

APPROXIMATING CORRELATION STRUCTURES IN CLUSTERED BINARY DATA USING RANDOM EFFECTS MODELS

Andreas I. Sashegyi,¹ Patrick J. Farrell² and K. Stephen Brown³

ABSTRACT

In this paper we suggest applications of random effects models for correlated binary data which can be interpreted as modelling both the mean as well as the correlation structure of the data. These models can lead to a significant improvement in fit over simpler random effects models, by making the most effective use of covariates thought to be related to the correlation structure. The importance of such covariates in the model is interpreted in terms of a regression parameter γ . We investigate the power of hypothesis tests for γ and illustrate ideas by discussing a data example from a randomized smoking prevention trial.

KEY WORDS: Generalized Linear Models; Correlated Binary Responses; Clustered Data; Random Effects Models; Logistic-Normal Model; Overdispersion.

RÉSUMÉ

Dans cet article, nous suggérons des applications de modèles à effets aléatoires pour des données binaires corrélées qui peuvent être interprétées aussi bien en tant que modèles pour la moyenne que pour la structure de corrélation des données. Ces modèles peuvent mener à une amélioration significative de l'ajustement par rapport à des modèles plus simples, en utilisant des variables covariantes qu'on pense liées à la structure de corrélation. L'importance de telles variables covariantes dans le modèle est interprétée en terme de paramètre γ de régression. Nous examinons la puissance de tests d'hypothèse pour γ et illustrons ces idées en discutant un exemple d'un jeu de données provenant d'une étude sur la prévention du tabagisme.

MOTS CLÉS : Modèle linéaire généralisé; réponses binaires corrélées; données en grappes; modèles à effets aléatoires; modèle logistic-normal; dispersion excessive.

1. INTRODUCTION

Clustered data for which it is reasonable to assume independence between clusters, but correlation among observations within clusters, arise frequently in various settings. The omission of important cluster-level covariates in an analysis to model factors shared by all respondents in a given cluster is one explanation for this correlation. Alternatively, correlation within clusters could be due to the direct influence of some

members of the cluster on others in the same cluster. The presence of intra-cluster correlation leads to more variability between clusters than would be expected if individuals within clusters were independent. If ignored, this overdispersion can lead to overstating the effects of cluster-level covariates.

For binary responses intra-cluster correlation can be addressed using the logistic-normal random effects model. This assumes responses in a cluster to be

¹ Andreas I. Sashegyi, Eli Lilly and Company, Lilly Corporate Center, D.C. 2244, Indianapolis, Indiana, USA, 46285, sashegyi_andreas@lilly.com.

² Patrick J. Farrell, Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, Canada, B0P 1X0, pat.farrell@acadiau.ca.

³ K. Stephen Brown, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1, ksbrown@icarus.math.uwaterloo.ca.

conditionally independent, given a cluster-level random effect. In some cases, however, the assumption of a common random effect for all observations in a cluster may be too restrictive. Consider the Waterloo Smoking Prevention Project 3 (WSPP3), the third in a series of randomized, controlled smoking prevention trials (See Best et al. 1995 and Brown and Cameron 1997). In this study 100 Southern Ontario elementary schools were randomized, within strata defined by an estimated level of school risk, to either treatment or control conditions. Starting in grade 6, students in the treatment schools were exposed each year until grade 8 to the smoking prevention curriculum. A baseline measure of smoking status was taken prior to any intervention at the beginning of grade 6, and subsequently smoking status was measured on the same students at the end of grades 7 and 8, after which they moved on into secondary schools. One question of interest in the WSPP3 study concerns the effectiveness of the smoking intervention program. We address this issue here through the development of an appropriate model for describing the data, where it was discovered that, even after a careful modelling of the mean response, the data still show evidence of considerable school-to-school variability which is unexplained. Specifically, it appears that the school-to-school variation in smoking rates is considerably greater among students deemed to be at high-risk of smoking than among medium or low-risk students. Hence assuming that a common school-level random effect is associated with all responses in a given school is not appropriate, and we therefore consider an alternative modelling strategy.

