Variance Estimation Adjusted for Weight Calibration via the Generalized Exponential Model with Application to the National Household Survey on Drug Abuse

A. K. Vaish, H. Gordek and A. C. Singh, Research Triangle Institute
A. K. Vaish, P. O. Box 12194, Research Triangle Park, NC 27709 (email: avaish@rti.org)

Key Words: Estimating equations, Nonresponse, Poststratification, Sandwich variance, Taylor linearization

1 Introduction

In this paper we study the impact of weight adjustment factors for nonresponse (NR) and poststratification (PS) on the variance of the calibrated sample estimate when adjustment factors are modeled via the generalized exponential model (GEM) of Folsom and Singh (2000) with suitable predictors and bounding restrictions on the adjustment factors. Using the bias corrected estimating function (BCEF) approach of Singh and Folsom (2000), the estimation problem can be cast in the form of estimating equations, which in turn can be linearized to obtain sandwich-type variance estimate of the calibrated estimator. The method is applied to the 1999 National Household Survey on Drug Abuse (NHSDA), and numerical results comparing variance estimates with and without adjustments are presented. We use the following notations

U: Finite population of size N
s^*: Selected sample of size n^*
s: Subsample of respondents of size n
y: Study or outcome variables
x: Nonresponse (NR) predictor variables
α: Model parameters associated with x variables
z: Poststratification (PS) variables
β: Model parameters associated with z variables
T_v: Population total for an arbitrary variable v
$\hat{\theta}$: An estimator of θ
d: Design weight (inverse sample inclusion probability)

Suppose we want to estimate some population parameter, say $T_y = \sum_{k\in U} y_k$. An estimator of T_y is given by $\hat{T}_y = \sum_{k\in s} y_k\ d_k$. Generally, design weights are adjusted to reduce NR bias; and appropriate PS control totals are used for variance and coverage-bias reduction purposes, see Section 6 for a list of covariates for NR and PS adjustments. Let us assume that $f_k(\alpha)$ and $g_k(\beta)$ denote NR and PS adjustment factors, respectively.

In the presence of NR and PS adjustment factors, the estimator takes the following form:

$$\hat{T}_y(\alpha,\beta) = \sum_{k\in s} y_k\ d_k\ f_k(\alpha)\, g_k(\beta) \qquad (1)$$

where α and β can be estimated using well known techniques such as raking ratio, exponential or logistic regression or by the new GEM technique developed by Folsom and Singh. They showed that under $p\xi_1\xi_2$ model the estimator (1) is unbiased and the calibrated estimator, $\hat{T}_y(\hat{\alpha},\hat{\beta})$, is consistent where p denotes sampling design, ξ_1 and ξ_2 denote two independent super population models as defined below

$$E_{\xi_1}(\eta_1) = g_k^{-1}(\beta) \text{ for each } k \in U$$

where η_1= Number of times the kth unit in U appears in the sampling frame, and

$$E_{\xi_2}(\eta_2) = f_k^{-1}(\alpha) \text{ for each } k \in U$$

where η_2= 1 if the kth unit in U responds and zero otherwise.

Variance Estimation of $\hat{T}_y(\hat{\alpha},\hat{\beta})$

In estimating the variance of $\hat{T}_y(\hat{\alpha},\hat{\beta})$, it is often difficult in practice to account for the variability present due to the estimation of α and β. If the estimated values of α and β can be treated as fixed, one can use standard methods in sampling theory to obtain the variance as if the final calibrated weights, $w_k = d_k\, f_k(\hat{\alpha})\, g_k(\hat{\beta})$, were design weights. Clearly, $\hat{\alpha}$ and $\hat{\beta}$ are not fixed since they are based on the sample. Hence it is important to account for their variability in the estimation of variance of $\hat{T}_y(\hat{\alpha},\hat{\beta})$. To do this, we use the BCEF approach of Singh and Folsom.

