ANALYSIS OF 1990 DECENNIAL CENSUS CHECKIN-TIME DATA

Eric V. Slud
Census Bureau & University of Maryland

Abstract: 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 attempts have been made to understand the demographics of households which fail to respond by mail. However, relatively little work has been done to relate demographic variables to the duration of enumerator followup before a household is enumerated. The present research models (non-group-quarters) household decennial census response longitudinally with respect to enumerator checkin times relative to earliest checkin-times within ARA, using demographic explanatory variables which characterize census block-groups rather than households. The resulting models include statewise logistic regressions of several indicators of 1990 household response during quantile-intervals of ARA enumerator-followup time. Summary analyses are also given on a highly aggregated array of response rates by intervals of checkin time, cross-classified by state and 48 demographic strata.

This paper reports on research and analysis undertaken by Eric Slud, and is released to inform interested parties and to encourage discussion. Results and conclusions expressed are those of the author and have not been endorsed by the Census Bureau.

1. Introduction

The study of statistical models of household response to the decennial census is motivated by several potential applications. First, indications of which localities will be hard to count can improve various aspects of planning at the field office level (Robinson & Kobilarcik 1995). Second, understanding of which combinations of demographic characteristics are associated with homogeneous patterns of response can suggest an objective basis for the formation of *poststrata* with which to attempt to correct the census count through a post-enumeration survey (Hogan 1993). A third, related, type of application is to create or justify algorithms used to impute household characteristics missing from an incomplete form.

Previous research on models for individual (person or household) census response has been confined primarily to predictive variables derived from census forms of the persons or housing units whose responsiveness was being modelled. (See Slud 1998, 1999 for citations to Census Bureau reports and ASA Proceedings.) Such models of individual response all concerned characteristics for *enumerated* persons or households and were in that sense probability models for responses *conditional upon having been enumerated*.

By contrast, the objective of the present research was prospective or predictive modelling of census response, so that explanatory variables were formed using neighborhood-based aggregated information, together with items like housing-type or place-size which would be known before enumerating the household in the housing unit. The use of demographic variables to summarize neighborhood response to the census is related to the 'Targeting Database' research of Gregg Robinson and others, but differs in that their 'Hard-to-Count Scores' reflect only the rank of census tracts via mail-response rates in being difficult to enumerate, while the objective here is to see quantitatively whether the pattern of individual household response-probabilities varies systematically over enumerator checkin-times with respect to explanatory variables. All aspects of the present research are elaborated in greater detail in Slud (1999).

2. Data Sources, Variables & Preprocessing

The unit of study is the (non-group-quarters) housing unit (HU) listed under the 1990 **CENSAS** *100% Edited Detail File* database restricted to mailout-mailback areas (*type of enumeration area* 1, 2, or 4). 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*), **block-group**. In addition, *housing type*(**htyp**) for each HU was coded into Mobile-home = 0, Single-family home = 1, or Other = 2 (primarily Apartments). All 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. In addition, two re-coded indicator variables have been used in analyses: **Sing** indicating whether more than two-thirds of the HU's in a block-group are single-family homes; and **I(plsz=0)** indicating that the *plsz* code is 0. Mail-response rate (for block-group by *htyp* strata) is denoted **ymr**.

Beyond the demographic variables, data on check-in dates for enumerator-completed forms, were obtained from the *Operational Files* (**CDOP**), linked to the **CENSAS** data through unique HU identifiers. Checkin-time for each HU was re-coded by subtracting the minimum checkin-time for the ARA. Then the HU's were tallied as falling into one of the mutually exclusive response-categories: **MR** for mail-response, **rsp50** for form filled in by an enumerator and checked in before the median of checkin times for the same ARA, **rsp75** for HU's checked in between the median and upper quartile of checkin times for the ARA, **rsp90** for HU forms checked in between the 75'th and 90'th percentiles for the ARA, and **LT** for all other HU's. We adopt the convention that regardless of form-type or checkin-date, if a HU is recorded as having *all person-items imputed* it is treated as a non-responder, and is not included in the pool of HU's available to respond at later checkin times.

