Heumann:

Marginal regression modeling of correlated multicategorical response: A likelihood approach

Sonderforschungsbereich 386, Paper 19 (1996)

Marginal regression modeling of correlated multicategorical response: A likelihood approach

Christian Heumann*
Institut für Statistik
LMU München

Abstract

A full likelihood approach for marginal regression modeling of correlated multicategorical data is proposed. It is in fact an extension of the approach of Fitzmaurice and Laird (1993) for repeated binary response. The association is directly modeled in terms of conditional odds ratio parameters resulting in the fact that the maximum likelihood estimates of mean and association parameters are asymptotically independent. The technical details are worked out and the approach is illustrated with data previously analyzed by Miller, Davis and Landis (1993).

Keywords: conditional log odds ratios, correlated response, discrete data, longitudinal data, marginal model, maximum likelihood, ordinal response, repeated measures

1 Introduction

Semiparametric methods for marginal regression modeling of correlated response have found wide acceptance in the literature since the stimulating paper by Liang and Zeger (1986), which proposed the use of Generalized Estimating Equations (GEE) in such data situations. In the recent decade many proposals have been made for modeling the correlation, especially in the case of repeated binary observations. Prentice (1988) extended the approach of Liang and Zeger using a second score equation for estimating the association. Lipsitz, Laird and Harrington (1991) and Liang, Zeger and Qaqish (1992) proposed odds ratios for modeling the marginal association. Fitzmaurice and Lipsitz (1995) made use of serial odds ratios. Fahrmeir and Pritscher (1995) proposed a GEE method for ordinal response using global cross ratios.

More recently full likelihood methods have become available. They are useful if there are only few observations per sample unit. In this cases likelihood

*e-mail: chris@stat.uni-muenchen.de

approaches are a flexible tool for modeling the marginal distributions and the associations, simultaneously. Limitations of a fixed parameter likelihood approach are caused by the rapidely growing number of parameters when the number of repeated observations is high since the underlying distribution is the multinomial.

First we give a short review of likelihood approaches for multicategorical correlated response proposed in literature.

Molenberghs and Lesaffre (1994) extended the model of Dale (1986) for bivariate ordinal response using a multivariate Plackett-distribution. The marginal pairwise associations are parametrized with global cross-ratios, which reduce to common odds ratios in the binary case. A disadvantage of this approach is that it cannot be used in the case of a repeated nominal response or mixed ordinal and nominal response.

Liang, Zeger and Quaqish (1992) proposed another set of odds ratios for modeling marginal pairwise associations of multicategorical response. They give only detailed description how to extend this approach to a full likelihood approach for binary response in which their approach coincides with the approach of Molenberghs and Lesaffre.

Fitzmaurice and Laird (1993) and Fitzmaurice, Laird and Rotnitzky (1993) proposed a method based on the so-called loglinear representation of the joint distribution. They transformed the canonical parameters to describe the joint distribution in terms of the mean parameters and the canonical association parameters. The canonical association parameters can be interpreted as conditional log odds ratios and contrasts of conditional log odds ratios, where conditioning is done on those response variables not in the set which is conditioned. Hence this approach is in general limited to the case of a balanced number of observed responses per unit.

Lang and Agresti (1994) used a Lagrangian method for simultaneously estimating the marginal and the joint distribution. Balagtas, Becker and Lang (1995) gave an application of this method to a cross-over design. Their method seems to be limited practically if also continuous covariates are present, which is not the case for the other methods mentioned.

In this paper we extend the approach of Fitzmaurice and Laird (1993), based on conditional log odds ratios, to the case of correlated multicategorical response. As in the binary case we get the result to be expected, that the mean and the odds ratio parameters are orthogonal in the sense, that the Fisher information matrix is blockdiagonal. As a consequence the maximum likelihood estimators are asymptotically independent. This orthogonality property also results from the fact that the proposed model is a special case of the family of partly exponential models introduced by Zhao, Prentice and Self (1992).

Section 2 describes the data situation and a possible model for marginal modeling of ordinal response. In Section 3 we give a compact description of the loglinear representation of the joint distribution. Section 4 addresses the problem of computing the joint distribution from marginal and conditional log odds ratio parameters and the problem of deriving maximum likelihood estimates. In Section 5 we apply the method on a dataset previously analyzed by Miller, Davis and Landis (1993) using weighted least squares (WLS) and the GEE approach. Finally we discuss the method in Section 6.

2 Modeling the marginal response

In the following we assume to have $i = 1, \ldots, N$ subjects. For each of these subjects we observe T multicategorical response variables. We assume further that the response within subjects is correlated while the response between subjects is independent. Only for simplicity, not by the model to propose, assume that we are in a longitudinal data situation where we have T repeated measures of a response variable Y_{it} at T timepoints $t = 1, \ldots T$ with $R = r + 1$ categories. As usual the response Y'_{it} is represented as a row vector of r dummies $(Y_{it1}, \ldots, Y_{itj}, \ldots, Y_{itr})$ with $Y_{itj} = 1$ if $Y_{it} = j$ and $Y_{itj} = 0$ if $Y_{it} \neq j$, $j = 1, \ldots, r$. Along with the response we observe a $1 \times p$ vector x'_{it} of covariates for $t = 1, \ldots, T$. The response of subject i can then be arranged in the $T \times r$ matrix $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ and the corresponding covariates can be arranged in the $T \times p$ matrix $X_i = (x_{i1}, \ldots, x_{iT})'$. To avoid complications we assume that the covariates are nonstochastic, but are allowed to be time-varying. If useful for technical work we stack the rows of the $T \times r$-matrix Y_i in a $Tr \times 1$ column vector.

We restrict our attention to the cumulative logit model known as proportional odds model, which is parsimonious for ordinal response. The cumulative logit model with $r + p$ parameters

$$\beta = (\beta_{10}, \ldots, \beta_{r0}, \tilde{\beta}),$$

which are assumed to be constant over time, may be written as

$$P(Y_{it} \leq j | x_{it}, \beta) = \frac{\exp(\beta_{j0} + x'_{it}\tilde{\beta})}{1 + \exp(\beta_{j0} + x'_{it}\tilde{\beta})}. \tag{1}$$

Let

$$\mu_i(\beta) = (\mu_{i11}(\beta), \mu_{i12}(\beta), \ldots, \mu_{i1r}(\beta), \ldots, \mu_{iTr}(\beta))'$$

denote the vector of marginal expectations with elements $\mu_{itj}(\beta) = \mathrm{E}(Y_{itj}|x_{it}) = P(Y_{it} = j|x_{it})$. From (1) we get

$$\mu_{it1} = \frac{\exp(\beta_{10} + x'_{it}\tilde{\beta})}{1 + \exp(\beta_{10} + x'_{it}\tilde{\beta})}, \quad \text{for } j = 1 \tag{2}$$

and

$$\mu_{itj}(\beta) = \frac{\exp(\beta_{j0} + x'_{it}\tilde{\beta})}{1 + \exp(\beta_{j0} + x'_{it}\tilde{\beta})} - \frac{\exp(\beta_{(j-1)0} + x'_{it}\tilde{\beta})}{1 + \exp(\beta_{(j-1)0} + x'_{it}\tilde{\beta})}, \quad \text{for } j = 2, \ldots, r. \tag{3}$$

This and other useful models for multicategorical response are described in detail e.g. in Fahrmeir and Tutz (1994, ch. 3). These models can be extended to the case of time-varying parameters. This becomes useful when covariates do not change in time, which is the case, e.g., for a treatment indicator covariate. Note that, in principle, any other (sufficiently smooth) parametric link function can be chosen in our approach.

3 Loglinear modeling of the joint distribution

Let for any subject i

$$\pi_{i,j_1,j_2,\ldots,j_T} = P(Y_{i1} = j_1, Y_{i2} = j_2, \ldots, Y_{iT} = j_T)$$

$1 \leq j_t \leq R$, $t = 1, \ldots, T$ denote the probability of one of the R^T possible response profiles and let π_i denote the column vector containing all these probabilities. We use the usual order 'last index is varying fastest'. Then, because of the restriction $1'\pi_i = 1$, with $1' = (1, \ldots, 1)$ a $1 \times R^T$ vector, a saturated loglinear model for π_i can be written in compact matrix notation as

$$\pi_i = \frac{\exp(D^{\Psi}\Psi_i + D^{\Omega}\Omega_i)}{1' \exp(D^{\Psi}\Psi_i + D^{\Omega}\Omega_i)}, \tag{4}$$

where D^{Ψ} is a $R^T \times Tr$ and D^{Ω} is a $R^T \times (R^T - Tr - 1)$ design matrix corresponding to the main effect parameters Ψ_i and association parameters Ω_i, respectively, and the denominator in (4) is a scalar (see e.g. Agresti, 1990, p. 437). For D^{Ψ} we adopt the dummy coding scheme where the highest category R is the reference category. The columns of D^{Ω} are then the appropriate products of two up to T columns of D^{Ψ}. D^{Ψ}, D^{Ω} and an interpretation of the parameters Ψ_i and Ω_i in the case of $T = 3, R = 3$ are given in Appendix A. Here we only state that the parameters in Ψ_i are logits of conditional probabilities and the parameters in Ω_i are conditional log odds ratios and contrasts of conditional log odds ratios.

