Estimation of Time-Varying Markov Processes with Aggregate Data
Author(s): Elizabeth Chase MacRae
Source: *Econometrica*, Vol. 45, No. 1 (Jan., 1977), pp. 183-198
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/1913295
Accessed: 01-06-2015 14:00 UTC

Econometrica, Vol. 45, No. 1 (January, 1977)

ESTIMATION OF TIME-VARYING MARKOV PROCESSES WITH AGGREGATE DATA

BY ELIZABETH CHASE MACRAE

The exact stochastic character of observed data from a Markov process is derived for the case where only aggregate stocks, as opposed to individual transitions, are observed. Particular attention is devoted to the distinction between data generated by a panel study, where a single group of individuals is followed over time, and that generated by random sampling, where the observed groups are not identical over time. Several alternative estimators are developed which take into account the particular stochastic structure of the data.

1. INTRODUCTION

THE THEORY OF MARKOV PROCESSES has been commonly used to describe the behavior of individual economic actors who move among a number of discrete states over time. If observations are available over time on individual transitions, then the transition probabilities, whether constant or functions of exogenous variables, may be readily estimated. However, much of the data relevant for Markov processes is not in the form of individual transitions but is aggregate stock data showing only the proportion of individuals in each state at each moment of time. Moreover, the available data are very often calculated from relatively small samples of individuals and are only estimates of the true proportions.

Estimation from aggregate data in the case where the transition probabilities are constant over time has been extensively analyzed by Lee, Judge, and Zellner [**3**]. They do not derive the actual probability distribution of the observed aggregate data, but instead present a number of models based on different assumptions regarding the stochastic character of the data. Estimators are then suggested for each of the alternative models. However, none of their stochastic assumptions corresponds to the true nature of the aggregate data.[1]

The purpose of this paper is to extend the analysis of Lee, Judge, and Zellner in two directions. First, the individual transition probabilities are permitted to vary over time in response to aggregate exogenous variables. The parametrization technique suggested for this purpose has the added advantage that the estimated probabilities are automatically constrained to be nonnegative and to sum to one. Second, the actual probability distributions of the observed aggregate data are derived and models based on the derived stochastic specifications are presented. The derivations distinguish between unconditional probability distributions and distributions which are conditional upon past observations, and make a further distinction between perfectly and imperfectly observed data. Several estimators are suggested for the new models with special attention given to the "errors in variables" problem that arises with imperfectly observed data, a problem which may be significant where group proportions are calculated from small samples.

[1] I wish to thank J. Kadane for alerting me to this fact.

Section 2 of the paper presents a formal statement of the Markov probability process and discusses the choice of functional form to represent the time-varying individual transition probabilities. Section 3 derives the stochastic characteristics of the observed aggregate data for both perfect and imperfect observation cases. Section 4 then develops the appropriate least squares, iterative generalized least squares, and maximum likelihood estimators for the perfect observations case, while Section 5 deals with instrumental variables, limited information least squares, and limited information maximum likelihood estimators for the imperfect observations case.

2. THE TIME-VARYING MARKOV PROCESS

This section describes the probabilistic behavior of individuals and characterizes the transition probabilities in terms of exogenous variables.

Individual State Probabilities

The first order Markov process with which this paper deals is characterized by a finite number of mutually exclusive and exhaustive states to which an individual may belong. Let $\theta_j(t)$ be the probability that an individual will be in state j at time t, and let $P_{ij}(t)$ be the probability of moving from state i at time $t-1$ to state j at time t. Then, assuming there are s states, the probabilities θ_j are related over time by

$$\theta_j(t) = \sum_{i=1}^{s} P_{ij}(t)\theta_i(t-1) \qquad (j = 1, 2, ,\ldots, s) \tag{2.1}$$

or in matrix notation,

$$\theta(t) = P'(t)\theta(t-1), \tag{2.2}$$

where $\theta(t)$ is an s-dimensional vector of state probabilities, and $P(t)$ is an s by s matrix of transition probabilities, $P_{ij}(t)$. Since the states are mutually exclusive and exhaustive, the elements of θ and the rows of P must sum to 1 in every period:

$$\sum_{j=1}^{s} \theta_j(t) = 1 \tag{2.3}$$

and

$$\sum_{j=1}^{s} P_{ij}(t) = 1 \qquad (i = 1, 2, \ldots, s). \tag{2.4}$$

The implication of identities (2.3) and (2.4) is that one of the equations in system (2.1) is redundant and may be omitted. Without loss of generality, the last equation is chosen to be omitted so that the Markov process is now described by an alternative system of $s-1$ equations,

$$\theta_*(t) = P_*'(t)\theta(t-1) \tag{2.5}$$

together with identity (2.3). The $s-1$ dimensional vector $\theta_*(t)$ is formed from $\theta(t)$ by dropping the last element and the s by $s-1$ matrix $P_*(t)$ is the same as $P(t)$ without its last column.

Time-Varying Transition Probabilities

An individual's transition probabilities are assumed to vary over time in response to aggregate exogenous variables. That is,

$$P_{ij}(t)=f_{ij}(z(t-1),\beta_{ij}) \qquad (i,j=1,2,\ldots,s), \tag{2.6}$$

where $z(t-1)$ is a vector of exogenous or predetermined variables at time $t-1$ and β_{ij} is a vector of parameters which relates the exogenous variables to the transition probabilities. Variables $z(t-1)$ appear lagged because they are related to transitions which begin at time $t-1$; it seems reasonable to assume that transition probabilities cannot depend on variables which are available only after the transition has occurred. However, the notation for $z(t-1)$ is not meant to exclude variables lagged more than one period before t. Since the transition probabilities must be nonnegative and satisfy identity (2.4), the choice of functional forms for f_{ij} is somewhat limited. In this paper, only the multinomial logit formulation will be discussed although the estimation techniques may obviously be extended to other functional forms.

