Bayesian nonparametric imputation of missing design information under informative survey samples

Anonymous Author(s)
Affiliation
Address
email

Introduction

Surveys are conducted at local, national, and international levels, to gather information and aid public and private sectors in effective policy making. The objective of a survey is usually to obtain summary statistics for the finite population or for specific subgroups. Sampling theory is used to design surveys that achieve best possible accuracy under a fixed budget. Probability sampling, which each member of the population has a non-zero probability of inclusion in the sample, is desirable in that it eliminates selection bias (Särndal et al., 2003). Often times though, the optimal designs are those where the units are chosen with unequal probability of selection.

In populations with differing size units, it is often the case that larger units contribute more to population quantities of interest than smaller units. A popular design that includes larger units with higher probability is sampling with probability proportional to size (PPS). The size variables x, govern the selection mechanism, and must be known for all units in the population at the time of sampling. However, since the nonsampled sizes are not required for the classical Horvitz-Thompson or Hajek estimators, this information is rarely included in public use data files (Pfeffermann et al., 1998). We ultimately aim at estimating finite population quantities $Q(Y)$ of a continuous outcome Y, from a PPS sample of size n (fixed), in situations where the size measures are only reported for the sampled cases, but the number of nonsampled cases $N - n$ and $T_x = \sum_{i=1}^{N} X_i$, is known. Under PPS sampling, the sample design becomes informative (Sugden and Smith, 1984), when the sizes of the nonsampled units are unknown, and hence need to be adjusted for the effects of selection. Studies have shown that incorporating the auxiliary or design information could lead to more efficient estimates of $Q(Y)$. When the design variables are missing for the nonsampled units, an essential intermediate step is imputing the design information for the nonsampled units.

The goal in imputation, is to 'fill in' the missing values using the mean, or draws from a predictive distribution of the missing values, with the objective of preserving distributional properties of the original data. Predictive distributions of the missing values may be obtained by explicit, implicit, or composite models based on the observed data (Little and Rubin, 1987). Explicit modeling relies on a formal statistical model, while the focus of implicit modeling is primarily on algorithms. Both modeling approaches require careful assessment of some underlying assumptions.

Methods

We propose a general imputation framework which is based on considering two separate exchangeable models for the sampled and nonsampled units. Denoting the parameter vector corresponding to the entire population as θ, the inclusion indicator for unit i as I_i, and defining $f(x, e|\theta)$ to be the joint density of (X_i, I_i)

given T_x and indexed by θ, the sampling design, relates the marginal density of the sampled and nonsampled units via

$$f(x \mid 0, \theta) = \frac{1}{N-n}\left(\frac{T_x}{x} - n\right) f(x \mid 1, \theta). \tag{1}$$

According to (1), we show that if the sampling fraction n/N is small and the population distribution of size variables is log-normal, then the size variables observed in the sample will also follow a log-normal distribution, with the mean of the distribution shifted to the right by the variance. Thus, assuming a very small sampling fraction, we use the estimated sample mean and variance of the sampled values to impute the nonsampled sizes by random draws from the parametric predictive distribution of the nonsampled sizes.

When the size variables do not follow a log-normal distribution, the distribution of the nonsampled units is no longer known. In this case, we use Bayesian nonparametric inference via Dirichlet process mixture models (DPMM) to estimate the density of the sampled units; we then obtain the density of the nonsampled units via (1), and use Importance Sampling to draw the nonsampled size variables from the posterior predictive distribution of interest. Nonparametric methods are desirable to predict the sizes of the nonsampled units, as they are robust to model mis-specification and impose minimal prior assumptions on the distribution of the size variables. This is especially desirable, when modeling a marginal distribution, as conditional draws are not possible due to the lack of existing covariate information. Moreover, multiple imputation based on Bayesian principles is a powerful tool for missing data problems, due to its ability to propagate imputation uncertainty (Little and Rubin, 1987).

In this study, we focus on two specific examples of DPMMs. By assuming a discrete distribution for X_i and a Dirichlet/multinomial distribution for θ, we get the Bayesian bootstrap (BB) model (Rubin, 1981), whereas a normal distribution for X_i and a Normal/Inverse Wishart distribution for θ gives rise to a Dirichlet process mixture of normals (DPMN) (Escobar and West, 1995). While the BB model gives rise to a known parametric model for the posterior predictive distribution of nonsampled sizes under PPS sampling (Little and Zheng, 2007), this does not hold in the DPMN model, and hence the latter can be regarded as a composite imputation method. We compare these models along with the parametric model for different population structures through numerical investigations, both in terms of point and interval estimation. Our studies suggest that for general underlying distributions of size variables, DPMMs yield reliable imputations. We also show that while the BB method is superior for moderate and large sample sizes, the DPMN is suitable for small samples.

Finally, we use the imputed size variables within a two-step model, to obtain predictions for the nonsampled survey outcomes $Y = (Y_{N+1}, \ldots, Y_N)$ via a Bayesian penalized spline model, to estimate the finite population mean and median. We show that imputing the missing design information and utilizing them in predicting the survey outcomes gives significant gains in efficiency over traditional methods.

References

Escobar, M. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430).

Little, R. and Rubin, D. (1987). Statistical analysis with missing data.

Little, R. and Zheng, H. (2007). The Bayesian Approach to the Analysis of Finite Population Surveys. *Bayesian Statistics*, 8(1):1–20.

Pfeffermann, D., Krieger, A., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8:1087–1114.

Rubin, D. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

Särndal, C., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Verlag.

Sugden, R. and Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3):495.