MAXIMUM LIKELIHOOD WITH AUXILIARY INFORMATION

RAY CHAMBERS, SUOJIN WANG

ABSTRACT

Analysis of survey data does not happen in a vacuum. We typically know more about the target population than just the data observed in the survey. In some cases this extra information can be incorporated via calibration of survey weights. However, model fitting using weights often leads to increased standard errors. Also, weights are usually calibrated to a relatively small set of variables, while population data may be known for many more variables. Here we use the general approach to maximum likelihood estimation for complex surveys described in Breckling et al. (1994) to develop methods for efficiently incorporating external population information into model fitting using survey data. In particular, we focus on two simple, but very popular, models fitted to survey data. These are the linear regression model and the logistic regression model.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M06/08

Maximum Likelihood With Auxiliary Information

Ray Chambers

Centre for Statistical and Survey Methodology, University of Wollongong

(ray@uow.edu.au)

&

Suojin Wang

Department of Statistics, Texas A&M University

(sjwang@stat.tamu.edu)

May 2005

Abstract

Analysis of survey data does not happen in a vacuum. We typically know more about the target population than just the data observed in the survey. In some cases this extra information can be incorporated via calibration of survey weights. However, model fitting using weights often leads to increased standard errors. Also, weights are usually calibrated to a relatively small set of variables, while population data may be known for many more variables. Here we use the general approach to maximum likelihood estimation for complex surveys described in Breckling et. al. (1994) to develop methods for efficiently incorporating external population information into model fitting using survey data. In particular, we focus on two simple, but very popular, models fitted to survey data. These are the linear regression model and the logistic regression model.

1. Introduction

Analysis of survey data does not happen in a vacuum. A model for the number of children ever born to a woman from a particular target population could depend on a number of factors, e.g. her age, her education level, her labour force status, her household income, her ethnic background and her access to family planning information, perhaps measured by presence or absence of a family planning clinic within a specified distance of her home. All of these variables are measured for women taking part in the survey, and the classical approach is to consider them 'in isolation' in the modelling process, implicitly assuming that the model fitted to these sample data is also appropriate for the population from which the sample is drawn. Sometimes, if this is felt to be too big an assumption, and survey weights are available, these are included in the model fitting process, assuming that they correct the parameter estimation process for potential sample selection bias.

However, we typically know a lot more about the target population than just the data observed in the survey. In particular we may know the total number of women in the population, their average number of children, their average age, their labour force participation rate and their ethnic distribution in the population. By 'know' here we mean either the actual population value or at least an accurate estimate. The question here is how to integrate this auxiliary population information into the model fitting process described above.

In some cases, this information is incorporated in the survey weights, through the process of calibration (Deville and Särndal, 1992). That is, these weights are constructed so that weighted averages for selected variables measured in the survey equal corresponding known (or highly accurate estimates of) population values. One approach to using this auxiliary information would therefore be to use such calibrated weights in estimation. However, this has two major problems. First, such weights typically lead to increased standard errors compared to unweighted analysis.

Second, weights are usually calibrated to a fixed and relatively small set of variables (e.g. age by sex population distributions, regional population distributions), while population data are often known for many more variables.

Alternative, more model-based, ways of incorporating auxiliary population information when modelling survey data have been explored in the econometrics literature, mainly in the context of analysis of linked data sets. An early example is Imbens and Lancaster (1994), who suggest a generalised method of moments approach to the problem of incorporating knowledge of the population expected value of the response variable Y into a sample-based linear regression of Y on an explanatory variable X. More recently, Qin (2000) has considered the same problem using a combination of empirical and parametric likelihood.

This paper focuses on developing methods for efficiently using auxiliary population information when survey data are used to fit a statistical model for a target population. In particular, we look at how maximum likelihood methods can be modified to incorporate this information. The approach we take is based on the general approach to maximum likelihood estimation for complex surveys described in Breckling *et. al.* (1994), hereafter referred to as BCDTW. In particular, we focus on two simple, but very popular, models fitted to survey data. These are the linear regression model and the linear logistic regression model.

2. MLE for a linear model given auxiliary population information

Consider the following situation. A sample survey measures the values y_i and x_i of two scalar variables, Y and X respectively, for a sample s of n units from a population U of N units. The variable X is a population covariate, i.e. we know the values of X for every unit in the population and the sampling method is non-informative given these values. Our aim is to use the

sample survey data to fit a simple normal linear model to the population values of *Y* and *X*. That is, we want to use the survey data to estimate the parameters α, β and σ^2 that characterise the population model

$$\sigma_i^{-1}(y_i - \alpha - \beta x_i) \sim iid\ N(0,1). \quad (1)$$

Given this set-up, the maximum likelihood estimates (MLEs) for α, β and σ^2 are

$$\hat{\beta}_{smle} = \left(\sum_s x_i(x_i - \bar{x}_s)\right)^{-1} \sum_s x_i(y_i - \bar{y}_s)$$

$$\hat{\alpha}_{smle} = \bar{y}_s - \hat{\beta}_{smle}\bar{x}_s$$

$$\hat{\sigma}^2_{smle} = n^{-1}\sum_s (y_i - \hat{\alpha}_{ols} - \hat{\beta}_{ols}x_i)^2 .$$

We use a subscript of *smle* above to indicate that these MLEs are just based on the *sample* values of *Y* and *X*. However, suppose we also know the population mean $\bar{y}_U$ of *Y*. This can happen, for example, if the variable *Y* is also measured in a census, and census tabulations are published. In this case the OLS estimators above are no longer the MLEs for α, β and σ^2. In order to obtain the 'full information' MLEs that include this additional information, we first observe that the population level score function for $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$ is defined by the components

$$sc_1(\boldsymbol{\theta}) = \sigma^{-2}\sum_U (y_i - \alpha - \beta x_i) \quad (2a)$$

$$sc_2(\boldsymbol{\theta}) = \sigma^{-2}\sum_U x_i(y_i - \alpha - \beta x_i) \quad (2b)$$

$$sc_3(\boldsymbol{\theta}) = -N/2\sigma^2 + \sum_U (y_i - \alpha - \beta x_i)^2 / 2\sigma^4 . \quad (2c)$$

In what follows we let E_s and Var_s denote the expectation and variance operators respectively that condition on the 'available data' for use in analysis. In this case these data correspond to the sample values of *Y* and *X*, the non-sample values of *X* and the population mean of *Y*. We refer to the score function for α, β and σ^2 given these data as the *full information* score function for

these parameters. BCDTW show that this full information score function is the conditional expectation of the corresponding population level score function given these data. Denoting the components of this full information score function by an additional subscript of *s*, we have

$$sc_{1s}(\boldsymbol{\theta}) = \sigma^{-2}\sum\nolimits_U (E_s(y_i) - \alpha - \beta x_i) \tag{3a}$$

$$sc_{2s}(\boldsymbol{\theta}) = \sigma^{-2}\sum\nolimits_U x_i(E_s(y_i) - \alpha - \beta x_i) \tag{3b}$$

$$sc_{3s}(\boldsymbol{\theta}) = -N/2\sigma^2 + \left[\sum\nolimits_U (E_s(y_i) - \alpha - \beta x_i)^2 + \sum\nolimits_U Var_s(y_i)\right]/2\sigma^4 . \tag{3c}$$

Since $E_s(y_i) = y_i$ and $Var_s(y_i) = 0$ for sampled population units, all we need to do is to determine these conditional moments for population units not in sample. To do this, we note that for non-sample unit *i*,

$$\begin{pmatrix} y_i \\ \bar{y}_r \end{pmatrix} \Big| \mathbf{x}_U \sim N\left[\begin{pmatrix} \alpha + \beta x_i \\ \alpha + \beta \bar{x}_r \end{pmatrix}, \begin{bmatrix} \sigma^2 & (N-n)^{-1}\sigma^2 \\ (N-n)^{-1}\sigma^2 & (N-n)^{-1}\sigma^2 \end{bmatrix}\right].$$

Here $\mathbf{x}_U$ denotes the population values of *X*, $\bar{y}_r$ denotes the non-sample population average of *Y* and $\bar{x}_r$ denotes the corresponding non-sample average of *X*. Hence

$$y_i \,|\, \mathbf{x}_U, \bar{y}_r \sim N\left[\bar{y}_r + \beta(x_i - \bar{x}_r), \sigma^2\left(1 - (N-n)^{-1}\right)\right]. \tag{4}$$