The multinomial logit formulation, as suggested by Theil [**5**], expresses the log of a ratio of probabilities as a function of the exogenous variables. Since the Markov model has s sets of probabilities, one for each row of the transition matrix, there will be s sets of ratios, each of which will use, for convenience, the transition probability from the last column of P as the denominator:

$$\ln\,(P_{ij}(t)/P_{is}(t))=F_{ij}(t)=F_{ij}(z(t-1),\beta_{ij}) \tag{2.7}$$
$$(j=1,2,\ldots,s-1;\, i=1,2,\ldots,s).$$

These equations, together with identities (2.4) to close the system, comprise a transformation from the space of exogenous variables to the space of transition probabilities such that all elements of P are nonnegative and the rows of P sum to 1. Rewriting (2.7) so as to express each probability separately as a function of the exogenous variables and coefficients yields

$$P_{is}(t)\sum_{j=1}^{s-1}\exp F_{ij}(t)=\sum_{j=1}^{s-1}P_{ij}(t)=1-P_{is}(t) \qquad (i=1,2,\ldots,s) \tag{2.8}$$

and hence

$$P_{ij}(t)=[\exp F_{ij}(t)]\Big/\left[1+\sum_{j=1}^{s-1}\exp F_{ij}(t)\right] \qquad (j=1,\ldots,s-1), \tag{2.9}$$
$$P_{is}(t)=1\Big/\left[1+\sum_{j=1}^{s}\exp F_{ij}(t)\right].$$

If the functions $F_{ij}(t)$ are simply constants, then equations (2.9) represent the special case of constant transition probabilities.

As can be seen from (2.7) and (2.9), all the transition probabilities in a row of $P(t)$ depend on exactly the same set of parameters. This will also be true in general for formulations other than the multinomial logit, since the constraints on transition probabilities imply that no probability in a row of $P(t)$ can be determined completely independently of all the other probabilities in the same row. The implication of this interdepedence of transition probabilities is that it is not possible to estimate the $s-1$ equations of system (2.9) separately.

The advantage of using a parameterization such as the multinomial logit to characterize the transition probabilities is that not only does it allow for variation over time but it also constrains the probabilities to be nonnegative and to sum to unity by rows, since the probabilities are not estimated directly but only through the estimation of the parameters β.

3. STOCHASTIC CHARACTER OF AGGREGATE DATA

The aggregate data available for estimation of individual Markov transition probabilities consist of the number of individuals in each state in each period, or equivalently, the proportion of individuals in each state. It is useful to distinguish between two classes of aggregate data depending upon the nature of the observations. In the first type, that of perfect state observations, an entire group is perfectly and completely observed over time. The data consist of a set of vectors of state proportions, $x(t)$, over time, for a particular group of size N. In the second and more common type of data, imperfect state observations, only subsets of the group are observed over time, so that the state proportions for the entire group are never known with certainty. The data in this case consist of a set of vectors of state proportions, $y(t)$, calculated from samples of size $M(t)$ drawn randomly in each period from the full group.

It is also useful to distinguish between conditional and unconditional probability distributions for $x(t)$ and $y(t)$. The unconditional distribution for $x(t)$ is multinomial about a mean of $\theta(t)$, the vector of state probabilities for the underlying Markov process. For, if no additional information is considered, it is as if the N individuals in the group were drawn at random from the infinite Markov population. Similarly, the unconditional distribution of $y(t)$ is also multinomial about $\theta(t)$ (but with size parameter $M(t)$ instead of N), since with no information about $x(t)$ or N considered, $y(t)$ is distributed as though it were drawn directly from the basic Markov population. Thus, from the point of view of unconditional distributions, there is nothing to distinguish the perfect observation data, $x(t)$, from the imperfect observation data, $y(t)$, except for the size of N and $M(t)$.

It is with respect to their conditional distributions that the distinction between perfect and imperfect observations becomes important. The conditional distribution of $y(t)$ given $x(t)$ is obviously multinomial with mean $x(t)$ since the sample is assumed to be drawn randomly (with replacement) from the group. However, the two conditional distributions of interest for estimation, that of $x(t)$ given

$x(t-1), \ldots, x(0)$ and that of $y(t)$ given $y(t-1), \ldots, y(0)$, are definitely not multinomial but have more complicated forms.