2 Calibration Estimation via BCEF

Let us denote

$$\begin{aligned} h_1(T_y,\alpha,\beta) &= \sum_{k\in s} y_k\, d_k\, f_k(\alpha)\, g_k(\beta) - T_y, \\ h_2(\alpha) &= \sum_{k\in s} x_k\, d_k\, f_k(\alpha) - \sum_{k\in s^*} x_k\, d_k \text{ and} \\ h_3(\alpha,\beta) &= \sum_{k\in s} z_k\, d_k\, f_k(\alpha)\, g_k(\beta) - T_z. \end{aligned}$$

Now, for simplicity we assume that α, β and T_y are scalars. Later on α, β and T_y will be treated as vector valued parameters. Under the joint design and super population model, $p\xi_1\xi_2$, the estimating functions have the following mean and variance-covariance matrices

$$\begin{aligned} E_{p\xi_1\xi_2}\begin{bmatrix} h_1(T_y,\alpha,\beta) \\ h_2(\alpha) \\ h_3(\alpha,\beta) \end{bmatrix} &= \mathbf{0} \\ V_{p\xi_1\xi_2}\begin{bmatrix} h_1(T_y,\alpha,\beta) \\ h_2(\alpha) \\ h_3(\alpha,\beta) \end{bmatrix} &= V(\alpha,\beta). \end{aligned}$$

The variance-covariance matrix, $V(\alpha,\beta)$, for given α and β can be estimated using standard methods in sampling theory. To solve for T_y, α and β we simply solve the following equations

$$\begin{aligned} h_1(T_y,\alpha,\beta) &= 0 \\ h_2(\alpha) &= 0 \\ h_3(\alpha,\beta) &= 0 \end{aligned}$$

and get solutions $\hat{T}_y$, $\hat{\alpha}$ and $\hat{\beta}$. To find the variance of $\hat{T}_y(\hat{\alpha},\hat{\beta})$ we use the Taylor series method and linearize $h_1 = h_1(T_y,\alpha,\beta)$, $h_2 = h_2(\alpha)$ and $h_3 = h_3(\alpha,\beta)$ about $(\hat{T}_y,\hat{\alpha},\hat{\beta})$ and obtain

$$\begin{bmatrix} \hat{T}_y - T_y \\ \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{bmatrix} \doteq -\mathbf{H}^{-1} \begin{bmatrix} h_1(T_y,\alpha,\beta) \\ h_2(\alpha) \\ h_3(\alpha,\beta) \end{bmatrix}$$

where the (i,j)th element of $\mathbf{H}$ is given by $H(i,j) = \frac{\partial h_i}{\partial \theta_j}|_{(\hat{T}_y,\hat{\alpha},\hat{\beta})}$, and $\theta_1 = T_y$, $\theta_2 = \alpha$, $\theta_3 = \beta$. Note that $\mathbf{H}$ is not in general a symmetric matrix.

Using standard techniques in variance estimation we get

$$\begin{aligned} V\begin{bmatrix} \hat{T}_y - T_y \\ \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{bmatrix} &\doteq \mathbf{H}^{-1} V \begin{bmatrix} h_1(T_y,\alpha,\beta) \\ h_2(\alpha) \\ h_3(\alpha,\beta) \end{bmatrix} (\mathbf{H}^{-1})' \\ &= \mathbf{H}^{-1}\, \hat{\mathbf{V}}\, (\mathbf{H}^{-1})' \end{aligned}$$

where $\hat{\mathbf{V}}$ is obtained by standard variance estimation techniques, in which $\hat{\alpha}$ and $\hat{\beta}$ are plugged in for unknown α and β.

So far we have assumed that T_y, α and β are scalars. In practice, these parameters are vector valued. For example, in NHSDA, there are more than 20 study variables, 100-260 PS control totals and 140-300 predictors for NR adjustment. Suppose that

$\mathbf{Y}_{n\times r} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_r]$ denotes matrix of r outcome variables and the corresponding vector of population totals is denoted by $\mathbf{T}'_{\mathbf{y}} = [T_{\mathbf{y}_1}, T_{\mathbf{y}_2}, \ldots, T_{\mathbf{y}_r}]$,

$\mathbf{X}_{n^*\times p} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ denotes matrix of p predictor variables used for NR adjustment and the associated parameters are denoted by
$\boldsymbol{\alpha}' = [\alpha_1, \alpha_2, \ldots, \alpha_p]$,

$\mathbf{Z}_{n\times q} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q]$ denotes matrix of q predictor variables used for PS adjustment and the associated parameters are denoted by $\boldsymbol{\beta}' = [\beta_1, \beta_2, \ldots, \beta_q]$.

