INTERVAL ESTIMATION IN THE PRESENCE OF NONRESPONSE

Jan F. Bjørnstad and Frode Skjold, The University of Trondheim
Jan F. Bjørnstad, Dept. of Math. and Statistics, College of Arts and Science, University of Trondheim, N- 7055 Dragvoll,Norway

KEY WORDS: Predictive likelihood, stochastic censuring.

1. INTRODUCTION

The paper deals with a likelihood-based approach to survey sampling when nonresponse is present . Consider a finite population consisting of N units where N is assumed known. The units are labelled 1,..., N, and y_i is the value of a univariate variable of interest for unit i. The aim is to make inference about $\mathbf{y} = (y_1,....,y_N)$, usually in the form of a function of $\mathbf{y}$. In this paper we are concerned with estimating the total $t=\Sigma y_i$. A sample s (a subset of $\{1, \ldots,N\}$) of size n is chosen according to some sampling design $p(s \mid \mathbf{y})$, a probability distribution over all subsets of $\{1, \ldots, N\}$. We shall assume that $p(s \mid \mathbf{y}) = p(s)$, i.e. the probability of choosing s does not depend on $\mathbf{y}$.

We regard $\mathbf{y}$ as a realized value of a random vector $\mathbf{Y}$ with distribution characterized by unkown parameters θ. 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 almost 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 nonresponse 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 nonresponse.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 independent for i =1, . . . ,N. We regard situations where auxiliary information $\mathbf{x}_i = (x_{i1}, \ldots,x_{ip})$ is available for all units in the population. The Y_i's are assumed to be normally distributed with mean $\mathbf{x}_i\beta_1$ and variance σ^2 . Here, β_1 is a p-dimensional column-vector . The response mechanism, i.e. the conditional distribution of R_i given $Y_i = y_i$, is modelled by latent variables, a model that was first considered by Heckman (1976).

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 $\mathbf{y_r} = (y_i: i \in s \ \& \ r_i = 1)$. 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 $Z = Z_1 + Z_2$. We shall focus on constructing confidence intervals for z, based on a predictive likelihood approach.

Section 2 reviews the the general concept of predictive likelihood and shows how predictors and confidence intervals can be constructed from a predictive likelihood.

Section 3 describes the response model. The profile predictive likelihood for z is considered. Section 4 contains the results of four simulated cases and Section 5 considers evaluation of the predictive intervals based on the profile predictive likelihood.

2. PREDICTIVE LIKELIHOOD

The main aim of the paper is 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 $f_\theta(y,z)$. This is the joint likelihood for the two unknown quantities z and θ ; $l_y(z,\theta) = f_\theta(y,z)$. The aim is to develop a likelihood for z, $L(z \mid y)$, by elminating 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. An application of L_p to nonresponse problems is given by Bjørnstad & Walsøe (1991). If θ consists of many parameters L_p can 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).

We shall assume that any L considered is 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. $L(z \mid y)$ 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 $(1-\alpha)$-predictive interval based on L if

$$\int_{a_y}^{b_y} L(z \mid y)dz = 1 - \alpha.$$

If L is unimodal the shortest $(1-\alpha)$ predictive interval is of the form $I_y = \{z: L(z \mid y) \geq c\}$.

3. A LATENT MODEL FOR NONRESPONSE

Response for unit i is assumed to be controlled by an unobserved, latent, variable λ_i :

$$R_i = 1 \Leftrightarrow \lambda_i > 0 .$$

The joint distribution of (Y_i,λ_i) is assumed to be bivariate normal with $E(Y_i) = \mathbf{x}_i\beta_1$, $Var(Y_i) = \sigma^2$, $E(\lambda_i) = \mathbf{x}_i\beta_2$, $Var(\lambda_i) = 1$ and $Cov(Y_i,\lambda_i) = \rho\sigma$. It follows that the conditional distribution of R_i given y_i is given by

$$P(R_i = 1 \mid y_i) = \Phi\left(\frac{\mathbf{x}_i\beta_2 + \rho\sigma^{-1}(y_i - \mathbf{x}_i\beta_1)}{\sqrt{1 - \rho^2}}\right) .$$

Let $\theta = (\beta_1,\beta_2,\sigma,\rho)$ and $\mathbf{r}_s = (r_i : i \in s)$. In general, $f(\cdot)$ and $f(\cdot|\cdot)$ denote the distribution and the conditional distribution of the enclosed variables.The profile predictive likelihood is then

$$L_p(z \mid y) = \max_\theta f_\theta(\mathbf{y}_r,\mathbf{r}_s,z), \text{ where}$$

$$f_\theta(\mathbf{y}_r,\mathbf{r}_s,z) = \{\prod_{i \in s_r} f_\theta(y_i)P(R_i = 1|y_i)\}\times \{\prod_{i \in s-s_r} P(R_i = 0)\}f_\theta(z \mid \mathbf{r}_s) .$$

