ORDER SELECTION, RANDOM EFFECTS, AND MULTILEVEL PREDICTORS IN MODELLING DECENNIAL CENSUS RESPONSE

Eric V. Slud, Census Bureau & Univ. of Maryland
Mathematics Department, University of Maryland, College Park MD 20742

Key words: logistic regression; multi-level models; nonresponse adjustment; random effects, targeting database; weighting.

Abstract: Models of household (HU) response to the decennial census and other large national surveys, by mail and at various stages of enumerator followup, are important both in targeting survey resources and in weighting adjustments for nonresponse. Targeting involves modelling of response rates at an aggregated level, such as block-group, and may involve predictors at the level either of the individual HU or of an aggregate. Weighting requires individual-level models of HU response, by mode.

Previous work on modelling of response to the 1990 decennial census has focused on only one level, individual response in terms of individual HU characteristics or aggregated response rates in terms of block-group or tract predictors. The present research goes farther in two ways, with results reported for DE, MD, and GA:

(i) Statewise logistic regression models of individual propensity to respond by mail are developed in terms of the individual HU characteristics together with block-group predictors, and both kinds of predictors turn out to be important in the models.

(ii) Selection of predictors is aided by inclusion of random effects, which dramatically diminish overfitting by reducing the contribution (deviance) of interaction-terms among predictors.

This paper describes research and analysis undertaken by Eric Slud, and is released to inform interested parties and encourage discussion. Results and conclusions are the author's and have not been endorsed by the Census Bureau.

1. INTRODUCTION

The 5–15 % segment of the US population which fails to respond to the Decennial Census either by mail or to early attempts at enumerator followup is the source of much of the cost and difficulty of the Decennial Census. Many systematic attempts have been made to understand the demographics of households which fail to respond by mail (of which the Targeting Database of G. Robinson and co-workers in Population Division of the Census Bureau is an ongoing example, *cf.* Robinson and Kobilarcik 1995). The ultimate objective is the efficient management of Census Bureau data-collection resources, both in future Decennial Censuses and in other surveys such as the American Community Survey. The targeting of resources on the basis of demographic differences could be implemented either in the form of increased duration of followup enumeration in hard-to-count neighborhoods or, less controversially, in expanded advertising and outreach. However, such targeting might best be accomplished with respect to population segments which are judged hard-to-count in terms of rates of response to the Census at later stages of enumerator followup and not simply of response by mail. Some differences between the demographic models for response by mail and at later stages of followup have previously been studied by the author in Slud (1999).

The demographics of response to large national surveys can be understood in two distinct ways: in terms of characteristics of individual households, and in terms of aggregated characteristics of larger areas such as block-groups. Aggregated census short-form responses over block-groups, along with geographic and housing-type information, can serve as neighborhood-level predictor variables, while enumerated-household characteristics can serve as unit-level predictors. Models of national mail response in terms only of enumerated unit-level variables were previously studied by Word (1997) and Causey (1998). The Census Bureau's Targeting Database for Census 2000 involved modelling the propensity of enumerated HU's to respond by mail, by tract, in terms of short- and long-form Census characteristics aggregated to tract level and recoded. Models of response by mail and to enumerators within intervals of followup time, in terms only of block-group aggregated short-form predictors, were studied by Slud (1998b, 1999).

The purpose of this paper is to examine detailed models of mail response to the census, using demographic and geographic predictors from 1990 decennial data by state. The two related issues addressed here which go beyond previous research are: (a) how to use mixed-effect logistic models to eliminate large numbers of predictors and interactions which, while 'statistically significant' in models with fixed-effects only, are neither interpretable nor predictive; and (b) assessment of the relative importance to mail response within the same model of characteristics of individual households versus neighborhood (block-group) level predictors.

Acknowledgment. The comments of Mary Mulry as discussant were very helpful in focusing the objectives of the paper. The research was supported by Census Bureau contract 43-YA-BC-038053.