In this case, the dimension of $\mathbf{H}$ matrix will be $(r+p+q)\times(r+p+q)$. If it is computationally prohibitive to invert $\mathbf{H}$ then we can use a nonstandard version of the **Inverse Partitioned Matrix** formula. It is nonstandard in the sense that we are not trying to find the full $\mathbf{H}^{-1}$ from known inverse of the lower dimensional principal submatrix of $\mathbf{H}$. This is so because we are generally interested in the variance-covariance matrix corresponding to the r outcome variables which is a $(r\times r)$ submatrix of $\mathbf{H}^{-1}\,\hat{\mathbf{V}}\,(\mathbf{H}^{-1})'$. A recursion formula to obtain the desired variance-covariance is given below. Let $\mathbf{H}^{(0)} = \mathbf{H}$ and $\mathbf{V}^{(0)} = \mathbf{V}$. Let ν denote the ν-th recursion step. Consider the following conformal partition separating the last row and the last column.

$$\mathbf{H}^{(\nu)} = \begin{bmatrix} \mathbf{G}_{11}^{(\nu)} & \mathbf{G}_{12}^{(\nu)} \\ \mathbf{G}_{21}^{(\nu)} & \mathbf{G}_{22}^{(\nu)} \end{bmatrix}, \quad \mathbf{V}^{(\nu)} = \begin{bmatrix} \mathbf{U}_{11}^{(\nu)} & \mathbf{U}_{12}^{(\nu)} \\ \mathbf{U}_{21}^{(\nu)} & \mathbf{U}_{22}^{(\nu)} \end{bmatrix}$$

where $\mathbf{G}_{22}^{(\nu)}$ and $\mathbf{U}_{22}^{(\nu)}$ are scalar matrices. Let $\mathbf{H}^{(\nu+1)} = \mathbf{G}_{11}^{(\nu)} - (\mathbf{G}_{12}^{(\nu)}\,\mathbf{G}_{21}^{(\nu)})/\mathbf{G}_{22}^{(\nu)}$ and $\mathbf{V}^{(\nu+1)} = [\mathbf{I}, -\mathbf{G}_{12}^{(\nu)}/\mathbf{G}_{22}^{(\nu)}]\,\mathbf{V}^{(\nu)}\,[\mathbf{I}, -\mathbf{G}_{12}^{(\nu)}/\mathbf{G}_{22}^{(\nu)}]'$ then the desired variance-covariance is given by $(\mathbf{H}^{(p+q)})^{-1}\,\mathbf{V}^{(p+q)}\,[(\mathbf{H}^{(p+q)})^{-1}]'$. Note that we only need to invert the matrices of order $(r\times r)$. Next, we discuss variance estimation using the GEM.

3 GEM Calibration Adjusted Variance Estimation for Total Estimators

Folsom and Singh (2000) introduced an innovative approach for calibration of sampling weight for extreme values, nonresponse and poststratification using the GEM. The model for the adjustment factor a_k of the kth unit in the sample is defined as:

$$a_k(\lambda_1, \lambda_2, \ldots, \lambda_m | (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m), l, c, u)$$

$$= \frac{l_k\,(u_k - c_k) + u_k\,(c_k - l_k)\,exp\,(A_k \sum_{i=1}^{m} v_{ki}\,\lambda_i)}{(u_k - c_k) + (c_k - l_k)\,exp\,(A_k \sum_{i=1}^{m} v_{ki}\,\lambda_i)}$$

where

$$A_k = \frac{(u_k - l_k)}{(u_k - c_k)(c_k - l_k)},\ l_k < c_k < u_k,$$

$v_{k1}, v_{k2}, \ldots, v_{km}$ are the m predictor values for the kth unit and

$\boldsymbol{\lambda}' = [\lambda_1, \lambda_2, \ldots, \lambda_m]$ are the associated parameters.

