The Statistician (1994)
43, *No.* 4, *pp.* 495–504

The multinomial–Poisson transformation

By STUART G. BAKER†

National Institutes of Health, Bethesda, USA

[Received June 1993. Revised March 1994]

SUMMARY

The multinomial–Poisson (MP) transformation simplifies maximum likelihood estimation in a wide variety of models for multinomial data. On the basis of specialized derivations, investigators have applied the MP transformation to various models. Here we present a general derivation, which is simpler than the specialized derivations and allows investigators to use the MP transformation readily in new models. We also show how the MP transformation can accommodate incomplete multinomial data and how it can assist in finding closed form maximum likelihood estimates and variances. Previous applications include log-linear models, capture–recapture models, proportional hazards models with categorical covariates and generalizations of the Rasch model. New applications include computing the variance of the logarithm of the odds ratio, a model for voter plurality, conditional logistic regression for matched sets and two-stage case–control studies.

Keywords: Case–control; Categorical data; Conditional likelihood; Incomplete data; Rasch model; Survival analysis

1. Introduction

Many problems in statistics give rise to a multinomial likelihood, either unconditional or conditional (Andersen, 1970), in which the multinomial probabilities are a ratio of a function of parameters to a sum of the function of parameters. Because of the complex functional form of the multinomial probabilities, it is often difficult to maximize these likelihoods and to obtain asymptotic variances. To simplify maximum likelihood estimation and computation of asymptotic variances, we can transform the multinomial likelihood into a Poisson likelihood, with additional parameters. We call this the multinomial–Poisson (MP) transformation. The Poisson likelihood is much easier to maximize than the multinomial likelihood and yields identical estimates and asymptotic variances.

Heretofore, such transformations have been conducted on an *ad hoc* basis, for specific problems, each with its own derivation (Brookmeyer and Damiano, 1989; Conaway, 1989, 1992; Cormack, 1990; Cressie and Holland, 1983; Palmgren, 1981; Tjur, 1982; Whitehead, 1980). The general result is simpler to derive. It also extends readily to the analysis of incomplete data and makes it easier to apply the MP transformation to new problems.

The paper is organized as follows. Section 2 derives the MP transformation, and Section 3 extends it to incomplete data. Section 4 discusses closed form and numerical maximum likelihood computation using the MP transformation. Section 5 reviews previous applications, and Section 6 presents new applications.

†*Address for correspondence*: National Cancer Institute, EPN 344, 9000 Rockville Pike, Bethesda, MD 20892, USA.
E-mail: xpv@helix.nih.gov

0039–0526/94/43495

2. Multinomial–Poisson transformation for complete data

To introduce the notation and the MP transformation, we begin with a simple model.

Assume that $\mathbf{Y} = \{Y_1, \ldots, Y_j, \ldots, Y_J\}$ follow a multinomial distribution with parameters proportional to $\exp\beta_j$. Let $\{y_1, \ldots, y_j, \ldots, y_J\}$ denote a realization of $\mathbf{Y}$. The likelihood kernel is

$$L_{\mathrm{M}}(\{\beta_j\}) = \prod_{j=1}^{J} \left(\exp \beta_j \Big/ \sum_{j=1}^{J} \exp \beta_j\right)^{y_j}. \tag{1}$$

The MP transformation of equation (1) is

$$L_{\mathrm{P}}(\phi, \{\beta_j\}) = \prod_{j=1}^{J} \{\exp(\phi + \beta_j)\}^{y_j} \exp\{-\exp(\phi + \beta_j)\}. \tag{2}$$

As we shall prove, maximizing equation (2) over ϕ and $\{\beta_j\}$ yields the same maximum likelihood estimates for $\{\beta_j\}$ as maximizing equation (1) over $\{\beta_j\}$. The reason that the transformation contains the word Poisson is that equation (2) is the likelihood kernel for a vector of independent Poisson random variables: $Y_j \sim P\{\exp(\phi + \beta_j)\}, j = 1, 2, \ldots, J$.

Now consider a more general class of models. Let $\mathbf{Y}_i = \{Y_{i1}, \ldots, Y_{ij}, \ldots\}$, for $i = 1, 2, \ldots, I$, and $j \in J_i$ denote a vector of random variables with a realization $\mathbf{y}_i = \{y_{i1}, \ldots, y_{ij}, \ldots, \}$. The subscript i indexes levels of a categorical covariate or a cross-classification of categorical covariates. We assume that $\mathbf{Y}_i$ follows a multinomial distribution with parameters $\{g_{ij}(\boldsymbol{\beta})/G_i(\boldsymbol{\beta})$, for $j \in J_i\}$, where

$$G_i(\boldsymbol{\beta}) = \sum_{j \in J_i} g_{ij}(\boldsymbol{\beta})$$

and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_q\}$. In the introductory example, $g_{ij}(\boldsymbol{\beta}) = \exp\beta_j$ and $G_i(\boldsymbol{\beta}) = \Sigma_{j=1}^{J} \exp \beta_j$. The kernel of the likelihood is

$$L_{\mathrm{M}}(\boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j \in J_i} \left\{\frac{g_{ij}(\boldsymbol{\beta})}{G_i(\boldsymbol{\beta})}\right\}^{y_{ij}}. \tag{3}$$

