Analysis of Survey Data. Edited by R. L. Chambers and C. J. Skinner

ISBN: 0-471-89987-9

CHAPTER 12

Fitting Generalized Linear Models under Informative Sampling

Danny Pfeffermann and M. Yu. Sverchkov

12.1. INTRODUCTION

Survey data are frequently used for analytic inference on population models. Familiar examples include the computation of income elasticities from household surveys, the analysis of labor market dynamics from labor force surveys and studies of the relationships between disease incidence and risk factors from health surveys.

The sampling designs underlying sample surveys induce a set of weights for the sample data that reflect unequal selection probabilities, differential nonresponse or poststratification adjustments. When the weights are related to the values of the outcome variable, even after conditioning on the independent variables in the model, the sampling process becomes informative and the model holding for the sample data is different from the model holding in the population (Rubin, 1976; Little, 1982; Sugden and Smith, 1984). Failure to account for the effects of informative sampling may yield large biases and erroneous conclusions. The books edited by Kasprzyk *et al.* (1989) and by Skinner, Holt and Smith (SHS hereafter) (1989) contain a large number of examples illustrating the effects of ignoring informative sampling processes. See also the review articles by Pfeffermann (1993, 1996).

In fairly recent articles by Krieger and Pfeffermann (1997), Pfeffermann, Krieger and Rinott (1998) and Pfeffermann and Sverchkov (1999), the authors propose a new approach for inference on population models from complex survey data. The approach consists of approximating the parametric distribution of the sample data (or moments of that distribution) as a function of the population distribution (moments of this distribution) and the sampling weights, and basing the inference on that distribution. The (conditional) sample

probability density function (*pdf*) $f_s(y_t|x_t)$ of an outcome variable Y, corresponding to sample unit t, is defined as $f(y_t|x_t; t \in s)$ where s denotes the sample and $x_t = (x_{t0}, x_{t1}, \ldots, x_{tk})'$ represents the values of auxiliary variables $X_0 \ldots X_k$ (usually $x_{t0} = 1$ for all t). Denoting the corresponding population *pdf* (before sampling) by $f_U(y_t|x_t)$, an application of Bayes' theorem yields the relationship

$$f_s(y_t|x_t) = f(y_t|x_t; t \in s) = \Pr(t \in s|y_t, x_t) f_U(y_t|x_t)/\Pr(t \in s|x_t). \qquad (12.1)$$

Note that unless $\Pr(t \in s|y_t, x_t) = \Pr(t \in s|x_t)$ for all possible values y_t, the sample and population *pdfs* differ, in which case the sampling process is informative. Empirical results contained in Krieger and Pfeffermann (1997) and Pfeffermann and Sverchkov (1999), based on simulated and real datasets, illustrate the potentially better performance of regression estimators and tests of distribution functions derived from use of the sample distribution as compared to the use of standard inverse probability weighting. For a brief discussion of the latter method (with references) and other more recent approaches that address the problem of informative sampling, see the articles by Pfeffermann (1993, 1996).

The main purpose of the present chapter is to extend the proposed methodology to likelihood-based inference, focusing in particular on situations where the models generating the population values belong to the family of generalized linear models (GLM; McCullagh and Nelder, 1989). The GLM consists of three components:

1. The *random component*; independent observations y_t, each drawn from a distribution belonging to the exponential family with mean μ_t.
2. The *systematic component*; covariates $x_{0t} \ldots x_{kt}$ that produce a linear predictor $\sum_{j=0}^{k} \beta_j x_{jt}$.
3. A *link function* $h(\mu_t) = \sum_{j=0}^{k} \beta_j x_{jt}$ between the random and systematic components.

The unknown parameters consist of the vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$ and possibly additional parameters defining the other moments of the distribution of y_t. A trivial example of the GLM is the classical regression model with the distribution of y_t being normal, in which case the link function is the identity function. Another familiar example considered in the empirical study of this chapter and in Chapters 6, 7, and 8 is logistic regression, in which case the link function is $h(\mu_t) = \log[\mu_t/(1 - \mu_t)]$. See the book by McCullagh and Nelder (1989) for many other models in common use. The GLM is widely used for modeling discrete data and for situations where the expectations μ_t are nonlinear functions of the covariates.

In Section 12.2 we outline the main features of the sample distribution. Section 12.3 considers three plausible sets of estimating equations for fitting the GLM to the sample data that are based on the sample distribution and discusses their advantages over the use of the randomization (design-based) distribution. A simple test statistic for testing the informativeness of the sampling process is developed. Section 12.4 proposes appropriate variance

estimators. Section 12.5 contains simulation results used to assess the performance of the point estimators and the test statistic proposed in Sections 12.3 and 12.4. We close in Section 12.6 with a brief summary and possible extensions.

12.2. POPULATION AND SAMPLE DISTRIBUTIONS

12.2.1. Parametric distributions of sample data

Let $U = \{1 \dots N\}$ define the finite population of size N. In what follows we consider single-stage sampling with selection probabilities $\pi_t = \Pr(t \in s)$, $t = 1 \dots N$. In practice, π_t may depend on the population values $y_U = (y_1 \dots y_N)'$ of the outcome variable, the values $x_U = [x_1 \dots x_N]'$ of the auxiliary variables and values $z_U = [z_1 \dots z_N]'$ of design variables used for the sample selection but not included in the working model under consideration. We express this by writing $\pi_t = g_t(y_U, x_U, z_U)$ for some function g_t.

The probabilities π_t are generally not the same as the probabilities $\Pr(t \in s|y_t, x_t)$ defining the sample distribution in (12.1) because the latter probabilities condition on (y_t, x_t) only. Nonetheless, by regarding π_t as a random variable, the following relationship holds:

$$\Pr(t \in s|y_t, x_t) = E_U(\pi_t|y_t, x_t) \tag{12.2}$$

where $E_U(.)$ defines the expectation operator under the corresponding population distribution. The relationship (12.2) follows by defining I_t to be the sample inclusion indicator and writing $\Pr(t \in s|y_t, x_t) = E(I_t|y_t, x_t) = E[E(I_t|y_u, x_u, z_u)|y_t, x_t] = E(\pi_t|y_t, x_t)$. Substituting (12.2) in (12.1) yields an alternative, and often more convenient, expression for the sample pdf,

$$f_s(y_t|x_t) = E_U(\pi_t|y_t, x_t) f_U(y_t|x_t)/E_U(\pi_t|x_t). \tag{12.3}$$

It follows from (12.3) that for a given population *pdf*, the marginal sample *pdf* is fully determined by the conditional expectation $\pi(y_t, x_t) = E_U(\pi_t|y_t, x_t)$. The paper by Pfeffermann, Krieger and Rinott (1998, hereafter PKR) contains many examples of pairs $[f_U(y, x), \pi(y, x)]$ for which the sample *pdf* is of the same family as the population *pdf*, although possibly with different parameters. In practice, the form of the expectations $\pi(y_t, x_t)$ is often unknown, but it can generally be identified and estimated from the sample data (see below).

