ARE SURVEY WEIGHTS NECESSARY? THE MAXIMUM LIKELIHOOD APPROACH TO SAMPLE SURVEY INFERENCE

R. L. Chambers[(1)], A. H. Dorfman[(2)] & SuojinWang[(3)]
(1) Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, UK
(2) Office of Survey Methods Research, U.S. Bureau of Labor Statistics
2 Massachusetts Ave NE, Washington DC 20212
(3) Department of Statistics, Texas A&M University, College Station, Texas 77843

The authors dedicate this paper to Richard Royall in this year of his retirement, for the profound influence of his thought on survey sampling.

Key Words: stratified sampling; estimating equations; pseudo-likelihood; sample likelihood

In the present work we explicate the application of maximum likelihood inference in the analysis of surveys which are the result of (possibly informative) stratified sampling. In Section 1 we review basic ideas, including two general results useful for applying maximum likelihood to sample data. Ideas are illustrated by a simple through the origin regression model. In Section 2, we discuss the application of these ideas to the situation of (possibly) informative stratified sampling. The variable of interest *Y* depends linearly on covariates ***x***, and the stratification variable *T* depends linearly on ***x*** and *Y*. For simplicity, we focus on the through the origin model, taking *T* = *Y*. Section 3 gives results of a simulation study, and Section 4 states conclusions.

1. Maximum likelihood estimation in survey sampling.

Survey sampling is said to have two goals: **analysis** and **enumeration** (Deming 1950.) We illustrate the distinction with a simple example.

Example 1 A through the origin regression model
Suppose we have a population *P* of size *N*¸ in which the variable of interest *Y* follows the model

$$Y_i = x_i\beta + x_i^{1/2}\varepsilon_i, \qquad (1)$$

with $\varepsilon_i \sim N(0,\sigma^2)$ independently for *i* = 1,…,*N*. Suppose the values of the auxiliary variable x_i are available, for *i* = 1,…,*N*, and we take a sample *s* of size *n* < *N*, and determine the values of the Y_i`s on *s*. In *enumeration*, we aim to estimate a Population Value, like $T = \sum_P Y_i$. In a*nalysis* we are instead concerned with estimating a model parameter, like β.

A major tool of analysis (in general of course, not just in the sampling context) is *maximum likelihood estimation*: Suppose the available data *D* are the realization of a random variable *D* ~ $f(D;\Omega)$, where Ω is a vector of unknown parameters and *f* is a probability function or density. (In *Example* 1, $\Omega = \{\beta, \sigma^2\}$.) The likelihood function is just $f(D;\Omega)$ taken as function of Ω. A Maximum Likelihood estimate $\hat{\Omega}$ maximizes $f(D;\Omega)$, or, equivalently, $\log f(D;\Omega)$. Towards this end, it is convenient to calculate the *score function* (with respect to *D*) $sc_D(\Omega) = \dfrac{\partial \log f(D;\Omega)}{\partial \Omega}$; this is a vector with (in our example) components

$$sc_D(\beta) = \frac{\partial \log f(D;\Omega)}{\partial \beta}, \text{ etc.}$$

We set $sc_D(\beta) = 0$, etc. and solve, to get the maximum likelihood estimates.

In the sampling context, the data include not only the data on sampled units, but often also auxiliary information outside the sample, a variable *I* indicating whether particular units are sampled or not, inclusion probabilities π, and a response indicator variable *R* telling whether a sampled unit gets measured. If we broaden the population to include the vectors *I, R,* then the available data lies as it were between the sample data and the full population data: $s \subset D \subset P$. For example, we might have $D = \{Y_s, x_P, I_P, \pi_s, R_s\}$, with distributions of the additional variables

parameterized by nuisance parameters Ω_{nuis}, estimation of which can complicate estimation of Ω. If R is superfluous when we solve for $\hat{\Omega}$, then it is said that "non-response is non-informative", otherwise "informative ". If I, π are superfluous, then "sampling is non-informative", otherwise "informative" (Rubin 1976). In what follows we shall assume non-informative non-response, and ignore it.

