A simple two–stage method as an alternative to random effects models for binary clustered data

Guillermina Eslava[1] and Tue Tjur[2]

Abstract

The following situation is considered. A single binary response and a number of explanatory variables are given, and we want to estimate the influence of these on the response. But a simple logistic regression model can not be used because the observations occur in clusters, and responses in the same cluster are not likely to be independent. We study and compare two different approaches to the analysis of such data. (1) A two–stage model, where within cluster effects are estimated by a logistic regression model conditioned on sums of responses over clusters, and between cluster effects are estimated by an analysis of these sums by a binomial overdispersion model. (2) A logistic regression model with random (normal) cluster effect. Results from these two models are compared. We introduce these models with reference to a small example from marketing, where clusters are households, the single events are purchases of chocolate bars, the responses are indicators for the event that the brand selected was Mars, and the only explanatory variable is a measure of exposure for Mars commercials. In section 5 we illustrate by a more substantial data set from a survey concerning malnourishment of children of preschool age in rural areas of Mexico, where the clusters are towns, the response is an indicator of undernourishment, and the explanatory variables represent socioeconomic information about the children and their families.

Key words: Conditional logistic regression; Clustered survey data; Generalized linear mixed models; Binomial regression; Overdispersion.

1. Introduction

It is wellknown (see e.g. Birch 2002, Tjur 2002) that logistic regression in clustered data — in some contexts also called *longitudinal data, panel data,* or *single source data* — can produce serious inference errors when heterogeneity between clusters is ignored. In its most extreme form, this kind of error occurs when a few persons are asked essentially the same question again and again, and the data coming out of this are analysed as a representative sample, pretending that the question was posed to a new person each time. Obviously, a relevant model for such data must take into account that the persons are likely to give the same answer each time they are asked, which means that the responses can not be regarded as independent, unless

[1]Dept. of Mathematics, Faculty of Sciences, UNAM, Ciudad Universitaria, 04510, D.F. Mexico. E–mail eslava@matematicas.unam.mx. On Sabbatical leave at the Copenhaguen Business School for the academic year 2007-2008.

[2]Center for Statistics, Copenhagen Business School, Solbjerg Plads 3, DK–2000 Frederiksberg, Denmark. E–mail tuetjur@cbs.dk

the model includes an individual parameter for each person. But in more complex designs this is less transparent, and for marketing applications in particular it is not always recognised how serious this error can be. What happens is typically that explanatory variables with no effect at all on the binary response appear to be strongly significant. See Tjur (2002) for a simulation study illustrating this.

The data set we will use for illustration is taken from the British AdLab data base, created by Central Independent Television 1985–90 (see Moseley and Parfitt 1987), kindly made available to us for research purposes by Flemming Hansen, Center for Marketing Communication at Copenhagen Business School. The data set, which has been extracted from the data base by Lotte Yssing Hansen and further prepared by Kristina Birch, consists of all purchases of chocolate bars over observation periods of varying lengths, made by 560 households, adding up to a total of 11,246 such purchases, i.e. around 20 purchases per household on average. The binary response is defined as 1 if the chocolate bar happens to be a Mars Bar, 0 otherwise. The only explanatory variable considered here — called x in the following — is constructed as a weighted average of the counts of television and radio advertisements for Mars Bar that the household was exposed to on day 1, 2, ..., 28 before the purchase. The weights used in this averaging are proportional to 0.95^d, where d is the number of days passed since the advertisement. A lot of details concerning data structure, other background variables, the choice of 0.95 as the "retention rate" etc. are left out here, because they are irrelevant to the general ideas discussed.

The result of an ordinary logistic regression analysis with "Mars/not Mars" as the binary response y and a logit–linear structure consisting only of a constant term and a linear effect of the above mentioned variable x (ignoring differences between households) results in the following conclusions. The estimated probability that the hth household's ith purchase result in the choice of a Mars Bar is

$$P(y_{hi} = 1) = \Lambda(-0.8548 + 0.02007 \times x_{hi})$$

where — here and in the following — Λ denotes the function $\Lambda(z) = \exp(z)/(1 + \exp(z))$, the c.d.f. of the logitistic distribution. The approximate standard deviation of the estimate 0.02007 is reported to be 0.0044. Thus, the coefficient appears to be strongly significantly positive (p=0.000005).

2. Conditional logistic regression

The obvious problem with the analysis above is that it does not take into account that households have different preferences. One can easily imagine that some buy Mars Bars all the time and some buy other brands all the time, rather independently of the number of Mars commercials they happen to have seen. This would not be a problem in a design where each household is observed once, but when several purchases are observed for each household, the differences between households will

most likely be a lot larger than the random variation can explain.

