*Discussiones Mathematicae*
*Probability and Statistics 24 (2004) 77–84*

# LIKELIHOOD AND QUASI-LIKELIHOOD ESTIMATION OF TRANSITION PROBABILITIES

Ewa Bakinowska and Radosław Kala*

*Department of Mathematical and Statistical Methods*
*Agricultural University of Poznań*
*Wojska Polskiego 28, PL 60637 Poznań, Poland*

**Abstract**

In the paper two approaches to the problem of estimation of transition probabilities are considered. The approach by McCullagh and Nelder [5], based on the independent model and the quasi-likelihood function, is compared with the approach based on the marginal model and the standard likelihood function. The estimates following from these two approaches are illustrated on a simple example which was used by McCullagh and Nelder.

**Keywords:** likelihood estimation, quasi-likelihood estimation, transition probabilities, quasi-information matrix.

**2000 Mathematics Subject Classification:** 62F10, 62F05.

## 1. Problem

Longitudinal studies in biology, medicine or sociology often lead to estimation of transition probabilities from aggregate data. As pointed out by Hawkins and Han [1], this idea goes back to Lee, Judge and Zellender [3] and Kalbfleisch, Lawless and Vollomer [2]. A simple example of this kind was considered by McCullagh and Nelder [4], [5]. It concerns the estimation of voter transition probabilities for two parties based on the vote totals in two successive elections. Although the complete data for

*Partially supported by the State Committee for Scientific Research, Grant KBN No. 3 PO6A 021 24.

one constituency can be formed as in Table 1, only the marginal sums are observed.

Table 1. Complete results of two successive elections.

El 2

| El 1 | | 1 | 2 | |
|---|---|---|---|---|
| | 1 | $m_{11}$ | $m_{12}$ | $m_{1\cdot}$ |
| | 2 | $m_{21}$ | $m_{22}$ | $m_{2\cdot}$ |
| | | $m_{\cdot 1}$ | $m_{\cdot 2}$ | $m_{\cdot\cdot}$ |

The problem consists in estimating the probabilities

$$\pi_{r|s} = P(X_2 = r \mid X_1 = s), \quad r, s = 1, 2,$$

where $X_j = r$ denotes the vote for $r$-th party in $j$-th election. Since each voter can take only one from two contrary decisions, we can focus only on $\pi_{1|1}$ and $\pi_{1|2}$.

## 2. Independent distribution model

In the approach of McCullagh and Nelder [5], the rows of Table 1 are regarded as realizations of two independent binomial random variables conditioned by the observed vote sums at the first election. It means that

$$m_{11} \sim B(m_{1\cdot}, \pi_{1|1}) \text{ and } m_{21} \sim B(m_{2\cdot}, \pi_{1|2}). \tag{1}$$

In consequence, for the sum $m_{\cdot 1} = m_{11} + m_{21}$, we have

$$E(m_{\cdot 1} \mid m_{1\cdot}, m_{2\cdot}) = m_{1\cdot}\pi_{1|1} + m_{2\cdot}\pi_{1|2}$$

and

$$Var(m_{\cdot 1} \mid m_{1\cdot}, m_{2\cdot}) = m_{1\cdot}\pi_{1|1}(1 - \pi_{1|1}) + m_{2\cdot}\pi_{1|2}(1 - \pi_{1|2}).$$

The same formulas can be derived considering the convolution of two successive binomials in the frame of the Markov process (see e.g. Kalbfleisch, Lawless and Vollomer, [2]).

Assuming now that there are data from $n$ constituencies for which the same transition probabilities are valid, we obtain the model

$$E(\mathbf{Y} \mid \mathbf{M}) = \mathbf{M}\boldsymbol{\pi}\ , \ \ \mathrm{Cov}(\mathbf{Y} \mid \mathbf{M}) = \mathbf{V}(\boldsymbol{\pi}\ ), \tag{2}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of the totals $m^i_{.1}$, $i = 1, 2, ..., n$, $\mathbf{M}$ is $n \times 2$ matrix formed from the rows $(m^i_{1.}, m^i_{2.})$, $\boldsymbol{\pi}\ = (\pi_{1|1}, \pi_{1|2})^T$, while $\mathbf{V}(\boldsymbol{\pi}\ )$ is an $n \times n$ diagonal matrix with diagonal elements

$$v_i = m^i_{1.}\pi_{1|1}(1 - \pi_{1|1}) + m^i_{2.}\pi_{1|2}(1 - \pi_{1|2}).$$

If $\mathbf{M}$ is of full column rank, an unbiased estimate of $\pi$ can be determined by the root $\hat{\pi}$ of the nonlinear equation

