Model-assisted Weighting for Survey Data with Multiple Response Modes

Eric Slud, Census Bureau & University of Maryland

May 28, 2002

Abstract

Several large surveys administered by the Census allow responses in several different modes (mail, telephone, personal interview), in some cases — as in the American Community Survey (ACS) — after sub-sampling at some stages from that part of the sample which has not responded at any earlier stage. Analysis of the survey results usually proceeds by weighting responses roughly by inverse unconditional probability of response to the survey. But often, nonresponse in the different modes can be modelled meaningfully in terms of demographic and geographic variables (such as dwelling type and aggregated characteristics related to ethnicity and socioeconomic status), and the results of such models derived from previous recent surveys could conceivably be used to improve the estimation of population totals and domain subtotals which is ordinarily the goal of large surveys. This paper applies the concepts of sample-survey theory to investigate the theoretical improvements possible assisted with both correctly and incorrectly specified models, and illustrates the issues and improvements using data from the 1990 decennial census. In the setting of models no worse than those fitted to one state's decennial-census data and applied to another similar and neighboring state, the mean-squared error is definitely improved by incorporating the model into sample-weighted estimators, through an adjustment factor which is constant across PSU's.

1 Introduction and Problem Statement

Suppose that individuals in a sampling frame U are selected for inclusion in a large survey S with single and joint inclusion probabilities respectively $\pi_i,\ \pi_{ij}$, and that each individual can respond to the survey in a succession of K possible modes, with mode 1 being the most direct, such as mail-response in the decennial census or American Community Survey. (In the ACS, the modes are: Mail-, Telephone-, and Personal-Interview response. In the decennial census, the modes are Mail- and Interviewer- response, but the census data can be analyzed further, as in Slud (1998, 1999, 2001) by treating as different modes the responses to interviewers within successive quantile-intervals of interviewer followup time within ARA.) Of those individuals selected for initial inclusion and not responding in any of modes $1, \ldots, k$, where $k = 1, \ldots, K-1$, a fraction a_k are randomly sub-sampled for attempted enumeration under mode $k+1$. For each individual $i \in U$, denote by J_{ik} the indicator that individual i would, if selected and followed up, respond to the survey in mode k and no earlier mode. Then $\sum_{k=1}^{K} J_{ik} = J_i$ is the indicator that individual i responds in any mode to the survey.

Suppose also that each individual in the sampling frame comes equipped with a vector $\mathbf{X}_i$ of predictor variables for response, which are observable or known in advance of actual enumeration. Such predictors would include geographic area, along with variables such as housing type which would be known from a master address list and aggregates from recent previous surveys of demographic characteristics for the neighborhood (such as census block-group or tract) containing the individual. A list of such predictors based on block-group aggregated census long-form data is provided in Slud (1998). Assume that the conditional probabilities h_{ik} of survey response by individual i in mode k, *given nonresponse in each of modes* $1, \ldots, k-1$, are modelled and have been estimated as parametric functions $h_k(\mathbf{X}_i) = h_k(\mathbf{X}_i,\ \beta^{(k)})$ of the predictor variables, where the estimated parameters (e.g., regression coefficients in generalized linear models) are denoted $\beta^{(k)}$. The particular model implemented below in analysis of 1990 decennial census data, following Slud (1998, 1999, 2001), is the logistic model

$$h_k(\mathbf{x},\ \beta) \ = \ \frac{e^{\mathbf{x}\cdot\beta}}{1 + e^{\mathbf{x}\cdot\beta}}$$

