Single Stage Generalized Raking Weight Adjustment in the Current Population Survey*

Eric V. Slud[1,2], Christopher Grieves[3], Reid Rottach[3]
[1]U.S. Census Bureau, CSRM, 4600 Silver Hill Road, Washington, DC 20233
[2]Mathematics Department, University of Maryland, College Park, MD 20742
[3]U.S. Census Bureau, DSMD, 4600 Silver Hill Road, Washington, DC 20233

Abstract

This research concerns the adaptation to the Current Population Survey (CPS) of single-stage weight adjustment techniques developed in a 2010 Census research report by Slud and Thibaudeau. Those techniques involved weight optimization with respect to a loss function in the spirit of Deville and Särndal (1992, JASA), subject to population-control constraints, with additive penalty terms for discrepancies between weight-adjusted survey totals and corresponding known or base-weighted estimated totals for certain survey attributes, and with an additional nonlinear penalty term designed to force weights not to be too different from the design weights scaled to the population total. The novel elements of the current research include: defining several appropriate additive quadratic penalty terms corresponding to the current multistage CPS nonresponse adjustment; developing a methodology to define penalty multipliers by tracking properties of the current CPS weights across weighting stages; enforcing weight compression by a penalty term in place of the current CPS approach based on cell collapsing; and implementing the method on CPS data for detailed comparison with the weights as currently adjusted in CPS.

Key Words: calibration equations, optimization, linearized variance, loss function, population controls, quadratic programming, weight tracking.

*This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

1. Introduction

1.1 Notation and Assumptions

Consider a sample survey with a frame $\mathcal{U}$ from which a probability sample $\mathcal{S}$ is drawn according to a plan with known single and double inclusion probabilities π_i, π_{ij}, for $i, j \in \mathcal{U}$. Assume that the total $Y = t_y = \sum_{i \in \mathcal{U}} y_i$ of a scalar attribute is of primary interest, and that $(y_i, \mathbf{x}_i, \mathbf{z}_i \; i \in \mathcal{S})$ is (potentially) observable, i.e., the sample data include the auxiliary q-dimensional vector $\mathbf{z}_i$ and p-vector $\mathbf{x}_i$. This setting corresponds to the *InfoS* sampling framework of Särndal and Lundström (2005), with auxiliary data available at sample but not frame level.

In the present context, the vector of attributes $\mathbf{x}_i$ is composed of p_k dimensional subvectors $\mathbf{x}_i^{(k)}$ which are assumed to be used in adjusting survey weights through a succession of stages, with the subvector $\mathbf{x}_i^{(k)}$ used in the k'th stage, $k = 1, 2, \ldots, K$, where $\sum_{k=1}^{K} p_k = p$. The vector $\mathbf{z}_i = (z_{1i}, \ldots, z_{qi})$ of survey variables defined for each unit $i \in \mathcal{U}$ is used in final-stage population calibration. These vectors are generally observable only for survey responders, with the exception of the vector $\mathbf{x}_i^{(1)}$ observed for all sample units which is used in first-stage nonresponse adjustment.

Assume that each sampled individual in the survey decides independently whether or not to respond. Without loss of generality, denote by r_i for all $i \in \mathcal{U}$ the indicator which is 1 or 0 respectively if the i'th individual *would* or *would not* have responded if sampled, and assume that these random variables are independent of each other and of the sample selection mechanism. (This is the Oh and Scheuren (1983) *quasi-randomization* model. In some surveys this assumption could be applied only with 'individuals' replaced by households.) The observable data are now taken to be

$$\Big(y_i \cdot r_i, \; r_i, \; \mathbf{x}_i^{(1)}, \; r_i \cdot \mathbf{z}_i, \; (r_i \cdot \mathbf{x}_i^{(k)}, \;\; 2 \le k \le K), \quad i \in \mathcal{S} \Big)$$

No restriction other than positivity is placed on the probabilities

$$P(r_i = 1) \;=\; Er_i \;\equiv\; \rho_i$$

with which individual units respond.

At the data-collection stage, individuals $i \in \mathcal{S}$ are sampled with *design weights* $w_i^o \;=\; 1/\pi_i \;=\; 1/P(i \in \mathcal{S})$. These weights are altered to a final set of weights w_i, often in an elaborate series of stages. The premise of the present research is that in major surveys like the CPS there are three different ways in which weights are adjusted, which can all be viewed as constraints, either exact or approximate, on weighted survey totals. The first, which we call *hard constraints*, is typically imposed at the end of the weight-adjustment cycle as precise population controls in order to enforce conformity of published survey totals with those known and published from the best known source. That is, the final weights must exactly satisfy

$$\sum_{i \in \mathcal{S}} r_i \, w_i \, \mathbf{z}_i \;=\; t_{\mathbf{z}}^* \tag{1}$$

where the totals $t_{\mathbf{z}}^*$ are obtained from an external source like a (possibly updated) census and are assumed known. As a second form of adjustment, a series of balance equations

$$\frac{1}{N} \Big(\sum_{i \in \mathcal{S}} r_i \, w_i \, \mathbf{x}_i^{(k)} \;-\; t_{\mathbf{x}^{(k)}}^* \Big) \;\approx\; 0 \tag{2}$$

