PREDICTIVE LIKELIHOOD IN NONRESPONSE PROBLEMS

Jan F. Bjørnstad, University of Trondheim and Heidi K. Walsøe, MMI, Oslo
Jan F. Bjørnstad, Inst. of Math. and Statistics, College of Arts and Science, University
of Trondheim, N-7055, Dragvoll, Norway

KEY WORDS: Nonresponse, callbacks, imputation, likelihood

1. INTRODUCTION

The aim of the paper is to discuss a likelihood-based approach to survey sampling when nonresponse is present. Consider a finite population consisting of N units where N is known. The units are labelled 1, ... , N, and y_i is the value of a univariate variable of interest for unit i. A sample s (a subset of $\{1, \ldots, N\}$) of size n is chosen according to some sampling design p(s | y), a probability distribution over all subsets of $\{1, \ldots, N\}$. We shall assume that p(s | y) = p(s) , i.e. the probability of choosing s does not depend on y. The aim is to make inference about y = $(y_1, \ldots, y_N)$, usually in the form of a function of y. In this paper we are concerned with estimating the total $t = \Sigma y_i$. We regard y as a realized value of a random vector Y with distribution characterized by unkown parameters θ. Such *modelling* of the population is in principle no different from the usual type of modelling we regularly do in statistical analysis. Under a population model inference about t becomes a *prediction* problem about the unobserved part of t. The sampling design is ignorable according to the likelihood principle. Hence, all analysis is done conditional on the actual s chosen

In practically all sample surveys one has to expect that some units in the survey do not respond, i.e., we have *nonresponse* in the survey. The non-response will usuallly be at least 5-10%, and it is not uncommon with a nonresponse of 30-40%. In order to perform a realistic and relevant statistical analysis it is therefore necessary to include into the population model a model of the process that leads to non-response. To describe the response pattern we define the response variables $R_i = 1$ if unit i responds and 0 otherwise. (Y_i, R_i) are assumed to be i.i.d. for i =1, . . . ,N,and the Y_i's are discrete with possible values 0,1,2,... . The common distribution is denoted by $f_\theta(y) = P_\theta(Y_i = y)$, and the conditional distribution of R_i given $Y_i = y_i$ is $f_\psi(r_i|y_i)$. In general, $f(\cdot)$ and $f(\cdot|\cdot)$ denote the distribution and conditional distribution of the enclosed variables. Let $\mu(\theta)$ and $\sigma^2(\theta)$ denote mean and variance for Y_i. We shall consider surveys where call-backs are made to the nonrespondents, and the non-response model suggested by Thomsen and Siring (1983) is studied in Section 4.

The response sample is $s_r = \{i \in s : r_i = 1\}$, and n_r is the size of s_r .The observed y_i - values in s are denoted by $y_r = (y_i: i \in s \ \& \ r_i = 1)$. The problem is to make inference about the total t which is a realized value of $T = \sum_{i \in s_r} Y_i + Z_1 + Z_2$, where $Z_1 = \sum_{i \in s-s_r} Y_i$ and $Z_2 = \sum_{i \notin s} Y_i$. Since $\sum_{i \in s_r} Y_i$ is observed, estimating t can be regarded as the problem of predicting the value z of $Z = Z_1 + Z_2$. To predict z we shall use a predictive likelihood approach and the mean imputation method.

Section 2 reviews the the general concept of predictive likelihood and shows how predictors and prediction intervals can be constructed from a predictive likelihood. Section 3 describes the mean imputation method for estimating the total t. Section 4 considers surveys where callbacks are made to nonrespondents. The mean imputation method and the marginal profile predictive likelihoods for y_i, $i \notin s_r$, are considered

Section 5 considers a fertility study in Norway from 1977 with n = 5047 women. After 3 calls the mean number of live births in the response sample (of size 3438) equals 1.662. By fitting the data with a mixed Poisson model the mean imputation method gives the estimate 1.550 of t/N, while the mean predictor based on the marginal predictive likelihoods gives the estimate 1.526. From registers it is found that for the whole sample s the mean is 1.50. Hence, the likelihood method and the imputation method reduce the bias by 84.0 % and 69.1 % , respectively.

2. PREDICTIVE LIKELIHOOD

The main aim of the paper is to apply the likelihood approach to the prediction of the unobserved part z of the total t. This section gives a short description of likelihood prediction generally. For a more complete exposition we refer to Bjørnstad (1990). Let Y = y be the data. The problem is to predict the unobserved or future value z of a random variable Z, usually by a predictor and a prediction interval. (Y,Z) has a density or discrete probability function $f_\theta(y,z)$. The joint likelihood for the two unknown quantities z and θ is defined to be $l_y(z,\theta) = f_\theta(y,z)$. The aim is to develop a likelihood for z, L(z | y), by eliminating θ from l_y. Any such likelihood is called a predictive likelihood.

