Statistica Neerlandica (2001) Vol. 55, nr. 2, pp. 249–262

Estimating transition probabilities from a time series of independent cross sections

Ben Pelzer*

Research Technical Department, University of Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

Rob Eisinga

Department of Social Science Research Methods, University of Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

Philip Hans Franses

Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

This paper considers the implementation of a nonstationary, heterogeneous Markov model for the analysis of a binary dependent variable in a time series of independent cross sections. The model, previously considered by MOFFITT (1993), offers the opportunity to estimate entry and exit transition probabilities and to examine the effects of time-constant and time-varying covariates on the hazards. We show how ML estimates of the parameters can be obtained by Fisher's method-of-scoring and how to estimate both fixed and time-varying covariate effects. The model is exemplified with an analysis of the labor force participation decision of Dutch women using data from the Socio-economic Panel (SEP) study conducted in the Netherlands between 1986 and 1995. We treat the panel data as independent cross sections and compare the employment status sequences predicted by the model with the observed sequences in the panel. Some open problems concerning the application of the model are also discussed.

Key Words and Phrases: repeated cross sections, pseudo-panel data, Markov model.

1 Introduction

The increasing availability of repeated cross-sectional (RCS) surveys not only provides researchers with a growing opportunity to analyze over-time change but also

* b.pelzer@maw.kun.nl. Our thanks to Marno Verbeek for his helpful comments on a previous draft of this paper. The data for *Socio-economic Panel* (*SEP*) utilized in this paper were collected by STATISTICS NETHERLANDS (NCBS) and were made available by the SCIENTIFIC STATISTICAL AGENCY (WSA) of the NETHERLANDS ORGANIZATION FOR SCIENTIFIC RESEARCH (NWO).

raises questions about new methodology for exploiting these data for longitudinal study. RCS data contain information on different cross-sectional units (typically individuals) independently drawn from a population at multiple points in time and aim to provide a representative cross section of the population at each sample point. A limitation of this type of data for longitudinal research is that the sample units are not retained from one time period to the next. RCS data are therefore, in the context of dynamic modeling, generally regarded as inferior to genuine panel data, that is, repeated observations on the same units across occasions. Obviously, an important advantage to using a matched panel file is that it provides a measure of gross individual change for each sample unit and that it enables us to use each unit as its own control. Panel data, however, may also be inferior to repeated cross sections in terms of sample size, representativeness, and time period covered. The size of a panel is commonly reduced over time by the process of selective attrition, which may create serious biases in the analysis. Especially in the case of long-term panel surveys the panel may become unrepresentative as time proceeds. Moreover, logistical constraints often preclude tracking individual units through long periods of time, so that analyzing rolling cross-sectional data for the assessment of long-run change is the best we can do.

This paper discusses, for the case of a binary dependent variable, a dynamic model previously treated briefly by MOFFITT (1990, 1993) that permits the estimation of entry and exit transition rates from a time series of RCS samples. The model also offers the opportunity to examine the effects of covariates on the hazards. It is therefore likely to be useful to researchers seeking to explain over-time change at the micro level in the absence of microlevel data. The paper is organized as follows. Section 2 discusses the model, parameter estimation and some refinements of the model. Section 3 provides an example application using panel data on female labor force participation taken from the Socio-economic Panel (SEP) study conducted in the Netherlands between 1986 and 1995. We treat the panel data as independent cross sections and compare the predictions of the Markov model for RCS data with the observations in the panel. Section 4 concludes.

2 Dynamic model for RCS data

The problem of analyzing repeated cross-sectional data has attracted increasing attention in econometrics and other disciplines in the last several years. One class of models considered is the linear fixed effect model (BALTAGI, 1995; COLLADO, 1997; DEATON, 1985; GIRMA, 2000, 2001; NIJMAN and VERBEEK, 1990; VERBEEK, 1996; VERBEEK and NIJMAN, 1992, 1993; VERBEEK and VELLA, 2000). In this approach individual observations are grouped into cohorts based on a time-invariant characteristic (typically date of birth) which results in a so-called pseudo panel with cohort aggregates. The studies are concerned with the conditions under which we can validly ignore the cohort nature of the averaged data and treat the pseudo panel of cohorts as if it were a panel of individuals. MOFFITT (1993) has generalized this

approach by considering models with a more dynamic structure and binary dependent variables. In his method actual grouping of the data into cohorts need not be done and the variation in the micro data is utilized as part of the analytic procedure. This section elaborates his method. It is assumed in the sequel that the responses are observed at equally spaced discrete time intervals $t = 1, 2, \ldots$ and that the samples at periods t_j and t_k are independent if $j \neq k$. Other discussions of the model include FELTEAU *et al.* (1997) and MEBANE and WAND (1997).