Distribution of Perfectly Observed Proportions

To derive the conditional distribution of $x(t)$ given $x(t-1), \ldots, x(0)$, let $n(t-1)$ be a vector showing the number of individuals in each state in the previous period, $t-1$, so that $n(t-1) = Nx(t-1)$. Consider now only the $n_i(t-1)$ individuals who were in state i at time $t-1$. Let $X^i(t)$ be a vector of proportions for this group in the current period; that is, of the $n_i(t-1)$ individuals who were in state i at time $t-1$, $X^i(t)$ shows the proportions in each of the states at time t. Since the individuals which make up $X^i(t)$ are all independently and identically multinomially distributed with probability $P^i(t)$, defined as row i of $P(t)$, the vector $X^i(t)$ has a multinomial distribution about mean $P^i(t)$ with size $n_i(t-1)$. The observed group proportion vector $x(t)$ is simply a weighted sum of the unobserved vectors $X^i(t)$ and hence is distributed as a weighted sum of independent but not identical multinomial random variables[2] where the weights are equal to the group proportions at time $t-1$:

$$(3.1) \qquad x(t) = \sum_{i=1}^{s} X^i(t) x_i(t-1).$$

The mean of $x(t)$, conditional on $x(t-1), . . , x(0)$, is $p(t)$, given by

$$(3.2) \qquad p(t) = E\{x(t) \,|\, x(t-1), \ldots, x(0)\} = \sum_{i=1}^{s} P^i(t) x_i(t-1) = P'(t) x(t-1),$$

and its covariance matrix $\Omega(t)$ is

$$(3.3) \qquad \Omega(t) = \operatorname{cov}\{x(t) \,|\, x(t-1), \ldots, x(0)\} = \sum_{i=1}^{s} \Omega^i(t) x_i^2(t-1),$$

where $\Omega^i(t)$ is the covariance matrix of $X^i(t)$. Since each of the vectors $X^i(t)$ has a multinomial distribution about $P^i(t)$, the elements of the individual covariance matrices $\Omega^i(t)$ have the form:

$$(3.4) \qquad \Omega^i_{jj}(t) = [1 - P_{ij}(t)] P_{ij}(t) / n_i(t-1)$$

and

$$\Omega^i_{jk}(t) = -P_{ik}(t) P_{ij}(t) / n_i(t-1), \qquad k \neq j.$$

Combining (3.4) and the fact that $x(t) = n(t)/N$, yields weighted sums of products of transition probabilities for the elements of $\Omega(t)$:

$$(3.5) \qquad \Omega_{jj}(t) = \sum_{i=1}^{s} [1 - P_{ij}(t)] P_{ij}(t) x_i(t-1)/N$$

[2] For an alternative demonstration of this fact for the binomial case, see MacRae and MacRae [**4**].

and

$$\Omega_{jk}(t)=\sum_{i=1}^{s}-P_{ij}(t)P_{ik}(t)x_i(t-1)/N.$$

Matrix $\Omega_*(t)$, the (conditional) covariance matrix of $x_*(t)$, is obtained by dropping the last row and column of $\Omega(t)$, so that in matrix notation $\Omega_*(t)$ is given by

$$\Omega_*(t)=[\operatorname{diag}\{P'_*(t)x(t-1)\}-P'_*(t)\ \operatorname{diag}\{x(t-1)\}P_*(t)]/N, \tag{3.6}$$

where diag $\{\cdot\}$ is a square matrix with the argument vector as the main diagonal and zeros elsewhere.

Comparison wih Multinomial Distribution

One of the assumptions used by Lee, Judge, and Zellner in developing estimators is that the vector $x(t)$ is distributed multinomially (given $x(t-1)$) about the mean of $p(t)$. This is in contrast to the above derived correct characterization of $x(t)$ as a sum of multinomials with mean $p(t)$. Although both distributions have the same mean, the correct distribution of $x(t)$ has a smaller variance than a multinomial distribution. The covariance matrix for a multinomial distribution, $\Phi(t)$, is equal to products of sums of transition probabilities:

$$\begin{aligned}(3.7)\qquad \Phi_{jj}(t)&=[1-p_j(t)]p_j(t)/N\\ &=\left[1-\sum_{i=1}^{s}P_{ij}(t)x_i(t-1)\right]\left[\sum_{i=1}^{s}P_{ij}(t)x_i(t-1)\right]\Big/N,\\ \Phi_{jk}(t)&=-p_j(t)p_k(t)/N\\ &=-\left[\sum_{i=1}^{s}P_{ji}(t)x_i(t-1)\right]\left[\sum_{i=1}^{s}P_{ki}(t)x_i(t-1)\right]\Big/N,\end{aligned}$$

so that

$$\Phi_*(t)=[\operatorname{diag}\{P'_*(t)x(t-1)\}-P'_*(t)x(t-1)x'(t-1)P_*(t)]/N. \tag{3.8}$$

The above covariance matrix $\Phi(t)$ based on a multinomial $x(t)$ overstates the true variance of the observed data; that is, matrix $\Phi_*(t)$ is larger than the correct covariance $\Omega_*(t)$ in the sense that $\Phi_*(t)-\Omega_*(t)$ is positive semidefinite. To see this, note that the difference between $\Phi_*(t)$ and $\Omega_*(t)$ is

$$\Phi_*(t)-\Omega_*(t)=P'_*(t)[\operatorname{diag}\{x(t-1)\}-x(t-1)x'(t-1)]P_*(t)/N \tag{3.9}$$

which is a positive semidefinite matrix since it can be shown that the expression in square brackets is positive semidefinite. The only case where $\Phi_*(t)$ is identical to the correct covariance matrix, $\Omega_*(t)$, is when all rows of $P(t)$ are identical. But this situation represents a non-dynamic model where an individual's past state does not influence his present state, so that each of the identical rows of $P(t)$ is also equal to $p(t)$.

