Generalised Linear Models Incorporating Population Level Information: An Empirical Likelihood Based Approach[1]

Sanjay Chaudhuri
University of Washington, Seattle

Mark S. Handcock
University of Washington, Seattle

Michael S. Rendall
Rand Corporation, Santa Monica

Working Paper no. 48
Center for Statistics and the Social Sciences
University of Washington

May 5, 2005

[1] Sanjay Chaudhuri is a graduate student, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322 E-mail:sanjay@stat.washington.edu; Web: www.stat.washington.edu/sanjay; Mark S. Handcock is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. E-mail: handcock@stat.washington.edu; Web: www.stat.washington.edu/handcock; Michael S. Rendall is with the Rand Corporation, Labor and Population Programme and Pennsylvania State University, Population Research Institute. E-mail: mrendall@rand.org.
This work was funded by the National Institute of Child Health and Human Development grant R01-HD043472-01. The authors would like to thank Thomas S. Richardson, Antar Bandyopadhyay and Ryan Admiral for their extremely useful comments and suggestions.

Abstract

In many situations information from a sample of individuals can be supplemented by population level information on the relationship between a dependent and explanatory variables. Inclusion of the population level information can reduce bias and increase the efficiency of the parameter estimates.

Population level information can be incorporated via constraints on the model parameters. In general the constraints are nonlinear making the task of maximum likelihood estimation harder. In this paper we develop an alternative approach exploiting the notion of an empirical likelihood. It is shown that within the framework of generalised linear models, the population level information corresponds to linear constraints, which are comparatively easy to handle. We provide a two-step algorithm that produces parameter estimates using only unconstrained estimation. We also provide computable expressions for the standard errors. We give an application to demographic hazard modeling by combining panel survey data from the British Household Panel Survey (BHPS) with birth registration data.

Key Words: Empirical Likelihood, Constrained Optimisation, Generalised Linear Models.

1 Introduction

In many statistical demographic applications, some information on the relationship of explanatory variable with the dependent variables is known from the population level data. Sources of population level data include a census, registration systems, and governmental filing. Population level data are comparatively free of sampling error and are typically less biased due to the effects of non response. However standard regression modelling techniques cannot be appropriately used with these data sets, as there are too few variables to estimate the model of dependence between the dependent variable and the explanatory variables. On the other hand the data obtained from sample surveys are subject to sampling errors and bias due to non-response. However they usually contain information about more variables and thus is essential to infer scientifically interesting specifications of the model. With the increasing popularity of likelihood based methods, more emphasis has been placed on estimation from sample data alone, with the information from complete enumeration procedures ignored in practice. It is not surprising that the population level information contains valuable information for statistical analysis whose inclusion can lead to statistically more accurate estimates and better inference.

The traditional method for combining the population level information in statistical analysis is through post stratification. The literature goes back to Deming and Stephan (1942) and Ireland and Kullback (1968), but the statistical literature on the statistical properties of post-stratified estimates has been relatively sparse (Holt and Smith, 1979). The post stratification approach is usually applied in the design-based inference and becomes inappropriate in the presence of non-response, unless strong restrictions on the nature of non-response is assumed. In the alternative model based approach, one can view non-response and attrition as further sample selection. Here the the combination of survey and population data is operationalised via a model and the composite sampling scheme is inferred from a selection model and the observed data. In this case post-stratification has been used to make the non-response ignorable for inference (Little and Rubin, 1987). Little (1993) develops a Bayesian model-based approach and Little and Wu (1991) compare the design-based and model based approaches when the sampled population differs from the population data due to non-response or coverage errors.

The use of both sample and population information within the framework of likelihood principles and more specifically for *generalised linear models* has been proposed by Handcock, Huovilainen, and Rendall (2000). This method does not provide post-stratification weights but directly constraints the likelihood. Imbens and Lancaster (1994) use a similar method for combining macro data in microeconomic models. In Handcock, Rendall, and Cheadle (2003) the methodology has been further developed. They express the population level data as (usually non-linear) functions of the model parameters and use them as restrictions to the likelihood. The *constrained maxi-*

mum likelihood estimates can then be obtained by maximising the likelihood function under these population level constraints.

This method can be implemented using any of the widely available procedures for numerical optimisation with equality constraints. It is also known that the estimates are asymptotically normal and unbiased. An explicit form of the standard error of the parameter estimates can also be obtained. Further one can show analytically that the standard error of the parameter estimates are guaranteed to be smaller compared to those with no population restriction.

Even though the method uses intuitively straight-forward procedures it has its own limitations. Non-linear equality constraints are in general numerically difficult to handle and time-consuming to code in applications involving multiple population level restrictions and even moderate numbers of regressors. Also knowledge about the distribution of the explanatory variables is required.

Data combination is a active research area in econometrics. Imbens and colleagues(Imbens and Lancaster, 1994; Hellerstein and Imbens, 1999) explore the benefits of combining population with survey data, using economic data in a *generalised method of moment* (GMM) regression framework. Imbens and Lancaster (1994) consider the estimation of parameters in the regression model under moment restrictions on the data contributed by population data, and report large gains in efficiency by incorporating marginal moments from census data with sample-survey joint distributions. They restrict on the conditional moment information if distributional knowledge on the explanatory variables is not available. For a review of recent developments see Moffitt and Ridder (2005).

The article is structured as follows. In Section 2 we introduce the notation and formulate the problem. It also contains a brief discussion of the empirical likelihood and translates the problem to a constrained empirical likelihood problem. Section 3 describes the methods for estimating the model parameter through a nested maximisation problem. We show that in many cases a two-step maximisation problem can be devised to estimate the parameters. The latter method generalises the weighted least squared estimation described in Hellerstein and Imbens (1999). We also provide the details of solving the linearly constrained optimisation problem associated with this estimation scheme. In Section 4 some asymptotic properties of the parameter estimates are derived. We show that if the sample is representative of the population then our parameter estimates are asymptotically consistent and unbiased. Under usual regularity conditions we further show that the estimates are asymptotically normal and give an explicit expression for their standard errors. We show that the use of population information reduces the standard errors. We indicate that these standard errors can be calculated using either the asymptotic weights or more accurately using the calculated weights through a *sandwich* estimator. In Section 5 we apply the methodology to combine panel survey data from the British Household Panel Survey (BHPS) with birth registration data to estimate annual birth probabilities by parity.

2 Notation and Formulation

2.1 Notation and problem description

Suppose Y denote the response variable dependent on the explanatory variables $\underline{X} = \left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right)$. Suppose $g_j\left(Y, X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right)$, $j = 1, 2, \ldots, m$ be functions of the response and p explanatory variables such that the average value of each g_j over the population represented by $Y, X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ is known to be γ_j, for all $j = 1, 2, \ldots, m$. Moreover assume that each γ_j is known without any sampling error.

Let us have a sample of size n. The ith sample point is denoted by

$$\left(y_i, \underline{x}_i\right) \equiv \left(y_i, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}\right) \qquad i = 1, 2, \ldots, n \tag{1}$$

We assume that $n > p + m$. We assume that there are no missing values.

Suppose we fit a generalised linear model $\mathcal{M}$ to the data and incorporate the population level information in the known population mean of g_j, for all $j = 1, 2, \ldots, m$. The model is specified as

$$E\left(Y|\underline{X} = \underline{x}_i, \underline{\beta}\right) = \eta_{\mathcal{M}}\left(x_i \underline{\beta}^T\right) \equiv \eta_{\mathcal{M}}\left(\sum_{k=1}^{p} \beta_k x_i^{(k)}\right) \tag{2}$$

where $\eta_{\mathcal{M}}(\cdot)$ is the inverse link function corresponding to the family $\mathcal{M}$ and $\underline{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_p\right)$ is the vector of regression coefficients. We assume that the marginal distribution of $\underline{X}$ does not depend on $\underline{\beta}$.