Here are two Results, which hold in general, but are especially useful in the sampling context:

Result 1. $D \subset U \Rightarrow sc_D(\Omega) = E(sc_U(\Omega) \mid D)$
(Breckling et al. 1992, Orchard & Woodbury, 1972,…)

Result 2. $D \subset U$. Suppose
$sc_U(\Omega) = 0 \Rightarrow \hat{\Omega} = g(D)$. Then $sc_D(\hat{\Omega}) = 0$.
(Chambers, et al 1998)

Result 1 is basic. It says that if we have the score function with respect to data *U*, we can derive the score function with respect to data *D* included in *U* by conditioning on *D*. *Result 2* says that if the maximum likelihood estimator based on a data set needs only information available from a smaller data set included within it, then the maximum likelihood estimators for the smaller and larger data sets are the same.

Consider again the through the origin regression model with normal errors.

Example 1 (continued)
The density of Y_i given x_i is

$$f(y_i \mid x_i) = (2\pi x_i \sigma^2)^{-1/2} \exp\left(-\frac{(y_i - x_i\beta)^2}{2x_i\sigma^2}\right),$$

leading to the score function for β with respect to the population

$$sc(\beta) = \sigma^{-2}\sum_P (Y_i - \beta x_i).$$

Then $sc(\beta) = 0 \Rightarrow \hat{\beta} = \frac{\sum_P Y_i}{\sum_P x_i} = \frac{\overline{Y}}{\overline{x}}$,

the maximum likelihood estimate of β, if $D = P$.

Suppose what is available is the data $D = \{x_P, Y_s, I_P, \pi_s\}$. We consider two cases:

Case 1 Sampling is done probability proportional to size (*pps*) with size variable *x*, that is, $\pi_i = \frac{nx_i}{N\overline{x}}$, i = 1, 2, …, N. This is an instance of non-informative sampling. For units in the sample, it is clear that $E(Y_i \mid D) = Y_i$, since D contains Y_i. For non-sample units, we have $f(Y_i \mid D) = f(Y_i \mid x_P) = f(Y_i \mid x_i)$, since the inclusion probabilities add no information beyond what is in *x*. Thus we get a score function with respect to the data

$$sc_D(\beta) = \sum_P E(Y_i - \beta x_i \mid D)$$
$$= \sum_s (Y_i - \beta x_i) + \sum_r (E(Y_i \mid x_i) - \beta x_i).$$

But $E(Y_i \mid x_i) = \beta x_i$, so

$sc_D(\beta) = \sum_s (Y_i - \beta x_i)$ and, setting this score function to zero, we get $\hat{\beta}_D = \frac{\sum_s Y_i}{\sum_s x_i}$.

Case 2 (informative sampling) Suppose now *pps* sampling with $\pi_i = nY_i / N\overline{Y}$, I = 1, 2,…, N. This is an extreme case of informative sampling. (In practice this might arise approximately if we did *pps* sampling with respect to a 3rd variable highly correlated with *Y* that is not available at the time of analysis.) Then
$sc_D(\beta) = \sum_P E(Y_i - \beta x_i \mid x_P, Y_s, I_P, \pi_s)$
$= \sum_s Y_i + \sum_r E(Y_i \mid x_i, \pi_s) - \sum_P \beta x_i$. At first sight, the 2nd term looks difficult to deal with. However, since we know $\pi_i = nY_i / N\overline{Y}$ and Y_i, for each sample unit, any such unit determines for us what $\overline{Y}$ is. That is, implicitly $\overline{Y} \in D$. And we were assuming also that $\overline{x}$ is available. Then, since maximum likelihood estimator with respect to the full data P was $\hat{\beta} = \frac{\overline{Y}}{\overline{x}}$, and the ingredients of this expression are available from *D*,

Result 2 implies that $\hat{\beta}_D = \frac{\overline{Y}}{\overline{x}}$ as well. Could any other sample-based estimator be more efficient?

Note bene: Neither in the non-informative or informative case, did we arrive at an estimator that explicitly incorporates the sample weights. In the non-informative case they are ignored; in the informative case they are merely exploited, unconventionally.

1.1 Two s-based approaches to maximum likelihood

Typically the available data *D* contains information beyond what is available on the sample units. We here review two approaches to maximum likelihood that rely only on the sample data (even when "extra-sample" information is available.) The first relies on the sample weights in classic fashion, and is a special case of the use of weighted estimating equations (Binder 1991; Godambe and Thompson 1986).

