Journal of Official Statistics
Vol. 5. No. 3, 1989, pp. 223–239

Generalized Linear Modeling of Sample Survey Data

Lennart Nordberg[1]

Abstract: The theme of this paper is regression analysis – extended to Generalized linear models (GLMs) – of sample survey data, with the data obtained by a more or less complex survey design and possibly affected by nonresponse.

The suggested approach is neither purely model based nor purely design based. In fact we consider, simultaneously, three sources of random variation, specified by a superpopulation model (a GLM), the sampling design and a response model.

Ordinary (ML-based) inference – being based on the assumption of independent observations – is not automatically valid in this situation. It is, however, shown that ordinary inference does apply under certain conditions. It is demonstrated – and illustrated by simulations – how these conditions can be checked and met by incorporating variables associated with the design and the response pattern into the model.

Furthermore, it is demonstrated by simulation results that ordinary, unweighted GLM inference – when valid – can be considerably more efficient than inference based on Horvitz–Thompson weighting.

Key words: Analysis of survey data; generalized linear models; superpopulation models; nonresponse.

1. Introduction

The theme of this paper is regression analysis of sample survey data, the data being obtained by a more or less complex survey design and possibly affected by nonresponse. The word regression should be interpreted in a fairly wide sense here. We will consider generalized linear models (GLMs), c.f. McCullagh and Nelder (1983), which include, e.g., linear, logistic, probit, and Poisson regressions.

Smith (1981) – who considers linear regression for complex surveys – distinguishes between descriptive and analytic inference, and we will use this terminology.

In descriptive inference, the objective is to estimate a parameter, **B**, say, which is a specified function of the elements of a given finite population. In this approach one pays attention only to random variation which emanates from the sampling and the nonresponse mechanism. If all the elements of the whole population were to respond there would be no uncertainty. A descriptive approach to generalized linear modeling of survey data is found in Binder (1983).

[1] Senior Statistician, Statistical Research Unit, Statistics Sweden, S-70189 Örebro, Sweden.

Acknowledgements: I am indepted to Bengt Rosén, David Binder, Eva Elvers, a referee and the editor for their helpful comments on earlier versions of this manuscript. Naturally, the author must claim the full responsibility for whatever shortcomings that may still remain in the paper.

Our approach in this paper will be analytic. Then the interest focuses on the (unknown) relation between some variables y and $\mathbf{x}$. This relation is assumed to be of interest not only as a description of the structure in the particular population at the time of the survey, but also to have a more general interpretation. The relation between y and $\mathbf{x}$ is assumed to be expressable by a family of statistical distributions, a superpopulation model, indexed by a parameter β. This model parameter – rather than the fixed population quantity $\mathbf{B}$ – is in the focus of interest in the analytic approach.

The present paper treats three main topics. The first one concerns conditions under which ordinary unweighted GLM inference can apply to data obtained by complex survey designs. It is shown in Section 3 that the ordinary inference – which is not automatically valid in this situation – does apply under certain conditions, and various ways to check these are suggested.

The following is a rough characterization of these conditions. If the design – expressed by first and second order inclusion probabilities – does not contain any "relevant" information on y which is not already accounted for by $\mathbf{x}$ then the ordinary GLM inference is valid. The conditions are expressed in terms of correlations between inclusion probabilities (regarded as random variables) and model residuals.

The second topic is a comparison of ordinary unweighted GLM inference to inference based on Horvitz-Thompson weighting – the latter is summarized in Section 4. The unweighted GLM inference – when valid – is more efficient than Horvitz-Thompson based inference. The magnitude of this efficiency gain can be substantial as demonstrated by the simulation study in Section 4.

If unweighted GLM inference does not apply, then Horvitz–Thompson weighting can be recommended in some cases. However, an alternative and sometimes superior approach is the following one. Since the design now contains some residual information on y, not accounted for by $\mathbf{x}$, it may be possible to use this information to improve the model rather than just to account for it by weighting. Suppose there is a variable $\mathbf{z}$, say, which is highly correlated with the inclusion probabilities. Bring $\mathbf{z}$ into the model building process if it makes sense from a subject matter point of view. This should result in an improved model, and the validity conditions for unweighted inference may now be met by the new model.

Although our emphasis is on model building rather than estimation of a fixed population quantity, our approach is not purely model based. In fact we consider three sources of random variation, specified by a superpopulation model, the sampling design, and a response model. This is the third main topic of the paper. We will assume that data are generated by a three-step process as follows.

i. A population of N elements is generated by a specified superpopulation model.
ii. From the population generated in (i) a sample of prescribed size n, $n \leqslant N$, is drawn according to a specific sampling design.
iii. An element of the sample generated in (ii) may or may not respond according to a specific response model.

A more detailed and technical specification of the above steps (i) through (iii) follows. In particular we will concentrate on GLM superpopulation models. To avoid burdening the text by too many technical details the introduction of nonresponse is postponed until Section 5. In Section 2, steps (i) and (ii) are specified and this frame-

work is used in Sections 3 and 4. Step (iii) is specified in Section 5 and the previous results are then generalized to the full three-step specification.

Related approaches – although restricted to linear regression and without step (iii) – are found in Du Mouchel and Duncan (1983), Nathan and Holt (1980), and Ten Cate (1986). Treatment of nonresponse within the general framework of (i) through (iii) is found in Rubin (1976, 1987), Little (1982) and Little and Rubin (1987). Other relevant references are Fay (1986) and Scott (1987).

The approach advocated in the present paper contains as special cases classical regression – where data are generated as independent observations from a family of distributions – as well as the descriptive approach to regression, c.f. Binder (1983), which essentially builds on sample survey theory. We may then have a way to bridge the gap between the two approaches.

2. Specification of Superpopulation and Sampling Mechanisms

Let $\{Y_i\}_{i=1}^N$ be independent random variables, taking values in $\psi \subseteq R$, following a generalized linear model, c.f. McCullagh and Nelder (1983). The probability density g_i for Y_i takes the form

$$g_i(y, \beta, \phi) = \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi w_i} + c_i(y, \phi)\right], \quad y \in \psi, \tag{2.1}$$

where $\beta = (\beta_0, \ldots \beta_m)'$ and ϕ are unknown parameters while θ_i, $i = 1, 2, \ldots, N$, depends on β through a relation of the type

$$\theta_i = f\left(\sum_{k=0}^{m} \beta_k x_{ki}\right). \tag{2.2}$$

The ws are known scale factors and the xs are known covariates playing the role of explanatory variables. (We could also regard x as random and then make the inference conditional on x.) Furthermore the functions $b(\cdot)$, $c_i(\cdot)$ and $f(\cdot)$ are known and sufficiently regular. Notice that w may be contained in c_i while b and f are free of w. Let $\mu_i = E(Y_i)$ and $\sigma_i^2 = \text{Var}(Y_i)$. The following relations are straightforward consequences of (2.1)–(2.2), c.f. McCullagh and Nelder (1983, pp. 20–21).

