Evaluation and Selection of Models for Attrition Nonresponse Adjustment

Eric V. Slud[1,2] and Leroy Bailey[1]

[1]Census Bureau, SRD, and [2]Univ. of Maryland College Park

Eric.V.Slud@census.gov, Leroy.Bailey@census.gov

Abstract. The setting of this paper is a longitudinal survey like SIPP, with successive "waves" of data collection from sampled individuals,in which nonresponse attrition occurs and is treated by weighting adjustment, either through adjustment cells or a model like logistic regression in terms of auxiliary covariates. Following Bailey (2004) and Slud and Bailey (2006), we measure the discrepancy in estimated initial-wave ("Wave 1") attribute totals between the survey-weighted estimator in the first wave and for the corresponding weight-adjusted estimator for the same Wave-1 item total based on later-wave respondents. The present research defines a composite metric of quality of a model used for nonresponse adjustment of a longitudinal survey. The metric combines the magnitudes of estimated between-wave adjustment biases based on subsets of the sample, relative to the estimated total, for various survey items. The maximum of the adjustment biases for estimated totals of a survey item are calculated from the first j sample units, as j ranges from 1 to the size of the entire (Wave-1) sample, after each of a number of random re-orderings either of the whole sample or of the units within specified cells (which are then also randomly re-ordered); and the average over re-orderings of the maximal adjustment bias is divided by the estimated wave-1 attribute total to give the metric value. Confidence bands for the metric are estimated, and the metric is applied to judge the quality of and to select among a collection of logistic-regression models for nonresponse adjustment in SIPP 96.

Keywords: Adjustment cells, Logistic regression, Model Selection, Nonresponse Weighting Adjustment, Random Re-ordering, Subdomains.

1 Introduction

Slud and Bailey (2006) studied the estimated biases between Wave 1 totals of various Survey of Income and Program Participation (SIPP) 1996 cross-sectional survey items and their estimates based on nonresponse-adjusted totals of the same Wave 1 items using only response data from a later Wave (4 or 12). The nonresponse adjustments studied were derived either by an adjustment cell method (using the 149 standard SIPP adjustment cells described by Tupek 2002) or a parsimonious logistic regression model for the later-wave response probabilities. One of the tentative conclusions of that work was that the relative and standardized magnitudes of the estimated biases varied considerably and somewhat erratically from one adjustment model to another. Many competing adjustment models could be defined, depending on which attribute variables were used in constructing adjustment cells or as logistic regression predictors. Slud and Bailey (2006) noted that using `Poverty` as a logistic regression predictor did have the artificial effect, akin to raking, of making the sample-wide estimated Wave 1 total of `Poverty` particularly small. However, since that effect directly stems from the sample-wide estimating equation defining the logistic regression coefficients, it was conjectured that this artificiality could be removed by considering estimated Wave 1 bias within a number of different subdomains.

In Slud and Bailey (2006) the possibility was also considered of customizing the adjustment model in order to remove between-wave adjustment biases as far as possible. This suggests creating a composite metric defined by combining the magnitudes of estimated between-wave adjustment biases for various SIPP items. The considerations of the previous paragraph suggest also including in the metric the estimated biases on multiple subdomains of the SIPP target population. We present below two constructions of such a composite metric, and we exhibit its values for the models studied in Slud and Bailey (2006). The ultimate objective of this research is then to use the metrics

defined to choose an optimal model – which might be either of adjustment-cell or logistic-regression form, although we restrict attention in this paper to the latter – based on SIPP 96 data.

2 Background & Previous Work

To measure the quality of nonresponse adjustment in a longitudinal survey, one would certainly try to evaluate the biases of adjustments using external data on the sample frame and the same variables whenever such external data are available. But that will seldom be the case.

There seems not to have been much published methodological work on how to measure the biases of adjustment, from the internal evidence of a longitudinal survey. One important paper on this topic is that of Dufour et al. (2001). That paper starts from the perspective that large longitudinal national studies will almost always be raked or calibrated to population totals derived from a past or current census or survey of high quality. The paper specifically considers calibration, and proposes to measure magnitudes of adjustment through a metric the authors define for tracking weight change through several stages of a weight-adjusted longitudinal study. By conducting a large simulation study within which they randomly subsample from a large longitudinal survey dataset (SLID, the Canadian Survey of Labor and Income Dynamics), the authors compare the weight-changes experienced from nonresponse weighting adjustment done by two main model-based adjustment approaches (Logistic regression with stepwise variable selection and Response Homogeneity Groups — what we call below the adjustment-cell method — with cells defined using a CHAID-based Segmentation Model). Calibration optimally adjusts weights according to a model (of adjustment-cell or logistic-regression type), in order that estimated population totals in designated subsets perfectly match the totals from an external study. Then the estimated adjustment biases (as in Bailey 2004 and Slud and Bailey 2006) for population totals of other early-stage variables could be used to judge the overall success of the modelling approach used in adjustment. This could have been, but was not, done in Dufour et al. 2001, nor were effects of weighting adjustment on population subdomains examined.

By contrast, we propose to adjust simultaneously, using models for later-stage response (also of adjustment-cell or logistic-regression type), the later-stage estimates of totals of population and other survey variables. We will then measure the biases of later-stage subtotals for population and the early-stage survey variables, for an array of different population subdomains including the cells to which calibration would have been done. The metric for effectiveness will combine the magnitudes of relative biases of specific survey variables over an interesting array of population subdomains. The results will then be assembled for a set of different survey variables along with population count into a weighted loss function. Unlike the calibration-first approach, this method provides the possibility of giving heavy but not overwhelming weight to population adjustment biases as opposed to biases in totals of other survey variables.