In this paper we suggest an extension of the standard logistic-normal random effects model which can be interpreted as modelling both the mean and the correlation structure of the data. It incorporates covariates thought to be related to the correlation structure through a regression-like function depending on a parameter γ . We discuss the model in section 2, and in section 3 examine the results of a simulation study to investigate the power of relevant hypothesis tests for γ . An illustration using the WSPP3 data is provided in section 4.

2. EXTENDING THE LOGISTIC-NORMAL RANDOM EFFECTS MODEL

Numerous applications of various random effects models, for binary data in particular, have been proposed in the literature. An appreciation of these models is conveyed in Breslow and Clayton (1993). Many other useful references are to be found in

Sashegyi (1998). Consider binary observations Y_{ij} collected in K clusters (Y_{ij} referring to the j th observation in cluster i), such that Y_{ij} and $Y_{i'j}$, $i \neq i'$ are independent, but that marginally $\text{Corr}(Y_{ij}, Y_{ij'}) > 0$. Assume the following model for Y_{ij} :

$$\begin{aligned} Y_{ij} \mid b_i &\sim \text{Bin}(1, p_{ij}), \\ \text{logit}(p_{ij}) &= x'_{ij}\beta + \varphi(z_{ij}; \gamma)b_i, \\ b_i &\sim \text{i.i.d. } N(0, \sigma^2), \end{aligned} \quad (2.1)$$

where $j = 1, \dots, n_i$, $i = 1, \dots, K$, and x_{ij} represents a vector of covariates for the j th observation in cluster i with associated parameter vector β . Similarly, z_{ij} is a vector of covariates thought to be related to the correlation structure of the data through some function φ and parameter vector γ . Finally, b_i is a random effect associated with the i th cluster, assumed to be normally distributed. Assume also that two observations from the same cluster are independent conditional on the random effect for that cluster. Naturally one could write down a similar formulation for other link functions or data types; for dichotomous data in particular, (2.1) offers a convenient means of addressing both variation in the mean response, as well as allowing for an indirect modelling of the correlation structure through the function $\varphi(z_{ij}; \gamma)$. In multiplying the cluster-specific random effect, this function, which can be specific to the individual, either inflates or attenuates the random effect, thus tailoring its impact for each specific cluster, perhaps even each individual. The term $\varphi(z_{ij}; \gamma)$ relaxes the assumption of a common random effects variance; the variance of the random component associated with observation (i, j) is $\varphi^2(z_{ij}; \gamma)\sigma^2$, so the function φ serves to explain some of the extra heterogeneity in the data, over and above that which can be captured in a simpler model assuming $\varphi = 1$. For fixed β and σ^2 , varying degrees of correlation between Y_{ij} and $Y_{ij'}$ can be accommodated through the values of $\varphi(z_{ij}; \gamma)$ and $\varphi(z_{ij'}; \gamma)$; see Sashegyi (1998) and Sashegyi, Brown, and Farrell (1998b). In choosing a mixing distribution for b_i whose range of support is the entire real line, we can without loss of generality impose the restriction $\varphi(z_{ij}; \gamma) > 0$. In general, $\text{Corr}(Y_{ij}, Y_{ij'})$ is an increasing function of $\varphi^2(z_{ij}; \gamma)\sigma^2$ and $\varphi^2(z_{ij'}; \gamma)\sigma^2$. Model (2.1) gives us a straightforward framework for assessing the impact of covariate(s) z_{ij} thought to be associated with the correlation structure in the data. In letting $\varphi(z_{ij}; \gamma)$ depend on parameter(s) γ we are able to use the data to estimate the best fitting relationship, for a given function φ , between z_{ij} and a sample of random effects $\{b_1, \dots, b_K\}$ from a common distribution. Throughout this paper we will model assuming that $\gamma = 0$ implies

that $\varphi(z_{ij}; \gamma) = 1$, although other formulations are possible.

Model (2.1) is useful in that it provides a straightforward tool with which aspects of the correlation structure of the data can be captured in a parsimonious fashion. Instead of resorting to several variance components to explain the extraneous variability, covariate information (z) is used in conjunction with associated parameters (γ) to adjust the impact of random effects from a single univariate distribution. Added flexibility is gained by the fact that γ , which determines the impact of z , is estimated from the data. Maximum likelihood estimation is reasonably straightforward in this model, and the estimate of γ allows us to quantify in a convenient manner the degree of departure from the standard model having $\varphi = 1$.