Since γ_j is the known value of g_j in the population, we can assume that for $j = 1, 2 \ldots, m$,

$$E_{Y,\underline{X}}\left[g_j\left(Y, \underline{X}\right)\right] = \gamma_j. \tag{3}$$

Note that in (3) the L.H.S. is in general a nonlinear function of $\underline{\beta}$. Thus using these expression directly as constraints on the population parameters (as in Handcock et al. (2003)) imposes nonlinear constraints on the problem. Thus in order to find the parameter estimates we need to maximise the likelihood corresponding to $\mathcal{M}$ under those constraints.

In the alternative approach we estimate the joint distribution of Y and $\underline{X}$ explicitly by its *empirical distribution* from the sample, incorporating the conditional moment information from (2) and the population level information from (3). We show that this distribution can be estimated by solving a maximisation problem constrained only by linear constraints. The parameter estimates can then be obtained from this estimated distribution.

2.2 Empirical distribution and likelihood

We only introduce the concepts of empirical likelihood and empirical distribution for univariate random variables in this section. Extensions for multivariate random variables are natural.

Suppose $Z_1, Z_2, \ldots, Z_n$ are i.i.d. univariate random variable with a common distribution F_0. Let $\mathcal{F}$ be the set of all univariate distribution functions. In particular $F_0 \in \mathcal{F}$.

Definition 1 (Owen (2001)). *Suppose $F \in \mathcal{F}$, then the non-parametric likelihood of F is defined as*

$$L(F) = \prod_{i=1}^{n} \left(F(Z_i) - F(Z_i-)\right), \tag{4}$$

where $F(Z_i-) = \lim_{\delta \downarrow 0} F(Z_i - \delta)$.

In Definition 1. we use the word "likelihood" to mean that $L(F)$ in (4) is the probability of the sample $Z_1, Z_2, \ldots, Z_n$ from the distribution F. Also we note that, if F is continuous $\lim_{\delta \downarrow 0} F(Z_i - \delta) = F(Z_i)$, for all i. So if F is continuous $L(F) = 0$. Thus for $L(F)$ to be positive F has to put positive mass on every sample point $Z_1, Z_2, \ldots, Z_n$ and be discrete.

The *empirical distribution* function of the sample $Z_1, Z_2, \ldots, Z_n$ is defined as

$$\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Z_i \le z\}}, \qquad -\infty < z < \infty. \tag{5}$$

Owen (2001) shows that the non-parametric likelihood in (4) is maximised over $\mathcal{F}$ for $F(z) = \hat{F}_n(z)$ for all $-\infty < z < \infty$. Therefore F_n is the *non-parametric maximum likelihood estimator* (NPMLE) of F. Thus in a completely non-parametric setting the best estimator of the underlying distribution assigns equal weight of $1/n$ to each of its sample points.

The non-parametric likelihood in (4) can be used for parametric distribution also. Suppose it is known that the distribution depends on a parameter $\theta \in \Theta$ and let

$$\mathcal{F}_\Theta = \{F_\theta \text{ s.t. } F_\theta \text{ is an univariate distribution function and } \theta \in \Theta.\} \tag{6}$$

Clearly $\mathcal{F}_\Theta \subseteq \mathcal{F}$.

The NPMLE, $\hat{F}_{n,\theta} \in \mathcal{F}_\Theta$ in this case can be obtained by maximising $L(F)$ in (4) over $F \in \mathcal{F}_\Theta$. This means finding n sample weights $\hat{w}_i = \left\{\hat{F}_{n,\theta}(Z_i) - \hat{F}_{n,\theta}(Z_i-)\right\}$ s.t. $\hat{w}_i \ge 0$, for all $i = 1, 2, \ldots, n$, $\sum_{i=1}^{n} \hat{w}_i = 1$ and

$$L\left(F_{n,\theta}\right) = \prod_{i=1}^{n} \left\{F_{n,\theta}(Z_i) - F_{n,\theta}(Z_i-)\right\} = \prod_{i=1}^{n} \hat{w}_i \tag{7}$$

is the maximum value of $L(F)$ for $F \in \mathcal{F}_\Theta$.

From the asymptotic point of view, a scaled version of the non-parametric likelihood is more interesting to consider. One can use the *non-parametric likelihood ratio* of $F \in \mathcal{F}_\Theta$ to $F \in \mathcal{F}$.

This is given by

$$R\left(F_{\Theta}\right)=\frac{\max_{F\in\mathcal{F}_{\Theta}}L\left(F\right)}{\max_{F\in\mathcal{F}}L\left(F\right)}=\frac{\max_{F\in\mathcal{F}_{\Theta}}\prod_{i=1}^{n}\left\{F\left(Z_i\right)-F\left(Z_i-\right)\right\}}{\prod_{i=1}^{n}\frac{1}{n}} \tag{8}$$

$$=\max_{F\in\mathcal{F}_{\Theta}}\prod_{i=1}^{n}n\left\{F\left(Z_i\right)-F\left(Z_i-\right)\right\}. \tag{9}$$

From (7) it follows that

$$R\left(F_{\Theta}\right)=\prod_{i=1}^{n}n\left\{\hat{F}_{n,\theta}\left(Z_i\right)-\hat{F}_{n,\theta}\left(Z_i-\right)\right\}=\prod_{i=1}^{n}n\hat{w}_i. \tag{10}$$

In a sample n is a constant, so considering $R\left(F_{\theta}\right)$ is actually same as considering $L\left(F_{\theta}\right)$.

From the maximising weights $\underline{\hat{w}}=\left(\hat{w}_1,\hat{w}_2,\ldots,\hat{w}_n\right)$ the constrained estimate of the distribution function is determined by

$$\hat{F}_{n,\theta}(z)=\hat{F}_{\underline{w}}\left(z\right)=\sum_{i=1}^{n}\hat{w}_i 1_{\{Z_i\leq z\}},\qquad \text{for } -\infty<z<\infty. \tag{11}$$

So in order to estimate $\hat{F}_{n,\theta}$ it is enough to determine the maximising weights $\hat{w}_i$ for the *empirical likelihood* $\prod_{i=1}^{n}nw_i$ under the constraint that the resulting distribution $\hat{F}_{\underline{w}}\in\mathcal{F}_{\Theta}$. In other words the NPMLE $\hat{F}_{n,\theta}$ is determined by

$$\underline{\hat{w}}=\arg\max_{\underline{w}}\left\{\prod_{i=1}^{n}nw_i\ :\ w_i\geq 0,\forall i,\sum_{i=1}^{n}w_i=1,F_{\underline{w}}\in\mathcal{F}_{\Theta}\right\}. \tag{12}$$

All the concepts described above can be extended to multivariate random variables. The extension is natural and we refrain from repeating it. Details can be found in Owen (2001, Chapter 3).

2.3 Formulation of the empirical likelihood with population level information

In the problem under consideration we not only have parameters describing the relationship of Y and $\underline{X}$ in (2), moreover there are m population level information in (3) imposing more constraints on the set of allowable joint distribution of Y and $\underline{X}$. Thus in order to estimate this joint distribution one needs to maximise the corresponding non-parametric likelihood in (4) over the set of all joint distributions satisfying both sets of constraints imposed by the parametric model and the population information.