Distribution of Imperfectly Observed Proportions

In the perfect observations case discussed above, the conditional distribution of the vector $x(t)$, given $x(t-1), \ldots, x(0)$, is identical to the distribution of $x(t)$ given only $x(t-1)$; earlier observations contain no additional information about $x(t)$. With imperfect observations, however, the immediately preceding observation $y(t-1)$ does not completely characterize the distribution of $y(t)$; the distribution of a sample proportion depends upon all the preceding observed sample proportions, $y(t-1), \ldots, y(0)$. The conditional probability of $m(t)$ given $m(t-1), \ldots, m(0)$ may be written in general terms as

$$(3.10) \quad \begin{aligned} &\mathrm{pr}\,\{m(t)\,|\,m(t-1), \ldots, m(0)\} \\ &\quad = \sum_{n(t)} \sum_{n(t-1)} \mathrm{pr}\,\{m(t)\,|\,n(t)\} \cdot \mathrm{pr}\,\{n(t)\,|\,n(t-1)\} \\ &\quad\quad \cdot \mathrm{pr}\,\{n(t-1)\,|\,m(t-1), \ldots, m(0)\}, \end{aligned}$$

where $m(t) = y(t) \cdot M(t)$ and $n(t) = x(t) \cdot N$. Using Bayes' rule, equation (3.10) may be expanded to

$$(3.11) \quad \begin{aligned} &\mathrm{pr}\,\{m(t)\,|\,m(t-1), \ldots, m(0)\} \\ &\quad = \Bigg(\sum_{n(t), \ldots, n(0)} \!\!\cdots \sum \mathrm{pr}\,\{m(t)\,|\,n(t)\} \\ &\quad\quad \cdot \mathrm{pr}\,\{n(t)\,|\,n(t-1)\}\, \mathrm{pr}\,\{m(t-1)\,|\,n(t-1)\} \\ &\quad\quad \cdot \mathrm{pr}\,\{n(t-1)\,|\,n(t-2)\} \ldots \mathrm{pr}\,\{m(1)\,|\,n(1)\} \\ &\quad\quad \cdot \mathrm{pr}\,\{n(1)\,|\,n(0)\} \cdot \mathrm{pr}\,\{n(0)\,|\,m(0)\}\Bigg) \Big/ (\mathrm{pr}\,\{m(t-1), \ldots, m(0)\}). \end{aligned}$$

The probabilities in (3.11) are all either multinomial or sums of multinomials. Therefore it is conceptually possible to evaluate (3.11) and to construct estimators based on this conditional distribution, but it would not be practical to implement them numerically. Subsequent development of estimators in this paper will not rely on the explicit form of the conditional distribution of $y(t)$, but rather will use the fact that $y(t)$, unlike $x(t)$, depends upon all previous observations.

4. ESTIMATION WITH PERFECT OBSERVATIONS

Least Squares Estimation

As was pointed out in the preceding section, the vector of perfectly observed group proportions, $x(t)$, is distributed as a sum of multinomials about $x(t-1)$. This means that, using standard regression equation format, $x(t)$ may be written as

$$(4.1) \quad x(t) = P'(t)x(t-1) + \varepsilon(t),$$

where $\varepsilon(t)$ is distributed as a sum of multinomials, but with mean zero. Since one of the equations in (4.1) is redundant by virtue of the fact that the elements of $x(t)$ must sum to 1, estimation techniques will actually be applied to the modified model of $s-1$ equations,

$$x_*(t)=P'_*(t)x(t-1)+\varepsilon_*(t) \tag{4.2}$$

where $_*$ indicates that the last element of x and ε, and the last column of P, is omitted.

Although the variance of $\varepsilon_*(t)$, $\Omega_*(t)$, depends upon the magnitude of $x(t-1)$, the expected value of $\varepsilon_*(t)$ is zero, independent of the value of $x(t-1)$, so that nonlinear least squares estimation will produce consistent but not efficient estimates. If the transition probability matrix is written as a function of $z(t-1)$ and β, the nonlinear least squares estimator β is chosen so as to minimize the sum of squared residuals over all observations:

$$\sum_{t=1}^{T} e'_*(t)\, e_*(t), \tag{4.3}$$

where

$$e_*(t)=x_*(t)-P'_*(t)x(t-1). \tag{4.4}$$

The phrase "non-linear least squares" is used to emphasize the fact that the residuals $e_*(t)$ are non-linear functions of the parameters, β, which are to be estimated. As was described in Section 2, the matrix of individual transition probabilities is taken to be a function of the underlying parameters β in such a way that the parameter estimates, $\hat{\beta}$, automatically yield an estimated $P(t)$ with the property that all elements are nonnegative and the rows sum to 1.

A more efficient estimator may be obtained by correcting for heteroscedasticity in the error term through the use of generalized least squares (GLS) in which β is chosen to minimize

$$\sum_{t=1}^{T} e'_*(t)\Omega_*^{-1}(t)e_*(t). \tag{4.5}$$

The elements of $\Omega_*(t)$ depend upon the true but unknown transition probabilities so that in the actual minimization of (4.5), matrix $\Omega_*(t)$ must be replaced by a consistent estimate $\hat{\Omega}_*(t)$. This suggests an iterative GLS procedure in which nonlinear least squares is applied first, and fitted values of the transition probabilities are obtained and used to calculate $\hat{\Omega}_*(t)$. Expression (4.5), with $\hat{\Omega}_*(t)$ in place of $\Omega_*(t)$ is then minimized with respect to β to obtain a new set of parameter estimates. At each subsequent iteration the value of $\Omega_*(t)$ is reestimated using the parameter estimates of the preceding iteration.

