Analysis of Categorical Data Structures with Repeated Measurements and Possibly Clustering

Eliana Marques and Gary G. Koch
Department of Biostatistics, University of North Carolina
Chapel Hill, N.C. 27599

Abstract

In the health and social sciences, researchers often encounter categorical data for which complexities come from the involvement of a nested hierarchy and/or cross-classification for sampling structure. In data collection, a common feature of these studies is a non-standard data structure with repeated measurements which may have some degree of clustering. In this paper, an application of an extension of methods for analysis of clustered attribute data from a two-stage nested design is discussed. Quantities of interest in this context are the mean value π of an observed dichotomous response for a certain condition or time point and the correlation coefficients: ρ, that measures the strength of clustering in a set of data within a condition or time period; γ, that reflects time correlation; and ν, that reflects correlation between times for different subjects in the same cluster. An example dealing with dental practices is provided for illustrative purposes.

1. Introduction

Longitudinal categorical data studies involving a clustering structure have often been analyzed focusing on the cluster sub-unit instead of the cluster as the basic unit of analysis. Analysis accounting for the clustered structure of the data is relevant so that variances of measures of association or group differences are not underestimated. Only if it is possible to demonstrate negligible correlation among cluster sub-units is an analysis based on the cluster sub-units as units of analysis justified.

This paper is concerned with describing an application of weighted least squares methods to clustered attribute data. The setting of interest is a two-stage nested design with a partially balanced structure. More complete discussion of the methods outlined in this paper is given in Marques (1988).

2. The Data

The data used to illustrate aspects of analysis are from dental practices located in 14 states of the United States. Visits were made during 1984 and 1985 to 300 offices to gather information regarding the structure, process and outcome for the practices.

The sample of 300 offices represents about 15 percent of the dentists who volunteered their participation in response to letters of solicitation, subsequent to their selection from the American Dental Association Directory. Even though the practices visited do not constitute a strictly random sample of American dentists, they do provide useful information about the character of the dentist-patient interface.

Data were available from 200 urban non-group practices, 50 urban group practices and 50 rural non-group practices. A questionnaire dealing with dental outcomes was to be completed by about eight patients from each practice. In all, 2234 patients answered questions about preventive orientation of dental practices and of themselves, dental patients. The responses to these questions were for brushing: never, once, twice, three times a day or more; for flossing: never or rarely, usually once a month, usually once a week, usually on a daily basis. It is likely that patients within the same clinic may tend to provide relatively similar responses, and so these subjects should not be viewed as independent of each other. The objective of analysis presented here is to assess the effect of clustering of patients within centers and its implications to the interpretation of results. An outline of the methodology to account for the possible correlation structure of patients within a practice, with practice as the unit of analysis, is presented next.

3. Methods

The theoretical framework for the analysis of clustered attribute bivariate data in a two-stage nested, and partially balanced design is briefly reviewed. The basic data structure in this setting involves clusters, with several sizes, one sub-population (treatment) per cluster and two binary responses from each subject in a cluster; see Kempthorne and Koch (1983) for discussion of the case with one binary response.

In some experimental situations, modules may be formed with clusters of different sizes grouped accordingly; for example, dental practices providing 6, 7, or 8 patients to the sample constitute distinct modules. Each module is balanced so that there is one set of estimates for each module.

Suppose the m-th module has n_m clusters each of size d_m. Let

$$y_{ijkm}=\begin{cases}1 & \text{if k-th response for subject j in the i-th} \\ & \text{cluster of the m-th module is favorable} \\ 0 & \text{Otherwise}\end{cases}$$

where $i=1,\ 2,\ \ldots,\ n_m$; $j=1,\ 2,\ \ldots,\ d_m$; $k=1,\ 2$; $m=1,\ 2,\ \ldots,\ m$.