Inference is usually based on the method of maximum likelihood. There are four approaches to maximizing equation (3). One is to use a Newton–Raphson algorithm without any transformations or reparameterizations. A second is to use a modification of the EM algorithm for truncated categorical data (Baker, 1991; Dempster *et al.*, 1977; Hartley, 1958; Tu *et al.*, 1993; Turnbull, 1976). A third approach, especially for discrete time survival data, is to use a Newton–Raphson algorithm after first reparameterizing with reverse time hazard functions (Brookmeyer and Liao, 1990; Kalbfleisch and Lawless, 1991). The fourth and simplest approach is the MP transformation.

Let $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_i, \ldots, \phi_I)'$. The MP transformation of equation (3) is the following likelihood kernel:

$$L_{\mathrm{P}}(\boldsymbol{\phi}, \boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j \in J_i} \{g_{ij}(\boldsymbol{\beta}) \exp \phi_i\}^{y_{ij}} \exp\{-g_{ij}(\boldsymbol{\beta}) \exp \phi_i\}. \tag{4}$$

The derivative of the logarithm of equation (4) with respect to ϕ_i is $\Sigma_j\, y_{ij} - G_i(\boldsymbol{\beta}) \exp \phi_i$. Setting the derivative to 0, solving for $\hat{\phi}_i(\boldsymbol{\beta})$ and substituting into equation (4) give $L_{\mathrm{P}}(\boldsymbol{\phi}(\boldsymbol{\beta}), \boldsymbol{\beta}) \propto L_{\mathrm{M}}(\boldsymbol{\beta})$. On the basis of results in Richards (1961) concerning profile likelihoods, it follows that the maximum likelihood estimates of $\boldsymbol{\beta}$ and their asymptotic variances (based on the observed information matrix) are identical for $L_{\mathrm{P}}(\boldsymbol{\phi}, \boldsymbol{\beta})$ and $L_{\mathrm{M}}(\boldsymbol{\beta})$. Therefore we can use equation (4) for maximum likelihood inference about $\boldsymbol{\beta}$. Because of the elimination of $G_i(\boldsymbol{\beta})$ from the denominator of the multinomial probabilities and the creation of a diagonal covariance matrix, it is easier to maximize equation (4) than equation (3).

It is convenient to specify equation (4) by using the following shorthand notation:

$$Y_{ij} \sim P\{g_{ij}(\boldsymbol{\beta}) \exp \phi_i\}, \qquad j \in J_i, \tag{5}$$

since equation (4) is the likelihood kernel corresponding to expression (5). Technically, expression (5) is not correct because, under the multinomial distribution, the sample space for Y_{ij} is finite, whereas the expression assumes an infinite sample space for Y_{ij}. However, expression (5) is a useful mnemonic for equation (4), which helps in specifying the correct model when using software for maximizing a Poisson likelihood.

3. Multinomial–Poisson transformation for incomplete data

The results of the MP transformation are valid when data are incomplete in the second index j. The missing data mechanism can be either ignorable or non-ignorable (Little and Rubin, 1987). Using terminology from Dempster *et al.* (1977), let $\mathbf{z}_i = \{z_{i1}, \ldots, z_{im}, \ldots\}$ denote the complete data, and let $\mathbf{y}_i$ denote the incomplete data. The incomplete and complete data are related as follows:

$$y_{ij} = \sum_{m \in M_j} z_{im}(\boldsymbol{\beta}),$$

where M_j denotes the set of complete data indices associated with incomplete data index j. We assume that z_i is a realization from a multinomial distribution with parameters $\{f_{im}(\boldsymbol{\beta})/F_i(\boldsymbol{\beta})\}$, where

$$F_i(\boldsymbol{\beta}) = \sum_{j \in J_i} \sum_{m \in M_j} f_{im}(\boldsymbol{\beta}).$$

Setting $g_{ij}(\boldsymbol{\beta}) = \Sigma_{m \in M_j} f_{im}(\boldsymbol{\beta})$, and therefore $G_i(\boldsymbol{\beta}) = F_i(\boldsymbol{\beta})$, we see that the likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i=1}^{I} \prod_{j \in J_i} \sum_{m \in M_j} \left\{\frac{f_{im}(\boldsymbol{\beta})}{F_i(\boldsymbol{\beta})}\right\}^{y_{ij}} \xrightarrow{\mathrm{MP}} Y_{ij} \sim P\left\{\sum_{m \in M_j} f_{im}(\boldsymbol{\beta}) \exp \phi_i\right\}. \tag{6}$$

