A Bayesian analysis of the four-year follow-up data of the Wisconsin epidemiologic study of diabetic retinopathy*

Jean-François Angers† Atanu Biswas‡

CRM-2757

September 2001

*The work is partially supported by a grant from NSERC. A part of the work of the second author was carried out when he was visiting the Département de mathématiques et de statistique, Université de Montréal. The second author thanks the department for its hospitality. The authors wish to thank Drs. Ronald Klein and Barbara E. K. Klein of the University of Wisconsin for providing both the baseline and 4-year follow-up data of the WESDR study. The WESDR project was originally supported in part by grant EY 03083 (R. Klein) from the National Eye Institute, NIH.

†Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, Succ. "Centre-ville", Montréal, Québec H3C 3J7, angers@dms.umontreal.ca

‡Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta – 700 035, India, atanu@isical.ac.in

Abstract

The 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. Most of the analyses were carried out on the baseline data, although there were two follow-up studies on the same population. In this present paper we provide a Bayesian analysis of the first follow-up data which were taken four years after the baseline study. Our Bayesian analysis provides estimates of the associated covariate effects. Choice of the best model in terms of the covariate inclusion is also done. The baseline data was used to set the prior for the parameters. Extensive numerical computations illustrate our present methodology.

Some Key Words and Phrases: bivariate ordinal data; latent variable; global odds ratio; bias-variance tradeoff; sensitivity analysis; Bayesian model selection.

AMS 2000 subject classification: Primary 62J12; Secondary 62F15, 62-07.

1 Introduction

In different studies related to biomedical and social sciences, bivariate or multivariate ordinal data is a common outcome. Ordinal scales for measurement are often used in the absence of well defined non-invasive direct measurements. The response of each component is measured in an ordinal scale, for example, mild, moderate, severe, etc. (*cf.* Ashford [1]; Cox [2]; Macullagh [3] and Snell [4]). Eversince Dale [5] proposed the analysis of bivariate ordinal categorical data, a considerable studies have been done in this fascinating research area in statistics to develop a flexible model which describes the relationship between bivariate ordered categorical responses and the various available covariates.

Historically such a study was important to psychometricians. For example, Arminger and Kusters [6] assumed that each observed ordinal categorical outcome is a manifestation of an underlying continuous variable that is linearly related to a normal 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. Perhaps the most natural example of bivariate ordinal data structure is the retinopathy levels of two eyes. In the present paper we discuss such a much used and much cited retinopathy study where measurements on both eyes are in an ordered categorical fashion with some person specific (which are demographic, clinical and laboratory information) and some eye specific (which are clinical and laboratory information) covariates. The study is called the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR).

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.* [7]. Of the younger-onset persons (less than 30 years of age), 996 participated in the baseline examination (1980 to 1982). The baseline and the two 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 (*cf.* Klein *et al.* [8]) 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 (*cf.* Hamman [9]; Rand [10]). 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 (*cf.* Rand, [10]; Palmberg *et al.* [11]). There were 4-year and 10-year follow-up examinations (*cf.* Klein *et al.* [12-[13]). In this present paper, we analyze the 4-year follow-up data only. But we use the baseline data to fix the priors for different parameters in our Bayesian study. It is, of course, an interesting task to look at the 10-year follow-up data in this connection. But we could not access that raw data.

In the present paper our objective is to look into the 4-year follow-up data of the well known WESDR study and analyze the dataset in a Bayesian view point using the concept of underlying latent continuous variable, in some sense. In Section 2, we describe the nature of the data available from the WESDR experiment and provide a brief review on some of the past analyses of this dataset available in statistical and medical literature. In Section 3, we provide our model and the proposed Bayesian technique of analysis of such a data. In Section 4, the problem of model selection in terms of inclusion of covariates is discussed. In Section 5, the detailed computational results are given and discussed. The results are then compared with some of the earlier analyses. Finally Section 6 ends with a discussion.

2 A Review

In the first part of this section we discuss the WESDR data structure and in the subsequent part we make a brief review of the available literature.

The retinopathy scale (RS) provided in the dataset is a more current one than the one used in some earlier works. Both the right and 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, 85 corresponding to increasing levels of severity of retinopathy within an eye. A commonly used grouping is 10, 21-37, 43-53, and 60-85 which corresponds to no retinopathy, mild nonproliferative retinopathy, moderate to severe nonproliferative retinopathy, and proliferative retinopathy, respectively. Although such a grouping is used to reduce the computational burden, in this present paper we do not consider this grouping, instead we consider the original 13 ordered values.

