Generalized Linear Modeling of Sample Survey Data

Lennart Nordberg

R & D Report
Statistics Sweden
Research - Methods - Development
1988:8

Från trycket	Juli 1988
Producent	Statistiska centralbyrån, Utvecklingsavdelningen
Ansvarig utgivare	Staffan Wahlström
Förfrågningar	Lennart Nordberg, tel. 019-176012

ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1988

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1988:8. Generalized linear modeling of sample survey data / Lennart Nordberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1988-X101OP8808

GENERALIZED LINEAR MODELING OF SAMPLE SURVEY DATA

by

Lennart Nordberg

ABSTRACT

The theme of this paper is regression analysis - extended to Generalized linear models (GLIMs) - of sample survey data, the data being obtained by a more or less complex survey design and possibly affected by non-response.

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 GLIM), the sampling design and a response model.

Ordinary (ML-based) GLIM inference - being based on the assumption of independent observations - is not automatically valid in this situation. It is, however, shown that ordinary GLIM inference does apply under certain conditions. It is demonstrated - and illuminated 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 GLIM inference - when valid - can be considerably more efficient than inference based on Horvitz-Thompson weighting.

Key words: Analysis of survey data, Generalized linear regression, superpopulation models, non-response.

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 non-response.

The word regression should be interpreted in a fairly wide sense here. We will consider generalized linear models (GLIMs), c.f. McCullagh & Nelder (1983), which include linear-, logistic-, probit- and Poisson regression among others.

Smith (1981) - who considers linear regression for complex surveys - makes the distinction between descriptive and analytic inference, and we adhere to this terminology.

In a descriptive inference the objective is to estimate a parameter, $\underline{B}$ 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 non-response 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).

Our approach in this paper will be analytic. Then the interest focuses on the (unknown) relation between some variables y and $\underline{x}$. 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 $\underline{x}$ is assumed to be expressable by a family of statistical distributions - usually referred to as the superpopulation model. This family of distributions is assumed to be indexed by a parameter $\underline{\beta}$, and this model parameter is in the focus of interest in the analytic approach. Emphasis is on model building whereas in the descriptive approach the main problem is estimation of a fixed-population quantity $\underline{B}$.

It should be emphasized that our approach is neither purely model based nor purely design based. In fact we shall consider three sources of random variation, specified by the superpopulation model, the sampling design and the response model respectively. A fourth relevant source of error, measurement errors, could be introduced, but we abstain from doing so in this paper.

We will assume data as generated by a three step process as follows.

(i) A population of N elements is generated by N independent observations from a specified family of distributions.

(ii) From the population generated in (i) a sample of prescribed size n, $n<N$, 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)-(iii) follows in Section 2 ahead. In particular we will concentrate on GLIM superpopulation models.

Related approaches - although restricted to linear regression and without step (iii) - are found in Du Mouchel & Duncan (1983), Nathan & Holt (1980) and Ten Cate (1986).

Treatment of non-response within the general framework of (i)-(iii) is found in Rubin (1976, 1987), Little (1982) and Little & Rubin (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 separate approaches.

2 Specification of superpopulation, sampling and response 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 & Nelder (1983), i.e. the probability density g_i for Y_i takes the form

$$g_i(y,\underline{\beta},\Phi) = \exp\left[\frac{y\theta_i - b(\theta_i)}{\Phi w_i} + c_i(y,\Phi)\right], \; y \varepsilon \Psi, \tag{2.1}$$

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

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

We assume that $\{w_i\}_{i=1}^{N}$ are known scale factors and that $\{x_{ki}, \; k=0,1,\ldots,m, \; i=1,\ldots,N\}$ are known covariates playing the role of explanatory variables. (We could also regard x as random and then make the inference conditioned on x.) Furthermore, we assume that $b(\cdot)$, $c_i(\cdot)$ and $f(\cdot)$ are known, three times continously differentiable, functions which satisfy

$$b''(\cdot) > 0, \tag{2.3}$$

$$f(\cdot) \text{ is strictly monotone,} \tag{2.4}$$

$$\Phi > 0. \tag{2.5}$$

The following relations are straightforward consequences of (2.1)-(2.2):

For $\mu_i = E(Y_i)$ and $\sigma_i^2 = Var(Y_i)$ we have

$$\mu_i(\underline{\beta}) = b'(\theta_i), \tag{2.6}$$

$$\sigma_i^2(\underline{\beta}) = b''(\theta_i)\cdot\Phi w_i. \tag{2.7}$$

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

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

$$\delta_i = \begin{cases} 1 & \text{if } i\varepsilon\Sigma_n \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

We assume that all the relevant information about the sampling design is contained in a set of - possibly multidimensional - random variables $\underline{z}_i$, $i=1,2,\ldots,N$ - the design variables.

Two extreme special cases can be noticed here. The first one occurs if $\underline{z}$ can be expressed by a known function of $\underline{x}$ (exogenous sampling). This implies - by assumptions made above - that $\underline{z}$ is non-random and known.

The second case occurs if $\underline{z}$ can be expressed by a known function of Y (endogenous sampling). Then $\underline{z}$ is random since Y is random. Although these two cases will be covered our main interest will be cases where $\underline{z}$ cannot - without random error - be expressed through $\underline{x}$ and/or Y.

The first and second order inclusion probabilities, being functions of $\underline{z}$, are defined as follows.

$$\pi_i = P(\delta_i=1|\underline{z}) \qquad i=1,2,\ldots,N. \qquad (2.9)$$

$$\pi_{ij} = P(\delta_i=1,\delta_j=1|\underline{z}) \qquad i,j,=1,2,\ldots N. \qquad (2.10)$$

Next we discuss the response mechanism. In order to model the response pattern we assume a set of functions $P_i(\underline{u}_i,\underline{\alpha})$, $i=1,2,\ldots,N$, where $\underline{\alpha}$ is an unknown parameter vector and $\underline{u}_i$ $i=1,2,\ldots,N$ is a set of random vectors.

For notational convenience we introduce the following vectors for $i=1,2,\ldots,N$

$$\left.\begin{aligned} t_i &= (Y_i,\underline{x}_i,\underline{z}_i,\underline{u}_i)\\ \underline{t} &= (t_1,t_2,\ldots,t_N)\\ &\text{where}\\ \underline{x}_i &= (x_{oi},\ldots,x_{mi}). \end{aligned}\right] \qquad (2.11)$$

Let for i=1,2,...,N

$$r_i = \begin{cases} 1 & \text{if } i\varepsilon\Sigma_n \text{ and element } i \text{ responds} \\ 0 & \text{otherwise} \end{cases} \qquad (2.12)$$

We assume pairwise conditionally independent responses in the following sense.