4. Maximum likelihood estimates and asymptotic variances for multinomial–Poisson transformation

4.1. *Closed form solutions*

The MP transformation simplifies the derivation of closed form maximum likelihood estimates and asymptotic variances, which is particularly useful with incomplete data. The likelihood equations for the incomplete data set the expected values of the sufficient statistics, given the complete data, equal to their unconditional expected values (Dempster *et al.*, 1977). Because the Poisson distribution involves independent parameters, it is easier to solve the likelihood equations after making the MP transformation (Baker, 1989, 1991; Baker *et al.*, 1992). Let $\pi(\hat{\boldsymbol{\beta}})$ denote the estimate of interest and let μ_{ij} denote $E(y_{ij})$. To obtain a closed form approximation for the variance of $\pi(\hat{\boldsymbol{\beta}})$, we apply the delta method with the Poisson distribution

$$\widehat{\mathrm{var}}\{\pi(\hat{\boldsymbol{\beta}})\} = \sum_i \sum_j \left(\frac{\partial \pi(\hat{\boldsymbol{\beta}})}{\partial y_{ij}}\right)' \widehat{\mathrm{var}}(y_{ij}) \left(\frac{\partial \pi(\hat{\boldsymbol{\beta}})}{\partial y_{ij}}\right) = \sum_i \sum_j \left(\frac{\partial \pi(\hat{\boldsymbol{\beta}})}{\partial y_{ij}}\right)' \hat{\mu}_{ij} \left(\frac{\partial \pi(\hat{\boldsymbol{\beta}})}{\partial y_{ij}}\right). \tag{7}$$

The advantage of equation (7), instead of a delta method for a multinomial distribution, is the simplification arising from the independence of the y_{ij}. Although the y_{ij} do not really follow a Poisson distribution because the sample space is finite, the approximation works well in practice. In the example in Section 6.1, the variance in equation (7) equalled the asymptotic variance for the multinomial likelihood. With the incomplete data in Baker (1991), the standard errors using equation (7) agree, to five decimal places, with the asymptotic standard errors based on the observed information matrix from a multinomial likelihood. We caution that both equation (7) and the asymptotic variance based on the observed information matrix are inappropriate when parameter estimates lie on the boundary of the parameter space, as may occur with non-ignorable non-response models for incomplete data, or when the sample size is small or moderate.

4.2. *Numerical solutions*

For complete data, we can maximize the Poisson likelihood by using Poisson regression, generalized linear models (McCullagh and Nelder, 1983) or, for hierarchical log-linear models, iterative proportional fitting (Bishop *et al.*, 1975). For incomplete data, we can maximize the Poisson likelihood by using a composite linear model (Baker, 1992, 1994), an exponential family non-linear model (Palmgren and Ekholm, 1987) or the method of Lang (1992).

5. Previous applications of multinomial–Poisson transformation

5.1. *Log-linear models*

Palmgren (1981) derived a special case of the MP transformation for a log-linear model for complete data. Let y_{ij} denote the cell counts in an $I \times J$ table. Under the log-linear model, $g_{ij} = \exp(x_{ij}\boldsymbol{\beta})$, where x_{ij} is a row of a design matrix. Suppose that $\{Y_{ij}\}$ follow a multinomial distribution with parameters $\{g_{ij}/\Sigma_j\, g_{ij}\}$. The likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i=1}^{I} \prod_{j=1}^{J} \left\{\exp(x_{ij}\hat{\boldsymbol{\beta}}) \Big/ \sum_{u=1}^{J} \exp(x_{iu}\boldsymbol{\beta})\right\}^{y_{ij}} \xrightarrow{\mathrm{MP}} Y_{ij} \sim P\{\exp(\phi_i + x_{ij}\boldsymbol{\beta})\}. \tag{8}$$

5.2. *Capture–recapture models*

Baker (1990) proposed a simple EM algorithm for maximizing a conditional likelihood associated with a log-linear model for capture–recapture data (Bishop *et al.*, 1975). In the discussion, Cormack (1990) reanalysed the data by using the Poisson likelihood which arises from the MP transformation. See also Cormack and Jupp (1991). Let i index the stratum of the covariate (age × screen). Let j index the observed cells in $J_i = \{$(positive mammogram, positive self-examination), (positive mammogram, negative self-examination), (negative mammogram, positive self-examination)$\}$ among women with breast cancer. For a log-linear model, the likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i=1}^{I} \prod_{j \in J_i} \left\{\exp(x_{ij}\boldsymbol{\beta}) \Big/ \sum_{u \in J_i} \exp(x_{iu}\boldsymbol{\beta})\right\}^{y_{ij}} \xrightarrow{\mathrm{MP}} Y_{ij} \sim P\{\exp(\phi_i + x_{ij}\boldsymbol{\beta})\}.$$

