Discussion Papers No. 274, June 2000
Statistics Norway, Division for Statistical Methods and Standards

Ib Thomsen, Li-Chun Zhang and Joseph Sexton

Markov Chain Generated Profile Likelihood Inference under Generalized Proportional to Size Non-ignorable Non-response

Abstract:
We apply two non-ignorable non-response models to the data of the Norwegian Labour Force Survey, the Fertility Survey and the Alveolar Bone Loss Survey. Both models focus on the marginal effect which the object variable of interest has on the non-response, where we assume the probability of non-response to be generalized proportional to the size of the object variable. We draw the inference of the parameter of interest based on the first-order theory of the profile likelihood. We adapt the Markov chain sampling techniques to efficiently generate the profile likelihood inference. We explain and demonstrate why the resampling approach is more flexible for the likelihood inference than under the Beyesian framework.

Keywords: Non-ignorable non-response, profile likelihood, Markov chain sampling.

Address: Li-Chun Zang, Statistics Norway, Division for Statistical Methods and Standards.
E-mail: li.chun zang @ssb.no

Discussion Papers	comprise research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborate than a standard journal article by including intermediate calculation and background material etc.

Abstracts with downloadable PDF files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
N-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

Missing data is practically unavoidable in sample surveys, clinical trials, and various social studies. Statistical analysis of data subject to non-response has received increasing attention. Models of the non-response mechanism are often classified as ignorable or non-ignorable (Rubin, 1976). Opinions differ as to the appropriateness of the one type over the other — see Scharfstein, Rotnitzky, and Robins (1999) and the accompanying discussions. In a simple missing-not-completely-at-random setting, we have one *object* variable of interest, one *auxiliary* variable, and one *non-response indicator*. While the auxiliary variable is all known, the object variable will be missing from the non-respondents. If we treat the non-response as independent of the object variable conditional to the auxiliary one, the model is said to be ignorable. Whereas it is non-ignorable if, for instance, we regard the non-response as independent of the auxiliary variable conditional to the object one. In either case, the non-response depends on the value, i.e. *size*, of the variable to be conditioned on, and its probability may considered to be generalized proportional to the size. More generally, of course, the size could also be multiple.

Such *generalized proportional to size (GPS)* ignorable non-response models are common in sample surveys, e.g. through the use of post-stratification and calibration (Thomsen and Holmøy, 1998; Lundström and Särndal, 1999). Whereas a recent example of a GPS non-ignorable model can be found in Troxel, Harrington, and Lipsitz (1998). In this paper we apply two special instances of the GPS non-ignorable models to the data of the Norwegian Labour Force Survey (LFS), the Fertility Survey and the Alveolar Bone Loss Survey (ABLS). In the first two cases, we assume a Bernoulli distribution of the non-response indicator from each sampled unit. In the ABLS data, multiple measurements are taken from each unit, to which we shall apply a Truncated Poisson distribution to the total number of non-response. In all the three cases, the non-response is linked to the object variables there, making the models non-ignorable. As such the GPS models focus on the marginal effect of the object variable on the non-response. While this is necessary if only from a sensitivity analysis point of view, over-dispersion may occur if the variation in the non-response pattern can not be sufficiently explained through the object variable alone. We shall give some discussions based on our data, although we know little about its consequences in general.

Interpretational aspects aside, the non-ignorable non-response model frequently brings out the two basic numerical problems in statistics, namely integration and optimization. The EM algorithm (Dempster, Laird, and Rubin, 1977) gets around the first one if the missing data can be integrated out of the complete latent log-likelihood in a closed form. Otherwise, Monte Carlo methods can be applied in combination (Ibrahim, Lipsitz, and Chen, 1999; Ibrahim, Chen, and Lipsitz, 1999), though it may not always be numerically efficient. We found that the Laplace approximation (Tierney and Kadane, 1986) worked well for the ABLS data. Worse is the situation, however, when the non-ignorable non-response model leads to sensitive point estimation (Smith, Skinner, and Clarke, 1999). We believe that this is by no means uncommon with this type of model, and shall offer a heuristic explanation when we present our models in more details.

Like Smith, Skinner, and Clarke (1999), we draw inference of the parameter of interest based on the first-order theory of the profile likelihood (Barndorff-Nielsen and Cox, 1994). We do not employ any of the various higher order corrections (Stern, 1997; Severini, 1998; Davison and Stafford, 1998). We concentrate on adapting the Markov chain sampling techniques (Brooks, 1998;

Robert and Casella, 1999) to efficiently generate the profile likelihood inference. Our approach has been anticipated by Geyer (1996), who rightly pointed out that the explosive development in the MCMC methods should also liberate the likelihood-based inference. For the Bayesian inference it is required that the marginal distribution of the Markov chain converges to the posterior, in order to perform the various Monte Carlo calculations. For the likelihood inference, as will be explained and demonstrated below, we only need the chain to visit the high-likelihood parameter subspace with reasonable frequency. Otherwise we do not even need to know if, or where, the chain converges. This gives us extra flexibility which the Bayesian approach does not enjoy.

The rest of the paper will be organized as follows. In Section 2 we present our particular GPS non-response models, indicating possible problems of over-dispersion and model identification. We outline the basic approach of Markov chain sampling generated profile likelihood inference in Section 3, and explain how we deal with the two numerical problems of integration and optimization. In Section 4 to 6 we apply the methods to the data of, respectively, the Norwegian LFS, the Fertility Survey and the ABLS. Finally, Section 7 contains a short summary.

2 Generalized proportional to size non-response

Suppose that non-response is suspected to be influenced by the size (i.e. value) of a univariate variable, denoted by X where $x \geq 0$. We could, as we do in this paper, model its marginal effect on non-response through a generalized proportional to size (GPS) predictor, i.e.

$$\eta = \alpha(x+1)^{\beta} \qquad \text{where} \quad \alpha > 0. \tag{1}$$

Let $R_i = 1$ denote non-response from unit i, and $R_i = 0$ otherwise. We may, for instance, put a Bernoulli distribution with parameter p_i on $R_i|x_i$, i.e. conditional to $X_i = x_i$, where

$$p_i = \eta_i/(1+\eta_i) \qquad \Leftrightarrow \qquad \log p_i - \log(1-p_i) = \log\alpha + \beta\log(x_i+1), \tag{2}$$

and parameter β takes positive value if non-response is more severer with larger x_i; it is negative the other way around. In particular, the GPS mechanism above is a logistic regression of R_i on $\log(x_i+1)$. The transformation of X_i is of course optional, which could be appropriate if X has many levels or is continuous. Troxel, Harrington, and Lipsitz (1998), however, applied the logistic model directly to X. In any case, we have, for p_i in (2),

$$\partial p_i/\partial\alpha = p_i(1-p_i)/\alpha \qquad \text{and} \qquad \partial p_i/\partial\beta = p_i(1-p_i)\log(x_i+1).$$