Combining (3) and (4) leads to

$$sc_{1s}(\boldsymbol{\theta}) = \sigma^{-2}\left[\sum\nolimits_s (y_i - \alpha - \beta x_i) + (N-n)(\bar{y}_r - \alpha - \beta \bar{x}_r)\right] \tag{5a}$$

$$sc_{2s}(\boldsymbol{\theta}) = \sigma^{-2}\left[\sum\nolimits_s x_i(y_i - \alpha - \beta x_i) + (N-n)\bar{x}_r(\bar{y}_r - \alpha - \beta \bar{x}_r)\right] \tag{5b}$$

$$sc_{3s}(\boldsymbol{\theta}) = -(n+1)/2\sigma^2 + \left[\sum\nolimits_s (y_i - \alpha - \beta x_i)^2 + (N-n)(\bar{y}_r - \alpha - \beta \bar{x}_r)^2\right]/2\sigma^4 . \tag{5c}$$

Setting these score components to zero and solving for α, β and σ^2 gives the full information MLEs in this case. They are

$$\hat{\beta}_{fimle} = \frac{\sum_s x_i(y_i - \bar{y}_s) + n\bar{x}_s(\bar{y}_s - \bar{y}_U) + (N-n)\bar{x}_r(\bar{y}_r - \bar{y}_U)}{\sum_s x_i(x_i - \bar{x}_s) + n\bar{x}_s(\bar{x}_s - \bar{x}_U) + (N-n)\bar{x}_r(\bar{x}_r - \bar{x}_U)} \tag{6a}$$

$$\hat{\alpha}_{fimle} = \bar{y}_U - \hat{\beta}_{fimle}\bar{x}_U \tag{6b}$$

$$\hat{\sigma}^2_{fimle} = (n+1)^{-1}\sum_s (y_i - \hat{\alpha}_{fimle} - \hat{\beta}_{fimle}x_i)^2 + (N-n)(\bar{y}_r - \hat{\alpha}_{fimle} - \hat{\beta}_{fimle}\bar{x}_r)^2 . \tag{6c}$$

These estimators are identical to the estimators defined by a weighted least squares (WLS) fit to an extended sample consisting of the data values in *s* (each with weight equal to one) plus an additional data value (with weight equal to $N - n$) defined by the non-sample means $\bar{y}_r$ and $\bar{x}_r$.

Intuitively, one expects the extra information from knowing $\bar{y}_U$ to contribute mainly to estimation of α in (1). To see that this is the case we now write down the variances of (6a) and (6b). This can be done by differentiating the score functions (5), changing signs and evaluating at the MLEs (6) to get the observed information matrix for these parameters. This matrix can then be inverted to get the (asymptotic) variances and covariances of these MLEs. Alternatively, exploiting their equivalence to a WLS fit, we can obtain the variances of the regression coefficients (6a) and (6b) directly. These are

$$Var(\hat{\alpha}_{fimle}) = n^{-1}\sigma^2\left(\frac{\bar{x}_s^{(2)} - (1 - nN^{-1})(\bar{x}_s^{(2)} - \bar{x}_r^2)}{\bar{x}_s^{(2)} - \bar{x}_r^2 + Nn^{-1}(\bar{x}_r^2 - \bar{x}_U^2)}\right)$$

$$Var(\hat{\beta}_{fimle}) = \frac{n^{-1}\sigma^2}{\bar{x}_s^{(2)} - \bar{x}_r^2 + Nn^{-1}(\bar{x}_r^2 - \bar{x}_U^2)} .$$

Here $\bar{x}_s^{(2)}$ is the mean of the squares of the sample *X*-values. In an *X*-balanced sample ($\bar{x}_s = \bar{x}_r = \bar{x}_U$) it is easy to see that $Var(\hat{\beta}_{fimle}) = Var(\hat{\beta}_{smle})$ while $Var(\hat{\alpha}_{fimle}) = Var(\hat{\alpha}_{smle}) - n^{-1}(1 - nN^{-1})\sigma^2$, confirming our intuition above.

As noted earlier, the full information MLE approach used to derive (6) is not necessarily the only way one might attempt to use the fact that we know $\bar{y}_U$. From a survey estimation point

of view, the situation set out above is one where we have three calibration identities. We know the population size *N*, population total of *X* and the population total of *Y*. We could therefore calibrate the survey weights to recover these population totals. That is, if w_i denotes the initial survey weight for sample unit *i* (e.g. the inverse of its sample inclusion probability), we replace this weight by w_i^*, where $\sum_s w_i^* = N$, $\sum_s w_i^* x_i = N\bar{x}_U$ and $\sum_s w_i^* y_i = N\bar{y}_U$. There are standard methods for doing this (e.g. Deville and Särndal, 1992; Chambers, 1996). For simple random sampling, a least squares calibration criterion leads to weights $\mathbf{w}^* = (w_i^*)$, where

$$\mathbf{w}^* = \frac{N}{n}\mathbf{1}_n + N[\mathbf{1}_n\ \mathbf{y}_s\ \mathbf{x}_s]\begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\mathbf{y}_s & \mathbf{1}_n'\mathbf{x}_s \\ \mathbf{y}_s'\mathbf{1}_n & \mathbf{y}_s'\mathbf{y}_s & \mathbf{y}_s'\mathbf{x}_s \\ \mathbf{x}_s'\mathbf{1}_n & \mathbf{x}_s'\mathbf{y}_s & \mathbf{x}_s'\mathbf{x}_s \end{bmatrix}^{-1}\begin{pmatrix} 0 \\ \bar{y}_U - \bar{y}_s \\ \bar{x}_U - \bar{x}_s \end{pmatrix}.$$

Here $\mathbf{1}_n$ denotes an *n*-vector of ones, and $\mathbf{y}_s$, $\mathbf{x}_s$ are vectors containing the sample values of *Y* and *X* respectively. The calibrated weights are then used to estimate α, β and σ^2 by weighted least squares. That is, we estimate these parameters via

$$\hat{\beta}_{calw} = \left(\sum_s w_i^* x_i (x_i - \bar{x}_{ws})\right)^{-1} \sum_s w_i^* x_i (y_i - \bar{y}_{ws}) \qquad (7a)$$

$$\hat{\alpha}_{calw} = \bar{y}_{ws} - \hat{\beta}_{calw}\bar{x}_{ws} \qquad (7b)$$

$$\hat{\sigma}^2_{calw} = N^{-1}\sum_s w_i^* (y_i - \hat{\alpha}_{calw} - \hat{\beta}_{calw} x_i)^2 . \qquad (7c)$$

Here $\bar{y}_{ws} = \sum_s w_i^* y_i / \sum_s w_i^* = \bar{y}_U$ and $\bar{x}_{ws} = \sum_s w_i^* x_i / \sum_s w_i^* = \bar{x}_U$.

Although use of calibrated weights may seem natural from a survey statistician's point of view, such an approach is not the most obvious if one considers the problem from a standard statistical modelling perspective. Here it makes sense to incorporate our population information (the values of $\bar{y}_U$ and $\bar{x}_U$) via constraints on the estimates of the parameters of interest. Under (1) $E(Y) = \alpha + \beta E(X)$, so an obvious constraint is $\bar{y}_U = \hat{\alpha} + \hat{\beta}\bar{x}_U$. This is the general approach

described in Handcock, Rendall and Cheadle (2005), where the likelihood generated by the sample values of *Y* and *X* is maximised subject to this constraint. In the context of (1) this is the same as estimating α and β by minimising the sum of squared errors subject to this constraint. It is not difficult to see that this leads to the estimators

$$\hat{\beta}_{con} = \left[\sum_s (x_i - \bar{x}_r)^2 + n(\bar{x}_s - \bar{x}_U)^2\right]^{-1}\left[\sum_s (x_i - \bar{x}_s)(y_i - \bar{y}_s) + n(\bar{x}_s - \bar{x}_U)(\bar{y}_s - \bar{y}_U)\right] \quad (8a)$$

$$\hat{\alpha}_{con} = \bar{y}_U - \hat{\beta}_{con}\bar{x}_U \quad (8b)$$

$$\hat{\sigma}^2_{con} = n^{-1}\sum_s (y_i - \hat{\alpha}_{con} - \hat{\beta}_{con}x_i)^2 \,. \quad (8c)$$

A slight generalisation of this approach (Li-Chun Zhang, private communication) is to maximise the sample-data likelihood subject to the predictive mean $E(\bar{y}_U \mid \bar{y}_s, \bar{x}_s, \bar{x}_r)$ of $\bar{y}_U$ equalling its known value. This is equivalent to requiring that our estimates of α and β satisfy $\hat{\alpha} = \bar{y}_r - \hat{\beta}\bar{x}_r$. Maximising the sample-data likelihood subject to this constraint leads to estimators of the form

$$\hat{\beta}_{pred} = \left[\sum_s (x_i - \bar{x}_s)^2 + n(\bar{x}_s - \bar{x}_r)^2\right]^{-1}\left[\sum_s (x_i - \bar{x}_s)(y_i - \bar{y}_s) + n(\bar{x}_s - \bar{x}_r)(\bar{y}_s - \bar{y}_r)\right] \quad (9a)$$

$$\hat{\alpha}_{pred} = \bar{y}_r - \hat{\beta}_{pred}\bar{x}_r \quad (9b)$$

$$\hat{\sigma}^2_{pred} = n^{-1}\sum_s (y_i - \hat{\alpha}_{pred} - \hat{\beta}_{pred}x_i)^2 \,. \quad (9c)$$

In a balanced sample ($\bar{x}_s = \bar{x}_r = \bar{x}_U$), $\hat{\beta}_{fimle}$, $\hat{\beta}_{con}$ and $\hat{\beta}_{pred}$ all reduce to the sample-based MLE $\hat{\beta}_{smle}$ and $\hat{\alpha}_{fimle} = \hat{\alpha}_{con}$. In general, the differences between the constraint-based estimators (8) and (9) and the full information MLEs defined by (6) will be small.

