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

Sanjay Chaudhuri, Mark S. Handcock, and Michael S. Rendall

Introduction

In most demographic applications apart form 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 Hessian and standard errors derived therefrom are separately computed.

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 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.

Methodology

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

$$(1.1) \qquad \log it(y_i = 1 \mid X = x) = \beta_0 + \beta_m I^m + \sum_{j=18}^{30} \beta_j I^j + \beta_D D + \beta_{D^2} D^2 .$$

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, $i = 1, 2, ..., n$ for each sample point such that $w_i \geq 0$, $i = 1, 2, ..., n$, $\sum_{i=1}^{n} w_i = 1$, which maximize

$$(1.2) \qquad \prod_{i=1}^{n} nw_i \equiv \log\left(\prod_{i=1}^{n} nw_i\right) \equiv \sum_{i=1}^{n} \log(nw_i)$$

The calculation of the maximizing weights depends on the (logistic) functional form (1.1) 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 (1.2) is maximized for $w_i = n^{-1}$, for all $i = 1, 2, ..., n$.

The restrictions imposed by the model in (1.1) are through its score functions, which are known to have zero expectation.

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

$$\vartheta_i = \beta_0 + \beta_m I_i^m + \sum_{j=18}^{30} \beta_j I_i^j + \beta_D d_i + \beta_{D^2} d_i^2 \tag{1.3}$$

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

$$\sum_{i=1}^{n} w_i \{ y_i - \eta(\vartheta_i) \} = 0 \,. \tag{1.4}$$

$$\sum_{i=1}^{n} w_i I_i^j \left\{ y_i - \eta\left(\vartheta_i\right) \right\} = 0, \qquad j = m, 18, \ldots, 30 \,, \tag{1.5}$$

$$\sum_{i=1}^{n} w_i d_i \left\{ y_i - \eta\left(\vartheta_i\right) \right\} = 0, \tag{1.6}$$

$$\sum_{i=1}^{n} w_i d_i^2 \left\{ y_i - \eta\left(\vartheta_i\right) \right\} = 0 \,. \tag{1.7}$$

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

$$\mathrm{E}\left(I^j Y \right) = \phi_j \,. \tag{1.8}$$

From (1.8) the constraints imposed on the weights by the population level information are given by

$$\frac{\sum_{i=1}^{n} w_i I_i^j y_i}{\sum_{i=1}^{n} w_i I_i^j} = \phi_j \,, \quad j = 17, 18, \ldots, 30 \,, \tag{1.9}$$

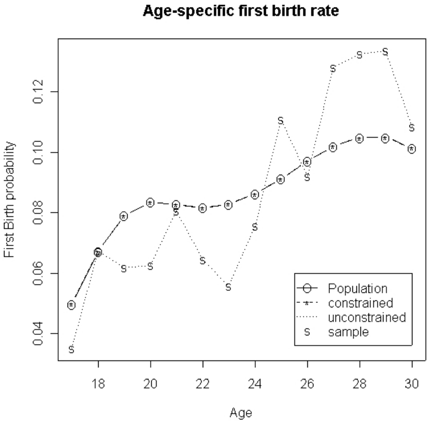
which may be simplified to

$$\sum_{i=1}^{n} w_i I_i^j \left(y_i - \phi_j \right) = 0 \,. \tag{1.10}$$

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 (1.2) for the non negative weights w_i', $i = 1, 2, \ldots, n$ s.t. $\sum_{i=1}^{n} w_i' = 1$ and the equality constraints in (1.10) are satisfied. The second step is to compute the vector of parameter estimates $\underline{\hat{\beta}}$. They can be readily obtained from a weighted logistic regression with w_i' as weight for the i th observation for $i = 1, 2, \ldots, n$. The first step only involves maximizing (1.2) with linear equality constraints. This has been studied in details by Owen (2001). Instead of solving the primal problem one solves the dual one. This by itself reduces the constraints to linear equality constraints. Moreover by defining a convex pseudo-logarithmic function over the whole space, one can maximize (1.2) unconstrained.

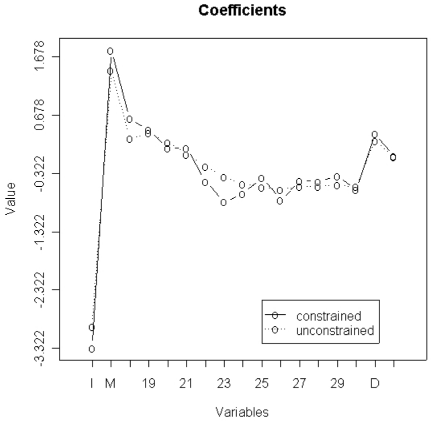
The alternative and more general procedure is to get the parameter estimates through a nested two step maximization, where the outer maximization is unconstrained and the inner one is constrained by linear constraints given in (1.10). The standard errors are obtained from the Hessian matrix calculated at the value of the parameter estimates.