For k,ℓ=0,1 , i,j=1,2,...,N , i≠j

$$\underline{P}(r_i=k,r_j=\ell|\delta_i=1,\delta_j=1,\underline{t})=\underline{P}(r_i=k|\delta_i=1,\underline{t})\cdot\underline{P}(r_j=\ell|\delta_j=1,\underline{t}). \qquad (2.13)$$

Furthermore, we assume that

$$\underline{P}(r_i=1|\delta_i=1,\underline{t})=1-\underline{P}(r_i=0|\delta_i=1,\underline{t})=P_i(\underline{u}_i,\underline{\alpha}), \qquad (2.14)$$

where the response probabilities $P_i(\underline{u}_i,\underline{\alpha})$, i=1,2,...,N, are the functions $P_i(\underline{u}_i,\underline{\alpha})$ introduced earlier. Notice that (2.13) and (2.14) mean that we assume that all the relevant information about the response pattern which is contained in $\underline{t}$ is in fact carried by $\underline{u}$.

Although in many applications it will be reasonable to assume that $\underline{u}$ can be expressed by a known function of $(Y,\underline{x},\underline{z})$ we will also cover cases where $\underline{u}$ cannot - without random error - be expressed through $(Y,\underline{x},\underline{z})$.

As a response model $P_i(\underline{u}_i,\underline{\alpha})$ 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 $P_i(\underline{u}_i,\underline{\alpha})$ in this paper, except that it can be expressed by a parametric model.

We close this section by deriving some conditional expectations which will be useful later. It is easy to see by (2.12) that

$E(r_i|\underline{t})=\underline{P}(r_i=1|\underline{t})=\underline{P}(r_i=1,\delta_i=1|\underline{t})$.

Thus, by (2.14),

$$E(r_i|\underline{t})=P_i(\underline{u}_i,\underline{\alpha})\cdot P(\delta_i=1|\underline{t}). \qquad (2.15)$$

We assumed earlier that all the relevant information about the sampling mechanism is contained in $\underline{z}$. More specifically we make the following assumptions:

$$\underline{P}(\delta_i=1|\underline{t})=\underline{P}(\delta_i=1|\underline{z})=\pi_i(\underline{z}). \qquad (2.16)$$

$$\underline{P}(\delta_i=1,\delta_j=1|\underline{t})=\underline{P}(\delta_i=1,\delta_j=1|\underline{z})=\pi_{ij}(\underline{z}) \qquad (2.17)$$

Combination of (2.15) and (2.16) yields

$$E(r_i|\underline{t})=P_i(\underline{u}_i,\underline{\alpha})\cdot\pi_i(\underline{z}). \qquad (2.18)$$

Furthermore, for $i\neq j$,

$$E(r_ir_j|\underline{t})=\underline{P}(r_i=1,r_j=1|\underline{t})=\underline{P}(r_i=1,r_j=1,\delta_i=1,\delta_j=1|\underline{t})=$$

$$=\underline{P}(r_i=1,r_j=1|\delta_i=1,\delta_j=1,\underline{t})\cdot\underline{P}(\delta_i=1,\delta_j=1|\underline{t}).$$

By (2.13), (2.14) and (2.17) we have

$$E(r_i r_j|\underline{t})=P_i(\underline{u}_i,\underline{\alpha})P_j(\underline{u}_j,\underline{\alpha})\cdot\pi_{ij}(\underline{z}),\ i\neq j\ . \qquad (2.19)$$

Notice that the P's and π's, being functions of $\underline{u}$ and $\underline{z}$ respectively, must generally be treated as random variables in our setup.

3 Unweighted estimation

Suppose for a moment that, in the data generation process of Section 2, sampling is done by simple random sampling without replacement, with such a small (known) sampling fraction that the sampling, as a good approximation, can be regarded as being done with replacement. Furthermore, suppose that the response probabilities P_i all equal a (common) positive known constant. If the distribution of $(r_1,r_2,\dots,r_N)$ is unrelated to $\underline{\beta}$, then the

$n_r = \sum_{i=1}^{N} r_i$ Y-observations may, as a good approximation,

be regarded as independent random variables, distributed according to (2.1)-(2.2). The likelihood equation for $\underline{\beta}$ (conditional on $r_1,r_2,\dots,r_N$)) can be written on the form (notice (2.6)):

$$\sum_{i=1}^{N} \left(\frac{y_i-\mu_i(\underline{\beta})}{\Phi w_i}\right) f'(\underline{x}_i\underline{\beta})\cdot x_{ji}\cdot r_i = 0 \ , \ j=0,1,\dots m. \qquad (3.1)$$

In cases when the data may be regarded as independent observations - although in general non-i.i.d. due to the GLIM form - the following results hold under various regularity conditions (see e.g. Habermann (1977), Nordberg (1980) and Fahrmeir & Kaufmann (1985)).

The likelihood equation has - with a probability tending to one as the sample size tends to infinity - one root being a consistent estimator of $\underline{\beta}$. (Multiple roots may exist but only one yielding consistency.) This estimator is asymptotically efficient and asymptotically normally distributed.

However, the line of analysis indicated above does not cover the situation in Section 2 where data cannot even as an approximation a priori be regarded as independent observations. Furthermore, it is obvious that, due to the random nature of r_i, (3.1) is not in general the likelihood equation for $\underline{\beta}$.

Nevertheless, as seen by Proposition 1 ahead, equation (3.1) does, under certain general conditions, have a consistent and 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 within the framework of Section 2.

$$\left.\begin{array}{l} \text{Set } S_n(\underline{\beta},\underline{Y})=(S_n^{(o)}(\underline{\beta},\underline{Y}),\ldots,S_n^{(m)}(\underline{\beta},\underline{Y}))' \\ \text{where - c.f. (3.1) -} \\ S_n^{(j)}(\underline{\beta},\underline{Y})=\frac{1}{n}\sum_{i=1}^{N}(\frac{y_i-\mu_i(\beta)}{\Phi w_i})f'(\underline{x}_i\underline{\beta})x_{ji}r_i. \end{array}\right] \qquad (3.2)$$

Let

$$D_n(\underline{\beta},\underline{Y})=-\{\frac{\partial S_n^{(j)}}{\partial\beta_k},\ j,k=0,1,\ldots m\} \qquad (3.3)$$

and

$$A_n(\underline{\beta})=E(D_n(\underline{\beta},\underline{Y})). \qquad (3.4)$$

Furthermore, let $V_n(\underline{\beta})$ be the variance - covariance matrix (with respect to the total variation) of $\sqrt{n}S_n(\underline{\beta},\underline{Y})$. We are now ready to formulate the following result on consistency and asymptotic normality.