For estimating λ parameters we refer to Folsom and Singh (2000).

So far we have not assumed any model for NR and PS adjustment factors. Next we assume that $f_k(\boldsymbol{\alpha})$ and $g_k(\boldsymbol{\beta})$ follow the GEM and obtain variance-covariance matrix of

$$\hat{\mathbf{T}}'_{\mathbf{y}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = [\hat{T}_{\mathbf{y}_1}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}), \hat{T}_{\mathbf{y}_2}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}), \ldots, \hat{T}_{\mathbf{y}_r}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})].$$

To do so, we first write $r+p+q$ estimating functions as given below

$$h_i(T_{\mathbf{y}_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{k \in s} y_{ki}\, d_k\, f_k(\boldsymbol{\alpha})\, g_k(\boldsymbol{\beta}) - T_{\mathbf{y}_i}$$

for i=1 to r,

$$h_{r+j}(\boldsymbol{\alpha}) = \sum_{k \in s} x_{kj}\, d_k\, f_k(\boldsymbol{\alpha}) - \sum_{k \in s^*} x_{kj}\, d_k$$

for j=1 to p, and

$$h_{r+p+t}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{k \in s} z_{kt}\, d_k\, f_k(\boldsymbol{\alpha})\, g_k(\boldsymbol{\beta}) - T_{\mathbf{z}_t}$$

for t=1 to q, and where $\mathbf{T}'_{\mathbf{z}} = [T_{\mathbf{z}_1}, T_{\mathbf{z}_2}, \ldots, T_{\mathbf{z}_q}]$ are known population control totals for PS adjustments and $f_k(\boldsymbol{\alpha})$ and $g_k(\boldsymbol{\beta})$ are defined below

$$f_k(\boldsymbol{\alpha}) = a_k(\alpha_1, \alpha_2, \ldots, \alpha_p | (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p), l_1, c_1, u_1)$$

$$g_k(\boldsymbol{\beta}) = a_k(\beta_1, \beta_2, \ldots, \beta_q | (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q), l_2, c_2, u_2).$$

Also, let $A_{1k} = A_k(l_1, c_1, u_1)$ and $A_{2k} = A_k(l_2, c_2, u_2)$.

We set $h_i = 0$ for $i = 1, 2, \ldots, (r+p+q)$ and obtain $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{T}}'_{\mathbf{Y}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. Now the variance-covariance matrix of the estimated parameters is given by

$$V \begin{bmatrix} \hat{\mathbf{T}}_{\mathbf{Y}} - \mathbf{T}_{\mathbf{Y}} \\ \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{bmatrix} \doteq \mathbf{H}^{-1}\, V \begin{bmatrix} \mathbf{h}_1(\mathbf{T}_{\mathbf{y}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \mathbf{h}_2(\boldsymbol{\alpha}) \\ \mathbf{h}_3(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{bmatrix} (\mathbf{H}^{-1})'$$

$$\doteq \mathbf{H}^{-1}\, \hat{\mathbf{V}}\, (\mathbf{H}^{-1})'$$

where
$\mathbf{h}'_1 = [h_1, h_2, \ldots, h_r]$,
$\mathbf{h}'_2 = [h_{r+1}, h_{r+2}, \ldots, h_{r+p}]$,
$\mathbf{h}'_3 = [h_{r+p+1}, h_{r+p+2}, \ldots, h_{r+p+q}]$,

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \mathbf{H}_{23} \\ \mathbf{H}_{31} & \mathbf{H}_{32} & \mathbf{H}_{33} \end{bmatrix}$$

and $\mathbf{H}_{11}$ is $(r \times r)$, $\mathbf{H}_{22}$ is $(p \times p)$ and $\mathbf{H}_{33}$ is $(q \times q)$ matrices. The analytic expression for $\mathbf{H}$ may seem to be tedious but as shown below, it is fairly simple and follow certain patterns. The components of $\mathbf{H}$ matrix are defined below

