FITTING A MULTILEVEL LINEAR MODEL TO A SAMPLE OF CONTINGENCY TABLES USING GENERALIZED LEAST SQUARES

Minerva Montero*[1], Ernestina Castell**, Mario Miguel Ojeda***
*Instituto de Cibernética, Matemática y Física, La Habana, Cuba
**Universidad de La Habana. La Habana, Cuba
***Universidad Veracruzana. Veracruz, México.

ABSTRACT
In this work we propose a new approach for estimating multilevel models in contingency tables. This approach is based mainly on the use of the linear model as fundamental base to elaborate inference methods and the application of algorithms of iterative generalized least squares for the estimation of the fixed and random parameters. As illustration it is considered a logistic regression model in two levels and the proposed procedure is applied to a real problem from the literature. Finally, we carried out a brief simulation study to examine the behaviour of the estimates.

KEY WORDS: Logistic regression, Categorical data, Generalized Least Squares, Hierarchical linear models.

MSC : 62H17

RESUMEN
En este trabajo se propone un nuevo enfoque para la estimación de modelos multinivel en tablas de contingencia. Este enfoque se basa principalmente en el uso del modelo lineal como base fundamental para elaborar métodos de inferencia y el empleo de algoritmos de mínimos cuadrados generalizados iterativos para la estimación de los parámetros fijos y aleatorios. Como ilustración se considera un modelo de regresión logística en 2 niveles y se aplica el procedimiento propuesto a un problema tomado de la literatura. Finalmente se realiza un breve estudio de simulación para examinar el comportamiento de las estimaciones.

1. INTRODUCTION

The analysis of a sample of contingency tables plays an important role in many fields of research. For example, samples of contingency tables are obtained in educational studies, where the students are grouped in schools; in genetic studies, where the individuals are grouped in families; or in meta-analysis, where the data give rise to correlated responses within each study reported. The similarity among the individuals within the same group establishes a structure of 'intra-group' correlation and hence the independence assumption underlying the use of many standard statistical methods is violated. An appropriate approach is to use multilevel analysis (Goldstein, 1995), which explicitly takes into account within and between variance. The multilevel models also are referred as random coefficient models (Longford, 1995) or hierarchical models (Bryk and Raudenbush, 2002). General ideas, methodological formulation and guidelines for implementing hierarchical linear modeling are available in Ojeda *et al*. (1999).

The class of nonlinear multilevel models enables a more realistic process for modeling discrete data in many common situations, such as mentioned above. Various methods have been developed. Liang and Zeger (1986) propose the estimation of models with correlated binary responses using generalized estimating equations (GEE). Goldstein (1991) proposed a procedure for the analysis of multilevel nonlinear models using a linearization. Lee and Nelder (2001) introduced hierarchical generalized linear models which extend generalized linear models to include random components in the linear predictor with arbitrary distribution. There are several computational algorithms for fitting multilevel models. Longford

[1] minerva@icmf.inf.cu

(1994) gives a brief summary of computational algorithms for binary data. In addition, there is a number of specialized software implements multilevel modeling (Raundenbush *et al*., 2000; Rasbash *et al*., 2000; Hox, 2002).

The flexibility of multilevel models has resulted in an increasingly important role of this technique in both statistical theory and applications. However, some troubles are frequently encountered in the analysis of non-normal multilevel models. This class of models frequently leads to intractable likelihood functions (Breslow and Clayton, 1993). To avoid this, various approximated solutions have been proposed (Goldstein, 1991; Schall, 1991; Goldstein and Rasbash, 1996). In the binary case the applications of many approximated methods show occurrence of large biases in parameter estimates (Rodriguez and Goldman, 1995). In the recent years, several alternatives methods to eliminate the biases have been applied (Rodriguez and Goldman, 2001; Yun and Lee, 2004). Some of them are computationally intensive and therefore are not appropriate for exploratory work. Of particular value will be to explore new methods for providing the researcher with adequate tools for the efficient estimation of multilevel models for binary data.