is generally imposed not on the final weights but as an equality on an earlier, intermediate set of adjusted weights, but which in effect result only in approximate equalities on the final weights. (Here $N = |\mathcal{U}|$ is the population size, assumed known.) The farther removed that stage k is from the final stage, the less one may regard exact equality in (2) as being important to the survey analysts. Therefore we refer to (2) as *soft constraints* on the final weights. The totals $t^*_{\mathbf{x}^{(k)}}$ that appear in (2) for $k \geq 2$ generally arise externally, either from a survey or updated census which is believed to be more accurate than the current survey, and the corresponding attributes $\mathbf{x}_i^{(k)}$ are generally observable only for responders to the survey (i.e., for indices $i \in \mathcal{S}$ for which $r_i = 1$. However, at a first *nonresponse adjustment* stage $k = 1$ the $\mathbf{x}_i^{(1)}$ variables are sometimes assumed known for all sampled units, and the totals $t^*_{\mathbf{x}^{(1)}}$ obtained as an estimate $\sum_{i\in\mathcal{S}} r_i\, w_i^o\, \mathbf{x}_i^{(1)}$ using design weights and data from the current survey. In the terminology of Särndal and Lundström (2005), the external census source of the $\mathbf{x}_i^{(k)}$ totals for $k \geq 2$ is called the *infoU* setting, while the internal estimated source of totals $t^*_{\mathbf{x}^{(1)}}$ is called *infoS*. Thus, we distinguish the $k = 1$ case of (2) as a third type of weight adjustment. From now on, we index the approximate balance equations as needed by the subscript k, e.g., (2_k).

To clarify an important element of the notation, in succeeding formulas we use the index i for units within the frame $\mathcal{U}$ or sample $\mathcal{S}$, k for the stage index for the soft controls, and j for components of one of the control-subsets, i.e. for components of vectors $\mathbf{x}_i^{(k)}$ or $\mathbf{z}_i$.

1.2 Single-stage Weight-Adjustment via Optimization

The exact constraints (1) and approximate constraints (2) fall far short of determining the set of final weights w_i. In large government surveys such as CPS or Survey of Income and Program Participation (SIPP), the sample size $n = |\mathcal{S}|$ is of the order 10^5 while the numbers q of controls are of the order 100–200 while the intermediate constraints or soft controls p number up to 1000–2000. Moreover, the survey design is put in place with the idea that final weights should be maintained as similar as possible to the design weights. Starting from the seminal paper of Deville and Särndal (1992), it has been known that ratio adjustment, raking and linear calibration can all be viewed as weight adjustment methods in which (hard) constraints are met while the weight vector $(w_i,\, i \in \mathcal{S})$ are determined as close as possible to $(w_i^o,\, i \in \mathcal{S})$ subject to a *loss function* $\sum_{i\in\mathcal{S}} r_i\, w_i^o\, G(w_i/w_i^o - 1)$.

As described in Kott (2006) and Slud and Thibaudeau (2010), there is a stream of papers from 1992 to the present in which final-stage survey weights w_i are determined by optimizing a loss function, possibly including a penalty term enforcing that the weight ratios w_i/w_i^o never or rarely depart from a bounded interval (a, b) containing 1, subject to constraints (1). Many of these papers, from Deville and Särndal (1992) up through Slud and Thibaudeau (2010), show that under some superpopulation regularity conditions guaranteeing that the response propensities ρ_i can be consistently estimated and that the great majority of changes from w_i^o to w_i are quite small, the survey estimators

$$\hat{t}_y/N \;=\; N^{-1} \sum_{i\in\mathcal{S}} r_i\, w_i\, y_i \tag{3}$$

are design-consistent for t_y/N. Many of these same papers establish asymptotic normality of $(\sqrt{n}/N)\,(\hat{t}_y - t_y)$ and provide asymptotic variance formulas based

on joint inclusion probabilities based on Taylor linearization, i.e., based on a proof that $(\sqrt{n}/N)(\hat{t}_y - t_y)$ differs asymptotically negligibly from a sum of linear weighted (Horvitz-Thompson) estimators.

A key feature of the stream of survey methodology papers cited by Slud and Thibaudeau (2010) and Kott (2006) is that any two of the three goals of nonresponse adjustment, population controls, and weight compression (i.e., keeping weight-ratios w_i/w_i^o to a prescribed bounded interval) had already been seen to be achievable simultaneously in a single optimization step, and Slud and Thibaudeau showed that one can actually accomplish all three. The method of Slud and Thibaudeau did allow the choice of a tuning parameter to control how closely the balance relation used in nonresponse adjustment would be satisfied by the final adjusted weights. However, their method did not allow the possibility contemplated here that there might be a series of approximate equalities (2_k) to be satisfied simultaneously to differing degrees which might be chosen by the survey analyst or survey client.

The goal of the present research is to show how a linear-calibration loss-function can be combined with quadratic penalty terms quantifying inequality in (2_k) and possibly also a penalty term enforcing that $L \le w_i/w_i^o \le U$ can be optimized computably in the context of the CPS. The objective function to be minimized over $\underline{\mathbf{w}} = \{w_i : i \in \mathcal{S},\ r_i = 1\}$ subject to (1) is $J(\underline{\mathbf{w}}) \equiv$

$$\sum_{i\in\mathcal{S}} r_i \frac{(w_i - w_i^o)^2}{2\, w_i^o} + \sum_{k=1}^{K} \frac{\alpha_k}{2} \Big\| \sum_{i\in\mathcal{S}} r_i\, w_i\, \mathbf{x}_i^{(k)} - t_k^* \Big\|^2 + \sum_{i\in\mathcal{S}} r_i\, w_i^o\, Q(\frac{w_i}{w_i^o}) \qquad (4)$$

