MODELING NONRESPONSE IN SURVEYS WITH CALLBACKS

J. H. Drew and W. A. Fuller, Iowa State University

I. The Model

We shall study a model for the response-nonresponse of individuals given repeated opportunities to respond to a questionnaire. Models of this general type have been considered by Proctor (1977) and Thomsen and Siring (1979). Other models have been considered by Politz and Simmons (1949, 1950), Simmons (1954), Deming (1953), and more recently by Cassel, Särndal and Wretman (1979), Särndal and Hui (1980), and Frankel and Dutka (1979).

Let a simple random sample of n units be selected from a population of N units. Suppose that the population is partitioned into K categories corresponding to the K values of a discrete random variable. Associated with each unit of the k^{th} category is a response probability $q_k \in [0, 1]$, which is the conditional probability that the unit furnishes a response when sampled. If some $q_k \neq 1$, $n_1 \leq n$ responses are obtained on the first call. The $n - n_1$ nonrespondents are contacted in a second call, and n_2 of the $n - n_1$ respond. Calls continue in this way until, after R calls, n_0 units have not responded. One observes $n_{rk} (r = 1, 2, \ldots, R; k = 1, 2, \ldots, K)$, where n_{rk} is the number of units from the k^{th} category responding on the r^{th} call, and n_0.

It is assumed that a proportion $1 - \gamma$ of the population is composed of hard-core nonrespondents who will never answer the survey. In our initial treatment of the model we assume that the fraction of hard-core nonrespondents is the same in each category. Let the population proportions in categories 1, 2, ..., K be $f_1, f_2, \ldots, f_K$, respectively, where $\sum_{k=1}^{K} f_k = 1$. Under these assumptions the data $(n_{11}, n_{12}, \ldots, n_{1K}, \ldots, n_{RK}, n_0)$ satisfy the multinomial model with probabilities

$$\pi_{rk} = \gamma(1 - q_k)^{r-1} q_k f_k , \tag{1}$$

$$\pi_0 = (1 - \gamma) + \gamma \sum_{k=1}^{K} (1 - q_k)^R f_k , \tag{2}$$

where π_{rk} is the probability that an individual in category k will respond on call r, and π_0 is the probability that an individual will not have responded after R calls. The log likelihood, $\log L(\underset{\sim}{n}; \underset{\sim}{f}, \underset{\sim}{q}, \gamma)$ is proportional to

$$\sum_{r=1}^{R} \sum_{k=1}^{K} n_{rk} \log \pi_{rk} + n_0 \log \pi_0 . \tag{3}$$

The likelihood can be maximized by the method of scoring (see, for example, Rao, 1973), or by other methods, to give estimates of the parameters $\underset{\sim}{f}$, $\underset{\sim}{q}$, γ.

The given model requires data grouped into K categories. If the response probability is hypothesized to depend upon a continuous variable, it is necessary to postulate a parametric form for the response probability or to group the data on the basis of the continuous variable. See Brewer (1979).

Because many survey variables are continuous, it is of interest to describe the relative efficiency of the discretized model to the continuous model under some parametric form for q.

To develop the model, assume that the response probability for an individual is a function of the variable X. The example considered in the next section uses the model

$$q(X) = \beta_0 + \beta_1 X + \beta_2 X^2 .$$

Assume an infinite population and let the probability density function (p.d.f.) of X be $f_{\underset{\sim}{\theta}}(X)$, where $\underset{\sim}{\theta}$ is the vector of parameters defining the p.d.f. The log likelihood of this model is proportional to

$$\sum_{r=1}^{R} \sum_{j=1}^{n_r} \{\log \gamma + \log f_{\underset{\sim}{\theta}}(X_{rj}) + \log[\{1 - q(X_{rj})\}^{r-1} q(X_{rj})] - \log(m_r^*)\} + n_0 \log[(1-\gamma) + \gamma m_R^{**}] , \tag{4}$$

where $m_r^* = E_{\underset{\sim}{\theta}}[\{1-q(X)\}^{r-1} q(X)]$, $m_R^{**} = E_{\underset{\sim}{\theta}}[\{1-q(X)\}^R]$, and n_r is the number of units observed on the r^{th} call.

