Brazilian Journal of Probability and Statistics (2004), 18, pp. 1–18.

Bivariate ordinal data in the perspective of the Wisconsin epidemiologic study of diabetic retinopathy: different statistical directions

Atanu Biswas
Indian Statistical Institute

Abstract: Several available approaches of analyzing bivariate/multivariate ordinal categorical data are discussed in the present paper using the example of an ophthalmologic data. Wisconsin epidemiologic study of diabetic retinopathy (WESDR) is a population-based epidemiologic study carried out in Southern Wisconsin during the eighties of the last century. The resulting data were analyzed by different statisticians and ophthalmologists during the last two decades. We can primarily divide the statistical analysis in three categories. Some ophthalmologists and statisticians involved in the trial analyzed the data to understand the epidemiologic behavior. Some statisticians analyzed the data in the context of smoothing spline. Some other statisticians considered the data as a nice example of bivariate ordinal dataset, due to the very nature of the data, and analyzed the data in that direction. There are, of course, different angles of study within a particular direction. The present paper reviews the WESDR study and the resulting data. Different directions of studies are then described with due references.

Key words: Bivariate ordinal categorical data; contingency table; generalized estimating equations; Gibbs sampler; global odds ratio; latent variable; Markov chain Monte Carlo; multivariate Placket distribution; Newton-Raphson method; ophthalmologic study; smoothing spline ANOVA, truncated bivariate ordinal data.

1 Introduction

In different studies related to medical and social sciences, bivariate or multivariate ordinal data is a common outcome. Here the response of each component is measured in an ordinal scale, for example, mild, moderate, severe, etc. (see Ashford, 1959, Cox, 1970, McCullagh, 1980, and Snell, 1964). Eversince Dale (1986) proposed the analysis of bivariate ordinal categorical data, a considerable studies have been made in this fascinating research area in statistics. The primary concern is to develop a flexible model which describes the relationship between bivariate/multivariate ordered categorical responses and the various available covariates.

Historically such a study was important to psychometricians. For example, Arminger and Kusters (1988) assumed that each observed outcome is a manifestation of an underlying continuous variable that is linearly related to a normal latent trait. Each type of outcome is related to its own latent trait, and the set of latent traits is assumed to be multivariate normal with expectation potentially dependent on covariates. In different situations dealing with pain, tenderness, post-operative conditions, (multivariate) ordinal data structure are quite common. Multivariate ordinal categorical data is a natural outcome in many biomedical studies. Pre-and-post-operative conditions can often be modeled as bivariate ordinal categorical scale. In several arthritis studies, experimenters are often interested to study several aspects like pain, tenderness, swelling, joint stiffness, etc. (see Biswas and Dewanji, 2004), which are often measured in ordinal categorical scale like nil, mild, moderate, severe, etc. Perhaps the most natural example of bivariate ordinal data structure is the retinopathy levels of two eyes. The technique of analysis for such multivariate ordinal categorical data is almost same in any such case. We discuss the several available techniques by means of an eye data. In the present paper we discuss such a retinopathy study and provide different analysis of the resulting data.

The rest of the paper is organized as follows. In Section 2, we discuss the Wisconsin epidemiologic study of diabetic retinopathy. In the present paper we discuss several available analysis of this study. It has a two-fold goal. Firstly, to enlist different approaches of study of the resulting data of this study. Secondly, to illustrate the technique of data analysis of bivariate ordinal data, in general. In Section 3 a description of the data is given. Sections 4 - 7 illustrate the analysis of this data in different directions. Section 4 describes the immediate and straightforward analysis by the group who collected the data. In Section 5, some smoothing spline approach is discussed. Section 6 covers the analysis where the data is treated as a bivariate ordinal data only, and the global odds ratio and the latent variable based approaches are mostly considered. Section 7 provides a comparison of the results available by bivariate and some (transformed) univariate data analysis. Section 8 gives a tabular comparison of different available methods in literature. Finally, Section 9 ends with some concluding remarks.

2 Wisconsin epidemiologic study of diabetic retinopathy

One much used and much cited epidemiologic study in the statistical literature over the last two decades is the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR). This is, in fact, one real life epidemiologic study which drew attention of the statisticians so much that we feel that they should be documented under a single heading. In a population-based study in Southern Wisconsin between 1980 and 1982, a total of 996 insulin-taking, younger onset diabetic persons were examined using standard protocols to determine the prevalence and severity of diabetic retinopathy and associated risk variables. The population of the study consisted of a probability sample selected from 10135 diabetic persons

who received primary care in an 11-county area in southern Wisconsin from 1979 to 1980. A detailed description of the population is given by Klein et al. (1983). Of the younger-onset persons (less than 30 years of age), 996 participated in the baseline examination (1980 to 1982). The baseline and follow-up examinations (after 4 and 10 years) were performed in a mobile examination van in or near the city where the participants lived. The ocular and physical examinations included taking stereoscopic color fundus photographs of seven standard fields.

