ADJUSTING FOR NONRESPONSE BIAS OF CORRELATED ITEMS USING LOGISTIC REGRESSION[1]

Pao-Sheng and Robin Fisher
Robin Fisher, U.S. Bureau of the Census, Washington, D.C. 20233

1. Introduction

Item nonresponse in sample surveys is the failure to obtain a specific question that should have been answered. In particular, item nonresponse rate on the survey of the economic is usually high. This can result for many reasons, the most frequent being "refusals to answer", which can relate to the underlying data value (non-ignorable nonresponse) and human behavior. Most designs of the survey questionnaire incorporate procedures for following up on missing responses to items identified as either especially important to the overall quality of the survey data or with previously noted high nonresponse rate. For example, the design of the Survey of Income and Program Participation (SIPP) questionnaire incorporated procedures for following up on missing responses to the items of wage and salary income, income received from self-employment and interest and dividend income. The response status on these items by the same individual are most likely correlated. The problem of missing items for categorical variable has been examined from the perspective of modeling the mechanisms of nonresponse by Fay (1986), Chambers and Welsh (1993), Alho (1990), and Särndal (1981).

This paper proposes a method of adjusting item nonresponse in presence of callback based on a generalized logistic regression model that can account for the correlation among responses on items. The probability of response for any item is represented by a logistic regression model, in which the value of that item, the response status of the rest of the items and the available covariates, which may include the observed item variables for all the individuals by the last callback, are explanatory variables. The respondents are assumed to answer some or all of items after one or more call-backs. The parameters of our model can be estimated by taking a conditional maximum likelihood approach based on the respondents. This approach has the advantage of the simple expression of conditional logistic model. The estimated individual probabilities of responding are used in a Horvitz-Thompson type estimator to reduce bias in the estimation of sample means for every single item.

2. The Logistic Regression for Correlated Responses

2.1 A Class of Conditional Logistic Models

Let $I = \{1, ..., n\}$ be a set of indices for n individuals selected in a simple random sample. Let $\mathbf{X}_i = (X_{i1}, ..., X_{iL})$ be the set of item outcomes from individual i and they suffer the nonresponse, i=1, ..., n, where X_{il} expresses the outcome of the l^{th} item from individual i, the value of which becomes known when individual i responds for the item l. The vector of covariates of individual i is denoted by $\mathbf{Z}_i$. Suppose up to $J \geq 2$ attempts are made to capture the data for an individual. Define the nonresponse indicator vector $\mathbf{U}_{ij} = (U_{ij1}, ..., U_{ijL})^T$, where, for l=1, ..., L with $U_{ijl} = 1$ if individual i was captured at the j^{th} attempt for the l^{th} item, and $U_{ijl} = 0$ otherwise.

Define $y_{ijl} = \sum_{k=1}^{j} U_{ikl}$ (i=1, ..., n; l=1, ..., L) for short. Then $y_{ijl} = 1$, if and only if individual i was captured by the j^{th} attempt for the l^{th} item. If U_{ijl} are correlated, the probability for $U_{ijl}=1$ not only depends upon X_{il} and $\mathbf{Z}_i$ but also depends on the responses for the rest of the items. First, consider the class of conditional logistic models when j=1

$$\begin{aligned} logit\ Pr\ (U_{il1} = 1\ |\mathbf{U}_{il}^{-1}, X_{il}, \mathbf{Z}_i) \\ = F_{1l}\ (\mathbf{U}_{il}^{-1}) + G_{1l}(X_{1l}, \mathbf{Z}_i) \end{aligned} \quad (1)$$

when j > 1

$$\begin{aligned} logit\ Pr\ (U_{ijl} = 1\ |\mathbf{U}_{il}^{-1}, \mathbf{S}_{ij-1}, X_{il}, \mathbf{A}_i) \\ = F_{jl}\ (\mathbf{U}_{ij}^{-1}) + G_{jl}(X_{il}, \mathbf{Z}_i)\ if\ y_{ijl} = 0 \end{aligned} \quad (2)$$