$$\mathbf{H}_{11} = \frac{\partial}{\partial \mathbf{T}_{\mathbf{y}}} \mathbf{h}_1(\mathbf{T}_{\mathbf{y}}, \boldsymbol{\alpha}, \boldsymbol{\beta})|_{(\hat{\mathbf{T}}_{\mathbf{y}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = -\mathbf{I}(r \times r)$$

$$\mathbf{H}_{12} = \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{h}_1(\mathbf{T}_{\mathbf{y}}, \boldsymbol{\alpha}, \boldsymbol{\beta})|_{(\hat{\mathbf{T}}_{\mathbf{y}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = \mathbf{Y}'\, \mathbf{X}^*$$

$$\mathbf{H}_{13} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{h}_1(\mathbf{T}_{\mathbf{y}}, \boldsymbol{\alpha}, \boldsymbol{\beta})|_{(\hat{\mathbf{T}}_{\mathbf{y}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = \mathbf{Y}'\, \mathbf{Z}^*$$

$$\mathbf{H}_{21} = \frac{\partial}{\partial \mathbf{T}_{\mathbf{y}}} \mathbf{h}_2(\boldsymbol{\alpha})|_{(\hat{\boldsymbol{\alpha}})} = \mathbf{O}(p \times r)$$

$$\mathbf{H}_{22} = \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{h}_2(\boldsymbol{\alpha})|_{(\hat{\boldsymbol{\alpha}})} = \mathbf{X}'\, \mathbf{X}^{**}$$

$$\mathbf{H}_{23} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{h}_2(\boldsymbol{\alpha})|_{(\hat{\boldsymbol{\alpha}})} = \mathbf{O}(p \times q)$$

$$\mathbf{H}_{31} = \frac{\partial}{\partial \mathbf{T}_{\mathbf{y}}} \mathbf{h}_3(\boldsymbol{\alpha}, \boldsymbol{\beta})|_{(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = \mathbf{O}(q \times r)$$

$$\mathbf{H}_{32} = \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{h}_3(\boldsymbol{\alpha}, \boldsymbol{\beta})|_{(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = \mathbf{Z}'\, \mathbf{X}^*$$

$$\mathbf{H}_{33} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{h}_3(\boldsymbol{\alpha}, \boldsymbol{\beta})|_{(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = \mathbf{Z}'\, \mathbf{Z}^*$$

where
$\mathbf{X}^*(k, j) = \mathbf{X}(k, j)\, d_k\, g_k(\boldsymbol{\beta})\, D_{1k}$

$\mathbf{X}^{**}(k, j) = \mathbf{X}(k, j)\, d_k\, D_{1k}$

$\mathbf{Z}^*(k, j) = \mathbf{Z}(k, j)\, d_k\, f_k(\boldsymbol{\alpha})\, D_{2k}$ and
$D_{1k} = D_k(\alpha_1, \alpha_2, \ldots, \alpha_p | (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p), l_1, c_1, u_1)$
$D_{2k} = D_k(\beta_1, \beta_2, \ldots, \beta_q | (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q), l_2, c_2, u_2)$

and

$$D_k(\lambda_1, \lambda_2, \ldots, \lambda_m | (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m), l, c, u) = \frac{(u_k - l_k)(c_k - l_k)(u_k - c_k) A_k \, exp(A_k \sum_{i=1}^{m} v_{ki} \lambda_i)}{[(u_k - c_k) + (c_k - l_k) exp(A_k \sum_{i=1}^{m} v_{ki} \lambda_i)]^2}.$$

Note that the $\mathbf{H}$ matrix involves $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and is therefore computable, whereas the covariance matrix $\mathbf{V}$ involves the unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are replaced by $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ to get an approximate answer.

