1.086	0.906	–0.732	0.881	–0.817	1.195	0.838	–0.802
	(.075)	(.050)	(.054)	(.040)	(.040)	(.073)	(.036)	(.038)
Black head	1.529	0.246	–0.871	0.713	–0.498	1.379	0.407	–0.493
	(.071)	(.048)	(.069)	(.044)	(.039)	(.077)	(.036)	(.040)
Random effects (mass points and probabilities)								
ν_1	–2.148	–2.403	0.332	–2.610	1.161	–2.592	–2.154	1.003
	(.107)	(.082)	(.111)	(.071)	(.063)	(.110)	(.069)	(.069)
ν_2	–0.901	–1.678	–0.615	–1.143	0.121	–1.180	–0.870	0.029
	(.094)	(.092)	(.088)	(.059)	(.047)	(.112)	(.081)	(.055)
π_1					0.590			0.677
					(.024)			(.030)
Statistics								
No. persons					5,248			5,248
No. years					20			16
Log L					–43,444.5			–34,621.8

d_{it}: elapsed duration in current spell at the end of year t; IC: initial conditions equation; Entry: structural equation for entering poverty; Exit: structural equation for exiting poverty.

Table 6: Differences between the DBR(1b) and MSD(6a) models

	Model description	Pars	Log L	LR	(df)
1	DBR(1b) model, same as a MSD model with no duration dependence, and symmetric entry and exit effects	18	–45,060.2		
2	MSD model with no duration dependence, and flexible entry and exit effects (test 1 against 2)	25	–44,968.7	183.6	(7)
3	MSD model with duration dependence, and symmetric duration dependence, covariate and unobserved heterogeneity effects on entry and exit (test 1 against 3)	37	–44,466.5	1,187.4	(19)
4	MSD model with duration dependence, and symmetric covariate and unobserved heterogeneity effects on entry and exit (test 3 against 4)	47	–43,647.3	1,638.4	(10)
5	MSD(6a) model, i.e. with duration dependence, and flexible effects on entry and exit (test 4 against 5)	61	–43,444.5	405.6	(14)

Pars: number of parameters; LR: likelihood ratio statistics; df: degrees of freedom.

Table 7: Predictions of years in poverty and transitions

No. years poor	No. transitions						
					4+	5+	
	0	1	2	3	Even	Odd	Total
Actual data							
0	255	0	0	0	0	0	255
1	0	201	735	0	0	0	936
2–5	0	232	315	291	526	168	1,532
6–10	0	88	68	209	325	322	1,012
11–15	0	55	38	95	299	290	777
16–19	0	88	155	92	190	62	587
20	149	0	0	0	0	0	149
Total	404	664	1,311	687	1,340	842	5,248
DBR(1b) predictions							
0	473.2	0	0	0	0	0	473.2
1	0	123.3	388.6	0	0	0	511.8
2–5	0	131.4	320.1	369.0	580.0	205.1	1,605.5
6–10	0	31.0	62.4	194.6	475.5	458.6	1,221.9
11–15	0	23.0	53.7	131.8	365.8	274.0	848.2
16–19	0	49.9	179.2	79.1	172.2	37.4	517.7
20	69.9	0	0	0	0	0	69.9
Total	543.0	358.5	1,003.8	774.4	1,593.4	975.0	5,248.0
Pearson GOF-statistic = 973.5 (23*df*)							
DBR(2) predictions							
0	527.5	0	0	0	0	0	527.5
1	0	113.8	440.7	0	0	0	554.5
2–5	0	166.8	228.2	334.0	588.0	212.9	1,529.9
6–10	0	68.0	84.0	208.0	406.1	402.6	1,168.6
11–15	0	44.4	72.8	136.4	315.5	265.0	834.1
16–19	0	58.6	179.1	76.7	179.6	45.1	539.0
20	94.6	0	0	0	0	0	94.6
Total	622.1	451.5	1,004.7	755.0	1,489.1	925.6	5,248.0
Pearson GOF-statistic = 613.6 (23*df*)							
MSD(6a) predictions							
0	336.4	0	0	0	0	0	336.4
1	0.0	156.4	625.0	0	0	0	781.4
2–5	0.0	217.0	318.9	306.5	582.9	167.3	1,592.5
6–10	0.0	105.2	73.4	187.8	341.6	350.5	1,058.4
11–15	0.0	68.5	54.4	121.5	288.2	262.0	794.5
16–19	0.0	67.6	175.0	75.2	188.1	43.1	548.9
20	136.1	0	0	0	0	0	136.1
Total	472.5	614.5	1,246.6	690.9	1,400.8	822.8	5,248.0
Pearson GOF-statistic = 98.9 (23*df*)							

GOF: goodness of fit; *df*: degrees of freedom. Poverty rates and incidence rates can be computed by taking the number of years in poverty and the number of transitions and divide by the number of years under observation.

Table 8: Predictions of spells by initial state

No. spells	Actual data Initial state		DBR(1b) predictions Initial state		MSD(6a) predictions Initial state	
	Not poor	In-poverty	Not poor	In-poverty	Not poor	In-poverty
1	255	149	473.2	69.9	336.4	136.1
2	159	505	99.2	259.3	132.7	481.8
3	1,089	222	739.0	264.8	989.7	257.0
4	214	473	230.4	544.0	203.0	487.9
5	455	271	600.0	370.8	518.5	265.3
6	129	398	211.2	458.3	158.7	323.3
7	219	155	261.8	225.5	227.4	175.4
8	81	139	84.4	174.2	79.3	155.6
9	119	73	60.9	61.1	81.8	79.5
10	31	47	14.9	28.5	31.9	51.0
11	21	16	6.2	6.2	21.2	22.2
12	6	10	1.2	2.5	8.2	11.7
13	7	3	0.5	0.5	3.5	4.8
14	0	1	0.0	0.0	1.1	1.9
15	0	1	0.0	0.0	0.4	0.8
16	0	0	0.0	0.0	0.1	0.2
17	0	0	0.0	0.0	0.0	0.1
Total	2785	2463	2,782.7	2,465.4	2,793.7	2,454.3
GOF			564.6	481.9	66.1	32.6
(df)			(11)	(11)	(11)	(11)

GOF: Pearson goodness-of-fit statistic conditional on the initial state, with cells 12–17 combined.

Table 9: Predictions spell type and poverty status

	Actual data		DBR(1b) prd		MSD(6a) prd	
	Initial spells	Fresh spells	Initial spells	Fresh spells	Initial spells	Fresh spells
Status: not poor						
Avg spell duration	8.23	4.85	8.43	4.70	8.10	4.92
No. spells	2,785	9,277	2,783	9,510	2,794	9,266
Status: in-poverty						
Avg spell duration	5.86	2.72	4.48	2.96	5.88	2.67
No. spells	2,463	8,324	2,465	8,684	2,454	8,368

Prd: predictions; avg: average.