A simple way of accounting for this is by the introduction of 560 household parameters α_h, describing these differences. Thus, a much more relevant model would assume that the probability of success in the ith purchase for household h takes the form

$$P(y_{hi} = 1) = \Lambda(\alpha_h + \beta \times x_{hi}). \tag{1}$$

The simplest and best way of estimating this model — since we are not particularly interested in the 560 household parameters — is by *conditional logistic regression*. By conditioning on the numbers of successes for each household, we obtain an expression for the conditional likelihood, where the household parameters α_h have cancelled out. This approach is usually recommended when it is numerically possible, see e.g. Breslow and Day (1980). This analysis results in the following conclusion. The estimated probability that a given purchase of chocolate bar results in the choice of a Mars Bar becomes

$$P(y_{hi} = 1) = \Lambda(\alpha_h + 0.00802 \times x_{hi}).$$

But here, an approximate standard error of 0.00795 on the coefficient 0.00802 is reported, which means that it is *not* significantly positive (p=0.31).

This conclusion is obviously more reliable than the one we ended up with in section 1. But the drawback of this model is that it can only measure the effect of the variation of x from purchase to purchase *within* households, where it is relatively constant. The effects of the (probably much more pronounced) differences between x levels for different households can not be measured, because these differences are confounded with the differences between the individual household parameters. This problem would be even more obvious if we had a covariate that was constant within households (like number of children, age of mother or whatever). The effect of this would simply cancel out when we form the conditional likelihood.

3. An overdispersion model for the cluster totals

There are two ways in which the exposure to Mars commercials could influence the household's tendency to buy Mars Bars, namely

1) *Within households.* If a household is heavily exposed to Mars advertisements, it tends to buy more Mars Bars than usual in a period thereafter.

2) *Between households.* Those households that are exposed to many Mars commercials buy, on average, more Mars Bars than those exposed to few Mars commercials.

These two types of exposure effect correspond closely to "within blocks" and "between blocks" effects in classical analysis of variance. There is no reason to

expect that they are equal in any sense. The conditional logistic regression model of section 2 is the natural analogue to the classical "intra–blocks analysis" in a block design. What remains to be done is the "between blocks analysis" or "recovery of inter–blocks information", which in the classical ANOVA setup can be done by a linear model for the block totals. The analogy suggests that we supplement the conditional logistic regression model with a model for the cluster totals, in this case the total numbers of Mars purchases for each household (i.e. the numbers conditioned on in section 2),

$$y_{h.} = \text{ the number of Mars purchases for household } h\,.$$

A simple, almost canonical (but slightly too naive) model is the one that assumes these counts to be independent, binomial with binomial totals n_h = the total number of purchases for household h and probability parameters depending logit–linearly of the households' average exposures to Mars commercials. Indeed, if the probability of success varies only little from purchase to purchase, the response $y_{h.}$ will be a sum of n_h indicators that are approximately identically distributed. The only problem is — again — that the success probability can be expected to vary much more from household to household than the variation in advertisment exposure and the binomial variation can account for. However, this is a standard problem with a standard solution; a simple model that takes this extra variation into account is the *overdispersion model* corresponding to this binomial model, where the expected responses are assumed to be of the same form as in the binomial model, but the binomial variances are modified by a common scale factor (the overdispersion parameter). See McCullagh and Nelder (1989). Following Tjur (1998), we can formally think of this as a non–linear regression model that assumes the household totals $y_{h.}$ to be approximately normal with expectations of the form

$$E(y_{h.}) = n_h\pi_h = n_h\Lambda(\alpha + \beta \times x_{h.})$$

and variances $\lambda n_h\pi_h(1-\pi_h)$, where λ, the *overdispersion factor*, is a common squared scale parameter. For $\lambda = 1$ we have the binomial model. Other variance structures might be considered, but this is by far the simplest, and the nice thing about it is that it results (for a suitable choice of estimation method, the IRLS method or quasi likelihood) in the same estimates of the covariate effects as the binomial model.

The conclusion of this model is as follows. The estimated expectations of the responses $y_{h.}$ become

$$\hat{E}(y_{h.}) = n_h\Lambda(-0.965 + 0.0437 \times x_h)$$

where x_h denotes a suitably defined average over time of household h's exposure to Mars advertisements. The test for "no effect of x" shows weak significance (p = 0.034 two–sided). This test is a T–test, correcting for overdispersion and estimation error for the overdispersion parameter. There is a strong overdispersion ($\hat{\lambda}$=8.66),

corresponding to a standard deviation which is almost three times that of the binomial model.

Thus, the overall conclusion of the two analyses is that an effect of Mars advertisments on the tendency to buy Mars bars is hardly visible in this study.

4. A logistic regression model with random cluster effect

Another way of modelling variation between households, which is more in accordance with modern thinking around random effects in generalized linear (or general nonlinear) regression models is to introduce the household effect as a random effect in the logistic regression model. That is, to replace the 560 household parameters α_h in formula 1 with 560 i.i.d. normal variables A_h with expectation α_0 and variance σ^2. This is a straightforward generalization of the variance component model known from the linear case; which, in turn, is the modern way of justifying the classical method for recovery of interblock information in the linear case.