Proposition 1: Let assumptions be as in Section 2 and let $\underline{\beta}_o$ be the true parameter point. Suppose that

$$(S_n(\underline{\beta}_o,\underline{Y})-E(S_n(\underline{\beta}_o,\underline{Y}))\xrightarrow{P}0 \quad \text{as } n\to\infty. \tag{3.5}$$

$$\sqrt{n}V_n^{-1/2}(\underline{\beta}_o)(S_n(\underline{\beta}_o,\underline{Y})-E(S_n(\underline{\beta}_o,\underline{Y})))\Longrightarrow N(0,I) \text{ as } n\to\infty. \tag{3.6}$$

$$\lim_{n\to\infty} \sqrt{n}E(S_n(\underline{\beta}_o,\underline{Y}))=0. \tag{3.7}$$

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 $n\to\infty$ - exaktly one consistent root $\hat{\underline{\beta}}^{(n)}$ i.e.

$$\hat{\underline{\beta}}^{(n)}\xrightarrow{P}\underline{\beta}_o. \tag{3.8}$$

Furthermore,

$$\sqrt{n}V_n^{-1/2}(\underline{\beta}_o)A_n(\underline{\beta}_o)(\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o)\Longrightarrow N(0,I) \text{ as } n\to\infty. \tag{3.9}$$

__Remark__: Conditions (3.5)-(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 invertibility of certain matrices etc. A set of such regularity conditions is specified in appendix where a more precise version of Proposition 1 is proved.□

Conditions (3.5) and (3.6) state that $S_n(\boldsymbol{\beta}_o,\underline{Y})$ 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 - and response mechanisms. 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 $V_n(\boldsymbol{\beta}_o)$. If $V_n(\boldsymbol{\beta}_o)$ and $A_n(\boldsymbol{\beta}_o)$ coincide then (3.9) takes the classical form i.e. that $\sqrt{n}A_n{}^{1/2}(\boldsymbol{\beta}_o)(\hat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta}_o)$ is asymptotically N(O,I). Classical asymptotic theory can then be applied. We will derive conditions which are sufficient for (3.7) and for $V_n(\boldsymbol{\beta}_o)$ and $A_n(\boldsymbol{\beta}_o)$ to coincide.

By applying $E(\cdot)=EE(\cdot|\underline{t})$ to (3.2) and (2.18) we have

$$\sqrt{n}E(S_n^{(j)}(\boldsymbol{\beta}_o,\underline{Y}))=\frac{1}{\sqrt{n}}\,E(\sum_{i=1}^{N}(\frac{y_i-\mu_i(\boldsymbol{\beta}_o)}{\Phi w_i})f'(\underline{x}_i\boldsymbol{\beta}_o)x_{ji}\pi_iP_i),j=0,1,\ldots m. \tag{3.10}$$

Hence (3.7) is equivalent to

$$\lim_{n\to\infty}\frac{1}{\sqrt{n}}\,E(\sum_{i=1}^{N}(\frac{y_i-\mu_i(\boldsymbol{\beta}_o)}{\Phi w_i})f'(\underline{x}_i\boldsymbol{\beta}_o)x_{ji}\pi_iP_i)=0\ ,\ j=0,1,\ldots,m. \tag{3.11}$$

Relation (3.11) holds if $\pi_i P_i$ is uncorrelated with the residual $Y_i-\mu_i(\beta_o)$. Another way of expressing this is that (3.11) holds if $\pi_i P_i$ does not carry any information on Y, not already accounted for by the covariates $\underline{x}$ of model (2.1)-(2.2). We will return to this point later but first we consider the relation between $V_n(\beta_o)$ and $A_n(\beta_o)$.

As pointed out earlier, classical inference can be applied if - in addition to (3.7) - $V_n(\beta_o)$ and $A_n(\beta_o)$ coincide. We will now derive sufficient conditions for $V_n(\beta_o)\equiv A_n(\beta_o)$.

By (3.2)

$$V_n(\beta_o)=\frac{1}{n}\ \{\mathrm{Cov}(\sum_i(\frac{Y_i-\mu_i(\beta_o)}{\Phi w_i})f'(\underline{x}_i\beta_o)x_{ki}r_i,$$

$$,\sum_j(\frac{Y_j-\mu_j(\beta_o)}{\Phi w_j})\ f'(\underline{x}_j\beta_o)x_{\ell j}r_j)\quad k,\ell=0,\ldots,m\} \tag{3.12}$$

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

$$V_n^{(k\ell)}(\beta_o)=\frac{1}{n}(E(\sum_{ij}\sum(\frac{Y_i-\mu_i(\beta_o)}{\Phi w_i})(\frac{Y_j-\mu_j(\beta_o)}{\Phi w_j})f'(\underline{x}_i\beta_o)f'(\underline{x}_j\beta_o)x_{ki}x_{\ell j}r_ir_j)-$$

$$-\ E(\sum_i(\frac{Y_i-\mu_i(\beta_o)}{\Phi w_i})f'(\underline{x}_i\beta_o)x_{ki}r_i)E(\sum_j(\frac{Y_j-\mu_j(\beta_o)}{\Phi w_j})f'(\underline{x}_j\beta_o)x_{\ell j}r_j))$$

(3.13)