$$\mu_i(\beta) = b'(\theta_i), \tag{2.3}$$

$$\sigma_i^2(\beta) = b''(\theta_i)\phi w_i. \tag{2.4}$$

Let the Y values generated through (2.1) and (2.2) make up a population $\Omega_N = \{i : i = 1, 2, \ldots, N\}$. Now, a sample Σ_n of prescribed size n, $n \leqslant N$, is drawn from the elements of Ω_N. Let for $i = 1, 2, \ldots, N$

$$\delta_i = \begin{cases} 1 \text{ if } i \in \Sigma_n \\ 0 \text{ otherwise.} \end{cases} \tag{2.5}$$

We assume that all the relevant information about the sampling design is contained in a set of, possibly multidimensional, random variables $\mathbf{z}_i$, $i = 1, 2, \ldots, N$ – the design variables. The first and second order inclusion probabilities are defined as follows.

$$\pi_i = P(\delta_i = 1 | \mathbf{z}) \quad i = 1, 2, \ldots, N. \tag{2.6}$$

$$\pi_{ij} = P(\delta_i = 1, \quad \delta_j = 1 | \mathbf{z}) \quad i, j, = 1, 2, \ldots, N. \tag{2.7}$$

Notice that the πs, being functions of $\mathbf{z}$, must here be regarded as random variables. For notational convenience we introduce the following vectors for $i = 1, 2, \ldots, N$

$$\mathbf{t}_i = (Y_i, \mathbf{x}_i, \mathbf{z}_i)$$

$$\mathbf{t} = (t_1, t_2, \ldots, t_N)'$$

where

$$\mathbf{x}_i = (x_{0i}, \ldots, x_{mi}). \tag{2.8}$$

Since $\mathbf{z}$ holds all relevant information about the design, we have the following relations which will be useful later.

$$\mathrm{E}(\delta_i|\mathbf{t}) = \mathrm{E}(\delta_i|\mathbf{z}) = \pi_i(\mathbf{z})$$

$$i = 1, 2, \ldots, N. \tag{2.9}$$

$$\mathrm{E}(\delta_i\delta_j|\mathbf{t}) = \mathrm{E}(\delta_i\delta_j|\mathbf{z}) = \pi_{ij}(\mathbf{z})$$

$$i, j = 1, 2, \ldots, N. \tag{2.10}$$

3. Unweighted Estimation

We will now derive sufficient conditions under which ordinary unweighted GLM inference is valid within the framework of Section 2. The point of departure will be the following equation

$$\sum_{i=1}^{N} \left(\frac{y_i - \mu_i(\boldsymbol{\beta})}{\phi w_i} \right) f'(\mathbf{x}_i\boldsymbol{\beta}) x_{ji} \delta_i = 0,$$

$$j = 0, 1, \ldots, m, \tag{3.1}$$

which can be identified as the likelihood equation for $\boldsymbol{\beta}$ if sampling were done by simple random sampling with such a small sampling fraction that the sampled ys could be regarded as independent.

In the special case of independent observations, although in general non-i.i.d. due to the GLM form, it can be shown, under various regularity conditions (see e.g., Habermann (1977), Nordberg (1980), and Fahrmeir and Kaufmann (1985)), that the likelihood equation has – with probability tending to one as the sample size tends to infinity – one root being arbitrarily close to the true $\boldsymbol{\beta}$. (Multiple roots may exist but only one is arbitrarily close.) This root is an asymptotically efficient and asymptotically normally distributed estimator of $\boldsymbol{\beta}$.

However, it is obvious that this theory does not apply directly to the situation in Section 2 where data cannot, even as an approximation, a priori be regarded as independent observations. Nevertheless, as seen by Proposition 1, Equation (3.1) does, under certain general conditions, have a consistent and asymptotically normally distributed root.

Before proceeding we need some further notation. It should be emphasized that, in the sequel, when calculating probabilities, expectations, etc., we consider the total random variation induced by the GLM and the design.

Set $\mathbf{S}_n(\boldsymbol{\beta}, \mathbf{Y}) = (S_n^{(0)}(\boldsymbol{\beta}, \mathbf{Y}), \ldots, S_n^{(m)}(\boldsymbol{\beta}, \mathbf{Y}))'$

where, c.f. (3.1),

$$S_n^{(j)}(\boldsymbol{\beta}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i - \mu_i(\boldsymbol{\beta})}{\phi w_i} \right) f'(\mathbf{x}_i\boldsymbol{\beta}) x_{ji} \delta_i. \tag{3.2}$$

Let

$$D_n(\boldsymbol{\beta}, \mathbf{Y}) = -\left\{ \frac{\partial S_n^{(j)}}{\partial \beta_k}, \; j, k = 0, 1, \ldots, m \right\} \tag{3.3}$$

and

$$A_n(\boldsymbol{\beta}) = \mathrm{E}(D_n(\boldsymbol{\beta}, \mathbf{Y})). \tag{3.4}$$

We are now ready to formulate the following result on consistency and asymptotic normality.

Proposition 1: Let assumptions be as in Section 2, let $\boldsymbol{\beta}_0$ be the true parameter point and let $V_n(\boldsymbol{\beta}_0)$ be the variance-covariance matrix of $\sqrt{n}\mathbf{S}_n(\boldsymbol{\beta}_0, \mathbf{Y})$.

Suppose that

$$(\mathbf{S}_n(\boldsymbol{\beta}_0, \mathbf{Y}) - \mathrm{E}(\mathbf{S}_n(\boldsymbol{\beta}_0, \mathbf{Y})) \xrightarrow{P} 0$$

$$\text{as } n \to \infty. \tag{3.5}$$

$$\sqrt{n}V_n^{-1/2}(\beta_0)(\mathbf{S}_n(\beta_0, \mathbf{Y}) - \mathrm{E}(\mathbf{S}_n(\beta_0, \mathbf{Y}))$$

$$\to N(0, I) \quad \text{as } n \to \infty. \tag{3.6}$$

$$\lim_{n\to\infty} \sqrt{n}\mathrm{E}(\mathbf{S}_n(\beta_0, \mathbf{Y})) = 0. \tag{3.7}$$

If (3.5)–(3.7) as well as some additional regularity conditions (to be discussed later) are fulfilled then the following conclusions hold.

Equation (3.1) has – with probability tending to one as $n \to \infty$ – exactly one root $\hat{\beta}^{(n)}$ such that

$$|\hat{\beta}^{(n)} - \beta_0| \leqslant \delta \text{ for every } \delta > 0. \tag{3.8}$$

Furthermore,

$$\sqrt{n}V_n^{-1/2}(\beta_0)A_n(\beta_0)(\hat{\beta}^{(n)} - \beta_0)$$

$$\to N(0, I) \text{ as } n \to \infty. \tag{3.9}$$

Remark: Conditions (3.5) through (3.7) are vital for (3.8) and (3.9) while the additional regularity conditions mentioned in the proposition are of a more technical nature such as the invertibility of certain matrices, etc. A set of such regularity conditions is specified in the Appendix where a more precise version of Proposition 1 is proved.

