INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

Research and Publications

Inference on Categorical Survey Response: A Predictive Approach

Sumanta Adhya
Tathagata Banerjee
Gaurangadeb Chattopadhyay

W.P. No.2007-05-07
May 2007

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA

Inference on Categorical Survey Response: A Predictive Approach

Sumanta Adhya
Calcutta University, India
Tathagata Banerjee
Indian Institute of Management Ahmedabad, India
Gaurangadeb Chattopadhyay
Calcutta University, India.

Abstract

We consider the estimation of finite population proportions of categorical survey responses obtained by probability sampling. The customary design-based estimator does not make use of the auxiliary data available for all the population units at the estimation stage. We adopt a model-based predictive approach to incorporate this information and make the estimates more efficient. In the first part of our paper we consider a multinomial logit type model when logit function is a known parametric function of the covariates. We then use it for the prediction of non-sampled responses. This together with sampled responses is used to obtain the estimates of the proportions. The asymptotic biases and variances of these estimators are obtained. The main drawback of this approach is, being a parametric model it may suffer from model misspecification and thus, may lose it's efficiencies over the usual design-based estimates. To overcome this drawback, in the next part of this paper we replace the multinomial logit type model by a nonparametric model using recently developed random coefficients splines models. Finally, we carry out a simulation study. It shows that the nonparametric approach may lead to an appreciable improvement over both parametric and design-based approaches when the regression function is quite different from multinomial logit.

Key Wards: Auxiliary information, Model-based inference, Finite population estimation, Multinomial logit, Random coefficients splines models, Laplace approximation.

Acknowledgement: The work of the first author was supported by a research fellowship from Council of Scientific and Industrial Research (CSIR) (Sanction no.: 9/28(610)/2003-EMR-I), India

---

Note: If one is interested in the proofs, write directly to the author at tathagata@iimahd.ernet.in

1. Introduction

We consider the analysis of survey data obtained from a finite population of size N consisting of a single categorical response d along with a vector of covariates x that may include design variables. The value of the response d is observed for the sampled units. The values of x are assumed to be known for all the units of the population. The customary design-based estimator of the finite population proportions does not make use of the auxiliary data available for all population units at the estimation stage. To utilize this extra information in the estimation stage we adopt the model-based predictive approach (Sarndal and Wright (1984), Firth and Bennett (1998)). We assume that the finite population responses $d_1,...,d_N$ represent a random sample from a superpopulation described by a model (Royall (1970)). In the following we propose estimators based on two models. The first is a multinomial logit type model. It differs from the usual multinomial logit model by assuming that the logit function is a known function of the covariates and is not necessarily linear. The other is based on a purely nonparametric model.

We now introduce the following notations. Suppose each population unit belongs to exactly one of the p categories. The categorical response is $d_i = (d_{i1},...,d_{ip})^T$ for *i*-th unit where $d_{ih} = 1$, if it belongs to the *h*-th ($h = 1,...,p$) category and = 0, otherwise. Also $x_i = (x_{i1},...,x_{iq})^T$, $q \geq 1$ is the vector of auxiliary variables corresponding to *i-th* unit and is assumed to be the same for all categories. Let S be a subset of $\{1,...,N\}$ of size n denoting the set of indices of sampled units and $\overline{S}$, the set of nonsampled units. We let $P_h = N^{-1}\sum_{k=1}^{N} d_{kh}$. Thus P_h denotes the finite population proportion of h-th category.

2.1 Estimator Based on Multinomial Logit Type Model

We assume that d_i 's are independent with

$$P(d_{ih} = 1 \mid x_i) = \pi_h(x_i; \beta), h = 1,..., p \qquad (1)$$

where,

$$\pi_h(x_i; \beta) = \pi_{ih}(\beta) = \exp\{g_h(x_i; \beta_h)\}[1 + \sum_{u=1}^{p-1} \exp\{g_u(x_i; \beta_u)\}]^{-1}, \qquad (2)$$

$\sum_{h=1}^{p} \pi_{ih}(\beta) = 1$, $\beta_h^T = (\beta_{h1},..., \beta_{ha_h})$, $h = 1,..., p-1$, $\beta = (\beta_1^T,..., \beta_{p-1}^T)^T$ and $g_h(.)$,

$h = 1,...,p-1$ are known but arbitrary functions of x_i. Thus for any realization d_i of the response variable, we obtain

$$\mathrm{P}\{d_i \mid x_i; \beta\} = \prod_{h=1}^{p} \{\pi_{ih}(\beta)\}^{d_{ih}} .$$

A standard choice for $g_h(.)$ is linear that is $g_h(x_i; \beta_h) = x_i^T \beta_h$, for all i and h. This gives the well-known multinomial logit model.

The log-likelihood for the sample S is given by

$$l(\beta) = \sum_{i \in S} \sum_{h=1}^{p} d_{ih} \ln \pi_{ih}(\beta) . \qquad (3)$$

Denoting by $\hat{\beta}$ the maximum likelihood estimator (MLE) based on sample observations, the multinomial type model-based predictive estimator of P_h is

$$\hat{P}_{h,m} = N^{-1}[\sum_{i \in S} d_{ih} + \sum_{j \in \bar{S}} \pi_{jh}(\hat{\beta})], \qquad (4)$$

where $\pi_{jh}(\hat{\beta}) = E[d_{jh} \mid \{d_i : i \in S\}, x_1,..., x_N; \beta]_{\beta=\hat{\beta}}$ is the predictor of d_{jh}, the h-th component of the j-th non sampled unit; $E(.)$ is the expectation with respect to the superpopulation model.