Maximum Likelihood Estimation

Another possibility is to take explicit account of the form of the probability distribution of observations by using maximum likelihood estimation. Let $N^i(t)$ be a vector showing the numbers in each state at time t of those who were in state i the previous period, $N^i(t)=n_i(t-1)X^i(t)$. The vector $N^i(t)$ has a multinomial distribution about vector $n_i(t-1)P^i(t)$, expressed by

$$\text{(4.6)} \qquad \operatorname{pr}\{N^i(t)\,|\,n_i(t-1)\} = (n_i(t-1)!)\left(\prod_{j=1}^{s} P_{ij}(t)^{n_{ij}(t)}/n_{ij}(t)!\right),$$

where $n_{ij}(t)$ is the number of individuals who made the transition from state i at time $t-1$ to state j at time t. Now let matrix $N(t)$ show the number of individuals making each possible transition, so that the vectors $N^i(t)$ are rows of $N(t)$, and $n_{ij}(t)$ are elements of $N(t)$. Then the probability of obtaining a particular set of transitions is the product of the probabilities of each of the rows $N^i(t)$:

$$\text{(4.7)} \qquad \operatorname{pr}\{N(t)\,|\,n(t-1)\} = \prod_{i=1}^{s} (n_i(t-1)!)\left(\prod_{j-1}^{s} P_{ij}(t)^{n_{ij}(t)}/n_{ij}(t)!\right).$$

To obtain the probability of a particular observed vector of states, $n(t)$, given $n(t-1)$, it is necessary to add together expressions of the form of (4.7) for all possible matrices $N(t)$ which could have resulted in the observed change from $n(t-1)$ to $n(t)$. That is,

$$\text{(4.8)} \qquad \operatorname{pr}\{n(t)\,|\,n(t-1)\} = \sum_{\mathcal{N}(t)} \prod_{i=1}^{s} (n_i(t-1)!)\left(\prod_{j=1}^{s} P_{ij}(t)^{n_{ij}(t)}/n_{ij}(t)!\right),$$

where the summation is over set $\mathcal{N}(t)$ of all possible matrices $N(t)$ of nonnegative integers such that the rows of $N(t)$ sum to $n(t-1)$ and the columns sum to $n(t)$:

$$\text{(4.9)} \qquad \mathcal{N}(t) = \{N(t)\,|\,\sum_k n_{ik}(t) = n_i(t-1), \quad \sum_k n_{kj}(t) = n_j(t), \quad \text{for all } i, j\}.$$

The likelihood function for all T of the group observations is:

$$\text{(4.10)} \qquad \begin{aligned} L(\beta) &= \operatorname{pr}\{x(1), x(2), x(3), \ldots, x(T)\} \\ &= \operatorname{pr}\{x(1)\}\cdot\operatorname{pr}\{x(2)\,|\,x(1)\}\cdot\operatorname{pr}\{x(3)\,|\,x(2)\}\ldots\operatorname{pr}\{x(T)\,|\,x(T-1)\}. \end{aligned}$$

Each of the conditional probabilities which comprise the product in (4.10) is conditional upon only the immediately preceding observation, a feature of perfectly observed data. The likelihood function which is to be maximized by vector β is still a rather formidable expression, the product of a sum of products,

$$\text{(4.11)} \qquad L(\beta) = \prod_{t=1}^{T} \sum_{\mathcal{N}(t)} \prod_{i=1}^{s} (n_i(t-1)!)\left(\prod_{j=1}^{s} P_{ij}(t)^{n_{ij}(t)}/n_{ij}(t)!\right).$$

While it is possible to develop computational algorithms to search for a maximum of (4.11), the iterative generalized least squares estimator may represent a better combination of numerical and statistical efficiency.

5. ESTIMATION WITH IMPERFECT OBSERVATIONS

Least Squares Estimation

Consider now the case where the available data are vectors of sample proportions, $y(t)$, from samples of size $M(t)$ drawn randomly each period from the parent population of size N and proportions $x(t)$. As pointed out in Section 3, the conditional distribution of $y(t)$, given $x(t)$, is multinomial with mean equal to $x(t)$, and may therefore be written as

$$(5.1) \qquad y(t)=x(t)+\eta(t),$$

where $\eta(t)$ is a zero mean vector of sampling noise. Combining (5.1) with the perfect observation model, (3.8), yields the system

$$(5.2) \qquad y_*(t)=P'_*(t)y(t-1)+[\eta_*(t)-P'_*(t)\eta(t-1)+\varepsilon_*(t)],$$

or

$$(5.3) \qquad y_*(t)=P'_*(t)y(t-1)+\omega_*(t).$$

The composite noise term, $\omega_*(t)$, in (5.3) is clearly serially correlated. This fact, along with the existence of lagged dependent variables on the right-hand side of (5.3) means that least squares estimates will *not* be consistent. The inconsistency is, of course, simply a manifestation of the well-known "errors in variables" problem; use of an incorrectly measured right-hand side variable in a regression equation brings about a correlation between that variable and the disturbance term of the equation which in turn causes least squares estimates to be inconsistent.

An alternative way of viewing this problem is to focus on the dependence of $y(t)$ upon past observations $y(t-1),\ldots,y(0)$. As was mentioned in Section 3, the distribution of $y(t)$, unlike that of $x(t)$, depends upon all the preceding observed proportions. Therefore, describing $y(t)$ with a first-order autoregressive scheme like (5.3) is a misspecification in the sense that some variables, namely $y(t-2)$, $y(t-3)$, etc., have been omitted. As is well known, if omitted variables are correlated with variables remaining in a model, least squares estimates will be biased. From two points of view, then, that of errors in variables and that of omitted variables, it can be seen that least squares estimation applied to the imperfect observation model (5.3) will give inconsistent estimates.