2.1 *First-order Markov model*

Suppose we have the following two-state first-order Markov matrix of transition rates in which the cell probabilities sum to unity across rows

$$\begin{array}{cc} & \begin{array}{cc} y_{it} & \\ 0 & 1 \end{array} \\ \begin{array}{c} y_{it-1} \\ \begin{array}{c} 0 \\ 1 \end{array} \end{array} & \left(\begin{array}{cc} 1-\mu_{it} & \mu_{it} \\ \lambda_{it} & 1-\lambda_{it} \end{array} \right) \end{array}$$

This expression records the probabilities of making each of the possible transitions from one time period to the next; e.g., μ_{it} represents the probability that the unit satisfying $y_i = 0$ at time $t-1$ subsequently satisfies $y_i = 1$ at time t. Recall that the first-order Markov process assumes that the underlying process of change can be described in terms of one-step transitions, i.e., the probability of occupying a state at time t depends only on the state occupied at time $t-1$. This assumption implies that the dependency between successive transitions can be eliminated by conditioning on the previous state. Operationally this can be achieved by including the previous state in the model as a covariate predicting y_{it}. Also note that, if we let

$$p_{it} = P(Y_{it} = 1),\ \mu_{it} = P(Y_{it} = 1 | Y_{it-1} = 0),\ \text{ and } \lambda_{it} = P(Y_{it} = 0 | Y_{it-1} = 1)$$

then we have

$$E(Y_{it}) = p_{it} = \mu_{it}(1 - p_{it-1}) + (1 - \lambda_{it})p_{it-1} = \mu_{it} + \eta_{it} p_{it-1}, \tag{1}$$

where $\eta_{it} = 1 - \lambda_{it} - \mu_{it}$. The accounting identity in (1) is the elemental equation for estimating dynamic models with repeated cross-sectional samples as it relates the marginal probabilities p_{it} and p_{it-1} to the probabilities of inflow (μ_{it}) and outflow (λ_{it}) from each of the two states. Obviously, the difficulty with using cross-sectional surveys is that the state-to-state transitions over time for each sample unit are not observed, but rather one observes at each of a number of times a distinct cross section of units and their current states. This implies that identification of the unobserved transitions over time in RCS data is only possible with the imposition of certain restrictions over *i* and/or *t*.

A popular restriction is to assume that the transition probabilities are both time-stationary and unit-homogeneous, hence $\mu_{it} = \mu$ and $\lambda_{it} = \lambda$ for all i and t. Using $\eta = 1 - \lambda - \mu$, it is easy to show that the long-run outcome of p_{it} based on t sets of successive transitions is $p_{it} = (\mu/(\mu + \lambda))(1 - \eta^t) + \eta^t p_{i0}$, which collapses to

$p_{it} = \mu/(\mu + \lambda)$ as t goes to infinity. The limiting result for p_{it} gives the long-run probability of being in a state, i.e., for a time point sufficiently far in the future the probability that the state is 1 is $\mu/(\mu + \lambda)$. Note that this probability does not depend on the initial probability p_{i0}. Hence there is a tendency as time passes for the probability of being in a state to be independent of the initial condition. Moreover, as noted by MOFFITT (1993), the initial probability refers to the value of the state prior to the beginning of the Markov process, for example the state of being unemployed at the beginning of an unemployment spell, rather than to the first observed outcome (which is p_{i1}). It is therefore assumed in many applications to finite-horizon situations that $p_{i0} = 0$ (see, e.g., BISHOP, FIENBERG and HOLLAND, 1975). This time-invariant steady state model is the standard approach to the problem of estimating transition rates from aggregate frequency data in the statistical literature (see, e.g., FIRTH, 1982; HAWKINS, HAN and EISENFELD, 1996; KALBFLEISH and LAWLESS, 1984, 1985; LAWLESS and MCLEISH, 1984; LEE, JUDGE and ZELLNER, 1970; LI and KWOK, 1990). The formulation has been applied in several economic studies, for example, by TOPEL (1983) in his study on employment duration and by MCCALL (1971) in his Markovian analysis of earnings mobility. Similar uses occur in the social science literature on intra-generational job mobility processes where it has come to be known as the 'mover-stayer' model (see, e.g., BARTHOLOMEW, 1996; GOODMAN, 1961).

Because the assumption of stationarity and homogeneity is generally not plausible and frequently violated in applications, it is desirable to relax this restriction. If we define the model as in (1) and let $p_{i0} = 0$ (or $t \to \infty$), it is easy to verify that p_{it} has the representation

$$p_{it} = \mu_{it} + \sum_{\tau=1}^{t-1} \left[\mu_{i\tau} \prod_{s=\tau+1}^{t} \eta_{is} \right], \tag{2}$$

where $\eta_{is} = 1 - \lambda_{is} - \mu_{is}$. This reduced form equation for p_{it} accounts for time-dependence and heterogeneity in a flexible manner and it will therefore be maintained in the ensuing method.

To estimate the model in (2) with RCS data, MOFFITT (1990, 1993) uses the following estimation procedure. While repeated cross-sections lack direct information on the individual transitions, they often do provide a set of time-invariant or time-varying covariates X_{it} that affect the hazards. If so, the history of these covariates (X_{it}, X_{it-1}, ..., X_{i1}) can be employed to generate backward predictions for the transition probabilities (μ_{it}, μ_{it-1}, ..., μ_{i1} and λ_{it}, λ_{it-1}, ..., λ_{i2}) and thus for the marginal probabilities (p_{it}, p_{it-1}, ..., p_{i1}). Hence the basic idea is to model the current and past μ_{it}'s and λ_{it}'s in a regression setting as functions of current and backcasted values of time-invariant and time-varying covariates X_{it}. Parameter estimates of the covariates are thereupon obtained by substituting the hazard functions into (2).