In this work we propose a new estimation method based on the use of the Generalized Least Squares. Up to now the use of the Generalized Least Squares for analysis of categorical data has been restricted to the case where the model parameters are fixed (Grizzle, Starmer and Koch, 1969). In this paper we extend this strategy to the case of multilevel model for the analysis of a sample of contingency tables.

In this class of analysis a large number of functions of the unknown true cell probabilities may be of interest. The values of these functions become realizations of the dependent variable in a multilevel linear model. Dependencies between the observations are modeled via random effects. Once the model is formulated, it is possible to apply the asymptotic theory of estimation in the framework of the general linear model. The estimation procedure is based on iterative Generalized Least Squares.

The validity of the procedure presented in this paper is explored by means of the logit function. Other functions of the unknown true cell probabilities (see, *e.g*., Wickens, 1989) could also be considered permitting a greater flexibility in the data modeling for a sample of contingency tables. We develop a multilevel model, in which the heterogeneous treatment effects as measured by the log-odds ratio are regarded as random effects from a population of contingency tables. We motivate the application of the proposed procedure with a published example. A brief simulation study has been carried out to examine the efficiency of the estimates.

The use of the general linear model with categorical data permitted to unify the handling of an extensive range of problems that previously had been dealt by different techniques (see, *eg.*, Forthofer and Lehnen, 1981; Drew, 1985). In the same way, we feel that the procedure presented here can be utilized to develop a unified approach to modeling of a sample of contingency tables.

2. A MULTILEVEL MODEL FOR PROPORTIONS

Suppose a sample of *m* contingency tables where the rows, called subpopulations, represent *s* levels of an explanatory variable or combinations of levels of several explanatory variables. Independent random samples of size n_{ij} (i=1,...,s; j=1,...m) are selected from the rows,. The responses are classified according to two categories.

Let

$$\boldsymbol{\pi}_j = \left(\boldsymbol{\pi}_{1j}^T, \boldsymbol{\pi}_{2j}^T, \cdots \boldsymbol{\pi}_{sj}^T\right)^T$$

be, where

$$\boldsymbol{\pi}_{ij} = \left(\pi_{ij}, 1-\pi_{ij}\right)^T$$

is a vector of probabilities for the *j*-th table.

In the analysis of categorical data it is interesting to examine the relationship between a function of the probabilities and certain explanatory variables. The function can be simple (e.g., the same probability) or complex (e.g., a rank correlation, any coefficient between two response variables, etc). A set of different types of functions of interest may be represented in a relatively simple manner using matrix notation (Forthofer and Koch, 1973). In this paper we will be mainly concerned with the logit function. Now we consider a 2-level hierarchical data structure assuming a set of level 1 units (subpopulations) nested within level 2 units (tables).

Logit response models for the data of the *j*th table have response functions of the form

$$\boldsymbol{F}_j(\boldsymbol{\pi}_j) = \boldsymbol{B}_j \log(\boldsymbol{\pi}_j) ,$$

where the elements of $\boldsymbol{B}_j$ are the coefficients of the natural logarithms of the vector $\boldsymbol{\pi}_j$. In section 5 we will present discusses an application of the logit function.

Once the function $\boldsymbol{F}_j(\pi_j)$ has been specified, it can be used as dependent variable in a multilevel linear model. In 2-level models, separate level 1 models are developed for each of the *m* tables.

For the *j*th table the level 1 model can be expressed as:

$$\boldsymbol{F}_j(\pi_j) = \boldsymbol{X}_j \boldsymbol{\beta}_j , \qquad j=1,2,...,m; \qquad (1)$$

where $\boldsymbol{X}_j$ is a *s* x *t* design matrix with rank *t* and $\boldsymbol{\beta}_j$ is a *t* x 1 vector of unknown coefficients.