From the discussion above the estimator is given by the weights $\hat{w}_i$ such that

$$\hat{w}_i\geq 0 \text{ for all } i=1,2,\ldots,n \text{ and } \sum_{i=1}^{n}\hat{w}_i=1, \tag{13}$$

maximising the empirical likelihood $\prod_{i=1}^{n} nw_i$, or equivalently

$$\log\left(\prod_{i=1}^{n} nw_i\right) = \sum_{i=1}^{n} \log(nw_i) \tag{14}$$

over $\underline{w}$, under some more constraints imposed by the model $\mathcal{M}$ and population level information.

The constraints imposed by the model are based on the score functions of the likelihood. It is well known (McCullagh and Nelder, 1989) that for any generalised linear model for $k = 1, 2, \ldots, p$

$$E\left[X^{(k)}\left(Y - \eta_{\mathcal{M}}\left(\sum_{k=1}^{p} \beta_k X^{(k)}\right)\right)\right] = 0. \tag{15}$$

This *expectation* is over the joint distribution of the sample. In our formulation we replace the joint distribution by its *empirical* estimate. Thus the *score constraints* imposed by the model $\mathcal{M}$ on (14) through its score functions are given by

$$\sum_{i=1}^{n} w_i x_i^{(k)} \left\{y_i - \eta_{\mathcal{M}}\left(\sum_{k=1}^{p} \beta_k x_i^{(k)}\right)\right\} = 0 \qquad k = 1, 2, \ldots, p. \tag{16}$$

The *population constraints* on (14) can be expressed similarly from (3) as

$$\sum_{i=1}^{n} w_i \left\{g_j\left(y_i, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}\right) - \gamma_j\right\} = 0 \qquad j = 1, 2, \ldots, m. \tag{17}$$

Unlike the direct constraints method (Handcock et al., 2003), in (17) the population level constraints do not explicitly depend on the model parameters $\underline{\beta}$. However as the score constraints involve all the weights and the model parameters, eventually same constraints are imposed on the parameter estimation in both the procedures.

3 Estimation of model parameters

The empirical joint distribution is determined by the vector of maximising weights $\underline{\hat{w}} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n)$ satisfying (13) and other $m+p$ constraints in (16) and (17). In this section we discuss the numerical methods for maximising the empirical likelihood ratio and estimating the model parameters.

3.1 Estimation of the model parameters

The main goal of this discussion is to estimate the model parameters incorporating the population constraints. Under the assumption of (14), (16) and (17) the parameter estimates are given by

$$\begin{aligned} \underline{\hat{\beta}} = \arg\max_{\underline{\beta}} \Bigg[\max_{\underline{w}} \Bigg\{ & \sum_{i=1}^{n} \log(nw_i),\, w_i \geq 0, \forall i = 1, 2, \ldots, n, \sum_{i=1}^{n} w_i = 1, \qquad (18) \\ & \sum_{i=1}^{n} w_i x_i^{(k)} \left\{ y_i - \eta_{\mathcal{M}} \left(\underline{x}_i \underline{\beta} \right) \right\} = 0, \;\; k = 1, 2, \ldots, p, \\ & \sum_{i=1}^{n} w_i \left\{ g_j \left(y_i, \underline{x}_i \right) - \gamma_j \right\} = 0, \;\; j = 1, 2, \ldots, m \Bigg\} \Bigg]. \end{aligned}$$

Thus the parameter estimates are obtained from a nested maximisation procedure and the model parameters only influence the estimation of weights through the score constraints.

Note that in (18), the outer maximisation over $\underline{\beta}$ is unconstrained. The inner one over $\underline{w}$ is constrained by linear constraints. So the use of empirical likelihood requires maximising over linear constraints only, which is numerically much easier to handle, than non-linear constraints.

However this nested maximisation may be slow, but in most cases an alternative and faster two-step estimation procedure can be devised to obtain the parameter estimates instantly.

3.2 A two-step procedure for estimating the model parameters

The two step estimator for $\underline{\beta}$ takes the advantage of the fact that the population constraints do not depend on the model parameters. This method is a generalisation of the weighted estimator described in Hellerstein and Imbens (1999). In the first step we get the weights $\underline{\hat{w}}$ maximising (14) under (13) and the population constraints (17). That is we compute

$$\begin{aligned} \underline{\hat{w}} = \arg\max_{\underline{w}} \Bigg[& \sum_{i=1}^{n} \log(nw_i),\, w_i \geq 0,\; i = 1, 2, \ldots, n, \sum_{i=1}^{n} w_i = 1, \qquad (19) \\ & \sum_{i=1}^{n} w_i \left\{ g_j \left(y_i, \underline{x}_i \right) - \gamma_j \right\} = 0, \;\; j = 1, 2, \ldots, m \Bigg]. \end{aligned}$$

In the second step we solve the score constraints with $\hat{w}_i$ as the sample weights for the parameter estimates. That is we solve

$$\sum_{i=1}^{n} \hat{w}_i x_i^{(k)} \left\{ y_i - \eta_{\mathcal{M}} \left(\underline{x}_i \underline{\beta} \right) \right\} = 0, \;\; k = 1, 2, \ldots, p, \qquad (20)$$

for the parameter estimates $\underline{\hat{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \right)$.

The second step consists of the usual estimation of the model parameters for the model $\mathcal{M}$ with $\underline{\hat{w}}$ as the vector of sample weights. So we can use standard methods for estimating model parameters for a generalised linear model to perform this step (McCullagh and Nelder, 1989).

A rationale validating the two-step estimation procedure can be outlined as follows. Suppose we define

$$\begin{aligned}\mathcal{A} = \Bigg\{ \underline{w} : w_i \geq 0,\ i = 1, 2, \ldots, n, \sum_{i=1}^{n} w_i = 1, \\ \sum_{i=1}^{n} w_i \left\{ g_j \left(y_i, \underline{x}_i \right) - \gamma_j \right\} = 0, \quad j = 1, 2, \ldots, m \Bigg\}.\end{aligned} \tag{21}$$

and

$$\begin{aligned}\mathcal{B} = \Bigg\{ \underline{w} : w_i \geq 0,\ i = 1, 2, \ldots, n, \sum_{i=1}^{n} w_i = 1, \\ \sum_{i=1}^{n} w_i \left\{ g_j \left(y_i, \underline{x}_i \right) - \gamma_j \right\} = 0, \quad j = 1, 2, \ldots, m, \\ \sum_{i=1}^{n} w_i x_i^{(k)} \left\{ y_i - \eta_{\mathcal{M}} \left(\underline{x}_i \underline{\beta} \right) \right\} = 0, \quad k = 1, 2, \ldots, p \Bigg\}\end{aligned} \tag{22}$$

Clearly both $\mathcal{A}$ and $\mathcal{B}$ are convex and $\mathcal{B} \subseteq \mathcal{A}$. However $\sum_{i=1}^{n} \log(n w_i)$ is concave on $\mathcal{A}$. Thus there is an unique maximising $\underline{\hat{w}}$ in $\mathcal{A}$. If $\underline{\hat{w}} \in \mathcal{B}$, i.e. there is $\underline{\beta}$ satisfying the score constraints $\underline{\hat{w}}$ is the intended maximising weights.