For independent population measurements PKR establish an asymptotic independence of the sample measurements with respect to the sample distribution, under commonly used sampling schemes for selection with unequal probabilities. The asymptotic independence assumes that the population size increases holding, the sample size fixed. Thus, the use of the sample distribution permits in principle the application of standard statistical procedures like likelihood-based inference and residual analysis to the sample data.

The representation (12.3) enables also the establishment of relationships between moments of the population and the sample distributions. Denote the expectation operators under the two distributions by E_U and E_s and let

$w_t = 1/\pi_t$ define the sampling weights. Pfeffermann and Sverchkov (1999, hereafter PS) develop the following relationship for pairs of (vector) random variables (u_t, v_t):

$$E_U(u_t|v_t) = E_s(w_t u_t|v_t)/E_s(w_t|v_t). \tag{12.4}$$

For $u_t = y_t; v_t = x_t$, the relationship (12.4) can be utilized for *semi-parametric* estimation of moments of the population distribution in situations where the parametric form of this distribution is unknown. The term semi-parametric estimation refers in this context to the estimation of population moments (for example, regression relationships) from the corresponding sample moments defined by (12.4), using least squares, the method of moments, or other estimation procedures that do not require full specification of the underlying distribution. See PS for examples. Notice, also, that by defining $u_t = \pi_t$ and $v_t = (y_t, x_t)$ or $v_t = x_t$ in (12.4), one obtains

$$E_U(\pi_t|y_t, x_t) = 1/E_s(w_t|y_t, x_t); E_U(\pi_t|x_t) = 1/E_s(w_t|x_t). \tag{12.5}$$

Equation (12.5) shows how the conditional population expectations of π_t that define the sample distribution in (12.3) can be evaluated from the sample data. The relationship (12.4) can be used also for *nonparametric* estimation of regression models; see Chapter 11. The relationships between the population and the sample distributions and between the moments of the two distributions are exploited in subsequent sections.

12.2.2. Distinction between the sample and the randomization distributions

The sample distribution defined by (12.3) is different from the randomization (design) distribution underlying classical survey sampling inference. The randomization distribution of a statistic is the distribution of the values for this statistic induced by all possible sample selections, with the finite population values held fixed. The sample distribution, on the other hand, accounts for the (superpopulation) distribution of the population values and the sample selection process, but the sample units are held fixed. Modeling of the sample distribution requires therefore the specification of the population distribution, but it permits the computation of the joint and marginal distributions of the sample measurements. This is not possible under the randomization distribution because the values of the outcome variable (and possibly also the auxiliary variables) are unknown for units outside the sample. Consequently, the use of this distribution for inference on models is restricted mostly to estimation problems and probabilistic conclusions generally require asymptotic normality assumptions.

Another important advantage of the use of the sample distribution is that it allows conditioning on the values of auxiliary variables measured for the sample units. In fact, the definition of the sample distribution already uses a conditional formulation. The use of the randomization distribution for conditional inference is very limited at present; see Rao (1999) for discussion and illustrations.

The sample distribution is different also from the familiar $p\xi$ distribution, defined as the combined distribution over all possible realizations of the finite population measurements (the population ξ distribution) and all possible sample values for a given population (the randomization p distribution). The $p\xi$ distribution is often used for comparing the performance of design-based estimators in situations where direct comparisons of randomization variances or mean square errors are not feasible. The obvious difference between the sample distribution and the $p\xi$ distribution is that the former conditions on the selected sample (and values of auxiliary variables measured for units in the sample), whereas the latter accounts for all possible sample selections.

Finally, rather than conditioning on the selected sample when constructing the sample distribution (and hence the sample likelihood), one could compute instead the joint distribution of the selected sample and the corresponding sample measurements. Denote by $y_s = \{y_t, t \in s\}$ the outcome variable values measured for the sample units and by $x_s = \{x_t, t \in s\}$ and $x_{\tilde{s}} = \{x_t, t \notin s\}$ the values of the auxiliary variables corresponding to the sampled and nonsampled units. Assuming independence of the population measurements and independent sampling of the population units (Poisson sampling), the joint *pdf* of $(s, y_s)|(x_s, x_{\tilde{s}})$ can be written as

$$f(s, y_s|x_s, x_{\tilde{s}}) = \prod_{t\in s}[\pi(y_t, x_t)f_U(y_t|x_t)/\pi(x_t)]\prod_{t\in s}\pi(x_t)\prod_{t\notin s}[1-\pi(x_t)] \qquad (12.6)$$

where $\pi(y_t, x_t) = E_U(\pi_t|y_t, x_t)$ and $\pi(x_t) = E_U(\pi_t|x_t)$. Note that the product of the terms in the first set of square brackets on the right hand side of (12.6) is the joint sample *pdf*, $f_s(y_s|x_s, s)$, for units in the sample as obtained from (12.3). The use of (12.6) for likelihood-based inference has the theoretical advantage of employing the information on the sample selection probabilities for units outside the sample, but it requires knowledge of the expectations $\pi(x_t) = E_U(\pi_t|x_t)$ for all $t \in U$ and hence the values $x_{\tilde{s}}$. This information is not needed when inference is based on the sample *pdf*, $f_s(y_s|x_s, s)$. When the values $x_{\tilde{s}}$ are unknown, it is possible in theory to regard the values $\{x_t, t \notin s\}$ as random realizations from some *pdf* $g_{\tilde{s}}(x_t)$ and replace the expectations $\pi(x_t)$ for units $t \notin s$ by the unconditional expectations $\pi(t) = \int \pi(x_t)g_{\tilde{s}}(x_t)\mathrm{d}x_t$. See, for example, Rotnitzky and Robins (1997) for a similar analysis in a different context. However, modeling the distribution of the auxiliary variables might be formidable and the resulting likelihood $f(s, y_s|x_s, x_{\tilde{s}})$ could be very cumbersome.

12.3. INFERENCE UNDER INFORMATIVE PROBABILITY SAMPLING

12.3.1. Estimating equations with application to the GLM

In this section we consider four different approaches for defining estimating equations under informative probability sampling. We compare the various approaches empirically in Section 12.5.

Suppose that the population measurements $(y_U, x_U) = \{(y_t, x_t), t = 1 \ldots N\}$ can be regarded as N independent realizations from some *pdf* $f_{y,x}$. Denote by $f_U(y|x;\beta)$ the conditional *pdf* of y_t given x_t. The true value of the vector parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$ is defined as the unique solution of the equations

$$W_U(\beta) = \sum_{t=1}^{N} E_U[d_{Ut}|x_t] = 0 \tag{12.7}$$

where $d_{Ut} = (d_{Ut,0} d_{Ut,1} \ldots d_{Ut,k})' = \partial \log f_U(y_t|x_t;\beta)/\partial\beta$ is the tth score function. We refer to (12.7) as the 'parameter equations' since they define the vector parameter β. For the GLM defined in the introduction with the distribution of y_t belonging to the exponential family, $f_U(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}$ where $a(.) > 0, b(.)$ and $c(.)$ are known functions, and ϕ is known. It follows that $\mu(\theta) = E(y) = \partial b(\theta)/\partial\theta$ so that if $\theta = h(x\beta)$ for some function $h(.)$ with derivative $g(.)$,

$$d_{Ut,j} = \{y_t - \mu(h(x_t'\beta))\}[g(x_t'\beta)]x_{t,j}. \tag{12.8}$$

The 'census parameter' (Binder, 1983) corresponding to (12.7) is defined as the solution B_U of the equations

$$W_U^*(\beta) = \sum_{t=1}^{N} d_{Ut} = 0. \tag{12.9}$$

Note that (12.9) defines the maximum likelihood estimating equations based on all the population values.