The variability of the *m* coefficients

$$(\beta_{1k,\cdots},\beta_{mk})$$

pertaining to the *k*th variable (*k=1, ... ,t*) can be explained by an additional set of variables $Z_1, Z_2, \cdots, Z_q$ measured at level 2. It then follows that:

$$\beta_{jk} = \boldsymbol{Z}_{jk}^T \boldsymbol{\Gamma}_k + \boldsymbol{u}_{jk} , \qquad j=1,2,..., m; \qquad (2)$$

where $\boldsymbol{Z}_{jk}$ are the values of the q_k variables at level 2 in the *j*th table, $\boldsymbol{\Gamma}_k$ is the $q_k \times 1$ vector of the coefficients associated with explanatory variables at level 2 and u_{jk} are the non observable level 2 random errors.

The equation (2) can be more succinctly expressed as:

$$\boldsymbol{\beta}_j = \boldsymbol{Z}_j \boldsymbol{\Gamma} + \boldsymbol{u}_j , \qquad j=1, 2,..., m; \qquad (3)$$

where

$$\boldsymbol{Z}_j = \mathrm{diag}(\boldsymbol{Z}_{j1}^T, \cdots \boldsymbol{Z}_{jt}^T)$$

is a $t \times \mathrm{Q}$ block-diagonal matrix, $\mathrm{Q} = q_1 + q_2 + \cdots q_t$,

$$\boldsymbol{\Gamma} = \left(\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \cdots, \boldsymbol{\Gamma}_t\right)$$

is the $\mathrm{Q} \times 1$ vector of coefficients and $\boldsymbol{u}_{\mathrm{j}}$ is the $t \times 1$ vector of level 2 random errors.

Substituting equation (3) into equation (1) we obtain a single expression for the model given by

$$F\left(\pi_j\right) = \boldsymbol{A}_j \boldsymbol{\Gamma} + \boldsymbol{X}_j \boldsymbol{u}_j, \quad j=1,2,..., m;$$

where $\boldsymbol{A}_j = \boldsymbol{X}_j \boldsymbol{Z}_j$.

3. THE ESTIMATION PROCEDURE

Let $\boldsymbol{p}_j$ be the vector of observed proportions, given in the same way as $\boldsymbol{\pi}_j$, associated with the sample from the *j*th table.

The model for the observed proportions is:

$$\boldsymbol{F}\left(\boldsymbol{p}_j\right) = \boldsymbol{A}_j \boldsymbol{\Gamma} + \boldsymbol{X}_j \boldsymbol{u}_j + \boldsymbol{e}_j, \quad j=1,2,..., m; \qquad (4)$$

where $\boldsymbol{e}_j$ is a $s \times 1$ vector of level 1 random errors.

Now we assume that:

$$\mathrm{E}\left(\boldsymbol{e}_j\right) = \boldsymbol{0}; \quad \mathrm{E}\left(\boldsymbol{u}_j\right) = \boldsymbol{0},$$
$$\mathrm{E}\left(\boldsymbol{e}_j \boldsymbol{e}_j^T\right) = \boldsymbol{\Omega}_{\boldsymbol{e}_j} \quad , \quad E\left(\boldsymbol{u}_j \boldsymbol{u}_j^T\right) = \boldsymbol{\Omega}_{u_j} \quad \text{and} \quad E\left(\boldsymbol{e}_j \boldsymbol{u}_j^T\right) = \boldsymbol{0}.$$

The model (4) can then be expressed more compactly as:

$$\boldsymbol{F}\left(\boldsymbol{p}\right) = \boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{X}\boldsymbol{u} + \boldsymbol{e}, \qquad (5)$$

where

$$\boldsymbol{F}\left(\boldsymbol{p}\right) = \left(\boldsymbol{F}\left(\boldsymbol{p}_1\right)^T, \boldsymbol{F}\left(\boldsymbol{p}_2\right)^T, \cdots, \boldsymbol{F}\left(\boldsymbol{p}_m\right)^T\right)^T,$$

$$\boldsymbol{A} = \left(\boldsymbol{X}_1 \boldsymbol{Z}_1, \boldsymbol{X}_2 \boldsymbol{Z}_2, \cdots, \boldsymbol{X}_m \boldsymbol{Z}_m\right)^T$$

and

$$\boldsymbol{X} = \mathrm{diag}\left(\boldsymbol{X}_j\right)$$

is a diagonal block matrix with $\boldsymbol{X}_j$ in the *j*th diagonal block. The index *j* (*j*=1,2,..., *m*) is some what redundant to $\boldsymbol{X}_j$, since such a variable is constant within tables. Here

$$\boldsymbol{e} = \left(\boldsymbol{e}_1^T, \boldsymbol{e}_2^T \cdots, \boldsymbol{e}_m^T\right)^T$$

and

$$\boldsymbol{u} = \left(\boldsymbol{u}_1^T, \boldsymbol{u}_2^T \cdots, \boldsymbol{u}_m^T\right)^T$$

