On Response Model Selection and its Use for Testing Bias of Mean Estimates in Case of Not Missing at Random Nonresponse

Michael Sverchkov

Bureau of Labor Statistics, Washington DC, USA

Thanks to

Stephen Ash, Jeffrey Gonzalez, Steven Paben, Barry Steinberg and David Swanson

for helpful comments and suggestions

INTRODUCTION

There exists almost no survey without nonresponse.

In practice most ***survey*** methods that ***adjust for nonresponse*** assume either explicitly or implicitly that the missing data are ‘missing at random’ (MAR).

In many practical situations, this assumption is not valid. In such cases, the use of methods that assume nonresponse is MAR can lead to large bias of parameter estimators and distort subsequent inference.

The case where the missing data are not MAR (NMAR) can be treated by postulating a parametric model for the distribution of the outcomes before nonresponse and a model for the response mechanism. These two models define a parametric model for the observed outcomes, so that the parameters of these models can be estimated from the observed data. Once the parameters are estimated, the first model can be used for inference.

Modeling the distribution of the outcomes before nonresponse is difficult since only the observed data are available.

Sverchkov (2008) proposed an alternative approach, which allows the parameters of the response model to be estimated without postulating a parametric model for the distribution of the outcomes before nonresponse. To account for the nonresponse, Sverchkov (2008) assumes a response model and estimates the response probabilities by applying the missing information principle (MIP), which consists of defining the likelihood as if there was complete response, and then integrating out the unobserved outcomes from the likelihood, employing the relationship between the distributions of the observed and unobserved data.

In this talk I show how this approach can be applied to testing whether probability weighted estimators, corrected for MAR nonresponse, are still biased due to ignoring that nonresponse is NMAR actually.

Finally, I illustrate this approach on real data (Consumer Expenditure Survey) example.

NOTATION AND MODELS

$\{y_i, \mathbf{x}_i; i = 1,..., N\}$ represent the data in a finite population of N units, y_i is the value of the outcome variable for unit i,

$\mathbf{x}_i' = (x_{i,1},..., x_{i,K})$ is a vector of corresponding K covariates.

Population outcomes follow (unknown) model:

$$y_i \mid \mathbf{x}_i \sim f(y_i \mid \mathbf{x}_i), \ i = 1,..., N$$

The target is to estimate the population mean $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$, based on a sample s of n units with inclusion probabilities $\pi_i = \Pr(i \in s)$.

Denote by I_i the sample indicator; $I_i = 1$ if unit i is selected in the sample and 0 otherwise. Let $w_i = 1/\pi_i$ denote the sampling weights.

In practice, not every unit in the sample responds.

Define the response indicator; $R_i = 1$ if unit $i \in s$ responds and $R_i = 0$ otherwise.

The sample of respondents is thus $R = \{i : I_i = 1, R_i = 1\}$ and the sample of nonrespondents among the sampled units is $R^c = \{i : I_i = 1, R_i = 0\}$.

Assumption 1. The response process is assumed to occur stochastically, independently between units.

The sample of respondents defines therefore a second, self-selected stage of the sampling process with unknown response probabilities.

Then the observed data follow 'respondents' model: $f_R(y_i \mid \mathbf{x}_i) = f(y_i \mid \mathbf{x}_i, i \in R)$.

The 'respondents' model is again general and all that we state at this stage is that under informative sampling and/or NMAR nonresponse, the models for the respondents and for the population differ; $f_R(y_i \mid \mathbf{x}_i) \neq f(y_i \mid \mathbf{x}_i)$.

Remark 1. The respondents' model refers to the observed data and hence can be estimated and tested by standard methods.

Let $p_r(y_i,\mathbf{x}_i)=\Pr(R_i=1|\,y_i,\mathbf{x}_i,i\in s)$. If the probabilities $p_r(y_i,\mathbf{x}_i)$ were known, the sample of respondents could be considered as a sample from the finite population with known sampling probabilities $\tilde{\pi}_i=\pi_i p_r(y_i,\mathbf{x}_i)$. In this case, the population mean $\bar{Y}$ can be estimated, for example, by probability weighted estimator, $\hat{\bar{Y}}=\left(\sum_{i=1}^{n}1/\tilde{\pi}_i\right)^{-1}\sum_{i=1}^{n}y_i/\tilde{\pi}_i$.

