MEANINGFUL REGRESSION AND ASSOCIATION MODELS FOR CLUSTERED ORDINAL DATA

JUKKA JOKINEN, JOHN W. MCDONALD, PETER W. F. SMITH

ABSTRACT

Many proposed methods for analyzing clustered ordinal data focus on the regression model and consider the association structure within a cluster as a nuisance. However, often the association structure is of equal interest, for example, temporal association in longitudinal studies and association between responses to similar questions in a survey. We discuss the use, appropriateness and interpretability of various latent variable and Markov models for the association structure and propose a new structure that exploits the ordinality of the response. The models are illustrated with a study concerning opinions regarding government spending and an analysis of stability and change in teenage marijuana use over time, where we reveal different behavioral patterns for boys and girls through a comprehensive investigation of individual response profiles.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/08

MEANINGFUL REGRESSON AND ASSOCIATION MODELS FOR CLUSTERED ORDINAL DATA

Jukka Jokinen
John W. McDonald
Peter W.F. Smith

Many proposed methods for analyzing clustered ordinal data focus on the regression model and consider the association structure within a cluster as a nuisance. However, often the association structure is of equal interest, for example, temporal association in longitudinal studies and association between responses to similar questions in a survey. We discuss the use, appropriateness and interpretability of various latent variable and Markov models for the association structure and propose a new structure that exploits the ordinality of the response. The models are illustrated with a study concerning opinions regarding government spending and an analysis of stability and change in teenage marijuana use over time, where we reveal different behavioral patterns for boys and girls through a comprehensive investigation of individual response profiles.

1. INTRODUCTION

In sociological applications, the response variables of interest are often measured on an ordinal scale, such as opinions or attitudes toward sociological issues (e.g. strongly disagree, ... , strongly agree). The research hypothesis typically addresses the question whether these responses differ in various subgroups of the population of interest. There are well-established methods for the regression analysis of ordinal responses, see e.g. Agresti (2002). However, further complication to the modeling process arises when the ordered responses are clustered in some way, such as responses to similar questions on the same individual, or repeated responses to the same question in longitudinal studies. In order to obtain correct inferences, the association between responses within a cluster has to be taken into account in the analysis. Compared to simple cross-sectional studies with independent observations, further insights can also be gained by investigating the structure of this association.

Agresti and Natarajan (2001) surveyed various strategies for analyzing these type of clustered ordinal categorical data, focusing on marginal models and cluster-specific models. Our emphasis here, in the context of re-analyses of clustered ordinal datasets, is to present a method that combines a marginal regression model with a meaningful model for the association structure, and to relate it to methods surveyed in Agresti and Natarajan (2001).

The data presented in Table 1, previously analyzed by Lang, McDonald, and Smith (1999) and Vermunt and Hagenaars (2004), comes from the US National Youth Survey, where 237 teenagers (117 boys and 120 girls), aged 13 at the beginning of the study, filled in a questionnaire yearly for five consecutive years. At the end of each year they were asked about their marijuana use during that year. The response is ordinal, with values 1=never (non-user), 2=less than once a month (occasional user) and 3=more than once a month (frequent user). Obvious substantive research questions involve comparisons of the prevalences of marijuana use by age and sex of the respondent. However, investigation of the entire response profiles of teenagers may reveal other forms of dissimilarities rather than just prevalence differences.

Table 1: Marijuana Use Data: Observed Response Profiles and Counts.

Boys						Girls					
Age					Count	Age					Count
13	14	15	16	17		13	14	15	16	17	
1	1	1	1	1	48	1	1	1	1	1	63
1	1	1	1	2	8	1	1	1	1	2	10
1	1	1	1	3	4	1	1	1	1	3	3
1	1	1	2	1	2	1	1	1	2	1	4
1	1	1	2	2	4	1	1	1	2	2	2
1	1	1	2	3	1	1	1	1	3	1	1
1	1	1	3	1	1	1	1	1	3	2	1
1	1	1	3	3	1	1	1	1	3	3	3
1	1	2	1	3	1	1	1	2	1	1	2
1	1	2	2	1	2	1	1	2	1	2	2
1	1	2	2	2	2	1	1	2	2	1	3
1	1	2	2	3	1	1	1	2	2	2	7
1	1	2	3	3	5	1	1	2	2	3	1
1	1	3	1	2	1	1	1	2	3	3	1
1	1	3	2	2	1	1	2	1	1	1	1
1	1	3	3	3	3	1	2	1	1	2	2
1	2	1	1	2	1	1	2	2	1	2	1
1	2	1	2	1	1	1	2	2	2	1	1
1	2	2	1	1	2	1	2	2	3	3	2
1	2	2	2	1	1	1	2	3	1	2	1
1	2	2	1	3	1	1	2	3	3	2	1
1	2	2	3	3	1	1	2	3	3	3	1
1	2	3	2	2	1	1	3	3	2	2	1
1	2	3	2	3	1	2	1	1	3	3	1
1	2	3	3	2	1	2	1	2	2	2	1
1	2	3	3	3	4	2	1	3	3	3	1
1	3	1	3	3	1	2	3	3	3	3	1
1	3	2	2	2	1	2	3	3	3	2	1
1	3	3	3	3	2	3	3	3	2	3	1
1	3	3	2	2	1						
2	1	1	1	1	3						
3	1	1	1	1	1						
2	2	2	2	2	1						
2	2	3	3	3	1						
2	3	2	1	1	1						
2	3	2	3	3	1						
2	3	3	3	3	2						
3	2	3	3	3	1						
3	3	3	3	1	1						
3	3	3	3	3	1						

Source: National Youth Survey, Elliott et al. (1983). Marijuana use response: 1=Never, 2=Less than once a month, 3=More than once a month in the past year.

Lang and Agresti (1994) analyzed a dataset from the 1989 US General Social Survey, where 607 adults, aged over 18 years, were asked their opinion concerning government spending on (a) The Environment, (b) Health, (c) Assistance to Big Cities, and (d) Law Enforcement. Each response is ordinal with levels 1=too little, 2=about right, 3=too much. Lang and Agresti (1994) did not consider any covariates. In order to illustrate the potential of our method, we include the covariates age, sex, race, and political party affiliation. In addition, we include three more questions concerning government spending with a common response: (e) National Defence, (f) Education, and (g) Assistance to the Poor. An interesting regression task is to compare how the marginal distributions of these 7 targets of government spending differ by political party affiliation, after adjusting for age, sex, and race of the respondent. It is also equally interesting to study whether individuals have different tendencies when answering questions concerning government spending in general. In order to achieve both goals, the joint probability of a response profile needs to be parametrized in terms of univariate probabilities and suitable dependence measures that facilitate comprehensive modeling of the association structure.

For a multivariate binary response, Ekholm, Smith, and McDonald (1995) parametrized the joint probability with univariate probabilities and dependence ratios. The dependence ratio is defined as the joint success probability divided by the joint success probability assuming independence. For example, the second-order dependence ratio is

$$\tau_{12} = \frac{\mathrm{pr}(Y_1 = 1, Y_2 = 1)}{\mathrm{pr}(Y_1 = 1)\,\mathrm{pr}(Y_2 = 1)}. \tag{1}$$

The dependence ratio was extended to the ordinal case by Ekholm et al. (2003), who also demonstrated how meaningful association structures can easily be specified using dependence ratios. We present the dependence ratio parametrization for a multivariate ordinal response in Section 2. In Section 3, we analyze the dataset from the 1989 US General Social Survey, regarding opinions about government spending, and discuss the appropriateness of various exchangeable association structures presented in Ekholm et al. (2003). In Section 4, a model for temporal association is fitted to the dataset concerning teenage marijuana use, and exchangeable association structures are extended further by exploiting the ordinality of the response. In Section 5 we consider maximum likelihood estimation and computational aspects. In Section 6 our approach is related to methods surveyed in Agresti and Natarajan (2001).