A common specification for the hazard functions in panel studies uses a separate

binary logistic regression for $P(Y_{it} = 1|Y_{it-1} = y_{it})$, $y_{it} = 0, 1$. That is, we assume that

$$\text{logit}\, P(Y_{it} = 1|Y_{it-1} = 0) = \text{logit}\,(\mu_{it}) = X_{it}\beta, \text{ and}$$

$$\text{logit}\, P(Y_{it} = 1|Y_{it-1} = 1) = \text{logit}\,(1 - \lambda_{it}) = X_{it}\beta^*,$$

where β and β^* are two potentially different sets of parameters. Hence the model assumes that the effects of the covariates will differ depending on the previous response. A condensed form for the same general model is

$$\text{logit}\, P(Y_{it} = 1|Y_{it-1} = y_{it-1}) = X_{it}\beta + y_{it-1}X_{it}\alpha, \tag{3}$$

where $\alpha = \beta^* - \beta$. This equation expresses the two regressions as a single dynamic model that includes as predictors both the previous response y_{it-1} (given that the intercept vector is included in X_{it}) and the interaction of y_{it-1} and the covariates X_{it}. Note that the transition matrix varies across both individuals and time periods because the hazards depend on the current and backcasted values of the covariates. Theoretical uses of (3) for panel data occur in AMEMIYA (1985), DIGGLE, LIANG and ZEGER (1994), and HAMERLE and RONNING (1995). BOSKIN and NOLD (1975) offer an application of a heterogeneous but stationary model with exogenous variables to the case of turnover in welfare based on panel data. See TOIKKA (1976) for an application of a three-state Markov model with exogenous variables to labor market choices (employed, unemployed and searching for a job, and withdrawal from employment) in which the transitions are estimated using frequency data disaggregated by sex.

According to equation (3) the transition rates are $\mu_{it} = F(X_{it}\beta)$ and $\lambda_{it} = 1 - F[X_{it}(\alpha + \beta)]$, where F – in this article – is the logistic function. Maximum likelihood estimates of α and β can be obtained by maximization of the log likelihood function

$$LL = \sum_{t=1}^{T}\sum_{i=1}^{n_t}[y_{it}\log(p_{it}) + (1 - y_{it})\log(1 - p_{it})], \tag{4}$$

with respect to the parameters, with p_{it} defined by (2). As indicated by MOFFITT (1993), obtaining p_{it} by means of the reduced form equation is equivalent to 'integrating out' over all possible transition histories for each individual i at time t to derive an expression for the marginal probability p_{it}. A graphical presentation of the model illustrating this is given in Figure 1, omitting the subscript i for clarity.

The marginal probability p_{it} depends on the set of all possible transition histories for each individual i up to time t. That is, p_{it} is a polynomial in μ_{it} and λ_{it}. The unobserved transition probabilities themselves are modeled as functions of current and backcasted values of time-invariant and time-varying covariates X_{it}. Hence an important feature of the model is that the transition probabilities and the marginal probabilities are estimated as a function of all the available cross sections rather than simply the observations from the current time period. Thus estimates of the

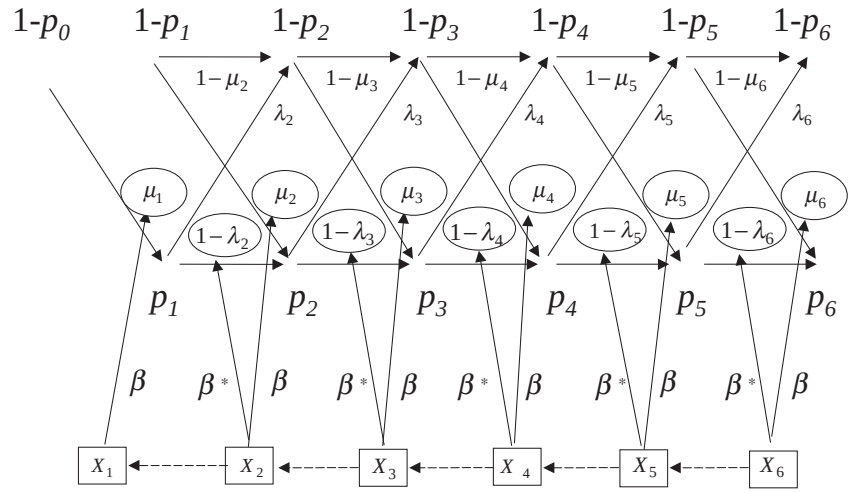

Fig. 1. Graphical illustration of Markov model for RCS data

transitions at the beginning of the Markov chain, for example, are not determined solely by the sample obtained for the first time period but by all the samples.

2.2 *ML estimation*

Maximum likelihood fitting of the model in equation (2) requires the derivatives of the likelihood function (4) with respect to the parameters. For ease of exposition, subscript i is omitted in the expressions of the derivatives and equation (2) is re-written as

$$p_t = \sum_{\tau=1}^{t} \left[\mu_\tau \left(\prod_{s=\tau}^{t} \eta_s \right) \eta_\tau^{-1} \right], \tag{5}$$

where $\mu_\tau = (1 + e^{-x_\tau \beta})^{-1}$, $\eta_s = 1 - \lambda_s - \mu_s$, $\lambda_s = (1 + e^{x_s(\alpha+\beta)})^{-1}$, and x_τ and x_s the current and backcasted values of the covariates at $t = \tau$ and $t = s$, respectively. The first order partial derivatives of p_t in equation (5) with respect to the parameters β and α are

$$\frac{\partial p_t}{\partial \beta} = \sum_{\tau=1}^{t} \frac{\partial \mu_\tau}{\partial \beta} \left(\prod_{s=\tau}^{t} \eta_s \right) \eta_\tau^{-1} + \sum_{\tau=1}^{t-1} \sum_{s=\tau+1}^{t} \eta_\tau \frac{\partial \eta_s}{\partial \beta} \left(\prod_{\gamma=\tau+1}^{t} \eta_\gamma \right) \eta_s^{-1}, \text{ and}$$

