Comparing Weighting Methods When Adjusting for Logistic Unit Nonresponse

Phillip S. Kott and Dan Liao

RTI International
pkott@rti.org; dliao@rti.org

One way to adjust for unit nonresponse in a sample survey is by fitting a logistic response function using some variant of maximum likelihood and then treating the inverses of the fitted response probabilities as second-phase sampling weights. An slightly different approach fits the logistic response model using a calibration equation. Theory suggests that the calibration approach should be more efficient than its maximum-likelihood-based competition when estimating the population total (or mean) of a survey variable that is roughly a linear function of the covariates in the response model. Suppose, however, as may be the case in some establishment surveys, the survey variable is indeed roughly a linear function of a single size measure, but unit response is a logistic function of the *log* of that measure, that is, a one percent increase in the size measure results in a *q* percent change in the odds of responding. What estimation strategy should then be used? We will investigate this question empirically using data from public-use files of the Drug Awareness Warning Network (DAWN) survey of drug-related emergency-room visits. Prediction (or outcome) models will be employed in interpreting the sometimes surprising results.

Key Words: Bias; Calibration Weighting; Quasi-random response model; Prediction model; WTADJUST.

1. Introduction

Several methods of adjusting for unit nonresponse begin by fitting a logistic response function to the original sample using either maximum-likelihood (ML) or weighted-maximum likelihood (WML). In the next step, either the fitted probabilities of response are inverted and treated like second-phase probability-sampling weights or the sample is sorted by the fitted probabilities of response into (say) five reweighting classes.

A similar approach fits the logistic response model indirectly using a calibration equation. When the survey variable of interest is roughly a linear function of the covariates in the logistic-response model, Kim and Riddles (2012) argue that this method can be more efficient than a weighting scheme derived using maximum-likelihood methods. They do not suppy an intutive reason for this, but one explanation is that if the survey variable really obeyed a prediction (or outcome) model in which its expectation were a linear funtion of the covariates, then only calibration weighting would always produce an unbiased estimator under the model.

We will compare various methods of the unit nonresponse adjustments based on a simple, one-covariate logistic response model using data from the 2008 public use file of the Drug Awareness Warning Network or DAWN (Substance Abuse and Mental Health Services Administration, 2011). The DAWN is an annual survey of drug-related visits to hospital emergency rooms based on a stratified simple random sample of hospitals.

As is in many establishment surveys, there is an auxiliary size variable attached to each hospital on the DAWN frame – all emergency-room visits in a previous year. Using this auxiliary variable in estimation can improve the statistical efficiency of estimated totals for DAWN survey variables having a rough linear relationship with it. We will investigate some of the estimators, like the simple ratio, that exploit this relationship.

In one set of simulations, we will generate unit response as a logistic function of the *log* of the auxiliary variable, so that a one percent increase is the size of the auxiliary variable results in a *q* percent increase (or decreases) in the odds of response. Some preliminary exploration of DAWN data reveal this better mimics the behavior of survey response than a model in which unit response is a logistic function of the auxiliary variable. Unfortunately, it may also remove the theoretical advantage of calibration weighting, which exploits the near linear relationship of the survey variable and the auxiliary.

In a second set of simulations, we will generate unit response as a logistic function of the *square root* of the auxiliary variable, which also fits the actual response behavior of DAWN survey data reasonably well, but construct the estimators wrongly assuming response to be a function of the *log* of the auxiliary variable. This will let us assess how robust the estimators are when this failure of the response-model assumption occurs.

We will interpret the results drawing on both quasi-probability (often called "quasi-design-based") and prediction-model (often called "model-based") sampling theory. In the first, response is treated as a second phase of probability sampling. In the second, the survey value of interest is treated like a random variable with an expectation that is a function of the auxiliary variable.

Our goals here are very modest. We are not attempting to undercover the best model for for adjusting for the unit nonresponse in establishments surveys in general or the DAWN in particular. We are simply exploring how well various adjustment methods based on fitting a logistic model work under relatively clean circumstances with a single cause of nonresponse. Nevertheless, some of our results provide useful insights. In particular, we see that in out setup, near unbiasedness under the prediction model is a particularly valuable property.

After reviewing some of the quasi-probability sampling theory underpinning response modeling in Section 2, we provide a description of the data to be used and estimators to be computed in Section 3. We display and interpret our empirical results in Section 4 and offer some concluding remarks in Section 5.

All the calibration weighting in this paper is done using the WTADJUST procedure in SUDAAN 10® (RTI International, 2008), which produces only positive weights for respondents.

2. Some Theory

Suppose we have a randomly drawn sample S of size n and no nonresponse. Using probability-sampling principle (also called "design-based inference"), we can estimate a population total, $T_y = \sum_{i \in U} y_k = \sum_U y_k$, where U denotes the population, with the expansion estimator $t_y^E = \sum_S y_k/\pi_k = \sum_U y_k I_k/\pi_k$, where $I_k = 1$ when $k \in S$ and 0 otherwise. We can also write $t_y^E = \sum_U d_k y_k = \sum_S d_k y_k$, where $d_k = I_k/\pi_k$ is the design weight of element k. Treating the I_k as random variables, it is easy to see that t_y^E is an unbiased estimator for T_y.