Let $s = \{1 \ldots n\}$ denote the sample, assumed to be selected by some probability sampling scheme with first-order selection probabilities $\pi_t = \Pr(t \in s)$. (The sample size n can be random.) The first approach that we consider for specifying the estimating equations involves redefining the parameter equations with respect to the sample distribution $f_s(y_t|x_t)$ (Equation (12.3)), rather than the population distribution as in (12.7). Assuming that the form of the conditional expectations $E_U(\pi_t|y_t, x_t)$ is known (see Section 12.3.2) and that the expectations $E_U(\pi_t|x_t) = \int_y E_U(\pi_t|y, x_t) f_U(y|x_t;\beta)\mathrm{d}y$ are differentiable with respect to β, the parameter equations corresponding to the sample distribution are

$$\begin{aligned} W_{1s}(\beta) &= \sum_s E_s\{[\partial \log f_s(y_t|x_t;\beta)/\partial\beta]|x_t\} \\ &= \sum_s E_s\{[d_{Ut} + \partial \log E_s(w_t|x_t)/\partial\beta]|x_t\} = 0. \end{aligned} \tag{12.10}$$

(The second equation follows from (12.3) and (12.5).) The parameters β are estimated under this approach by solving the equations

$$W_{1s,e}(\beta) = \sum_s [d_{Ut} + \partial \log E_s(w_t|x_t)/\partial\beta] = 0. \tag{12.11}$$

Note that (12.11) defines the sample likelihood equations.

The second approach uses the relationship (12.4) in order to convert the population expectations in (12.7) into sample expectations. Assuming a random sample of size n from the sample distribution, the parameter equations then have the form

$$W_{2s}(\beta) = \sum_s E_s(q_t d_{Ut}|x_t) = 0 \tag{12.12}$$

where $q_t = w_t/E_s(w_t|x_t)$. The vector β is estimated under this approach by solving the equations

$$W_{2s,e}(\beta) = \sum_s q_t d_{Ut} = 0. \tag{12.13}$$

The third approach is based on the property that if β solves Equations (12.7), then it solves also the equations

$$\tilde{W}_U(\beta) = \sum_{t=1}^{N} E_U(d_{Ut}) = E_x\left[\sum_{t=1}^{N} E_U(d_{Ut}|x_t)\right] = 0$$

where the expectation E_x is over the population distribution of the x_t. Application of (12.4) to each of the terms $E_U(d_{Ut})$ (without conditioning on x_t) yields the following parameter equations for a random sample of size n from the sample distribution:

$$W_{3s}(\beta) = \sum_s E_s(w_t d_{Ut})/E_s(w_t) = 0. \tag{12.14}$$

The corresponding estimating equations are

$$W_{3s,e}(\beta) = \sum_s w_t d_{Ut} = 0. \tag{12.15}$$

An interesting feature of the equations in (12.15) is that they coincide with the pseudo-likelihood equations as obtained when estimating the census equations (12.9) by the Horvitz–Thompson estimators. (For the concept and uses of pseudo-likelihood see Binder, 1983; SHS; Godambe and Thompson, 1986; Pfeffermann, 1993; and Chapter 2.) Comparing (12.13) with (12.15) shows that the former equations use the adjusted weights, $q_t = w_t/E_s(w_t|x_t)$ instead of the standard weights w_t used in (12.15). As discussed in PS, the weights q_t account for the net sampling effects on the target conditional distribution of $y_t|x_t$, whereas the weights w_t account also for the sampling effects on the marginal distribution of x_t. In particular, when w is a deterministic function of x so that the sampling process is noninformative, $q_t \equiv 1$ and Equations (12.13) reduce to the ordinary likelihood equations (see (12.16) below). The use of (12.15) on the other hand may yield highly variable estimators in such cases, depending on the variability of the x_t.

The three separate sets of estimating equations defined by (12.11), (12.13), and (12.15) all account for the sampling effects. On the other hand, ignoring the

sampling process results in the use of the ordinary (face value) likelihood equations

$$W_{4s,e}(\beta) = \sum_s d_{Ut} = 0. \tag{12.16}$$

We consider the solution to (12.16) as a benchmark for the assessment of the performance of the other estimators.

Comment The estimating equations proposed in this section employ the scores $d_{Ut} = \partial \log f_U(y_t|x_t;\beta)/\partial\beta$. However, similar equations can be obtained for other functions d_{Ut}; see Bickel *et al.* (1993) for examples of alternative definitions.

12.3.2. Estimation of $E_s(w_t|x_t)$

The estimating equations defined by (12.11) and (12.13) contain the expectations $E_s(w_t|x_t)$ that depend on the unknown parameters β. When the w_t are continuous as in probability proportional to size (PPS) sampling with a continuous size variable, the form of these expectations can be identified from the sample data by the following three-step procedure that utilizes (12.5):

1. Regress w_t against (y_t, x_t) to obtain an estimate of $E_s(w_t|y_t, x_t)$.
2. Integrate $\int_y E_U(\pi_t|y, x_t) f_U(y|x_t;\beta)\mathrm{d}y = \int_y [1/E_s(w_t|y, x_t)] f_U(y|x_t;\beta)\mathrm{d}y$ to obtain an estimate of $E_U(\pi_t|x_t)$ as a function of β.
3. Compute $E_s(w_t|x_t) = 1/E_U(\pi_t|x_t)$.

(The computations in steps 2 and 3 use the estimates obtained in the previous step.) The articles by PKR and PS contain several plausible models for $E_U(\pi_t|y_t, x_t)$ and examples for which the integral in step 2 can be carried out analytically. In practice, however, the specific form of the expectation $E_U(\pi_t|y_t, x_t)$ will usually be unknown but the expectation $E_s(w_t|y_t, x_t)$ can be identified and estimated in this case from the sample. (The expectation depends on unknown parameters that are estimated in step 1.)

Comment 1 Rather than estimating the coefficients indexing the expectation $E_s(w_t|y_t, x_t)$ from the sample (step 1), these coefficients can be considered as additional unknown parameters, with the estimating equations extended accordingly. This, however, may complicate the solution of the estimating equations and also result in identifiability problems under certain models. See PKR for examples and discussion.

Comment 2 For the estimating equations (12.13) that use the weights $q_t = w_t/E_s(w_t|x_t)$, the estimation of the expectation $E_s(w_t|x_t)$ can be carried out by simply regressing w_t against x_t, thus avoiding steps 2 and 3. This is so because in this case there is no need to express the expectation as a function of the parameters β indexing the population distribution.