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 8 person-specific covariates are recorded. The first one is the duration of diabetes (DuD) in years. The second one is glycosylated hemoglobin (GH) in percent, which 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 urine protein (UP) (present/absent), doses of insulin (DI) per day are also recorded.

A lot of statisticians looked at the analysis of the WESDR data, primarily because it is a nice real life dataset of bivariate ordinal structure with a number of covariates, on a large number of individuals. But, unfortunately most of the available analyses were done using the baseline data only, and only a few papers dealt with the follow-up datasets.

Some of the early analyses were done by Klein *et al.* [14] and by Klein *et al.* [15]. Klein *et al.* [15] studied the scenario with a concatenated score that incorporates data from the severity levels of both eyes into one score for a person. Williamson *et al.* [16] considered the generalized estimating equations approach as an alternative to computationally expensive likelihood methods of Dale [5] and Molenberghs and Lessafre [17]. They considered 720 subjects of the WESDR dataset with complete response and covariate data, and fitted cumulative probit margins and a global odds ratio association model. While analyzing the WESDR data, Kim [18] extended the concept of a continuous normally distributed latent variable to the bivariate set up. He obtained the maximum likelihood estimators of the underlying parameters including the typically unknown cut off points of the latent variables employing Newton-Raphson iteration method. Williamson and Kim [19] considered the bivariate latent variable regression model and modeled the dependency between the fellow eyes with the global odds ratio, where no specific choice of underlying latent distribution is needed except its continuity, and one assumes no specific structure of the correlation. Recently Williamson *et al.* [20] discussed the applicability of their computer program GEEGOR (generalized estimating equation using global odds ratio) through the WESDR dataset once again. Das and Sutradhar [21] developed an approach to model the association between the bivariate responses by a Pearson type correlation. Biswas and Das [22] considered a model using normally distributed latent variables similar to that of Kim [18], but the analysis has been carried out in the Bayesian paradigm. The merit of the approach of Biswas and Das [22] is that through the well known Gibbs sampler one may easily arrive at a consistent solution of the underlying regression parameter and may draw inference based on that. Note that all the above mentioned works were done using the baseline data only.

As mentioned earlier, not much work has been done with the follow-up data. Wahba and her colleagues ([23]-[31]) have done some works in this direction. Primarily they looked at the disease progression and the role of different covariates on it. 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. Wahba *et al.* [26] worked on 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, DuD, BMI and the age at diagnosis at baseline, and the response (progression of retinopathy or not) at follow-up. Then Wahba *et al.* [25] carried out their analysis on a subgroup of the younger onset population, consisting of 669 subjects with no or nonproliferative retinopathy. Some exploratory GLIM modeling using the SAS procedure LOGISTIC [SAS Institute [32]] were carried out and after some exploratory considerations they took the model. Bayesian confidence intervals were also obtained. See also Wahba *et al.* [26] and Wang *et al.* [27] in this connection.

3 Methodology

Let y_{Li} and y_{Ri} denote the bivariate ordered categorical responses for the ith individual corresponding to left and right eye respectively. Note that, $y_{Li}, y_{Ri} \in \{10, 21, 31, 37, 43, 47, 53, 60, 61, 65, 71, 75, 85\}$. Let y_L and y_R be the vectors combining y_{Li}'s and y_{Ri}'s for all the individuals. In order to assume normality of the error terms, we add z_{Li} and z_{Ri} to y_{Li} and y_{Ri}, where $(z_{Li}, z_{Ri})^T \sim N_2(0, \sigma_Z^2 I_\rho)$ with I_ρ is the 2×2 correlation matrix with unknown correlation ρ, and σ_z^2 is such that $y_{Li} \pm 3\sigma_z$ will not change categories. Let z_L (z_R) be the vector combining all the z_{Li}'s (z_{Ri}'s).

Let $u_L = y_L + z_L$ and $u_R = y_R + z_R$ and these u_L and u_R are the true values and we observe y_L and y_R in place of them. We model u_L and u_R as follows:

$$\begin{aligned} u_L &= X_0\beta_0 + X_1\beta_1 + \epsilon_L, \\ u_R &= X_0\beta_0 + X_2\beta_2 + \epsilon_R, \end{aligned}$$

where $\epsilon_L \sim N_n(0, \sigma^2 I)$ and $\epsilon_R \sim N_n(0, \sigma^2 I)$, independently of each other, if there are n individuals. Here β_0 is the vector of parameters associated with the covariates common to both u_L and u_R, and X_0 is the related design matrix;

β_1 is the covariate effects for the left eye only and X_1 is the associated design matrix; and β_2 and X_2 are the same for the right eye.