Although raking or calibration to updated-census population totals in defined cells will ultimately be done in practice whenever a weighting adjustment is applied to a large national longitudinal study, it may not be best to do all comparisons of adjustment methods with respect to bias in the presence of such raking/calibration adjustments, which often function as a black-box method superimposed on otherwise simple and understandable adjustment models. Therefore, in this paper as in Bailey (2004) and Slud and Bailey (2006), we restrict attention to weighting adjustments based on SIPP weights before adjustment-cell raking.

There has been a great deal of work on calibrating, reconciling, and benchmarking time series of differing reporting periods and accuracies, c.f. Dagum and Cholette (2006). But our literature search has yielded few papers (Dufour et al. 2001 in particular) explicitly recognizing survey and weighting-calibration aspects in this regard.

There has also been previous theoretical work on the large-sample behavior of model-based nonresponse weight adjustment methods. One recent example is Kim and Kim (2007); and these same authors, in an unpublished 2007 preprint, have considered the problem of choosing between alternative parametric models for survey nonresponse using the same data on which the estimated parameters will be applied to adjust the weights.

The goal of this research is to devise metrics to aid in the comparison of different model-based methods of adjustment for nonresponse due to attrition, which will provide a basis for choosing among adjustment methods. Several earlier comparative investigations related to adjustment methods have been conducted, even within the SIPP survey structure, but they seem not to have resulted in clear advantage for any adjustment method over others. (See Rizzo et al. 1994 for example.)

3 Formal Development: Metrics and Bounds

Let $\mathcal{S}$ denote the sample of $n = |\mathcal{S}|$ persons drawn from sampling frame $\mathcal{U}$, with known single inclusion probabilities $\{\pi_i\}_{i\in\mathcal{U}}$, and responding in Wave 1. For a series of cross-sectional survey measurements indexed by $k = 1, \ldots K$, such as the $K = 11$ items studied by Bailey (2004) and Slud and Bailey (2006), denote by $y_i^{(k)}$ the Wave 1 item values and $\mathbf{x}_i$ a vector of auxiliary variable values for all $i \in \mathcal{U}$. Let r_i denote individual response indicators (observed for all $i \in \mathcal{S}$) in a specified later Wave of the same survey, and let $p_i = P(r_i = 1 \,|\, \mathcal{S})$ denote the (unknown) conditional probabilities of later-wave response. Let $\hat{p}_i = g(\mathbf{x}_i, \hat{\vartheta})$ denote estimators of these unknown probabilities derived (using a known function g) from a parametric model using auxiliary data $\mathbf{x}_i$, within which parameter-estimators $\hat{\vartheta}$ are obtained via estimating equations (Kim and Kim, 2007). For any population attribute z_i, $i \in \mathcal{U}$, the frame-population total is denoted $t_z = \sum_{i\in\mathcal{U}} z_i$, and the corresponding Horvitz-Thompson estimator is $\hat{t}_z = \sum_{i\in\mathcal{S}} z_i/\pi_i$.

For each survey item $y_i^{(k)}$, $i \in \mathcal{U}$, with respect to the specific strategy of adjustment embodied in the estimated response probabilities $\hat{p}_i$, and for each domainsubset $\mathcal{D} \subset \mathcal{U}$ of the population, define the estimated nonresponse bias

$$\hat{B}_k(\mathcal{D}) \;=\; \sum_{i\in\mathcal{D}\cap\mathcal{S}} \Big(\frac{r_i}{\hat{p}_i} - 1\Big)\, y_i^{(k)} \,/\, \pi_i \tag{1}$$

In Slud and Bailey (2006) and earlier papers of Bailey, the domain $\mathcal{D}$ was all of $\mathcal{U}$, and the quantity $\hat{B}_k(\mathcal{U})$ was interpreted as the difference between an adjusted estimator of $t_{y^{(k)}}$ using only the data $(y_i^{(k)}, \mathbf{x}_i,\ i \in \mathcal{S})$ and the ordinary Horvitz-Thompson estimator $\hat{t}_{y^{(k)}}$, and was regarded as an estimator of attrition nonresponse bias due to the method of adjustment.

3.1 Relative Subdomain Bias

We now propose a measure of the typical relative bias in estimating item totals over subdomains. The idea is to consider the largest value of absolute relative bias $\hat{B}_k(\mathcal{D})/\hat{t}_{y^{(k)}}$ over a collection of different subsets $\mathcal{D} \subset \mathcal{U}$. Suppose that we re-order the elements of $\mathcal{U}$, inducing a re-ordering $\tau = (\tau(1), \tau(2), \ldots, \tau(n))$ of the n elements of $\mathcal{S}$. The largest absolute bias in survey variable k over consecutively τ-indexed subdomains of $\mathcal{S}$ is

$$\max_{1\le a\le b\le n} |\hat{B}_k(\{\tau(i) :\ a \le i \le b\})| \;\le\; 2 \cdot \max_{1\le a\le n} |\hat{B}_k(\{\tau(1), \ldots, \tau(a)\})|$$

To measure the overall relative bias in estimating item k totals over subdomains, we define

$$m_k \;=\; E_\tau\Big(\max_{1\le a\le n} |\hat{B}_k(\{\tau(1), \ldots, \tau(a)\})|\Big) / \,\hat{t}_{y^{(k)}} \tag{2}$$

where the expectation is taken, for a fixed sample, over random permutations τ chosen equiprobably from the $n!$ permutations of the elements of $\mathcal{S}$. The quantity m_k is smaller than the largest relative bias $|\hat{B}_k(\mathcal{D})|/\hat{t}_{y^{(k)}}$ over *all* subsets $\mathcal{C} \subset \mathcal{U}$ — which is too large an estimate of error, and also too expensive to calculate — but does represent the typical magnitude of the worst relative bias in a random scanning order of the sampled population.

