The Canadian Journal of Statistics
Vol. 12, No. 3, 1984, Pages 169–182
La Revue Canadienne de Statistique

Least-squares estimation of transition probabilities from aggregate data

J. D. KALBFLEISCH and J. F. LAWLESS

University of Waterloo

Key words and phrases: Aggregate data, conditional least squares, asymptotics, Markov chains, estimation.
AMS 1980 *subject classifications*: Primary 62M05; secondary 60J10, 60J20.

ABSTRACT

Consider a population of n individuals that move independently among a finite set $\{1, 2, \ldots, k\}$ of states in a sequence of trials, $t = 0, 1, 2, \ldots, m$, each according to a Markov chain with transition probability matrix $\mathbf{P}$. This paper deals with the problem of estimating $\mathbf{P}$ on the basis of aggregate data which record only the numbers of individuals that occupy each of the k states at times $t = 0, 1, 2, \ldots, m$. Estimation is accomplished using conditional least squares, and asymptotic results are verified for the case $n \to \infty$. A weighted least-squares estimator is introduced and compared with previous estimators. Some comments are made on estimability questions that arise when only aggregate data are available.

RÉSUMÉ

Considérons une population composée de n individus évoluant indépendamment dans le temps à l'intérieur d'un système fini à k états, chacun selon un processus de Markov dont $\mathbf{P}$ est la matrice commune des probabilités de passage. Dans cet article, nous nous penchons sur le problème qui consiste à estimer $\mathbf{P}$ à partir de données regroupées n'indiquant que le nombre d'individus qui occupent chacun des k états aux temps $t = 0, 1, \ldots, m$. L'approche adoptée est celle des moindres carrés conditionnels et les propriétés asymptotiques des estimateurs proposés sont étudiées dans le cas où n tend vers l'infini. Nous suggérons en particulier des estimateurs des moindres carrés pondérés que nous comparons avec d'autres estimateurs déjà connus. Nous discutons aussi des problèmes d'estimabilité qui surgissent lorsque les seules données disponibles sont regroupées.

1. INTRODUCTION

Markov chains are widely used as models in many areas such as economics, population theory, and the social sciences (e.g. Bartholomew 1973, Coleman 1964, Lee et al. 1970). In this paper, we consider estimation of a Markov chain when only aggregate or "macro" data are available, so that, at each observation time, the data specify only the number of individuals in the sample that occupy each state.

Consider a population of individuals that move among a finite set $\{1, 2, \ldots, k\}$ of states in a sequence of trials, $t = 0, 1, 2, \ldots$. For a random sample of size n, let X_{st} represent the state occupied by the sth individual at time t, and suppose that $\{X_{st} : t = 0, 1, 2, \ldots\}$ is a time-homogeneous Markov chain with transition-probability matrix $\mathbf{P} = (p_{ij})_{k \times k}$, $s = 1, 2, \ldots, n$ independently. [See, for example, Cox and Miller (1965, Ch. 3) or Karlin and Taylor (1975, Ch. 2)]. If the data consist of complete transition histories X_{st}, $t = 0, 1, \ldots, m$, $s = 1, 2, \ldots, n$, then methods of estimation and hypothesis testing are well known (e.g. Anderson and Goodman 1957, Bartholomew 1973, Ch. 2). If, on the

other hand, only aggregate data,

$$N_{jt} = \#\{s : X_{st} = j\}, \qquad j = 1, \ldots, k, \quad t = 0, 1, \ldots, m, \tag{1}$$

are available, then estimation is more difficult. The latter problem has been studied by several authors, including Miller (1952), Madansky (1959), and Lee et al. (1970), but numerous problems remain, especially concerning the properties of estimators and methods of interval estimation.

In this paper, least-squares estimation is examined. A weighted least-squares estimator is introduced and compared with previous approaches. Section 2 discusses the various estimation procedures, and Section 3 examines asymptotic properties of estimators and provides large-sample results suitable for estimation and testing. The methods are exemplified in Section 4. Section 5 considers the important but neglected problem of evaluating the information in aggregate data about the parameters of a Markov chain. Finally, Section 6 concludes with some remarks about problems requiring further study.

2. ESTIMATION OF TRANSITION PROBABILITIES

For the moment, we continue to assume that the same $n = \Sigma_{j=1}^{k} N_{jt}$ individuals are observed at each trial t. Let $\mathbf{W}_t = n^{-1}(N_{1t}, \ldots, N_{kt})'$ and $\mathbf{Z}_t = n^{-1}(N_{1t}, \ldots, N_{k-1,t})'$, and define the (unobserved) variates

$$Y_{ijt} = \#\{s : X_{s,t-1} = i, X_{st} = j\}$$

Given $\mathbf{W}_{t-1}$, $\mathbf{Y}_{it} = (Y_{i1t}, \ldots, Y_{ikt})'$ has a k-class multinomial distribution with parameters $N_{i,t-1}; p_{i1}, \ldots, p_{ik}$. Since

$$n\mathbf{W}_t = \sum_{i=1}^{k} \mathbf{Y}_{it}, \qquad t = 1, \ldots, m, \tag{2}$$

the distribution of $\mathbf{W}_t$ given $\mathbf{W}_{t-1}$ is obtainable as a convolution of multinomials. Since $\{\mathbf{W}_t : t = 0, 1, 2, \ldots\}$ is a Markov process, the joint probability function of $\mathbf{W}_1, \ldots, \mathbf{W}_m$ given $\mathbf{W}_0$ is obtained as a product of conditional probabilities, $\Pi_{t=1}^{m}$ pr $(\mathbf{W}_t | \mathbf{W}_{t-1})$. The conditional probability functions are, however, computationally intractable, which prohibits direct maximum-likelihood estimation of $\mathbf{P}$.

The convolution structure (2) allows for simple computation of conditional moments. Since the conditional covariance matrix of $\mathbf{W}_t$ given $\mathbf{W}_{t-1}$ is singular, it is more convenient to work with the equivalent distribution of $\mathbf{Z}_t$ given $\mathbf{W}_{t-1}$. It can be seen that

$$\begin{aligned} \mathscr{E}(\mathbf{Z}_t | \mathbf{W}_{t-1}) &= \mathbf{P}_1' \mathbf{W}_{t-1}, \\ n\,Cov(\mathbf{Z}_t | \mathbf{W}_{t-1}) &= \boldsymbol{\Sigma}_{t-1} = \operatorname{diag}(P_1' \mathbf{W}_{t-1}) - \mathbf{P}_1' \operatorname{diag}(\mathbf{W}_{t-1}) \mathbf{P}_1, \end{aligned} \tag{3}$$

$t = 1, 2, \ldots, m$, where $\mathbf{P}_1$ is the $k \times (k-1)$ matrix obtained by deleting the last column of $\mathbf{P}$. We assume throughout that $\boldsymbol{\Sigma}_{t-1}$ is nonsingular, $t = 1, 2, \ldots, m$.