Also, if known, the response probabilities could be used for imputation of the missing data within the selected areas, by applying the relationship between the sample and sample-complement distributions, (Sverchkov and Pfeffermann, 2004);

$$f(y_i \mid \mathbf{x}_i, i \in R^c) = \frac{[p_r^{-1}(y_i, \mathbf{x}_i) - 1] f(y_i \mid \mathbf{x}_i, i \in R)}{E\{[p_r^{-1}(y_i, \mathbf{x}_i) - 1] \mid \mathbf{x}_i, i \in R\}}$$

Remark 2. This equation is a key relationship for derivation of the main result of the next section. It allows the distribution of the unobserved outcomes to be written as a function of the distribution of the observed outcomes and response and/or selection model.

ESTIMATION OF RESPONSE PROBABILITIES

Assume a parametric model of response probabilities, which is allowed to depend on the outcome and the covariate values; $\Pr[R_i = 1 | y_i, \mathbf{x}_i, i \in s; \boldsymbol{\gamma}]$ $= p_r(y_i, \mathbf{x}_i; \boldsymbol{\gamma})$, where γ is a vector of unknown coefficients. We assume that $p_r(y_i, x_i; \boldsymbol{\gamma})$ is differentiable with respect to $\boldsymbol{\gamma}$. Under these assumptions and Assumption 1, if the missing outcome values were observed, $\boldsymbol{\gamma}$ could be estimated by solving the likelihood equations:

$$\sum_{i \in R} \frac{\partial \log p_r(y_i, \mathbf{x}_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} + \sum_{\mathrm{k} \in R^c} \frac{\partial \log[1 - p_r(y_k, \mathbf{x}_k; \boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}} = 0.$$

Missing Information Principle (MIP)

In practice, the missing data are unobserved, but one may apply in this case the **MIP**: since no observations are available for $i \in R^c$, solve instead,

$$E\left\{\left[\sum_{i\in R}\frac{\partial \log p_r(y_i,\mathbf{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+\sum_{\mathrm{k}\in R^c}\frac{\partial \log[1-p_r(y_k,\mathbf{x}_k;\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\right]\middle| O\right\}=0 \ ,$$

$$O=\{y_i, n, i\in R;\ \mathbf{x}_t,\ \mathrm{t}=1,\ldots,n\}.$$

It can be shown that, by applying the relationship between the sample and sample-complement distributions, the later can be solved by maximizing the log-likelihood,

$$
\begin{aligned}
l(\boldsymbol{\gamma}) &= E\left(\sum_{i\in R} \log p_r(y_i,\mathbf{x}_i;\boldsymbol{\gamma}) + \sum_{i\in R^c} [1-\log p_r(y_i,\mathbf{x}_i;\boldsymbol{\gamma})] \middle| O \right) \\
&= \sum_{i\in R} \log p_r(y_i,\mathbf{x}_i;\boldsymbol{\gamma}) \\
&+ \sum_{k\in R^c} \frac{E\{[p_r^{-1}(y_k,\mathbf{x}_k;\boldsymbol{\gamma}^*)-1]\log[1-p_r(y_k,\mathbf{x}_k;\boldsymbol{\gamma})]\mid \mathbf{x}_k, k\in R\}}{E_{re}\{[p_r^{-1}(y_k,\mathbf{x}_k;\boldsymbol{\gamma}^*)-1]\mid \mathbf{x}_k, k\in R\}}.
\end{aligned}
$$

Notice that the expectations in the last expression are with respect to the model holding for the observed data for the respondents.

Remark 3. A fundamental question regarding the solution of the MIP equations is the existence of a unique global solution or more generally, the identifiability of the response model. Riddles et al. (2016) propose a similar approach to deal with NMAR nonresponse in the general context of survey sampling inference and establish the following fundamental condition for the response model identifiability: the covariates $\mathbf{x}$ can be decomposed as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ with $dim(\mathbf{x}_2) \geq 1$, such that $\Pr(R_i = 1 | y_i, \mathbf{x}_i) = \Pr(R_i = 1 | y_i, \mathbf{x}_{1i})$. In other words, the covariates in $\mathbf{x}_2$ that appear in the outcome model do not affect the response probabilities, given the outcome and the other covariates.