If the model is discretized by grouping the data on the basis of X into K groups the approximate log likelihood is proportional to

$$\sum_{r=1}^{R} \sum_{k=1}^{K} n_{rk} \log \pi_{rk} + n_0 \log \pi_0 , \tag{5}$$

where π_{rk}, π_0, n_{rk}, n_0 are as previously defined, and

$$q_k = \beta_0 + \beta_1 A_k + \beta_2 A_k^2 , \tag{6}$$

A_k = median of the k^{th} category,

$$f_k = \int_{X_{(k-1)}}^{X_{(k)}} f_{\underset{\sim}{\theta}}(X) dx , \tag{7}$$

and $X_{(i)}$, $i = 0, 1, \ldots, K$ are the category boundaries.

The approximate asymptotic relative efficiency for estimating a parameter θ_i is given by the ratio of the appropriate terms for θ_i in the two information matrices.

II. Estimation of Means

The estimated population proportions $\hat{f}_k$ calculated by the maximum likelihood method based on (3) of the previous section can be used

to construct estimates of the means of other variables in the survey.

Call the variable of interest Y, and let $\bar{y}_k$ be the sample mean of Y for the respondents (over all calls) in category k. In this section we assume that the probability that an individual responds on any particular call is independent of the y-value of the individual.

This implies that

$$E(\bar{y}_k | n_{11}, n_{12}, \ldots, n_{RK}, n_0) = \bar{Y}_k \quad . \tag{8}$$

Therefore, an asymptotically model unbiased estimator of the mean of Y is

$$\hat{\bar{Y}} = \sum_{k=1}^{K} \hat{f}_k \bar{y}_k \quad , \tag{9}$$

where $\hat{f}_k$ are the maximum likelihood estimators of the population fraction in category k.

The variance of $\hat{\bar{Y}}$ can be evaluated by using the result (8) and noting that $\bar{y}_i$ is uncorrelated with $\bar{y}_j$ for $i \neq j$. It follows that

$$V(\hat{\bar{Y}}) = \sum_{k=1}^{K} f_k^2 V(\bar{y}_k) + \underset{\sim}{\bar{Y}} \underset{\sim}{V}(\hat{\underset{\sim}{f}}) \underset{\sim}{\bar{Y}}' + O(n^{-2}) \quad , \tag{10}$$

where $V(\bar{y}_k)$ is the variance of the sample mean for category k, $\underset{\sim}{\bar{Y}} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_K)$ is the vector of population means and $\underset{\sim}{V}(\hat{\underset{\sim}{f}})$ is the K x K covariance matrix of $\hat{\underset{\sim}{f}}$. The variance of $\hat{\bar{Y}}$ can be estimated by substituting the appropriate estimators for the parameters,

$$\hat{V}(\hat{\bar{Y}}) = \sum_{k=1}^{K} \hat{f}_k^2 \hat{V}(\bar{y}_k) + \underset{\sim}{\bar{Y}} \hat{\underset{\sim}{V}}(\hat{\underset{\sim}{f}}) \underset{\sim}{\bar{Y}}' \quad . \tag{11}$$

III. Example [1]

A survey of households in several communities in north-central Iowa was taken in 1975 to determine people's views of the community in which they lived. We consider the variables: "Age of Respondent" and "Number of Years Residing in Community". An initial mailing was made to 1023 households. After two additional mailings a total of 787 eligible units had responded. We analyze the sample of 1023 households as a simple random sample.

The respondents were divided into seven categories on the basis of age. The age categories and the number of responses to each mailing are given in Table 1.