We shall not report the results of this model because there is a conceptual problem with it. Namely that it does not split up the effect of x in a within household and a between household component. In the examples that we know of, this makes no particular sense. We would prefer, as the starting point at least, a model with separate sets of parameters describing the two types of effects. A solution to this problem, suggested by Neuhaus and Kalbfleisch (1998), is to replace the covariate x_{hi} by its two components, the "between households component" $\bar{x}_{h.}$ and the "within household component" $x_{hi} - \bar{x}_{h.}$. The coefficient to the first of these will then somehow measure the variation between households, and the coefficient to the second will measure the variation within households. Thus, our model states that, conditionally on the random household effects A_h, the binary observations y_{hi} are independent with

$$P(y_{hi} = 1) = \Lambda(A_h + \beta_{\text{between}} \times \bar{x}_{h.} + \beta_{\text{within}} \times (x_{hi} - \bar{x}_{h.})) \qquad (2)$$

where, in turn, $A_1, A_2, \ldots, A_{560}$ are i.i.d. normal with mean α_0 and variance σ^2.

This model was fitted by an approximate maximum likelihood method integrated with an adaptive Gaussian Quadrature as integration method implemented in SAS procedure NLMIXED. The result of this, together with the corresponding results from section 2 and 3, are found in table 1. The conclusion from this seems to be as follows. There is a close agreement between the results from the conditional logistic regression model and the estimated coefficient to $x_{hi} - \bar{x}_{h.}$ in the random effects model. The reason for this is probably that the conditional logistic regression model can also be interpreted as the conditional model in the random effects logistic model, because here the contribution from the household averages $\bar{x}_{h.}$ and its coefficient cancels out, leaving us with exactly the same conditional likelihood as in

section 2. If we imagine the total likelihood written as a product of this conditional likelihood and the marginal likelihood based on the observed household sums $y_{h.}$, it is rather obvious that the last factor can not contribute much to the estimation of the coefficient to $x_{hi} - \bar{x}_{h.}$. Intuitively, at least, it is hard to imagine how the household sums $y_{h.}$ can provide us with any information at all about the advertisment exposure effect which is due to variation over time within households. Consequently, the information that the model gives us about this comes almost exclusively from the conditional likelihood, which is the same as in section 2.

Conversely, columns 2 and 4 in table 1 show no overwhelming agreement between the results from the overdispersion model of section 3 and the random effects model's estimate of the coefficient to $\bar{x}_{h.}$. Not that the results are conflicting, in the sense of significantly different conclusions, but the figures indicate that we have two quite different models here. Analyzing this a little further, it can be seen that there is actually no reason at all to expect similar results from these two models. The distributions of the sums $y_{h.}$ in the two models differ both in mean and variance. For the variance component model, the expected value of $y_{h.}$ is approximately, under the simplifying assumption that the effect of $x_{hi} - \bar{x}_{h.}$ is small, so that the observations within households are almost i.i.d.,

$$n_h F(\alpha_0 + \beta_{\text{between}} \bar{x}_{h.})$$

where F is the c.d.f. of the convolution of the logistic distribution with a normal distribution with mean 0 and variance σ^2. Whereas the expression for the mean of $y_{h.}$ in the overdispersion model of section 3 is (approximately, under the same simplifying assumption)

$$n_h \Lambda(\alpha_0 + \beta_{\text{between}} \bar{x}_{h.}) \,.$$

From this it follows that if the parameters were the same in the two models, the expected value of $y_{h.}$ as a function of $\bar{x}_{h.}$ would be a more flat function in the random effects model than in the overdispersion model, in particular for large σ^2. Moreover, the variances are also different in a fundamental way. For example, the variance of the relative frequency $y_{h.}/n_h$ tends to zero as n_h tends to infinity in the overdispersion model, whereas in the random effects model this limit is positive, because the contribution from the random variation of A_h does not disappear in the limit. Thus, for a number of good reasons, the estimated regression coefficients from the two models are not comparable.

5. A second example: clustered survey data

The problem that partly motivated the present report was the statistical analysis of data from the 1996 National Survey of Nutrition in preschool children in Rural Mexico (ENAL96). One of the objectives of the study was to determine the relation between undernourished children of preschool age and some socioeconomic aspects

of the family, in rural areas of the country.

There are three standardized measures on the children that could be used to reflect undernourishment: weight for age (*wfa*), weight for height (*wfh*), and height for age (*hfa*). With the advice of the specialist interested in the study, we selected *wfa* as response variable, and selected 20 explanatory variables, which reflected measures on the child, the mother, the father, and the house. We analyzed the data and constructed some composed indices that summarized aspects of the house, the mother, and the father, respectively. These composed variables showed that most of the information from the parents is conveyed by the mother alone, and also that it was preferable to work with the original variables rather than with composed variables.

