Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 139 a 166

Aggregation of Explanatory Factor Levels in a Binomial Logit Model: Generalization to the Multifactorial Unsaturated Case

La agregación de niveles en un factor explicativo del modelo logit binomial: generalización al caso multifactorial no saturado

Ernesto Ponsot-Balaguer[1,a], Surendra Sinha[2,b], Arnaldo Goitía[2,c]

[1]Departamento de Estadística, Facultad de Ciencias Económicas y Sociales (FACES), Universidad de Los Andes (ULA), Mérida, Venezuela

[2]Programa de Doctorado en Estadística, Instituto de Estadística Aplicada y Computación (IEAC), FACES-ULA, Mérida, Venezuela

Abstract

We discuss a situation in which, once a logit model is fitted to the data in a contingency table, some factor levels are grouped. Generally, researchers reapply a logit model on the pooled data, however, this approach leads to the violation of the original distributional assumption, when the probabilities of success of the random variables of aggregation differ. In this paper we suggest an alternative procedure that operates under the unsaturated, multifactorial, binomial, logit model. Based on asymptotic theory and taking advantage of the decrease in the variance when the correct distributional assumption is made, the suggested procedure significantly improves the estimates, reduces the standard error, produces lower residuals and is less likely to reject the goodness of fit test on the model. We present the necessary theory, the results of an extensive simulation designed for this purpose, and the suggested procedure contrasted with the usual approach, through a complete numerical example.

***Key words*:** Contingency tables, Generalized linear model, Levels sets, Logit model.

Resumen

Se discute la situación en la que, una vez ajustado un modelo logit a los datos contenidos en una tabla de contingencia, se selecciona un factor cualquiera de los participantes y se agregan algunos de sus niveles. Generalmente los investigadores proceden a postular nuevamente un modelo logit

[a]Associate Professor. E-mail: ernesto@ula.ve
[b]Professor. E-mail: sinha32@yahoo.com
[c]Professor. E-mail: goitia@ula.ve

sobre los datos agrupados, sin embargo, este proceder conduce a la violación del supuesto distribucional original, cuando las probabilidades de éxito de las variables aleatorias de la agregación, son disímiles. En este trabajo se sugiere un procedimiento alternativo que opera en el marco del modelo logit binomial no saturado, multifactorial. Con base en la teoría asintótica y aprovechando la disminución en la varianza cuando se postula el modelo distribucional correcto, el procedimiento sugerido mejora apreciablemente las estimaciones, reduce el error estándar, produce valores residuales más cercanos al cero y menores probabilidades de rechazo en la prueba de bondad del ajuste del modelo. Sustentan tales afirmaciones tanto los desarrollos teóricos necesarios, como los resultados de una extensa simulación diseñada al efecto. También se expone el procedimiento sugerido contrastado con el habitual, mediante un ejemplo numérico completo.

***Palabras clave*:** conjuntos de niveles, modelo lineal generalizado, modelo logit, tablas de contingencia.

1. Introduction

Assume a Bernoulli phenomenon, that is, an experiment whose outcome regarding an individual can only be a success or a failure (or equivalently, the presence or absence of a feature, membership to a particular group or other similar forms). Assume also that a researcher wants to test whether the outcome of the experiment is determined by certain characteristics, measurable in each individual and possibly the direction of the relationship if it exists. For this, the researcher collects data from a previous study or by sampling, for example, and builds a contingency table including the levels of the factors under study, the number of cases in which tests the response of interest (success or failure) and total individuals examined, for each combination of these levels.

A statistical model is related to a contingency table in order to capture the essence of the phenomenon of study in a manageable way and to draw valid conclusions for the population regarding about the causal relationships between the observed response and the measured characteristics.

Now, assuming that the responses are distributed as independent binomials, a model that postulates a certain function of the probability of success of the response and relates linearly with the measured characteristics in individuals looks suitable for analysis. Thus, taking the probit model as a precursor, are the logistic regression for continuous variables and its counterpart, the logit model for categorical explanatory variables or factors introduced by Joseph Berkson in 1944 (Hilbe 2009, p. 3).

In the case of a logit model, the link function considered is $\text{logit}(p) = \log(\frac{p}{1-p})$. Applied to the probability of success p of a Bernoulli random variable, $\text{logit}(p)$ represents the logarithm of the possibility. This, in turn, is defined as the ratio between the probability of success p and its complement, the probability of failure $1 - p$.

Moreover, suppose that the researcher, after fitting a logit model to the data, decides to add some levels of one factor, and repeat the analysis, i.e., fit a new logit model on a contingency table resulting from the aggregation.

It happens that in reiterating a logit model on a second contingency table, with grouped levels of the factors, generally the original binomial assumption is violated, with important implications on the estimated variances (Ponsot, Sinha & Goitía 2009)[1].

In seeking to address this problem and keep the situation under the generalized linear model frame (Nelder & Wedderburn 1972), this paper postulates the problem of aggregation of factor levels in a broad context, i.e., in multifactorial unsaturated logit model situation, and proposes and demonstrates some theorems needed to suggest a procedure, alternative to the usual, that takes advantage of the true variance of the random variables added. It is shown theoretically by asymptotic means, and by simulation, that the suggested procedure is appropriate and in many cases, better than the usual procedure.

This paper continues with the next section presenting a summary of the main background of the work. The third section presents the problem and its solution, including the theorems that support the suggested procedure and their proofs. The fourth section illustrates the suggested procedure with a numerical example. The fifth section summarizes the extensive simulation results comparing the two procedures (normal and suggested). The sixth section is devoted to conclusions, and the work ends with the acknowledgments, references and a brief appendix on the design matrix for the saturated and unsaturated models.

2. Backgrounds

Ponsot et al. (2009) present the problem of aggregation levels of an explanatory factor in the saturated logit model. The authors study the affectation of the binomial distributional assumption and show that, once factor levels are grouped, which involves adding independent binomial random variables (RV's), in the general case where the probabilities of success are different, the random variable (RA) resulting from the aggregation does not follow a binomial distribution. Proper distribution is as follows:

Let X_1 and X_2 be two independents RV's such that $X_1 \sim \text{Bin}(n_1, p_1)$ and $X_2 \sim \text{Bin}(n_2, p_2)$ with $n_1 \leq n_2$. Then, the RV $Z = X_1 + X_2$ is distributed as follows:

$$\text{P}[Z = k] = \left(\frac{p_1}{1 - p_1}\right)^k (1 - p_1)^{n_1} (1 - p_2)^{n_2} S(k) \tag{1}$$

[1]This is central in the doctoral thesis of first author (Ponsot 2011), one of whose results is this paper.

where

$$S(k) = \begin{cases} \sum_{i=0}^{k} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = 0, \ldots, n_1 \\ \sum_{i=k-n_1}^{k} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_1+1, \ldots, n_2 \\ \sum_{i=k-n_1}^{n_2} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_2+1, \ldots, n_1+n_2 \end{cases}$$

The authors also prove that as the difference between the probabilities of success of the RV's involved in the aggregation increases, the correct variance of the resulting RV [distributed as in (1)], becomes less than the variance calculated assuming that the RV resulting is binomially distributed.

In general, let $X_1, X_2, \ldots, X_a$ be independents RV's such that $X_i \sim \text{Bin}(n_i, p_i)$ for $i = 1, \ldots, a$. Let $X_{a-k+1}, X_{a-k+2}, \ldots, X_a$, the k last RV's being added $(1 < k < a)$ forming the RV $Z = X_{a-k+1} + X_{a-k+2} + \cdots + X_a$. Due to the independence of the originals RV's, $\text{V}[Z]$ is the simple sum of $\text{V}[X_i]$ for $i = a-k+1, \ldots, a$. However, if Z is assumed (incorrectly) binomial, the variance (V_{Bin}) should be calculated differently, making assumptions about the probability of success. By studying the difference $\Delta\text{V} = \text{V}_{\text{Bin}}[Z] - \text{V}[Z]$, it follows that:

$$\Delta\text{V} = \frac{\sum_{i=a-k+1}^{a-1} \sum_{j=i+1}^{a} n_i n_j (p_i - p_j)^2}{\sum_{i=a-k+1}^{a} n_i} \tag{2}$$

Clearly $\Delta\text{V} \geq 0$, then the correct variance is generally smaller than the binomial (equal if and only if $p_i = p_j, \forall i, j$).

Based on these facts and using arguments of asymptotic nature, these authors suggest an alternative procedure to the reiteration of the logit model fitting when factor levels are added. This procedure improves the precision of the estimates, using the true variance of the RV's involved.

Now, as mentioned, the entire development applies in the univariate situation and saturated model exclusively, leaving pending the study of unsaturated logit model in the multifactorial situation. Such an extension is the aim in this work.

Besides, it must be mentioned that there are different courses of action than the asymptotic approach to the problem. For example, we may include (1) as a factor in the likelihood function; however, clearly an analytically intractable expression is obtained, and therefore, very difficult to derivate.

Another possible course of action is to postulate the exact distribution for each given data set, from the contingency table. This way to avoid the assumption of

binomial populations, leading to the hypergeometric distribution and combinatorial analysis. This path has been explored successfully in the theory of generalized linear model; however, it is not of very frequent application because it imposes considerable computational challenges.