The model of interest is characterized by:

$E(y_{ijkm}) = \pi_{km} =$ Pr (k-th response for a given subject in a cluster of the m-th module is favorable)

$E(y_{ijkm}\ y_{ij'km}) = \lambda_{km} =$ Pr (k-th responses for two distinct subjects in a cluster of the m-th module are favorable)

$E(y_{ijkm}\ y_{ijk'm}) = \theta_m =$ Pr (both responses for a subject in a cluster of the m-th module are favorable)

$E(y_{ijkm}\ y_{ijk'm}) = \phi_m =$ Pr (response k from subject j and response k′ from subject j′ in a cluster of the m-th module are favorable)

Other parameters of interest are

$$\text{Cov}\ (y_{ijkm}\ y_{ij'km}) = \lambda_{km} - \pi^2_{km}$$
$$\text{Cov}\ (y_{ijkm}\ y_{ijk'm}) = \theta_m - \pi_{km}\pi_{k'm}$$
$$\text{Cov}\ (y_{ijkm}\ y_{ij'k'm}) = \phi_m - \pi_{km}\pi_{k'm}$$

and the corresponding correlation coefficients

$\rho_{km} = (\lambda_{km} - \pi^2_{km})/\pi_{km}(1-\pi_{km}) =$ intraclass correlation of subjects in same cluster for k-th response

$\gamma_{km} = (\theta_m - \pi_{km}\pi_{k'm})/\{\pi_{km}(1-\pi_{km})\pi_{k'm}(1-\pi_{k'm})\}^{\frac{1}{2}}$ = correlation of k-th and k′-th response for same subject

$\nu_{km} = (\phi_m - \pi_{km}\pi_{k'm})/\{\pi_{km}(1-\pi_{km})\pi_{k'm}(1-\pi_{k'm})\}^{\frac{1}{2}}$ = correlation of k-th and k′-th response from distinct subjects within the same cluster

Grouping the observations y_{ijkm} in an appropriate way, expressions can be formulated for the parameter estimators. The estimates $\hat{\pi}_{km}$ and $\hat{\theta}_m$ may be written as

$$\hat{\pi}_{1m} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{d} y_{ij1m}}{nd} = \tfrac{1}{n}\sum_{i=1}^{n} F_{i1m} = F_{1m},$$

$$\hat{\pi}_{2m} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{d} y_{ij2m}}{nd} = \tfrac{1}{n}\sum_{i=1}^{n} F_{i2m} = F_{2m},$$

$$\hat{\theta}_m = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{d} y_{ijkm}\ y_{ijk'm}}{nd} = \tfrac{1}{n}\sum_{i=1}^{n} F_{i3m} = F_{3m},$$

while

$$(\hat{\pi}_{1m} - \hat{\lambda}_{1m}) = s^2_{1m} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{d} (y_{ij1m} - F_{i1m})^2}{n(d-1)}$$
$$= \tfrac{1}{n}\sum_{i=1}^{n} F_{i4m} = F_{4m}$$

$$(\hat{\pi}_{2m} - \hat{\lambda}_{2m}) = s^2_{2m} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{d} (y_{ij2m} - F_{i2m})^2}{n(d-1)}$$
$$= \tfrac{1}{n}\sum_{i=1}^{n} F_{i5m} = F_{5m}$$

and

$$(\hat{\theta}_m - \hat{\phi}_m) = s_{12m}$$

$$= \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{d} (y_{ij1m} - F_{i1m})(y_{ij2m} - F_{i2m})}{n(d-1)}$$
$$= \tfrac{1}{n}\sum_{i=1}^{n} F_{i6m} = F_{6m}.$$