By (2.18) and (2.19) relation (3.13) takes the form

$$V_n^{(k\ell)}(\underline{\beta}_o)=\frac{1}{n}\,E(\sum_i\frac{(Y_i-\mu_i(\underline{\beta}_o))^2}{\Phi^2 w_i^2}(f'(\underline{x}_i\underline{\beta}_o))^2 x_{ki}x_{\ell i}\pi_i P_i)+$$

$$+\frac{1}{n}\,E(\sum_{i\neq j}\sum(\frac{Y_i-\mu_i(\underline{\beta}_o)}{\Phi w_i})(\frac{Y_j-\mu_j(\underline{\beta}_o)}{\Phi w_j})f'(\underline{x}_i\underline{\beta}_o)f'(\underline{x}_j\underline{\beta}_o)x_{ki}x_{\ell j}\pi_{ij}P_iP_j)-$$

$$-\frac{1}{n}\,(E(\sum_i(\frac{Y_i-\mu_i(\underline{\beta}_o)}{\Phi w_i})f'(\underline{x}_i\underline{\beta}_o)x_{ki}\pi_iP_i)\;E(\sum_j(\frac{Y_j-\mu_j(\underline{\beta}_o)}{\Phi w_j})f'(\underline{x}_j\underline{\beta}_o)x_{\ell j}\pi_jP_j))$$

$$(3.14)$$

By differentiating (3.2) and noting (3.3), (3.4), (2.6), (2.7) and (2.18) it is seen that entry (k,ℓ) of $A_n(\underline{\beta}_o)$ is

$$A_n^{(k\ell)}(\underline{\beta}_o)=-\,\frac{1}{n}\,E(\sum_i(\frac{Y_i-\mu_i(\underline{\beta}_o)}{\Phi w_i})f''(\underline{x}_i\underline{\beta}_o)x_{ki}x_{\ell i}\pi_iP_i)\;+$$

$$+\frac{1}{n}\,E(\sum_i(\frac{\sigma_i(\underline{\beta}_o)f'(\underline{x}_i\underline{\beta}_o)}{\Phi w_i})^2x_{ki}x_{\ell i}\pi_iP_i) \qquad (3.15)$$

Suppose now that π_iP_i is uncorrelated with the residual i.e.

$$E((Y_i-\mu_i(\underline{\beta}_o))\cdot\pi_iP_i)=0 \qquad i=1,2,\ldots N. \qquad (3.16)$$

Then (3.11) holds and the last term of (3.14) and the first one of (3.15) will vanish. If, in addition to (3.16), the following conditions hold

$$E(Y_i-\mu_i(\underline{\beta}_o))^2\pi_iP_i=E((Y_i-\mu_i(\underline{\beta}_o))^2E(\pi_iP_i) \qquad (3.17)$$

and

$$E(Y_i-\mu_i(\underline{\beta}_o))(Y_j-\mu_j(\underline{\beta}_o))\pi_{ij}P_iP_j=0, \quad i\neq j=1,2,\ldots N, \qquad (3.18)$$

then it is easily seen that $V_n(\underline{\beta}_o)$ and $A_n(\underline{\beta}_o)$ coincide since the last term of (3.15) equals the first one of (3.14) while all other terms of (3.14) and (3.15) disappear.

Hence, if (3.16)-(3.18) are fulfilled we can (under very general additional conditions) apply classical inference to survey data generated as in Section 2.

Next we compare (3.16)-(3.18) to the more general ignorability conditions as discussed by Rubin (1976, 1987), Little (1982), Little & Rubin (1987). The essence of ignorability is as follows.

- The sampling distribution of $\underline{\delta}$, conditioned on $\underline{t}$, must not depend on $\underline{\beta}$ or on y-values (of the population) which are unobserved due to sampling or non-response.

- The response distribution of sampled units must not depend on $\underline{\beta}$ or on unobserved y-values.

It is straightforward to establish (3.16)-(3.18) as consequences of these conditions. Hence (3.16)-(3.18) are weaker than the general ignorability conditions. It should be kept in mind though that (3.16)-(3.18) were derived under a specific class of superpopulations - GLIMs - while ignorability applies to more general situations.

The main advantage, however, of using (3.16)-(3.18) to check if classical theory applies is that (3.16)-(3.18) should be easier to check in practise by such devices as residual plots etc. We return to this in Section 6 ahead. Next we will discuss Horvitz-Thompson weighting applied to the framework of Section 2.

4 Horvitz-Thompson weighting

Consider the function

$$F(\underline{\beta},\Phi)=\sum_{i=1}^{N}\frac{1}{\pi_i P_i}\left(\frac{y_i\theta_i-b(\theta_i)}{\Phi w_i}\right)+c_i(y_i,\Phi))r_i, \qquad (4.1)$$

where θ_i depends on $\underline{\beta}$ through (2.2).

$F(\underline{\beta},\Phi)$ 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 $\partial F/\partial\underline{\beta}=0$ i.e. (c.f. (3.1))

$$\sum_{i=1}^{N}\frac{1}{\pi_i P_i^*}\left(\frac{y_i-\mu_i(\underline{\beta})}{\Phi w_i}\right)f'(\underline{x}_i\underline{\beta}_i)x_{ji}r_i=0 \quad , \quad j=0,1,\ldots m \qquad (4.2)$$

for $\underline{\beta}$ we get the Horvitz-Thompson weighted $\underline{\beta}$-estimator $\tilde{\underline{\beta}}^{(n)}$.

Notice that, in order to make (4.2) operational, P_i has been replaced by P_i^*. The quantity P_i^* is obtained by plugging a suitable estimator $\underline{\alpha}^*$ of $\underline{\alpha}$ into the function $P_i(\underline{u}_i,\underline{\alpha})$ i.e. $P_i^*=P_i(\underline{u}_i,\underline{\alpha}^*)$ (c.f. 2.14).

In correspondence with (3.2)-(3.4) the following quantities are introduced.

$$\left.\begin{array}{l}\tilde{S}_n(\underline{\beta},\underline{Y})=(\tilde{S}_n^{(o)}(\underline{\beta},\underline{Y}),\ldots,\ \tilde{S}_n^{(m)}(\underline{\beta},\underline{Y}))' \\ \text{where} \\ \tilde{S}_n^{(j)}(\underline{\beta},\underline{Y})=\frac{1}{n}\sum\limits_{i=1}^{N}\frac{1}{\pi_i P_i^*}\left(\frac{Y_i-\mu_i(\underline{\beta})}{\Phi w_i}\right)f'(\underline{x}_i\underline{\beta})x_{ji}r_i\end{array}\right]\ (4.3)$$

$$\tilde{D}_n(\underline{\beta},\underline{Y})=-\{\frac{\partial\tilde{S}_n^{(j)}}{\partial\beta_k},\ j,k=0,1,\ldots,m\} \qquad (4.4)$$

$$\check{A}_n(\underline{\beta})=E(\tilde{D}_n(\underline{\beta},\underline{Y})) \qquad (4.5)$$