$$Pr\ (U_{ijl} = 1\ |\ \mathbf{U}_{ij}^{-l}, \mathbf{S}_{ij-1}, X_{il}, \mathbf{Z}_i) = 0 \quad \text{if } y_{ij-1,l}=1$$

where $\mathbf{U}_{ij}^{-l}$ is $\mathbf{U}_{ij}$ with the exclusion of U_{ijl}, for j=1, ..., J, $\mathbf{S}_{ij-1} = (\mathbf{U}_{i1}^T, ..., \mathbf{U}_{ij-1}^T)^T$.

---

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

F_{1l} is an arbitrary function of U_{il}^{-1} such that $\sum_{l=1}^{L} U_{ill} F_{1l}(I_{il}^{-1})$ *is* invariant under permutation of U_{i1l}'s, where I_{il}^{-1} is U_{il}^{-1} with $U_{ilk} = 0$ for $k > 1$, F_{jl} is an arbitrary function of U_{ij}^{-1} such that $\sum_{l=1}^{L} U_{ijl} F_{jl}(I_{ij}^{-1}, S_{i,j-1})$ *is* invariant under permutation of U_{ijl}'s, and where I_{ij}^{-1} is U_{ij}^{-1} with $U_{ijk} = 0$ for $k > 1$

Thus F_{jl} is a function describing the dependence of item l on the response status of the other items in and before call-back attempt j. The function G describes the dependence on the outcomes **X** and the covariates **Z**.

For given F_{1l} and F_{jl}, from (1) and (2) we have the joint probability of U_{ijl}'s uniquely defined as when j = 1

$$Pr(U_{il} \mid X_i, Z_i) = \exp\left[\sum_{l=1}^{L} U_{ill}[F_{1l}(I_{il}^{-1}) + G_{1l}(X_{ip}, Z_i)]\right] \Big/ d_{1i} \quad (3)$$

where

$$d_{1i} = \sum_{u_{i1},\ldots,u_{iL}} \exp\left[\sum_{l=1}^{L} u_{ill}[F_{1l}(I_{il}^{-1}) + G_{1l}(X_{ip}, Z_i)]\right],$$

$$u_{ill} = 0,1,\ l=1,\ldots,L$$

when j > 1

$$Pr(U_{ij} \mid S_{i,j-1}, X_i, Z_i) = \exp\left[\sum_{l=1}^{L} U_{ijl}[F_{jl}(I_{ij}^{-1}, S_{i,j-1}) + G_{jl}(X_{ip}, Z_i)]\right] \Big/ d_{ji} \quad (4)$$

where

$$d_{ji} = \sum_{u_{i1},\ldots,u_{iL}} \exp\left[\sum_{l=1}^{L} u_{ijl}[F_{jl}(I_{ij}^{-1}, S_{i,j-1}) + G_{jl}(X_{ip}, Z_i)]\right]$$

$$u_{ijl} = \begin{cases} 0,\ 1 & \text{if } y_{i,j-1,l} = 0 \\ 0 & \text{if } y_{i,j-1,l} = 1 \end{cases}$$

Equations (3) and (4) follow from an argument similar to that given Liang and Zeger (1989) in the appendix. Note that d_{1i} and d_{ji} are the normalizing constants which involve a sum of 2^L exponential terms.

2.2 An example

1. A special case is that where the response probability of item l in attempt j depends on the responses of the other item only through their number in attempt j (denoted by $r_{ijl} = \sum_{k \neq l} U_{ijl}$ and their number by attempt j-1 (denoted by $t_{i,j-1,l} = \sum_{k \neq l} y_{i,j-1,k}$). That is, when j = 1

$$logit\ Pr(U_{ill} = 1 \mid U_{il}^{-1}, X_{ip}, Z_i) = F_1(r_{ill}) + G_{1l}(X_{1p}, Z_i)$$

when j > 1

$$logit\ Pr(U_{ijl} = 1 \mid U_{ij}^{-1}, S_{i,j-1}, X_{ip}, Z_i) = F_j(r_{ijl}, t_{i,j-1,l}) + G_{jl}(X_{ip}, Z_i) \quad if\ y_{i,j-1,l} = 0$$

$$Pr(U_{ijl} = 1 \mid U_{ij}^{-1}, S_{i,j-1}, X_{ip}, Z_i) = 0 \quad if\ y_{i,j-1,l} = 1$$