2.2. Design-based Estimator

The design-based estimator for this problem is given by

$$\hat{P}_{h,d} = (\sum_{i\in S} \tau_i^{-1})^{-1} \sum_{i\in S} \tau_i^{-1} d_{ih} \,, \tag{5}$$

where $\tau_i (>0)$ is the inclusion probability for the *i*-th sampled unit. In (5), the auxiliary information available through x_i's cannot be incorporated into the estimation process. In theory this is achieved at the survey design stage using appropriate definition of inclusion probabilities; for example, in stratified random sampling, stratification may depends on the known design variable. In multipurpose survey, this is not always possible and one might also like to introduce the auxiliary information at the estimation stage.

On the other hand, assuming some standard design conditions ($\sum_{i\in S} \tau_i^{-1} = N$), the above estimator may shown to be asymptotically design unbiased irrespective of any model assumption. This raises the question: could we protect the model based estimator $\hat{P}_{h,m}$ from the model uncertainty? We follow up it by proposing a nonparametric predictive estimator of categorical proportion P_h which is obtained to (4) except that $g_h(.)$'s are now unknown and additive smooth functions of the form $g_h(x_i) = \sum_{\alpha=1}^{q} g_{h\alpha}(x_{i\alpha})$ and $g_{h\alpha}(.)$'s are estimated from data by using splines (Brumback et al. (1999)).

The model is thus given by

$$\pi_h(x_i) = \exp\{g_h(x_i)\}[1 + \sum_{u=1}^{p-1} \exp\{g_u(x_i)\}]^{-1}, h = 1,...,p-1 \tag{6}$$

subject to the usual constraint $\sum_{h=1}^{p} \pi_h(x) = 1$, for every x. Letting $\hat{g}_h(.)$ the estimate of $g_h(.)$ and writing

$$\hat{\pi}_h(x_i) = \exp\{\hat{g}_h(x_i)\}[1 + \sum_{u=1}^{p-1} \exp\{\hat{g}_u(x_i)\}]^{-1},$$

the estimate of P_h becomes

$$\hat{P}_{h,np} = N^{-1}[\sum_{i \in S} D_{ih} + \sum_{j \in \bar{S}} \hat{\pi}_h(x_j)]. \quad (7)$$

In Section 2, without loss of generality we obtain the expressions for asymptotic bias and asymptotic variance of $\hat{P}_{h,m}$ assuming $g_h(.) = g(.)$ for all h. The assumption is made to simplify the presentation avoiding unnecessary notational complexity. The estimator is found to be asymptotically model unbiased as well as model consistent. We also derive expression for the asymptotic variance of the model-based estimator (4) and its consistent estimator. Simultaneous confidence intervals for the population proportions based on asymptotic normality are proposed. We introduce the random coefficients splines model in Section 3. To obtain $\hat{g}_h(.)$ we adopt the likelihood approach. We discuss this approach in detail in Section 3. But finding maximum likelihood estimate (MLE) by direct maximization of the likelihood function is simply not practical in our setup. It involves too many integrals. In Section 4 we adopt and extend the EM methodology developed by Steele (1996) to our set up for finding MLE. Steele (1996), in fact, develops it for finding MLE in generalized linear mixed models (GLMM). This could be used directly in our set up if we had only binary responses, an application of which is considered by French and Wand (2004) in a different context. In Section 5, we extend our methods for multiple auxiliary variables. We present simulation studies in Section 6 to compare the performances of the three estimators given by (4)-(5) and (7). The results show marked improvements in some cases. Finally, in Section 7, we give the concluding remarks.

2. Properties of $\hat{P}_{h,m}$

In order to find asymptotic bias and variance of $\hat{P}_{h,m}$ we make the following assumptions. Our assumption $g_h(.) = g(.)$ for all h entails $a_h = r$ for all h; r is some positive integer.

A1. The parameter space Θ is a compact subset of $\Re^{r(p-1)}$, where $\Re = (-\infty, \infty)$.

A2. $N^{-1}\sum_{i=1}^{N} \| x_i \|^{\delta} = O(1)$, $\delta = 1,2$.

A3. Consider a sequence of finite populations of size N_ν and corresponding samples of sizes n_ν, indexed by ν. Assume that as $\nu \to \infty$, both n_ν and $N_\nu - n_\nu \to \infty$ such that sampling fraction $f_\nu = n_\nu / N_\nu \to \rho \in [0,1)$.

For simplicity, we drop the suffix ν in the rest of the paper. We now define

$$\pi'_{jh}(\beta) = \partial \pi_{jh}(\beta)/\partial\beta,\ \pi''_{jh}(\beta) = \partial^2 \pi_{jh}(\beta)/\partial\beta\partial\beta^T,\ \bar{\pi}'_h(\beta) = (N-n)^{-1}\sum_{j\in\bar{S}} \pi'_{jh}(\beta),$$

$$\bar{\pi}''_h(\beta) = (N-n)^{-1}\sum_{j\in\bar{S}} \pi''_{jh}(\beta). \qquad (8)$$

The assumptions A2 and A3 then imply that $\bar{\pi}'_h(\beta)$ and $\bar{\pi}''_h(\beta)$ are $O(1)$.

A4. The sample and nonsample design points have a common asymptotic distribution function G; that is,

$$n^{-1}\sum_{i\in S} I(x_i \le x) \to G(x),\ (N-n)^{-1}\sum_{j\in\bar{S}} I(x_j \le x) \to G(x) \text{ for all } x.$$

Assuming the maximum likelihood estimator $\hat{\beta}$ exists, the above assumptions (A2-A4) entail $\sqrt{n}$ consistency of $\hat{\beta}$. Now we state the following theorems. The proofs are given in the appendix.