Alternatively, model (4) can be written as

$$\log(\pi_i) = D^{\Psi}\Psi_i + D^{\Omega}\Omega_i - A_i \tag{5}$$

with $A_i = (a_i, a_i, \ldots, a_i)'$ a $R^T \times 1$ vector with elements

$$a_i = \log(1' \exp(D^{\Psi}\Psi_i + D^{\Omega}\Omega_i)). \tag{6}$$

Now let W_i be a vector containing all double, threefold and higher products of the dummies of Y_i in complete analogy to D^{Ω} and let $\nu_i = \mathrm{E}(W_i)$. Then, as shown in Appendix A, we have the following formulas for the moments of Y_i

and W_i:

$$\begin{aligned}
\mu_i &= \mathrm{E}(Y_i) = D'^{\Psi}\pi_i\,, & (7)\\
\nu_i &= \mathrm{E}(W_i) = D'^{\Omega}\pi_i\,, & (8)\\
\mathrm{Cov}(Y_i) &= D'^{\Psi}\,[\mathrm{diag}(\pi_i) - \pi_i{\pi_i}']\,D^{\Psi}\,, & (9)\\
\mathrm{Cov}(Y_i, W_i) &= D'^{\Psi}\,[\mathrm{diag}(\pi_i) - \pi_i{\pi_i}']\,D^{\Omega}\,, & (10)\\
\mathrm{Cov}(W_i, Y_i) &= \mathrm{Cov}(Y_i, W_i)'\,, & (11)\\
\mathrm{Cov}(W_i) &= D'^{\Omega}\,[\mathrm{diag}(\pi_i) - \pi_i\pi_i']\,D^{\Omega}\,. & (12)
\end{aligned}$$

As π_i depends on Ψ_i and Ω_i (formula (4)), we may conclude from formula (7) that μ_i can be written as an explicit function of the canonical parameters (Ψ_i, Ω_i). Since we want to model the joint distribution in terms of $\mu_i(\beta)$ and the association parameters $\Omega_i(\alpha)$, e.g. $\Omega_i(\alpha) = Z_i\alpha$, where Z_i is a design matrix and α is a $q \times 1$ vector of parameters, we adapt the idea of Fitzmaurice and Laird (1993) and proceed a transformation from the canonical parameters (Ψ_i, Ω_i) to (μ_i, Ω_i). As in the binary case the problem arises that the joint distribution cannot be represented explicitly in terms of (μ_i, Ω_i) which complicates the derivation of the score equations for β and α and the task of maximum likelihood estimation. The problem of computing the joint distribution is shortly addressed in the following section. The derivation of the score equations is given in Appendix B. Since at least μ_i is modeled as a function of covariates we are indeed modeling the conditional distribution of Y_i given X_i with parameters (β, α).

4 Maximum likelihood estimation

For each individual i we observe one of the R^T possible response profiles. To compute the likelihood function, expression (8) and the covariance matrices given in (9) to (12), we need the cell probabilities stacked in the vector π_i. For example, in the case $R = 3, T = 3$, the individual contribution to the likelihood function of an individual i with observation $y_i = (y_{i1}, y_{i2}, y_{i3}) = (2, 3, 1)$ – or $y_i = ((0,1), (0,0), (1,0))$ in dummy coding – is $\pi_{i231} = P(Y_{i1} = 2, Y_{i2} = 3, Y_{i3} = 1)$. As there is in general no explicit formula for computing π_i, $i = 1, \ldots, N$ given (β, α) and therefore (μ_i, Ω_i), one can use the Iterative Proportional Fitting (IPF) algorithm to solve this problem. The idea is to create a start table of probabilities which have the desired conditional log odds ratios, e.g. by $\pi_i = \exp(D^{\Omega}\Omega_i)$ and then dividing each cell probability by $1'\pi_i$ to ensure the restriction $1'\pi_i = 1$. Then apply the IPF routine for table standardization as described in Agresti (1990, p. 196) in order to get a table which has the desired margins μ_i without having changed the odds ratios.

The task of deriving the score equations for (β, α) is solved by use of the chain rule, as shown in Appendix B. This results in the following likelihood equations in (β, α):

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)' \operatorname{Cov}(Y_i)^{-1} (y_i - \mu_i) = 0 \tag{13}$$

$$\sum_{i=1}^{N} \left(\frac{\partial \Omega_i}{\partial \alpha}\right)' \left[w_i - \nu_i - \operatorname{Cov}(W_i, Y_i) \operatorname{Cov}(Y_i)^{-1} (y_i - \mu_i)\right] = 0. \tag{14}$$

The equations are the same as in Fitzmaurice and Laird (1993) for the binary case. If $\Omega_i = Z_i \alpha$ the second score equation reduces to

$$\sum_{i=1}^{N} Z_i' \left[w_i - \nu_i - \operatorname{Cov}(W_i, Y_i) \operatorname{Cov}(Y_i)^{-1} (y_i - \mu_i)\right] = 0. \tag{15}$$

Because, as shown in Appendix C, the expected information matrix has block structure, the likelihood equations can be solved by iterating through the following three steps until convergence:

1. Update $\widehat{\beta}$ by the formula

$$\widehat{\beta}^{(s+1)} = \widehat{\beta}^{(s)} + \operatorname{Cov}(\widehat{\beta})^{(s)} \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^{\prime(s)} \operatorname{Cov}(Y_i)^{-1(s)} \left(y_i - \mu_i^{(s)}\right) \right) \tag{16}$$

2. Update $\widehat{\alpha}$ by the formula

$$\widehat{\alpha}^{(s+1)} = \widehat{\alpha}^{(s)} + \operatorname{Cov}(\widehat{\alpha})^{(s)} \times \left(\sum_{i=1}^{N} \left(\frac{\partial \Omega_i}{\partial \alpha}\right)^{\prime(s)} \left[w_i - \nu_i^{(s)} - \operatorname{Cov}(W_i, Y_i)^{(s)} \operatorname{Cov}(Y_i)^{-1(s)} \left(y_i - \mu_i^{(s)}\right)\right] \right) \tag{17}$$

3. For $i = 1, \ldots, N$ compute
   (a) $\mu_i^{(s+1)}$ from $\widehat{\beta}^{(s+1)}$ as specified by the marginal model.
   (b) $\Omega_i^{(s+1)}$ from $\widehat{\alpha}^{(s+1)}$, e.g. $\Omega_i^{(s+1)} = Z_i \widehat{\alpha}^{(s+1)}$.
   (c) the vector of joint probabilities $\pi_i^{(s+1)}$ for each individual i from $(\mu_i^{(s+1)}, \Omega_i^{(s+1)})$ using IPF.
   (d) the updates $\nu_i^{(s+1)}$, $\operatorname{Cov}(Y_i)^{(s+1)}$, $\operatorname{Cov}(W_i, Y_i)^{(s+1)}$ and $\operatorname{Cov}(W_i)^{(s+1)}$ by use of formulas (8) to (12) with $\pi_i^{(s+1)}$ from (c) as current estimate.

The asymptotic covariance matrices of $\hat{\beta}$ and $\hat{\alpha}$ are given by the inverse of the expected fisher information matrix (Appendix C). Fitzmaurice, Laird and

Rotnitzky (1993) proposed a robust estimate for the covariance matrix of $\hat{\beta}$ by using the so-called sandwich formula

$$\widehat{\mathrm{Cov}(\hat{\beta})} = \hat{F}_1^{-1}\hat{F}_2\hat{F}_1^{-1} \tag{18}$$

with

$$\begin{aligned}
\hat{F}_1 &= \sum_{i=1}^{N} \left(\widehat{\frac{\partial\mu_i}{\partial\beta}}\right)' \widehat{\mathrm{Cov}(Y_i)}^{-1} \left(\widehat{\frac{\partial\mu_i}{\partial\beta}}\right) \\
\hat{F}_2 &= \sum_{i=1}^{N} \left(\widehat{\frac{\partial\mu_i}{\partial\beta}}\right)' \widehat{\mathrm{Cov}(Y_i)}^{-1}(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'\widehat{\mathrm{Cov}(Y_i)}^{-1} \left(\widehat{\frac{\partial\mu_i}{\partial\beta}}\right)
\end{aligned}$$

where $\hat{F}_1^{-1}$ is the estimated asymptotic covariance matrix of $\hat{\beta}$ under the assumed model.

5 Respiratory disorder example

We illustrate our approach by applying it on data described in Miller, Davis and Landis (1993), previously analyzed by Koch et al. (1989). 111 patients were randomly assigned to one of two treatments (active, placebo). A 5-point ordinal scale was originally used to classify the response measured at four time points (terrible=0, poor=1, fair=2, good=3, excellent=4). We use the 3-point ordinal data given in Table 1 in Miller et al. The response is denoted as poor (0-1), good (2-3) and excellent (4). Note that only one treatment indicator covariate is given in their table which can be included in the marginal model. 22 of the possible $3^4 = 81$ response profiles were observed in the active group, 23 in the placebo group. Since the two separate response tables, one for the active group and one for the placebo group, are sparse, we only considered models involving second order interactions. Additionally we assumed equal interactions in both treatment groups. The treatment indicator variable x_{it} is effect coded as 1 'active' and -1 'placebo'.

In Table 1 we give parameter estimates with estimated model and robust standard errors for the marginal model

$$\log\left(\frac{P(Y_{it(j)} \leq j|x_{it})}{P(Y_{it(j)} > j|x_{it})}\right) = \beta_j + \tau x_{it}, \qquad j = 1,2; t = 1,2,3,4$$

with constant thresholds β_1, β_2 and a constant treatment effect τ.

In Table 2 we give parameter estimates and estimated standard errors for the marginal model

$$\log\left(\frac{P(Y_{it(j)} \leq j|x_{it})}{P(Y_{it(j)} > j|x_{it})}\right) = \beta_j + \tau_t x_{it}, \qquad j = 1,2; t = 1,2,3,4$$

with constant thresholds β_1, β_2 and visit-specific treatment effects τ_t.