In settings where the relative estimated bias

$$\delta^{(k)} \;\equiv\; \hat{B}_k(\mathcal{U})/\hat{t}_{y^{(k)}} \;=\; \sum_{i\in\mathcal{S}} \Big(\frac{r_i}{\hat{p}_i} - 1\Big) \frac{y_i^{(k)}}{\pi_i} \;\Big/\, \hat{t}_{y^{(k)}} \tag{3}$$

is large, we will see below that m_k and its estimator $\hat{m}_k$ are not much different from $|\delta^{(k)}|$. However, if $\delta^{(k)}$ is small — which may be true for artificial reasons if the model used to define $\hat{p}_i$ prominently features the attributes $\{y_j^{(k)},\ j \in \mathcal{S}\}$), then $\hat{m}_k$ will often be meaningfully large, reflecting the fact that the model-fitting does not simultaneously adjust for weighted $y_i^{(k)}$ totals over arbitrary subsets of the sample. This is an attempt to penalize models which directly adjust the population-wide total of an attribute.

The bias measure m_k cannot be calculated directly from the sample data, but can be estimated by evaluating its defining expectation over random permutations τ using a Monte Carlo simulation strategy. For each of a set $1, \ldots, R$ of indices c denoting Monte Carlo replicates, we define independent random permutations τ_c of the

indices $i \in \mathcal{S}$. For each b, $(\tau_c(j),\ 1 \le j \le n)$ is equiprobably chosen from the $n!$ possible re-orderings of $\mathcal{S}$, which is easily implemented in a Monte Carlo simulation by defining a sample of independent Uniform$(0,1)$ variates $\mathbf{V}_c = (V_{ci},\ i \in \mathcal{S})$ and to let $\tau_c(j)$ be the sequence of indices i of the V_{ci} observations written in increasing order. Then the estimator $\hat{m}_k$ defined in (2) is

$$\begin{aligned} \hat{m}_k &= \frac{1}{R}\sum_{c=1}^{R} \max_{1\le j\le n} |\hat{B}_k(\{\tau_c(1),\ldots,\tau_c(j)\})| \Big/ \hat{t}_{y^{(k)}} \\ &= \frac{1}{R\,\hat{t}_{y^{(k)}}} \sum_{c=1}^{R} \max_{0<x\le 1} \Big| \sum_{i\in\mathcal{S}} I_{[V_{ci}\le x]}\,(\frac{r_i}{\hat{p}_i} - 1)\,\frac{y_i^{(k)}}{\pi_i} \Big| \end{aligned} \tag{4}$$

As a metric for nonresponse bias combined over all survey variables indexed by $k = 1,\ldots,K$, we propose a simple weighted average and estimator

$$M = \sum_{k=1}^{K} w_k\, m_k = \sum_{k=1}^{K} w_k \cdot \Big(\sup_{\mathcal{D}\subset\mathcal{S}} |\hat{B}_k(\mathcal{D})| \Big/ \hat{t}_{y^{(k)}} \Big) \tag{5}$$

$$\hat{M} = \sum_{k=1}^{K} w_k\, \hat{m}_k \tag{6}$$

where $\mathbf{w} = \{w_k\}_{k=1}^{K}$ is a fixed vector of positive weights summing to 1. If all survey variables are considered equally important then, as below, we would use $w_k = 1/K$.

The quality of estimation of m_k in terms of $\hat{m}_k$, and relationships between these and $|\delta^{(k)}|$, are addressed in Section 3.3 below. We turn first to the modification of (2) and (4) to allow expected and estimated maximum absolute relative discrepancies with respect only to those random re-orderings which preserve specific cells of the population, such as the cells to which population totals would be raked or calibrated.

3.2 Metric for Subdomain Bias with Distinguished Cells

Most random permutations of the sample completely shatter any meaningful sample subdomains. Yet the idea behind raking or calibration is precisely that certain estimated subdomain totals — usually, the estimated population totals over the *cells* A_j of a specified geographic-demographic partition $\mathcal{U} = \cup_{j=1}^{J} A_j$ of the frame population — must be constrained equal to those of a current (updated) census. For that reason, it makes sense to measure bias estimates $\hat{B}_k(A_j)$ over these cells, where we assume from now on that a partition $\mathcal{A}$ of $\mathcal{U}$ into cells A_j, $j = 1,\ldots,J$, has been fixed. The idea is to modify (2) so that the allowed permutations must retain the consecutive indexing of elements in each cell A_j.

One approach would be to aggregate these biases into an *relative accumulated absolute bias*

$$m_k^{Cum} = \sum_{j=1}^{J} \omega_j^{(k)}\, |\hat{B}_k(A_j)| \Big/ \hat{t}_{y^{(k)}} \tag{7}$$

where $\omega_j^{(k)}(\mathcal{S})$ are a set of cell- and item-specific weights. A related approach would be to replace each term $|\hat{B}_k(A_j)|$ in (7) by the expectation of $\max_{l\in A_j} |\hat{B}_k(\{i \in A_j : \tau_j(i) \le \tau_j(l)\}|$ with respect to a random permutation τ_j of the elements of $\mathcal{S}\cap A_j$. However, in either of these two forms, the relative accumulated absolute bias is likely too conservative to be very useful, because it aggregates across cells the worst sub-cell biases, as though all domain totals in all cells could be simultaneously badly biased.