It should also be mentioned that the aggregation of factor levels and subsequent repetition of a logit setting is of common practice among statisticians. Hosmer & Lemeshow (2000, p. 136) suggested as a strategy to overcome the drawback of responses with very low or no representation in the contingency table. Examples abound in which the researcher adds factor levels, simply to reduce the complexity of the analysis or because wish to concentrate *posteriori* on some levels and try the other anonymously. An exercise that illustrates this approach can be seen in Hilbe (2009, pp. 74 y 88). In his text the author develops models from the Canada's National Cardiovascular Registry, using a first opportunity to age with four levels as an explanatory factor, and another time, this factor grouping up to only two levels. Another example of the latter type is shown in Menard (2010). In his text the author uses data from the National Center for Opinion Research (University of Chicago, USA), taken from the General Social Survey. In some instances, operates with three or even more levels for the factor "race" (Caucasian, African descent and others), while in alternative examples, it does so with only two levels (not Caucasian and other), grouping the original levels.

3. The Problem and Its Solution

Let T a contingency table for a binary response with s crossed factors $A_1, A_2, \ldots, A_s$, each with $t_1, t_2, \ldots, t_s$ levels, respectively. Each combination of factor levels has an observed response ($y_{i_1 i_2 \cdots i_s}$) as the number of successes, all assumed independently binomially distributed with a total number of observations ($n_{i_1 i_2 \cdots i_s}$), $i_j = 1, \ldots, t_j$ and $j = 1, \ldots, s$. On T, an unsaturated logit model is fitted with the reference parameterization [see for example Rodríguez (2008, cap. 2, p. 29) or SAS Institute Inc. (2004, p. 2297)], then let:

$$\eta_{i_1 i_2 \cdots i_s} = \text{logit}(p_{i_1 i_2 \cdots i_s}) = \boldsymbol{x}^T_{i_1 i_2 \cdots i_s} \boldsymbol{\beta}, \quad \begin{array}{l} i_j = 1, \ldots, t_j; \\ j = 1, \ldots, s \end{array} \tag{3}$$

be the univariate version of the logit model for crossed factors $A_1, \ldots, A_s$. To simplify the treatment of the subscripts of the model, assume that each combination of factors is reindexed orderly, making it correspond to a single value as:

$$1 \equiv (1, 1, \ldots, 1), \ldots, i \equiv (i_1, i_2, \ldots, i_s), \ldots, k \equiv (t_1, t_2, \ldots, t_s)$$

so as to produce $k = t_1 \times \cdots \times t_s$ sequenced indexes. In turn, the response is reindexed as $y_1, y_2, \ldots, y_k$ and so the totals as $n_1, n_2, \ldots, n_k$. Then (3) is expressed in the usual way as:

$$\eta_i = \text{logit}(p_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad i = 1, \ldots, k \tag{4}$$

In (4) $\boldsymbol{x}_i^T$ is the row vector corresponding to the combination of levels $i_1, i_2, \ldots, i_s$ of the design matrix $\boldsymbol{X}_{k\times m}$ and $\boldsymbol{\beta}_{m\times 1}$ is the vector of parameters to be estimated. Let $\boldsymbol{\eta} = [\eta_1 \ \cdots \ \eta_k]^T$ be the vector that groups the logit elements, then the multivariate version of the binomial logit model can be expressed as $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$.

Suppose that after fitting the model to the data, we decide to group some levels of a factor. In the multifactorial situation, the grouping of levels of a factor occurs in several separate clusters, whose number is directly related to the number of levels of other factors of the model. For example, let $s = 3$, A_1, A_2, A_3 be crossed ordered factors and $t_1 = t_2 = t_3 = 3$ its levels. This factor structure contains the tuples $(1,1,1), (1,1,2), \ldots, (3,3,3)$, resulting in $3 \times 3 \times 3 = 27$ tuples.

Let examine the following situation for illustrative purposes: Levels 2 and 3 of A_3 are grouped. In this situation, the new number of levels of A_3 is $t_3^* = 2$ and factor structure is reduced to $3 \times 3 \times 2 = 18$ tuples. For $i = 1, 2, 3$ and $j = 1, 2, 3$, original tuples $(i, j, 2)$ and $(i, j, 3)$ collapse in the new tuples $(i, j, 2^*)$ by adding the corresponding values of the response variables $y_{ij2} + y_{ij3}$ and the totals $n_{ij2} + n_{ij3}$. It is easy to notice that 9 aggregation sets are required, $c_k, k = 1, 2, \ldots, 9$, each one with two elements or levels $c_1 = \{(1,1,2), (1,1,3)\}, \ldots, c_9 = \{(3,3,2), (3,3,3)\}$.

If the proposed model is saturated ($k = m$), i.e. the number of available observations equals the number of model parameters, the $\boldsymbol{X}$ matrix is a square, full rank, and therefore invertible matrix. Moreover, when assuming an unsaturated logit model, generally $k > m$, the design matrix $\boldsymbol{X}$ is no longer square and it has no inverse.

It has been proved by McCullagh & Nelder (1989, p. 119) that

$$V[\widehat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$$

where $\boldsymbol{W} = \text{diag}[n_i p_i (1-p_i)]$. These authors also discuss that the problems of over or under dispersion, deserve detailed study and that they can be solved by simply scaling $V[\widehat{\boldsymbol{\beta}}]$ by a constant, obtained from the deviance or Pearson's statistics and residual degrees of freedom ratio.

Thus, assuming no over or under dispersion (which simply involves the appropriate scaling of the estimated variance-covariance matrix), an immediate consequence of the fact that $\boldsymbol{X}$ has no inverse is that, once parameters have been estimated by iterative reweighted least squares (Searle, Casella & McCulloch 2006, p. 295), $V[\boldsymbol{X}\widehat{\boldsymbol{\beta}}] = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$, do not support further simplification.

Let be $\boldsymbol{\Sigma} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$, with elements $[\sigma_{ij}]$, $i, j = 1, \ldots, k$. In general, though not necessarily, $\sigma_{ij} \neq 0$. Then, due to the central limit theorem (Lehmann 1999, p. 73) and asymptotic properties of maximum likelihood estimators:

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \sim \text{AN}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{\Sigma}_{k\times k}) \tag{5}$$

In (5), "AN" is the abbreviation for "Asymptotically Normal", commonly used in the statistical literature. Moreover, it is necessary the asymptotic distribution of the $\widehat{p}_i$. It is developed in the following theorem:

Theorem 1. *If* $\widehat{\boldsymbol{\eta}} = [logit(\widehat{p}_1)\ logit(\widehat{p}_2)\ \cdots\ logit(\widehat{p}_k)]^T$ *is distributed as in* (5)*, then* $\widehat{\boldsymbol{p}} = [\widehat{p}_1\ \widehat{p}_2\ \cdots\ \widehat{p}_k]^T$*, such that:*

$$\widehat{p}_i = \frac{e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}}}, \quad i=1,\ldots,k$$

is asymptotically distributed as multivariate normal with $E[\widehat{p}_i] = p_i = e^{\boldsymbol{x_i^T\beta}}/(1+e^{\boldsymbol{x_i^T\beta}})$ *and variance covariance matrix* $\boldsymbol{\Psi} = [\psi_{ij}]$ *with elements* $\psi_{ij} = \sigma_{ij}p_i(1-p_i)p_j(1-p_j)$*,* $i,j=1,\ldots,k$*.*

Proof. Let g_i^{-1} for $i=1,\ldots,k$ be real-valued functions defined as

$$g_i^{-1}(\widehat{\eta}_1,\ldots,\widehat{\eta}_i,\ldots,\widehat{\eta}_k) = \frac{e^{\widehat{\eta}_i}}{1+e^{\widehat{\eta}_i}}$$

then,

$$\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_j} = \begin{cases} 0 & ,i\neq j \\ e^{\widehat{\eta}_i}/(1+e^{\widehat{\eta}_i})^2 & ,i=j \end{cases}$$

$$\begin{aligned}\psi_{ij} &= \sum_{s=1}^{k}\sum_{t=1}^{k}\sigma_{st}\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_s}\frac{\partial g_j^{-1}}{\partial\widehat{\eta}_t}\bigg|_{\widehat{\boldsymbol{\eta}}=\boldsymbol{\eta}} = \sum_{s=1}^{k}\sigma_{sj}\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_s}\frac{\partial g_j^{-1}}{\partial\widehat{\eta}_j}\bigg|_{\widehat{\boldsymbol{\eta}}=\boldsymbol{\eta}} \\ &= \sigma_{ij}\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_i}\frac{\partial g_j^{-1}}{\partial\widehat{\eta}_j}\bigg|_{\widehat{\boldsymbol{\eta}}=\boldsymbol{\eta}} = \sigma_{ij}\frac{e^{\eta_i}}{(1+e^{\eta_i})^2}\frac{e^{\eta_j}}{(1+e^{\eta_j})^2} \\ &= \sigma_{ij}p_i(1-p_i)p_j(1-p_j)\end{aligned}$$

Thus, given the existence of the partial derivatives around $\widehat{\boldsymbol{\eta}}$, multivariate version of the delta method (Lehmann 1999, p. 315) ensures that $\widehat{\boldsymbol{p}} = [\widehat{p}_1\ \widehat{p}_2\ \cdots\ \widehat{p}_k]^T$ is asymptotically normal with $\mathrm{E}[\widehat{p}_i] = p_i = e^{\boldsymbol{x_i^T\beta}}/(1+e^{\boldsymbol{x_i^T\beta}})$ and variance covariance matrix $\boldsymbol{\Psi} = [\psi_{ij}]$ with $\psi_{ij} = [\sigma_{ij}p_i(1-p_i)p_j(1-p_j)]$, $i,j=1,\ldots,k$. □

