Adjusting for Unit Nonresponse in an Establishment Survey Under a Logistic Response Model

Phillip S. Kott
pkott@rti.org

Outline

- Motivating example
- Quasi-randomization
- Weighting-class adjustments and poststratification
- A logistic response model
- Calibration weighting
- Instrumental variables
- SUDAAN 11 (featuring a numerical example)
- Concluding remarks

Motivating Example

A fictional *stsrs* of 364 hospital emergency departments (EDs), stratified by region, size class, ownership (public/private), urbanicity.

Survey weight for an ED is the inverse of its selection probability (d_k)

Key survey variable: drug-related ED visits in current year (y_k)

A size measure is available on the frame:

all ED visits in a previous year (q_k)

Unit (whole ED) nonresponse is generated as a logistic function of the *log* of the survey variable (roughly 45% response).

Quasi-randomization

- Unit response is treated as an additional phase of random sampling.

Each unit k has an estimable probability of response, p_k,

which is a function of covariates and independent across units.

Under mild conditions, a nearly unbiased for $T = \sum_U y_k$ is

$$t = \sum_R d_k \left(\frac{1}{\hat{p}_k} \right) y_k ,$$

where $\hat{p}_k$ is a nearly unbiased estimator for p_k

Weighting-class Adjustment and Poststratification

Suppose the population can be divided into G mutually exclusive groups or classes, like the design strata, such that each unit in a group has the same probability of response if sampled.

When the probability of response within group g is estimated by

$$\hat{p}_k = \frac{\sum_{R_g} d_k \leftarrow \text{estimated number of units in } g \text{ that would respond if sampled}}{\sum_{S_g} d_k \leftarrow \text{estimated number of units in } g \text{ computed from the sample } (\hat{N}_g)},$$

we have a *weighting-class estimator:*

$$t_{wc} = \sum_{g=1}^{G} \sum_{k \in R_g} d_k \left(\frac{\sum_{j \in S_g} d_j}{\sum_{j \in R_g} d_j} \right) y_k = \sum_{g=1}^{G} \hat{N}_g \frac{\sum_{R_g} d_k y_k}{\sum_{R_g} d_k}$$

When the probability of response within group g is estimated by

$$\hat{p}_k = \frac{\sum_{R_g} d_k \leftarrow \text{estimated number of units in } g \text{ that would respond if sampled}}{N_g \leftarrow \text{number of units in } g},$$

we have a *post-stratified estimator*.

$$t_{ps} = \sum_{g=1}^{G} \sum_{k \in R_g} d_k \left(\frac{N_g}{\sum_{j \in R_g} d_j} \right) y_k = \sum_{g=1}^{G} N_g \frac{\sum_{R_g} d_k y_k}{\sum_{R_g} d_k}.$$

Although r_g/n_g is a better estimator of the group-g response rate, the weighting-class estimator usually provides a better estimator for T than using the $\hat{p}_k = r_g/n_g$ within t. The poststratified estimator is better still.

A Logistic Response Model

A more general unit response model allows response to be a logistic function of a vector of covariates: $\mathbf{z}_k$

$$\log\left(\frac{p_k}{1-p_k}\right)=\text{logit}(p_k)=\mathbf{z}_k{}^T\boldsymbol{\gamma} \quad \text{or}$$

$$p_k=\text{logit}^{-1}(\mathbf{z}_k{}^T\boldsymbol{\gamma})=\frac{\exp(\mathbf{z}_k{}^T\boldsymbol{\gamma})}{1+\exp(\mathbf{z}_k{}^T\boldsymbol{\gamma})}.$$

where $\boldsymbol{\gamma}$ is unknown but can be estimated.

For example, suppose $\mathbf{z}_k$ were the vector $(1 \;\; I_{public,k} \;\; q_k)^T$, where $I_{public,k} = 1$ when k is publicly owned, 0 otherwise.

Then

$$\log\left(\frac{p_k}{1-p_k}\right) = \gamma_1 + \gamma_2 I_{public,k} + \gamma_3 q_k$$

would mean that, given the ownership status, a 1 unit increase in q_k produces a γ_3 percent increase in the odds of k responding; while, given the size of q_k, being public results in a γ_2 percent increase in the odds of responding.