A less extreme modification of (2), which we adopt below, would combine cellwise biases within a partition $\mathcal{A}$ so that the permutations τ — now denoted σ — leave the cells $A_j \cap \mathcal{S}$ invariant. To explain this invariance, we assume that the sample is indexed in such a way that the $n_j \equiv |A_j \cap \mathcal{S}|$ sampled elements in the j'th cell A_j appear consecutively in the enumerated sample $\mathcal{S}$. The invariance of the cells under σ means that for all $1 \le j \le J$, the elements $\{\sigma(i) :\ i \in A_j \cap \mathcal{S}\}$ also form a consecutively indexed block in the indexed sample $\mathcal{S}$. Now the allowed random permutations σ of the sample elements are chosen equiprobably from the $J!\,\prod_{j=1}^{J} n_j!$ permutations which first permute the J complete blocks $A_j \cap \mathcal{S}$ of n_j elements each and then permute the elements within the

re-ordered blocks. Finally, we define the expectation over σ of the maximum absolute cumulative weighted sum of cellwise biases relative to $\hat{t}_{y^{(k)}}$, as follows:

$$m_k^* \equiv E_\sigma\Big(\max_{1\le q\le n} |\hat{B}_k(\{\sigma(1), \ldots, \sigma(q)\})|\Big) / \hat{t}_{y^{(k)}} = \tag{8}$$

$$E_\sigma\Big(\max_{1\le j\le J,\, b\in A_j} \Big| \sum_{l=1}^{j-1} \omega_{\sigma(l)}\, \hat{B}_k(A_{\sigma(l)}) + \omega_{\sigma(j)}\, \hat{B}_k(\{\tau(a):\ a\in A_{\sigma(j)},\ a\le b\})\Big|\Big)$$

Here $\omega_j = \omega_j^{(k)}(\mathcal{S})$ are again a set of cell- and item-specific weights (usually taken to be 1) which may depend on the sample, but which satisfy the relation $J^{-1}\sum_{j=1}^{J}\omega_j^{(k)} = 1$.

An estimator for the modified quantity (8) can be implemented in terms of a collection of random batches $\mathbf{V}_c$ of n independent Uniform$(0,1)$ random variates, along with independent batches $\mathbf{U}_c$ of J independent Uniform$(0,1)$ variates, for $1\le c\le R$. For each fixed batch-index c, we use the ordering of the variates $U_{c1}, U_{c2}, \ldots, U_{cJ}$ to determine the c'th random ordering of the blocks $A_j\cap\mathcal{S}$, $1\le j\le J$. Next, the c'th reordering of the elements i within the re-ordered block $A_j\cap\mathcal{S}$, is given by the order of the variates $(V_{ci},\ i\in A_j\cap\mathcal{S})$. For each $q = 1,\ldots,n$ indexing an element of the sample $\mathcal{S}$, denote by $j(q)$ the index j for which $q\in A_j$. With these notations in mind, we express the estimator for (8) as

$$\hat{m}_k^* \equiv (R\,\hat{t}_{y^{(k)}})^{-1}\sum_{c=1}^{R}\max_{1\le q\le n}\Big|\sum_{l:U_{c,l}<U_{c,j(q)}}\omega_l^{(k)}\,\hat{B}_k(A_l) + \omega_{j(q)}^{(k)}\,\hat{B}_k(\{i:\ i\in A_{j(q)},\ V_{ci}\le V_{cq}\})\Big| \tag{9}$$

or equivalently,

$$\hat{m}_k^* \equiv (R\,\hat{t}_{y^{(k)}})^{-1}\sum_{c=1}^{R}\max_{1\le q\le n}\Big|\sum_{i:\,U_{c,j(i)}\le U_{c,j(q)},\,V_{ci}\le V_{cq}}(\frac{r_i}{\hat{p}_i}-1)\,\frac{y_i^{(k)}}{\pi_i}\Big| \tag{10}$$

We next show how to place confidence bounds on the differences between the quantities m_k, m_k^* and their estimates $\hat{m}_k$, $\hat{m}_k^*$ and on the differences between these quantities and $|\delta^{(k)}|$.

3.3 Confidence Intervals and Bounds for m_k

In the first part of this Section, we provide a theoretical development of confidence intervals by bounding $\hat{m}_k - m_k$ and $m_k - |\delta^{(k)}|$ probabilistically. However, all of these quantities are functions of the sampled survey data, and the probability statements made at this stage concern only the chance element introduced by the random permutations τ, τ_c used in defining (2) and (4). At the end of the Section, we interpret the meaning of sample-based metric-estimators $\hat{m}_k$ for the survey population and adjustment model.

We begin with the simplest and clearest confidence statement. Since $\hat{m}_k$ is calculated as the empirical average over quantities calculated from a series of R random permutations of the sample, its sampling variability due to those permutations can be assessed by empirical standard errors

$$se(\hat{m}_k) = \frac{1}{|\hat{t}_{y^{(k)}}|}\Big[\frac{1}{R(R-1)}\sum_{c=1}^{R}\big(\max_{0<x\le 1}\Big|I_{[V_{ci}\le x]}\,(\frac{r_i}{\hat{p}_i}-1)\,\frac{y_i^{(k)}}{\pi_i}\Big| - \hat{t}_{y^{(k)}}\,\hat{m}_k\big)^2\Big]^{1/2}$$

Thus, with approximate 99% confidence when R is large,

$$|m_k - \hat{m}_k| \le 2.576\cdot se(\hat{m}_k) \tag{11}$$

and similar confidence statements with respect to the randomness of the permutations σ_c can be given bounding $m_k^* - \hat{m}_k^*$.