It is straightforward (by modification of Proposition 1) to show that conclusions (3.8) and (3.9) hold for $\tilde{\underline{\beta}}^{(n)}$ under conditions that are very similar to those of Proposition 1. As seen from (4.3) the consistency condition

$$\lim_{n\to\infty}\sqrt{n}\,E(\tilde{S}_n(\underline{\beta}_o,\underline{Y}))=0 \qquad (4.6)$$

corresponding to (3.7) is not automatically fulfilled since P_i is replaced by P_i^*. However, (4.6) holds under mild regularity conditions. A sufficient condition is (c.f. (2.18)).

$$\left.\begin{array}{l}E((Y_i-\mu_i(\underline{\beta}_o))e_i)=0 \\ \text{where} \\ E(\frac{r_i}{\pi_i P_i(\underline{u}_i,\underline{\alpha}^*)}|\underline{t})=e_i.\end{array}\right]\ (4.7)$$

Condition (4.7) is - assuming that $\underline{\alpha}^*$ is a good estimator of $\underline{\alpha}$ - milder than (3.16) or (3.11) and this is the main advantage of $\tilde{\underline{\beta}}^{(n)}$ as compared to $\hat{\underline{\beta}}^{(n)}$.

The counterpart of (3.9) for $\underline{\beta}^{(n)}$ is

$$\sqrt{n}\tilde{V}_n^{-1/2}(\underline{\beta}_o)\tilde{A}_n(\underline{\beta}_o)(\tilde{\underline{\beta}}^{(n)}-\underline{\beta}_o) ==> N(0,I) \text{ as } n->\infty, \qquad (4.8)$$

where $\tilde{V}_n(\underline{\beta}_o)$ is the variance-covariance matrix of $\sqrt{n}\tilde{S}_n(\underline{\beta}_o,\underline{Y})$.

By expressing $\tilde{V}_n$ and $\tilde{A}_n$ in forms analogous to (3.14) and (3.15) it is seen that $\tilde{V}_n$ and $\tilde{A}_n$ do not coincide except in some very special cases. This means that likelihood ratios - based on (4.1) - and their associated χ^2-tests do not apply without extensive modifications, see e.g. Rao & Scott (1984), Roberts et al (1987), Hidiroglou & Rao (1987 a, b).

There is, however, a way to circumvent this obstacle as suggested by Binder (1983). Suppose that $\underline{\beta}$ is partitioned into $(\underline{\beta}_1, \underline{\beta}_2)$ and we want to test H_o:$\underline{\beta}_2=0$. Calculate the estimate $\tilde{\underline{\beta}}=(\tilde{\underline{\beta}}_1, \tilde{\underline{\beta}}_2)$ and its variance $\tilde{C}= \frac{1}{n}\tilde{A}_n^{-1}(\tilde{\underline{\beta}})\cdot\tilde{V}_n(\tilde{\underline{\beta}})\cdot\tilde{A}_n^{-1}(\tilde{\underline{\beta}})$. Let $\tilde{C}_2$ be the part of $\tilde{C}$ which corresponds to $\underline{\beta}_2$. Then $Q=\tilde{\underline{\beta}}_2'\tilde{C}_2^{-1}\tilde{\underline{\beta}}_2$ is asymptotically χ^2 under H_o.

It can be seen as a drawback of this procedure that the additivity under nested sets of models (see e.g. McCullagh & Nelder (1983), p. 26) which applies to the log likelihood is lost here.

5 Unweighted versus weighted estimation

5.1 General remarks

As pointed out in Section 4, $\tilde{\underline{\beta}}$ is consistent under mild condi- It should be emphasized though that the consistency of $\tilde{\underline{\beta}}$ is not "global" but in fact conditional on the superpopulation model as demonstrated ahead.

By Section 3 it is seen that, under certain conditions, $\hat{\underline{\beta}}$ is consistent, and classical inference methods do apply. However, these conditions are not always fulfilled and in such cases $\hat{\underline{\beta}}$ may be inconsistent.

One objective of the simulation study presented ahead is to illustrate, empirically, the above results. A second one is to make some efficiency comparisons. Although it can be shown that, when classical inference applies, $\hat{\underline{\beta}}$ is asymptotically more efficient than $\tilde{\underline{\beta}}$, it is of interest to get an idea of the magnitude of this efficiency gain. A third objective is to demonstrate how the conditions for classical inference on $\hat{\underline{\beta}}$ can be met by incorporating design variables into the model building process.

5.2 A Simulation Study

5.2.1 Background to the choice of superpopulation model

The choice of the superpopulation model used in the simulations to be presented ahead 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. The background was that - due to a surplus of diary products - various means had been introduced by the government as to encourage farmers to diminish milk production. We summarize here only those parts of the study which are relevant in the present context.

For each farm which - according to the Swedish Farm Register - had at least one and at most nine milk cows in June 1983 (roughly 12 000 farms), it was checked whether it still had milk cows in June 1984. In case it did have cows a variable y was set to zero, otherwise y=1. A logit analysis was then performed with y as dependent variable. This analysis was based on the full population with a very small non-response rate so there was no problem here related to sampling or non-response. Eight explanatory factors were tried and the following four turned out to be the most "relevant".

- The size (S) measured as number of milk cows in 1983. S=0 if this number was between 4 and 9 while S=1 if the number was between 1 and 3.

- Age (A) of the farmer - classified as [-49], [50-59], [60-] (years) and denoted by A=1, 2 and 3 respectively.

- 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.

- The region (R) where R=1 in the most productive farming areas in the south and the middle of the country, R=2 in the rest of the south and the middle, while R=3 in the north.

In the sequel we represent R by the two dummy variables

$$R_2 = \begin{cases} 1 & \text{if } R=2 \\ 0 & \text{otherwise} \end{cases}$$

$$R_3 = \begin{cases} 1 & \text{if } R=3 \\ 0 & \text{otherwise} \end{cases}$$

while A_2 and A_3 are defined in a completely analogous fashion.

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

$P(Y=1)=\exp(H)/(1+\exp(H))$

where

$$H=-2.5+1.6S-0.3A_2+0.8A_3-0.8A_3xS+1.0T-0.3R_2xS-0.5R_3xS \qquad (5.1)$$

Table 5.1 presents P(Y=1) according to (5.1) as well as the number of observations, N, in the population for the different combinations of the explanatory factors.

		S (Size)							
		S=0 (4-9 cows)				S=1 (1-3 cows)			
A (Age)	R (Region)	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

TABLE 5.1 P(Y=1) by (5.1) and number of observations (N) in the population for different combinations of Age, Region, Size and Type.