$$\frac{\partial p_t}{\partial \alpha} = \sum_{\tau=1}^{t-1} \sum_{s=\tau+1}^{t} \mu_\tau \frac{\partial \eta_s}{\partial \alpha} \left(\prod_{\gamma=\tau+1}^{t} \eta_\gamma \right) \eta_s^{-1}, \tag{6}$$

respectively, where $\partial \mu_\tau / \partial \beta = x_\tau (1 - \mu_\tau) \mu_\tau$, $\partial \eta_s / \partial \beta = x_s (1 - \lambda_s) \lambda_s - x_s (1 - \mu_s) \mu_s$,

and $\partial\eta_s/\partial\alpha = x_s(1-\lambda_s)\lambda_s$. Using these expressions we can calculate the derivatives of the log likelihood function with respect to the parameters. The ML estimates are the values of the parameters for which the efficient scores (RAO, 1973) are zero. To obtain a solution to the equations resulting from setting $\partial LL/\partial\beta = \partial LL/\partial\alpha = 0$, we use Fisher's method-of-scoring which provides an iterative search procedure for the estimation of β and α. Let θ be the vertical concatenation of the column vectors β and α, then the iteration scheme is $\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} + \varepsilon[\hat{\mathbf{I}}(\hat{\theta}^{(i)})]^{-1}(\partial LL(\hat{\theta}^{(i)})/\partial\theta)$ (see, e.g., AMEMIYA, 1981). The parameter ε denotes an appropriate step length that scales the parameter increments and $\hat{\mathbf{I}}(\hat{\theta}^{(i)})$ is an estimate of the Fisher information matrix $\mathbf{I}(\theta) = -\mathrm{E}[\partial^2 LL(\theta)/\partial\theta_j\partial\theta_k]$ evaluated at $\hat{\theta}^{(i)}$, where $\partial^2 LL(\theta)/\partial\theta_j\partial\theta_k$ is the Hessian matrix. As a by-product of this iterative scheme, the method-of-scoring produces an estimate of the asymptotic variance-covariance matrix of the model parameters, being the inverse of the information matrix $\mathbf{I}^{-1}(\theta)$ evaluated at the values of the maximum likelihood estimates.

2.3 *Some model extensions*

A potential drawback to the model presented by MOFFITT (1990, 1993) is that it assumes that the effects of the covariates are fixed over time, implying that they are expected to have much the same impact over the period of time during which the observations were obtained. This restriction may not be valid for long time periods and potentially biases the estimated effects. An alternative model that could be considered is to allow the regression coefficient to become polynomials in t using the expression $\beta_t = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \ldots + \gamma_d t^d$, where d is a positive integer specifying the degree of the polynomial. Obviously, in practice it will be desirable to have models with low degree polynomials that avoid problems of overparametrization (i.e., nonexistence of unique ML estimates) and that combine parsimony of parametrization with fidelity to data. Another way in which we may accommodate the model is that whereas Moffitt defined the first observed outcome of the process $P(Y_{i1} = 1)$ to equal the transition probability μ_{i1}, we take $P(Y_{i1} = 1)$ to equal the state probability p_{i1}. That is, we assume that the Y_{i1}'s are random variables with a probability distribution $P(Y_{i1} = 1) = F(X_{it}\delta)$, where δ is a set of parameters to be estimated and F is the logistic function. The δ-parameters for the first observed outcome at $t = 1$ are estimated simultaneously with the entry and exit parameters of interest at $t = 2, \ldots, T$. Moreover, recall that the probability vector at the beginning of the Markov chain is estimated as a function of all cross-sectional data, rather than simply the observations at $t = 1$. Finally, we may also relax the implicit assumption that the cross sections at each time t are of the same sample size. To ensure a potentially equal contribution of the cross-sectional samples to the likelihood, we use the weighted log likelihood function $LL^* = \sum_{t=1}^{T}\sum_{i=1}^{n_t} w_i[y_{it}\log(p_{it}) + (1-y_{it})\log(1-p_{it})]$, where $w_i = \bar{n}/n_t$, with $\bar{n} = \sum_{t=1}^{T} n_t/T$, n_t is the number of observations of cross section t and T is the number of cross sections.

3 Application

Our empirical application employs panel data on female labour force participation of Dutch women aged 20–64 drawn from the Socio-economic Panel (SEP) study conducted by STATISTICS NETHERLANDS in the period 1986–1995. The panel data were treated as if they were a temporal sequence of cross sections of unrelated women (i.e., no estimate of cov(y_t, y_{t-1}) is available in the data used to estimate the Markov model). These data were used because they allow us to verify the results of the Markov model. The labor market status y_{it} is defined to equal 1 if the woman participates in the labor force at time t and 0 otherwise.

Table 1 gives the number of observations (including panel inflow and outflow), the marginal distribution of participation over time, and the observed annual entry and exit transitions rates in the panel. The table shows that over the period considered the female participation rate in the panel increased from about 40% in 1986 to around 56% in 1995. It also shows that both the panel entry and exit transition rates are relatively low. The analysis uses only covariates that are generally available in repeated cross-sectional surveys. As time-varying covariates, the analysis employs age in four different age categories (20–34, 35–44, 45–54, 55–64 years of age), the number of children at three different age categories (< 5, 5–17, $\geqslant 18$ years of age), and the annual nationwide unemployment rate (in %). The covariate completed education is taken to be fixed over time. Next to these variables the analysis also includes three initial conditions variables that capture the first entry into the process at age 20, the interaction of first entry with education and its interaction with the aggregate unemployment rate. The potentially important interaction of first entry with number of children was not included, as the number of mothers aged 20 was insufficient to allow reliable estimation. It is of interest to note that the individual observations were backcasted until the minimum age of 20, at which the first entry into the participation process is taken to have occurred. If for an observation the backcasted value of age in a particular cross section was less than 20, the entry and exit rates at that time period were fixed to zero.