3. A SIMULATION STUDY

3.1 Motivation

We illustrate the consequences of model misspecification by simulating data from a model with $\varphi \neq 1$, and comparing its fit to such data with that of a

misspecified model where $\varphi = 1$. Consider (3.1) below, describing responses from 40 individuals in each of 30 clusters, where in each cluster an individual belongs to one of two groups:

$$Y_{ij} \mid b_i \sim \text{Bin}(1, p_{ij}),$$

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + \exp(\gamma x_{ij}) b_i, \quad (3.1)$$

$$x_{ij} = \begin{cases} 1 & j = 1, \dots, 20 \\ 0 & j = 21, \dots, 40 \end{cases},$$

$$\beta_0 = -2, \quad \beta_1 = 1, \quad \gamma = 1, \quad b_i \sim \text{i.i.d. } N(0, 0.25).$$

Under (3.1) both the mean and variability in the responses differ among the two groups. We choose $\varphi(z_{ij}; \gamma) = \varphi(x_{ij}; \gamma) = \exp(\gamma x_{ij})$, as there are no restrictions on γ under this formulation; this approach is also adopted when a model of similar structure is used to analyze the WSPP3 data in Section 4. We simulated 300 data sets from (3.1) and to each fit the true model using maximum likelihood, and an incorrect simpler model, assuming that $\gamma = 0$. The results are given in Table 1.

Term	True Value	Correct Model Fit		Misspecified Model Fit	
		Mean	Mean s.e.	Mean	Mean s.e.
β_0	-2.0	-2.0274	0.1653	-2.1896	0.2381
β_1	1.0	1.0195	0.3209	1.2369	0.3196
γ	1.0	1.2108	0.5764	0.0	
$\exp(2\gamma)\sigma^2$	1.8473	1.8580		1.0863	
σ^2	0.25	0.2546		1.0863	
<i>llik</i>		-550.28		-554.63	

Table 1: Summary of correct and misspecified model fit to data generated from (3.1).

The mean of the estimates over the 300 data sets is reported for each parameter, along with the average of the model-based standard errors (s.e.). The two values of the variance of the $\exp(\gamma x_{ij}) b_i$ term are given, corresponding to $x_{ij} = 1$ and $x_{ij} = 0$. The average value of the maximized log-likelihood, across the 300 data sets, is also reported.

The model implies that the variance of the cluster-level random effect is more than 7 times greater for observations for which $x_{ij} = 1$. Fitting the correct model reflects this; the misspecified model underestimates the random variability in observations from the first group and overestimates it for the second group. The fit of the correct model appears to improve significantly on that of the misspecified model, as judged from the average increase in log-likelihood.

Estimators for β_0 and β_1 under the correct model appear to have little bias; in addition, the empirical root mean square errors of these two estimators over the 300 simulated data sets are approximated well by the average model-based standard errors in Table 1. Further, an investigation of normal probability plots and histograms of the 300 estimates for β_0 and β_1 indicate no serious departures from normality. By contrast, the latter model appears to produce biased estimates of β_0 and β_1 . This has important implications when modelling the WSPP3 data, as the effectiveness of the intervention program in reducing smoking rates will be quantified through a parameter in β , which would be associated with a covariate in x_{ij} . Since it will be shown in Section 4 that the data show evidence of considerable school-to-school variability in smoking rates which is unexplained even after a

careful modelling of the mean response using (2.1) with $\varphi(z_{ij}; \gamma) = 1$ (resulting from much greater school-to-school variation in smoking rates for students deemed to be at high individual level risk of smoking than among medium or low risk students), the effectiveness of the intervention program might be inappropriately assessed using (2.1) with $\varphi(z_{ij}; \gamma) = 1$ due to the bias in the estimator of β . Thus, the potential for bias in the estimator of β further motivates the importance of identifying the correct model, which a test of $\gamma = 0$ facilitates.