*Theorem*1. Under assumptions (A2)-(A4), the bias of $\hat{P}_{h,m}$ is

$$E(\hat{P}_{h,m} - P_h) = O(n^{-1}),\ h = 1,\ldots, p.$$

Letting

$$S(\beta) = (S_1^T(\beta),\ldots, S_{p-1}^T(\beta))^T,\ I(\beta) = (I_{hh'}(\beta))_{h,h'=1,\ldots,p-1}$$

where

$$S_h(\beta) = \partial l(\beta)/\partial\beta_h = \sum_{i\in S} x_i (d_{ih} - \pi_{ih}(\beta)),\ h = 1,\ldots, p-1, \qquad (9)$$

$$I_{hh'}(\beta) = -\partial S_h(\beta)/\partial \beta_{h'} = \begin{cases} \sum_{i \in S} x_i x_i^T \pi_{ih}(\beta)(1-\pi_{ih}(\beta)), h = h', \\ -\sum_{i \in S} x_i x_i^T \pi_{ih}(\beta)\pi_{ih'}(\beta), h \neq h', \end{cases} \quad (10)$$

we write

$$\bar{I}(\beta) = \lim_{n\to\infty} n^{-1} I(\beta) = \lim_{n\to\infty} n^{-1} \sum_{i \in S} I_i(\beta) = \lim_{n\to\infty} n^{-1} \sum_{i \in S} G_1(x_i) = \int G_1(u) dG(u),$$

$$\bar{V}_{hh}(\beta) = \lim_{n\to\infty} (N-n)^{-1} \sum_{j \in \bar{S}} \pi_{jh}(\beta)(1-\pi_{jh}(\beta))$$

$$= \lim_{n\to\infty} (N-n)^{-1} \sum_{j \in \bar{S}} G_{2,hh}(x_j) = \int G_{2,hh}(u) dG(u),$$

$$\bar{V}_{hh'}(\beta) = \lim_{n\to\infty} (N-n)^{-1} \sum_{j \in \bar{S}} (-1)\pi_{jh}(\beta)\pi_{jh'}(\beta)$$

$$= \lim_{n\to\infty} (N-n)^{-1} \sum_{j \in \bar{S}} G_{2,hh'}(x_j) = \int G_{2,hh'}(u) dG(u),$$

$$\bar{D}_h(\beta) = \lim_{n\to\infty} \bar{\pi}'_h(\beta) = \lim_{n\to\infty} (N-n)^{-1} \sum_{j \in \bar{S}} \pi'_{jh}(\beta)$$

$$= \lim_{n\to\infty} (N-n)^{-1} \sum_{j \in \bar{S}} G_3(x_j) = \int G_3(u) dG(u),$$

where $I_i(\beta) = G_1(x_i)$, say, is the contribution of the i-th observation to $I(\beta)$. Similarly, we define $G_{2,hh}(.), G_{2,hh'}(.)$ and $G_3(.)$. Moreover we define

$\mathsf{v}_{hh}(\beta) = \mathrm{Var}(\hat{P}_{h,m} - P_h)$ and

$\mathsf{v}_{hh'}(\beta) = \mathrm{Cov}(\hat{P}_{h,m} - P_h, \hat{P}_{h',m} - P_{h'}), h \neq h'$.

Theorem 2. Under assumptions (A2)-(A4),

(a). the asymptotic variance of the prediction error for the h-th category is given by

$$\mathrm{Var}(\hat{P}_{h,m} - P_h) = n^{-1}(1-\rho)^2 \bar{D}_h(\beta)^T \bar{I}(\beta)^{-1} \bar{D}_h(\beta) + n^{-1}\rho(1-\rho)\bar{V}_{hh}(\beta) + o(n^{-1}),$$

$h = 1,..., p$, and

(b). the asymptotic covariance $v_{hh'}(\beta)$ is given by

$\mathrm{Cov}(\hat{P}_{h,m} - P_h, \hat{P}_{h',m} - P_{h'})$

$= n^{-1}(1-\rho)^2 \overline{D}_h(\beta)^T \bar{I}(\beta)^{-1} \overline{D}_{h'}(\beta) + n^{-1}\rho(1-\rho)\overline{V}_{hh'}(\beta) + o(n^{-1})$,

$h, h'(h \neq h') = 1,..., p-1$.

Theorem 3. Under assumptions (A1)-(A4), a consistent estimate of the asymptotic variance (asymptotic covariance) $\mathrm{Var}(\hat{P}_{h,m} - P_h)$ ($\mathrm{Cov}(\hat{P}_{h,m} - P_h, \hat{P}_{h',m} - P_{h'})$) in Theorem 2 is given by

$\mathsf{v}_{hh}(\hat{\beta}) = n^{-1}(1-f)^2 \overline{D}_h(\hat{\beta})^T \bar{I}(\hat{\beta})^{-1} \overline{D}_h(\hat{\beta}) + n^{-1} f(1-f)\overline{V}_{hh}(\hat{\beta})$,

($\mathsf{v}_{hh'}(\hat{\beta}) = n^{-1}(1-f)^2 \overline{D}_h(\hat{\beta})^T \bar{I}(\hat{\beta})^{-1} \overline{D}_{h'}(\hat{\beta}) + n^{-1} f(1-f)\overline{V}_{hh'}(\hat{\beta})$)

$h = 1,..., p$ ($h, h'(h \neq h') = 1,..., p-1$).

Further, let

$P = (P_1,..., P_{p-1})^T$, $\hat{P}_m = (\hat{P}_{1,m},..., \hat{P}_{p-1,m})^T$ and the asymptotic variance-covariance matrix $\mathsf{v}(\beta) = (\mathsf{v}_{hh'}(\beta))_{h,h'=1,...,p-1}$. Following theorem gives asymptotic normality of $\hat{P}_m - P$.