In most applications it is unlikely that individual population data on the explanatory variable *X* in (1) will be available. It is far more likely that only sample data for *Y* and *X* will be available, along with the corresponding population means of these variables. Following the

BCDTW approach in this case then requires us to condition on this more limited information set, rather than on the information set assumed in the previous section. However, from (5) we see that the full information score functions in the complete *X* data case actually only depend on the non-sample *X*-values through their average $\bar{x}_r$. This average is known given $\bar{x}_U$ and $\bar{x}_s$. Using Result 2 of Chambers, Dorfman and Wang (1998) we conclude that the full information MLEs for this case (only $\bar{x}_U$ known) are also given by (6).

Suppose now that $\bar{x}_U$ is also unknown (so $\bar{x}_r$ is unknown). That is, the only population level data we have is the value of $\bar{y}_U$. The formal BCDTW framework for calculating the MLEs of the parameters of (1) still applies in this 'limited information' case, however, and the component score functions (5) become

$$sc_{1s}(\boldsymbol{\theta}) = \sigma^{-2}\left[\sum_s (y_i - \alpha - \beta x_i) + (N-n)(\bar{y}_r - \alpha - \beta E_s(\bar{x}_r))\right] \tag{10a}$$

$$sc_{2s}(\boldsymbol{\theta}) = \sigma^{-2}\left[\begin{array}{c}\sum_s x_i(y_i - \alpha - \beta x_i) + \\ (N-n)\left\{E_s(\bar{x}_r)(\bar{y}_r - \alpha - \beta E_s(\bar{x}_r)) - \beta Var_s(\bar{x}_r)\right\}\end{array}\right] \tag{10b}$$

$$sc_{3s}(\boldsymbol{\theta}) = -\frac{n+1}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\begin{array}{c}\sum_s (y_i - \alpha - \beta x_i)^2 + \\ (N-n)\left\{(\bar{y}_r - \alpha - \beta E_s(\bar{x}_r))^2 + \beta^2 Var_s(\bar{x}_r)\right\}\end{array}\right] \tag{10c}$$

where $E_s(\bar{x}_r)$ and $Var_s(\bar{x}_r)$ denote the expected value and variance of $\bar{x}_r$ conditional on the available data, i.e. the sample values of *Y* and *X* and the value $\bar{y}_r$. The solutions to the estimating equations defined by (10) are then

$$\hat{\beta}_{limmle} = \frac{\sum_s x_i(y_i - \bar{y}_s) + n\bar{x}_s(\bar{y}_s - \bar{y}_U) + (N-n)E_s(\bar{x}_r)(\bar{y}_r - \bar{y}_U)}{\sum_s x_i(x_i - \bar{x}_s) + n\bar{x}_s(\bar{x}_s - E_s(\bar{x}_U)) + (N-n)\left\{E_s(\bar{x}_r)(E_s(\bar{x}_r) - E_s(\bar{x}_U)) + Var_s(\bar{x}_r)\right\}} \tag{11a}$$

$$\hat{\alpha}_{limmle} = \bar{y}_U - \hat{\beta}_{limmle}E_s(\bar{x}_U) \tag{11b}$$

$$\hat{\sigma}^2_{limmle} = \frac{\sum_s (y_i - \hat{\alpha}_{limmle} - \hat{\beta}_{limmle} x_i)^2 + (N-n)\left\{(\bar{y}_r - \hat{\alpha}_{limmle} - \hat{\beta}_{limmle} E_s(\bar{x}_r))^2 + \hat{\beta}^2_{limmle} Var_s(\bar{x}_r)\right\}}{n+1} \quad (11c)$$

where $E_s(\bar{x}_U) = N^{-1}\left[n\bar{x}_s + (N-n)E_s(\bar{x}_r)\right]$. Note that mutual independence of population units under (1) implies $E_s(\bar{x}_r) = E(\bar{x}_r \mid \bar{y}_r)$ and $Var_s(\bar{x}_r) = Var(\bar{x}_r \mid \bar{y}_r)$. Assuming random sampling and a sample size *n* large enough to ensure that the joint distribution of $\bar{y}_r$, $\bar{y}_s$, $\bar{x}_r$ and $\bar{x}_s$ can be well approximated by multivariate normal distribution, we can then write down the approximations

$$E_s(\bar{x}_r) \approx \bar{x}_s + E(\bar{x}_r - \bar{x}_s \mid \bar{y}_r - \bar{y}_s) = \bar{x}_s + \beta\sigma_x^2\left(\sigma^2 + \beta^2\sigma_x^2\right)^{-1}(\bar{y}_r - \bar{y}_s) \quad (12)$$

and

$$Var_s(\bar{x}_r) \approx (N-n)^{-1}\left[\sigma_x^2 - \beta^2\sigma_x^4\left(\sigma^2 + \beta^2\sigma_x^2\right)^{-1}\right]. \quad (13)$$

Here σ_x^2 denotes the population marginal variance of *X*. Estimated values of $E_s(\bar{x}_r)$ and $Var_s(\bar{x}_r)$ can be calculated by substituting the sample-based estimates $\hat{\beta}_{smle}$ and $\hat{\sigma}^2_{smle}$ for β and σ^2, and the sample variance of *X* for σ_x^2, in the right hand sides of (12) and (13). Substituting these estimates into (11) then leads to simple approximations to the maximum likelihood estimates for the parameters of (1) in this limited information situation.

Although there is no obvious extension of the prediction estimators (9) to where only population mean of *Y* is known, it is relatively easy to modify the calibration approach (7) for this case. Here there are two, rather than three, constraints defined by our knowledge of the population size (*N*) and the population mean of *Y* ($\bar{y}_U$), and so the calibrated weights become

$$\mathbf{w}^*_{lim} = \frac{N}{n}\mathbf{1}_n + N[\mathbf{1}_n\ \mathbf{y}_s]\begin{bmatrix}\mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\mathbf{y}_s \\ \mathbf{y}_s'\mathbf{1}_n & \mathbf{y}_s'\mathbf{y}_s\end{bmatrix}^{-1}\begin{pmatrix}0 \\ \bar{y}_U - \bar{y}_s\end{pmatrix}.$$

The calibration estimators defined by these 'limited information' weights are denoted by LIMCAL in Table 1, where we show simulation results for the performances of the different

estimators defined so far (with names given by their corresponding subscripts). The simulations are model-based, with population values first simulated, then sample values drawn from this population using simple random sampling without replacement (SRSWOR). A total of 1000 simulations were carried out for each scenario.

Not surprisingly, the results set out in Table 1 support our earlier comment that estimation of α should benefit most from inclusion of the extra information about the population mean of *Y*. It is also clear that the full information MLEs (6) perform well (although their results are omitted, the constrained predictive estimators (9) were almost as efficient). With respect to RMSE, the estimators (7) based on full information calibrated weights are inefficient, even relative to the unconstrained sample-based MLEs that ignore the auxiliary information, while the limited information calibration and MLE estimators performed relatively poorly at small sample sizes. In the case of the MLE this was due to outlying estimates generated in a small number of samples where the estimation error for the population mean of *X* was large and negative. In the case of the calibration estimators this was due to negative weights being generated in these samples. A better assessment of the comparative efficiencies of the various estimators is therefore obtained by looking at their median absolute errors (MAE) in Table 1. Here we see a more consistent picture, with increased amounts of auxiliary population information leading to better inference, at least as far as α is concerned, with MLE-based methods that incorporate this inference clearly preferable.