Conditions (3.5) and (3.6) state that $\mathbf{S}_n(\beta_0, \mathbf{Y})$ obeys the law of large numbers and the central limit theorem. Sufficient conditions for (3.5) and (3.6) to hold will differ in appearance depending – among other things – on the nature of the sampling mechanism. There is a large literature on this subject which we will not try to cover here. We simply assume that (3.5) and (3.6) are satisfied. We will, however, take a closer look at (3.7) and also at $V_n(\beta_0)$. If $V_n(\beta_0)$ and $A_n(\beta_0)$ coincide then (3.9) takes the "classical" form, i.e., $\sqrt{n}A_n^{1/2}(\beta_0)(\hat{\beta}^{n)} - \beta_0)$ is asymptotically $N(0, I)$. Ordinary GLM inference can then be applied. We will derive conditions which are sufficient for (3.7) and for $V_n(\beta_0)$ and $A_n(\beta_0)$ to coincide. The essence of these conditions is that the mean and variance structure of the model should not be affected by the design variable.

Proposition 2: Let $\mathbf{z}$ be the design variable and consider the following conditions.

$$\mathrm{E}(Y_i - \mu_i(\beta_0)|\mathbf{z}) = \mathrm{E}(Y_i - \mu_i(\beta_0)) = 0,$$

$$i = 1, 2, \ldots, N \tag{3.10}$$

$$\mathrm{E}((Y_i - \mu_i(\beta_0))^2|\mathbf{z}) = \mathrm{E}(Y_i - \mu_i(\beta_0))^2,$$

$$i = 1, 2, \ldots, N \tag{3.11}$$

$$\mathrm{E}((Y_i - \mu_i(\beta_0))(Y_j - \mu_j(\beta_0))|\mathbf{z})$$

$$= \mathrm{E}(Y_i - \mu_i(\beta_0))(Y_j - \mu_j(\beta_0))$$

$$= 0, \quad i \neq j. \tag{3.12}$$

a. If (3.10) is fulfilled, then (3.7) holds.

b. If (3.10) through (3.12) are fulfilled, then $V_n(\beta_0) \equiv A_n(\beta_0)$ where entry (k, l) is

$$\frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{N}\left(\frac{\sigma_i(\beta_0)f'(\mathbf{x}_i\beta_0)}{\phi w_i}\right)^2 x_{ki}x_{li}\pi_i\right). \tag{3.13}$$

c. $\hat{A}_n^{(kl)}(\hat{\beta})$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{\sigma_i(\hat{\beta})f'(\mathbf{x}_i\hat{\beta})}{\hat{\phi}w_i}\right)^2 x_{ki}x_{li}\delta_i \tag{3.14}$$

is an asymptotically unbiased estimator of $A_n^{(kl)}(\beta_0)$ if $\hat{\beta}$ and $\hat{\phi}$ are consistent estimators of β and ϕ.

Remark 1: Notice that (3.14) is the variance estimator that one would use if ordinary GLM inference is applied to the sample.

Remark 2: A consistent ϕ estimator is required in (3.14). For many widely used

models, e.g., logit, probit, and log-linear Poisson, this problem can be ignored since $\phi = 1$. For normal regression ϕ equals the residual variance. Although there might be some GLMs for which ϕ estimation must be treated with care we will not discuss this problem any further in the present paper.

Proof:
By applying $E(\cdot) = EE(\cdot|\mathbf{t})$ to (3.2) and (2.9) it is seen that (3.7) is equivalent to

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} E\left(\sum_{i=1}^{N} \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times f'(\mathbf{x}_i\beta_0)x_{ji}\pi_i \right) = 0, \quad j = 0, 1, \ldots, m. \tag{3.15}$$

Since π is a function of $\mathbf{z}$, (3.10) implies (3.15) and hence (3.7).
By (3.2)

$$V_n(\beta_0) = \frac{1}{n}\left\{ \mathrm{Cov}\left(\sum_i \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times f'(\mathbf{x}_i\beta_0)x_{ki}\delta_i, \sum_j \left(\frac{Y_j - \mu_j(\beta_0)}{\phi w_j} \right) \times f'(\mathbf{x}_j\beta_0)x_{lj}\delta_j \right); k, l = 0, \ldots, m \right\}. \tag{3.16}$$

Entry (k, l) of $V_n(\beta_0)$ can thus be expressed as follows

$$V_n^{(kl)}(\beta_0) = \frac{1}{n}\left(E\left(\sum_i \sum_j \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times \left(\frac{Y_j - \mu_j(\beta_0)}{\phi w_j} \right) \times f'(\mathbf{x}_i\beta_0)f'(\mathbf{x}_j\beta_0)x_{ki}x_{lj}\delta_i\delta_j \right) - E\left(\sum_i \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times f'(\mathbf{x}_i\beta_0)x_{ki}\delta_i \right) E\left(\sum_j \left(\frac{Y_j - \mu_j(\beta_0)}{\phi w_j} \right) \times f'(\mathbf{x}_j\beta_0)x_{lj}\delta_j \right) \right). \tag{3.17}$$

By (2.9) and (2.10) relation (3.17) takes the form

$$V_n^{(kl)}(\beta_0) = \frac{1}{n} E\left(\sum_i \frac{(Y_i - \mu_i(\beta_0))^2}{\phi^2 w_i^2} \times (f'(\mathbf{x}_i\beta_0))^2 x_{ki}x_{li}\pi_i \right) + \frac{1}{n} E\left(\sum_{i\neq j}\sum \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times \left(\frac{Y_j - \mu_j(\beta_0)}{\phi w_j} \right) \times f'(\mathbf{x}_i\beta_0)f'(\mathbf{x}_j\beta_0)x_{ki}x_{lj}\pi_{ij} \right) - \frac{1}{n}\left(E\left(\sum_i \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times f'(\mathbf{x}_i\beta_0)x_{ki}\pi_i \right) \times E\left(\sum_j \left(\frac{Y_j - \mu_j(\beta_0)}{\phi w_j} \right) \times f'(\mathbf{x}_j\beta_0)x_{lj}\pi_j \right) \right). \tag{3.18}$$

By differentiating (3.2) and noting (3.3), (3.4), (2.3), (2.4) and (2.9) it is seen that entry (k, l) of $A_n(\beta_0)$ is

$$A_n^{(kl)}(\beta_0) = -\frac{1}{n} E\left(\sum_i \left(\frac{Y_i - \mu_i(\beta_0)}{\phi w_i} \right) \times f''(\mathbf{x}_i\beta_0)x_{ki}x_{li}\pi_i \right) + \frac{1}{n} E\left(\sum_i \left(\frac{\sigma_i(\beta_0)f'(\mathbf{x}_i\beta_0)}{\phi w_i} \right)^2 x_{ki}x_{li}\pi_i \right). \tag{3.19}$$

Now, (3.10) through (3.12) imply (3.13) since the last term of (3.19) equals the first one of (3.18) while all other terms of (3.18) and (3.19) disappear. The claim in (3.14) holds if σ_i and f' are sufficiently regular functions.