2. MEAN PARAMETRIZATION AND DEPENDENCE RATIOS

To formally present the connection between the joint probability of a response profile and our parameters of interest, some notation is necessary. For a simple illustration of the parametrization, see Section 2.1, and for a comprehensive representation of the relationship, see Ekholm et al. (2003). Consider a response profile of length q in cluster i, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iq})$, for $i = 1, \ldots, n$, where the realizations of Y_{ik}, denoted a_k, are ordered, $a_k = 1, \ldots, f$ and $k = 1, \ldots, q$. For the government spending data, $n = 607$, $q = 7$ and $f = 3$, and for the marijuana data, $n = 237$, $q = 5$ and $f = 3$. There are f^q possible realizations of the profile with $\mathrm{pr}(\boldsymbol{Y}_i = \{a_1, \ldots, a_q\}) = \pi_i(a_1, \ldots, a_q)$, and when there is a time ordering, we refer to the response profiles as paths. For example, out of $3^5 = 243$ possible realizations, the observed path of a girl in the bottom right corner of Table 1 is $\{3, 3, 2, 3, 3\}$, with probability $\pi_i(3, 3, 3, 2, 3)$. To specify the probability distribution of $\boldsymbol{Y}_i$, denote the $1 \times f^q$ vector of profile or path probabilities by $\boldsymbol{\pi}_i$, with $\boldsymbol{\pi}_i \mathbf{1}^T = 1$. Furthermore, define dummy variables $Y_{ik}^{(a_k)} = 1$

if $Y_{ik} = a_k$, else 0, for $a_k = 2, \ldots, f$, and $Y_{ik}^{(1)} = 1 - Y_{ik}^{(2)} - \cdots - Y_{ik}^{(f)}$. The $1 \times (f^q - 1)$ vector of mean parameters is given by

$$\boldsymbol{\mu}_i = (\mu_{i1}^{(2)}, \ldots, \mu_{iq}^{(f)}, \mu_{i12}^{(2,2)}, \ldots, \mu_{i1\ldots q}^{(f,\ldots,f)}), \tag{2}$$

where $\mu_{ik}^{(a_k)} = E(Y_{ik}^{(a_k)})$ and $\mu_{i1\ldots k}^{(a_1,\ldots,a_k)} = E(Y_{i1}^{(a_1)} \cdots Y_{ik}^{(a_k)})$. Ekholm et al. (2003) showed that there exists a one-to-one correspondence $\boldsymbol{\mu}_i \to \boldsymbol{\pi}_i$ for specifying the joint distribution in terms of these mean parameters.

Second- and higher-order means capture the information about the association between the responses within a cluster. For a more interpretable measure of association, we replace these by the corresponding dependence ratios. For example, the second-order dependence ratio is a generalization of (1):

$$\tau_{kl}^{(a_k,a_l)} = \frac{\mathrm{pr}(Y_{ik} = a_k, Y_{il} = a_l)}{\mathrm{pr}(Y_{ik} = a_k)\,\mathrm{pr}(Y_{il} = a_l)} = \frac{\mu_{ikl}^{(a_k,a_l)}}{\mu_{ik}^{(a_k)}\mu_{il}^{(a_l)}}, \tag{3}$$

for $a_k, a_l = 2, \ldots, f$ and $k, l = 1, \ldots, q$, $k \neq l$, that is, the joint probability divided by joint probability assuming independence.

To specify the joint distribution through the mean parameters in (2), the second-order means can be expressed as a simple transformation of the marginal probabilities and the second-order dependence ratios: $\mu_{ikl}^{(a_k,a_l)} = \mu_{ik}^{(a_k)}\mu_{il}^{(a_l)}\tau_{kl}^{(a_k,a_l)}$. Higher-order transformations can be expressed similarly. In what follows, we regress, using the most appropriate link function, the univariate marginal or cumulative probabilities on explanatory variables with intercept parameters $\boldsymbol{\theta}$ and regression coefficients $\boldsymbol{\beta}$. In order to find an underlying association model generating dependence within a cluster, we impose a structure on $f^q - q(f-1) - 1$ dependence ratios $\boldsymbol{\tau} = (\tau_{12}^{(2,2)}, \ldots, \tau_{1\ldots q}^{(f,\ldots,f)})$ through association parameters

$$\boldsymbol{\tau} = g(\boldsymbol{\alpha}). \tag{4}$$

Equation (4) can also be extended to include explanatory variables. Note that an explicit expression $\boldsymbol{\pi}_i = f(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ now exists for maximum likelihood (ML) estimation of the combined regression and association model.

2.1. *Illustration of the parametrization*

Consider a bivariate ordinal dataset, previously analyzed by Hout, Duncan, and Sobel (1987), concerning responses of married couples to the questionnaire item: 'Sex is fun for me and my partner'. The response is ordinal with levels 1=Almost Always, 2=Very Often, 3=Fairly Often, 4=Never or Occasionally. Responses of 91 couples are summarized in Table 2. It is of interest whether husbands and wives differ in the way they are distributed into these four categories. In addition, it is presumable that the married couples are in some (imperfect) way associated in their responses. Hout, Duncan, and Sobel (1987) called the former as marginal dissimilarity, and the latter as structural dissimilarity.

Altogether $2(4-1) = 6$ univariate marginal probabilities, that measure marginal dissimilarity, and $4^2 - 2(4-1) - 1 = 9$ dependence ratios, that measure structural dissimilarity, are required to specify the joint distribution of this bivariate ordinal response. Table 3 summarizes the estimates for the saturated model with 15 parameters. Marginal distributions do not notably differ between husbands and wives. It also seems that there is little association

Table 2: Responses of married couples: Sex is fun for me and my partner.

	Wife				
Husband	Always fun	Very Often	Fairly Often	Never Fun	Total
Always fun	14	9	8	2	33
Very Often	9	4	5	1	19
Fairly Often	7	3	8	2	20
Never Fun	3	2	7	7	19
Total	33	18	28	12	91

in the joint responses of the married couples: Most of the dependence ratios are close to one, which implies near independence. However, the estimated joint probability that both of the couples respond 'Never or Occasionally' is rather high, $7/91 = 0.077$, compared to the joint probability assuming independence: $(12/91) \times (19/91) = 0.0275$. This results in a dependence ratio estimate of $\hat{\tau}^{(4,4)} = 0.077/0.0275 = 2.79$.

Table 3: Estimates of the saturated model: Sex is fun for me and my partner.

	Wife				
Husband	Always fun	Very Often	Fairly Often	Never Fun	Marginal
Always fun					
Very Often		$\hat{\tau}^{(2,2)} = 1.06$	$\hat{\tau}^{(2,3)} = 0.86$	$\hat{\tau}^{(2,4)} = 0.40$	$\hat{\mu}^{(2)} = 0.21$
Fairly Often		$\hat{\tau}^{(3,2)} = 0.76$	$\hat{\tau}^{(3,3)} = 1.30$	$\hat{\tau}^{(3,4)} = 0.76$	$\hat{\mu}^{(3)} = 0.22$
Never Fun		$\hat{\tau}^{(4,2)} = 0.53$	$\hat{\tau}^{(4,3)} = 1.20$	$\hat{\tau}^{(4,4)} = 2.79$	$\hat{\mu}^{(4)} = 0.21$
Marginal		$\hat{\mu}^{(2)} = 0.20$	$\hat{\mu}^{(3)} = 0.31$	$\hat{\mu}^{(4)} = 0.13$	1

In order to explore the underlying mechanisms of marginal dissimilarity, restrictions can be imposed for the univariate probabilities in a regression model, using explanatory variables and various choices of link functions. In addition, structural dissimilarity can be explored by imposing a structure on the dependence ratios. This example is revisited in Section 6, where a marginal regression model is combined with a meaningful model for the association that explains the dependence between the responses of married couples.

2.2. *Properties of the dependence ratio*

The odds ratio is the most commonly used measure of association for a multivariate categorical response; see, for example: Fitzmaurice and Laird (1993); Glonek and McCullagh (1995); Molenberghs and Lesaffre (1994); and Lang and Agresti (1994). To elaborate the dependence ratio measure and to relate it to the odds ratio and the relative risk in the context of one our main examples, consider the second-order dependence ratio $\tau_{12}^{(2,3)}$ at ages 13 and 14 in the teenage marijuana use example.

(1) **Interpretation:** $\tau_{12}^{(2,3)}$ compares how many times more probable than under independence it is that the teenager is first an occasional user and the next year a frequent user. The construction and the interpretation of a higher-order dependence ratio is a straightforward generalization. For example, the fifth-order dependence ratio $\tau_{12345}^{(3,3,3,3,3)}$ measures the probability of a teenager being a frequent user throughout the ages from 13 to 17 compared to the probability assuming independence.

(2) **Relationship with the relative risk:** Like the relative risk (risk ratio), $\tau_{12}^{(2,3)}$ is a ratio of probabilities. Several authors favor the relative risk over the odds ratio (Greenland, 1987; Sackett, Deeks, and Altman, 1996; Davies, Crombie, and Tavakoli, 1998; King and Zeng, 2002), especially because risks and relative risks are easier to interpret than ratios of odds. In reply to Deeks et al. (1998), Davies, Crombie, and Tavakoli go as far as saying: 'On one thing we are in clear agreement: odds ratios can lead to confusion and alternative measures should be used when these are available.'

(3) **Invariance:** Similar to the relative risk, $\tau_{12}^{(2,3)}$ is not invariant to the coding chosen. This means that using a reverse coding for the response results in a different set of associ-

ation measures. However, one advantage of this asymmetry is that $\tau_{12}^{(2,3)}$ only measures the association of a teenager being first an occasional user and then a frequent user. Therefore, if the association of marijuana use, rather than non-use, is of interest, the dependence ratio is focused on the association of interest.

(4) **Range:** While the odds ratio takes values between zero and infinity, the range of the dependence ratio depends on the marginal probabilities. The upper bound of $\tau_{12}^{(2,3)}$ is $\min(1/\mu_1^{(2)}, 1/\mu_2^{(3)})$. This is again similar to the properties of the relative risk: For example, if the baseline risk is 0.8, the maximum relative risk is $1/0.8 = 1.25$. However, if the proportion of marijuana users was really excessive, say 0.8, it is reasonable to argue that it would be of more interest to model non-use, rather than use of marijuana. Therefore, to exploit fully the preciseness of the dependence ratio as a measure of the association (see point 3 above), it is advisable to use as the reference category the category that contains the higher observed frequency of the first and last categories.