When j=1, we have the joint probability of U_{il} uniquely defined as

$$Pr(U_{il} \mid X_i, Z_i) = \exp\left[\sum_{l=1}^{L} U_{ill}[F_1(B_{ill}) + G_{1l}(X_{ip}, Z_i)]\right] \Big/ d_{1i} \quad (5)$$

where $B_{ill} = \sum_{k=1}^{l-1} U_{ilk}$ *and we assume that* $B_{il1} = 0$

Similarly, when j > 1, the joint probability of U_{ij} conditional on $S_{i,j-1}$ is uniquely defined as

$$PR(U_{ij} \mid S_{i,j,p}, X_i, Z_i = \exp\left[\sum_{l=1}^{L} U_{ijl}\left[F_j(B_{ijl}, t_{i,j-1,l}) + G_{jl}(X_{ip}, Z_i)\right]\right] \Big/ d_{ji}$$

where $U_{ijl}=0$ when $y_{i,j-1,l} = 1$

$= 0$ otherwise (6)

, $B_{ijl} = \sum_{k=1}^{l-1} U_{ijk}$ and we assume that $B_{ij1} = 0$

We might choose the model to describe a situation in which some respondents are more willing to respond than others. Consider item l in attempt j. If the subject has already responded to a large number of items, we might expect him or her to be more likely to respond to item l. We index the function F with the call-back attempt to allow the dependence to change with the call-back number.

This type of model has been considered by Qu, William, Beck and Goormastic (1987) for the case J=1.

If we let $\sum_{t=0}^{T_i} F_1(t) = 0$ when $T_{i1} = 0$, and

$\sum_{t=t_{ij-1}}^{T_{ij-1}} F_j(t)=0$ when $T_{ij} = t_{i,j-1}$ and $F_j(B_{ijl}, t_{i,j-1}) = F_j(B_{ijl} + t_{i,j-1})$, l=1,...,L,

We get the following class of conditional logistic models.

When j=1

$$Pr(U_{i1l}=1 \mid U_{i1}^{-l}, X_{ip}, Z_i) = \frac{e^{F_1(r_{i1l}) + G_{1l}(X_{ip}, Z_i)}}{1 + e^{F_1(r_{i1l}) + G_{1l}(X_{ip}, Z_i)}}$$

When j > 1

$$Pr(U_{ijl} = 1 \mid U_{ij}^{-l}, S_{i,j-1}, X_{ip}, Z_i) = \frac{e^{F_j(r_{ijl}^*) + G_{jl}(X_{ip}, Z_i)}}{1 + e^{F_j(r_{ijl}^*) + G_{jl}(X_{ip}, Z_i)}} \text{ if } y_{i,j-1,l} = 0$$

$$= 0 \text{ if } y_{i,j-1,l} = 1$$

Where $r_{ijl}^* = r_{ijl} + t_{i,j-1,l}$

Here is one simple model.

$$G_{jl}(X_{ip}, Z_i) = \alpha_{jl} + X_{ip}\beta_{1l} + Z_i^T\beta_{2l} \quad (7)$$

$$F_1(r_{i1l}) = r_{i1l}\delta, \text{ and } F_j(r_{ijl}^*) = r_{ijl}^*\delta \text{ for } j>1 \quad (8)$$

Therefore the characteristics X_i and covariates Z_i affect the capture probabilities for the same way for each attempt. The effect of X_i is felt on P_{ijl} only through the l^{th} characteristics X_{il}. Different attempts may have different capture probabilities depending on the α_{jl}'s. This may reflect varying methods at callback or the possibility that the respondent's probability of response changes after a number of calls. We can imagine, for example, that a respondent may develop some resistance after even a small number of attempts have been made in which case α_{jl} decreases in j. Also notice that the number of responses in other items affect each item the same way.