For the present, assume that the parametric model is a fixed-effect model only. The case where individuals share a random effect within the same neighborhood over which the aggregated predictor-components in $\mathbf{X}_i$ are common, will be treated later. Then prospectively, before sampling,

$$p^0_{ik} \;=\; \prod_{j<k} (1-h_{ij}) \;\cdot\; h_{ik}$$

gives the probability with which individual i would respond to the survey in mode k and not earlier, if selected for inclusion and followed up that far. (Here we adopt the convention that $h_{i0}=0$.) Taking into account the probabilities with which nonrespondents at each stage are included in later stages of followup, we obtain the probabilities

$$p_{ik} \;=\; \prod_{j=1}^{k-1} \{(1-h_{ij})\,a_j\} \;\cdot\; h_{ik} \;=\; \left(\prod_{1\le j<k} a_j\right)\; p^0_{ik}$$

with which individuals assumed to be included in the initial sample are sampled up to and respond within the k'th response mode. Note that as a practical matter, models and estimates for the final-stage response probabilities h_{iK} are largely speculative, because they reflect rates of interview-refusal and omission (e.g., because of failure of interviewers to make personal contact or find proxies) concerning which there is no direct data. Only a followup or post-enumeration study could give more than a hypothetical cast to estimates of these probabilities.

The data recorded from the survey will be, for each individual $i \in S$, the response indicator vector $(J_{ik},\; k=1,\ldots,K)$ together with a label A_i for the latest mode under which individual i is selected for followup. In case $J_i = 1$, the label A_i is equal to that mode k for which $J_{ik}=1$, and an attribute value y_i (such as number in household, or total household income) is also recorded.

The survey data are to be used to estimate the frame population total $t = \sum_U y_i$ for the attribute in question, and the estimators to be considered all take the form of a weighted linear combination $\widehat{t}_w \;=\; \sum_S \sum_{k=1}^K w_{ik}\, J_{ik}\, y_i$. The objective of the present research is to investigate what the optimal weights w_{ik} are, from the vantage point of minimum mean-squared error; how much difference it makes to use them by comparison with the weights one would use if the probabilities p_{ik} were completely unknown, and how sensitive the weights are to correct specificaton of the model $h_k(\mathbf{X}_i,\, \beta^{(k)})$.

Remark 1 *A further objective of this effort at modelling and analysis could be to study the problem of optimally designing sub-sampling rates* a_k *and allowing them to depend on location and neighborhood demographics. In the absence of dependence upon geographic or demographic characteristics, this problem has been studied in Elliott, Little & Lewitzky (2000). Further work in this direction, involving dependencies* $a_k(\mathbf{X}_i)$, *is probably most applicable in business market surveys, but could also be useful in some government surveys, unless there are constraints precluding some types of inequalities among inclusion probabilities based on demographics.*

2 Formulas for Variance of Estimation and Optimal Weights

We begin by deriving formulas for the expectation and variance of the statistic $\widehat{t}_w$. First, recalling that $E\, J_{ik} = p_{ik}$ by definition, we find

$$E(\widehat{t}_w) \;=\; E\left(\sum_{i\in U}\sum_k I_{[i\in S]}\, y_i\, w_{ik}\, J_{ik}\right) \;=\; \sum_{i\in U}\sum_k \pi_i\, w_{ik}\, p_{ik}\, y_i$$

Thus the bias in estimating $\sum_U y_i$ by $\widehat{t}_w$ is

$$\sum_{i\in U} y_i \left[\pi_i \sum_k w_{ik}\, p_{ik} \;-\; 1\right]$$

If the quantities p_{ik} were in fact known or accurately estimated, then for the estimator $\widehat{t}_w$ to be approximately unbiased for all possible attribute-values y_i, the weights w_{ik} must evidently satisfy

$$\pi_i \sum_k w_{ik}\, p_{ik} \;=\; 1 \tag{1}$$

To simplify concepts and notations in what follows, define

$$q_i \;\equiv\; 1 - \sum_{k=1}^{K} p_{ik}$$

to be the probability of *non*response to the survey for an included unit i.