Data were preprocessed so that (a) all block-group covariate proportions p were re-coded into *logit* scores $\log(p/(1-p))$; (b) all block-groups which did not contain at least 50 HU's were discarded; and (d) all block-group-by-htyp strata which did not contain at least 21 HU's were discarded.

3. Preliminary Analyses

There turns out to exist a weak but significant relationship between minimum checkin-time for an ARA and demographic variables. (Correlations between fitted and observed minimum checkin-time, in terms of averages over ARA of the block-group aggregated fractions *fb, fspou*, etc., were e.g. 0.65 for DE, 0.38 for NC, and 0.42 for NY.) Variables and interactions in the models were selected by step-down from the third-order model for DE data, and models with the same selected terms were then re-fitted for the other states. The terms which proved most significant in the 30-variable linear models, which give considerably better fits than the 11-variable models, are:

```
fspou, Sing, ymr, fhisp, fb, fown, focc, [htyp=2], fb:fown, fb:fnp7,
funr:fown, fb:[htyp=2], fhisp:fnp7, funr:fb, ymr:fown, ymr:fspou
```

We are interested in statewise relationships between numbers of HU's uncounted at various stages of the census. Table 1 of Slud (1999) gives the statewise fractions of HU's for which either no form was received or all person-items were imputed, out of all HU's on the **CENSAS** list falling in retained block-group-by-htyp strata. These ratios range from 0.00845 in IO, 0.00942 in NEB, 0.00989 in IDA, 0.0132 in MN, to 0.0359 in GA, 0.0373 in RI, 0.0417 in MI, 0.0453 in FL, and 0.0557 in DC. Figures 5 and 6 of Slud (1999) (not reproduced here) show both the considerable difference in statewise rates of responding to the census at times after 30 and 60 days of followup within ARA, and also that these differences show regional patterns. In particular the Midwestern states, like most of the southern and western states, show low rates of late response, while New England and middle Atlantic states (especially DE) show surprisingly high rates. Indeed, these figures make DE appear a notable outlier because of its high rate of response after 30 days.

4. Models, Methods, & Results

Logistic Regression models were used to fit the census-response data: the response-indicators y_{ij} for the j'th HU among the n_i within the i'th block-group-by-*htyp* stratum are assumed to be independent binomial random variables, each with the same heads-probability π_i such that the log-odds or *logit* score is a linear combination of the entries of an explanatory row-vector x_i for the stratum. With Y_i as the aggregated response-count for stratum i, the model is

$$Y_i \;\equiv\; \sum_{j=1}^{n_i} y_{ij} \;\sim\; Binom(n_i,\, \pi_i) \quad , \qquad \log\,(\frac{\pi_i}{1-\pi_i}) \;=\; x_i\,\beta \qquad (1)$$

The unknown regression-coefficient vector β is fitted by maximum likelihood for each dataset, separately for each state and response variable. These models are *longitudinal* in the sense that the observed response rates at earlier stages enter explicitly as covariates and within interaction terms at later stages. Thus, when analyzing HU responses **rsp75** to census enumerators betweeen the 50'th and 75'th percentiles of checkin times within the HU's ARA, the cell-count n_i denotes the number of HU's 'at risk' of such a response, which means the number of HU's in the stratum which have failed to respond by mail or before the median checkin time, and Y_i is the observed tally of responses before the upper quartile of ARA checkin times. Among the explanatory covariates for analyzing this response would be the observed rates of mail-response and before-median response for HU's in the same stratum, as well as the interactions between these rates and the demographic and *htyp* covariates.

A set of 52 explanatory variables (including interactions) was selected for DE Mail-Response rates; then after further variable-selection from these variables and interactions with earlier-response rates, respectively 51-variable, 30-variable, and a 49-variable regressor sets were chosen for the DE responses **rsp50**, **rsp75**, and **rsp90**. For full details of the variable-selection, see Slud (1998, 1999). Logistic regression models were then fitted, using the DE-selected regressors for each of the response variables **MR**, **rsp50**, **rsp75**, and **rsp90**, to the household-response data, for all 50 states plus DC.