One popular way of handling unit nonresponse is to assume a model where whether (or not) a unit responds is treated as an additional phase of probability sampling. In the simplest version of this quasi-probability-sampling approach, each element $k \in U$ is assumed to have a probability of response ρ_k if it is sampled. The probability elements k and j jointly respond if sampled is $\rho_k \rho_j$, and the magnitude of ρ_k is independent of whether k is chosen for the original sample. The value of ρ_k is itself unknown, but it is assumed to be expressible as a function $\rho_k = f(\mathbf{x}_k^T \boldsymbol{\gamma})$, where $\mathbf{x}$ is a vector of characteristics for k that is known when $k \in S$. Although the form of f (.) is assumed to be known, its governing parameter vector $\boldsymbol{\gamma}$ is not; it needs to be estimated from the sample.

When a consistent estimator $\mathbf{g}$ is found for $\boldsymbol{\gamma}$, it is not hard to show that under the assumptions of the response model and the original probability-sampling mechanism, the double expansion estimator $t_y^{E*} = \sum_U y_k I_k Q_k/(\pi_k p_k) = \sum_R w_k y_k$ is a nearly (i.e., asymptotically) unbiased estimator for T_y, where $Q_k = 1$ when k responds if sampled and 0 otherwise, $p_k = f(\mathbf{x}_k^T \mathbf{g})$, R is the subset of S that responds, and $w_k = 1/(\pi_k p_k) = d_k/p_k$ is the sampling weight for unit k, the unit's design weight adjusted for nonresponse. The interested reader is referred to, for example, Chang and Kott (2008) for a rigorous treatment of the underlying theory.

A commonly used response model assumes that the log odds of unit k responding is linear function of $\mathbf{x}_k$. This means that f(.) is the logistic function $f(\mathbf{x}_k^T \boldsymbol{\gamma}) = 1/[1 + \exp(-\mathbf{x}_k^T \boldsymbol{\gamma})]$. One can then estimate $\boldsymbol{\gamma}$ through maximum likelihood (ML), which comes done to using Newton's method (i.e., successive linearizations) to find a $\mathbf{g}$ that satisfies the estimating equation:

$$\sum_{k\in S}\left[Q_k - f\left(\mathbf{x}_k{}^T\mathbf{g}\right)\right]\mathbf{x}_k = \mathbf{0}. \tag{1}$$

The ML method is not only consistent under mild conditions on the population and original sample design, but also produces the most efficient estimator for $\boldsymbol{\gamma}$. That is why it is arguably superior to its "design-based" alternative that finds a solution to the weighted estimating equation:

$$\sum_{k\in S} d_k\left[Q_k - f\left(\mathbf{x}_k{}^T\mathbf{g}\right)\right]\mathbf{x}_k = \mathbf{0}. \tag{2}$$

The weighted-maximum-likelihood (WML) method in equation (2), which also produces a consistent estimator for $\boldsymbol{\gamma}$, may be more commonly used in practice. See, for example, Diaz-Tina *et al.* (2002). From here on, we drop the cumbersome phrase "under mild conditions on the population and original sample design" for convenience. The reader should be aware not only that the phrase is missing but also that we always assume those mild conditions to be met.

Kim and Riddles (2012) assert that if the goal is estimating T_y, and y_k is roughly linear in $\mathbf{x}_k$, then a consistent estimator for $\mathbf{g}$ that is likely to be more efficient than either ML or WML when creating a nearly unbiased estimator for T_y is the solution of the equation:

$$\sum_{k\in S} d_k\left[\frac{Q_k}{f\left(\mathbf{x}_k{}^T\mathbf{g}\right)} - 1\right]\mathbf{x}_k = \mathbf{0}. \tag{3}$$

Equation (3) is equivalent to the second equality in the *calibration equatio*n:

$$\sum_{k\in R} w_k\mathbf{x}_k = \sum_{k\in S} d_k\frac{Q_k}{f\left(\mathbf{x}_k{}^T\mathbf{g}\right)}\mathbf{x}_k = \sum_{k\in S} d_k\mathbf{x}_k \tag{4}$$

It is easy to see from equation (4) that if y_k were exactly equal to a linear function of $\mathbf{x}_k$, say $\mathbf{x}_k{}^T\boldsymbol{\beta}$, then adjusting for nonresponse by solving for $\mathbf{g}$ in equation (3) and then computing $t_y^{E*} = \sum_R w_k y_k$ would produce the same estimator for T_y as if there were no nonresponse (since $\sum_R w_k\mathbf{x}_k{}^T\boldsymbol{\beta} = \sum_S d_k\mathbf{x}_k{}^T\boldsymbol{\beta}$). The same cannot be said, in general, if $\mathbf{g}$ were estimated using ML or WML. That is the reasoning behind the assertion in Kim and Riddles.