5.2.2 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 3x3x2x2 cells of Table 5.1. The number of elements in a particular group equals the value of N in the corresponding cell of Table 5.1. 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 (5.1)

(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 5.1. 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 per cent 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) Response mechanism: The option of non-response was not considered in this experiment, i.e. $P_i=1$, $i=1,2,\ldots,N$.

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

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

5.2.3 Results

Suppose that the chosen explanatory vector includes Region and Age only. The results of the simulations in this case are presented in Table 5.2. A comparison of the $\underline{\beta}$-estimators of Table 5.2 to the true $\underline{\beta}$ (see (5.1)) shows that all three estimators, including $\hat{\underline{\beta}}_{pop}$, 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 as strong "design-bias", as seen by comparing MEAN($\hat{\beta}$) to MEAN($\hat{\beta}_{pop}$). Notice that $\tilde{\beta}$ lacks the latter type of bias. It might then be argued that $\tilde{\beta}$ is preferable to $\hat{\beta}$ in the present situation.

However, it is possible to get rid of the design bias and much of the model bias simultaneously as seen by the following argument.

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 5.2.2 (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 5.3 and 5.4 support this conclusion. Table 5.4 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 5.3 the explanatory vector contains only the most significant variables, i.e. S,T,A_3 and SxA_3.

It is also seen from Tables 5.3 and 5.4 that

- MEAN($\hat{\sigma}$)≈STD($\hat{\beta}$) which means that the classical variance estimator is approximately unbiased.

- $\hat{\beta}$ is considerably more efficient than $\tilde{\beta}$.

Explanatory vector	MEAN($\hat{\beta}_{pop}$)	MEAN($\hat{\beta}$)	MEAN($\tilde{\beta}$)	MEAN($\hat{\sigma}(\hat{\beta}$)	STD($\hat{\beta}$)	STD($\tilde{\beta}$)	$\left[\frac{STD(\tilde{\beta})}{STD(\hat{\beta})}\right]^2$
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 5.2 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($\tilde{\beta}$)	$\left[\frac{STD(\tilde{\beta})}{STD(\hat{\beta})}\right]^2$
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
SxA_3	-0.78	-0.77	-0.77	0.19	0.21	0.25	1.42

TABLE 5.3 Explanatory vector contains the most significant variables only.

Explanatory vector	MEAN($\hat{\underline{\beta}}_{pop}$)	MEAN($\hat{\underline{\beta}}$)	MEAN($\tilde{\underline{\beta}}$)	MEAN($\hat{\sigma}(\hat{\underline{\beta}})$	STD($\hat{\underline{\beta}}$)	STD($\tilde{\underline{\beta}}$)	$\left[\frac{STD(\tilde{\underline{\beta}})}{STD(\hat{\underline{\beta}})}\right]^2$
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
SxT	0.01	-0.01	-0.02	0.23	0.21	0.22	1.11
SxR_2	-0.30	-0.29	-0.30	0.24	0.24	0.28	1.33
SxR_3	-0.50	-0.49	-0.51	0.29	0.30	0.33	1.22
SxA_2	0.00	0.02	0.03	0.30	0.30	0.35	1.31
SxA_3	-0.79	-0.80	-0.80	0.24	0.25	0.27	1.21
TxR_2	-0.01	-0.00	-0.01	0.24	0.24	0.24	1.03
TxR_3	0.00	-0.00	-0.01	0.28	0.28	0.28	1.03
TxA_2	-0.01	0.01	0.01	0.30	0.32	0.32	1.01
TxA_3	-0.02	-0.01	-0.00	0.24	0.24	0.24	1.00

TABLE 5.4 Explanatory vector contains all main effects and S- and T-interactions.

6 Concluding discussion

As pointed out earlier, the main consistency condition (3.11) for $\hat{\underline{\beta}}$ holds if $\pi_i P_i$ does not carry any information on Y, not already accounted for by the model. By this fact - and with the simulation results above in mind - we are led to the following procedure, previously suggested in a special case by DuMouchel & Duncan (1983).

Extend the explanatory vector $\underline{x}_i$ by bringing in $\pi_i P_i^*$ and its interactions $\pi_i P_i^* x_{ji}$, j=1,...,m, together with $\underline{x}_i$ as additional variables. (We use the convention $x_{0i} \equiv 1$, the constant term corresponding to the intercept). Then test model (2.1)-(2.2) against the extended model (using classical methods, i.e. performing as if data were independent observations). If the extended model does not significantly improve the fit, then this suggests that (3.11) is reasonbly satisfied and that unweighted estimation will be consistent (provided of course, that the other conditions of Proposition 1 hold). On the other hand, if the extended model significantly improves the fit then this suggests that there is useful information in the design or response model which may be used to improve model (2.1)-(2.2). By bringing variables associated with πP - such as S and T in the simulation example - into the model building process it is possible to remove the design bias and much of the model bias simultaneously.

However, sometimes the residual information carried by ΠP cannot, for one reason or another, be utilized.

An obvious case is endogenous sampling where the sampling is based on y itself. In such a case Horvitz-Thompson weighting is usually the proper procedure.

There may also be cases where the variables involved in ΠP should not, for "subject matter reasons", be included in the model. For instance, suppose that y is income and that sampling is based on last year's paid income tax. If the latter variable is not of interest as explanatory variable although correlated with the residual, then, again, Horvitz-Thompson weighting may be recommended.

However, we believe - with experience from Statistics Sweden in mind - that often enough, model improvement by including variables associated with ΠP (such as S and T in the simulation example) is possible and indeed worthwhile.

To make sure that classical inference applies we must also check the variance conditions (3.17) and (3.18) in the model building process outlined above. Condition (3.17) can be checked by residual analysis. Prevalent methods for residual analysis with special reference to GLIMs are reviewed in McCullagh & Nelder (1983).

Condition (3.18) seems at first to be more complex than (3.17). However, suppose that (3.16) (which should hold if (3.11) holds but which could also be checked by residual analysis) and (3.17) hold and that the GLIM framework as specified in Section 2 applies, in particular

$$E(Y_i-\mu_i(\beta_o))(Y_j-\mu_j(\beta_o))=0 \quad i\neq j \quad i,j=1,2,\dots,N. \qquad (6.1)$$

Then (3.18) appears to be quite harmless while (3.16) and (3.17) are the more crucial conditions.