We assume that

$$\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}; \qquad \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0},$$
$$E(\boldsymbol{e}\boldsymbol{e}^T) = \boldsymbol{\Omega}_e, \quad E(\boldsymbol{u}\boldsymbol{u}^T) = \boldsymbol{\Omega}_u \text{ and } E(\boldsymbol{e}\boldsymbol{u}^T) = \boldsymbol{0},$$

where

$$\boldsymbol{\Omega}_e = \mathrm{diag}\left[\boldsymbol{I}_{e_j} \otimes \boldsymbol{\Omega}_{e_j}\right] \text{ and } \boldsymbol{\Omega}_u = \mathrm{diag}\left[\boldsymbol{I}_{u_j} \otimes \boldsymbol{\Omega}_{u_j}\right],$$

with $\otimes$ representing the Kronecker product.

Assuming independence, we then say that the corresponding variance-covariance matrix has the general form:

$$\boldsymbol{V}_{\boldsymbol{F}(\boldsymbol{p})} = \boldsymbol{X}\boldsymbol{\Omega}_u \boldsymbol{X}^T + \boldsymbol{\Omega}_e.$$

It should be noted now that the model (5) is a special case of the General Linear Model:

$$\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{e}^*,$$

where

$$\boldsymbol{e}^* = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{e}, \ \mathrm{E}(\boldsymbol{e}^*) = 0$$

and

$$E\left(\left(\boldsymbol{e}^*\right)\left(\boldsymbol{e}^*\right)^T\right) = \boldsymbol{V}_F. \qquad (6)$$

For brevity we write: $\boldsymbol{V}_F = \boldsymbol{V}_{F(p)}$.

If the variance-covariance matrix is known, $\boldsymbol{\Gamma}$ can be estimated by the Generalized Least Squares (GLS), i.e.

$$\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{A}^T \boldsymbol{V}_F^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{V}_F^{-1} \boldsymbol{F}(\boldsymbol{p}) \qquad (7)$$

However, $\boldsymbol{V}_F$ is usually unknown and no reasonable structure can be postulated because of lack of any information about it. A common practice is to replace it in the expression (6) by an estimate $\hat{\boldsymbol{V}}_F$. This is done iteratively.

The procedure starts with initial values of the fixed parameters. We propose those obtained from Generalized Least Squares (GLS) for categorical data ignoring the random errors at level 2 (see, *eg.* Wickens, 1989). A consistent estimator for the covariance matrix of $\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{B}\ln(\boldsymbol{p})$ (Forthofer and Koch, 1973) is the $sm \times sm$ matrix:

$$\hat{\boldsymbol{V}}_{\boldsymbol{F}} = \boldsymbol{B}\boldsymbol{D}^{-1}\left[\hat{\boldsymbol{V}}(\boldsymbol{p})\right]\boldsymbol{D}^{-1}\boldsymbol{B}^T,$$

where

$\boldsymbol{B} = \text{diag}(\boldsymbol{B}_1, \cdots, \boldsymbol{B}_m)$,

$\boldsymbol{D}$ contains the elements of the vector

$\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots \boldsymbol{p}_m)^T$

on the main diagonal, and $\hat{\boldsymbol{V}}_p$ is an estimation of the covariance matrix of $\boldsymbol{p}$.