The discussion so far focuses on the case where the sample selection probabilities are continuous. The evaluation of the expectation $E_s(w_t|x_t)$ in the case of discrete selection probabilities is simpler. For example, in the empirical study of this chapter we consider the case of logistic regression with a discrete independent variable x and three possible values for the dependent variable y. For this case the expectation $E_s(w_t|x_t = k)$ is estimated as

$$\begin{aligned} E_s(w_t|x_t = k) &= 1/E_U(\pi_t|x_t = k), \\ E_U(\pi_t|x_t = k) &= \sum_{a=1}^{3} \Pr_U(y_t = a|x_t = k) E_U(\pi_t|y_t = a, x_t = k) \\ &= \sum_{a=1}^{3} \Pr_U(y_t = a|x_t = k)/E_s(w_t|y_t = a, x_t = k) \\ &\hat{=} \sum_{a=1}^{3} \Pr_U(y_t = a|x_t = k)/\overline{w}_{ak} \end{aligned} \tag{12.17}$$

where $\overline{w}_{ak} = [\sum_s w_t I(y_t = a, x_t = k)]/[\sum_s I(y_t = a, x_t = k)]$. Here $I(A)$ is the indicator function for the event A. Substituting the logistic function for $\Pr(y_t = a|x_t = k)$ in the last expression of (12.17) yields the required specification. The estimators $\overline{w}_{ak}$ are considered as fixed numbers when solving the estimating equations.

For the estimating equations (12.13), the expectations $E_s(w_t|x_t = k)$ in (12.13) in this example are estimated by

$$E_s(w_t|x_t = k) \hat{=} \overline{w}_k = \left[\sum_s w_t I(x_t = k)\right] \Big/ \left[\sum_s I(x_t = k)\right] \tag{12.18}$$

rather than using (12.17) that depends on the unknown logistic coefficients (see Comment 2 above).

For an example of the evaluation of the expectation $E_s(w_t|x_t)$ with discrete selection probabilities but continuous outcome and explanatory variables, see PS (section 5.2).

12.3.3. Testing the informativeness of the sampling process

The estimating equations developed in Section 12.3.1 for the case of informative sampling involve the use of the sampling weights in various degrees of complexity. It is clear therefore that when the sampling process is in fact noninformative, the use of these equations yields more variable estimators than the use of the ordinary score function defined by (12.16). See Tables 12.2 and 12.4 below for illustrations. For the complex sampling schemes in common use, the sample selection probabilities are often determined by the values of several design variables, in which case the informativeness of the selection process is not always apparent. This raises the need for test procedures as a further indication of whether the sampling process is ignorable or not. Several tests have been proposed in the past for this problem. The common

feature of these tests is that they compare the probability-weighted estimators of the target parameters to the ordinary (unweighted) estimators that ignore the sampling process, see Pfeffermann (1993) for review and discussion. For the classical linear regression model, PS propose a set of test statistics that compare the moments of the sample distribution of the regression residuals to the corresponding moments of the population distribution. The use of these tests is equivalent to testing that the correlations under the sample distribution between powers of the regression residuals and the sampling weights are all zero. In Chapter 11 the tests developed by PS are extended to situations where the moments of the model residuals are functions of the regressor variables, as under many of the GLMs in common use.

A drawback of these test procedures is that they involve the use of a series of tests with dependent test statistics, such that the interpretation of the results of these tests is not always clear-cut. For this reason, we propose below a single alternative test that compares the estimating equations that ignore the sampling process to estimating equations that account for it. As mentioned before, the question arising in practice is whether to use the estimating equations (12.16) that ignore the sample selection or one of the estimating equations (12.11), (12.13), or (12.15) that account for it, so that basing the test on these equations is very natural.

In what follows we restrict attention to the comparison of the estimating equations (12.13) and (12.16) (see Comment below). From a theoretical point of view, the sampling process can be ignored for inference if the corresponding parameter equations are equivalent or $\sum_s E_s(d_{Ut}|x_t) = \sum_s E_s(q_t d_{Ut}|x_t)$. Denoting $R(x_t) = E_s(d_{Ut}|x_t) - E_s(q_t d_{Ut}|x_t)$, the null hypothesis is therefore

$$H_0: R_n = n^{-1} \sum_s R(x_t) = 0. \tag{12.19}$$

Note that $\dim(R_n) = k + 1 = \dim(\beta)$. If β were known, the hypothesis could be tested by use of the Hotelling test statistic,

$$H(R) = \frac{n - (k+1)}{k+1} \hat{R}_n' S_n^{-1} \hat{R}_n \sim^{H_0} F_{k+1, n-(k+1)} \tag{12.20}$$

where

$$\hat{R}_n = n^{-1} \sum_s \hat{R}(x_t); \quad \hat{R}(x_t) = (d_{Ut} - q_t d_{Ut}) \text{ and}$$

$$S_n = n^{-1} \sum_s (\hat{R}(x_t) - \hat{R}_n)(\hat{R}(x_t) - \hat{R}_n)'.$$

In practice, β is unknown and the score d_{Ut} in $\hat{R}(x_t)$ has to be evaluated at a sample estimate of β. In principle, any of the estimates defined in Section 12.3.1 could be used for this purpose since under H_0 all the estimators are consistent for β, but we find that the use of the solution of (12.16) that ignores the sampling process is the simplest and yields the best results.

Let $\hat{d}_{Ut}$ define the value of d_{Ut} evaluated at $\hat{\beta}$ – the solution of (12.16) – and let $\tilde{R}(x_t)$, $\tilde{R}_n$, and $\tilde{S}_n$ be the corresponding values of $\hat{R}(x_t)$, $\hat{R}_n$, and S_n obtained after substituting $\hat{\beta}$ for β in (12.20). The test statistic is therefore

$$\tilde{H}(R) = \frac{n-(k+1)}{k+1}\tilde{R}'_n\tilde{S}_n^{-1}\tilde{R}_n \approx^{H_0} F_{k+1,\,n-(k+1)}. \tag{12.21}$$

Note that $\sum_s \hat{d}_{Ut} = 0$ by virtue of (12.16), so $\tilde{R}_n = -n^{-1}\sum_s q_t\hat{d}_{Ut}$. The random variables $q_t\hat{d}_{Ut}$ are no longer independent since $\sum_s \hat{d}_{Ut} = 0$, but utilizing the property that $E_s(q_t|x_t) = 1$ implies that under the null hypothesis $\text{var}_s[\sum_s q_t\hat{d}_{Ut}] = \text{var}_s[\sum_s (q_t\hat{d}_{Ut} - \hat{d}_{Ut})] = \sum_s \text{var}_s(\hat{d}_{Ut} - q_t\hat{d}_{Ut})$, thus justifying the use of $\tilde{S}_n/(n-1)$ as an estimator of $\text{var}(\tilde{R}_n)$ in the construction of the test statistic in (12.21).