The results shown in Table 1 mask another story, however, which is the change in the bias of the different estimators as the *Y*-balance of the sample changes. In Figure 1 we illustrate this by plotting the estimation errors for α against the corresponding rank of the sample mean $\bar{y}_s$ for one of the scenarios considered in Table 1. Here we see that the sample-based MLE has a

substantial conditional bias, while the two limited information estimators also exhibit evidence of a conditional bias. This bias essentially disappears under full information ML estimation.

So far, our analysis has focussed on the improvement in efficiency that can be obtained when we include auxiliary information about the distribution of the model variables in the target population. Another advantage when this information is included, however, is that it can help protect inference from bias in cases where sample inclusion probabilities depend on these variables. To illustrate this, in Table 2 we report simulation results for the same scenarios explored in Table 1 but now where sample inclusion probabilities are either approximately proportional to X (PPX sampling) or approximately proportional to Y (PPY sampling).

The gains from using the full information MLEs under both PPX and PPY sampling are clear in Table 2. In contrast, the calibration-based estimators LIMCAL and CALW become quite unstable. The limited information MLE (LIMMLE) performs comparably with the sample-based MLE (SMLE) under PPX sampling, but is superior under PPY sampling. Although we do not show it here, the conditional bias properties of the different estimators of α under PPX and PPY sampling are qualitatively similar to those under SRSWOR (see Figure 1). In particular, the sample-based MLE is clearly conditionally biased, particularly under PPY sampling, while the limited information MLE has reduced conditional bias. The full information MLE of this parameter has essentially zero conditional bias.

3. MLE for a linear logistic model given auxiliary population information

Here Y is a zero-one variable but X is an arbitrary real-valued variable. As in the previous section we initially assume sample values of Y and X are available, together with auxiliary information corresponding to the non-sample total t_{ry} of Y and the non-sample values of X. We

wish to combine the sample data and this auxiliary information in order to model the relationship between Y and X in the population using a linear logistic model. For simplicity we assume independent population elements and simple random sampling.

For population element i, put $\pi(x_i) = \Pr(y_i = 1 \mid x_i) = \exp(\alpha + \beta x_i)\left(1 + \exp(\alpha + \beta x_i)\right)^{-1}$. The population level component score functions for $\theta = (\alpha, \beta)$ are then

$$sc_1(\theta) = \sum_U (y_i - \pi(x_i))$$

$$sc_2(\theta) = \sum_U x_i (y_i - \pi(x_i))$$

so the full information component score functions become

$$sc_{1s}(\theta) = \sum_U y_i - \sum_U \pi(x_i) \tag{14a}$$

$$sc_{2s}(\theta) = \sum_s x_i (y_i - \pi(x_i)) + E_s\left(\sum_r x_i y_i\right) - \sum_r x_i \pi(x_i). \tag{14b}$$

For arbitrary non-sample population element i, let $r(i)$ denote the remaining $N - n - 1$ non-sampled population elements. Without loss of generality we assume $t_{ry} > 0$, so the conditional expectation in (14b) can be written

$$\begin{aligned} E\left(\sum_r y_i x_i \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right) &= \sum_r x_i E\left(y_i \mid \sum_r y_j = t_{ry}, \mathbf{x}_r\right) \\ &= \sum_r x_i \Pr\left(y_i = 1 \mid \sum_r y_j = t_{ry}, \mathbf{x}_r\right) \\ &= \frac{\sum_r x_i \Pr\left(y_i = 1, \sum_{r(i)} y_j = t_{ry} - 1 \mid \mathbf{x}_r\right)}{\Pr\left(\sum_r y_j = t_{ry} \mid \mathbf{x}_r\right)} \\ &= \sum_r x_i \pi(x_i) R_{1i} \end{aligned}$$

where $R_{1i} = \left(\Pr\left(\sum_r y_j = t_{ry} \mid \mathbf{x}_r\right)\right)^{-1} \Pr\left(\sum_{r(i)} y_j = t_{ry} - 1 \mid \mathbf{x}_{r(i)}\right)$. The full information score function components defined by (14) are therefore

$$sc_{1s}(\theta) = \sum_U (y_i - \pi(x_i)) \tag{15a}$$

$$sc_{2s}(\boldsymbol{\theta}) = \sum_s x_i(y_i - \pi(x_i)) - \sum_r x_i\pi(x_i)(1 - R_{1i}). \qquad (15b)$$

A saddlepoint approximation to the second term on the right hand side of (15b) is developed in the Appendix. This is

$$sc_{2s}(\boldsymbol{\theta}) \approx \sum_s x_i\left(y_i - \pi(x_i)\right) - \sum_r x_i\pi(x_i)\left(1 - [1 + (1 - \pi(x_i))\{b(t_{ry}) - 1\}]^{-1}\right) \qquad (15c)$$

with $b(t_{ry}) = \exp\left\{\left[\sum_r \pi(x_j)\left(1 - \pi(x_j)\right)\right]^{-1}\left[\sum_r \pi(x_j) - t_{ry}\right]\right\}$.

As noted already in section 2, it is extremely unlikely in practice that the actual non-sample *X* values will be known. Since the full information score function (15) depends directly on these values, we need to revise this function when non-sample *X* values are unavailable. In general, the score function for α and β is then defined by

$$sc_{1s}(\boldsymbol{\theta}) = \sum_U y_i - \sum_s \pi(x_i) - E_s\left(\sum_r \pi(x_i)\right) \qquad (16a)$$

$$sc_{2s}(\boldsymbol{\theta}) = \sum_s x_i(y_i - \pi(x_i)) + E_s\left(\sum_r x_i y_i\right) - E_s\left(\sum_r x_i\pi(x_i)\right) \qquad (16b)$$

where E_s denotes expectation after conditioning on the actual auxiliary information that we have (we continue to assume that t_{ry} is known). Suppose we know the non-sample mean $\bar{x}_r$ of *X*. We can then approximate the conditional expectations $E_s\left(\sum_r \pi(x_i)\right)$ and $E_s\left(\sum_r x_i\pi(x_i)\right)$ using a smearing approach (Duan, 1983). This is based on the assumption that, for an arbitrary function f of *x* that depends on some parameter θ, we can write

$$\frac{1}{N-n}\sum_r f(x_i, \theta) = \frac{1}{N-n}\sum_r f\left(\bar{x}_r + (x_i - \bar{x}_r), \theta\right) \approx \frac{1}{n}\sum_s f\left(\bar{x}_r - \bar{x}_s + x_i, \theta\right).$$

Put $\Delta = \bar{x}_r - \bar{x}_s$. The smearing approximation to $E_s\left(\sum_r \pi(x_i)\right)$ is then

$$E_s\left(\sum_r \pi(x_i)\right) \approx \frac{N-n}{n}\sum_s \pi(\Delta + x_i).$$

We therefore replace the score component (16a) by

$$sc_{1smear}(\theta) = \sum_U y_i - \sum_s \pi(x_i) - \frac{N-n}{n}\sum_s \pi(\Delta + x_i)\,. \tag{17a}$$

A corresponding smearing approximation to (16b) that includes a saddlepoint approximation is given by (A.7) in the Appendix. This allows us to replace this component score by

$$\begin{aligned} sc_{2smear}(\theta) = {} & \sum_s x_i\left(y_i - \pi(x_i)\right) - \left(\frac{N-n}{n}\right)\sum_s (\Delta + x_i)\pi(\Delta + x_i) \\ & + \left(\frac{N-n}{n}\right)\sum_s (\Delta + x_i)\pi(\Delta + x_i)\left[1 + \left\{1 - \pi(\Delta + x_i)\right\}\left\{b_{smear}(t_{ry}) - 1\right\}\right]^{-1} \end{aligned} \tag{17b}$$

where

$$b_{smear}(t_{ry}) = \exp\left\{\left[\sum_s \pi(\Delta + x_i)\left(1 - \pi(\Delta + x_i)\right)\right]^{-1}\left[\sum_s \pi(\Delta + x_i) - \frac{n}{N-n}t_{ry}\right]\right\}.$$

Finally, there is the case where even $\bar{x}_r$ is unknown. In this case we can still use (17), but replace $\bar{x}_r$ by an appropriate sample-based estimate. This will depend on the characteristics of the sample design and the nature of the auxiliary population information available to us. For the case of simple random sampling and no auxiliary information it is natural to estimate $\bar{x}_r$ by $\bar{x}_s$, i.e. use expansion estimation. This is equivalent to setting $\Delta = 0$ in (17). To avoid confusion with the full information MLEs approximated by (15a) and (15c), we refer to estimators of α and β obtained by setting (17) to zero and solving for these parameters as smearing MLEs when the actual value of $\bar{x}_r$ is used (subscript *smear*) and as expansion MLEs when $\bar{x}_r$ is replaced by $\bar{x}_s$ (subscript *exp*).