Suppose then that the researcher wants to add r levels $(1<r<t_i)$ of i-th factor A_i and, therefore, $a = t_1\times\cdots\times t_{i-1}\times t_{i+1}\times\cdots\times t_s$ sets are produced, whose elements are each r of the indexes $1,\ldots,k$ without repetition, affected by the aggregation. Let the sets (called "aggregation sets"), be defined by:

$$\begin{aligned} c_\nu = \{\xi_1^i,\xi_2^i,\ldots,\xi_r^i\}, \qquad & \xi_j^i\in\{1,\ldots,k\};\ j=1,\ldots,r; \\ & i=1,\ldots,a;\ \ \nu=\min\{\xi_1^i,\xi_2^i,\ldots,\xi_r^i\}\text{ for each }i; \\ & c_\nu\cap c_{\nu'}=\phi,\forall\,\nu,\nu'\end{aligned}$$

for each of which, in turn is defined:

$$n_\nu^* = \sum_{c_\nu} n_i, \qquad \widehat{p}_\nu^* = \frac{\sum_{c_\nu} n_i\widehat{p}_i}{n_\nu^*} \tag{6}$$

Since $\widehat{p}^*_\nu$ is the weighted sum of asymptotically normal RV's, $\widehat{p}^*_\nu$ is an asymptotically normal RV for all ν and covaries with the other probability estimators. It is easy to verify that $\mathrm{E}[\widehat{p}^*_\nu] = p^*_\nu = (\sum_{c_\nu} n_i p_i)/n^*_\nu$, however, the variance and covariance associated with $\widehat{p}^*_\nu$ are more complex, as is proved in the following theorem:

Theorem 2. *Given* $\widehat{\boldsymbol{p}} = [\widehat{p}_1\ \widehat{p}_2\ \cdots\ \widehat{p}_k]^T$ *distributed as in Theorem 1, if* $\widehat{p}^*_\nu = (\sum_{c_\nu} n_i \widehat{p}_i)/n^*_\nu$ *with* $n^*_\nu = \sum_{c_\nu} n_i$*, then:*

$$V[\widehat{p}^*_\nu] = \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2\psi_{ii} + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \psi_{ij}\right\}$$

$$Cov[\widehat{p}^*_\nu, \widehat{p}_j] = \frac{\sum_{i\in c_\nu} n_i\psi_{ij}}{n^*_\nu},\ \textit{for all } j \notin \bigcup_{i=1}^{a} c_i$$

$$Cov[\widehat{p}^*_\nu, \widehat{p}^*_{\nu'}] = \frac{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \psi_{ij}}{n^*_\nu n^*_{\nu'}}$$

for any two aggregation sets $c_\nu, c_{\nu'}$*.*

Proof.

$$(\widehat{p}^*_\nu)^2 = \left\{\frac{\sum_{c_\nu} n_i\widehat{p}_i}{n^*_\nu}\right\}^2$$

$$= \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2\widehat{p}_i^2 + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \widehat{p}_i\widehat{p}_j\right\}$$

$$(\mathrm{E}[\widehat{p}^*_\nu])^2 = \left\{\frac{\sum_{c_\nu} n_i\mathrm{E}[\widehat{p}_i]}{n^*_\nu}\right\}^2$$

$$= \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2(\mathrm{E}[\widehat{p}_i])^2 + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j]\right\}$$

$$\mathrm{E}[(\widehat{p}^*_\nu)^2] = \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2\mathrm{E}[\widehat{p}_i^2] + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \mathrm{E}[\widehat{p}_i\widehat{p}_j]\right\} \Rightarrow$$

$$\mathrm{V}[\widehat{p}^*_\nu] = \mathrm{E}[(\widehat{p}^*_\nu)^2] - (\mathrm{E}[\widehat{p}^*_\nu])^2$$

$$= \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2(\mathrm{E}[\widehat{p}_i^2] - \mathrm{E}[\widehat{p}_i]^2)\right.$$

$$\left. +2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j(\mathrm{E}[\widehat{p}_i\widehat{p}_j] - \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j])\right\}$$

$$= \frac{1}{(n_\nu^*)^2}\left\{\sum_{c_\nu} n_i^2 \mathrm{V}[\widehat{p}_i] + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\sum_{j\in c_\nu > i} n_i n_j \mathrm{Cov}[\widehat{p}_i,\widehat{p}_j]\right\}$$
$$= \frac{1}{(n_\nu^*)^2}\left\{\sum_{c_\nu} n_i^2 \psi_{ii} + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\sum_{j\in c_\nu > i} n_i n_j \psi_{ij}\right\}$$

Furthermore, for $j \notin c_\nu$:

$$\begin{aligned}\mathrm{Cov}[\widehat{p}_\nu^*,\widehat{p}_j] &= \mathrm{E}[\widehat{p}_\nu^*\widehat{p}_j] - \mathrm{E}[\widehat{p}_\nu^*]\mathrm{E}[\widehat{p}_j] \\ &= \frac{\sum_{c_\nu} n_i \mathrm{E}[\widehat{p}_i\widehat{p}_j]}{n_\nu^*} - \frac{\sum_{c_\nu} n_i \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j]}{n_\nu^*} \\ &= \frac{\sum_{c_\nu} n_i \mathrm{Cov}[\widehat{p}_i,\widehat{p}_j]}{n_\nu^*} = \frac{\sum_{c_\nu} n_i \psi_{ij}}{n_\nu^*}\end{aligned}$$

Finally:

$$\widehat{p}_\nu^*\widehat{p}_{\nu'}^* = \left(\frac{\sum_{c_\nu} n_i\widehat{p}_i}{n_\nu^*}\right)\left(\frac{\sum_{c_{\nu'}} n_i\widehat{p}_i}{n_{\nu'}^*}\right) = \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \widehat{p}_i\widehat{p}_j\right\} \Rightarrow$$

$$\begin{aligned}\mathrm{Cov}[\widehat{p}_\nu^*,\widehat{p}_{\nu'}^*] &= \mathrm{E}[\widehat{p}_\nu^*\widehat{p}_{\nu'}^*] - \mathrm{E}[\widehat{p}_\nu^*]\mathrm{E}[\widehat{p}_{\nu'}^*] \\ &= \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \mathrm{E}[\widehat{p}_i\widehat{p}_j]\right\} \\ &\qquad - \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j]\right\} \\ &= \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \mathrm{Cov}[\widehat{p}_i,\widehat{p}_j]\right\} \frac{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \psi_{ij}}{n_\nu^* n_{\nu'}^*}\end{aligned}$$

□

Note that the cardinality of the index range of the model (originally k) has been reduced given the aggregation levels and is now $k^* = k - a(r-1)$. Each group of r originals RV's, for each of the a different combinations of the levels of the other factors $(A_1,\ldots,A_{i-1},A_{i+1},\ldots,A_k)$, gives way to a single random variable constructed from the sum, renamed in its index at the lower value of the aggregation set that corresponds. So, having both likelihood estimators not affected by aggregation, as those who effectively are, we can settle the new vector:

$$\widehat{\boldsymbol{p}}_{k^*\times 1}^* \sim \mathrm{AN}(\boldsymbol{p}_{k^*\times 1}^*; \boldsymbol{\Psi}_{k^*\times k^*}^*) \tag{7}$$

where:

$$\widehat{\boldsymbol{p}}^{*} = \begin{bmatrix} \widehat{p}_1^{*} \\ \vdots \\ \widehat{p}_k^{*} \end{bmatrix}, \ \boldsymbol{p}^{*} = \begin{bmatrix} p_1^{*} \\ \vdots \\ p_k^{*} \end{bmatrix}, \ \boldsymbol{\Psi}^{*} = \begin{bmatrix} \psi_{11}^{*} & \psi_{12}^{*} & \cdots & \psi_{1k}^{*} \\ \psi_{21}^{*} & \psi_{22}^{*} & \cdots & \psi_{2k}^{*} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{k1}^{*} & \psi_{k2}^{*} & \cdots & \psi_{kk}^{*} \end{bmatrix}$$

except that the range of the index $1, \ldots, k$ in $\boldsymbol{\Psi}^{*}$, although ordered, is not correlated with $\mathbb{N}$, that is, some of their values are no longer present.

Also, $\widehat{p}_i^{*} \equiv \widehat{p}_i$, $p_i^{*} \equiv p_i$, $\psi_{ij}^{*} \equiv \psi_{ij}$ for all $i, j \notin \cup c_\nu$ and $\widehat{p}_i^{*}$, p_i^{*}, ψ_{ij}^{*} are as in the definition and Theorem 2 for the remaining i, j.

Example 1. Let T be a contingency table with two factors A_1 and A_2, the first with 2 levels $(1, 2)$ and the second with three $(1, 2, 3)$. Reindexing the original subscripts properly, we have:

$$1 \equiv (1,1); 2 \equiv (1,2); 3 \equiv (1,3); 4 \equiv (2,1); 5 \equiv (2,2); 6 \equiv (2,3)$$

with the original logit model estimates $\widehat{\boldsymbol{p}} = [\widehat{p}_1 \quad \widehat{p}_2 \quad \widehat{p}_3 \quad \widehat{p}_4 \quad \widehat{p}_5 \quad \widehat{p}_6]^T$.