The difference between the metric value m_k and the overall relative bias $|\delta^{(k)}|$ is due to the fluctuations with varying x of the quantities

$$Z_k(x) = \sum_{i\in\mathcal{S}} I_{[V_i\le x]}\,(\frac{r_i}{p_i}-1)\,\frac{y_i^{(k)}}{\pi_i}$$

being maximized in (4), where $V_i = V_{ci}$ denote independent identically distributed Uniform$(0,1)$ variates. If these quantities were replaced by their expectations (i.e., if $I_{[V_i \le x]}$ were replaced by x), then the expression (4) would become $|\delta^{(k)}|$. Thus, the discrepancy $\hat{m}_k - |\delta^{(k)}|$ can be bounded by the maximum absolute value of the random *weighted empirical* process indexed by a continuous argument $x \in [0,1]$,

$$\beta_k(x) \;=\; \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{S}} \Big(I_{[V_i \le x]} \;-\; x \Big) \, (\frac{r_i}{\hat{p}_i} \;-1) \, \frac{y_i^{(k)}}{\pi_i} \tag{12}$$

conditionally with all sample data $\Big\{ i,\; r_i,\; x_i,\; (y_i^{(k)},\; 1 \le k \le K) \;:\quad i \in \mathcal{S} \Big\}$ fixed and with only the variates $V_i,\; i \in \mathcal{S},$ regarded as random. The process $\beta_k(\cdot)$ has mean 0 , and according to a slight extension of the Donsker Theorem (Pollard 1980), which can be proved as Corollary of the Martingale Central limit Theorem (Hall and Heyde 1980), has approximate distribution for large n the same as

$$\sqrt{\gamma^{(k)}} \; W^{\circ}(x) \;\equiv\; \Big[\frac{1}{n} \sum_{i \in \mathcal{S}} (\frac{r_i}{\hat{p}_i} \;-1)^2 \, \frac{(y_i^{(k)})^2}{\pi_i^2} \Big]^{1/2} W^{\circ}(x) \tag{13}$$

as a random continuous function of $x \in [0,1]$, where $W^{\circ}(x)$ denotes a *tied-down Wiener process* or Gaussian process with mean 0 and

$$\mathrm{Cov}(W^{\circ}(v), W^{\circ}(u)) \;=\; \min(v,u) \;-\; v \cdot u$$

The scaling constants governing the amplitude of fluctuations of $\beta_k(\cdot)$,

$$\gamma^{(k)} \;=\; \frac{1}{n} \sum_{i \in \mathcal{S}} (\frac{r_i}{\hat{p}_i} \;-1)^2 \, (y_i^{(k)})^2 / \pi_i^2 \tag{14}$$

can readily be computed from the sample data, and under general assumptions remain bounded for large n.

By definition of m_k and the remark that (4) would become $|\delta^{(k)}|$ if $I_{[V_i \le x]}$ were replaced by x,

$$\begin{aligned} \Big| m_k \;-\; |\delta^{(k)}| \Big| \;&\le\; \frac{1}{\hat{t}_{y^{(k)}}} \, E_{\mathbf{V}} \Big(\max_{0 < x \le 1} \Big| \sum_{i \in \mathcal{S}} (I_{[V_i \le x]} - x) \, (\frac{r_i}{\hat{p}_i} \;-1) \, \frac{y_i^{(k)}}{\pi_i} \Big| \Big) \\ &=\; \frac{\sqrt{n \, \gamma^{(k)}}}{\hat{t}_{y^{(k)}}} \, E_{\mathbf{V}} \big(\max_{0 < x \le 1} |\beta_k(x)| \big) \;\approx\; 1.2286 \; \frac{\sqrt{n \, \gamma^{(k)}}}{\hat{t}_{y^{(k)}}} \end{aligned} \tag{15}$$

since 1.2286 is the expectation of $\sup_{x \in [0,1]} |W^{\circ}(x)|$ which arises in calculating percentage points of the one-sample Kolmogorov-Smirnoff statistic, readily calculated using the density of this random variable given by Kolmogorov and reproduced by Feller (1948).

For specific items k, we find when n is large that for moderate numbers R of random permutations, the difference $\hat{m}_k - m_k$ (or $\hat{m}_k^* - m_k^*$) is generally very small compared to m_k (respectively m_k^*). Then, by calculating the right-hand side of (15), we find roughly how small the value $\hat{m}_k$ must be in order that the sample data be compatible with a zero relative bias $\delta^{(k)}$. The objective of this kind of analysis is first of all to flag as 'inadequately adjusted' those items for which model-based attrition nonresponse adjustment has resulted in estimated metric values $\hat{m}_k$ greater than the sum of the right-hand sides of (11) and of (15). Since we will find generally that the values of $\hat{m}_k^*$ and $\hat{m}_k$ are roughly the same, we will use the same threshold for metric values $\hat{m}_k^*$.

Next, we compare the estimated metric values $\hat{m}_k$ and $\hat{m}_k^*$, individually or in their aggregated form (6), across different adjustment models with a view to choosing a 'best' model in a specific survey application.

4 Adjustment Metric Values in SIPP 96

For the case of SIPP 96, with $K = 11$ cross-sectional items, response probabilities $\hat{p}_i$ were estimated by the specific adjustment-cell and logistic-regression models mentioned above, all as described in detail by Slud and Bailey (2006). Briefly, the cross-sectional survey items $y_i^{(k)}$ studied are: indicators that the individual lives in a Household which

Table 1: Logistic regression models used to adjust Wave 4 or Wave 12 nonresponse in SIPP 96. `Df` is the number of independent coefficients in each model, including Intercept, and `Dev` the deviance for the 94444-record SIPP 96 sample data.