Table 1. Marginal fraction of women's employment and observed annual entry and exit transition rates

year	n_t	inflow (age 20)	outflow (age 64)	$\bar{y}_t$	$\bar{y}_t \mid y_{t-1} = 0$	$\bar{y}_t \mid y_{t-1} = 1$
1986	2,302	52	21	0.400		
87	2,299	18	33	0.406	0.076	0.109
88	2,306	39	28	0.425	0.097	0.106
89	2,308	30	28	0.432	0.087	0.109
90	2,316	36	36	0.448	0.107	0.113
91	2,288	8	47	0.476	0.127	0.105
92	2,241	0	41	0.515	0.128	0.074
93	2,200	0	39	0.525	0.097	0.086
94	2,161	0	48	0.526	0.077	0.082
95	2,113	0	36	0.557	0.121	0.066

First a simple time-stationary Markov model with constant terms only was applied to the data using the software program ***CrossMark*** (which is available upon request). This model produced a $\beta(\mu_{t>1})$ of -0.222 and a $-\beta^*(\lambda_{t>1})$ of -0.078. These estimates imply constant transition rates of $\mu = 0.445$ and $\lambda = 0.480$; hence implausibly high values that amply exceed those reported in Table 1. The model was thereupon extended to a nonstationary, heterogeneous Markov model by including the covariates reported above. The results are shown in Table 2.

The parameters in the first column show the effect of the variables on the employment state probability p_{i1} at $t = 1$, estimated for all observations in the model. As can be seen, the parameters are well determined, with employment positively affected by education and negatively by age and the number of children (particularly preschool children) in the household. The second column in Table 2 presents the effect of the variables on the transition from non-employment to employment. Whereas education is significant in encouraging entry into the labor force, young children in the household and the aggregate unemployment rate negatively affect the entry decision. We also find that age has a negative effect on entry implying that the entry rates decline with age. The initial conditions variables indicate that higher unemployment rates and higher education increase the probability of entry at age 20. According to the standard errors, however, these variables have little impact on the hazards. The third column gives the effect of the variables on the transition into non-employment. We find that the exit rates are negatively affected by education and positively by the number of school and preschool children in the household. The

Table 2. Markov repeated cross-section estimates for women's transition into and out of employment, $n = 22{,}534$

	$\delta(p_{t=1})$[a]	$\beta(\mu_{t>1})$	$-\beta^*(\lambda_{t>1})$
Intercept	−0.027 (*0.099*)	−0.684 (*0.468*)	−1.877* (*0.670*)
Education	0.322* (*0.031*)	0.347* (*0.043*)	−0.570* (*0.067*)
Age[b]:			
35–44 years old	−0.199* (*0.079*)	−1.287* (*0.127*)	−2.190* (*0.287*)
45–54 years old	−1.198* (*0.095*)	−1.592* (*0.203*)	−0.311 (*0.309*)
55–64 years old	−2.187* (*0.115*)	−3.139* (*0.439*)	1.290* (*0.240*)
Number of children:			
<5 years old	−1.543* (*0.094*)	−0.214* (*0.089*)	2.066* (*0.151*)
5–17 years old	−0.438* (*0.036*)	−0.017 (*0.052*)	0.220* (*0.107*)
$\geqslant 18$ years old	−0.176* (*0.054*)	0.091 (*0.105*)	−0.253 (*0.179*)
Unemployment rate		−0.225* (*0.067*)	0.052 (*0.093*)
Age20[b]		0.853 (*1.599*)	
Age20 × education		0.306 (*0.209*)	
Age20 × unemployment rate		0.283 (*0.191*)	
Log likelihood (LL^*)			−12760.67

* Significant at 5% level (based on the estimated information matrix).
[a] Estimates of standard errors in parentheses. The β-parameters represent the effect on μ_t, the β^*-parameters the effect on $(1 - \lambda_t)$, and thus $-\beta^*$ the effect on λ_t.
[b] Reference category Age = 20–34 years; Age20: 1 if age = 20, 0 if not.

coefficients of the age terms imply that the incentives to end a job initially decrease with age but they are forced up again (presumably by occupational pension) after the age of 54. The effect of the aggregate unemployment rate on the transition into non-employment is insignificant.

Because there are substantive arguments to anticipate that the effect of some of the covariates (intercept, number of young children, education) may vary over time, several tests with different time-varying-coefficient models were applied to the data. These models, however, describe the data only slightly better (in terms of goodness-of-fit) than the time-constant-coefficient model and their results are therefore not reported here. We instead concentrate on an examination of the fit of the estimated model presented in Table 2 in terms of predictions. There are several ways to do so. One is to compare the actual sample frequency of all possible labor force participation sequences from 1986 to 1995 with the estimated expected frequency of each sequence. The latter were computed as follows. With T sample periods, we have $\sum_{t=1}^{T} 2^t$ different sequences (which in the present application equals 2,046) ranging in length from 1 (e.g., '0') to T (e.g., '0101010101') . We define the probability of a sequence of length t for each observation i of cross section t as