Comment The Hotelling test statistic uses the estimating equations (12.13) for the comparison with (12.16) and here again, one could use instead the equations defined by (12.11) or (12.15): that is, replace $q_t d_{Ut}$ in the definition of $\hat{R}(x_t)$ by $d_{Ut} + \partial \log[E_s(w_t|x_t)]/\partial\beta$, or by $w_t d_{Ut}$ respectively. The use of (12.11) is more complicated since it requires evaluation of the expectation $E_s(w_t|x_t)$ as a function of β (see Section 12.3.2). The use of (12.15) is the simplest but it yields inferior results to the use of (12.13) in our simulation study.

12.4. VARIANCE ESTIMATION

Having estimated the model parameters by any of the solutions of the estimating equations in Section 12.3.1, the question arising is how to estimate the variances of these estimators. Unless stated otherwise, the true (estimated) variances are with respect to the sample distribution for a given sample of units, that is, the variance under the *pdf* obtained by the product of the sample *pdfs* (12.3). Note also that since the estimating equations are only for the β-parameters, with the coefficients indexing the expectations $E_s(w_t|y_t, x_t)$ held fixed at their estimators of these values, the first four variance estimators below do not account for the variability of the estimated coefficients.

For the estimator $\hat{\beta}_{1s}$ defined by the solution to the estimating equations (12.11), that is, the maximum likelihood estimator under the sample distribution, a variance estimator can be obtained from the inverse of the information matrix evaluated at this estimator. Thus,

$$\hat{V}(\hat{\beta}_{1s}) = \{-E_s[\partial W_{1s,\,e}(\beta)/\partial\beta']_{\beta=\hat{\beta}_{1s}}\}^{-1}. \tag{12.22}$$

For the estimators $\hat{\beta}_{2s}$ solving (12.13), we use a result from Bickel *et al.* (1993). By this result, if for the true vector parameter β_0, the left hand side of an estimating equation $W_n(\beta) = 0$ can be approximated as $W_n(\beta_0) = n^{-1}\sum_s \varphi(y_t, x_t; \beta_0) + O_p(n^{-1/2})$ for some function φ satisfying $E(\varphi) = 0$ and $E(\varphi^2) < \infty$, then under some additional regularity conditions on the order of convergence of certain functions,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_s [\dot{W}(\beta_0)]^{-1} \varphi(y_t, x_t; \beta_0) + o_p(1) \tag{12.23}$$

where $\hat{\beta}_n$ is the solution of $W_n(\beta) = 0$, $\dot{W}(\beta_0) = [\partial W(\beta)/\partial \beta']_{\beta=\beta_0}$ and $W(\beta) = 0$ is the parameter equation with $\dot{W}(\beta_0)$ assumed to be nonsingular.

For the estimating equations (12.13), $\varphi(y_t, x_t; \beta) = q_t d_{Ut}$, implying that the variance of $\hat{\beta}_{2s}$ solving (12.13) can be estimated as

$$\hat{V}_s(\hat{\beta}_{2s}) = [\dot{W}_{2s,e}(\hat{\beta}_{2s})]^{-1} \left\{ \sum_s [q_t d_{Ut}(\hat{\beta}_{2s})]^2 \right\} [\dot{W}_{2s,e}(\hat{\beta}_{2s})]^{-1} \tag{12.24}$$

where $\dot{W}_{2s,e}(\hat{\beta}_{2s}) = [\partial W_{2s,e}(\beta)/\beta']_{\beta=\hat{\beta}_{2s}}$ and $d_{Ut}(\hat{\beta}_{2s})$ is the value of d_{Ut} evaluated at $\hat{\beta}_{2s}$. Note that since $E_s(q_t d_{Ut} | x_t) = 0$ (and also $\sum_s q_t d_{Ut}(\hat{\beta}_{2s}) = 0$), the estimator (12.24) estimates the conditional variance $V_s(\hat{\beta}_{2s} | \{x_t, t \in s\})$; that is, the variance with respect to the conditional sample distribution of the outcome y.

The estimating equations (12.15) have been derived in Section 12.3.1 by two different approaches, implying therefore two separate variance estimators. Under the first approach, these equations estimate the parameter equations (12.14), which are defined in terms of the unconditional sample expectation (the expectation under the sample distribution of $\{(y_t, x_t), t \in s\}$). Application of the result from Bickel *et al.* (1993) mentioned before yields the following variance estimator (compare with (12.24)):

$$\hat{V}_s(\hat{\beta}_{3s}) = [\dot{W}_{3s,e}(\hat{\beta}_{3s})]^{-1} \left\{ \sum_s [w_t d_{Ut}(\hat{\beta}_{3s})]^2 \right\} [\dot{W}_{3s,e}(\hat{\beta}_{3s})]^{-1} \tag{12.25}$$

where $\dot{W}_{3s,e}(\hat{\beta}_{3s}) = [\partial W_{3s,e}(\beta)/\partial \beta']_{\beta=\hat{\beta}_{3s}}$. For this case $E_s(w_t d_{Ut}) = 0$ (and also $\sum_s w_t d_{Ut}(\hat{\beta}_{3s}) = 0$) so that (12.25) estimates the unconditional variance over the joint sample distribution of $\{(y_t, x_t), t \in s\}$.

Under the second approach, the estimating equations (12.15) are the randomization unbiased estimators of the census equations (12.9). As such, the variance of $\hat{\beta}_{3s}$ can be evaluated with respect to the randomization distribution over all possible sample selections, with the population values held fixed. Following Binder (1983), the randomization variance is estimated as

$$\hat{V}_R(\hat{\beta}_{3s}) = [\hat{\dot{W}}^*_U(\hat{\beta}_{3s})]^{-1} \hat{V}_R \left[\sum_s w_t d_{Ut}(\hat{\beta}_{3s}) \right] [\hat{\dot{W}}^*_U(\hat{\beta}_{3s})]^{-1} \tag{12.26}$$

where $\hat{\dot{W}}^*_U(\hat{\beta}_{3s})$ is design (randomization) consistent for $[\partial W^*_U/\partial \beta']_{\beta=\beta_0}$ and $\hat{V}_R[\sum_s w_t d_{Ut}(\hat{\beta}_{3s})]$ is an estimator of the randomization variance of $\sum_s w_t d_{Ut}(\beta)$, evaluated at $\hat{\beta}_{3s}$.

In order to illustrate the difference between the variance estimators (12.25) and (12.26), consider the case where the sample is drawn by Poisson sampling such that units are selected into the sample independently by Bernoulli trials

with probabilities of success $\pi_t = \Pr(t \in s)$. Simple calculations imply that for this case the randomization variance estimator (12.26) has the form

$$\hat{V}_R(\hat{\beta}_{3s}) = [\hat{\dot{W}}^*_U(\hat{\beta}_{3s})]^{-1}\left\{\sum_s (1-\pi_t)[w_t d_{Ut}(\hat{\beta}_{3s})]^2\right\}[\hat{\dot{W}}^*_U(\hat{\beta}_{3s})]^{-1} \tag{12.27}$$

where $\hat{\dot{W}}^*_U(\hat{\beta}_{3s}) = \dot{W}_{3s,e}(\hat{\beta}_{3s})$. Thus, the difference between the estimator defined by (12.25) and the randomization variance estimator (12.26) is in this case in the weighting of the products $w_t d_{Ut}(\hat{\beta}_{3s})$ by the weights $(1-\pi_t)$ in the latter estimator. Since $0 < (1-\pi_t) < 1$, the randomization variance estimators are smaller than the variance estimators obtained under the sample distribution. This is expected since the randomization variances measure the variation around the (fixed) population values and if some of the selection probabilities are large, a correspondingly large portion of the population is included in the sample (in high probability), thus reducing the variance.