The basic goal of the study (Klein et al., 1984a) was to find the associated risk factors which are important in planning a well-coordinated approach to the public health problem posed by the complications of diabetes (Hamman, 1982, Rand, 1981). Identifying the patients who may be at high risk of severe retinopathy is important in advising ophthalmologic care. Such data and the related analysis are also helpful in planning future studies such as controlled clinical trials of treatment of diabetes and diabetic retinopathy (Rand, 1981, Palmberg et al., 1981). Thus the objectives of the WESDR were two-fold (Klein et al., 1984a): (1) to describe the prevalence and severity of diabetic retinopathy and its component lesions, and to determine the frequency of visual impairment in a total population of diabetic patients who were under physicians' care in a defined geographic area; and (2) to determine the relationships between risk factors, prevalence, and severity of diabetic retinopathy of these patients.

There were 4-year and 10-year follow-up examinations. Four-year follow-up (1984 to 1986) examinations were performed in a mobile examination van in or near the city where the participants resided (Klein et al., 1989a). A similar 10-year follow-up examination was also conducted (Klein et al., 1994a).

Note that taking the older onset persons into two groups - older onset taking insulin and older onset not taking insulin - data from these groups were also analyzed and the results were reported in Wang (1994). See also Wang et al. (1995) in this connection. But this is not in the purview of our present study. We focus our attention to the younger onset data only.

3 Description of the data

The WESDR dataset includes the retinopathy levels and several associated covariate information on 996 individuals. Of course, in the dataset, some components of the covariates corresponding to some individuals are missing.

Both the right eye and the left eye retinopathy severity levels are recorded as two components of the bivariate response. Possible values are 10, 21, 31, 37, 43, 47, 53, 60, 61, 65, 71, 75 and 85 corresponding to increasing levels of severity of retinopathy within an eye. A commonly used grouping is:

- 10: no retinopathy,
- 21-37: mild nonproliferative retinopathy,
- 43-53: moderate to severe nonproliferative retinopathy, and
- 60-85: proliferative retinopathy.

These groupings are usually done to analyze the present data and most of the retinopathy data.

Three eye-specific covariates are recorded. These are

- right and left eye macular edema (ME) (present/absent),
- right and left eye refractive error (RE) in diopters, the values can be negative or positive, negative values represent myopia (nearsightedness), and positive values represent hyperopia (farsightedness),
- right and left eye intraocular pressure (IOP) in mmHg.

In addition 11 person-specific covariates are recorded. The first two are age at diagnosis (AgD) of diabetes in years and duration of diabetes (DuD) in years. To get current age at the time of the examination, one has to add age at diagnosis and duration of diabetes. The third one is glycosylated hemoglobin (GH) in percent. This is a measure of control of blood sugar. Lower values are considered better. Systolic and diastolic blood pressures (SBP & DBP) are measured in mmHg. Body mass index (BMI) in kilograms per meter squared, using weight and height, and pulse rate (PR) in beats per 30 seconds are also observed. Further sex, urine protein (UP) (present/absent), doses of insulin (DI) per day and area of residence (AR) (urban/rural) are recorded in addition.

The presence of both person-specific and eye-specific covariates extends the scope of the applicability of the present discussion to the more general situation that might possibly occur in other biomedical applications.

4 Analysis of WESDR data

Relations between various covariates and the retinopathy scores have been extensively analyzed by standard statistical methods including categorical data analysis and parametric GLIM models, and the results have been reported in the various WESDR manuscripts. See Klein, Klein, Moss, Davis and DeMets (1984a, 1984b, 1989a, 1989b), Klein, Klein, Moss, DeMets, Kauffman and Voss (1984) and Klein, Klein, Moss and Cruickshanks (1994b). Note that not much sophisticated statistical techniques were used in these analysis. Here we highlight some of the findings.

- Women had a higher frequency of retinopathy than did men. But men had higher prevalences of the more severe levels of proliferative diabetic retinopathy. Thus sex is an important covariate.
- Retinopathy increases with the increase in duration of diabetes. Prevalence rose sharply from 2% in those with diabetes of less than two years' duration to 97.5% in those with diabetes of 15 or more years of duration. Thus duration of diabetes is also an important covariate in this context.
- Retinopathy of moderate severity or worse increases with age.

- Significant association (with a P-value < 0.05) were found between presence of retinopathy and age at diagnosis, age at examination, glycosylated hemoglobin level, diastolic blood pressure and body mass index.
- No statistically significant association was found between the presence of retinopathy and systolic blood pressure, urine protein, pulse rate, family history of diabetes, insulin type, number of daily insulin doses, or amount of daily insulin.