4 GEM Calibration Adjusted Variance Estimation for Ratio Estimators

Suppose we are interested in finding the variance of $\hat{R}_i = \frac{\hat{T}_{y_i}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})}{\hat{T}_{\nu_i}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})}$ for i=1 to r. The estimating equations in this case can be written as

$$\sum_{k \in s} y_{ki} \, d_k \, f_k(\boldsymbol{\alpha}) \, g_k(\boldsymbol{\beta}) \; - \; R_i \sum_{k \in s} \nu_{ki} \, d_k \, f_k(\boldsymbol{\alpha}) \, g_k(\boldsymbol{\beta}) = 0$$

for i=1 to r,

$$\sum_{k \in s} x_{kj} \, d_k \, f_k(\boldsymbol{\alpha}) - \sum_{k \in s^*} x_{kj} \, d_k = 0$$

for j=1 to p, and

$$\sum_{k \in s} z_{kt} \, d_k \, f_k(\boldsymbol{\alpha}) \, g_k(\boldsymbol{\beta}) - T_{\mathbf{z}_t} = 0$$

for t=1 to q, and $T_{\mathbf{z}_t}$, $f_k(\boldsymbol{\alpha})$, and $g_k(\boldsymbol{\beta})$ are defined in the previous section. The components of $\mathbf{H}$ matrix for estimating functions corresponding to the estimating equations given above are defined below $\mathbf{H}_{11}(i) = -\sum_{k \in s} \nu_{ki} \, d_k \, f_k(\boldsymbol{\alpha}) \, g_k(\boldsymbol{\beta})$, i=1 to r where $\mathbf{H}_{11}$ is a diagonal matrix, $\mathbf{H}_{12} = \boldsymbol{\Psi}' \mathbf{X}^*$, $\mathbf{H}_{13} = \boldsymbol{\Psi}' \mathbf{Z}^*$, $\mathbf{H}_{22} = \mathbf{X}' \mathbf{X}^{**}$, $\mathbf{H}_{32} = \mathbf{Z}' \mathbf{X}^*$, $\mathbf{H}_{33} = \mathbf{Z}' \mathbf{Z}^*$, and $\mathbf{H}_{21}$, $\mathbf{H}_{23}$, $\mathbf{H}_{31}$ are null matrices, and $\boldsymbol{\Psi} = [(\mathbf{y}_1 - \hat{R}_1 \mathbf{x}_1), \ldots, (\mathbf{y}_r - \hat{R}_r \mathbf{x}_r)]$. The sandwich variance of $\hat{R}_i$'s can now be obtained by using results described in the previous section.

5 SAS-IML Macro

To obtain sandwich variance a SAS-IML macro was developed and is currently being used at RTI. It requires several parameters as described below

indata: input data set of respondents and nonrespondents
y-var: list of study variables
domn-var: list of domain variables for ratio estimates
x-nonrsp: list of NR predictor variables
x-pststr: list of PS adjustment variables
l-nonrsp, c-nonrsp, u-nonrsp: lower, center and upper bounds for NR adjustment factor
l-pststr, c-pststr, u-pststr: lower, center and upper bounds for PS adjustment factor
w-design: design weight
w-nonrsp: NR adjustment factor
w-pststr: PS adjustment factor
resp-ind: response indicator variable
v-str: stratum variable
v-rep: replicate within stratum variable.

In our experience the macro is easy to use and takes only few minutes on a PC equipped with Pentium III-450 Mhz chip with 128 Mb of RAM for large data sets such as that for NHSDA.

6 Applications to the NHSDA

We demonstrate the methodology for the East South Central (AL, MS, KY, TN) Census Division. The data is taken from the 1999 NHSDA. We used 96 predictors for PS adjustment, and 138 for NR adjustment. Total size (respondents and nonrespondents) of the data set was 4664 and there were 3688 respondents in the data sets, see Chen, Penne, and Singh (2000).

Predictors for the NR adjustment: Typically, we use State/region, quarter, group quarters indicator, population density, % Hispanic in segment, % Black in segment, % owner occupied dwelling units in segment, and socio-economic status indicators, age group, gender, race, Hispanicity, relation to head of household and various interaction terms.

Predictors for the PS adjustment: Predictors, typically, used are State/region, age, race, gender, Hispanicity, quarter, and the model consists of main effects and some interactions of these predictors.