Now suppose we add levels 2 and 3 of factor A_2. Aggregation sets that arise are $c_2 = \{2, 3\}$ and $c_5 = \{5, 6\}$, and the new model estimates $\widehat{\boldsymbol{p}}^{*} = [\widehat{p}_1^{*} \quad \widehat{p}_2^{*} \quad \widehat{p}_4^{*} \quad \widehat{p}_5^{*}]^T$, where:

$$\begin{aligned} \widehat{p}_1^{*} &= \widehat{p}_1 \\ \widehat{p}_2^{*} &= \frac{n_2 \widehat{p}_2 + n_3 \widehat{p}_3}{(n_2 + n_3)} \\ \widehat{p}_4^{*} &= \widehat{p}_4 \\ \widehat{p}_5^{*} &= \frac{n_5 \widehat{p}_5 + n_6 \widehat{p}_6}{(n_5 + n_6)} \end{aligned}$$

In addition, the variance covariance matrix of $\widehat{\boldsymbol{p}}$ is

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} & \psi_{15} & \psi_{16} \\ \psi_{21} & \psi_{22} & \psi_{23} & \psi_{24} & \psi_{25} & \psi_{26} \\ \psi_{31} & \psi_{32} & \psi_{33} & \psi_{34} & \psi_{35} & \psi_{36} \\ \psi_{41} & \psi_{42} & \psi_{43} & \psi_{44} & \psi_{45} & \psi_{46} \\ \psi_{51} & \psi_{52} & \psi_{53} & \psi_{54} & \psi_{55} & \psi_{56} \\ \psi_{61} & \psi_{62} & \psi_{63} & \psi_{64} & \psi_{65} & \psi_{66} \end{bmatrix}$$

while the variance covariance matrix of $\widehat{\boldsymbol{p}}^{*}$ (symmetric) is

$$\boldsymbol{\Psi}^{*} = \begin{bmatrix} \psi_{11}^{*} & \psi_{12}^{*} & \psi_{14}^{*} & \psi_{15}^{*} \\ \psi_{21}^{*} & \psi_{22}^{*} & \psi_{24}^{*} & \psi_{25}^{*} \\ \psi_{41}^{*} & \psi_{42}^{*} & \psi_{44}^{*} & \psi_{45}^{*} \\ \psi_{51}^{*} & \psi_{52}^{*} & \psi_{54}^{*} & \psi_{55}^{*} \end{bmatrix}$$

where, following the Theorem 2:

$$\begin{aligned}
\psi^*_{11} &= \psi_{11}\\
\psi^*_{12} &= (n_2\psi_{12} + n_3\psi_{13})/(n_2 + n_3)\\
\psi^*_{14} &= \psi_{14}\\
\psi^*_{15} &= (n_5\psi_{15} + n_6\psi_{16})/(n_5 + n_6)\\
\psi^*_{22} &= (n_2^2\psi_{22} + n_3^2\psi_{33} + 2n_2n_3\psi_{23})/(n_2 + n_3)^2\\
\psi^*_{24} &= (n_2\psi_{24} + n_3\psi_{34})/(n_2 + n_3)\\
\psi^*_{25} &= (n_2n_5\psi_{25} + n_2n_6\psi_{26} + n_3n_5\psi_{35} + n_3n_6\psi_{36})/[(n_2 + n_3)(n_5 + n_6)]\\
\psi^*_{44} &= \psi_{44}\\
\psi^*_{45} &= (n_5\psi_{45} + n_6\psi_{46})/(n_5 + n_6)\\
\psi^*_{55} &= (n_5^2\psi_{55} + n_6^2\psi_{66} + 2n_5n_6\psi_{56})/(n_5 + n_6)^2
\end{aligned}$$

Now, returning to the theoretical development, the following theorem shows the required distribution of $\text{logit}(\widehat{p}^*_i), i = 1, \ldots, k$, prior to the estimation of the parameters associated with the factors.

Theorem 3. *If* $\widehat{\boldsymbol{p}}^*$ *is distributed as in* (7)*, then:*

$$\textit{logit}(\widehat{\boldsymbol{p}}^*) = \begin{bmatrix}\textit{logit}(\widehat{p}^*_1) & \cdots & \textit{logit}(\widehat{p}^*_k)\end{bmatrix}^T$$

is asymptotically distributed multivariate normal with $E[\textit{logit}(\widehat{p}^*_i)] = \textit{logit}(p^*_i)$ *and variance covariance matrix* $\boldsymbol{\Sigma}^* = [\sigma^*_{ij}] = [\psi^*_{ij}[p^*_i(1-p^*_i)p^*_j(1-p^*_j)]^{-1}]$.

Proof. Lets $g_i(i = 1, \ldots, k)$, real-valued functions defined as

$$g_i(\widehat{p}^*_1, \ldots, \widehat{p}^*_i, \ldots, \widehat{p}^*_k) = \text{logit}(\widehat{p}^*_i)$$

then,

$$\frac{\partial g_i}{\partial \widehat{p}^*_j} = \begin{cases} 0 & , i \neq j \\ [\widehat{p}^*_i(1-\widehat{p}^*_i)]^{-1}, & i = j \end{cases}$$

and

$$\sigma^*_{ij} = \sum_{s=1}^{k}\sum_{t=1}^{k}\psi^*_{st}\frac{\partial g_i}{\partial \widehat{p}^*_s}\frac{\partial g_j}{\partial \widehat{p}^*_t}\bigg|_{\widehat{\boldsymbol{p}}^* = \boldsymbol{p}^*} = \psi^*_{ij}[p^*_i(1-p^*_i)p^*_j(1-p^*_j)]^{-1}$$

And because in this case also there are the partial derivatives around $\widehat{\boldsymbol{p}}^*$, using again a multivariate version of the delta method, $\textbf{logit}(\widehat{\boldsymbol{p}}^*)$ is asymptotically distributed multivariate normal with $\text{E}[\text{logit}(\widehat{p}^*_i)] = \text{logit}(p^*_i)$ and variance covariance matrix $\boldsymbol{\Sigma}^* = [\sigma^*_{ij}] = [\psi^*_{ij}[p^*_i(1-p^*_i)p^*_j(1-p^*_j)]^{-1}]$. □

Finally, the following theorem shows the distribution of the new parameters $\widehat{\boldsymbol{\beta}}^*$, from the new design matrix $\boldsymbol{X}^*$. Its proof is omitted since it is easily obtained by appealing to the results included in the Appendix.

Theorem 4. *Given the model*

$$\boldsymbol{Y} = logit(\widehat{\boldsymbol{p}}^{*}) = \boldsymbol{X}^{*}\boldsymbol{\beta}^{*} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim AN(\boldsymbol{0}, \boldsymbol{\Sigma}^{*})$$

in which, $\boldsymbol{Y} = logit(\widehat{\boldsymbol{p}}^{*})$ *is a column vector whose elements are* $logit(\widehat{p}_i^{*}), i = 1, \ldots, k$ *and* $\boldsymbol{\Sigma}^{*}$ *is the variance covariance matrix, both constant and known, calculated according to the Theorem 3. Let* $\boldsymbol{X}^{*}$ *be the new design matrix*[2]*, using the reference parameterization, proposed after the process of aggregation of factor levels, constrained to include the same factors that included the original design matrix* $\boldsymbol{X}$*. And let* $\boldsymbol{\beta}^{*}$ *be the new vector of parameters to be estimated by maximum likelihood (*$\widehat{\boldsymbol{\beta}}^{*}$*). Then:*

- $\widehat{\boldsymbol{\beta}}^{*} = [(\boldsymbol{X}^{*})^{T}\boldsymbol{X}^{*}]^{-1}(\boldsymbol{X}^{*})^{T}\boldsymbol{Y}$.
- $V[\widehat{\boldsymbol{\beta}}^{*}] = [(\boldsymbol{X}^{*})^{T}\boldsymbol{X}^{*}]^{-1}(\boldsymbol{X}^{*})^{T}\boldsymbol{\Sigma}^{*}\boldsymbol{X}^{*}[(\boldsymbol{X}^{*})^{T}\boldsymbol{X}^{*}]^{-1}$.
- $\widehat{\boldsymbol{\beta}}^{*}$ *is distributed asymptotically normal.*

A modification in this asymptotical distribution has been induced for the original and transformed RV's, by applying aggregation some factor levels and some required sets of aggregation (c_ν). Thus, the suggested procedure is as follows:

1. Fit a logit model by preserving the calculation of the vector of estimates of p_i and the variance covariance matrix $\boldsymbol{\Sigma}$ estimated for $\widehat{\boldsymbol{\beta}}$.
2. Define the required aggregation sets, in order to calculate point estimates for the p_ν^{*} as in the Theorem 2 and the variance covariance matrix $\boldsymbol{\Psi}^{*}$ of $\widehat{\boldsymbol{p}}^{*}$, like in (7).
3. Compute $\text{logit}(\widehat{p}_i^{*})$ for the resulting range of values i and its variance covariance matrix $\boldsymbol{\Sigma}^{*}$, following Theorem 3.
4. Build the new design matrix $\boldsymbol{X}^{*}_{k^{*}\times m^{*}}$ according to the new desired parameters vector $\boldsymbol{\beta}^{*}_{m^{*}\times 1}$.
5. In general, setting a generalized least squares regression (Christensen 2002, pp. 33, 86) to estimate $\widehat{\boldsymbol{\beta}}^{*}_{m^{*}\times 1}$ with a new model formulated as follows:

   $$\boldsymbol{Y} = \text{logit}(\widehat{\boldsymbol{p}}^{*}) = \boldsymbol{X}^{*}\boldsymbol{\beta}^{*} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim \text{AN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{*})$$

   in which the matrix $\boldsymbol{\Sigma}^{*}$ is the result of step 3. However, using the reference parameterization, the computation of both, the vector of parameters to be estimated and the variance covariance matrix, is greatly simplified by using the Theorem 4.