$$\tilde{p}_i(\tilde{y}_1, \ldots, \tilde{y}_t) = P(Y_{i1} = \tilde{y}_1 \cap \ldots \cap Y_{it} = \tilde{y}_t),$$

where $\tilde{y}_1, \ldots, \tilde{y}_t = 0,1$. Hence

$$\tilde{p}_i(\tilde{y}_1) = P(Y_{i1} = \tilde{y}_1) = \tilde{y}_1 p_{i1} + (1 - \tilde{y}_1)(1 - p_{i1}),$$

where p_{i1} is $P(Y_{i1} = 1)$. For $t > 1$, we have

$$\tilde{p}_i(\tilde{y}_1, \ldots, \tilde{y}_t) = \tilde{p}_i(\tilde{y}_1) \prod_{\tau=2}^{t} (p_{00} + p_{01} + p_{10} + p_{11}),$$

where

$$p_{00} = (1 - \tilde{y}_{\tau-1})(1 - \tilde{y}_\tau)(1 - \mu_{i\tau}),\ p_{01} = (1 - \tilde{y}_{\tau-1})\tilde{y}_\tau \mu_{i\tau},\ p_{10} = \tilde{y}_{\tau-1}(1 - \tilde{y}_\tau)\lambda_{i\tau},$$

and $p_{11} = \tilde{y}_{\tau-1}\tilde{y}_\tau(1 - \lambda_{i\tau})$. The mean value of $\tilde{p}_i(\tilde{y}_1, \ldots, \tilde{y}_t)$ for all observations of cross section t was obtained as $\tilde{p}(\tilde{y}_1, \ldots, \tilde{y}_t) = \sum_{i=1}^{n_t} \tilde{p}_i(\tilde{y}_1, \ldots, \tilde{y}_t)/n_t$. The estimated expected absolute frequency $\tilde{f}(\tilde{y}_1, \ldots, \tilde{y}_t)$ of each participation sequence was thereupon computed by evaluating $\tilde{f}(\tilde{y}_1, \ldots, \tilde{y}_t) = \tilde{p}(\tilde{y}_1, \ldots, \tilde{y}_t)n_t$.

An initial examination is to compare the expected with the observed first-order transitions over the time period of our data. Table 3 shows the relative frequencies of the estimated expected $(\tilde{y}_{t-1}, \tilde{y}_t)$ transitions and the differences between the expected and the observed relative frequencies. As can be seen, the predicted frequencies are concentrated in the continuous work (11) and continuous nonwork (00) categories. Further, while for some time periods the discrepancies between the predicted and the observed proportions are significant at the 0.05 level, most differences are very small. This implies that both the mover and the stayer frequencies are predicted fairly well.

A further examination of the fit of the model reported here is to compare the estimated expected and the actually observed absolute frequencies of all 2,046

Table 3. Relative frequencies of estimated expected $(\tilde{y}_{t-1}, \tilde{y}_t)$ transitions at sample period T and estimated expected minus observed proportions, $n = 22{,}534$

T	n_t	estimated expected				expected - observed				χ^2
		(00)	(01)	(11)	(10)	(00)	(01)	(11)	(10)	
2	2,299	0.556	0.053	0.354	0.037	0.006	0.008	−0.007	−0.007	6.03
3	2,306	0.540	0.056	0.365	0.039	0.009	−0.001	−0.003	−0.005	1.78
4	2,308	0.521	0.060	0.381	0.038	0.000	0.010	−0.002	−0.009	8.57
5	2,316	0.495	0.067	0.400	0.038	−0.008	0.007	0.012	−0.011	10.49
6	2,288	0.469	0.059	0.432	0.040	−0.007	−0.010	0.025	−0.008	10.96
7	2,241	0.448	0.053	0.460	0.039	0.000	−0.013	0.011	0.003	8.41
8	2,200	0.441	0.038	0.482	0.039	0.011	−0.008	0.003	−0.006	6.12
9	2,161	0.441	0.031	0.491	0.037	0.011	−0.005	0.001	−0.007	5.47
10	2,113	0.435	0.033	0.500	0.032	0.028	−0.023	−0.001	−0.003	36.68

employment status sequences. Because it is unfeasible to tabulate all frequencies, they are graphically displayed in Figure 2 together with the OLS regression lines.

The top part of the figure displays the predicted and the actual frequencies of all possible employment profiles, but highlights the relatively small number of sequences with high frequencies. These sequences concern the continuous participation and continuous nonparticipation categories. The bottom part of the figure zooms in on the employment sequences with relatively low frequencies in the 0−140 range. Visual inspection suggests close agreement between the estimated expected frequencies predicted by the RCS Markov model and the observed frequencies of the spells in the panel. The (unreported) longitudinal profiles indicate that most women remain employed or non-employed throughout the observation interval and that proportionally few women move into and out of the labor force frequently.

4 Conclusion

The overall conclusion that we draw from this example is that the proposed model can be a useful tool in applied work. It obviously does not supersede genuine panel designs, but it definitely puts a series of one-shot surveys into perspective and it provides more refined results than would be available from a single cross-sectional study. Microdata panel sets offer the potential for the construction of more flexible and richer statistical models of transition dynamics than do those based upon cross-sectional information. However, while there has been a substantial increase of data archives holding vast collections of repeated cross-sectional data, panel data represent the exception of these collection efforts, rather than the rule. RCS data are cheaper to collect and they do not suffer from problems of non-random attrition which plague panel data. Moreover, a disadvantage to using pure panel surveys is the limited number of units followed and the limited number of time points at which these units are usually re-interviewed. These limitations have to be balanced against the lack of direct information on the transitions in long-run RCS data.