Another advantage of using the calibration approach in adjusting for unit nonresponse is that it can be used when $\mathbf{x}_k$ is not known for every sampled unit as long as the vector of population totals $T_{\mathbf{x}} = \sum_U \mathbf{x}_k$ is known (or there is an unbiased estimator for $T_{\mathbf{x}}$). This can be done by finding a $\mathbf{g}$ such that

$$\sum_{k\in R} w_k\mathbf{x}_k = \sum_{k\in S} d_k\frac{Q_k}{f\left(\mathbf{x}_k{}^T\mathbf{g}\right)}\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k = T_{\mathbf{x}}. \tag{5}$$

Calibration weighting can also be used in the absence of nonresponse – or after the sampling weights have been adjusted for nonresponse – to increase the efficiency the estimator when y_k is roughly linear in $\mathbf{x}_k$. In fact, Deville and Särndal (1992) developed calibration weighting for that purpose.

The inverse of the logistic function can be written generically as $1/f(\theta) = 1 + \exp(-\theta)$, and is always greater than 1. When calibrating in the absense of nonresponse to increase efficiency, $1/f(\theta)$ is replaced by an $h(\theta)$ such that $h(0) = h'(0) = 1$. One then finds a vector **g** "estimating" **0** such that

$$\sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in S} d_k h\left(\mathbf{x}_k{}^T\mathbf{g}\right)\mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k = T_{\mathbf{x}}. \qquad (6)$$

Although **g** has a known target (**0**), a more efficient estimator for T_y comes from solving for **g** in the calibration equation (6).

All choices for $h(\theta)$ are asymptotically equivalent; that is, combining them with equation (6) lead to estimators for T_y that have asymptotically identical mean squared errors. Two common ones are $1+\theta$ and $\exp(\theta)$. When the former is used, it is possible to solve for **g** using matrix algebra. This convenience is offset by the possibility that a calibration weights, w_k, will fall below zero. When $\exp(\theta)$ is used, that can never happen. It is possible, however, that no **g** satisfying the calibration equation exists.

Deville and Särndal noted that when the components of $\mathbf{x}_k$ are binary, and $h(\theta) = \exp(\theta)$, **g** can be solved using iterative proportional fitting or raking. Folsom and Singh (2000) allowed the $h(\theta)$ to vary across the units. Observe that, if we replace $h(\mathbf{x}_k{}^T\mathbf{g})$ in equation (6) by $h(\mathbf{x}_k{}^T\mathbf{g})$, where

$$h_k(\theta) = \begin{cases} 1 & \text{when } d_k = 1 \\ 1/d_k + [1-(1/d_k)]\exp([1-(1/d_k)]^{-1}\theta) & \text{otherwise,} \end{cases}$$

then w_k will never fall below 1, a property many find desirable.

We do not presently know how useful calibration weighting is when there is unit nonresponse but the survey variable is not roughly a linear function of the covariates of the response model. Moreover, we do not know whether it is better to make a single calibration adjustment in this situation or two (one to adjust for nonresponse and one to increase statistical efficiency). Those issues will be addressed in the following sections.

3. Data and Estimators

We generate a synthetic population, U, of hospitals from the 2008 DAWN public-use file in the following manner. The file is at the case level. We take the weighted sums across all cases in a hospital to create hospital-level variables attached to a hospital record k. We then treat that record as a_k independent records in U, where a_k is the minimum of the case weights in the hospital randomly rounded to an integer.

After creating U, we independently draw 2,500 stratified simple random samples of size 400 from U using the strata definitions on the public-use file. These definitions incorporate information on location and hospital ownership (public or private) not directly provided on the file.

We set the stratum sample sizes roughly proportional to a size measure, but never less than 4. For the size measure, z, we use total drug-related emergency-room visits. The actual DAWN frame size variable, total emergency-room visits in a previous year according to the American Hospital Association, is not on the file. Design weights in our simulations vary between 4.375 and 48.

For each simulated sample, we generate a respondent sample R based on Bernoulli draw from the logistic function:

$$\rho_k = \rho(z_k) = (1 + \exp(3.735 - .4\log(z_k))^{-1}, \qquad (7)$$

We also create alternative respondent samples using

$$\rho_k = \rho(z_k) = (1 + \exp(.597 - .005 z_k{}^{1/2})^{-1}. \qquad (8)$$

Both response models produce unweighted overall response rates of around 54%, which is similar to actual DAWN experience, where response is also a mildly increasing function of size.

Finally, we treat drug-related emergency-room visits with an adverse pharmaceutical reaction (y_{1k}) and deaths (y_{2k}) as two survey variables of interest and compute $T_y = \sum_U y_k$ for each (suppressing the indices 1 or 2 for convenience). We also create a third synthetic variable for investigative purposes: $y_{3k} = z_k^{1.3}$.