Both models were combined with five models for the association structure. Model I is the independence model. Model II assumes an exchangeable association structure, i.e. $\Omega_i = (\omega_{11}^{12}, \omega_{21}^{12}, \ldots, \omega_{22}^{34}, 0, \ldots, 0)$ with $\omega_{j_1 j_2}^{tt'} = \alpha_{j_1 j_2}$, for all $t, t' = 1, \ldots, 4$, $t < t'$, $j_1, j_2 = 1, 2$. Model III is model II supplementary assuming that $\omega_{j_1 j_2}^{14}$ is zero. In Model IV we assume that the conditional pairwise associations depend on the distance between two visits. Especially we assume for $j_1, j_2 = 1, 2$

$$\begin{aligned} \omega_{j_1 j_2}^{12} = \omega_{j_1 j_2}^{23} = \omega_{j_1 j_2}^{34} &= \alpha_{j_1 j_2}^{1} \\ \omega_{j_1 j_2}^{13} = \omega_{j_1 j_2}^{24} &= \alpha_{j_1 j_2}^{2} \\ \omega_{j_1 j_2}^{14} &= 0 \end{aligned}$$

The last model is a linear by linear association model (Agresti, 1990, p. 263) for the six pairwise conditional associations using equally spaced scores 1,2 and 3. The latter might not be a sensible choice but is often used for ordinal variables to reflect the ordering and to get more parsimonious models but to avoid problems which arise when scores are treated as additional parameters which have to be estimated. Only one parameter is needed for modeling the four log odds ratios for each pair. The conditional log odds ratios then can be written as

$$\omega_{j_1 j_2}^{t_1 t_2} = \alpha^{tt'}(3 - j_1)(3 - j_2) \quad t, t' = 1, 2, 3, 4, t < t'; j_1, j_2 = 1, 2.$$

The model assumes equal conditional *local* odds ratios which are build from adjacent rows and columns. Note that our set of odds ratios differs from the set of local odds ratios. Only the odds ratios denoted by $\omega_{22}^{tt'}$ in our set are local. The expression given above takes this into account appropriately.

From Table 1 model Ia, we see that the model estimates assuming independence underestimate the standard errors for the threshold and the treatment effect. For IIa, IIIa, IVa and Va the estimated model standard errors were, for practical purposes, equal to the robust estimates and equal to the robustified estimates of model Ia, calculated with the sandwich formula (18). Noting that a negative value of the treatment effect parameter says, that the treatment improves the response, it is interesting that, despite the fact that the estimates of the standard errors are equal, the estimated treatment effect is in absolute value equal for Ia and IIa (-0.45) but is lower for model IIIa, IVa and Va (-0.39). This effect nearly disappears for the estimates of the marginal model in Table 2. The variation of the estimates is smaller there over all five models than that of the weighted least square estimates and the estimates obtained by the GEE I correlation method used by Miller et al. The WLS estimates were given (without standard errors) as $\hat{\tau}_1 = -0.25$, $\hat{\tau}_2 = -0.66$, $\hat{\tau}_3 = -0.5$ and $\hat{\tau}_4 = -0.34$, the estimates of the GEE I correlation method were given as $\hat{\tau}_1 = -0.31$, $\hat{\tau}_2 = -0.76$, $\hat{\tau}_3 = -0.54$ and $\hat{\tau}_4 = -0.39$ by the authors.

For models II, III, IV and V all association parameters were significant on the 5%-level with the exception of α^{14} in model Va and Vb. The association models seem to scatch the dependence "relevant" for obtaining "correct" estimates for the standard errors of the marginal parameter estimates although they are very restrictive and, as not shown here, are in fact no good models when we compare the estimated cell probabilities with the observed ones e.g. by standardized residuals. Due to zero cells in the marginal tables, some conditional odds ratio parameters are only identified because of the restrictions in the association models, i.e. we are confronted with the problem of sparse tables in this dataset. More refined association models would also take into account the fact that 42 of the 111 patients have no change in their response over the four time points.

Model		Est.	mod. SE	rob. SE		Est.	mod. SE
Ia	β_1	-1.59	0.13	0.19			
	β_2	0.68	0.10	0.16			
	τ	-0.45	0.09	0.15			
	β_1	-1.57	0.19	0.19	α_{11}	3.33	0.51
IIa	β_2	0.69	0.16	0.16	α_{12}	1.91	0.48
	τ	-0.45	0.15	0.15	α_{21}	1.47	0.37
					α_{22}	1.10	0.14
	β_1	-1.58	0.19	0.19	α_{11}	4.33	0.71
IIIa	β_2	0.69	0.16	0.16	α_{12}	2.37	0.60
	τ	-0.39	0.15	0.15	α_{21}	2.03	0.50
					α_{22}	1.34	0.17
	β_1	-1.58	0.19	0.19	α_{11}^1	4.98	1.24
IVa	β_2	0.69	0.16	0.16	α_{12}^1	2.44	1.01
	τ	-0.39	0.15	0.15	α_{21}^1	2.13	0.89
					α_{22}^1	1.41	0.29
					α_{11}^2	3.42	1.33
					α_{12}^2	2.24	1.04
					α_{21}^2	1.89	0.90
					α_{22}^2	1.25	0.40
	β_1	-1.60	0.19	0.19	α^{12}	1.21	0.37
Va	β_2	0.69	0.16	0.16	α^{13}	0.62	0.36
	τ	-0.39	0.15	0.15	α^{14}	0.16	0.35
					α^{23}	1.33	0.42
					α^{24}	1.20	0.41
					α^{34}	1.70	0.41

Table 1: Marginal model with constant treatment effect

Model		Est.	mod. SE	rob. SE		Est.	mod. SE
	β_1	-1.60	0.13	0.19			
Ib	β_2	0.69	0.10	0.16			
	τ_1	-0.19	0.18	0.17			
	τ_2	-0.71	0.18	0.18			
	τ_3	-0.56	0.18	0.19			
	τ_4	-0.38	0.18	0.19			
	β_1	-1.59	0.19	0.19	α_{11}	3.45	0.51
IIb	β_2	0.70	0.16	0.17	α_{12}	1.96	0.47
	τ_1	-0.20	0.18	0.17	α_{21}	1.55	0.38
	τ_2	-0.71	0.19	0.18	α_{22}	1.16	0.14
	τ_3	-0.55	0.18	0.19			
	τ_4	-0.37	0.18	0.19			
	β_1	-1.60	0.19	0.19	α_{11}	4.53	0.71
IIIb	β_2	0.70	0.16	0.16	α_{12}	2.46	0.59
	τ_1	-0.20	0.18	0.17	α_{21}	2.18	0.51
	τ_2	-0.71	0.19	0.18	α_{22}	1.42	0.18
	τ_3	-0.55	0.18	0.19			
	τ_4	-0.37	0.18	0.19			
	β_1	-1.60	0.19	0.19	α_{11}^1	5.16	1.27
IVb	β_2	0.70	0.16	0.16	α_{12}^1	2.62	1.02
	τ_1	-0.21	0.18	0.17	α_{21}^1	2.22	0.89
	τ_2	-0.71	0.19	0.18	α_{22}^1	1.51	0.31
	τ_3	-0.56	0.18	0.19	α_{11}^2	3.64	1.36
	τ_4	-0.36	0.18	0.19	α_{12}^2	2.23	1.03
					α_{21}^2	2.10	0.96
					α_{22}^2	1.29	0.42
	β_1	-1.61	0.19	0.19	α^{12}	1.50	0.40
Vb	β_2	0.70	0.16	0.16	α^{13}	0.70	0.38
	τ_1	-0.20	0.18	0.17	α^{14}	0.01	0.36
	τ_2	-0.71	0.19	0.18	α^{23}	1.22	0.44
	τ_3	-0.56	0.18	0.19	α^{24}	1.38	0.44
	τ_4	-0.36	0.18	0.20	α^{34}	1.76	0.42

Table 2: Marginal model with time-varying treatment effect

Finally we calculated likelihood ratio statistics for the hypothesis of a constant treatment effect under the five different assumptions on the association. That means we compare Ia with Ib, IIa with IIb, etc. The statistics are given in Table 3. The hypothesis of a constant treatment effect is rejected in model II, III, IV and V at the 5%-level, but not rejected by model I. The statistics under II-V are comparable with the generalized Wald and score statistics given in

Table 3 (rows two to five) in Miller et al. for the GEE I independence estimator and the GEE I correlation method with a saturated correlation model using 48 correlation parameters.

I	II	III	IV	V	df
4.70	10.75	12.53	12.19	12.09	3

Table 3: Likelihood ratio statistics

6 Discussion

We have shown that the approach of Fitzmaurice and Laird can be extended to the case of a multicategorical response. The theoretical results coincide with those obtained by them for multivariate binary data – as expected – since in both cases the underlying joint distribution is the multinomial. Loosely speaking, things are only more complicate. This is because we have a double multivariate situation: each response is multivariate and we have correlated multivariate responses while in the binary case we have 'only' correlated univariate responses. For a detailed discussion of the advantages and disadvantages of this approach we refer the reader to Fitzmaurice, Laird and Rotnitzky (1993). In the following we want to mention some other points and look for further interesting research points.