Conditions (3.10) through (3.12) can be checked by various types of residual plots. Prevalent methods for residual analysis with special reference to GLMs are reviewed in McCullagh and Nelder (1983). There are also other ways to check the mean and variance structure here. It follows from the proof that (3.10) can be replaced by the slightly weaker (3.15) as consistency criterion. Notice that (3.15) can be identified as the normal equations for the parameters corresponding to π, $\pi x_1, \ldots, \pi x_m$ if these variables were included in the explanatory vector. We are then led to the following procedure, previously suggested in a special case by Du Mouchel and Duncan (1983).

Extend the explanatory vector $\mathbf{x}_i$ by bringing in π_i and its interactions $\pi_i x_{ji}$, $j = 1, \ldots, m$, together with $\mathbf{x}_i$ as additional variables. ($x_{0i} \equiv 1$ for convenience). Then test the model specified by (2.1) and (2.2) against this extended model using regular GLM technique. If the extended model does not significantly improve the fit, then this suggests that (3.15) is reasonably satisfied. On the other hand, if the extended model significantly improves the fit then this suggests that there is useful information in the design which may be used to improve model (2.1) and (2.2). Although one would usually not want to keep π itself as an explanatory variable there may be other variables associated with $\mathbf{z}$, which are highly correlated with π, and which make more sense – from a subject matter point of view – as explanatory variables. By bringing such variables into the model building process it is possible to remove design bias and model bias simultaneously. This will be illustrated by the simulation study in Section 4.

The assumed variance expression $\sigma_i^2(\cdot)$ can be checked by estimating the first term of (3.18) and compare it to (3.14) by checking the following relation. Notice that ϕ can be ignored here.

$$\frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i - \mu_i(\hat{\beta})}{w_i}\right)^2 (f'(\mathbf{x}_i\hat{\beta}))^2 x_{ki}x_{li}\delta_i$$

$$\approx \frac{1}{n}\sum_{i=1}^{N}\left(\frac{\sigma_i(\hat{\beta})}{w_i}\right)^2 (f'(\mathbf{x}_i\hat{\beta}))^2 x_{ki}x_{li}\delta_i. \tag{3.20}$$

Let $\tilde{V}_n$ be a suitable design-based variance estimate for $\sqrt{n}\mathbf{S}_n$. A rough way of checking the variance condition (3.13) is to compare $\tilde{V}_n$ with $\hat{A}_n$.

We end this section by comparing (3.10) through (3.12) to the more general ignorability conditions discussed by Rubin (1976, 1987), Little (1982), and Little and Rubin (1987). The essence of ignorability is that the sampling distribution of δ, conditioned on $\mathbf{t}$, must be independent of ys which are unobserved due to sampling, and its parameters must not depend on β.

One advantage of conditions (3.10) through (3.12) compared to ignorability is that they involve means and variances only instead of full distributions.

4. Unweighted Versus Weighted Estimation

4.1. General remarks

As shown in Section 3 the unweighted $\hat{\beta}$ is, under certain conditions, consistent, and ordinary GLM methods do apply. However, these conditions are not always fulfilled and $\hat{\beta}$ may then be inconsistent.

One alternative to $\hat{\beta}$ is the following

Horvitz–Thompson weighted estimator. Consider the function

$$F(\beta, \phi) = \sum_{i=1}^{N} \frac{1}{\pi_i} \left(\left(\frac{Y_i \theta_i - b(\theta_i)}{\phi w_i} \right) + c_i(Y_i, \phi) \right) \delta_i, \qquad (4.1)$$

where θ_i depends on β through (2.2).

F can be interpreted as a Horvitz–Thompson weighted estimator of the log-likelihood for the complete data $(Y_1, Y_2, \ldots, Y_N)$. By solving $\delta F/\delta\beta = 0$ for β we get the Horvitz–Thompson weighted β estimator $\tilde{\beta}^{(n)}$.

The estimator $\tilde{\beta}$ is design consistent under general conditions. However, as demonstrated by the simulations ahead, $\tilde{\beta}$ can be misleading in spite of design consistency if the superpopulation model is not properly specified.

It can be shown that $\hat{\beta}$ – when valid – is asymptotically more efficient than $\tilde{\beta}$. The efficiency gain can be substantial as shown by the simulation study ahead. This study will also illuminate the following point. If $\hat{\beta}$ is not consistent then it is sometimes possible to find one or more new variables, highly correlated with the inclusion probabilities π, to incorporate into the model. This may remove design bias and model bias simultaneously and be more efficient than weighting which removes design bias only.

4.2. A simulation study

4.2.1. Background to the choice of superpopulation model

The choice of the superpopulation model used in the simulations was inspired by a study of structural changes among Swedish milk producing farms, c.f. Nordberg (1985). The aim of this study was to determine factors which affect the tendency among farmers to give up milk production. We summarize only those parts of the study which are relevant in the present context.

Each farm which had at least one and at most nine milk cows in 1983 – 12 195 farms – was checked to see whether it still had milk cows in 1984. In the case it did have cows, the variable y was set to zero, otherwise $y = 1$. A logit analysis was then performed with y as the dependent variable. This analysis was based on the full population so there were no sampling errors. The following four explanatory factors turned out to be the most "relevant."

- The size depicted by (S). $S = 0$ if the number of cows was between 4 and 9 in 1983 while $S = 1$ otherwise.
- The type (T) of the farm where $T = 0$ if milk production is the major production branch according to the Swedish typology system while $T = 1$ otherwise.
- Age (A) of the farmer – classified as [–49], [50–59], [60–] years and represented by A_2 and A_3 where $A_2 = 1$ if age is 50–59 and zero otherwise while $A_3 = 1$ if age is 60 and over and zero otherwise.
- The region (R) being either the most productive farming area in the south and the middle of the country or the rest of the south and the middle or the north. R is represented by R_2 and R_3 where $R_2 = 1$ if the farm is in the second region and zero otherwise while $R_3 = 1$ in the third region, zero otherwise.