5.3. *Truncated discrete time survival data*

Brookmeyer and Damiano (1989) discussed an MP transformation for a conditional multinomial likelihood associated with reporting delays for acquired immune deficiency syndrome. Let y_{ij} denote the number of cases reported diagnosed in interval i with reporting delay of j intervals. Let J_i be the longest observed reporting delay associated with a diagnosis in interval i. If $\exp \beta_j$ denotes the unconditional probability of a reporting delay of j intervals, the probability of a reporting delay of j intervals, given a maximum reporting delay of J_i intervals, is $\exp\beta_j/\Sigma_{u=1}^{J_i}\exp\beta_u$. The likelihood kernel and MP transformation are

$$L_{\mathrm{M}}=\prod_{i=1}^{I}\prod_{j=1}^{J_i}\left\{\exp\beta_j\bigg/\sum_{u=1}^{J_i}\exp\beta_u\right\}^{y_{ij}}\xRightarrow{\mathrm{MP}}Y_{ij}\sim P\{\exp(\phi_i+\beta_j)\}.$$

5.4. *Proportional hazards model with categorical covariates*

Whitehead (1980) proposed an MP transformation for use with a proportional hazards model (Cox, 1972) with categorical fixed time covariates. Let i index the risk sets and let $j=1,2,\ldots,J$ index the levels of the covariate. Let $y_{ij}=1$ if the subject who fails in risk set i has covariate in level j, and let $y_{ij}=0$ otherwise. Let N_{ij} denote the number of subjects in risk set i with covariate at level j. The likelihood kernel is

$$L^{0}=\prod_{i=1}^{I}\prod_{j=1}^{J}\left\{\exp(x_{ij}\boldsymbol{\beta})\bigg/\sum_{u=1}^{J}N_{ij}\exp(x_{iu}\boldsymbol{\beta})\right\}^{y_{ij}}.$$

To apply the MP transformation, we multiply L^0 by a constant $\Pi_i\,\Pi_j\,N_{ij}^{y_{ij}}$. This does not change the maximum likelihood estimate or asymptotic variance. The revised likelihood kernel and MP transformation are

$$L_{\mathrm{M}}=\prod_{i=1}^{I}\prod_{j=1}^{J}\left\{N_{ij}\exp(x_{ij}\boldsymbol{\beta})\bigg/\sum_{j=1}^{J}N_{ij}\exp(x_{ij}\boldsymbol{\beta})\right\}^{y_{ij}}\xRightarrow{\mathrm{MP}}Y_{ij}\sim P\{N_{ij}\exp(\phi_i+x_{ij}\boldsymbol{\beta})\}.$$

5.5. *Generalizations of the Rasch model*

Cressie and Holland (1983), Duncan (1984), Kelderman (1984, 1989) and Tjur (1982) derived an MP transformation for the Rasch model (Rasch, 1960). See also Darroch (1981) and McCullagh (1982). More recently, Conaway (1989, 1992) and Agresti (1993) have derived MP transformations for generalizations of the Rasch model.

Consider the generalized Rasch model for repeated polychotomous data (Conaway, 1989) with

$$\Pr(\text{response category } c\,|\text{ subject } s,\text{ wave } k)=\exp(\alpha_{sc}+x_k\boldsymbol{\beta}_c)\bigg/\left\{1+\sum_{u=2}^{C}\exp(\alpha_{su}+x_k\boldsymbol{\beta}_u)\right\},$$

where α_{sc} is a nuisance parameter representing the effect of subject s on category c, $\boldsymbol{\beta}_c$ is a vector of parameters associated with response category $c=1,2,\ldots,C$ and x_k is a vector of covariates corresponding to wave k. Conditioning on the number of responses at each level eliminates the nuisance parameters. For example, consider three levels of response on each of three waves, and suppose that subject 1 responds at levels 2, 3 and 2 on waves 1, 2 and 3 respectively. The likelihood kernel for subject 1 is

$$L^{(1)}=\frac{\exp(\alpha_{12}+x_1\boldsymbol{\beta}_2)}{1+\sum_{u=2}^{3}\exp(\alpha_{1u}+x_1\boldsymbol{\beta}_u)}\,\frac{\exp(\alpha_{13}+x_2\boldsymbol{\beta}_3)}{1+\sum_{u=2}^{3}\exp(\alpha_{1u}+x_2\boldsymbol{\beta}_u)}\,\frac{\exp(\alpha_{12}+x_3\boldsymbol{\beta}_2)}{1+\sum_{u=2}^{3}\exp(\alpha_{1u}+x_3\boldsymbol{\beta}_u)}.$$