Many of the estimators we analyze have the form $t_y^{E*} = \sum_R w_k y_k$. For two, $w_k = d_k / p_k$, where d_k is the inverse of the hospital's selection probability into the sample of 400, and p_k is either the solution to equation (1) or (2) with $f(\theta) = [1 + \exp(-\theta)]^{-1}$, and $\mathbf{x}_k = (1 \;\; \log(z_k))'$. We label these estimators t_y^{ML} and t_y^{WML}, respectively. The corresponding calibration estimator derived from solving equation (4) with $\mathbf{x}_k = (1 \;\; \log(z_k))'$ is labeled t_y^{Cal1}. It is computed using the *nonresponse* option in SUDAAN's WTADJUST, setting the lower bound to 1 and the center to 2.

All three of these estimators should be nearly unbiased when equation (7) is used to generate response. A fourth estimator, $t_y^{Cal1\_pop}$, makes use of both the response model and the presumption that the survey variable is roughly linear in *z*. Its weights come from solving the calibration equation in (5) with $\mathbf{x}_k = (1 \;\; \log(z_k) \;\; z_k)'$. The the *post* option in WTADJUST is used to compute those weighting with the lowerbound set to 1 and the center to 2.

A more conventional way to exploit the presumed rough linear relationship between the survey variable and *z* is with a ratio estimator of the form:

$$t_{y,r}^{E*} = t_z \frac{\sum_{k \in R} w_k y_k}{\sum_{k \in R} w_k z_k}.$$

We label the ratio version of the ML estimator $t_{y,r}^{ML}$ and define $t_{y,r}^{WML}$ and $t_{y,r}^{Cal1}$ analogously.

One can also apply a second calibration-weighting adjustment to the weights in t_y^{Cal1} to exploit the presumed rough linear relationship between the survey variable and z. The estimator t_y^{Cal2} begins by setting the d_k in equation (6) to the weights from t_y^{Cal1}, replaces $h(\theta)$ with $h_k(\theta) = 1/d_k + [1 - (1/d_k)]\exp([1 - (1/d_k)]^{-1}\theta)$, and sets $\mathbf{x}_k = z_k$ (note that because there is a previous nonresponse adjustment, no $d_k = 1$). This is done with the post option in WTADJUST by setting the lower bound to $1/d_k$ (the center takes on its default setting of 1). Alternative two-step calibration estimators, labeled $t_y^{Cal2\_1}$ and $t_y^{Cal2\_z\log z}$, set $\mathbf{x}_k = (1 \;\; z_k)'$ and $\mathbf{x}_k = (z_k \;\; z_k\log(z_k))'$, respectively.

Little (1986) suggested that instead of using fitted p_k-values directly from either (1) or (2), reweighting groups should be created based on the sorted p_k-values. In our context, this is the same as grouping by the sorted *z*-values. We sort a sample *S* into five nearly equal groups and compute these *grouped* estimators:

$$t_y^{Gr} = \sum_{c=1}^{5} \left(\frac{\sum_{k \in S_c} d_k}{\sum_{k \in R_c} d_k} \right) \sum_{k \in R_c} d_k y_k,$$

$$t_{y,r}^{Gr} = \left(\sum_{k\in U} z_k\right) \frac{\sum_{c=1}^{5}\left(\frac{\sum_{k\in S_c} d_k}{\sum_{k\in R_c} d_k}\right)\sum_{k\in R_c} d_k y_k}{\sum_{c=1}^{5}\left(\frac{\sum_{k\in S_c} d_k}{\sum_{k\in R_c} d_k}\right)\sum_{k\in R_c} d_k z_k}, \text{ and}$$

$$t_y^{SGr} = \sum_{c=1}^{5}\left(\sum_{k\in U_c} z_k\right)\frac{\sum_{k\in R_c} d_k y_k}{\sum_{k\in R_c} d_k z_k}.$$

where the subscript c denotes the group.

These estimators effectively estimates ρ_k with $\mathbf{u}_k'\mathbf{p}_+$, where $\mathbf{u}_k$ is a five-element vector of group-membership indicators (i.e., $u_{kc} = 1$ when $k \in U_c$, and 0 otherwise). When $k \in U_c$, ρ_k is implicitly estimated by $p_k = p_{+c} = \sum_{R_c} d_j / \sum_{S_c} d_j$ in both the *grouped* (t_y^{Gr}) and *grouped-ratio* ($t_{y,r}^{Gr}$) estimators. Little and Varitvarian (2003) argue for using the unweighted response rate in group c ($p_{+c} = r_c/n_c$) in this context, but Kott (2011) presents reasons for preferring our formulation.

The *separate-grouped-ratio* estimator (t_y^{SGr}) has the word "ratio" in its name for an obvious reason: there is a ratio in every group. Nevertheless, it can be expressed in double-expansion form as $t_y^{E*} = \sum_R w_k y_k = \sum_R (d_k / p_k) y_k$ if we view the estimate of the group-c response rate as $p_{+c} = \sum_{R_c} d_j z_j / \sum_{U_c} z_j$.