We have made explicit the connection between the loglinear expansion of the multinomial distribution and marginal models. The approach is not limited to the longitudinal data situation where we observe repeated outcomes of one response variable. Consider e.g. a crossectional study where a number of different variables is measured at each individual. If it is sensible to treat some of them as explanatory variables (discrete or continuous) and some of them as (nominal and/or ordinal) response variables, our extended approach can be used for a full likelihood marginal regression model after specifying appropriately the marginal mean model. For a repeated ordinal response our approach is competitive to the approach of Molenberghs et al. (1994) but can handle also more general data situations. The application on a sparse data situation illustrated that GEE methods have not always to be the first choice but maximum likelihood can be a feasible alternative. Although we saw little variability of the point estimates for the marginal model in the application in Table 2, simulations should clearify how much bias in the estimates of the marginal parameters is introduced in *finite* samples if we totally misspecify the association model. The results should be compared e.g. with the GEE I correlation method. Another consideration would be if efficiency in *finite* samples can be improved – compared with the GEE I independence estimator – even if the association is misspecified, e.g. because it varies moderate/considerable between individuals and we are only assuming a simple association model common for all individuals. Another interesting point would be to get pre-estimates of the "margin free" *conditional* associations and solving only the score equations for β. The covariance matrix would then be determined by the actual estimates for the mean and the pre-estimates of the conditional association.

A Example and formulas (7) to (12)

In the following we drop the subject index i. D^Ψ and D^Ω are constructed by the dummy coding scheme. In the case $T = 3, R = 3$, we have $3^3 = 27$ cell probabilities, where one is redundant. Then D^Ψ is the matrix, where each row corresponds to a possible outcome of Y_i. For example, the row associated with π_{223}, or $\log(\pi_{223})$, is $(0, 1, 0, 1, 0, 0)$, where we need two dummies for each time point. In an obvious notation we write $D^\Psi = (D_1^{\Psi,Y_1}, D_2^{\Psi,Y_1}, D_1^{\Psi,Y_2}, D_2^{\Psi,Y_2}, D_1^{\Psi,Y_3}, D_2^{\Psi,Y_3})$ where D_j^{Ψ,Y_t}, $j = 1, 2$, $t = 1, 2, 3$ is the column representing the j-th dummy at time point t. D^Ω is the matrix where the columns denote certain interactions between time points. Each row of D^Ω corresponds to a realization w_i of W_i, which depends on the outcome y_i. Let, for example, $D_1^{\Psi,Y_1} D_2^{\Psi,Y_3}$ denote a column vector built by an element by element multiplication of these two columns of D^Ψ. Then we use an obvious ordering and build D^Ω as

$$\begin{aligned}
D^\Omega \quad = \quad & (D_1^{\Psi,Y_1} D_1^{\Psi,Y_2}, D_1^{\Psi,Y_1} D_2^{\Psi,Y_2}, D_2^{\Psi,Y_1} D_1^{\Psi,Y_2}, D_2^{\Psi,Y_1} D_2^{\Psi,Y_2}, D_1^{\Psi,Y_1} D_1^{\Psi,Y_3}, \\
& D_1^{\Psi,Y_1} D_2^{\Psi,Y_3}, D_2^{\Psi,Y_1} D_1^{\Psi,Y_3}, D_2^{\Psi,Y_1} D_2^{\Psi,Y_3}, D_1^{\Psi,Y_2} D_1^{\Psi,Y_3}, D_1^{\Psi,Y_2} D_2^{\Psi,Y_3}, \\
& D_2^{\Psi,Y_2} D_1^{\Psi,Y_3}, D_2^{\Psi,Y_2} D_2^{\Psi,Y_3}, D_1^{\Psi,Y_1} D_1^{\Psi,Y_2} D_1^{\Psi,Y_3}, D_1^{\Psi,Y_1} D_1^{\Psi,Y_2} D_2^{\Psi,Y_3}, \\
& D_1^{\Psi,Y_1} D_2^{\Psi,Y_2} D_1^{\Psi,Y_3}, D_1^{\Psi,Y_1} D_2^{\Psi,Y_2} D_2^{\Psi,Y_3}, D_2^{\Psi,Y_1} D_1^{\Psi,Y_2} D_1^{\Psi,Y_3}, \\
& D_2^{\Psi,Y_1} D_1^{\Psi,Y_2} D_2^{\Psi,Y_3}, D_2^{\Psi,Y_1} D_2^{\Psi,Y_2} D_1^{\Psi,Y_3}, D_2^{\Psi,Y_1} D_2^{\Psi,Y_2} D_2^{\Psi,Y_3}) \\
= \quad & (D_{11}^{\Omega,Y_1Y_2}, \ldots, D_{222}^{\Omega,Y_1Y_2Y_3})
\end{aligned}$$

For example with $m_1, m_2, m_3 = 1, 2, 3$ and $D_{1l}^{\Psi,Y_2}, D_{2l}^{\Psi,Y_2}$, denoting the elements in row l of the column vectors $D_1^{\Psi,Y_2}, D_2^{\Psi,Y_3}$, the rules

$$D_{1l}^{\Psi,Y_2} \quad = \quad \begin{cases} 1 & \text{if for the associated } \pi_{m_1m_2m_3} \text{ holds: } \pi_{m_1m_2m_3} = \pi_{m_1,1,m_3} \\ 0 & \text{else} \end{cases}$$

$$D_{2l}^{\Psi,Y_3} \quad = \quad \begin{cases} 1 & \text{if for the associated } \pi_{m_1m_2m_3} \text{ holds: } \pi_{m_1m_2m_3} = \pi_{m_1,m_2,2} \\ 0 & \text{else} \end{cases}$$

determine columns 3 and 6 of D^Ψ. The tenth column of D^Ω results from the product of these two columns or from the rule

$$D_{12l}^{\Omega,Y_2,Y_3} \quad = \quad \begin{cases} 1 & \text{if for the associated } \pi_{m_1m_2m_3} \text{ holds: } \pi_{m_1m_2m_3} = \pi_{m_1,1,2} \\ 0 & \text{else} \end{cases}$$

π, D^Ψ and D^Ω are therefore given by

row	π	D^{Ψ}						D^{Ω}																			
		Y_1		Y_2		Y_3		Y_1Y_2				Y_1Y_3				Y_2Y_3				$Y_1Y_2Y_3$							
								11	12	21	22	11	12	21	22	11	12	21	22	111	112	121	122	211	212	221	222
1	π_{111}	1	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
2	π_{112}	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0
3	π_{113}	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	π_{121}	1	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
5	π_{122}	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
6	π_{123}	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	π_{131}	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	π_{132}	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	π_{133}	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	π_{211}	0	1	1	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0
11	π_{212}	0	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
12	π_{213}	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	π_{221}	0	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
14	π_{222}	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
15	π_{223}	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	π_{231}	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
17	π_{232}	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
18	π_{233}	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	π_{311}	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
20	π_{312}	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
21	π_{313}	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	π_{321}	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
23	π_{322}	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
24	π_{323}	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	π_{331}	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	π_{332}	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	π_{333}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Interpretation of the parameters

In analogy to the notation of the columns of D^Ψ ans D^Ω we denote the parameters by $\Psi = (\Psi_1^{Y_1}, \Psi_2^{Y_1}, \Psi_1^{Y_2}, \Psi_2^{Y_2}, \Psi_1^{Y_3}, \Psi_2^{Y_3})$ and $\Omega = (\omega_{11}^{Y_1 Y_2}, \ldots, \omega_{222}^{Y_1 Y_2 Y_3})$. With a_i from (6) we get, e.g.:

$$\begin{aligned}\log(\pi_{133}) &= \Psi_1^{Y_1} - a_i \\ \log(\pi_{333}) &= -a_i\end{aligned}$$

Substracting the second equation from the first leeds to

$$\log(\pi_{133}) - \log(\pi_{333}) = \Psi_1^{Y_1}$$

or

$$\begin{aligned}\Psi_1^{Y_1} &= \log\left(\frac{\pi_{133}}{\pi_{333}}\right) \\ &= \log\left(\frac{P(Y_1 = 1, Y_2 = 3, Y_3 = 3)}{P(Y_1 = 3, Y_2 = 3, Y_3 = 3)}\right) \\ &= \log\left(\frac{P(Y_1 = 1 | Y_2 = 3, Y_3 = 3)}{P(Y_1 = 3 | Y_2 = 3, Y_3 = 3)}\right)\end{aligned}$$

The $\Psi_j^{Y_t}$, $j = 1, 2$, $t = 1, 2, 3$ can be interpreted as logits of conditional probabilities:

$$\Psi_j^{Y_t} = \log\left(\frac{P(Y_t = j | Y_{\neq t} = 3)}{P(Y_t = 3 | Y_{\neq t} = 3)}\right)$$

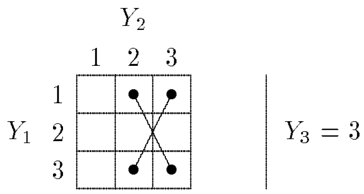
Considering e.g. the system of equations

$$\begin{array}{lclclclcl}\log(\pi_{123}) &=& \Psi_1^{Y_1} &+& \Psi_2^{Y_2} &+& \omega_{12}^{Y_1 Y_2} &-& a \\ \log(\pi_{333}) &=& & & & & &-& a \\ \log(\pi_{323}) &=& & & \Psi_2^{Y_2} &+& \omega_{12}^{Y_1 Y_2} &-& a \\ \log(\pi_{133}) &=& \Psi_1^{Y_1} & & & & &-& a\end{array}$$

it follows, that

$$\begin{aligned}\omega_{12}^{Y_1 Y_2} &= \log(\pi_{123}) + \log(\pi_{333}) - \log(\pi_{323}) - \log(\pi_{133}) \\ &= \log\left(\frac{\pi_{123}\pi_{333}}{\pi_{133}\pi_{323}}\right) \\ &= \log\left(\frac{P(Y_1 = 1, Y_2 = 2 | Y_3 = 3) P(Y_1 = 3, Y_2 = 3 | Y_3 = 3)}{P(Y_1 = 1, Y_2 = 3 | Y_3 = 3) P(Y_1 = 3, Y_2 = 2 | Y_3 = 3)}\right)\end{aligned}$$