where the notation t_k^* is a shortened form of the vector notation $t^*_{\mathbf{x}^{(k)}} = (t^*_{x_j^{(k)}},$ $1 \le j \le p_k)$, and the tuning constants $\alpha_k \equiv a_k/\|t_k^*\|_1$ are to be chosen by the survey analyst. The norm-square $\|\cdot\|^2$ in the middle term of (4) denotes the Euclidean norm-square, or sum of squared vector entries. Scaling the α_k coefficients down by the factor $\|t_k^*\|_1 = \sum_{j=1}^{p_k} |t^*_{x_j^{(k)}}|$ makes the aggregate soft-constraint penalty terms roughly of the same order as the weight-change loss function value, when the constraint totals t_k^* are properly specified. The nonlinear function Q might be chosen to take the following form (Deville and Särndal 1992, 'Case 6' loss function). Let $L < c_1 < 1 < c_2 < U$ and positive constants A_1, A_2 be fixed. Then $Q(x)$ is 0 on $[c_1, c_2]$ and infinite on $[0, L] \cup [U, \infty)$ and is convex throughout (L, U) from the definition

$$Q(x) \;=\; A_1\, I_{[x\le c_1]}\, \frac{(c_1 - x)^2}{x - L} \;+\; A_2\, I_{[x\ge c_2]}\, \frac{(x - c_2)^2}{U - x} \qquad (5)$$

Objective functions for weight adjustment with penalty terms like the soft-constraint terms with coefficients α_k in (4) have previously been considered in a simplified survey calibration setting by Fuller (2009, p. 164) and in the somewhat different setting of "Bayesian benchmarking" by Datta et al. (2011).

For simplicity of notation, from now on we adopt the convention that adjusted weights $w_i = r_i\, w_i$ are nonzero only for responding sampled units, and denote by $\mathcal{R} = \{\, i \in \mathcal{S} : \quad r_i = 1 \,\}$ the set of such units, and (somewhat unusually) let $n = |\mathcal{R}|$ be the number of responding (rather than of sampled) units. In addition, in order to allow weight-changes from initial (design) weights w_i^o to final adjusted weights w_i to be as small as possible, we adopt the (slightly unusual) convention that the design weights are already calibrated to the known population total N, so that $\sum_{i\in\mathcal{S}} w_i^o = N$. (This is done by multiplying each design weight by the scalar ratio $N/\sum_{i\in\mathcal{S}} w_i^o$, which can for some large surveys differ from 1 by $10 - 20\%$.)

Finally, some of the same survey variables $x_{j,i}^{(k)}$ might arise with more than one distinct (k, j) combination, and it might make sense to remove such redundancy by retaining the corresponding balance equation index combinations (k, j) for only the one with the largest k. However, we do not take this step in the data illustrations of Sec. 4 below.

1.3 Weight Adjustment and Quadratic Programming

Fuller (2002, 2009 p. 164) remarked that weight compression to satisfy linear inequalities $L \le w_i/w_i^o \le U$ can be accomplished by quadratic programming with linear equality and inequality constraints, under any objective function (4) omitting the Q term. That is, the problem to be solved becomes

$$\min_{\mathbf{w}} \left(\frac{1}{2}\,\mathbf{w}'H\mathbf{w} - \mathbf{d}'\mathbf{w}\right) \quad \text{subject to} \quad Z'\mathbf{w} = t_{\mathbf{z}}^* \;,\; L\mathbf{w}^o \le \mathbf{w} \le U\mathbf{w}^o \qquad (6)$$

where Z' is the $q \times n$ matrix with i'th column $\mathbf{z}_i$, $\mathbf{w}$ is the vector of n nonzero weights, the final inequalities in (6) are interpreted componentwise,

$$H \;=\; \mathrm{diag}(\{1/w_i^o\}_{i=1}^n) \;+\; \sum_{k=1}^{n} \alpha_k\, X_k\, X_k' \;\;, \quad \mathbf{d} \;=\; \mathbf{1} \;+\; \sum_{k=1}^{K} \alpha_k\, X_k\, t_{\mathbf{x}^{(k)}}^*$$

X_k' is the $p_k \times n$ matrix with i'th column $\mathbf{x}_i^{(k)}$, and $\mathbf{1}$ is the n-vector of 1's.

For the CPS weight-adjustment application that we have in mind, the total number $n = \sum_{i \in \mathcal{S}} r_i$ of responders is the number of rows of the square quadratic-form matrix being optimized, and is of the order of 130,000, while the numbers of hard and soft contraints will be roughly $q \le 200$, $p \in [1000, 2000]$. So these quadratic programming problems will be very large.

A great deal is known about the solution of quadratic programming problems with linear equality and inequality constraints. The topic is admirably treated in the book of Nocedal and Wright (1999, Ch. 16). However, not all methods are suitable for extremely large problems, in which the best methods are determined by the special structure of the problem. Two distinguishing features of the problems generally arising in survey weight adjustment are the following:

(i) The matrix H is the sum of a diagonal matrix of reciprocal design-weights and a small finite linear combination of matrices $X_k X_k'$ such that for all k, each row of the $n \times p_k$ matrix X_k contains only one nonzero element.

(ii) Problem (6) with all matrices X_k replaced by 0 has a *pure-calibration* solution for which all of the inequality constraints hold with strict inequality.

