Discussion: Inference with Non-Probability Sample Through Data Integration

Phillip S.Kott[1]
[1]RTI International; 6110 Executive Blvd; Rockville, MD 20852

Let me start by thanking Sixia Chen for inviting me to discuss these four fine papers. Two (the ones presented by Chen and by David Haziza) are on imputing for an absent variable on a probability survey given a nonprobability sample with that variable and a vector of common variables. One paper (presented by Lingxiao Wang) is on weighting a nonprobability sample when there is a probability sample with a vector of common covariates. The final (presented by Danhyang Lee), an outlier, is on imputing for missing variables in a probability survey. Because of time constraints and the limited abilities of this discussant, I will not give any of the papers the justice it deserves.

The outlier paper (Lee and Kim; our chair Jae-Kwang Kim had a hand in writing two of the four papers) imputes for missing values using a conditional Gaussian mixture model. Survey data is rarely Gaussian, but to me that assumption may be a good enough approximation for imputation. I have a bigger problem in assuming data is multivariate Gaussian, but that may not be so much of a problem when the population in divided up like it is.

One suggestion I have is to use the cross-validation variance estimator in You (2009) as a method for choosing G rather than the Bayesian Information Criterion (BIC). You's variance estimator is the usual linearization variance estimator for an estimator from a complex sample but it replaces differences between sampled values and predicted values in the usual linearization variance estimator with differences between sampled values and predicted values computed without the associated sampled value.

The paper by Chen, Yang, and Kim first points out that the simplest way to impute for a missing variable in a probability sample is to use its predicted value from the nonprobability sample. This requires one to assume that a known parametric model holds in the population from which they both derive. When, as is often the case, the nonprobability sample is much larger than the probability sample, a more robust alternative applies kernel estimation to the nonprobability sample. One of the contributions of this paper, which I will not discuss, is how to estimate the variance of the resulting estimation.

Because kernel estimation suffers from the curse of dimensionality an alternative using a general additive model is also proposed by Chen and company. I wonder whether the two approach could be combined with the predicted value from the general additive model used as the single covariate in kernel estimation.

The paper from Wang and her coauthors (Graubard, Katki, and Li) is itself a bit or an outlier. It is not so much on parameter estimation but on fitting a model: $y_k = f(\mathbf{x}_k^T\boldsymbol{\beta}) + \varepsilon_k$. Here, the nonprobability sample will be used to fit the model, that is, finding a $\boldsymbol{\beta}$

such that $E(\varepsilon_k \mid \mathbf{x}_k) = 0$ for all possible $\mathbf{x}_k$, but it needs to be weighted – with the help of the probability sample – for the results to apply to the population. The authors correctly point out that even if we assume the same model fits the nonprobability sample and the population, weights are needed if either the probability a unit has been selected for the nonprobability sample is informative (i.e., depends on the dependent variable of the model, y_k, even after accounting for the independent variables, $\mathbf{x}_k$) or the model fails. In the latter case, the question becomes, how should one fit a model when it fails? The authors treat fitting a maximum-likelihood estimating equation for the entire population as the goal of model fitting when a model fails. This can be done by weighting the maximum-likelihood estimating equation. I would instead replace the *standard model* assumption $E(\varepsilon_k \mid \mathbf{x}_k) = 0$ with the less binding *general model* assumption, $E(\varepsilon_k \mathbf{x}_k) = \mathbf{0}$ (Kott 2018). My approach and the authors' produce identical results for the linear model with independent and identically distributed errors and for a logistic model with independent errors, but not always (see, for example, Kott and Frechtel 2018). Under either approach, weights are needed to find a consistent estimator for $\boldsymbol{\beta}$.