Because there are three possible sets of two responses at level 2 and one at level 3, $\{2,2,3\}$, $\{2,3,2\}$ and $\{3,2,2\}$, the conditional likelihood kernel for subject 1 is

$$L_{\mathrm{M}}^{(1)}=\frac{\exp(x_1\boldsymbol{\beta}_2+x_2\boldsymbol{\beta}_3+x_3\boldsymbol{\beta}_2)}{\exp(x_1\boldsymbol{\beta}_2+x_2\boldsymbol{\beta}_2+x_3\boldsymbol{\beta}_3)+\exp(x_1\boldsymbol{\beta}_2+x_2\boldsymbol{\beta}_3+x_3\boldsymbol{\beta}_2)+\exp(x_1\boldsymbol{\beta}_3+x_2\boldsymbol{\beta}_2+x_3\boldsymbol{\beta}_2)}.$$

The conditional likelihood kernel for multiple subjects is

$$L_{\mathrm{M}} = \prod_{s} L_{\mathrm{M}}^{(s)},$$

which can be simplified by grouping subjects in the following manner. Let i denote a set of the form {number of responses at level 1, number at level 2, number at level 3}. Let J_i denote the subsets (indexed by j) of response levels for waves 1–K which correspond to set i. In the above example $i = \{0, 2, 1\}$ and $J_{\{0,2,1\}} = \{\{2, 2, 3\}, \{2, 3, 2\}, \{3, 2, 2\}\}$. Let y_{ij} denote the number of subjects corresponding to subset $j \in J_i$. Let $\delta_{kc} = 1$, for $k \in j$, if the response for wave k in subset j equals c, and let $\delta_{kc} = 0$ otherwise. The conditional likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i=1}^{I} \prod_{j \in J_i} \left\{ \frac{\exp\left(\sum\limits_{k \in j} \sum\limits_{c} x_k \boldsymbol{\beta}_c \delta_{kc} \right)}{\sum\limits_{u \in J_i} \exp\left(\sum\limits_{k \in u} \sum\limits_{c} x_k \boldsymbol{\beta}_c \delta_{kc} \right)} \right\}^{y_{ij}} \xrightarrow{\mathrm{MP}} Y_{ij} \sim P\left\{ \exp\left(\phi_i + \sum_{k \in j} \sum_{c} x_k \boldsymbol{\beta}_c \delta_{kc} \right) \right\}. \qquad (9)$$

In some cases, particularly with the simple Rasch model, it is easier to fit the Poisson distribution in equation (9) if the data are arranged in a C^K-array (Conaway, 1989; Tjur, 1982).

6. New applications of multinomial–Poisson transformation

6.1. *Derivation of variance of logarithm of odds ratio*

In a 2 × 2 contingency table, a basic measure of association is the odds ratio or its logarithm. We can use the MP transformation to derive easily the asymptotic variance of the logarithm of the odds ratio. Surprisingly, many basic texts do not derive the asymptotic variance of the logarithm of the odds ratio (Breslow and Day, 1980; Fienberg, 1980; Schlesselman, 1982). Instead they present it without derivation or cite Woolf (1955) who also did not give a derivation. Tanner (1991) gives a more complicated derivation based on the multinomial distribution. Let y_{11}, y_{12}, y_{21} and y_{22} denote the counts in a 2×2 contingency table. Under a log-linear model, the likelihood kernel and MP transformation, which are a special case of equation (8), are

$$L_{\mathrm{M}} = \left(\frac{1}{1 + \exp \beta_Y} \right)^{y_{11}} \left(\frac{\exp \beta_Y}{1 + \exp \beta_Y} \right)^{y_{12}} \left\{ \frac{1}{1 + \exp(\beta_Y + \beta_{XY})} \right\}^{y_{21}} \left\{ \frac{\exp(\beta_Y + \beta_{XY})}{1 + \exp(\beta_Y + \beta_{XY})} \right\}^{y_{22}}$$

$$\xrightarrow{\mathrm{MP}} \{ y_{11} \sim P(\exp \phi_1), y_{12} \sim P\{\exp(\phi_1 + \beta_Y)\}, y_{21} \sim P(\exp \phi_2), y_{22} \sim P\{\exp(\phi_2 + \beta_Y + \beta_{XY})\}\}.$$

Because the Poisson formulation involves four parameters and four independent cell counts, it is easy to show that

$$\hat{\beta}_{XY} = \log y_{11} - \log y_{12} - \log y_{21} + \log y_{22}.$$

Since $\hat{\mu}_{ij} = y_{ij}$, invoking equation (7) gives

$$\begin{aligned} \operatorname{var}(\hat{\beta}_{XY}) &= \sum_{i=1}^{2} \sum_{j=1}^{2} (\partial \hat{\beta}_{XY} / \partial y_{ij})^2 \hat{\mu}_{ij} \\ &= (1/y_{11})^2 y_{11} + (-1/y_{12})^2 y_{12} + (-1/y_{21})^2 y_{21} + (1/y_{22})^2 y_{22} \\ &= \sum_{i} \sum_{j} 1/y_{ij}. \end{aligned}$$