There are certainly situations where (6.1) would be in doubt. There may for instance be cluster effects in the population in the sense that neighbors are "more alike" than non-neighbors, e.g. reading ability among classmates due to common factors such as the same teacher etc. Similar cluster effects can also arise from measurement errors due to systematic interviewer effects. These problems are not primarily caused by the sampling mechanism and they would not go away even if we were to sample the whole population. Horvitz-Thompson weighting is not a solution here. The proper way - in our view - to deal with such deviations from (6.1) is to find a reasonable model specification - perhaps a nested variance component model - which incorporates these cluster effects. This search for a proper model may very well end outside the GLIM class. However, if the GLIMs apply then the model building procedure as outlined above can be of great value.

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

7 References

Binder, D.A. (1983): On the Variances of Asymptotically Normal Estimators from Complex Surveys. Int. statist. Rev., 51, pp., 279-292.

DuMouchel, 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. Ann. statist., 13, pp. 342-368.

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. Ann. Statist., 5, pp. 815-841.

Hidiroglou, M.A. and Rao, J.N.K. (1987 a): Chi-Squared Tests with Categorical Data from Complex Surveys. I. Journal of Official statistics, 2, pp. 117-132.

Hidiroglou, M.A. and Rao, J.N.K. (1987 b): Chi-Squared Tests with Categorical Data from Complex Surveys. II. Journal of Official Statistics, 2, pp. 133-140.

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.

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

Nathan, G. and Holt, D. (1980): The Effect of Survey Design on Regression Analysis. J.R. Statist. Soc. B, 42, pp. 377-386.

Nordberg, L. (1980): Asymptotic Normality of Maximum Likelihood Estimators Based on Independent, Unequally Distributed Observations in Exponential Family Models. Scand. J. Statist., 7, pp. 27-32.

Nordberg, L. (1985): Analys av Avgångar från Mjölkproduktion. Memo, Statistics Sweden (in Swedish).

Rao, J.N.K. and Scott, A.J. (1984): On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Ann. Statist., 12, pp. 46-60.

Roberts, G., Rao, J.N.K. and Kumar, S. (1987): Logistic Regression Analysis of Sample Survey Data. Biometrika, 74, pp. 1-12.

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

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

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.

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

APPENDIX

Let assumptions and notation be as in Section 1-3. Notice that $E(\cdot)$ and $P(\cdot)$ below are supposed to take account of all the random variation induced by the set up in Section 2.

Theorem: Let $\underline{\beta}_0$ be the true parameter point and consider the following conditions

(i) $S_n(\underline{\beta}_o, \underline{Y}) \xrightarrow{P} \underline{0}$ as $n \to \infty$.

(ii) There is a $\lambda > 0$ and an $n_0 > 0$ such that the smallest eigenvalue of $A_n(\underline{\beta}_0) \geq \lambda > 0$ for $n > n_0$.

(iii) $|D_n(\underline{\beta}_o, \underline{Y}) - A_n(\underline{\beta}_o)| \xrightarrow{P} \underline{0}$ (component wise convergence).

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

$$P\left(\left|\frac{\partial^2 S_n^{(j)}}{\partial\beta_k \partial\beta_\ell}\right| < C_\varepsilon \text{ for every } |\underline{\beta}-\underline{\beta}_o| < \delta_o\right) \geq 1-\varepsilon,\ n \geq n_\varepsilon$$

$j,k,\ell = 0,1,\ldots,m$.

(v) $\sqrt{n} V_n^{-1/2}(\underline{\beta}_o) S_n(\underline{\beta}_o, \underline{Y}) \Longrightarrow N(0,I)$ as $n \to \infty$.

(vi) All components of $V_n(\underline{\beta}_o)$ are uniformly bounded for $n>n_o$.

(vii) There is a $\lambda''>0$ and an $n_o>0$ such that the smallest eigenvalue of $V_n(\underline{\beta}_o)\geqslant\lambda''>0$, $n>n_o$.

(a) If conditions (i)-(iv) are fulfilled then with a probability tending to one as the sample size n tends to infinity equation (3.1) has exactly one consistent root $\hat{\underline{\beta}}^{(n)}$ i.e. $\hat{\underline{\beta}}^{(n)}\xrightarrow{P}\underline{\beta}_o$.

(b) If conditions (i)-(vii) are fulfilled then

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

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 so as 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).

<u>Lemma</u>: Let $g_n(\underline{u}) = (g_n^{(1)}(u_1...u_m), \; ... \; g_n^{(m)}(u_1...u_m)$, $n=1,2,...$

be a sequence of three times differentiable functions from R^m to R^m.

Set $H_n(\underline{u}) = \{\frac{\partial g_n^{(j)}(\underline{u})}{\partial u_k}, j,k=1,\dots m\} \quad n=1,2\dots$

Let $\underline{a}\varepsilon R^m$ and suppose that there is a $\lambda>0$ and an $n_o>0$ such that $|H_n(\underline{a})\underline{x}|\geq\lambda|\underline{x}|$ for any $\underline{x}\varepsilon R^m$, $n>n_o$. (A1)

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

$|\frac{\partial^2 g_n^{(j)}(\underline{u})}{\partial u_k \partial u_\ell}|\leq G<\infty$ for every $\underline{u}\varepsilon d(\underline{a},\delta_o)$, $n>n_o$

where $d(\underline{a},\delta_o) =\{\underline{u}:|\underline{u}-\underline{a}|\leq\delta_o\}$ (A2)

Then there is a $\delta_1>0$ such that the restriction of $g_n(\underline{u})$, $n>n_o$, to $d(\underline{a},\delta_1)$ is one-to-one.

Furthermore, if $0<\delta<\delta_1$ and $\underline{z}\varepsilon d(g_n(\underline{a}), \lambda\delta/2)$ then there is exactly one $\underline{u}\varepsilon d(\underline{a},\delta)$ such that $g_n(\underline{u})=\underline{z}$, $n>n_o$.