Remark 2 *The unbiasedness condition (1) cannot generally be satisfied, even approximately, in model-free fashion. That is, it forces weights* w_{ik} *to be different for different* i *as long as the quantities* $\pi_i(1-q_i)$ *are. Since the* q_i *are not known, getting them approximately right involves models and estimates.*

If we let the indicator variables for initial sample inclusion be ϵ_i (with $\pi_i = E\epsilon_i$ and $\pi_{ij} = E(\epsilon_i\epsilon_j)$), then the design variance is calculated as

$$\sum_{i,j\in U} \sum_k \sum_l w_{ik}\, w_{jl}\, y_i\, y_j \, \mathrm{Cov}(J_{ik}\epsilon_i,\ J_{jl}\epsilon_j) \ =$$

$$\sum_{i,j\in U} \sum_k \sum_l w_{ik}\, w_{jl} y_i\, y_j\, p_{ik} \Big(\pi_i\, \delta_{i,j}\, (\delta_{k,l} - \pi_i\, p_{il}) \ + \ (1-\delta_{i,j})\, p_{jl}\, (\pi_{ij} - \pi_i\, \pi_j) \Big)$$

which, after some further algebra, can be expressed as

$$\begin{aligned} & \sum_U \sum_k \pi_i\, w_{ik}\, p_{ik}\, y_i^2 \left(w_{ik} - \sum_l w_{il} p_{il} \right) \\ + & \sum_{i,j\in U} y_i y_j \left(\sum_k w_{ik} p_{ik} \right) \left(\sum_l w_{jl} p_{jl} \right) (\pi_{ij} - \pi_i \pi_j) \end{aligned} \tag{2}$$

Let us examine formula (2) to understand better its dependence upon the probabilities p_{ik} and weights w_{ik}. Denote by $\widehat{t}_\pi$ the ordinary Horvitz-Thompson estimator (Särndal et al. 1997) for the population total of the attributes y based upon a single response-mode (the usual case) with the same inclusion-probabilities $\pi_i,\ \pi_{ij}$ as above, and recall that the formula for the theoretical variance of $\widehat{t}_\pi$ is given by

$$V(\widehat{t}_\pi) \ = \sum_{i,\,j\in U} y_i\, y_j\, \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

Then, as long as the weights w_{ik} are chosen as in (1) to make the estimator $\widehat{t}_w$ unbiased, formula (2) can be re-written

$$V(\widehat{t}_w) \ - \ V(\widehat{t}_\pi) \ = \ \sum_U \sum_k w_{ik}^2\, y_i^2\, \pi_i\, p_{ik} \ - \ \sum_U \frac{y_i^2}{\pi_i} \tag{3}$$

More generally, if the weights w_{ik} are fixed, and if individual bias-terms are defined as

$$b_i \ = \ \pi_i \sum_{k=1}^K w_{ik}\, p_{ik} \ - \ 1$$

then $V(\widehat{t}_w) - V(\widehat{t}_\pi)$ is given by the right-hand side of (3) plus

$$\sum_U \frac{y_i^2}{\pi_i} - \sum_U \frac{y_i^2}{\pi_i}(1+b_i)^2 + \sum_{i,j \in U} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \{(1+b_i)(1+b_j) - 1\}$$

$$= \sum_{i,j \in U} y_i\, y_j \{b_i\, b_j + b_i + b_j\} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i\, \pi_j} - \frac{\delta_{ij}}{\pi_i} \right) \qquad (3')$$

But the sum of of (3) and (3′) is easily minimized subject to fixed b_i, for each i, by equalizing the weights w_{ik} over all k, leading to

$$w_{ik}^{eq} = \left(\pi_i \sum_{k=1}^{K} p_{ik} \right)^{-1} = \frac{1}{\pi_i\,(1 - q_i)} \qquad (4)$$

A preliminary conclusion of this analysis is that even if the models on which we base weights are misspecified, there is still no benefit in letting the weights vary with response-mode: all responses in a particular demographic stratum should receive the same weight, regardless of response mode. But perhaps there continues to be some benefit in keeping response modes separate from the point of view of generating models for probabilities of nonresponse.

