PREDICTIVE MODELS FOR DECENNIAL CENSUS HOUSEHOLD RESPONSE

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

Key words: block-group aggregated covariates, enumerator checkin times, logistic regression, longitudinal data, mail response, variable selection.

Abstract. Data-preparation and fitting for a comprehensive model of statewise household response to the 1990 census is described, using a methodology of successive logistic regressions for longitudinally defined response variables, including indicators of response by mail, and enumerator checkin within quantile intervals of enumerator operational time for the ARA containing the household. The explanatory variables consist of geographic and housing-type data aggregated over census block-groups. Results of the data analysis are given for Delaware and North Carolina. Models are validated by refitting models including random effects, and by applying models with variables selected from DE to data for NC. Indicators of response by mail show a much stronger relationship than the checkin-time responses with the explanatory variables, and the indicator of late checkin-times (between the 75^{th} and 90^{th} percentiles) appear slightly more predictable than the earlier checkin-time indicators.

This paper reports on research and analysis undertaken by the author, and is released to inform interested parties and 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 response to or cooperation with 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, Alho et al. 1993). Third, and more speculatively, detection of possible systematic demographic patterns in households or persons enumerated later rather than earlier might supply a basis on which to estimate characteristics of the unenumerated population, serving as a check on models for undercount adjustment. Two simple instances of this sort of pattern are the often-asserted relationship between undercount and mail nonresponse in post-strata (Hogan 1993) and the observation (documented by Gbur 1996 from a followup study on the 1995 Oakland Test Census) that household size is smaller among households supplying later enumerator-completed returns.

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. Thus, Alho et al. (1993) modelled the response variables of enumeration in the E-sample (regular census) but not P-sample (post-enumeration sample), in the P-sample but not E-sample, or in both the E- and P- samples, in terms of demographic (short-form) characteristics; Word (1997) and many studies he cites treated response via mailed-in census forms, in terms of short-form characteristics; Krenzke (1997) modelled the status of being enumerated within the last 10% of census forms (either within census tract, or within the nation); and Causey (1998) modelled mail-response for the subpopulation of enumerated 1990 long forms. These individual-response models all concerned response characteristics for *enumerated* persons or households and were in that sense probability models for responses *conditional upon having been enumerated*. Other models of E- and P- sample response, such as those of Ericksen & Kadane (1985) and Isaki, Huang & Tsay (1991) for 1980 and 1990 undercount adjustment, differed in modelling aggregated response rates at the post-stratum level. Other Census Bureau literature, some appearing in Survey Methods ASA Proceedings — analyzes census omissions and errors by separate but not cross-classified demographic characteristics and form types.

In the present research, the unit of study is the (non-group-quarters) housing unit (HU) in the 1990

CENSAS *100% Edited Detail File* database. This database represents a final version of the list of HU's, including vacant units, after application of rules eliminating duplicates and inappropriate addresses. The only information about a HU available without a census form is the *housing type* (**htyp**) which we code into Mobile-home = 0, Single-family = 1, or Apartment = 2. Other explanatory variables are either geographic (**plcod** for reservation vs. rural vs. small- or large-urban, and a numerical place-size code **plsz** from 0 to 19), or are aggregated over Census *block-groups* from all census short-forms collected from HU's. As in Robinson & Kobilarcik (1995), the motivation for aggregating demographic variables over block-groups or tracts is to summarize neighborhood characteristics, but unlike those authors who were interested in ranking Census tracts in being 'Hard-to-Count', we restrict to short-form variables. Thus we have aggregated:

fspou: the fraction of enumerated HU's which are **spousal** (following Word 1997), i.e., contain the spouse of the head-of-household or a head-of-household aged at least 50;
fown: the fraction of enumerated HU's units which own (rather than rent) their unit;
focc: the fraction of HU's ascertained *occupied*;
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 containing any person unrelated to the head-of-household;
fhisp: the fraction of enumerated HU's with *Hispanic* head-of-household.