Property (i) reflects the fact that nonresponse ratio adjustment or raking, at each of the one or more adjustment stages of large surveys, is generally based on one or two partitions of the population into cells. Because of this property, the matrix H may be somewhat sparse. (In the CPS example below, a little less than 10% of the n^2 entries of H are positive.) Property (ii) is a reflection of the guidance – given in Fuller (2009) and other expositions of calibration in surveys – that the set of columns (of the matrix Z) used in calibration not be allowed to be large enough for weight ratios w_i/w_i^o to vary by large factors, since such variation leads to undesirably large variances for weighted survey totals.

Several methods of calibration-based weight adjustment in the survey literature can be viewed as approximate quadratic-programming solution methods.

(a) The default method of weight adjustment in surveys (cf. Särndal and Lundström 2005) is to perform a cell-based ratio or raking adjustment, providing an intermediate set of weights to which a separate stage of calibration to population controls is applied. This corresponds to the case $K = 1$ in (6), but without an explicit penalty term with $\alpha_1 > 0$ in H. Thus, the nonresponse-adjustment balance relations are not assumed to hold for the final weights $\mathbf{w}$.

(b) Many large national surveys, such as the CPS, SIPP, and American Community Survey (ACS) in the US, adjust weights in a series of nonresponse adjustment stages, each of which is based on cells defined through partitions of the frame population, before a final step of calibration to population controls. Here also, only the balance equation (1) for the last calibration step is treated as an exact constraint for the final weights, and the balance equations for earlier nonresponse adjustments do not hold exactly for the final weights, and so correspond to our soft constraints. Nevertheless, when performed in successive stages by ratio adjustment and raking, the calibrated adjusted weights can be viewed as approximate solujtions to the quadratic programming problem (6).

(c) With the quadratic-programming reformulation (6) given above, in surveys with n not too large, it will often be possible to solve directly for the adjusted and calibrated weights using standard quadratic programming software like the function `quadprog` in `R`. However, problems with large n require a different strategy. One approach, which has apparently been implemented successfully in large ($n > 10^4$) problems at the A.C. Nielsen company (Daehmen 2013), is to remove the first loss-function term $\sum_{i\in\mathcal{R}} (w_i - w_i^o)^2/w_i^o$ from the objective function and constrain it to lie between fixed bounds (b_1, b_2) chosen by the analyst. Then the remaining soft-penalty quadratic form $\tilde{H} = \sum_{k=1}^{K} \alpha_k \, \mathbf{w}' \, X_k X_k' \mathbf{w}$ can be re-expressed as a $p \times p$ quadratic form in the new variables $\omega_j \equiv \sum_{i\in\mathcal{R}} w_i \mathbf{x}_{i,j}^{(k)}$ together with an additional set of $n - p$ variables w_i appearing in the weight-ratio constraints. The smaller quadratic programming problem so obtained can be solved by off-the-shelf, or slightly modified, versions of quadratic programming codes like `ipop` in `R`.

(d) We have begun experimenting with *gradient-projection* quadratic programming codes to solve (6) in survey settings with large n where the special conditions (i)–(ii) above hold, as in CPS. The *gradient projection algorithm* for quadratic programming with some linear equality constraints and purely upper- and lower-bound inequality constraints on the variables w_i, is beautifully explained in Section 16.6 of the book of Nocedal and Wright (1999), and we are implementing it in `R` code. In doing this, we can benefit in two ways from the very special form of the soft-calibration matrices X_k. First, the fact that H differs from a diagonal matrix by a matrix of the relatively low rank p allows inversion of H even though n is large. Moreover, dramatic computational speedups stem from the fact that the rows of X_k are dummy-variables labeling which element $j(i) \in \{1, \ldots, p_k\}$ of the population partition-cells the sampled units i fall into. For example, right-multiplication by vectors $\mathbf{v} \in \mathbf{R}^{p_k}$ is simply expressed in the form $(X_k \mathbf{v})_i = h_i^{(k)} \, v_{j(i)}$, where generally $h_i^{(k)} = 1$, but in the CPS setting as described below in Sec. 4.1, the $h_i^{(1)}$ value is the reciprocal of the number of persons in the household of person i).

In our present research, we experimented also with a simpler algorithm for minimization of the original objective function (4). A resubstitution-algorithm described in Section 3 converts this convex optimization problem to a tractable iterative search (for the Lagrange multiplier λ of the hard constraints) in a q-dimensional space, and was implemented in some of our computational illustrations in Section 4.

2. Weight Tracking and Penalty Factors α_k

One observation motivating the present research is that the balance equations $\sum_{i\in\mathcal{S}} w_i\, r_i\, \mathbf{x}_i^{(k)} = t_k^*$ which are imposed on weights w_i at the k'th stage of a multi-stage adjustment procedure are often found to hold only very imprecisely at the final stage of weight-adjustment. There have been some published studies tracking the changes of weights across stages of adjustment, such as Dufour et al. (2001).

The idea of tracking weight-changes by the mean-square discrepancy between the left-hand sides of such balance equations as weights w_i change from stage to stage, provides us with an approach to defining the soft-control penalty coefficients α_k in (4) . Indeed, we choose the coefficients α_k so that, in terms of a final set of weights w_i^F generated by multistage adjustment, the quantities $\alpha_k \cdot \| \sum_{i\in\mathcal{S}} r_i\, w_i^F\, \mathbf{x}_i^{(k)}\ t_k^*\|^2$ are roughly equated across k. That is, based on the final weights w_i^F obtained by multi-stage adjustment in current practice, the penalty coefficients α_k used to define a new objective-function (4) are chosen proportional to $\|t_k^*\|_1 \,/\, \| \sum_{i\in\mathcal{S}} r_i\, w_i^F\, \mathbf{x}_i^{(k)} - t_k^*\|^2$.