SELECTION OF A RESPONSE MODEL

The above likelihood suggests at least two procedures for the selection of the response model under NMAR nonresponse. One procedure involves comparing different models based on information criteria such as the Akaike information criterion, $\text{AIC}=-2l(\boldsymbol{\gamma})+2\dim(\boldsymbol{\gamma})$, or Schwarz information criterion, $\text{BIC}=-2l(\boldsymbol{\gamma})+\dim(\boldsymbol{\gamma})\log(n)$, $n=\sum_{i\in s} n_i$; a second procedure involves testing a saturated versus a nested model based on the likelihood ratio test.

(See technical details and simulation study in Sverchkov, M. and Pfeffermann, D. (2023) Response Model Selection in Small Area Estimation Under Not Missing at Random Nonresponse. To appear in Calcutta Statistical Association Bulletin)

TESTING BIAS OF MEAN ESTIMATES DUE TO NOT MISSING AT RANDOM NONRESPONSE

Let $\hat{w}_i^{MAR}$ be a good estimate of $w_i[p_r(\mathbf{x}_i)]^{-1} = w_i[\Pr(R_i = 1 \mid \mathbf{x}_i, i \in s)]^{-1}$, in other words $\hat{w}_i^{MAR}$ is a good estimate of final inclusion weight if (or “assuming”) the nonresponse is MAR. Then the obvious estimate for the mean will be $\hat{\bar{Y}}^{MAR} = \left(\sum_{i=1}^{n} \hat{w}_i^{MAR} \right)^{-1} \sum_{i=1}^{n} \hat{w}_i^{MAR} y_i$ and if the response is really MAR then this estimate is consistent.

How one can check in real situation that the latter estimate is not biased? Formally one can estimate response probabilities $\Pr[R_i = 1 | y_i, \mathbf{x}_i, i \in s; \boldsymbol{\gamma}]$ $= p_r(y_i, \mathbf{x}_i; \boldsymbol{\gamma})$ following Sverchkov (2008) approach and then compare model $p_r(y_i, \mathbf{x}_i; \boldsymbol{\gamma})$ with $p_r(\mathbf{x}_i; \boldsymbol{\gamma})$ by some information criterion (as in previous slide). But the model, $\Pr[R_i = 1 | y_i, \mathbf{x}_i, i \in s; \boldsymbol{\gamma}] = p_r(y_i, \mathbf{x}_i; \boldsymbol{\gamma})$, can be hard to identify and to estimate.

$Bias(\hat{\hat{Y}}^{MAR}) = N^{-1}\sum_{j=1}^{N} y_j - \hat{\hat{Y}}^{MAR} \cong -\left(\sum_{i=1}^{n}\hat{\hat{w}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\hat{w}}_i\varepsilon_i^{MAR}$, where $\varepsilon_i^{MAR} = y_i - \left(\sum_{i=1}^{n}\hat{w}_i^{MAR}\right)^{-1}\sum_{i=1}^{n}\hat{w}_i^{MAR}y_i$, $\hat{\hat{w}}_i = w_i[p_r(y_i,\varepsilon_i^{MAR};\hat{\boldsymbol{\gamma}})]^{-1}$, and therefore one can test whether $\hat{\hat{Y}}^{MAR}$ is biased or not by testing if $\Pr(R_i = 1 | \varepsilon_i^{MAR})$ depends on ε_i^{MAR} or not.

Note that although $\Pr(R_i = 1 | \varepsilon_i^{MAR};\boldsymbol{\gamma})$ does not involve any covariates, we still need a covariate $\mathbf{x}_{2i}$ such that ε_i^{MAR} correlates with $\mathbf{x}_{2i}$, but $\Pr(R_i = 1 | \varepsilon_i^{MAR}, \mathbf{x}_{2i}) = \Pr(R_i = 1 | \varepsilon_i^{MAR})$, otherwise $\boldsymbol{\gamma}$ estimate can be not unique, see Remark 3 above.