Write

$$u = \begin{pmatrix} u_L \\ u_R \end{pmatrix}, \quad X = \begin{pmatrix} X_0 & X_1 & 0 \\ X_0 & 0 & X_2 \end{pmatrix},$$
$$\theta = (\beta_0^T, \beta_1^T, \beta_2^T)^T, \quad \epsilon = (\epsilon_L^T, \epsilon_R^T)^T.$$

Then we represent

$$u = X\theta + \epsilon.$$

If the dependence on $z = (z_L^T, z_R^T)^T$ has to be made explicit, we will write

$$u(z) = \begin{pmatrix} y_L + z_L \\ y_R + z_R \end{pmatrix}.$$

Now we need to set some suitable prior for the parameters under consideration. Suppose θ is a p-component vector. We consider

$$\theta \sim N_p\left(\theta_0, \frac{\sigma^2}{\kappa} V\right), \tag{3.1}$$

where the hyperparameters θ_0 and κ ($\kappa \leq 1$) are assumed to be known and σ^2 is unknown. It is assumed that the prior of σ^2 is an inverted gamma with parameters $\alpha/2$ and $\gamma/2$. Note that $\gamma = 0$ would lead to standard noninformative prior on σ^2. The prior of ρ is chosen to be an uniform density on the interval $(-1, 1)$. We use the baseline data in this context to estimate θ_0. To denote the baseline data we just put "*" to y_L, y_R and X. Thus θ_0 is estimated using y_L^*, y_R^* and X^* as follows:

$$\widehat{\theta}_0 = (X^{*T} X^*)^{-1} X^{*T} y^*,$$

where $y^* = (y_L^{*T}, y_R^{*T})^T$. We also need to do a sensitivity analysis on κ (see Section 5).

Using standard technique, after some routine steps, it can be shown that the posterior of θ, σ^2 and ρ are

$$\begin{aligned} \theta|u, \sigma^2, \rho &\sim N_p((\kappa V^{-1} + X^T W^{-1} X)^{-1}(\kappa V^{-1}\theta_0 + X^T W^{-1} X \theta_{LS}(\rho)), \\ &\qquad \sigma^2(\kappa V^{-1} + X^T W^{-1} X)^{-1}), \end{aligned}$$

$$\begin{aligned} \sigma^2|\rho, u &\sim I|G\left(\frac{n-2}{2}, r^T W^{-1} r + (\theta_{LS}(\rho) - \theta_0)^T [\kappa^{-1} V + (X^T W^{-1} X)^{-1}]^{-1} \right. \\ &\qquad \left. \times (\theta_{LS}(\rho) - \theta_0)\right), \end{aligned}$$

$$\begin{aligned} \pi_3(\rho|u) &\propto \frac{1}{|D - \rho E + (1-\rho^2)\kappa V^{-1}|^{1/2}} \\ &\qquad \times \left(\frac{(1-\rho^2)}{r^T r - 2\rho t + (1-\rho^2)[r + h(\rho)]}\right)^{\frac{\alpha+n}{2}}, \end{aligned} \tag{3.2}$$

where

$$\begin{aligned} r &= u - X\theta_{LS}(\rho) = (r_L^T, r_R^T)^T, \\ \theta_{LS}(\rho) &= (X^T W^{-1} X)^{-1} X^T W^{-1} u, \\ W &= \begin{pmatrix} I & \rho I \\ \rho I & I \end{pmatrix}, \\ D &= 2X_0^T X_0 + X_1^T X_1 + X_2^T X_2, \\ E &= 2X_0^T X_0 + X_1^T X_2 + X_2^T X_1, \\ t &= r_L^T r_R, \\ h(\rho) &= (\theta_{LS}(\rho) - \theta_0)^T (\kappa^{-1} V + (1-\rho^2)(D - \rho E)^{-1})^{-1} (\theta_{LS}(\rho) - \theta_0). \end{aligned}$$