To produce quasi-probability weights for a nonprobability sample, denoted here by S_0, it has become popular to combine it with a probability sample S_1 covering the same population U. After that, one estimates the probability γ_k that a population unit k in the blended sample $S = S_0 \cup S_1$ was originally from the nonprobability sample given a vector of covariates $\mathbf{x}_k$ available for members from both samples (when used here, "sample" always refers to a respondent sample). Valliant and Dever (2011) suggest that the inverse of this estimated probability – which they, following much of the literature, call a "propensity" – can be used as a quasi-probability sampling weight either directly, $w_k = 1/\hat{\gamma}_k$, or indirectly after some form or poststratification. For example, Lee (2006) proposes sorting the blended sample by their $\hat{\gamma}_k$ values, then breaking the sample into cells of nearly equal size, and finally assigning to k the poststratified weight $w_k = \hat{N}_{1c}/n_{0c}$, where $\hat{N}_{1c}$ is the estimated-from-the-probability-sample population size of cell c containing nonprobability-sampled unit k, and n_{0c} is the number of members of S_0 in c .

Although estimating $\gamma_k = \Pr(k \in S_0 \mid k \in S\ ;\ \mathbf{x}_k)$ by fitting a logistic regression on $\mathbf{x}_k$ in the blended sample is often treated as an estimate for $\Pr(k \in S_0 \mid \mathbf{x}_k)$, Robbins (2017) argues that a better estimate for the quasi-probability that k is in the nonprobability sample when $S_0 \cap S_1 = \varnothing$ is $p_{0k} = \Pr(k \in S_0 \mid \mathbf{x}_k) = \pi_k \hat{\gamma}_k/(1-\hat{\gamma}_k)$, where π_k is the known probability that k is chosen for the probability sample (which can include an adjustment for unit nonresponse when needed). It is assumed π_k is also known for members in the population that are not in S_1.

To see how $p_{0k} = \Pr(k \in S_0 \mid \mathbf{x}_k)$ is derived, start with $\pi_k = \Pr(k \in S_1 \mid k \in S\ ;\ \mathbf{x}_k) \times \Pr(k \in S \mid \mathbf{x}_k)$, and $\Pr(k \in S_0 \mid \mathbf{x}_k) = \Pr(k \in S_0 \mid k \in S;\ \mathbf{x}_k) \times \Pr(k \in S \mid \mathbf{x}_k)$), then solve for $\Pr(k \in S_0 \mid \mathbf{x}_k)$. Elliott and Valliant (2017) make a similar point, suggesting a more sophisticated method could be used in estimating γ_k.

In discussions with the authors, they pointed out that for them estimating the γ_k was only a means to an end – creating kernel weights (with $\hat{\gamma}_j$ and $1 - \hat{\gamma}_i$ employed in defining a distance measure between $\mathbf{x}_j$ for $j \in S_0$ and $\mathbf{x}_i$ for $i \in S_1$). I would have used the inverse of Robbins's p_{0k} in place of their kernel weight for $k \in S_0$. (The paper discusses using

either a direct or a poststratified version of an estimated probability of k being in S_0 based on replacing S in $\gamma_k = \Pr(k \in S_0 \mid k \in S; \mathbf{x}_k)$ with the union of S_0 and a weighted version of S_1, the poststratified method attributed to Lee and Valliant 2009. That union, however, is purely mythical.) Better still would be to estimate and then invert the probability of drawing a unit into the nonprobability sample with a calibration equation as described in Kott (2019) and below.

Suppose the selection model for unit k being in the nonprobability sample is a logistic function of $\mathbf{x}_k$:

$$\Pr(k \in S_0 \mid \mathbf{x}_k) = 1/[1 + \exp(\mathbf{x}_k^T \boldsymbol{\gamma})].$$

One can estimate $\boldsymbol{\gamma}$ in is a consistent manner under mild conditions we assume to hold by a finding a $\mathbf{g}$ that satisfies the calibration equation:

$$\sum_{k \in S_0} [1 + \exp(\mathbf{x}_k^T \mathbf{g})]\ \mathbf{x}_k = \sum_{k \in S_1} \mathbf{x}_k / \pi_k \qquad (1)$$

The quasi-probability weight for $k \in S_0$ is then $w_k = 1 + \exp(\mathbf{x}_k^T \mathbf{g})$. The WTADJST procedure in SUDAAN 11 (RTI International 2012) can be used to fit equation (1). There are also packages in R that can be used for this purpose.