Different ways of eliminating θ give rise to different L. One way is by maximizing $l_y(z,\theta)$ with respect to θ, giving us the so-called profile predictive likelihood: $L_p(z \mid y) = \max_\theta f_\theta(y,z)$.

L_p typically works well when θ has low dimension. If θ consists of many parameters L_p will usually be misleadingly precise and needs to be modified. Such modifications have been suggested by Butler (1986 rejoinder, 1989) and are also considered in Bjørnstad (1990).

Any L considered is assumed to be normalized as a probability distribution in z. The mean of L is called the predictive expectation, $E_p(Z)$, and is a natural predictor for z. L(z | y) gives us an idea of how likely different z-values are in light of the data, and can be used to construct prediction intervals for z. An interval $I_y = (a_y,b_y)$ is a (1-α) *predictive* interval based on L if $\int_{a_y}^{b_y} L(z \mid y)dz = 1 - \alpha$. If L is unimodal the shortest (1-α) predictive interval is of the form $I_y = \{z: L(z \mid y) \geq c\}$.

3. IMPUTATION

A general and common method for estimating t = Σy_i when we have nonresponse is by imputation of the missing values in the nonresponse group. Let $\hat{t}$ be a predictor for t based on a complete sample s, $\hat{t} = \sum_{i \in s} y_i + \hat{z}_2$, where $\hat{z}_2$ is a predictor for $z_2 = \sum_{i \notin s} y_i$, based on $\mathbf{y}_s = (y_i : i \in s)$. In case of nonresponse we must modify $\hat{t}$ since only $\mathbf{y}_r$ is available. *One* general way of adjusting $\hat{t}$ is to *impute* the missing values in $s\text{-}s_r$ with values y_i^* based on the observed part of s. Let now $y_c(s)$ denote the *constructed* complete y-sample, i.e.: $y_c(s) = (y_{ci} : y_{ci} = y_i$ for $i \in s_r$ & $y_{ci} = y_i^*$ for $i \in s\text{-}s_r)$.

The imputation predictor $\hat{t}_I$ is then given by $\hat{t}$ based on $y_c(s)$. I.e.,

$$\hat{t}_I = \sum_{i \in s_r} y_i \quad + \quad \sum_{i \in s\text{-}s_r} y_i^* \quad + \hat{z}_{2c}$$

where $\hat{z}_{2c}$ is $\hat{z}_2$ based on $y_c(s)$. The imputation predictor for $z = \sum_{i \notin s_r} y_i$ equals $\hat{z}_I = \sum_{i \in s\text{-}s_r} y_i^* + \hat{z}_{2c}$.

There are different methods of imputation that are used in practice.(See Little and Rubin (1987, ch. 4.5) for a review.) We consider a method used by Greenlees et al. (1982) that is natural under a population model. The expected value of Y_i for $i \in s\text{-}s_r$ is $E_{\theta,\psi}(Y_i \mid r_i = 0)$. Let $\hat{\theta},\hat{\psi}$ be maximum likelihood estimates (mle) of θ,ψ . The mean imputed value for y_i is then $y_i^* = E_{\hat{\theta},\hat{\psi}}(Y_i \mid r_i = 0)$.

4. A NONRESPONSE MODEL FOR SURVEYS WITH CALLBACKS

We shall look at a nonresponse model appropriate when callbacks are made in an interview survey. Callbacks are necessary in order to reduce the sensitivity of the statistical analysis to nonignorable nonresponse. Revisiting nonrespondents reduces the amount of nonresponse and at the same time provides information about the nonresponse group, making it possible to check and possibly adjust the model for the response mechanism.

Let k denote the maximum number of revisits. The model we shall consider has been proposed by Thomsen and Siring (1983). For every attempt to make an interview there are three possible outcomes:

(I) response

(II) no response, but it is decided to make a callback

(III) no response and classified as nonresponse (refusal) .

For unit i with $y = y_i$ it is assumed that $P(I) = p_y$ depends on y and $P(III) = f$. For callbacks it is assumed that a larger effort is made to get in contact with group (II), such that at the second attempt or later, $P(I) = \Delta_y p_y$ with $\Delta_y > 1$ expected. P(III) is assumed to equal f on all visits.

