Rendall, M.S., Admiraal, R., De Rose, A., Di Giulio, P., Handcock, M.S. and Racioppi, F. (2006) Estimating first birth probabilities in Italy by combining population and survey data. In: Proceedings of the 43rd Scientific Meeting of the Italian Statistical Association, 14 - 16 June, Turin, Italy, pp 151 - 154

Estimating first birth probabilities in Italy by combining population and survey data(*)

Stima delle probabilità di prima nascita in Italia integrando dati di indagini campionarie con statistiche correnti

Michael S. Rendall,
RAND Corporation

Ryan Admiraal,
University of Washington

Alessandra De Rose,
Univ. of Rome "La Sapienza"
alessandra.derose@uniroma1.it

Paola Di Giulio,
Max Planck Demographic Research Institute, Rostock

Mark S. Handcock,
University of Washington

Filomena Racioppi
Univ. of Rome "La Sapienza"

Riassunto: In questo lavoro si propone un uso integrato di dati provenienti da due indagini campionarie (FFS Italia 1995/96 e Indagine Multiscopo 1998) e delle informazioni sulla fecondità fornite dalle statistiche correnti. Si dimostra che l'errore di stima delle probabilità di prima nascita si riduce sostanzialmente introducendo il dato di popolazione come vincolo nei modelli di regressione logistica applicati per stimare l'effetto di età, generazione ed istruzione, indipendentemente dalla numerosità campionaria, e che un'ulteriore riduzione dell'errore si ottiene combinando le due indagini campionarie.

Keywords: data sources integration, birth probabilities, constrained maximum likelihood estimator

1. Introduction

Statistical methods for using population information to improve sample-survey- based estimates have a long history of development in statistics (Deming and Stephan 1942, Ireland and Kullback 1968). More recently, they have been applied to economic data (Imbens and Lancaster 1994). The availability of population counts of both demographic events in registration-data system and of cross-sectional population sizes in population censuses may allow for even greater scope for efficiency gains to be realized also in demographic applications.
Use of population data alone limits the amount of socio-economic information that can be incorporated into the analysis. Large scale survey data are also increasingly considered undesirable, either for their lack of a longitudinal dimension or for their lack of certain variables needed for specific applications, and an increasing reliance of small specialized survey data alone has been seen. Small survey data, however, have major disadvantages with respect to statistical precision.

(*) The authors are grateful for financial support from grants from the National Institute of Child Health and Human Development (R01-HD34484-01A1, R01-HD043472-01, and R24 HD41025), and to Piero Giorgi for providing us with population first-birth probabilities by age and cohort.

These are the concerns that have led to the development of methods for combining population and large-scale data with small-sample survey data. In previous work (Handcock, Houvilainen, and Rendall 2000), we introduced and implemented a constrained maximum likelihood estimator (MLE) for demographic applications. In Handcock, Rendall, and Cheadle 2005), we extended this to the case of multiple population constraints.

The present study applies the same estimation procedure and statistical software in a realistic demographic application to see how age at first birth has changed over time for Italian women with and without higher educational qualification. It also introduces a further statistical innovation by pooling observations from a second large-scale survey dataset with observations from the specialist demographic survey dataset. We thus evaluate the gains from combining data between survey samples and between survey and population collections.

2. Data and Method

Italy has two survey datasets that collected women's fertility histories in the 1990s: the smaller, 1995-96 Italian Fertility and Family Survey ("FFS", De Sandre *et al*. 2000); and the larger, 1998 Italian Multipurpose Survey ("Multiscopo", ISTAT 2000). Both are approximately equal probability samples. Differential probabilities of selection are adjusted for throughout our analyses by using the sample weights provided for each dataset.

The FFS includes approximately 4,800 female sample members aged 20 to 49 at survey date, and their fertility and family histories. The Multiscopo includes more than 20,000 households with approximately 54,000 individuals. A fertility history was collected for female sample members aged between 25 and 49 years old. From both the surveys, we use data from female respondents born in the years 1951-55 and 1961-65. We used the variable "year of first birth" to assemble the data into person-years of exposure to first birth from age 25 and above. Both surveys include a question on highest educational qualification obtained, from which we coded "higher education" for women with any tertiary education qualification.

For the entire period of our analyses, the Italian birth registration system collected details including age of mother and how many children the mother has previously given birth to. Using these data, Giorgi (1993) calculated first birth probabilities by single-year cohort. We use these calculations as our population-level estimates of first-birth probabilities by single-year age.

Estimation of the probability of first birth by age, education, and birth cohort is by logistic regression.

Let X be a vector of regressors, and θ be a vector consisting of an intercept β_0 plus a vector of coefficients β_1 for each of the regressors. The binomial logit model for the first birth probability $P(Y=1 \mid X=x)$ is then:

$$P(y)= 1/\{1+\exp(-\theta'x)\} \qquad (1)$$

This is the standard logistic regression model. Here, we refer to it as "unconstrained" model. Denote the survey data by $D=(y_i, x_i)$, $i=1,\ldots,n$. Under standard regularity conditions, the value of θ that maximizes the likelihood:

$$L(\theta; y, x) = \prod_{i=1}^{n} P(Y = y_i, X = x_i \mid \theta) = \prod_{i=1}^{n} P(Y = y_i \mid X = x_i, \theta) P(X = x_i) \qquad (2)$$

is an asymptotically efficient estimator of θ_0. Under these conditions, the estimator is also asymptotically unbiased and Gaussian.