In the special case where the weights (4) are used, we can view inclusion in the sample as consisting both of random selection *and* response, so that $\widehat{t}_w$ becomes the standard Horvitz-Thompson or π-estimator which is unbiased with variance

$$\sum_{i,\, j \in U} y_i\, y_j \left[\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i\, \pi_j} + \delta_{i,j} \frac{q_i}{\pi_i (1 - q_i)} \right] = \sum_{i,j \in U} y_i\, y_j \left(\frac{\widetilde{\pi}_{ij}}{\widetilde{\pi}_i \widetilde{\pi}_j} - 1 \right) \qquad (5)$$

and the modified inclusion probabilities are $\widetilde{\pi}_i = \widetilde{\pi}_{ii} = \pi_i(1 - q_i)$ and $\widetilde{\pi}_{ij} = \pi_{ij}(1 - q_i)(1 - q_j)$, for $i \neq j$.

We develop the variance formula (2) now, together with formulas for bias and MSE, in the setting where the weights $w_{ik} = w_i$ are constant over response-mode k, and can therefore be expressed in the form

$$w_i = [\, \pi_i\,(1 - \tilde{q}_i)\,]^{-1} \qquad (6)$$

By direct substitution, we obtain in this case $1 + b_i = (1 - q_i)/(1 - \tilde{q}_i)$, and

$$\text{Bias}(\hat{t}_w) \;=\; \sum_U \frac{y_i\,(\tilde{q}_i - q_i)}{1 - \tilde{q}_i} \tag{7}$$

and then, using (2), we obtain $\text{Var}(\hat{t}_w) \;=$

$$\sum_U \frac{y_i^2(1-q_i)}{\pi_i\,(1-\tilde{q}_i)^2} - \sum_U \frac{y_i^2(1-q_i)^2}{\pi_i\,(1-\tilde{q}_i)^2} + \sum_{i,j \in U} y_i\, y_j\, \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i\, \pi_j} \left[\frac{(1-q_i)(1-q_j)}{(1-\tilde{q}_i)(1-\tilde{q}_j)} \right]$$

which leads directly to the expression

$$\text{Var}(\hat{t}_w) \;=\; \text{Var}\left(\hat{t}_{\pi,\, y(1-q)/(1-\tilde{q})}\right) + \sum_U \left(y_i \, \frac{1-q_i}{1-\tilde{q}_i} \right)^2 \frac{q_i}{\pi_i\,(1-q_i)} \tag{8}$$

both terms of which can conveniently be expressed in terms of a modified 'attribute'

$$\tilde{y}_i \;=\; y_i \, \frac{1-q_i}{1-\tilde{q}_i}$$

Remark 3 *The approach followed here accommodates both general attributes* y_i *and survey-weights* π_i, *but the specific numerical MSE comparisons below will be made only for response-indicator attributes and simple-random-sampling (SRS) weights* $(\pi_i \equiv \pi)$.

Two special cases of the variance formulas (2) and (8) are of particular interest, in the setting where equal weights (balanced random sampling without replacement) have been applied. Let $\pi_i \equiv \pi$ denote the probability with which each individual i is included in the initial sample; let $s^2_{yU} = \frac{1}{|U|-1} \sum_U (y_i - \overline{y}_U)^2$ denote the frame-population attribute variance; and let $s^2_{\tilde{y}U}$ denote the analogous frame-population variance for $\tilde{y}_i$ in place of y_i. First, consider the case where variations of the probabilities p_{ik} by individual (or demographic group such as neighborhood) are not available. In that case, denote by p^*_k the population-wide probability of responding to the survey in mode k. Then

$$q^* \;\equiv\; 1 - \sum_{k=1}^{K} p^*_k \;=\; 1 \;-\; |U|^{-1} \sum_U \sum_{k \le K} p_{ik}$$