If we consider the *s* subpopulations of the *j*th table as being uncorrelated with one another and we assume that the observed data follow a binomial distribution; then, a consistent estimator for the covariance matrix of $\boldsymbol{p}_j$ is the matrix:

$$\hat{\boldsymbol{V}}_j(\boldsymbol{p}_j) = \text{diag}\left(\hat{\boldsymbol{V}}_{1j}(\boldsymbol{p}_{ij}), \hat{\boldsymbol{V}}_{2j}(\boldsymbol{p}_{ij}), \cdots, \hat{\boldsymbol{V}}_{sj}(\boldsymbol{p}_{ij})\right),$$

with the matrices

$$\hat{\boldsymbol{V}}_{ij}(\boldsymbol{p}_{ij}) = \frac{1}{n_{ij}} \begin{bmatrix} p_{ij}(1-p_{ij}) & -p_{ij}(1-p_{ij}) \\ -p_{ij}(1-p_{ij}) & p_{ij}(1-p_{ij}) \end{bmatrix}, \qquad i=1,2,\ldots,s.$$

Then, under the assumed independence, the estimated covariance matrix of the vector *p* is:

$$\hat{\boldsymbol{V}}(\boldsymbol{p}) = diag\left(\hat{\boldsymbol{V}}_1(\boldsymbol{p}_j), \hat{\boldsymbol{V}}_2(\boldsymbol{p}_j), \cdots, \hat{\boldsymbol{V}}_m(\boldsymbol{p}_j)\right).$$

Once obtained suitable starting value for the fixed parameters we carry out an iterative Generalized Least Squares analysis, analogous to the described in Goldstein (1995) to fit (5). In each iteration we use the current estimates of the fixed and random parameters until convergence.

4. A MOTIVATING EXAMPLE

At present, it is remarkable the use of complexes data structures in scientific researches of different branches of knowledge. This tendency has stimulated the interest of the scientific community for multilevel modelling and as a consequence the application areas of the multilevel models have become considerably multiplied in the last years.

In the area of the analysis of contingency tables the theoretical contribution on the multilevel modelling is not very abundant; however, in the last years the use of multilevel logistic regression has been increased for the analysis of such data [Efron, 1996, Hatzel et al, 2001, Lee and Nelder, 2002]. An important area of application of these models is meta-analysis.

Meta-analysis can be considered as a multilevel statistical problem, because information within the studies is combined in presence of a potential heterogeneity between studies. (Thompson et al, 2001). In this paper a problem of meta-analysis de clinical trials with binary outcomes is put into the proposed multilevel framework. The standard models for a set of tables with binary response suppose a common effect within or between tables, given, for example, by a certain type of odds ratio. However, in practice, there exists heterogeneity among such odds ratios. The multilevel models treat the true log odds ratios of the specific studies as a sample with any unknown mean and standard deviation.

As a numerical illustration we use the data from Sacks *et al*. (1990) of 41 randomized trials of a new surgical treatment for stomach ulcers. With binary outcomes, the data form several 2×2 contingency tables.

n_{1j}^{0} and n_{1j}^{1} are the respective numbers of non occurrence and occurrence of recurrent bleeding in the traditional surgery (treatment group) and n_{2j}^{0} and n_{2j}^{1} the non occurrences and occurrences for the new surgery (control group).

Let θ_j be the specific log-odds ratio of the *j*th trial:

$$\theta_j = log\left(\frac{P_j(\text{occurrence}/\text{treatment})/P_j(\text{no occurrence}/\text{treatment})}{P_j(\text{occurrence}/\text{control})/P_j(\text{no occurrence}/\text{control})}\right)$$

where P_j represents the probability for the *j*th trial.

The estimated log- odds ratio for trial *j,*

$$\hat{\theta}_j = log\left(\frac{n_{1j}^{0}}{n_{1j}^{1}}\Big/\frac{n_{2j}^{0}}{n_{2j}^{1}}\right)$$

measures the excess of occurrence of treatment over control. In the ulcer data, the estimated log-odds ratios vary from $-\infty$ to ∞, indicating heterogeneities among trials. The potential heterogeneity can be explored by using models that include parameters describing the variability in odds ratios among trials. Standard models for a set of contingency tables assume a common effect within or between tables described by a certain type of odds ratio. Other more realistic models (Hartzel et al., 2001; Lee and Nelder, 2001) as the multilevel model discussed in this paper, uses random effects terms to describe the variability in conditional associations. We consider a multilevel logistic regression model, in which the heterogeneous treatment effects as measured by the log-odds ratio are regarded as random effects from a population of contingency tables.