(5) **Connection with the transition probability:** When the responses within a cluster have a natural ordering, such as age in the teenage marijuana use example, the transition probabilities between the states may be of interest. A simple connection exists:

$$\tau_{12}^{(2,3)} = \frac{\mathrm{pr}(Y_{i2} = 3 | Y_{i1} = 2)}{\mathrm{pr}(Y_{i2} = 3)}. \tag{5}$$

Thus an alternative interpretation of $\tau_{12}^{(2,3)}$ is, in terms of conditional probability, given the teenager was an occasional user at the age of 13, how many times more probable it is that she/he is a frequent user at the age of 14, compared to the marginal probability.

Regardless of the above mentioned differences between the odds ratio and the dependence ratio, we find that one of the most important advantages in using dependence ratios is a convenient formulation of various plausible association mechanisms. We illustrate this in the following two sections in the context of our two main examples.

3. OPINIONS CONCERNING GOVERNMENT SPENDING

We set ourselves a regression task to compare whether opinions concerning government spending on each of the 7 targets differ according to political party affiliation. Marginal frequencies and percentages of the responses by political party affiliation are summarized in Table 4.

Table 4: Marginal frequencies (percentages) of responses regarding government spending by political party affiliation.

Government Spending	Party	Too Little	About Right	Too Much	Missing
National Defence	Democrat	35 (15.4)	86 (37.7)	98 (43.0)	9 (3.9)
	Independent	29 (16.7)	65 (37.4)	76 (43.7)	4 (2.3)
	Republican	35 (17.1)	108 (52.7)	59 (28.8)	3 (1.5)
Assistance to Big Cities	Democrat	61 (26.8)	94 (41.2)	73 (32.0)	0 (0.0)
	Independent	38 (21.8)	66 (37.9)	70 (40.2)	0 (0.0)
	Republican	35 (17.1)	80 (39.0)	90 (43.9)	0 (0.0)
Law Enforcement	Democrat	156 (68.4)	58 (25.4)	14 (6.1)	0 (0.0)
	Independent	106 (60.9)	55 (31.6)	13 (7.5)	0 (0.0)
	Republican	116 (56.6)	76 (37.1)	13 (6.3)	0 (0.0)
Education	Democrat	175 (76.8)	51 (22.4)	2 (0.9)	0 (0.0)
	Independent	134 (77.0)	33 (19.0)	7 (4.0)	0 (0.0)
	Republican	137 (66.8)	59 (28.8)	9 (4.4)	0 (0.0)
Enviromnent	Democrat	164 (71.9)	49 (21.5)	15 (6.6)	0 (0.0)
	Independent	137 (78.7)	29 (16.7)	8 (4.6)	0 (0.0)
	Republican	143 (69.8)	50 (24.4)	12 (5.9)	0 (0.0)
Assistance to Poor	Democrat	176 (77.2)	37 (16.2)	14 (6.1)	1 (0.4)
	Independent	125 (71.8)	36 (20.7)	10 (5.7)	3 (1.7)
	Republican	101 (49.3)	72 (35.1)	29 (14.1)	3 (1.5)
Health	Democrat	186 (81.6)	32 (14.0)	10 (4.4)	0 (0.0)
	Independent	125 (71.8)	35 (20.1)	14 (8.0)	0 (0.0)
	Republican	123 (60.0)	71 (34.6)	11 (5.4)	0 (0.0)

The original categories for party affiliation were (a) Strong Democrat, (b) Not very strong Democrat, (c) Independent, close to Democrat, (d) Independent, (f) Independent, close to Republican, (g) Not very strong Republican, and (h) Strong Republican. This variable was collapsed to a 3-level factor with levels: Democrat (a, b), Independent (c, d, e), and Republican (f, g). In order to adjust for possible confounders, age (in years), sex (male, female), and race (white, black, other) were also included as explanatory variables into the regression model, denoted by R. There were 4 missing values for party affiliation, and 2 for age. These missing values were imputed using regression imputation (Little and Rubin, 2002). Missing values in the responses, the rightmost column in Table 4, were assumed to be missing at random (Little and Rubin, 2002), and handled accordingly; see Section 5.

The marginal regression model, for $i = 1, \ldots, 607$, $k = 1, \ldots, 7$ and $a = 1, 2$, is of form

$$\eta_{ik}(a) = \theta_a + \boldsymbol{\beta}_a \boldsymbol{x}_{ik}^T, \tag{6}$$

where $\boldsymbol{\beta}_a$ is a vector of regression coefficients, constant with respect to i and k, $\boldsymbol{x}_{ik}$ a vector of explanatory variables and θ_1, θ_2 are the intercepts. For ordinal variables, the link function η is usually a logit, probit or complementary log-log function of the cumulative probabilities $\text{pr}(Y_{ik} \leq a)$. In this case it is often plausible to assume $\boldsymbol{\beta}_a = \boldsymbol{\beta}$. However, other link functions that operate on marginal reponse categories may be useful in some situations, such as adjacent category logit and baseline category logit (Hartzel, Agresti, and Caffo, 2001). For the government spending dataset, we use the most common one, the logit function for cumulative probabilities, with regression parameters constant with respect to $a = 1$ and 2, that is, the proportional odds assumption.

An individual profile, concerning opinions about government spending, has $3^7 = 2187$ possible realizations: $\{1,1,1,1,1,1,1\}, \{1,1,1,1,1,1,2\}, \ldots, \{3,3,3,3,3,3,3\}$. In addition to marginal probabilities, altogether $3^7 - 7(3-1) - 1 = 2172$ dependence ratios are required to specify the profile probabilities. Consequently, a strong structure needs to be imposed on these measures. Ekholm et al. (2003) presented a set of exchangeable association models where the association is parametrized in terms of dependence ratios. We discuss the appropriateness of these models in the context of our analysis of the government spending dataset. In what follows, full independence of the responses within a cluster is referred to as the null association model, denoted by $\mathcal{I}$. For the technical details of these structures, we refer the reader to Ekholm et al. (2003).

3.1. *Necessary Factor* $\mathcal{N}$

Suppose that there is a subgroup of people that, regardless of the question at hand, always answers that the government is spending too little. This kind of association can be captured by imposing a latent structure, denoted by $\mathcal{N}$, where all responses of a subject either do or do not carry a factor necessary for the response to be greater than one, that is, $a_{ik} > 1$. Conversely, if there is a subgroup that always answers 'too much' regardless of the question, this can be captured similarly by using the same association structure but with reverse coding of the response; see also Section 2, point 4.

Denote the absence and presence of this kind of necessary factor by, respectively, $\{N_i = 0\}$ and $\{N_i = 1\}$ and suppose that $\text{pr}(N_i = 1) = \nu_1$, for $i = 1, \ldots, 607$. Furthermore, suppose that the responses within a cluster are conditionally independent given N_i. If $N_i = 0$, an individual i will always answer 'too little', irrespective of the covariate values. Therefore it is usually more appropriate to regress the effect of the covariates on the univariate probability conditional on the presence of necessary factor. Using the conditional univariate probabilities, the profile probabilities can now be expressed, for $a_k = 1, 2, 3$, as

$$\pi_i(a_1, \ldots, a_7) = \nu_1\{\mathrm{pr}(Y_{i1} = a_1|N_i = 1)\cdots\mathrm{pr}(Y_{i7} = a_7|N_i = 1)\} + \mathbf{1}_{\{a=1\}}(1 - \nu_1), \qquad (7)$$

where $\mathbf{1}_{\{a=1\}} = 1$ if $\{a_1 = \cdots = a_7 = 1\}$, else 0. The association model has a single parameter ν_1, and $1 - \nu_1$ quantifies the proportion of people that will always answer that the government is spending too little. To elaborate further the association model, other structures can be imposed on the conditional probabilities in (7). Note also from (7) that the probability of the profile $\{a_1 = \cdots = a_7 = 1\}$ is a sum of probabilities with and without the necessary factor. In other words, a subject may answer 'too little' to all questions but still be capable of answering otherwise.

When fitted to government spending data, $\mathcal{N}$ combined with a regression model R, gives $\hat{\nu_1} = 0.99$. In other words, there is little evidence that there exists a notably big subgroup that always answers 'too little'.

3.2. *Latent Binary Factor $\mathcal{L}$*

Suppose that the population is divided into two groups with different response category probabilities. This may happen if an important dichotomous covariate has not been recorded or cannot be observed. Suppose that each subject i has a realization of a latent factor $L_i = 1$ or 0, and that all 7 responses are conditionally independent given L_i. This association model, called $\mathcal{L}$, has 3 parameters $\boldsymbol{\alpha} = (\nu_2, \kappa^{(2)}, \kappa^{(3)})$, where $\nu_2 = \mathrm{pr}(L_i = 1)$ and the $\kappa^{(a_k)} = \mathrm{pr}(Y_{ik} = a_k|L_i = 0)/\mathrm{pr}(Y_{ik} = a_k|L_i = 1)$, $a_k = 2, 3$. In other words, ν_2 quantifies the proportion of subjects in the latent group 1, and κ-parameters are ratios of conditional univariate probabilities for those in the latent groups 0 and 1 respectively.