Finally, it is clear that the deductions have been made for the aggregate of r levels of a single factor. However, this approach does not diminish generality. If the researcher wants to group two or more factors, we just weed to iteratively apply the suggested procedure, one factor at a time. In other words, we simply applies the suggested procedure, repeatedly.

[2]Note that $\boldsymbol{X}^{*}$ has a smaller number of columns than $\boldsymbol{X}$ (hence $\boldsymbol{\beta}^{*}$ has fewer elements $\boldsymbol{\beta}$), because it models a smaller number of junctions in the levels of the factors.

4. Illustration of the Suggested Procedure

Table 1 presents a situation where the interest lies in studying the relationship between a response variable Y and two explanatory factors A_1 and A_2 with 2 and 3 levels, respectively. The observed frequencies or number of successes for each levels combination are shown in Table 1 above.

TABLE 1: Example $Y(0,1)$ vs. $A_1(1,2)$, $A_2(1,2,3)$.

i	A_1	A_2	No. of successes	Total
1	1	1	53	133
2	1	2	11	133
3	1	3	127	133
4	2	1	165	533
5	2	2	41	533
6	2	3	476	533
Total			873	1998

In this particular case, the proposed logit model omits interactions between the factors, and therefore is not saturated. Using the first level of each factor as a reference, the equations are as follows:

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \text{logit}(p_3) \\ \text{logit}(p_4) \\ \text{logit}(p_5) \\ \text{logit}(p_6) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \tag{8}$$

In (8), β_1 represents the intercept effect, β_2 represents the level 2 of A_1 effect, β_3 represents the level 2 of A_2 effect and β_4 the level 3 of A_2 effect. The levels 1 of A_1 and 1 of A_2 are not explicitly represented in the model (are the references levels).

Preliminary tests and goodness of fit for this model are shown in Table 2. The model fits the data appropriately, as it is deduced from the deviance and Pearson's statistics. Also, according to the Pearson's statistic, the overdispersion is negligible.

Table 3 contains the parameter estimates with their corresponding standard tests for $\text{H}_0 : \beta_i = 0, i = 1,2,3$ and 95% confidence intervals (CI) for β_i. The predicted probabilities and their CI's are also shown in Table 4.

Now suppose that we cluster levels 2 and 3 of the factor A_2 in Table 1, producing the Table 5 with aggregate data, and postulate the usual procedure: a new logit model.

Table 2: Original model. Preliminary tests and goodness of fit.

Test	Value
Residual deviance:	2.5065
Residual degrees of freedom (DF):	2
Deviance χ^2/DF:	0.2856
Deviance test:	No reject
Pearson's statistic:	2.3104
Pearson's χ^2/DF:	0.315
Pearson's test:	No reject
Deviance/DF:	1.2533
Pearson's/DF:	1.1552

Table 3: Original model. $\widehat{\beta}_i$ and normal tests ($H_0 : \beta_i = 0$).

	Estimation of β_i					95% CI	
i	$\widehat{\beta}_i$	SE	z-Value	p ($>\|z\|$)	Conclusion	Ll	Ul
1	−0.39857	0.14688	−2.71369	0.00666	Reject	−0.68644	−0.11070
2	−0.40725	0.15536	−2.62130	0.00876	Reject	−0.71176	−0.10275
3	−1.75603	0.16676	−10.53046	0.00000	Reject	−2.08286	−1.42919
4	2.99348	0.15665	19.10956	0.00000	Reject	2.68646	3.30051

SE: Standard Error. Ll: Lower limit. Ul: Upper limit.

Table 4: Original model. Predicted probabilities and 95% CI's.

i	$\widehat{p_i}$	Ll	Ul	i	$\widehat{p_i}$	Ll	Ul
1	0.4017	0.3325	0.4708	4	0.3088	0.2713	0.3463
2	0.1039	0.0702	0.1376	5	0.0716	0.0521	0.0912
3	0.9305	0.9068	0.9542	6	0.8991	0.8752	0.9231

Table 5: Example $Y(0,1)$ vs. $A_1(1,2)$, $A_2(1,2^*)$.

i	A_1	A_2	No. of success (y)	Total (t)	y/t
1	1	1	53	133	0.3984
2	1	2^*	138	266	0.5188
3	2	1	165	533	0.3096
4	2	2^*	517	1066	0.4850
Total			873	1998	

The new unsaturated model (ignoring interactions), using the reference parameterization (with level 1 of both factors by reference), unfolds as follows:

$$\begin{bmatrix} \text{logit}(p_1^*) \\ \text{logit}(p_2^*) \\ \text{logit}(p_3^*) \\ \text{logit}(p_4^*) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{bmatrix} \tag{9}$$

Now in (9), β_1^* represents the intercept effect, β_2^* level 2 of A_1 effect and β_3^* the level 2^* of A_2 effect. Some measures of goodness of fit for this model are reproduced in Table 6.

The model fits the data with negligible overdispersion. Re-adjusting a logit model over the resulting contingency table, the new estimates are shown in Table

7. Without regard the parameters significance, the probabilities predicted by the model are reproduced in Table 8.

TABLE 6: Usual procedure. Testing goodness of fit of the aggregate data model.

Test	Value
Residual deviance:	1.0986
Residual degrees of freedom (DF):	1
Deviance χ^2/DF:	0.2946
Deviance test:	No reject
Pearson's statistic:	1.1051
Pearson's χ^2/DF:	0.2932
Pearson's test:	No reject
Deviance/DF:	1.0986
Pearson's/DF:	1.1051

TABLE 7: Usual procedure. $\widehat{\beta}_i^*$ and normal tests ($H_0 : \beta_i^* = 0$).

	β_i^* estimation					95% CI	
i	$\widehat{\beta}_i$	SE	z-Value	p ($>|z|$)	Conclusion	Ll	Ul
1	−0.5485	0.1220	−4.4965	0.0000	Reject	−0.7876	−0.3094
2	−0.2162	0.1137	−1.9014	0.0573	No reject	−0.4390	0.0067
3	0.6885	0.0992	6.9401	0.0000	Reject	0.4941	0.8830

TABLE 8: Usual procedure. Predicted probabilities and 95% CI without regard to model parameters statistical significance.

i	$\widehat{p}_i^*$	Ll	Ul
1	0.3662	0.3107	0.4217
2	0.5349	0.4831	0.5868
3	0.3176	0.2811	0.3542
4	0.4810	0.4519	0.5100

The new parameter vector $\boldsymbol{\beta^*}$ is estimated differently in both models (original and aggregated data). With $\alpha = 0.05$, Table 7 suggests the absence of sufficient evidence to reject the null hypothesis about β_2^*. This finding has important implications for the analysis: Since it is not possible to conclude that β_2^* is significantly different that 0, the predicted probabilities in Table 8, in strict statistical sense, should not be considered valid. Statistical valid predictions are as follows:

$$\widehat{p}_1^* = \frac{e^{\widehat{\beta}_1^*}}{1 + e^{\widehat{\beta}_1^*}} = 0.3662$$

$$\widehat{p}_2^* = \frac{e^{\widehat{\beta}_1^* + \widehat{\beta}_3^*}}{1 + e^{\widehat{\beta}_1^* + \widehat{\beta}_3^*}} = 0.5349$$

$$\widehat{p}_3^* = \frac{e^{\widehat{\beta}_1^* + \widehat{\beta}_2^*}}{1 + e^{\widehat{\beta}_1^* + \widehat{\beta}_2^*}} = \frac{e^{\widehat{\beta}_1^* + 0}}{1 + e^{\widehat{\beta}_1^* + 0}} = 0.3662 \tag{10}$$

$$\widehat{p}_4^* = \frac{e^{\widehat{\beta}_1^*+\widehat{\beta}_2^*+\widehat{\beta}_3^*}}{1+e^{\widehat{\beta}_1^*+\widehat{\beta}_2^*+\widehat{\beta}_3^*}} = \frac{e^{\widehat{\beta}_1^*+0+\widehat{\beta}_3^*}}{1+e^{\widehat{\beta}_1^*+0+\widehat{\beta}_3^*}} = 0.5349 \tag{11}$$

Finally, Table 9 contains the estimates and the 95% CIs for the logit model postulated in (9), obtained by the procedure suggested in this paper. Note that the point estimates of the standard procedure and the suggested procedure are slightly but significantly different. Using the suggested procedure, the Pearson's goodness of fit of the model produces a χ^2 of 0.0104 that leaves a probability of 0.9188 at right. Then, the model analyzed by the suggested procedure properly fits the data; in fact it fits in a better way than with the usual procedure, which produces a Pearson's statistic 1.1051 that leaves a probability of 0.2932 at right (see Table 6).

Table 9: Suggested procedure. $\widehat{\beta}_i^*$ and normal tests ($H_0 : \beta_i^* = 0$).

	β_i^* estimation					95% CIs	
i	$\widehat{\beta}_i$	SE	z-Value	p ($>\|z\|$)	Conclusion	Ll	Ul
1	−0.4685	0.1257	−3.7280	0.0002	Reject	−0.7149	−0.2222
2	−0.2673	0.1019	−2.6236	0.0087	Reject	−0.4670	−0.0676
3	0.6074	0.0931	6.5234	0.0000	Reject	0.4249	0.7899

Table 9 shows that the estimated standard errors for the parameters β_2^* and β_3^*, using the suggested procedure are lower than those found by conventional procedure (Table 7).