For each unit $i \in s$ define, for $j=1,\dots, k$,

$R_{ij} = 1$ if unit i responds on visit j

0 otherwise .

The random vectors $(R_{i1},\dots,R_{ik})$,$i = 1,\dots, N$ are assumed to be independent.

$$P(R_{ij} = 1|Y_i = y)$$
$$= p_y \qquad \text{if } j=1$$
$$= (1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y \qquad \text{if } 2\le j\le k .$$

(Y_i,R_i) are independent and $P(R_i = 1 \mid Y_i = y) = p_y + \sum_{j=2}^{k} (1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y$. The response mechanism (RM) is ignorable if and only if Δ_y,p_y are independent of y. The data consists of $(\mathbf{y}_r,\mathbf{r}_s)$ where $\mathbf{r}_s = (r_{i1},\dots,r_{ik})$, $i \in s$.

We shall consider a stratified model for the RM by assuming that Y_i rarely takes values > m for some m . Define then m+1 strata where, for $j = 0, \dots,m-1$, stratum $j = \{i : y_i = j \}$, and stratum $m = \{i : y_i \ge m\}$. It is assumed that p_y and Δ_y are constant within each stratum. Hence, $p_y = p_m$ and $\Delta_y = \Delta_m$ if $y \ge m$, and the RM parameters are $\psi = (f, (p_0,\dots p_m),(\Delta_0,\dots, \Delta_m))$.

Imputation. We shall first consider the mean imputation method described in section 3. Let, for $i=0,\dots,m$ og $j=1,\dots,k$, v_{ij} = #(responses in post-stratum i at visit no. j) . With the Y_i's i.i.d., the imputation predictor is based on $\hat{t} = N\bar{y}_s$, and given by

$$\hat{t}_I = \frac{N}{n}\Big(\sum_{i\in s_r} y_i + \sum_{i\in s - s_r} y_i^*\Big)$$

where , with $\mu_{Y|0}(\theta,\psi) = E_{\theta,\psi}(Y_i \mid r_i = 0)$, $y_i^* = \mu_{Y|0}(\hat{\theta},\hat{\psi})$. Here $(\hat{\theta},\hat{\psi})$ are mle and

$$\mu_{Y|0} = \sum_{y=1}^{\infty} yP(Y_i=y \mid R_i=0) = \sum_{y=1}^{\infty} y\frac{f_\theta(y)P(R_i = 0 |y)}{P(R_i = 0)}$$
$$= \sum_{y=1}^{m-1} y\frac{f_\theta(y)(1 - p_y - \sum_{j=2}^{k} (1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y)}{P(R_i= 0)} +$$
$$\{\sum_{y=m}^{\infty} yf_\theta(y)\}\frac{(1- p_m - \sum_{j=2}^{k} (1-p_m-f)(1-\Delta_m p_m-f)^{j-2}\Delta_m p_m)}{P(R_i= 0)} . \quad (1)$$

Here,

$$P(R_i = 0) = \sum_{y=0}^{m-1} f_\theta(y)(1 - p_y - \sum_{j=2}^{k} (1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y)$$
$$+ \quad \{\sum_{y=m}^{\infty} f_\theta(y)\}(1 - p_m - \sum_{j=2}^{k} (1-p_m-f)(1-\Delta_m p_m-f)^{j-2}\Delta_m p_m).$$

Let $s_{rj} = \{i \in s_r : r_{ij} = 1\}$.$(\hat{\theta},\hat{\psi})$ maximizes the likelihood $l(\theta,\psi) = f_{\theta,\psi}(\mathbf{y}_r,\mathbf{r}_s) =$

$$\prod_{j=1}^{k} \prod_{i\in s_{rj}} P(Y_i=y_i \cap R_{ij}=1) \prod_{i\in s-s_r} P(R_i=0)$$
$$=$$
$$\{\prod_{i\in s_r} f_\theta(y_i)\} [P(R_i=0)]^{n-n_r}\prod_{y=0}^{m} p_y^{v_{y1}}$$
$$\prod_{j=2}^{k} \prod_{y=0}^{m} [(1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y]^{v_{yj}}. \quad (2)$$