Sometimes, several measurements are taken from the same unit. Let R_{ij} be the non-response indicator for the jth measurement from the ith unit, where $j = 1, ..., m$. It may no longer be reasonable to consider $R_{ij}|x_{ij}$ as independent. Let $R_i = \sum_j R_{ij}$, i.e. the total number of non-response from unit i. Let $x_i = h(x_{ij})$ be a scalar function of $(x_{i1}, ..., x_{im})$, such as the mean $x_i = \bar{x}_{ij} = \sum_j x_{ij}/m$. We may, for instance, put a Truncated Poisson distribution on $R_i|x_i$, i.e.

$$R_i|x_i \;\sim\; TrnPoisson(\lambda_i) \qquad \text{where} \quad 0 \leq r_i \leq m \quad \text{and} \quad \lambda_i = mp_i = m\eta_i/(1+\eta_i). \tag{3}$$

Equation (1) to (3) are all instances of GPS non-response, if the term "generalized" is taken in a wide sense. They are ignorable (Rubin, 1976) if the x_i's are known; whereas they are said to be non-ignorable if x_i is missing from the non-respondents, such as when X is the object variable of interest. In the latter case the GPS mechanism models the marginal effect of X on non-response. Over-dispersion may occur if the observed variation in the non-response pattern is larger than what can be explained through X. This is similar to the case of complete-data logistic regression, when studying the marginal effect of the covariates on some binary outcome (Cox and Snell, 1989, Chapter 3). In our applications, the matter seems to have caused little problem to the inference of interest , although we can not say much about its consequences in general.

One problem which arises when dealing with non-ignorable non-response models concerns the uncertainty of the models themselves, since they are "unexaminable in a fundamental sense" (Molenberghs, Kenward, and Lesaffre, 1997). Other times we may face unstable or sensitive point estimators. Smith, Skinner, and Clarke (1999) gave the matter geometric presentations. Heuristically, under a non-ignorable model, the estimation of the parameter of interest may be considered as bias-correction of the estimator under an ignorable non-response model. However, this adjustment can be highly uncertain or sensitive because (i) the underlying outcomes of X are more or less concentrated on a short interval, (ii) different parameter values of a non-ignorable model, or even different models, may appear rather similar over the highly densed region of (X, R), (iii) to choose or identify between them, i.e. to determine the actual bias-correction, we need to heavily rely on data from the low density area of (X, R), which are subject to the largest sampling variability. The GPS non-ignorable models above are no exceptions here, in which respect we shall focus on the consequences these models have on the inference of interest, instead of the goodness-of-fit of the models *per se*.

3 Resampling generated profile likelihood inference

Generically, let $X = (X_{obs}, X_{mis})$ be a sample of the variable of interest, where X_{obs} is its observed part and X_{mis} the part which is missing due to non-response. Let $f(x;\xi)$ be the model function of X. Let R denote non-response, and $f(r;\gamma|x)$ its model function conditional to $X = x$. The complete, latent likelihood, denoted by L_1, has the following factorization w.r.t. $\theta = (\xi, \gamma)$, i.e.

$$L_1(\theta; x, r) \quad \propto \quad f(x, r; \theta) \quad = \quad f(x_{obs}; \xi) f(x_{mis}; \xi | x_{obs}) f(r; \gamma | x_{obs}, x_{mis}). \tag{4}$$

The (observed) likelihood, denoted by L, is obtained from integrating out x_{mis} in L_1, i.e.

$$L(\theta; x_{obs}, r) \propto f(x_{obs}; \xi) \int f(x_{mis}, r; \xi, \gamma | x_{obs}) \, dx_{mis} = f(x_{obs}; \xi) H(\xi, \gamma; r | x_{obs}). \tag{5}$$

The likelihood L no longer factorizes w.r.t. (ξ, γ); neither may $H(\xi, \gamma)$ be available in a closed form. To get around the integral H, the E-step of the EM algorithm (Dempster, Laird, and Rubin, 1977) calculates, at the present parameter estimate $\tilde{\theta}$ and for $l_1 = \log L_1$,

$$E[l_1(\theta; x_{obs}, X_{mis}, r); \tilde{\theta}] = Q(\theta; \tilde{\theta} | x_{obs}, r).$$

Maximizing $Q(\theta)$ is usually considerably easier than that of $l = \log L$, provided it is available in closed form. Otherwise, we may apply the EM algorithm in combination with the Monte Carlo method (Ibrahim, Lipsitz, and Chen, 1999; Ibrahim, Chen, and Lipsitz, 1999), possibly at the some expense of simplicity. A worse situation, however, occurs when the non-ignorable non-response model leads to sensitive point estimation as noted earlier. We encountered such a case in the Alveolar Bone Loss data to which we shall return.

We have thus outlined two possible numerical difficulties of (a) integration, i.e. the calculation of the integral H in (5) and possibly Q, and (b) optimization, i.e. maximizing Q. In response we shall apply Markov chain sampling generated profile likelihood inference. However, unless the model and data have been worked out previously, the process of Markov chain resampling typically involves a number of trial-and-errors. Efficient evaluation of the likelihood L is therefore crucial. In case the integral H in (5) is not available in closed form, the various Monte Carlo methods can be used (Geweke, 1989). In what Geyer (1996) called the "many-samples method", each evaluation of L requires a new sample, which could be very inefficient in many cases. Geyer (1996) suggested a "single-sample method" through importance sampling which, in the present context, would give the following approximation to (5):

$$L(\theta; x_{obs}, r) \propto f(x_{obs}; \xi)\{\frac{1}{m}\sum_{i=1}^{m} f(r; \gamma | x_{obs}, x_{mis}^{(i)}) \frac{f(x_{mis}^{(i)}; \xi | x_{obs})}{\pi(x_{mis}^{(i)})}\} \quad \text{where} \quad x_{mis}^{(i)} \sim \pi(\cdot),$$

where π does not depend on θ such as when $\pi = f(x_{mis}; \hat{\xi} | x_{obs})$. The $(x_{mis}^{(1)}, ..., x_{mis}^{(m)})$, once generated, are held fixed for all values of θ to be evaluated. Hence the term "single-sample". The summation involved may nevertheless be time-consuming unless all the terms can be calculated parallelly. On the other hand, the Laplace approximation (Tierney and Kadane, 1986) may prove to be helpful, provided the mode of $f(x_{mis}; \xi | x_{obs})$ in x_{mis} can easily be derived. We shall apply the Laplace approximation to the Alveolar Bone Loss data.