However it is possible that there is no $\underline{\beta}$ satisfying the score constraints for the weights $\underline{\hat{w}}$, i.e. $\underline{\hat{w}} \notin \mathcal{B}$. In that situation the two-step procedure described above fails and the nested maximisation in (18) has to be done. The parameter estimates from the two step procedure are still useful as initial values for the nested maximisation. Hellerstein and Imbens (1999) estimate the model parameters by a similar two-step estimation procedure even if $\underline{\hat{w}} \notin \mathcal{B}$.

3.3 Computing the maximising weights in the first step

In the two-step estimation procedure discussed above, only the first step requires constrained maximisation.

We follow a similar discussion in Owen (2001) and instead of solving the primal problem for the maximising weights, solve its dual problem, which is more convenient to deal with.

Without loss of generality we can assume that $\hat{w}_i > 0, \forall i = 1, 2, \ldots, n$. The situation when some weights are 0 can be tackled fairly easily (Owen, 2001).

The primal objective function is given by

$$L\left(\underline{w}, \underline{\beta}\right) = \sum_{i=1}^{n} \log(nw_i) - \alpha\left(\sum_{i=1}^{n} w_i - 1\right) + \sum_{j=1}^{m} n\lambda_j \sum_{i=1}^{n} w_i \left\{g_j\left(y_i, \underline{x}_i\right) - \gamma_j\right\}. \tag{23}$$

Here $\alpha, \lambda_j,\ j = 1, 2, \ldots, m$ are the *Lagrangian multipliers*.

By differentiating with respect to w_i for each i and equating the resulting expression to 0 we get, for all $i = 1, 2, \ldots, n$

$$\frac{\partial L}{\partial w_i} = \frac{1}{w_i} - \alpha + \sum_{j=1}^{m} n\lambda_j \left\{g_j\left(y_i, \underline{x}_i\right) - \gamma_j\right\} = 0. \tag{24}$$

Note that $\sum_{i=1}^{n} w_i \frac{\partial L}{\partial w_i} = 0$. Thus using the facts that the weights sum to 1 and the weighted sum of the restrictions equals 0 for all $j = 1, 2, \ldots, m$ we get

$$\alpha = n. \tag{25}$$

By substituting the value of α in (24) and solving for the weights, we get

$$w_i = \frac{1}{n} \times \frac{1}{1 + \sum_{j=1}^{m} \lambda_j \left\{g_j\left(y_i, \underline{x}_i\right) - \gamma_j\right\}}. \tag{26}$$

Since $0 < w_i < 1$ for all $i = 1, 2, \ldots, n$, it follows that

$$1 + \sum_{j=1}^{m} \lambda_j \left\{g_j\left(y_i, \underline{x}_i\right) - \gamma_j\right\} \geq \frac{1}{n} \qquad\qquad i = 1, 2, \ldots, n. \tag{27}$$

By substituting the weights in (23) (or equivalently in (14)) we get

$$\sum_{i=1}^{n} \log(nw_i) = -\sum_{i=1}^{n} \log(1 + \sum_{j=1}^{m} \lambda_j \left\{g_j\left(y_i, \underline{x}_i\right) - \gamma_j\right\}). \tag{28}$$

Under the assumption that there is no duality gap, maximising (14) under constraints for weights is equivalent to minimising (28) under the constraints in (27) for the optimal values of $\lambda_j,\ j = 1, 2, \ldots, m$.

Note that the dimension of the dual problem is m which is much less than the dimension of the primal problem by assumption. Also the dual problem is constrained by linear inequality constraints and thus numerically far easier to solve.

Owen (2001, Chapter 3) defines a continuous and twice differentiable pseudo-logarithmic function over the whole real line such that the above minimisation can be performed without any constraint using a simple Newton-Raphson Algorithm.

The procedure for the inner constrained maximisation in (18) which includes the score constraints, when necessary, is similar. We note that in that case the dimension of the corresponding dual problem is $m + p$, which is still less than the dimension of the primal problem.

4 Asymptotic properties of the estimator

In this section we investigate the asymptotic properties of the constrained estimator of the model parameters. The main emphasis will be given on the two-step estimator in section 3. We shall show that this estimator is consistent and asymptotically normal. Also we shall find the form of the asymptotic Variance-Covariance matrix and show that the standard errors of the constrained estimator is less than that of the unconstrained one.

4.1 Consistency and asymptotic normality of the parameter estimators

For brevity of notation let us denote for all $i = 1, 2, \ldots, n$, $h_j(y_i, \underline{x}_i, \gamma_j) = g_j\left(y_i, \underline{x}_i\right) - \gamma_j$, $\mu_i = \sum_{k=1}^{p} \beta_k x_i^{(k)}$.

Define function f over the observations and model parameters as

$$f\left(y, \underline{x}, \underline{\beta}, \underline{\lambda}\right)_{(p+m)\times 1} = \begin{pmatrix} \frac{x^{(1)}(y-\eta_{\mathcal{M}}(\mu))}{1+\sum_{j=1}^{m}\lambda_j h_j(y_i,\underline{x}_i,\gamma_j)} \\ \vdots \\ \frac{x^{(p)}(y-\eta_{\mathcal{M}}(\mu))}{1+\sum_{j=1}^{m}\lambda_j h_j(y_i,\underline{x}_i,\gamma_j)} \\ \frac{h_1(y,\underline{x},\gamma_j)}{1+\sum_{j=1}^{m}\lambda_j h_j(y_i,\underline{x}_i,\gamma_j)} \\ \vdots \\ \frac{h_m(y,\underline{x},\gamma_j)}{1+\sum_{j=1}^{m}\lambda_j h_j(y_i,\underline{x}_i,\gamma_j)} \end{pmatrix} \tag{29}$$

and

$$\underline{h}\left(y, \underline{x}, \underline{\gamma}\right)_{m\times 1} = \left(h_1\left(y, \underline{x}, \gamma_1\right), h_2\left(y, \underline{x}, \gamma_2\right), \ldots, h_m\left(y, \underline{x}, \gamma_m\right)\right)^t. \tag{30}$$

We start by showing the strong consistency of the estimator of the model parameters. In Lemma 1 we show that asymptotically as $n \to \infty$, each Lagrangian multipliers minimising the dual problem in (19) tend to 0 almost surely. The second step involves a weighted maximum likelihood estimation, the consistency of the parameter estimates follow. We outline a proof of consistency and the asymptotic normality of the estimator in Theorem 1.

Lemma 1. *Suppose* $\left(Y_1, \underline{X}_1\right), \left(Y_2, \underline{X}_2\right), \ldots, \left(Y_n, \underline{X}_n\right)$ *are i.i.d. random vectors in* $\mathbb{R}^{p+1}$, $E\left[\underline{h}\left(Y, \underline{X}, \underline{\gamma}\right)\right] = \underline{0}$ *and* $Var\left[\underline{h}\left(Y_1, \underline{X}_1, \underline{\gamma}\right)\right]$ *has rank larger than* 0 *. Let* $\hat{\underline{\lambda}}$ *be the vector of Lagrangian multipliers minimising the dual in* (28)*. Then with probability* 1 *as* $n \to \infty$, $\hat{\underline{\lambda}} \to \underline{0}$.

Proof. Note that from the population level restrictions the vector of Lagrange multipliers solving the dual problem (28) satisfy

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)}{1 + \hat{\underline{\lambda}}^t \underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)} = \underline{0}. \tag{31}$$

The rest of the proof follows from Owen (1990)(Theorem 1) mutatis mutandis. □

We now prove the consistency and get an explicit expression for the asymptotic Variance - Covariance matrix of the parameter estimates.