1. *Pseudo-likelihood* ("weighted maximum likelihood") Let $\pi_i = f(I_i = 1 | P)$ be the probability the *i*th unit is in *s*, and $w_i = \pi_i^{-1}$. Then, in our example, the weighted sample-based score function

$$sc_w(\beta) = \sum_P w_i I_i (Y_i - \beta x_i) = \sum_s w_i (Y_i - \beta x_i)$$

is a design-unbiased estimator of $sc(\beta)$. Setting this to zero, yields the estimator $\hat{\beta}_w = \frac{\sum_s w_i Y_i}{\sum_s w_i x_i}$, that is,

$$\hat{\beta}_w = \begin{cases} \frac{\sum_s Y_i / x_i}{n}, & in\ Case\ \ 1 \\ \frac{n}{\sum_s x_i / Y_i}, & in\ Case\ \ 2 \end{cases}.$$

are, as one would expect, (considerably) less efficient than the corresponding maximum likelihood estimators above.

2. *Sample Likelihood* ("Weighted Distribution Maximum Likelihood") (Krieger and Pfeffermann 1992)

The *sample density* of Y_i is the density of Y_i conditional on unit *i* being in the sample:

$$f_s(y_i | x_i) \equiv f(y_i | x_i, I_i = 1) = \frac{f(I_i = 1 | x_i, y_i) f(y_i | x_i)}{f(I_i = 1 | x_i)}.$$

To this there corresponds the sample likelihood: $L_s(\beta) = \prod_s f_s(y_i | x_i)$ and a corresponding score function. In *Case 1* (the non-informative case) we have $f(I_i = 1 | x_i, y_i) = f(I_i = 1 | x_i)$

$$\Rightarrow f_s(y_i | x_i) = f(y_i | x_i)$$

$$\Rightarrow L_s(\beta) = \prod_s f(y_i | x_i) \Rightarrow \hat{\beta}_s = \frac{\sum_s Y_i}{\sum_s x_i},$$ thus getting the same estimator as in the full information case.

In *Case 2* (the informative case), $f_s(y_i | x_i) \equiv f(y_i | x_i, I_i = 1) = \frac{f(I_i = 1 | x_i, y_i) f(y_i | x_i)}{f(I_i = 1 | x_i)}$, with

$f(I_i = 1 | x_i, y_i) = \frac{nY_i}{N\overline{Y}} \approx \frac{nY_i}{N\mu_Y}$, where μ_Y is the marginal mean of *Y*. It follows that $f(I_i = 1 | x_i) \approx \int \frac{nY}{N\mu_Y} f(Y | x_i) dY = \frac{n\beta x_i}{N\mu_Y}$.

Thus

$$f(Y_i | x_i) = Y_i (2\pi x_i \sigma^2)^{-1/2} (\beta x_i)^{-1} \exp\left(-\frac{(Y_i - \beta x_i)^2}{2 x_i \sigma^2} \right),$$

and $sc_s(\beta) = \sigma^{-2} \sum_s (Y_i - \beta x_i) - n/\beta$.

Setting this equal to 0, and multiplying through by β, we get a quadratic equation, with solutions

$$\hat{\beta} = \frac{\sum_s Y_i \pm \sqrt{(\sum_s Y_i)^2 - 4n\sigma^2 \sum_s x_i}}{2\sum_s x_i}.$$

The negative solution corresponds to a *minimum* of the sample likelihood function.

The downward σ adjustment (we here assume for simplicity that σ is known.) under the radical seems appropriate, since *pps* sampling with size variable *y* would tend to select disproportionately the larger *Y*'s. We note the estimate converges to

the *Case 1* solution as $\sigma^2 \to 0$. In our experience, this estimator tends to be more efficient than the pseudo-likelihood estimator and less efficient than the full information maximum likelihood estimator (Chambers et al. 1998). This is certainly true for the informative case here.

2. ML under Informative Stratification

The general regression situation we would consider is

$Y_{hi} = x_{hi}^T \beta + v_{hi}^{1/2} \varepsilon_{hi}$, $h = 1,\ldots,H$, $i = 1,\ldots,N_h$,

with $\varepsilon_{hi} \sim N(0, \sigma^2)$ independently, and independently of the x's.