More generally, if response-rate constants $\tilde{q}$ are used in weights $\tilde{w} = 1/(\pi(1-\tilde{q}))$, then the estimator variance (8) with $\tilde{y}_i = (1-q_i)y_i/(1-\tilde{q})$ becomes

$$V(\widehat{t}_{\tilde{w}}) \;=\; \frac{|U|(1-\pi)}{\pi}\, s^2_{\tilde{y}U} \;+\; \frac{1}{(1-\tilde{q})^2\,\pi} \sum_U y_i^2\, q_i\,(1-q_i) \tag{9}$$

For comparison, the variance based upon the stratumwise weights w_i^{eq} defined in (4), based upon (correctly specified) probabilities p_{ik} which vary with demographics, is

$$V(\widehat{t}_{w^{eq}}) \;=\; \frac{|U|(1-\pi)}{\pi}\, s^2_{yU} \;+\; \sum_U \frac{q_i}{(1-q_i)\pi}\, y_i^2 \tag{10}$$

Note that the ratio of these final formulas depends not only upon the initial sampling fraction and the population coefficient of variation of the attribute y, but also on population characteristics relating the attributes to the q_i .

Assuming that the weights (6) will not vary with response mode, the formulas (7) for bias and (8) for variance of $\widehat{t}_w$ together lead to the following formula for Mean-Squared Error:

$$\mathrm{MSE}(\hat{t}_w) \;=\; \left(\sum_U \frac{y_i\,(q_i-\tilde{q}_i)}{1-\tilde{q}_i}\right)^2 \;+\; \mathrm{Var}(\hat{t}_{\pi,\tilde{y}}) \;+\; \sum_U \frac{\tilde{y}_i^2\, q_i}{\pi_i\,(1-q_i)} \tag{11}$$

Remark 4 *In the case of constant weights, where $\tilde{q}_i \equiv \tilde{q}$ does not vary with i, it is not obvious that the overall average response rate q^* or a direct model-based estimate of it are the best constants to use. In fact, since we will find below that bias squared is usually the dominant term in MSE, the best constant to use should be approximately the one which zeroes out the bias. In the case where y_i is a response indicator, this value is easily seen to be*

$$\tilde{q}_{opt} \;=\; \sum_{i\in U} y_i^2 \Big/ \sum_{i\in U} y_i \tag{12}$$

In terms of a modelled set of response probabilities $q_i^{mod} = 1 - \sum_{k\le K} p_{ik}$, the analogous constant is

$$\tilde{q}_{mod} \;=\; \sum_{i\in U} y_i\, q_i \Big/ \sum_{i\in U} y_i \tag{13}$$

If the constant weights $\tilde{w}_{opt} = 1/(\pi(1 - \tilde{q}_{opt}))$ *based on (12) were used, the bias would be* 0 *by definition, and the MSE would be given by (9). Since these constant weights will never be available, we will consider below using* $\tilde{w}_{mod} = 1/(\pi(1 - \tilde{q}_{mod}))$ *based on (13) instead, in which case the variance is again given by (9), to which we must add the bias-squared* $(\sum_U y_i(q_i - \tilde{q}_{mod})/(1 - \tilde{q}_{mod}))^2$ *to get MSE.*

2.1 *Effect of Weighting in a Census Survey: Data Analysis*

We first specialize the comparison between variances (9) and (10) to one of the cases most relevant to the ACS, where the attribute y_i of interest may be regarded as the indicator of a valid household enumeration for the index i on the Master Address File (MAF) if household i were followed up without time-limitation. (Different, but also relevant, choices would be to view y_i as the number of valid census persons within each household, or the household income or indicator of poverty or of participation in a program like Food Stamps, but we do not pursue these possibilities for now.) Then the population total $\sum_U y_i$ should be interpreted as the true count of households, and $q_i = 1 - \sum_k p_{ik}$ is the probability within a sample survey like ACS that a household on the MAF would *not* be enumerated up to and including the final (K'th) response mode. Although undercount and refusal-rates are usually small — of the order of one percent up to a few percent — it may still happen that differences among q_i for addresses i with different demographic and geographical characteristics can differ by sizeable factors. When this is true, there is some hope that model-assisted weighting can improve the accuracy of estimates from the survey.