The following model $P(Y = 1) =$

Table 1. $P(Y = 1)$ by (4.2) and number of observations (N) in the population for different combinations of Age, Region, Size, and Type

A (Age)	R (Region)	S (Size)							
		$S = 0$ (4–9 cows)				$S = 1$ (1–3 cows)			
		T (Type)				T (Type)			
		$T = 0$		$T = 1$		$T = 0$		$T = 1$	
		$P(Y = 1)$	N	$P(Y = 1)$	N	$P(Y = 1)$	N	$P(Y = 1)$	N
[–49]	1	0.08	369	0.18	114	0.29	12	0.53	108
1	2	0.08	962	0.18	40	0.23	57	0.45	182
	3	0.08	923	0.18	20	0.20	159	0.40	109
[50–59]	1	0.06	553	0.14	120	0.23	36	0.45	82
2	2	0.06	1 092	0.14	30	0.18	103	0.38	148
	3	0.06	795	0.14	10	0.16	197	0.33	66
[60–]	1	0.16	854	0.33	130	0.29	137	0.53	310
3	2	0.16	1 866	0.33	49	0.23	410	0.45	529
	3	0.16	1 009	0.33	8	0.20	429	0.40	177
			8 423		521		1 540		1 711

$\exp(H)/(1 + \exp(H))$, where

$$H = -2.5 + 1.6S - 0.3A_2 + 0.8A_3 - 0.8A_3 \times S + 1.0T - 0.3R_2 \times S - 0.5R_3 \times S \quad (4.2)$$

was found to fit data well, c.f. Nordberg (1985).

Table 1 presents $P(Y = 1)$ according to (4.2) as well as the number of observations, N, in the population for the different combinations of the explanatory factors.

4.2.2. Design of the simulation experiment

i. *Superpopulation mechanism*: A population of 12 195 elements, divided into 36 groups was created. Each group corresponds to one of the $3 \times 3 \times 2 \times 2$ cells of Table 1. The number of elements in a particular group equals the value of N in the corresponding cell of Table 1. In each group (cell) N_k independent 0–1 random variables $Y_1, Y_2, \ldots, Y_{N_k}$ (N_k being the N value of the cell) were generated by the model (4.2).

ii. *Sampling mechanism*: The population generated in (i) was then grouped into four strata corresponding to the combinations of (S, T), (0, 0), (0, 1), (1, 0) and (1, 1), i.e., the main columns of Table 1. In each stratum a sample was drawn by simple random sampling without replacement. The number of observations drawn in each stratum was 840, 521, 920, and 720 respectively. These correspond to inclusion probabilities 10, 100, 60, and 42% respectively. Notice that the design vector (S, T) here is a function of the true explanatory vector. The reason is that we want to demonstrate the effects of incorporating versus deleting from the model such design variables which carry important information on Y.

iii. *Parameter estimation*: A logit model was fitted to the data generated through (i) and (ii). Several choices of explanatory vector were considered and are discussed later. The unweighted estimator $\hat{\beta}$ and its classical variance-covariance matrix $1/n\ \hat{A}^{-1}(\hat{\beta})$ were evaluated. In particular the vector of estimated standard errors $\hat{\sigma}(\hat{\beta})$ (square roots of the diagonal elements of $1/n\ \hat{A}^{-1}$) was calculated. Finally the Horvitz–Thompson weighted $\tilde{\beta}$ and $\hat{\beta}_{pop}$, the latter being the MLE based on the full population, were evaluated.

iv. *Replications*: The above steps (i) through (iii) were repeated 500 times. The means – over the 500 replications – of $\hat{\beta}$, $\tilde{\beta}$, $\hat{\beta}_{pop}$, and $\hat{\sigma}(\hat{\beta})$, denoted MEAN ($\hat{\beta}$), MEAN ($\tilde{\beta}$), etc. were calculated. In addition, the standard deviations (over the 500 replications) denoted STD ($\hat{\beta}$), etc. were calculated.

4.2.3. Results

Suppose that the chosen explanatory vector includes Region and Age only. The results of the simulations in this case are presented in Table 2. A comparison of the β estimators of Table 2 to the true β (see (4.2)) shows that all three estimators, including $\hat{\beta}_{pop}$, are biased. The main reason, of course, is that Size and Type, which both have very strong effects on Y, are not included in the model.

The unweighted $\hat{\beta}$ has – in addition to this model bias – also a strong design bias, as seen by comparing MEAN ($\hat{\beta}$) to MEAN ($\hat{\beta}_{pop}$). Notice that $\tilde{\beta}$ lacks design bias. It

Table 2. Explanatory vector contains R and A only

Explanatory vector	MEAN $(\hat{\beta}_{pop})$	MEAN $(\hat{\beta})$	MEAN $(\tilde{\beta})$	MEAN $\hat{\sigma}$ $(\hat{\beta})$	STD $(\hat{\beta})$	STD $(\tilde{\beta})$	$\left[\frac{\text{STD}(\tilde{\beta})}{\text{STD}(\hat{\beta})}\right]^2$
Intcpt	−1.57	−1.06	−1.58	0.11	0.11	0.15	1.99
R_2	−0.22	−0.18	−0.22	0.10	0.10	0.14	1.44
R_3	−0.38	−0.46	−0.38	0.11	0.11	0.16	2.10
A_2	−0.38	−0.38	−0.39	0.14	0.13	0.18	1.74
A_3	0.54	0.35	0.54	0.11	0.11	0.14	1.63

Table 3. Explanatory vector contains the most significant variables only

Explanatory vector	MEAN $(\hat{\beta}_{pop})$	MEAN $(\hat{\beta})$	MEAN $(\tilde{\beta})$	MEAN $\hat{\sigma}$ $(\hat{\beta})$	STD $(\hat{\beta})$	STD $(\tilde{\beta})$	$\left[\frac{\text{STD}(\tilde{\beta})}{\text{STD}(\hat{\beta})}\right]^2$
Intcpt	−2.66	−2.70	−2.67	0.13	0.14	0.17	1.57
S	1.25	1.28	1.26	0.15	0.15	0.18	1.43
T	1.09	1.08	1.08	0.09	0.09	0.09	1.00
A_3	0.96	0.96	0.95	0.16	0.17	0.22	1.64
$S \times A_3$	−0.78	−0.77	−0.77	0.19	0.21	0.25	1.42

Table 4. Explanatory vector contains all main effects and S and T interactions

Explanatory vector	MEAN ($\hat{\beta}_{pop}$)	MEAN ($\hat{\beta}$)	MEAN ($\tilde{\beta}$)	MEAN $\hat{\sigma}$ ($\hat{\beta}$)	STD ($\hat{\beta}$)	STD ($\tilde{\beta}$)	$\left[\frac{\text{STD}(\tilde{\beta})}{\text{STD}(\hat{\beta})}\right]^2$
Intcpt	−2.51	−2.53	−2.55	0.29	0.28	0.33	1.37
S	1.59	1.60	1.62	0.31	0.31	0.34	1.18
T	1.01	1.01	1.03	0.29	0.29	0.30	1.11
R_2	0.01	−0.00	0.02	0.24	0.24	0.30	1.54
R_3	0.00	−0.00	0.02	0.27	0.28	0.32	1.36
A_2	−0.30	−0.31	−0.33	0.30	0.31	0.36	1.36
A_3	0.81	0.81	0.81	0.24	0.23	0.27	1.22
$S \times T$	0.01	−0.01	−0.02	0.23	0.21	0.22	1.11
$S \times R_2$	−0.30	−0.29	−0.30	0.24	0.24	0.28	1.33
$S \times R_3$	−0.50	−0.49	−0.51	0.29	0.30	0.33	1.22
$S \times A_2$	0.00	0.02	0.03	0.30	0.30	0.35	1.31
$S \times A_3$	−0.79	−0.80	−0.80	0.24	0.25	0.27	1.21
$T \times R_2$	−0.01	−0.00	−0.01	0.24	0.24	0.24	1.03
$T \times R_3$	0.00	−0.00	−0.01	0.28	0.28	0.28	1.03
$T \times A_2$	−0.01	0.01	0.01	0.30	0.32	0.32	1.01
$T \times A_3$	−0.02	−0.01	−0.00	0.24	0.24	0.24	1.00

might then be argued that $\tilde{\beta}$ is preferable to $\hat{\beta}$ in the present situation. However, it is possible to remove the design bias and much of the model bias simultaneously as the following argument shows.