In Section 4.2 below, we illustrate this choice using so-called *second-stage weights* defined (CPS documentation 2006) for CPS monthly survey data. In that section on CPS numerical results, we will also discuss the choice of the overall tuning constant α of proportionality, for which the rough guideline is that the weight-change loss function (the first summation in (4)) should be of the order of 10 times as large as the total soft-control penalty terms (the middle summations involving α_k in (4)).

3. Resubstitution Approach to Optimizing (4)

The objective function (4) is to be minimized over $\mathbf{w} = \{w_i\}_{i\in\mathcal{R}}$ subject to (1). This variable set is generally large for a survey like CPS, but as we will see, a Lagrange-multiplier form of the optimization, i.e. a minimization of the *Lagragian*, defined as the objective function minus $\lambda' (\sum_{i\in\mathcal{R}} w_i\, \mathbf{z}_i - t_{\mathbf{z}}^*)$, in which the unknown q-vector λ is determined from the constraint equation (1), can be performed as an iterative search over the q-dimensional Lagrange multiplier vector λ. That is, the number q of hard constraints is much more important than the number $p = p_1 + \cdots + p_K$ of soft constraints in determining the convergence and numerical stability of the optimization method. The specific algorithmic approach to optimization is as follows.

First, for all $i \in \mathcal{R}$, the partial derivative of the Lagrangian with respect to w_i is set to 0 in the equation

$$\frac{w_i - w_i^o}{w_i^o} + \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{p_k} x_{j,i}^{(k)} \Big(\sum_{l\in\mathcal{R}} w_l\, x_{j,l}^{(k)} - t_k^*\Big) + Q'\Big(\frac{w_i}{w_i^o}\Big) - \lambda'\, \mathbf{z}_i = 0$$

Summing both sides of this equation multiplied by $w_i^o\, \mathbf{z}_i$, and applying (1), imme-

diately gives

$$t^*_{\mathbf{z}} - \hat{t}^o_{\mathbf{z}} \;=\; M_{\mathbf{z}}\,\lambda \;-\; \sum_{i\in\mathcal{R}} w_i^o\,\mathbf{z}_i\,Q'(\frac{w_i}{w_i^o}) \;-\; \sum_{k=1}^{K}\sum_{j=1}^{p_k}\sum_{i\in\mathcal{R}} \alpha_k\,w_i^o\,x_{j,i}^{(k)}\,\mathbf{z}_i\,(\hat{t}_{x_j^{(k)}} \;-\; t_k^*) \qquad (7)$$

where the q-vector $\hat{t}^o_z$ and the $q\times q$ matrix $M_{\mathbf{z}}$ are defined in terms of the initial weights by

$$\hat{t}^o_z \;\equiv\; \hat{t}^{(0)}_z \;=\; \sum_{i\in\mathcal{R}} w_i^o\,\mathbf{z}_i \quad , \qquad M_{\mathbf{z}} \;\equiv\; \sum_{i\in\mathcal{R}} w_i^o\,\mathbf{z}_i\,\mathbf{z}_i'$$

Now we define an iterative resubstitution sequence for the preceding equations, as follows. First let $w_i^{(0)} = w_i^o$ be the initial settings in a sequence of values $w_i^{(m)}$, $m \geq 0$. For ease of notation, inductively define via (7) :

$$\lambda^{(m)} \;=\; M_{\mathbf{z}}^{-1}\left\{t^*_{\mathbf{z}} - \hat{t}^{(0)}_{\mathbf{z}} + \sum_{i\in\mathcal{R}} w_i^o\,\mathbf{z}_i\,Q'(\frac{w_i^{(m)}}{w_i^o}) + \sum_{k=1}^{K}\alpha_k \sum_{j=1}^{p_k} \hat{t}^{(0)}_{x_j^{(k)}\,\mathbf{z}}\,(\hat{t}^{(m)}_{x_j^{(k)}} - t_k^*)\right\} \quad (8)$$

for $m \geq 0$, where

$$\hat{t}^{(m)}_{x_j^{(k)}} \;\equiv\; \sum_{i\in\mathcal{R}} w_i^{(m)}\,x_{j,i}^{(k)} \quad , \qquad \hat{t}^{(m)}_{x_j^{(k)}\,\mathbf{z}} \;\equiv\; \sum_{i\in\mathcal{R}} w_i^{(b)}\,x_{j,i}^{(k)}\,\mathbf{z}_i$$

The inductively updated weights $w_i^{(m+1)}$ are then defined for $i \in \mathcal{R}$ by:

$$w_i^{(m+1)} \;=\; w_i^o\left\{\,1 \;+\; (\lambda^{(m)})'\,\mathbf{z}_i \;-\; Q'(\frac{w_i^{(m)}}{w_i^o}) \;-\; \sum_{k,j}\alpha_k\,x_{j,i}^{(k)}\,(\hat{t}^{(m)}_{x_j^{(k)}} - t_k^*)\,\right\} \qquad (9)$$