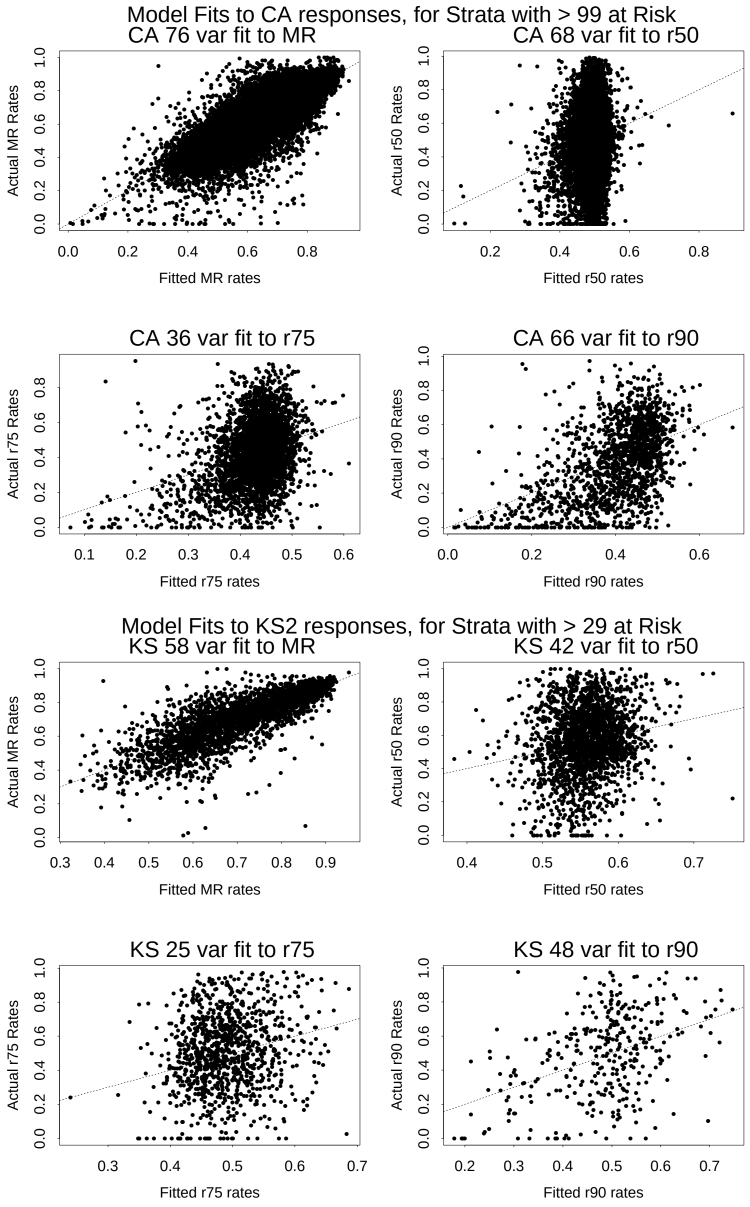

Figure 1: Plots of observed versus predicted response rates within CA (4 upper plots) and mailout-mailback areas of KS (lower plots) for block-group-by-*htyp* strata within different logistic regression models fitted to each state's data, respectively using **MR**, **rsp50**, **rsp75**, and **rsp90** as response indicator variables. In each case the model was fitted to the full dataset, but points are plotted only for strata with at least 100 HU's for CA (and 30 for KS). Correlations between fitted and observed rates for the four response variables are respectively (0.81, 0.12, 0.21, 0.43) for CA and (0.83, 0.19, 0.22, 0.53) for KS.

Figure 1 exhibits the fit of the models for the four response variables on the CA and KS data, and shows that, while **MR** is strongly correlated with state-by-state predicted values, **rsp50** and **rsp75** are not, while the association between **rsp90** and block-group-by-*htyp* fitted values is strong enough to be interesting but not highly predictive. The correlations between fitted and observed responses for all states and the four response variables show the general pattern, repeated in virtually all states, of very high (0.8 to 0.9) values for **MR**, rather low values (0.15 to 0.3) for **rsp50** and **rsp75**, and intermediate correlations (sometimes as low as 0.3 or 0.35, but mostly in the range 0.45 to 0.60). for **rsp90**. The progression from low to higher correlations as one moves from **rsp50** to **rsp75** to **rsp90** suggests that a high rate of late-responding HU's to enumerator followup is usually associated, but not with sufficient strength to allow high predictability at the block-group level, with block-group-by-*htyp* demographic and geographic covariates.