Let us first consider the i.i.d. case where there are no auxiliary variables and $E(Y_i) = \mu$ and $E(\lambda_i) = c$. $Z = Z_1 + Z_2$ where

$Z_1 = \sum_{i \in s-s_r} Y_i$ depends on $\mathbf{r}_s$, and $Z_2 = \sum_{i \notin s} Y_i$ does not. Z_2 is normally distributed with $E(Z_2) = (N-n)\mu$ and $Var(Z_2) = (N-n)\sigma^2$.

Let $\phi(x)$ and $\Phi(x)$ denote the density and distribution function of the N(0,1) distribution. Given $\mathbf{r}_s$, Y_i for $i \in s-s_r$ are i.i.d. with

$$\mu^* = E(Y_i \mid r_i = 0) = \mu - \rho\sigma\frac{\phi(c)}{1-\Phi(c)}$$

$$\sigma^{*2} = Var(Y_i \mid r_i = 0)$$

$$= \sigma^2 + \sigma^2\rho^2\frac{\phi(c)}{1-\Phi(c)}(c - \frac{\phi(c)}{1-\Phi(c)}).$$

It follows that approximately, $Z_1 \mid \mathbf{r}_s$ is $N((n-n_r)\mu^*, (n-n_r)\sigma^{*2})$ and $Z \mid \mathbf{r}_s$ is $N((n-n_r)\mu^* + (N-n)\mu, (n-n_r)\sigma^{*2} + (N-n)\sigma^2)$ such that, with $\theta = (\mu,c,\sigma,\rho)$,

$$f_\theta(\mathbf{y}_r,\mathbf{r}_s,z) \approx$$

$$\{\prod_{i \in s_r}\frac{1}{\sigma}\phi(\frac{y_i-\mu}{\sigma})\,\Phi(\frac{c + \rho\sigma^{-1}(y_i - \mu)}{\sqrt{1-\rho^2}})\}\times$$

$$(1 - \Phi(c))^{n-n_r}\frac{1}{\sqrt{(n-n_r)\sigma^{*2} + (N-n)\sigma^2}}\times$$

$$\phi(\frac{z - (n-n_r)\mu^* - (N-n)\mu}{\sqrt{(n-n_r)\sigma^{*2} + (N-n)\sigma^2}}).$$

For the general regression model, under regularity conditions sufficient for Lindebergs condition to hold we have again that, approximately, Z_1 given $\mathbf{r}_s$ is normal, now with

$$E(Z_1|\mathbf{r}_s) = \sum_{i \in s-s_r} E(Y_i|r_i = 0)$$

$$= \sum_{i \in s-s_r}(\mathbf{x}_i\beta_1 - \rho\sigma\frac{\phi(\mathbf{x}_i\beta_2)}{1 - \Phi(\mathbf{x}_i\beta_2)})$$

and

$$Var(Z_1|\mathbf{r}_s) = \sum_{i \in s-s_r} Var(Y_i|r_i = 0) =$$

$$\sum_{i \in s-s_r}\{\sigma^2 + \sigma^2\rho^2\frac{\phi(\mathbf{x}_i\beta_2)}{1-\Phi(\mathbf{x}_i\beta_2)}(\mathbf{x}_i\beta_2 - \frac{\phi(\mathbf{x}_i\beta_2)}{1-\Phi(\mathbf{x}_i\beta_2)})\}.$$

Z_2 is normally distributed with $E(Z_2) = \sum_{i \notin s}\mathbf{x}_i\beta_1$ and $Var(Z_2) = (N-n)\sigma^2$.

Let $\mu_{z|0} = E(Z \mid \mathbf{r}_s) = E(Z_1|\mathbf{r}_s) + E(Z_2)$ and $\sigma^2_{z|0} = Var(Z \mid \mathbf{r}_s) = Var(Z_1 \mid \mathbf{r}_s) + Var(Z_2)$.