Theorem 4. Under assumption (A2)-(A5), the vector of finite population proportion estimators $\hat{P}_m$ satisfies

$\mathsf{v}(\beta)^{-1/2}(\hat{P}_m - P) \xrightarrow{d} N(0,1)$ for all $\beta \in \Theta$.

Corollary 1. Under the assumptions (A1)-(A4), from theorems 3 and 4, we have

$\mathsf{v}(\hat{\beta})^{-1/2}(\hat{P}_m - P) \xrightarrow{d} N(0,1)$.

The proof of the above corollary is obvious from theorems 3 and 4.

From the corollary1, we can find the simultaneous confidence intervals for the finite population proportion estimators $\hat{P}_{h,m}$ s'.

3. Random coefficients splines model

From a model-based perspective, design unbiasedness property of an estimator lacks appeal. This property holds over repeated sampling. The survey statistician has only one sample and one set of sampled data. The worry is how to protect against incorrect inference given this data. Why should then dividing by the sample inclusion probabilities protect one against model uncertainty in this special case? A natural alternative, first suggested by Kuo (1988), is to adopt a nonparametric model-based approach, that is, replace the parametric working model by a nonparametric working model linking $\pi_h(x)$ to x. In binary case, it is tantamount to replacing a parametric link function by a nonparametric link function. As noted in the literature (Chiou and Muller (1998), Carroll, Gijbels and Wand (1997), Weisberg and Welsh (1994)) parametric specification is quite inadequate in many data applications leading to the biased estimates of regression parameters and thus resulting in incorrect inference. We consider now a nonparametric formulation of the model introduced in (6).

We let

$$\pi_h(x) = \exp\{g_h(x)\}[1 + \sum_{h=1}^{p-1} \exp(g_h(x))]^{-1} \qquad (11)$$

where $g_h(x)$ is an unknown smooth function of x. In what follows we confine our discussion to a single covariate x. In Section 5, we discuss the extension of our approach to multiple covariates assuming an generalized additive model (GAM) for $g_h(x)$ (Hastie and Tibshirani (1986)).

In generalized linear model (GLM) set-up there are a number of mixed model representations of smoothing that can be used to subsume the $g_h(x)$ into GLMM (Chen and Ibrahim (2006), French and Wand (2004), Verbyla et al., (1999), Lin and Zhang (1999), Brumback et al., (1999), Brumback and Rice (1998), Wang (1998)). Here we consider a generalization of similar smoothing technique for the multinomial logit model.

The random coefficients splines model that we use for obtaining a smooth estimate of $g_h(x)$ is given by

$$g_h(x) = \beta_{h0} + \beta_{h1}x + ... + \beta_{hr_h}x^{r_h} + \sum_{k=1}^{K_h} b_{hk}(x - \kappa_{hk})_+^{r_h}, \quad (12)$$

where $b_h = (b_{h1},...,b_{hK_h})^T$ is $N(0, \sigma_h^2 I_{K_h})$, $(t)_+^a = t^a$ if $t > 0$ and 0 otherwise, r_h is the degree of spline and κ_{hk}'s are the knots. Typically, r_h is fixed and low, usually $r_h \leq 3$. We assume K_h to be fixed but sufficiently large (e.g. 25) to ensure the desired flexibility in the choice of knots. Note that the unknown variance component σ_h^2 controls the amount of smoothing; larger the value of σ_h^2 smoother is the function and $\sigma_h^2 = 0$ corresponds to the case of no smoothing.

4. Likelihood Estimation

We obtain an estimate of $g_h(x)$ using the likelihood estimate of the model parameters. For writing the likelihood function we need to introduce the following notations:

$$b_h = (b_{h1},...,b_{hK_h})^T, b = (b_1^T,...,b_{p-1}^T)^T, \beta_h = (\beta_{h0},...,\beta_{hr_h})^T, \beta = (\beta_1^T,...,\beta_{p-1}^T)^T,$$

$$\theta_h = \ln\sigma_h^2, \Sigma_h = \exp(\theta_h)I_{K_h}, \theta = (\theta_1,...,\theta_{p-1})^T, \Sigma = diag(\Sigma_1,...,\Sigma_{p-1}),$$

$$\nu = (\beta^T, \theta^T)^T = (\nu_1,...,\nu_R)^T, r = \sum_{h=1}^{p-1} r_h, R = (p-1)(r+2), K = \sum_{h=1}^{p-1} K_h.$$

Thus the likelihood function is given by

$$l(\nu) = (2\pi)^{-K_0/2}|\Sigma|^{-1/2}\int \exp\left(\sum_{i\in S}\sum_{h=1}^{p} d_{ih}\ln\pi_h(x_i;\beta,b) - \frac{1}{2}b^T\Sigma^{-1}b\right) db \quad (13)$$

where $\pi_h(x_i;\beta,b)$ is related to $g_h(x_i)$ by equation (11) and $g_h(x_i)$ is given by equation (12).

In most cases vis-à-vis our application the above integral is intractable. Thus we cannot find the maximum likelihood estimate using observed data likelihood. One option to overcome this problem could be to use the penalized quasi-likelihood that essentially replaces (13) by a first order Laplace approximation. Another approach to finding

maximum likelihood estimate of ν is to use EM algorithm (Dempster et al.(1977)) by making use of a second order Laplace approximation at E-step (Steele (1996)). EM algorithm is a general purpose algorithm for finding the mode of the likelihood or posterior density function. Viewing the random effects b as latent variables, the EM algorithm iterates between calculating the conditional expectation of the complete-data log likelihood given the observed data and maximizing this expected value as a function of ν. Dempster et al. (1977) have shown that EM algorithm will lead to maximum likelihood estimates based on the observed data likelihood given in equation (13). In fact, since the observed-data likelihood $l(\nu)$ is log concave, the EM algorithm will work quite well and will not get stuck in local modes.