Instrumental Variables Estimation

One possible approach to the errors in variables problem is simply to ignore it. This is the method used by Lee, Judge, and Zellner in their estimation of constant transition probabilities. Either a simple nonlinear regression can be carried out on system (5.3), or some attempt can be made to adjust for heteroscedasticity in the disturbance term ω_* by using a GLS technique with a covariance matrix constructed iteratively from fitted values for the vectors $y(t)$ and the sample sizes $M(t)$. The resulting estimates will naturally be inconsistent, both for the coefficients and for the covariance matrices (so that it is not clear that GLS provides a gain in efficiency), but this may not be serious if the sample size $M(t)$ is large.

A second technique for dealing with an errors-in-variables problem is instrumental variables estimation. Following the method of Durbin as described in Johnson [**2**], an instrument may be constructed for each of the lagged sample proportions, $y_j(t-1)$, $t=1,2,\ldots,T$. The constructed instrument, $z_j(t)$,

$t=1, 2, \ldots, T$, is equal to the rank ordering of the $y_j(t-1)$ observations, divided by T. That is, z_j consists of the fractions $1/T, 2/T, \ldots, T/T$ arranged in order by size in exactly the way that the elements of lagged y_j are ordered. These s instruments may be used alone or may be combined with other exogenous variables or expressions involving exogenous variables. As shown by Amemiya [1], the use of instrumental variables in a nonlinear model involves choosing the vector of coefficients β to minimize a transformed sum of squared residuals,

$$(5.4) \quad e'_*(I_{s-1} \otimes H(H'H)^{-1}H')e_*,$$

where e_* is the $T \cdot (s-1)$ vector of residuals arranged by equation, and H is a T by k matrix of instruments. The instrumental variable approach will, of course, lead to consistent parameter estimates.

Limited Information Least Squares

The preceding discussion in this section has been based on the conditional distribution of $y(t)$ given $x(t)$. This approach led to a first-order autoregressive model which may be estimated consistently with an instrumental variable technique. As an alternative to the instrumental variable technique, an attempt could be made to incorporate explicitly all of the information contained in the omitted variables by constructing a model based on the conditional distribution of $y(t)$ given $y(t-1), \ldots, y(0)$. This would be very cumbersome and will not be pursued in this paper. One other approach remains, however, based on the *unconditional* distribution of $y(t)$. That is, rather than use more of the available information, a model may be formulated which uses less information, thereby sacrificing some efficiency for consistency and a gain in computational ease.

The last approach mentioned above leads to a rather different regression model specification. As described in Section 3, the unconditional distribution of $y(t)$ is multinomial, with sample size $M(t)$ and mean $\theta(t)$. The vector $y(t)$ may therefore be written as

$$(5.) \quad y(t)=\theta(t)+\delta(t),$$

where $\delta(t)$ is a vector with (unconditional) mean equal to zero. Since $y(t)$ is correlated with past values of y, the vector $\delta(t)$ must be serially correlated, but not in any straightforward fashion. The vector $\theta(t)$ is non-stochastic and varies over time in accordance with the underlying Markov process,

$$(5.6) \quad \theta(t)=P'(t)\theta(t-1).$$

Essentially, the model given by (5.5) solves the problem of omitted variables by omitting all the variables except for a constant. The disturbance term $\delta(t)$, which includes all the lagged y's, is not correlated with the non-stochastic vector $\theta(t)$, and hence presents no consistency problems for least squares estimation.

The alternative model (5.5) may also be derived directly from (5.3) by recursively substituting (5.3) for its lagged right-hand variable, finally replacing $y(0)$ by its expected value $\theta(0)$. The resulting equation involves a product of the transition matrices and is given by

$$(5.7)\qquad \begin{aligned} y(t) = {} & P'(t)P'(t-1)\dots P'(1)\theta(0) \\ & +[\omega(t)+P'(t)\omega(t-1)+P'(t)P'(t-1)\omega(t-2)\dots \\ & +P'(t)\dots P'(1)\delta(0)], \end{aligned}$$

which, given (5.6), is identical to (5.5). This technique of recursive substitution will provide a trade-off of efficiency in return for consistency as long as the product of the coefficients, $P'(t)P'(t-1)\dots P'(1)$, has a finite limit as t becomes infinitely large. This will be true in the present case because of the particular characteristics of the transition matrices.

Consistent, but of course inefficient, estimates of the parameter vector β and the initial state probabilities $\theta(0)$ may be obtained by applying nonlinear least squares to

$$(5.8)\qquad y_*(t) = \theta_*(t) + \delta_*(t),$$

which involves only parameters and the exogenous variables which influence the transition probabilities.