The design bias of $\hat{\beta}$ indicates that π carries relevant information about y which has not been accounted for by R and A. The stratification discussed earlier – Section 4.2.2. (ii) – implies that $\pi = \gamma_0 + \gamma_1 S + \gamma_2 T + \gamma_3 ST$ for some $\gamma_0, \gamma_1, \gamma_2, \gamma_3$. Now if S and T are included in the model building process we can expect the model bias (due to the missing variables S and T) and the design bias (due to varying π) to disappear simultaneously. Tables 3 and 4 support this conclusion. Table 4 presents the case where the explanatory vector contains all main effects of R, A, S, and T as well as all first order interactions involving S and T. In Table 3 the explanatory vector contains only the most significant variables, i.e., S, T, A_3 and $S \times A_3$.

It is also seen from Tables 3 and 4 that MEAN $(\hat{\sigma}) \approx$ STD $(\hat{\beta})$ which means that the classical variance estimator is approximately unbiased. $\hat{\beta}$ is considerably more efficient than $\tilde{\beta}$.

5. Nonresponse

Next we specify the response mechanism. To each element in the sample we assign a response probability $P_i(\mathbf{u}_i, \alpha)$, where $\mathbf{u}_i$, $i = 1, 2, \ldots, N$, is a set of – possibly multidimensional – explanatory variables and α is a parameter. As a response model $P_i(\mathbf{u}_i, \alpha)$ we may use, e.g., a logistic model or a response homogenity groups model to mention some possibilities, but we will not make any specific assumptions about the form of $P_i(\mathbf{u}_i, \alpha)$ in this paper, except that it can be expressed by a parametric model.

Let for $i = 1, 2, \ldots, N$

$$r_i = \begin{cases} 1 \text{ if } \delta_i = 1 \text{ and element } i \text{ responds} \\ 0 \text{ otherwise.} \end{cases} \tag{5.1}$$

Extend $\mathbf{t}_i$ (c.f. (2.8)) by the explanatory vector $\mathbf{u}_i$ of the response model. Hence $\mathbf{t}_i = (Y_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)$ and $\mathbf{t} = (\mathbf{Y}, \mathbf{x}, \mathbf{z}, \mathbf{u})'$. Although it will be reasonable in many applications to assume that $\mathbf{u}$ can be expressed by a known function of $(\mathbf{Y}, \mathbf{x}, \mathbf{z})$ we will also cover cases where $\mathbf{u}$ cannot – without random error – be expressed through $(\mathbf{Y}, \mathbf{x}, \mathbf{z})$. We will assume, however, that $\mathbf{u}$ holds all the relevant information about the response pattern. We then have

$$\begin{aligned} \mathbf{P}(r_i = 1 | \delta_i = 1, \mathbf{t}) &= 1 - \mathbf{P}(r_i = 0 | \delta_i = 1, \mathbf{t}) \\ &= P_i(\mathbf{u}_i, \alpha). \end{aligned} \tag{5.2}$$

Set

$$P_{ij}(\mathbf{u}, \alpha) = P(r_i = 1, r_j = 1 | \delta_i = 1, \delta_j = 1, \mathbf{t}). \tag{5.3}$$

In many applications one would assume conditionally independent responses, i.e., $P_{ij} = P_i P_j$. However, this will not be necessary in order to reach our conclusions.

It is now straightforward to generalize (2.9) and (2.10) as follows.

$$\mathrm{E}(r_i | \mathbf{t}) = \pi_i(\mathbf{z}) \cdot P_i(\mathbf{u}_i, \alpha), \tag{5.4}$$

$$\mathrm{E}(r_i r_j | \mathbf{t}) = \pi_{ij}(\mathbf{z}) \cdot P_{ij}(\mathbf{u}, \alpha), \quad i \neq j. \tag{5.5}$$

The results of Sections 3 and 4 were derived under the two-step specification in Section 2 of the superpopulation mechanism and the sampling design. Now, with δ_i replaced by r_i, π_i by $\pi_i P_i$, π_{ij} by $\pi_{ij} P_{ij}$ and $\mathbf{z}$ by $(\mathbf{z}, \mathbf{u})$ it is straightforward to show that the same results hold also under the full three-step specification which includes the response mechanism as introduced earlier.

Appendix

Let assumptions and notation be as in Sections 1 and 2. Notice that $E(\cdot)$ and $P(\cdot)$ below are supposed to take account of all the random variation induced by the superpopulation and the design.

Theorem: Let β_0 be the true parameter point and consider the following conditions.

i. $\mathbf{S}_n(\beta_0, \mathbf{Y}) \xrightarrow{P} \mathbf{O}$ as $n \to \infty$.

ii. There is a $\lambda > 0$ and an $n_0 > 0$ such that the smallest eigenvalue of $A_n(\beta_0) \geqslant \lambda > 0$ for $n > n_0$.

iii. $|D_n(\beta_0, \mathbf{Y}) - A_n(\beta_0)| \xrightarrow{P} \mathbf{O}$ (component wise convergence) as $n \to \infty$.

iv. For some $\delta_0 > 0$ there is for every $\varepsilon > 0$ a $C_\varepsilon < \infty$ and an $n_\varepsilon < \infty$ such that

$$P\left(\left|\frac{\partial^2 S_n^{(j)}}{\partial\beta_k \partial\beta_l}\right| \leqslant C_\varepsilon \text{ for every } |\beta - \beta_0| \leqslant \delta_0\right) \geqslant 1 - \varepsilon, \quad n \geqslant n_\varepsilon \quad j, k, l = 0, 1, \ldots, m.$$

v. $\sqrt{n} V_n^{-1/2}(\beta_0)\mathbf{S}_n(\beta_0, \mathbf{Y}) \to N(0, I)$ as $n \to \infty$.

vi. All components of $V_n(\beta_0)$ are uniformly bounded for $n > n_0$.

vii. There is a $\lambda'' > 0$ and an $n_0 > 0$ such that the smallest eigenvalue of $V_n(\beta_0) \geqslant \lambda'' > 0$, $n > n_0$.

a. If conditions (i) through (iv) are fulfilled then, with probability tending to one as the sample size n tends to infinity, equation (3.1) has exactly one root $\hat{\beta}^{(n)}$ such that $|\hat{\beta}^{(n)} - \beta_0| \leqslant \delta$ for every $\delta > 0$.

b. If conditions (i) through (vii) are fulfilled then

$$\sqrt{n} V_n^{-1/2}(\beta_0) A_n(\beta_0)(\hat{\beta}^{(n)} - \beta_0) \to N(0, I) \text{ as } n \to \infty.$$

This theorem and its proof are quite similar to Theorem 1 in Nordberg (1980). Since the latter does not apply directly to the situation treated in this paper, various modifications have been made to cover the present situation. The main ingredient of the proof is a version of the implicit function theorem. See also Foutz (1977) for similar ideas. Foutz treats a broad class of models but confines himself to i.i.d. observations. We will use the following version of the implicit function theorem (and also prove it for completeness).