Theorem 1. *Suppose that the assumptions of Lemma 1 hold. Also assume that*

1. *The Jacobian matrix* $\frac{\partial f\left(y,\underline{x},\underline{\beta},\underline{\lambda}\right)}{\partial(\underline{\beta},\underline{\lambda})}$ *and the Hessian* $\frac{\partial^2 f\left(y,\underline{x},\underline{\beta},\underline{\lambda}\right)}{\partial^2\underline{\beta}\partial^2\underline{\lambda}}$ *exists for all* $\underline{\beta}$ *and* $\underline{\lambda}$*.*
2. $f\left(y,\underline{x},\underline{\beta},\underline{\lambda}\right)$, $\frac{\partial f\left(y,\underline{x},\underline{\beta},\underline{\lambda}\right)}{\partial(\underline{\beta},\underline{\lambda})}$ *and* $\frac{\partial^2 f\left(y,\underline{x},\underline{\beta},\underline{\lambda}\right)}{\partial^2\underline{\beta}\partial^2\underline{\lambda}}$ *are elementwise bounded by integrable function in a neighbourhood nbd* $\left(\underline{\beta},\underline{\lambda}\right)$ *of the true value* $\left(\underline{\beta}^\star,\underline{0}\right)$*.*
3. $E\left[f\left(y,\underline{x},\underline{\beta}^\star\right)f^t\left(y,\underline{x},\underline{\beta}^\star\right)\right]$ *is positive definite,*
4. $E\left[\frac{\partial f\left(y,\underline{x},\underline{\beta}^\star\right)}{\partial\underline{\beta}}\right]$ *has full rank.*

Suppose that $\mu_i^\star=\sum_{k=1}^{p}\beta_k^\star X_i^{(k)}$, $\eta'_{\mathcal{M}}\left(\mu_i\right)=\frac{\partial\eta_{\mathcal{M}}(\mu_i)}{\partial\mu_i}$, *for all* $i=1,2,\ldots,n$ *and denote*

$$\mathbf{X}_{n\times p}=\left(\underline{X}_1^t,\ldots,\underline{X}_n^t\right)^t \tag{32}$$

$$\mathbf{h}_{n\times m}=\left(\underline{h}\left(Y_1,\underline{X}_1,\underline{\gamma}\right),\ldots,\underline{h}\left(Y_n,\underline{X}_n,\underline{\gamma}\right)\right)^t \tag{33}$$

Assume that

$$G_{p\times p}=E_{\underline{X},\underline{\beta}^\star}[\mathbf{X}^t diag\left[\eta'_{\mathcal{M}}(\mu_1^\star),\ldots,\eta'_{\mathcal{M}}(\mu_n^\star)\right]\mathbf{X}], \tag{34}$$

$$G^\star_{p\times p}=E_{\underline{X},\underline{\beta}^\star}\left[\mathbf{X}^t diag\left[\left(Y_1-\eta_{\mathcal{M}}(\mu_1^\star)\right)^2,\ldots,\left(Y_n-\eta_{\mathcal{M}}(\mu_n^\star)\right)^2\right]\mathbf{X}\right], \tag{35}$$

$$T_{p\times m}=E_{\underline{X},\underline{\beta}^\star}\left[\mathbf{X}^t diag\left[\left(Y_1-\eta_{\mathcal{M}}(\mu_1^\star)\right),\ldots,\left(Y_n-\eta_{\mathcal{M}}(\mu_n^\star)\right)\right]\mathbf{h}\right], \tag{36}$$

$$H_{m\times m}=E_{\underline{X},\underline{\beta}^\star}[\mathbf{h}^t\mathbf{h}]. \tag{37}$$

where $diag[\cdot]$ *is the diagonal matrix with the arguments as its diagonal elements. Also assume that* H *is non-singular. Then as* $n\to\infty$

$$1.\ \hat{\underline{\beta}}\longrightarrow\underline{\beta}^\star \textit{ almost everywhere}, \tag{38}$$

$$2.\ \sqrt{n}\left(\hat{\underline{\beta}}-\underline{\beta}^\star\right)\Longrightarrow N\left(0,G^{-1}\left\{G^\star-TH^{-1}T^t\right\}G^{-1}\right), \tag{39}$$

$$3.\ \sqrt{n}\hat{\underline{\lambda}}\Longrightarrow N\left(0,H\right), \tag{40}$$

$$4.\ \hat{\underline{\beta}} \textit{ and } \hat{\underline{\lambda}} \textit{ are asymptotically independent.} \tag{41}$$

Proof. From the discussion in the previous section we note that $\forall k = 1, 2, \ldots, p$ and $\forall j = 1, 2, \ldots, m$

$$\sum_{i=1}^{n} w_i x_i^{(k)} \left(y_i - \eta_{\mathcal{M}}(\mu_i)\right) = \sum_{i=1}^{n} \frac{x_i^{(k)} \left(y_i - \eta_{\mathcal{M}}(\mu_i)\right)}{1 + \sum_{j=1}^{m} \lambda_j h_j(y_i, \underline{x}_i, \gamma_j)} = 0. \tag{42}$$

$$\sum_{i=1}^{n} w_i h_j(y_i, \underline{x}_i, \gamma_j) = \sum_{i=1}^{n} \frac{h_j(y_i, \underline{x}_i, \gamma_j)}{1 + \sum_{j=1}^{m} \lambda_j h_j(y_i, \underline{x}_i, \gamma_j)} = 0. \tag{43}$$

Suppose $\hat{\underline{\beta}}$ and $\hat{\underline{\lambda}}$ are the vector of estimates of the model parameters and the Lagrangian multipliers. then by assumption

A Taylor series expansion of $\frac{1}{n}\sum_{i=1}^{n} f\left(y_i, \underline{x}_i, \underline{\beta}, \underline{\lambda}\right)$ around $(\underline{\beta}^{\star}, \underline{0})$ gives

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} f\left(y_i, \underline{x}_i, \underline{\beta}, \underline{\lambda}\right) - \frac{1}{n}\sum_{i=1}^{n} f(y_i, \underline{x}_i, \underline{\beta}^{\star}, \underline{0}) \\ = &\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial f(y_i, \underline{x}_i, \underline{\beta}, \underline{0})}{\partial(\underline{\beta}, \underline{\lambda})}\right]\left[\begin{array}{c} \underline{\beta} - \underline{\beta}^{\star} \\ \underline{\lambda} \end{array}\right] + \frac{\zeta}{2n}\left[(\underline{\beta} - \underline{\beta}^{\star})^t, \underline{\lambda}^t\right]\sum_{i=1}^{n}\mathcal{H}(y_i, \underline{x}_i)\left[\begin{array}{c} \underline{\beta} - \underline{\beta}^{\star} \\ \underline{\lambda} \end{array}\right], \end{aligned} \tag{44}$$

where $|\zeta| \le 1$ and $\mathcal{H}(Y, \underline{X})$ is integrable and is a bound to the Hessian matrix in $nbd(\underline{\beta}^{\star}, \underline{0})$.