Using the model (1), (2) and the method of maximum likelihood, the estimated fractions in the age categories are

$$\hat{\underset{\sim}{f}} = \begin{pmatrix} 0.081 & 0.130 & 0.166 & 0.167 \\ (0.016) & (0.014) & (0.017) & (0.014) \\ 0.167 & 0.130 & 0.159 \\ (0.014) & (0.013) & (0.029) \end{pmatrix} \quad , \tag{12}$$

where the numbers in parentheses are the estimated standard errors of the estimates. The estimated response probabilities are

$$\hat{\underset{\sim}{q}} = \begin{pmatrix} 0.376 & 0.518 & 0.464 & 0.642 \\ (0.108) & (0.065) & (0.064) & (0.046) \\ 0.627 & 0.594 & 0.313 \\ (0.048) & (0.057) & (0.088) \end{pmatrix} \quad , \tag{13}$$

and the estimated fraction of hard-core non-respondents is

$$1 - \hat{\gamma} = \underset{(0.035)}{0.109} \quad . \tag{14}$$

Table 1. Responses by age and fitted value (in parentheses) for Community Study

Age	First mailing	Second mailing	Third mailing
15-24	28 (27.82)	17 (17.36)	11 (10.82)
25-34	63 (61.24)	26 (29.52)	16 (14.24)
35-44	73 (70.18)	32 (37.64)	23 (20.18)
45-54	97 (97.52)	36 (34.95)	12 (12.53)
55-64	97 (95.24)	32 (35.52)	15 (13.24)
65-74	72 (70.65)	26 (28.70)	13 (11.65)
75+	47 (45.41)	28 (31.18)	23 (21.41)
	No response after 3 calls	236 (236)	

The estimates given in (12), (13) and (14) are substituted for the associated parameters in the model (1), (2) to give $\hat{\pi}_{rk}$ ($r = 1, 2, \ldots, R$, $k = 1, 2, \ldots, K$) and $\hat{\pi}_0$. Using $\hat{n}_{rk} = n\hat{\pi}_{rk}$ and $\hat{n}_0 = n\hat{\pi}_0$ the likelihood ratio chi-square statistic is

$$\chi^2 = 2 \sum_{r=1}^{R} \sum_{k=1}^{K} n_{rk} \log(\hat{n}_{rk}^{-1} n_{rk}) \quad .$$

For the model given by (1), (2), the calculated value is 3.702 and has 22-14-1 = 7 degrees of freedom. This value is not significant ($.950 \leq p \leq .975$), and indicates that the model is compatible with the data. The observed number of responses and the responses estimated by (12), (13), and (14) are given in Table 1.

Inspection of the estimates in (13) suggests a quadratic relationship between $\hat{q}_k$ and the median age in the k^{th} category, say A_k.

Replacing q_k in (13) and (2) by

$$q_k = \beta_0 + \beta_1 A_k + \beta_2 A_k^2 \quad , \tag{15}$$

and applying the method of maximum likelihood gives the following estimate of $\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2)$:

$$\hat{\underset{\sim}{\beta}} = \begin{pmatrix} -0.166 & 0.029 & -0.00027 \\ (0.205) & (0.008) & (0.00007) \end{pmatrix} \quad .$$

[1] We are grateful to Professor Willis Goudy of the Department of Sociology at Iowa State University for making these data available to us.

Since β_1 and β_2 are both more than three times as large as their standard errors, we can reject the hypothesis that the components of $\underset{\sim}{q}$ are identical. The generalized likelihood ratio test given by $\sup_{q_k \in \Omega} \log L - \sup_{q_k \in \Omega_0} \log L$, where Ω is the set of unrestricted q_k and $\Omega_0 = \{q_k : q_k$ has the form (15)$\}$ gives $-2 \log \lambda = 5.74$, which is asymptotically distributed as X^2_{14-10}. Therefore, the quadratic model is an acceptable model for the response probabilities.

This example can be approximated by the model (1), (2), and (6) with $f_{\underset{\sim}{\theta}}(X) = N(51, (15)^2)$, $\gamma = 0.9$, and $\underset{\sim}{\beta}' = (0, 0.027, -0.00026)$, where the data are categorized by the boundaries defined in Table 1. We set $A_1 = 7.5$, $A_9 = 83$, and the remaining A_i equal to the midrange of the categories. Then the efficiency of the discretized model to the continuous model for estimating the mean of X is 0.942; that is, about 6% of the information of the continuous model is lost by discretizing.