It follows that

$$\underset{\sim}{F}_m = \begin{bmatrix} \hat{\pi}_{1m} \\ \hat{\pi}_{2m} \\ \hat{\theta}_m \\ \hat{\pi}_{1m}\text{-}\hat{\lambda}_{1m} \\ \hat{\pi}_{2m}\text{-}\hat{\lambda}_{2m} \\ \hat{\theta}_m\text{-}\hat{\phi}_m \end{bmatrix} = \begin{bmatrix} F_{1m} \\ F_{2m} \\ F_{3m} \\ F_{4m} \\ F_{5m} \\ F_{6m} \end{bmatrix} = \tfrac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} F_{i1m} \\ F_{i2m} \\ F_{i3m} \\ F_{i4m} \\ F_{i5m} \\ F_{i6m} \end{bmatrix} = \tfrac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_{im} \qquad (1)$$

is a (6 x 1) vector of modular estimates. Thus, variation among the elements of $\underset{\sim}{F}_{(6M\times 1)}$, where

$$\underset{\sim}{F}_{(6M\times 1)} = [\underset{\sim}{F}_1, \underset{\sim}{F}_2, \ldots, \underset{\sim}{F}_m, \ldots, \underset{\sim}{F}_M]'$$

can be studied by using the extensions outlined in Koch, *et al.* (1977) for the methodology of Grizzle, *et al.* (1969). The model $E(\underset{\sim}{F}) = \underset{\sim}{X}\underset{\sim}{\beta}$, where $\underset{\sim}{X}' = [I_6,\ I_6,\ \ldots,\ I_6]'$ and

$\underset{\sim}{\beta} = [\pi_1, \pi_2, \theta, \pi_1-\lambda_1, \pi_2-\lambda_2, \theta-\phi]'$ is fit. The weighted least squares estimate $\hat{\underset{\sim}{\beta}}$ is obtained from $\hat{\beta} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'V_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}$ and its estimated covariance matrix is $V_{\hat{\underset{\sim}{\beta}}}=(X'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}$. Here $\underset{\sim}{V}_{F}$ is a block diagonal matrix with diagonal blocks $\underset{\sim}{V}_{\underset{\sim}{m}}$ where $\underset{\sim}{V}_{\underset{\sim}{m}}=\frac{1}{n^2}\sum_{i=1}^{n}(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)'$. Lack of fit of the model is assessed through $Q = (\underset{\sim}{F}-\underset{\sim}{X}\hat{\underset{\sim}{\beta}})'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}(\underset{\sim}{F}-\underset{\sim}{X}\hat{\underset{\sim}{\beta}})$ which approximately has the chi-square distribution d.f.=.6(M−1).

If the number of clusters *in* a module is small ($n_m \leq 10$), then estimates derived by the methodology of Grizzle, *et al.* (1969) may not have the advantageous properties which follow from asymptotic arguments, and so alternative approaches such as direct estimates like mean of means or ratio means may be used.

The estimates $\hat{\underset{\sim}{\beta}}$ can be transformed to estimates for the probabilities π_1 and π_2 of favorable response and intraclass correlation coefficients for the effect of clustering through the application of a series of linear, logarithmic, and exponential transformations; see Koch, *et al.* (1985) for background discussion. This specification, which is used to facilitate computation of linear Taylor series based estimates of the covariance matrix, is as follows:

$$\begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\rho}_1 \\ \hat{\rho}_2 \\ \hat{\gamma} \\ \hat{\nu} \end{bmatrix} = \underset{\sim}{\exp}\ [\underset{\sim}{A}_4\underset{\sim}{\ln}[\underset{\sim}{A}_3[\underset{\sim}{C}_2+\underset{\sim}{\exp}[\underset{\sim}{A}_2\underset{\sim}{\ln}(\underset{\sim}{C}_1+A_1\hat{\underset{\sim}{\beta}})]]]] \tag{2}$$

where $\underset{\sim}{\ln}$ and $\underset{\sim}{\exp}$ apply natural logarithms or exponential transformations to the elements of a vector and

$$A_1= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}, C_1= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$A_2= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \end{bmatrix}, C_2= \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ 0 \\ -1 \\ 0 \end{bmatrix},$$