The variables *fspou, fown, fb* and their second and third order interactions were good explanatory variables in the statewise **MR** logistic regression models. The interaction of the indicator of *(plsz=0)* with *htyp* was also an important variable. The variable *Sing* showed important interactions with *plcod, htyp, fown, focc, plcod*, and *fb*. Indeed, apart from the variable *fnp7* measuring the prevalence of units with 7 or more occupants, *Sing* and *plcod* were the only variables showing significant interaction with the variable *fb* measuring the fraction of black block-group residents. The variable *fnp7* measuring the prevalence of large households within a block-group, while not particularly important in the **MR** or **rsp50** analyses, *is* important in the models for **rsp75** and **rsp90**. This remark gibes well with the observation of Gbur (1996) on the differing size of interviewed and noninterviewed HU's.

4.1. Further Cross-Tabulations

For further analysis of the data-files described above, the geographic and demographic variables were replaced by a much coarser partition into 48 strata defined, within each state, by the cross-classification of the variables **htyp** (3 levels) by **Sing** (2 levels) by three more binary variables: the indicators FB that **fb** (before *logit* transform) was at least 0.5; FOWN that the homeowning fraction for a block-group was at least 0.6; and FSPOU that **fspou** for a block-group (before *logit* transform) was at least 0.75. Only the 39 strata which have 5000 or more HU's nationally are used below.

The ratios of observed over predicted response-rates were first examined, for each of the four response variables **MR**, **rsp50**, **rsp75**, **rsp90**, each state, and each of the 39 strata described above. There were no clear regional patterns, so to a first approximation the logistic regression models do capture the observed state-by-covariate differences in response rates.

Next studied was an array $(51 \times 48 \times 18)$ of response counts cross-classified by state, stratum, and time (i.e, response by mail, within successive intervals of 10 days starting at the minimum checkin-time for the ARA, or no response at all). Table 1 displays, by stratum, the nationally aggregated rates of mail-response, overall non-response to the census (in any form, as a fraction of non-mail-responding HU's), as well as the fractions of non-mail-responding HU's which responded respectively later than 30 or 60 days into enumerator followup in the HU's ARA. Only the 39 strata containing at least 5000 HU's nationally were included in the tabulation, and these are listed together with their defining covariate-values. Across all of the (39 significant) covariate categories, the rate of response to enumerators by the thirtieth day following initiation of enumerator followup within ARA is in the range (0.60, 0.82), with the rate of response by the sixtieth day ranging from 0.82 to 0.93, where the ultimate rate of enumerator-followup response among all HU's is seen to range from 0.84 to 0.94. Visual inspection of the Table confirms what was found in the model-fitting, that rates of late responses to enumerators are not directly related to the rates of mail-response.

Table 1 shows the definite but weak positive association across (nationally aggregated) strata between rates of late response and the rate of final *non*response to the census. An elaboration of the same relationship, to reflect differences among selected large states for Late-response Rates over the 39 nationally aggregated covariate-defined strata which contain more than 5000 HU's, is given in Figures 16 and 17 of Slud (1999) (not reproduced here). The stratum-specific rates of followup response later than 30 days seem to be reasonably strongly related to final nonenumeration rates in NY, IL, WASH, positively but less strongly related in TX, NJ, and MICH, and unconvincingly related in CA and NC.

Finally, in Figure 2 we show the fractions of non-mail-responding HU's which have not responded to followup-enumerators, as a function of ARA followup-time and demographic stratum. From this Figure, we can see directly that Strata such as 3, 4, 6, 25, 27, 30, 37, 39, and 40 show a consistently strong pattern of response to enumerator followup, while Strata such as 2, 8, 14, 17, 20, 23, and 47 show consistently weak response. Thus, faster response to enumerators is seen to be positively associated with the combination FSPOU=0, HTYP=0 or 2, and mostly FB=1 (and with FOWN=0 in strata with FB=0). By contrast, slow response to enumerators occurs in strata with HTYP=1 and either Sing=0 or the combination of Sing=1 and FOWN=1.