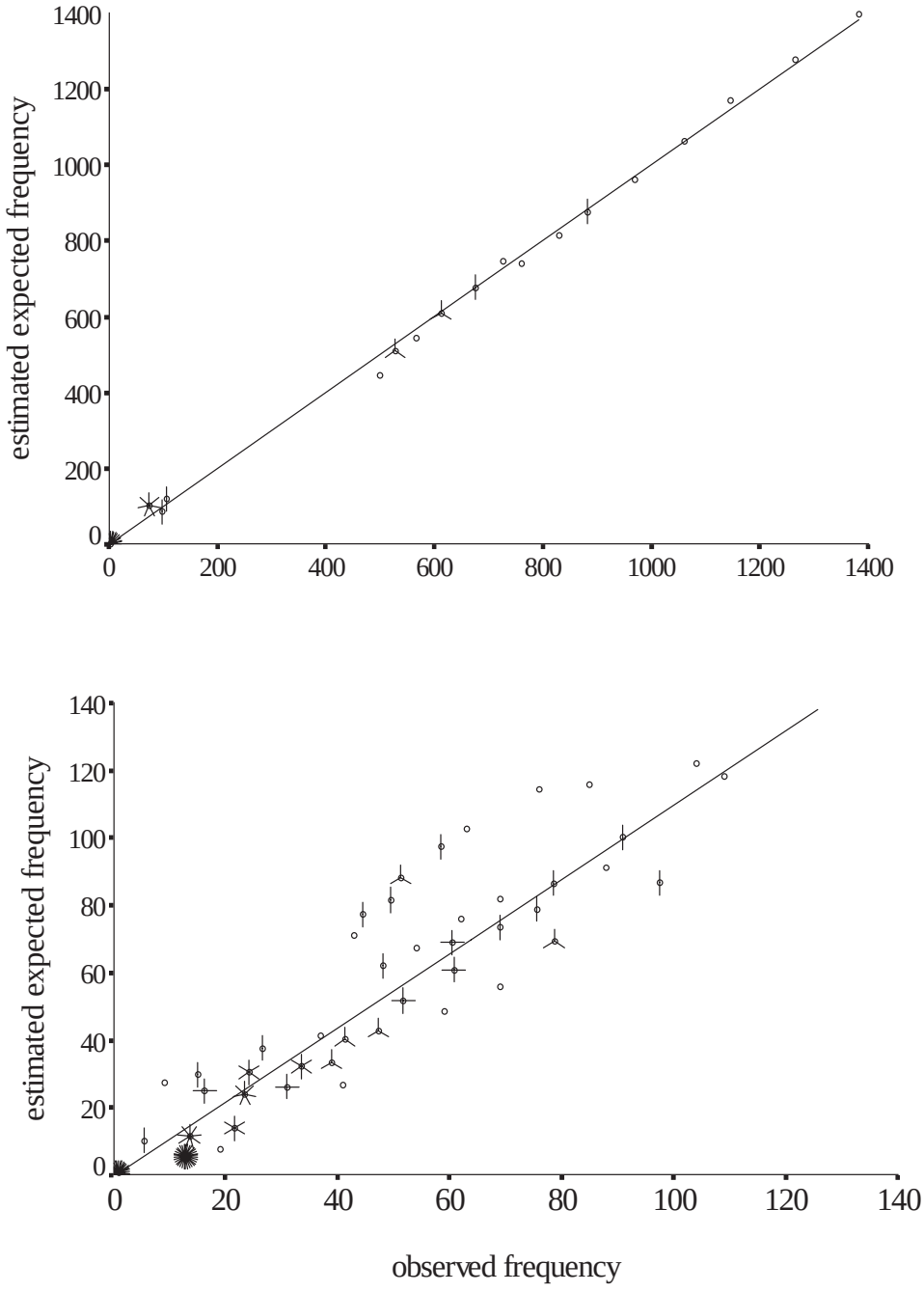

Fig. 2. Estimated expected versus observed frequencies of 2,046 employment status sequences and OLS regression lines

Some problems we encountered in trying to model unobserved transitions over time using RCS data deserve to be mentioned. The application of the model presented here requires knowing the history of the explanatory variables for the individuals in the samples. We often have characteristics for which the history is unknown however. These characteristics may be relevant explanatory variables, but in many applications the analysis would omit them. Nevertheless, it is our believe that relatively rich dynamic models can be developed with a time series of RCS data. Many individual variables can be backcasted with considerable accuracy and many aggregate indicators are also measurable in the past. Moreover, our experiments have shown that it is also possible to specify a model with two different sets of parameters for both μ and λ, i.e., one for the past transition rates and a separate one for the transition at the current time period. This offers the opportunity to also include relevant non-backcastable covariates in the (current part of the) Markov model.

A somewhat related problem, common to all duration analyses, is that the model specification assumes that individual heterogeneity is due to the observed variables. It is likely, however, that unobserved and possibly unobservable variables including initial conditions are also a source of population heterogeneity. The presample history is lost by imposing an arbitrary survey window on the behavioral process, thus left-censoring the process and omitting events of interests associated with, or arising from, the periods prior to the first survey. The potential effect of this uncontrolled heterogeneity can bias the estimated effects of the explanatory variables included in the model. It is unknown, however, how serious the consequences of misspecification are if we have sufficiently flexible models for baseline hazards and time-varying covariates. Hence further investigation is needed on how much of the evidence is censored.