3.2 Results

We examine here the properties of the test $H_0: \gamma = 0$ versus $H_a: \gamma > 0$ based on the test statistic $\hat{\gamma}/s.e.(\hat{\gamma})$. We examine five different models (containing only cluster-level covariates to increase the speed of computation) and for each determine power curves associated with the above test. Each model can be expressed in the form

$$\begin{aligned} Y_{ij} &| b_i \sim \text{Bin}(1, p_i), \\ \text{logit}(p_i) &= \beta_0 + \beta_1 x_i + \varphi(z_i; \gamma) b_i, \\ b_i &\sim \text{iid } N(0, \sigma^2), \end{aligned} \quad (3.2)$$

where $j = 1, \dots, n_i$, and $i = 1, \dots, 50$. We consider the following five formulations:

M1: $\varphi(z_i; \gamma) = (n_i/10)^{-\gamma}$,
 $(z_i = n_i)$,
 $x_i = \begin{cases} 1 & i = 1, \dots, 25 \\ 0 & i = 26, \dots, 50 \end{cases}$
 $(n_1, n_2, \dots, n_{25}) = (20, 40, 60, 80, \dots, 500)$,
 $n_{i+25} = n_i$, $\beta_0 = -1$, $\beta_1 = 2$, $\sigma^2 = 4$;

M2: as *M1*, but
 $(n_1, n_2, \dots, n_{25}) = (20, 28, 35, 42, 50, \dots, 200)$
(cluster sizes increasing linearly from 20 to 200, with n_i rounded to the nearest integer);

M3: as *M1*, but
 $(n_1, n_2, \dots, n_{25}) = (20, 23, 27, 30, 33, \dots, 100)$
(cluster sizes increasing linearly from 20 to 100, with n_i rounded to the nearest integer);

M4: $\varphi(z_i; \gamma) = \exp(\gamma x_i)$,
 $(z_i = n_i)$,
 $x_i \sim \text{Uniform}(0,1)$,
 $n_i = 40$, $i = 1, \dots, 50$, $\beta_0 = -1$, $\beta_1 = 2$, $\sigma^2 = 0.25$;

M5: as *M4*, but $\sigma^2 = 1.0$.

Models *M1* to *M3* describe situations in which intra-cluster correlation is a function of the cluster size n_i , for various ranges of n_i . In these models the covariate associated with the correlation structure of the data has no impact on the mean. Models *M4* and *M5* examine cases where intra-cluster correlation is a function of a continuous covariate, which also appears in the mean formulation of the model.

For each model, 400 data sets were simulated for each of various values of γ , including $\gamma=0$. The empirical distribution under H_0 of the test statistic $\hat{\gamma}/s.e.(\hat{\gamma})$ showed no significant departure from normality in any of the cases studied (see Sashegyi 1998). For various significance levels α , an estimate of the power was computed as the proportion of times an observed value $\hat{\gamma}_r/s.e.(\hat{\gamma}_r)$, $r = 1, \dots, 400$ exceeded $Z_{1-\alpha}$. We summarize the findings here; for details see Sashegyi (1998). As one might expect intuitively, for models *M1* to *M3*, the power to detect a positive value of γ decreases as the range of cluster sizes becomes smaller. Furthermore, it seems that the power depends not so much on the absolute range of the variance of $\varphi_i b_i$ as on the ratio of the upper to the lower endpoint. This is illustrated well in models *M4* and *M5*, which have very similar power curves; the model specifications are the same with the exception that the constant prior variance σ^2 in *M5* is four times greater than in *M4*. Of course one must bear in mind that in testing $H_0: \gamma = 0$ versus $H_a: \gamma > 0$ we are trying to determine whether the extraneous variation in the data is of a particular form. If σ^2 is very small such a test will be neither powerful nor relevant. The more pertinent question then is whether there is evidence for overdispersion at all.