In implementing the steps (9), it turns out to be necessary to make a modification. Ideally, all of the newly defined $w_i^{(m+1)}$ weights will fall inside the permissible range (L, U) used in defining the penalty function Q for extreme weight ratios. But in case some of the quantities $w_i^{(m+1)}$ defined by (9) would fall outside the range $[L+\epsilon, U-\epsilon]$, they are replaced by

$$\min\left\{U-\epsilon\,,\;\max\{\,w_i^{(m)},\;L+\epsilon\,\}\,\right\}$$

where $\epsilon > 0$ is a fixed constant chosen as an input to the optimization code. With this modification, the hard-constraints (1) would be violated, so a further linear calibration of these weights must be applied to enforce those constraints at each iteration of the algorithm. These modifications turn out to have essentially no effect in convergent instances of the resubstitution algorithm, and in fact are an effective diagnostic for the failure of the algorithm when the constant factor C for the soft-constraint penalty terms is taken too large.

4. Numerical Results on CPS Data

We illustrate the objective-function optimization and results, and the tracking of weights in a multi-stage weight adjustment framework, using CPS data.

4.1 CPS Background

In the CPS data from January 2012, there were 164386 sampled persons, and 131978 responders, corresponding to the standard monthly total of roughly 72,000 sample households. We treat as 'base' weights the household control weights (`hcwwgt` on CPS files) derived after a preliminary re-scaling to force the weights of sampled and responding units to add to the same control population size (307,567,803) used later in calibrating second-stage weights. The variables $\mathbf{x}_j^{(k)}$ and z_j used to define the soft and hard controls were as follows. The Non-Interview stage (`NIntv`, p. 10-3 of CPS documentation 2006) is based upon a partition (`niclcode`) of the population into 202 cells defined by partitioning the entire sampled population into PSU's by metropolitan (central and non-central city) and non-metropolitan subsets. Since the adjustment done at this stage is a household-level adjustment, the variables $x_{j,i}^{(1)}$ entering balance equations are dummy indicators of `NIntv` partition cell j divided by the number of persons in the household containing person i. The variables $x_{j,i}^{(2)}$ appearing in balance relations for the National Coverage adjustment stage (`NatCov`) are dummy variables for 180 cells of a partition of the sample defined by cross-classification according to sex and age category, with different age groupings used within 6 different Race/Hispanic subsets. The State Coverage stage (`StCov`) also uses dummy indicators $x_{j,i}^{(3)}$, in this case for a partition of the population into 515 cells by sex, age, Race/Hispanic subgroup, and 53 state indicators (including District of Columbia, with California and New York each split into major city and balance of state). Finally, in the Second-stage adjustments (`2ndStg`), the population is partitioned three different ways. The 337 soft-control variables are broken down into three stages ($k = 4, 5, 6$) corresponding to the respectively 159, 52, and 126 dummy variables from the three partitions, which are used as variables $x_{j,i}^{(k)}$, for $k = 4, 5, 6$ in balance equations for adjustment. The 1234 variables $x_j^{(k)}$ mentioned so far are used for soft-calibration and are intended to cause final weights to reproduce approximately the updated-census population estimates for Tables 10-1 to 10-4 in the CPS (2006) documentation. In addition, CPS second-stage weights are raked to 79 non-redundant hard constraints defined by 53 state-category totals and 26 other Race/Hispanic by age-group by sex cells within the population aged 16 and older. We treat these constraints as exact linear balance equations defined through variables $z_{j,i}$.

CPS actually goes beyond second-stage weight adjustments to combine weights from monthly data to provide a 'composite' weight taking account of different numbers of persons' months in sample within the CPS rotating panel design, but we ignore that further complication in the present research.

4.2 Tracking Weight Adjustments in CPS Practice

It was mentioned above that the choice of soft-constraint penalty coefficients α_k in our objective function (4) should be informed by calculations of the extent to which early-stage weight-adjustment balance equations fail to hold with later- or final-stage weights. We do such a calculation on CPS data for the root-mean-squares $\texttt{RMS}_k$ of balance-equation discrepancies for $k = 1, \ldots, 3$, i.e. the square roots of

$$\texttt{RMS}_k^2 \equiv \frac{1}{p_k} \| \hat{t}_{\mathbf{x}^{(k)}} - t_k^* \|^2 = \frac{1}{p_k} \sum_{j=1}^{p_k} \Big(\sum_{i \in \mathcal{R}} w_i \, x_{j,i}^{(k)} - t_k^* \Big)^2 \quad , \quad k = 1, 2, 3$$

Table 1: Root-mean-square soft-control balance equation discrepancies for Jan. 2012 CPS data with CPS-calculated second-stage and composite weights.

	$\texttt{RMS}_1$	$\texttt{RMS}_2$	$\texttt{RMS}_3$	$\texttt{RMS}_4$
`psswgt`	98155	20813	4740	0
`pwcmpwgt`	198635	603232	349491	1462551
	$p_1 = 202$	$p_2 = 180$	$p_3 = 515$	$p_4 = 337$

and for a fourth stage in which the 337 second-stage adjustment variables are lumped together to create a root-mean-square equal to the square root of

$$\texttt{RMS}_4^2 = (p_4 + p_5 + p_6)^{-1} \sum_{k=4}^{6} \|\hat{t}_{\mathbf{x}^{(k)}} - t_k^*\|^2$$