We compute three Taylor-type variance estimates (i) accounting for the nonresponse and poststratification (ii) accounting for poststratification only and finally (iii) the unadjusted variance estimates. If the estimator is a ratio (such as the drug prevalence rate) then the unadjusted variance estimate considered here does take into account for the nonlinear nature of the statistic by using residuals obtained by linearizing the ratio estimator. This is so even if

the denominator (representing a domain total) becomes constant after PS adjustment. In Table 1, the suffixes FLAG, YR, MON denote lifetime, year and monthly usage respectively, and prefixes ALC, CIG and MRJ and COC denote Alcohol, Cigarette, Marijuana and Cocaine respectively. The data in Table 1 represent the percent relative standard errors (standard error/estimate) for the chosen drug recency variables. We use similar notations in Table 2. The data in Table 2 represent the percent relative standard error for the ratio estimators of the chosen drug recency variables.

In Table 1 and Table 2 the numbers in the unadjusted row are based on SUDAAN variances, assuming that $f_k(\hat{\alpha})$'s, $g_k(\hat{\beta})$'s are fixed and do not contribute to the variances of the estimators i.e. we treat $d_k\, f_k(\hat{\alpha}), g_k(\hat{\beta})$ as the basic design weight. The numbers in PS-adj and NRPS-adj rows are obtained by the methods described in Sections 3 and 4. The PS-adj row corresponds to the assumption that $f_k(\hat{\alpha})$'s are fixed and do not contribute to the variances of the estimators. Whereas the numbers in the NRPS-adj row do not assume that $f_k(\hat{\alpha})$'s and $g_k(\hat{\beta})$'s are fixed and hence their contribution to the variance is accounted.

The numerical results given in this section show that the variance estimates are remarkably stable and, in general, show gains in efficiency after calibration. It may be noted that if there are too many parameters, then the sandwich variance may become unstable and may be more than the unadjusted variance. With nonresponse adjustment, we generally expect that the variance may go up although the variance inflation due to random controls (corresponding to predictors used in NR adjustment) may be offset by the correlation between the outcome and predictor variables. Also, after the poststratification adjustment we expect the variance to go down because the controls are nonrandom and the corresponding predictors are expected to be well correlated with the outcome variable.

The variances were computed using the standard methods in sampling theory based on treating PSUs as iid. In the 1999 NHSDA, there are 12 strata or Field Interviewer regions per small state, each having two (pseudo) PSUs defined by grouping four segments–one from each quarter. Since the number of PSUs is not that large compared to the number of parameters, it suggests that the degrees of freedom available for variance estimation is probably more than the total number of PSUs minus the number of strata; which is 4(24-12)=48 for the East South Central Census division. Although the notion of degrees of freedom with survey data is still not well understood, it is conjectured that the effective sample size might be a better starting point for assessing the available degrees of freedom.

7 Final Remarks and Future Research

The main points are listed below.

- Although the **H** matrix looks complicated but it follows a nice pattern and can easily be calculated and inverted by using SAS IML.
- Even with a large number of predictor variables used in calibration, the variance estimates adjusted for calibration seem remarkably stable and, in general, show gains in efficiency after calibration.
- The BCEF methodology is very general and can easily be adapted to other types of calibration techniques.

In future we plan to do a validation study by computing resampling variance estimates using Jackknife and compare the results with the BCEF approach based on the Taylor method. It may be noted that the Jackknife method for obtaining adjusted variance could be quite tedious for large data sets, and with somewhat elaborate NR/PS models.

Acknowledgments: This work is partially supported by Substance Abuse and Mental Health Services Administration, Department of Health and Human Services.

References

Folsom, R. E. and Singh, A. C. (2000). "A Generalized Exponential Model for Sampling Weight calibration for Extreme Values, Non-response and Post-stratification," *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*(to appear)

Chen. P., Penne, M.A., and Singh, A.C. (2000). "Experience with the generalized exponential model for weight calibration for the National Household Survey on Drug Abuse," *Proceedings of the Section on Survey Research Methods of the American Statistical Association.* (to appear)

Singh, A. C. and Folsom, R. E. (2000). "Bias Corrected Estimating Function Approach for Variance Estimation Adjusted for Post-Stratification," *Proceedings of the Section on Survey Research Methods of the American Statistical Association.* (to appear)

