On checking whether response is ignorable or not

Michail Sverchkov, Bureau of Labor Statistics

The opinions expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Introduction and Notation:

$\{Y_i, X_i; i \in U\}$ - finite population from unknown *pdf* $f(Y_i | X_i)$

("*pdf*" - probability density function when Y_i is continuous or the probability function when Y_i is discrete)

$\{Y_i, X_i; i \in S\}$ - sample drawn from finite population U with known inclusion probabilities $\pi_i = \Pr(i \in S)$

Y_i - target variable, $X_i = (X_i^1, ..., X_i^K)$ - covariates (observed for entire sample). R - sample of respondents (the sample with observed outcome values)

Let $p(Y_i, X_i) = \Pr(i \in R \mid Y_i, X_i, i \in S)$. If $p(Y_i, X_i)$ were known then the sample of respondents could be considered as a sample from the finite population with known selection probabilities $\tilde{\pi}_i = \pi_i p(Y_i, X_i)$ $\Rightarrow$ population model parameters (or finite population parameters) could be estimated as if there was no non-response.

Also, if known, the response probabilities could be used to impute the missing sample data via the relationship between the sample and sample-complement distributions (Sverchkov & Pfeffermann 2004);

$$f(Y_i = y \mid X_i = x, i \notin R, i \in S) =$$

$$\frac{[p^{-1}(y,x)-1] f(Y_i = y \mid X_i = x, i \in R)}{E\{[p^{-1}(y,x)-1] \mid X_i = x, i \in R\}} \qquad \textbf{(1)}$$

Note that $f(Y_i = y \mid X_i = x, i \in R)$ refers to the *observed data* and therefore can be estimated using classical statistical inference procedures.

Most methods of estimating the response probabilities assume (explicitly or implicitly) that the missing data are 'missing at random', $\Pr(i \in R \mid Y_i, X_i, i \in S) = \Pr(i \in R \mid X_i, i \in S)$.

In this case, if auxiliary data is not missing, $\Pr(i \in R \mid X_i, i \in S)$ refers to *fully observed* data and can be estimated by use of classical regression techniques.

In many practical situations, MAR assumption is not valid: the probability of responding often depends directly (or indirectly) on the outcome value. In this case the use of methods that assume MAR can lead to large bias of population parameter estimators and large imputation bias.

The case where the missing data are NMAR can be treated by postulating a parametric model for the distribution of the outcomes *before non-response,* $f[Y_i \mid X_i, i \in S; \alpha]$, and a model for the response mechanism, $p(Y_i, X_i; \gamma)$,

$\Rightarrow$ two models define a parametric model for the joint distribution of the outcomes and the response indicators

$\Rightarrow$ the parameters of these models can be estimated by maximization of the likelihood based on this joint distribution.

Problems: i) Modeling the distribution of the outcomes *before non-response* can be problematic since it refers to the partly unobserved data.

ii) The same problem with the response mechanism.

iii) Estimators assuming NMAR are usually much less stable than estimators assuming MAR. (Moreover, often NMAR estimator does not exist - too many unknown parameters).

Sverchkov JSM 2008 suggested an approach that allows estimation of the parameters of the response model without modeling the distribution of the outcomes *before non-response*:

For simplicity assume that auxiliary variables are not missing. Let $p(Y_i, X_i; \gamma) = \Pr(i \in R \mid Y_i, X_i, i \in S; \gamma)$ and suppose that p is differentiable with respect to the (vector) parameter γ.

If the missing data were later observed, γ could be estimated by solving:

$$0 = \sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \quad \textbf{(2)}$$

Denote the observed data by $O = \{Y_i, i \in R; X_k, k \in S\}$.

Missing Information Principle: since the outcome values are missing for $i \notin R,\ i \in S$, we propose to solve instead,

$$0 = E\{[\sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma}] \mid O\} =$$

$$\sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} E\{\frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \mid O, i \in R^c\} \overset{\textbf{Eq.1}}{=}$$

$$\sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} +$$

$$+ \sum_{i \in R^c} \frac{E\{[p^{-1}(Y_i, X_i; \gamma) - 1] \dfrac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \mid X_i, i \in R\}}{E\{[p^{-1}(Y_i, X_i; \gamma) - 1] \mid X_i, i \in R\}} = 0 \quad \textbf{(3)}$$

Parameter γ can be estimated by solving **(3)**.

Note that the second sum in **(3)** predicts the unobserved second sum in **(2)**.

Note also that if $p(Y_j, X_j; \gamma)$ is a function of X_j and γ only (missing data is MAR) then **(3)** reduces to the common log-likelihood equations,

$$0 = \sum_{i \in R} \frac{\partial \log p(X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{\partial \log[1 - p(X_i; \gamma)]}{\partial \gamma}. \qquad \textbf{(4)}$$

The proposed approach can be generalized to the case when auxiliary variables are partly missing. See also Sverchkov JSM 2010 for similar approaches.

The proposed approach requires knowledge of the parametric form of the response model which refers to the unobserved data in the case of NMAR.