2. DATA FILES

The unit of study is the (non-group-quarters) housing unit (HU) in the 1990 **CENSAS** *100% Edited Detail File* database restricted to mailout-mailback areas (*type of enumeration area* 1, 2, or 4, where forms are delivered and can be mailed back before an enumerator visits). Geographic variables collected were: **plcod** for reservation/rural/small-urban/larger-urban, and a numerical place-size code **plsz** from 0 to 19), in addition to location identifiers (**state**, **county**, **district office**, **tract**, **ARA** = *Address Register Area*), and **BG**=*block-group*). Variables collected by HU include *housing type* (**htyp**) which is coded as Mobile-home = 0, Single-family home = 1, and Other (mostly Apartments) = 2. Other explanatory variables were aggregated over census short-forms collected from HU's in block-groups. The aggregated variables were **fspou** (the fraction of enumerated HU's which contain the spouse of the head-of-household or a head-of-household aged at least 50); **fown** (the fraction of enumerated HU's owned rather than rented); **focc** (the fraction of *occupied* HU's); **fb** (the fraction of enumerated *persons* with racial category *black*); **fnp7** (the fraction of enumerated HU's containing at least 7 enumerated persons); **funr** (the fraction of enumerated HU's with at least one person unrelated to the head-of-household); and **fhisp** (the fraction of enumerated HU's with Hispanic head-of-household). Each of these demographic BG proportion-covariates p was also re-coded into a *logit* score $\log(p/(1-p))$. Two further re-coded indicator variables have been used in analyses: **Sing** indicating whether more than two-thirds of the HU's in a BG are single-family homes; and **I(plsz=0)** indicating that the *plsz* code is 0. These are the BG aggregated variables which were previously studied in Slud (1998b).

In addition, statewise and national data-files from the 1990 census were created to cross-tabulate the following responses by *enumerated* HU: indicators of ownership (**own**), black race (**black**), spousal unit (**spou**), hispanic ethnicity (**hisp**), and unrelated HU member (**unr**). Thus, for each state, predictor variables fall into the geographic or aggregated BG category, or into the category of individual-household descriptors, all of which other than **htyp** are available only when a HU is enumerated. The response variable treated in this study is Mail-response.

3. MODELS & ESTIMATION

The primary statistical tool used in model-fitting is the logistic model for mail-response. Within a single state, let $i = 1, \ldots, m$ index the **BG**×**htyp** strata described above, and let $j = 1, \ldots, n_i$ index the individual enumerated households within the i'th stratum. Let $y_{ij} = 0, 1$ denote the indicator of mail-response for the j'th HU within the i'th stratum, X_{ij} denote the vector of predictors for that HU (including a first component of 1), and $y_i = \sum_j y_{ij}$. The predicctors will include some components particular to the HU, and some 'BG-level' components which are common to all HU's within the same block group. Initially, as in Slud (1998b), we fit only the fixed-effect logistic model which posits a constant coefficient-vector β (of the same dimension as each X_{ij}) for which all y_{ij} are independent with

$$P(y_{ij} = 1) \;=\; e^{\beta' X_{ij}}/(1 \,+\, e^{\beta' X_{ij}})$$

Slud (1998b) describes how the model including *only* BG-level predictors and interactions up to third order was reduced, using step-down and BIC upon DE data, to a set of 52 predictors (including interactions), and the same predictors were used in fitting models for all other states.

In order to make further progress, we adopt the *mixed-effect logistic* model which assumes the existence of independent random intercept variables $u_i \sim \mathcal{N}(0, \sigma^2)$, one for each stratum, such that the responses y_{ij} are conditionally independent given the u_i, with

$$P(y_{ij} = 1 \mid u_i) \; = \; e^{\beta' X_{ij} + u_i} / (1 \, + \, e^{\beta' X_{ij} + u_i}) \tag{1}$$

Other mixed-effect models have been tried, with independent random effects for BG instead of stratum, and with multidimensional random-effects entering as coefficients of observed HU variables such as **htyp**, but those models were not found to fit better than (1) and are not further discussed here. The model (1) already accommodates some model-prediction errors by allowing purely random differences in the propensity to respond through the independent identically distributed random intercepts u_i.