The standard way to estimate $\boldsymbol{\gamma}$ is with weighted logistic regression, which finds an $\mathbf{h}$ (to estimate $\boldsymbol{\gamma}$) such that

$$\sum_S d_k \left[R_k - \text{logit}^{-1}(\mathbf{z}_k{}^T \mathbf{h}) \right] \mathbf{z}_k = \mathbf{0},$$

where R_k is 1 when k responds and 0 otherwise.

We can then set $\hat{p}_k = \text{logit}^{-1}(\mathbf{z}_k{}^T \mathbf{h})$.

Estimating the standard error of the resulting t (assuming the response model has the correct form) is not trivial.

Calibration Weighting

Alternatively, we can find an **h** such that

$$\sum_S d_k \frac{R_k}{\text{logit}^{-1}(\mathbf{z}_k{}^T\mathbf{h})}\mathbf{z}_k = \sum_S d_k \mathbf{z}_k \quad \text{or} \quad \sum_U \mathbf{z}_k .$$

↑ sample ↑ population

This is called a *calibration equation* (with calibration to the sample or to the population).

$$w_k = d_k \frac{R_k}{\text{logit}^{-1}(\mathbf{z}_k{}^T\mathbf{h})}$$ is a *calibration weight.*

When, for example, $\mathbf{z}_k = (1 \;\; I_{public,k} \;\; q_k)^T$,

there is an individual calibration equation for each component of the vector $\mathbf{z}_k$.

$$\sum_R w_k = \sum_S d_k \;\; (\text{or } N)$$

$$\sum_R w_k I_{public,k} = \sum_S d_k I_{public,k} \;\; (\text{or } N_{public})$$

$$\sum_R w_k q_k = \sum_S d_k q_k \;\; (\text{or } \sum_U q_k)$$

Calibration weighting will produce an estimator for T with a smaller standard error that using the result of weighted-logistic-regression fit when the survey variable is roughly a linear function of the components of $\mathbf{z}_k$.

Calibration to the population will have less standard error than calibration to the sample.

Instrumental Variables

Suppose a more reasonable response model is

$$p_k = \text{logit}^{-1}(\mathbf{x}_k{}^T \boldsymbol{\gamma}),$$

where *some* components of the **model** vector $\mathbf{x}_k$

do not coincide with the **calibration** vector $\mathbf{z}_k$

(but the two vectors have the same size). We can solve:

$$\sum_R w_k \mathbf{z}_k = \sum_R \frac{d_k}{\text{logit}^{-1}(\mathbf{x}_k{}^T \mathbf{h})} \mathbf{z}_k = \sum_S d_k \mathbf{z}_k \text{ or } \sum_U \mathbf{z}_k .$$

In establishment surveys, it often makes sense to calibrate to a size variable (like ED visits in a previous year) because the main survey variable (drug-related ED visits in the survey year) is nearly linear in the size variable.

But response is better modeled as a logistic function of the ***log of the size variable***, so that a one percent increase in the size variable results in a c percent change in the odds of response.

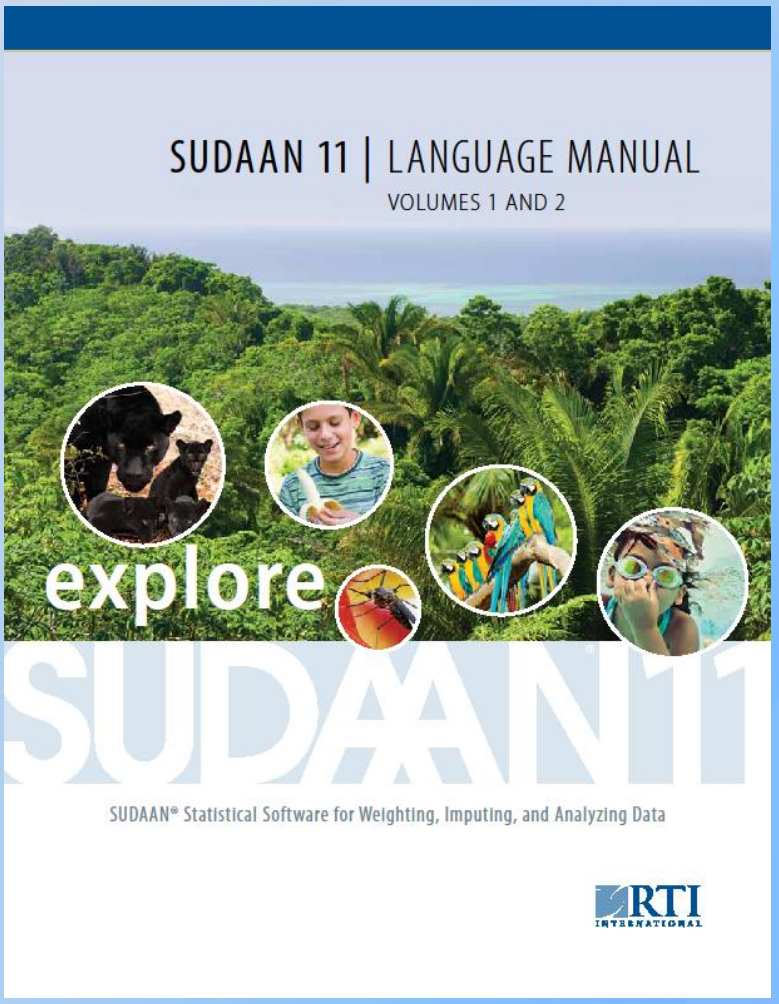
Thus, $\log(q_k)$ should be an instrumental variable used in place of q_k.