Although t_y^{Gr}, $t_{y,r}^{Gr}$, and t_y^{SGr} are not nearly unbiased under the logistic response model when the z-values vary within each group, their use should remove a good deal of the potential for response bias when response is generated by *either* equation (7) or (8) because the z-values do not vary within a group as much as they do across the entire population.

For comparison purposes we also compute the following two estimators that naively estimate p_k by the overall unweighted response rate r/n:

$$t_y^0 = \frac{n}{r}\sum_{k\in R} d_k y_k, \text{ and } t_{y,r}^0 = \left(\sum_{k\in U} z_k\right)\frac{\sum_{k\in R} d_k y_k}{\sum_{k\in R} d_k z_k}.$$

4. The Results and their Interpretation

Table 1 contains the formulae for all the estimators. Table 2 displays the relative empirical biases and root mean squared errors of the competing estimators when response is generated using equation (7). Not surprisingly, none of the ML, WML and calibrated estimators designed to remove response bias have an empirical bias that is more than 12% of the corresponding empirical root mean squared error, which means that the contribution of bias to mean squared error is always less than 1.5% (since $0.12^2 < 0.015$), usually much less.

Among the three one-step estimators t_y^{ML}, t_y^{WML}, and t_y^{Cal1}, the calibration estimator is always the least efficient (i.e., has the largest mean squared error). This efficiency disadvantage usually goes away when the estimators are in ratio form ($t_{y,r}^{ML}$, $t_{y,r}^{WML}$, $t_{y,r}^{Cal1}$), except for the contrived third survey variable. This is likely because that variable is not as nearly linear in the size measure *z* as the other two.

Applying a second calibration adjustment to t_y^{Cal1} in place of a ratio, as t_y^{Cal2} does, improves efficiency especially for the third variable (it is not shown, but adding the restriction on the weights for t_y^{Cal2} that none fall below unity has virtually no effect on its empirical bias or root mean squared error). Using a single calibration to adjust for nonresponse and exploit the rough linear relationship between survey variable and size measure simultaneously , as $t_y^{Cal1\_pop}$ does, appears less efficient than the ratio estimators for the first two survey variables (adverse reaction to pharmaceuticals and deaths).

Adding a constant to the **x**-vector ($t_y^{Cal2\_1}$) usually has a negative impact on efficiency in Table 2. Adding $z_k\log(z_k)$ in its place ($t_y^{Cal2\_z\log z}$) has the opposite effect. This is likely because each of the survey variables is better fit as a linear model of *z* and zlog(*z*) than as a function of *z* and an intercept or of *z* alone. This is confirmed by running regressions in *U* (not shown).

The naïve expansion estimator t_y^0 performs poorly both in terms of bias and mean squared error as expected. It ratio version, however, $t_{y,r}^0$, appears more efficient than any of the ML, WML, or calibration estimators for deaths. Let us explain why we think that happens.

The expectation of $t_{y,r}^0$ under quasi-probability theory is roughly:

$$E_{I,Q}(t_{y,r}^0) = \left(\sum_{k\in U} z_k\right) E_{I,Q}\left(\frac{\sum_{k\in U} d_k I_k Q_k y_k}{\sum_{k\in U} d_k I_k Q_k z_k}\right) \approx \left(\sum_{k\in U} z_k\right) \frac{\sum_{k\in U} \rho(z_k) y_k}{\sum_{k\in U} \rho(z_k) z_k}. \qquad (9)$$

Now suppose y_k itself is a random variable such that $E_{pr}(\mathrm{y_k}|\mathrm{z_k}) = \beta \mathrm{z_k}$. Under this *prediction model*, the expression on the far right of equation (9) has the same expectation as T_y; that is βT_z. Since T_y is being treated as a random variable, strictly speaking, $t_{y,r}^0$ is an predictor of T_y, which is origin of the term "prediction model."

In point of fact, the prediction-model asumption $E_{pr}(y_k|z_k) = \beta z_k$ does not appear to hold for any of our three survey variables, but it does not fail too badly for deaths. As a consequence, $t_{y,r}^0$ is not noticeably biased, and a relatively low empirical mean squared error results.

The near equality between the estimator and its target under the prediction-model assumption $E_{pr}(y_k|z_k) = \beta z_k$ can be shown to apply to all our ratio and two-step calibration estimators. Moreover, this near equality, although dependent on the prediction-model assumption, does not require the response response function $\rho(\mathrm{z}_k)$ to be a logistic in $\log(z_k)$; it can be any function of z_k. In other words, the actually shape of the response function is ignorable as long as it is a function of z_k. Notice that we have not made the same assumption about the probability-sampling mechanism, which incorporates information about location and ownership not reflected in the assumed prediction model.