To introduce the "constrained" model, let the proportion of women with a higher education qualification at each age *a* and birth cohort *c* be given by $\pi(a,c)$. For each age and cohort, the probability of a first birth $P(a,c)$ can be specified as the weighted sum of the probability of a first birth for a woman with higher education qualification $P(a,c,1)$ and the probability of a first birth for a woman with no higher education qualification $P(a,c,0)$. For a given set of constants $\{\pi(a,c)\}$ the constrained function depends on regression parameters θ and so may be expressed as $C(\theta)$:

$$C(\theta)=P(a,c)=P(a,c,1)\pi(a,c) + P(a,c,0)[1-\pi(a,c)] \qquad (3)$$

The values $P(a,c)$ are known from population data, as described above. The constrained MLE solves equation (1) subject to constrained functions (3). If we maximize the above likelihood subject to this constraint, the estimator is still asymptotically efficient, unbiased and Gaussian. It may be demonstrated that the inclusion of the population information leads to an improvement in the estimation of θ_0. In particular, the standard error of the estimator will always be less than the one that ignores the constraints. A further result is that the asymptotic ratio of variances of constrained and unconstrained parameters is independent of the survey sample size. This means that the additional efficiency gains realized by pooling survey samples in unconstrained estimation will be preserved in the constrained estimation.

3. Results

Our formal statistical comparisons are limited to the regression parameter estimates between constrained and unconstrained version of the same specification. In Table 1 parameter estimates and standard errors are presented for the small survey (FFS) only, the large survey (Multiscopo) only, and the two surveys with their observations pooled.

Consistent with the statistical theory presented above, all standard errors in the constrained version are as low or lower than the corresponding standard errors of the unconstrained version. The standard errors of the age parameters are in the range of 80 to 90 percent less in the constrained version than in the unconstrained version. The standard errors of the education parameter are reduced by negligible amounts, as no direct constraint is placed on it by the population data (see also Handcock *et al*. cit.).

As expected, the percentage reduction in the standard errors is seen to be approximately the same for the FFS and the Multiscopo, and the standard errors for the pooled sample are reduced by similar amounts in percentage terms than are the standard errors for either of the two surveys alone. Pooling the two surveys results in substantial reductions in the standard errors of the parameters for education.

Table 1: *Constrained versus unconstrained estimates of first birth probabilities in Italy*

CONSTRAINED MODEL						
	FFS		*Multiscopo*		*FFS + Multiscopo*	
Covariate	*parameter*	*Std. error*	*parameter*	*Std. error*	*parameter*	*Std. error*
Intercept	-1.661**	0.028	-1.676**	0.017	-1.674**	0.014
High Education	-1.203**	0.337	-1.177**	0.213	-1.162**	0.178
Age (ref.25)						
26	-0.016	0.011	-0.023**	0.004	-0.021**	0.004
27	-0.098**	0.024	-0.091**	0.010	-0.092**	0.010
..						
Cohort 1961-65	-0.598**	0.039	-0.560**	0.017	-0.566**	0.016
UNCONSTRAINED MODEL						
	FFS		*Multiscopo*		*FFS + Multiscopo*	
Covariate	*parameter*	*Std. error*	*parameter*	*Std. error*	*parameter*	*Std. error*
Intercept	-1.282**	0.139	-1.868**	0.097	-1.700**	0.079
High Education	-1.215**	0.338	-1.180**	0.213	-1.164**	0.178
Age (ref.25)						
26	-0.184	0.209	0.344**	0.131	0.187	0.110
27	-0.258	0.222	0.148	0.141	0.023	0.118
..						
Cohort 1961-65	-0.718**	0.199	-0.433**	0.132	-0.519**	0.110

** significant at p-value <0.01

References

Deming W.E., Stephan F.F. (1942) On the last squares adjustament of a sampled frequency table when the expected marginal tables are known, *The Annals of Mathematical Statistics*, 11, 417-424.

De Sandre P., Ongaro F., Rettaroli R., Salvini S. (2000) *Fertility and Family Surveys in Countries of the ECE Regions. Standard Country Report : Italy*, New York-Geneve, UNECE.

Giorgi P. (1993) Una rilettura della fecondità del momento per ordine di nascita in Italia nel periodo 1950-90 considerando la struttura per parità, *Genus*, 40(3-4), 177-204.

Handcock M.S., Houvilainen S.M., Rendall M.S. (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 (2005) Improved regression estimation of a multivariate relationship with population data on the bivariate relationship, *Sociological Methodology,* 35(1), 291-334.

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

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

ISTAT (2000) *Indagine Statistica Multiscopo sulla Famiglia 1998*, Roma.