For $w = 2, \ldots, 7$, equation (4) is now defined using the following connection:

$$\tau^{(a_1,\ldots,a_w)} = \frac{\nu_2 + (1-\nu_2)\kappa^{(a_1)} \cdots \kappa^{(a_w)}}{\{\nu_2 + (1-\nu_2)\kappa^{(a_1)}\} \cdots \{\nu_2 + (1-\nu_2)\kappa^{(a_w)}\}}. \tag{8}$$

Estimates for the combined model $\{R; \mathcal{L}\}$, fitted to the government spending dataset, are reported in Table 5 and discussed in Section 3.5.

3.4. *Latent Dirichlet-distributed Propensities* $\mathcal{D}$

Suppose that there is continuous variability in the way individuals respond to questions regarding government spending. This may be caused by an unobserved continuous explanatory variable, or a combination of unobserved variables, resulting in different underlying propensities for each individual. This kind of association structure can be captured by utilizing a continuous Dirichlet distribution (Kotz, Balakrishnan, and Johnson, 2000, Chap. 49), which is an extension of the Beta distribution to more than two categories.

Denote the propensities of subject i by $\boldsymbol{P}_i = (P_i^{(2)}, P_i^{(3)})$, $P_i^{(2)}, P_i^{(3)} \geq 0$, $P_i^{(2)} + P_i^{(3)} \leq 1$, and suppose that $\boldsymbol{P}_i$, $i = 1, \ldots, 607$, follow independently the same Dirichlet distribution with parameters $\xi_1, \xi_2, \xi_3 > 0$. In other words, each subject has an individual realization of propensities $\{1-(p_i^{(2)}+p_i^{(3)}),\ p_i^{(2)},\ p_i^{(3)}\}$ for, respectively, categories 1='too little', 2='about right', and 3='too much'. Further suppose that all 7 responses concerning government spending are conditionally independent given the propensities. We call this association model $\mathcal{D}$.

Equation (4) can now be expressed with $\boldsymbol{\alpha} = (\xi_1, \xi_2, \xi_3)$ using the following connection:

$$\tau^{(a_1,\ldots,a_w)} = \frac{E(P^{(a_1)} \cdots P^{(a_w)})}{E(P^{(a_1)}) \cdots E(P^{(a_w)})}. \tag{9}$$

To illustrate this association model in terms of dependence ratios, consider e.g. $\tau^{(2,2)}$, $\tau^{(2,2,2)}$, $\tau^{(2,3)}$, and $\tau^{(2,2,3)}$. Assuming latent Dirichlet-distributed propensities, the corresponding equations are:

$$\tau^{(2,2)} = \frac{\xi_2(\xi_1+\xi_2+\xi_3+1)+(\xi_1+\xi_3)}{\xi_2(\xi_1+\xi_2+\xi_3+1)}; \quad \tau^{(2,2,2)} = \tau^{(2,2)} \times \frac{\xi_2(\xi_1+\xi_2+\xi_3+2)+2(\xi_1+\xi_3)}{\xi_2(\xi_1+\xi_2+\xi_3+2)};$$

$$\tau^{(2,3)} = \frac{(\xi_1+\xi_2+\xi_3)}{(\xi_1+\xi_2+\xi_3+1)}; \quad \tau^{(2,2,3)} = \tau^{(2,3)} \times \frac{\xi_2(\xi_1+\xi_2+\xi_3)+(\xi_1+\xi_2+\xi_3)}{\xi_2(\xi_1+\xi_2+\xi_3)+2\xi_2}.$$

Note that $\tau^{(2,2,2)} > \tau^{(2,2)} > 1$ and $\tau^{(2,3)} < 1$, regardless of the values of $\xi > 0$. This model, arising from the properties of a Dirichlet distribution, is therefore appropriate for studies where it is assumed that repetition of certain response categories is more probable than under independence. When fitted to the government spending dataset, the estimates of the association parameters for the combined model $\{R; \mathcal{D}\}$ are $\hat{\xi}_1 = 0.467$, $\hat{\xi}_2 = 3.010$, $\hat{\xi}_3 = 4.804$. From the above equations, we get $\hat{\tau}^{(2,2)} = 1.19$, $\hat{\tau}^{(2,2,2)} = 1.59$, $\hat{\tau}^{(2,3)} = 0.89$ and $\hat{\tau}^{(2,2,3)} = 0.96$.

3.5. *Results for the combined regression and association model*

The results of the regression model are summarized in Table 5. The parameter estimates for the government spending targets of the Democrats are contrasted with spending on National Defence. The view of the Democrats is that the government should spend more on all other targets than on National Defence, with the greatest emphasis on Health.

The parameter estimates for the Independents and the Republicans are contrasted with the estimates for the Democrats. In other words, 0 indicates no difference with the views of the Democrats on that specific target. People who call themselves Independent, have opinions somewhere in between those of Democrats and Republicans, with views generally slightly closer to Democrats than Republicans. The exception is The Environment, where the Independents seem to feel the government should spend more, compared to Democrats and Republicans, whose views do not differ.

The greatest differences in the viewpoints of the Democrats and the Republicans are in government spending regarding Assistance to the Poor, and Health. For both these targets the Democrats expect much stronger financial involvement. In contrast, for National Defence, Republicans expect the government to spend more.

Since the association models presented in Sections 3.1 to 3.3 are not nested, we use Akaike's information criteria (AIC) to compare which of the models fit the data best. AIC for models $\{R;\mathcal{I}\}$, $\{R;\mathcal{N}\}$, $\{R;\mathcal{L}\}$, and $\{R;\mathcal{D}\}$ are, respectively, 7039.9, 7029.8, 6971.9 and 6972.2. In terms of the model fit, there is virtually no difference between the fit assuming a latent binary variable, or assuming a latent continuous, Dirichlet-distributed variable. One needs to resort to subject-matter judgement in order to distinguish which one of the two is the more plausible mechanism. In Table 5, we report the regression model combined with a model assuming a latent binary factor for the association.

The interpretation of the association model $\mathcal{L}$ is that the population is divided into two groups, denoted by 0 and 1, with different response category probabilities concerning government spending. The percentage of subjects in group 0 is $100 \times (1 - \hat{\nu}_2) \approx 40\%$, and for those in that group, the probability of answering 'about right' is approximately one third (0.362) of the probability in group 1. However, answering 'too much' is 1.541 times more probable than in group 1. In search for explanation for the latent variable, one needs to look for a subgroup that constitutes 40% of the target population, and is more prone to a view that the government is spending too much, compared to their counterparts.

Table 5: Estimates of a model $\{R; \mathcal{L}\}$ concerning government spending

Effects		Estimate	Std. Error
Sex	Female	−0.214	0.067
Race	Black	−0.437	0.119
	Other	0.431	0.192
Age	(in years)	0.0038	0.0019
Democrat	National Defence	-	-
	Assistance to Big Cities	−0.564	0.174
	Law Enforcement	−2.471	0.189
	Education	−2.919	0.201
	Enviromnent	−2.625	0.194
	Assistance to Poor	−2.904	0.204
	Health	−3.170	0.213
Independent (Contrasted to Democrats)	National Defence	−0.047	0.193
	Assistance to Big Cities	0.278	0.191
	Law Enforcement	0.264	0.207
	Education	−0.024	0.239
	Enviromnent	−0.426	0.236
	Assistance to Poor	0.191	0.235
	Health	0.525	0.240
Republican (Contrasted to Democrats)	National Defence	−0.506	0.175
	Assistance to Big Cities	0.503	0.191
	Law Enforcement	0.348	0.196
	Education	0.398	0.214
	Enviromnent	−0.017	0.210
	Assistance to Poor	1.126	0.209
	Health	0.931	0.220
Association	ν_2	0.596	0.103
	κ_2	0.362	0.057
	κ_3	1.541	0.138

The association structures $\mathcal{N}$, $\mathcal{L}$, and $\mathcal{D}$ are all exchangeable, that is, independent of the ordering of the responses $Y_{i1}, \ldots, Y_{iq}$. For longitudinal studies, it is more appropriate to assume that the association between repeated measurements has mainly a temporal rather than an exchangeable structure. In the next example, we present a model for temporal association (Ekholm et al., 2003), and extend further the exchangeable structures by introducing a new hierachical latent structure that exploits the ordinality of the response.

4. STABILITY AND CHANGE IN TEENAGE MARIJUANA USE

Marginal frequencies and percentages of marijuana use for boys and girls at ages 13 to 17 are reported in Table 6. These show a monotone increase in marijuana use for both girls and boys with age, with use being consistently more frequent for boys at each age. None of the link functions described in Section 3 is superior over the others, so a proportional odds model is used for estimation of the effect of age in the regression model.