The fitting of the unknown coefficients β and random-effect variance σ_u^2 by Maximum Likelihood in state datasets, which at a minimum (for DE) have 727 strata and 52 predictors, is already a difficult computational problem using available commercial software (Slud 1998, 2000). However, a very efficient computational method can be based on the adaptive Gaussian quadratures idea of Pinheiro and Bates (1995) implemented in their Splus **nlme** function for nonlinear regression, and has been implemented by the author in Splus (Slud 2000).

4. RESULTS

Logistic models with and without random intercepts were fitted to enumerated HU data for individual states, with the objective of understanding differences between the statewise models and the magnitude of contributions due to predictors at the BG as opposed to HU level. The primary goal was to find not only ‘statistically significant’ predictors in models, but effects which are predictively important. We report results for DE, MD, and GA.

First, in each of the three states two models (corresponding to the first two rows in each of Tables 1–3) were fitted to the **BG**×**htyp** stratified data using BG level predictors. The first of these is the complete model of Slud (1998b), including many BG-level interaction terms; the second is obtained by stepping out most of those interactions which, although apparently significant from the decrement in *logLik* for the fixed-effect logistic model (the first numerical column of each Table), are not highly significant in view of the much smaller difference between *logLik*'s (given in column 2 of Tables 1– 3) for the mixed-effect logistic model (1) with a single random intercept.

Table 1: Summary of model-fitting comparisons for DE 1990 Census data. In these data, 212767 enumerated HU's were included, with 727 **BG**×**htyp** strata in first two models and 6371 **HU**×**BG** strata for enumerated HU's. The number out of the 96 HU-covariate-defined cells containing more than 50 HU's was 50, and the Weighted Sum of Squares in the Enumerated-HU data due to HU-covariate stratification within **BG**×**htyp** is 3332.7.

Model	logLik	logLik	chisq	WtSS	AbSS
	Fixed	Mixed	#> 50		
BG.52	-122332	-120772	*	1021	10544
BG.22	-122550	-120817	*	1151	11210
BG.21	-112446	-106927	2171	4430	22336
HU.09	-109150	-105916	680	3214	17668
HU.18	-108642	-105778	240	3014	17340
HBG22	-107821	-105627	226	2679	16207

Next, for each state a BG-level model (corresponding to the third row in each Table) was fitted on enumerated HU data using the stepped-down set of predictors from the second model (except for **fnp7**, which was not retained in the more finely stratified HU-level file), namely **intercept** plus

```
fb          fown        fown^2      fspou
fspou^2     plcod       htyp        plsz
I(plsz=0)   focc        Sing        focc^2
fown^3      plcod*htyp  fown*htyp
plcod*fown  plcod*focc
```

where **htyp** is a 3-level factor and **plcod** is also in states (MD and GA but not DE) with large cities. Then a model (the first of the HU models, the fourth row in each Table) was fitted on the enumerated-HU

Table 2: Summary of model-fitting comparisons for MD 1990 Census data. In these data, 1741210 enumerated HU's were included, with 5561 **BG**×**htyp** strata in first two models and 46270 **HU**×**BG** strata for enumerated HU's. The number out of the 96 HU-covariate-defined cells containing more than 50 HU's was 65; and the Weighted Sum of Squares in the Enumerated-HU data due to HU-covariate stratification within **BG**×**htyp** is 22536.5.