$$U(\boldsymbol{\pi}\ ) = \mathbf{M}^T\mathbf{V}^{-1}(\boldsymbol{\pi}\ )(\mathbf{Y} - \mathbf{M}\boldsymbol{\pi}\ ) = \mathbf{0}, \tag{3}$$

where $U(\boldsymbol{\pi}\ )$ is the quasi-likelihood score function introduced by Wedderburn [6]. As shown by McCullagh and Nelder ([5], p. 336–339), the vector $U(\boldsymbol{\pi}\ )$ under model (2), cannot be the gradient vector of any scalar function $Q(\boldsymbol{\pi}\ )$. Nevertheless, the inverse of the quasi-information matrix

$$-E\left(\frac{\partial U(\boldsymbol{\pi}\ )}{\partial \boldsymbol{\pi}\ ^T}\right) = \mathbf{M}^T\mathbf{V}^{-1}(\boldsymbol{\pi}\ )\mathbf{M}$$

is, under suitable conditions, the asymptotic covariance matrix of $\hat{\boldsymbol{\pi}}\ ,$

$$Cov(\hat{\boldsymbol{\pi}}\ ) = \left(\mathbf{M}^T\mathbf{V}^{-1}(\boldsymbol{\pi}\ )\mathbf{M}\right)^{-1}.$$

## 3. Marginal model

A different approach follows by assuming that the entries of Table 1 are governed by two marginal distributions

$$m_{1\cdot} \sim B(m, \pi_1), \qquad \pi_1 = P(X_1 = 1)$$

and

$$m_{\cdot 1} \sim B(k, \pi_2), \qquad \pi_2 = P(X_2 = 1), \tag{4}$$

where $m$ is not necessarily equal to $k$. In consequence

$$E(m_{\cdot 1}) = k\pi_2$$

and

$$Var(m_{\cdot 1}) = k\pi_2(1 - \pi_2).$$

On the other hand, we have

$$\pi_2 = P(X_1 = 1)P(X_2 = 1 \mid X_1 = 1) + P(X_1 = 2)P(X_2 = 1 \mid X_1 = 2)$$

$$= \pi_1\pi_{1|1} + (1 - \pi_1)\pi_{1|2}.$$

This means that the main source of information on the vector $\boldsymbol{\pi} = (\pi_{1|1}, \pi_{1|2})^T$ is provided by the distribution of $m_{\cdot 1}$. The role of $m_{1\cdot}$ is only auxiliary, supplying the information on $\pi_1$. The probability $\pi_1$ is estimated by the observed proportion $p_1$, $p_1 = m_{1\cdot}/m$. Therefore, replacing $\pi_2$ in (4) by $p_1\pi_{1|1} + q_1\pi_{1|2}$, with $q_1 = 1 - p_1$, we can establish the conditional distribution of $m_{\cdot 1}$ given the fraction $p_1$. In consequence

$$E(m_{\cdot 1} \mid p_1) = (p_1\pi_{1|1} + q_1\pi_{1|2})k$$

and

$$Var(m_{\cdot 1} \mid p_1) = (p_1\pi_{1|1} + q_1\pi_{1|2})(1 - p_1\pi_{1|1} - q_1\pi_{1|2})k.$$

This, in the case of $n$ independent constituencies, leads to the model

$$E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{N}\boldsymbol{\pi}\ , \qquad \mathrm{Cov}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{W}(\boldsymbol{\pi}\ ), \tag{5}$$

where $\mathbf{Y}$ and $\boldsymbol{\pi}$ are defined as in (2), $\mathbf{X}$ is an $n \times 1$ vector of fractions $p_1^i = m_{1.}^i/m^i$, $i = 1, 2, ..., n$, $\mathbf{N}$ is an $n \times 2$ matrix formed from rows $(p_1^i k^i, q_1^i k^i)$, $q_1^i = 1 - p_1^i$, while $\mathbf{W}(\boldsymbol{\pi}\ )$ is an $n \times n$ diagonal matrix with elements

$$w_i = \left(p_1^i \pi_{1|1} + q_1^i \pi_{1|2}\right)\left(1 - p_1^i \pi_{1|1} - q_1^i \pi_{1|2}\right) k^i.$$

When $k^i = m^i$ for all $i$, i.e., when the system is closed, then $\mathbf{N} = \mathbf{M}$. But $w_i \geq v_i$, which implies that

$$\left(\mathbf{M}^T \mathbf{W}^{-1}(\boldsymbol{\pi}\ )\mathbf{M}\right)^{-1} - \left(\mathbf{M}^T \mathbf{V}^{-1}(\boldsymbol{\pi}\ )\mathbf{M}\right)^{-1} \tag{6}$$

is a non-negative definite matrix.