We follow Schema D2, Krieger and Pfeffermann (1992). Stratification is determined by a variate T, which is possibly in part determined by Y:

$T_{hi} = bY_{hi} + x_{hi}^T c + w_{hi}^{1/2} \eta_{hi} \equiv z_{hi}^T \gamma + w_{hi}^{1/2} \eta_{hi}$, with $\eta_{hi} \sim N(0, \tau^2)$ independent, and independent of the Y's and x's. Strata are determined by the H+1 stratum boundaries

$-\infty = t^{(0)} < t^{(1)} < t^{(2)} < \ldots < t^{(H-1)} < t^{(H)} = \infty$, so that

$t^{(h-1)} < t_i \le t^{(h)} \Rightarrow$ population unit i in stratum h

N_h = number of units in the hth stratum

n_h = number of units sampled from this stratum using SRSWOR

s_h : sample labels in stratum h,

r_h : corresponding non-sample labels.

The population score functions are:

$$sc(\beta) = \sigma^{-2} \sum_{1}^{H} \sum_{i=1}^{N_h} \frac{(Y_{hi} - x_{hi}^T \beta) x_{hi}}{v_{hi}}$$

$$sc(\sigma^2) = -\frac{N}{\sigma^2} + \frac{1}{2} \sum_{1}^{H} \sum_{i=1}^{N_h} \frac{(Y_{hi} - x_{hi}^T \beta)^2}{(\sigma^2)^2 v_{hi}}$$

$$sc(\gamma) = \tau^{-2} \sum_{1}^{H} \sum_{i=1}^{N_h} \frac{(t_{hi} - z_{hi}^T \gamma) z_{hi}}{w_{hi}}$$

$$sc(\tau^2) = -\frac{N}{\tau^2} + \frac{1}{2} \sum_{1}^{H} \sum_{i=1}^{N_h} \frac{(t_{hi} - z_{hi}^T \gamma)^2}{(\tau^2)^2 w_{hi}},$$

the last two equations accommodating the nuisance parameters. To keep technical matters relatively simple, we again focus on the through-the-origin-model, assume the variance constant σ^2 is known, and take Y itself as T, the stratification variable:

Example 2

$Y_{hi} = x_{hi} \beta + x_{hi}^{1/2} \varepsilon_{hi}$, $\varepsilon_{hi} \sim N(0, \sigma^2)$

independently $h = 1,\ldots,H$, $i = 1,\ldots,N_h$,

x_i's are available, $h = 1,\ldots,H$, $i = 1,\ldots,N_h$, $T_i = Y_i$, and the data is thus

$D = (Y_s, x_P, I_P, \{y^{(h)}, n_h, N_h; h = 1,\ldots,H\}, \sigma^2)$.

The population score function is given by $sc(\beta) = \sigma^{-2} \sum_P (Y_i - \beta x_i)$. A weighted sample version of this leads at once to the pseudo-likelihood estimator of β,

$$\hat{\beta}_w = \frac{\sum_h N_h n_h^{-1} \sum_{hi \in s} Y_{hi}}{\sum_h N_h n_h^{-1} \sum_{hi \in s} x_{hi}}.$$

To get the maximum likelihood estimate given the data D, we note

$$sc_D(\beta) = \sigma^{-2} \sum_h \left\{ \sum_{sh} (Y_{hi} - \beta x_{hi}) + \left(\sum_{r_h} E(Y_{hi} \mid y^{(h-1)} \le Y_{hi} \le y^{(h)}, x_{hi}) - \beta x_{hi} \right) \right\}$$

implying

$$\hat{\beta}_D = \left\{ \sum_s Y_i + \sum_h \sum_{r_h} E(Y_{hi} \mid y^{(h-1)} \le Y_{hi} \le y^{(h)}, x_{hi}) \right\} / \sum_P x_l .$$

Thus we need the expectation of the non-sample Y's conditional on their being within the stratum bounds. We have

$$E(Y_{hi} \mid y^{(h-1)} \le Y_{hi} \le y^{(h)}, x_{hi})$$
$$= \int y f(y \mid y^{(h-1)} \le y \le y^{(h)}, x_{hi}) dy$$
$$= \frac{\int_{y^{(h-1)}}^{y^{(h)}} y f(y \mid x_{hi}) dy}{\int_{y^{(h-1)}}^{y^{(h)}} f(y \mid x_{hi}) dy}$$

$$= \beta x_{hi} + x_{hi}^{1/2} \sigma (\varphi(A_{hi}^{(h-1)}) - \varphi A_{hi}^{(h)}) / (\Phi(A_{hi}^{(h)}) - \Phi A_{hi}^{(h-1)}),$$