Another plausible variance estimation procedure is the use of bootstrap samples. As mentioned before, under general conditions on the sample selection scheme listed in PKR, the sample measurements are asymptotically independent with respect to the sample distribution, implying that the use of the (classical) bootstrap method for variance estimation is well founded. In contrast, the use of the bootstrap method for variance estimation under the randomization distribution is limited, and often requires extra modifications; see Sitter (1992) for an overview of bootstrap methods for sample surveys. Let $\hat{\beta}_s$ stand for any of the preceding estimators and denote by $\hat{\beta}^b_s$ the estimator computed from bootstrap sample b $(b = 1 \ldots B)$, drawn by simple random sampling with replacement from the original sample (with the same sample size). The bootstrap variance estimator of $\hat{\beta}_s$ is defined as

$$\hat{V}_{boot}(\hat{\beta}_s) = B^{-1}\sum_{b=1}^{B}(\hat{\beta}^b_s - \overline{\beta}_{boot})(\hat{\beta}^b_s - \overline{\beta}_{boot})' \tag{12.28}$$

where

$$\overline{\beta}_{boot} = B^{-1}\sum_{b=1}^{B}\hat{\beta}^b_s.$$

It follows from the construction of the bootstrap samples that the estimator (12.28) estimates the unconditional variance of $\hat{\beta}_s$. A possible advantage of the use of the bootstrap variance estimator in the present context is that it accounts in principle for all the sources of variation. This includes the identification of the form of the expectations $E_s(w_t|y_t, x_t)$ when unknown, and the estimation of the vector coefficient λ indexing that expectation, which is carried out for each of the bootstrap samples but not accounted for by the other variance estimation methods unless the coefficients λ are considered as part of the unknown model parameters (see Section 12.3.2).

12.5. SIMULATION RESULTS

12.5.1. Generation of population and sample selection

In order to assess and compare the performance of the parameter estimators, variance estimators, and the test statistic proposed in Sections 12.3 and 12.4, we designed a Monte Carlo study that consists of the following stages:

A. Generate a univariate population of x-values of size $N = 3000$, drawn independently from the discrete $U[1, 5]$ probability function, $\Pr(X = j) = 0.2, j = 1 \ldots 5$.

B. Generate corresponding y-values from the logistic probability function,

$$\begin{aligned} \Pr(y_t = 1|x_t) &= [\exp(\beta_{10} + \beta_{11}x_t)]/C \\ \Pr(y_t = 2|x_t) &= [\exp(\beta_{20} + \beta_{21}x_t)]/C \\ \Pr(y_t = 3|x_t) &= 1 - \Pr(y_t = 1|x_t) - \Pr(y_t = 2|x_t) \end{aligned} \tag{12.29}$$

where $C = 1 + \exp(\beta_{10} + \beta_{11}x_t) + \exp(\beta_{20} + \beta_{21}x_t)$.

Stages **A** and **B** were repeated independently $R = 1000$ times.

C. From every population generated in stages **A** and **B**, draw a single sample using the following sampling schemes (one sample under each scheme):

Ca. Poisson sampling: units are selected independently with probabilities $\pi_t = nz_t/\sum_{u=1}^{N} z_u$, where $n = 300$ is the expected sample size and the values z_t are computed in two separate ways:

$$\begin{aligned} &\textbf{Ca(1):}\ z_t(1) = Int[(5/9)y_t^2u_t + 2x_t]; \\ &\textbf{Ca(2):}\ z_t(2) = Int[5u_t + 2x_t]. \end{aligned} \tag{12.30}$$

The notation $Int[\cdot]$ defines the integer value and $u_t \sim U(0, 1)$.

Cb. Stratified sampling: the population units are stratified based on either the values $z_t(1)$ (scheme **Cb(1)**) or the values $z_t(2)$ (scheme **Cb(2)**), yielding a total of 13 strata in each case. Denote by $S_{(h)}(j)$ the strata defined by the values $z_t(j)$ such that for units $t \in S_h(j), z_t(j) \equiv z_{(h)}(j), j = 1, 2$. Let $N_{(h)}(j)$ represent the corresponding strata sizes. The selection of units within the strata was carried out by simple random sampling without replacement (SRSWR), with the sample sizes $n_{(h)}(j)$ fixed in advance. The sample sizes were determined so that the selection probabilities are similar to the corresponding selection probabilities under the Poisson sampling scheme and $\sum_h n_h(j) = 300, j = 1, 2$.

The following points are worth noting:

1. The sampling schemes that use the values $z_t(1)$ are *informative*, as the selection probabilities depend on the y-values. The sampling schemes that

use the values $z_t(2)$ are *noninformative* since the selection probabilities depend only on the x-values and the inference is targeted at the population model of the conditional probabilities of $y_t|x_t$ defined by (12.29).

2. For the Poisson sampling schemes, $E_U[\pi_t|\{(y_u, x_u), u = 1 \ldots N\}]$ depends only on the values (y_t, x_t) when $z_t = z_t(1)$, and only on the value x_t when $z_t = z_t(2)$. (With large populations, the totals $\sum_{t=1}^{N} z_t(j)$ can be regarded as fixed.) For the stratified sampling schemes, however, the selection probabilities depend on the strata sizes $N_{(h)}(j)$ that are random (they vary between populations), so that they depend in principle on all the population values $\{(y_t, x_t), t = 1 \ldots N\}$, although with large populations the variation of the strata sizes between populations will generally be minor.
3. The stratified sampling scheme **Cb** corresponds to a *case–control study* whereby the strata are defined based on the values of the outcome variable (y) and possibly some of the covariate variables. See Chapter 8. (Such sampling schemes are known as *choice-based sampling* in the econometric literature.) For the case where the strata are defined based only on y-values generated by a logistic model that contains intercept terms and the sampling fractions within the strata are fixed in advance, it is shown in PKR that the standard MLE of the *slope* coefficients (that is, the MLE that ignores the sampling scheme) coincides with the MLE under the sample distribution. As illustrated in the empirical results below, this is no longer true when the stratification depends also on the x-values. (As pointed out above, under the design **Cb** the sampling fractions within the strata have some small variation. We considered also the case of a stratified sample with fixed sampling fractions within the strata and obtained almost identical results as for the scheme **Cb**.) In order to assess the performance of the estimators derived in the previous sections in situations where the stratification depends only on the y-values, we consider also a third stratified sampling scheme:

   Cc. Stratified sampling: The population units are stratified based on the values y_t (three strata); select n_h units from stratum h by SRSWR with the sample sizes fixed in advance, $\sum_h n_h = 300$.