In any case, suppose now that we no longer have problem (a). To explore the likelihood, we basically need to know, for which parameter values should we calculate L? Markov chain sampling can be adapted here to our advantage. Let the target distribution be $\pi(x_{mis}, \theta; x_{obs}, r) \propto L_1$. Suppose we manage to generate a sample of (x_{mis}, ξ, γ) by means of Markov chain sampling, whose marginal distribution follows π. Simply omitting x_{mis}, we obtain by (4) and (5) the remaining sample of (ξ, γ), with a marginal distribution proportional to L. That is,

$$(x_{mis}, \xi, \gamma) \sim \pi(x_{mis}, \xi, \gamma; x_{obs}, r) \propto L_1 \quad \Rightarrow \quad (\xi, \gamma) \sim \int \pi \, dx_{mis} \propto L.$$

Since these (ξ, γ) cover the high likelihood region of θ reasonably well, they form a basis for our exploration of L. In particular, two relaxations make this approach flexible: (i) the target distribution π needs not to be exactly proportional to L_1 — its job is to generate high likelihood parameter values with reasonable frequency, and (ii) neither do we need to bother much about if, or where, the Markov chain converges — its performance is judged from the $L(\theta)$ it generates.

Suppose we have a sample of θ, regardless how it has come about. Let $\psi = h(\theta)$ be a scalar function of the parameter θ of interest. In particular, ψ may simply be a component of θ. We may plot $L(\theta)$ against ψ: the *contour* gives an approximation to the profile likelihood of ψ, denoted

by $L_{Pro}(\psi)$, and the mode an approximation to $\hat{\psi}$, i.e. the maximum likelihood estimate (mle) of ψ. How quickly the contour takes shape as the resample grows larger, provides an indication of how well the resampling is working for our purpose. Neither the convergence nor autocorrelation of the Markov chain matter otherwise. We derive an approximation to the confidence interval of ψ using the first-order χ^2_1-approximation to $2L_{Pro}(\hat{\psi}) - 2L_{Pro}(\psi)$ (Barndorff-Nielsen and Cox, 1994, Chapter 3). Such simple approximations are consistent, provided the probability for the resampling scheme to visit any particular point-mass in the parameter space tends to unity as the resample grows to infinity. However, a better approximation can be obtained: we simply hold ψ at the value to be evaluated, and resample the rest of the parameters under the constraint. Often no change to the sampling scheme is required otherwise. The highest resampled $L(\theta)$ can be called the *Monte Carlo (MC) profile likelihood* of ψ. This is a "many-samples method" since each value of ψ requires a separate sample. A Rao-Blackwellization-like "single-sample" short-cut is also available. Simply dropping ψ from the resampled θ, the rest of the parameters has a marginal distribution proportional to $\int L d\psi$. Recalculating $L(\theta)$ over the same resample, with ψ held fixed at the value to be evaluated, the highest $L(\theta)$ can be called an *R-B MC profile likelihood* of ψ. Having obtained either the MC or R-B MC profile likelihood over a grid of values of ψ, we may *calibrate* both $\hat{\psi}$ and the end points of its confidence interval.

4 The Norwegian LFS

The Norwegian LFS is a quarterly national survey comprising about 24000 people (Table 1). Let the LFS-Employment Status be the object variable, which is only available from the respondents. The auxiliary variable, i.e. the Register-Employment Status, is independently constructed from administrative registers, and is known throughout the population. The post-stratified estimator is unbiased under the ignorable model, short-handed obviously as "Non-response $\perp$ Object | Auxiliary". Its bias under the non-ignorable model, i.e. "Non-response $\perp$ Auxiliary | Object", has a particular form and can be estimated based on the respondents directly (Zhang, 1999).

	LFS-Employment	LFS-Unemployment	Non-response
Register-Employment	12881	1158	518
Register-Unemployment	1829	6726	796

Table 1: The data of the Norwegian Labour Force Survey in the 1st quarter of 1995.

The object variable being binary, the GPS formulation (1) and (2) is rather trivial since, let $\gamma_x = P[R = 1|X = x]$ where $x = 1$ for LFS-Employment and $x = 0$ otherwise, we have

$$\alpha = \gamma_0/(1-\gamma_0) \qquad \text{and} \qquad \beta \log 2 = \log\{\gamma_1/(1-\gamma_1)\} - \log\{\gamma_0/(1-\gamma_0)\}.$$

Let $\xi_y = P[X = 1|Y = y]$ where $y = 1$ for Register-Employment and $y = 0$ otherwise. The complete, latent data and its distribution are given as in Table 2. In particular, $m_{11} + m_{10} = 518$, i.e. the number of non-respondents among the Register-Employment's, and similarly $m_{01} + m_{00} =$

	$(R,X)\|Y$				$P[(R,X)\|Y]$			
	R = 0		R = 1		R = 0		R = 1	
	X = 1	X = 0	X = 1	X = 0	X = 1	X = 0	X = 1	X = 0
Y = 1	n_{11}	n_{10}	m_{11}	m_{10}	$\xi_1(1-\gamma_1)$	$(1-\xi_1)(1-\gamma_0)$	$\xi_1\gamma_1$	$(1-\xi_1)\gamma_0$
Y = 0	n_{01}	n_{00}	m_{01}	m_{00}	$\xi_0(1-\gamma_1)$	$(1-\xi_0)(1-\gamma_0)$	$\xi_0\gamma_1$	$(1-\xi_0)\gamma_0$

Table 2: The complete, latent data of the Norwegian LFS, and its distribution.

796. The complete, latent likelihood is

$$L_1 \propto \xi_1^{n_{11}+m_{11}}(1-\xi_1)^{n_{10}+m_{10}}\xi_0^{n_{01}+m_{01}}(1-\xi_0)^{n_{00}+m_{00}} \\ \gamma_1^{m_{11}+m_{01}}(1-\gamma_1)^{n_{11}+n_{01}}\gamma_0^{m_{10}+m_{00}}(1-\gamma_0)^{n_{10}+n_{00}}.$$

Resampling targeted at $\pi(m_{yx}, \xi_y, \gamma_x) \propto L_1$ is a standard worked-out case (Tanner, 1993, "The Genetic Linkage Example"). At each Gibbs iteration, we draw m_{y1} from Binomial$(m_{y1}+m_{y0}, p_y)$ where $p_y = \xi_y\gamma_1/(\xi_y\gamma_1+(1-\xi_y)\gamma_0)$, and ξ_1 from Beta$(n_{11}+m_{11}+1, n_{10}+m_{10}+1)$, and so on from the respective Beta distributions of ξ_0, γ_1 and γ_0. We have plotted the resampled l against the overall LFS-Employment Rate, denoted by p (Figure 1), using respectively 500 and 5000 Gibbs iterations. To calibrate the sample mle by means of the MC profile likelihood, we resample under