[eqtn (2)]

where $A_{hi}^{(h^*)} = (y^{(h^*)} - \beta x_{hi}) / \sigma x_{hi}^{1/2}$, and

$\Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) du$,

$\varphi(z) = (2\pi)^{-1/2} \exp\left(-\frac{u^2}{2}\right)$ are the standard normal *cdf* and density respectively (In the simulations described below these were readily calculated in *Splus*© using the functions *pnorm* and *dnorm*.). The derivation of equation (2) is given in *Appendix* 1.

Substituting (2) into the expression above for $\hat{\beta}_D$ we have

$\hat{\beta}_D = \left(\sum_P x_l\right)^{-1} \left[\sum_h \sum_{s_h} Y_{hi} + \sum_h \sum_{r_h} \{\beta x_{hi} + x_{hi}^{1/2} \sigma\left(\varphi\left(A_{hi}^{(h-1)}\right) - \varphi A_{hi}^{(h)}\right) / \left(\Phi\left(A_{hi}^{(h)}\right) - \Phi A_{hi}^{(h-1)}\right)\}\right]$

Since the very quantity β we seek appears on the right hand side (explicitly and also as part of the *A* terms), we proceed iteratively:

$\hat{\beta}_D^{(k+1)} = \left(\sum_P x_l\right)^{-1} \left[\sum_h \sum_{s_h} Y_{hi} + \sum_h \sum_{r_h} \{\beta^{(k)} x_{hi} + x_{hi}^{1/2} \sigma\left(\varphi\left(A_{hi}^{(h-1)k}\right) - \varphi A_{hi}^{(h)k}\right) / \left(\Phi\left(A_{hi}^{(h)k}\right) - \Phi A_{hi}^{(h-1)k}\right)\}\right]$,

with $A_{hi}^{(h^*)k} = \left(y^{(h^*)} - \beta^{(k)} x_{hi}\right) / \sigma x_{hi}^{1/2}$. We begin the process by setting $\hat{\beta}_D^{(0)} = \hat{\beta}_w$.

3. Simulation study

A series of simulation studies was carried out on populations generated in accord with the model of *Example 2*. Each study consisted of 200 runs, in each of which such a population of size was generated having size $N = 1000$, $\beta = 10$, $\sigma = 3$, with the *x*'s the *square* of realizations of a normally distributed random variable *z* (so *x* is positive and skewed). The population was stratified into $H = 2$ strata, bounded by $y^{(h)}$: $-\infty$, median(Y_i), ∞ (so $N_1 = N_2 = 500$.) From these, samples were taken of size $n_1 = 20$, and $n_2 = 80$.

Case 1. Population values available for auxiliary variable x

In this case,

$D = \left(Y_s, x_P, I_P, \left\{y^{(h)}, n_h, N_h; h = 1,..., H\right\}, \sigma^2\right)$.

Simulation results are given in

Table 1

estimator	*means*	*rmse*
pop	10.005	0.114
crude	10.820	0.880
pseudo	10.032	0.438
max1	10.022	0.278
max5	10.006	0.164
max10	10.004	0.162
max20	10.004	0.162

Here "pop" refers to ideal estimation using all the population data, "crude" to an unweighted ratio estimator, "pseudo" to the pseudo-likelihood estimator, and "max*k*" to the *k*th iteration estimate of the maximum likelihood estimator based on equation (2); estimates appear to level out at about the 10th iteration. The maximum likelihood estimator is about 7 times as efficient as the pseudo-likelihood estimator.

Case 2. Population density available for auxiliary variable x

Everything is the same as in *Case 1,* except that the non-sample *x* values are missing, and a density function representing the distribution of the *x*'s is available (in practice, of course, this would be unusual). Thus we have

$D = \left(Y_s, x_s, I_P, \left\{y^{(h)}, n_h, N_h; h = 1,..., H\right\}, \sigma^2, f_x(\ )\right)$,

and the maximum likelihood estimator gets modified to

$\hat{\beta}_D = \frac{\sum_s Y_i + \sum_h (N_h - n_h) E\left(Y_{hi} \mid y^{(h-1)} < Y_{hi} \le y^{(h)}\right)}{\sum_s x_i + \sum_h (N_h - n_h) E\left(x_{hi} \mid y^{(h-1)} < Y_{hi} \le y^{(h)}\right)}$

.