For illustration we consider the following multilevel model:

$$\text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{10}x_{ij} + u_{1j}x_{ij} \qquad (8)$$

where π_{ij} is the expected proportion of occurrences for the patients exposed to the *i*th type of surgery (0=traditional surgery, 1=new surgery) in the *j*th experimental population (level 2 unit), x_{ij} is a binary variable describing the structure of the level-1 units or subpopulations (traditional surgery = 0, new surgery = 1). γ_{00} and γ_{10} are the overall mean intercept and the treatment effect, respectively. At level 2, the random errors u_{1j} represent the heterogeneity among the treatment effects of individual trials.

The log-odds ratio between the responses of the individuals of two subpopulations i and i' for a table j is:

$$\text{logit}(\pi_{ij}) - \text{logit}(\pi_{i'j}) = \gamma_{10} + u_{1j}.$$

The fit of the multilevel model provides a single summary such as an estimated mean and standard deviation of log-odds ratios for the population of tables. That is, γ_{10} is the expected trial-specific log-odds ratio between treatment and response and

$$\mathrm{var}(u_{1j}) = \sigma_u^2$$

describes the variability in these log-odd ratios. So the odds ratio is a random variable rather than a fixed parameter, and this should be reflected when interpreting the model. However, in practice, it is not possible to condition on the random effects because these are unobservable. Larsen *et al,* (2000) proposed median-based interpretation for both fixed effects and random effects.

In this paper the main focus is on the accuracy of the estimates of the parameters (γ_{00}, γ_{10} and $\mathrm{var}(u_{1j}) = \sigma_u^2$) following the proposed procedure[1] rather that describing the heterogeneity of association.

In the next section a brief simulation study has been carried out to examine the efficiency of the estimates of the fixed and random parameters.

5. ANALYSIS OF SIMULATED DATA

Using the example described in Section 5 as a framework for the simulation we generated data for a 2-level hierarchical structure and fit the model (8) by the proposed procedure:

i) to compare the parameter estimates to the true value.

ii) to assess the effect of two different values of the variance on parameter estimates.

For the simulations we kept the number of tables and subpopulation sizes as in the example. According estimates produced by the procedure for the example data, we chose the magnitudes for the parameters, γ_{00}, γ_{10} and σ_u^2. In data generating the values of fixed parameters γ_{00} and γ_{10} were set to 0.5 and 1.0, respectively. Small and large level 2 variances were assumed ($\sigma_u^2 = 0.5$ and $\sigma_u^2 = 1.0$); therefore, there are two different designs in the study.

The independent random effects u_{1j} were generated from a normal distribution with mean 0 and the variances before specified. The explanatory variable is fixed following the requirements of the contingency tables. $\mathrm{logit}(\pi_{ij})$ is obtained by adding the fixed part and level 2 random effects. Finally, the values of the response p_{ij} are generated from a binomial distribution with parameter π_{ij} and n_{ij}. For each condition 100 simulated data sets were generated. The estimation procedure converged in all 200 simulated data sets.

Table 1 displays for each parameter the *true* value and the values of the estimated fixed and random parameter averaged over the 100 simulations conducted for each of two designs. The mean of the correspondent mean squared errors (MSE), and standard deviations of the estimates are also given.

[1] The analyses were carried out using Matlab [Math Works Inc 2000], MLwiN [Rassbash et al. 2000] and Statistica [StatSoft Inc 2001]. The interested ones can request the author (minerva@icmf.inf.cu) the executable code in Matlab specially made for this paper.