The data used in the present comparison are the 1990 decennial-census files previously used by Slud (1998, 1999, 2001) in modelling response to the census, by mail or by later 'modes' of personal response to followup enumerators within successive quantile intervals of followup time within ARA. These files include tallies by block-group of the numbers (of HU's) responding by mail, before the 50^{th} percentile followup time within the ARA, between the 50^{th} and 75^{th}, between the 75^{th} and 90^{th}, and after the 90^{th}. In addition, the files contain geographic and demographic information on the block-group, plus the one variable housing-type (*htyp*) which referred to individual HU's (and was known from the address-file, before enumeration).

These predictive variables were used in Slud (1998, 1999, 2001) to fit logistic models for response-rates by state, with numbers of predictors ranging from about 20 to about 70, depending on the size of the state and the mode of response. (The greater numbers of variables arose, from BIC-like penalized-deviance model-fitting, in models in large states either for mail-response of for response before the 90^{th} percentile of enumerator checkin times, among HU's which had not responded by the 75^{th} percentile. The fitted models are used here as the models $h_k(\mathbf{X}_i, \beta)$ specifying conditional response probabilities h_{ik} for individual HU's in the i'th *htyp*-by-block-group stratum within a state to respond in mode k, given that it had not responded earlier. The response-modes chosen for illustration are $k = 1$ for Mail-response, and then for HU's which did *not* respond by mail, $k = 2$ for response before the 50^{th} percentile of ARA checkin time, $k = 3$ for response between the 50^{th} and 75^{th} percentiles of checkin time, and $k = 4$ for response between the 75^{th} and 90^{th} percentiles. For present purposes, we treat the HU's which did not respond by the 90^{th} percentile of checkin times as though they did not respond at all.

To mimic the sub-sampling scenario of this paper, we define a vector of sub-selection probabilities π, a_1, a_2, a_3, for example an inclusion probability of $\pi = 1/2$ followed by sub-sampling rates $a_k = 1/3$ for later stages among HU's which have not responded in mode $\leq k-1$. (These are the inclusion and subsampling rates tenatively projected for areas sampled in the ACS in a given year.) We compare the elements and total of the MSE formula (11) for several choices of sub-selection probabilities, states, and choices of weights based either on a 'correct' model (ie one fitted to the same state for the same census data), no model, or on an incorrect model, such as one based on a neighboring state. Recall that the attribute y_i of primary interest here is the indicator of response (by mode $K = 4$ or earlier). We display for each state, subselection probability vector, set of 'true' probabilities p_{ik}, and choice of weights w_i, the following quantities: *Bias* as given by formula (7), variance-term 1 or *Var1* equal to $\mathrm{Var}(\hat{t}_{\pi,\tilde{y}})$ in (11), variance-term 2 or *Var2* equal to the last summation-term in (11), and the total MSE as given by (11). Note that always $MSE = Bias^2 + Var1 + Var2$.

Consider first the state of Delaware, with $(\pi, a_1, a_2, a_3) = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. There were a total of 219509 HU's on the (final) address-list within the 727 *htyp* by block-group strata containing at least 21 HU's, of which 10662 did not respond by the 90^{th} percentile of their ARA's checkin times. Since the total count $\sum_{i \in U} y_i$ of responders is likely to have estimation error of order 500 to 2000, we can expect total MSE to be in the range 10^5 to $5 \cdot 10^6$.

First assume that the conditional probabilities h_{ik} of response at each stage are as given by the empirical fractions of households who did respond at that stage in that *htyp* by block-group stratum. The results are as follows.