Beyond the demographic variables, the analyses below use data on check-in dates for enumerator-completed forms, which are available from the so-called *CDOP* or *Operational Files* (Katzoff & McLaughlin 1994), and which have been linked to the **CENSAS** data through unique HU identifiers.

An important issue in the longitudinal analysis of census response was the choice of a time-to-response variable. Census Bureau operational literature, and the discussion of checkin times by Krenzke (1997), indicate that enumerators in 1990 were assigned ARA's (*Address Register Areas*, in size between Census block-group, with an average of 400 HU's, and tract, with an average of 1500) to cover at somewhat haphazard times, not obviously related to demographics. The bulk of local enumerations within the ARA were then checked in over several weeks; but visits to the ARA for Last-Resort and Close-out enumerations could take place much later. For mailout areas within North Carolina, the earliest checkin times for ARA's ranged from the 127'th to 181'th day of 1990, and the latest checkin day minus the earliest, by ARA, ranged from 0 to 127, with 49-90 as interquartile range. Examination of the data showed that the earliest checkin day, by ARA, is virtually unrelated to the mail-response rates for block-group-by-*htyp* strata within the ARA.

The data have been preprocessed as follows for all of the analyses reported here: (o) the only block-groups included are within so-called *mailout areas* where forms are delivered and can be mailed back over some time before an enumerator visits, (i) the demographic block-group proportion-covariates p have all been re-coded into *logit* scores $\log(p/(1-p))$; (ii) all block-groups which did not contain at least 50 HU's have been discarded; (iii) all *block-group-by-htyp* strata which did not contain at least 21 housing units have been discarded; (iv) the HU's in all other *block-group-by-htyp* strata are 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 enumerator checkin times for the same ARA, **rsp75** for enumerator-completed form checked in between the median and upper quartile of checkin times for the ARA, **rsp90** for form checkin 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 **LT** (non-responding).

2. STATISTICAL METHODS

The statistical models used to fit the census-response data are Logistic Regression models, in which response-indicators y_{ij} for the j^{th} HU 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 has the linear-regression form

$$logit(\pi_i) \;=\; \log\left(\frac{\pi_i}{1-\pi_i}\right) \;=\; x_i\beta$$

Here x_i is a row-vector constructed from the explanatory variables, including recoded variables and interaction-terms, for the i^{th} stratum, and β is an unknown column vector of regression coefficients which is fitted by maximum likelihood for each dataset (one analysis for each state and response-variable) under study. The strata have been defined

in such a way that the explanatory variables x_i are the same for all n_i HU's within stratum i, so the model can be more simply expressed for the i^{th} stratum response-count Y_i in the form

$$Y_i \equiv \sum_{j=1}^{n_i} y_{ij} \sim Binom\Big(n_i, \frac{e^{x_i\beta}}{1+e^{x_i\beta}}\Big) \quad (1)$$

The data tallied in preprocessing step (iv) can also be regarded as *longitudinal data*, with each HU providing a categorical variable with ordered levels **MR**, **rsp50**, **rsp75**, **rsp90**, and **LT** indicating whether and at what stage (i.e., mail-in or quantile interval of enumerator checkin times) response occurred. While there are generalized-linear and other models which can be fitted to such data (Diggle et al. 1994), we instead analyzed the responses at each stage separately, without pre-judging whether the same explanatory variables should enter, possibly with the same coefficients, at different stages. Moreover, we explicitly allowed the observed response rates at earlier stages to enter 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, in other words, 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 mail-response rates and rates of before-median response to enumerators for individuals in the same stratum, as well as their interactions with the demographic and *htyp* covariates.

We begin by examining in Figure 1 the scatterplots of **MR** versus the other response variables for Delaware, showing no immediate relationship. All analyses and figures were done with version 3.4 of **Splus**.

Variable-selection within each of the logistic regression models (1) was based upon likelihood ratio tests and analysis of deviance tables, and scrutiny of plots versus covariates and predictors of deviance residuals and differences between actual response rates and corresponding fitted rates. Details of the mail-response modeling for Delaware, and of more detailed logistic modelling of Krenzke's (1997) data, can be found in Slud (1998).