5 Smoothing spline approach

Smoothing spline ANOVA (SS-ANOVA) models are endowed with some useful features like adaptively controlling the complexity or degrees of freedom of the model (sometimes called the bias-variance tradeoff) and for comparing different candidate models in the same or related families of models. These models represent a function $f(t)$, $t = (t_1, \ldots, t_d)$, of d variables as

$$f(t) = C + \sum_{\alpha} f_\alpha(t_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots ,$$

where the components satisfy side conditions which generalize the usual side conditions for parametric ANOVA to function spaces, and the series is truncated in some manner. Independent observations y_i, $i = 1, \ldots, n$, are assumed to be distributed as $h(y_i, f(t(i)))$ with parameter of interest $f(t(i))$ where $t(i)$ is the value of t for the subject i, and $f(\cdot)$ is assumed to be "smooth" in some sense; f is estimated as the minimizer, in an appropriate function space, of

$$L(y, f) + \sum_{\alpha} \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots ,$$

where $L(y, f)$ is the negative log likelihood of $(y_1, \ldots, y_n)$ given f, the $J_\alpha, J_{\alpha,\beta}, \ldots$ are quadratic penalty functionals and the λ_α, $\lambda_{\alpha\beta}$ are smoothing parameters to be chosen.

Wahba et al. (1995) worked on $d > 1$ case for exponential families and demonstrated its usefulness by analyzing data from the WESDR. They built an SS-ANOVA model to estimate the risk of progression of diabetic retinopathy, an important cause of blindness, at follow-up, given values of the predictor variables GH, AgD, DuD and BMI at baseline, at baseline, and the response (progression of retinopathy or not) at follow-up.

Wahba et al. (1995) carried out their analysis on a subgroup of the younger onset population, consisting of 669 subjects with no or nonproliferative retinopathy. In several works in this connection, Klein, Klein, Moss, Davis and DeMets reported that GH is a strong predictor of progression of diabetic retinopathy in the younger onset group. Wahba et al. (1995) observed that DuD has nonlinear effect on the probability of progression. Four individual univariate spline fits for risk of progression as functions, respectively, of GH, AgD, DuD and BMI suggested that

the effect of GH was very strong and fairly linear in the logit and that the effects of AgD, DuD and BMI were strong and nonlinear. Some exploratory GLIM modeling using the SAS procedure LOGISTIC [SAS Institute (1989)] suggested (AgD, BMI) and/or (DuD,BMI) interactions might be present. After some exploratory considerations they took the model as

$$f(\text{AgD}, \text{DuD}, \text{GH}, \text{BMI}) = f_1(\text{AgD}) + f_2(\text{DuD}) + f_3(\text{GH}) + f_4(\text{BMI}) + f_{14}(\text{AgD}, \text{BMI}).$$

Bayesian confidence intervals were also obtained. In their analysis, it was observed that

- Increases in GH at baseline are associated with increases in the risk of progression of diabetic retinopathy over the first four years of the study.
- The risk increases with increasing BMI at baseline until a value of about 25 kg/m^2 , after which there was flattening, except at the longer durations, where risk of progression continues to increase with BMI.
- This risk as a function of DuD at baseline increases up to a duration of about 10 years, and it then decreases.

See also Wahba et al. (1994) and Wang et al. (1997) in this connection.

6 WESDR data as bivariate ordinal data

6.1 Frequentist's view point

The WESDR and the resulting data drew attention of some statisticians, mainly due to its nature it is a large dataset involving bivariate ordinal responses with several covariates. Ordered structures become complex when a bivariate response is observed, the categories for each margin being ordinal. Dale (1986) has done the pioneering work on the analysis of bivariate ordinal categorical data. He expressed the joint probabilities of the bivariate responses in terms of marginal cumulative probabilities and the global odds ratios, and used the multinomial cell count based likelihood approach to estimate the association as well as the regression parameters involved in the model. Molenberghs and Lessafre (1994) used a multivariate Placket distribution for the extension of Dale model. The first analysis of WESDR data in this direction was due to Williamson, Kim and Lipsitz (1995) who provided their approach as an alternative to the computationally expensive likelihood methods of Dale (1986) and Molenberghs and Lessafre (1994). They defined the global odds ratio in terms of the marginal and the bivariate distribution function of two jointly distributed ordinal categorical random variables are defined. For detailed mathematical formulation one can see the paper by Williamson, Kim and Lipsitz (1995). The log of this global odds ratio is then expressed as a linear model involving the covariates. The generalized estimating equations approach is then carried out to get an estimate of the regression parameters. The numerical computations of Williamson, Kim and Lipsitz (1995) was done based on the data on 720

subjects with complete response and covariate data. From their computations, the person-specific covariates, found to be significantly associated with diabetic retinopathy, are DuD, GH, DBP, UP, Sex (male gender) and ME were found to be associated with worse diabetic retinopathy. No eye- specific covariates were examined in the global odds ratio model. The person-specific covariates found to be significantly related to the association between eyes are gender and doses of insulin per day.