Table 6: Marginal frequencies (percentages) of marijuana use for girls and boys at ages 13 to 17.

Age	Girls			Boys		
	Never	Less than once a month	More than once a month	Never	Less than once a month	More than once a month
13	114 (95.0)	5 (4.2)	1 (0.8)	104 (88.9)	9 (7.7)	4 (3.4)
14	106 (88.3)	10 (8.3)	4 (3.3)	89 (76.1)	17 (14.5)	11 (9.4)
15	91 (75.8)	21 (17.5)	8 (6.7)	76 (65.0)	20 (17.1)	21 (17.9)
16	85 (70.8)	21 (17.5)	14 (11.7)	71 (60.7)	20 (17.1)	26 (22.2)
17	75 (62.5)	30 (25.0)	15 (12.5)	63 (53.8)	22 (18.8)	32 (27.4)

The clustering of the repeated responses for each teenager occurs in time. Consequently, Markov structures are a natural choice for modeling the temporal association of teenage marijuana use in age.

4.1. *Markov Structures $\mathcal{M}$*

Consider a path containing five consecutive annual responses for teenager i, $(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5})$, and suppose that these satisfy the first-order Markov assumption. For $i = 1, \ldots, 237$ and $k = 1, \ldots, 4$,

$$\mathrm{pr}(Y_{i(k+1)} = a | Y_{i1}, \ldots, Y_{ik}) = \mathrm{pr}(Y_{i(k+1)} = a | Y_{ik}). \qquad (10)$$

Equation (10) implies that this association structure, denoted by $\mathcal{M}$, has $(5-1)(3-1)^2 = 16$ adjacent second-order dependence ratio parameters: $\tau_{k\,k+1}^{(a_k, a_{k+1})}$, $a_k = 2, 3$ and $k = 1, \ldots, 4$. For such large-dimensional problems, this still leads to an unneccesarily complicated association structure. For clearer interpretability, it is advisable to strengthen the $\mathcal{M}$ assumption by further restrictions, such as functional forms for the adjacent dependence ratios, and introducing equality of parameters for certain pairs of categories or time points.

The set of association models can also be further elaborated by assuming that the $\mathcal{M}$-structures operate, independently, within the latent classes of the exchangeable structures $\mathcal{N}$ and $\mathcal{L}$.

4.2. *Hierarchical Necessary Factors $\mathcal{N}2$*

Consider the structure $\mathcal{N}$ presented in Section 3.1 in the teenage marijuana example, dividing the teenage population into those who would never use marijuana, and to those that are susceptible to marijuana use. Further suppose that there is another subpopulation among the susceptibles that might try marijuana or use it occasionally, but categorically refuse to become frequent users. We can capture this kind of behavior in the population by nesting the latent structures $\mathcal{L}$ and $\mathcal{N}$.

Suppose that $\mathcal{L}$ is defined conditionally on $\{N_i = 1\}$, with $\nu_2 = \text{pr}(L_i = 1|N_i = 1)$ and $\kappa^{(a_k)} = \text{pr}(Y_{ik} = a_k|L_i = 0, N_i = 1)/\text{pr}(Y_{ik} = a_k|L_i = 1, N_i = 1)$, for $a_k = 2, 3$. Furthermore, impose a restriction $\kappa^{(3)} = 0$. We denote our extension of the $\mathcal{N}$ association model by $\mathcal{N}2$. There are three latent classes and two necessary factors that operate in a hierarchical manner. The profile probabilities follow from equations (7) and (8) where $\mathcal{L}$ is imposed on the product of conditional probabilities, with $\kappa^{(3)} = 0$. Note that $\tau^{(a_1,\ldots,a_w)}$ is here defined conditionally on $\{N_i = 1\}$ and thus ν_2 quantifies the proportion of potential frequent users within the susceptibles. This association model has three parameters, $\boldsymbol{\alpha} = (\nu_1, \nu_2, \kappa^{(2)})$, but can be further simplified by assuming $\kappa^{(2)} = 1$, which implies that the susceptibles and potential frequent users have the same probability for occasional marijuana use.

4.3 *Results of the combined regression and association model*

Regardless of similar shape of the marginal probablilities, the observed paths show quite different patterns for boys and girls. This stresses the importance of modeling the whole path probability instead of only the marginal probabilities. It can be seen from Table 1 that the observed paths for boys are more dispersed than for girls, with the majority of girls staying as a non-user throughout the follow-up. This suggests that the association structures, while temporal, are different for boys and girls. In fact, modeling girls and boys together would require seven interaction terms with sex. Therefore, we analyze girls and boys separately.

Model for the marijuana use of boys

As was noted in Section 4.1, assuming $\mathcal{M}$, some sensible constraints need to be imposed on the 16 second-order dependence ratios, without losing the essence of the form of the association.

The dashed lines in Figure 1 summarize the 16 observed adjacent second-order dependence ratios for boys with bootstrap 95% confidence intervals. The solid lines in Figure 1 present the fitted values, where the dependence ratios for movers $\tau^{(2,3)}$ and $\tau^{(3,2)}$ are assumed to follow a linear relationship for the consecutive pairs of ages, and dependence ratios of the stayers $\tau^{(2,2)}$ and $\tau^{(3,3)}$ are assumed to be stationary across the pairs of ages. This proves to be a superior fit over the exhangeable structures. The parameter estimates of this model $\mathcal{M}_6$, where the subscript indicates the number of association parameters, combined with a regression model R for the age-effect are reported in Table 7.

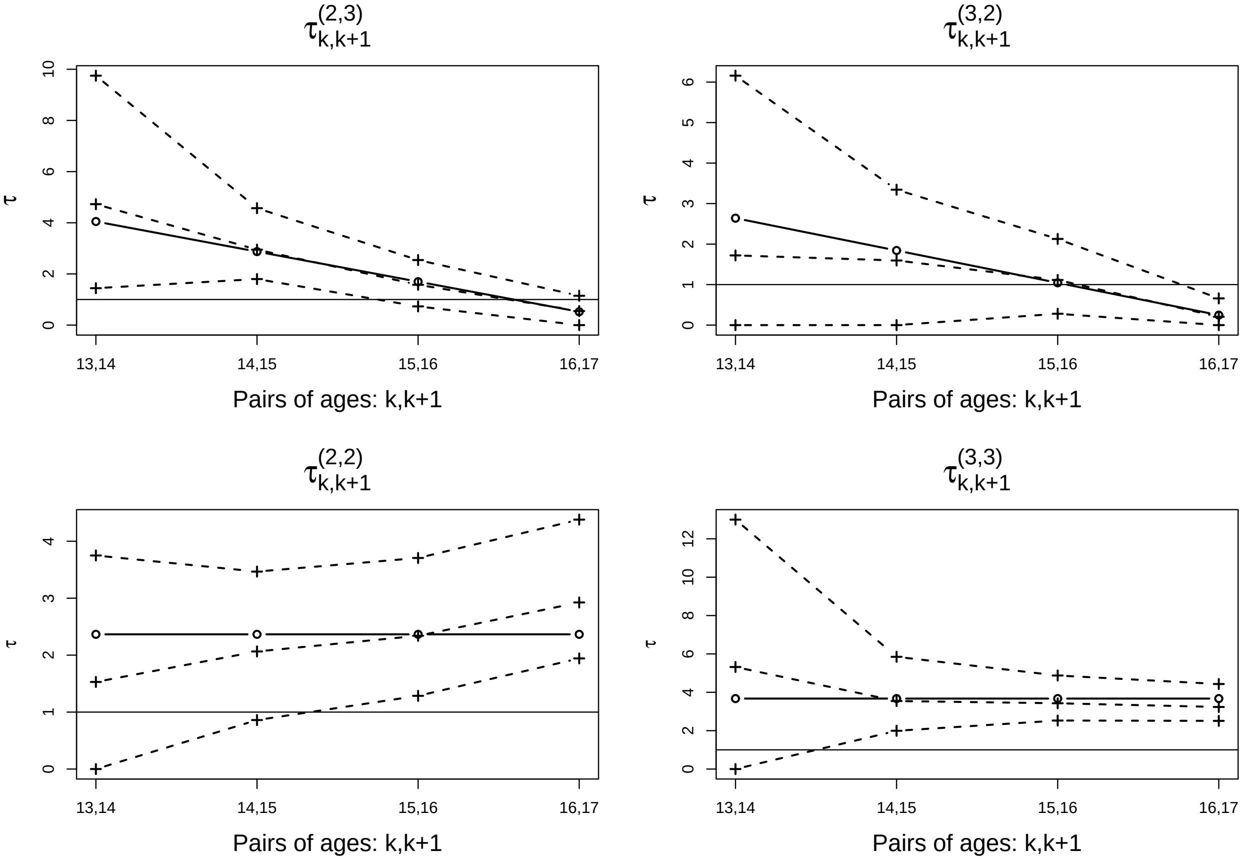

Figure 1: Observed second-order dependence ratios with bootstrap 95% confidence intervals (dashed lines) compared with the fitted ones (solid lines) for model $\{R; \mathcal{M}_6\}$ for the marijuana use of boys. Horizontal line at value 1 corresponds to independence.