6.2. *Model for voter plurality*

Esty (1992) presented a model for voter plurality in which each voter evaluates a subset of the nominees. To obtain maximum likelihood estimates, Esty proposed the method of steepest ascent; the MP transformation provides a simpler alternative. Let J_i denote the set of nominees evaluated by individual i. Let β_j denote the probability of voting for nominee j, and let $y_{ij} = 1$ if voter i votes for nominee j, and $y_{ij} = 0$ otherwise. In Esty's model, each voter picks a favourite so, $\Sigma_j y_{ij} = 1$. The likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i=1}^{I} \prod_{j\in J_i} \left(\beta_j \Big/ \sum_{u\in J_i} \beta_u\right)^{y_{ij}} \xrightarrow{\mathrm{MP}} Y_{ij} \sim P(\beta_j \exp \phi_i).$$

6.3. *Conditional logistic regression for matched sets with categorical exposures*

A common method for analysing matched case–control data is conditional logistic regression (Breslow *et al.*, 1978; Breslow and Day, 1980; Farewell, 1979; Prentice and Breslow, 1978; Prentice and Pyke, 1979; Thompson, 1986). With multiple cases in a matched set, maximization of the conditional likelihood is computationally burdensome (Breslow and Day, 1980) despite computational advances (Gail *et al.*, 1981). A topic for future research is whether we can take advantage of the MP transformation to reduce the computational burden with multiple cases and categorical exposures.

The probability that subject k in matched set s is a case is $\exp(\alpha_s + x_{sk}\boldsymbol{\beta})/\{1 + \exp(\alpha_s + x_{sk}\boldsymbol{\beta})\}$, where α_s is a nuisance parameter for matched set s and x_{sk} is a vector of categorical exposures for subject k in set s. Conditioning on the number of cases in each matched set eliminates the nuisance parameters. For example, suppose that matched set 1 consists of a case, a case and a control, with x_{11}, x_{12} and x_{13} respectively. The likelihood kernel for set 1 is

$$L^{(1)} = \frac{\exp(\alpha_1 + x_{11}\boldsymbol{\beta})}{1 + \exp(\alpha_1 + x_{11}\boldsymbol{\beta})} \frac{\exp(\alpha_1 + x_{12}\boldsymbol{\beta})}{1 + \exp(\alpha_1 + x_{12}\boldsymbol{\beta})} \frac{1}{1 + \exp(\alpha_1 + x_{13}\boldsymbol{\beta})}.$$

Conditioning on the event of two cases and one control gives

$$L_{\mathrm{M}}^{(1)} = \frac{\exp(x_{11}\boldsymbol{\beta} + x_{12}\boldsymbol{\beta})}{\exp(x_{11}\boldsymbol{\beta} + x_{12}\boldsymbol{\beta}) + \exp(x_{11}\boldsymbol{\beta} + x_{13}\boldsymbol{\beta}) + \exp(x_{12}\boldsymbol{\beta} + x_{13}\boldsymbol{\beta})}.$$

The conditional likelihood kernel for multiple subjects is

$$L_{\mathrm{M}} = \prod_{s} L_{\mathrm{M}}^{(s)}.$$

The conditional likelihood can be simplified by grouping subjects into exposure patterns (e.g. the number of binary exposures, 0 or 1, equal to 0) indexed by stratum i. Let J_i, indexed by j, denote the possible assignments of exposures to the cases in stratum i. In the above example $J_i = \{\{1, 2\}, \{1, 3\}\ \{2, 3\}\}$. Let y_{ij} denote the number of matched sets corresponding to $j \in J_i$. Let k index the cases in set j. The conditional likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i=1}^{I} \prod_{j\in J_i} \left\{ \frac{\exp\left(\sum_{k\in j} x_{ik}\boldsymbol{\beta}\right)}{\sum_{u\in J_i} \exp\left(\sum_{k\in u} x_{ik}\boldsymbol{\beta}\right)} \right\}^{y_{ij}} \xrightarrow{\mathrm{MP}} Y_{ij} \sim P\left\{\exp\left(\phi_i + \sum_{k\in j} x_{ik}\boldsymbol{\beta}\right)\right\}.$$

6.4. *Two-stage case–control study*

Let i, e and s index disease state (case or control), exposure and stratum respectively. The sampling involves two stages. In stage 1, investigators identify cases and controls and classify them by stratum. In stage 2, investigators take a random sample of the subjects in each stratum and classify them by exposure. The strata variable might represent a surrogate exposure (Carroll *et al.*, 1993) or an additional covariate (Fears and Brown, 1986; Breslow and Cain, 1988; Breslow and Zhao, 1988; Scott and Wild, 1991; Wacholder *et al.*, 1994) which is expensive or difficult to obtain.