4. AN EXAMPLE

We now examine some aspects of elementary school smoking behaviour, as revealed by the WSP3 study. We consider here the responses (smoking status) of students in grades 7 and 8 who were non-smokers at baseline. Considering a complete-case analysis, this corresponded to two observations on each of 3380 students, attending 99 schools. Sashegyi, Brown, and Farrell (1998a) proposed an approach for modelling data of this type, which are simultaneously correlated cross-sectionally by school, and longitudinally due to repeated observations on students. They applied their approach to the WSP3 data described above, discovering that the intra-individual correlation between grade 7 and 8 observations on the same student was extremely small. However, they did not attempt to explicitly model the correlation structure of

the data. We therefore shall ignore the longitudinal clustering in the data here, and focus on an examination of the school-to-school variability. Let $Y_{ij} = 1$ if observation j in school i refers to a smoker, and 0 otherwise, where $j = 1, \dots, n_i$, and $i = 1, \dots, 99$. Note that Y_{ij} and $Y_{ij'}$ may refer to the same student observed in school i in grades 7 and 8; smoking status is time-dependent and for a given student may change from one grade to the next.

The WSPP3 elementary schools were randomized to one of five study conditions. Four of these were treatment conditions, corresponding to the combinations of the type of provider who administered the intervention curriculum (nurse or teacher) and the type of training the provider received (workshop or mediated training through printed material). The fifth was a control condition. Initially, we modelled Y_{ij} using the standard logistic regression model (no random effects) to determine the covariates of most relevance to the prediction of individual smoking status. These variables, included in x_{ij} in model (2.1) for subsequent analyses, were:

Cond: study condition (*Cond* = 1 for schools in one of the four treatment conditions, and 0 otherwise);

Risk: an individual-level smoking risk score (*Risk* = 1 for students classified on the basis of external factors to be at low risk for smoking, *Risk* = 2 for students at medium risk and 3 for students at high risk);

Gr8surv: a school-level risk score, coded as a continuous covariate ranging between 0 and 100, with larger values indicative of higher-risk schools; this was the percentage of smokers among Grade 8 students in each school when the cohort was in grade 6, and was used to form the stratification variable for school risk; and

Gr8: a grade effect (*Gr8* = 1 for a grade 8 observation, 0 otherwise).

The interaction between *Cond* and *Gr8surv* ($C \times Gr8surv$) was also included. We then fit the standard logistic-normal random effects model, letting $\phi(z_{ij}; \gamma) = 1$ in (2.1). The results, given in the first panel of Table 2, indicate that this model improves significantly on the fit of the logistic, (log-likelihood of -2303 versus -2321); thus there is considerable overdispersion in the data due to school-to-school variability. However, there is much school-to-school variation in smoking rates that remains unexplained with the standard random effects model.

Term	Random Effects Models					
	est.	s.e.	est.	s.e.	est.	s.e.
<i>Intercept</i>	-4.4820	(0.2820)	-4.3936	(0.2760)	-4.3557	(0.2720)
<i>Cond</i>	0.3640	(0.2802)	0.3135	(0.2655)	0.3089	(0.2668)
<i>Risk</i>	0.8600	(0.0550)	0.8398	(0.0609)	0.8261	(0.0597)
<i>Gr8surv</i>	0.0293	(0.0135)	0.0283	(0.0128)	0.0278	(0.0129)
<i>Gr8</i>	0.8764	(0.0816)	0.8711	(0.0819)	0.8728	(0.0819)
<i>C x Gr8surv</i>	-0.0261	(0.0149)	-0.0284	(0.0141)	-0.0251	(0.0142)
γ_1			0.3184	(0.3560)		
γ_2			0.7508	(0.3159)	0.6335	(0.2615)
σ^2	0.1950		0.0993		0.1300	
<i>llik</i>	-2302.76		-2299.93		-2300.31	

Table 2: Models addressing school-to-school variability in the WSPP3 elementary school data, based on formulation (2.1).

