Two (Not So) Different Approaches for Dealing with Survey Nonresponse That Is Not Missing at Random

Phillip S. Kott

Outline

Developing My Favorite Equation

The Outcome (Prediction) Model with Ignorable Nonresponse

The Response (Quasi-random) Model with Missingness at Random

The Response Model with Missingness Not at Random

The Outcome Model with Nonignorable Nonresponse

The Pattern Mixture Model

Discussion

Developing My Favorite Equation

Bias from Nonresponse Under Simple Random Sampling:

$$\text{Bias} = \sum_{k\in R} \frac{y_k}{r} - \sum_{k\in S} \frac{y_k}{n} = \sum_{k\in S} \frac{\left(R_k - \frac{r}{n}\right) y_k}{r} =$$

$$\sum_{k\in S} \frac{\left(R_k - \frac{r}{n}\right)}{r} (y_k - \bar{y}) = \sum_{k\in S} \frac{1}{n} \frac{\left(R_k - \frac{r}{n}\right)}{\frac{r}{n}} (y_k - \bar{y}).$$

Adding weights ($d_k = \frac{1/\pi_k}{\sum_{i\in S} 1/\pi_i}$):

$$\text{Bias} = \sum_{k\in S} d_k \frac{(R_k - \tilde{E}(R_k))}{\tilde{E}(R_k)} (y_k - \bar{y}).$$

RTI INTERNATIONAL

Developing My Favorite Equation (2)

$$\text{Bias} = \sum_{k \in S} d_k \frac{(R_k - \tilde{E}(R_k))}{\tilde{E}(R_k)} (y_k - \bar{y})$$

$$= \sum_{k \in S} d_k \left(\frac{R_k}{\tilde{E}(R_k)} - 1 \right) (y_k - \bar{y})$$

$$= \sum_{k \in S} d_k (R_k g_k - 1) \left(y_k - \mathbf{z}_k^T \boldsymbol{\delta} \right),$$

where $g_k = 1/\tilde{E}(R_k)$, and the *calibration equation*:

$\sum_{k \in S} d_k R_k g_k \mathbf{z}_k = \sum_{k \in S} d_k \mathbf{z}_k$ holds.

An Example

Consider a probability sample of hospital emergency departments.

Let y_k = the current number of annual drug-related visits to k

$$\mathbf{z}_k = \begin{pmatrix} z_{k1} \\ \vdots \\ z_{k5} \end{pmatrix},$$

z_{kj} = 1 if k in region j, 0 otherwise (j = 1, …, 4), and

z_{k5} = the number of annual ED visits to k in the frame year

The Outcome Model

The estimator $\sum_{k \epsilon S} d_k R_k g_k y_k$ is nearly unbiased in some sense when the *output model* with ignorable nonresponse holds:

$y_k = \mathbf{z}_k^T \boldsymbol{\beta} + \varepsilon_k$; $\mathrm{E}(\varepsilon_k | \mathbf{z}_k, R_k) = 0$,

and $g_k = 1/\tilde{E}(R_k)$ can be *any* function of $\mathbf{z}_k$,

but not $y_k | \mathbf{z}_k$.

For example:

$$g_k = \left[1 + \left(\sum_{i \epsilon S} d_i \mathbf{z}_i^T \sum_{i \epsilon S} d_i R_i \mathbf{z}_i^T\right)\left(\sum_{i \epsilon S} d_i R_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \mathbf{z}_k\right]$$

-

$$g_k = \left[1 + \left(\sum_{i \in S} d_i \mathbf{z}_i^T - \sum_{i \in S} d_i R_i \mathbf{z}_i^T\right)\left(\sum_{i \in S} d_i R_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \mathbf{z}_k\right]$$

implies

$$\sum_{k \in S} d_k R_k g_k y_k = \left(\sum_{k \in S} d_k \mathbf{z}_k\right)^T \left(\sum_{i \in S} d_i R_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \sum_{i \in S} d_i R_i \mathbf{z}_i y_i$$

$\mathbf{b}_{dR\mathbf{z}}$

So long as $\mathbf{z}_k$ contains a 1 or the equivalent.

The Response Model

The estimator $\sum_{k \in S} d_k R_k g_k y_k$ is nearly unbiased in some sense when $E(R_k) = 1/\gamma(\mathbf{x}_k)$, and $\gamma(\mathbf{x}_k)$ is consistently estimated by $g(\mathbf{x}_k) = g_k$,

but

$$g_k = \left[1 + \left(\sum_{i \in S} d_i \mathbf{z}_i^T - \sum_{i \in S} d_i R_i \mathbf{z}_i^T\right)\left(\sum_{i \in S} d_i R_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \mathbf{z}_k\right]$$

$= \quad 1 + \lambda^T \mathbf{z}_k$ is not always sensible

$g_k = \exp(1 + \lambda^T \mathbf{z}_k)$,

where the g_k satisfy $\sum_{k \in S} d_k R_k g_k \mathbf{z}_k = \sum_{k \in S} d_k \mathbf{z}_k$,

is more so.