Then some statisticians thought of applying the latent variable approach. The use of latent variable in the analysis of categorical data is quite old. Long back, in the context of logistic model for ordered categorical data, Snell (1964) used such a variable. Also McCullagh (1980) considered regression models for univariate categorical data by postulating some continuous latent variables. While analyzing the WESDR data, Kim (1995) extended the concept of a continuous latent variable to the bivariate set up. He considered the bivariate latent variables to follow a bivariate normal distribution. Denoting the four categories of retinopathy as 0, 1, 2 and 3, suppose y_{1i} and y_{2i} denote the bivariate ordered categorical responses for the i-th individual corresponding to right and left eye respectively. Here y_{1i} takes values 0, 1, 2, 3, and similarly y_{2i} takes values 0, 1, 2, 3. The latent variables y_{1i}^* and y_{2i}^* are postulated for the two eyes which are unobserved and continuous. We assume that

$$\begin{aligned} y_{1i} = j \text{ if } \gamma_{1j} < y_{1i}^* \le \gamma_{1,j+1}, \quad j = 0, 1, 2, 3, \\ y_{2i} = l \text{ if } \gamma_{2j} < y_{2i}^* \le \gamma_{2,j+1}, \quad l = 0, 1, 2, 3, \end{aligned} \tag{6.1}$$

where the two sets of cut-off (break) points $\gamma_1 = (\gamma_{10}, \gamma_{11}, \ldots, \gamma_{14})$ and $\gamma_2 = (\gamma_{20}, \gamma_{21}, \ldots, \gamma_{24})$ are unknown. To avoid complications one can take $\gamma_{10} = \gamma_{20} = -\infty$ and $\gamma_{14} = \gamma_{24} = \infty$. Due to the symmetry of the problem (as it is dealing with two eyes), Kim considered $\gamma_1 = \gamma_2$. We assume that such a latent variable actually exists, and in practice, it is not possible to observe this. We can only observe whether the value belongs to a particular interval or not. Thus the pair of responses (y_{1i}, y_{2i}) can take the values (j, l) where (j, l) can take any of the 16 possible pairs (0,0), (0,1), ..., (3,3). The cut-off points $\gamma_1 (= \gamma_2)$ and the regression vector corresponding to the person-specific and eye- specific covariates were obtained by maximizing the log-likelihood function using Newton-Raphson iteration technique. Kim (1995) also analyzed the 720 complete observations. The estimates of the cut-off points and the regression parameters were obtained according to the latent variable scale. In the analysis

- DuD, GH, DBP, UP, Sex and ME were identified as the important covariates.
- Dependency between the left and right eyes is rather strong as evidenced by the estimate of the correlation coefficient which is 0.922.

The maximization of the log-likelihood function and the evaluation of the negative hessian matrix in Kim's (1995) method in quite computationally intensive. Kim used the Newton-Raphson method for this purpose. As the Newton-Raphson method for function maximization in multi-dimensional space is known to be very sensitive to the choice of initial guesses, for the initial guesses of the cut-off points,

Kim used the average of the initial estimates of the cut-off points from the marginal distributions.

Then Williamson and Kim (1996) came up with a global odds ratio regression model using bivariate latent variables. For a covariate vector x_i, the (j,l)-th cell probability of the contingency table is

$$\pi_{jl}(x_i) = P(y_{1i} = j, y_{2i} = l) = P(\gamma_{1j} < y^*_{1i} \leq \gamma_{1,j+1}, \gamma_{2l} < y^*_{2i} \leq \gamma_{2,l+1}),$$

which can be written as

$$F(\gamma_{1,j+1}, \gamma_{2,l+1}) - F(\gamma_{1,j+1}, \gamma_{2l}) - F(\gamma_{1j}, \gamma_{2,l+1}) + F(\gamma_{1j}, \gamma_{2l})$$

or as

$$F_{j+1,l+1}(x_i) - F_{j+1,l}(x_i) - F_{j,l+1}(x_i) + F_{jl}(x_i).$$

They interpreted each person's response as represented in a 4×4 contingency table of sample size one. As there is no reason to assume different marginal distributions for the left and right eyes, they denoted the cumulative marginal probabilities as

$$\eta_{j+1}(x_i) = P(y_{ti} \leq j), \quad t = 1, 2.$$

Then the global odds ratio is represented as

$$\psi_{jl}(x_i) = \frac{F_{jl}(x_i)\{1 - \eta_j(x_i) - \eta_l(x_i) + F_{jl}(x_i)\}}{\{\eta_l(x_i) - F_{jl}(x_i)\}\{\eta_j(x_i) - F_{jl}(x_i)\}}.$$

Global odds ratio model is fitted with both the cumulative logit link function and the cumulative probit link function.

Kim (1995) used the bivariate normal distribution as the natural choice of the underlying latent distribution since its marginal and conditional distributions are also normal. He modelled the association between the responses of the two eyes with the bivariate normal correlation coefficient. The choice of the bivariate normal or any other specific bivariate continuous distribution as the underlying latent distribution, however, is a strong assumption, and the potential impact of assuming a specific underlying bivariate distribution has not been studied. Other than the normal distribution, there appears no natural family of bivariate distributions for consideration. Alternatively, with a global odds ratio model, one need not make a specific choice for the underlying latent distribution; one needs only to assume that the distribution is bivariate and continuous. Therefore, the global odds ratio regression model may have more flexibility in analyzing bivariate ordered categorical response data. Also, the bivariate cumulative probit regression model makes the implicit assumption that the association between eyes is correctly modelled with the correlation coefficient, whereas the global odds ratio regression model assumes no specific structure to the correlation. A disadvantage of the proposed global odds ratio model is that there is no straightforward estimate of the correlation between eyes.