Hence, given a fixed value of ρ, we can estimate θ by

$$\widehat{\theta}(\rho) = (\kappa V^{-1} + X^T W^{-1} X)^{-1}(\kappa V^{-1}\theta_0 + X^T W^{-1} X \theta_{LS}(\rho)).$$

Note that $\widehat{\theta}(\rho)$ does not depend on σ^2. Thus, even if σ^2 is unknown, we will obtain the same estimator. However, $\widehat{\theta}(\rho)$ depends on ρ. Under the squared error loss, the estimator of θ is given by

$$\widehat{\theta} = E^{\pi_3(\rho|u)}[\widehat{\theta}(\rho)].$$

This expectation can be computed using Monte Carlo integration technique. Since u depend on z, which is random, we use EM algorithm (see Dempster *et al.* [33]) to find $\widehat{\theta}$ as

$$\widehat{\theta} = \theta_0 + E^{\pi_3(\rho|u)}[(\kappa V^{-1} + X^T W^{-1} X)^{-1} X^T W^{-1}](\bar{u} - X\theta_0), \tag{3.3}$$

where

$$\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u(z_i) = y + \frac{1}{m}\sum_{i=1} m z_i = y + \bar{z}.$$

4 Model Selection

In this section we carry out Bayesian model selection in the term of keeping only the relevant covariates in the model and analysis. Note that, although a lot of covariates were collected in the WESDR study, the inclusion of covariates in the study models in different studies were mostly done in an *ad hoc* manner without much statistical justification. Only Wahba [25] have done some studies for the model selection and variable inclusion. Here our object is to choose the "best" model given the data in hand.

Let $\theta = (\theta_{(1)}^T, \theta_{(2)}^T)^T$, and we are interested to test

$$H_0 : \theta_{(2)} = 0 \quad \text{against} \quad H_1 : \theta_{(2)} \neq 0,$$

to decide whether we would include the covariates corresponding to $\theta_{(2)}$ in the model or not. It is a well-known result that if

$$V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}\right),$$

with V_1 is of p_1 dimension, then

$$V_1 | V_2 \sim N_{p_1}(\mu_1 + BC^{-1}[V_2 - \mu_2], A - BC^{-1}B^T).$$

Hence, if the above H_0 is true, then

$$\theta_{LS}(\rho) = \begin{pmatrix} \theta_{LS(1)}(\rho) \\ \theta_{LS(2)}(\rho) \end{pmatrix} \sim N_p\left(\begin{pmatrix} \theta_{(1)} \\ \theta_{(2)} = 0 \end{pmatrix}, \sigma^2 (X^T W^{-1} X)^{-1}\right).$$

If we represent $(X^T W^{-1} X)^{-1}$ as

$$(X^T W^{-1} X)^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix},$$

we have

$$\theta_{LS(1)}(\rho) | \theta_{LS(2)}(\rho), \rho, \sigma^2 \sim N_{p_1}(\theta_{(1)} + BC^{-1}\theta_{LS(2)}(\rho), \sigma^2 F),$$

where $F = A - BC^{-1}B^T$. One need to take expectation with respect to σ^2 and ρ, and that can be tackled by Monte Carlo integration technique. Using standard technique, if $\theta_{(1)} \sim N_{p_1}\left(\theta_{0(1)}, \frac{\sigma^2}{\kappa} V_{(1)}\right)$, the distribution of $\theta_{LS(1)}$ given σ^2 and ρ is

$$\theta_{LS(1)}(\rho) | \sigma^2, \rho \sim N_{p_1}(\theta_{0(1)} + BC^{-1}\theta_{LS(2)}(\rho), \sigma^2 G),$$

where $G = F^{-1} - F^{-1}(\kappa V_{(1)}^{-1} + F^{-1})^{-1} F^{-1}$.

Let

$$m_0(u) = E^{\pi_3(\rho|u)}[m_{0,1}(\theta_{LS(1)}(\rho)|\theta_{LS(2)}(\rho)).m_{0,2}(\theta_{LS(2)}(\rho))]$$

be the marginal density of u under the null hypothesis. Then

$$\begin{aligned} m_0(u) \quad = \quad & E^{\pi_3(\rho|u)}\left[\frac{1}{(2\pi\sigma^2)^{p/2}|G|^{1/2}|C|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}[\theta_{LS(2)(\rho)}^T C^{-1} \theta_{LS(2)(\rho)}\right.\right. \\ & + (\theta_{LS(1)}(\rho) - \theta_{0(1)} - BC^{-1}\theta_{LS(2)}(\rho))^T G^{-1} \\ & \left.\left. \times (\theta_{LS(1)}(\rho) - \theta_{0(1)} - BC^{-1}\theta_{LS(2)}(\rho))]\right\}\right]. \end{aligned} \tag{4.1}$$