We do this first for w_i equal to the second-stage CPS weights `pwsswgt` calculated in CPS for the January 2012 data, and then again for the even later-stage 'composite' weights `pwcmpwgt` which make use of the month-in-sample information for persons in sample. The results are displayed in Table 1. Additional calculations (not shown) confirm that the `psswgt` results are very stable across months of CPS data. The 0 value for $\texttt{RMS}_4$ with second-stage weights arises because these weights are actually raked to the totals t_4^*, t_5^*, and t_6^*, and many of the aggregated cell totals `StCov` are actually the same as for `2ndStg`, so the $\texttt{RMS}_3$ value for `psswgt` is also in some sense artificially small. The behavior of the $\texttt{RMS}_k$ values with composite weights do not show the same kind of steady decrease as with second-stage weights. Since our concern here is how to define soft-control penalties for second-stage weights, we suppose that apart from the coincidence of aggregate-level cells from the stages 3 and 4 with fixed calibration cells for `psswgt` there might be decay of $\texttt{RMS}_k$ over k by a factor of 5 to 10, and we choose penalty coefficients $(a_1, \ldots, a_4) = C \cdot (1, 2, 4, 8)$ in our following optimization calculations, with $a_4 = a_5 = a_6$ reflecting that the $k = 4, 5, 6$ adjustment 'stages' together constitute the single second-stage adjustment in CPS. The a_k sequence for each run is therefore determined by the value $a_1 = C$, and recall that $\alpha_k = a_k / \|t_k^*\|_1$.

4.2.1 Characteristics of Optimized CPS 2nd Stage Weights

We implemented the algorithm of Section 3 using the CPS January 2012 data, based on $K = 6$ stages of soft controls with $C = 1.25$ and a total of 1234 variables, and with 79 hard controls. We set the Q penalty-function parameters as before, at $c_1 = .5$, $c_2 = 2$, $L = .2$, $U = 5$, and $A_1 = A_2 = 20$. We used as initial weights the CPS housing control weights `hccwgt` scaled (roughly by the factor 1.27) so that when summed over all 131978 CPS responding persons they yield the US control population total of $307,567,803$.

For comparison, we calculated for $C = 500$ and 2500 the minimum of the quadratic form in (6) without regard to any bounds on weight-ratios w_i/w_i^o, respectively denoting by `Qmin1` and `Qmin2` the two optimized sets of weights.

In Table 2, we display the objective-function components

$$\texttt{Loss} = \sum_{i \in \mathcal{R}} (w_i - w_i^o)^2/(2w_i^o) \quad , \qquad \texttt{Soft.Pen} = \sum_{k=1}^{K} \frac{\alpha_k}{2} \|\hat{t}_{x^{(k)}} - t_k^*\|^2$$

along with `range` of weight-ratios w_i/w_i^o, and – since the soft-penalty terms are a little difficult to interpret directly – we also provide the root-mean-square discrepancies RMS_k. balance-equation discrepancies. These quantities are displayed for the initial and second-stage weights together with the three sets optimized by resubstitution and by the minimum of the linearly constrained quadratic form in (6) without weight-ratio bounds.

It turned out that the resubstitution-algorithm runs (the one shown and many others not shown) converged only when the penalty factor C was sufficiently small that the weight-ratio bounds $L\mathbf{w}_i^o \le \mathbf{w}_i \le U\mathbf{w}_i^o$ had no effect. This was an unanticipated deficiency of the resubstitution method, which we think is due to the artificial re-calibration needed to restore the hard constraints when the resubstitution steps resulted in weights w_i set equal to Lw_i^o or Uw_i^o. In fact, the resubstitution-generated weights were essentially identical to the weights fitted to (6) with the Q penalty omitted when $C = 1.25$

Table 2: Comparison of features of design weights, CPS second-stage weights, two sets of weights obtained by linearly constrained quadratic minimizer without weight compression, and one set generated by the method of Sec. 3, based on the CPS Jan. 2012 data. `Loss`, `Soft.Pen` and $\mathtt{RMS}_k$ values given in units of 10^5. Column C denotes weight-factor multiplying $\mathbf{a}$ coefficients $(1, 2, 4, 8, 8, 8)$ used in optimization.All `Soft.Pen` terms were calculated with $a_1 = 1.25$.

Weights	C	`Loss`	`Soft.Pen`	`range`	$\mathtt{RMS}_1$	$\mathtt{RMS}_2$	$\mathtt{RMS}_3$	$\mathtt{RMS}_4$
`init.`	$*$	0.00	3.01	(1,1)	0.96	2.01	0.67	2.56
`Resub`	1.25	12.42	0.81	(0.80,1.58)	1.02	1.08	0.46	0.95
`Qmin1`	500	35.50	.05	(0.26,2.27)	0.59	0.15	0.14	0.10
`Qmin2`	2500	71.36	.01	(0.01,2.68)	0.29	0.05	0.05	0.04
`2nd-St`	$*$	68.37	0.11	(0.41,4.01)	0.98	0.21	0.04	0.00

In the weight-optimization runs `Qmin1` and `Qmin2` summarized in the Table, the 79 fitted Lagrange-multipliers λ_j respectively fall in the ranges $(-0.32, 0.43)$ and $(-0.51, 0.65)$. Generally speaking, the large-sample theory of Deville and Särndal (1992), developed slightly further in Slud and Thibaudeau (2010), requires that the Lagrange multipliers fall in a small neighborhood of 0 (which does not vary with growing superpopulation and sample size) in order that the weights minimizing (4) subject to (1) produce design-consistent weighted survey totals. So the slightly larger λ_j's under $C = 2500$ indicate that C cannot be taken much larger than that if design-consistency is to remain important. Both optimized weight-sets `Qmin1` and `Qmin2` contained some small weights, for example, respectively 4 and 590 (out of about 132,000) falling below 1/3.