Let y_{1is} denote the number of subjects classified as (i, s) in stage 1 and not selected in stage 2. Let y_{2ies} denote the number of subjects classified as (i, e, s) in stage 2. Let

$$f_{ies}(\boldsymbol{\beta}) = \Pr(\text{disease state } i \mid \text{exposure } e \text{ and stratum } s) \Pr(\text{exposure } e, \text{ stratum } s).$$

Following equation (6), the likelihood kernel and MP transformation are

$$L_{\mathrm{M}} = \prod_{i \in \{\text{case, control}\}} \prod_{s} \left\{ \frac{\sum_{e} f_{ies}(\boldsymbol{\beta})}{\sum_{e}\sum_{s} f_{ies}(\boldsymbol{\beta})} \right\}^{y_{1is}} \prod_{s}\prod_{e} \left\{ \frac{f_{ies}(\boldsymbol{\beta})}{\sum_{e}\sum_{s} f_{ies}(\boldsymbol{\beta})} \right\}^{y_{2ies}}$$

$$\xrightarrow{\mathrm{MP}} Y_{1is} \sim P\left\{\sum_{e} f_{ies}(\boldsymbol{\beta}) \exp \phi_i\right\} \quad \text{and} \quad Y_{2ies} \sim P\{f_{ies}(\boldsymbol{\beta}) \exp \phi_i\}.$$

7. Discussion

The MP transformation requires categorical covariates for the asymptotic theory of maximum likelihood estimates to be valid. If we tried to use the MP transformation with continuous covariates, we would need a separate parameter ϕ_i corresponding to the distinct covariate value for the ith subject. In this situation, when the dimension of the parameter space is comparable with the number of observations, the maximum likelihood estimate may be inconsistent or biased (Andersen, 1973; Cox and Hinkley, 1974). On a related matter, the motivation for using a conditional multinomial likelihood in Sections 5.5 and 6.3 was to eliminate the nuisance parameter for each subject and thereby to avoid the problem of too many parameters. The MP transformation does not reintroduce those parameters; instead it introduces many fewer parameters corresponding to sets of responses. Thus the asymptotic theory is still valid.

In summary, the MP transformation can simplify maximization of multinomial likelihoods by substituting a Poisson likelihood with additional parameters for the multinomial likelihood. An important property of the MP transformation is that it is valid when data on the response variable are incomplete. For likelihood-based inference with incomplete multinomial data, the MP transformation is particularly useful.

Acknowledgements

The author thanks Charles Brown, Laurence S. Freedman, Barry Graubard, David Pee and Sylvan Green for reviewing an early draft of this manuscript.

References

Agresti, A. (1993) Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scand. J. Statist.*, **20**, 63–71.

Andersen, E. B. (1970) Asymptotic properties of conditional maximum-likelihood estimators. *J. R. Statist.*.*Soc.* B, **32**, 283–301.

——— (1973) *Conditional Inferences and Models for Measuring*, p. 69. Copenhagen, Mental Hygiensk Forlag.

Baker,. S. G. (1989) Innovations in screening: evaluating periodic screening without using data from a control group. In *Advances in Cancer Control: Innovations and Research*. New York: Liss.

——— (1990) A simple EM algorithm for capture–recapture data with categorical covariates (with discussion). *Biometrics*, **46**, 1193–1200.

——— (1991) Evaluating a new test using a reference test with estimated sensitivity and specificity. *Communs Statist. Theory Meth.*, **20**, 2739–2752.

——— (1992) A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *J. Comput. Graph. Statist.*, **1**, 63–76.

——— (1994) Composite linear models for incomplete categorical data. *Statist. Med.*, **13**, 609–622.

Baker, S. G., Rosenberger, W. and Dersimonian, R. (1992) Closed-form estimates for missing counts in two-way contingency tables. *Statist. Med.*, **11**, 643–657.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: Massachusetts Institute of Technology Press.

Breslow, N. and Cain, K. C. (1988) Logistic regression for two-stage case–control data. *Biometrika*, **75**, 11–20.

Breslow, N. E. and Day, N. E. (1980) *Statistical Methods in Cancer Research*, p. 134. Lyon: International Agency for Research on Cancer.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. and Sabai, C. (1978) Estimation of multiple relative risk functions in matched case–control studies. *Am. J. Epidem.*, **108**, 299–307.

Breslow, N. and Zhao, L. P. (1988) Logistic regression for stratified case–control studies. *Biometrics*, **44**, 891–898.

Brookmeyer, R. and Damiano, A. (1989) Statistical methods for short-term projections of AIDS incidence. *Statist. Med.*, **8**, 23–34.