Table 1: Mean values of estimates for 100 simulated data sets for model (8) assuming $\sigma_u^2 = 0.5$ and $\sigma_u^2 = 1.0$

Parameters	True σ_u^2 = 0.5 Estimate (s.e) (MSE)	True σ_u^2 = 1.0 Estimate (s.e) (MSE)
γ_{00}	0.519 (0.064) (0.004)	0.519 (0.070) (0.005)
γ_{10}	1.046 (0.173) (0.032)	1.008 (0.203) (0.041)
σ_u^2	0.908 (0.198) (0.206)	1.301 (0.287) (0.173)

As we can see from Table 1 the procedure produced reasonably unbiased estimates for the fixed parameters γ_{00} and γ_{10}. It is clear that the fixed parameter estimates are close to their true value, that is, the bias of the estimates is small. Table 1 show that the estimation procedure results in very small MSE for the fixed parameters.

Note, however, that biased estimates are expected for the random parameters. We found that, on the situations considered, the proposed approach behaves poorly in estimating random parameter, and this situation is particularly bad when the variance of the random effects is small, the random parameters estimates are subject to large biases. When the variance is large, the mean of the estimates improves, nearing to the true value. The values of MSE reported show that behavior of the procedure is better for estimating the fixed parameters. We see that, in general sense, the procedure overestimate the random parameters.

Theoretical reasons that justify the unbiased estimations for the random parameters must be analyzed. When we refer to contingency tables empty cells and sparse tables can cause problems with severe bias in estimation of descriptive measures such as odds ratios (Agresti, 2002) On the other hand we know in statistical procedure, the sample size can strongly influence the results. Especially, when we use multilevel model, it is preferable to have large sample sizes (Hox and Mass, 2002). These facts suggest some cautionary remarks and supplementary researches are necessary.

6. CONCLUSIONS

In this paper we consider the linear model as a tool to formulate multilevel models for analyses a sample of contingency tables and we introduce an estimation procedure that may be applied to fit these models. The proposed approach relegates the analysis of a sample of contingency tables to a class of problem that can be handled by Generalized Squared Least. One of the main advantages of this procedure is its similarity with the case of the multilevel linear model; hence it can be used in situations where other methods impose the solution of complicated mathematical expressions.

We focused on the application of the proposed procedure to logit response models but this approach is more general and can be used for handle other functions of the probabilities. However, a further analysis of more complex models and extreme data sets is necessary to recommend this approach as a unified approach of modeling of a sample of contingency tables.

A Generalized Least Squares analysis for hierarchical categorical data is simple but practical considerations become this approach less appropriated for sparse data. The sample response functions may then be ill-defined or have a singular estimated covariance matrix. Of particular value will be do a further research in this area. An additional analysis concerned with the effects of the sample size and different variance distributions on the efficiency of estimates also is necessary. Moreover it is important to say that some numerical modifications can be analyzed for improve the estimates of the random parameters

Received September 2005

REFERENCES

AGRESTI, A. (2002). **Categorical Data Analysis**. 2nd. Ed. Wiley, New York.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear models. **Journal of the American Statistical Association**, 88**,** 9-25**.**

BRYK, A. S. and RAUDENBUSH, S. W. (2002). **Hierarchical Linear Models: Applications and Data Analysis Methods**. Second Edition Sage Publications, Thousand Oaks, California, USA.

DREW, J. H. (1985). Performance of the weighted least squares approach to categorical data analysis. **Communication in Statistics Theory and Methods**, 14(8), 1963-1979.

EFRON B. (1996) Empirical bayes methods for combining likelihoods. **Journal of the American Statistical Association.** 91(434), 538-565.

FORTHOFER, R. N. and KOCH, G. G. (1973). An analysis for compounded functions of categorical data. **Biometrics,** 29**,** 143-159.

FORTHOFER, R. N. and LEHNEN, P. (1981). **Public Program Analysis, A New Categorical Data Approach**. Lifetime Learning Publications. Belmont, California.

GRIZZLE, J. E., STARMER, F. and KOCH, G. (1969). Analysis of categorical data by linear models. **Biometrics**, 25, 489-504.

GOLDSTEIN, H. (1991). Nonlinear multilevel models, with an application to discrete response data. **Biometrika***.* 78(1), 45-51.

GOLDSTEIN, H. (1995). **Multilevel Statistical Models***.* 2nd. Ed. Halsted Press, New York.