Deville (COMPSTAT - Proceedings in Computational Statistics: 14th Symposium held in Utrecht, 2000) observed that the values of an instrumental variable need only be known for respondents.

That means by using instrumental variables in the calibration weighting *one can adjust for nonresponse that is not be missing at random* – as long as there are as many calibration variables as there are explanatory variables in the response model (i.e., instrumental variable).

Chang and Kott (*Biometrika*, 2008) expanded on that idea.

Instrumental-variable (IV) calibration under a logistic response model can be done using the WTADJX procedure in SUDAAN 11.



SUDAAN 11

SUDAAN 11 will also produce appropriate large-sample standard errors when there is one round of calibration or logistic-regression reweighting.

When the response model is assumed to be logistic, one can use WTADJUST (when the calibration variables are the model variables) or WTADJX (otherwise) with a lower bound of 1, a center of 2, and no upper bound.

Other bounds can be used to fit a truncated logistic response model.

Results 1

Assuming first that response is a logistic function of the log of the *size measure*, we estimated the survey-variable total and its large-sample standard error using the following methods:

Method 1: Logistic regression (RLOGIST) with $\mathbf{z}_k = (1 \;\; \log(q_k))^T$

Method 2: Calibration (WTADJUST) to the sample with same $\mathbf{z}_k$

Method 3a: IV Calibration (WTADJX) to the sample with $\mathbf{x}_k = (1 \;\; \log(q_k))^T$ and $\mathbf{z}_k = (1 \;\; q_k)^T$

Method 3b: IV Calibration (WTADJX) to the population with same $\mathbf{x}_k$ and $\mathbf{z}_k$

We computed the large-sample standard errors in SUDAAN 11 and converted them into CVs

(one need not collapse strata in WTADJUST or WTADJX even when less than two respondents in a stratum).

Using RLOGIST	CV = 7.33
Using WTADJUST	CV = 8.30
Using WTADJX	
calibrating to the sample	CV = 6.39
calibrating to the population	CV = 3.40

Results 2

We can also test whether there is a significant difference between estimates derived under different assumed response models.

In this case, the estimated bias (roughly 1.2%) from incorrectly assuming response is a logistic function of the log of the *frame* variable (EDs visits in a previous year) rather than the log of the *survey* variable (drug-related visits in the survey year) is significant at the .08 level.

Even when we don't know the true response model, the test – duplicating each record, assigning the first version to a domain governed by one assumed response model and the second to a domain governed by a different assumed model *while keeping both in the same PSU* – can be used to determine whether different response models lead to significantly different estimates.

Replicate-based variance estimation could also be used to test whether different response model produce significantly different estimates.

The SUDAAN website contains the mostly made-up sample data orginally derived from the Drug Abuse Warning Network (DAWN) public-use file and used to produce in the numerical results featured here.

The second WTADJX example on the site develops the SAS callable SUDAAN code employed to generate those results.

Concluding Remarks

Although the adjusted weights are the same regardless of the survey variable, the effect of weight adjustment on standard error varies across survey variables.

Weight adjustment is less appealing for item nonresponse.

Calibrating to the population is more efficient than calibrating to the full sample. Nevertheless, it is often better to calibrate in two steps.

Kott and Day (ICES IV *Proceedings*, 2012) describes how that could be done with an actual DAWN survey.