Table 10 presents the predicted probabilities, now fitting the data according to the procedure suggested in this paper. Note that the predicted probabilities are considerably closer to those expected for the new data set (see the column and y/t in the Table 5), than those predicted with the usual procedure.

Table 10: Suggested procedure. Predicted probabilities and 95% CIs.

i	$\widehat{p}_i^*$	Ll	Ul
1	0.4017	0.3325	0.4708
2	0.5172	0.4927	0.5417
3	0.3088	0.2713	0.3463
4	0.4854	0.4696	0.5012

Also, note in Table 9 that the conclusion about the significance of β_2^* is no lunger the same. The standard procedure statistically valid estimates (10) and (11) and look considerably different from using the suggested procedure. The predictions on Table 10 are statistically valid, approaching in a better way than would be expected from the available data.

Finally, Figure 1 presents the Pearson's standardized residuals, calculated using both methods. Clearly, the estimates produced by the suggested procedure are much closer to the expected value than those produced by the conventional procedure.

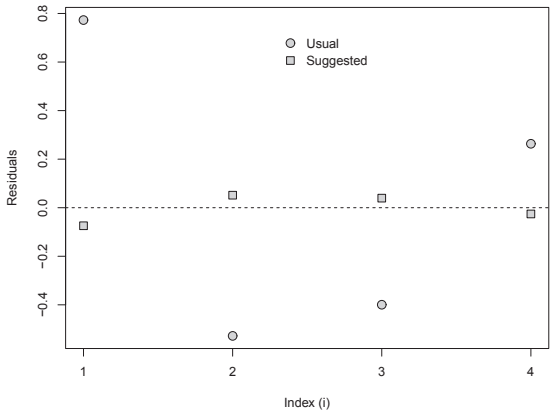

FIGURE 1: Pearson's standardized residuals calculated by both procedures.

5. Comparison of Alternative Procedures Through Simulation

In a situation with a binary response Y and two explanatory factors A_1, A_2, the first with 2 levels and the second with 3 leves, we propose a simulation in order to study the effect of the aggregation of levels 2 and 3 for the factor A_2, using pseudo-random generation of a large number of contingency tables of the type shown in Table 11.

TABLE 11: Original arrangement for simulation Y vs. $A_1(1,2)$, $A_2(1,2,3)$.

i	A_1	A_2	No. of successes (y_i)	Total (n_i)
1	1	1	y_1	n_1
2	1	2	y_2	n_2
3	1	3	y_3	n_3
4	2	1	y_4	n_4
5	2	2	y_5	n_5
6	2	3	y_6	n_6
Total			$y.$	$n.$

An unsaturated logit model is fitted to of the generated tables signoring the effect of interactions, using the first level of each factor as a reference, by

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \text{logit}(p_3) \\ \text{logit}(p_4) \\ \text{logit}(p_5) \\ \text{logit}(p_6) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \tag{12}$$

In (12), p_i represents the probability of success of the i-th combination of levels of the two explanatory factors identified in Table 11 ($i = 1, \ldots, 6$), while β_j are

the parameters to be fitted ($j = 1, \ldots, 4$). Specifically, β_1 represents the effect of the intercept, β_2 is the effect of level 2 of factor A_1, β_3 is the effect of level 2 of the factor A_2 and β_4 is the effect of the level 3 of the factor A_2.

The Table 12 is formed by grouping the last two levels of the second factor in the Table 11.

TABLE 12: Aggregated data for simulation Y vs. $A_1(1,2)$, $A_2(1,2^*)$.

i	A_1	A_2	No. of successes (y_i)	Total (n_i)
1	1	1	y_1	n_1
2	1	2^*	$y_2 + y_3$	$n_2 + n_3$
3	2	1	y_4	n_4
4	2	2^*	$y_5 + y_6$	$n_5 + n_6$
Total			$y_\cdot$	$n_\cdot$

Following the usual procedure, we set a new unsaturated logit model for the Table 12, than also ignores the effect of interactions and uses the first level of each factor as a reference:

$$\begin{bmatrix} \text{logit}(p_1^*) \\ \text{logit}(p_2^*) \\ \text{logit}(p_3^*) \\ \text{logit}(p_4^*) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{bmatrix} \tag{13}$$

Now, in (13) p_i^* represents the probability of success of the i-th combination of levels of the two explanatory factors identified in the Table 12 ($i = 1, \ldots, 4$), with β_1^* representing the effect of the intercept, β_2^* the effect of level 2 of the factor A_1 and β_3^* the effect of level 2^* of the factor A_2.

Lastly, following the suggested procedure, from the original model results (12) we estimate the parameters of the new model with aggregate levels of the factor. The comparison between the two procedures (usual and suggested) is then used to analyze the resulting performance statistics in each case.

5.1. Design of the Experiment of Simulation

That total in the experiment set $n_\cdot = 2000$ is distributed in two randomized numbers to each level of A_1 and, within these, in three randomized numbers for each level of A_2. In this particular study, it is not of interest to compare the effect of both procedures on the levels of factor A_1, or on the first level of factor A_2, p_1 y p_4. Then, independent pseudo-randomly uniform $(0, 1)$ samples are generated. Using the generated values of n_1, n_4 (selected randomly from $n_\cdot$) and p_1 and p_4, the samples $Y_1 \sim \text{Bin}(n_1, p_1)$ and $Y_4 \sim \text{Bin}(n_4, p_4)$ are generated.

For the factor levels being compared in A_2, the samples $Y_2 \sim \text{Bin}(n_2, p_2)$, $Y_3 \sim \text{Bin}(n_3, p_3)$, $Y_5 \sim \text{Bin}(n_5, p_5)$ and $Y_6 \sim \text{Bin}(n_6, p_6)$, are generated n_j randomized as before and sequentially using combinations of $\Delta_p = |p_2 - p_3| = |p_5 - p_6| = 0.0, 0.2, 0.4, 0.6, 0.8$. Such combinations are obtained by maintaining

the values $p_2 = p_5 = 0.1$ as constant and varying by the values of $p_3 = p_6 = 0.1, 0.3, 0.5, 0.7, 0.9$.

For each combinations of Δ_p to experiment several, contingency tables are produced, regrading to by the binomials generated, which are independent within each table and between tables. We only incorporate samples that meet the following conditions:

1. Lead to acceptance of the original logit model, as assessed by the Pearson's goodness of fit.

2. Lead to an original logit model the does not present important problems on subdispertion. That is, that produces a statistical ratio of the Pearson's and residual degrees of freedom in the range $(0.75; 1.25)$.

3. Lead to acceptance of the logit model with levels 2 and 3 of the factor added A_2, also following the Pearson's test of goodness of fit.

4. Lead to a logit model with aggregate levels of the factor, which does not have important problems of subdispertion. This in order to produce a statistical ratio of the Pearson's and residual degrees of freedom in the range $(0.75; 1.25)$.

Finally, there are 10,000 valid samples, 2,000 for each combination of Δ_p, and significance level is set up with for testing $\alpha = 0.05$. The performance measures considere were:

a) Firstly, we examine descriptive statistics of the differences the Pearson's χ^2 goodness of fit test, obtained using standard procedures and suggested (in that order).

b) We compare the absolute differences in point estimates of β_1^*, β_2^* y β_3^*, obtained by the standard and suggested procedures, regardless to their statistical significance.

c) Compare the differences in the lengths of the calculated CIs using the usual and suggested procedure. It uses the average ratio between the lengths of the first and the second (in that order). These ratios are calculated for the CIs accompanying the parameter estimates β_1^*, β_2^* and β_3^*.

d) We study the absolute frequency of occurrence of the change in the conclusion of the analysis of variance (acceptance to rejection, or vice versa) for testing hypotheses about the parameters $H_0 : \beta_1^* = 0$, $H_0 : \beta_2^* = 0$ y $H_0 : \beta_3^* = 0$, when they are contrasted by the usual way, and when they are contrasted by the suggested procedure.

e) Finally, for each sample Pearson's standardized residuals produced by both methods were calculated. Also, analysis of each value of Δ_p, we construct boxplots their corresponding.

5.2. Results of the Simulation Experiment

a) Firstly, Table 13 shows means and standard deviations (SD) of the simple differences between the probabilities that leaves to the right Pearson's χ^2 test, in the examination of the goodness of fit of the model, obtained by the usual and suggested procedure.

Table 13: Mean and standard deviations of the differences to the Pearson's χ^2 probabilities (Usual - Suggested).

Δ_p	Mean	SE
0.0	−0.0002	0.0006
0.2	−0.0104	0.0214
0.4	−0.0219	0.0615
0.6	−0.0314	0.1148
0.8	−0.1173	0.1884

Since the average values in Table 13 are all negative, it is clear that the suggested procedure fits the data consistently better than the usual, with the increase in the differences Δ_p.

As evidence of goodness of fit of the model, the Pearson's statistic is particularly suitable in this case, since it is based on the accumulation of the standardized residuals. Although the variability is high, Table 13 that steadily as there are greater differences between the probabilities of the variables involved in the aggregation, the probability to the right of Pearson's χ^2 goodness of fit test increases in the suggested procedure compared with the usual.