Model	Df	Variables	Dev
A	7 8	Wnotsp Renter College RefPer Black Renter*College Black*College	76558.0
B		same as **A**, plus Pov	76544.6
C	13	same as **B**, plus Foodst Mdcd Heins UnEmp Div	76299.3
D	13	same as **B**, minus Black*College plus Mdcd Heins UnEmp Pov*Heins Mdcd*Heins Heins*College	76242.4
E	17	same as **D**, plus hisp + Famtyp	76017.1
F	18	same as **C**, plus Afdc SocSec Emp Mar	76279.9

receives (i) Food Stamps (`Foodst`), or (ii) Aid to Families with Dependent Children (`AFDC`); or indicators that the individual receives (iii) Medicaid (`Mdcd`), or (iv) Social Security (`SocSec`); and indicators that the individual (v) has health insurance (`Heins`), (vi) is in poverty (`Pov`), (vii) is employed (`Emp`), (viii) is unemployed (`UnEmp`), (ix) is not in the labor force (`NILF`), (x) is married (`MAR`), or (xi) is divorced (`DIV`).

In this data example, nonresponse is adjusted in one of two ways: either using an adjustment-cell model based on 149 standard cells (Tupek 2002) defined in terms of variables including race, hispanicity, and family-type; or using one of a series of logistic regression models A–F summarized in Table 1. (Of these models Model A and B were the ones used in Slud and Bailey 2006.) The models C–E were selected to have progressively better fit, using an indicator of Wave 4 response as response-variable within the 94444 SIPP Wave-1 sample records with positive base-weights. The variables used in these regression models include race, hispanic origin, Renter versus Owner of housing unit, indicator that individual is the Household Reference Person, indicator of College education, a 4-category variable of Family type, plus some or all of the 11 SIPP survey items listed above.

The method followed in this data analysis, as described and justified in the previous Section, is based on searching for metric values $\hat{m}_k$, $\hat{m}^*_k$ which are large compared to the bounds obtained by adding the right-hand sides of (11) and (15). This contrasts with the approach of Slud and Bailey (2006) which, in the present notation compared estimated population-wide adjustment biases $\delta^{(k)}$ with their design-based standard errors as found by a Balanced Repeated Replication method. A summary of the results of Slud and Bailey (2006) in the present notation is given in Tables 6 and 7 of the Appendix. Despite the differences in method, the two sets of results from the two different methods seem quite consistent.

Calculations of $\hat{m}_k$ have been made with $R = 100$ random-permutation Monte Carlo replications, with the results for Model B presented in Table 2 below. (Because n=94444 is so large, the between-replication differences are small and this choice of R is ample.) The final columns of Table 2 respectively display the bounds $b_{4,k}$, $b_{12,k}$ on the right-hand sides of (15) (which turn out to be virtually identical for the adjustment-cell and logistic-regression adjustment methods) for adjustments of Wave 4 and 12 nonresponse. It also turns out that for all items and combinations 4C, 4L, 12C, and 12L, the bounds on the right-hand side of (11) are much smaller, ranging from 1–5% of the corresponding bounds (15). The analogous Table with logistic models A and D and F, also calculated with $R = 100$ iterations, are displayed as Table 3. However, the columns of bounds $b_{4,k}$, $b_{12,k}$ are included in the latter Table only for model D, because the bounds for the other models are virtually identical with these, and again the bounds from (11) are only a few percent of the bounds (15).

Inspection of Tables 2 and 3 reveals that the metric $\hat{m}_k$ with very few exceptions in Wave 12 clearly exceeds the corresponding bounds b_k for the adjustment-cell model and all of the logistic regression models A and B. One notable exception is `Pov`, where as seen by Slud and Bailey (2006), model B includes `Pov` as a predictor and does adjust effectively both in Waves 4 and 12. Similarly, we see that Model D which includes variables `Pov, Mdcd, Heins`, and `UnEmp` as predictors, does a particularly good job of adjusting the totals of these same variables as measured by the metric $\hat{m}_k$. Indeed, the most striking preliminary conclusion from examining the tables of metric values under these various logistic regression models is that including a variable as a predictor generally results in very good adjustment

Table 2: Quantities $\hat{m}_k$ in (4) estimated from SIPP96 data, for later-wave nonresponse adjustment either to wave 4 or 12, and by either the Adjustment-Cell (**C**) or Logistic-Regression (**L**) method (Model B) and based on $R = 100$ replications. The last two columns are the bounds in (15), with $\alpha = .01$.

Item	$\hat{m}^{4C}$	$\hat{m}^{4L}$	$\hat{m}^{12C}$	$\hat{m}^{12L}$	$b_{4,k}$	$b_{12,k}$
Foodst	.0052	.0186	.0442	.0130	.0056	.0123
AFDC	.0067	.0248	.1040	.0350	.0078	.0173
Mdcd	.0066	.0279	.0163	.0426	.0053	.0119
SocSec	.0191	.0116	.1118	.1038	.0041	.0086
Heins	.0085	.0065	.0197	.0133	.0019	.0040
Pov	.0187	.0033	.0372	.0091	.0047	.0097
Emp	.0016	.0017	.0082	.0122	.0020	.0041
UnEmp	.0534	.0594	.1176	.1280	.0131	.0250
NILF	.0032	.0034	.0333	.0462	.0033	.0069
MAR	.0111	.0018	.0508	.0226	.0025	.0051
DIV	.0124	.0201	.0235	.0390	.0067	.0133

Table 3: Quantities $\hat{m}_k$ estimated from SIPP96 data based on $R = 100$ random permutations, for wave 4 or 12 nonresponse adjustment by logistic regression model A (first two columns) or model D (next two columns). The last two columns are the bounds $b_{4,k}$, $b_{12,k}$ from (15) using model D.