$= \frac{\sum_s Y_i + \sum_h (N_h - n_h) D_h^{-1} (\beta B_h + \sigma C_h)}{\sum_s x_i + \sum_h (N_h - n_h) D_h^{-1} B_h}$,

with

$D_h = \int_{-\infty}^{\infty} \left(\Phi(A_h(x)) - \Phi(A_{h-1}(x))\right) f(x) dx$,

$B_h = \int_{-\infty}^{\infty} x\left(\Phi(A_h(x)) - \Phi(A_{h-1}(x))\right) f(x) dx$

$C_h = \int_{-\infty}^{\infty} x^{1/2} \left(\phi(A_{h-1}(x)) - \phi(A_h(x))\right) f(x) dx$,

$A_h(x) = \left(y^{(h)} - \beta x\right) / \left(\sigma x^{1/2}\right)$. In the case of the squared normal density of our simulations, we have

$$f(x)=\frac{1}{2\sigma_u\sqrt{x}}\left\{\phi\left(\frac{\sqrt{x}-\mu_u}{\sigma_u}\right)+\phi\left(\frac{-\sqrt{x}-\mu_u}{\sigma_u}\right)\right\},$$

where μ_u, σ_u are the mean and standard deviation of the root of *X*, assumed *known.* The integrals were calculated using the function *integrate* in *Splus©.* Simulation results are in

Table 2

estimator	*means*	*rmse*
pop	10.000	0.128
crude	10.870	0.955
pseudo	10.066	0.466
max1	10.057	0.381
max10	10.026	0.242

The maximum likelihood estimator is now about 4 times as efficient as the pseudo-likelihood estimator.

Case 3. Moments, but not the form, of the population density available for auxiliary variable
We assume we know the mean and variance μ_x, σ_x^2 respectively of *x*, but *not* the form of the actual density. On the supposition that *x* has a skewed distribution, we model the density with the (incorrect) gamma distribution with these (correct) moments. Results are given in

Table 3

estimator	*means*	*rmse*
pop	9.988	0.113
crude	10.787	0.841
pseudo	10.007	0.351
max1	10.023	0.282
max*10*	10.055	0.199

The relative efficiency of the maximum likelihood estimator to the pseudo-likelihood estimator is now about 3.

Case 4. Moments available of the population values for the auxiliary variable x.
The form of the density is unknown; we use the gamma density and finite population moments $\bar{x}$, s_x^2 as estimates of the unknown "super-population" parameters μ_x, σ_x^2. Results are in

Table 4

estimates	*means*	*rmse*
pop	9.997	0.114
crude	10.813	0.874
pseudo	10.028	0.390
max1	10.038	0.321
max10	10.057	0.199

The relative efficiency of the maximum likelihood estimator to the pseudo-likelihood estimator is about 4. The improvement over *Case 3* is probably due to random variation. In other words, case 3 and case 4 are roughly equivalent.

Case 5. Mean of the population values for the auxiliary variable x is available.

Here we lack information on the population 2nd moment. We use the gamma density, with the finite population estimate $\bar{x}$ of μ_u, and the weighted estimate of variance $\hat{\sigma}_x^2 = N^{-1}\sum_h \frac{N_h}{n_h}\sum_{s\cap h}\left(x_i-\bar{x}\right)^2$ of σ_x^2. Results are given in

Table 5

estimator	*means*	*rmse*
pop	10.011	0.124
crude	10.830	0.904
pseudo	10.052	0.379
max1	10.054	0.314
max*10*	10.056	0.215

The relative efficiency of the maximum likelihood estimator to the pseudo-likelihood estimator is about 3.