$\omega_{12}^{Y_1 Y_2}$ is one of the four conditional log odds ratios of Y_1, Y_2 given $Y_3 = 3$:



We have in general $(t \neq t'; j_1 = 1, 2; j_2 = 1, 2)$

$$\omega_{j_1 j_2}^{Y_t Y_{t'}} = \log \left(\frac{P(Y_t = j_1, Y_{t'} = j_2 | Y_{\neq t, \neq t'} = 3) P(Y_t = 3, Y_{t'} = 3 | Y_{\neq t, \neq t'} = 3)}{P(Y_t = j_1, Y_{t'} = 3 | Y_{\neq t, \neq t'} = 3) P(Y_t = 3, Y_{t'} = j_2 | Y_{\neq t, \neq t'} = 3)} \right)$$

From the system of equations

$$\begin{array}{rcl}
\log(\pi_{222}) & = & \Psi_2^{Y_1} + \Psi_2^{Y_2} + \Psi_2^{Y_3} + \omega_{22}^{Y_1Y_2} + \omega_{22}^{Y_1Y_3} + \omega_{22}^{Y_2Y_3} + \omega_{222}^{Y_1Y_2Y_3} - a \\
\log(\pi_{233}) & = & \Psi_2^{Y_1} - a \\
\log(\pi_{323}) & = & \Psi_2^{Y_2} - a \\
\log(\pi_{332}) & = & \Psi_2^{Y_3} - a \\
\log(\pi_{223}) & = & \Psi_2^{Y_1} + \Psi_2^{Y_2} + \omega_{22}^{Y_1Y_2} - a \\
\log(\pi_{232}) & = & \Psi_2^{Y_1} + \Psi_2^{Y_3} + \omega_{22}^{Y_1Y_3} - a \\
\log(\pi_{322}) & = & \Psi_2^{Y_2} + \Psi_2^{Y_3} + \omega_{22}^{Y_2Y_3} - a \\
\log(\pi_{333}) & = & - a
\end{array}$$

follows further, that

$$\begin{aligned}
\omega_{222}^{Y_1Y_2Y_3} &= \log \left(\frac{\pi_{222}\pi_{233}\pi_{323}\pi_{332}}{\pi_{223}\pi_{232}\pi_{322}\pi_{333}} \right) \\
&= \log \left(\frac{\pi_{222}\pi_{332}}{\pi_{232}\pi_{322}} \right) - \log \left(\frac{\pi_{223}\pi_{333}}{\pi_{233}\pi_{323}} \right) \\
&= \log \left(\frac{P(Y_1 = 2, Y_2 = 2 | Y_3 = 2) P(Y_1 = 3, Y_2 = 3 | Y_3 = 2)}{P(Y_1 = 2, Y_2 = 3 | Y_3 = 2) P(Y_1 = 3, Y_2 = 2 | Y_3 = 2)} \right) - \\
&\quad \log \left(\frac{P(Y_1 = 2, Y_2 = 2 | Y_3 = 3) P(Y_1 = 3, Y_2 = 3 | Y_3 = 3)}{P(Y_1 = 2, Y_2 = 3 | Y_3 = 3) P(Y_1 = 3, Y_2 = 2 | Y_3 = 3)} \right)
\end{aligned}$$

$\omega_{222}^{Y_1Y_2Y_3}$ is the contrast of two conditional log odds ratios. It is clear that conditioning can also be done on Y_1 or Y_2, i.e. it holds also

$$\begin{aligned}
\omega_{222}^{Y_1Y_2Y_3} &= \log \left(\frac{\pi_{222}\pi_{233}}{\pi_{223}\pi_{232}} \right) - \log \left(\frac{\pi_{322}\pi_{333}}{\pi_{323}\pi_{332}} \right) \\
&= \log \left(\frac{P(Y_2 = 2, Y_3 = 2 | Y_1 = 2) P(Y_2 = 3, Y_3 = 3 | Y_1 = 2)}{P(Y_2 = 2, Y_3 = 3 | Y_1 = 2) P(Y_2 = 3, Y_3 = 2 | Y_1 = 2)} \right) - \\
&\quad \log \left(\frac{P(Y_2 = 2, Y_3 = 2 | Y_1 = 3) P(Y_2 = 3, Y_3 = 3 | Y_1 = 3)}{P(Y_2 = 2, Y_3 = 3 | Y_1 = 3) P(Y_2 = 3, Y_3 = 2 | Y_1 = 3)} \right)
\end{aligned}$$

Formulas (7) to (12)

Formula (7): In the following we linearize the multiindex into a single index, e.g. π_{123} is attached as π_6. The underlying distribution of each individual i is multinomial $M(1, \pi_i)$ (but, as already noted, we drop the subject index i in the following). Let $U = (U_1, \ldots, U_{R^T=27})'$ be a vector of random indicator variables, where $\{U_l = 1\}$ denotes the event that the observation i falls in cell l ($l = 1, \ldots, 27$). Then it holds for the indicator functions $I_{\{U_l=1\}}$: $\mathrm{E}\, I_{\{U_l=1\}} = P(U_l = 1) = \mathrm{E}\, U_l = \pi_l \quad l = 1, \ldots, 27$ or, in vector notation, $\mathrm{E}\, U = \pi$. We now look at the indicator functions $I_{\{Y_{t(j)}=1\}}, t = 1, 2, 3, j = 1, 2$. Then, as the events $\{U_l = 1\}$ are disjoint,

$$\begin{aligned}
I_{\{Y_{1(1)}=1\}} &= I_{\{U_1=1\}} + I_{\{U_2=1\}} + I_{\{U_3=1\}} + I_{\{U_4=1\}} + I_{\{U_5=1\}} + I_{\{U_6=1\}} \\
&\quad + I_{\{U_7=1\}} + I_{\{U_8=1\}} + I_{\{U_9=1\}} \\
I_{\{Y_{1(2)}=1\}} &= I_{\{U_{10}=1\}} + I_{\{U_{11}=1\}} + I_{\{U_{12}=1\}} + I_{\{U_{13}=1\}} + I_{\{U_{14}=1\}} + I_{\{U_{15}=1\}} \\
&\quad + I_{\{U_{16}=1\}} + I_{\{U_{17}=1\}} + I_{\{U_{18}=1\}} \\
\vdots &= \vdots \\
I_{\{Y_{3(2)}=1\}} &= I_{\{U_2=1\}} + I_{\{U_5=1\}} + I_{\{U_8=1\}} + I_{\{U_{11}=1\}} + I_{\{U_{14}=1\}} + I_{\{U_{17}=1\}} \\
&\quad + I_{\{U_{20}=1\}} + I_{\{U_{23}=1\}} + I_{\{U_{26}=1\}}
\end{aligned}$$

or, using D^Ψ,

$$I_{\{Y_{t(j)}=1\}} = (D_j^{\Psi,Y_t})'(I_{\{U_1=1\}}, \ldots, I_{\{U_{27}=1\}})'.$$

Thus we have

$$\begin{aligned}
\mathrm{E}(Y_{t(j)}) &= P(Y_{t(j)} = 1) = \mathrm{E}\, I_{\{Y_{t(j)}=1\}} \\
&= (D_j^{\Psi,Y_t})'\, \mathrm{E}(I_{\{U_1=1\}}, \ldots, I_{\{U_{27}=1\}})' \\
&= (D_j^{\Psi,Y_t})'(\mathrm{E}\, I_{\{U_1=1\}}, \ldots, \mathrm{E}\, I_{\{U_{27}=1\}})' \\
&= (D_j^{\Psi,Y_t})'(\pi_1, \ldots, \pi_{27})' \\
&= (D_j^{\Psi,Y_t})'\pi
\end{aligned}$$

or, in matrix notation,

$$\mathrm{E}(Y) = D'^{\Psi}\pi \; .$$

Formula (8) is the analogous version of formula (7) for the expectations of the products in W of the dummies $Y_{t(j)}$. For example $W_{12}^{Y_1Y_2} = Y_{1(1)}Y_{2(2)}$ and $W_{121}^{Y_1Y_2Y_3} = Y_{1(1)}Y_{2(2)}Y_{3(1)}$. For different time points t, t' and for $j_1, j_2, j_3 = 1, 2$ we have

$$I_{\{Y_{t(j_1)}=1; Y_{t'(j_2)}=1\}} = (D_{j_1j_2}^{\Omega,Y_tY_{t'}})'(I_{\{U_1=1\}}, \ldots, I_{\{U_{27}=1\}})'$$

and

$$I_{\{Y_{1(j_1)}=1; Y_{2(j_2)}=1; Y_{3(j_3)}\}} = (D_{j_1j_2j_3}^{\Omega,Y_1Y_2Y_3})'(I_{\{U_1=1\}}, \ldots, I_{\{U_{27}=1\}})'.$$