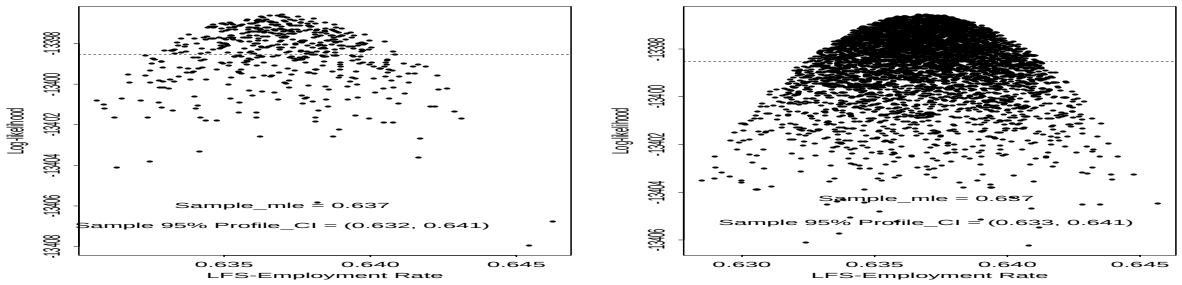

Figure 1: Illustration of the log-likelihood generated by the Gibbs sampler. (With 500 iterations on the left, and 5000 iterations on the right.)

the constraint of $p = 0.613 \cdot \xi_1 + 0.387 \cdot \xi_0$ over a grid of p, i.e. (0.636, 0.637, 0.638), where 0.613 is the known marginal probability of Register-Employment. At each iteration, we still resample ξ_1 from Beta$(n_{11}+m_{11}+1, n_{10}+m_{10}+1)$, but calculate ξ_0 directly from the constraint. This is the only change from the unconstrained resampling we made. Notice that the modified Markov chain does *not* converge to a marginal distribution proportional to L_1 under the constraint. However, as the plots of $L(\theta)$ against ξ_1 and ξ_0 (Figure 2) illustrate, this hardly matters for our purpose. In any case, based on the Monte Carlo profile likelihood of 0.636 - 0.638, we retained 0.637 as the mle of the overall LFS-Employment Rate, which is the same as that from the proper EM algorithm (Zhang, 1999). Applying the same calibration procedure, we obtained $(0.634, 0.640)$ as its 95% confidence interval derived from the profile likelihood. We note in general that, since the confidence interval here does not address the uncertainty about the non-response mechanism, it should be treated with caution. However, compared to the post-stratified estimate, i.e. 0.645, and the simple sample mean, i.e. 0.651, it is clear that the bias caused by non-response dominates the sampling error in the Norwegian LFS, at which attention should be directed in the future.

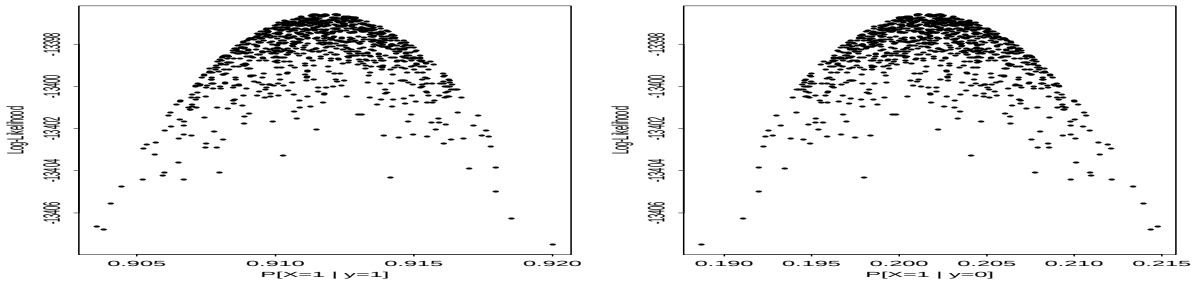

Figure 2: Illustration of the constrained Gibbs sampler: $P[X = 1|y = 1]$ on the left and $P[X = 1|y = 0]$ on the right. (With 1000 iterations at $p = 0.637$.)

5 The Fertility Survey 1977

A primary interest of the Norwegian Fertility Survey 1977 (Table 3) was the distribution of the number of live-births per female member of the population. Since the households with few or no

Number of Live-Births	0	1	2	3	4	5	≥ 6	Missing
The Sample	886	640	1065	548	216	61	22	535

Table 3: The data of the Norwegian Fertility Survey 1977.

children are more difficult to reach than otherwise, a GPS mechanism is here intuitive, as it is in many household surveys. We notice that some of the responses were obtained through call-backs. Indeed, Thomsen and Siring (1983) considered an ignorable non-response model which allowed the response rate to vary between the calls. Bjørnstad (1995) applied the predictive likelihood approach under the same model, where (a) the numbers of live-births of the non-respondents were obtained from the administrative registers which, however, were not available at the time of the survey, and (b) the following modified Poisson distribution was suggested for the number of live-births, denoted by X, in the population:

$$P(X = 0) = \psi \qquad \text{and} \qquad P(X = x) = (1 - \psi)\lambda^{x-1}e^{-\lambda}/(x - 1)! \quad \text{for } x = 1, 2, ...$$

Let R be the non-response indicator. We apply model (1) and (2) to $R|x$, where we assume $P(R = 1|x \geq 6)/P(R = 0|x \geq 6) = \alpha(1 + 6)^{\beta}$. In particular, we expect β to be negative, such that the non-response probability decreases as the number of live-births increases. Let A_x be the number of respondents with x live-births, and B_x that of the non-respondents, and $C_x = A_x + B_x$. That is, $\sum_x b_x = 535$, i.e. the total number of missing and denoted by m, and $\sum_x c_x = 3973$, i.e. the sample size and denoted by n. The complete, latent likelihood is, for $\zeta_\lambda = P[X \geq 6; \lambda|X > 0]$,

$$L_1(\theta) = \psi^{c_0}(1 - \psi)^{n-c_0} \cdot \lambda^{\sum_{x=1}^{5} c_x(x-1)} e^{-\lambda(n-c_0-c_6)} \zeta_\lambda^{c_6} \cdot \prod_{x=0}^{6} \frac{\{\alpha(x + 1)^{\beta}\}^{b_x}}{\{1 + \alpha(x + 1)^{\beta}\}^{c_x}}.$$

We put up the following Gibbs sampler, where each step is to be carried out conditional to the present values of the remaining variables and parameters:

1. The augmentation step: $b_x \sim Multinomial(m, p_x)$, where p_x is the cell-probability condi-

tional to missing, i.e. $p_x = P[X = x|R = 1]$ for $x = 0, 1, ..., 5$, and $p_6 = P[X \geq 6|R = 1]$.

2. Draw $\lambda \sim Gamma(u, v)$ with shape parameter $u = 1 + \sum_{x=1}^{6} c_x(x-1)$ and scale parameter $v = 1/(n - c_0)$.