Write the marginal of u under the alternative hypothesis as

$$m_1(u) = E^{\pi_3(\rho|u)}\left[\frac{|X^TW^{-1}X|^{1/2}}{(2\pi\sigma^2)^{r/2}} \times \exp\left\{-\frac{1}{2\sigma^2}(\theta_{LS}(\rho)-\theta_0)^TX^TW^{-1}X(\theta_{LS}(\rho)-\theta_0)\right\}\right].$$

Then we accept H_0 if $m_0(u)/m_1(u) > 1$ and accept H_1 otherwise.

In our present work, we have tried several models starting from one component equal to zero to all but one equal to zero. To implement we ordered the standardized Bayesian estimates in the following way. Writing $Q_{(i)}$ as the ith diagonal element of the square matrix Q, we write

$$\mu_i = \frac{|\widehat{\theta}_i|}{E^{\pi_3(\rho|u)}[\sigma^2(\kappa V^{-1}+X^TW^{-1}X)^{-1}]_{(i,i)}},$$

and suppose $\mu_{(1)} \leq \mu_{(2)} \leq \cdots \leq \mu_{(p)}$ be the ordered arrangement. The different possible models are then $M_l : \mu_{(1)} = \cdots = \mu_{(l)} = 0$ for $l = 1, 2, \cdots, p-1$. If m_l denotes the marginal probability of M_l, and if $B_l = m_l/m_0$, with m_0 being the marginal probability of the full model, then we accept M_{l^*} as the correct model for our purpose if

$$B_{l^*} = \max_{0\leq l\leq p-1} B_l.$$

Note that here $B_0 = 1$.

5 Numerical Calculations

In this section, the data for the first follow-up are analyzed. The data set contained information described in Section 2 on 629 subjects.

It is to be noted that with the chosen *a priori* model, only numerical integration with respect to ρ is needed. In order to evaluate equation (3), the Monte Carlo method with importance sampling is used. The importance sampling function used to generate the ρ values is

$$g(\rho) \propto (1-\rho^2)^{(\alpha+n)/2},$$

and 5000 iterations were made. For each value of ρ, the $\overline{z}$ is computed using 1000 iterations. In practice, it is generated from $\overline{z} \sim N_2(0, \sigma_Z^2 I_\rho/1000)$.

In Table 1, we present the posteriors corresponding to the full model (*i.e.* the model with all the covariates) for several values of κ (*cf.* equation (1)). The posterior standard deviation are given in parenthesis. Note that RSRBR (RSRBL) correspond to the retinopathy scale of the right eye from the baseline study while RSLBR (RSLBL) correspond to the one of the left eye. These covariates were used in both X_1 and X_2 in order to measure their influence separately on the retinophaty scale of each eye. From Table 1, it can be seen that the coefficient of RSRBR (RSRBL) is similar to the one of RSLBL (RSLBR). Hence it can be concluded that the retinophaty scale of the right eye in baseline study has a similar effect of the retinopathy scale of the right eye in the current study than the RS of the left eye in both studies. It can also be seen from this table that the estimated values for the covariates are not influenced by the choice of κ. This is due to the fact that the large dataset is quite large. For the full model, the correlation coefficient between the observation on both eyes and the predicted values is 0.786 for all values of κ.

In Table 2, the estimated values of the covariates for the "best" model are given. The values of the correlation between the observations and the predicted values are also given. The "best" model is defined as the model maximizing the marginal probability of the observations. Since it is quite difficult to compare all models, the models tested are obtained in the following way:

1. Compute the posterior mean and variance for all covariate in the model;
2. Compute the ratio of the posterior mean over the posterior standard deviation;
3. Delete the covariate with the smallest ratio from the model and repeat steps 1 to 3.

Applying this algorithm we obtain the numerical figures in Table 2. From this table, it can be seen that the choice of the best model depends heavily on the choice of κ. However, the covariates RSRBL, RSLBL and GH are included in almost all model. Based on the results obtained from several models, we decided to fit our model with the covariates RSRBL, RSLBL, GH and a constant term. This model is adjusted in Table 3 for several values of κ. From this table,

Table 1: Results for the full model.