The simulation results set out in Table 3 allow one to compare the root mean squared errors and median absolute errors of the sample-based MLEs $\hat{\alpha}_{smle}$ and $\hat{\beta}_{smle}$ of α and β (i.e. the estimators that only use the sample values of *Y* and *X*, denoted SMLE) with those of the MLEs that use the auxiliary information in t_{ry} as well as differing amounts of information about the

population distribution of *X*. These are the full information MLEs $\hat{\alpha}_{fimle}$ and $\hat{\beta}_{fimle}$ (FIMLE) that assume knowledge of the non-sample values of *X*, the smearing estimators $\hat{\alpha}_{smear}$ and $\hat{\beta}_{smear}$ (SMEAR) that only require the non-sample mean of *X* and the expansion estimators $\hat{\alpha}_{exp}$ and $\hat{\beta}_{exp}$ (EXP) that do not require any information about the non-sample distribution of *X*. The sample-based MLEs were computed using the *glm* function in R, with its default options, while the MLEs utilising auxiliary information were calculated using the *nlm* function in R, with starting values $\alpha = \log(\bar{y}_U) - \log(1 - \bar{y}_U)$ and $\beta = 0$. In each of 1000 independent simulations, a population of *N* independent and identically distributed values for *X* was generated from the standard lognormal distribution and corresponding values for *Y* generated under the linear logistic model. A sample of size *n* was then taken from this population using SRSWOR.

We see that there can be substantial gains when auxiliary population information is included in the modelling process, particularly when the probability that *Y* = 1 is small. We also note in passing that these gains become even more substantial as the sample size *n* decreases, however then greater care has to be taken with solution of the ML estimating equations. Observe that the expansion MLE sometimes provides the best RMSE performance, although this is not the case when one looks at MAE. However, the expansion MLE is conditionally biased, as is evident when one looks at the plots in Figure 2. This also shows that the sample-based MLE has a strong conditional bias, while both the smearing and full information MLEs are much better behaved.

4. MLE for a linear logistic model under case-control sampling

In the previous section we assumed simple random sampling from the population of interest. However, in many important applications of logistic modelling, particularly in medicine,

the sample data are obtained via some form of case-control sampling. In such cases the assumptions underpinning the saddlepoint and smearing approximations used in the development in the previous section are no longer valid. However, the basic strategy of using the approach of BCDTW to incorporate auxiliary population information into inference can still be used, provided the fact that the sample data are obtained via an informative sampling method (case-control sampling) is allowed for when taking conditional expectations. More specifically, we adopt the setup described in Scott and Wild (1997), and assume the existence of two sampling frames, one for the N_1 population units with values $Y = 1$ and one for the N_0 units with $Y = 0$. Independent simple random samples of size n_1 and n_0 respectively are then taken from these frames. Values of X are observed on the sample, and the aim again is to fit a linear logistic model to these data. By definition, we know N_1 and hence $t_{ry} = N_1 - n_1$.

Again, we consider the same three situations corresponding to different levels of knowledge of X. The first is where we know the non-sample values of this variable. In the standard case-control situation this is highly unlikely. However, it could correspond to a situation where a separate administrative register contains these values, and the case-control study is being used to forge a link between the Y registers and the X register. The second is where no X register exists, but the value of $\bar{x}_r$ (or an accurate estimate of this quantity) is known. The third is the conventional case-control situation, where no X knowledge is available outside the sample. In all three cases, the ML estimating equations for the parameters α and β of the assumed population level linear logistic model are theoretically defined as the conditional expectations of the population level ML estimating equations given the sample data and the known population information. However, in this case the random variables underpinning these conditional

expectations no longer follow the same logistic model as in the population, so the approximations to the ML score function derived in the previous section need modification.

To start, consider the first situation described above, where individual X values for non-sample population units are known, but the corresponding values of Y are not. We continue to use the notation introduced in the previous section. From (14), we see that the key unknown quantity in the score function is $E_s\left(\sum_r x_i y_i\right)$, where now, because of the case-control sampling, the y_i values in the summation no longer follow the assumed population level logistic model. Following Scott and Wild (1997), we use Bayes Theorem to approximate the distribution of these values as $N - n$ independent Bernoulli realisations with

$$\pi_r(x_i) = \Pr\left(y_i = 1 \mid i \in r, x_i\right) = \frac{N_1^{-1}(N_1 - n_1)\pi(x_i)}{N_1^{-1}(N_1 - n_1)\pi(x_i) + N_0^{-1}(N_0 - n_0)\left(1 - \pi(x_i)\right)}.$$

With this set up, we can use the same saddlepoint arguments as in the previous section to approximate $E_s\left(\sum_r x_i y_i\right)$, replacing $\pi(x_i)$ in that development by $\pi_r(x_i)$ above. This leads to a 'full information' score function with component (15a) as before, but with (15c) replaced by

$$sc_{2s}(\theta) = \sum_s x_i\left(y_i - \pi(x_i)\right) + \sum_r x_i \pi_r(x_i)\left[1 + \left(1 - \pi_r(x_i)\right)\left(b_r(t_{ry}) - 1\right)\right]^{-1} - \sum_r x_i \pi(x_i) \qquad (18)$$

where $b_r(t_{ry}) = \exp\left(\left[\sum_r \pi_r(x_i)\left(1 - \pi_r(x_i)\right)\right]^{-1}\left[\sum_r \pi_r(x_i) - t_{yr}\right]\right)$.

In the previous section, we used smearing to approximate the score function in the case where the individual non-sample X values are unknown, but their mean $\bar{x}_r$ is known. This approach needs modification under case-control, because sample and non-sample averages no longer have the same expected values. In particular, for the case-control design assumed here, we need to apply smearing approximations separately for cases and controls. That is, for an arbitrary function f of x characterised by a parameter θ, we use the approximation

$$\sum_r f(x_i,\theta) \approx M_1 n_1^{-1} \sum_{s1} f\left(\Delta_1 + x_i, \theta\right) + M_0 n_0^{-1} \sum_{s0} f\left(\Delta_0 + x_i, \theta\right).$$

Here sd denotes the sample units with $Y = d$ and Δ_d denotes our best estimate of the difference between the non-sample and sample means of X for those units with $Y = d$. Since we know the overall non-sample mean $\bar{x}_r$ of X, we calculate Δ_d using a regression type estimate, i.e.

$$\Delta_d = \lambda_d n_d^{-1} s_{xd}^2 \left(\lambda_1^2 n_1^{-1} s_{x1}^2 + \lambda_0^2 n_0^{-1} s_{x0}^2\right)^{-1} \left(\bar{x}_r - \lambda_1 \bar{x}_{s1} - \lambda_0 \bar{x}_{s0}\right)$$

where $\lambda_d = \left(N_d - n_d\right)/(N-n)$ and $\bar{x}_{sd}$, s_{xd}^2 denote the mean and variance of X for the sample units with $Y = d$. The case-control version of the smearing approximation (17a) is then

$$sc_{1smear}(\theta) = \sum_U y_i - \sum_s \pi(x_i) - \sum_{d=0}^{1} \frac{N_d - n_d}{n_d} \sum_{sd} \pi(\Delta_d + x_i) \tag{19a}$$

while the corresponding case-control version of (17b) is

$$\begin{aligned} sc_{2smear}(\theta) = {} & \sum_s x_i \left(y_i - \pi(x_i)\right) - \sum_{d=0}^{1} \left(\frac{N_d - n_d}{n_d}\right) \sum_{sd} \left(\Delta_d + x_i\right) \pi(\Delta_d + x_i) \\ & + \sum_{d=0}^{1} \left(\frac{N_d - n_d}{n_d}\right) \sum_s \left(\Delta_d + x_i\right) \pi_r(\Delta_d + x_i) \left[1 + \left\{1 - \pi_r(\Delta_d + x_i)\right\} \left\{b_{smear}^{cc}(t_{ry}) - 1\right\}\right]^{-1} \end{aligned} \tag{19b}$$

where

$$b_{smear}^{cc}(t_{ry}) = \exp\left(\left[\sum_{d=0}^{1} \frac{N_d - n_d}{n_d} \sum_{sd} \pi_r(\Delta_d + x_i)\left(1 - \pi_r(\Delta_d + x_i)\right)\right]^{-1} \left[\sum_{d=0}^{1} \frac{N_d - n_d}{n_d} \sum_{sd} \pi_r(\Delta_d + x_i) - t_{yr}\right]\right).$$

When $\bar{x}_r$ is also unknown, we replace $\bar{x}_{rd}$ by $\bar{x}_{sd}$ above. This is equivalent to setting $\Delta_d = 0$ in (19) and corresponds to using stratified expansion estimators for the expected values of the unknown non-sample components of the score function.