Lemma: Let $\mathbf{g}_n(\mathbf{u}) = (g_n^{(1)}(u_1 \ldots u_m), \ldots g_n^{(m)}(u_1 \ldots u_m)$, $n = 1, 2, \ldots$ be a sequence of three times differentiable functions from R^m to R^m.

$$\text{Set } H_n(\mathbf{u}) = \left\{\frac{\partial g_n^{(j)}(\mathbf{u})}{\partial u_k}, \quad j, k = 1, \ldots m\right\} \quad n = 1, 2 \ldots$$

Let $\mathbf{a} \in R^m$ and suppose that there is a $\lambda > 0$ and an $n_0 > 0$ such that $|H_n(\mathbf{a})\mathbf{x}| \geqslant \lambda|\mathbf{x}|$ for any $\mathbf{x} \in R^m$, $n > n_0$. (A1)

Furthermore, suppose that for some $\delta_0 > 0$ and some $G < \infty$

$$\left|\frac{\partial^2 g_n^{(j)}(\mathbf{u})}{\partial u_k \partial u_l}\right| \leqslant G < \infty \text{ for every } \mathbf{u} \in d(\mathbf{a}, \delta_0), \quad n > n_0$$
$$\text{where } d(\mathbf{a}, \delta_0) = \{\mathbf{u}:|\mathbf{u} - \mathbf{a}| \leqslant \delta_0\}. \tag{A2}$$

Then there is a $\delta_1 > 0$ such that the restriction of $\mathbf{g}_n(\mathbf{u})$, $n > n_0$, to $d(\mathbf{a}, \delta_1)$ is one to one. Furthermore, if $0 < \delta < \delta_1$ and $\mathbf{z} \in d(\mathbf{g}_n(\mathbf{a}), \lambda\delta/2)$ then there is exactly one $\mathbf{u} \in d(\mathbf{a}, \delta)$ such that $\mathbf{g}_n(\mathbf{u}) = \mathbf{z}$, $n > n_0$.

Proof of lemma: Let $\mathbf{u}'$ and $\mathbf{u}'' \in d(\mathbf{a}, \delta_0)$ and suppose that $\mathbf{u}' \neq \mathbf{u}''$. By (A2) we have for $n > n_0$

$$\mathbf{g}_n(\mathbf{u}') - \mathbf{g}_n(\mathbf{u}'') = H_n(\mathbf{a})(\mathbf{u}' - \mathbf{u}'') + (H_n(\mathbf{u}'') - H_n(\mathbf{a}))(\mathbf{u}' - \mathbf{u}'') + \mathbf{R} \tag{A3}$$

where for some $C < \infty$

$$|\mathbf{R}| \leqslant C|\mathbf{u}' - \mathbf{u}''|^2.$$

But due to (A2) the following relation holds for some $C' < \infty$

$$|(H_n(\mathbf{u}'') - H_n(\mathbf{a}))(\mathbf{u}' - \mathbf{u}'')| \leqslant C'|\mathbf{u}' - \mathbf{u}''||\mathbf{u}'' - \mathbf{a}| \tag{A4}$$

and thus by (A1)

$$|\mathbf{g}_n(\mathbf{u}') - \mathbf{g}_n(\mathbf{u}'')| \geqslant \lambda|\mathbf{u}' - \mathbf{u}''| - C'|\mathbf{u}' - \mathbf{u}''||\mathbf{u}'' - \mathbf{a}| - C|\mathbf{u}' - \mathbf{u}''|^2. \tag{A5}$$

$$\text{Set } \delta_1 = \min(\delta_0, \lambda/(2C' + 4C)). \tag{A6}$$

Then the following relation holds as soon as $\mathbf{u}'$ and $\mathbf{u}'' \in d(\mathbf{a}, \delta_1)$:

$$|\mathbf{g}_n(\mathbf{u}') - \mathbf{g}_n(\mathbf{u}'')| \geqslant |\mathbf{u}' - \mathbf{u}''|\left(\lambda - \frac{\lambda C'}{2C' + 4C} - \frac{2\lambda C}{2C' + 4C}\right) = \frac{\lambda}{2}|\mathbf{u}' - \mathbf{u}''|. \tag{A7}$$

We have thus proved that if $\mathbf{u}' \neq \mathbf{u}''$ then $g_n(\mathbf{u}') \neq \mathbf{g}_n(\mathbf{u}'')$, which means that if $\mathbf{z} \in d(\mathbf{g}_n(\mathbf{a}), \lambda\delta/2)$ where $0 < \delta < \delta_1$ then there is at most one $\mathbf{u} \in d(\mathbf{a}, \delta)$ such that $\mathbf{g}_n(\mathbf{u}) = \mathbf{z}$.

We will now prove that if $\mathbf{z} \in d(\mathbf{g}_n(\mathbf{a}), \lambda\delta/2)$ where $0 < \delta < \delta_1$ then there is exactly one $\mathbf{u} \in d(\mathbf{a}, \delta)$ such that $\mathbf{g}_n(\mathbf{u}) = \mathbf{z}$. Consider the function $\mathbf{h}_n(\mathbf{u}) = |\mathbf{g}_n(\mathbf{u}) - \mathbf{z}|^2$ for $\mathbf{u} \in d(\mathbf{a}, \delta)$. Since $\mathbf{h}_n(\mathbf{u})$ is defined on a closed set it has a minimum at $\bar{\mathbf{u}}$, say, and $\bar{\mathbf{u}}$ satisfies the equation

$$H_n(\bar{\mathbf{u}})(\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z}) = 0.$$

Thus

$$H_n(\mathbf{a})(\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z}) = -(H_n(\bar{\mathbf{u}}) - H_n(\mathbf{a}))(\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z}).$$

By (A4)

$$|(H_n(\bar{\mathbf{u}}) - H_n(\mathbf{a}))(\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z})| \leqslant C'\delta|\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z}|$$

where, by A(6), $C'\delta < \dfrac{\lambda C'}{2C' + 4C} < \dfrac{\lambda}{2}$.