The prediction-model for $t_y^{Cal2\_1}$ can be expanded to $E_{pr}(y_k|z_k) = \beta_0 + \beta_1 z_k$. Similarly, the prediction-model can be expanded to $E_{pr}(y_k|z_k) = \beta_1 z_k + \beta_2 z_k \log(z_k)$ for $t_y^{Cal2\_z\log z}$ and to $E_{pr}(y_k|z_k) = \beta_1 z_k + \beta_2 \log(z_k)$ for $t_y^{Cal1\_pop}$. Although none of these expanded prediction-model assumptions may hold for any of our three survey variables, they are often closer to being true than the simple ratio-model assumption. This can explain the results in Table 3. When we simulate responses using equation (8) rather than (7), these three estimators tend to have smaller empirical biases and often smaller empirical mean squared errors than their non-grouped competitors.

The poor relative showing of these three estimators for deaths is likely the result of the expectation of deaths being nearly linear in the size measure, drug-related hospital-room visits. One should also note that although $t_y^{Cal2\_z\log z}$ does an excellent job estimating T_y for the third variable, its empirical bias to root mean squared error ratio in Table 3 is over 20%.

What stands out most in both Tables 2 and 3 is that the separate-grouped-ratio estimator $t_{y,r}^{SGr}$ has the smallest mean squared error among the competitors for adverse pharmaceutical reactions and deaths. To see why that may be, observe that the triple expectation of $t_{y,r}^{SGr}$ and its target (T_y) would be equal under a prediction model where $E_{pr}(y_k|z_k) = z_k\mathbf{u}_k'\boldsymbol{\delta}$ (recall that $\mathbf{u}_k$ itself depends on z_k). This may be viewed as another way of expanding the assumption of a simple ratio prediction model since $z_k\mathbf{u}_k'\boldsymbol{\delta} = \beta z_k + z_k\mathbf{v}_k'\boldsymbol{\phi}$, where $\mathbf{v}_k = (u_{k1} \ldots u_{k4})'$, and $\phi_c = \delta_c - \beta$. Even though the model lacks plausibility when $\boldsymbol{\phi} \neq \mathbf{0}$, it does feature four parameters in addition to simple-ratio slope β, so it is not too surprising that $t_{y,r}^{SGr}$ is often superior to its competitors.

Notice that in both tables, the empirical bias for the separate-grouped-ratio estimator is always less than that for the grouped estimator t_y^{Gr} even though both estimators rely on the same approximate response model: a constant response rate within each group. This suggests that the prediction model supporting the latter provides a closer fit to the population data than the prediction model supporting the former: $E_{pr}(y_k|z_k) = \mathbf{u}_k'\boldsymbol{\delta}$.

The grouped-ratio estimator $t_{y,r}^{Gr}$ is supported by the simple ratio-model assumption, which, as we noted before, is not too bad for deaths. This may be why it has less empirical bias for deaths than the separate-grouped-ratio in both tables, and its empirical mean squared error tends to be among the lowest of the competiting estimators.

In DAWN as in many establishment surveys, the impact on mean squared error of the increased variability of the weights due to calibration adjustments is not as clear as in surveys where ideally every unit has the same weight. Still, too much unproductive weight adjustment from a prediction-model viewpoint does seem to have an adverse effect on empirical mean squared error. We have already noted that $t_{y,r}^{0}$ is more efficient than many of its competitors for deaths. In addition, the nonresponse-adjusted-only calibration estimator t_y^{Cal1} is less efficient than its WML counterpart in all cases, which in turn is usually less efficient than its ML analogue.

5. Some Concluding Remarks

One of the take-away messages from the the last section is that calibration weighting appears most useful for nonresponse adjustment when it is also employed to reduce prediction-model bias under an assumed model. In addition, a single-step calibration-weighting routine that includes variables to remove both response-model bias and prediction-model bias (as used in $t_y^{Cal1-pop}$) does not appear to be as effective as employing separate calibration steps for each purpose.

Another take-away message is that a calibration-weighting step that reduces the prediction-model bias under a roughly-holding model can decrease overall mean squared error whether or not the bias due to unit nonresponse has been (asymptotically) removed by the nonresponse-adjustment step. The same can be said about a ratio adjustment when the simple-ratio model roughly holds.

We should not be too impressed by the strong performance of the separate-ratio estimator in the simulations of the *previous* two sections because both the true response model and the true prediction model were monotonic functions of a single variable (although we do not know for certain that this monotonicity holds for deaths and adverse pharmaceutical reactions, it seems likely). That will rarely be the case in practice. In the real DAWN, for example, both response and survey variables are functions of hospital size, ownership, geographic location, and level of urbanization. Separate modeling steps seem prudent especially since survey variables tend to be a function of size, other variables, and interactions between some of those other variables and size , while unit response in better modeled as a function of the log of size, other variables (but not necessarily the same list of other variables as in the prediction model), and interaction between some of those variables and log of size.

We end with a caveat about drawing strong conclusion from limited simulations. We also need to point out that the issue of mean-squared-error *estimation* has not been addressed. One reason to focus on the relative size of the empirical bias compared to empirical mean squared error is that when bias plays too big a role in mean squared error, large-sample variance estimators based on the assumption that bias is asymptotically ignorable, like those for one-step calibration in Kott (2006), will fail.

