

# A 2-step Empirical Likelihood approach for combining sample and population data in regression estimation

Chaudhuri, Sanjay

*National University of Singapore, Department of Statistics and Applied Probability*

*6 Science Drive 2,*

*Singapore 117546, Singapore*

*E-mail: sanjay@stat.nus.edu.sg*

Handcock, Mark S.

*University of Washington, Department of Statistics*

*B313 Padelford Hall, Box 354322*

*Seattle, WA 98105, USA*

*E-mail: handcock@stat.washington.edu*

Rendall, Michael S.

*Rand Corporation, Population research Center*

*1776 Main Street, Box 2138*

*Santa Monica, CA 90407, USA*

*E-mail: mrendall@rand.org*

## 1 Introduction

In most demographic applications apart from the samples, some information of the relationship of explanatory variable with the dependent variables are known from the population level data, obtained as for example from census, registration system, etc. However, with the increasing popularity of likelihood based methods, over the years more and more emphasis has been placed on estimation from sample data alone, with the information from complete enumeration procedures ignored in this estimation. Clearly the population level information contains valuable information for statistical analysis whose inclusion can lead to statistically more accurate estimates and better inference.

The use of both sample and population information within the framework of likelihood principles and more specifically for *generalized linear models* has been proposed by Handcock, Huovilainen, and Rendall (2000). 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 *maximum likelihood estimates* can then be obtained by maximizing the likelihood function under the population level constraints. The method can be implemented using any of the widely available procedures for numerical optimization with equality constraints. It also known that the estimates are asymptotically normal, unbiased. An explicit form of the Hessian at the parameter estimates can also be thus 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. However, non-linear equality constraints are in general numerically difficult to compute and the constraints time-consuming to code in applications involving multiple population level restrictions and even moderate numbers of regressors.

In this note we introduce an alternative method to combine the population level information with the sample observations based on the method empirical likelihood. This is a 2-step method similar to that of Hellerstein and Imbens (1999), who conducted a 2-step GMM estimation. In the first step, new sample weights are computed such that the population expectation of the dependent variable given a subset of the explanatory variables is reproduced in the re-weighted sample. This population

expectation is known from population-level data. In the second step, unconstrained estimation is conducted using the weights from the first step. The asymptotic standard errors can be computed from their analytic expressions.

## 2 Application and Data

The relationship of first childbearing to age and other covariates is modeled using a combination of sample data from the Panel Study of Income Dynamics (PSID, Hill 1992) and population data on first childbearing by age only from NCHS, compiled by Schoen (2003). The sample data are person-years of exposure to first childbearing from 1986 to 1997 at ages 17 to 30, with additional covariates for marital status, marital duration, and standard socio-economic variables available in panel data such as race/ethnicity, education, and earnings. The population data are age-specific first-birth probabilities. The shape of the first-birth probability function with age is complex, attaining an early peak at age 20 and second, higher peak at age 28 (see Figure 1(a) below). Such a function is difficult to model with sample data only, as it is not easily amenable to a parametric specification as, for example, using a polynomial representation of birth by age. The sample data, moreover, are too sparse for a statistically reliable non-parametric estimation of the birth-by-age function. The application is therefore one for which the population information may be especially useful in deriving an estimated relationship of births by age and other covariates.

## 3 Methodology

Suppose  $Y$  is the indicator of the birth,  $I^m$  is the indicator of marriage,  $I^j$  is the indicator for the  $j$ th age,  $j = 17, 18, \dots, 30$ . Also let  $D$  denote the marital duration. We ignore other covariates in this exposition. We fit a logistic regression model with canonical link to the data, specified as

$$(1) \quad \log \text{it}(y_i = 1 | X = x) = \beta_0 + \beta_m I^m + \sum_{j=18}^{30} \beta_j I^j + \beta_D D + \beta_{D^2} D^2.$$

We use *empirical likelihood* based methodology to incorporate the age-specific first birth rates obtained from the population level information in fitting a weighted logistic regression to our data.