Some gain in efficiency may be obtained by taking account of the interequation covariances in (5.8) while still ignoring the serial correlation. That is, generalized least squares is applied to choose β and $\theta(0)$ to minimize the transformed sum of squared residuals

$$(5.9)\qquad \sum_{t=1}^{T} e_*'(t)\Psi_*^{-1}(t)e_*(t)$$

subject to constraint (5.6), where $e_*(t)$ is the vector of residuals from (5.8) and $\Psi_*(t)$ is the (unconditional) covariance matrix of $y_*(t)$ (and $\delta_*(t)$). Since $y(t)$ is multinomially distributed about $\theta(t)$, the elements of $\Psi(t)$ have the form

$$(5.10)\qquad \begin{aligned} &\Psi_{ii}(t) = (1-\theta_i(t))\theta_i(t)/M(t), \\ &\Psi_{ij}(t) = -\theta_i(t)\theta_j(t)/M(t), \quad \text{for } i \neq j, \end{aligned}$$

and the inverse $\Psi_*^{-1}(t)$ has a particularly simple form as shown by Lee, Judge, and Zellner. Letting superscripts represent elements of the inverse, then

$$(5.11)\qquad \begin{aligned} &\Psi_*^{ii}(t) = M(t)/\theta_s(t) + M(t)/\theta_i(t), \\ &\Psi_*^{ij}(t) = M(t)/\theta_s(t), \end{aligned}$$

where Ψ_* is formed from Ψ by deleting the sth row and column.

An iterative procedure may be adopted to carry out the constrained GLS estimation. An initial guess is made for the unknown matrices $\Psi_*^{-1}(t)$, say $\hat{\Psi}_*^{-1}(t) = I$, and the constrained minimization of (4.10) is carried out with $\hat{\Psi}_*^{-1}(t)$ in place of $\Psi_*^{-1}(t)$. The resulting fitted values $\hat{y}(t) = \hat{\theta}(t)$ are then used to calculate new estimates $\hat{\Psi}_*^{-1}(t)$. At each subsequent iteration, a constrained GLS estimation is performed using estimates $\hat{\Psi}_*^{-1}(t)$ calculated at the end of the previous iteration.

Limited Information Maximum Likelihood Estimation

A full information maximum likelihood estimator for the imperfect observations case would choose β and $\theta(0)$ to maximize the likelihood function of the sample observations,

$$(5.12)\qquad \mathscr{L}\{\beta, \theta(0)\} = \text{pr}\,\{y(1), y(2), \ldots, y(T-1), y(T)\},$$

which may be rewritten as a product of conditional probabilities,

$$(5.13)\qquad \mathscr{L}\{\beta, \theta(0)\} = \text{pr}\,\{(1)\}\cdot\text{pr}\,\{y(2) \mid y(1)\}\ldots\text{pr}\,\{y(T)\,|\,y(t-1), \ldots, y(1)\}.$$

Since the conditional probabilities in (4.14) are complicated expressions, a simplifying approach will be adopted whereby the dependence of each $y(t)$ upon previous observations is ignored, resulting in the limited information likelihood function,

$$(5.14)\qquad \mathscr{L}^*\{\beta, \theta(0)\} = \text{pr}\,\{y(1)\}\cdot\text{pr}\,\{y(2)\}\ldots\text{pr}\,\{y(T)\}.$$

As was discussed above, the unconditional distribution of each $y(t)$ is multinomial about $\theta(t)$ with sample size $M(t)$, which results in the limited information log likelihood function

$$(5.15)\qquad \begin{aligned} L^*\{\beta, \theta(0)\} &= \ln \mathscr{L}^*\{\beta, \theta(0)\} \\ &= \sum_{t=1}^{T}\sum_{j=1}^{s} (\ln M(t)!)m_j(t)(\ln \theta_j(t))/(\ln m_j(t)!), \end{aligned}$$

where $m_j(t) = M(t)y_j(t)$ is the number of individuals observed in state j at time t. The limited information maximum likelihood (LIML) estimate of vector β is obtained by maximizing L^* with respect to β, subject to constraint (5.6).

It will now be shown that the iterative GLS estimator described previously is identical to the limited information maximum likelihood estimator. The first-order conditions for the maximization of the log likelihood function with respect to a vector of parameters β, ignoring constraint (5.6), for the moment, is obtained by differentiating (4.16), taking into account the fact that the θ_j's must sum to 1 in each period:

$$(5.16)\qquad \partial L^*/\partial\beta' = \sum_{t=1}^{T}\sum_{j=1}^{s-1}[\partial L^*/\partial\theta_j(t)][\partial\theta_j(t)/\partial\beta'].$$

The first element on the right of (5.16) is

$$(5.17)\qquad \partial L^*/\partial\theta_j(t) = m_j(t)/\theta_j(t) - m_s(t)/\theta_s(t)$$

which may be rewritten as

$$(5.18)\qquad \begin{aligned} \partial L^*/\partial\theta_j(t) &= [m_j(t) - M(t)\theta_j(t)]/\theta_j(t) - [m_s(t) - M(t)\theta_s(t)]/\theta_s(t) \\ &= [M(t)/\theta_j(t)][y_j(t) - \theta_j(t)] \\ &\quad -[M(t)/\theta_s(t)][y_s(t) - \theta_s(t)]. \end{aligned}$$

Then, using the fact that $y_s(t)-\theta_s(t)=-\Sigma_{k=1}^{s-1}(y_k(t)-\theta_k(t))$,

$$\partial L^*/\partial\theta_j(t)=[M(t)/\theta_s(t)+M(t)/\theta_j(t)][y_j(t)-\theta_j(t)] + \sum_{k\neq j}^{s-1}[M(t)/\theta_s(t)][y_k(t)-\theta_k(t)]. \tag{5.19}$$