Table 1: Comparison of Percent Relative Standard Error for Different Types of Calibration Adjustment in Taylor Variance						
Drug Recency	Type	Age Group				
		Overall	12-17	18-25	26-34	35+
ALCFLAG	Unadjusted	1.94	4.43	2.23	1.76	3.03
	PS-adj	1.55	4.10	2.16	1.70	2.46
	NRPS-adj	1.47	4.18	2.00	1.65	2.31
ALCYR	Unadjusted	3.18	4.94	3.54	3.83	5.30
	PS-adj	3.55	4.51	3.28	3.77	6.23
	NRPS-adj	2.47	4.58	3.12	3.47	4.35
ALCMON	Unadjusted	4.58	7.56	5.02	6.24	6.77
	PS-adj	5.79	6.75	4.31	5.31	9.54
	NRPS-adj	3.46	7.19	4.14	5.16	5.32
CIGFLAG	Unadjusted	1.99	4.57	2.28	3.05	2.95
	PS-adj	1.67	4.15	1.96	3.06	2.52
	NRPS-adj	1.64	4.15	1.79	2.88	2.49
CIGYR	Unadjusted	4.38	5.55	3.84	7.14	7.37
	PS-adj	4.49	5.49	3.75	6.16	7.16
	NRPS-adj	3.88	5.31	3.47	6.10	6.51
CIGMON	Unadjusted	4.43	6.68	4.31	7.83	7.20
	PS-adj	4.52	7.15	4.18	6.51	7.11
	NRPS-adj	3.93	6.71	3.94	6.48	6.38
MRJFLAG	Unadjusted	3.48	7.36	4.36	5.30	6.16
	PS-adj	2.98	7.20	3.88	4.83	5.16
	NRPS-adj	2.60	7.05	3.77	5.00	4.21
MRJYR	Unadjusted	6.81	8.81	6.59	16.57	17.02
	PS-adj	6.88	8.63	6.55	16.89	18.31
	NRPS-adj	6.29	8.58	6.24	16.37	17.30
MRJMON	Unadjusted	14.06	14.20	10.78	21.51	51.12
	PS-adj	10.76	16.03	10.34	21.38	37.53
	NRPS-adj	10.73	13.70	9.76	21.94	35.54
COCFLAG	Unadjusted	10.33	20.94	12.12	11.86	16.25
	PS-adj	9.84	30.77	11.41	9.81	16.10
	NRPS-adj	8.80	20.84	11.00	9.11	14.06
COCYR	Unadjusted	21.73	29.26	15.60	31.12	48.08
	PS-adj	19.90	42.94	15.34	30.31	45.81
	NRPS-adj	17.69	27.49	14.81	29.38	38.87
COCMON	Unadjusted	40.27	48.05	37.18	51.27	86.13
	PS-adj	31.71	86.07	35.12	48.54	79.65
	NRPS-adj	35.34	43.91	32.83	54.87	74.07

Table 2: Comparison of Percent Relative Standard Error of Ratio Estimators for Different Levels of Calibration Adjustment in Taylor Variance							
Ratio of Population Totals of		Type	Age Group				
			Overall	12-17	18-25	26-34	35+
ALCFLAG	CIGFLAG	Unadjusted	1.61	4.69	2.17	2.96	2.40
		PS-adj	1.57	4.47	2.20	2.96	2.37
		NRPS-adj	1.58	4.39	2.13	2.97	2.34
MRJFLAG	CIGFLAG	Unadjusted	4.04	6.63	3.64	5.62	6.58
		PS-adj	3.23	6.60	3.47	4.81	5.10
		NRPS-adj	3.06	6.08	3.36	4.91	4.58
COCFLAG	CIGFLAG	Unadjusted	10.38	20.16	11.99	11.10	16.36
		PS-adj	9.39	22.15	11.39	8.58	15.08
		NRPS-adj	8.81	20.11	11.06	8.26	14.19
COCFLAG	MRJFLAG	Unadjusted	9.11	20.63	12.11	9.89	14.52
		PS-adj	8.60	22.53	11.68	7.85	13.51
		NRPS-adj	8.07	21.15	11.30	7.20	12.90