<u>Proof of lemma</u>: Let $\underline{u}'$ and $\underline{u}''\varepsilon d(\underline{a},\delta_o)$ and suppose that $\underline{u}'\neq\underline{u}''$. By (A2) we have for $n>n_o$

$$g_n(\underline{u}')-g_n(\underline{u}'')=H_n(\underline{a})(\underline{u}'-\underline{u}'')+(H_n(\underline{u}'')-H_n(\underline{a}))(\underline{u}'-\underline{u}'')+R \tag{A3}$$

where for some $C<\infty$

$$|R|\leq C|\underline{u}'-\underline{u}''|^2$$

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

$$|(H_n(\underline{u}'')-H_n(\underline{a}))(\underline{u}'-\underline{u}'')|\leq C'|\underline{u}'-\underline{u}''||\underline{u}''-\underline{a}| \tag{A4}$$

and thus by (A1)

$$|g_n(\underline{u}')-g_n(\underline{u}'')|\geqslant\lambda|\underline{u}'-\underline{u}''|-C'|\underline{u}'-\underline{u}''||\underline{u}''-\underline{a}|-C|\underline{u}'-\underline{u}''|^2 \tag{A5}$$

Set $\delta_1=\min(\delta_o,\ \lambda/(2C'+4C))$ (A6)

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

$$|g_n(\underline{u}')-g_n(\underline{u}'')|\geqslant|\underline{u}'-\underline{u}''|(\lambda-\frac{\lambda C'}{2C'+4C}-\frac{2\lambda C}{2C'+4C})=\frac{\lambda}{2}|\underline{u}'-\underline{u}''| \tag{A7}$$

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

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

$$H_n(\bar{\underline{u}})(g_n(\bar{\underline{u}})-\underline{z})=0$$

Thus

$$H_n(\underline{a})(g_n(\bar{\underline{u}})-\underline{z})\ =\ -(H_n(\bar{\underline{u}})-H_n(\underline{a}))(g_n(\bar{\underline{u}})-\underline{z})$$

By (A4)

$$|(H_n(\bar{\underline{u}})-H_n(\underline{a}))(g_n(\bar{\underline{u}})-\underline{z})|<C'\delta|g_n(\bar{\underline{u}})-\underline{z}|$$

where - by A(6) - $C'\delta< \frac{\lambda C'}{2C'+4C} < \frac{\lambda}{2}$

Therefore

$$|H_n(\underline{a})(g_n(\bar{\underline{u}})-\underline{z})|<\frac{\lambda}{2}\ |g_n(\bar{\underline{u}})-\underline{z}|$$

But due to (A1)

$$|H_n(\underline{a})(g_n(\bar{\underline{u}})-\underline{z})|>\lambda|g_n(\bar{\underline{u}})-\underline{z}|$$

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

<u>Proof of theorem</u>: Let - $S_n(\beta,\underline{Y})$ correspond to $g_n(\underline{u})$, $\underline{\beta}$ and $\underline{\beta}_o$ correspond to $\underline{u}$ and $\underline{a}$ respectively.

Suppose that the smallest eigenvalue of $D_n(\underline{\beta},\underline{Y})>\lambda'>0$, $n>n_o$ (A8)

$$\left|\frac{\partial^2 S_n^{(j)}}{\partial\beta_k\partial\beta_\ell}\right|<G \text{ for every } \underline{\beta}\varepsilon d(\underline{\beta}_o,\delta_o), n>n_o, j,k,1,=0,\ldots m \qquad \text{(A9)}$$

$$\underline{O}\varepsilon d(-S_n(\underline{\beta}_o,\underline{Y}), \lambda'\delta/2) \qquad \text{(A10)}$$

where δ_o, δ and n_o are defined in the proof of the lemma.

If (A8)-(A10) hold then, due to the lemma, there is exactly one root $\hat{\underline{\beta}}^{(n)} \varepsilon d(\underline{\beta}_o,\delta)$ of equation (3.1). Now (ii) and (iii) imply that (A8) is true with a 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)-(A10) hold true simultaneously with a probability tending to one as $n \to \infty$ and this implies conclusion (a) of the theorem.

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

$$\lim_{n \to \infty} P(|\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o|<C'|S_n(\underline{\beta}_o,\underline{Y})|)=1$$

and this relation combined with conditions (v) and (vi) yields for some $C''<\infty$

$$\lim_{n \to \infty} P(\sqrt{n}|\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o|<C''<\infty)=1 \qquad \text{(A11)}$$

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

$$\sqrt{n}|\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o|^2 \xrightarrow{P} 0 \text{ as } n \to \infty \qquad \text{(A12)}$$

Taylor-expansion of $S_n(\underline{\beta},\underline{Y})$ around $\underline{\beta}_o$ (note that $S_n(\hat{\underline{\beta}}^{(n)},\underline{Y})=0$) yields

$$S_n(\underline{\beta}_o,\underline{Y})=A_n(\underline{\beta}_o)(\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o)+(D_n(\underline{\beta}_o,\underline{Y})-A_n(\underline{\beta}_o))(\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o)+R_n \qquad \text{(A13)}$$

where

$$\sqrt{n}|R_n| \xrightarrow{P} 0 \text{ as } n \to \infty \qquad \text{(A14)}$$

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

Condition (iii) and (A11) yields

$$\sqrt{n}|(D_n(\underline{\beta}_o,\underline{Y})-A_n(\underline{\beta}_o))(\hat{\underline{\beta}}^{(n)}-\underline{\beta}_o)| \xrightarrow{P} 0 \text{ as } n \to \infty \qquad \text{(A15)}$$

By condition (v)

$$\sqrt{n}V_n^{-1/2}(\underline{\beta}_o)S_n(\underline{\beta}_o,\underline{Y}) \Longrightarrow N(0,I)$$

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

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

which completes the proof of the theorem.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports, Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown (beige) covers).

Reports published earlier during 1988 are:

Nummer	**Titel (författare)**
1988:1 (beige)	Abstracts I - Sammanfattningar av metodrapporter från SCB
1988:2 (grön)	Coverage Probabilities for Confidence Intervals Based on Stratified Random Sampling (Jörgen Dalén)
1988:3 (gul)	Base Operators as a Tool for Systems Development (Bo Sundgren)
1988:4 (gul)	Development of Systems Design for National Household Surveys - Report from a short-term mission to Harare, Zimbabwe, 12th-28th January, 1988 (Birgitta Lagerlöf)
1988:5 (grön)	Några råd och synpunkter för rationalisering av produktionsmomentet granskning (Leopold Granquist)
1988:6 (grön)	Hur möta energianvändningen och dess utveckling - några alternativa beräkningar (Urban Aspén)
1988:7 (grön)	Bortfallsbarometer nr 3 (Peter Lundquist)

Kvarvarande BEIGE och GRÖNA exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 Stockholm, eller per telefon 08-7834178.

Dito GULA exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 Stockholm, eller per telefon 08-7834147.