Observe that $t_{y0} = \sum_{S0} w_k y_k$ is an unbiased estimator for $T_y = \sum_U y_k$ in some sense if either the selection model holds or the following linear prediction (outcome) model holds: $E(y_k) = \mathbf{x}_k^T \boldsymbol{\beta}$, whether or not unit k is in the sample. This property is sometimes called "double robustness," although I prefer "double protection against selection bias." The property also obtains when some of the estimated totals for some of the components of $\mathbf{x}_k$ on the right-hand size of the calibration equation in (1) are replaced by the actual population totals for by estimates for a different probability sample.

The paper by Chen and Haziza (presented by David Haziza; observe that our organizer like our chair also had a hand in writing two of the papers in this section) is on multiple robustness. My own view – based on my interpretation of D'Arrigo and Skinner (2010) where the wrong selection model coupled with a flawed prediction model appears to remove most of the selection bias due to nonresponse – is that double protection may be enough for handling nonresponse. For nonprobability samples, however, multiple protection may indeed be needed. Rather than discuss that the Chen and Haziza paper *per se*, I will offer my own version of multiple protection concentrating entirely on weighting the nonprobability sample in order to estimate T_y.

Suppose we have J candidate selection models and J candidate outcome models. If anyone of the $2J$ candidate models holds, then t_{y0} is nearly unbiased for T_y in some sense (adding dummy candidate models when needed to equalize the number of selection and prediction models should not be difficult). Multiple protection obtains under mild conditions we assume to hold by finding an $\mathbf{h}$ that satisfies the calibration equation:

$$\sum_{k \in S_0} \exp\left[\sum_{j=1}^{J} h_j \log(\hat{p}_{kj})\right] \mathbf{m}_k = \sum_{k \in S_1} \mathbf{m}_k / \pi_k \qquad (2)$$

where p_{kj} is the estimated probability that unit $k \in S_0$ in the j^{th} potential selection model, and $\mathbf{m}_k = (m_{k1} \ \ldots, \ m_{kJ})^T$, where each component fits a different one of $j = 1, \ldots, J$

prediction models; that is $E_j(y_k) = m_j(\mathbf{x}_k) = m_{kj}$. The model fits are made in S_0 without weights. The multiply-protected weight for $k \in S_0$ is $w_k = \exp\left[-\sum_{j=1}^{J} h_j \log(\hat{p}_{kj})\right]$.

Each of the J prediction models implicitly assumes that selection into the nonprobability model is ignorable (given $\mathbf{x}_k$) and that the error structure of the prediction model and the probability-sample design is such that weighted sum in the probability sample S_1 of the m_{kj} is a consistent estimator for the weighted sum of the y_k; that is,

$$\text{rel dif}\left(\sum_{k\in S_1}\frac{y_k}{\pi_k}, \sum_{k\in S_1}\frac{m_{kj}}{\pi_k}\right) = \frac{\sum_{k\in S_1}\left(\frac{y_k}{\pi_k} - \frac{m_{kj}}{\pi_k}\right)}{\sum_{k\in S_1}\frac{m_{kj}}{\pi_k}}$$

converges to 0 in probability as the sample grows arbitrarily large.

Recall that at most only one model needs to hold in the population. Note that even if an ignorable prediction model holds, weights may be needed to estimate T_y with t_{y0} in a nearly unbiased fashion.