In the first step, we derive sample weights that result in age-specific first-birth probabilities from the weighted sample matching those of the population probabilities. We find the weights  $w_i$ ,  $i = 1, 2, \dots, n$  for each sample point such that  $w_i \geq 0$ ,  $i = 1, 2, \dots, n$ ,  $\sum_{i=1}^n w_i = 1$ , which maximize

$$(2) \quad \prod_{i=1}^n n w_i \equiv \log \left( \prod_{i=1}^n n w_i \right) \equiv \sum_{i=1}^n \log(n w_i)$$

The calculation of the maximizing weights depends on the (logistic) functional form (1) between  $Y$  and the explanatory variables and on the population level information through the additional restrictions imposed on the optimization problem. Without these extra constraints (2) is maximized for  $w_i = n^{-1}$ , for all  $i = 1, 2, \dots, n$ .

The restrictions imposed by the model in (1) are through its score functions, which are known to have zero expectation.

Suppose  $(y_i, I_i^m, I_i^j, j = 17, \dots, 30, d_i)$  be the observations in the sample,  $i = 1, 2, \dots, n$ . By defining

$$(3) \quad \vartheta_i = \beta_0 + \beta_m I_i^m + \sum_{j=18}^{30} \beta_j I_i^j + \beta_D d_i + \beta_{D^2} d_i^2$$

the constraints on the weights  $w_i$  through the score is given by

$$(4) \quad \sum_{i=1}^n w_i \{y_i - \mu(\vartheta_i)\} = 0.$$

$$(5) \quad \sum_{i=1}^n w_i I_i^j \{y_i - \mu(\vartheta_i)\} = 0, \quad j = m, 18, \dots, 30,$$

$$(6) \quad \sum_{i=1}^n w_i d_i^k \{y_i - \mu(\vartheta_i)\} = 0, \quad k = 1, 2$$

where  $\mu$  is the mean function of the logistic model.

Suppose for  $j = 17, 18, \dots, 30$ ,  $\phi_j$  denote the age-specific first birth rates. Then assuming that  $\phi_j$  was obtained without any sampling error we can assume that for  $j = 17, 18, \dots, 30$

$$(7) \quad E(I^j Y) = \phi_j.$$

From (7) the constraints imposed on the weights by the population level information are given by

$$(8) \quad \frac{\sum_{i=1}^n w_i I_i^j y_i}{\sum_{i=1}^n w_i I_i^j} = \phi_j, \quad j = 17, 18, \dots, 30,$$

which for each  $j$  may be simplified to

$$(9) \quad \sum_{i=1}^n w_i I_i^j (y_i - \phi_j) = 0.$$

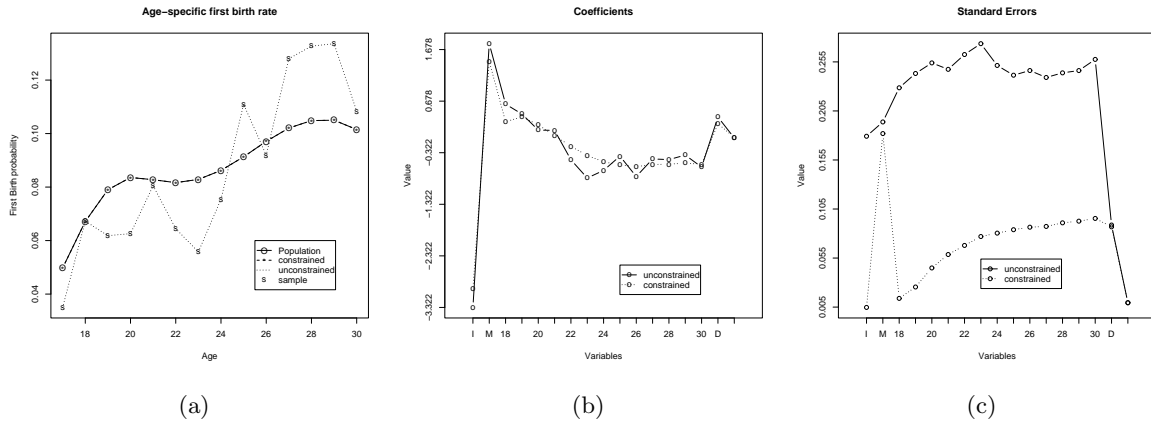
In most of the cases the model parameters can be estimated fairly easily through a two step estimation procedure. In the first step of the procedure we maximize (2) for the non negative weights  $w'_i$ ,  $i = 1, 2, \dots, n$  s.t.  $\sum_{i=1}^n w'_i = 1$  and the linear equality constraints in (9) are satisfied. The second step is to compute the vector of parameter estimates  $\hat{\beta}$ . They can be readily obtained from a weighted logistic regression with  $w'_i$  as weight for the  $i$ th observation for  $i = 1, 2, \dots, n$ . The first step only involves maximizing (2) with linear equality constraints. This has been studied in details by Owen (2001) and Chen, Sitter and Wu (2002). Instead of solving the primal problem both solve the dual one. This by itself reduces the constraints to linear inequality constraints. Moreover by defining a convex pseudo-logarithmic function (Owen, 2001) over the whole space, one can maximize (2) unconstrained.