In practice this means that, on average, the estimated parameters using the suggested procedure are closer to the expected for a given dataset in comparison to the estimates produced by the usual procedure. It also means that the model fitted using the suggested procedure is less likely to be rejected than the other model.

b) Without considering the significance of the estimated parameters, the Table 14 contains the ranges obtained by both methods (standard and suggested) for each estimate. It can be seen that these ranges are very similar in general and, as should be verified, the same when $\Delta_p = 0$, and slightly more dissimilar as Δ_p increase.

Table 15 contains the averages and standard deviations of the absolute differences between the parameters. As seen there, both the average and the standard deviation of the differences between the parameters estimated by the usual procedure (u) and suggested (s), $|\beta_i^*(\mathrm{u}) - \beta_i^*(\mathrm{s})|, i = 1, 2, 3$ behave similarly. This is, they grow as the probabilities of the variables involved in the aggregation are more dissimilar.

Nevertheless, given the ranges shown in Table 14, these differences do not seem important on average. The conclusion here is that both procedures

(usual and suggested) essentially estimate the same values of model parameters, in most situations.

TABLE 14: Ranges of $\widehat{\beta}_i^*$ according to the usual and suggested procedures.

	$\widehat{\beta}_1^*$				$\widehat{\beta}_2^*$				$\widehat{\beta}_3^*$			
	Usual		Suggested		Usual		Suggested		Usual		Suggested	
Δ_p	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
0.0	-2.43	2.51	-2.43	2.51	-1.31	1.20	-1.31	1.20	-4.69	0.43	-4.69	0.43
0.2	-2.57	2.50	-2.58	2.50	-0.70	0.92	-0.71	0.92	-4.04	1.22	-4.04	1.22
0.4	-2.38	2.65	-2.38	2.65	-0.53	0.54	-0.54	0.55	-3.51	1.63	-3.50	1.63
0.6	-2.51	2.70	-2.51	2.69	-0.57	0.45	-0.66	0.50	-3.13	2.09	-3.13	2.09
0.8	-2.44	2.52	-2.47	2.59	-0.38	0.32	-0.50	0.39	-2.54	2.52	-2.54	2.45

TABLE 15: Average of absolute differences and deviations between the parameters estimated by both methods.

	$\lvert\widehat{\beta}_1^*(u) - \widehat{\beta}_1^*(s)\rvert$		$\lvert\widehat{\beta}_2^*(u) - \widehat{\beta}_2^*(s)\rvert$		$\lvert\widehat{\beta}_3^*(u) - \widehat{\beta}_3^*(s)\rvert$	
Δ_p	Mean	SD	Mean	SD	Mean	SD
0.0	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.002	0.002	0.001	0.001	0.002	0.003
0.4	0.004	0.005	0.004	0.004	0.006	0.007
0.6	0.008	0.010	0.009	0.008	0.011	0.013
0.8	0.012	0.018	0.014	0.014	0.016	0.022

c) Regarding to the lengths of the CIs for each estimator, Table 16 presents the results of the average rates and standard deviations obtained. In general terms, the CI length for the intercept effect shows no appreciable variations in both procedures. However, for the other parameters, the higher Δ_p is the higher the average ratio of the CIs lengths estimated by both methods. Then, it consistently appears that the confidence intervals related to the suggested procedure are narrower and therefore preferable than those estimated by the usual procedure.

TABLE 16: Averages of the ratio between the lengths of confidence intervals (LCI) obtained by the usual method (u) and the suggested method (s).

	β_1^*: LCI(u)/LCI(s)		β_2^*: LCI(u)/LCI(s)		β_3^*: LCI(u)/LCI(s)	
Δ_p	Mean	SD	Mean	SD	Mean	SD
0.0	1.00	0.00	1.00	0.00	1.00	0.00
0.2	1.00	0.00	1.01	0.00	1.01	0.00
0.4	1.01	0.01	1.03	0.01	1.03	0.01
0.6	1.01	0.02	1.06	0.01	1.05	0.02
0.8	1.00	0.04	1.09	0.03	1.06	0.05

Another aspect to note is that while in average terms the conclusion is clear, the differences for the unsaturated case are not as significant as they were in the saturated case developed by Ponsot et al. (2009). The introduction of the covariance and the fact that it examines a larger number of factors have somewhat dampened these differences.

d) Table 17 shows the absolute frequencies of occurrence of the change in the conclusions on the significance of model parameters (H_0: $\beta_i^* = 0$ for $i = 1, 2, 3$), when they are examined with the usual procedure and when they are examined with the suggested procedure.

β_1^* changes occur in similar frequency and any direction. This indicates that is not possible to suggest preferences between the two procedures for intercept estimation. On the other hand, for the remaining two parameters, the conclusion about the statistical significance of not rejecting the null hypothesis and its rejection, greatly promotes the suggested procedure. Improvements in the results on β_2^* are remarkable. There was no change from rejection to acceptance of the null hypothesis, however, there were considerable changes to the contrary, i.e., acceptance to rejection of this hypothesis. The suggested procedure allows us to reject the null hypothesis of model parameters, in a higher proportion of cases, generally increasing with Δ_p.

TABLE 17: Change of the conclusions for H_0: $\beta_i^* = 0$ from the suggested procedure, compared to usual.

Δ_p	Rejection to acceptance	without changes	Acceptance to rejection
	For H_0: $\beta_1^* = 0$		
0.0	0	2000	0
0.2	1	1998	1
0.4	1	1997	2
0.6	3	1989	8
0.8	9	1983	8
	For H_0: $\beta_2^* = 0$		
0.0	0	2000	0
0.2	0	1982	18
0.4	0	1946	54
0.6	0	1951	49
0.8	0	1901	99
	For H_0: $\beta_3^* = 0$		
0.0	0	2000	0
0.2	0	2000	0
0.4	1	1992	7
0.6	1	1990	9
0.8	2	1973	25

e) Finally, the Figures from 2 to 6 contain Pearson's standardized residuals boxplots, grouped according to the procedure that gave rise to (usual and suggested), for each $y_i, i = 1, \ldots, 4$. Observe that for Δ_p from 0.0 to 0.4, boxplots not vary appreciably, indicating that the residuals produced by both procedures are very similar. However, when Δ_p is greater, although the variability of residuals becomes less stable, their averages are closer to 0 in those settings using the suggested procedure. This confirms that in trend terms, the suggested procedure produces better fits than the usual procedure[3].

[3]All programs, both for example, as for the simulation, were made with R statistical system (R Development Core Team 2007).

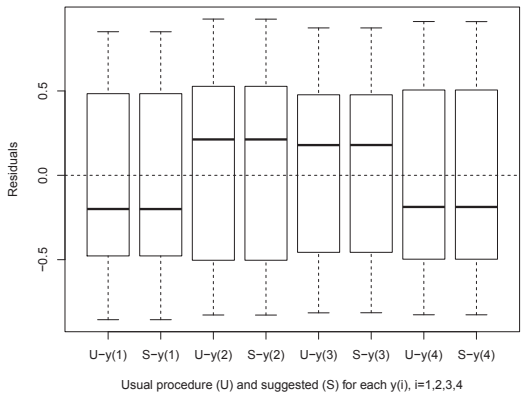

FIGURE 2: Pearson's standardized residuals boxplots for$\Delta_p = 0.0$.

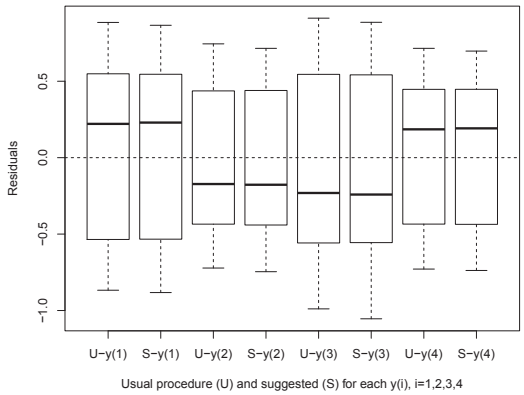

FIGURE 3: Pearson's standardized residuals boxplots for $\Delta_p = 0.2$.

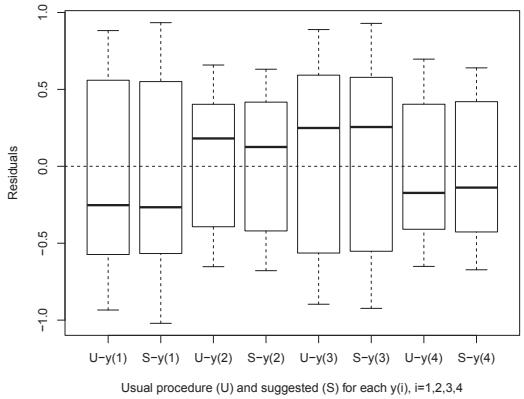

FIGURE 4: Pearson's standardized residuals boxplots for $\Delta_p = 0.4$.

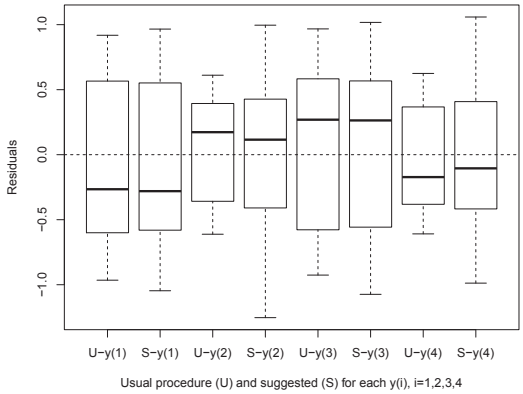

FIGURE 5: Pearson's standardized residuals boxplots for $\Delta_p = 0.6$.