2.1. Least-Squares Estimation.

To estimate the transition probability matrix $\mathbf{P}$, we consider (conditional) least-squares estimates obtained by minimizing quadratic forms of the type [cf. (3)]

$$S_{\mathbf{Q}} = n \sum_{t=1}^{m} (\mathbf{Z}_t - \mathbf{P}_1' \mathbf{W}_{t-1})' \mathbf{Q}_{t-1} (\mathbf{Z}_t - \mathbf{P}_1' \mathbf{W}_{t-1}). \tag{4}$$

In this, $\mathbf{Q}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_{m-1}$ are suitably chosen positive definite matrices of dimension

$k-1$. The entries of $\mathbf{Q}_{t-1}$ may depend on $\mathbf{W}_l$, $l = 0, \ldots, m$, but are functionally independent of the transition probabilities p_{ij}. In what follows, we consider three such estimators: (i) ordinary least squares, in which $\mathbf{Q}_{t-1} = \mathbf{I}_{k-1}$, where $\mathbf{I}_{k-1}$ is the identity matrix of dimension $k-1$; (ii) weighted least squares, in which $\mathbf{Q}_{t-1} = \hat{\mathbf{\Sigma}}_{t-1}^{-1}$, where $\hat{\mathbf{\Sigma}}_{t-1}^{-1}\mathbf{\Sigma}_{t-1}$ converges in probability to $\mathbf{I}_{k-1}$, and (iii) a modified version of weighted least squares, in which $\mathbf{Q}_{t-1}^{-1} = \text{Diag}(\hat{\mathbf{\Sigma}}_{t-1})$, where $\text{Diag}(A) = \text{diag}(a_{11}, \ldots, a_{kk})$ for $\mathbf{A} = (a_{ij})_{k\times k}$. It is convenient, however, to rewrite (4) to exhibit its properties more clearly before proceeding further.

Suppose that $\mathbf{P}$ contains $r \leq k(k-1)$ linearly independent transition probabilities p_{ij} to be estimated. Note that $r < k(k-1)$ is possible if $\mathbf{P}$ contains structural 0's (i.e. elements that are known to be 0 *a priori*). We write the p_{ij}'s to be estimated as an $r \times 1$ column vector $\boldsymbol{\gamma}$ by proceeding through $\mathbf{P}_1$ columnwise. For example, if there are no structural 0's, so that $r = k(k-1)$, then

$$\boldsymbol{\gamma}' = (\mathbf{p}_1', \ldots, \mathbf{p}_{k-1}'), \tag{5}$$

where $\mathbf{p}_j = (p_{1j}, \ldots, p_{kj})'$ is the jth column of $\mathbf{P}$ (or $\mathbf{P}_1$). We also define $(k-1)\times r$ matrices $\mathbf{B}_t$ such that

$$\mathbf{B}_t\boldsymbol{\gamma} = \mathbf{P}_1'\mathbf{W}_t, \qquad t = 1, \ldots, m. \tag{6}$$

When $r = k(k-1)$ and $\boldsymbol{\gamma}$ is defined as in (5), then

$$\mathbf{B}_t = \mathbf{I}_{k-1} \otimes \mathbf{W}_t', \tag{7}$$

where $\otimes$ denotes the direct, or Kronecker, product of two matrices. It is now clear that (4) can be rewritten as

$$S_{\mathbf{Q}} = n\sum_{t=1}^{m}(\mathbf{Z}_t - \mathbf{B}_{t-1}\boldsymbol{\gamma})'\mathbf{Q}_{t-1}(\mathbf{Z}_t - \mathbf{B}_{t-1}\boldsymbol{\gamma}). \tag{8}$$

For ordinary least squares $\mathbf{Q}_{t-1} = \mathbf{I}_{k-1}$ and (8) is minimized by

$$\tilde{\boldsymbol{\gamma}}_{\text{OLS}} = \left(\sum_{t=1}^{m}\mathbf{B}_{t-1}'\mathbf{B}_{t-1}\right)^{-1}\left(\sum_{t=1}^{m}\mathbf{B}_{t-1}'\mathbf{Z}_t\right), \tag{9}$$

where it is assumed that $\Sigma\mathbf{B}_{t-1}'\mathbf{B}_{t-1}$ is of full rank so that all parameters in $\boldsymbol{\gamma}$ are estimable. The estimate (9) has been given by Madansky (1959) and by Lee et al. (1970, Ch. 3) but in a different form. Madansky notes that (9) gives rise to functionally independent estimates of the columns $\mathbf{p}_j$ of $\mathbf{P}_1$, $j = 1, \ldots, k-1$. For example, if $r = k(k-1)$, then $\mathbf{B}_{t-1}'\mathbf{B}_{t-1} = \mathbf{I}_{k-1} \otimes (\mathbf{W}_{t-1}\mathbf{W}_{t-1}')$ from (7), and (9) gives

$$\tilde{\mathbf{p}}_{j(\text{OLS})} = \left(\sum_{t=1}^{m}\mathbf{W}_{t-1}\mathbf{W}_{t-1}'\right)^{-1}\sum_{t=1}^{m}\mathbf{W}_{t-1}\mathbf{Z}_{jt}, \qquad j = 1, \ldots, k-1. \tag{10}$$

Thus the $\mathbf{p}_j$ estimates are obtained by solving $k-1$ sets of k equations each.

In general, (8) is minimized by

$$\tilde{\boldsymbol{\gamma}}_{\mathbf{Q}} = \left(\sum_{t=1}^{m}\mathbf{B}_{t-1}'\mathbf{Q}_{t-1}\mathbf{B}_{t-1}\right)^{-1}\left(\sum_{t=1}^{m}\mathbf{B}_{t-1}'\mathbf{Q}_{t-1}\mathbf{Z}_t\right), \tag{11}$$

provided again that $\Sigma\mathbf{B}_{t-1}'\mathbf{B}_{t-1}$ is nonsingular. A natural approach is to consider replacing $\mathbf{Q}_{t-1}$ with an estimate of $\mathbf{\Sigma}_{t-1}^{-1}(\boldsymbol{\gamma})$. Thus we consider $\mathbf{Q}_{t-1} = \hat{\mathbf{\Sigma}}_{t-1}^{-1} = \mathbf{\Sigma}_{t-1}^{-1}(\tilde{\boldsymbol{\gamma}}_{\text{OLS}})$. The weighted least-squares estimate is then

$$\tilde{\boldsymbol{\gamma}}_{\text{WLS}} = \left(\sum_{t=1}^{m}\mathbf{B}_{t-1}'\hat{\mathbf{\Sigma}}_{t-1}^{-1}B_{t-1}\right)^{-1}\left(\sum_{t=1}^{m}\mathbf{B}_{t-1}'\hat{\mathbf{\Sigma}}_{t-1}^{-1}\mathbf{Z}_t\right). \tag{12}$$

This procedure can be iterated by estimating $\boldsymbol{\Sigma}_{t-1}(\boldsymbol{\gamma})$ at each stage using $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ from the previous iteration. There is, however, no evidence that such iteration past the first step yields any gains in finite samples, and as $n \to \infty$, the asymptotic properties of the one-step estimator are identical to those of estimators based on further iterations.

An alternative approach, due to Madansky (1959), employs $\mathbf{Q}_{t-1} = \mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{t-1})^{-1}$ in (11), although Madansky does not give the estimator in this form. The resulting estimator $\tilde{\boldsymbol{\gamma}}_{\mathrm{M}}$ is less efficient than $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$, but this approach does facilitate functionally independent estimation of $\mathbf{p}_j$ for $j = 1, \ldots, k-1$ as in ordinary least squares. Lee et al. (1970) give other weighted estimators which, like $\tilde{\boldsymbol{\gamma}}_{\mathrm{M}}$, are less efficient than $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$, but these will not be considered here. It is of some interest that $\tilde{\boldsymbol{\gamma}}_{\mathrm{M}}$ and $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ agree if $k = 2$.