In the analysis presented in this report, we considered a binary response variable: y=0 when $wfa > -2$ which is considered a score for a normal or mildly undernourished child, and Y=1 when $wfa \leq -2$, a score for an undernourished child. We selected eight explanatory variables of which three are considered as continuous: food expenses per capita per week (*Food expenses*), number of persons per room in the house (*Persons per room*), and child's mother age at birth (*Mother's age*); and five considered as binary: material on the house's floor (*No floor layer*: 1 no layer, 0 some material); availability of latrine with running water (*No wc*: 1 no availability, 0 yes); availability of gas cooker (*No gas cooker*: 1 no gas cooker, 0 yes); formal schooling of the mother (*Mother with no schooling*: 1 none or incomplete primary school, 0 primary school or more); and mother's language (*Bilingual mother*: 1 dialect or, dialect and Spanish; 0 only Spanish). The response variable was originally measured in a scale that goes from -5 to 5; the five explanatory binary variables were originally defined for more than two categories, and adjacent categories were collapsed to obtain binary variables.

The sample design and estimated proportions for three categories of malnourished preschool children for each state and for the whole country are presented in Avila et al. (1997). Dr. A. Avila Curiel kindly gave permission to use the data for research purposes. The sample design was stratified with three sampling stages. The target population was the population in rural areas of Mexico, defined as the population contained in towns of less than 2,500 inhabitants, excluding new small towns developed as residential areas. Roughly speaking and according to the 1995 population count, there were about 95 million inhabitants in the country distributed by town size as follows: 10 millions in towns of size 1–499, 13 millions in towns of size between 500 and 2,499, and 72 millions in towns larger than 2,499. For practical and cost reasons the sample was taken from the population contained in towns of size between 500 and 2,499, and we will refer to this as the rural population.

The target population was stratified in 372 strata. One strata contained only one primary sampling unit and it was collapsed with one of its neighbours' strata.

Within each stratum between two to four towns were selected with simple random sampling; a total of 854 towns or clusters were sampled and correspond to the primary sampling units. Due to missing values the analysis is based on 853 towns. Within each town, between 2 and 49 households were sampled, this produced a total of 38,232 households, we refer to them also as families and they constitute the secondary sampling units. Within each family, a sample of one to three preschool children were selected, this gives a total of 31,601 children in the sample. In the three sampling stages, the selection of sampling units was done by simple random sampling. The data set was filtered, considering cases with children satisfying three conditions: age less than or equal to 5 years, mother's age between 12 and 50 years, and weight for age score between -5 and 5. That makes a total of 26,819 children.

For the estimation of parameters of the models presented in this report, we considered only one randomly selected child per family, which gave 18,774 children or families, after deleting missing values for the 8 selected explanatory variables, the models were adjusted using 17,865 children contained in 853 towns.

The sampling weights or expansion factors, and stratification, though available and used to compute descriptive statistics like means and proportions, were not used for the determination of associations between variables.

To analyze the data, as a first step we adjusted a logistic regression conditioning on the sums of malnourished children for each cluster or town. Here we assume that the observations on children from the same town are correlated. The number of malnourished children within a town varies from 0 up to the total number of children in the town, n_h, though in the likelihood function of the conditional regression, only towns with malnourished children varying from 1 up to $n_h - 1$ are considered, these were 714 out of 853. The coefficients adjusted in the conditional logistic regression, formula (1), measure the association between the explanatory variables and the probability of one child being undernourished in a given town, that is the within towns effects.

The estimated coefficients, displayed in column 3 in table 2, show that five variables have a significant effect on the response variable ($p < .01$), variable *No wc* has a non significant effect, whereas *Food expenses* and *Mother's age* are only sligthly significant ($.01 < p < .05$).

In a second step, in order to estimate the between towns effects, we adjusted an overdispersion binomial model on the cluster sums for the response and cluster means for the explanatory variables, that is on 853 observations. The squared scale parameter λ was estimated as 1.3 reflecting a slightly larger dispersion than a binomial model could account for. Estimated coefficients, displayed in column 2 in table 2, show the following. Two variables have no significant effect on the response variable, *Mother's age* and *No floor layer,* variable *No gas cooker* is only slightly

significant ($.01 < p < .05$), and the other five variables are significant ($p < .01$) including *No wc* whose within towns effect is not significant.

Considering the effects of the explanatory variables on the response, estimated separately as within and between towns, columns 2 and 3 in table 2, we observe that although estimated within and between towns effects for each variable have equal sign, four different patterns actually occur:

a) Both effects are significant, $p < .01$, and their confidence intervals overlap; as in variables *Persons per room*, and *Bilingual mother*

b) One effect is significant or slightly significant and the other is not. As in variables *Mother's age, No floor layer,* and *No wc.*

c) One effect is significant, $p < .01$, and the other is only slightly significant ($.01 < p < .05$); as in variables *Food expenses,* and *No gas cooker.*

d) Both effects are significant and, though with equal sign, their confidence intervals do not overlap; as in variable *Mother with no schooling* where the between towns effect is about 3 times larger than the within towns.