An easier imputation method (and typically a good approximation of $y_i^* = \mu_{Y|0}(\hat{\theta},\hat{\psi})$ for large n_r,n) can be derived by expresssing $E_{\theta,\psi}(Y_i \mid r_i = 0)$ in the following way:

$$E(Y_i \mid r_i = 0) = \frac{E(Y_i) - P(R_i=1)E(Y_i|R_i=1)}{P(R_i=0)} .$$

Now, $P(R_i = 1)$ can be estimated by n_r/n, and $E(Y_i|R_i=1)$ by $\bar{y}_r$. An estimate of $E(Y_i \mid r_i = 0)$ is therefore given by

$$y_i^* = \frac{\mu(\hat{\theta}) - \frac{n_r}{n}\bar{y}_r}{\frac{n-n_r}{n}} = \frac{n\mu(\hat{\theta}) -n_r \bar{y}_r}{n-n_r} . \quad (3)$$

Based on (3) :

$$\tilde{t}_I = \frac{N}{n}\Big(\sum_{i\in s_r} y_i + \sum_{i\in s - s_r} y_i^*\Big) =$$
$$\frac{N}{n}\Big(n_r\bar{y}_r + (n-n_r)\frac{n\mu(\hat{\theta}) -n_r \bar{y}_r}{n-n_r}\Big) = N\mu(\hat{\theta}) . \quad (4)$$

Predictive likelihood. A direct predictive likelihood for z is based on $f_{\theta,\psi}(\mathbf{y}_r,\mathbf{r}_s,z)$ which is extremely complex. In order to derive a likelihood based predictor we need, however, only to consider marginal likelihoods for each y_i , $i \in s-s_r$, and each y_i, $i \notin s$. These likelihoods are based on $f_{\theta,\psi}(\mathbf{y}_r,\mathbf{r}_s,y_i)$ for each $i \notin s_r$. We shall derive the profile predictive likelihood for each y_i , $i \notin s_r$. Let L_{p1} be the likelihood for y_q, $q \in s-s_r$, and L_{p2} for y_q, $q \notin s$. Then

$$L_{p1}(y_q \mid \mathbf{y}_r,\mathbf{r}_s) = \max_{\theta,\psi} f_{\theta,\psi}(\mathbf{y}_r,\mathbf{r}_s,y_q) =$$

$$\max_{\theta,\psi}[\{\prod_{i\in s_r} f_\theta(y_i)\} [P(R_i=0)]^{n-n_r-1} \prod_{y=0}^{m} p_y{}^{V_{y1}}$$

$$\prod_{j=2}^{k} \prod_{y=0}^{m} [(1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y]^{V_{yj}}\} f_\theta(y_q) x$$

$$\{1 - p_{y_q} - \sum_{j=2}^{k} (1-p_{y_q}-f)(1-\Delta_{y_q}p_{y_q}-f)^{j-2}\Delta_{y_q}p_{y_q}\}] .$$

$$L_{p2}(y_q \mid \mathbf{y}_r,\mathbf{r}_s) = \max_{\theta,\psi} f_{\theta,\psi}(\mathbf{y}_r,\mathbf{r}_s,y_q) =$$

$$\max_{\theta,\psi}[\{\prod_{i\in s_r} f_\theta(y_i)\} [P(R_i=0)]^{n-n_r} \prod_{y=0}^{m} p_y{}^{V_{y1}}$$

$$\prod_{j=2}^{k} \prod_{y=0}^{m} [(1-p_y-f)(1-\Delta_y p_y-f)^{j-2}\Delta_y p_y]^{V_{yj}} f_\theta(y_q)].$$

L_{p1} and L_{p2} must be derived numerically for each possible value of y_q, and then normalized. As an approximation we shall assume that the likelihood of values of $y_q \geq m+1$ equals zero, compute $L_{p1}(y_q \mid \mathbf{y}_r,\mathbf{r}_s)$ and $L_{p2}(y_q \mid \mathbf{y}_r,\mathbf{r}_s)$ for $y_q = 0,1,\ldots, m$ and normalize such that

$$\sum_{y=0}^{m} L_{p1}(y \mid \mathbf{y}_r,\mathbf{r}_s) = \sum_{y=0}^{m} L_{p2}(y \mid \mathbf{y}_r,\mathbf{r}_s) = 1.$$

Let E_{p1} and E_{p2} be the mean of these normalized L_{p1} and L_{p2}, respectively. Since $Z = Z_1 + Z_2 = \sum_{i \in s-s_r} Y_i + \sum_{i \notin s} Y_i$, a predictor for Z based on these marginal predictive likelihoods is

$$E^*_p(Z) = (n-n_r)E_{p1} + (N-n)E_{p2} . \qquad (5)$$

We have used the notation $E^*_p(Z)$ to indicate that this predictive mean of Z was not derived from a normalized predictive likelihood for Z.