The expression on the left-hand side of the calibration equation in (2) was chosen because it can be fit using the WTADJX routine in SUDAAN 11 (2012) (the routine would be applied to S_0 with the $\log(p_{kj})$ as the model variables, the m_{kj} as the calibration variables, and the summation on right-hand side of (2) serving as the poststratification totals). Observe that if the j^{th} selection model is correct *and the only selection model fitting all J* $\sum_{k\in S1} m_{kj}/\pi_k$, then h_j should be close to -1, while the other h_g values should be close to 0 (technically, h_j converges to -1, and h_g, for $g \neq j$, converges to 0 as the sample grows arbitrarily large).

For an example of equation (2), let y_k be a binary (0/1) variable, while $\mathbf{x}_k = (1\ z_k)^T$, where z is continuous. Now suppose we have two candidate selection and two candidate prediction models:

$$p_k = 1/[1 + \exp(\gamma_{11} + \gamma_{12} z_k)] \quad (3)$$
$$p_k = 1/[1 + \exp(\gamma_{21} + \gamma_{22} \log[z_k])] \quad (4)$$

$$y_k = 1/[1 + \exp(\beta_{11} + \beta_{12} z_k)] \quad (5)$$
$$y_k = 1/[1 + \exp(\beta_{21} + \beta_{22} \log[z_k])]. \quad (6)$$

We can estimate the parameters of the first selection model (3) using equation (1), and then estimate the second selection model (4) by again using equation (1) but with $\mathbf{x}_k$ replaced by $(1\ \log[z_k])^T$. The fitted values from equation (3) $(1/[1 + \exp(g_{11} + g_{12} z_k)]$, where g_{ab} is the estimate for γ_{ab}) are the p_{k1} in equation (2) while the fitted values from equation (4) are the p_{k2}. We can estimate the parameters of the two prediction models by running logistic regressions on the nonprobability sample. The fitted values from equation (5) $(1/[1 + \exp(b_{11} + b_{12} z_k)]$, where b_{ab} is the estimate for β_{ab}) are the m_{k1} in both the nonprobability and probability samples of equation (2), while the fitted values from equation (6) are the m_{k2} in both samples.

A limitation of this approach to multiple protection is that just as it is possible to find a $\mathbf{g}$ that fits equation (1) without selection into S_0 really being a logistic function of $\mathbf{x}_k$, it is possible that more than one of the J candidate selection models fits all J $\sum_{k\in S1} m_{kj}/\pi_k$

even though only one, if that many, is really the selection model for members of S_0. As a result, equation (2) will likely not have a single solution.

References

D'Arrigo, J. and C. Skinner (2010). "Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse." *Survey Methodology*, 36, 181-192.

Elliott, M. and R. Valliant (2017). "Inference for Non-Probability Samples." *Statistical Science*, 32, 249–264.

Kott, P. (2019). "Partially Successful Attempt to Integrate a Web-recruited Cohort into an Address-based Sample." *Survey Research Methods*, 13, 95-101.

Kott, P. (2018). "A Design-sensitive Approach to Fitting Regression Models with Complex Survey Data." *Statistics Surveys*, 12, 1–17.

Kott, P. and P. Frechtel (2019). "An Alternative Way of Estimating a Cumulative Logistic Model with Complex Survey Data." *Survey Methodology*, 45, 339-347.

Lee, S. (2006). "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics*, 22, 329-49.

Lee, S. and R. Valliant (2009). "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research*, 37, 319–343.

Robbins, M. (2017). "Blending of Probability and Convenience Samples: Applications to a Survey of Military Caregivers." Presented as the NISS/WSS Workshop on Inference from Nonprobability Samples. Available at https://www.niss.org/sites/default/files/event_attachments/RobbinsSlidesHandout.pdf (a paper is available from the author upon request).

RTI International (2012). *SUDAAN Language Manual, Release 11.0*. Research Triangle Park, NC: RTI International.

Valliant, R. and J. Dever. (2011). "Estimating Propensity Adjustments for Volunteer Web Surveys." *Sociology, Methods and Research*, 40, 105–137.

You, L. (2009). *Cross-validation in Model-assisted Eestimation. Graduate Theses and Dissertations.* Iowa State University. Available at https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1509&context=etd