3. Draw $\psi \sim Beta(1 + c_0, 1 + n - c_0)$.

4. Draw $\alpha = \{z/(1-z)\}/\tilde{x}_\beta$, where $z \sim Beta(1+m, 1+n-m)$, and $\tilde{x} = \sum_x c_x(x+1)^\beta / \sum_x c_x$.

5. Draw $\beta \sim Multinomial(1, p_i)$, where $p_i = L_1(\beta_i)/\sum_{j=1}^{k} L_1(\beta_j)$ for an equi-distance grid of β, denoted by $(\beta_1, ..., \beta_k)$, with sufficiently large k.

Notice that, apart from step 1 and 3, the Gibbs sampler does not aim at the exact conditional distributions. For λ, we substitute for $P[X \geq 6|X > 0]$ by $P[X = 6|X > 0]$, which results into the conditional Gamma distribution at step 2. The grid-sampling of β was deviced for the griddy Gibbs sampler by Ritter and Tanner (1992). We could, of course, sample β continuously, if we include one extra sub-step to draw uniformly around the sampled β. Finally, for α, we expand the involved part of L_1 around $\tilde{x}_\beta \approx E[(X+1)^\beta]$, and obtain the leading term $z^m(1-z)^{n-m}$, where $z = \alpha\tilde{x}_\beta/(1+\alpha\tilde{x}_\beta)$ and $z \sim Beta(1+m, 1+n-m)$.

Figure 3 was based on 5000 Gibbs iterations, where we used a grid of 1100 points over $\beta \in (-10, 1)$. While the profile likelihood based confidence intervals were derived directly from the

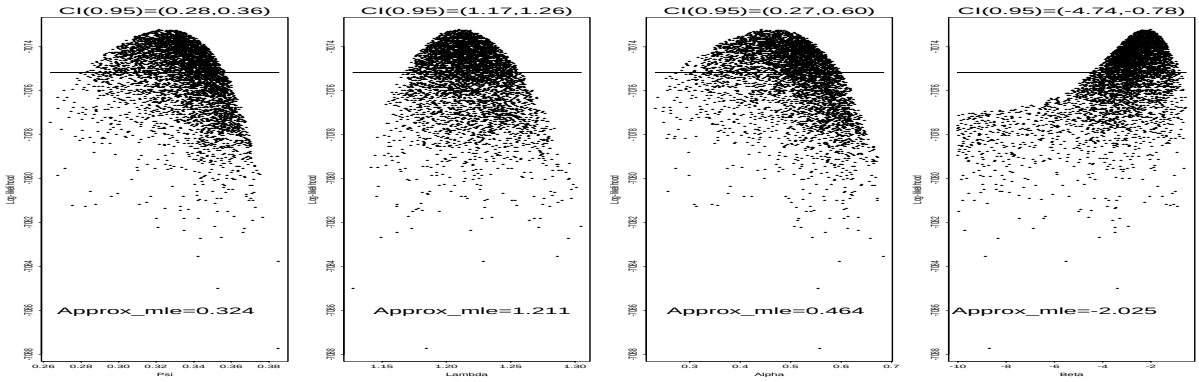

Figure 3: Illustration of the log-likelihood generated by 5000 Gibbs iterations.

sample, the approximate mle was calibrated by means of the R-B MC profile likelihood. The Markov chain covered the high-likelihood region reasonably well. The profile log-likelihood of β appears very flat for $\beta < -5$, where just about all non-response is attributed to women with no live-births, whose response probability is determined by α alone. Under the GPS non-ignorable non-response model, we have $\hat{c}_x/n = a_x/n + \hat{b}_x/n$, where $\hat{b}_x = E[B_x|\sum_x B_x = 535]$ is evaluated at the approximate mle. In Table 4 we compare this to $a_x/(n-m)$, i.e. based on the respondents alone, as well as the true c_x/n acquired from the registers. The respondents alone contained large bias due to non-response, which was greatly reduced under the GPS non-ignorable model despite its relatively poor fit. The deviance, i.e. twice the difference between the maximum reachable and the fitted log-likelihood, was 33.52 on 3 degrees of freedom. Over-dispersion could be one of the reasons, now that no distinction was made among the calls. This, however, seems to have mattered little regarding the inference of interest.

The Sample Distribution (%)	Number of Live-Births						
	0	1	2	3	4	5	≥ 6
Sample Proportion among the Respondents	25.8	18.6	31.0	15.9	6.3	1.8	0.6
Conditional Expectation under the GPS Model	32.7	17.9	27.7	14.0	5.5	1.6	0.6
True Distribution Acquired from the Registers	32.2	18.4	27.5	14.4	5.5	1.6	0.6

Table 4: The distribution of the number of live-births within the sample of the Fertility Survey 1997.

6 The Alveolar Bone Loss data with missing teeth

Clinical studies are commonly hampered by the problem of missing sites (Lawrence, Beck, Hunt, and Koch, 1996; Crawford, Tennstedt, and Mckenlay, 1995). The ABLS data (Schüller, Thomsen, and Holst, 1998) contained 813 persons of Age 45 to 64. From each of them, the Alveolar Bone Loss was intended to be measured for 24 teeth, i.e. the sites. The Mean Bone Loss is considered to be a personal health index. Among other things one is interested in how this varies with Age. Since it is suspected that the non-response is positively correlated with the degree of Bone Loss, the analysis could be biased without appropriate adjustment. Figure 4 contained results of some exploratory data analysis (EDA) based on the observed data. In the bottom plot of Figure 4,