$$A_3= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{and}$$

$$A_4= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

4. Analysis of Example

In this example, patients from dental practices answered questions about: i) brushing their teeth, and ii) using dental floss. Answers to these questions are first transformed to the indicators

$$y_{ij1m}=\begin{cases} 1 & \text{if response } y_{ij1m} \text{for brushing teeth is ``twice-a-day'' or ''three times or more''} \\ 0 & \text{Otherwise} \end{cases}$$

and

$$y_{ij2m}=\begin{cases} 1 & \text{if response } y_{ij2m} \text{for flossing teeth is ''usually once a week'' or ''usually on a daily basis''} \\ 0 & \text{Otherwise.} \end{cases}$$

Then means are calculated for each variable by calculating the proportion of patients within a clinic brushing twice or more and the proportion flossing usually once a week or more often, and then averaging these proportions across clinics. Data from 289 practices were analyzed; these included urban group clinics of sizes 7 and 8 grouped in modules with 16 and 30 clinics respectively; and urban and rural non-group clinics of sizes 6, 7, and 8 grouped in modules of 33, 79 and 131 clinics respectively. Note that one practice with one patient and three with five patients each were excluded from the analysis. There were three clinics with missing data on the variables of interest for the analysis and there were only 4 urban group clinics with 6 patients each; these 7 practices were also excluded from the analysis.

The number of clusters in each module is considered good for the methodology to be applied in order to obtain estimates of the parameters of interest.

The following model was fit to the vector of module estimates $\hat{\pi}_{1m}$, $\hat{\pi}_{2m}$, $\hat{\theta}_m$, $(\hat{\pi}_{1m}-\hat{\lambda}_{1m})$, $(\hat{\pi}_{2m}-\hat{\lambda}_{2m})$ and $(\hat{\theta}_m-\hat{\phi}_m)$:

$$\underset{(30\ x\ 1)}{E(\underset{\sim}{F})} = \begin{bmatrix} \underset{\sim}{I}_6 & \underset{\sim}{0} \\ \underset{\sim}{I}_6 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{I}_6 \\ \underset{\sim}{0} & \underset{\sim}{I}_6 \\ \underset{\sim}{0} & \underset{\sim}{I}_6 \end{bmatrix} \begin{bmatrix} (\pi_1)_1 \\ (\pi_2)_1 \\ \theta_1 \\ (\pi_1-\lambda_1)_1 \\ (\pi_2-\lambda_2)_1 \\ (\theta-\phi)_1 \\ (\pi_1)_2 \\ (\pi_2)_2 \\ \theta_2 \\ (\pi_1-\lambda_1)_2 \\ (\pi_2-\lambda_2)_2 \\ (\theta-\phi)_2 \end{bmatrix} = \underset{\sim}{X}\underset{\sim}{\beta}. \qquad (3)$$

The model fits well (Q=15.3 with 18 d.f., p=0.64). Parameter estimates for this model are presented in Table 1. By applying a series of linear, logarithmic, and exponential functions as in (2) (with matrices $\underset{\sim}{A}_1$, $\underset{\sim}{A}_2$, $\underset{\sim}{A}_3$ and $\underset{\sim}{A}_4$ as basic blocks of block diagonal matrices of two blocks each and corresponding concatenated vectors $\underset{\sim}{C}_1$ and $\underset{\sim}{C}_2$) to the estimated parameter vector $\hat{\underset{\sim}{\beta}}$ from (3) as obtained by weighted least squares, estimates of the intraclass correlation coefficients and between condition correlation coefficients were obtained for the two types of practices: urban and non-urban. These results are shown in Table 2.