Item	$\hat{m}^{4,A}$	$\hat{m}^{12,A}$	$\hat{m}^{4,D}$	$\hat{m}^{12,D}$	$\hat{m}^{4,F}$	$\hat{m}^{12,F}$	$b^D_{4,k}$	$b^D_{12,k}$
Foodst	.0120	.0086	.0076	.0110	.0039	.0093	.0056	.0123
AFDC	.0175	.0446	.0067	.0624	.0053	.0134	.0077	.0170
Mdcd	.0219	.0346	.0035	.0078	.0037	.0084	.0052	.0114
SocSec	.0117	.1040	.0125	.1066	.0027	.0073	.0041	.0086
Heins	.0076	.0148	.0013	.0027	.0012	.0028	.0019	.0039
Pov	.0123	.0127	.0032	.0074	.0032	.0085	.0047	.0098
Emp	.0021	.0116	.0015	.0161	.0014	.0034	.0020	.0041
UnEmp	.0626	.1322	.0095	.0207	.0098	.0184	.0139	.0288
NILF	.0026	.0447	.0029	.0456	.0023	.0063	.0033	.0069
MAR	.0023	.0236	.0018	.0213	.0017	.0037	.0025	.0051
DIV	.0201	.0390	.0168	.0334	.0011	.0026	.0068	.0139

as measured either by metric $\hat{m}_k$ or $\hat{m}^*_k$. This is true even under Model F, where we can see from Table 1 that the last batch of variables entered between model D and F are not very significant as measured by an increase in maximized loglikelihood, or equivalently in decreased Deviance.

Recall that we devised the metrics $\hat{m}_k$, $\hat{m}^*_k$ in part to penalize model-based adjustment which, like raking, removes bias directly in terms of population totals. Recall also that $\hat{m}^*_k$ differed only by finding maximum absolute discrepancies over consecutive sequences of re-ordered indices which keep adjustment-cells consecutively indexed. In fact, the metric values $\hat{m}^*_k$ turn out to be only slightly larger than $\hat{m}_k$, and they follow a very similar pattern across the different models. Consider Table 4 charting the progression of averaged $\hat{m}^*_k$ metrics (over $k = 1, \ldots, 11$ and Population Count) as the adjustment model varies over the Adjustment Cell model and the six Logistic Regression models described in Table 1, and for brevity let $\hat{M}^*$ denote the average of these metric values analogous to $\hat{M}$ in (6), with equal weights $w_k = 1/12$. The logistic regression models are all clearly better than the cell-based model in adjusting at Wave 12, but at Wave 4, models A and B actually seem a little worse than the cell-based method. Since the models A–E are listed in order of decreasing Deviance or AIC, there is no strict relationship between decreasing AIC and decreasing $\hat{M}^*$. Model C looks to be the best among A–E, and might have been chosen also for parsimony from examination of deviances; but the metric $\hat{M}^*$ rewards model F for including essentially all of the SIPP items as predictors.

Although we would not have chosen model F from likelihood considerations, it may well be that this model is a good choice from the vantage point of nonresponse adjustment. The SIPP dataset is large enough (n=94444)

Table 4: Metric values $\hat{m}_k^*$ calculated on SIPP 96 data for Adjustment-cell model and for logistic regression models A–F and averaged over $k = 1, \ldots, 12$, where 'item' 12 is Population Count ($y_i^{(12)} \equiv 1$).

Model	Wave-4	Wave-12
Adj.Cell	0.01228	0.04741
LReg, A	0.01451	0.03942
LReg, B	0.01504	0.03893
LReg, C	0.00426	0.02475
LReg, D	0.00571	0.02812
LReg, E	0.00481	0.02654
LReg, F	0.00342	0.00782

Table 5: Metric (8) values for Wave 4 adjustment, based on the Adjustment cell and logistic regression models, using SIPP 96 data with adjustment cells as partition elements A_j.

item	ModA	ModB	ModC	ModD	ModE	ModF	Adj.Cell
Fdst	.0667	.0700	.0615	.0627	.0608	.0609	.0594
AFDC	.0800	.0840	.0728	.0734	.0705	.0722	.0697
Mdcd	.0588	.0620	.0517	.0516	.0514	.0511	.0527
SocS	.0300	.0300	.0305	.0310	.0315	.0287	.0332
Hins	.0249	.0247	.0238	.0241	.0226	.0237	.0224
Pov	.0627	.0632	.0574	.0576	.0556	.0566	.0572
Emp	.0276	.0281	.0263	.0261	.0234	.0259	.0227
UnEmp	.1005	.0988	.0868	.0866	.0878	.0867	.0946
NILF	.0315	.0315	.0312	.0315	.0313	.0304	.0327
MAR	.0207	.0206	.0206	.0207	.0198	.0216	.0225
DIV	.0498	.0498	.0459	.0476	.0443	.0451	.0480
POP	.0260	.0264	.0243	.0244	.0222	.0238	.0221

that all of the SIPP survey items except `AFDC` and `Emp` have highly significant coefficients. Moreover, the highly parametrized adjustment models C–F are accomplishing something that simple raking cannot: they are generating response probabilities with good behavior over adjustment cells considered as subdomains. To see this more clearly, consider the unforgiving metric (8): for each item k, we sum the absolute estimated biases $|\hat{B}_k(A_j)|$ over all adjustment cells and form the ratio of the total to $\hat{t}_{y^{(k)}}$. The result on the SIPP 96 data is given in Table 5. Although the Adjustment Cell model is given an advantage by evaluating adjustment effectiveness over exactly the same adjustment cells used to form ratio weighting-adjustments, a few of the logistic regression models (especially models D–F) do at least as well, item by item, with far fewer parameters than the 149 adjustment-cell response fractions. Nevertheless, with respect to this metric none of these models except possibly Model F shows much advantage over the others.