In what follows we use the same notation as in the previous section, denoting estimates obtained by setting (15a) and (18) to zero by FIMLE , and referring to them as full information MLEs. Estimates obtained by setting (19) to zero and solving are referred to as smearing MLEs

and are denoted by SMEAR. Finally, those obtained by solving (19) with $\Delta_d = 0$ are referred to as expansion MLEs and are denoted by EXP.

Table 4 sets out simulation results for the above approximate MLEs as well as for the standard sample-based MLEs $\hat{\alpha}_{smle}$ and $\hat{\beta}_{smle}$ (SMLE). Prentice and Pyke (1979) showed that $\hat{\beta}_{smle}$ provides a good approximation to the actual MLE of this parameter under case-control sampling. In addition we show results for the maximum pseudo-likelihood estimates, defined by solving weighted versions of the sample-based MLE estimating equations, with weights given by $w_i = N_0 n_0^{-1} I(y_i = 0) + N_1 n_1^{-1} I(y_i = 1)$, and are denoted by WTD. We also computed the maximum 'pseudo-model' likelihood estimates proposed by Scott and Wild (1997) for case-control sampling, but do not show results for them since these were almost identical to those for SMLE for β and tended to be unstable for α.

The simulation methodology used to obtain the results in Table 4 is identical to that used in Table 3, with the exception that sampling here is carried out using the stratified case-control design described at the start of this section. Note that SMLE and WTD estimates were computed using the *glm* function in R (without and with weights respectively) and with default settings. The FIMLE, SMEAR and EXP approximations to the MLEs that utilised auxiliary information were all computed by using the *nlm* function in R to solve the relevant estimating equations.

The results set out in Table 4 confirm once again that inclusion of population level auxiliary information can bring substantial gains in maximum likelihood-based inference. This is particularly the case where this information is strong, as in the FIMLE. However, there are still gains when the auxiliary information used is much weaker, as in SMEAR. Not surprisingly, we see that the SMLE is biased for α but well behaved for β.

4. Discussion

The two most important conclusions that we draw from the results set out in this paper is that it pays to include population level auxiliary information when modelling sample survey data, and that the BCDTW likelihood framework offers a viable approach to achieving this aim. Obviously, the more auxiliary information one has available, the more significant the improvement in one's inference. However, even marginal information (e.g. knowledge of population means for the model variables) can be extremely useful when integrated with the sample data within this framework. In general, use of the BCDTW framework requires the evaluation of conditional expectations that depend both on the assumed population model as well as on the method used to select the sample. For the important case of a logistic population model, the saddlepoint and smearing approximations to these conditional expectations that we describe in this paper seem to work well and should be useful in extending our results in practice.

This paper does not include results on interval estimation when auxiliary population data are integrated into likelihood inference. The BCDTW framework also covers this situation, and in the Appendix we show how the information function can be extended to allow for the auxiliary information in the case of a logistic model, including appropriate saddlepoint approximations. An important use of this function is in evaluating the extra information for parametric inference provided by the auxiliary information, e.g. along the lines set out in Steel *et. al.* (2004).

Finally, we note that the auxiliary population information is assumed to be known precisely. In reality population marginal information may in fact be estimated, typically from another, larger, survey. The impact of the resulting imprecision on our results requires further research.

References

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

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, **12**, 3 - 32.

Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B*, **60**, 397 - 412.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376 - 382.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation estimate. *Journal of the American Statistical Association*, **78**, 605 - 610.

Handcock, M., Rendall, M. and Cheadle, J. (2005). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociological Methodology*, **35**, 291 - 334.

Imbens, G.W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies*, **61**, 655 - 680.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403 - 411.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, **87**, 484 - 490.

Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57 - 71.

Steel, D.G., Beh, E. J. and Chambers, R.L. (2004). The information in aggregate data. In *Ecological Inference: New Methodological Strategies*. (eds. G. King, O. Rosen and M. Tanner). Cambridge University Press: Cambridge.

Appendix

A. Saddlepoint Approximations

We first consider approximation of R_{1i}. Let $\bar{y}_v$ be the mean of *Y* over the set *v*, with N_v the corresponding number of observations. Further, let $g_v(d) = \Pr(\bar{y}_v = d \mid \mathbf{x}_v)$ and $\pi_i = \pi(x_i)$. Then, for $t_{ry} > 0$

$$R_{1i} = \frac{g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}}{\pi_i g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\} + (1-\pi_i) g_{r(i)}(t_{ry}/N_{r(i)})} = \left(1 + (1-\pi_i)\left[\frac{g_{r(i)}(t_{ry}/N_{r(i)})}{g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}} - 1\right]\right)^{-1} . \quad \text{(A.1)}$$

It follows that the major problem is to approximate $\left[g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}\right]^{-1} g_{r(i)}(t_{ry}/N_{r(i)})$ accurately. Now the cumulant generating function of $\sum_v y_j$ is $K_v(u) = \sum_v \log\{\pi_j e^u + (1-\pi_j)\}$. For any $d \in (0,1)$ the saddlepoint approximation to $g_v(d)$ is then

$$h_v(d) = \frac{N_v}{\{2\pi K_v''(u_d)\}^{1/2}} \exp\{K_v(u_d) - N_v u_d d\}$$

where u_d is called the saddlepoint, and is defined as the solution of

$$K_v'(u)/N_v = d . \quad \text{(A.2)}$$

Standard arguments can be used to show that $h_v(d) = g_v(d)\{1 + O(\frac{1}{N_v})\}$ under general regularity conditions. That is, the saddlepoint approximation has relative error of order N_v^{-1}. Substituting $d = d_1 = t_{ry}/N_{r(i)}$ or $d = d_2 = (t_{ry}-1)/N_{r(i)}$ in $h_{r(i)}(d)$, we then have

$$\frac{g_{r(i)}(t_{ry}/N_{r(i)})}{g_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}} = \frac{h_{r(i)}(t_{ry}/N_{r(i)})}{h_{r(i)}\left\{(t_{ry}-1)/N_{r(i)}\right\}}\{1 + O(\tfrac{1}{N})\} = \exp\{-u_{d_1}\}\{1 + O(\tfrac{1}{N})\} \quad \text{(A.3)}$$

where the last equation is due to the identity

$$K_{r(i)}(u_{d_1}) - N_{r(i)} u_{d_1} d_1 - \left\{K_{r(i)}(u_{d_2}) - N_{r(i)} u_{d_2} d_2\right\} = N_{r(i)} u_{d_1}(d_2 - d_1) + O(\tfrac{1}{N}) = -u_{d_1} + O(\tfrac{1}{N}) .$$

From the central limit theorem $N_v^{-1/2}\sum_v (y_j - \pi_j) \to N(0,\gamma^2)$ as $N_v \to \infty$, where $\gamma^2 = \lim N_v^{-1}\sum_v \pi_j(1-\pi_j)$. It follows that we can focus on the normal deviation values of t_{ry}: $t_{ry} - \sum_{r(i)} \pi_j = O\left(\sqrt{N}\right)$. For such values of t_{ry}, $u_{d_1} = O\left(N^{-1/2}\right)$. In fact, from (A.2), it can be seen that

$$u_{d_1} = \frac{t_{ry} - \Sigma_{r(i)}\pi_j}{\Sigma_{r(i)}\pi_j(1-\pi_j)} + O\left(\frac{1}{N}\right) = \frac{t_{ry} - \Sigma_r \pi_j}{\Sigma_r \pi_j(1-\pi_j)} + O\left(\frac{1}{N}\right). \tag{A.4}$$

By (A.1), (A.3) and (A.4), an approximation to R_{1i} is then

$$R_{1i} = \left[1 + (1-\pi_i)\left\{b(t_{ry}) - 1\right\}\right]^{-1}\left\{1 + O\left(\frac{1}{N}\right)\right\} \tag{A.5}$$

with $b(t_{ry}) = \exp\left\{\left[\sum_r \pi_j(1-\pi_j)\right]^{-1}\left[\sum_r \pi_j - t_{ry}\right]\right\}$. It immediately follows that (15b) can be approximated by

$$sc_s(\beta) \approx \sum_s x_i(y_i - \pi_i) - \sum_r x_i \pi_i \left(1 - [1 + (1-\pi_i)\{b(t_{ry}) - 1\}]^{-1}\right). \tag{A.6}$$