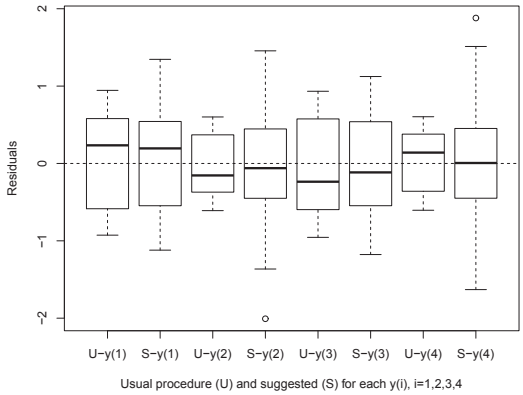

FIGURE 6: Pearson's standardized residuals boxplots for $\Delta_p = 0.8$.

6. Conclusions

This paper addresses and resolves a problem rarely studied, which arises from the practical application of the binomial logit model. We discuss the situation in which, once fitted a logit model to the data in a contingency table, a factor from any of the participants is selected and some levels are added as a new level, to reiterate a logit setting.

In general, there is a problem in the logit model fit with aggregate levels of the factor, particularly when the probabilities of success of RV's involved in aggregation are far from each other. Consequently, this paper suggests a procedure that operates in a broader context, i.e., under the binomial unsaturated multifactorial logit model, and with arguments of asymptotic nature, taking advantage of the reduction in variance when postulates proper distributional model instead of the binomial model, significantly improves the estimates, while lowering the standard error.

As the difference in the probabilities of success accentuates, it becomes better supported by the suggested procedure, instead of the usual. The model fitted by the suggested procedure, also produces closer to zero residuals and less chance of rejection in the goodness of fit test.

In summary, it is proposed to the researcher logit model user, an alternative procedure that can provide theoretical correctness, greater accuracy and less computational effort in the state of aggregation levels of a factor, especially when they involve sample proportions which are markedly dissimilar.

Acknowledgements

The authors thank the Council of Scientific, Humanistic, Technology and the Arts (CDCHTA) of the Los Andes University, the financial support to carry out this work, registered with the code E-303-09-09-ED.

[Recibido: junio de 2011 — Aceptado: febrero de 2012]

References

Christensen, R. (2002), *Plain Answers to Complex Questions. The Theory of Linear Models*, 3 edn, Springer-Verlag, Nueva York, Estados Unidos.

Graybill, F. (1969), *Introduction to Matrices with Applications in Statistics*, 1 edn, Wadsworth Publishing, California, Estados Unidos.

Hilbe, J. M. (2009), *Logistic Regression Models*, 1 edn, Chapman & Hall, Florida, Estados Unidos.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied Logistic Regression*, 2 edn, John Wiley & Sons, Nueva York, Estados Unidos.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, 1 edn, Springer-Verlag, Nueva York, Estados Unidos.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, 2 edn, Chapman & Hall, London, United Kingdom.

Menard, S. (2010), *Logistic Regression: From Introductory to Advanced Concepts and Applications*, 1 edn, SAGE Publications, Inc., California, Estados Unidos.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society. Serie A* (135), 370–384.

Ponsot, E. (2011), Estudio de la Agrupación de Niveles en el Modelo Logit, Unpublisehd PhD Thesis, Instituto de Estadística Aplicada y Computación, Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes, Mérida, Venezuela.

Ponsot, E., Sinha, S. & Goitía, A. (2009), 'Sobre la agrupación de niveles del factor explicativo en el modelo logit binario', *Revista Colombiana de Estadística* **32**(2), 157–187.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Rodríguez, G. (2008), 'Lectures notes about generalized linear models'. *Http://data.princeton.edu/wws509/notes

SAS Institute Inc. (2004), *SAS/STAT(R) 9.1 User's Guide*, SAS Institute Inc., Carolina del Norte, Estados Unidos.

Searle, S., Casella, G. & McCulloch, C. (2006), *Variance Components*, 1 edn, John Wiley and Sons, Inc., Nueva Jersey, Estados Unidos.

Appendix. Study of the Design Matrix $\boldsymbol{X}$ for Saturated and Unsaturated Models

Theorem 5. *Using the reference parameterization, the design matrix of the saturated logit model is invertible.*

Proof. We prove the invertibility of the matrices of design, both in the univariate situation, as in the multifactorial situation, then:

1. Let the saturated univariate logit model design matrix be:

$$\boldsymbol{X}_{k\times k} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

   The calculation of the determinant of $\boldsymbol{X}$ by cofactors decomposition (X_{ij}) is $|\boldsymbol{X}| = (-1)^{k+1}|I| = (-1)^{k+1} \neq 0$, pivoting the last row of the matrix, since the only one nonzero element is x_{k1}. Then, since $|\boldsymbol{X}| \neq 0$, $\boldsymbol{X}^{-1}$ exists.

2. Suppose $A_1, A_2, \ldots, A_s$ factors, each one $t_1, t_2, \ldots, t_s$ levels, respectively. Using the reference parameterization, the followings are postulated:

   a) 1 parameter for the effect of the intercept.

   b) $t_1 - 1$ parameters for the main effects of A_1 factor levels, except the reference; $t_2 - 1$ parameters for the main effects of A_2 factor levels, except the reference; and so on until $t_s - 1$ parameters for the main effects of A_s factor levels, except the reference; in total, $\sum_{i=1}^{s}(t_i - 1)$ parameters for the main effects.

c) $(t_1-1)(t_2-1)$ parameters for the double interaction effects between levels of the factors A_1 and A_2; $(t_1-1)(t_3-1)$ parameters for the double interaction effects between levels of the factors A_1 and A_3; so on until $(t_{s-1}-1)(t_s-1)$ parameters for the double interaction effects between levels of the factors A_{s-1} and A_s; in total$\sum_{i=1}^{s-1}\sum_{j=i+1}^{s}(t_i-1)(t_j-1)$.

d) $(t_1-1)(t_2-1)(t_3-1)$ parameters for the triple effects of interaction between levels of the factors A_1, A_2 and A_3; $(t_1-1)(t_2-1)(t_4-1)$ parameters for the triple effects of interaction between levels of the factors A_1, A_2 and A_4; so on until $(t_{s-2}-1)(t_{s-1}-1)(t_s-1)$ parameters for the triple effects of interaction between levels of the factors A_{s-2}, A_{s-1} and A_s; in total $\sum_{i=1}^{s-2}\sum_{j=i+1}^{s-1}\sum_{k=j+1}^{s}(t_i-1)(t_j-1)(t_k-1)$.

In general, for order a interactions $(1 \leq a \leq s)$ the followings parameters are postulated

$$\sum_{i_1=1}^{s-a+1}\sum_{i_2=i_1+1}^{s-a+2}\cdots\sum_{i_a=i_{a-1}+1}^{s}\prod_{j=1}^{a}(t_{i_j}-1)$$

As the model is saturated, the k total number of postulated parameters equals the number of observations in the contingency table ($k = t_1 \times t_2 \times \cdots \times t_s$). Now, including the interaction of order a $(1 \leq a \leq s)$ in its $i_1, i_2, \ldots, i_a$ levels, requires a row of $\boldsymbol{X}$ like:

$$\begin{bmatrix} 1 & x_1 \cdots x_s & x_{12} \cdots x_{(s-1)s} & \cdots & x_{i_1 i_2 \cdots i_a} & 0 \cdots 0 \end{bmatrix}$$

where

$$x_i = \begin{cases} 1, & i \in \{i_1, i_2, \ldots, i_a\} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1, & i \text{ y } j \in \{i_1, i_2, \ldots, i_a\} \\ 0, & \text{otherwise} \end{cases}$$

and so on until $x_{i_1 i_2 \cdots i_a} = 1$. In other words, the equation that introduces a new parameter representing the interaction of any kind involves only the parameter representing this interaction, and those representing the lower-order interactions contained in it.

Appropriately arranging the rows of $\boldsymbol{X}$ thus constructed, it is easy to verify that a triangular matrix is formed, whose diagonal consists of ones only. Then, using Theorems 1.5.3 and 8.6.5 of Graybill (1969, pp. 8, 191), $|\boldsymbol{X}| \equiv \pm 1 \neq 0$, and therefore $\boldsymbol{X}^{-1}$ exist.

□

Corollary 1. *With reference to parameterization, the design matrix $\boldsymbol{X}$ of the logit model is such that there is $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.*

Proof. Clearly in the saturated model situation, as there is $\boldsymbol{X}^{-1}$, $(\boldsymbol{X}^T)^{-1}$ exist and then $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ also exist.

In the unsaturated model situation, the design matrix $\boldsymbol{X}$ is no longer square and has no inverse. However, the unsaturated model starts from the saturate model ignoring parameters in reverse order of the interactions (high order to low order) as desired by the researcher, always following the construction rules described in item 2 of Theorem 5. Therefore, the construction of an unsaturated model is produced by simply removing

columns in the design matrix of the corresponding saturated model. However, as the columns of the saturated model design matrix are linearly independent, any subset of the columns in it (in this case $\boldsymbol{X}$) is also such that its columns are linearly independent, whereby the unsaturated model matrix is columns full range and following the corollary B.53 of Christensen (2002, p. 415), $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exist. □