Later Kim et al. (1996) discussed regression models for bivariate ordered categorical data from ophthalmologic studies. The paper by Kim et al. (1996) is

essentially a review of the three regression models for bivariate ordered categorical data from ophthalmologic studies as proposed in Kim (1995), Williamson and Kim (1996) and Williamson, Kim and Lipsitz (1995). Then Williamson et al. (1999) developed some relevant computer programs for the analysis of correlated categorical response data and they reported them in a paper. Earlier Williamson et al. (1995) presented two sets of generalized estimating equations for the analysis of repeated ordered categorical responses. The first set of estimating equation models the marginal distribution following Lipsitz et al. (1994), who extended the Liang and Zeger's (1986) generalized estimating equation (GEE) method. The second set of estimating equations models the joint association between responses using the global odds ratio as a measure of association. All these analysis were done with the first data of 1980-82.

In another effort of analyzing the WESDR dataset, Das and Sutradhar (2002) distinguished the ordinal categories in a general way so that the covariate effects are generally different under different ordinal categories. This allows to model the cumulative margins through certain non-linear regression functions which does not require any introduction of the cut-off points. The association between the bivariate responses are modeled by using the Pearson type correlation parameters. These consistent estimates of the correlation parameters are then used in the generalized estimating equations for the regression parameters to obtain their consistent estimates. It was observed that the probability that a female would belong to the 'no retinopathy' group is larger as compared to a male, and the probability that a female would belong to the highly severe group is smaller as compared to that of a male. It is also observed that those with higher doses of insulin per day have higher probabilities to have no severe effects on their right and left eyes respectively, whereas they have smaller chances to be in the worst group if high doses of insulin are taken per day.

6.2 Bayesian view point

A first Bayesian analysis of the WESDR data (of the first experiment of 1980-82) was carried out by Biswas and Das (2002). They considered the set up (6.1) and assumed (y_{1i}^*, y_{2i}^*) to follow a bivariate model

$$y_{1i}^* = x_{1i}'\beta_1 + \varepsilon_{1i},$$
$$y_{2i}^* = x_{2i}'\beta_2 + \varepsilon_{2i},$$

where x_{1i} and x_{2i} are the vector of covariates for the i-th individual, the disturbance vector $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})$ follows a bivariate normal distribution with mean vector 0 and dispersion matrix Σ, independently of each other. Here γ_{1j}'s and γ_{2l}'s are typically unknown cut-off points. Some mild conditions on these γ_{1j}'s and γ_{2l}'s, such as (i) $\gamma_{10} = \gamma_{20} = -\infty$, $\gamma_{14} = \gamma_{24} = \infty$, and (ii) $\gamma_{11} = \gamma_{12} =$ a known constant, are needed for identifiability (see Chen and Shao, 1999, for details). The analysis is carried out with data on 691 individuals only for which complete information were obtained. Suitable priors for β_1, β_2, Σ and γ_1, γ_2 are considered according to the prior belief of the experimenter. Posterior summary statistics are obtained by implementing the popular Gibbs sampler technique

(Geman and Geman, 1984, Gelfand et al., 1990) using the computer package WinBUGS for computation (see http://www.mrc-bsu.cam.ac.uk/bugs/ for details; see also Spiegelhalter et al., 1994). The analysis shows that the severity of retinopathy among the younger onset diabetic persons is directly affected by DuD and DBP. Also GH was one vital covariate for retinopathy. The estimate of the correlation coefficient came out as 0.8281 for normal prior and 0.8066 for noninformative prior for the vector of the regression parameters.

It is clear that a full understanding of epidemiology and progression of diabetic retinopathy in WESDR is only possible with the analysis of bivariate ordinal data collected longitudinally. Das and Biswas (2004) considered the analysis of bivariate ordinal data repeated over time. They took the WESDR data for two time points only (baseline and 4-year follow-up data). The analysis was carried out with data on 548 individuals for whom we have complete data at both the time points, as some of the individuals of the baseline study were missing in the follow-up study. A random effect model (in the frequentist's sense) is assumed as:

$$y_{it}^{*} = X_{it}\beta + Z_{it}b_i + \varepsilon_{it},$$

where the additional suffix t represents time (for baseline $t = 0$ and for 4-year follow-up $t = 1$). The i-th subject effect b_i is responsible for the longitudinal dependence. A Dirichlet Process (DP) prior for the unknown distribution of b_i's are taken (see Ferguson, 1973, Kleinman and Ibrahim, 1998). Markov Chain Monte Carlo (MCMC) computations were carried out. GH, DBP and DuD came out as significant covariates. The estimate of the correlation was observed to be 0.8378.

Note that, Bayesian analysis of such kind of data is doable using the MCMC approach and also the available softwares. The Bayesian computations need several points to note. For some of the parameters the Metropolis-Hastings algorithm (Hastings, 1970) is needed to be carried out. Again, prior elicitation is a delicate question, convergence has to be assessed. Standard approaches can be carried out for this (Gelman and Rubin, 1992). Qiu et al. (2002) and Biswas (2002) discussed several theoretical and computational issues in connection of Bayesian analysis of multivariate categorical data.