Upon comparison of (5.19) with (5.11), it is readily apparent that the derivative may be written in matrix notation as

$$\partial L^*/\partial\theta_*(t)=\Psi_*^{-1}(t)e_*(t) \tag{5.20}$$

so that the derivative $\partial L^*/\partial\beta'$ becomes

$$\partial L^*/\partial\beta'=\sum_{t=1}^{T}e_*'(t)\Psi_*^{-1}(t)[\partial\theta_*(t)/\partial\beta']. \tag{5.21}$$

This derivative is similar to that obtained in the GLS case from (5.9), where $\Psi_*^{-1}(t)$ is held constant at a value calculated from the previous iteration. This means that the GLS derivative at the ith iteration is

$$\begin{aligned}\partial J/\partial\beta' &= \sum_{t=1}^{T}e_*'(t)\hat{\Psi}_*^{-1}(t)[\partial e_*(t)/\partial\beta'] \\ &= -\sum_{t=1}^{T}e_*'(t)\hat{\Psi}_*^{-1}(t)[\partial\theta_*(t)/\partial\beta'].\end{aligned} \tag{5.22}$$

The differences between (5.21) and (5.22) lie entirely in the differences between $\Psi_*^{-1}(t)$ and $\hat{\Psi}_*^{-1}(t)$, the first of which depends upon the β's being estimated and the second of which depends upon the β's estimated in the previous iteration. It can easily be seen that, when the iterative GLS estimators have converged, the derivatives $\partial L^*/\partial\beta'$ and $\partial J/\partial\beta'$ are identical, except for sign which captures the fact that one problem involves maximization and the other involves minimization. Since the unconstrained derivatives, (5.21) and (5.22), converge, and since both GLS and LIML face the same parameter constraint, (5.6), the iterative GLS estimator coincides with the limited information maximum likelihood estimator. This is analogous to the fact that for systems of equations with normally distributed disturbances, the iterative Zellner estimates are the same as maximum likelihood estimates.

The limited information estimation methods which ignore the serial correlation in the observed data, maximum likelihood and least squares, produce consistent but not efficient estimates. In addition to being applicable to models where the group proportions are not perfectly observable, they may also be used in the case of perfect observations. The limited information maximum likelihood estimator discussed in this section will of course not be the same for perfect observations as the maximum likelihood estimator described in Section 4. The latter takes explicit account of the serial correlation in the observed data, which, in the case of perfect observations, can be represented by a simple first-order autoregressive scheme.

The nature of the serial correlation in the imperfect observation case is so complex that no attempt was made in this section to develop a full information maximum likelihood estimator. In theory, it could be done in a manner analogous to that used for perfect observations, but that would not be particularly illuminating.

6. CONCLUSION

In this paper a variety of estimators for time-varying Markov processes have been developed. These estimators are based on the true probability distribution of the observed aggregate proportions and thus augment the set of estimators presented by Lee, Judge, and Zellner which are based on simpler assumptions concerning the stochastic nature of the aggregate data. In particular, a careful distinction is made in this paper between the case of perfect observations in which the entire group is perfectly observed in each period, and the case of imperfect observations where the observations are from different, randomly chosen subgroups in each period.

In the case of perfect observations, the group proportions are distributed as a weighted sum of multinomials, where the weights are the group proportions of the previous period and the individual multinomials have means equal to the rows of the transition probability matrix. In regression equation format, the group proportions may be written as a first-order autoregressive system. Since the associated disturbance term is not correlated with past proportions, nonlinear generalized least squares yield consistent parameter estimates. A full information maximum likelihood estimator is also developed for the perfect observations case, but since the computation involved is formidable, the iterative generalized least squares method may represent the best combination of numerical and statistical efficiency.

With imperfect observations, the distribution of the observed proportions depends upon all past observations, and cannot be characterized as a simple sum of multinomials. Although a maximum likelihood estimator could be constructed, it would be so cumbersome as not to warrant serious consideration. Instead, two different approaches are pursued. The first approach is to characterize the observed proportions by a first-order autoregressive system with serially correlated disturbances. This results in the well-known "errors in variables" problem, so that both nonlinear least squares and nonlinear generalized least squares, as proposed by Lee, Judge, and Zellner, will yield inconsistent estimates. Hence, a consistent instrumental variables estimator is developed, where, following the technique of Durbin, a separate instrument is construction for each lagged proportion.

The second approach is to sacrifice some efficiency to gain consistency by ignoring the information contained in the observed lagged proportions. This limited information approach treats the observed proportions as being independently distributed about their unconditional means. Nonlinear least squares and nonlinear generalized least squares estimators are developed in this limited information framework. A limited information maximum likelihood estimator is also derived and is shown to be identical to the iterative limited information generalized least squares estimation.

Federal Energy Administration, Washington, D.C.

Manuscript received June, 1975; revision received February, 1976.

REFERENCES

[1] AMEMIYA, T.: "The Nonlinear Two Stage Least Squares Estimator," Technical Report No. 116, Institute for Mathematical Studies in the Social Sciences, Stanford University, December, 1973.

[2] JOHNSTON, J.: *Econometric Methods*, 2nd ed. New York: McGraw-Hill, 1972.

[3] LEE, T. C., G. G. JUDGE, AND A. ZELLNER: *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*. Amsterdam: North-Holland, 1970.

[4] MACRAE, C. D., AND E. C. MACRAE: "A Stochastic Model of Job Search and Labor Turnover," Urban Institute Working Paper 350-56, August, 1971.

[5] THEIL, H.: "A Multinomial Extension of the Linear Logit Model," *International Economic Review*, 10 (1969), 251–259.