The Jacobian matrix $\frac{\partial f(y_i, \underline{x}_i, \underline{\beta}, \underline{\lambda})}{\partial(\underline{\beta}, \underline{\lambda})}$ is given by

$$\frac{\partial f(y_i, \underline{x}_i, \underline{\beta}^{\star}, \underline{0})}{\partial(\underline{\beta}, \underline{\lambda})} = \left[\begin{array}{cc} -\frac{\underline{x}_i^t \underline{x}_i \eta'_{\mathcal{M}}(\mu_i)}{1+\underline{\lambda}^t \underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)} & \frac{\underline{h}\left(y_i \underline{x}_i \underline{\gamma}\right)\underline{x}_i\{y_i - \eta_{\mathcal{M}}(\mu_i)\}}{\left\{1+\underline{\lambda}^t \underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)\right\}^2} \\ 0 & \frac{\underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)\underline{h}^t\left(y_i, \underline{x}_i, \underline{\gamma}\right)}{\left\{1+\underline{\lambda}^t \underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)\right\}^2} \end{array}\right] \tag{45}$$

The expression of the Jacobian matrix at $\left(\underline{\beta}^{\star}, \underline{0}\right)$ is given by

$$\frac{\partial f(y_i, \underline{x}_i, \underline{\beta}^{\star}, \underline{0})}{\partial(\underline{\beta}, \underline{\lambda})} = \left[\begin{array}{cc} -\underline{x}_i^t \underline{x}_i \eta'_{\mathcal{M}}(\mu_i) & \underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}\right)\underline{x}_i^t\left\{y_i - \eta_{\mathcal{M}}(\mu_i)\right\} \\ 0 & \underline{h}\left(y_i, \underline{x}_i, \underline{\gamma}_i\right)\underline{h}^t\left(y_i, \underline{x}_i, \underline{\gamma}\right) \end{array}\right] \tag{46}$$

Clearly by assumption of the problem

$$E\left[f\left(Y_1, \underline{X}_1, \underline{\beta}^{\star}, \underline{0}\right)\right] = \underline{0}. \tag{47}$$

$E\left[\mathcal{H}(Y_1, \underline{X}_1)\right]$ is finite and

$$E\left[\frac{\partial f\left(Y_1, \underline{X}_1, \underline{\beta}, \underline{0}\right)}{\partial(\underline{\beta}, \underline{\lambda})}\right] = \left[\begin{array}{cc} -G & T \\ 0 & H \end{array}\right]. \tag{48}$$

So by the *S.L.L.N.* with probability 1

$$\frac{1}{n}\sum_{i=1}^{n} f\left(y_i, \underline{x}_i, \underline{\beta}^\star, \underline{0}\right) \longrightarrow 0 \tag{49}$$

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial f\left(y_i, \underline{x}_i, \underline{\beta}, \underline{0}\right)}{\partial(\underline{\beta}, \underline{\lambda})} \longrightarrow \begin{bmatrix} -G & T \\ 0 & H \end{bmatrix}, \tag{50}$$

and $\frac{1}{n}\sum_{i=1}^{n} \mathcal{H}\left(y_i, \underline{x}_i\right) \to E\left[\mathcal{H}(Y_1, \underline{X}_1)\right]$. Also by the *C.L.T.* it follows that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f\left(y_i, \underline{x}_i, \underline{\beta}^\star, \underline{0}\right)\right) \longrightarrow N\left(\underline{0}, V\right). \tag{51}$$

where

$$V = Var\left[f\left(Y_1, \underline{X}_1, \underline{\beta}^\star, \underline{0}\right)\right] = \begin{bmatrix} G^\star & T \\ T^t & H \end{bmatrix} \tag{52}$$

From this by following the steps in Serfling (1980, Chapter 4.2) the strong consistency of $\hat{\underline{\beta}}$ follows.

To show the asymptotic normality we note that

$$\sum_{i=1}^{n} f\left(y_i, \underline{x}_i, \hat{\underline{\beta}}, \hat{\underline{\lambda}}\right) = 0. \tag{53}$$

Thus for large n with probability 1, (44) holds in a neighbourhood of $\left(\underline{\beta}^\star, \underline{0}\right)$. So it follows that

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^{n} f(y_i, \underline{x}_i, \underline{\beta}^\star, \underline{0}) + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial f(y_i, \underline{x}_i, \underline{\beta}, \underline{0})}{\partial(\underline{\beta}, \underline{\lambda})}\right]\begin{bmatrix} \hat{\underline{\beta}} - \underline{\beta}^\star \\ \hat{\underline{\lambda}} \end{bmatrix} \\ + \frac{\zeta}{2n}\left[(\hat{\underline{\beta}} - \underline{\beta}^\star)^t, \hat{\underline{\lambda}}^t\right]\sum_{i=1}^{n}\mathcal{H}(y_i, \underline{x}_i)\begin{bmatrix} \hat{\underline{\beta}} - \underline{\beta}^\star \\ \hat{\underline{\lambda}} \end{bmatrix}. \end{aligned} \tag{54}$$

Rearranging the terms it follows that for large n, with probability 1

$$\begin{aligned} \sqrt{n}\begin{bmatrix} \hat{\underline{\beta}} - \underline{\beta}^\star \\ \hat{\underline{\lambda}} \end{bmatrix} + \left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial f(y_i, \underline{x}_i, \underline{\beta}, \underline{0})}{\partial(\underline{\beta}, \underline{\lambda})}\right] + \frac{\zeta}{2n}\left[(\hat{\underline{\beta}} - \underline{\beta}^\star)^t, \hat{\underline{\lambda}}^t\right]\sum_{i=1}^{n}\mathcal{H}(y_i, \underline{x}_i)\right]^{-1} \\ \times\left[\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(y_i, \underline{x}_i, \underline{\beta}^\star, \underline{0})\right)\right] \to \underline{0}. \end{aligned} \tag{55}$$

Now from this, the consistency of $\hat{\underline{\beta}}$, $\hat{\underline{\lambda}}$, (48), (51), (52) and Slutsky's theorem it is evident that

$$\sqrt{n}\begin{pmatrix} \hat{\underline{\beta}} - \underline{\beta}^\star \\ \hat{\underline{\lambda}} \end{pmatrix} \longrightarrow N\left(\underline{0}, V^\star\right), \tag{56}$$

where

$$V^\star = \begin{pmatrix} -G & T \\ 0 & H \end{pmatrix}^{-1} \begin{pmatrix} G^\star & T \\ T^t & H \end{pmatrix} \begin{pmatrix} -G & 0 \\ T^t & H \end{pmatrix}^{-1} \tag{57}$$

$$= \begin{pmatrix} -G^{-1} & G^{-1}TH^{-1} \\ 0 & H^{-1} \end{pmatrix}^{-1} \begin{pmatrix} G^\star & T \\ T^t & H \end{pmatrix} \begin{pmatrix} -G^{-1} & 0 \\ H^{-1}T^tG^{-1} & H^{-1} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} G^{-1}\left[G^\star - TH^{-1}T^t\right]G^{-1} & 0 \\ 0 & H \end{pmatrix}. \tag{58}$$

From (58) the rest of the assertions of the theorem follow. □

4.2 Reduction of standard errors of the model parameters

It is well known that without any population constraints the asymptotic variance covariance matrix of $\sqrt{n}\left(\underline{\hat{\beta}} - \underline{\beta}^\star\right)$ is given by $G^{-1}G^\star G^{-1}$. Also $G^{-1}TH^{-1}T^tG^{-1}$ is positive definite. Thus inclusion of the population constraints reduces the standard error of the model parameters. However we should note that this result is asymptotic and for finite samples the reduction of standard error for some coefficients may be trivial. We shall illustrate this fact in Section 5 below.