It follows that

$$f_\theta(\mathbf{y}_r,\mathbf{r}_s,z) \approx \{\prod_{i \in s_r}\frac{1}{\sigma}\phi(\frac{y_i-\mathbf{x}_i\beta_1}{\sigma})\times$$

$$\Phi(\frac{\mathbf{x}_i\beta_2 + \rho\sigma^{-1}(y_i - \mathbf{x}_i\beta_1)}{\sqrt{1-\rho^2}})\}\times$$

$$\prod_{i \in s-s_r}(1 - \Phi(\mathbf{x}_i\beta_2))\frac{1}{\sigma_{z|0}}\phi(\frac{z - \mu_{z|0}}{\sigma_{z|0}}).$$

$L_p(z \mid \mathbf{y}_r,\mathbf{r}_s)$ is computed numerically for a finite sufficient set C of chosen z-values and then normalized to be

$$L_{p(z \mid \mathbf{y}_r,\mathbf{r}_s)} / \sum_{z^* \in C} L_{p(z^* \mid \mathbf{y}_r,\mathbf{r}_s)}.$$

95% predictive intervals $I_y = (a_y,b_y)$ are computed by letting a_y and b_y be the lower and upper 2.5% points in L_p:

$$\int_{-\infty}^{a_y} L_p(z \mid \mathbf{y}_r,\mathbf{r}_s)dz = \int_{b_y}^{\infty} L_p(z \mid \mathbf{y}_r,\mathbf{r}_s)dz = .025$$

computed as sums from the discrete numerical version of L_p.

To evaluate the intervals we estimate, by simulation, the unconditional level

$Cl(\theta) = P_\theta(a_Y \le Z \le b_Y)$

and the conditional level

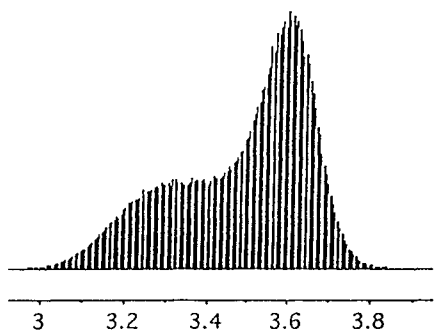
$C_\theta(y) = P_\theta(a_y \le Z \le b_y \mid y)$.

4. SIMULATION CASES

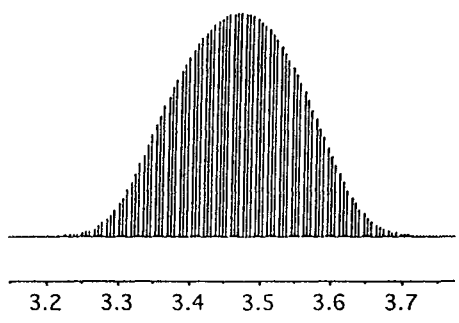
Four simulation cases are presented, two i.i.d. cases (with N = 4637,n = 300 and N = 50, n = 30) and two simple regression cases with the same (N,n) - values. CPU time for computing L_p on a CRAY X-MP/216 was about 8 minutes for the large sample cases and 35 seconds for the small sample cases. In each i.i.d. case one simulaton is made for the parameter values $\mu = 3.5$, $c = .7$, $\sigma = 1$, $\rho = -.5$. The regression cases with $E(Y_i) = \alpha_1 + \beta_1 x_i$ and $E(\lambda_i) = \alpha_2 + \beta_2 x_i$ uses the same y-data as in the corresponding iid cases and x_i - values are simulated from the $N(y_i,.5)$-distribution. It follows that in the regression cases $\alpha_1 = .7$, $\beta_1 = .8$, $\alpha_2 = 2.1$ and $\beta_2 = -.4$. In all four cases $E(T/N) = \mu = 3.5$.

Plots of $L_p(z \mid \underline{y}_r, \underline{r}_s)$ on the scale $z/(N-n_r)$

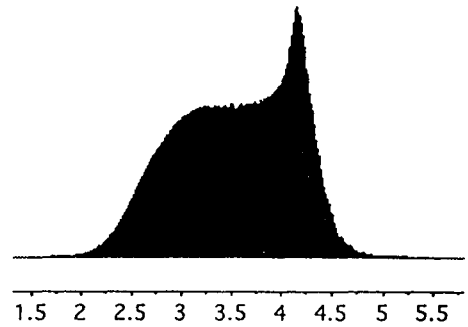
(I) I.i.d.case. N= 4637 , n = 300 , n_r = 230.



(II). Regression case. N= 4637 , n = 300 , n_r = 230.



(III) I.i.d.case. N = 50, n = 30 , n_r = 21.



(IV). Regression case. N = 50, n = 30 , n_r = 21.