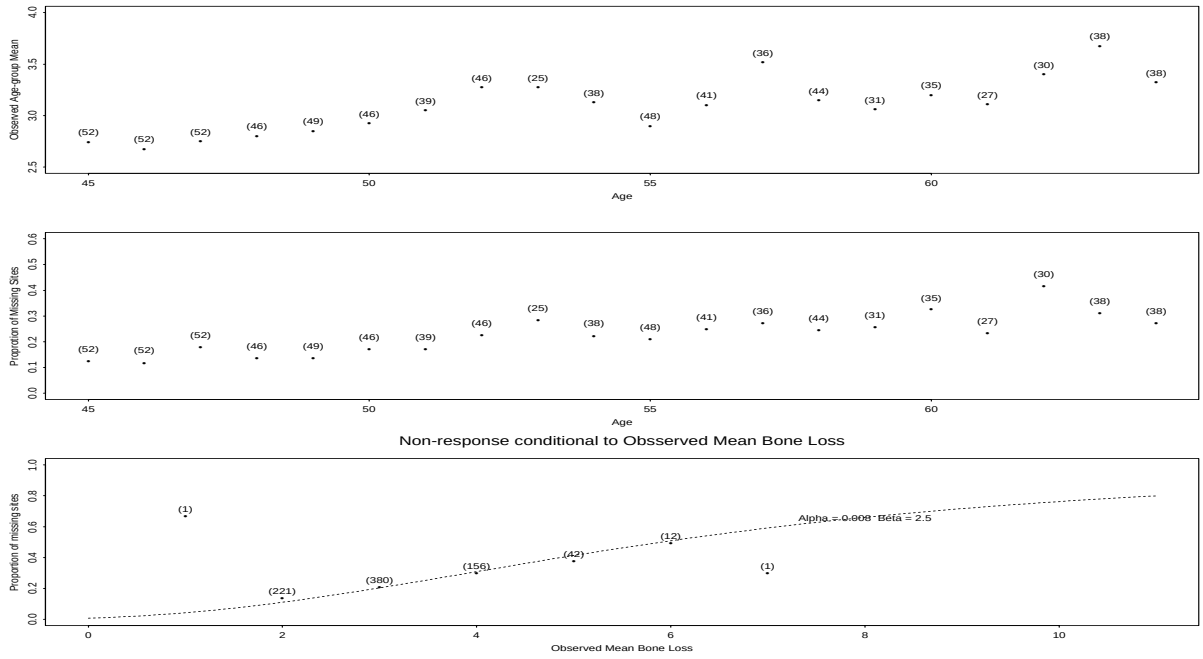

Figure 4: (a) Age-group mean based on the observed personal Mean Bone Loss (top), (b) mean proportion of missing sites within each Age-group (middle), (c) mean proportion of missing sites conditional to observed Mean Bone Loss (bottom). (In the parentheses: the number of persons for each point.)

we grouped the persons according to their observed Mean Bone Loss, i.e. rounded to the nearest integer, to explore the proportions of missing sites within such groups. The dotted model-curve was calculated under (1) and (2), i.e. at $(\alpha, \beta) = (0.008, 2.5)$ based on guessing. The same plot has also suggested two clear anomalies in the non-response pattern, who were identified and removed

from the subsequent analysis.

Let X_{ij} be the Bone Loss at site j of person i, for $i = 1, ..., n$, $j = 1, ..., 24$, and $n = 811$. Let R_{ij} be the corresponding non-response indicator. It seems unrealistic to treat $R_{ij}|x_{ij}$ as independent of each other. Let $R_i = \sum_j R_{ij}$ and $\bar{x}_i = \sum_j x_{ij}/24$. We assume, following (3),

$$R_i|x_i \sim TrnPoisson(\lambda_i) \quad \text{where } 0 \leq r_i \leq 24, \text{ and } \eta_i = \alpha(\bar{x}_i + 1)^\beta, \text{ and } \lambda_i = 24\eta_i/(1+\eta_i).$$

Divide the sample into successive 5-year Age-groups, and let $a_i = 1, 2, 3, 4$ be the group-index of person i. We model X_{ij} as a convolution of two independent variance components, i.e.

$$X_{ij} = \mu_i + u_i + e_{ij} \quad \text{where } \mu_i = \psi_0 + \psi_1(a_i - 1), \text{ and } u_i \sim N(0, \sigma_u^2), \text{ and } e_{ij} \sim N(0, \sigma_e^2),$$

and $u_i \perp u_k$ for $i \neq k$, and $e_{ij} \perp e_{kl}$ for $i \neq k$ or $j \neq l$, and $\{u_i\} \perp \{e_{ij}\}$. In this way, $(X_{i,1}, ..., X_{i,24})^T$ are independent outcomes from the Truncated Multivariate Normal distribution, with mean $(\mu_i)_{j=1}^{24}$ and covariance matrix (τ_{jk}) where $\tau_{jk} = \sigma_u^2 + \sigma_e^2$ if $j = k$ and σ_u^2 if $j \neq k$, for $i = 1, ..., n$ and $0 \leq x_{ij} \leq 11$, where 11 is the maximum possible Bone Loss. Partition $(x_{ij})^T = (x_{i,obs}^T, x_{i,mis}^T)$ into the observed and missing parts. The complete, latent likelihood is given as

$$L_1 \propto \prod_{i=1}^{n} \omega_i^{-1} \cdot \Phi_{24-r_i}(x_{i,obs}; \xi|a_i) \cdot \Phi_{r_i}(x_{i,mis}; \xi|x_{i,obs}, a_i) \cdot \varpi_i^{-1} \cdot \zeta(r_i; \gamma|\bar{x}_i),$$

where $\xi = (\psi_0, \psi_1, \sigma_u, \sigma_e)^T$, and $\gamma = (\alpha, \beta)^T$, and Φ_m the probability density function (pdf) of the corresponding m-variate Normal distribution, and ω_i the joint probability of $X_{ij} \in [0, 11]$ for $1 \leq j \leq 24$, and ζ the probability of the corresponding Poisson distribution, and ϖ_i the probability of $R_i \in [0, 24]$. Let $\bar{x}_{i,obs}$ and $\bar{x}_{i,mis}$ be the respective means of $x_{i,obs}$ and $x_{i,mis}$. We have

$$L \propto \prod_{i=1}^{n} \omega_{i,obs}^{-1} \cdot \Phi_{24-r_i}(x_{i,obs}; \xi|a_i) \cdot \int \varpi_i^{-1} \cdot \zeta(r_i; \gamma|\bar{x}_i) \cdot \omega_{i,mis}^{-1} \cdot \phi(\bar{x}_{i,mis}; \xi|x_{i,obs}, a_i)\ d\bar{x}_{i,mis},$$

where ϕ is the pdf of the corresponding univariate Normal distribution, and $\omega_{i,obs}$ the truncation factor of $X_{i,obs}$, and $\omega_{i,mis}$ that of $\bar{X}_{i,mis}|x_{i,obs}$. Notice that $x_{i,mis}$ has been marginalized into $\bar{x}_{i,mis}$. The integral over $\bar{x}_{i,mis}$ differs according to i, and a Monte Carlo evaluation, "many-sample" or not, is highly inefficient. We employed the Laplace approximation. The factor $\omega_{i,mis}$ being very close to 1, we take $\mu_{i,mis} = E[\bar{x}_{i,mis}; \xi|\bar{x}_{i,obs}, a_i]$ as the mode, which gives us, omitting $\omega_{i,obs}$,

$$\tilde{L} \propto \prod_{i=1}^{n} \Phi_{24-r_i}(x_{i,obs}; \xi|a_i) \cdot \omega_{i,mis}^{-1} \cdot \varpi_i^{-1} \cdot \zeta(r_i; \gamma|\bar{x}_{i,obs}, \mu_{i,mis}).$$

We put up the following Gibbs sampler, where each step is to be carried out conditional to the present values of the remaining variables and parameters:

1. The augmentation-step: $\bar{x}_{i,mis} \sim \phi(\bar{x}_{i,mis}; \xi|\bar{x}_{i,obs}, a_i) \Rightarrow \bar{x}_i = \{(24 - r_i)\bar{x}_{i,obs} + r_i\bar{x}_{i,mis}\}/24$.