When non-sample values of X are unavailable, but their mean $\bar{x}_r$ is known, we can combine the saddlepoint approximation developed above with a smearing approximation to again approximate the logistic score function. In particular, this procedure can be used together with (A.6) to approximate the second part of (16b). We continue to use (17a) to approximate (16a). By (A.6),

$$\begin{aligned} sc_s(\beta) &\approx \sum_s x_i(y_i - \pi_i) - \sum_r \left\{\bar{x}_r + (x_i - \bar{x}_r)\right\}\pi_i \left(1 - \left[1 + (1-\pi_i)\left\{b(t_{ry}) - 1\right\}\right]^{-1}\right) \\ &\approx \sum_s x_i(y_i - \pi_i) - \left(\frac{N-n}{n}\right)\sum_s \left(\bar{x}_r - \bar{x}_s + x_i\right)\pi_{i,adj}\left(1 - \left[1 + (1-\pi_{i,adj})\left\{b(t_{ry}) - 1\right\}\right]^{-1}\right) \quad \text{(A.7)} \\ &\approx \sum_s x_i(y_i - \pi_i) - \left(\frac{N-n}{n}\right)\sum_s \left(\bar{x}_r - \bar{x}_s + x_i\right)\pi_{i,adj}\left(1 - \left[1 + (1-\pi_{i,adj})\left\{b_{adj}(t_{ry}) - 1\right\}\right]^{-1}\right) \end{aligned}$$

where

$$\pi_{i,adj} = \exp\left\{\beta(\bar{x}_r - \bar{x}_s) + \alpha + \beta x_i\right\} / \left[1 + \exp\left\{\beta(\bar{x}_r - \bar{x}_s) + \alpha + \beta x_i\right\}\right]$$

and

$$b_{adj}(t_{ry}) = \exp\left\{\left[\sum_s \pi_{i,adj}\left(1 - \pi_{i,adj}\right)\right]^{-1}\left[\sum_s \pi_{i,adj} - \frac{n}{N-n} t_{ry}\right]\right\}.$$

Note that the last two approximation steps in (A.7) used smearing approximations repeatedly.

B. The Information Function in the Logistic Case

Within the BCDTW framework the information function for parametric likelihood inference is the conditional expectation of the population level information function minus the conditional variance of the population level score function. As always, conditioning here is with respect to the observed survey data as well as the auxiliary information. In the logistic case the information function components are therefore given by

$$\begin{aligned} info_s(\alpha,\alpha) &= E_s\left(info(\alpha,\alpha)\right) - Var_s\left(sc(\alpha)\right) \\ &= E_s \sum_U \pi(x_i)(1-\pi(x_i)) - Var_s\left(\sum_U (y_i - \pi(x_i))\right) \\ &= \sum_U \pi(x_i)(1-\pi(x_i)) \end{aligned}$$

$$\begin{aligned} info_s(\alpha,\beta) &= E_s\left(info(\alpha,\beta)\right) - Cov_s\left(sc(\alpha), sc(\beta)\right) \\ &= E_s \sum_U x_i \pi(x_i)(1-\pi(x_i)) - Cov_s\left(\sum_U (y_i - \pi(x_i)), \sum_U x_i (y_i - \pi(x_i))\right) \\ &= \sum_U x_i \pi(x_i)(1-\pi(x_i)) \end{aligned}$$

$$\begin{aligned} info_s(\beta,\beta) &= E_s\left(info(\beta,\beta)\right) - Var_s\left(sc(\beta)\right) \\ &= \sum_U x_i^2 \pi(x_i)(1-\pi(x_i)) - Var_s\left(\sum_U x_i (y_i - \pi(x_i))\right) \\ &= \sum_U x_i^2 \pi(x_i)(1-\pi(x_i)) - Var_s\left(\sum_U x_i y_i\right) \end{aligned}$$

where

$$\begin{aligned} Var_s\left(\sum_U y_i x_i\right) &= Var\left(\sum_r y_i x_i \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right) \\ &= E\left(\sum_{i\in r}\sum_{j\in r} y_i y_j x_i x_j \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right) - \left[E\left(\sum_r y_i x_i \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right)\right]^2 \end{aligned}$$

with

$$E\left(\sum_{i\in r}\sum_{j\in r} y_i y_j x_i x_j \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right) = \sum_r x_i^2 E\left(y_i \mid \sum_r y_k = t_{ry}, \mathbf{x}_r\right) + \sum_{i\in r}\sum_{j\neq i\in r} x_i x_j E\left(y_i y_j \mid \sum_r y_j = t_{ry}, \mathbf{x}_r\right) = \sum_r x_i^2 \pi(x_i) R_{1i} + \sum_{i\in r}\sum_{j\neq i\in r} x_i x_j \pi(x_i)\pi(x_j) R_{2ij}$$

$$\left[E\left(\sum_r y_i x_i \mid \sum_r y_i = t_{ry}, \mathbf{x}_r\right)\right]^2 = \left[\frac{\sum_r x_i \pi(x_i) \Pr\left(\sum_{r(i)} y_j = t_{ry} - 1 \mid \mathbf{x}_{r(i)}\right)}{\Pr\left(\sum_r y_k = t_{ry} \mid \mathbf{x}_r\right)}\right]^2 = \sum_r x_i^2 \pi^2(x_i) R_{1i}^2 + \sum_{i\in r}\sum_{j\neq i\in r} x_i x_j \pi(x_i)\pi(x_j) R_{1i} R_{1j}$$

and $R_{2ij} = \left[\Pr\left(\sum_r y_k = t_{ry} \mid \mathbf{x}_r\right)\right]^{-1} \Pr\left(\sum_{r(ij)} y_k = t_{ry} - 2 \mid \mathbf{x}_{r(ij)}\right)$. It follows

$$Var_s\left(\sum_U y_i x_i\right) = \sum_r x_i^2 \pi(x_i) R_{1i}\left(1 - \pi(x_i) R_{1i}\right) + \sum_{i\in r}\sum_{j\neq i\in r} x_i x_j \pi(x_i)\pi(x_j)(R_{2ij} - R_{1i} R_{1j}).$$

A saddlepoint approximation to R_{2ij} similar to that developed above for R_{1i} can be written down. This is based on the fact that the denominator of R_{2ij} can be expressed as

$$\Pr\left(\sum_r y_k = t_{ry} \mid \boldsymbol{x}_r\right) = \pi_i \pi_j \Pr\left(\sum_{r(ij)} y_k = t_{ry} - 2 \mid \boldsymbol{x}_{r(ij)}\right) + \left\{\pi_i(1-\pi_j) + (1-\pi_i)\pi_j\right\} \Pr\left(\sum_{r(ij)} y_k = t_{ry} - 1 \mid \boldsymbol{x}_{r(ij)}\right) + (1-\pi_i)(1-\pi_j) \Pr\left(\sum_{r(ij)} y_k = t_{ry} \mid \boldsymbol{x}_{r(ij)}\right)$$

leading to

$$R_{2ij} = \left\{\pi_i \pi_j + (\pi_i + \pi_j - 2\pi_i \pi_j) \frac{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} - 1 \mid \boldsymbol{x}_{r(ij)}\right)}{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} - 2 \mid \boldsymbol{x}_{r(ij)}\right)} + (1-\pi_i)(1-\pi_j) \frac{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} \mid \boldsymbol{x}_{r(ij)}\right)}{\Pr\left(\Sigma_{r(ij)} y_k = t_{ry} - 2 \mid \boldsymbol{x}_{r(ij)}\right)}\right\}^{-1}.$$

Using the same saddlepoint approximation technique as that used for R_{1i}, the two ratios in this expression can be approximated by $b(t_{ry} - 1)$ and $b^2(t_{ry} - 1)$ respectively. That is,

$$R_{2ij} = \left\{\pi_i \pi_j + (\pi_i + \pi_j - 2\pi_i \pi_j) b(t_{ry} - 1) + (1-\pi_i)(1-\pi_j) b^2(t_{ry} - 1)\right\}\left\{1 + O\left(\tfrac{1}{N}\right)\right\}.$$

Table 1 Root mean squared errors (RMSE) and median absolute errors (MAE) under SRSWOR and a linear population model with $\alpha = 5,\ \beta = 1$ and $\sigma^2 = 1$. Values of X drawn from the standard lognormal distribution.