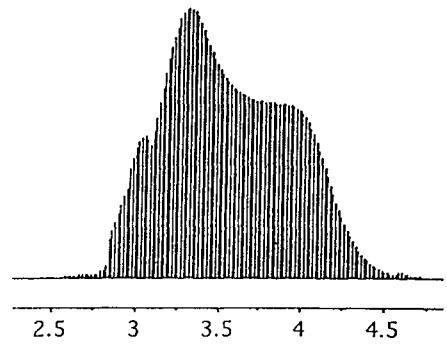


Table 1 summarizes the results of the four cases . The estimate of t/N is

$$\hat{t}_p/N = \frac{1}{N}(n_r\bar{y}_r + E_p(Z)).$$

For imputation, the imputed values in s - s_r are $E(Y_i \mid r_i = 0)$ with the unknown parameters substituted by maximum likelihood estimates, computed by the EM-algorithm. For iid cases 1 and 3, the imputation estimator is based on $\bar{y}_s = \frac{1}{n}\sum_{i \in s} y_i$, while for the regression cases 2 and 4 with $E(Y_i) = \alpha_1 + \beta_1 x_i$, the imputation estimator is based on

$$\hat{t}/N = \frac{1}{N}(\sum_{i \in s} y_i + \sum_{i \notin s}(\hat{\alpha}_1 + \beta_1 x_i))$$

where $(\hat{\alpha}_1 , \beta_1)$ are the least squares estimates based on (x_i,y_i) , $i \in s$.

The predictive interval for t/N based on L_p is given by

$$[\frac{1}{N}(n_r\bar{y}_r + a_y) , \frac{1}{N}(n_r\bar{y}_r + b_y)]$$

Table 1a

	Case 1	Case 2
N	4637	4637-reg
n	300	300
n_r	230	230
$\bar{y}_r$	**3.345**	**3.345**
$E_p(Z)/(N-n_r)$	3.486	3.466
$\hat{t}_p/N$	**3.479**	**3.460**
Imputation	3.652	3.525
$\frac{t}{N}$ - interval	(3.15,3.70)	(3.31,3.62)

Table 1b

	Case 3	Case 4
N	50	50-reg
n	30	30
n_r	21	21
$\bar{y}_r$	**3.353**	**3.353**
$E_p(Z)/(N-n_r)$	3.532	3.553
$\hat{t}_p/N$	**3.457**	**3.469**
Imputation	3.858	3.553
$\frac{t}{N}$ - interval	(3.15,3.70)	(3.11,3.89)

It seems that the likelihood method does better than imputation.

5. EVALUATION OF PREDICTIVE INTERVALS

Estimated confidence levels are computed only for i.i.d. cases. Unconditional levels for the 95% predictive interval are estimated for the following parameter configurations:

$\theta_1 : \mu = 3.5 , c = .7, \sigma = 1 , \rho = -.5$
$\theta_2 : \mu = 3.0 , c = .7, \sigma = 1.5 , \rho = - 5/6$
$\theta_3 : \mu = 3.0 , c = 0, \sigma = 1.5 , \rho = - 5/6$

Table 2. Confidence levels for 95% L_p- intervals

	N	n	$Cl(\theta)$
θ_1	4637	300	$\frac{496}{508} = .977$
θ_1	50	30	$\frac{348}{364} = .956$
θ_2	50	30	$\frac{186}{201} = .925$
θ_2	4637	300	$\frac{193}{205} = .941$
θ_3	4637	300	$\frac{337}{353} = .955$

CPU time for 500 simulations in large sample cases on a CRAY X-MP/216 was about 45 hours.For θ_1 with N = 50 and n = 30, 6 cases of data were simulated and for each case the conditional coverage $C_\theta(y)$ were estimated from 100 000 simulations of z. The results are given in table 3.

Table 3. Conditional coverage probabilities for 95% L_p- intervals

Case	C_{θ_1}
1	.9903
2	.9979
3	.7696
4	.3020
5	.5421
6	.9997

For the case θ_3, N = 4637, n = 150, 32 datacases were simulated and the conditional coverage estimated by 1000 simulations of z. In 50% of the cases the conditional coverage is larger than the nominal level of 0.95. Specifically, in 11 cases $C_{\theta_3} = 0$, in 14 cases $C_{\theta_3} = 1$, and the remaining results are in table 4.

Table 4. Conditional coverage probabilities for 95% L_p- intervals

Case	C_{θ_3}
1	.969
2	.189
3	.117
4	.002
5	.933
6	.999
7	.900

Table 2 indicates that the nominal 95% likelihood intervals have a confidence level close to .95 , although very few cases are considered. The conditional probability of coverage seems to be less than .95 more often than desired, suggesting that L_p may lead to intervals that are too narrow if we use the conditional $C_\theta(y)$ as evaluation criterion.

REFERENCES

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

Bjørnstad, J.F. and Walsøe, H.K. (1991).Predictive likelihood in nonresponse problems. American Statistical Association 1991 Proceedings of the Section on Survey Research Methods, 152-156.

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

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

Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurements*, 5, 475-492.