7 Bivariate versus univariate ordinal categorical data

Klein et al. (1984b, 1989b) analyzed the WESDR data separately for left and right eyes. They observed significant effects of DuD, AgD, GH, Sex, DBP and BMI on the retinopathy. Thus the nature of results in the univariate and bivariate cases are almost same. But, quite naturally, the polychoric correlation between the two eyes cannot be estimated by marginal analysis.

The very high correlation estimates in the bivariate analysis motivate us to consider some univariate transformation of the data. As one can easily understand that an analysis with bivariate ordinal categorical dataset is quite complicated

and computationally expensive too. Moreover, one has to take care of the polychoric correlation. The question naturally arises is: *Can we get the same kind of results with a proper univariate transform of the data?* Biswas and Angers (2003) carried out a study to see whether one can suitably transform the bivariate ordinal categorical data in the sense that we can clearly catch the major features of the data in such univariate version of the data, whose analysis is quite simpler.

The first hardle is, of course, to make a suitable transformation maintaining the ordinal nature and flavor of the data. An univariate transformation of the bivariate data with 13 ordered categories is done by the experimenters themselves. This is a retinopathy scale (RS) in which the retinopathy levels of both the eyes are concatenated into a person-level scale. The RS used in the present discussion is a more current one than the one used in some earlier works by different authors. In finding RS, the worse eye is given greater weight. The fellow eye has either the same level or a lower level. All levels of proliferative retinopathy (60-85) are grouped together. This results in a 15-level scale: 10/10, 21/<21, 21/21, 31/<31, 31/31, 37/<37, 37/37, 43/<43, 43/43, 47/<47, 47/47, 53/<53, 53/53, 60+/<60+, 60+/60+, which are numbered 0 through 14. The ordinal nature is maintained by simply putting arbitrarily larger weight to the worse eye (see Klein et al., 1984).

Biswas and Angers (2004) used latent variable-type modeling and carried out Bayesian analysis using both (i) the original bivariate data and (ii) the reduced concatenated data. The analysis were done using a normal error model for the latent score vector and normal prior $N_p\left(\theta_0, \sigma^2 V/\kappa\right)$ for the parameters under consideration, where the hyperparameters θ_0 and κ ($\leq$ 1) are assumed to be known and σ^2 is unknown. An inverted gamma prior for σ^2 is considered. The posterior mean of different parameters with $\kappa = 1$ are given in Table 1, both for bivariate and univariate cases. Although the estimates are often in the same direction in terms of signs, the estimates vary significantly in magnitude. Thus the results differ remarkably in terms of magnitude if we ignore the bivariate data sacrificing some information.

Table 1 *Comparison of bivariate and univariate posterior means.*

Parameters	Bivariate	Univariate	Parameters	Bivariate	Univariate
Right ME	11.777	1.709	DuD	0.156	0.032
Right RE	-0.172	0.036	GH	0.942	0.220
Right IOP	-0.006	-0.080	SBP	0.006	0.004
Left ME	13.399	3.478	DBP	0.123	0.032
Left RE	0.126	-0.048	BMI	0.340	0.072
Left IOP	-0.013	0.081	PR	0.061	0.017
UP	-0.268	-0.088	DI	-0.004	-0.097

8 Comparison of results

In this section, we list the nature of results using different approaches in Table 2. We only list the significant covariates and the direction of their effects (positive/negative). The exact numerical figures of the estimates are not reported as the numerical figures are not comparable due to different type of modelling, number of parameters in the models, scales, etc.

9 Concluding remarks

In some situations, some of the cells of the two-way ordered classification may remain empty. The analysis is much complicated in such a situation. Weiss (1993) observed such a situation in a dataset on head/neck injury and body injury in motorcycle accidents in Los Angeles. In a frequentist's approach, Weiss (1993) observed that the log-likelihood function is not globally concave resulting serious difficulty in estimation. Das and Biswas (2000) carried out a Bayesian analysis in such a situation and observed that the Bayesian solution does not have the drawbacks of the frequentist's approach.

Although a lot of statistical analysis have been done on the WESDR data, it is not yet complete. In this present article we have paid attention only on the younger onset data part. Note that most of the analysis were done on the baseline data, only a few were done on the 4-year and 10-year progression data to analyze the progression structure of retinopathy. Thus, a lot more analysis combining the three datasets are pertinent to obtain a clear picture of the underlying story of the WESDR datasets. Several detailed study on the older onset data is also pertinent, some of which are done by Wang and others (1994, 1995).

In a recent study on model selection, we used some exploratory GLIM modeling using the SAS procedure LOGISTIC. We observed that DuD, GH, DBP and the interaction of GH and DBP have significant effect on the retinopathy levels. In addition, there may be several unobserved causes which might have significant effect on the responses. For example, as the two eyes of the same human being are considered, several nerves, tissues, same reading habit and work habit, same environmental exposure, etc. are responsible for similar effect on both the eyes. These unobserved factors can be better modeled in terms of random effect. Some works are done, but a lot of future study is pertinent.