3. **Estimation Procedure (Conditional Maximum Likelihood Approach)**

Without loss of generality we can order the data so that by observation 1 through n_i are the responded items for the i^{th} individual and n_i+1 through L are the nonresponded items, we can estimate the probabilities based on the following 'working' conditional likelihood.

$$L^\circ = \prod_{i\in(I_A \cup I_S)} \prod_{l=1}^{n_i} \prod_{j=1}^{J} \nu_{ijl}^{U_{ijl}}$$

where I_A denotes the set of individuals who answer all the items, I_S denotes the set of individuals who only answer some of the items,and

$$\nu_{ijl} = \frac{\mu_{ijl}}{1-\mu_{i,J+1,l}}$$

where $\mu_{i,J+1,l} = 1 - \sum_{k=1}^{J} \mu_{ikl}$, and where

$$\mu_{i1l} = P_{i1l} = \frac{e^{F_{1l}(U_i^{-l}) + G_{1l}(X_{ip}, Z_i)}}{1 + e^{F_{1l}(U_i^{-l}) + G_{1l}(X_{ip}, Z_i)}},$$

$$\mu_{ijl} = P_{ijl} \prod_{k=1}^{j-1} (1-P_{ikl}), \quad j=2,...,J, \quad \text{and where}$$

$$P_{ijl} = \frac{F_{jl}(U_{ij}^{-l}, S_{i,j-1}) + G_{jl}(X_{ip}, Z_i)}{1 + F_{jl}(U_{ij}^{-l}, S_{i,j-1}) + G_{jl}(X_{ip}, Z_i)}.$$

Maximum 'working' conditional likelihood estimates of the parameters can be found by numerically maximizing the log of this function with respect to the parameters involved. Consider the assumptions of (9) and (10). Does not have a unique maximum. One way to solve this problem is to use the available additional information in conjunction with the likelihood equation. For computational advantage, we use the approach proposed by Alho (1991).

Let $I_{1l} \subset I$ be the set of individuals captured at the first attempt for the l^{th} item ($l=1,..., L$), I_{2l} the set of individuals captured at the second attempt for the l^{th} item, etc. Let $I_{J+1,l}$ be the set of individuals that are not captured for the l^{th} item at all. Define $n_{jl} = \text{card}(I_{jl})$, for $l=1,...,L$ and $j=1,...,J+1$; thus $n = n_{1l} + ... + n_{J+1,l}$ for $l=1,...,L$.

Note that for the l^{th} item, given $S_{i,j-1}$ and U_{ij}^{-l}, we have, for $j=1,...,J$

$$E\left(\sum_{i \in \bigcup_{k=j}^{J+1} I_{kl}} U_{ijl} \frac{1-P_{ijl}}{P_{ijl}} \,\middle|\, S_{i,j-1},\ U_{ij}^{-l} \right) = (n_{jl} + ... + n_{J+1,l}) - \sum_{i \in \bigcup_{k=j}^{J+1} I_{kl}} P_{ijl}$$

Notice $E(n_{jl} \mid S_{i,j-1},\ U_{ij}^{-l}) = \left(\sum_{i \in \bigcup_{k=j}^{J+1} I_{kl}} P_{ijl} \right)$; suppose we use n_{jl} to estimate this. We estimate the expectation on the left-hand side with the observed value also. This yields the likelihood equations

$$\sum_{i \in I_{jl}} \exp(-\alpha_{jl} - W_{il}^T \beta_l - r_{ijl}\delta) = n - (n_{1l} + ... + n_{jl}),\ l=1,...,L$$