4.1 *Implementation of EM Algorithm*

Let $\mathcal{D}_o = \{(D_i, x_i) : i \in S\}$ and $\mathcal{D}_c = \{\mathcal{D}_o, b\}$ represent the observed and complete data respectively if we consider b as missing. The kernel of the complete data log-likelihood is given by

$$l_c(\nu \mid \mathcal{D}_c) = \sum_{i \in S} \sum_{h=1}^{p} d_{ih} \ln \pi_h(x_i; \beta, b) - (1/2) \ln|\Sigma| - (1/2) b^T \Sigma^{-1} b \qquad (14)$$

Note that the observed data log-likelihood is $l_c(\nu \mid \mathcal{D}_c) = \ln l(\nu)$, where $l(\nu)$ is given by (13).

EM algorithm iterates between two steps viz., E-step and M-step. Start with an initial value of the parameter ν, say, $\nu^{(0)}$. At the $(t+1)$-th iteration:

E-step: Compute conditional expectation

$$Q(\nu; \nu^{(t)}) = E[l_{\mathbf{c}}(\nu \mid \mathcal{D}_{\mathbf{c}}) \mid \mathcal{D}_o; \nu^{(t)}] = \int l_{\mathbf{c}}(\nu \mid \mathcal{D}_{\mathbf{c}}) f(b \mid \mathcal{D}_o; \nu^{(t)}) db ,$$

where conditional density of b is given by

$$f(b \mid \mathcal{D}_o; \nu) = \exp\{l_{\mathbf{c}}(\nu \mid \mathcal{D}_{\mathbf{c}})\} / \int \exp\{l_{\mathbf{c}}(\nu \mid \mathcal{D}_{\mathbf{c}})\} db .$$

M-step: Maximize the conditional expectation $Q(\nu;\nu^{(t)})$ with respect to ν over the parameter space to obtain an updated estimate $\nu^{(t+1)}$.

Iterate between E-step and M-step until convergence.

Implementation of E step requires computation of $Q(\nu;\nu^{(t)})$ which involves a high dimensional integral. To overcome this problem some authors use Monte Carlo EM (MCEM) algorithm (Wei and Tanner (1990), Walker (1996), McCulloch (1997), Booth and Hobert (1999), Ibrahim et al.(2001)) which replace these integrals by its monte carlo approximation based on samples from $f(b \mid \boldsymbol{\mathcal{D}}_o, \nu^{(t)})$. However, the method is computationally intensive and more suited for smaller number of random effects and sample sizes. In our case, even for three categories and a single covariate the number of random effects may be 50 or more and the sample sizes may be prohibitively large. MCEM algorithm does not seem to be practical and an approximation to the integral seems to be in order.

Steele (1996) describes an EM algorithm that alternates between calculating $\hat{D}_\nu Q(\nu;\nu^{(t)})$ a second order Laplace approximation to $D_\nu Q(\nu;\nu^{(t)})$ and solving $\hat{D}_\nu Q(\nu;\nu^{(t)}) = 0$ where $D_\nu Q(\nu;\nu^{(t)})$ represents the derivative of $Q(\nu;\nu^{(t)})$ with respect to ν. Assuming that the differentiation under the integral sign is permissible, a standard assumption of the EM algorithm, we obtain

$$D_\nu Q(\nu;\nu^{(t)}) = \int D_\nu l_{\mathbf{c}}(\nu \mid \boldsymbol{\mathcal{D}}_{\mathbf{c}}) f(b \mid \boldsymbol{\mathcal{D}}_o; \nu^{(t)}) db \,. \tag{15}$$

Regularity conditions allow the fully exponential Laplace approximation (Tierny et al. (1989)) to the complete expected data score vector (15), but not the expected complete data log-likelihood (Steele (1996)). The Laplace approximation to $D_\nu Q(\nu;\nu^{(t)})$ is given by

$$\hat{D}_\nu Q(\nu;\nu^{(t)}) = \{D_\nu l_c(\nu \mid \boldsymbol{\mathcal{D}}_{\mathbf{c}}) + C(\nu \mid \boldsymbol{\mathcal{D}}_{\mathbf{c}})\}\big|_{b=\tilde{b}(\nu^{(t)})} \,, \tag{16}$$

where

$\tilde{b}(\nu) \equiv \arg\max_{b} l_c(\nu \mid D_c)$,

and the term $C(\nu \mid \mathfrak{D}_c) = (C_{\beta_1}(\nu \mid \mathfrak{D}_c),...,C_{\beta_{p-1}}(\nu \mid \mathfrak{D}_c), C_{\theta_1}(\nu \mid \mathfrak{D}_c),...,C_{\theta_{p-1}}(\nu \mid \mathfrak{D}_c))^T$ is an adjustment factor that allows $\hat{D}_\nu Q(\nu;\nu^{(t)})$ to differ from $D_\nu Q(\nu;\nu^{(t)})$ by $O(n^{-2})$. The exact expressions for C_{β_l} and $C_{\theta_l}, l = 1,.., p-1$ and the detailed derivation of $C(\nu \mid \mathfrak{D}_c)$ are given in the appendix.

Steele (1996) notes that a useful first-order approximation to $D_\nu Q(\nu;\nu^{(t)})$ is obtained by ignoring the last term in (16). Laplace EM algorithm (Steele (1996)) with this first order approximation leads to the estimating equations for β that are identical to PQL algorithm (Breslow and Clayton (1993)). The two algorithms essentially differ in the manner in which θ is estimated. However, when the estimates of θ are similar, the Laplace EM algorithm should yield more accurate estimates of the fixed effects since they are based on a second-order approximation rather than a first-order approximation (Steele (1996)). The details of the M-step are given in the appendix.