As this was a large scale epidemiologic study, the resulting information on the effect of covariates and the disease progression nature can be used for other populations, in general. With such a large dataset, the WESDR data is likely to be an excellent example of bivariate (repeated) ordinal data with several covariates, some of which are time-dependent. The present paper focuses the need for a proper bivariate/multivariate categorical distribution and modeling which will incorporate different logistics like the general covariate structure. This can be done in a particular problem by careful comparison of different available approaches.

Table 2 *Different available methods for the WESDR dataset.*

Methods by	F/B	Data type	Method/Model	Significant covariates	Estimate of correlation
Klein et al. (1984b)	F	Univ.	Student's t, χ^2, Cox regression	DuD(+), GH(+), DBP(+), AgD(+), UP(+), Sex(-)	
Klein et al. (1989b)	F	Univ.	Student's t, χ^2	DuD(+), AgD (+), Sex(-)	
Wahba et al. (1995)	F	Univ.	Smoothing spline ANOVA, GLIM	GH(+), DuD(+), AgD(+), BMI(+)	
Williamson et al. (1995)	F	Biv.	Global odds ratio, GEE	DuD(+), GH (+), DBP(+), UP(+), Sex(-), ME(+)	not obtained
Kim (1995)	F	Biv.	Bivariate cumulative Probit regression, MLE	DuD(+), GH(+), DBP(+), UP(+), Sex(-), ME(+)	0.922
Williamson and Kim (1996)	F	Biv.	Global odds ratio, regression, bivariate latent variables, cumulative logit/ probit link	DuD(+), GH(+), DBP (+), UP(+), Sex(-), ME(+)	not obtained
Sutradhar and Das (2001)	F	Biv.	Pearson type correlation, GEE	DI(+), Sex(-)	not obtained
Biswas and Das (2002)	B	Biv.	Bivariate latent variables, MCMC	DuD(+), DBP (+), GH(+)	0.8281* 0.8066**
Das and Biswas (2004)	B	Biv.	Bivariate latent variables, Dirichlet process prior, MCMC	DuD(+), DBP (+), GH(+)	0.8378
Biswas and Angers (2004)	B	Univ. &Biv.	Bivariate latent variable type distribution	DBP(+), GH(+)	not obtained

+ (-) indicates higher value will result higher (lower) retinopathy levels, i.e. worse (better) condition.
F: Frequentist approach. B: Bayesian approach.
* using normal prior for regression parameters.
** using noninformative prior for regression parameters.

Acknowledgements

The author wishes to thank two referees for their careful reading and valuable comments which led some improvement over an earlier version of the paper.

(Received May, 2003. Accepted February, 2004.)

References

Arminger, G. and Kusters, U. (1988). Latent trait models with indicators of mixed measurement level. In Langeheine, R. and Rost, J. eds. *Latent Trait and Latent Class Models*. New York: Plenum, 51–53.

Ashford, J. R. (1959). An approach to the analysis of data for semiquantal responses in biological assay. *Biometrics*, **15**, 573–581.

Biswas, A. (2002). Theoretical and computational issues in the Bayesian analysis of multivariate ordinal data. In *Recent Advances in Statistical Methods* (Proceedings of Statistics 2001 Canada conference). Ed. Chaubey, Y. P. World Scientific Publishing Inc., London. To appear.

Biswas, A. and Angers, J.-F. (2004). Analyzing a bivariate ordinal data set on diabetic retinopathy. *Craiova Medicala*, **5**, 290–291. (Proceedings of the 1st MEDINF International Conference on Medical Informatics & Engineering "MEDINF 2003").

Biswas, A. and Das, K. (2002). A Bayesian analysis of bivariate ordinal data: Wisconsin epidemiologic study of diabetic retinopathy revisited. *Statistics in Medicine*, **21**, 549–559.

Biswas, A. and Dewanji, A. (2004). A randomized longitudinal play-the-winner design for repeated binary data. *Australian and New Zealand Journal of Statistics*, **46**, 675–684.

Chen, M.-H. and Shao, Q.-M. (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *Journal of Multivariate Analysis*, **71**, 277–296.

Cox, D.R. (1970). *The Analysis of Binary Data*. London: Chapman and Hall.

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Das, K. and Biswas, A. (2000). *Bayesian analysis of truncated bivariate ordinal data in regression set up*. Technical Report. Applied Statistics Division, Indian Statistical Institute.

Das, K. and Biswas, A. (2004). *Dirichlet process mixed model for bivariate ordered categorical data with application to the Wisconsin epidemiologic study of diabetic retinopathy.* Under revision.

Das, K. and Sutradhar, B. C. (2002). Analyzing bivariate ordinal polytomous data: a marginal multinomial logistic approach. Proceedings of the Fourth International Triennial Calcutta Symposium on Probability and Statistics (2000). *Calcutta Statistical Association Bulletin*, **52**, (2002), no. 205–208, 181–204.

Ferguson, T. S. (1973). A Bayesian analysis of somenon-parametric problems. *Ann. Statist.*, **1**, 209–230.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–985.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Hamman, R.F. (1982). Data assessment and problem identification: reviewing the experience. In *Proceedings of the Diabetes Control Conference.* Atlanta, Centers of Disease Control, 32–40.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine*, **14**, 1341–1352.