We also observe a larger variability in the between towns effects than in the within towns in all variables except in variable *Bilingual mother.*

As an alternative to measure within and between towns effects, we adjusted a logistic regression model with random cluster effect (intercepts), as it has been suggested before for data from cluster sampling, by e.g. Agresti et al. (2000, sec. 3.8), but with two sets of parameters, one for the between towns effects attached to cluster mean values, and the other for the within towns attached to deviations from cluster mean values. The estimated variance for the normal distribution of the random effect was $\hat{\sigma}^2 = .28$, which shows a small variation. Estimated parameter values and standard deviations are presented in the last column of table 2 and estimated 95% confidence intervals are displayed in figure 1. We observe that estimated effects and standard deviations are similar to the corresponding ones estimated by a conditional regression and an overdispersion model, as described in a) - d); except for the between towns effect in variable *No gas cooker* which is not significant under the random effects model and is slightly significant under the overdispersion model ($.01 < p < .05$).

Numerically, parameters measuring within towns effects estimated under a conditional logistic regression are very close to the corresponding ones under the logistic regression with random intercepts, and even more similar are their standard deviations. Parameters measuring between town effects estimated under an overdispersion binomial model are similar to the corresponding ones under the logistic regression with random intercepts, and standard deviations are slightly larger under the logistic regression model with random intercepts.

When we adjusted a random intercept model, we made a statistical test for equal

between and within towns effects for each of the eight explanatory variables, and we found the following. i) The effects are statistically different for two variables: *No wc* (two sided test, $p = .004$, delta method for variance estimation) and *Mother with no schooling* ($p = .001$), though in the first one the within towns effect is not significant; ii) the effects are statistically different with $.01 < p < .05$, for the two variables *Food expenses* and *Persons per room* and iii) the effects can be considered as equal for the other four variables, *Mother's age, No floor layer, No gas cooker,* and *Bilingual mother,* though, in the first three the between towns effect is not significant.

6. Discussion

Our study suggests that, as far as within cluster effects are concerned, there is very little difference between the two models considered. But when it comes to between cluster effects, the two models are different and difficult to compare, in particular if the overdispersion and the variance σ^2 of the random cluster parameters are large. Both models have some advantages and disadvantages.

The binomial overdispersion model is easy to understand because it has a simple formula for the expected cluster totals. But it also has the property that the variance of a cluster average tends to zero when the cluster size tends to infinity, which seems somewhat unrealistic for the kind of cluster effects we have in mind. Another disadvantage may be that it does not make much sense, in this model, to ask whether the between and within cluster effects are the same.

The random effects model is more complicated, because it has random normal effects on the logit scale, which results in complicated expressions for mean and variance of cluster sums etc. Moreover, situations where the between and within effects are expected to be different can only be handled by the somewhat artificial split of explanatory variables in two. It has, on the other hand, the advantage that it makes sense to ask (and perform a statistical test to answer) the question of whether a covariate has "the same" effect within and between clusters.

If the latter property is considered meaningful, the random effects model should probably be preferred. If not, we would tend to prefer the simpler "two–stage method".

References

Agresti, A., Booth, J.G., Hobert, J.P. and Caffo, B. (2000). Random–effects Modeling of Categorical Response Data. Sociological Methodology, 30, 27–80.

Avila Curiel, A. Shamah, T, and Chávez, A. (1997). Encuesta Nacional de Ali-

mentación y Nutrición en el Medio Rural 1996. Instituto Nacional de la Nutrición Salvador Zubirán (Internal document in Spanish).

Birch, K. (2002). Analyzing effects of advertising using conditional logistic regression. Preprint no. 2, Dept. of Management Science and Statistics, Copenhagen Business School.

Breslow, N. E. and Day, N. E. (1980). Statistical Methods in Cancer Research WHO, International Agency for Research on Cancer, Lyon.

Hansen, L.Y. and Hansen, F. (2001). Advertising and promotion effectiveness — learnings from a five year study. Research Paper no. 18, Advertising Research Group, Dept. of Marketing, Copenhagen Business School.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman and Hall.

Moseley, S. and Parfitt, J. (1987). Measuring advertising effect from single–source data: the first year of the AdLab panel. Admap, June 1987, 26–33.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between– and Within–Cluster Covariate Effects in the anlysis of Clustered Data. Biometrics, 54, 638–645.

Nordmoe, E.D. and Jain, D.C. (2000). Drawing inferences from logit models for panel data. Applied Stochastic Models in Business and Industry, 16, 127–145.

SAS (2007). The SAS system for windows. Release 9, Cary, NC: SAS Institute INC.

Tjur, T. (1998). Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models. The American Statistician, 52, 222–227.