Parameter	$\kappa = 1$	$\kappa = 0.75$	$\kappa = 0.5$	$\kappa = 0.25$	$\kappa = 0.1$
Constant	-19.280 (6.76e-3)	-19.664 (6.06e-3)	-20.100 (3.30e-3)	-20.450 (5.28e-4)	-20.700 (6.67e-3)
Right ME	11.777 (7.61e-3)	11.786 (1.00e-2)	11.800 (2.88e-3)	11.791 (1.11e-3)	11.800 (3.26e-3)
Right RE	-0.172 (1.42e-4)	-0.173 (4.53e-4)	-0.172 (7.71e-5)	-0.174 (1.53e-5)	-0.174 (1.43e-4)
Right IOP	-0.006 (1.23e-4)	-0.003 (3.27e04)	0.001 (1.08e-4)	0.003 (1.95e-5)	0.004 (8.63e-5)
RSRBR	0.366 (1.03e-4)	0.366 (1.33e-4)	0.366 (1.32e-5)	0.366 (4.93e-6)	0.366 (3.86e-5)
RSLBR	0.278 (2.93e-5)	0.278 (2.09e-4)	0.278 (2.01e-5)	0.277 (3.89e-6)	0.277 (8.30e-5)
Left ME	13.399 (2.10e-2)	13.399 2.20e-2)	13.300 (1.10e-2)	13.403 (3.47e-3)	13.500 (2.13e-2)
Left RE	0.126 (3.96e-4)	0.125 (2.17e-4)	0.127 (2.05e-4)	0.124 (3.66e-5)	0.123 (3.62e-4)
Left IOP	-0.013 (1.6e-4)	-0.011 (3.87e-4)	-0.007 (8.76e-5)	-0.004 (1.37e-5)	-0.003 (1.43e-4)
RSRBL	0.293 (1.79e-4)	0.292 (1.21e-4)	0.293 (2.73e-5)	0.292 (1.11e-5)	0.292 (1.17e-4)
RSLBL	0.361 (3.29e-5)	0.361 (1.46e-4)	0.360 (3.34e-5)	0.360 (9.46e-6)	0.360 (5.07e-5)
DuD	0.156 (1.51e-4)	0.157 (8.91e-5)	0.158 (2.64e-5)	0.159 (5.63e-6)	0.159 (7.72e-5)
GH	0.942 (3.15e-4)	0.947 (5.45e-4)	0.956 (1.76e-4)	0.962 (3.22e-5)	0.966 (2.37e-4)
SBP	0.006 (4.75e-5)	0.006 (9.93e-5)	0.006 (6.38e-6)	0.007 (1.97e-6)	0.007 (4.52e-5)
DBP	0.123 (4.68e-5)	0.124 (2.45e-5)	0.125 (8.80e-6)	0.125 (2.81e-6)	0.126 (5.74e-5)
BMI	0.340 (1.18e-4)	0.342 (2.47e-4)	0.344 (3.79e-5)	0.347 (5.51e-6)	0.349 (1.62e-4)
PR	0.061 (8.99e-5)	0.062 (1.29e-4)	0.063 (3.11e-5)	0.064 (8.83e-6)	0.065 (5.02e-5)
UP	-0.268 (1.89e-3)	-0.290 (1.64e-3)	-0.306 (4.16e-4)	-0.327 (1.67e-4)	-0.345 (8.62e-4)
DI	-0.004 (2.64e-4)	0.017 (1.47e-3)	0.041 (2.70e-4)	0.063 (3.29e-5)	0.078 (6.01e-4)

Table 2: Results for the best model chosen using the maximum marginal probability.

Parameter	$\kappa = 1$	$\kappa = 0.75$	$\kappa = 0.5$	$\kappa = 0.25$	$\kappa = 0.1$
Constant	—	—	—	—	2.165 (2.45e-3)
RSRBR	0.732 (1.72e-5)	—	—	—	—
RSLBR	0.377 (3.04e-4)	—	—	—	—
RSRBL	—	—	1.090 (5.48e-5)	1.090 (1.37e-5)	0.741 (6.39-5)
RSLBL	0.372 (3.38e-4)	—	1.100 (5.01e-5)	1.100 (3.32e-5)	0.746 (5.97e-5)
GH	—	2.890 (1.04e-5)	—	—	0.977 (2.17e-4)
DBP	0.156 (1.85e-5)	—	—	—	—
Correlation	0.734	0.098	0.712	0.712	0.721

Table 3: Results for the chosen model.