RTI INTERNATIONAL

Missing (Not) at Random

- Suppose $E(R_k) = 1/\gamma(\mathbf{x}_k) = R(\mathbf{x}_k)$, such as $1/\exp(1 + \lambda^T \mathbf{x}_k)$

  If the components of $\mathbf{x}_k$ are functions of the components of $\mathbf{z}_k$ but not of $y_k|\mathbf{z}_k$, then missingness is said to be at random.

  If a component of $\mathbf{x}_k$ is a function of $y_k|\mathbf{z}_k$, then missingness is said to be *not* at random.

  So long as the dimension of $\mathbf{x}_k$ (the *model variables*) equals that of $\mathbf{z}_k$ (the *calibration variables*) and
  $\sum_{k \in S} d_k R_k \frac{1}{E(R_k)} \mathbf{z}_k = \sum_{k \in S} d_k \mathbf{z}_k$ is solvable, then
  $\sum_{k \in S} d_k R_k g_k y_k$ is nearly unbiased.

Nonignorable Nonresponse

Suppose this outcome model holds

(with y_k a component of $\mathbf{x}_k$, which is otherwise as before):

$$y_k = \mathbf{z}_k^T \boldsymbol{\beta} + \varepsilon_k; \;\; E(\varepsilon_k | \mathbf{x}_k, R_k) = 0 \text{ , then}$$

$$\sum_{k \in S} d_k R_k g_k y_k = \left(\sum_{k \in S} d_k \mathbf{z}_k\right)^T \left(\sum_{i \in S} d_i R_i \mathbf{x}_i \mathbf{z}_i^T\right)^{-1} \sum_{i \in S} d_i R_i \mathbf{x}_i y_i$$

$$\mathbf{b}_{dR\mathbf{x}}$$

is nearly unbiased so long as a component of $\mathbf{x}_k$ is 1.

Pattern-Mixture Modeling

Recall: $\sum_{k \in S} d_k \frac{(R_k - \tilde{E}(R_k))(y_k - \mathbf{z}_k^T \boldsymbol{\delta})}{\tilde{E}(R_k)}$

Let $\hat{y}_k = \mathbf{z}_k^T \mathbf{b}_{dRz}$

Pattern 1: Nonresponse is a function of $\hat{y}_k$ but not $y_k | \hat{y}_k$

Pattern 2: Nonresponse is a function of y_k (but not $\hat{y}_k |\, y_k$)

Pattern Mixture: Nonresponse is a function of $\alpha y_k + (1 - \alpha)\, \hat{y}_k$

Discussion

The Outcome Model (and Pattern-Mixture Model) depends on a single survey variable.

When the lone survey variable is not linear, we can replace $\mathbf{z}_k$ with $\begin{pmatrix} 1 \\ \hat{y}_k^{fitted} \end{pmatrix}$ and (at the nonignorable extreme) $\mathbf{x}_k$ with $\begin{pmatrix} 1 \\ y_k \end{pmatrix}$.

SUDAAN generalizes logistic response to:

$$R(\mathbf{x}_k) = \frac{1+\exp(\boldsymbol{\gamma}^T\mathbf{x}_k)/U_w}{L_w+\exp(\boldsymbol{\gamma}^T\mathbf{x}_k)}$$

Discussion

Often the vector of calibration variables contains totals (or means) for the entire population, as well as estimated totals (or means) for the sample before nonresponse.

SUDAAN's calibration-weighting PROCs can handle either, but not both.

What is the purpose of the design (inverse probability) weights?

They are like the $1/R(x_k)$, except they don't have to be estimated, but we do not know if the are functions of the y_k.

Adding a Randomized Device to the Response Model

Suppose we gave a random fraction of the probability sample an incentive to respond.

Let $c_k = 1$ if k gets the incentive, 0 otherwise.

Calibrating to the full sample:

Let the new $\mathbf{z}_k$ vector be $\begin{pmatrix} \mathbf{z}_k^{old} \\ c_k \mathbf{z}_k^{old} \\ y_k(c_k - \bar{c}) \end{pmatrix}$

$\mathbf{z}_k^{old}$ includes a 1; $\sum_{k \in S} d_k\, y_k(c_k - \bar{c}) = 0$

and the new $\mathbf{x}_k$ vector be $\begin{pmatrix} \mathbf{z}_k^{old} \\ c_k \mathbf{z}_k^{old} \\ y_k \end{pmatrix}$.

Some of the components of $c_k \mathbf{z}_k^{old}$ may need to be dropped