Tjur, T. (2002). Logistic regression models for single–source data — a simulation study. Preprint no. 4, Dept. of Management Science and Statistics, Copenhagen Business School. (http://staff.cbs.dk/tuetjur/02-4.pdf)

NB At the end of the text, in this report, we include some material not referred in the text: figures 2 and 3, sas syntax.

```
%Mars data;
PROC genmod DATA=WORK.long_11246 descending;
model y = expo/ dist=binomial; run;
PROC genmod DATA=WORK.long_11246 descending;
class househ;
model y = expo househ/    dist=binomial; run;
proc logistic DATA=WORK.long_11246;
      strata househ;
      model y(event='1')= expo/clodds=Wald; run;
proc nlmixed DATA=WORK.long_11246 start;
parms b0= -0.8548 b1=0.0201 sigma=1.0;
eta = b0 + b1*expo  + u;
p=exp(eta)/(1 + exp(eta));
```

```
model y ~ binary(p);
random u ~ normal(0,sigma) subject =househ; run;
proc nlmixed DATA=WORK.long_11246 start;
parms b0= -0.9653  b1=0.0437 b2=0.0042  sigma=1.0;
eta = b0 + b1*expo_househ + b2*expo_d + u;
p=exp(eta)/(1 + exp(eta));
model y ~ binary(p);
random u ~ normal(0,sigma) subject =househ; run;
PROC genmod DATA=WORK.long_household;
model y_sum/n_break = expo_househ/dist=binomial scale=pearson;
run;
%Children data;
libname datos 'c:\Eslava_07\clogit';
PROC genmod DATA=datos.finfil1f_m17865 descending;
model wfa2 = foodexp persroom agemoth floor01 wc01_psu wc01_d cooker01
             schoolm01_psu schoolm01_d languagem01
             /dist=binomial;
 run;
proc nlmixed DATA=datos.finfil1f_m17865 start;
parms b0= -2.3299 b1=-0.006 b2= 0.0543 b3=-0.0083
      b4= 0.2045  b5= 0.5122  b6=0.0472 b7=0.3422 b8=0.7601
      b9=0.1683 b10=0.4143
      sigma=0.5;
eta = b0 + b1*foodexp + b2*persroom + b3*agemoth +
      b4*floor01 + b5*wc01_psu + b6*wc01_d +
      b7*cooker01 +
      b8*schoolm01_psu + b9*schoolm01_d +
      b10*languagem01 + u;
p=exp(eta)/(1 + exp(eta));
model wfa2 ~ binary(p);
random u ~ normal(0,sigma) subject =psu;
estimate 'wc01_psu-wc01_d' b5 - b6;
estimate 'schoolm01_psu-schoolm01_d' b8 - b9;
run;
*Rerun Conditional and binomial to save parameter values on outfile;
proc logistic data=datos.finfil1f_m17865 covout outest=datos.cond_pars simple;
      strata psu/info;
      model wfa2(event='1')= foodexp persroom agemoth floor01
      wc01 cooker01 schoolm01 languagem01/clodds=Wald;
run;
*Rerun nonlinear regression to save parameter values on outfile;
PROC genmod DATA=datos.finfil1f_m17865_psu;
model event/n_break = foodexp_psu persroom_psu agemoth_psu floor01_psu wc01_psu
  cooker01_psu schoolm01_psu languagem01_psu/dist=binomial scale=pearson;
run;
```

Table 1: Panel data. Between and within households effects. Parameter estimates and standard errors for three models. An overdispersion binomial model, a conditional logistic regression, and a logistic regression with random intercepts with explanatory variable values split into household mean and deviation to its household mean

Variable	Two–stage modelling		One–stage modelling
	Overdispersed[1] binomial model	Conditional[2] logistic reg.	Logistic regression[3] with rand. int.
Intercept	-.9653 (.1143)		-1.0468** (.1804)
$\bar{x}_{h.}$	.0437 (.0205)		-.01061 (.0318)
$x_{hi} - \bar{x}_{h.}$		.00802 (.00795)	.00797 (.00793)

[1] Model adjusted using 560 observations (households) sumarized from 11,246 purchases. Estimated scale parameter $\sqrt{\hat{\lambda}} = \sqrt{\chi^2/df} = \sqrt{4832.29/558}$=2.9428

[2] Model adjusted using 9,757 observations (purchases) grouped into 302 households.

[3] Model adjusted using 11,246 observations (purchases) grouped into 560 households. Estimated variance for the random intercept $\hat{\sigma}^2$ =4.56.