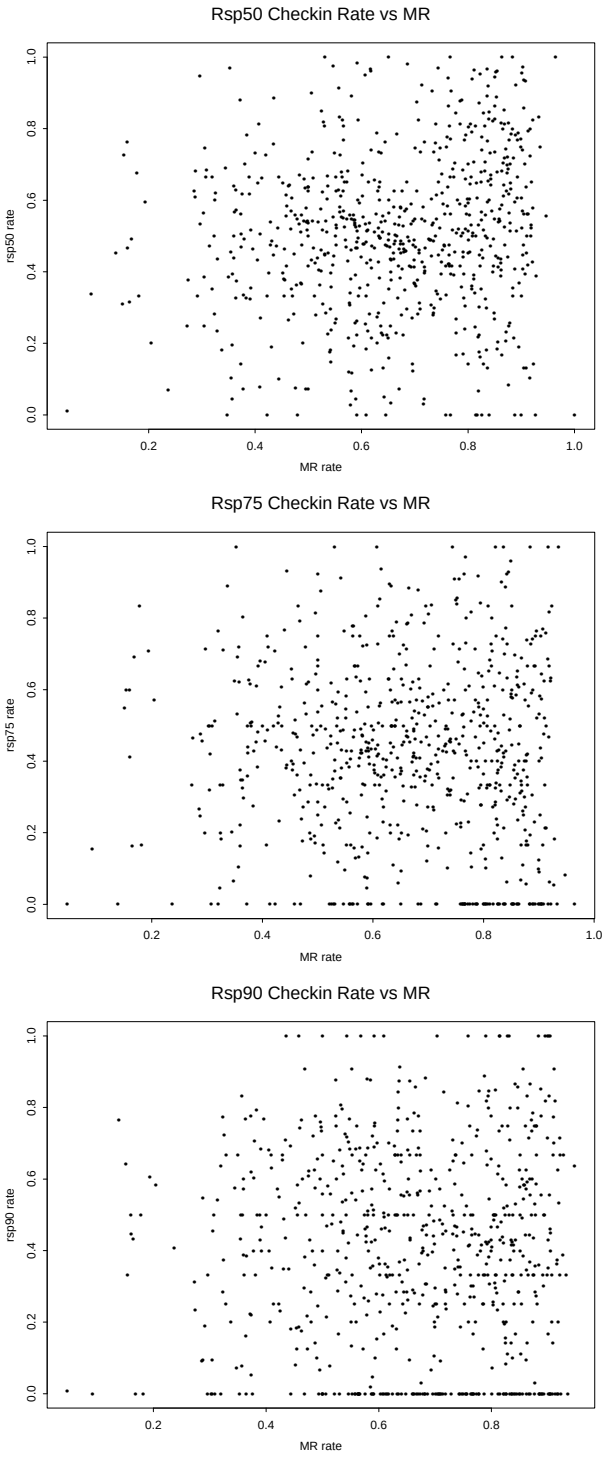

Figure 1: Scatterplots of rates of response **rsp50**, **rsp75**, **rsp90** versus **MR** for DE block-group-by-*htyp* strata.

3. SUMMARY OF RESULTS

In presenting results, we restrict attention to the household-response data from the states of Delaware (DE) and North Carolina (NC), and in the latter we restrict to mailout-mailback areas (*TEA* numbers 1, 2, or 4). In Delaware, the preprocessing steps (o), (ii) and (iii) resulted in deletion of 7 block-groups, with a total of 149 housing-units, and of small *htyp*-by-block-group strata totalling an additional 1,359 housing-units. There were 727 retained block-group-by-*htyp* strata in Delaware (with a total of 219,509 housing-units), and 8388 strata in the mailout-mailback areas of NC.