By comparison of the objective-function components in the optimized-weight cases with those for the CPS initial and second-stage weights, it becomes clear that optimization allows one to trade off the spread of allowed weights against the desired degree of agreement between weighted estimates of soft-control totals and their fixed targets. The resubstitution-generated weights remain close to the design weights but allow substantial reductions in $\mathtt{RMS}_k$ values for $k \ge 2$ by comparison with the design weights. The `Qmin1` and `Qmin2` weights allow much broader departures from the design weights but achieve remarkable reductions in $\mathtt{RMS}_k$, down to levels below those of the second-stage CPS weights, especially in view of their much narrow ranges of weight-ratios w_i/w_i^o as compared with the second-stage weights. Further progress

in optimization technique for (6) will likely provide similar soft-control balance equation discrepancies with still narrower ranges of weight ratios. If hard-control constrained minima of (4) also turn out to be accurately and rapidly computable in the CPS setting, then the use of convex penalty-functions Q in place of constant upper and lower weight-ratio bounds promises a still more favorable distribution of weight ratios without appreciable enlargement of soft-control discrepancies.

5. Conclusions and Future Research

The feasibility of the quadratic-programming solution in (d) of Sec. 1.3 makes us optimistic that in many large surveys, single-stage nonresponse and hard-calibration weight adjustments incorporating quadratic soft-calibration penalty terms can be solved rapidly and accurately and can become part of the Census Bureau's repertoire of weight-adjustment methods. The clear benefits of such optimized weights will include the ease of documentation of weight adjustments, the customization of soft-control penalties providing weight-adjustments with small (but nonzero) soft-control discrepancies where they are most desired, and also the availability of linearized analytical variance estimates, along the lines of those developed in Thibaudeau and Slud (2010).

We plan to continue our research in this area, in the following directions:

- `Improved optimization technique` As mentioned in paragraph (d) within Sec. 1.3, we have already begun to develop gradient-projection techniques of weight-optimization within (6), which can be tailored to the special features (i)-(ii) of quadratic programs arising in large national surveys. Further research is also needed into the development of numerical optimization code for (4) subject to (1). Due to the Q penalty terms, this is a convex and not a quadratic programming problem, but it may be that starting from the solutions to (6) could make the further optimization much easier.

- `Linearized variance estimates` Design-based superpopulation asymptotic theory is available in this setting under regularity conditions, as in Deville and Särndal (1992) and Slud and Thibaudeau (2010), establishing that when all of the soft- and hard-calibration totals (apart fron the internal *InfoS* controls used in the $k = 1$ stage for nonresponse adjustment) are correct, the converged Lagrange multipliers $\lambda^{(b)}$ and centered weighted survey estimators $\sum_{i\in\mathcal{R}} w_i^{(b)} y_i$, will be $O_P(N/\sqrt{n})$ and asymptotically normally distributed. The resulting linearized variance estimators will provide a useful alternative to the replication-based variance estimation methodology currently used for CPS and other major surveys by the Census Bureau. However, as in the other cited references, the asymptotic theory relies on the asymptotic smallness of Lagrange multipliers λ and weight-changes $w_i/w_i^o - 1$, a requirement which seems to fail in many real applications where adjustments change weights markedly. If the large weight-changes happen rarely enough, then it should be possible to show as in Thibaudeau and Slud (2010) that design-consistency of survey-weighted totals is still possible.

- Future detailed comparative studies are needed to evaluate the new optimization-based weight adjustments in CPS and resulting survey estimates versus the current, standard method. One ultimate goal is a detailed cross-classified

comparisons over subdomains of weighted totals of the important employment-related attributes in CPS data.

6. References

CPS (Oct. 2006) *Design and Methodology: Current Population Survey*, Technical Paper 66, `http://www.bls.census.gov/cps/tp/tp66.htm`

Daehmen, Ludo (2013), *personal communication.*

Datta, G., Ghosh, M., Steorts, R. and Maples, J. (2011) *Bayesian benchmarking with applications to small area estimation.* Test, 20, 574-588.

Deville, J.-C. and Särndal, C.-E. (1992), *Calibration estimators in survey sampling.* Jour. Amer. Statist. Assoc., 87, 376-382.

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

Fuller, W. (2002), *Regression estimation for survey samples.* Survey Methodology, 28, 5-23.

Fuller, W. (2009) *Sampling Statistics.* Hoboken: John Wiley.

Kott, P. (2006), *Using calibration weighting to adjust for nonresponse and coverage errors.* Survey Methodology, 32, 133-142.

Nocedal, J. and Wright, S. (1999) *Numerical Optimization.* Springer: New York.

Oh, H. and Scheuren, F. (1983) *Weighting adjustment for unit nonresponse.* In: Incomplete Data in Sample Surveys, vol. 2, Eds. Madow, W., Olkin, I. and Rubin, D. New York: Academic Press, 143-184.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-07-0, `http://www.R-project.org`.

Särndal, C.-E. and Lundström, S. (2005) *Estimation in Surveys with Non-response.* Wiley: Chichester.

Slud, E. and Thibaudeau, Y. (2010), *Simultaneous Calibration and Nonresponse Adjustment.* Census Bureau CSRM Research Report, RRS2010/03.