It should be noted that some of the estimates of p_{ij} may lie outside the interval (0, 1), and the estimation should be constrained to take account of this range restriction. Lee et al. (1970) discuss some methods of solving constrained minimization problems in this context, but the methods are complicated. If the Markov model is appropriate and n is large, inadmissible estimates occur with low probability. When they do, there is either evidence against the Markov model, or else an ad hoc adjustment of the unrestricted estimate yields satisfactory results. We do not discuss the constrained problem here, though more work needs to done on this.

2.2. *Immigration and Emigration.*

Frequently the individuals under study differ over time, but such variations cause little additional difficulty in the estimation procedures outlined above. Suppose that additions and deletions occur immediately following each trial, and reinterpret $N_{j,t-1}$ in Section 2 as the number of individuals in the jth state when emigration and immigration are completed after the $(t-1)$th trial. Let M_{jt} represent the number of individuals in state j following the tth trial, but prior to any additions or deletions. Let $n_{t-1} = \sum_{j=1}^{k} N_{j,t-1}$ denote the total number "at risk" at trial t, and define $\mathbf{W}_t = n_{t-1}^{-1}(N_{1t}, \ldots, N_{kt})'$ and $\mathbf{Z}_t = n_{t-1}^{-1}(M_{1t}, \ldots, M_{k-1,t})'$. All results of the previous section now apply with only minor changes.

Markov models are often used when transitions occur in continuous time, and a continuous-time Markov process is then more appropriate. Nonetheless, the simpler discrete-time process is sometimes used in such situations. If immigration and emigration are substantial and occur between observation times (between trials), problems do arise with the discrete model. Ad hoc methods of producing an effective number "at risk" $(N_{1t}, \ldots, N_{kt})$ can be devised, but it is preferable to use the continuous model, where immigration and emigration are easily accommodated (Kalbfleisch et al. 1983).

In the remainder of this paper, it is assumed that the same cohort of n individuals is followed at all time points $t = 0, 1, 2, \ldots, m$.

3. ASYMPTOTIC PROPERTIES

There are two limiting situations of interest: In the first, and the one most appropriate for applications, the number of trials m is fixed and the number of individuals $n \to \infty$. In the second, n is fixed and $m \to \infty$. The latter is of interest in a long sequence of observations on an ergodic chain. We concentrate, in the main, on the former.

3.1. *Limiting Distributions for m Fixed and $n \to \infty$.*

Suppose that as $n \to \infty$, $\mathbf{W}_0 \to \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is a fixed vector of constants. It is readily

shown (see Appendix A) that provided plim $(\Sigma_{t=0}^{m-1}\mathbf{B}_t'\mathbf{Q}_t\mathbf{B}_t)$ is positive definite, the estimate $\tilde{\boldsymbol{\gamma}}_Q$ in (11) is consistent and $\sqrt{nm}(\tilde{\boldsymbol{\gamma}}_Q - \boldsymbol{\gamma})$ has a limiting normal distribution with covariance matrix

$$\mathbf{V}_Q = m\Big(\sum_{t=0}^{m-1}\boldsymbol{\beta}_t'\mathbf{Q}_t\boldsymbol{\beta}_t\Big)^{-1}\sum_{t=0}^{m-1}\boldsymbol{\beta}_t'\mathbf{Q}_t\boldsymbol{\delta}_t\mathbf{Q}_t\boldsymbol{\beta}_t\Big(\sum_{t=0}^{m-1}\boldsymbol{\beta}_t'\mathbf{Q}_t\boldsymbol{\beta}_t\Big)^{-1} \tag{13}$$

where $\boldsymbol{\beta}_t = \operatorname{plim}\mathbf{B}_t$, $\boldsymbol{\delta}_t = \operatorname{plim}\boldsymbol{\Sigma}_t$, and the $\mathbf{Q}_t$'s are known (nonrandom) matrices. Thus $\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}}$ is consistent with covariance given by (13) with $\mathbf{Q}_t = \mathbf{I}_{k-1}$. The properties of the weighted least-squares estimator are obtained by consideration of $\mathbf{Q}_t = \boldsymbol{\delta}_t^{-1}$, $t = 1, \ldots, m$; since $\hat{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t(\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}})$ is a consistent estimator of δ_t, it follows that $\sqrt{nm}(\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}} - \boldsymbol{\gamma})$ is asymptotically normal with covariance matrix

$$\mathbf{V}_{\mathrm{WLS}} = m\Big(\sum_{t=0}^{m-1}\boldsymbol{\beta}_t'\boldsymbol{\delta}_t^{-1}\boldsymbol{\beta}_t\Big)^{-1}, \tag{14}$$

which is consistently estimated by $\tilde{\mathbf{V}}_{\mathrm{WLS}} = m(\Sigma_{t=0}^{m-1}B_t'\hat{\boldsymbol{\Sigma}}_t^{-1}B_t)^{-1}$. The asymptotic properties of $\tilde{\boldsymbol{\gamma}}_M$ are similarly determined. Since this work was developed, McLeish (1984) has shown that $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ is actually asymptotically fully efficient.

Madansky (1959) gives expressions only for the covariance matrices of the separate estimates $\tilde{\mathbf{p}}_i$ obtained from $\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}}$ and $\tilde{\boldsymbol{\gamma}}_{\mathrm{M}}$. The full covariance structure is, however, often necessary and has not been given before.

In showing consistency of $\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}}$, we have assumed that $\Sigma_{t=0}^{m-1}\boldsymbol{\beta}_t'\boldsymbol{\beta}_t = \operatorname{plim}\Sigma_{t=0}^{m-1}\mathbf{B}_t'\mathbf{B}_t$ is positive definite. If, for example, the chain is ergodic with equilibrium distribution $\boldsymbol{\pi}$ and $\boldsymbol{\mu}_0 = \boldsymbol{\pi}$, then this condition is not satisfied. To illustrate, consider the case in which $\mathbf{P}$ has no structural 0's. In this case $\mathbf{B}_t'\mathbf{B}_t = \mathbf{I}_{k-1} \otimes (\mathbf{W}_t\mathbf{W}_t')$ from (7), so that

$$\sum_{t=0}^{m-1}\boldsymbol{\beta}_t'\boldsymbol{\beta}_t = \mathbf{I}_{k-1}\otimes\Big(\sum_{t=0}^{m-1}\boldsymbol{\mu}_t\boldsymbol{\mu}_t'\Big), \tag{15}$$

where $\boldsymbol{\mu}_t = \operatorname{plim}\mathbf{W}_t = (\mathbf{P}')'\boldsymbol{\mu}_0$. If $\boldsymbol{\mu}_0 = \boldsymbol{\pi}$, then $\boldsymbol{\mu}_t = \boldsymbol{\pi}$, $t = 0, \ldots, m-1$, and (15) is not positive definite. If $\mathbf{W}_0$ is near $\boldsymbol{\pi}$, then $\Sigma\mathbf{B}_t'\mathbf{B}_t$ will be nearly singular and little information is available about certain aspects of $\mathbf{P}$. This is of importance and discussed further in Section 5.