5. AN EXAMPLE - THE NORWEGIAN FERTILITY SURVEY

In 1977 a fertility survey was performed in Norway (see also Thomsen and Siring (1983)). A sample of n = 5047 women between the ages of 18 and 44 was selected. N = 695909. A maximum of 8 calls were made. We shall use the data from the first 3 calls. The variable of interest is the number of live births for each woman in the survey. We choose m = 6 . The table below gives the number of responses in each stratum on each call.

Table 1. Number of responses in each call.

	Call			
Stratum	1	2	3	all calls
0	311	387	188	886
1	258	248	134	640
2	497	410	158	1065
3	261	199	88	548
4	107	79	30	216
5	37	15	9	61
6	12	7	3	22
Total	1483	1345	610	$n_r = 3438$

The observed sample mean is $\bar{y}_r = 1.662$. From registers it is found that, for the whole sample s, $\bar{y}_s = 1.50$ .Evidently the nonresponse was largest among women with few or no children. Thomsen and Siring (1983) use a moment-type estimation method for $\bar{y}_s$, (without assuming a population model) that gives the estimate $\hat{\bar{y}}_s = 1.593$. This can be regarded as a distribution-free estimate of $\mu(\theta)$.

We expect that the probability of response in stratum 1 to be at least not significantly lower than for stratum 0. By looking at the data it therefore seems clear that a pure Poisson model is not appropriate. We consider instead a mixed Poiison model given by

$$f_\theta(y) = \varepsilon\frac{\lambda_1^y e^{-\lambda_1}}{y!} + (1-\varepsilon)\frac{\lambda_2^y e^{-\lambda_2}}{y!} .$$

We find that the maximum likelihood estimates are $\hat{\varepsilon} = .0413$, $\widehat{\lambda_1} = 0$, $\widehat{\lambda_2} = 1.5990$, and $\hat{f} = .0475$. The other maximum likelihood estimates are in table 2.

Table 2. Maximum likelihood estimates of Δ_i and p_i .

Stratum	$\widehat{\Delta_i}$	$\hat{p}_i$
0	1.809	.262
1	1.079	.167
2	1.482	.392
3	1.362	.388
4	1.350	.410
5	.781	.437
6	1.069	.425

The mle imputation method gives $y_i{}^* = \mu_{Y|0}(\hat{\theta},\hat{\psi}) = 1.311$ with $\hat{t}_I/N = 1.550$. Hence the imputation method reduces the bias in $\bar{y}_r$ by 69.1 % . The simplified approach (4) gives $\tilde{t}/N = \mu(\hat{\theta}) = \hat{\varepsilon}\hat{\lambda}_1 + (1-\hat{\varepsilon})\hat{\lambda}_2 = 1.533$.

The *marginal* predictive likelihood approach gives the following normalized L_{p1} and L_{p2}.

Table 3. Normalized marginal predictive likelihoods

y	$L_{p1}(y\|y_r,r_s)$	$L_{p2}(y\|y_r,r_s)$
0	.1874	.2354
1	.5607	.3102
2	.1243	.2478
3	.0779	.1321
4	.0294	.0529
5	.0162	.0170
6	.0041	.0046
E_{pj}	1.2662	1.5261

From (5), a likelihood predictor of z is $E_p^*(Z) = (n-n_r)E_{p1} + (N-n)E_{p2} = 1{,}056{,}361.8$. This gives the following predictor of t/N: $t^*/N = [n_r\bar{y}_r + E_p^*(Z)]/N = 1.526$, reducing the bias in $\bar{y}_r$ by 84.0%.

The low estimate of p_1 indicates that even the mixed Poisson model for **Y** may not be appropriate. Another possible model that can be tried is a truncated Poisson model. Still, we are able to reduce the bias due to nonresponse by a substantial margin even with the mixed model.

REFERENCES

Bjørnstad, J.F. (1990). Predictive Likelihood: A review (with discussion). *Statistical Science*,5, 242-265.

Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion). *J.R. Statist. Soc.*, B48, 1-38.

Butler, R.W. (1989). Approximate predictive pivots and densities. *Biometrika* ,76, 489-501.

Greenlees, J.S., Reece,W.S., and Zieschang, K.D. Imputation of missing values when the probability of response depends on the variable being imputed. *J. Am.Statist. Ass*, 77, 251-261.

Little, R.J.A. and Rubin,D.B. (1987). *Statistical Analysis with Missing Data.* Wiley, N.Y.

Thomsen, I. and Siring, E. (1983). On the causes and effects of non-response. Norwegian experiences. In "*Incomplete Data in Sample Surveys, vol. 3 , Session 1.* Academic Press.