The mean Number of Years in Community is given by age category in Table 2.

Table 2. Years in Community by Age Category

Category	$\bar{y}_k$	s_k^2	$[\hat{V}(\bar{y}_k)]^{\frac{1}{2}}$
1	10.1	80.5	1.20
2	16.2	230.8	1.48
3	20.1	162.7	1.13
4	27.8	290.4	1.41
5	37.3	346.0	1.56
6	42.2	449.5	2.04
7	54.4	618.6	2.57

Using (9) and (11) and the estimated population fractions from (12), the estimated mean years in community is

$$\hat{\bar{Y}} = \Sigma_{k=1}^{7} \hat{f}_k \bar{y}_k = \underset{(1.14)}{31.40} \quad .$$

In contrast, the simple mean of the observations is

$$\bar{y} = \underset{(0.80)}{30.70} \quad ,$$

where the number in parenthesis is the standard error estimated under the (incorrect) assumption that the 787 observations are a simple random sample from the entire population. The weighted mean is larger than the simple mean because of the estimated low response rate of older people. Note that the model recognizing the nonresponse has a larger estimated standard error.

To illustrate the effect of allocation of the hard core nonrespondents on the estimators, we introduce an alternate assumption. Assume that the proportion of the population in category k that are hard core nonrespondents is equal to $\beta(1-q_k)$, where β is a parameter to be estimated. This assumption leads to cell probabilities

$$\pi_{rk} = [1-\beta(1-q_k)](1-q_k)^{r-1} q_k f_k \quad , \tag{16}$$

$$\pi_0 = \beta \sum_{k=1}^{K} (1-q_k) f_k + \sum_{k=1}^{K} [1-\beta(1-q_k)](1-q_k)^R f_k \quad . \tag{17}$$

Fitting this model to the data one obtains these estimates:

$$\hat{\underset{\sim}{f}} = \begin{pmatrix} \underset{(0.018)}{0.083} , & \underset{(0.015)}{0.130} , & \underset{(0.019)}{0.168} , & \underset{(0.014)}{0.161} , \\ \underset{(0.014)}{0.162} , & \underset{(0.013)}{0.128} , & \underset{(0.031)}{0.167} \end{pmatrix} , \tag{18}$$

$$\hat{\underset{\sim}{q}} = \begin{pmatrix} \underset{(0.108)}{0.376} , & \underset{(0.065)}{0.518} , & \underset{(0.064)}{0.464} , & \underset{(0.046)}{0.642} , \\ \underset{(0.047)}{0.627} , & \underset{(0.057)}{0.594} , & \underset{(0.088)}{0.313} \end{pmatrix} , \tag{19}$$

$$\hat{\beta} = \underset{(0.083)}{0.223} \quad . \tag{20}$$

The chi-square statistic for lack of fit is 3.702 with 7 degrees of freedom.

Note that the alternative hypothesis for hard core nonrespondents does not affect the model estimates of $\underset{\sim}{q}$, nor is the chi-square lack of fit statistic affected. Using the data in Table 2 and the estimates (18), (19), (20) we obtain

$$\sum_{k=1}^{7} \hat{f}_k y_k = \underset{(1.23)}{31.35} \quad .$$

One factor of the data collection procedure which has been ignored is the fact that not all callbacks were executed with the same intensity. The first call was the initial mailing of the questionnaire followed by a reminder postcard. The second call was the mailing of a second questionnaire, while the third call was a third copy of the questionnaire with a certified letter explaining the importance of being a part of the survey. It is reasonable to suppose that the conditional probability of a unit's response was larger for the third call. To formulate a modified model, let q_k be the probability that a unit in the k^{th} category will respond to the first call and let $\delta (\delta \leq 1)$ be the multiplicative effect of the certified letter. Then, $\delta(1-q_k)$ is the probability that a unit in the k^{th} category will not respond to the third call, given that it did not respond to the first call or to the second call. Returning to the initial assumption that the proportion of hard core is the same in each category, the cell probabilities of the multinomial model become