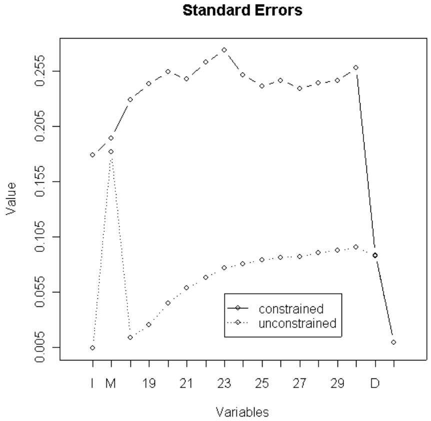
Results



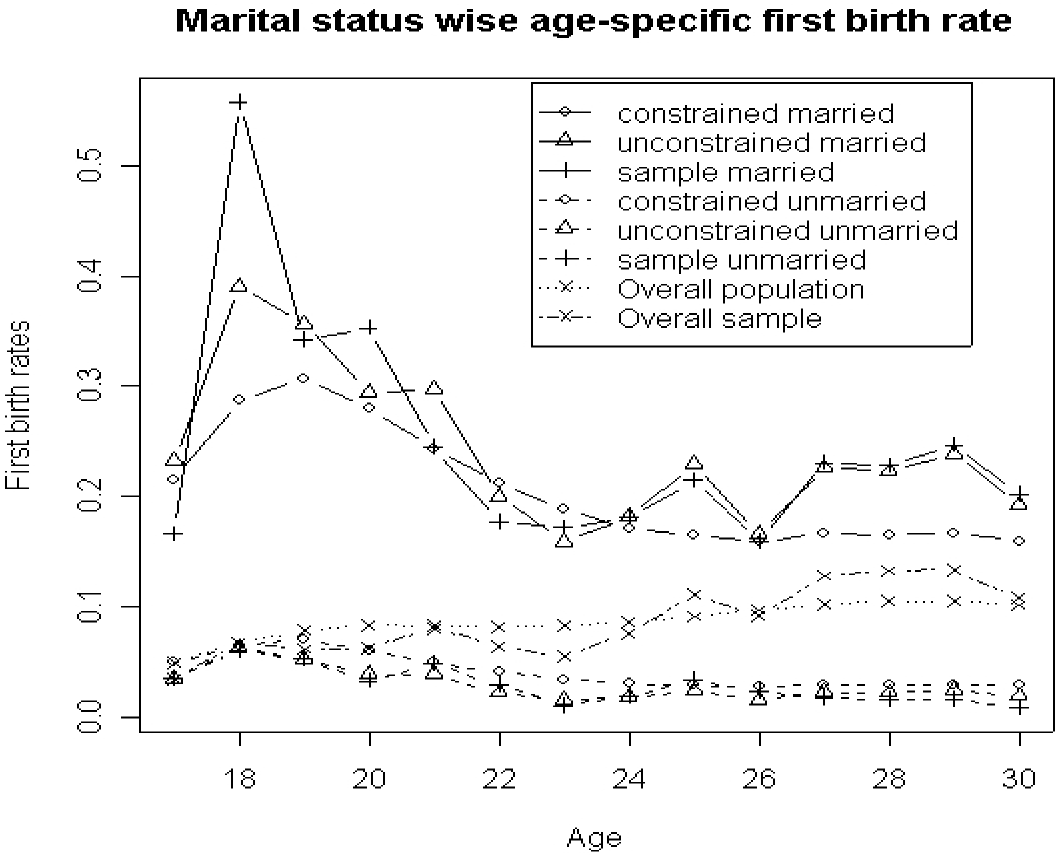
The performance of the constrained logistic regression model as described above can be compared with the unconstrained logistic regression using the sample data only. The plot in figure to the left compares the age-specific first birth rates predicted by the constrained and the unconstrained regression 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 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 right-hand plot) 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).



In the figure above, the first-birth probabilities for married versus unmarried women are plotted. The two lines with diamonds give our best estimates, using all the information available in both sample and population data. The weighted sum of the two lines always sums to the overall population line (with crosses joined by a dotted line). At the youngest ages, very few women are married, but those who are have a much higher first-birth probability. In this case, first birth (or conception) may be causing women to marry. Most early first births, however, are non-marital, and the predicted non-marital first-birth probability line is at these ages very close to the overall population constraint line. As age increases, more women are married, and the age-specific marital first-birth probability becomes increasingly close to the overall population first-birth probability, and the age-specific non-marital first-birth probability increasingly further from it. The constrained marital and non-marital first-birth probabilities appear to fit well the sample pattern (see the lines with the plus signs), while smoothing the variability shown in the sample especially in the case of marital first births.

References

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" CSSS Working Paper, University of Washington, Seattle.

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.