4.3 Estimating the asymptotic Variance-Covariance matrix

The asymptotic Variance-Covariance matrix can be easily estimated form the sample. If $\underline{\hat{\beta}}$ is the estimate of the model parameters then we estimate

$$\hat{V}_{asym} = \frac{1}{n}\sum_{i=1}^{n} f\left(y_i, \underline{x}_i, \underline{\hat{\beta}}, \underline{0}\right) f^t\left(y_i, \underline{x}_i, \underline{\hat{\beta}}, \underline{0}\right) \tag{59}$$

$$\hat{J}_{asym} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial f\left(y_i, \underline{x}_i, \underline{\hat{\beta}}, \underline{0}\right)}{\partial\left(\underline{\beta}, \underline{0}\right)}. \tag{60}$$

Then the estimated asymptotic Variance-Covariance matrix is given by the $p \times p$ leading principle submatrix

$$V\hat{a}r\left(\underline{\hat{\beta}}\right)_{asym} = \left(\frac{1}{n}\hat{J}_{asym}^{-1}\hat{V}_{asym}\hat{J}_{asym}^{-1}\right)_{(1,2,\ldots,p)\times(1,2,\ldots,p)}. \tag{61}$$

Instead of using the asymptotic weights $\frac{1}{n}$, one can use the estimated weights $\underline{\hat{w}}$.

4.4 A Sandwich estimator of the Variance-Covariance matrix

For samples of finite size a sandwich estimator of the Variance-Covariance matrix is more accurate. The theoretical expression closely follows the expression of the asymptotic Variance-Covariance matrix in (58), however the sample estimate of the vector of the Lagrangian multipliers is used in

(45) and (52) instead of $\underline{0}$. Suppose $\underline{\hat{\beta}}$ is the estimate of the model parameters and $\underline{\hat{w}}$ is the estimate of the maximising weights. Suppose $\mathbf{x}$ is the $n \times p$ matrix of the sample. Let us denote

$$\hat{\mu}_i = \sum_{k=1}^{p} \hat{\beta}_k x_i^{(k)} \tag{62}$$

$$\Delta_{\eta'} = diag\left[\hat{w}_1 \eta'_{\mathcal{M}}(\hat{\mu}_1), \ldots, \hat{w}_n \eta'_{\mathcal{M}}(\hat{\mu}_n)\right] \tag{63}$$

$$\Delta^{\star}_{(y-\eta)} = diag\left[(\hat{w}_1)^2\{y_1 - \eta_{\mathcal{M}}(\hat{\mu}_1)\}^2, \ldots, (\hat{w}_n)^2\{y - \eta_{\mathcal{M}}(\hat{\mu}_n)\}^2\right] \tag{64}$$

$$\Delta_{(y-\eta)} = diag\left[\hat{w}_1\{y_1 - \eta_{\mathcal{M}}(\hat{\mu}_1)\}, \ldots, \hat{w}_n\{y - \eta_{\mathcal{M}}(\hat{\mu}_n)\}\right], \tag{65}$$

where as before $diag[\cdot]$ is a diagonal matrix with the arguments as the diagonal entries. Also let

$$\mathfrak{h}^t = \left[\hat{w}_1 \underline{h}\left(y_1, \underline{x}_1, \underline{\gamma}\right) \ldots, \hat{w}_n \underline{h}\left(y_n, \underline{x}_n, \underline{\gamma}\right)\right]. \tag{66}$$

We estimate the necessary matrices by

$$\hat{G}_{sand} = \mathbf{x}^t \Delta_{\eta'} \mathbf{x} \tag{67}$$

$$\hat{G}^{\star}_{sand} = \mathbf{x}^t \Delta^{\star}_{(y-\eta)} \mathbf{x} \tag{68}$$

$$\hat{T}_{sand} = \mathbf{x}^t \Delta_{(y-\eta)} \mathfrak{h} \tag{69}$$

$$\hat{H}_{sand} = \mathfrak{h}^t \mathfrak{h}. \tag{70}$$

Then the sandwich estimate of the asymptotic Variance-Covariance matrix of $\underline{\hat{\beta}}$ is given by

$$V\hat{a}r_{sand}\left(\hat{\beta}\right) = \hat{G}^{-1}_{sand}\left[\hat{G}^{\star}_{sand} - \hat{T}_{sand}\hat{H}^{-1}_{sand}\hat{T}^{t}_{sand}\right]\hat{G}^{-1}_{sand}. \tag{71}$$

This estimate equals the value of the inverse of the Hessian matrix calculated at the parameter estimates.

5 Application to demographic hazard modeling

In the section we apply the methodology to combine survey data from the British Household Panel Survey (BHPS)(Taylor et al., 1995) with population level information from the birth registration system on the General Fertility Rate (GFR) to estimate annual birth probabilities by parity. This situation was considered by Handcock, Huovilainen, and Rendall (2000) using a constrained maximum likelihood framework.

The birth registration system in England and Wales collect birth data by parity only within marriage (hence we do not know the parity information for unmarried women). However we will show the system can be used to improve the efficiency of estimation, from the BHPS, of annual birth probabilities of all women by parity. From the registry we can determine the *general fertility rate* (GFR) for the years 1992 to 1996, of England and Wales(Office of National Statistics, 1998). The population level value of the GFR was found to be 0.06179(Handcock et al., 2000).

Our survey data are from the BHPS between the years 1991 to 1996 and we have excluded the women living in Scotland for consistency with the registration system. Births were not directly recorded in this dataset. Instead we code births for women aged 15 to 44 when the women's coresident family unit experiences an increase in the number of dependent-aged children from one year $(t-1)$ to the next (t). The details of the construction may be found in Handcock, Huovilainen, and Rendall (2000). The BHPS data uses an approximately equal-probability sample design so that we will assume that the weights of the samples are all equal. This simplifies the estimation and improves the standard error of the parameter estimates.

In the dataset the dependent variable Y represents the indicator of birth for a woman between times $(t-1)$ and t and the explanatory variable X is the indicator of existence of at least one child of that woman at time $(t-1)$.

We fit a logistic regression model to the data and constrain on the population information given by the GFR. The model is represented as

$$\log\left(\frac{Pr\,(Y=1|X=x)}{1-Pr\,(Y=1|X=x)}\right)=\beta_0+\beta_1 x. \tag{72}$$

We also assume that the marginal distribution of X does not involve β_0 or β_1.

The two score constraints corresponding to two model parameters are given by

$$\sum_{i=1}^{n} w_i \left\{y_i-\frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}}\right\}=\sum_{i=1}^{n} w_i \frac{y_i(1+e^{(\beta_0+\beta_1 x_i)})-e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}}=0 \tag{73}$$

$$\sum_{i=1}^{n} w_i x_i \left\{y_i-\frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}}\right\}=\sum_{i=1}^{n} w_i x_i \frac{y_i(1+e^{(\beta_0+\beta_1 x_i)})-e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}}=0. \tag{74}$$