Given ν, Laplace approximation entails

$$E(\pi_h(x;\beta,b) \mid \mathfrak{D}_O) = \pi_h(x;\beta,\tilde{b}(\nu)) + O_P(n^{-1}) \tag{17}$$

and hence the nonparametric estimator of $\pi_h(x)$ is $\pi_h(x;\beta,\tilde{b}(\nu))$. Thus the model-based predictive estimator of finite population proportion is given by (cf. (7))

$$\hat{P}_{h,np} = N^{-1}[\sum_{i\in S} D_{ih} + \sum_{j\in \bar{S}} \pi_h(x_j;\hat{\beta},\hat{b})], \tag{18}$$

where $\hat{b} = \tilde{b}(\hat{\nu})$.

5. Splines model for multiple auxiliary variables

In Section 4, we consider random coefficient splines model for a single auxiliary variable. An advantage of using this model is that, its extension to multiple auxiliary variables is conceptually quite straight forward. For q auxiliary variables $x_1,..,x_q$, we

consider a generalized additive model (GAM) (Hastie and Tibshirani (1990)) with $g_h(x_i)$ in (2) being replaced by $\sum_{\alpha=1}^{q} g_{h\alpha}(x_{i\alpha})$, where $g_{h\alpha}(.)'s$ are unknown smooth functions. For every choice of h and α ($h = 1,..., p-1$ and $\alpha = 1,..., q$), we use a random coefficient splines model for $g_{h\alpha}(x)$. The model is thus given by

$$g_{h\alpha}(x_{i\alpha}) = \beta_{h0} + \sum_{\alpha=1}^{q} \beta_{h\alpha}^T X_{ih\alpha} + \sum_{\alpha=1}^{q} b_{h\alpha}^T Z_{ih\alpha}, \tag{19}$$

where $\beta_{h\alpha} = (\beta_{h\alpha 1},..., \beta_{h\alpha r_{h\alpha}})^T$ is a fixed vector of regression coefficients, $X_{ih\alpha} = (x_{i\alpha},..., x_{i\alpha}^{r_{h\alpha}})^T$, $b_{h\alpha} = (b_{h\alpha 1},..., b_{h\alpha K_{h\alpha}})^T$ a random vector of spline coefficients and $Z_{ih\alpha} = ((x_i - \kappa_{\alpha 1})_+^{r_{h\alpha}},..., (x_i - \kappa_{\alpha K_{h\alpha}})_+^{r_{h\alpha}})^T$. Also we assume that $b_{h\alpha i}, i = 1,.., K_{h\alpha}$ are independently and normally distributed random variables with mean 0 and unknown variance $\sigma_{h\alpha}^2 (> 0)$ and as before we note that larger the values of $\sigma_{h\alpha}$'s smoother the functions $g_{h\alpha}(.)'s$ are. We now introduce the following notations:

$$b_h = (b_{h1}^T,..., b_{hq}^T)^T, b = (b_1^T,..., b_{p-1}^T)^T, \beta_h = (\beta_{h1}^T,..., \beta_{hq}^T)^T, \beta = (\beta_1^T,..., \beta_{p-1}^T)^T,$$

$$\theta_{h\alpha} = \ln \sigma_{h\alpha}^2, \Sigma_{h\alpha} = \exp(\theta_{h\alpha}) I_{K_{h\alpha}}, \theta_h = (\theta_{h1},..., \theta_{hq})^T, \theta = (\theta_1^{\mathrm{T}},..., \theta_{p-1}^T)^T,$$

$$\Sigma_h = diag(\Sigma_{h1},..., \Sigma_{hq}), \Sigma = diag(\Sigma_1,..., \Sigma_{p-1}), \nu = (\beta^T, \theta^T)^T = (\nu_1,..., \nu_R)^T,$$

$$r_h = \sum_{\alpha=1}^{q} r_{h\alpha}, r = \sum_{h=1}^{p-1} r_h, R = (p-1)(r+2), K_h = \sum_{\alpha=1}^{q} K_{h\alpha}, K = \sum_{h=1}^{p-1} K_h.$$

With the above notations the likelihood becomes exactly equal to (13). Then to find the likelihood estimates of the parameters we apply EM algorithm. The E and M steps are exactly similar to that of the single auxiliary variable case given above but the notations would become more involved. Finally our estimate of the finite population proportions would be given by (18).

6. Simulation study

Here we report the results of two limited simulation experiments. First, we consider a simple logit model to generate the data and compare the performance of the model-based estimator (4) relative to the design-based estimator (5). Also in the same set-up we study the performance of the estimator of the variance of (4) given by Theorem 3.

The steps of the simulation study are as follows:

1. We draw x_i, i = 1,…,1000 randomly from uniform (0, 1). For each x_i, we find the trinomial probabilities $\pi_h(x_i), h = 1,2$ using a simple logit model with link functions $g_1(x) = 0.031 + 0.13x$ and $g_2(x) = 0.012 + 0.043x$. Given these probabilities a trinomial trial is carried out to generate the response (d_{i1}, d_{i2}, d_{i3}), i = 1,…,1000. These responses along with the corresponding x_i values constitute the set of observations of a finite population of size N=1000.

2. A simple random sample of size $n = 100$ is drawn from this finite population without replacement

3. We then compute $\hat{P}_{h,d}$, $\hat{P}_{h,m}$ and v_h, $h=1,...,3$ on the basis of the sample observations.

4. We repeat step 1 to generate an independent set of finite population observations of size 1000 and then compute P_h on the basis of it. We then compute $\hat{P}_{h,m}$ based on a sample of size 100 generated from it. To make it distinct from $\hat{P}_{h,m}$ generated at step 3, we refer to it as $\hat{P}^*_{h,m}$.