Table 1. Summary of Estimators Being Analyzed

Estimators in double expansion form: $t_y^{E*} = \sum_R (d_k / p_k) y_k$

t_y^0	$p_k = r_c / n_c$ for $k \in R_c$
t_y^{ML}	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k)\right)\right]^{-1}$ solves $\sum_S \left(Q_k - p_k\right)\begin{pmatrix} 1 \\ \log(z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
t_y^{WML}	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k)\right)\right]^{-1}$ solves $\sum_S d_k \left(Q_k - p_k\right)\begin{pmatrix} 1 \\ \log(z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
t_y^{Cal1}	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k)\right)\right]^{-1}$ solves $\sum_S d_k \left(\frac{Q_k - p_k}{p_k}\right)\begin{pmatrix} 1 \\ \log(z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
$t_y^{Cal1\_pop}$	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k) - g_3 z_k\right)\right]^{-1}$ solves $\sum_S d_k \left(\frac{Q_k}{p_k}\right)\begin{pmatrix} 1 \\ \log(z_k) \\ z_k \end{pmatrix} = \sum_U \begin{pmatrix} 1 \\ \log(z_k) \\ z_k \end{pmatrix}$
t_y^{Gr}	$p_k = \sum_{R_c} d_j / \sum_{S_c} d_j$ for $k \in R_c$
t_y^{SGr}	$p_k = \sum_{R_c} d_j z_j / \sum_{U_c} z_j$ for $k \in R_c$

Estimators in ratio form: $t_{y,r}^{E*} = \sum_U z_k \frac{\sum_R (d_k / p_k) y_k}{\sum_R (d_k / p_k) z_k}$

$t_{y,r}^0$	$p_k = r_c / n_c$ and $k \in R_c$
$t_{y,r}^{ML}$	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k)\right)\right]^{-1}$ solves $\sum_S \left(Q_k - p_k\right)\begin{pmatrix} 1 \\ \log(z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
$t_{y,r}^{WML}$	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k)\right)\right]^{-1}$ solves $\sum_S d_k \left(Q_k - p_k\right)\begin{pmatrix} 1 \\ \log(z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
$t_{y,r}^{Cal1}$	$p_k = \left[1+\exp\left(-g_1 - g_2 \log(z_k)\right)\right]^{-1}$ solves $\sum_S d_k \left(\frac{Q_k - p_k}{p_k}\right)\begin{pmatrix} 1 \\ \log(z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
$t_{y,r}^{Gr}$	$p_k = \sum_{R_c} d_j / \sum_{S_c} d_j$ for $k \in R_c$

Ratio forms for $t_y^{Cal1\_pop}$ and t_y^{SGr} are not displayed because each is equal to its own ratio form.

Two-step Calibration Estimators: $t_y^{E*} = \sum_R (d_k / p_k) h_k y_k$, where p_k comes from t_y^{Cal1}

t_y^{Cal2}	$h_k = 1/d_k + [1 - (1/d_k)] \exp\{[1 - (1/d_k)]^{-1} z_k g\}$ solves $\sum_R (d_k / p_k) h_k z_k = \sum_U z_k$
$t_y^{Cal2\_1}$	$h_k = 1/d_k + [1 - (1/d_k)] \exp\{ [1 - (1/d_k)]^{-1} [a + z_k g] \}$ solves $\sum_R (d_k / p_k) h_k \begin{pmatrix} 1 \\ z_k \end{pmatrix} = \begin{pmatrix} \sum_U 1 \\ \sum_U z_k \end{pmatrix}$
$t_y^{Cal2\_z\log z}$	$h_k = 1/d_k + [1 - (1/d_k)] \exp\{ [1 - (1/d_k)]^{-1} [z_k g_1 + z_k \log(z_k g_2] \}$ solves $\sum_R (d_k / p_k) h_k \begin{pmatrix} z_k \\ z_k \log(z_k) \end{pmatrix} = \begin{pmatrix} \sum_U z_k \\ \sum_U z_k \log(z_k) \end{pmatrix}$

Table 2. Summary of Simulation Results When Equation (7) is Used to Generate Response