Table 1. Mail-response (MR), Late-Response and Followup Nonresponse (NR) Fractions, by Stratum. (Cell-counts given in rounded units of 10,000.)

#	FB	OWN	SPOU	Sing	Htyp	Cellct	MR	30+	60+	NR
1	0	0	0	0	0	487	0.58	0.22	0.090	0.080
2	0	0	0	0	1	776	0.75	0.26	0.088	0.079
3	0	0	0	0	2	1657	0.56	0.32	0.098	0.076
4	0	0	0	1	0	3	0.48	0.28	0.115	0.091
5	0	0	0	1	1	222	0.68	0.22	0.070	0.063
6	0	0	0	1	2	65	0.55	0.22	0.070	0.063
7	0	0	1	0	0	3	0.61	0.29	0.165	0.150
8	0	0	1	0	1	53	0.79	0.23	0.097	0.086
9	0	0	1	0	2	91	0.68	0.28	0.117	0.098
10	0	0	1	1	0	1	0.52	0.31	0.180	0.157
11	0	0	1	1	1	23	0.69	0.19	0.092	0.082
12	0	0	1	1	2	5	0.67	0.18	0.092	0.080
13	0	1	0	0	0	107	0.61	0.23	0.100	0.087
14	0	1	0	0	1	259	0.76	0.26	0.108	0.096
15	0	1	0	0	2	153	0.59	0.28	0.102	0.086
16	0	1	0	1	0	28	0.55	0.25	0.106	0.087
17	0	1	0	1	1	1012	0.76	0.22	0.084	0.076
18	0	1	0	1	2	162	0.59	0.24	0.087	0.076
19	0	1	1	0	0	111	0.60	0.25	0.134	0.117
20	0	1	1	0	1	183	0.73	0.26	0.136	0.119
21	0	1	1	0	2	94	0.61	0.22	0.113	0.101
22	0	1	1	1	0	113	0.59	0.22	0.106	0.088
23	0	1	1	1	1	2449	0.80	0.22	0.110	0.098
24	0	1	1	1	2	156	0.63	0.23	0.104	0.090
25	1	0	0	0	0	3	0.41	0.25	0.101	0.091
26	1	0	0	0	1	119	0.58	0.32	0.098	0.078
27	1	0	0	0	2	278	0.44	0.39	0.125	0.083
28	1	0	0	1	0	1	0.44	0.22	0.089	0.071
29	1	0	0	1	1	79	0.55	0.23	0.071	0.062
30	1	0	0	1	2	20	0.45	0.28	0.092	0.076
32	1	0	1	0	1	1	0.63	0.34	0.109	0.086
33	1	0	1	0	2	3	0.66	0.36	0.177	0.109
37	1	1	0	0	0	5	0.46	0.34	0.119	0.094
38	1	1	0	0	1	16	0.61	0.37	0.112	0.087
39	1	1	0	0	2	7	0.46	0.40	0.121	0.087
40	1	1	0	1	0	4	0.47	0.32	0.122	0.097
41	1	1	0	1	1	179	0.63	0.28	0.083	0.072
42	1	1	0	1	2	18	0.46	0.35	0.120	0.095
47	1	1	1	1	1	9	0.75	0.32	0.103	0.092

These figures and tables exhibit considerable spread in late-response rates across strata, and show further that there is generally a weak but definite positive relationship — whether nationally or by state — between large late-response rates at block-group level and above-normal proportions of nonenumerated HU's .

5. Conclusions

Two major findings of this research were:

(I) In all states, Mail-Response Rates to the decennial census are highly predictable using 50 to 76 covariates and interactions (with correlations of order 0.8 between fitted and observed rates); and rates of late response to enumerators, after 90'th percentile of forms within the same ARA have been checked in, among all housing units who have not been enumerated by the 75'th percentile of checkin times, are moderately predictable (with correlations of order 0.5 between fitted and observed rates); but intermediate-time response-rates to enumerators are highly unpredictable.