References

AMEMIYA, T. (1981), Qualitative response models: a survey, *Journal of Econometric Literature* **19**, 1483–1536.

AMEMIYA, T. (1985), *Advanced econometrics*, Basil Blackwell, Oxford.

BALTAGI, B. H. (1995), *Econometric analysis of panel data*, Wiley, Chicester.

BARTHOLOMEW, D. J. (1996), *The statistical approach to social measurement*, Academic Press, San Diego.

BISHOP, Y. M. M., S. E. FIENBERG and P. W. HOLLAND (1975), *Discrete multivariate analysis: theory and practice*, MIT Press, Cambridge MA.

BOSKIN, M. J. and F. C. NOLD (1975), A Markov model of turnover in aid to families with dependent children, *Journal of Human Resources* **10**, 476–481.

COLLADO, M. D. (1997), Estimating dynamic models from time series of independent cross-sections, *Journal of Econometrics* **82**, 37–62.

DEATON, A. (1985), Panel data from time series of cross-sections, *Journal of Econometrics* **30**, 109–126.

DIGGLE, P. J., K. Y. LIANG and S. L. ZEGER (1994), *Analysis of longitudinal data*, Clarendon Press, Oxford.

FELTEAU, C., P. LEFEBVRE, PH. MERRIGAN and L. BROUILLETTE (1997), *Conjugalité et fécondité*

des femmes canadiennes: un modèle dynamique estimé à l'aide d'une série de coupes transversales, CREFÉ, Université de Québec à Montréal, Montréal.

FIRTH, D. (1982), Estimation of voter transition matrices from election data, M.Sc. thesis, Department of Mathematics, Imperial College London: London.

GIRMA, S. (2000), A quasi-differencing approach to dynamic modelling from a time series of independent cross-sections, *Journal of Econometrics* **98**, 365–383.

GIRMA, S. (2001), A note on dynamic modelling from short and heterogeneous pseudo panels, *Statistica Neerlandica* **55**, 239–248.

GOODMAN, L. A. (1961), Statistical methods for the mover–stayer model, *Journal of the American Statistical Association* **56**, 841–868.

HAMERLE, A. and G. RONNING (1995), Panel analysis for qualitative variables, in: G. ARMINGER, C. CLOGG and M. E. SOBEL (eds.), *Handbook of statistical modeling for the social and behavioral sciences*, Plenum Press, New York, 401–451.

HAWKINS, D. L., C. P. HAN and J. EISENFELD (1996), Estimating transition probabilities from aggregate samples augmented by haphazard recaptures, *Biometrics* **52**, 625–638.

KALBFLEISH, J. D. and J. F. LAWLESS (1984), Least squares estimation of transition probabilities from aggregate data, *Canadian Journal of Statistics* **12**, 169–182.

KALBFLEISH, J. D. and J. F. LAWLESS (1985), The analysis of panel data under a Markovian assumption, *Journal of the American Statistical Association* **80**, 863–871.

LAWLESS, J. F. and D. L. MCLEISH (1984), The information in aggregate data from Markov chains, *Biometrika* **71**, 419–430.

LEE, T. C., G. G. JUDGE and A. ZELLNER (1970), *Estimating the parameters of the Markov probability model from aggregate time series data*, North-Holland, Amsterdam.

LI, W. K. and M. C. O. KWOK (1990), Some results on the estimation of a higher order Markov chain, *Communications in Statistics, Part B, Simulation and Computation* **19**, 363–380.

MCCALL, J. J. (1971), A Markovian model of income dynamics, *Journal of the American Statistical Association* **66**, 439–447.

MEBANE, W. R. and J. WAND (1997), *Markov chain models for rolling cross-section data: how campaign events and political awareness affect vote intentions and partisanship in the United States and Canada*, paper presented at the 1997 Annual Meeting of the Midwest Political Science Association, Chicago Il.

MOFFITT, R. (1990), The effect of the U.S. welfare system on marital status, *Journal of Public Economics* **41**, 101–124.

MOFFITT, R. (1993), Identification and estimation of dynamic models with a time series of repeated cross-sections, *Journal of Econometrics* **59**, 99–123.

NIJMAN, TH. E. and M. VERBEEK (1990), Estimation of time-dependent parameters in linear models using cross-sections, panels, or both, *Journal of Econometrics* **46**, 333–446.

RAO, C. R. (1973), *Linear statistical inference and its applications*, Wiley, New York.

TOIKKA, R. S. (1976), A Markovian model of labor market decision by workers, *American Economic Review* **66**, 821–834.

TOPEL, R. H. (1983), On layoffs and unemployment insurance, *American Economic Review* **73**, 541–559.

VERBEEK, M. (1996), Pseudo panel data, in: L. MÁTYÁS and P. SEVESTRE (eds.). *The econometrics of panel data* (2nd revised ed.), Kluwer Academic Publishers, Dordrecht, 280–292.

VERBEEK, M. and TH. NIJMAN (1992), Can cohort data be treated as genuine panel data?, *Empirical Economics* **17**, 9–23.

VERBEEK, M. and TH. NIJMAN (1993), Minimum MSE estimation of a regression model with fixed effects from a series of cross-sections, *Journal of Econometrics* **59**, 125–136.

VERBEEK, M. and F. VELLA (2000), *Estimating dynamic models from repeated cross sections*, paper presented at the International Conference on the Analysis of Repeated Cross-sectional Surveys, June 15–16, 2000, Nijmegen, Netherlands.

Received: September 2000. Revised: February 2001.