For example we have $I_{\{Y_{1(1)}=1;Y_{3(2)}=1\}} = I_{\{U_2=1\}} + I_{\{U_5=1\}} + I_{\{U_8=1\}}$ and $I_{\{Y_{1(1)}=1;Y_{2(2)}=1;Y_{3(1)}=1\}} = I_{\{U_4=1\}}$. Taking expectations yields

$$\mathrm{E}\, I_{\{Y_{t(j_1)}=1;Y_{t'(j_2)}=1\}} = \mathrm{E}(Y_{t(j_1)}Y_{t'(j_2)}) = (D^{\Omega,Y_tY_{t'}}_{j_1j_2})'\pi$$

and

$$\mathrm{E}\, I_{\{Y_{1(j_1)}=1;Y_{2(j_2)}=1;Y_{3(j_3)}\}} = \mathrm{E}(Y_{1(j_1)}Y_{2(j_2)}Y_{3(j_3)}) = (D^{\Omega,Y_1Y_2Y_3}_{j_1j_2j_3})'\pi.$$

For example $\mathrm{E}\, I_{\{Y_{1(1)}=1;Y_{2(2)}=1;Y_{3(1)}=1\}} = P(Y_{1(1)} = 1; Y_{2(2)} = 1; Y_{3(1)} = 1) = \mathrm{E}\, I_{\{U_4=1\}} = \pi_{121}$. In matrix notation we can write

$$\mathrm{E}\, W = D'^{\Omega}\pi.$$

Formula (9): A standard result for the multinomial distribution is (see e.g. Agresti, 1990, p.423) $\mathrm{Cov}(U) = \mathrm{diag}(\pi) - \pi\pi'$ where "diag" denotes a diagonal matrix. For $t, t' = 1, 2, 3$ and $j_1, j_2 = 1, 2$ we can write

$$\begin{aligned} I_{\{Y_{t(j_1)}=1;Y_{t'(j_2)}=1\}} &= (D^{\Omega,Y_tY_{t'}}_{j_1j_2})'(I_{\{U_1=1\}},\ldots,I_{\{U_{27}=1\}})' \quad \text{(if } t \neq t'\text{)} \\ &= (D^{\Psi,Y_t}_{j_1} \cdot D^{\Psi,Y_{t'}}_{j_2})'(I_{\{U_1=1\}},\ldots,I_{\{U_{27}=1\}})' \quad \text{(also if } t = t'\text{)} \\ &= (D^{\Psi,Y_t}_{j_1})'\mathrm{diag}(I_{\{U_1=1\}},\ldots,I_{\{U_{27}=1\}})(D^{\Psi,Y_{t'}}_{j_2})\ , \end{aligned}$$

Note that $\mathrm{E}(Y_{t(j_1)}Y_{t'(j_2)}) = 0$ for $t = t', j_1 \neq j_2$ which is also reflected by the fact that the element by element product of $D^{\Psi,Y_t}_{j_1}$ and $D^{\Psi,Y_t}_{j_2}$ is zero for $j_1 \neq j_2$. We get

$$\begin{aligned} \mathrm{E}(Y_{t(j_1)}Y_{t'(j_2)}) &= P(Y_{t(j_1)} = 1; Y_{t'(j_2)} = 1) = \mathrm{E}\, I_{\{Y_{t(j_1)}=1;Y_{t'(j_2)}=1\}} \\ &= \mathrm{E}(D^{\Psi,Y_t}_{j_1})'\,\mathrm{diag}(I_{\{U_1=1\}},\ldots,I_{\{U_{27}=1\}})(D^{\Psi,Y_{t'}}_{j_2}) \\ &= (D^{\Psi,Y_t}_{j_1})'\,\mathrm{diag}(\pi)(D^{\Psi,Y_{t'}}_{j_2}), \end{aligned}$$

and the result

$$\begin{aligned} \mathrm{Cov}(Y_{t(j_1)}Y_{t'(j_2)}) &= \mathrm{E}(Y_{t(j_1)}Y_{t'(j_2)}) - \mathrm{E}(Y_{t(j_1)})\,\mathrm{E}(Y_{t'(j_2)}) \\ &= (D^{\Psi,Y_t}_{j_1})'\,\mathrm{diag}(\pi)(D^{\Psi,Y_{t'}}_{j_2}) - ((D^{\Psi,Y_t}_{j_1})'\pi)((D^{\Psi,Y_{t'}}_{j_2})'\pi) \\ &= (D^{\Psi,Y_t}_{j_1})'\,\mathrm{diag}(\pi)(D^{\Psi,Y_{t'}}_{j_2}) - (D^{\Psi,Y_t}_{j_1})'\pi\pi'(D^{\Psi,Y_{t'}}_{j_2}) \\ &= (D^{\Psi,Y_t}_{j_1})'(\mathrm{diag}(\pi) - \pi\pi')(D^{\Psi,Y_{t'}}_{j_2}) \\ &= (D^{\Psi,Y_t}_{j_1})'\,\mathrm{Cov}(U)(D^{\Psi,Y_{t'}}_{j_2}) \end{aligned}$$

or, in matrix notation,

$$\mathrm{Cov}(Y) = D'^{\Psi}(\mathrm{diag}(\pi) - \pi\pi')D^{\Psi}.$$

Formulas (10), (11) and (12) can be derived in a similar manner. As an example for applying formula (12) let $W_{12}^{Y_1,Y_2} = Y_{1(1)}Y_{2(2)}$ and $W_{22}^{Y_2,Y_3} = Y_{2(2)}Y_{3(2)}$. Then

$$\begin{aligned}
\mathrm{Cov}(W_{12}^{Y_1,Y_2}, W_{22}^{Y_2,Y_3}) &= \mathrm{E}(Y_{1(1)}Y_{2(2)}^2 Y_{3(2)}) - \mathrm{E}(Y_{1(1)}Y_{2(2)})\,\mathrm{E}(Y_{2(2)}Y_{3(2)}) \\
&= \pi_{122} - \pi_{12+}\pi_{+22} \\
&= (D_{12}^{\Omega,Y_1Y_2} \cdot D_{22}^{\Omega,Y_2Y_3})'\pi - (D_{12}^{\Omega,Y_1Y_2})'\pi(D_{22}^{\Omega,Y_2Y_3})'\pi \\
&= (D_{12}^{\Omega,Y_1Y_2})'(\mathrm{diag}(\pi) - \pi\pi')(D_{22}^{\Omega,Y_2Y_3})
\end{aligned}$$

B Derivation of the score equations

We give a detailed description of the way to get the score equations although we could refer the reader to the Appendix of Fitzmaurice and Laird (1993). Alternatively, results in Zhao, Prentice and Self (1992) could be used by noting that our model is a special case of their *partly exponential model*. To see this choose $D^{\Omega}\Omega_i$ for the shape function $c(\cdot)$ in the partly exponential model.

We now calculate the derivation with respect to (β, α) of the contribution l_i of an individual i to the loglikelihood. let l denote the cell in which the response Y_i of individual i falls with probability π_{il}, $l = 1, \ldots, R^T$. Assume that we observed a response $Y_i = y_i$ and the observed index of the cell to which the observation belongs is l. Since the underlying distribution is multinomial, the contribution of individual i to the loglikelihood is $l_i(y_i) = \log(\pi_{il})$. Consider the parameter transformation from (Ψ_i, Ω_i) to (μ_i, Γ_i) with $\mu_i = \mu_i(\Psi_i, \Omega_i)$ (see Section 3) and $\Gamma_i = \Gamma_i(\Omega_i) \equiv \Omega_i$ (the identity function).

Treating the loglikelihood as a function of (μ_i, Ω_i), $l_i((\mu_i(\Psi_i, \Omega_i), \Gamma_i(\Omega_i))$, we get by use of the chain rule

$$\frac{\partial l}{\partial \Psi_{s_\Psi}} = \sum_{s_\mu} \frac{\partial l}{\partial \mu_{s_\mu}} \frac{\partial \mu_{s_\mu}}{\partial \Psi_{s_\Psi}} + \sum_{s_\Gamma} \frac{\partial l}{\partial \Gamma_{s_\Gamma}} \frac{\partial \Gamma_{s_\Gamma}}{\partial \Psi_{s_\Psi}} \tag{19}$$

$$\frac{\partial l}{\partial \Omega_{s_\Omega}} = \sum_{s_\mu} \frac{\partial l}{\partial \mu_{s_\mu}} \frac{\partial \mu_{s_\mu}}{\partial \Omega_{s_\Omega}} + \sum_{s_\Gamma} \frac{\partial l}{\partial \Gamma_{s_\Gamma}} \frac{\partial \Gamma_{s_\Gamma}}{\partial \Omega_{s_\Omega}} \tag{20}$$

where we have dropped the subject index i and $s_\Psi, s_\Omega, s_\mu, s_\Gamma$ represent vector indices running from 1 to the appropriate length of the parameter vectors. The partial derivatives $\frac{\partial l}{\partial \Psi_{s_\Psi}}$ and $\frac{\partial l}{\partial \Omega_{s_\Omega}}$ (lefthand side of (19) and (20)) can be computed directly from the loglinear representation of the joint distribution (as shown in the next section). The partial derivatives $\frac{\partial \mu_{s_\mu}}{\partial \Psi_{s_\Psi}}$, $\frac{\partial \Gamma_{s_\Gamma}}{\partial \Psi_{s_\Psi}}$, $\frac{\partial \mu_{s_\mu}}{\partial \Omega_{s_\Omega}}$, $\frac{\partial \Gamma_{s_\Gamma}}{\partial \Omega_{s_\Omega}}$ are elements from the Jacobian (matrix) of the applied parameter transformation. Let $\frac{\partial \mu_{s_\mu}}{\partial \Psi_{s_\Psi}}$ be the element on row position s_μ and column position s_Ψ of the matrix $\frac{\partial \mu}{\partial \Psi}$. Use the analogous notation for the other three derivatives. Now turn back to matrix notation. By use of the formula $\sum_a vector_a matrix_{ab} = (matrix' vector)_b$, (19) and (20) can be written as