** p <.01

Table 2: Clustered survey data. Between and within towns effects. Parameter estimates and standard errors for three models. An overdispersion binomial model, a conditional logistic regression, and a logistic regression with random intercepts with explanatory variable values split into group means and deviations to its group mean

Variable	Two–stage modelling		One–stage modelling
	Overdispersed[1] binomial model	Conditional[2] logistic reg.	Logistic regression[3] with rand. int.
Intercept	-2.2921**(.4091)		-2.3302**(.4370)
Food expenses $\bar{x}_1$	-.0108** (.0033)		-.01247**(.0034)
Persons per room $\bar{x}_2$	.1530** (.0447)		.1745** (.0497)
Mother's age $\bar{x}_3$	-.0161 (.0138)		-.0209 (.0145)
No floor layer $\bar{x}_4$	.1541 (.1317)		.1144 (.1484)
No wc $\bar{x}_5$	.4149** (.0926)		.4281** (.1043)
No gas cooker $\bar{x}_6$	.3213* (.1429)		.2873 (.1518)
Moth. with no sch. $\bar{x}_7$	.6165** (.1335)		.7395** (.1534)
Bilingual mother $\bar{x}_8$	.4088** (.0790)		.4581** (.0930)
Food expenses $x_1 - \bar{x}_1$		-.0034* (.0017)	-.0034* (.0017)
Persons per room $x_2 - \bar{x}_2$		.0470** (.0118)	.0468** (.0118)
Mother's age $x_3 - \bar{x}_3$		-.0071* (.0033)	-.0070* (.0033)
No floor layer $x_4 - \bar{x}_4$		.1982** (.0548)	.1987** (.0549)
No wc $x_5 - \bar{x}_5$		.0853 (.0575)	.0833 (.0575)
No gas cooker $x_6 - \bar{x}_6$		.2569** (.0674)	.2605** (.0675)
Moth. with no sch. $x_7 - \bar{x}_7$		.1967** (.0547)	.1980** (.0549)
Bilingual mother $x_8 - \bar{x}_8$		.2868** (.1050)	.3015** (.1065)

[1] Model adjusted using 853 observations (towns), summarized from 17,865 children. Estimated scale parameter $\sqrt{\hat{\lambda}} = \sqrt{\chi^2/df} = \sqrt{1425.17/844}$=1.2995

[2] Model adjusted using 15,727 observations (children), grouped into 714 towns.

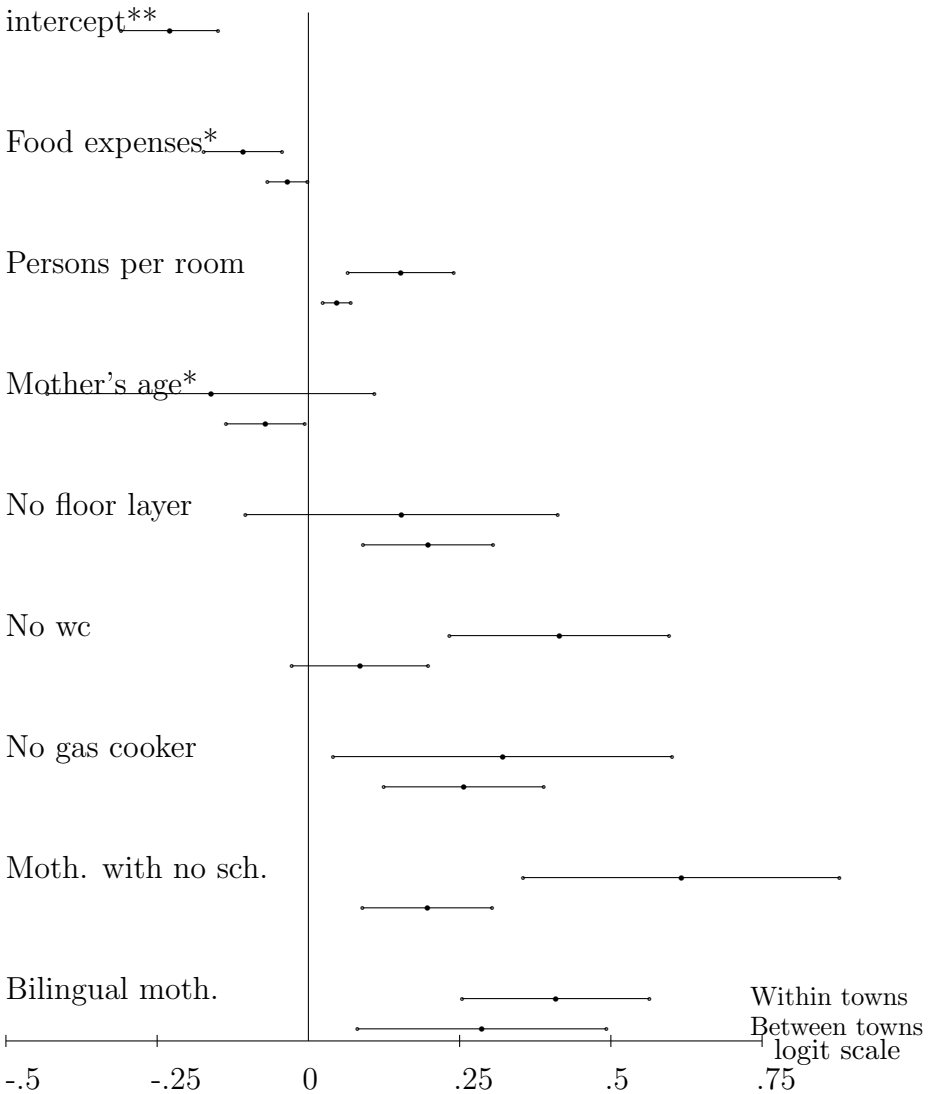
[3] Model adjusted using 17,865 observations (children) grouped into 853 towns. Estimated variance for the random intercepts $\hat{\sigma}^2$ =.2765