The steps 1-4 are repeated B=1000 times and let P_h^b, $\hat{P}_{h,d}$, $\hat{P}^b_{h,m}$, $\hat{P}^{*b}_{h,m}$ and v_h^b be the values of P_h, $\hat{P}_{h,d}$, $\hat{P}_{h,m}$, $\hat{P}^*_{h,m}$ and v_h obtained at the b-th (b=1,…,1000) repetition.

The performances of the estimators $\hat{P}_{h,d}$ and $\hat{P}_{h,m}$ are compared by computing their relative biases (RB), relative root mean squares (RRMSE) and finally finding the efficiency of one relative to the other. We define,

$$\bar{\hat{P}}_{h,d} = B^{-1}\sum_{b=1}^{B}\hat{P}_{h,d}^{b}\,,\ \bar{\hat{P}}_{h,m} = B^{-1}\sum_{b=1}^{B}\hat{P}_{h,m}^{b}\ \ \bar{P}_h = B^{-1}\sum_{b=1}^{B}P_h^b\,,$$

$$RMSE(\hat{P}_{h,d}) = [B^{-1}\sum_{b=1}^{B}(\hat{P}_{h,d}^{b} - \bar{P}_h)^2]^{1/2}, RMSE(\hat{P}_{h,m}) = [B^{-1}\sum_{b=1}^{B}(\hat{P}_{h,m}^{b} - \bar{P}_h)^2]^{1/2}.$$

Then the relative biases and relative root mean squares of $\hat{P}_{h,d}$ and $\hat{P}_{h,m}$ are obtained as

$$RB(\hat{P}_{h,d}) = \frac{\bar{\hat{P}}_{h,d}}{\bar{P}_h},\ RB(\hat{P}_{h,m}) = \frac{\bar{\hat{P}}_{h,m}}{\bar{P}_h},\ RRMSE(\hat{P}_{h,d}) = \frac{RMSE(\hat{P}_{h,d})}{\bar{P}_h}$$ and

$RRMSE(\hat{P}_{h,m}) = \frac{RMSE(\hat{P}_{h,m})}{\bar{P}_h}$. Finally, the efficiency of $\hat{P}_{h,m}$ relative to $\hat{P}_{h,d}$ is given by

$E(\hat{P}_{h,m}, \hat{P}_{h,d}) = (\frac{RRMSE(\hat{P}_{h,d})}{RRMSE(\hat{P}_{h,m})})^2$. For studying the performance of the variance estimator, we compute the following entities:

$$\bar{v}_h = B^{-1}\sum_{b=1}^{B}v_h^b\,,\ V_h = B^{-1}\sum_{b=1}^{B}(\hat{P}_{h,m}^{*b} - B^{-1}\sum_{b=1}^{B}P_h^b)^2\,, Var(v_h) = B^{-1}\sum_{b=1}^{B}(v_h^b - V_h)^2$$

Then we compute the relative bias (RB) and instability (INST), which are defined as RB $= (\bar{v}_h - V_h)/V_h$ and INST$= Var(v_h)/V_h$. Table 1 reports the results of the above simulation experiment.

Table 1

Category	RB			RRMSE		RE
	$\hat{P}_{h,d}$	$\hat{P}_{h,m}$	v_h	$\hat{P}_{h,d}$	$\hat{P}_{h,m}$	$E(\hat{P}_{h,m}, \hat{P}_{h,d})$
1	-0.00888	-0.00645	0.43400 (0.44077)*	0.13601	0.10745	1.6022
2	0.00452	0.00309	0.48071 (0.48931)*	0.14401	0.11356	1.6082
3	0.00510	-0.00391	0.34960 (0.36140)*	0.15083	0.11875	1.6133

* represents the figures for INST

In table 1, the relative bias of the analytical estimator v_h is found to be considerable and also the estimator always leads to an overestimate. This seems to be natural implications of the following facts: (*i*) the true variance itself is extremely small, (*ii*) the analytical expression is valid up to $O(n^{-1})$ and (*iii*) n is only 100. In fact, we check by running a few simulation experiments with larger population and sample sizes that the estimate becomes reasonable even with a sample size of 1000. Thus, for small sample sizes the analytical estimator leads to substantial overestimate.

The choices of the functions along with the x-values make the category probabilities varying between 0.25 to 0.4; ensuring sufficient no of observations for each category in the finite population. In terms of both relative bias and efficiency, the model based estimator clearly dominates over the usual design-based estimator.

Now we carry out a simulation experiment for studying the performance of nonparametric estimator (18) compared to the model-based and design-based estimators given by (4) and (5). Here we consider a set-up exactly similar to the above except that the probabilities $\pi_h(x_i), h = 1,2$ are linked to non-linear logit models with different choices of the smoothed functions $g_1(.)$ and $g_2(.)$ (cf. Breidt and Opsomer (2000) and Breidt et al. (2005)).

For, $x \in [0,1]$ different choices of $g_h(.)$'s are:

Linear (L): $m_{11}(x) = 0.475 + 0.05x$, $m_{12}(x) = 0.525 - 0.05x$, $m_{13}(x) = 1 + 2(x - 0.5)$

Quadratic (Q): $m_2(x) = 1 + 2(x - 0.5)^2$

Bump (B): $m_{31}(x) = 2(x - 0.5) + \exp(-200(x - 0.5)^2)$,

$m_{32}(x) = 2(x - 0.5) - \exp(-200(x - 0.5)^2)$

Jump (J): $m_4(x) = (0.35 + 2(x - 0.5))I_{\{x \le 0.65\}}$

Exponential (E) : $m_5(x) = \exp(-8x)$

Cycle (C): $m_6(x) = 2 + \mathrm{Sin}(2\pi x)$.