One factor related to the mean probability of smoking in both this model and the logistic which might also help explain the extraneous school-to-school variation in smoking rates is the individual-level risk profile of students. Thus, we considered the random effects model

$$\text{logit}(p_{ij}) = x'_{ij}\beta + \exp(\gamma_1 z_{ij1} + \gamma_2 z_{ij2})b_i, \quad (4.1)$$

$b_i \sim \text{i.i.d. } N(0, \sigma^2)$

for the response probability $P(Y_{ij} = 1 | b_i)$, where $z_{ij1} = 1$ for a low-risk observation (*Risk* = 1), and 0 otherwise, and $z_{ij2} = 1$ for a high-risk observation (*Risk* = 3), and 0 otherwise. This general formulation allows for varying dispersion in the responses of students according to their level of risk for smoking. The results of fitting this model are given in the second panel of Table 2. Note that the estimate of γ_2 is significant. Based on the simulation study conducted in Section 3.1, this would bring into question the bias

of the estimators for β in the standard logistic normal random effects model with $\varphi(z_{ij}; \gamma) = 1$. This finding suggests greater variability in the responses of the high-risk students as compared to those at medium and low risk. Combining the latter two risk groups, and fitting model (4.1) with $\gamma_1 = 0$ yields the results in the third panel of Table 2. We conclude from this fit that the standard deviation of the random effects distribution governing the school-to-school variability in smoking rates among high-risk students is $e^{0.6335}$ or about twice as large as that for the remaining students. Adjusting for this difference in dispersion also produces a significant increase in log-likelihood over the standard random effects model, from -2303 to -2300 , at the expense of only one degree of freedom. Regarding the effect of the intervention program, there is a marginally significant interaction between *Cond* and *Gr8surv* indicating that the intervention program seemed to induce lower smoking rates, particularly in high risk schools.

5. DISCUSSION

The class of random effects models presented here is ideally suited to drawing mixed-effects model inferences from cluster-correlated data, when specific hypotheses about the correlation structure are of interest, or when information about the correlation structure is available. The WSPP3 elementary intervention, for instance, was based on a social influences curriculum which increased students' awareness of various influences in their environment which might prompt smoking, such as peer pressure, and provided tools to help them resist these pressures. The nature of this intervention program underscores the importance of understanding behavioural patterns in the study population when examining an outcome such as smoking status. In mathematical terms this can be quantified in terms of a non-independent covariance structure, using covariates to determine a specific model. As seen in section 4, for example, an indicator of individual-level risk (high versus moderate or low) in the random component of a model of form (2.1) shows that school-to-school variability is much higher for high-risk students. This suggests that students whose individual profile puts them at high risk of smoking tend to be more similar in their smoking behaviour within a given school than students at low or moderate risk.

Apart from computational tractability, an attractive feature of this modelling approach is that continuous

covariates thought to be associated with the correlation structure in the data can be included. Taking recourse to a more general bivariate random effects distribution is an alternative to model (2.1) if a single dichotomous covariate z_{ij} is of interest in the formulation of $\varphi(z_{ij}; \gamma)$, but it becomes more difficult to match the flexibility and convenience of this model if z_{ij} is continuous and varying within or between clusters.

ACKNOWLEDGEMENTS

This work was supported through funds from the Natural Sciences and Engineering Research Council of Canada, the National Health Research and Development Program (Canada), and the National Heart, Lung and Blood Institute (US).

REFERENCES

- Best, J. A., Brown, K. S., Cameron, R., Manske, S. M. and Santi, S. (1995), "Gender and predisposing attributes as predictors of smoking onset: implications for theory and practice", *Journal of Health Education*, **26**, S52-S60.
- Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models", *Journal of the American Statistical Association*, **88**, 9-25.
- Brown, K. S. and Cameron, R. (1997), *Long-term evaluation of an elementary and secondary school smoking intervention*, Final Report to the NHRDP.
- Sashegyi, A. I. (1998), *Models for correlated binary responses: applications for the Waterloo Smoking Prevention Projects Data*, Ph.D. Dissertation, Department of Statistics and Actuarial Science, University of Waterloo.
- Sashegyi, A. I., Brown, K. S. and Farrell, P. J. (1998a) "Estimation in an empirical Bayes model for longitudinal and cross-sectionally clustered binary data", Submitted to the *Canadian Journal of Statistics*.
- Sashegyi, A. I., Brown, K. S. and Farrell, P. J. (1998b) "On the correspondence between population-averaged and a class of cluster-specific models for correlated binary data", Submitted to *Statistics and Probability Letters*.