Model	logLik Fixed	logLik Mixed	chisq #> 50	WtSS	AbSS
BG.60	-991972	-978735	*	8558	87888
BG.26	-993894	-978908	*	9401	92705
BG2.26	-903903	-864332	12114	30699	170539
HU.10	-883329	-857970	4558	23800	139920
HU.21	-879518	-856965	2130	22360	135143
HBG27	-874979	-856159	1729	20458	129604

Table 3: Summary of model-fitting comparisons for GA 1990 mailout-mailback Census data. In these data, 1651578 enumerated HU's were included, with 5752 **BG**× **htyp** strata in first two models and 49499 **HU**×**BG** strata for enumerated HU's. The number out of the 96 HU-covariate-defined cells containing more than 50 HU's was 65; and the Weighted Sum of Squares in the Enumerated-HU data due to HU-covariate stratification within **BG**×**htyp** is 23022.4.

Model	logLik Fixed	logLik Mixed	chisq #> 50	WtSS	AbSS
BG.53	-1060078	-1044743	*	10400	97064
BG.27	-1062436	-1044958	*	11429	101827
BG.26	-948453	-913288	13286	32108	171333
HU.10	-926855	-906732	3669	23691	139455
HU.21	-924092	-905914	1685	22616	136338
HBG31	-921289	-905301	1452	21508	132119

data using only the HU predictors

black, hisp, htyp, spou, own, unr, plcod

A more complicated model (the second HU model, the fifth row in each Table) included also the derived and interaction HU predictors

I(plsz $= 0$**)**, **I(plsz** > 12**)**, **black*own**, **htyp*own**, **black*spou**, **black*plcod**, **spou*plcod**, **spou*I(plsz**> 12**)**

Finally, the last row of each Table corresponds to a fitted 'HBG' model with fixed-effect predictors including all of those in the second HU model together with as many of the BG predictors

fb, fspou, fown, focc, fb^2**, black*fb, spou*fspou fb*plcod, own*fown, fown*spou, black*fown**

as were significant in the sense of producing deviance increments (that is, increments of the maximized log-likelihood multiplied by 2) of at least 80 per degree of freedom in the fixed-effect logistic model.

Within each fitted model, the number of fixed-effect predictors is indicated in Tables 1–3 by the suffix (e.g., 60 for model BG.60 in the MD **BG**×**htyp** stratified CENSAS data). Log-likelihoods are calculated both for the fixed-effect model and for the model (1) with a random intercept, but the likelihoods for the first 2 models in each Table are not comparable with those for the last 4 models, because the datasets are different. Throughout, the *logLik* increments are much less dramatic, but still important, between the models with random intercepts than between the fixed-effect models with the same sets of fixed-effect predictors. The fitted standard deviations in the random intercept models depend both on the quality of the fitted model and on the level of stratification in the data. These estimated standard deviations σ_u for random effects u_i are in the range 0.35–0.45 for the first two BG models in each state; around 0.5 for the third BG and the HBG model; and 0.55–0.7 for the HU models.

The goodness of fit of the models is assessed in three ways: by a chi-squared statistic, a stratum-size-weighted sum of squared deviations between fitted and observed mail response rates, and a sum of absolute deviations between fitted and observed numbers of mail-responders, by stratum. The chi-square statistic, applied only to the models for the enumerated HU data, was based on 96 cells generated by the categorical variables **htyp**, **own**, **urban**, **spou**, and a 4-level **race** variable encoding **black** and **hisp**. In the Tables, the chi-square terms are summed only over cells containing at least 50 HU's, but the results

Table 4: Coefficients in simplest HU model (1) with random intercept, for each of DE, MD, and GA. Estimated standard errors for the 10 fitted MD.HU coefficients are respectively: $.010, .009, .017, .009, .004, .009, .009, .010, .005, .004$.