3.2. *Limiting Distributions for n Fixed, $m \to \infty$.*

This case is more difficult with regard to $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$. For ordinary least squares, van der Plas (1983) has used the results of Klimko and Nelson (1978) to show consistency and asymptotic normality of $\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}}$ when $\mathbf{P}$ is ergodic. His results verify that as $m \to \infty$, $\sqrt{nm}(\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}} - \boldsymbol{\gamma})$ is asymptotically normal with covariance estimated by (13) with $\mathbf{Q}_t = \mathbf{I}_{k-1}$. Recently, McLeish (1984) has given an alternative proof of this result, and has also established asymptotic properties of $\boldsymbol{\gamma}_{\mathrm{WLS}}$ for the case in which $\mathbf{Q}_t$ can depend only on $\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_{t-1}$.

3.3. *Other Limiting Distributions.*

In many instances, a vector of s parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_s)'$ is of interest, where $\psi_i = g_i(\boldsymbol{\gamma})$ is a differentiable function of the transition probabilities, $i = 1, \ldots, s$. For example, Bartholomew (1973, p. 24) discusses some such parameters in the context of models for social mobility. If the derivatives $\partial\psi_i/\partial\gamma_j$ are easily calculated, then application of the δ-method (Rao 1973, p. 388) allows direct calculation of the asymptotic distribution of $\tilde{\boldsymbol{\psi}}$ from that of $\tilde{\boldsymbol{\gamma}}$, where $\tilde{\psi}_i = g_i(\tilde{\boldsymbol{\gamma}})$.

In some situations, calculation of $\partial\psi_i/\partial\gamma_j$ is complicated. An important example concerns the equilibrium or stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$, of an ergodic Markov chain. As is well known, $\boldsymbol{\pi}$ is the unique $k \times 1$ vector that satisfies $P'\boldsymbol{\pi} = \boldsymbol{\pi}$ subject to the constraint $\Sigma\pi_i = 1$. It is easily shown that (Guilbaud 1977)

$$\begin{aligned}\boldsymbol{\pi} &= [(\mathbf{P} - \mathbf{I})(\mathbf{P} - \mathbf{I})' + \mathbf{J}]^{-1}\mathbf{1} \\ &= \mathbf{A}\mathbf{1}\end{aligned}$$

where $\mathbf{I}$ is the $k \times k$ identity matrix, $\mathbf{J}$ is the $k \times k$ matrix of 1's, and $\mathbf{1}$ is a $k \times 1$ vector of 1's. Calculation of the derivatives $\partial\pi_i/\partial p_{ij}$ is possible, although somewhat complicated.

Guilbaud (1977) gives results which allow development of the limiting distribution of $\sqrt{nm}(\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi})$ (as $n \to \infty$, m fixed). Similar remarks hold as $m \to \infty$ with n fixed. He shows that as $n \to \infty$, $\sqrt{nm}(\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi})$ and $-\sqrt{nm}\mathbf{A}(\mathbf{P} - \mathbf{I})(\mathbf{P} - \mathbf{P})'\boldsymbol{\pi}$ have the same limiting distribution provided $\mathbf{P}$ is a $\sqrt{n}$ consistent estimator of P. Let

$$C = \begin{pmatrix} \boldsymbol{\pi}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \boldsymbol{\pi}' & \cdots & \mathbf{0}' \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \boldsymbol{\pi}' \\ -\boldsymbol{\pi}' & -\boldsymbol{\pi}' & \cdots & -\boldsymbol{\pi}' \end{pmatrix}_{k \times k(k-1)}.$$

Then $\sqrt{nm}\ \mathbf{A}(\mathbf{P} - \mathbf{I})(\tilde{\mathbf{P}} - \mathbf{P})\boldsymbol{\pi} = \sqrt{nm}\mathbf{A}(\mathbf{P} - \mathbf{I})\mathbf{C}(\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1)$, where $\boldsymbol{\gamma}_1' = (\mathbf{p}_1', \ldots, \mathbf{p}_{k-1}')$. Since there may be structural 0's in $\mathbf{P}$, some of the entries of $\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1$ may be 0. Let $\mathbf{M}$ be equal to $\mathbf{A}(\mathbf{P} - \mathbf{I})\mathbf{C}$ when there are no structural 0's in $\mathbf{P}$, and otherwise to $\mathbf{A}(\mathbf{P} - \mathbf{I})\mathbf{C}$ less the columns which correspond to the positions of the structural 0's in $\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1$. Then $\sqrt{nm}\mathbf{A}(\mathbf{P} - \mathbf{I})(\tilde{\mathbf{P}} - \mathbf{P})\boldsymbol{\pi} = \sqrt{nm}\mathbf{M}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$, where $\mathbf{M}$ is $k \times r$ and $\boldsymbol{\gamma}$ is $r \times 1$. If $\sqrt{nm}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is asymptotically normal with covariance matrix $\mathbf{V}$, it follows from Guilbaud's result that

$$\sqrt{nm}(\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{L} \mathbf{N}(\mathbf{0}, \mathbf{MVM}'). \tag{16}$$

Note that $\mathbf{MVM}'$ is singular. The covariance matrix in (16) can be estimated by replacing $\boldsymbol{\gamma}$ with $\tilde{\boldsymbol{\gamma}}$ in $\mathbf{M}$ and $\mathbf{V}$.

4. EXAMPLES

In this section, we consider two examples of aggregate data and apply the methods discussed above. The first example is based on simulated data, while the second is drawn from Lee, Judge, and Zellner (1970).

EXAMPLE 4.1. Consider a three state chain with transition matrix

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ 0 & p_{22} & p_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

and suppose that 200 individuals are observed at times 0, 1, 2, and 3 with corresponding frequencies

$$\begin{aligned}\mathbf{N}_0' &= (200, 0, 0), \\ \mathbf{N}_1' &= (123, 58, 19), \\ \mathbf{N}_2' &= (73, 84, 43), \\ \mathbf{N}_3' &= (42, 90, 68).\end{aligned}$$

TABLE 1: Market shares of three cigarette brands, 1933–1943

Year	Camel	Lucky Strike	Chesterfield	Total sales (billions)
1933	0.2794	0.4008	0.3198	91.6
1934	0.3418	0.3301	0.3281	97.6
1935	0.3867	0.3013	0.3120	101.9
1936	0.4074	0.2906	0.3020	113.9
1937	0.4084	0.2949	0.2967	116.9
1938	0.3842	0.3195	0.2963	113.8
1939	0.3746	0.3358	0.2896	114.2
1940	0.3708	0.3500	0.2792	120.0
1941	0.3579	0.3653	0.2768	135.5
1942	0.3527	0.3851	0.2622	154.5
1943	0.3276	0.3875	0.2849	175.5

In the notation of Section 2.1, $\boldsymbol{\gamma} = (p_{11}, p_{12}, p_{22})'$, $\mathbf{B}_{t-1}$ is a 2×3 matrix, and $\boldsymbol{\Sigma}_{t-1}(\boldsymbol{\gamma})$ is 2×2. The least-squares computations are easily performed to give $\tilde{\boldsymbol{\gamma}}_{\text{OLS}} = (0.606, 0.291, 0.823)$ and $\tilde{\boldsymbol{\gamma}}_{\text{WLS}} = (0.601, 0.301, 0.795)$. From (13) and (14), the estimated covariance matrices for $\sqrt{n}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ are

$$\mathbf{V}_{\text{OLS}} = \begin{pmatrix} .1340 & -.1146 & .0718 \\ & .1858 & -.2284 \\ & & .8449 \end{pmatrix}, \qquad \mathbf{V}_{\text{WLS}} = \begin{pmatrix} .1211 & -.0913 & .0000 \\ & .1380 & -.0886 \\ & & .4764 \end{pmatrix}.$$

The estimator $\tilde{\boldsymbol{\gamma}}_M$ of Madansky (1959) was also considered, and the covariance matrix $\tilde{\mathbf{V}}_M$ differs only slightly from $\tilde{\mathbf{V}}_{\text{OLS}}$. Note the $\tilde{\boldsymbol{\gamma}}_{\text{WLS}}$ is considerably more efficent than $\tilde{\boldsymbol{\gamma}}_{\text{OLS}}$, especially in the estimation of p_{22}.