Selection of explanatory variables, especially the two- and three- way interaction terms, was done using the DE data. A 52-variable logistic regression model for Delaware Mail-Response was chosen after a careful examination of all pairwise (and many three-way) interactions which seemed to generalize to data for other states. To fit the model with variables **rsp50**, **rsp75**, and **rsp90** as response, two backwards-selection steps were followed: first, the same 52 variables from the initial model, applied with each new response variable, were backwards-selected in **Splus** by stepping out variables which did not individually decrease deviance by at least 12. This resulted respectively in sets of 36, 16, and 30 explanatory variables. These sets were then augmented by earlier response-variables as follows: the 36 variables for **rsp50** were augmented by the 36 interactions of those variables with the mail-response rate **MR** ; the 16 variables for **rsp75** were augmented by the 16 interactions with **MR**, the 16 with **rsp50**, and the **MR:rsp50** interaction; and the 30 variables for **rsp90** were augmented by 30 interactions with each of **MR**, **rsp50**, **rsp75** along with the pairwise interactions **MR:rsp50**, **MR:rsp75**, **rsp50:rsp75**. In each of the datasets for **rsp50**, **rsp75** and **rsp90** responses, backward selection was again applied, respectively yielding a 51-variable logistic regression model for **rsp50**, a 30-variable model for **rsp75**, and a 49-variable model for **rsp90**. Finally, the corresponding model for each response in NC was fitted using the variables selected for DE.

Backward selection for large sets of explanatory variables can result in over-fitting, and this may well have happened in the fitted DE models displayed in Figure 2. This criticism does not apply when the same explanatory variables were used in fitting to NC data, although it was expected that the predictive model relationships for NC would be weaker. We did the stepdown analysis here because (a) the source of the 52 variables was the **MR** fit, (b) it was not known which of the previous variables would show interesting interactions with earlier-stage response variables, and (c) the resulting models were to be applied to data from other states.

3.1. Explanatory Variables

In the course of fitting the model to DE Mail-Response, powers up to the third of the variables found by Word (1997) to be most important (*fspou, fown, fb*) also turned out here to be good explanatory variables. The *plsz* code turned out to be 0 quite often, and the interaction of the indicator of *(plsz=0)* with *htyp* was an important variable. A further useful re-coded variable was *Bsing*, the indicator that single-family homes accounted for at least two-thirds of the housing units within the block-group. This variable showed significant interactions (in DE & NC) with *plcod, htyp, fown, focc, plcod*, and *fb*. Apart from the variable *fnp7*, *Bsing* and *plcod* were the only variables showing significant interaction with the racial variable *fb*.

The 52-variable logistic regression model for Delaware HU mail-response rates, pictured in the topmost plot of Figure 2, gives generally accurate but far from perfect predictions. (Correlation of the displayed data was 0.89.) Predictions were much better, with many fewer variables, in Word (1997) and Causey (1998), whose explanatory variables described the individual HU's whose responses were tallied, while here the components of x_i other than *htyp* are block-group aggregates. Consistently with the lack of relationship shown in Figure 1 between MR and later responses, the logistic regression models for **rsp50**, **rsp75**, and **rsp90** in Figure 2 show much weaker explanatory power than the **MR** model (with respective correlations 0.43, 0.55, and 0.60), and the fits would be worse still if the data were pictured for block-group-by-*htyp* strata with fewer than 50 HU's remaining to respond.

3.2. Patterns in Fitted Models

A few findings can be summarized from the enumerator-checkin-time models selected from DE data and fitted to the DE data and to strata with at least 50 HU's within the NC mailout areas. See Figure 3 for graphical display of the models fitted to NC data. The correlations between predicted and observed stratumwise response rates in the four plots of Figure 3 are respectively 0.82, 0.18, 0.20, and 0.49.