$$\pi_{rk} = \begin{cases} \gamma(1-q_k)^{r-1} q_k f_k & \text{if } r \leq 2 \\ \gamma(1-q_k)^{r-1}[1-\delta(1-q_k)] f_k & \text{if } r = 3 \end{cases}$$

$$\pi_0 = (1-\gamma) + \gamma\,\delta\,\Sigma_k (1-q_k)^R\, f_k \quad .$$

The estimates of the parameters are

$$\underset{\sim}{\hat{f}} = \begin{pmatrix} 0.074 & , & 0.133 & , & 0.165 & , & 0.179 & , \\ (0.010) & & (0.012) & & (0.014) & & (0.015) & \\ 0.178 & , & 0.138 & , & 0.133 & & & \\ (0.014) & & (0.013) & & (0.016) & & & \end{pmatrix}, \qquad (21)$$

$$\underset{\sim}{\hat{q}} = \begin{pmatrix} 0.474 & , & 0.564 & , & 0.524 & , & 0.665 & , \\ (0.086) & & (0.059) & & (0.061) & & (0.042) & \\ 0.650 & , & 0.624 & , & 0.433 & & & \\ (0.044) & & (0.052) & & (0.080) & & & \end{pmatrix},$$

$$\hat{\gamma} = \begin{matrix} 0.811 \\ (0.041) \end{matrix},$$

$$\hat{\delta} = \begin{matrix} 0.598 \\ (0.357) \end{matrix} \quad .$$

(Note that under asymptotic normality, δ is not significantly different from 1, at any reasonable α-level.)

The likelihood ratio chi-square for lack of fit is 1.158 with 6 degrees of freedom. Using the estimates (21) and the means of table 2, we have

$$\Sigma_k \hat{f}_k\, y_k = \begin{matrix} 30.94 \\ (0.82) \end{matrix} \quad .$$

ACKNOWLEDGEMENT

This research was supported by Joint Statistical Agreement 80-6 with the U.S. Bureau of the Census.

REFERENCES

Brewer, K. R. (1979), "Discussion of Cassel-Särndal-Wretman," Symposium on Incomplete Data, Washington, D.C., Preliminary Proceedings, 219-224.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1979), "Some Uses of Statistical Models in Connection with the Nonresponse Problem," Symposium on Incomplete Data, Washington, D.C., Preliminary Proceedings, 188-212.

Deming, W. E. (1953), "On a Probability Mechanism to Attain an Economic Balance between the Resultant Error of Response and the Bias of Nonresponse," Journal of the American Statistical Association, 48, 743-772.

Frankel, L. and Dutka, S. (1979), "Survey Design in Anticipation of Non-Response and Imputation," Symposium on Incomplete Data, Washington, D. C., Preliminary Proceedings, 72-94.

Politz, A. and Simmons, W. (1949), "An Attempt to Get the 'Not-at-Homes' into the Sample Without Callbacks," Journal of the American Statistical Association, 44, 9-31.

Politz, A. and Simmons, W. (1950), "Note on an Attempt to Get 'Not-at-Homes' into the Sample Without Callbacks," Journal of the American Statistical Association, 45, 136-137.

Proctor, C. (1977), "Two Direct Approaches to Survey Nonresponse: Estimating a Proportion with Callbacks and Allocating Effort to Raise the Response Rate," Proceedings of the Social Statistics Section, ASA, 1977, 284-290.

Rao, C. R. (1973), Linear Statistical Inference and its Applications. New York: Wiley.

Sarndal, C. E. and Hui, T. K. (1980), "Estimation for Nonresponse Situations: To What Extent Must We Rely on Models?", Statistics Report M543, Florida State University.

Simmons, W. (1954), "A Plan to Account for 'Not-at-Homes' by Combining Weighting and Callbacks," Journal of Marketing 42-53.

Thomsen, I. and Siring, E. (1979), "On the Causes and Effects of Non-Response: Norwegian Experiences," Symposium on Incomplete Data, Washington, D.C., Preliminary Proceedings, 21-64.