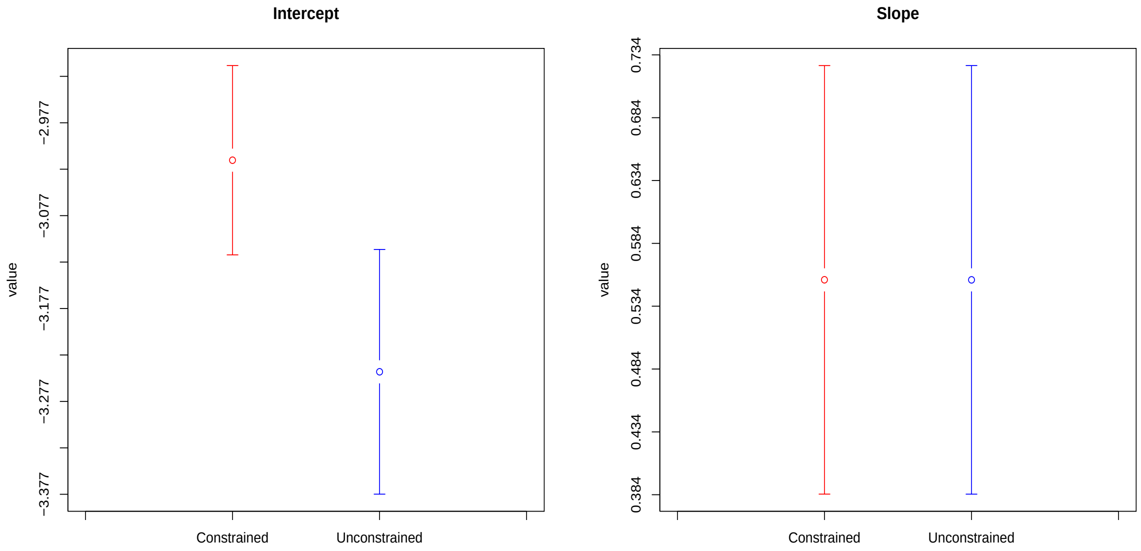
In the population we know that

$$GFR=E\,[Y]=0.06179. \tag{75}$$

This implies the the population restriction (assuming the restriction that $\sum_{i=1}^{n} w_i=1$) is given by

$$\sum_{i=1}^{n} w_i\,(y_i-0.06179)=0. \tag{76}$$

Thus from the discussion in Section 2, in order to determine the weights w_i, $i=1,2,\ldots,n$ we maximise $\sum_{i=1}^{n}\log(nw_i)$ subject to $w_i\geq 0$, for all $i=1,2,\ldots,n$, $\sum_{i=1}^{n} w_i=1$ and the restrictions in (73), (74), (76).



(a) Comparing $\hat{\beta}_0$.

(b) Comparing $\hat{\beta}_1$.

Figure 1: Comparison of the estimate and the standard errors of the estimator of the model parameters for constrained and unconstrained cases.

The estimation and standard errors of the parameter estimates was computed using an `R` package named `glmc` developed by Chaudhuri and Handcock (Chaudhuri et al., 2005). This package will be made available on CRAN. This package performs the two-step maximisation procedure (19) and (20) and if it fails the nested maximisation (18) as described in Section 3. The standard error of the parameter estimates are calculated using Theorem 1 from (71).

The estimated value and the standard errors of the estimator of the models parameters are shown in Figure 1. The figure also compares the estimates with and without population level constraints. We note that the estimate of the slope (i.e. β_1) is same in both constrained and unconstrained cases. The estimated value of the parameter is 0.55496 with a standard error of 0.08700. However the estimate of intercept (i.e. β_0) improves with the imposition of the population level constraints. Without the constraints the estimated value of β_0 is -3.24514 with a standard error of 0.06721. However with the constraints the estimated value changes to -3.01731 and the standard error reduces to 0.05199.

We note that our estimates differ from the corresponding estimates in Handcock, Huovilainen, and Rendall (2000). This can be ascribed to the fact that our estimates of the standard errors are asymptotically true, which may need a larger sample. However the constrained estimates reported here exactly matches the estimates obtained from putting constraints directly on the model parameters as described in Handcock, Rendall, and Cheadle (2003). In fact as expected in most of the cases these estimator gives the same results.

Discussion

In this paper we introduce a method to combine population level information with the sampled individual level information based using empirical likelihood. On one hand our approach can be seen as an extension of the post-stratification using a special case of empirical likelihood (Qin and Lawless, 1994; Imbens et al., 1998). On the other hand our method extends the weighted least squares estimator of the model parameters studied in Hellerstein and Imbens (1999). The method solves a two nested maximisation problem where in the outer step we maximise for the parameter estimates and in the inner step sample weights are computed such that the population expectation of the dependent variable given a subset of the explanatory variables and the restrictions imposed on the conditional expectation by the model are reproduced in the re-weighted sample. We further show that it is possible to simplify the procedure to a two-step estimator similar to Hellerstein and Imbens (1999). In the first step we find the weights satisfying only the population constraints. In the second, unconstrained estimation of the model parameters is conducted using the re-weighted sample from the first step. The asymptotic standard errors of the parameter estimates can be computed explicitly. The parameter estimates obtained from our method shares the asymptotic properties of the constrained maximum likelihood estimator.

In our method we do not place constraints on the parameters and can avoid non-linear constraints. This ensures improvement in terms of computational and implementational complexity. More importantly as the sample weights can be interpreted as an estimate of the joint probability of observing a particular sample, we don't need to specify a distribution of the explanatory variables.

Other than a more complete description using the empirical likelihood our method improves upon Hellerstein and Imbens (1999) in several ways. First we can specify several models of association between the dependent and explanatory variables. Our formulation allows for Bayesian extensions in several ways. We can allow for errors in the constraint values. Thus one can include expert opinions, information from a larger sample or some prior information about the constraints. On the other hand available prior information can be easily used in estimating the model parameters.

References

Chaudhuri, S., M. S. Handcock, and M. Rendall (2005). An empirical likelihood based approach to incorporate population information in sample based inference. Working paper, Center for Statistics in Social Sciences, University of Washington.

Deming, W. E. and F. F. Stephan (1942). On the least squares adjustment of a sampled frequency table when the expected marginal tables are known. *The Annals of mathematical Statistics 11*, 427–444.

Handcock, M. S., S. M. Huovilainen, and M. S. Rendall (2000). Combining registration-system and survey data to estimate birth probabilities. *Demography 37*(2), 187–192.

Handcock, M. S., M. S. Rendall, and J. E. Cheadle (2003). Improved regression estimation of a multivariate relationshipwith population data on the bivariate relationship. Working paper 36, Center for Statistics in Social Sciences, University of Washington.

Hellerstein, J. and G. W. Imbens (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics LXXXI*(1), 1–14.

Holt, D. and T. F. M. Smith (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A. 142*, 33–46.

Imbens, G. W. and T. Lancaster (1994). Combining micro and macro data in microeconomic models. *Review of Economic Studies 61*, 655–380.

Imbens, G. W., R. H. Spady, and P. Johnson (March, 1998). Information theoretic approches to inference in moment condition models. *Econometrica 66*(2), 333–357.

Ireland, C. T. and S. Kullback (1968). Contingency tables with given marginals. *Biometrika 55*, 179–188.

Little, R. J. A. (1993). Post-startification: Amodeler's perspective. *Journal of the American Statistical Association 88*(423), 1001–1012.

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

Little, R. J. A. and M. M. Wu (1991). Models for contingency tables with known margins when target and sampled population differ. *Journal of the American Statistical Association 86*(413), 87–95.

McCullagh, P. and J. Nelder (1989). *Generalised Linear Models*. Chapman& Hall/CRC.

Moffitt, R. and G. E. Ridder (2005). The econometrics of data combination. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. North-Holland, Amsterdam.

Office of National Statistics (1998). *1997 Birth Statistics*. London: Her Majesty's Stationery Office.

Owen, A. (2001). *Empirical Likelihood*. Chapman& Hall/CRC.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*, 300–325.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Willey & Sons.

Taylor, M. F., J. Bryce, N. Buck, and E. Prentice (1995). *British Household Panel Survey User Manual*. Colchester: University of Essex.