The alternative and more general procedure is to get the parameter estimates through a nested two step maximization, where the outer maximization is unconstrained and the inner one is constrained by linear constraints given in (9). The analytic expression of the standard errors are also known. The two step as well as the nested estimation has been implemented in an *R* package named *glmc* which is available on CRAN.

## 4 Results

The performance of the constrained logistic regression model as described above can be compared with the unconstrained logistic regression using the sample data only. The plot in figure 1(a) compares the age-specific first birth rates predicted by the constrained and the unconstrained regression procedure with the known population values of the same. It is clear that the age-specific first birth rates predicted by the constrained model are exactly equal to the population values. Thus these predicted first birth rates drop slightly for age 21, 22 and 23, which is also seen in the population. On the other hand

**Figure 1.**



(a) The population, constrained, unconstrained and sample age specific first birth rate, (b) Constrained and unconstrained estimated model coefficients, (c) their standard errors.

the values of the age-specific first birth rates predicted by the unconstrained model are same as the age-wise proportion of first births observed in the sample. Thus they vary unsystematically and are not close to the population values.

The main advantage of putting the constraints based on the population level information is in the reduction of standard error of the model parameters. This in turn allows for a non-parametric function of age (indicator variables for each single-year age) to be used in the estimation.

The standard errors are much lower for all variables in the constrained model (see Figure 1(c)) than the unconstrained model. The reduction is substantial for the intercept (related to the indicator for age 17) and age indicators. The standard errors about the other, unconstrained covariates (marital status and marital duration) are slightly below those of the unconstrained model. A theoretical explanation for this phenomenon is given in Handcock et al (2003).

## References

- Chaudhuri, S., M. S. Handcock, and M. S. Rendall (2005) Generalised Linear Models Incorporating Population Level Information: An Empirical Likelihood Based Approach. Working Paper no. 48, Center for Statistics in Social sciences, University of Washington, Seattle.
- Chaudhuri, S., M. S. Handcock, and M. S. Rendall (2006) glmc: An R package for generalized linear models subject to constraints.
- Chen, J., R. R. Sitter and C. Wu (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* 89(1), 230-237.
- Handcock, Mark S., Sami M. Huovilainen, and Michael S. Rendall (2000) Combining registration-system and survey data to estimate birth probabilities. *Demography* 37(2):187-192.
- Handcock, Mark S., Michael S. Rendall, and Jacob E. Cheadle (2003) "Improved regression estimation of a multivariate relationship with population data on the bivariate relationship" *Sociological Methodology* 35(1), 291-334.
- Hellerstein, J., and G.W. Imbens (1999) Imposing moment restrictions from auxiliary data by weighting *Review of Economics and Statistics* 81(1):1-14.
- Hill, M.S. 1992. *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, California: Sage.
- Owen, A.B. (2001) *Empirical Likelihood*. Chapman & Hall/CRC.
- Schoen, R. (2003) "Insights from Parity Status Life Tables for the 20th Century U.S." *Population Research Institute Working Paper 03-04*, Penn State University.