Brookmeyer, R. and Liao, J. (1990) The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome. *Am. J. Epidem.*, **132**, 355–365.

Carroll, R. J., Gail, M. H. and Lubin, J. H. (1993) Case–control studies with errors in covariates. *J. Am. Statist. Ass.*, **88**, 185–199.

Conaway, M. R. (1989) Analysis of repeated categorical measurements with conditional likelihood methods. *J. Am. Statist. Ass.*, **84**, 53–62.

——— (1992) The analysis of repeated categorical measurements subject to nonignorable nonresponse. *J. Am. Statist. Ass.*, **87**, 817–824.

Cormack, R. M. (1990) Discussion on A simple EM algorithm for capture–recapture data with categorical covariates (by S. G. Baker). *Biometrics*, **46**, 1193–1200.

Cormack, R. M. and Jupp, P. E. (1991) Inference for Poisson and multinomial models for capture–recapture experiments. *Biometrika*, **78**, 911–916.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cressie, N. and Holland, P. (1983) Characterizing the manifest probabilities of latent trait models. *Psychometrika*, **48**, 129–141.

Darroch, J. N. (1981) The Mantel–Haenszel test and tests of marginal symmetry; fixed-effects and mixed models for a categorical response. *Int. Statist. Rev.*, **49**, 285–307.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Duncan, O. D. (1984) Rasch measurement: further examples and discussion. In *Surveying Subjective Phenomena* (eds C. F. Turner and E. Martin), vol. 2, pp. 367–403. New York: Sage.

Esty, W. W. (1992) Votes or competitions which determine a winner by estimating expected plurality. *J. Am. Statist. Ass.*, **87**, 373–376.

Farewell, V. T. (1979) Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**, 27–32.

Fears, T. R. and Brown, C. C. (1986) Logistic regression methods for retrospective case–control studies using complex sampling procedures. *Biometrics*, **42**, 955–960.

Fienberg, S. E. (1980) *The Analysis of Cross-classified Categorical Data*, p. 18. Cambridge: Massachusetts Institute of Technology Press.

Gail, M. H., Lubin, J. H. and Rubinstein, L. V. (1981) Likelihood calculations for matched case–control studies and survival studies with tied death times. *Biometrika*, **68**, 703–707.

Hartley, H. O. (1958) Maximum likelihood estimation from incomplete data. *Biometrics*, **27**, 783–823.

Kalbfleisch, J. D. and Lawless, J. F. (1991) Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statist. Sin.*, **1**, 19–32.

Kelderman, H. (1984) Loglinear Rasch model tests. *Psychometrika*, **49**, 223–245.

——— (1989) Item bias detection using loglinear IRT. *Psychometrika*, **54**, 681–697.

Lang, J. B. (1992) Obtaining the observed information matrix for the Poisson log linear model with incomplete data. *Biometrika*, **79**, 405–407.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

McCullagh, P. (1982) Some applications of quasisymmetry. *Biometrika*, **69**, 303–308.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models.* London: Chapman and Hall.

Palmgren, J. (1981) The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika*, **68**, 563–566.

Palmgren, J. and Ekholm, A. (1987) Exponential family non-linear models for categorical data with errors of observation. *Appl. Stoch. Models Data Anal.*, **3**, 111–124.

Prentice, R. L. and Breslow, N. E. (1978) Retrospective studies and failure time models. *Biometrika*, **65**, 153–158.

Prentice, R. L. and Pyke, R. (1979) Logistic disease incidence models and case–control studies. *Biometrika*, **66**, 403–411.

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Richards, F. S. G. (1961) A method of maximum-likelihood estimation. *J. R. Statist. Soc.* B, **23**, 469–475.

Schlesselman, J. J. (1982) *Case–Control Studies Design, Conduct, Analysis*, p. 176. New York: Oxford University Press.

Scott, A. J. and Wild, C. J. (1991) Fitting logistic regression models in stratified case–control studies. *Biometrics*, **47**, 497–510.

Tanner, M. A. (1991) *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, p. 12. New York: Springer.

Thompson, S. G. (1986) Modelling in matched case–control studies in epidemiology. *Statistician*, **35**, 237–244.

Tjur, T. (1982) A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Statist.*, **9**, 23–30.

Tu, X. M., Ming, X. and Pagano, M. (1993) The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Am. Statist. Ass.*, **88**, 26–36.

Turnbull, B. W. (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc.* B, **38**, 290–295.

Wacholder, S., Carroll, R. J., Pee, D. and Gail, M. H. (1994) The partial questionnaire design for case–control studies. *Statist. Med.*, **13**, 623–634.

Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Appl. Statist.*, **29**, 268–275.

Woolf, B. (1955) On estimating the relation between blood group and disease. *Ann. Hum. Genet.*, **19**, 251–253.