$*$ $.01 < p < .05$ $**$ $p < .01$

Figure 1: Between and within towns effects. Parameter estimates and 95% confidence intervals, for each of the eight variables associated with an increase in the probability of a child being malnourished. Between and Within towns effects estimated simultaneously with a logistic regression with random cluster effect with explanatory variable values split into cluster means and deviations to its cluster mean. Eq. 2

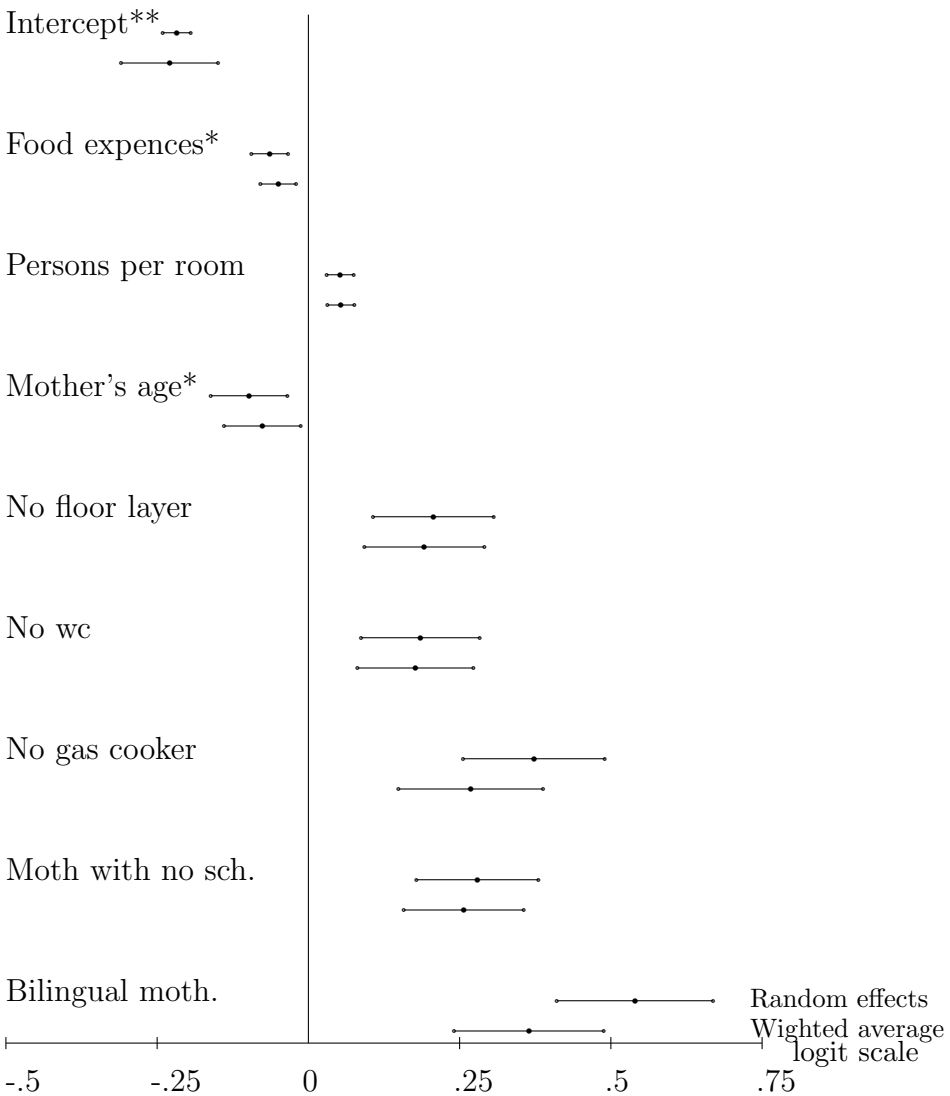
**values divided by 10; * values multiplied by 10.

Figure 2: Within- and between– towns effects. Parameter estimates and 95% confidence intervals, for each of the eight variables associated with an increase in the probability of a child being malnourished. Within– towns effects estimated with a conditional logistic regression, and between–towns effects with an overdispersed binomial model.



**values divided by 10; * values multiplied by 10.

Figure 3: Average population effects. Parameter estimates and 95% confidence intervals, for each of the eight variables associated with an increase in the probability of a child being malnourished. Estimates calculated with two models: a weighted average of estimates from within towns effects (conditional logistic regression) and estimates from between towns effects (overdispersed binomial model); and with a logistic regression with random intercepts.



**values divided by 10; * values multiplied by 10.