On the other hand, if the response is MAR, the propensity score, $p(X_i;\gamma) = \Pr(i \in R \mid X_i, i \in S;\gamma)$, can be estimated from the observed data, for example, by solving the log-likelihood equations **(4)**.

The latter estimators are much more stable than the estimators assuming NMAR.

Can we check whether the response is MAR or NMAR?

Testing whether the response is MAR or NMAR

<u>Step 1</u>. Fit the model for propensity score,

$$p(X_i;\alpha) = \Pr(i \in R \mid X_i, i \in S;\alpha),$$

and estimate the parameter γ from the observed data assuming MAR.

Step 2. Define a class of models for $p(Y_i, X_i; \gamma) = \Pr(i \in R \mid Y_i, X_i, i \in S; \gamma)$, $\gamma \in \Gamma$, in such way that for some $\tilde{\gamma} \in \Gamma$, $p(Y_i, X_i; \tilde{\gamma}) = p(X_i; \alpha)$. It is recommended to use models that include the *Y*-component in a simple form, **EXAMPLE:** if $\text{logit}[p(X_i; \alpha)] = g(X_i; \alpha)$ then one can consider $\text{logit}[p(Y_i, X_i; \gamma)] = \text{g}(\text{X}_i; \alpha) + cY_i, \quad \gamma = (\alpha, c)$, so in this case for $\tilde{\gamma} = (\alpha, 0)$, $p(Y_i, X_i; \tilde{\gamma}) = p(X_i; \alpha)$.

Step 3. Obtain estimating equations **(3)** based on the class of models defined in Step 2.

Step 4.1. Solve them and check whether *Y*-component is significant (in which case the response is NMAR) or not (the response is MAR or “not very informative”).

The latter can be done by a bootstrap procedure: one can take *B* simple random samples with replacement from the original sample and repeat steps 1 – 4 above in order to get a variance estimate for the *Y*-component.

Remark. Since the parametric family defined in Step 2 does not necessarily include the true response probability $\Pr(i \in R \mid Y_i, X_i, i \in S)$, we cannot conclude for sure that response is MAR even if the *Y*-component is insignificant. We recommend assuming MAR in this case. If response is very informative then one can expect that the *Y*-component will be significant even when fitting a simplified model.

Instead of Step 4.1 one can do

<u>Step 4.2</u>. Substitute $\tilde{\gamma}$ from Step 2 (which corresponds to MAR assumption) into **(3)** obtained in Steps 1 - 3 and check whether the result of this substitution is significantly non-zero (response is NMAR) or not (response “seems to be” MAR since $\tilde{\gamma}$ corresponds to the propensity score). The latter can also be done by use of a bootstrap.

Empirical illustration.

For simplicity assume that the finite population and the sample coincide, $U = S$. The simulation study consists of the following steps.

Step A: Generate independently 100 finite populations, each of size 1000, where X_i ~ Uniform (-1,1),

$$P(Y_i = 1 \mid X_i) = (\exp\{-0.1 - X_i\} + 1)^{-1},$$

$$P(Y_i = 0 \mid X_i) = 1 - P(Y_i = 1 \mid X_i).$$

Step B: For each population the response indicators were generated as:

$$P(R_i = 1 \mid Y_i, X_i) = [\exp(\gamma_0 + \gamma_1 X_i) + 1]^{-1} \times [\gamma_2 Y_i + 2]^{-1}.$$

We repeat the study for different values of parameter $\gamma = (\gamma_0, \gamma_1, \gamma_2)$.

Step C: For each sample of respondents estimate the response probabilities assuming the response is MAR and the response model is logistic, i.e. $(\hat{\gamma}_0, \hat{\gamma}_1)$ is a solution of the likelihood equations

$$\sum_{i \in R} \frac{\partial \log\{[\exp(\gamma_0 + \gamma_1 X_i^{(m)}) + 1]^{-1}\}}{\partial \gamma_d} + \sum_{i \in R^c} \frac{\partial \log\{1 - [\exp(\gamma_0 + \gamma_1 X_i^{(m)}) + 1]^{-1}\}}{\partial \gamma_d} = 0, \quad d = 0,1$$

These estimates were derived using Proc Logistics of SAS.

Step D: Define estimating equations **(3)** assuming that response follows the logistic model,

$$P(R_i = 1 \mid Y_i, X_i) = [\exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i) + 1]^{-1}$$

and substitute $(\hat{\gamma}_0, \hat{\gamma}_1)$ into the estimating equations.

Test:

If the result is significantly non-zero then response is NMAR

# of rejections MAR hypothesis, $\gamma_0 = -1$, $\gamma_1 = -1$

γ_2	95%	99%
-1	99	99
-0.9	99	99
-0.6	99	95
-0.3	50	15
0	4	0
0.3	32	5
0.6	68	32
0.9	91	71

# of rejections MAR hypothesis, $\gamma_0 = 1,\ \gamma_1 = -1$

γ_2	95%	99%
-1	99	99
-0.9	99	93
-0.6	73	35
-0.3	19	4
0	8	0
0.3	7	1
0.6	28	6
0.9	56	10

THANKS !!! (Sverchkov.Michael@bls.gov)