(a) If weights are chosen stratumwise, equally for all response modes $(w_i = w_i^{eq})$ as in (4), then we obtain

Bias	Var1	Var2	MSE
0	10144	144300	154444

(b) If weights are chosen as $w_i = w^* = 1/(\pi(1 - q^*))$ with the actual population value of q^*, the MSE components are

Bias	Var1	Var2	MSE
1261	15831	116041	1721488

(b′) If weights are chosen as $w_i = \tilde{w}_{opt} = 1/(\pi(1 - \tilde{q}_{opt}))$ with $\tilde{q}_{opt}$ as given by (12), the MSE components are

Bias	Var1	Var2	MSE
0	15641	114652	130294

(c) If weights are chosen using the *predictive* models, i.e., $w_i = 1/(\pi(1 - q_i))$ with the model-fitted value of q_i, the MSE components are

Bias	Var1	Var2	MSE
358	11520	137253	276711

(d) If weights are constant but q^* is taken to be not the true value but the one derived from the fitted-model values p_{ik}, then the MSE components are

Bias	Var1	Var2	MSE
1131	15811	115897	1410501

(d′) If constant model-based weights are chosen as $w_i = \tilde{w}_{mod} = 1/(\pi(1 - \tilde{q}_{mod}))$ with $\tilde{q}_{mod}$ as given by (13), the MSE components are

Bias	Var1	Var2	MSE
394	15700	115085	285771

Certainly it seems from these comparisons that it could be worthwhile to use predictively modelled response-probabilities to define weights, since the information to approach the truly optimal weights of settings (a) or (b′) will never be available. Stratum-dependent weights q_i derived from a good predictive model, as in (c), seem to provide nearly as good MSE's, but remarkably, constant weights $\tilde{q}_{mod}$ in (d′) are virtually just as good. Note that the relatively small changes among constant weights — $q^* = 0.23539$ in (b), to $q^* = 0.23586$ in (d), to $\tilde{q}_{opt} = 0.23125$ in (b′), and finally to $\tilde{q}_{mod} = 0.23270$ in (d′) — have sizeable consequences in MSE.

We illustrate with a further calculation on DE data, this time with subsampling probability vector $(\pi, a_1, a_2, a_3) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Now the results of the six plans (a)–(d′) itemized above can be found in the table

Weighting Plan	Bias	Var1	Var2	MSE
(a) strat, true	0	10144	112933	123078
(b) const, true	1152	14002	95072	1436854
(b') opt const	0	13848	94031	107880
(c) strat, pred	270	11082	108187	192374
(d) const, pred	949	13975	94888	1010369
(d') model const	332	13892	94331	218501

Here is a similar calculation, with the original subsampling probability vector $(\pi, a_1, a_2, a_3) = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, applied to 1990 Maryland state data and fitted models.

Weighting Plan	Bias	Var1	Var2	MSE
(a) opt, true	0	80886	1098353	1179239
(b) const, true	8337	120722	899262	70521374
(b') opt const	0	119558	890596	1010155
(c) opt, pred	2522	92122	1042125	7494569
(d) const, pred	34993	124479	927251	1.226e9
(d') model const	2801	119949	893504	8861252

The conclusions in these further tables are very much the same as in the first itemization of (a)–(d′): in the absence of knowledge of the true optimal sampling weights, stratumwise model-based weights from a good predictive model are very appealing. However, virtually all of the MSE benefits of the predictive model can already be obtained with constant weights $\tilde{w}_{opt} = 1/(\pi(1-\tilde{q}_{opt}))$ as prescribed by (13).