The assumption (4) together with model (5) allow the estimation of the vector $\boldsymbol{\pi}$ by solving the standard likelihood equation

$$U^*(\boldsymbol{\pi}\ ) = \mathbf{N}^T \mathbf{W}^{-1}(\boldsymbol{\pi}\ )(\mathbf{Y} - \mathbf{N}\boldsymbol{\pi}\ ) = \mathbf{0}, \tag{7}$$

in which $U^*(\boldsymbol{\pi}\ )$ is the score function for $n$ independent binomial distributions

$$B(k^i, (p_1^i \pi_{1|1} + q_1^i \pi_{1|2})).$$

The asymptotic covariance matrix of the root of (7), $\bar{\boldsymbol{\pi}}$ , has a form

$$Cov(\bar{\boldsymbol{\pi}}\ ) = \left(\mathbf{N}^T \mathbf{W}^{-1}(\boldsymbol{\pi}\ )\mathbf{N}\right)^{-1}.$$

Under condition $k^i = m^i$ and due to (6) the estimator $\bar{\boldsymbol{\pi}}$ is less efficient than $\hat{\boldsymbol{\pi}}$ . Nevertheless, it should be emphasized that according to the distributions specified in (1) the entries of $\mathbf{Y}$ and $\mathbf{M}$ in model (2) are obtained by aggregation of the complete data as in Table 1.

This can be done only by questioning, in the second election, $m_{1\cdot}$ voters from subpopulation of those who voted for the first party in the first election and, similarly, by questioning $m_{2\cdot}$ voters who voted for the second party in the first election. It means that model (2) applies only to the closed system, i.e., to the populations in which there are no migrations of voters between elections. Model (5) is different. It is based solely on the results of two successive elections,where migrations are permissible. In this sense, the marginal model is less restrictive what, however, is paid off in efficiency of the estimation.

## 4. Example

In order to compare the estimates following from the two models under discussion, we will consider an artificial example which was used by McCullagh and Nelder [5]. To obtain more realistic data, the numbers of voters were multiplied by 10. In this example, it is assumed that in three independent constituencies we have observed the results of the first and second election:

| El 1 | El 2 | |
|---|---|---|
| $m_{1\cdot}$ | $m_{2\cdot}$ | $\mathbf{Y}$ |
| 50 | 50 | 70 |
| 60 | 40 | 50 |
| 40 | 60 | 60 |

After a sequence of iterations the solution of equation (3), corresponding to the quasi-likelihood method, leads to the estimate $\hat{\boldsymbol{\pi}} = (0.3629, 0.8371)^T$, while the estimated covariance matrix is

$$\widehat{Cov}(\hat{\boldsymbol{\pi}}) = \begin{pmatrix} 0.0239 & -0.0223 \\ -0.0223 & 0.0232 \end{pmatrix}.$$

Using equation (7), following from the fully maximum likelihood method, we get:

$$\bar{\boldsymbol{\pi}} = \begin{pmatrix} 0.3597 \\ 0.8397 \end{pmatrix}, \qquad \widehat{Cov}(\bar{\boldsymbol{\pi}}) = \begin{pmatrix} 0.0308 & -0.0289 \\ -0.0289 & 0.0302 \end{pmatrix}.$$

It is easy to note that the estimates obtained are almost identical. However, the estimated variances confirm less efficiency of the second approach which, as was stated in the previous section, is a price for a better adequacy of the marginal model.

## References

[1] D.L. Hawkins and C.-P. Han, *Estimating Transition Probabilities from Aggregate Samples Plus Partial Transition Data*, Biometrics **56** (2000), 848–854.

[2] J.D. Kalbfleisch, J.F. Lawless and W.M. Vollomer, *Estimation in Markov Models from Aggregate Data*, Biometrics **39** (1983), 907–919.

[3] T.C. Lee, G.G. Judge and A. Zellender, *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*, New York, North Holland 1977.

[4] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, London 1983.

[5] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 2nd. ed. Chapman and Hall, London 1989.

[6] R.W.M. Wedderburn, *Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton method*, Biometrika Wedderburn, **61** (1974), 439–447.

Received 4 October 2003