Estimator	*Relative Empirical Bias (%)* $\frac{\frac{1}{2500}\sum_{a=1}^{2500}\left(t_y^{(a)}-T_y\right)}{T_y}$			*Relative Empirical Root Mean Squared Error (%)* $\frac{\sqrt{\frac{1}{2500}\sum_{a=1}^{2500}\left(t_y^{(a)}-T_y\right)^2}}{T_y}$		
	Adverse reactions	*Deaths*	*Synthentic variable*[1]	*Adverse reactions*	*Deaths*	*Synthentic variable*[1]
t_y^0	19.53	19.24	26.28	20.93	23.98	28.32
t_y^{ML}	-.0.03	-0.05	-0.01	4.51	11.80	5.59
t_y^{WML}	0.10	0.03	0.21	4.60	11.61	6.38
t_y^{Cal1}	-0.05	-0.18	0.02	5.27	11.95	7.32
$t_{y,r}^0$	-.068	-0.89	4.76	2.77	10.88	5.40
$t_{y,r}^{ML}$	0.01	-0.01	-0.03	2.41	11.18	2.08
$t_{y,r}^{WML}$	-0.01	-0.06	0.00	2.41	11.15	2.28
$t_{y,r}^{Cal1}$	-0.01	-0.13	-0.07	2.41	11.15	2.47
$t_y^{Cal1\_pop}$	-0.20	-0.27	0.11	2.47	11.28	0.98
t_y^{Cal2}	0.01	-0.11	-0.01	2.40	11.14	1.46
$t_{y,r}^{Cal2\_1}$	0.00	-0.13	0.03	2.43	11.19	1.21
$t_y^{Cal2\_z\log z}$	0.03	-0.16	-0.01	2.35	10.98	0.28
t_y^{Gr}	1.43	1.49	1.90	4.43	11.42	5.88
$t_{y,r}^{Gr}$	-0.24	-0.16	0.16	2.40	10.76	1.98
t_y^{Sgr}	-0.27	-0.62	0.45	2.29	10.00	1.02

[1] (Drug-related hospital visits)$^{1.3}$

Table 3. Summary of Simulation Results When Equation (8) is Used to Generate Response

Estimator	***Relative Empirical Bias (%)*** $\frac{\frac{1}{2500}\sum_{a=1}^{2500}\left(t_y^{(a)}-T_y\right)}{T_y}$			***Relative Empirical Root Mean Squared Error (%)*** $\frac{\sqrt{\frac{1}{2500}\sum_{a=1}^{2500}\left(t_y^{(a)}-T_y\right)^2}}{T_y}$		
	Adverse reactions	*Deaths*	*Synthentic variable*[1]	*Adverse reactions*	*Deaths*	*Synthentic variable*[1]
t_y^0	18.60	16.41	26.87	20.10	21.67	28.91
t_y^{ML}	0.09	-1.42	1.62	4.59	11.52	5.93
t_y^{WML}	2.06	0.55	4.20	5.10	11.48	7.84
t_y^{Cal1}	2.79	1.27	5.19	5.88	11.83	9.00
$t_{y,r}^0$	-1.09	-2.90	5.64	2.94	10.93	6.17
$t_{y,r}^{ML}$	-0.35	-1.84	1.10	2.49	11.06	2.34
$t_{y,r}^{WML}$	-0.43	-1.87	1.55	2.53	11.04	2.76
$t_{y,r}^{Cal1}$	-0.46	-1.89	1.75	2.54	11.03	2.98
$t_y^{Cal1\_pop}$	-0.05	0.40	-0.05	2.38	11.44	0.87
$t_{y,r}^{Cal2}$	-0.16	-1.12	0.93	2.42	11.01	1.69
$t_y^{Cal2\_1}$	-0.03	-0.72	0.48	2.41	11.08	1.27
$t_y^{Cal2\_z\log z}$	0.35	0.26	-0.06	2.31	11.06	0.26
t_y^{Gr}	1.16	1.45	2.32	4.37	11.55	6.02
$t_{y,r}^{Gr}$	-0.49	-0.19	0.57	2.44	10.94	1.97
t_y^{Sgr}	-0.48	-0.63	0.58	2.31	10.23	1.04

[1] (Drug-related hospital visits)$^{1.3}$

References

Chang, T. and Kott, P.S. (2008). Using Calibration Weighting to adjust for Nonresponse Under a Plausible Model, *Biometrika*, 95, 557-571.

Deville, J-C. Särndal, C-E. (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 87, 376-382.

Diaz-Tena, N., Potter, F., Sinclair, M. and Williams, S. (2002). Logistic Propenisty Models to Adjust for Nonresponse, ASA Proceedings of the Survey Research Methods Section, 776-781.

Folsom, R. E. and Singh, A. C. (2000). The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification, *Proceedings of the American Statistical Association, Survey Research Methods Section*, 598-603.

Kim, J.K. and Riddles, M. (2012). Some Theory for Propensity Scoring Adjustment Estimator, under review by *Survey Methodology*.

Kott, P.S. (2011). Why One Should Incorporate the Design Weights When Adjusting for Unit Nonresponse Using Response Homogeneity Groups, *Survey Methodology*, forthcoming.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors, *Survey Methodology,* 32, 133-142.

Little, R.J. (1986). Survey Nonresponse Adjustments, *International Statistical Review*, 54, 139-157.

Little, R., and Vartivarian, S. (2003). On Weighting the Rates in Non-response Weights. *Statistics in Medicine*, 22, 1589-1599.

RTI International (2008). *SUDAAN Language Manual, Release 10.0.* Research Triangle Park, NC: RTI International.

United States Department of Health and Human Services (2011). *Drug Abuse Warning Network (DAWN), 2008* (Computer file of survey conducted by the Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality). Ann Arbor, MI: Inter-university Consortium for Political and Social Research.