	DE.HU	MD.HU	GA.HU
Intercept	0.297	0.439	0.262
black	-0.855	-0.570	-0.498
hisp	-0.610	-0.426	-0.574
htyp1	0.318	0.328	0.405
htyp2	0.033	0.029	0.066
spou	0.426	0.352	0.300
own	0.802	0.830	0.760
unr	-0.342	-0.259	-0.210
plcod1	0.249	0.233	0.129
plcod2	0.000	0.042	0.044
σ_u	0.546	0.560	0.503

remain very similar as long as cells with fewer than 5 HU's are excluded. The other measures of goodness-of-fit are defined by

$$\mathbf{WtSS} \;=\; \sum_{i=1}^{m} n_i \, (\frac{y_i}{n_i} \;-\; \hat{P}(y_{i1}=1))^2$$

$$\mathbf{AbSS} \;=\; \sum_{i=1}^{m} |y_i \;-\; n_i \, \hat{P}(y_{i1}=1)|$$

While different states use essentially the same set of predictor variables in the fitted models, the models themselves do vary noticeably by state. We display this in two ways. First, in Figure 1 (plotting symbols **M** and **G**), the observed mail-response rates for each of MD and GA are shown cross-classified by the variables **htyp**, **urban**, and **own**. The classification is seen to give a similar pattern of low to high response rates in these two states, but the rates do show clear stratumwise differences in addition to the overall differences in mail-response rates (76.5% in MD, 71.0% in mailback areas of GA). Statewise differences between the 10-predictor HU models, without interactions but with random intercept, can be seen in Table 4. Although the HU models use exactly the same predictors (which have the same levels, in the case of categorical variables **htyp** and **plcod**), most of the coefficients are very significantly different across states. (See the estimated standard errors of MD.HU coefficients in the caption of Table 4.)

Figure 1 shows graphically the differences between the observed MD and GA mail-response rates and among the various GA fits, calculated as rates within the 12 covariate-defined strata defined by cross-classifying **htyp** (Mb, Sg, or Ap), **urban** (Ur or Ru), and **own** (Ow or Rt). The best of the models, the HBG model, corresponds to plotting symbol **3**, and generally, across almost all of the strata, these fitted rates are markedly closer to the actual GA rates (corresponding to plotting symbol **G**) than are the fitted rates from the other models. The discrepancies between the actual rates and those from the HU and BG models (symbols **1** and **2**) are generally of the same order as betwen the actual GA and MD rates (the latter plotted with symbol **M**).

5. Conclusions.

The Tables yield a few simple conclusions:

(a) the HU-level effects have much stronger predictive value than the BG-level effects,

(b) the 'HBG' model for each state, which incorporates both HU and BG level predictors, improves on the second HU model by at least as much as the second HU model improves on the first, and

(c) there are no comparable HU-level predictors beyond those in the second HU model which are as strong as the BG-level predictors.

The final goodness-of-fit **AbSS** column in Tables 1–3 indicates that roughly 5% of HU's appear to be misclassified with respect to Mail-response by the first two BG models, when data are tallied by **BG**×**htyp**, while the misclassification-rate for enumerated HU's tallied by **BG**×**htyp**×**urban**×**own**×**spou**×**race** ranges from 10% down to 7% as we progress from the third BG toward the HBG model.

Research on the combined multilevel modelling of 1990 Census response is continuing, with 'responses' defined as occurring within either the first 20 days of enumerator followup within ARA, or between 20 and 40 days of followup. Data on 2000 Decennial Census will soon be available to update these analyses.

6. References

Causey, B. (1998) A study of predictors of nonresponse in the Decennial Census. Census Bureau preprint.

CENSAS User Handbook (1996). U.S. Bureau of the Census Systems Support Division, July 1996.

Pinheiro, J. & Bates, D. (1995) **lme** and **nlme**: Mixed Effects, Models, Methods, and Classes for S and Splus, version 1.2. Documented at **http://lib.stat.cmu.edu**

Robinson, J. & Kobilarcik, E. (1995) Identifying differential undercounts at local geographic levels: a targeting database approach. Paper presented at Apr. 1995 annual meeting of the Population Assoc. of Amer.