EXAMPLE 4.2. Lee, Judge, and Zellner (1970) consider data on brand preferences of cigarette smokers. The data give market shares (based on total cigarette sales) for three brands: Camel, Lucky Strike, and Chesterfield. Although the data do not indicate preferences for individual smokers, Lee et al. fit a Markov chain in which the yearly proportions are treated as the $\mathbf{N}_t$'s in our notation. This approach would give estimates of preferences if the amount smoked were independent of brand and the number of smokers were nearly constant, with little immigration or emigration over time.

Lee et al. (1970) give data for 1925–43. There are, however, clear indications that the process is nonstationary over this period, and we consider only the data for 1933–43, given in Table 1. Besides going through the type of analysis given by Lee et al., we shall note some difficulties with this kind of data.

We began by fitting a Markov chain with three states and no structural 0's. In so doing, we treated the proportions as the $\mathbf{W}_t$'s in Section 2.1 [e.g. $\mathbf{W}_0 = (0.2794, 0.4008, 0.3198)'$]. This resulted in an estimate $\tilde{\boldsymbol{\gamma}}_{\text{OLS}}$ in which $\tilde{p}_{21}$ and $\tilde{p}_{32}$ were negative. The estimates were close to 0, and so we refitted the chain with $p_{21} = p_{32} = 0$. The estimated transition matrix is

$$\tilde{\mathbf{P}}_{\text{OLS}} = \begin{pmatrix} 0.458 & 0.252 & 0.290 \\ 0 & 0.723 & 0.277 \\ 0.687 & 0 & 0.313 \end{pmatrix}.$$

The corresponding weighted least-squares estimate is

$$\tilde{\mathbf{P}}_{\mathrm{WLS}} = \begin{pmatrix} 0.532 & 0.261 & 0.207 \\ 0 & 0.731 & 0.287 \\ 0.596 & 0 & 0.304 \end{pmatrix}$$

This model fits the data well in that the estimated proportions $(\tilde{\mathbf{P}}_{\mathrm{WLS}})'\mathbf{W}_0$ agree rather well with the observed proportions $\mathbf{W}_t$, $t = 1, 2, \ldots, 10$.

Although the estimation procedures of Section 2 can be used here, the data do not refer directly to individual smokers. Interpretation of $\mathbf{P}$ as a brand-preference transition matrix for smokers requires (as outlined above) that the number of smokers n be constant (with no immigration or emigration), and that the amount smoked be independent of brand. It is therefore somewhat academic to estimate the covariance matrix associated with $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$. Applying the formula in (14) directly, however, one obtains an estimated "covariance matrix" of $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}} - \boldsymbol{\gamma}$ (assuming $n = 1$) as

$$\tilde{\mathbf{V}}_{\mathrm{WLS}} = \begin{pmatrix} 6.4697 & -7.9168 & -0.8688 & 0.9025 \\ & 9.8719 & 1.0276 & -1.1161 \\ & & 2.5363 & -2.6975 \\ & & & 2.9905 \end{pmatrix},$$

where $\boldsymbol{\gamma} = (p_{11}, p_{31}, p_{12}, p_{22})'$. Although this is clearly not an estimate of variance for $\tilde{\boldsymbol{\gamma}}$, it does give some quantitative information on the relative precision of estimation of the different p_{ij}'s. One might also attempt to define an "effective number" of smokers n and so obtain an "absolute" variance estimate of $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$. On the other hand, there are other problems with these data: The assumption of no immigration or emigration is clearly false, and the estimates of transition probabilities will be highly influenced by preferences of new smokers; more information is needed to attempt to separate transition probabilities from initial preferences. In addition, there are other sources of variability in the data which make variance estimation based solely on the Markov chain variability inappropriate. For example, there would be substantial measurement error in the reported proportions. Any detailed evaluation of variability must take this into account.

On the assumption that we have a stationary, ergodic Markov chain, let us also estimate the chain's equilibrium distribution $\boldsymbol{\pi}$. Using $\tilde{P}_{\mathrm{WLS}}$, we find

$$\tilde{\boldsymbol{\pi}} = (0.371, 0.337, 0.292)'.$$

From (16), the corresponding estimated covariance matrix for $(\tilde{\pi}_1, \tilde{\pi}_2)'$ (again with $n = 1$) is

$$\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\pi}}} = \begin{pmatrix} 0.0387 & -0.0401 \\ & 0.0673 \end{pmatrix}.$$

As for $\tilde{P}_{\mathrm{WLS}}$, this cannot be taken as an absolute variance estimate for $\tilde{\boldsymbol{\pi}}$, but provides evidence on the relative information available. In particular, note that there is considerably more information about $\boldsymbol{\pi}$ than about the p_{ij}'s individually. That this would be the case with this sort of data is intuitively obvious.

5. INFORMATION CALCULATIONS

As is to some extent apparent on intuitive grounds, aggregate data may provide relatively little information on some aspects of the model. The formulae for asymptotic covariance matrices, (13) and (14), allow for some investigation of the precision of estimation, and

we here discuss some of the main points. We restrict attention to $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ and the corresponding formula (14), and consider one specific example. The results and remarks are, however, representative of other situations.

Suppose that the true transition-probability matrix is

$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \tag{17}$$

that m is fixed, and that $n \to \infty$ in such a way that $(1/n)\,\mathbf{N}_0 \to \boldsymbol{\mu}_0 = (\mu_1, \mu_2, \mu_3)'$. It follows that $\boldsymbol{\mu}_l = \operatorname{plim}(1/n)\mathbf{N}_l = (\mathbf{P}')^l\boldsymbol{\mu}_0$, and from (14), the asymptotic covariance matrix for $\sqrt{n}(\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}} - \boldsymbol{\gamma})$ is the 6×6 matrix

$$\mathbf{V}_{\mathrm{WLS}} = \Big(\sum_{l=1}^{m} \boldsymbol{\delta}_{l-1}^{-1} \otimes (\boldsymbol{\mu}_{l-1}\boldsymbol{\mu}_{l-1}') \Big)^{-1}. \tag{18}$$

To illustrate some key points, we consider some specific examples. First, suppose that $\boldsymbol{\mu}_0 = (1, 0, 0)'$, and consider (i) $m = 3$ and (ii) $m = 6$. From (18), the covariance matrices can be computed to give

(i)
$$\mathbf{V}_{\mathrm{WLS}} = \begin{pmatrix} 0.24 & -0.96 & 1.44 & -0.18 & 0.72 & -1.08 \\ & 25.29 & -53.01 & 0.72 & -18.43 & 37.97 \\ & & 114.4 & -1.08 & 37.97 & -81.43 \\ & & & .21 & -0.84 & 1.26 \\ & & & & 27.81 & -58.59 \\ & & & & & 128.6 \end{pmatrix},$$