Therefore

$$\left|H_n(\mathbf{a})(\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z})\right| < \frac{\lambda}{2}\left|\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z}\right|.$$

But due to (A1)

$$|H_n(\mathbf{a})(\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z})| \geqslant \lambda|\mathbf{g}_n(\bar{\mathbf{u}}) - \mathbf{z}|.$$

We have then arrived at a contradiction unless $\mathbf{g}_n(\bar{\mathbf{u}}) = \mathbf{z}$. This completes the proof of the existence and uniqueness of a $\mathbf{u} \in d(\mathbf{a}, \delta)$ such that $\mathbf{g}_n(\mathbf{u}) \in d(\mathbf{g}_n(\mathbf{a}), \lambda\delta/2)$.

Proof of theorem: Let $-\mathbf{S}_n(\beta, \mathbf{Y})$ correspond to $\mathbf{g}_n(\mathbf{u})$, β and β_0 correspond to $\mathbf{u}$ and $\mathbf{a}$ respectively.

Suppose that the smallest eigenvalue of $D_n(\beta_0, \mathbf{Y}) \geqslant \lambda' > 0, n > n_0$ (A8)

$$\left|\frac{\partial^2 S_n^{(j)}}{\partial\beta_k\,\partial\beta_l}\right| \leqslant G \text{ for every } \beta \in d(\beta_0, \delta_0),\ n > n_0,\ j, k, l = 0, \ldots m \tag{A9}$$

$$\mathbf{O} \in d(-\mathbf{S}_n(\beta_0, \mathbf{Y}), \lambda'\delta/2) \tag{A10}$$

where δ_0, δ and n_0 are defined in the proof of the lemma.

If (A8) through (A10) hold then, due to the lemma, there is exactly one root $\hat{\beta}^{(n)} \in d(\beta_0, \delta)$ of equation (3.1). Now (ii) and (iii) imply that (A8) is true with probability tending to one as $n \to \infty$ and the same conclusion about (A9) follows from (iv). Finally, the probability that (A10) is true converges to one as $n \to \infty$ by condition (i). Thus (A8) through (A10) hold true simultaneously with probability tending to one as $n \to \infty$ and this implies conclusion (a) of the theorem.

Consider $\hat{\beta}^{(n)}$ appearing in (a). It is seen from (A7) that for some $C' < \infty$

$$\lim_{n\to\infty} P(|\hat{\beta}^{(n)} - \beta_0| \leqslant C'|\mathbf{S}_n(\beta_0, \mathbf{Y}|) = 1$$

and this relation combined with conditions (v) and (vi) yields $\geqslant 1 - \varepsilon$ for every $\varepsilon > 0$.

$$\lim_{n\to\infty} P(\sqrt{n}|\hat{\beta}^{(n)} - \beta_0| \leqslant C''_\varepsilon < \infty) \tag{A11}$$

By conclusion (a) of the theorem and (A11) we have

$$\sqrt{n}|\hat{\beta}^{(n)} - \beta_0|^2 \xrightarrow{P} \mathbf{0} \text{ as } n \to \infty \tag{A12}$$

Taylor-expansion of $\mathbf{S}_n(\beta, \mathbf{Y})$ around β_0 (note that $\mathbf{S}_n(\hat{\beta}^{(n)}, \mathbf{Y}) = \mathbf{0}$) yields

$$\mathbf{S}_n(\beta_0, \mathbf{Y}) = A_n(\beta_0)(\hat{\beta}^{(n)} - \beta_0) + (D_n(\beta_0, \mathbf{Y}) - A_n(\beta_0))(\hat{\beta}^{(n)} - \beta_0) + \mathbf{R}_n \tag{A13}$$

where

$$\sqrt{n}\,|\mathbf{R}_n| \xrightarrow{P} \mathbf{0} \text{ as } n \to \infty. \tag{A14}$$

Relation (A14) follows from condition (iv) and (A12).

Condition (iii) and (A11) yields

$$\sqrt{n}|(D_n(\beta_0, \mathbf{Y}) - A_n(\beta_0)(\hat{\beta}^{(n)} - \beta_0)| \xrightarrow{P} \mathbf{0} \text{ as } n \to \infty. \tag{A15}$$

By condition (v)

$$\sqrt{n}\, V_n^{-1/2}(\beta_0)\mathbf{S}_n(\beta_0, \mathbf{Y}) \to N(0, I).$$

This relation, condition (vii), (A14) and (A15) imply that

$$\sqrt{n}\, V_n^{-1/2}(\beta_0)A_n(\beta_0)(\hat{\beta}^{(n)} - \beta_0) \to N(0, I),$$

which completes the proof of the theorem.

6. References

Binder, D.A. (1983): On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review, 51, pp. 279–292.

Du Mouchel, W.H. and Duncan, G.J. (1983): Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, pp. 535–543.

Fahrmeir, L. and Kaufmann, M. (1985): Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. Annals of Statistics, 13, pp. 342–368.

Fay, R.B. (1986): Causal Models for Patterns of Nonresponse. Journal of the American Statistical Association, 81, pp. 354–365.

Foutz, R.V. (1977): On the Unique Consistent Solution to the Likelihood Equations. Journal of the American Statistical Association, 72, pp. 147–148.

Habermann, S.J. (1977): Maximum Likelihood Estimates in Exponential Response Models. Annals of Statistics, 5, pp. 815–841.

Little, R.J.A. (1982): Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association, 77, pp. 237–250.

Little R.J.A. and Rubin, D.B. (1987): Statistical Analysis with Missing Data. Wiley, New York.

McCullagh, P. and Nelder, J.A. (1983): Generalized Linear Models. Chapman and Hall, London.

Nathan, G. and Holt, D. (1980): The Effect of Survey Design on Regression Analysis. Journal of the Royal Statistical Society, Series B, 42, pp. 377–386.

Nordberg, L. (1980): Asymptotic Normality of Maximum Likelihood Estimators Based on Independent, Unequally Distributed Observations in Exponential Family Models. Scandinavian Journal of Statistics, 7, pp. 27–32.

Nordberg, L. (1985): Analys av avgångar från mjölkproduktion. Memo, Statistics Sweden. (In Swedish.)

Rubin, D.B. (1976): Inference and Missing Data. Biometrika, 63, pp. 581–592.

Rubin, D.B. (1987): Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Scott, A.J. (1987): Generalized Linear Models with Survey Data. Proceedings of the Second International Tampere Conference in Statistics.

Smith, T.M.F. (1981): Regression Analysis for Complex Surveys. In Current Topics in Survey Sampling. Ed. Krewski, D., Platek, R. and Rao, J.N.K. Academic Press, New York.

Ten Cate, A. (1986): Regression Analysis Using Survey Data with Endogenous Design. Survey Methodology, 12, pp. 121–138.

Received June 1988
Revised July 1989