Case 6. No population information on x is available outside the sample.
Results using weighted sample estimates of mean *and* variance of are given in

Table 6

estimator	*means*	*rmse*
pop	9.992	0.116
crude	10.788	0.852
pseudo	9.987	0.416
max1	10.001	0.403
max*10*	10.038	0.452

There is clear deterioration of the maximum likelihood estimator. The weighted estimator indeed seems preferable. However,…

Case 7. Same set-up as Case 6.
We estimate *x*-mean using maximum likelihood estimator, given by

$$\hat{\mu}_x = N^{-1}\left(\sum_s x_i + \sum_h (N_h - n_h) D_h^{-1} B_h\right)$$

$$B_h = \int_{-\infty}^{\infty} x\left(\Phi(A_h(x)) - \Phi(A_{h-1}(x))\right) f(x)dx,$$

using the corresponding weighted estimator as seed. We get new estimates of both $\hat{\beta}$ and $\hat{\mu}_x$ in each iteration. Results are in

Table 7

estimator	*means*	*rmse*
pop	10.011	0.129
crude	10.868	0.928
pseudo	10.052	0.372
max1	10.077	0.354
max*10*	10.071	0.335

This seems to marginally improve things for the maximum likelihood estimator. However, it is clear that without some population information, it is as well to use the pseudo-likelihood estimator.

4. Conclusions

We have constructed the form of the maximum likelihood estimator of a regression coefficient in a particular case of informative stratified sample. The methodology is extendable to the general regression case described at the beginning of Section 2, although formulae can be complicated. The maximum likelihood estimator is much more efficient than the conventional pseudo-likelihood estimator which uses sample weights, when there exists information on the population beyond what is contained in the sample, as is often in practice the case.

Bibliography

Deming, W. E. (1950) *Some Theory of Sampling*, New York: Dover.

Binder, D.A. (1991) Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994) Maximum likelihood inference from sample survey data, *International Statistical Review*, 62, 349-363.

Chambers, R.L., Dorfman, A.H., and Wang, S. (1998) Limited information likelihood analysis of survey data, *Journal of the Royal Statistical Society B*, 60, Part 2, 397-411.

Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*, 54, 127-138.

Kreiger, A B. and Pfeffermann, D. (1992) Maximum likelihood estimation from complex surveys, *Survey Methodology*, 18, 225-239.

Orchard, T. and Woodbury, M.A. (1972) A missing information principle: theory and application. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*, vol 1, 697-715, Berkeley: University of California Press.

Rubin, D.R. (1976) Inference and missing data, *Biometrika*¸ 63, 605-614.

Appendix 1 *Proof of Equation* (2)

$$Den = \int_{y^{(h-1)}}^{y^{(h)}} f(y \mid x_{hi}) dy$$

$$= \left(2\pi x_{hi}\sigma^2\right)^{-1/2} \int_{y^{(h-1)}}^{y^{(h)}} \exp\left(-\frac{(y - x_{hi}\beta)^2}{2x_{hi}\sigma^2}\right) dy$$

$$= \Phi\left(A_{hi}^{(h)}\right) - \Phi\left(A_{hi}^{(h-1)}\right).$$

(gotten by substitution $u = \dfrac{y - x_{hi}\beta}{x_{hi}^{1/2}\sigma}$, etc.)

$$Num = \int_{y^{(h-1)}}^{y^{(h)}} y f(y \mid x_{hi}) dy$$

$$= \left(2\pi x_{hi}\sigma^2\right)^{-1/2} \int_{y^{(h-1)}}^{y^{(h)}} y \exp\left(-\frac{(y - x_{hi}\beta)^2}{2x_{hi}\sigma^2}\right) dy$$

$$= (2\pi)^{-1/2} \int_{A_{hi}^{(h-1)}}^{A_{hi}^{(h)}} \left(x_{hi}\beta + x_{hi}^{1/2}\sigma u\right) \exp\left(-\frac{u^2}{2}\right) du$$

$$= \beta x_{hi}\left(\Phi\left(A_{hi}^{(h)}\right) - \Phi\left(A_{hi}^{(h-1)}\right)\right) + \left(\varphi\left(A_{hi}^{(h-1)}\right) - \varphi\left(A_{hi}^{(h)}\right)\right) x_{hi}^{1/2}\sigma.$$

Hence

$$E\left(Y_{hi} \mid y^{(h-1)} \le Y_{hi} \le y^{(h-1)}, x_{hi}\right) =$$

$$\beta x_{hi} + x_{hi}^{1/2}\sigma\left(\varphi\left(A_{hi}^{(h-1)}\right) - \varphi\left(A_{hi}^{(h)}\right)\right) / \left(\Phi\left(A_{hi}^{(h)}\right) - \Phi\left(A_{hi}^{(h-1)}\right)\right).$$