12.5.2. Computations and results

The estimators of the logistic model coefficients $\beta_k = \{(\beta_{k0}, \beta_{k1}), k = 1, 2\}$ in (12.29), obtained by solving the estimating equations (12.11), (12.13), (12.15), and (12.16), have been computed for each of the samples drawn by the sampling methods described in Section 12.5.1, yielding four separate sets of estimators. The expectations $E_s(w_t|x_t)$ have been estimated using the procedures described in Section 12.3.2. For each point estimator we computed the corresponding variance estimator as defined by (12.22), (12.24), and (12.25) for the first three estimators and by use of the inverse information matrix (ignoring the sampling process) for the ordinary MLE. In addition, we computed for each of the point estimators the bootstrap variance estimator defined by (12.27). Due

to computation time constraints, the bootstrap variance estimators are based on only 100 samples for each of 100 parent samples. Finally, we computed for each sample the Hotelling test statistic (12.21) for sample informativeness developed in Section 12.3.3.

The results of the simulation study are summarized in Tables 12.1–12.5. These tables show for each of the five sampling schemes and each point estimator the mean estimate, the empirical standard deviation (Std) and the mean of the Std estimates (denoted Mean Std est. in the tables) over the 1000 samples; and the mean of the bootstrap Std estimates (denoted Mean Std est. Boot in the tables) over the 100 samples. Table 12.6 compares the theoretical and empirical distribution of the Hotelling test statistic under H_0 for the two noninformative sampling schemes defined by **Ca(2)** and **Cb(2)**.

The main conclusions from the results set out in Tables 12.1–12.5 are as follows:

Table 12.1 Means, standard deviations (Std) and mean Std estimates of logistic regression coefficients. Poisson sampling, informative scheme **Ca (1)**.

Coefficients	β_{10}	β_{11}	β_{20}	β_{21}
MLE (sample *pdf*, Equation (12.11)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.03	0.30	0.54	0.50
Empirical Std	0.62	0.18	0.61	0.18
Mean Std est.	0.58	0.18	0.58	0.18
Mean Std est. Boot	0.56	0.18	0.56	0.18
Q-weighting (Equation (12.13)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.99	0.31	0.51	0.51
Empirical Std	0.63	0.19	0.63	0.19
Mean Std est.	0.59	0.18	0.59	0.18
Mean Std est. Boot	0.68	0.21	0.66	0.20
W-weighting (Equation (12.15)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.00	0.31	0.53	0.51
Empirical Std	0.67	0.21	0.67	0.20
Mean Std est.	0.63	0.19	0.62	0.19
Mean Std est. Boot	0.65	0.21	0.64	0.20
Ordinary MLE (Equation (12.16)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.29	0.43	0.06	0.59
Empirical Std	0.60	0.19	0.60	0.18
Mean Std est.	0.59	0.18	0.58	0.18
Mean Std est. Boot	0.57	0.18	0.58	0.18

Table 12.2 Means, standard deviations (Std) and mean Std estimates of logistic regression coefficients. Poisson sampling, noninformative scheme **Ca (2)**.

Coefficients	β_{10}	β_{11}	β_{20}	β_{21}
MLE (sample *pdf*, Equation (12.11)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.01	0.31	0.53	0.50
Empirical Std	0.66	0.20	0.67	0.20
Mean Std est.	0.62	0.20	0.62	0.20
Mean Std est. Boot	0.62	0.20	0.62	0.20
Q-weighting (Equation (12.13)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.99	0.31	0.51	0.51
Empirical Std	0.67	0.20	0.68	0.20
Mean Std est.	0.63	0.19	0.63	0.20
Mean Std est. Boot	0.66	0.22	0.66	0.22
W-weighting (Equation (12.15)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.00	0.31	0.52	0.51
Empirical Std	0.71	0.22	0.72	0.22
Mean Std est.	0.65	0.21	0.66	0.21
Mean Std est. Boot	0.68	0.22	0.69	0.22
Ordinary MLE (Equation (12.16)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.99	0.31	0.50	0.51
Empirical Std	0.63	0.19	0.63	0.19
Mean Std est.	0.60	0.19	0.61	0.19
Mean Std est. Boot	0.62	0.20	0.63	0.20

1. The three sets of estimating equations defined by (12.11), (12.13), and (12.15) perform well in eliminating the sampling effects under informative sampling. On the other hand, ignoring the sampling process and using the ordinary MLE (Equation (12.16)) yields highly biased estimators.
2. The use of Equations (12.11) and (12.13) produces very similar results. This outcome, found also in PS for ordinary regression analysis, is important since the use of (12.13) is much simpler and requires fewer assumptions than the use of the full equations defined by (12.11); see also Section 12.3.2. The use of simple weighting (Equation (12.15)) that corresponds to the application of the pseudo-likelihood approach again performs well in eliminating the bias, but except for Table 12.5 the variances under this approach are consistently larger than under the first two approaches, illustrating the discussion in Section 12.3.1. On the other hand, the ordinary MLE that does not involve any weighting has in most cases the smallest standard deviation. The last outcome is known also from other studies.

Table 12.3 Means, standard deviations (Std) and mean Std estimates of logistic regression coefficients. Stratified sampling, informative scheme **Cb (1)**.

Coefficients	β_{10}	β_{11}	β_{20}	β_{21}
MLE (sample *pdf*, Equation (12.11)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.05	0.30	0.53	0.51
Empirical Std	0.56	0.17	0.59	0.17
Mean Std est.	0.56	0.17	0.56	0.17
Mean Std est. Boot	0.50	0.17	0.50	0.17
Q-weighting (Equation (12.13)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.05	0.30	0.53	0.50
Empirical Std	0.55	0.16	0.58	0.17
Mean Std est.	0.58	0.18	0.58	0.18
Mean Std est. Boot	0.66	0.21	0.64	0.20
W-weighting (Equation (12.15)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.07	0.29	0.55	0.50
Empirical Std	0.58	0.18	0.61	0.18
Mean Std est.	0.61	0.19	0.61	0.19
Mean Std est. Boot	0.63	0.21	0.63	0.21
Ordinary MLE (Equation (12.16)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.37	0.41	0.16	0.56
Empirical Std	0.52	0.15	0.55	0.16
Mean Std est.	0.57	0.18	0.57	0.18
Mean Std est. Boot	0.52	0.18	0.52	0.18

3. The MLE variance estimator (12.22) and the semi-parametric estimators (12.24) and (12.25) underestimate in most cases the true (empirical) variance with an underestimation of less than 8%. As anticipated in Section 12.4, the use of the bootstrap variance estimators corrects for this underestimation by better accounting for all the sources of variation, but this only occurs with the estimating equations (12.13) and (12.15). We have no clear explanation for why the bootstrap variance estimators perform less satisfactory for the MLE equations defined by (12.11) and (12.16) but we emphasize again that we have only used 100 bootstrap samples for the variance estimation, which in view of the complexity of the estimating equations is clearly not sufficient. It should be mentioned also in this respect that the standard deviation estimators (12.22), (12.24), and (12.25) are more stable (in terms of their standard deviation) than the bootstrap standard deviation estimators, which again can possibly be attributed to the relatively small number of bootstrap samples. (The standard deviations of the standard deviation estimators are not shown.)