(II) Mail Response rates alone do not predict late-response rates, but are significant predictors in conjunction with other variables and interactions. Late-response rates, in the sense of checkin times being more than 30 or 60 days after the earliest in the same ARA, vary considerably across covariate-defined strata, and are moderately (0.4 to 0.5) correlated with overall fractions of unenumerated housing units, across 39 national strata.

The models described here for block-group rates of late response to followup enumerators, in terms of explanatory variables at block-group level, could be effective management tools to aid in targeting those block-groups in which higher yields from late enumerator followup might be expected. Moreover, targeting for elevated late-response differs from targeting for high mail-response, which was the objective of the Targeting Database of Robinson & Kobilarcik (1995). The results obtained also suggest that carefully collected checkin-time data in the future — designed to reflect the level of effort expended in enumerating individual households — might be extremely useful in constructing more precise models.

References

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

Census Data Organization Project (CDOP), Operational File documentation, 1990 Decennial Census, Ellen Katzoff & George McLaughlin, March 24, 1994, Census Bureau.

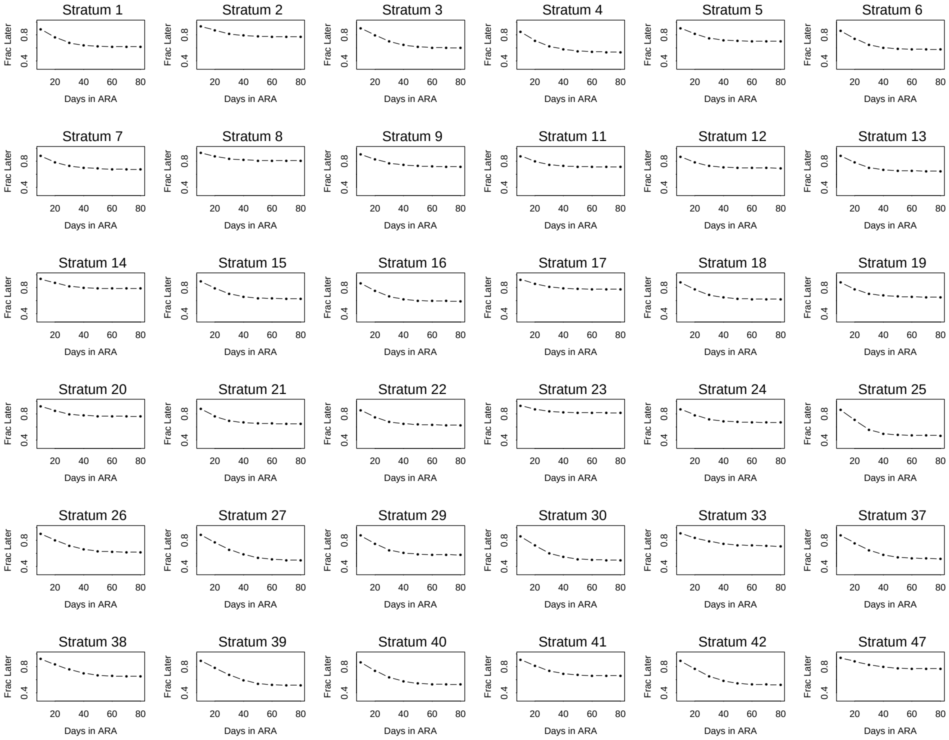

Figure 2: Series of survival curves showing, by the 36 nationally-aggregated demographic strata with at least 20000 HU's, the fraction of non-mail-responding HU's which have not yet had a form checked in after multiples of 10 days' enumerator-followup within ARA.

Gbur, P. (1996) Integrated Coverage Measurement evaluation project 3: noninterview followup. *1995 Census Test Results*, Memorandum **No. 44**, Census Bureau.

Hogan, H. (1993) The 1990 Post Enumeration Survey: operations and results. *Jour. Amer. Statist. Assoc.* **88**, 1047-60.

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

Slud, E. (1998) Predictive models for decennial census household response. *Proceedings of Amer. Statist. Assoc. on Survey Research Methodology*, 272-7.

Slud, E. (1999) Demographic models of decennial census household response. Census Bureau summary report on 1997/98 ASA/Census Fellowship project.