Given β_l and δ we can thus solve for α_{jl} by taking

$$\alpha_{jl} = -\log(n - n_{1l} - ... - n_{jl}) \Big/ \sum_{i \in I_{jl}} \exp(-W_{il}^T \beta_l - r_{ijl}\delta) \quad (j=1,...,J, l=1,...,L)$$

To solve for

$$\alpha = (\alpha_1^T, ..., \alpha_L^T)^T \text{ where } \alpha_l = (\alpha_{1l},...,\alpha_{Jl})^T, \text{ for } l = 1,...,L$$

$$\beta = (\beta_1^T, ..., \beta_L^T)^T \text{ and } \delta$$

we use an iteration based on Newton's method.

Differentiating the log likelihood L with respect to β_l we get

$$\frac{\partial L}{\partial \beta_l} = - \sum_{y_{il}=1} \left[\sum_{j=1}^{J} (U_{ijl} - \nu_{ijl}) \sum_{k=1}^{J} P_{ikl} \right] W_{il} = 0$$

We can solve numerically for α, β, and δ .

Having calculated the estimate $\hat{\alpha}$, $\hat{\beta}$, $\hat{\delta}$, the Horvitz-Thompson type estimator was considered based on the requirement of unbiasedness. Define $X_l = (X_{1l}, ..., X_{nl})^T$, $l=1, ..., L$.

The true sample mean for the item l is $\bar{X}_l = X_l^T 1_n / n$ where 1_n is a vector of n ones. By translating $\hat{\alpha}$, $\hat{\beta}$, $\hat{\delta}$, into (7) and (8), we can calculate the estimates $\hat{\mu}_{i,J+1,l}$ to get the conditional unbiased Horvitz-Thompson type estimator of $\bar{X}_l$ ($l=1,...L$) as

$$\overline{X}_l = \frac{1}{n} \sum_{i \in I_{jl} \cup I_{J+1,l}} X_{il}\, Y_{iJL} / \hat{\mu}_{i,J+1,l}$$

Let $\hat{\gamma} = (\hat{\alpha}, \hat{\beta}, \hat{\delta})$ be the estimates of $\gamma = (\alpha, \beta, \delta)$ subject to regularity conditions on the X_{il}'s and Z_i's, the constraints used to ensure identifiability, coverage to continuous relation between the α_{jl}o's and (β_l, δ) which is satisfied by the true parameter vector γ . The proof for the consistency of the parameter estimator $\hat{\gamma}$ can be given by emulating the standard argument from Fahrmeir & Kaufmann (1985).

4. References

Alho, J. (1990). "Adjusting for Nonresponse Bias Using Logistic Regression." Biometrika 77, 617-624.

Besag, J. (1974). "Spatial Interaction and the Statistical Analyses of Lattice Systems." Journal of the Royal Statistical Society, Ser. B, 36, 192-236.

Chambers, R. L. and Welsh A. H. (1993). "Log-linear Models for Survey Data with Non-ignorable non-response." J. R. Statis. Soc. B, 55, 157-170.

Fay, R. E. (1986). "Causal Model for Patterns of Nonresponse." J. Am. Statist. Ass., 81, 354-365.

Fahrmeir, L and Kaufman, H. (1985). "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models." Ann. Statist. 13, 342-368.

Liang, K. and Zeger, S.L. (1989). "A Class of Logistic Regression Models for Multivariate Binary Time Series." Journal of the Am. Statist. Assoc. 84, 447-451.

Qu, Y., Williams, G. W., Beck G. J., & Goormastic, M. (1987). "A Generalized Model of Logistic Regression for Clustered Data." Commun. Statist. - Theory Meth. 16, 3447-3476.

Särndal, C.E. (1981). "Frameworks for Inference in Survey Sampling with Applications to Small Area Estimation and Adjustment for Nonresponse." Bull. Int. Statist. Inst. 49, 494-513

Zhao, L.P. & Prentice, R.L. (1990). "Correlated Binary Regression Using a Quadratic Exponential Model." Biometrika 77, 642-648.