- The variable *fnp7* measuring the prevalence of large households within a block-group, while

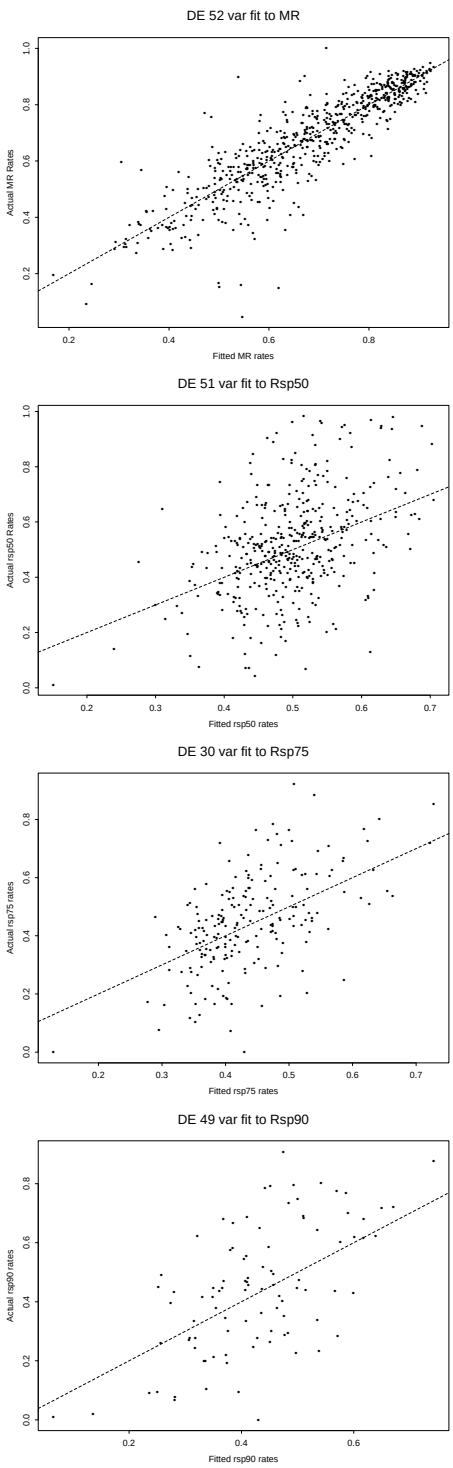

Figure 2: Observed vs. predicted response rates for block-group-by-*htyp* strata within different logistic regression models fitted to DE data, respectively using **MR**, **rsp50**, **rsp75**, **rsp90** as response indicator variables. One point is plotted for each stratum in which at least 50 HU's had not responded at previous stages.

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.

- Both in the DE and the NC fits to enumerator-checkin responses, there is essentially no demographic predictability of **rsp50**, i.e., of HU's responding before the median checkin time.

- Demographic relationships with **rsp90** are stronger than those with **rsp75**, in both DE and NC. In these relationships, terms involving the rates of earlier responses within the same strata play a prominent role, appearing in interactions with all of the major demographic variables.

The pattern of observed correlations seen in NC mailout data between block-group-by-*htyp* stratum-wise fitted and observed response rates has now been confirmed in about 25 additional states: the correlation is roughly 0.80 for **MR**, usually less than 0.2 for **rsp50**, somewhat higher (from 0.15 up to about 0.35) for **rsp75**, and in most cases substantially higher (between 0.4, 0.5) for **rsp90**.

3.3. Model Adequacy

The primary vehicle described so far for assessing the reality of demographic influence on the various types of census response has been the predictive value of models on one state (NC) using variables and model types selected from data on another state (DE). This is a visual assessment of model adequacy (compare Figures 2 and 3) which could also be made formal; a key point of Figure 3 is to confirm that the response-versus-predictor relationships in DE were not artifacts of extensive variable-selection.

Slud (1998) has described the fitting and interpretation of logistic regression models like (1) incorporating random stratum effects in the *logit* score, as a way of assessing model adequacy for logistic regressions with large stratum-counts and potentially large sets of predictor variables. The **MR** data-analysis for DE was used as a case study in that paper, and a similar analysis has been done on NC mailout-area data. Although the logistic regression models fitted to **MR** for DE and NC are not fully adequate in the sense of being able to pass likelihood-ratio and omnibus goodness-of-fit tests, the methods of Slud (1998) indicate that the lack of fit is more plausibly due to unmodelled random variation than to an insufficient number of interaction-terms among predictors.