		RMSE			MAE		
		$N = 500$ $n = 20$	$N = 1000$ $n = 50$	$N = 5000$ $n = 200$	$N = 500$ $n = 20$	$N = 1000$ $n = 50$	$N = 5000$ $n = 200$
α	SMLE	0.3217	0.1929	0.0922	0.2132	0.1339	0.0594
	LIMCAL	0.5935	0.2100	0.0977	0.2274	0.1276	0.0601
	LIMMLE	3.3769	0.3668	0.0676	0.1948	0.1015	0.0429
	CALW	1.3925	0.1658	0.0654	0.1803	0.0947	0.0421
	FIMLE	0.2554	0.1408	0.0631	0.1582	0.0869	0.0399
β	SMLE	0.1679	0.0867	0.0374	0.0935	0.0517	0.0234
	LIMCAL	0.4109	0.0977	0.0429	0.1018	0.0557	0.0246
	LIMMLE	3.2881	0.3327	0.0494	0.1270	0.0655	0.0310
	CALW	0.8008	0.0994	0.0391	0.1069	0.0553	0.0254
	FIMLE	0.1550	0.0843	0.0375	0.0884	0.0522	0.0234
σ^2	SMLE	0.3154	0.1975	0.1022	0.2350	0.1361	0.0741
	LIMCAL	0.4186	0.2033	0.1024	0.2557	0.1398	0.0743
	LIMMLE	107.4689	0.8051	0.1019	0.2440	0.1341	0.0738
	CALW	0.4258	0.2152	0.1036	0.2692	0.1509	0.0735
	FIMLE	0.3089	0.1957	0.1017	0.2315	0.1345	0.0737

Table 2 Root mean squared errors (RMSE) and median absolute errors (MAE) under PPX and PPY sampling and a linear population model with $\alpha = 5$, $\beta = 1$ and $\sigma^2 = 1$. Values of *X* drawn from the standard lognormal distribution.

		RMSE			MAE		
		$N = 500$ $n = 20$	$N = 1000$ $n = 50$	$N = 5000$ $n = 200$	$N = 500$ $n = 20$	$N = 1000$ $n = 50$	$N = 5000$ $n = 200$
				PPX Sampling			
α	SMLE	0.3285	0.1993	0.0927	0.2065	0.1312	0.0629
	LIMCAL	1.3901	3.3726	0.1824	0.3057	0.1939	0.1320
	LIMMLE	0.2359	0.1715	0.1170	0.1665	0.1224	0.0943
	CALW	5.0604	2.5311	0.0466	0.1552	0.0763	0.0270
	FIMLE	0.1116	0.0611	0.0263	0.0651	0.0410	0.0169
β	SMLE	0.0715	0.0369	0.0157	0.0368	0.0224	0.0102
	LIMCAL	0.7589	1.8295	0.0413	0.0934	0.0500	0.0187
	LIMMLE	0.0817	0.0414	0.0170	0.0386	0.0231	0.0109
	CALW	2.9475	1.6181	0.0272	0.0901	0.0411	0.0152
	FIMLE	0.0612	0.0316	0.0139	0.0319	0.0197	0.0091
σ^2	SMLE	0.3272	0.1984	0.1020	0.2429	0.1362	0.0675
	LIMCAL	0.6624	0.9780	0.1137	0.2723	0.1567	0.0786
	LIMMLE	0.3416	0.2110	0.1076	0.2280	0.1356	0.0698
	CALW	1.8410	0.5481	0.1155	0.2877	0.1579	0.0782
	FIMLE	0.3174	0.1954	0.1020	0.2347	0.1362	0.0677
				PPY Sampling			
α	SMLE	0.3558	0.2483	0.1906	0.2310	0.1825	0.1726
	LIMCAL	5.3714	1.6835	3.9844	0.2343	0.1723	0.1474
	LIMMLE	0.6531	0.1500	0.0939	0.1589	0.0945	0.0689
	CALW	2.2143	4.8633	16.6698	0.1626	0.1059	0.0873
	FIMLE	0.1953	0.0958	0.0408	0.0974	0.0558	0.0255
β	SMLE	0.1253	0.0580	0.0251	0.0619	0.0337	0.0158
	LIMCAL	3.8975	1.1988	2.4283	0.0957	0.0633	0.0519
	LIMMLE	0.5743	0.0880	0.0310	0.0671	0.0376	0.0183
	CALW	1.2909	2.9607	10.1346	0.0981	0.0644	0.0527
	FIMLE	0.1178	0.0555	0.0232	0.0603	0.0311	0.0136
σ^2	SMLE	0.3160	0.2035	0.0998	0.2343	0.1462	0.0682
	LIMCAL	0.9376	0.3779	0.5802	0.2552	0.1563	0.0759
	LIMMLE	2.1472	0.2027	0.0974	0.2252	0.1473	0.0672
	CALW	0.9910	1.1900	2.7521	0.2926	0.1802	0.0914
	FIMLE	0.3110	0.2031	0.0972	0.2253	0.1461	0.0666

Table 3 Root mean squared errors (RMSE) and median absolute errors (MAE) for the linear logistic model under SRSWOR and given different amounts of auxiliary information on *X*. In all cases $N = 5000$ and $n = 200$.Values of *X* drawn from the standard lognormal distribution.

True (α, β)		(–3, 1)	(–5, 2)	(–5, 1)	(–8, 2)
			RMSE		
α	SMLE	0.4150	0.8039	0.9372	143.0808
	EXP	4.5191	1.0254	0.7968	2.7469
	SMEAR	0.7845	2.4532	0.7735	2.6343
	FIMLE	0.3352	0.7060	0.7619	3.6909
β	SMLE	0.1899	0.3746	0.2497	39.8094
	EXP	2.2092	0.5105	0.2275	0.7696
	SMEAR	0.4121	1.2314	0.2329	0.7513
	FIMLE	0.1852	0.3605	0.2346	1.0223
			MAE		
α	SMLE	0.2519	0.4845	0.5040	1.1760
	EXP	0.2293	0.4826	0.4439	1.1852
	SMEAR	0.2152	0.4713	0.4312	1.1657
	FIMLE	0.2035	0.4382	0.3894	1.1216
β	SMLE	0.1165	0.2327	0.1309	0.3361
	EXP	0.1165	0.2342	0.1286	0.3388
	SMEAR	0.1112	0.2307	0.1283	0.3325
	FIMLE	0.1117	0.2332	0.1265	0.3281

Table 4 Root mean squared errors (RMSE) and median absolute errors (MAE) for the linear logistic model under case-control sampling and given different amounts of auxiliary information on X. In all cases N = 5000 and $n_1 = n_0 = 100$. Values of X drawn from the standard lognormal distribution.

True (α, β)		(–3, 1)	(–5, 2)	(–5, 1)	(–8, 2)
			RMSE		
α	SMLE	1.2100	1.2717	2.3204	2.2361
	WTD	0.2964	0.5971	0.4797	1.6615
	EXP	0.2828	0.5558	3.3436	1.2723
	SMEAR	0.2804	0.5483	0.3956	1.2928
	FIMLE	0.2738	0.5339	0.3241	0.9561
β	SMLE	0.1735	0.3122	0.1546	0.4282
	WTD	0.1827	0.3346	0.1983	0.5607
	EXP	0.1741	0.3117	10.3296	0.4406
	SMEAR	0.1621	0.3015	0.1521	0.4330
	FIMLE	0.1509	0.2730	0.1003	0.2760
			MAE		
α	SMLE	1.1911	1.1947	2.3248	2.0940
	WTD	0.2015	0.4063	0.3026	0.8445
	EXP	0.1910	0.3615	0.2480	0.6714
	SMEAR	0.1909	0.3619	0.2459	0.6623
	FIMLE	0.1864	0.3454	0.1983	0.5617
β	SMLE	0.1131	0.2026	0.0942	0.2207
	WTD	0.1178	0.2126	0.1225	0.2819
	EXP	0.1120	0.2002	0.1096	0.2320
	SMEAR	0.1069	0.1957	0.1015	0.2243
	FIMLE	0.1002	0.1734	0.0629	0.1660

Figure 1 Simulated estimation errors for α in the linear model (1). The true value of α is 5 and sampling is SRSWOR with N = 1000 and n = 50. Errors are ordered along the horizontal axis by the rank of the sample *Y*-mean $\bar{y}_s$. Solid red line shows median estimation error by decile group of these sample means. Errors greater than 0.5 in absolute value are not shown. Out of a total of 1000 simulated errors, there were 9 such values for SMLE, 22 for LIMCAL, 30 for LIMMLE, 9 for CALW and 4 each for PRED and FIMLE.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Figure 2 Simulated estimation errors for α in the linear logistic model. The true value of (α, β) is (–5, 1) and sampling is SRSWOR with N = 5000 and n = 200. Errors are ordered along the horizontal axis by the corresponding rank of the sample Y-mean $\bar{y}_s$. Solid red line shows median estimation error within each decile group of these sample Y-means.