GOLDSTEIN, H. and RASBASH, J. (1996). Improved approximations for multilevel models with binary responses. **Journal of the Royal Statistical Society**, Series B, 159, 505-513.

HARTZEL J., LIU I-M. and AGRESTI A. (2001). Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. **Computational Statistics & Data Analysis** 35, 429-499.

HOX, J. J. and MAAS, C. J. M. (2002). Sample sizes for multilevel modeling. In Blasius, J., Hox, J., De Leeuw, E. and Schmidt, P. (eds.) (2002). **Social Science Methodology in the New Millennium**. Proceedings of the Fifth International Conference on Logic and Methodology. Second expanded edition. Opladen, RG: Leske + Budrich Verlag (CD-ROM).

LARSEN, K., PETERSEN, J. H., BUDTZ-JORGENSEN, E. and ENDAHL, L. (2000). Interpreting parameters in the logistic regression model with random effects. **Biometrics**, 56, 909-914.

LEE, Y. and NELDER, J. A. (2001). Hierarchical generalized linear models: a synthesis of generalized linear model, random-effect models and structured dispersions. **Biometrika**, 88, 987-1006.

LEE Y. and NELDER J.A. (2002). Analysis of ulcer data using hierarchical generalized linear models. **Statistics in Medicine**, 21, 191-202.

LIANG, K. Y. and ZEGER, S. (1986). Longitudinal data analysis using generalized linear models. **Biometrika***,* 73(1), 13-22.

LONGFORD, N. (1994). Logistic regression with random coefficients. **Computational Statistics and Data Analysis**, 97, 1-15.

LONGFORD, N. (1995). Random coefficient models. In Arminger, G., Cogg, C. C., Sobel, M. E. (eds.), ***Handbook of Statistiscal Modeling for the Social and Behavioral Sciences***, Plenum Press, New York, 519-577.

MATH WORKS (2000). **MATLAB. The language of technical computing**. version 6.0.088. realese 12, september 22.

OJEDA, M. M., SAHAI, H. and JUÁREZ-CERRILLO, S. F. (1999). Multilevel data analysis with hierarchical linear models. **Statistica Aplicata**, 11(4), 577-590.

RASBASH, J., BROWNE, W., GOLDSTEIN, H., YANG, M., PLEWIS, I., HEALY, M., WOODHOUSE, G., DRAPER, D., LANGFORD, I. and LEWIS, T. (2000). **A User Guide to MLwiN**. Multilevel Models Project. University of London.

RAUNDENBUSH, S. W., BRYK, A. S, CHEANG, Y. F., and CONGDON, R. (2000). **HLM. 5. Hierarchical Linear and Nonlinear Modeling**. Scientific Software International, Chicago.

RODRIGUEZ, G. and GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary response. **Journal of the Royal Statistical Society,** Series A, 158, 73-89.

RODRIGUEZ, G. and GOLDMAN, N. (2001). Improved estimation procedures for multilevel models with binary response: a case-study**. Journal of the Royal Statistical Society,** Series A, 164, 339-355.

SACK, H. S., CHALMERS, T. C., BLUM, A. L., BERRIER, J. and PAGANO, D.(1990). Endoscopic hemostasis, an effective therapy for Bleeding Peptic Ulcers. **Journal of the American Medical Association**, 264, 496-499.

SCHALL, R. (1991). Estimation in generalized linear models with random effects. **Biometrika***,* 40**,** 719-727**.**

STATSOFT, INC. (2001) **STATISTICA** (data analysis software system), versión 6, www.statsoft.com.

THOMPSON, S., TURNER, R. and WARN D. (2001) Multilevel models for meta-analysis, and their application to absolute risk. **Statistical Methods in Medical Research**, 10 (6), 375-392.

WICKENS, T. D. (1989). **Multiway Contingency Tables Analysis for the Social Sciences**. Ed. Lawrence Erlbaum Associates. New Jersey.

YUN, S. and LEE, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. **Computational Statistics and Data Analysis***,* 45, 639-650.