Parameter	$\kappa = 1$	$\kappa = 0.75$	$\kappa = 0.5$	$\kappa = 0.25$	$\kappa = 0.1$
Constant	1.873	1.963	2.023	2.115	2.166
	(2.74e-3)	(2.73e-3)	(6.18e-3)	(1.72e-3)	(1.89e-3)
RSRBL	0.742	0.742	0.742	0.742	0.741
	(5.58e-5)	(4.65e-5)	(1.22e-7)	(1.45e-5)	(9.01e-6)
RSLBL	0.747	0.747	0.747	0.747	0.746
	(5.02e-5)	(5.41e-5)	(1.30e-7)	(1.62e-5)	(1.37e-5)
GH	1.003	0.994	0.989	0.981	0.977
	(1.46e-4)	(2.27e-4)	(1.66e-6)	(1.19e-4)	(2.60e-4)

it can be seen that the constant term is a decreasing function of κ, while the coefficient of GH is an increasing one. The estimated values of the coefficients of RSRBL and RSLBL do not depend on the choice of κ. The correlation between the observations and the predicted values is constant as a function of κ and is equal to 0.721.

In Figure 1, we provide the boxplot of the simulated values for the coefficient parameters of the covariates in the model fitted in Table 3 when $\kappa = 0.1$. It can be seen that, even if the boxplots have several outlying values, the simulated covariates are quite stable.

6 Concluding remarks

The present paper is an attempt to analyze bivariate ordinal data using a general linear model in a Bayesian framework. The model proposed in this paper is very general and flexible. In fact, the prior used to model the covariates can be made as noninformative (or informative) by choosing suitable values for the hyperparameters κ, α and γ. The observations from the baseline study was used to elicit the prior mean θ_0 of the covariates.

In Section 5, noninformative prior for σ^2 is used ($\alpha = 1$ and $\gamma = 0$), sensitivity analysis for the choice of κ is conducted in this section. From this study, it can be seen that κ does not have a significant influence the resulting estimates. The model selection approach provides an opportunity to deal with the significant covariates only. It is observed that although a lot of covariates were recorded, only a few of them have significant contribution to the retinopathy levels. Note that any other suitable standard criterion like the Bayes information criteria (BIC) could be used for model selection.

In the perspective of the WESDR study it could be of interest to analyze the 10 year follow-up data also by a similar technique. But we could not access the data in the form of raw data.