APPLICATION TO CONSUMER EXPENDUTURE SURVEY

We consider estimation of the following expenditures, Total, Food, Housing, Health, based on 2019, 2020 and 2021 samples. We assume truncated logit model for response probability

$$\Pr[R_i = 1 \mid \varepsilon_i^{MAR}, i \in s; \boldsymbol{\gamma}] = \left[0.0001 + \frac{\exp(\gamma_0 + \gamma_1 \varepsilon_i^{MAR})}{1 + \exp(\gamma_0 + \gamma_1 \varepsilon_i^{MAR})} \right] \Big/ 1.0001$$

where ε_i^{MAR} is defined by a particular expenditure and MAR final sample weight, FINLWT21, used currently by Consumer Expenditure Survey.

We use “median income” variable as instrumental variable $\mathbf{x}_{2i}$ (see Remark 3), correlation coefficients between $\mathbf{x}_{2i}$ and Total, Food, Housing, Health are equal to 0.31, 0.26, 0.35, 0.08 respectively for year 2021, and simple linear model explain relation between the expenditures and $\mathbf{x}_{2i}$ reasonably although with low R-square statistics, 0.15, 0.07, 0.16, 0.20. Parameter estimates for response model, $\Pr[R_i = 1 | \varepsilon_i^{MAR}, i \in s; \boldsymbol{\gamma}] = p_r(\varepsilon_i^{MAR}; \boldsymbol{\gamma})$, are all significant with p-value<0.0001, p-values are produced by SAS Proc NLIN used in maximization of the respective likelihood.

AIC and BIC criteria for all 4 expenditures and all years considered in this research prefer response model that include ε_i^{MAR} as a covariate (NMAR nonresponse) versus response model that assume MAR response, results summarized in the table below.

	AIC		BIC		$\frac{Bi\hat{a}s(\hat{\bar{Y}}^{MAR})}{s\hat{t}d(\hat{\bar{Y}}^{MAR})}$
	MAR	NMAR	MAR	NMAR	
Total 2019	68372.9	66794.6	68380.9	66810.6	1.00
Total 2020	70538.0	69337.4	70545.9	69353.3	0.49
Total 2021	71907.6	71082.4	71915.6	71098.2	1.67
Food 2019	68372.9	66872.5	68380.9	66888.4	0.71
Food 2020	70538.0	69337.4	70545.9	69353.3	2.50
Food 2021	71907.6	71159.3	71915.6	71175.1	0.65
Housing 2019	68372.9	66858.2	68380.9	66874.2	0.01
Housing 2020	70538.0	69366.4	70545.9	69382.0	2.86
Housing 2021	71907.6	71049.2	71915.6	71065.1	3.15
Health 2019	68372.9	66872.1	68380.9	66888.2	-1.11
Health 2020	70538.0	69337.4	70545.9	69353.3	-0.66
Health 2021	71907.6	71142.1	71915.6	71157.9	0.28

The last column of the table presents an estimate of normalized bias of MAR estimator, $\dfrac{Bi\hat{a}s(\hat{\bar{Y}}^{MAR})}{s\hat{t}d(\hat{\bar{Y}}^{MAR})}$, where $Bi\hat{a}s(\hat{\bar{Y}}^{MAR}) = \; -(\hat{\bar{Y}}^{NMAR} - \hat{\bar{Y}}^{MAR})$, and $s\hat{t}d(\hat{\bar{Y}}^{MAR})$ is a BRR standard error estimate of $\hat{\bar{Y}}^{MAR}$. Note that the later statistics can be used only as crude approximation of normalize bias $\dfrac{Bias(\hat{\bar{Y}}^{MAR})}{std[Bias(\hat{\bar{Y}}^{MAR})]}$, since it does not include errors due to $\hat{\bar{Y}}^{NMAR}$ estimation.

Conclusion: The above three statistics suggest that although current Consumer Expenditure Survey nonresponse adjustment of sample weights does not remove nonresponse bias (MAR estimates for the mean total expenditures are still biased after the adjustment), the bias is not or only slightly significant based on the standard error of the final estimates.

Thanks !

Sverchkov.Michael@bls.gov

REFERENCES

Riddles, K.M., Kim, J.K. and Im, J. (2016). A propensity-score adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, **4**, 215-245.

Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random. *Joint Statistical Meetings*, *Proceedings of the Section on Survey Research Methods*, 867-874.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of Finite Population Totals Based on the Sample Distribution. *Survey Methodology*, **30**, 79-92.