2. Draw $\lambda \sim Gamma(c, b)$ with shape parameter $c = 1 + \sum_i r_i$ and scale parameter $b = 1$, and set α to be the solution of $g(\alpha) = \sum_i \lambda_i = \lambda$.

3. Draw $\beta \sim N(\beta_0, \sigma_\beta^2)$, where $\beta_0 = 2.5$ and $\sigma_\beta = 0.1$.

4. Draw $\sigma_e = \nu - \sigma_0$, where $\sigma_0 = 0.12$, and $\nu^2 \sim InverseGamma(c, b)$ with shape parameter $c = \sum_i (24 - r_i)/2 - 1$, and scale parameter $b = \sum_{i=1}^{n} \sum_{j=1}^{24-r_i} (x_{ij} - \bar{x}_{i,obs})^2 / \{2 - 2/(24 - r_i)\}$.

5. Draw $\sigma_u = \nu - \sigma_e^2/24$, where $\nu^2 \sim InverseGamma\{n/2 - 1, \sum_i (\bar{x}_i - \mu_i)^2/2\}$.

6. Draw $(\psi_0, \psi_1)^T \sim \Phi_2(\mu, \tau_0)$, where $\mu = (A^T A)^{-1}(A^T \bar{x})$, and $\tau_0 = (\sigma_u^2 + \sigma_e^2/24)(A^T A)^{-1}$, and $\bar{x}^T = (\bar{x}_1, ..., \bar{x}_n)$, and the ith row of $A_{n\times 2}$ is given as $(1, a_i)$.

First of all, notice that the Markov chain by no means converges to a marginal distribution proportional to L_1. The generation of α was based on $\sum_i R_i \sim Poisson(\sum_i \lambda_i)$. A similar scheme for β, however, led to unstable Newton-Raphson algorithm. The Inverse Gamma distribution for σ_e^2 disregards the constraint of $\bar{x}_{i,obs}$ and covariances among $x_{i,obs} - \bar{x}_{i,obs}$, in which respect σ_0 was conceived as a tuning parameter for bias-adjustment. (Sampling from $\phi(\bar{x}_{i,obs} - \bar{x}_{i,mis})$, which is a pivotal of σ_e^2, was found to give divergent results.) Conditional sampling of $(\sigma_u^2, \psi_0, \psi_1)$ was the standard solution based on $\prod_i \phi(\bar{x}_i; \mu_i, \sigma_u^2 + \sigma_e^2/24)$. Finally, to tune in on the parameter, i.e. $(\beta_0, \sigma_\beta, \sigma_0)$, of the independence sampling of β and σ_e^2, we used the following procedure: (1) fix (σ_e^2, β) at some initial values, say, $\sigma_e^2 = \sum_i \sum_{j=1}^{24-r_i} (x_{ij} - \bar{x}_{i,obs})^2 / (\sum_i (24 - r_i) - 1)$ and $\beta = 2.5$ by EDA, and resample the rest of the parameters; (2) fix the rest of the parameters at the sampled mode and β at 2.5, now resample σ_e^2 alone, by means of random walk or independence uniform, to choose σ_0 so that Step 4 is approximately relocated about $\hat{\sigma}_e$, (3) tune in on (β_0, σ_β) similarly.

We have plotted $\tilde{L}$ against ξ (Figure 5) based on 5000 Gibbs iterations, together with the respective sample-based mle and the 95% profile likelihood confidence intervals. The truncation

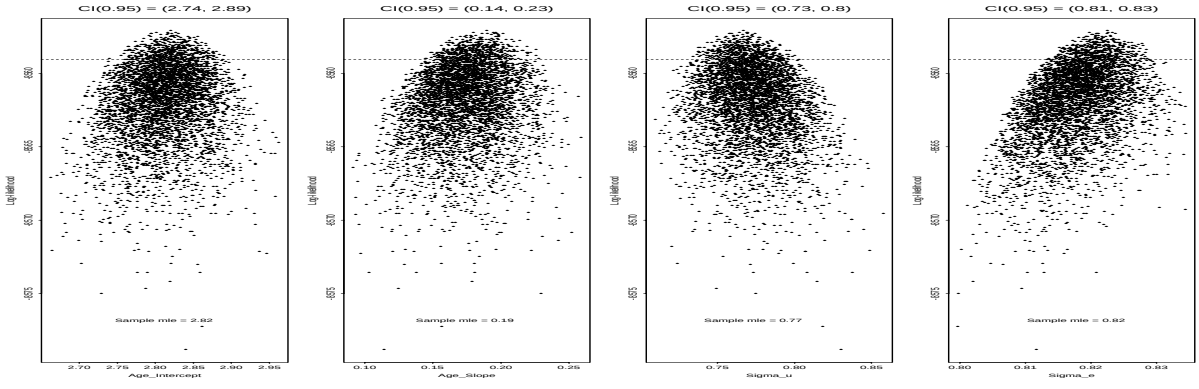

Figure 5: Illustration of the approximate log-likelihood generated by 5000 Gibbs iteration.

probability for $X_{i,obs}$ varies little around the mle $\hat{\xi}$, so that $\tilde{L}$ approximates the Monte Carlo L, denoted by L_{MC}, rather well. On our Sun Ultra5 Unix-server, each evaluation of L_{MC} (with 5000 points for the Monte Carlo involved) in Splus takes over 1000 times longer than that of $\tilde{L}$, so that the Laplace approximation was certainly helpful. The mle $\hat{\xi}$ can be calibrated as the following. For a fixed value of ψ_0, we calculate the R-B $\tilde{L}$, by varying the rest of the parameters over the same resample. We then calculate L_{MC} for a small number of θ which had the highest $\tilde{L}$. We take the highest L_{MC} as the Monte Carlo $L_{Pro}(\psi_0)$. Repeating the procedure over a grid values of ψ_0, we calibrate $\hat{\psi}_0$ as the one with the highest $L_{Pro}(\psi_0)$. The end points of the confidence interval can similarly be calibrated, where we approximate $\hat{L} = L(\hat{\theta})$ by $\max(L_{MC})$ over the resample.