(ii)
$$\mathbf{V}_{\mathrm{WLS}} = \begin{pmatrix} 0.23 & -0.46 & 0.39 & -0.17 & 0.34 & -0.29 \\ & 3.50 & -4.15 & 0.34 & -2.43 & 2.83 \\ & & 5.41 & -0.29 & 2.84 & -3.61 \\ & & & 0.20 & -0.42 & 0.35 \\ & & & & 3.90 & -4.76 \\ & & & & & 6.40 \end{pmatrix}.$$

Though these covariance matrices apply to the limiting distribution of $\sqrt{n}(\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}} - \boldsymbol{\gamma})$, they can be used with even moderately large n to indicate the precision of the various parameter estimates. For example, with $n = 400$, (i) gives standard deviation estimates for p_{11} and p_{32} (the first and sixth entries in $\tilde{\boldsymbol{\gamma}}$, respectively) as $\frac{1}{20}(0.24)^{\frac{1}{2}} = 0.024$ and $\frac{1}{20}(128.6)^{\frac{1}{2}} = 0.57$, respectively. Clearly, a sample of this size is totally inadequate for estimating p_{32} in this situation, though one does obtain a reasonably precise estimate of p_{11}. The longer observation period ($m = 6$) in (ii) improves the situation regarding p_{32} somewhat (standard deviation 0.126), but this is still unsatisfactory.

The initial vector $\boldsymbol{\mu}_0$ also exhibits considerable influence on the estimation of parameters. For example, if $\boldsymbol{\mu}_0$ is close to a steady-state probability vector, there may be little information present about many of the parameters in the model. To illustrate, consider (17) again, which is the transition-probability matrix for an ergodic chain with equilibrium distribution $\boldsymbol{\pi} = (0.20, 0.35, 0.45)'$. Consider the case with $m = 6$, but $\boldsymbol{\mu}_0 = (0.4, 0.4, 0.2)'$ instead of $(1, 0, 0)'$ as earlier. From (18), we now know that

$$\mathbf{V}_{\mathrm{WLS}} = \begin{pmatrix} 29.00 & -34.95 & 18.50 & -18.35 & 21.82 & -11.36 \\ & 44.48 & -24.88 & 21.82 & 27.45 & 15.15 \\ & & 14.82 & -11.37 & 15.15 & -8.92 \\ & & & 37.32 & -45.27 & 24.08 \\ & & & & 57.91 & -32.53 \\ & & & & & 19.44 \end{pmatrix}.$$

There is much less information about various paramters than when $\boldsymbol{\mu}_0 = (1, 0, 0)'$. With $n = 400$ the standard deviations of the $\tilde{p}_{ij}$'s are so large as to make estimation of them quite impractical.

Although the two initial conditions give quite different results with regard to estimation of the p_{ij}'s, both lead to precise estimation of the equilibrium probability vector $\boldsymbol{\pi}$. Using (16) in Section 3.3, we find that when $\boldsymbol{\mu}_0 = (0.4, 0.4, 0.2)'$ and $(1, 0, 0)'$, the asymptotic distribution of $\sqrt{n}(\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi})$ has covariance matrices

$$\begin{pmatrix} 0.8346 & -0.1928 & -0.6418 \\ & 1.3122 & -1.1194 \\ & & 1.7612 \end{pmatrix} \text{ and } \begin{pmatrix} 0.9437 & -0.2810 & 0.6627 \\ & 1.3070 & -1.0259 \\ & & 1.6887 \end{pmatrix}$$

respectively.

The above indicates that in many situations very large samples are necessary to estimate transition probabilities precisely. One consequence of this is that inadmissible $\tilde{p}_{ij}$'s (i.e. values outside of the range 0−1) are frequently encountered when working with aggregate data. An additional point is that long series of observations (i.e. fairly large m) can often only partly overcome the effect of a small number of study individuals (n), or an unfavourable initial disposition of individuals across states.

Finally, it is possible to compare the precision (in terms of asymptotic variance) of various least-squares estimators, using (13). The estimator $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ gives smallest asymptotic variance for the $\tilde{p}_{ij}$'s and linear functions of them, though in many situations $\tilde{\boldsymbol{\gamma}}_{\mathrm{OLS}}$ and $\tilde{\boldsymbol{\gamma}}_{\mathrm{M}}$ are not substantially less precise than $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$. To give general guidelines on the relative efficiencies of the estimators does not seem feasible, however, nor is it pressing to do so, since all of the estimators ($\boldsymbol{\gamma}_{\mathrm{WLS}}$ included) are easy to compute.

Lawless and McLeish (1984) examine the information in aggregate data from a Markov chain in more detail, for both of the cases $n \to \infty$, m fixed, and $m \to \infty$, n fixed.

6. ADDITIONAL REMARKS

When dealing with aggregate data, model assessment is restricted to questions related to the (time-homogeneous) Markov nature of the $\mathbf{N}_l$'s. For example $S_{\mathbf{Q}}$ given by (8) with $\mathbf{Q}_{l-1} = \boldsymbol{\Sigma}_{l-1}$ is asymptotically $\chi^2_{m(k-1)}$ as $n \to \infty$; if $\tilde{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ is substituted for $\boldsymbol{\gamma}$ in $\boldsymbol{\Sigma}_{l-1}$, it is $\chi^2_{m(k-1)-r}$. This can be used to assess fit, as can the examination of individual residual vectors $\mathbf{Z}_l - \mathbf{P}_l'\mathbf{W}_{l-1}$, $l = 1, \ldots, m$. With a sufficiently long series of observation times, one can also investigate possible time inhomogeneity in the transition probabilities by fitting different models to different sections of the data.

As noted above, aggregate data do not contain a lot of information about certain aspects of a Markov chain. As a result, large numbers of individuals are needed to obtain reliable parameter estimates. Thus, observational aggregate data may not be very helpful in some instances. If one has a choice of obtaining aggregate data on a great many individuals, or complete transition data on a few, the latter alternative would often be preferable. Lawless and McLeish (1984) study this further.

It should also be noted that measurement or other random error associated with the $\mathbf{N}_t$'s is often present, and should be taken into account when analyzing aggregate data. Many of the examples of Lee, Judge, and Zellner (1970) would involve substantial measurement error, and this would affect the standard errors of parameter estimation. Further investigation is needed in this area.

APPENDIX A

The following result, although very straightforward and similar to results of Chiang (1956), does not appear to be in the literature.

Let $\{\mathbf{X}_n\}$ and $\{\mathbf{Y}_n\}$ be sequences of random matrices with $\mathbf{X}_n$ and $\mathbf{Y}_n$ defined on the same probability space. Suppose that

$$\operatorname{plim} \mathbf{X}_n = \mathbf{G}, \quad \operatorname{plim} \mathbf{Y}_n = \mathbf{H} \tag{A1}$$

for constant matrices $\mathbf{G}$ and $\mathbf{H}$, where $\mathbf{X}_n$ and $\mathbf{G}$ have dimension $s \times 1$ and $\mathbf{Y}_n$ and $\mathbf{H}$ have dimension $s \times r$. Let $\boldsymbol{\theta}$ be an $r \times 1$ vector of constants, and suppose that

$$\sqrt{n}(\mathbf{X}_n - \mathbf{Y}_n\boldsymbol{\theta}) \xrightarrow{L} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{A2}$$

where $\mathbf{G} = \mathbf{H}\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ is an $r \times r$ positive definite matrix. Let $\tilde{\boldsymbol{\theta}}_n$ be the weighted least-squares estimate

$$\tilde{\boldsymbol{\theta}}_n = (\mathbf{Y}_n'\mathbf{Q}\mathbf{Y}_n)^{-} (\mathbf{Y}_n'\mathbf{Q}\mathbf{X}_n) \tag{A3}$$

obtained by minimizing

$$S = n(\mathbf{X}_n - \mathbf{Y}_n\boldsymbol{\theta})' \mathbf{Q}(\mathbf{X}_n - \mathbf{Y}_n\boldsymbol{\theta}),$$

where $\mathbf{Q}$ is an arbitrary positive definite asymmetric matrix of dimension $s \times s$ and $(\mathbf{Y}_n'\mathbf{Q}\mathbf{Y}_n)^{-}$ denotes a generalized inverse.