To make these calculations a little more realistic, we should consider some cases where a reasonable but *wrong* model is used, for example the model based on PA applied to MD data. The model fitted to PA data was applied to the 5523 MD *htyp* by block-group strata which had at least 21 HU's initially and which had at least one nonresponding HU as of the 75^{th} percentile of checkin times within ARA. The model was a reasonable one, but was also slightly misspecified. For data at the Mail-response stage, the model-fitted response rates at the PSU level had correlation 0.80 with the actual rates, where the analogous correlations for MD rates with the rates from the model fitted to MD was 0.83, and the correlation between PA rates and predicted rates from the model fitted to PA data was 0.85. For non-mail-responders, the MD response rates up to the 50^{th} percentile checkin within ARA correlated 0.18 with the PA model, where the MD model had correlation 0.23, and the correlation of corresponding PA rates with those predicted by the bPA-fitted model was 0.18. For nonresponders up to the 50^{th} percentile of checkins, the correlation between PA model predictions and actual response rates by the 75^{th} percentile was 0.15, while the MD model gave correlation 0.16 and the correlation of corresponding PA rates with those predicted by the PA-fitted model was 0.22. Finally, for nonresponders up to the 75^{th} percentile of checkin times within ARA, the correlation between PA model predictions and actual response rates by the 75^{th} percentile was 0.11, while the MD model gave correlation 0.09 and the correlation of corresponding PA rates with those predicted by the PA-fitted model was 0.15.

The following Table, representing the same weighting plans (a)–(d$'$) as previous tables, displays MSE and its components for MD data where the 'predictive' model is the misspecified one based on PA discussed in the previous paragraph. While greater degrees of model misspecification are very likely, this Table reinforces the conclusion of the previous tables. The central conclusion is that the predictive model if at all reliable *ought* to be used in determining weights for analysis of the multistage survey data, but only in the form $\tilde{w}_{opt} = 1/(\pi(1-\tilde{q}_{opt}))$ used in weighting-plan (d$'$), with $\tilde{q}_{opt}$

constant across PSU's, as given by (13).

TABLE 1. MSE AND COMPONENTS FOR MISSPECIFIED MODEL

Weighting Plan	Bias	Var1	Var2	MSE
(a) opt, true	0	80879	1097557	1178436
(b) const, true	8316	120634	898834	70177872
(b') opt const	0	119472	890178	1009650
(c) opt, pred	4696	95823	1065803	23215331
(d) const, pred	4188	120056	894533	18558040
(d') model cons	-1794	119222	888317	4224339

3 Conclusions

This paper has studied the sample-weighted estimation of population totals, based on survey data collected in several stages, in which nonresponders from earlier stages are sub-sampled at later stages. On theoretical and data-analytic grounds, comparisons among weighting plans with respect to Mean-Squared Error have been drawn. The clear but tentative conclusions, from both the theoretical formulas and data illustrations, are as follows:

- Weights for analyzing such survey data should be constant across response modes, although conceivably variable across demographically distinct PSU's.
- The weights chosen from a predictive demographic model for non-response model to be inversely proportional to sampling inclusion weights, $w_i \equiv 1/(\pi_i(1-\tilde{q}_{mod}))$, seem in all cases studied to be close to optimal. They are not strictly optimal from the point of view of MSE, especially if a very strongly predictive model for non-response by demographic predictors is available.
- It therefore seems advantageous to use a demographic model in weighting, but not in a form (such as (c) in the Tables) which applies different weighting adjustments in different PSU's. This kind of weighting adjustment appears robust to some degree of model misspecification. Further research is needed to investigate the degree of misspecification

which would make model-based weighting-adjustments (13) disadvantageous with respect to Mean-Squared Error.

4 References

Elliott, M., Little, R., and Lewitzky, S. (2000) Subsampling callbacks to improve survey efficiency. *Jour. Amer. Statist. Assoc.* **95**, 730-738.

Särndal, C., Swensson, B., and Wretman, J., **Model-Assisted Inference in Survey Sampling**. Springer-Verlag: New York 1997.

Slud, E. (1998) Predictive models for decennial census household response. *Proc. Amer. Statist. Assoc. on Survey Research Methodology*, 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. (2001) Order selection, random effects and multilevel predictors in modelling Decennial Census response. *Proc. Amer. Statist. Assoc. on Survey Research Methodology*, to appear.