Table 7: Model $\{R; \mathcal{M}_6\}$: ML estimates for marijuana use of boys.

Part of model	Effect	Estimate	Standard error
Regression model	age 14	0.862	0.282
	age 15	1.402	0.286
	age 16	1.649	0.295
	age 17	1.900	0.306
Association model	$\tau^{(2,2)}$	2.366	0.292
	$\tau^{(3,3)}$	3.673	0.463
	$\tau^{(2,3)}_{intercept}$	4.049	0.814
	$\tau^{(2,3)}_{slope}$	-1.176	0.288
	$\tau^{(3,2)}_{intercept}$	2.640	0.688
	$\tau^{(3,2)}_{slope}$	-0.798	0.257

The subscripts *intercept* and *slope* for the association parameters correspond to, respectively, intercept and slope of the fitted linear functions in Figure 1.

From the estimated regression parameters, with age 13 as the reference category, we conclude that the probability of marijuana use increases from one year to the next. The interpretation of the Markov association model is that remaining as an occasional or a frequent user at both ages is, respectively, over two (2.366) and over three (3.673) times more probable than under independence. However, the dependence ratios for changing their habit from frequent use to occasional use ($\tau^{(3,2)}_{slope}$) and vice versa ($\tau^{(2,3)}_{slope}$) at two consecutive ages decrease with increasing age, change being even less probable between ages 16 and 17 than under independence (see Figure 1). This can be interpreted as the boys who use marijuana, gradually develop a habit through the teenage years, which they are eventually reluctant to change.

Alternatively, if conditional dependencies are of interest, we can use equation (5) for the interpretation: given that the boy was an occasional user at age 16, the probability of being a frequent user at age 17 is only about half the marginal probability ($4.049 - 1.176 \times 3 = 0.52$). Similarly, if marijuana was used frequently at age 16, the probability of occasional marijuana use at age 17 is only a quarter the marginal probability $(2.640 - 0.798 \times 3 = 0.25)$.

Model for the marijuana use of girls

The observed adjacent second-order dependence ratios for girls did not reveal any clear structure. An interpretable fit assuming a Markov structure can be achieved by assuming stationarity over time for $\tau^{(a,b)}$, $a, b = 2, 3$, denoted by $\mathcal{M}_4$. The rightmost column in Table 8 summarize the observed frequencies for girls throughout the follow-up from age 13 to 17 in three exclusive classes: (i) a non-user at all ages, (ii) at most an occasional user at least once and (iii) a frequent user at least once. The majority of the girls, 53%, fall in category (i), suggesting that the association model, assuming first-order Markov structure, could be modified to include a necessary factor. In other words, within this 53%, there might be a subgroup of girls not susceptible to marijuana use. Furthermore, allowing $\mathcal{M}$ to operate independently within the two latent classes of a structure $\mathcal{L}$, gives an estimate of $\kappa^{(3)}$ near zero. Imposing the restriction $\kappa^{(3)} = 0$ gives rise to an association model $\mathcal{N}2\mathcal{M}_7$, which is significantly better than $\mathcal{M}_4$ (likelihood ratio test p-value = 0.016). This association model implies that the population is divided into the following latent groups: the non-susceptibles and the susceptibles. Furthermore, within the susceptibles, there is still a subgroup of potential frequent users. The following plausible constraints $\tau^{(2,2)} = \tau^{(3,3)}$ and $\tau^{(2,3)} = \tau^{(3,2)} = \kappa^{(2)} = 1$ have negligible effect on the likelihood, but make the interpretation of the association model, denoted by $\mathcal{N}2\mathcal{M}_3$, simpler. The improvement in the fit assuming $\mathcal{M}$, and further by introducing hierarchical necessary factors are epitomized in Table 8, where the fitted counts of the models with association structures $\mathcal{I}$, $\mathcal{M}_4$ and $\mathcal{N}2\mathcal{M}_3$ are compared with the observed counts.

Parameter estimates of the model $\mathcal{N}2\mathcal{M}_3$, combined with a regression model R for the

Table 8: Observed and fitted counts of models with different association structures for the marijuana use of girls.

Category	$\mathcal{I}$	$\mathcal{M}_4$	$\mathcal{N}2\mathcal{M}_3$	Observed
A non-user at all ages	33.77	58.17	63.00	63
At most an occasional user at least once	49.24	40.07	36.01	36
A frequent user at least once	36.99	21.77	20.99	21

age-effect are reported in Table 9. A similar conclusion can be made for the regression model for the girls as for the boys. However, the interpretation of the association model is quite different to boys: only 63% of the girls are susceptible to marijuana use through ages from 13 to 17, and out of these girls, only 42% are susceptible to frequent use. These latent structures, however, do not account for all the dependence between consecutive measurements, but temporal association still exists for the susceptibles; staying as an occasional user at two consecutive years is approximately 1.5 times more probable than under independence, as is staying as a frequent user within the latent class of potential frequent users.

Table 9: Model $\{R;\mathcal{N}2\mathcal{M}_3\}$: ML estimates for marijuana use of girls.

Part of model	Effect	Estimate	Standard error
Regression model	age 14	1.063	0.505
	age 15	2.015	0.475
	age 16	2.193	0.475
	age 17	2.435	0.469
Association model	ν_1	0.629	0.087
	ν_2	0.422	0.085
	$\tau^{(2,2)} = \tau^{(3,3)}$	1.553	0.281

5. MAXIMUM LIKELIHOOD ESTIMATION

We have presented a method that incorporates regression and association modeling for clustered ordinal responses. However, usefulness of any method is crucially dependent on its implementability. As noted by Agresti and Natarajan (2001), methods for maximum likelihood (ML) fitting of marginal and cluster-specific models are difficult to implement or require approximate methods. This is especially complicated for large-dimensional problems, such as the government spending data with a total of 7 clustered responses and several covariates, also one continuous.

As demonstrated in Section 2, there exists under our approach a closed form expression of the response profile probabilities in terms of the regression and association parameters and hence ML estimation is straightforward. By assuming an underlying multinomial distribution, the likelihood can be written as

$$\boldsymbol{l}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log\{\pi(a_{i1}, \ldots, a_{iq})\}. \tag{11}$$

In the case where some responses are missing for subject i, such as in the government spending example (see Table 4), the estimation is still straightforward by only including the observed values a_{ik} in equation (11). This is equivalent to assuming that the values are missing at random (MAR).

A significant computational advantage arises from the fact that when specifying $\boldsymbol{\pi}_i$ in terms of mean parameters, unconstrained maximization can be used, since this parametrization has a built-in unit-sum constraint (Ekholm et al., 2003). What this effectively means, is that one only needs to calculate the probability of the observed profile for each i with given parameters at each iteration. In addition, as a consequence of the closed form expression, the fitted probabilities for the unobserved profiles can also be easily calculated using the ML parameter estimates. However, when modeling sparse datasets, some of the fitted probabilities for the unobserved profiles may turn out negative. This can also happen for the conditional univariate probabilities under the $\mathcal{L}$-structure. Negative probabilities could be avoided by imposing restrictions on the model parameters. However, we do not recommend this unless it implies a meaningful model. We note that estimated negative probabilities are an indication of a poorly chosen model and can be used as a diagnostic test against misspecification of the model.

An important aspect of the applicability of any novel methodology is user-friendly and versatile software. For the freely available software R (Ihaka and Gentelman, 1996), a package has been developed for estimation of clustered binary and ordinal datasets using dependence ratios as measures of the association, which is available from the first author. This package allows the user to fit regression models, with several alternative link functions, combined with the association models described in Section 3 and 4. In the case of missing values, the software by default assumes that the data are MAR. Furthermore, for longitudinal studies with monotone missing patterns the package allows a selection model (Diggle and Kenward, 1994) to be specified on top of the regression and association model to investigate the sensitivity of the results to various models for the non-response mechanisms. See Ekholm et al. (2003) for an example.

6. OTHER APPROACHES TO THE ANALYSIS OF CLUSTERED ORDINAL DATA

There are three main approaches to analysis of clustered caregorical data: marginal models, random-effects models and transition models. Transition models are mainly applicable if the clustering occurs in time. A common way of specifying transition models is to include previous responses as explanatory variables into equation (6). Regression coefficients are then to be interpreted as effects conditional on previous responses. For a representation of different specifications of such models, see e.g. Lindsey (1999, pp. 63-67). For a more flexible approach to conditional modeling of ordinal longitudinal response, using generalization of the Kalman filter, see Lindsey and Kaufmann (2004).

If the population-averaged effects of the covariates are of interest, marginal models can be applied. A third option are random-effect models, where the parameters of the regression model are to be interpreted as subject-specific effects. For a survey of these two latter approaches, see Agresti and Natarajan (2001). We review these two approaches in relation to our method, using the sexual fun data in Table 2, and an example concerning reviews of 93 movies by four critics, previously analyzed by Hartzel, Agresti, and Caffo (2001).