Theorem 1. *If* (A1) *and* (A2) *hold and* $\mathbf{H}'\mathbf{Q}\mathbf{H}$ *is positive definite, then* $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ *is consistent and asymptotically normal with covariance matrix*

$$\mathbf{V}_\mathbf{Q} = (\mathbf{H}'\mathbf{Q}\mathbf{H})^{-1}\mathbf{H}'\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{H}(\mathbf{H}'\mathbf{Q}\mathbf{H})^{-1}. \tag{A4}$$

Proof. Let $\mathbf{U}_n = \sqrt{n}\mathbf{Y}_n'\mathbf{Q}(\mathbf{X}_n - \mathbf{Y}_n\boldsymbol{\theta})$, and note that $\mathbf{U}_n$ is asymptotically normal with covariance $\mathbf{H}'\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{H}$. Now

$$\mathbf{U}_n = \sqrt{n}\mathbf{Y}_n'\mathbf{Q}(\mathbf{X}_n - \mathbf{Y}_n\tilde{\boldsymbol{\theta}}_n) + \sqrt{n}\mathbf{Y}_n'\mathbf{Q}\mathbf{Y}_n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}), \tag{A5}$$

and the first term on the right side of (A5) is identically zero. Since $\operatorname{plim} \mathbf{Y}_n = \mathbf{H}$, the result follows. Q.E.D.

It can be easily seen that $\mathbf{V}_\mathbf{Q} - \mathbf{V}_{\boldsymbol{\Sigma}^{-1}}$ is nonnegative definite for all positive definite $\mathbf{Q}$, so that $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ produces an optimal $\tilde{\boldsymbol{\theta}}_n$. In addition, if $\mathbf{S}_n$ is a random matrix with $\operatorname{plim} \mathbf{S}_n = \mathbf{Q}$, then $\hat{\boldsymbol{\theta}}_n$ obtained by minimizing $n(\mathbf{X}_n - \mathbf{Y}_n\boldsymbol{\theta})'\mathbf{S}_n(\mathbf{X}_n - \mathbf{Y}_n\boldsymbol{\theta})$ has the same asymptotic properties as $\tilde{\boldsymbol{\theta}}_n$ based on $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$.

Theorem 1 is sufficient to establish the asymptotic results for $\tilde{\boldsymbol{\gamma}}_\mathbf{Q}$ in Section 3 as $n \to \infty$. We note that (8) can be written as

$$S_\mathbf{Q} = n(\mathbf{Z}^{(n)} - \mathbf{B}^{(n)}\boldsymbol{\gamma})'\mathbf{Q}(\mathbf{Z}^{(n)} - \mathbf{B}^{(n)}\boldsymbol{\gamma}),$$

where

$$\mathbf{Z}^{(n)} = \left(\mathbf{Z}_1' \vdots \cdots \vdots \mathbf{Z}_m'\right)',$$

$$\mathbf{Q} = \operatorname{diag}(\mathbf{Q}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_{m-1}),$$

$$\mathbf{B}^{(n)} = \left(\mathbf{B}_0' \vdots \mathbf{B}_1' \vdots \cdots \vdots \mathbf{B}_{m-1}'\right)'$$

with dimensions $m(k-1) \times 1$, $m(k-1) \times m(k-1)$ and $m(k-1) \times r$ respectively. By the law of large numbers, both $\mathbf{Z}^{(n)}$ and $\mathbf{B}^{(n)}$ have probability limits. As in Section 3, let

$$\operatorname{plim} \mathbf{B}^{(n)} = \boldsymbol{\beta} = \left(\boldsymbol{\beta}_0' \vdots \cdots \vdots \boldsymbol{\beta}_{m-1}'\right)'.$$

It is shown in Appendix B that

$$\sqrt{n}(\mathbf{Z}^{(n)} - \mathbf{B}^{(n)}\boldsymbol{\gamma}) \xrightarrow{\text{L}} \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \operatorname{plim} \operatorname{Diag}(\boldsymbol{\Sigma}_0, \ldots, \boldsymbol{\Sigma}_{m-1}) = \operatorname{Diag}(\boldsymbol{\delta}_0, \ldots, \boldsymbol{\delta}_{m-1}).$$

Thus Theorem 1 applies directly provided $\boldsymbol{\beta}'\boldsymbol{\beta} = \Sigma_{i=0}^{m-1} \boldsymbol{\beta}_i'\boldsymbol{\beta}_i$ and $\boldsymbol{\Sigma}$ are positive definite. The consistency of $\tilde{\boldsymbol{\gamma}}_{\text{OLS}}$ implies that $\hat{\boldsymbol{\Sigma}} = \operatorname{diag}(\boldsymbol{\Sigma}_0(\tilde{\boldsymbol{\gamma}}_{\text{OLS}}), \ldots, \boldsymbol{\Sigma}_{m-1}(\tilde{\boldsymbol{\gamma}}_{\text{OLS}}))$ is a consistent estimate $\boldsymbol{\Sigma}$, which establishes the limiting distribution and asymptotic optimality of $\tilde{\boldsymbol{\gamma}}_{\text{WLS}}$.

APPENDIX B

The arguments of Appendix A establish desired asymptotic results for estimates $\tilde{\boldsymbol{\gamma}}_{\mathbf{Q}}$, provided that $\sqrt{n}(\mathbf{Z}^{(n)} - \mathbf{B}^{(n)}\boldsymbol{\gamma}) \xrightarrow{\text{L}} \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, or equivalently, that the $\sqrt{n}(\mathbf{Z}_j - \mathbf{B}_{j-1}\boldsymbol{\gamma})$'s converge in law to independent $\mathbf{N}(\mathbf{0}, \boldsymbol{\delta}_{j-1})$'s for $j = 1, \ldots, m$. We shall outline a proof of this result here. The results presented are not given in the full generality possible; it will be apparent that many extensions are easily made. The convergence of the terms, $\mathbf{Z}_j - \mathbf{B}_{j-1}\boldsymbol{\gamma}$, in the conditional sum of squares (8) to the multivariate normal is fairly clear on intuitive grounds. It is likely that a proof of these results could also be developed using central limit theorems for martingales [see e.g. McLeish (1973)]. The following argument, however, gives an elementary demonstration in the univariate case; a simple extension handles the general situation.