The proportion of patients performing personal preventive care with regard to brushing their teeth at least twice a day is similar to that of flossing at least once a week in group practices but not in non-group practices ($p<0.05$). The intraclass correlation coefficients ρ_1 and ρ_2 as well as the correlation coefficient for the two conditions for different patients, ν in both practice types are essentially zero. This can also be confirmed through statistical tests directed at quantities like $\ln(\hat{\lambda}_k/\hat{\pi}_k^2)$ and $\ln(\hat{\theta}/\hat{\pi}_1\hat{\pi}_2)$. The condition correlation coefficients in both practice types, group and non-group, are different from zero but not significantly different from each other.

Findings from this example suggest that accounting for the correlation structure of multiple patients within a clinic by working with (within) clinic means may not be necessary since correlations due to clustering are nearly negligible. Thus, an analysis based on viewing patients as the basic unit of analysis could be appropriate, although an analysis based on clinics was required to verify this.

REFERENCES

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). The analysis of categorical data by linear models. Biometrics 25, 489-504.

Kempthorne, W.J., and Koch, G.G. (1983). A general approach for the analysis of attribute data from a two-stage nested design: one and two treatments per cluster. Contributions to Statistics: Essays in Honor of Norman L. Johnson., P.K. Sen (Ed.), North Holland Publishing Company, 259-280.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33, 133-158.

Koch, G.G., Imrey, P.B. Singer, J.M., Atkinson, S.S., Stokes, M.E. (1985). Analysis of Categorical Data. In Collection Seminaire de Mathematiques Superieures 96, G. Sabidussi (Ed.). Les Presses de l'Université de Montreal, Montreal.

Marques, E.H. (1988). Analysis of Categorical Data from Longitudinal Studies of Subjects with Possibly Clustered Structures. Unpublished Dr. P.H. thesis, Department of Biostatistics, University of North Carolina, Chapel Hill, NC.

Table 1: Parameter Estimates and Standard Errors for the Dental Practice Clinics Example

Parameter	Estimator:	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\theta}$	$(\hat{\pi}_1-\hat{\lambda}_1)$	$(\hat{\pi}_2-\hat{\lambda}_2)$	$(\hat{\theta}-\hat{\phi})$
Group	Modulel	0.723	0.705	0.518	0.190	0.223	0.016
		(0.048)	(0.032)	(0.045)	(0.025)	(0.014)	(0.025)
	Module2	0.746	0.717	0.575	0.202	0.205	0.039
		(0.021)	(0.028)	(0.027)	(0.012)	(0.013)	(0.014)
Nongroup	Modulel	0.763	0.641	0.535	0.171	0.217	0.037
		(0.034)	(0.039)	(0.042)	(0.017)	(0.013)	(0.015)
	Module2	0.736	0.696	0.552	0.193	0.204	0.036
		(0.019)	(0.022)	(0.025)	(0.011)	(0.010)	(0.010)
	Module3	0.759	0.677	0.556	0.175	0.211	0.040
		(0.015)	(0.016)	(0.017)	(0.026)	(0.007)	(0.007)
$\hat{\beta}_{group}$		0.738	0.706	0.557	0.207	0.218	0.040
		(0.018)	(0.019)	(0.020)	(0.010)	(0.008)	(0.011)
$\hat{\beta}_{nongroup}$		0.753	0.678	0.552	0.180	0.210	0.039
		(0.011)	(0.012)	(0.013)	(0.006)	(0.005)	(0.005)

Table 2: Correlation Coefficient Parameter Estimates for the Dental Clinics Example

	Practice Type			
	Group		NonGroup	
Overall Parameter	(WLS) Estimate	s.e.	(WLS) Estimate	s.e.
$\rho_1=\rho\binom{brush}{diff.\ patients}$	-0.071	0.017	0.034	0.013
$\rho_2=\rho\binom{floss}{diff.\ patients}$	-0.050	0.016	0.039	0.014
$\delta=\delta\binom{brush,\ floss}{same\ patients}$	0.181	0.051	0.210	0.025
$\nu=\nu\binom{brush,\ floss}{diff.\ patients}$	-0.016	0.011	0.019	0.010