Table 12.4 Means, standard deviations (Std) and mean Std estimates of logistic regression coefficients. Stratified sampling, noninformative scheme **Cb(2)**.

Coefficients	β_{10}	β_{11}	β_{20}	β_{21}
MLE (sample *pdf*, Equation (12.11)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.06	0.29	0.53	0.50
Empirical Std	0.63	0.19	0.64	0.20
Mean Std est.	0.60	0.19	0.61	0.19
Mean Std est. Boot	0.56	0.20	0.57	0.20
Q-weighting (Equation (12.13)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.04	0.30	0.51	0.51
Empirical Std	0.63	0.20	0.64	0.20
Mean Std est.	0.60	0.19	0.61	0.19
Mean Std est. Boot	0.62	0.22	0.63	0.22
W-weighting (Equation (12.15)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.05	0.30	0.52	0.51
Empirical Std	0.66	0.21	0.67	0.21
Mean Std est.	0.62	0.20	0.63	0.20
Mean Std est. Boot	0.63	0.22	0.65	0.23
Ordinary MLE (Equation (12.16)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.03	0.30	0.49	0.52
Empirical Std	0.60	0.19	0.61	0.20
Mean Std est.	0.58	0.19	0.59	0.19
Mean Std est. Boot	0.56	0.20	0.57	0.20

4. Our last comment refers to Table 12.5 that relates to the sampling scheme **Cc** by which the stratification is based only on the y-values. For this case the first three sets of parameter estimators and the two semi-parametric variance estimators perform equally well. Perhaps the most notable outcome from this table is that ignoring the sampling process in this case and using Equations (12.16) yields similar mean estimates (with smaller standard deviations) for the two slope coefficients as the use of the other equations. Note, however, the very large biases of the intercept estimators. As mentioned before, PKR show that the use of (12.16) yields the correct MLE for the slope coefficients under Poisson sampling, which is close to the stratified sampling scheme underlying this table.

Table 12.6 compares the empirical distribution of the Hotelling test statistic (12.21) over the 1000 samples with the corresponding nominal levels of the theoretical distribution for the two noninformative sampling schemes **Ca(2)** and **Cb(2)**. As can be seen, the empirical distribution matches almost perfectly the theoretical distribution. We computed the test statistic also under the

Table 12.5 Means, standard deviations (Std) and mean Std estimates of logistic regression coefficients. Stratified sampling, informative scheme **Cc**.

Coefficients	β_{10}	β_{11}	β_{20}	β_{21}
MLE (sample *pdf*, Equation (12.11)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.00	0.32	0.48	0.53
Empirical Std	0.42	0.16	0.42	0.15
Mean Std est.	0.40	0.14	0.38	0.14
Mean Std est. Boot	0.39	0.15	0.37	0.14
Q-weighting (Equation (12.13)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.99	0.31	0.48	0.52
Empirical Std	0.42	0.16	0.42	0.15
Mean Std est.	0.43	0.15	0.42	0.15
Mean Std est. Boot	0.44	0.16	0.43	0.16
W-weighting (Equation (12.15)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	1.00	0.31	0.48	0.52
Empirical Std	0.42	0.16	0.42	0.15
Mean Std est.	0.43	0.15	0.42	0.15
Mean Std est. Boot	0.44	0.16	0.43	0.16
Ordinary MLE (Equation (12.16)):				
True values	1.00	0.30	0.50	0.50
Mean estimate	0.01	0.31	−0.04	0.52
Empirical Std	0.36	0.14	0.36	0.13
Mean Std est.	0.40	0.14	0.38	0.14
Mean Std est. Boot	0.39	0.15	0.37	0.14

Table 12.6 Nominal levels and empirical distribution of Hotelling test statistic under H_0 for noninformative sampling schemes **Ca(2)** and **Cb(2)**.

Nominal levels	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
Emp. dist (**Ca(2)**)	0.01	0.02	0.04	0.09	0.89	0.95	0.98	0.99
Emp. dist (**Cb(2)**)	0.01	0.03	0.05	0.08	0.91	0.96	0.98	0.99

three informative sampling schemes and in all the 3×1000 samples, the null-hypothesis of noninformativeness of the sampling scheme had been rejected at the 1% significance level, indicating very high power.

12.6. SUMMARY AND EXTENSIONS

This chapter considers three alternative approaches for the fitting of GLM under informative sampling. All three approaches utilize the relationships

between the population distribution and the distribution of the sample observations as defined by (12.1), (12.3), and (12.4) and they are shown to perform well in eliminating the bias of point estimators that ignore the sampling process. The use of the pseudo-likelihood approach, derived here under the framework of the sample distribution, is the simplest, but it is shown to be somewhat inferior to the other two approaches. These two approaches require the modeling and estimation of the expectation of the sampling weights, either as a function of the outcome and the explanatory variables, or as a function of only the explanatory variables. This additional modeling is not always trivial but general guidelines are given in Section 12.3.2. It is important to emphasize in this respect that the use of the sample distribution as the basis for inference permits the application of standard model diagnostic tools so that the goodness of fit of the model to the sample data can be tested.

The estimating equations developed under the three approaches allow the construction of variance estimators based on these equations. These estimators have a small negative bias since they fail to account for the estimation of the expectations of the sampling weights. The use of the bootstrap method that is well founded under the sample distribution overcomes this problem for two of the three approaches, but the bootstrap estimators seem to be less stable. Finally, a new test statistic for the informativeness of the sampling process that compares the estimating equations that account for the sampling process with estimating equations that ignore it is developed and shown to perform extremely well in the simulation study.

An important use of the sample distribution not considered in this chapter is for prediction problems. Notice first that if the sampling process is informative, the model holding for the outcome variable for units outside the sample is again different from the population model. This implies that even if the population model is known with all its parameters, it cannot be used directly for the prediction of outcome values corresponding to nonsampled units. We mention in this respect that the familiar 'model-dependent estimators' of finite population totals assume noninformative sampling. On the other hand, it is possible to derive the distribution of the outcome variable for units outside the sample, similarly to the derivation of the sample distribution, and then obtain the optimal predictors under this distribution. See Sverchkov and Pfeffermann (2000) for application of this approach.

Another important extension is to two-level models with application to small-area estimation. Here again, if the second-level units (schools, geographic areas) are selected with probabilities that are related to the outcome values, the model holding for the second-level random effects might be different from the model holding in the population. Failure to account for the informativeness of the sampling process may yield biased estimates for the model parameters and biased predictions for the small-area means. Appropriate weighting may eliminate the first bias but not the second. Work on the use of the sample distribution for small-area estimation is in progress.