These choices, in a limited way, allow us to evaluate and compare the performance of the nonparametric estimator relative to the others. We consider five combinations of $g_1(.)$ and $g_2(.)$ in our simulation study, viz.,

Linear-Linear (L-L): m_{11} and m_{12}

Linear-Quadratic (L-Q): m_{13} and m_2

Bump-Exponential (B-E): m_{31} and m_5

Jump-Cycle (J-C): m_4 and m_6

Bump-Bump (B-B): m_{31} and m_{32}.

We report the results of our simulation study in Table 2.

Table 2

Category	Models	$E_{\hat{P}_{h,m},\hat{P}_{h,d}}$	$E_{\hat{P}_{h,np},\hat{P}_{h,d}}$
1	L-L	1.6075	1.5321
	L-Q	0.9740	1.2834
	B-E	1.0558	1.2043
	J-C	1.0159	1.1759
	B-B	1.0211	1.1849
2	L-L	1.6233	1.4375
	L-Q	0.9468	1.3833
	B-E	0.9580	1.5224
	J-C	1.0536	1.1364
	B-B	0.9921	1.2216
3	L-L	1.5884	1.6516
	L-Q	1.0290	1.3586
	B-E	1.0167	1.4473
	J-C	1.0237	1.1486
	B-B	1.0576	1.4095

Table 2 clearly shows that $\hat{P}_{h,np}$ always dominates $\hat{P}_{h,d}$ for all choices of the link functions. However, as expected when at least one of $g_1(.)$ and $g_2(.)$ is nonlinear, the estimator $\hat{P}_{h,np}$ is more efficient than $\hat{P}_{h,m}$ uniformly over all categories. For some of the categories the gain is substantial.

7. Concluding remarks

In this paper we use a predictive approach to improve upon the standard estimates of finite population proportions based on both parametric and nonparametric models incorporating the population information on the auxiliary variables. The question that may arise at this stage is: given a survey data set, how should one decide whether to use multinomial logit model or nonparametric model for modeling the category probabilities? In a recent work Goeman and Le Cessie (2006) propose a score test for testing the goodness of fit of multinomial logit model against the alternatives that nonlinearities or interaction effects may be present. This seems to be appropriate in our set-up for deciding which model should be used.

In survey literature this is possibly the pioneering attempt to estimate the population proportions using multinomial logit model following a predictive approach. More importantly, we are able to generalize this approach to nonparametric models by using the recently introduced random coefficients splines models. We also implement EM algorithm for finding likelihood estimates using second order Laplace approximation. Finally, we are able to come up with an asymptotic formula for its variance and then propose a re-sampling based estimate of it.

References

Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed models likelihoods with automated monte carlo EM algorithm, Journal of the Royal Statistical Association, Ser. B, 61, 265-285.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. Journal of American Statistical Association, 88, 9-25.

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for analysis of nested and crossed samples of curves. Journal of the American Statistical Association, 93, 961-1006.

Brumback, B.A., Ruppert, D. and Wand, M.P. (1999). Comment on Silvey, Kohn and Wood. JASA, 94, 794-797.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997) Generalized partially linear single-index models. Journal of American Statistical Association,92,477-489.

Chiou, J.-M. and Muller, H.-G. (1998) Quasi-likelihood regression with unknown link and variance function. Journal of American Statistical Association,93,1376-1387.

Chen, Q. and Ibahim, J.G. (2006). Semiparametric models for missing covariate and response data in regression models. Biometrics,62.

Cox, D.R. and Snell, E.J. (1968) A general definition of residuals, Journal of Royal Statistical Society, Ser. B, 30, 248-275.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, Journal of Royal Statistical Society, Ser. B, 39, 1-38.

Eilers, P.H.C. and Marx, B.D.(1996). Flexible smoothing with B-splines and penalties (with discussion). Statistical Science,11,89-121.

Firth, D. and Bennett, K.E. (1998) Robust models in probability sampling. Journal of the Royal Statistical Society, Series B, 60, 3-21.

French, J.L. and Wand, M.P. (2004). Generalized additive models for cancer mapping with incomplete covariates. Biostatistics,5,177-191.

Geoman, J.J. and Le Cessie S. (2006). A goodness-of-fit test for multinomial logistic regression. Biometrics (DOI: 10.1111/j.1541-0420.2006.00581.x)

Ibrahim, J.G., Lipsuitz, S.R. and Chen., M.-H. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. Biometrika, 88, 551-564.

Kuo, L.(1988). Classical and prediction approaches to estimating distribution functions From survey data. Proceedings of the Section on Survey Research Methods, American Statistical association, pp.280-285.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. Journal of the Royal Statistical society, Ser. B, 61, 381-400.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association, 92, 162-170.

Royall, R.M. (1970) On finite population sampling theory under certain linear regression model. Biometrika, 57, 377-387.

Ruppert,D. (2002) Selecting the number of knots for penalized spline.Journal of Computational and Graphical Statistics,11,735-757.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric regression. Cambridge University Press.

Sarndal, C.E. and Wright, R.L. (1984) Cosmetic form of estimators in survey sampling. Scandinavian Journal of Statistics, 11, 146-156.

Tierney, L., Kass, R. E. and Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpoisitive functions . Journal of the American Statistical association, 84, 710-716.

Walker, S. (1996). An EM algorithm for nonlinear random effects models. Biometrics, 52, 934-944.

Wand, M.P. (2003) Smoothing and mixed models. Computational Statistics,18,223-249.

Wang, Y. (1998). Smoothing spline models with correlated random errors. Journal of the American Statistical Association, 93, 341-348.

Weisberg, S. and Welsh, A.H. (1994) Adapting for the missing link. Annals of Statistics, 22,1674-1700.

Wei, G.C.G. and Tanner, M.A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical association, 85, 699-704.