5 Conclusions

This paper has developed metrics for nonresponse-adjustment effectiveness, calculated after randomly re-indexing the survey sample and calculating maximum discrepancies over consecutively indexed subdomains. The objective was to discount any advantage which an adjustment regression model might achieve toward eliminating whole-sample nonresponse biases by including survey attributes as predictors. However, when applied to SIPP 96 data, the metrics developed did not have the expected effect. Those regression models which incorporated most or all of the interesting survey attributes did exceptionally well with respect to the new metrics, even though some of those models would not have been preferred from examination of likelihood ratios or deviance. While the same adjustment strategy could not be tried if the selected set of 'interesting' survey attributes were very large, the strategy may actually be a good one in the setting chosen, where the selected set of attributes was still small enough to contain variables which were almost all highly predictive of response and yet not redundant (except for the triple `Emp, UnEmp, NILF` which

partitions the population by definition.)

One important check on the usefulness of the adjustment effectiveness metrics developed here remains to be pursued in depth: namely, an examination of the sampling variability of the metrics through calculation of their design-based variances. The variances could be calculated by a BRR method, although that computationally intensive calculation involves a design crossing many replicate weight-factors with many iterated random permutations of the sample. This kind of investigation of sampling variability of the metrics considered as statistics, may show that the present adjustment effectiveness study was statistically stable enough to justify a reliable adjustment strategy for SIPP.

6 References

Bailey, L. (2004), Weighting alternatives to compensate for longitudinal nonresponse in the Survey of Income and Program Participation. Census Bureau internal report, Nov. 16, 2004.

Billingsley, P. (1968) **Convergence of Probability Measures**. Wiley: New York.

Dagum, E. and Cholette, P. (2006) **Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series**. *Lect. Notes in Statist.* **186**, Springer: New York.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. and Särndal, C.-E. (2001), A better understanding og weight transformation through a measure of change. *Survey Methodology*, **27**, 97-108.

Kim, Jae-Kwang and Kim, Jay (2007), Nonresponse weighting adjustment using estimated response probability, *Canadian Jour. of Statist.*, to appear.

Feller, W. (1948) On th Kolmogorov-Smirnov limit theorem for empirical distributions. *Ann. Math. Statist.* **19**, 177-189.

Pollard, D. (1980) **Convergence of Stochastic Processes**. Springer-Verlag: New York.

Rizzo, L., Kalton, G., Brick, M. and Petroni, R. (1994), Adjusting for panel nonresponse in the Survey of Income and Program Participation, ASA Surv. Res. Methodology Proceedings paper, JSM 1994.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992) **Model Assisted Survey Sampling**. Springer-Verlag: New York.

Slud, E. and Bailey, L. (2006), Estimation of attrition biases in SIPP. ASA Surv. Res. Methodology Proceedings paper, JSM 2006, Seattle, WA.

Tupek, A. (2002) SIPP 96: specifications for the longitudinal weighting of sample people. Internal Census Bureau memorandum.

7 Appendix: Results from Slud & Bailey 2006

Table 6: Relative biases $\delta^{(k)}$ between waves 4 versus 1 and 12 versus 1 in SIPP 1996 survey-weighted estimated (adjusted) population total of 11 Wave 1 survey items, as found by Slud and Bailey (1996). The entries are indexed by wave 4 or 12 and `C` (cell-based adjustment model) or `L` (logistic regression model B). Entries are the biases given in Tables 1 and 2 of Slud and Bailey divided by `Wav1` item totals. Totals given here in 1000's.

Item	Total	μ_{4C}	μ_{4L}	μ_{12C}	μ_{12L}
Foodst	27268	-0.00315	0.01814	-0.04325	0.00956
AFDC	14030	-0.00394	0.02412	-0.10352	-0.03274
Mdcd	28173	0.00544	0.02764	-0.01410	0.04149
SocSec	37087	0.01885	0.01115	0.11168	0.10364
Heins	194591	0.00837	0.00634	0.01948	0.01299
Pov	41796	-0.01843	0.00071	-0.03655	0.00587
Emp	191201	0.00099	0.00113	-0.00758	-0.01172
UnEmp	6406	-0.05257	-0.05868	-0.11609	-0.12657
NILF	66647	0.00221	0.00245	-0.03290	0.04588
MAR	114367	0.01096	0.00083	0.05060	0.02231
DIV	18463	-0.01116	-0.01936	-0.02063	-0.03732

Table 7: Relative standard errors (rse) of the Wave 4 versus Wave 1 and Wave 12 versus Wave 1 differences between estimated totals of 11 SIPP96 Wave 1 survey items. Variances were calculated by Slud and Bailey (2006) using Fay's BRR method, and rse table entries are the respective SE columns in Tables 3 and 4 of Slud amd Bailey (2006) divided by the Wave 1 totals. The rse's given here should be multiplied by 2.576 to give 99% confidence interval half-widths comparable to the bounds b_k in Table 2.

Item	Total	rse_{4C}	rse_{4L}	rse_{12C}	rse_{12L}
Foodst	27268	0.00615	0.00655	0.01251	0.01509
AFDC	14030	0.01107	0.01150	0.02078	0.02375
Mdcd	28173	0.00481	0.00523	0.01050	0.01274
SocSec	37087	0.00335	0.00387	0.00682	0.00766
Heins	194591	0.00117	0.00121	0.00234	0.00258
Pov	41796	0.00429	0.00020	0.00901	0.00187
Emp	191201	0.00069	0.00086	0.00149	0.00191
UnEmp	6406	0.00874	0.00902	0.02051	0.02015
NILF	66647	0.00187	0.00244	0.00415	0.00549
MAR	114367	0.00170	0.00159	0.00365	0.00337
DIV	18463	0.00551	0.00523	0.01134	0.01072