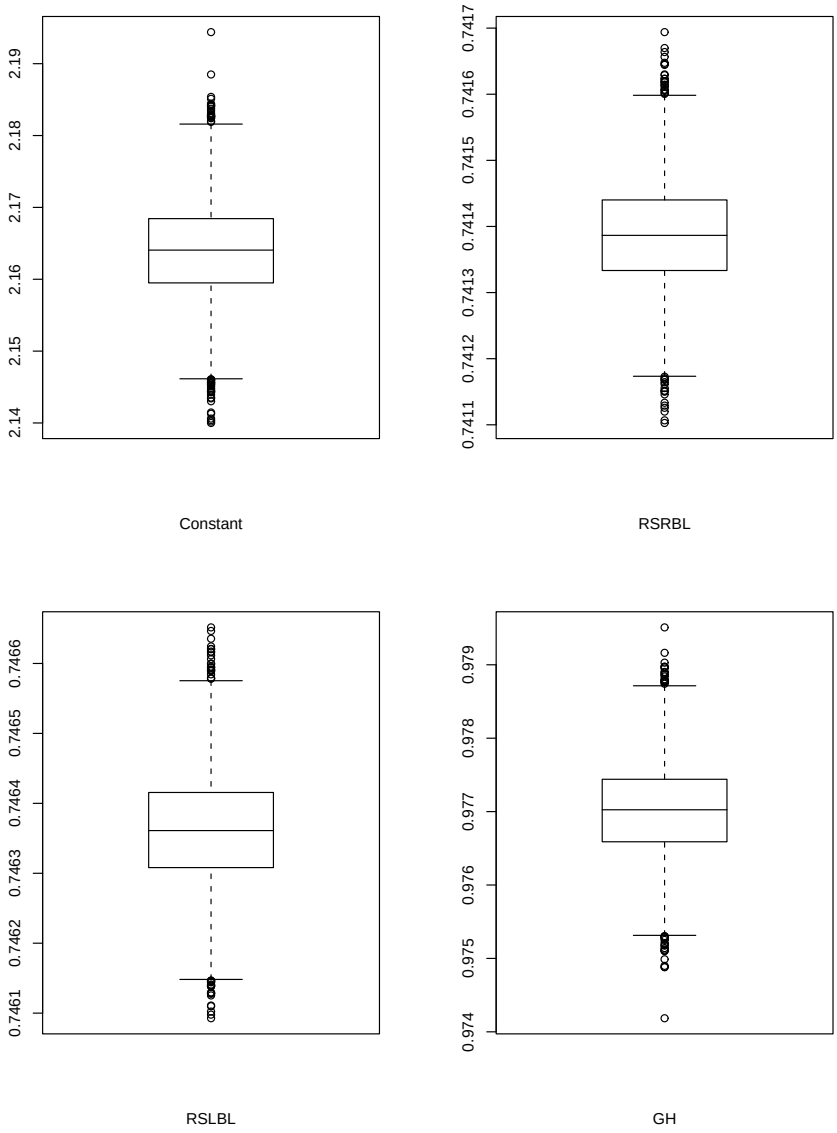

Figure 1: Boxplot of $\widehat{\theta}(\rho)$ for the constant term, RSRBL, RSLBL and GH when $\kappa = 0.1$.

References

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

[2] Cox, D. R. (1970). *The analysis of binary data*. Chapman and Hall, London.

[3] McCullagh, P. (1980). 'Regression models for ordinal data (with discussion).' *Journal of the Royal Statistical Society, Ser. B*, 42, 109-142.

[4] Snell, E. J. (1964). 'A scaling procedure for ordered categorical data.' *Biometrics*, 20, 592-607.

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

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

[7] Klein, R., Klein, B. E. K. and Davis, M. D. (1983). 'Is cigarette smoking associated with diabetic retinopathy?' *Amarican Journal of Epidemiology*, 118, 228-238.

[8] Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., DeMets, D. L. (1984). '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.

[9] Hamman, R. F. (1982). 'Data assessment and problem identification: reviewing the experience.' In Proceedings of the Diabetes Control Conference. Atlanta, Centers of Disease Control. pp. 32-40.

[10] Rand, L. I. (1981). 'Recent advances in diabetic retinopathy.' *Am. J. Med.*, 70, 595-602.

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

[12] Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1989). '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.

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

[14] 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.

[15] Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984). '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.

[16] 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.

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

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

[19] 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.

[20] 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.

[21] Das, K. and Sutradhar, B. C. (2001). 'Analyzing bivariate ordinal polytomous data: a marginal multinomial logistic approach.' Submitted in *Calcutta Statistical Association Bulletin*.

[22] Biswas, A. and Das, K. (2001). 'A Bayesian analysis of bivariate ordinal data: Wisconsin epidemiologic study of diabetic retinopathy revisited.' *Statistics in Medicine*, **20**.

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

[24] 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.

[25] 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.

[26] 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 *Advances in Neural Information Processing Systems 6*. Cowan, J., Tesauro, G. and Alspector, J. (eds). San Francisco, CA: Morgan Kaufmann Publishers.

[27] 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.

[28] Wahba, G., Lin, X., Gao, F., Xiang, D., Klein, R. and Klein, B. (1998). 'The bias-variance tradeoff and randomized GACV.' Technical Report No. 997, Department of Statistics, University of Wisconsin, Madison, WI. To appear *Advances in Information Processing Systems*, **11**.

[29] Gao, F., Wahba, G., Klein, R. and Klein, B. (1999) 'Smoothing spline ANOVA for multivariate Bernoulli observations, with application to ophthalmology data.' Technical Report No. 1009, Department of Statistics, University of Wisconsin, Madison, WI.

[30] Wang, Y., Wahba, G., Gu, C., Klein, R. and Klein, B. (1995). 'Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy.' Technical Report No. 956, Department of Statistics, University of Wisconsin, Madison, WI.

[31] Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (1998). 'Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV.' Technical Report No. 956, Department of Statistics, University of Wisconsin, Madison, WI.

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

[33] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). 'Maximum likelihood estimation from incomplete data via the EM algorithm.' *Journal of the Royal Statistical Society, series B*, **39**, 1-22.