6.1. *Marginal models*

Consider again the example concerning responses of married couples to the questionnaire item: ‘Sex is fun for me and my partner’. A parsimonious model is achieved with a regression model without explanatory variables, i.e., assuming that the marginal distributions of the couples responses are the same. Association model $\mathcal{L}$ with a parameter restriction $\kappa^{(4)} = 0$ fits the data well, yielding a likelihood ratio test statistic $L^2 = 4.92$ with 9 df, compared to the saturated model. This model implies that the observed weak association between couples, according to the saturated model in Table 3, may be a result of a latent binary factor. In other words, the observed association between couples is a consequence from the fact that the population of couples is divided into two groups with different response probabilities. Fitted marginals and dependence ratios of this model are summarized in Table 10.

In addition to marginal probabilities, conditional probabilities can also easily be obtained.

Table 10: Fitted parameters of model $\mathcal{L}$: Sex is fun for me and my partner

	Wife				
Husband	Always fun	Very Often	Fairly Often	Never Fun	Marginal
Always fun					
Very Often		$\tau^{(2,2)} = 1.21$	$\tau^{(2,3)} = 0.90$	$\tau^{(2,4)} = 0.45$	$\mu^{(2)} = 0.20$
Fairly Often		$\tau^{(3,2)} = 0.90$	$\tau^{(3,3)} = 1.05$	$\tau^{(3,4)} = 1.27$	$\mu^{(3)} = 0.26$
Never Fun		$\tau^{(4,2)} = 0.45$	$\tau^{(4,3)} = 1.27$	$\tau^{(4,4)} = 2.45$	$\mu^{(4)} = 0.17$
Marginal		$\mu^{(2)} = 0.20$	$\mu^{(3)} = 0.26$	$\mu^{(4)} = 0.17$	1

Table 11 summarizes the probabilities of couples in the two latent groups. Note that the interpretation of $\hat{\nu_2} = 0.41$ and restriction $\kappa^{(4)} = 0$ is that $100 \times (1 - \hat{\nu_2}) = 59\%$ of the couples would never answer 'Never or Occasionally' to a question concerning sexual fun.

Table 11: Fitted conditional probabilities of model $\mathcal{L}$: Sex is fun for me and my partner

Latent factor	$L_i = 1$	$L_i = 0$
Proportion in the population	$\hat{\nu_2} = 0.41$	$1-\hat{\nu_2} = 0.59$
Always fun	0.16	0.50
Very Often	0.09	0.28
Fairly Often	0.33	0.22
Never Fun	0.42	0.00

Marginal modeling is generally interpreted as an approach that focuses on the marginal probabilities and treats the dependence structure as a nuisance (Agresti and Natarajan, 2001). A popular approach is the GEE-methodology (Liang and Zeger, 1986), where models are specified only for marginal distributions and hence do not support maximum likelihood estimation. Without building models for the entire response profile, a comprehensive modeling of the association structure is not feasible.

For likelihood-based methods, log-linear modeling formulation is typically applied for the association. Fitzmaurice and Laird (1993), extended to the multicategorical case by Heumann

(1996), used conditional log odds ratio parameters to specify the association, which are the canonical parameters of the log-linear model. For the sexual fun data, Hout, Duncan, and Sobel (1987) fitted log-linear models with various quasi-symmetry constraints for the association. Other form of constraints, such as common local or global odds ratios (Lang and Agresti, 1994; Molenberghs and Lesaffre, 1994) have also been suggested for clustered ordinal data. Although useful in many situations for examining the form of the association, especially in the bivariate case, these constraints can be viewed as expedient ways of handling the association when the complexity of the data increase. For example, Fitzmaurice and Laird (1993) consider the association parameters as pure nuisance. In our case, symmetry constraints on the fitted dependence ratio parameters in Table 10 are a result of the fit for the underlying mechanism that has generated the association. Parameter interpretation is also straightforward, as proportions of the population or ratios of probabilities.

Agresti and Natarajan (2001) suggested that for purposes of conveying information, it is sometimes useful to provide results through univariate probabilities rather than parameter estimates. Depending on the research question, either the marginal probabilities or the conditional probabilities within the latent classes may be of interest. For models with latent structures $\mathcal{N}$ and $\mathcal{L}$, not only the marginal, but also the conditional probabilities of response categories can easily be obtained, as is shown in Table 11.

6.2. *Random-effects models*

Hartzel, Agresti, and Caffo (2001) analyzed a dataset from *Variety* magazine, containing reviews of 93 movies by four critics: Medved, Siskel, Ebert and Lyons. Each review is rated as 1=Pro (positive), 2=Mixed (mixture of positive and negative) or 3=Con (negative). An obvious regression task is to compare the distribution of the ratings for each of the critics. In addition, the investigation of the rater agreement may be considered to be equally important.

As the four critics do not have an ordering, it is natural to assume that the dependence between their ratings is exchangeable. A notable observation in the dataset, summarized in Hartzel, Agresti, and Caffo (2001), is that the 4 largest profile frequencies seem to indicate a moderate agreement between the raters (Medved, Siskel, Ebert, Lyons): 15 {1, 1, 1, 1}, 8 {3, 1, 1, 1}, 4 {3, 3, 3, 3}, and 4 {3, 1, 1, 3}. This suggests the appropriateness of association structures such as $\mathcal{N}$ or $\mathcal{D}$. The most parsimonious fit for the regression model, comparing the critics, is achieved with the adjacent category logit link with differing rater effects for $a = 1$ and 2. The left hand side of Table 12 compares the fits of the exchangeable association structures when the regression model for the critics effect, denoted by R, is specified with an adjacent category logit link. According to AIC, association model $\mathcal{D}$ produces the most parsimonious fit compared to the other structures. Regression and association parameter estimates for the combined model, using $\mathcal{D}$, are reported in the right hand side of Table 12.

Table 12: AIC of the exchangeable association models (left) and ML estimates and standard errors of model $\{R; \mathcal{D}\}$ (right) for movie critics.

AIC for five models		ML Estimates for a model $\{R; \mathcal{D}\}$			
Model	AIC		Effect	Estimate	Std Error
$\{R; \mathcal{I}\}$	766.3	Pro vs.	Siskel	-0.104	0.311
$\{R; \mathcal{N}\}$	747.1	Mixed	Ebert	-0.060	0.324
$\{R; \mathcal{L}\}$	745.8		Lyons	0.792	0.371
$\{R; \mathcal{N}2\}$	738.1	Mixed vs	Siskel	0.635	0.326
$\{R; \mathcal{D}\}$	734.8	Con.	Ebert	1.240	0.377
			Lyons	-0.050	0.379
			ξ_1	2.148	1.772
		Association	ξ_2	0.844	0.316
			ξ_3	1.252	0.509

From the regression model, with Medved as the reference category, our conclusions are the same as Hartzel, Agresti, and Caffo (2001), that is, the Pro versus Mixed effect is negligible when comparing Siskel and Ebert to Medved, and the Mixed versus Con effect is negligible when comparing Lyons to Medved. The association model suggests that there is a latent continuous variability in the quality of the movies reviewed by the critics. This can be interpreted as an indication that movies do possess certain objective criteria according to which they can be rated. One conclusion from this kind of interpretation is that judging the quality of a movie is not entirely a matter of taste.

Hartzel, Agresti, and Caffo (2001) fitted several random-effects models to account for the association of the ratings within a movie. Random-effects models take into account within subject heterogeneity in the linear predictor by imposing a mixing distribution for the mean parameter. By far the most popular choice for the mixing distribution is normal. Agresti et al. (2000) noted that this convention is possibly a controversial issue. As Lindsey (1999, p. 212) argued, the choice of the mixing distribution should be made on theoretical grounds in terms of how the mean is thought to vary in the population or, if necessary, empirically. As can be seen from Table 12, the exchangeable association models presented in Section 3 and 4 provide a rich set of meaningful association structures and, in the absence of theoretical knowledge, comparsion of models including and excluding within subject heterogeneity is straightforward through likelihood ratio tests and information criteria.

The choice of a parametric or non-parametric random-effects component is usually dependent on the choice of link function relating the mean and the linear predictor; see, for example, Hartzel, Agresti, and Caffo (2001). Since for parameter interpretation or for a more parsimonious fit, certain link functions may be preferable for specific studies, this limits the choice of random-effects models. In our case, all latent structures are imposed on the response category probabilities which are independent of the link function used. This is a natural approach if one wants to construct meaningful association models.

The parameter interpretation is the most important substantive distinction between random-effects models and regression models specified using equation (6). Regression estimates from random-effects models are cluster-specific, whereas estimates from (6) are population averaged. The merits and relevance of the effects of interest should be judged according to the problem at hand. For example, when comparing the movie critics, the question is whether the interest lies in how the critics differ when reviewing a movie, or how they differ on average across the movies. Agresti et al. (2000) noted that most of the discussion about this distinction has been in relation to epidemiological and clinical trial settings. However, they stressed that it is time to consider the practical implications also in social science applications. So far, latent variable models for clustered categorical data, where the regression model is of form (6), have received little attention in the social science literature.