Kim, K., Lipsitz, S. R. and Williamson, J. M. (1996). Regression models for bivariate ordered categorical data from ophthalmological studies. In Koo JY, Park BU, Lee KW, Jeon JW (eds). *Collected Papers in Honor of Retirement of Professor Chung Han Yung.* Seoul National University, Department of Computer Science and Statistics Alumni Association, 36–55.

Klein, B. E. K., Davis, M. D., Segal, P., Long, J. A., Harris, W. A., Haug, G. A., Magli, Y. and Syrjala, S. (1984). Diabetic retinopathy: assessment of severity and progression. *Ophthalmology*, **91**, 10–17.

Klein, R., Klein, B. E. K. and Davis, M. D. (1983). Is cigarette smoking associated with diabetic retinopathy? *Am. J. Epidemiol.*, **118**, 228–238.

Klein, R., Klein, B. E. K., Moss, S. E. and Cruickshanks, K. J. (1994a). The Wisconsin epidemiologic study of diabetic retinopathy XIV. Ten-year incidence and progression of diabetic retinopathy. *Arch. Ophthalmol.*, **112**, 1217–1228.

Klein, R., Klein, B. E. K., Moss, S. E. and Cruickshanks, K. J. (1994b). The relationship of hyperglycemia to long-term incidence and progression of diabetic retinopathy. *Archives of Internal Medicine*, **154**, 2169–2178.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984a). The Wisconsin Epidemiologic study of diabetic retinopathy, II: prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.*, **102**, 520–526.

Klein, R., Klein, B.E.K., Moss, S.E., Davis, M.D. and DeMets, D.L. (1984b). The Wisconsin Epidemiologic study of diabetic retinopathy, II: prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Archives of Ophthalmology*, **102**, 527–532.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1989a). The Wisconsin epidemiologic study of diabetic retinopathy IX. Four-year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.*, **107**, 237–243.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1989b). Is blood pressure a predictor of the incidence and progression of diabetic retinopathy? *Archives of Internal Medicine*, **149**, 2427–2432.

Klein, R., Klein, B. E. K., Moss, S. E., DeMets, D. L., Kaufman, I. and Voss, P. S. (1984). Prevalence of diabetes mellitus in southern Wisconsin. *Am. J. Epidemiol.*, **119**, 54–61.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effect model. *Biometrics*, **54**, 921–938.

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

Lipsitz, S. R., Kim, K. and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**, 1149–1163.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, B*, **42**, 109–142.

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

Palmberg, P., Smith, M., Waltman, S., et al. (1981). The natural history of retinopathy in insulin-dependent juvenile-onset diabetes. *Ophthalmology*, **88**, 613–618.

Qiu, Z., Song, P. and Tan (2002). Bayesian hierarchical analysis of multi-level repeated ordinal data using WinBUGS. *Journal of Biopharmaceutical Statistics*, **12**, 121–135.

Rand, L. I. (1981). Recent advances in diabetic retinopathy. *Am. J. Med.*, **70**, 595–602.

SAS Institute. (1989). *SAS/STAT User's Guide*, Version 6, 4th ed. SAS Institute, Inc. Cary, North Carolina.

Snell, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics*, **20**, 592–607.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1994). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.30.* Techmical report, University of Cambridge, MRC Biostatistics Unit.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1994). Structured machine learning for "soft" classification with smoothing spline ANOVA and stacked tuning, testing and evaluation. In Cowan J, Tesauro G, Alspector J, eds. *Advances in Neural Information Processing Systems 6.* San Francisco, CA: Morgan Kaufmann Publishers.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.*, **23**, 1865–1895.

Wang, Y (1994). *Smoothing spline analysis of variance of data from exponential families.* Ph. D. dissertation, Technical Report 928, Univ. Wisconsin-Madison.

Wang, Y., Wahba, G., Gu, C., Klein, R. and Klein. B. E. K. (1995). *Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy.* Technical Report 956, Univ. Wisconsin-Madison.

Wang, Y., Wahba, G., Gu, C., Klein, R. and Klein, B. (1997). Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. *Statistics in Medicine*, **16**, 1357–1376.

Weiss, A. A. (1993). A bivariate ordered probit model with truncation: helmet use and motorcycle injuries. *Applied Statistics*, **42**, 487–499.

Williamson, J. and Kim, K. (1996). A global odds ratio regression model for bivariate ordered categorical data from ophthalmologic studies. *Statistics in Medicine*, **15**, 1507–1518.

Williamson, J. M., Kim, K. and Lipsitz, S. R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association*, **90**, 1432–1437.

Williamson, J., Lipsitz, S. R. and Kim, K. (1999). GEECAT and GEEGOR: computer programs for the analysis of correlated categorical response data. *Computer Methods and Programs in Biomedicine*, **58**, 25–34.

Atanu Biswas
Applied Statistics Unit,
Indian Statistical Institute,
203 B. T. Road, Kolkata 700 108, India.
E-mail: atanu@isical.ac.in