We obtained the mle of interest $(\hat{\psi}_0, \hat{\psi}_1) = (2.812, 0.193)$, calibrated over (2.805, 2.810, 2.811, 2.812, 2.813, 2.814, 2.815, 2.820) for ψ_0 and (0.185, 0.190, 0.191, 0.192, 0.193, 0.194, 0.195, 0.200) for ψ. In comparison, the simple OLS estimates based on $\bar{x}_{i,obs}$ were (2.813, 0.176), which ignored the different variances of $\bar{x}_{i,obs}$; whereas the variance component approach under the "nestet-error regression model" (Fuller and Battese, 1973) gave us (2.809, 0.172). Both estimators require the likelihood to factorize for ξ and γ, which is the case under the ignorable non-response, i.e.

$$R_i|\bar{x}_{i,obs} \sim TrnPoisson(\lambda_i) \quad \text{where } 0 \leq r_i \leq 24, \text{ and } \eta_i = \alpha(\bar{x}_{i,obs} + 1)^{\beta}, \text{ and } \lambda_i = 24\eta_i/(1+\eta_i).$$

Non-response becomes now a nuisance of the inference of interest, so does over-dispersion. On the other hand, suppose we set up a hierarchical random-effect model, say, $\gamma_i \sim \Phi_2(\gamma, \tau_\gamma)$ for the non-response mechanism. The likelihood under the present marginal-effect model is then the Laplace approximation of that under the extended random-effect model. Finally, the mle $\hat{\gamma}$ is $(\hat{\alpha}, \hat{\beta}) = (0.008, 2.5)$ based on $\prod_i \varpi_i^{-1} \zeta(r_i; \gamma|\bar{x}_{i,obs})$, compared to $(\hat{\alpha}, \hat{\beta}) = (0.009, 2.4)$ under the non-ignorable model. Whereas $\bar{x}_{i,obs}$ alone constitutes the health index under the ignorable model, both r_i and $\bar{x}_{i,obs}$ are considered to be informative under the non-ignorable model. The two gave almost identical (marginal) account of the non-response, however, the non-ignorable model adjusted the estimated marginal effect of Age, i.e. $\hat{\psi}_1$, by about one standard deviation upwards. Since the assimilation of these results seems to require more detailed subject matter consideration, both should be included in a report of the statistical analysis.

7 Summary

As the Norwegian LFS data have shown, an ignorable non-response model may result into substantial reduction of the bias caused by non-response, provided good correlation between the auxiliary and object variables. Even in such cases, non-ignorable non-response models can be necessary if only from a sensitivity analysis point of view. However, non-response model selection differs from model selection in the complete-data case. Goodness-of-fit coupled with parsimonious parameterization seems neither sufficient nor necessary as a general criterion. The effect which the non-response models have on the inference of interest become more important than ever, in which respect it is sometimes appropriate to focus on the marginal effect which the object variable has on non-response. Over-dispersion may occur as a consequence of such marginal-effect modeling. In many cases, the likelihood can be considered a Laplace approximation to that under the model of random-effect non-response mechanism. However, we know little of how far the argument carries us in general. The particular form of the non-response models adopted in this paper, i.e. (1) to (3), can be useful if the size has many levels or is continuous. A generalized proportional to size non-response mechanism, on the other hand, is fairly universal. The term could be taken literally in almost all survey sampling. In an on-going project at Statistics Norway, we are studying the possibilities of applying the GPS non-ignorable model to the estimation of the household distribution in the population.

Non-ignorable non-response models may lead to numerical difficulties w.r.t. integration and maximization. In this paper we have concentrated on adapting the Markov chain Monte Carlo techniques to effectively generate the profile likelihood inference. The approach enjoys extra

flexibility compared to the standard applications of these methods under the Bayesian framework, because it is not required that the Markov chain converges exactly to some fixed target distribution. In fact, we sometimes do not know where, or if, it has converged at all. The profile likelihood inference can be drawn based on single or multiple samples, provided reasonable coverage of the high-likelihood parameter region. Basically, all we need is to be able to evaluate the likelihood.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.

Bjørnstad, J.F. (1995). Utvalgsundersøkelser og Prediksjon. Lecture Notes, University of Trondheim - AVH, In Norwegian.

Brooks, S.P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Crawford, S.L., Tennstedt, S.L., and Mckenlay, J.B. (1995). A comparison of analytic methods for non-random missingness of outcome data. *J. Clin. Epidemiol.*, **48**(2), 209–219.

Davison, A.C. and Stafford, J.E. (1998). The score function and a comparison of various adjustments of the profile likelihood. *The Canadian Journal of Statistics*, **26**, 139–148.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B*, **39**, 1–38.

Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *J. Amer. Statist. Assoc.*, **68**, 626–632.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integrations. *Econometrica*, **24**, 1317–1339.

Geyer, C.J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice*, chap. 14, pp. 241–258. Chapman and Hall.

Ibrahim, J.G., Chen, M.H., and Lipsitz, S.R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, **55**, 591–596.

Ibrahim, J.G., Lipsitz, S.R., and Chen, M.H. (1999). Missing covariates in generalized linear models when missing data mechanism is non-ignorable. *J. Roy. Statist. Soc. B*, **61**, 173–190.

Lawrence, H.P., Beck, J.D., Hunt, R.J., and Koch, G.G. (1996). Adjustment of the M-component of the DMFS index for prevalence studies of older adults. *Community Dent Oral Epidemiology*, **24**, 322–331.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *J. Off. Statist.*, **15**, 305–327.

Molenberghs, G., Kenward, M., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random drop-out. *Biometrika*, **84**, 33–44.

Ritter, C. and Tanner, M.A. (1992). The Gibbs stopper and the greedy Gibbs sampler. *J. Amer. Statist. Assoc.*, **87**, 861–868.

Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric nonresponse models. (With discussions). *J. Amer. Statist. Assoc.*, **94**, 1097–1146.

Schüller, A.A., Thomsen, I.O., and Holst, D. (1998). Adjusting estimates of alveolar bone loss for missing observations — developing and testing a general model. *Journal of Dental Research*, To appear.

Severini, T.A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika*, **85**, 507–522.

Smith, P.W.F., Skinner, C.J., and Clarke, P.S. (1999). Allowing for non-ignorable non-response in the analysis of voting intention data. *Appl. Statist.*, **48**, 563–577.

Stern, S.E. (1997). A second-order adjustment to the profile likelihood in the case of a multidimensional parameter of interest. *J. Roy. Statist. Soc. B*, **59**, 653–665.

Tanner, M.A. (1993). *Tools for Statistical Inference* (2nd edn). Springer-Verlag.

Thomsen, I. and Holmøy, A.M.K. (1998). Combining data from surveys and administrative record systems. The Norwegian experience. *Int. Statist. Rev.*, **66**, 201–221.

Thomsen, I. and Siring, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. In *Incomplete Data in Sample Surveys*, vol. **3**. New York: Academic Press.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82–86.

Troxel, A.B., Harrington, D.P., and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Appl. Statist.*, **47**, 425–438.

Zhang, L.-C. (1999). A note on post-stratification when analyzing binary survey data subject to nonresponse. *J. Off. Statist.*, **15**, 329–334.