6.3. *Combining the three approaches*

Vermunt and Hagenaars (2004, Chap. 15) discuss the implications of marginal, transition and random-effects approaches for ordinal longitudinal data, and fit several models applying each of the approaches to the dataset concerning teenage marijuana use, presented in Table 1. When summing up the deviances of the fitted models in Section 4.3 for boys and girls, we obtain L^2 of 177.38 with 464 df, which indicates significant improvement in terms of the model fit compared to all the alternative versions presented in Vermunt and Hagenaars (2004).

In another application of the marijuana use data, Vermunt, Rodrigo, and Ato-Garcia (2001) proposed a modification of the approach suggested by Lang and Agresti (1994) for combining marginal, transitional, and cluster-specific approaches in a single framework. However, the application was limited to four time-points rather than all five. Our modeling approach can also be viewed as a combination of the three methods. Consider the models presented in Section 4: marginal modeling is applied for the univariate probabilities; cluster-specific models are applied using the hierarchical necessary factor; and finally, the transitions are modeled using the Markov model specification, and interpreted using the link between transition probabilities and dependence ratios.

7. CONCLUDING REMARKS

A strong age effect in the prevalence of teenage cannabis use has been widely reported, see e.g. Johnston, O'Malley, and Bachman (2000). Also the higher prevalence among teenage boys, compared to the girls of same age, is well-known (Hall, Johnston, and Donnelly, 1999). However, in addition to marginal prevalences, investigation of individual profiles of cannabis use has not gained similar attention. Some of this may have been because suitable tools for this joint task have not been available. We have presented a methodology that allowed us to extract valuable additional information about the different behaviors of girls and boys through a comprehensive investigation of individual response profiles.

We presented in Sections 3 and 4 latent structures that provide a meaningful alternative to random-effects models for modeling the association structure. Alternatively for longitudinal studies, Markov-structures in Section 4.1 can be used to describe the temporal association. Furthermore, combining some of these structures is straightforward and therefore provides an ample set of association models for multivariate categorical responses.

There exists an explicit expression for the joint probability in terms of marginal probabilities and dependence ratios. In addition, the mean parametrization facilitates maximum likelihood estimation, and is therefore straightforward and feasible even for large-dimensional problems. However, sometimes fitted probabilities for non-observed response profiles are negative, which indicates model misspecification. Pursuing meaningful models for both the regression and association structures for a multivariate categorical response is a challenging task. Extra care is needed when fitting models to sparse datasets. Therefore, fitted response profile probabilities serve as a helpful tool for checking the plausibility of the fitted model.

Agresti (1999) stressed that with the continuing development of more complex models, an increasingly important but difficult task is communicating to non-statisticians the interpretation of the models and their parameters. The dependence ratio is a measure of association which, for researchers used to concepts like relative risks, is easy to grasp. In addition, the parameters of the latent variable models, such as those for the hierarchical necessary factor in the model of marijuana use by girls, have a clear and simple interpretation as proportions of the population. Compared to the multivariate normal case, the association structure of a multivariate categorical response is very complex. Further development of more complex models is therefore an admirable goal. Our view is, however, that this should not be made at the expense of the interpretability of the model.

References

Agresti, A. 1999. "Modelling Ordered Categorical Data: Recent Advances and Future Challenges." *Statistics in Medicine* 18:2191-2207.

Agresti, A., J.G. Booth, J.P. Hobert, and B. Caffo. 2000. "Random-Effects Modeling of Categorical Response Data." *Sociological Methodology* 30:27-81.

Agresti, A. and R. Natarajan. 2001. "Modeling Clustered Ordered Categorical Data: a Survey." *International Statistical Review* 69:345-71.

Agresti, A. 2002. *Categorical Data Analysis.* 2nd ed. New York: Wiley.

Davies, H.T.O., I.K. Crombie, and M. Tavakoli. 1998. "When Can Odds Ratios Mislead?" *British Medical Journal* 316:989-91.

Deeks, J.J., M.B. Bracken, J.C Sinclair, H.T.O. Davies, M. Tavakoli, and I.K. Crombie. 1998. "Letter to the Editor: When Can Odds Ratios Mislead?" *British Medical Journal* 317:1155.

Diggle, P. and M.G. Kenward. 1994. "Informative Drop-out in Longitudinal Data Analysis (with Discussion)." *Applied Statistics* 43:49-93.

Ekholm, A., P.W.F. Smith, and J.W. McDonald. 1995. "Marginal Regression Analysis of a Multivariate Binary Response." *Biometrika* 82:847-54.

Ekholm, A., J.W. McDonald, and P.W.F. Smith. 2000. "Association Models for a Multivariate Binary Response." *Biometrics* 56:712-18.

Ekholm, A., J. Jokinen, J.W. McDonald, and P.W.F. Smith. 2003. "Joint Regression and Association Modelling for Longitudinal Ordinal Data." *Biometrics* 59:795-803.

Elliott, D.S., S.S. Ageton, D. Huizinga, B.A. Knowles, and R.J. Canter. 1983. "The Prevalence and Incidence of Delinquent Behavior: 1976-1980." *Project Report No. 26, National Youth Survey, Boulder, CO: Behavioral Research Institute.*

Fitzmaurice, G.M. and N.M. Laird. 1993. "A Likelihood-based Method for Analysing Longitudinal Binary Responses." *Biometrika* 80:141-51.

Glonek, G.F.V. and P. McCullagh. 1995. "Multivariate Logistic Models." *Journal of the Royal Statistical Society,* ser. B, 57:533-46.

Greenland, S. 1987. "Interpretation and Choice of Effect Measures in Epidemiologic Analyses." *Americal Journal of Epidemiology* 125:761-68.

Hall, W., L.D. Johnston, and N. Donnelly. 1999. "Epidemiology of cannabis use and its consequences." Pp. 71-125 in *H. Kalant, W. Corrigall, W. Hall et al. (Eds.) The Health Effects of Cannabis.* Toronto: Centre for Addiction and Mental Health.

Hartzel, J., A. Agresti, and B. Caffo. 2001. "Multinomial Logit Random Effects Models." *Statistical Modelling* 1:81-102.

Heumann, C. 1996. "Marginal regression modeling of correlated multicategiorical response: A likelihood approach." *SFB386 - Discussion paper 19.* Universität München.

Hout, M., O.D. Duncan, and M.E. Sobel. 1987. "Association and Heterogeneity: Structural Models of Similarities and Differences." *Sociological Methodology* 17:145-184.

Ihaka, R. and R. Gentleman. 1996. "R: a Language for Data Analysis and Graphics." *Journal of Computational Graphics and Statistics* 5:299-314.

Johnston, L.D., P.M. O'Malley, and J.G. Bachman. 2000. *Monitoring the future: National Survey Results on Drug Use, 1975-1999.* Bethesda, MD: National Institute on Drug Abuse.

King, G., and L. Zeng. 2002. "Estimating Risk and Rate Levels, Ratios and Differences in Case-Control Studies." *Statistics in Medicine* 21:1409-27.

Kotz, S., N. Balakrishnan, and N.L. Johnson. 2000. "Dirichlet and Inverted Dirichlet distributions." Pp. 485-527 in *Continuous Multivariate Distributions, Volume 1: Models and Applications.* 2nd ed. Chichester: Wiley.

Lang, J.B. and A. Agresti. 1994. "Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses." *Journal of the Americal Statistical Association* 89:625-32.

Lang, J.B., J.W. McDonald, and P.W.F. Smith. 1999. "Association-marginal Modeling of Multivariate Categorical Responses: A Maximum Likelihood Approach." *Journal of the American Statistical Association* 94:1161-71.

Liang, K.Y. and S.L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73:13-22.

Lindsey, J.K. 1999. *Models for Repeated Measurements.* 2nd ed. New York: Oxford University Press.

Lindsey, P.J. and J. Kaufmann. 2004. "Analysis of a longitudinal ordinal response clinical trial using dynamic models." *Applied Statistics* 53:523-37.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data.* 2nd ed. New York: John Wiley.

Molenberghs, G. and E. Lesaffre. 1994. "Marginal Modelling of Correlated Ordinal Data Using a Multivariate Plackett Distribution." *Journal of the American Statistical Association* 89:633-44.

Sackett, D.L., J.J. Deeks, and D.G.Altman. 1996. "Down With Odds Ratios!" *Evidence-Based Medicine* 1:164-66.

Vermunt, J.K., M.F. Rodrigo, M. Ato-Garcia. 2001. "Modeling joint and marginal distributions in the analysis of categorical panel data." *Sociological Methods and Research* 30:170-196.

Vermunt, J.K. and J.A. Hagenaars. 2004. "Ordinal longitudinal data analysis." Pp. 374-393 in *R.C. Hauspie, N. Cameron and L. Molinari (Eds.). Methods in Human Growth Research.* Cambridge, UK: Cambridge University Press.