$$\left(\frac{\partial l_i}{\partial \Psi_i}\right)_{s_\Psi} = \left(\left(\frac{\partial \mu_i}{\partial \Psi_i}\right)' \left(\frac{\partial l_i}{\partial \mu_i}\right)\right)_{s_\Psi} + \left(\left(\frac{\partial \Gamma_i}{\partial \Psi_i}\right)' \left(\frac{\partial l_i}{\partial \Gamma_i}\right)\right)_{s_\Psi} \tag{21}$$

$$\left(\frac{\partial l_i}{\partial \Omega_i}\right)_{s_\Omega} = \left(\left(\frac{\partial \mu_i}{\partial \Omega_i}\right)' \left(\frac{\partial l_i}{\partial \mu_i}\right)\right)_{s_\Omega} + \left(\left(\frac{\partial \Gamma_i}{\partial \Omega_i}\right)' \left(\frac{\partial l_i}{\partial \Gamma_i}\right)\right)_{s_\Omega} \tag{22}$$

or in matrix notation

$$\begin{pmatrix} \frac{\partial l_i}{\partial \Psi_i} \\ \frac{\partial l_i}{\partial \Omega_i} \end{pmatrix} = \begin{pmatrix} \left(\frac{\partial \mu_i}{\partial \Psi_i}\right)' & \left(\frac{\partial \Gamma_i}{\partial \Psi_i}\right)' \\ \left(\frac{\partial \mu_i}{\partial \Omega_i}\right)' & \left(\frac{\partial \Gamma_i}{\partial \Omega_i}\right)' \end{pmatrix} \begin{pmatrix} \frac{\partial l_i}{\partial \mu_i} \\ \frac{\partial l_i}{\partial \Gamma_i} \end{pmatrix} \tag{23}$$

As shown below, $\left(\frac{\partial \mu_i}{\partial \Psi_i}\right) = \operatorname{Cov}(Y_i)$, a symmetric matrix, $\left(\frac{\partial \Gamma_i}{\partial \Psi_i}\right) = 0$, $\left(\frac{\partial \mu_i}{\partial \Omega_i}\right) = \operatorname{Cov}(Y_i, W_i)$, the transposed of which is $\operatorname{Cov}(W_i, Y_i)$, $\left(\frac{\partial \Gamma_i}{\partial \Omega_i}\right) = I$, $\frac{\partial l_i}{\partial \Psi_i} = y_i - \mu_i$ and $\frac{\partial l_i}{\partial \Omega_i} = w_i - \nu_i$. Substituting these expressions in (23) yields

$$\begin{pmatrix} y_i - \mu_i \\ w_i - \nu_i \end{pmatrix} = \begin{pmatrix} \operatorname{Cov}(Y_i) & 0 \\ \operatorname{Cov}(W_i, Y_i) & I \end{pmatrix} \begin{pmatrix} \frac{\partial l_i}{\partial \mu_i} \\ \frac{\partial l_i}{\partial \Gamma_i} \end{pmatrix} \tag{24}$$

or, by multiplication with the inverse from the left

$$\begin{pmatrix} \frac{\partial l_i}{\partial \mu_i} \\ \frac{\partial l_i}{\partial \Gamma_i} \end{pmatrix} = \begin{pmatrix} \operatorname{Cov}(Y_i) & 0 \\ \operatorname{Cov}(W_i, Y_i) & I \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ w_i - \nu_i \end{pmatrix} \tag{25}$$

Using a formula for inverting partioned matrices (e.g. Rao and Toutenburg (1995), p.291) we get

$$\begin{pmatrix} \frac{\partial l_i}{\partial \mu_i} \\ \frac{\partial l_i}{\partial \Gamma_i} \end{pmatrix} = \begin{pmatrix} \operatorname{Cov}(Y_i)^{-1} & 0 \\ -\operatorname{Cov}(W_i, Y_i) \operatorname{Cov}(Y_i)^{-1} & I \end{pmatrix} \begin{pmatrix} y_i - \mu_i \\ w_i - \nu_i \end{pmatrix} \tag{26}$$

Note that we have assumed that the inverse $\operatorname{Cov}(Y_i)^{-1}$ exists. Otherwise the transformation can not be applied. A sufficient but not necessary condition for $\operatorname{Cov}(Y_i)$ being positive definite is that all cell probabilities are greater than zero.

Applying the chain rule once more we get

$$\begin{pmatrix} \frac{\partial l_i}{\partial \beta} \\ \\ \frac{\partial l_i}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} \left(\frac{\partial \mu}{\partial \beta}\right)' & 0 \\ \\ 0 & \left(\frac{\partial \Omega}{\partial \alpha}\right)' \end{pmatrix} \begin{pmatrix} \frac{\partial l_i}{\partial \mu_i} \\ \\ \frac{\partial l_i}{\partial \Gamma_i} \end{pmatrix}$$

$$= \begin{pmatrix} \left(\frac{\partial \mu}{\partial \beta}\right)' & 0 \\ \\ 0 & \left(\frac{\partial \Omega}{\partial \alpha}\right)' \end{pmatrix} \begin{pmatrix} \mathrm{Cov}(Y_i)^{-1} & 0 \\ \\ -\mathrm{Cov}(W_i, Y_i)\,\mathrm{Cov}(Y_i)^{-1} & I \end{pmatrix} \begin{pmatrix} y_i - \mu_i \\ \\ w_i - \nu_i \end{pmatrix} \quad (27)$$

Multiplying the righthand side yields the two separate equations for β and α, (13) and (14), in section 4.

Derivation of the loglikelihood with respect to (Ψ_i, Ω_i)

We now derive the formulas

$$\frac{\partial l_i}{\partial \Psi_i} = y_i - \mathrm{E}(Y_i) = y_i - \mu_i \quad (28)$$

$$\frac{\partial l_i}{\partial \Omega_i} = w_i - \mathrm{E}(W_i) = w_i - \nu_i \quad (29)$$

Proof of (28) and (29) Let D_l^{Ψ} denote the l-th row of D^{Ψ} and $D_{ls_\Psi}^{\Psi}$ a specific element in D^{Ψ} in row l and column s_Ψ. Assume that the observation of individual i, y_i, falls in cell l. With (5) and (6), the contribution to the loglikelihood is

$$l_i = \log(\pi_{il}) = D_l^{\Psi}\Psi_i + D_l^{\Omega}\Omega_i - a_i$$

Following results in Agresti (1990, p. 437) we have for arbitrary elements Ψ_{is_Ψ} and Ω_{is_Ω} of Ψ_i and Ω_i

$$\frac{\partial \pi_{il}}{\partial \Psi_{is_\Psi}} = \pi_{il} D_{ls_\Psi}^{\Psi} - \pi_{il} \sum_{k=1}^{R^T} D_{ks_\Psi}^{\Psi} \pi_{ik}$$

$$\frac{\partial \pi_{il}}{\partial \Omega_{is_\Omega}} = \pi_{il} D_{ls_\Omega}^{\Omega} - \pi_{il} \sum_{k=1}^{R^T} D_{ks_\Omega}^{\Omega} \pi_{ik}$$

Now, $\partial l_i / \partial \Psi_{is_\Psi} = (1/\pi_{il})(\partial \pi_{il} / \partial \Psi_{is_\Psi})$ and $\partial l_i / \partial \Omega_{is_\Omega} = (1/\pi_{il})(\partial \pi_{il} / \partial \Omega_{is_\Omega})$. It follows that

$$\frac{\partial l_i}{\partial \Psi_{is_\Psi}} = D_{ls_\Psi}^{\Psi} - \sum_{k=1}^{R^T} D_{ks_\Psi}^{\Psi} \pi_{ik}$$

$$\frac{\partial l_i}{\partial \Omega_{is_\Omega}} = D_{ls_\Omega}^{\Omega} - \sum_{k=1}^{R^T} D_{ks_\Omega}^{\Omega} \pi_{ik}$$

Noting that by construction of D^{Ψ} and D^{Ω}, $y_i' \equiv D_l^{\Psi}$ and $w_i' \equiv D_l^{\Omega}$ this is equivalent to

$$\begin{aligned}
\frac{\partial l_i}{\partial \Psi_{is_\Psi}} &= y_{is_\Psi} - \sum_{k=1}^{R^T} D^{\Psi}_{ks_\Psi} \pi_{ik} = y_{is_\Psi} - \mu_{is_\Psi} \\
\frac{\partial l_i}{\partial \Omega_{is_\Omega}} &= w_{is_\Omega} - \sum_{k=1}^{R^T} D^{\Omega}_{ks_\Omega} \pi_{ik} = w_{is_\Omega} - \nu_{is_\Omega}
\end{aligned}$$

where we have identified the two sums on the right side as the elementwise versions of (7) and (8). Applying matrix notation yields the postulated result.