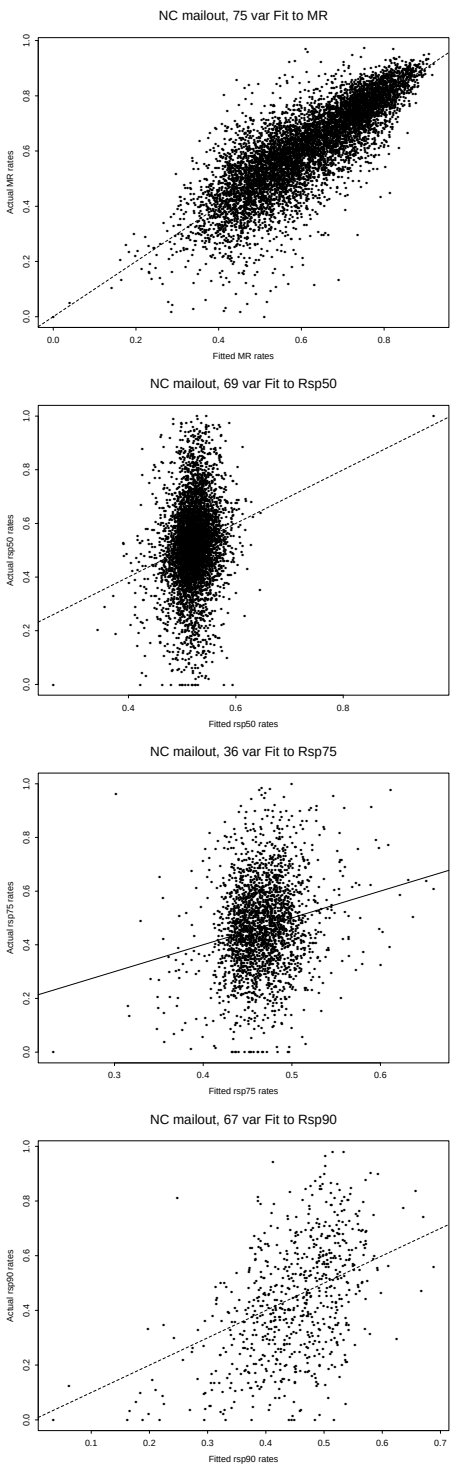

Figure 3: Observed vs. predicted response rates for block-group-by-*htyp* strata within different logistic regression models fitted to NC mailout-area data, respectively using **MR**, **rsp50**, **rsp75**, **rsp90** as response indicator variables. One point is plotted for each stratum in which at least 50 HU's had not responded at previous stages.

Research on these questions is continuing. The general findings of this modelling approach have now been confirmed on data from many other states. Moreover, inclusion of additional explanatory variables (for DE, NC) from among the long-form items aggregated at tract level causes almost no change in the fitted models or results. One remaining task is to study, through consideration of absolute enumerator checkin times, the effects of Last Resorts and Closeouts on response counts in the **rsp75** and **rsp90** models.

4. REFERENCES

Alho, J., Mulry, M. Wurdeman, K. & Kim, J. (1993) Estimating heterogeneity in the probabilities of enumeration for dual-system estimation *Jour. Amer. Statist. Assoc.* **88**, 1130-6.

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

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

Diggle, P., Liang, K.-Y., & Zeger, S. (1994) **Analysis of Longitudinal Data**. Oxford: Clarendon.

Ericksen, E. & Kadane, J. (1985) Estimating the population in a census year: 1980 and beyond. *Jour. Amer. Statist. Assoc.* **80**, 98-1131.

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.

Isaki, C., Huang, E. & Tsay, J. (1991) Smoothing adjustment factors from the 1990 Post Enumeration Survey. *Proc. Soc. Statist. Sec., Amer. Statist. Assoc.*, 338-43.

Krenzke, T. (1997) Profile of the Last 10% of Returns in the 1990 Census. Decennial Statist. Studies Div. Memorandum, Jan. 15, 1997, Census Bureau.

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 Association of America.

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

Word, D. (1997) Who Responds ? Who Doesn't ? *Census Bureau Population Div. Working Paper 19.*