Let $\{Y_{j,n} : j = 0, 1, \ldots, m\}$ be a Markov sequence of nonnegative integer-valued random variables, for $n = 1, 2, \ldots$. Suppose that $Y_{0n} = \mu_0(n)$ is a constant and that

$$E(Y_{j,n} \mid Y_{j-1,n}) = \mu_j(Y_{j-1,n})$$

$$Var(Y_{j,n} \mid Y_{j-1,n}) = \sigma_j^2(Y_{j-1,n})$$

where μ_j and σ_j do not depend on n. (Results below also hold when this is relaxed.) Suppose further that as $k \to \infty$,

$$\mu_j(k) = O(k^{c_{1j}}), \quad \sigma_j(k) = O(k^{c_{2j}}), \tag{B1}$$

where $0 \le c_{2j} < c_{1j}$, $j = 0, 1, 2, \ldots, m$.

Our interest centers on situations where $Y_{j,n}^* = [Y_{j,n} - \mu_j(Y_{j-1,n})]/\sigma_j(Y_{j-1,n})$ converges in a conditional sense to a normal random variable. To make this precise, we suppose that, for given $\varepsilon > 0$, $z_j \in (-\infty, \infty)$, there exist y_{j-1} such that for all integers $y \ge y_{j-1}$,

$$\left|\operatorname{pr}\{Y_{j,n}^* \ge z_j \mid Y_{j-1,n} = y\} - \{1 - \Phi(z_j)\}\right| < \varepsilon, \tag{B2}$$

where $\Phi(\cdot)$ is the $\mathbf{N}(0, 1)$ distribution function and $j = 1, 2, \ldots, m$.

LEMMA B1 *Let* (B2) *hold, and suppose that*

$$B \subseteq N \times N \times \ldots \times N \times \{y_{j-1}, y_{j-1}+1, \ldots\}$$

where $N = \{0, 1, 2, \ldots\}$. *Then*

$$\left|\operatorname{pr}\{Y^*_{j,n} \geq z_j \mid (Y_{1,n}, \ldots, Y_{j-1,n}) \in B\} - [1 - \Phi(z_j)]\right| < \varepsilon. \tag{B3}$$

Proof Let

$$C_y = \{(i_1, \ldots, i_{j-1}) \in B \mid i_{j-1} = y\}.$$

Then, for all $y \geq y_{j-1}$,

$$\left|\operatorname{pr}\{Y^*_{j,n} \geq z_j \mid C_y\} - \{1 - \Phi(z_j)\}\right| < \varepsilon$$

from the Markov property and (B2). Since $B = \bigcup_{y=y_{j-1}} C_y$, and the C_y's are disjoint, the result (B3) follows. Q.E.D.

Consider now

$$\operatorname{pr}\{Y^*_{j,n} \geq z_j; j = 1, \ldots, m\} = \operatorname{pr}\{Y^*_{1,n} \geq z_1\} \prod_{j=2}^{m} \operatorname{pr}\{Y^*_{j,n} \geq z_j \mid Y^*_{l,n} \geq z_l,\ l = 1, \ldots, j-1\}. \tag{B4}$$

The conditioning event in the jth term on the right side of (B4) can be written

$$Y_{l,n} \geq \mu_l(Y_{l-1,n}) + z_l \sigma_l(Y_{l-1,n}), \qquad l = 1, \ldots, j-1. \tag{B5}$$

It follows from (B1) that, for any given $z_1, \ldots, z_{j-1}$, there exists an N_j such that for any $n > N_j$ the condition (B5) implies that $Y_{j-1,n} > y_{j-1}$ for $j = 1, \ldots, m$. Lemma 1 then shows that the jth term in the product of (B4) is within ε of $1 - \Phi(z_j)$. It follows that for $n > \max(N_1, \ldots, N_m)$,

$$\left|\operatorname{pr}\{Y^*_{j,n} \geq z_j,\ j = 1, \ldots, m\} - \prod_{j=1}^{m} \{1 - \Phi(z_j)\}\right| < 1 - (1 - \varepsilon)^m.$$

Since ε is arbitrary, we have proved

THEOREM 2. *The random variables* $Y^*_{1,n}, \ldots, Y^*_{m,n}$ *converge in law (as* $n \to \infty$*) to independent* $\mathbf{N}(0, 1)$ *variates.*

In the present context, $Y_{j,n}$ is identified with $n\mathbf{Z}_j$ in Section 2 and $Y^*_{j,n}$ with $\sqrt{n}\boldsymbol{\Sigma}_{j-1}^{-\frac{1}{2}}(\mathbf{Z}_j - B_{j-1}\boldsymbol{\gamma})$, which converges, in the conditional sense, to $\mathbf{N}(\mathbf{0}, I)$. It then follows, by an extension of Theorem 2 to the vector case, that as $n \to \infty$

$$\sqrt{n}\boldsymbol{\Sigma}_{j-1}^{-\frac{1}{2}}(\mathbf{Z}_j - B_{j-1}\boldsymbol{\gamma}) \xrightarrow{L} \mathbf{N}(\mathbf{0}, I),$$

$j = 1, \ldots, m$, independently, and since $\boldsymbol{\delta}_{j-1} = \operatorname{plim} \boldsymbol{\Sigma}_{j-1}$,

$$\sqrt{n}(\mathbf{Z}_j \mathbf{B}_{j-1}\boldsymbol{\gamma}) \xrightarrow{L} \mathbf{N}(\mathbf{0}, \boldsymbol{\delta}_{j-1}),$$

as is used in Appendix A.

REFERENCES

Anderson, T.W., and Goodman, L.A. (1957). Statistical inference about Markov chains. *Ann. Math. Statist.*, 28, 89–110.

Bartholomew, D.J. (1973). *Stochastic Models for Social Processes*. Second edition. Wiley, London.

Chiang, C.L. (1956). On regular best asymptotically normal estimates. *Ann. Math. Statist.*, 27, 336–351.

Coleman, J.S. (1964). *Introduction to Mathematical Sociology*. Collier-Macmillan, London.

Cox, D.R., and Miller, H.D. (1965). *The Theory of Stochastic Processes*. Methuen, London.

Guilbaud, O. (1977). Estimating the probability eigenvector and related characteristics of an ergodic transition matrix. *Scand. J. Statist.*, 4, 97–104.

Kalbfleisch, J.D.; Lawless, J.F., and Vollmer, W.M. (1983). Estimation in Markov models from aggregate data. *Biometrics*, 39, 907–919.

Karlin, S., and Taylor, H.M. (1975). *A First Course in Stochastic Processes*. Second Edition. Academic Press, New York.

Klimko, L.A., and Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.*, 6, 629–642.

Lawless, J.F., and McLeish, D.L. (1984). The information in aggregate data from Markov chains. *Biometrika*, 71, No. 3.

Lee, T.C.; Judge, G.C., and Zellner, A. (1970). *Estimating the Parameters of the Markov Model from Aggregate Time Series Data*. North Holland, Amsterdam.

Madansky, A. (1959). Least squares estimation in finite Markov processes. *Psychometrika*, 24, 137–144.

McLeish, D.L. (1973). Dependent central limit theorems and invariance principles. *Ann. Probab.*, 2, 620–628.

McLeish, D.L. (1984). Estimation for aggregate models: The aggregate Markov chain. *Canadian J. Statist.*, 12.

Miller, G.A. (1952). Finite Markov processes in psychology. *Psychometrika*, 17, 149–167.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Second Edition. Wiley, New York.

van der Plas, A. (1983). On the estimation of the parameters of Markov probability models using macro data. *Ann. Statist.*, 11, 78–85.

Received 3 November 1982
Revised 31 January 1984
Accepted 26 February 1984

Faculty of Mathematics
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario N2L 3G1