Jacobian matrix of the transformation

Using the notation as in the subsection before and with s_μ and s_Ψ running both from 1 to Tr, it holds

$$\begin{aligned}
\frac{\partial \mu_{i,s_\mu}}{\partial \Psi_{i,s_\Psi}} &= \sum_{k=1}^{R^T} D^{\Psi}_{ks_\mu} \left(\frac{\partial \pi_{ik}}{\partial \Psi_{i,s_\Psi}} \right) = \sum_{k=1}^{R^T} D^{\Psi}_{ks_\mu} \left(\pi_{ik} D^{\Psi}_{ks_\Psi} - \pi_{ik} \sum_{m=1}^{R^T} D^{\Psi}_{ms_\Psi} \pi_{im} \right) \\
&= \sum_{k=1}^{R^T} D^{\Psi}_{ks_\mu} D^{\Psi}_{ks_\Psi} \pi_{ik} - \sum_{k=1}^{R^T} D^{\Psi}_{ks_\mu} \pi_{ik} \sum_{m=1}^{R^T} D^{\Psi}_{ms_\Psi} \pi_{im}
\end{aligned}$$

Converting to matrix notation and looking at the proof of formula (9) in Appendix A yields the result that $\frac{\partial \mu_{i,s_\mu}}{\partial \Psi_{i,s_\Psi}}$ is the element in row s_μ and column s_Ψ of the matrix $D'^{\Psi} \left[\text{diag}(\pi_i) - \pi\pi' \right] D^{\Psi} = \text{Cov}(Y_i)$. The result $\frac{\partial \mu_i}{\partial \Omega_i} = \text{Cov}(Y_i, W_i)$ can be obtained in an analogous manner. Finally, $\frac{\partial \Gamma_i}{\partial \Psi_i} = 0$ and $\frac{\partial \Gamma_i}{\partial \Omega_i} = I$ follow from the fact that $\Gamma_i(\Omega_i) = \Omega_i$ does by construction not depend on Ψ_i but is the identity function.

C Expected Fisher Information Matrix

It is

$$\text{E}\left[\begin{pmatrix} \frac{\partial l_i}{\partial \beta} \\ \frac{\partial l_i}{\partial \alpha} \end{pmatrix} \begin{pmatrix} \frac{\partial l_i}{\partial \beta} \\ \frac{\partial l_i}{\partial \alpha} \end{pmatrix}' \right] = \text{E} \begin{pmatrix} \left(\frac{\partial l_i}{\partial \beta}\right) \left(\frac{\partial l_i}{\partial \beta}\right)' & \left(\frac{\partial l_i}{\partial \beta}\right) \left(\frac{\partial l_i}{\partial \alpha}\right)' \\ \left(\frac{\partial l_i}{\partial \alpha}\right) \left(\frac{\partial l_i}{\partial \beta}\right)' & \left(\frac{\partial l_i}{\partial \alpha}\right) \left(\frac{\partial l_i}{\partial \alpha}\right)' \end{pmatrix} \quad (30)$$

with

$$\text{E}\left[\left(\frac{\partial l_i}{\partial \beta}\right) \left(\frac{\partial l_i}{\partial \beta}\right)' \right] =$$

$$
\begin{aligned}
&= \mathrm{E}\left[\left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathrm{Cov}(Y_i)^{-1}\,(Y_i-\mu_i)\,(Y_i-\mu_i)'\,\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right)\right] \\
&= \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathrm{Cov}(Y_i)^{-1}\,\mathrm{E}\left[(Y_i-\mu_i)\,(Y_i-\mu_i)'\right]\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right) \\
&= \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i)\,\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right) \\
&= \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{\partial l_i}{\partial \alpha}\right)\left(\frac{\partial l_i}{\partial \alpha}\right)'\right] \\
&= \mathrm{E}\Bigg[\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\left[(W_i-\nu_i)-\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,(Y_i-\mu_i)\right]\times \\
&\qquad \left[(W_i-\nu_i)-\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,(Y_i-\mu_i)\right]'\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)\Bigg] \\
&= \left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\Big\{ \\
&\qquad \mathrm{E}\left[(W_i-\nu_i)\,(W_i-\nu_i)'\right] \\
&\qquad -\mathrm{E}\left[(W_i-\nu_i)\,(Y_i-\mu_i)'\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i)\right] \\
&\qquad -\mathrm{E}\left[\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,(Y_i-\mu_i)\,(W_i-\nu_i)'\right] \\
&\qquad +\mathrm{E}\left[\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,(Y_i-\mu_i)\,(Y_i-\mu_i)'\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i)\right] \\
&\qquad \Big\}\left(\frac{\partial \Gamma_i}{\partial \alpha}\right) \\
&= \left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\Big\{ \\
&\qquad \mathrm{Cov}(W_i) \\
&\qquad -\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i) \\
&\qquad -\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i) \\
&\qquad +\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i) \\
&\qquad \Big\}\left(\frac{\partial \Gamma_i}{\partial \alpha}\right) \\
&= \left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\left\{\mathrm{Cov}(W_i)-\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i)\right\}\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{E}\left[\left(\frac{\partial l_i}{\partial \alpha}\right)\left(\frac{\partial l_i}{\partial \beta}\right)'\right] \\
= \; & \mathrm{E}\left[\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\left[(W_i-\nu_i)-\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,(Y_i-\mu_i)\right]\times\right. \\
& \left.\left[(Y_i-\mu_i)'\,\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right)\right]\right] \\
= \; & \left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right) - \\
& -\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right) \\
= \; & 0
\end{aligned}
$$

It follows that β und α are orthogonal parameters in the sense of the definition of Cox and Reid (1987, 1989) and the expected Fisher information matrix is

$$
\left(\begin{array}{c|c}
\sum_{i=1}^{N}\left(\frac{\partial \mu_i}{\partial \beta}\right)'\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right) & 0 \\
0 & \sum_{i=1}^{N}\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)' B\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)
\end{array}\right)
$$

with $B=\left\{\mathrm{Cov}(W_i)-\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i)\right\}$. Further, the asymptotic covariance matrix of $(\widehat{\beta},\widehat{\alpha})$, $\mathrm{Cov}(\widehat{\beta},\widehat{\alpha})$, is block diagonal with upper left block

$$
\mathrm{Cov}(\widehat{\beta})=\left(\sum_{i=1}^{N}\left(\frac{\partial \mu_i}{\partial \beta}\right)'\mathrm{Cov}(Y_i)^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right)\right)^{-1} \tag{31}
$$

as asymptotic covariance matrix of $\widehat{\beta}$ and lower right block

$$
\mathrm{Cov}(\widehat{\alpha}) = \left(\sum_{i=1}^{N}\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)'\left\{\mathrm{Cov}(W_i)-\mathrm{Cov}(W_i,Y_i)\,\mathrm{Cov}(Y_i)^{-1}\,\mathrm{Cov}(Y_i,W_i)\right\}\left(\frac{\partial \Gamma_i}{\partial \alpha}\right)\right)^{-1} \tag{32}
$$

as asymptotic covariance matrix of $\widehat{\alpha}$.

References

[1] Agresti, A. *Categorical data analysis.* Wiley, New York, 1990.

[2] Balagtas, C. C., Becker, M. P., and Lang, J. B. Marginal modelling of categorical data from crossover experiments. *Applied Statistics 44* (1995), 63–77.

[3] Cox, D. R., and Reid, N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Ser. C 49* (1987), 1–39.

[4] Cox, D. R., and Reid, N. On the stability of maximum–likelihood estimators of orthogonal parameters. *Canadian Journal of Statistics 17* (1989), 229–233.

[5] Dale, J. R. Global cross–ratio models for bivariate, discrete, ordered response. *Biometrics 42* (1986), 909–917.

[6] Fahrmeir, L., and Pritscher, L. Regression analysis of forest damage by marginal models for correlated ordinal data responses. *SFB 386 – Discussion Paper 9* (1995).

[7] Fahrmeir, L., and Tutz, G. *Multivariate statistical modelling based on generalized linear models.* Springer, Berlin, 1994.

[8] Fitzmaurice, G. M., and Laird, N. M. A likelihood-based method for analysing longitudinal binary responses. *Biometrika 80* (1993), 141–151.

[9] Fitzmaurice, G. M., Laird, N. M., and Lipsitz, S. R. Analysing incomplete binary responses: a likelihood–based approach. *Biometrics 50* (1994), 601–612.

[10] Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. Regression models for discrete longitudinal responses. *Statitical Science 8* (1993), 284–309.

[11] Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M., and Uryniak, T. J. Categorical data analysis. In *Statistical Methodology in the Pharmaceutical Sciences*, D. A. Berry, Ed. New York, Marcel Dekker, 1989, pp. 391–475.

[12] Lang, J. B., and Agresti, A. Simultaneously modeling joint and marginal distibutions of multivariate categorical responses. *Journal of the American Statistical Association 89* (1994), 625–632.

[13] Liang, K.-Y., and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika 73* (1986), 13–22.

[14] Liang, K.-Y., Zeger, S. L., and Quaqish, B. Multivariate regression analysis for categorical data. *J. Royal. Statist. Soc. B 54* (1992), 3–40.

[15] Lipsitz, S. R., Laird, N. M., and Harrington, D. P. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika 78* (1991), 153–160.

[16] Miller, M. E., Davis, S. C., and Landis, J. R. The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics 49* (1993), 1033–1044.

[17] Molenberghs, G., and Lesaffre, E. Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association 89* (1994), 633–644.

[18] Prentice, R. L. Correlated binary regression with covariates specific to each binary observation. *Biomerics 44* (1988), 1033–1048.

[19] Rao, C. R., and Toutenburg, H. *Linear Models – Least Squares and Alternatives.* Springer, New York, 1995.

[20] White, H. Maximum likelihood estimation of misspecified models. *Econometrica 50* (1982), 1–25.

[21] Zhao, L. P., Prentice, R. L., and Self, S. G. Multivariate mean parameter estimation by using a partly exponential model. *JRSS B 54* (1992), 805–811.