Slud, E. (1998) Logistic regression with large cell-counts and multiple-level random effects. Preprint.

Slud, E. (1998b) Predictive models for decennial census household response. *Proc. Amer. Statist. Assoc. on Survey Res. Meth.* 272-277.

Slud, E. (1999) Analysis of 1990 decennial census checkin-time data. Proc. Fed. Comm. Statist. Methodology Res. Conf., Pt. 2, pp. 635–44, June 2000, *Statistical Policy Working Paper 30, OMB.*

Slud, E. (2000) Accurate calculation and maximization of log-likelihood for mixed-effect logistic regression. Preprint.

Word, D. (1997) Who Responds ? Who Doesn't ? Analyzing variation in mail response rates during the 1990 Census. *Census Bureau Population Division Working Paper No. 19.*

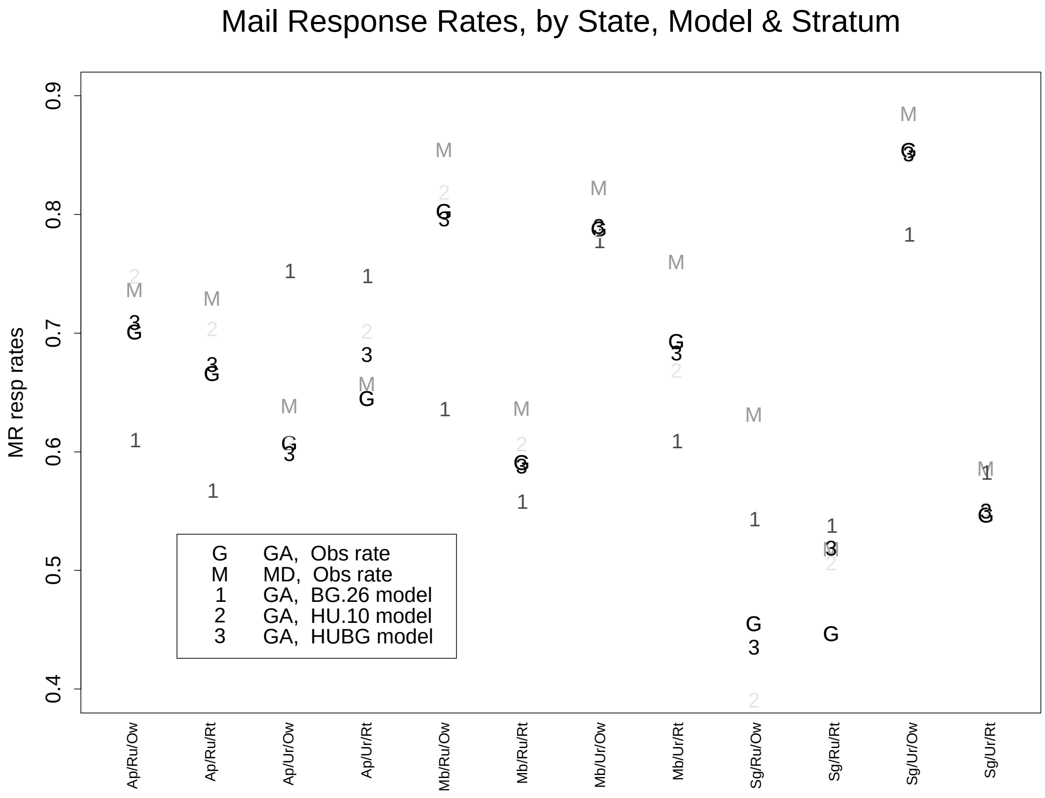

Figure 1. Plot of observed (MD and GA) and fitted (GA models BG.26, HU.10, and HBG31) mail-response rates within strata of enumerated HU's, indicated on x-axis, defined by HU variables **htyp**, **urban**, and **own**.