Stat Meth Appl (2008) 17:519–539
DOI 10.1007/s10260-008-0106-8

ORIGINAL ARTICLE

Population constraints on pooled surveys in demographic hazard modeling

Michael S. Rendall · Ryan Admiraal · Alessandra DeRose · Paola DiGiulio · Mark S. Handcock · Filomena Racioppi

Accepted: 12 December 2007 / Published online: 1 August 2008

Abstract In non-experimental research, data on the same population process may be collected simultaneously by more than one instrument. For example, in the present application, two sample surveys and a population birth registration system all collect observations on first births by age and year, while the two surveys additionally collect information on women's education. To make maximum use of the three data sources, the survey data are pooled and the population data introduced as constraints in a logistic regression equation. Reductions in standard errors about the age and birth-cohort parameters of the regression equation in the order of three-quarters are obtained by introducing the population data as constraints. A halving of the standard errors about the education parameters is achieved by pooling observations from the larger survey dataset with those from the smaller survey. The percentage reduction in the

M. S. Rendall (✉)
RAND Population Research Center, 1776 Main Street, Santa Monica, CA 90407-2138, USA
e-mail: mrendall@rand.org

R. Admiraal · M. S. Handcock
Center for Statistics and the Social Sciences, University of Washington, Padelford Hall, Seattle, WA 98195-4320, USA

A. DeRose
Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale, University of Rome, 'La Sapienza', via del Castro Laurenziano 9, 00161 Rome, Italy

P. DiGiulio
Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1, 18057 Rostock, Germany

F. Racioppi
Dipartimento di Scienze Demografiche, University of Rome, 'La Sapienza', Via Nomentana 41, 00161 Rome, Italy

standard errors through imposing population constraints is independent of the total survey sample size.

Keywords Combining data · Constrained estimation · Fertility

1 Introduction

Statistical methods for using population information to increase the efficiency of 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 and demographic data (Imbens and Lancaster 1994; Handcock et al. 2000). In demographic applications, the availability of population counts of both vital events (in registration-system data) and of population characteristics (in population censuses and inter-censal estimates) increases the scope for realizing efficiency gains. Moreover, because prediction is frequently a goal in demography, efficiency gains may be especially beneficial.

The alternatives of using either population or survey data alone each have their disadvantages. Use of population data alone limits the amount of socio-economic information that can be incorporated into the analysis. Data from large-scale, general purpose surveys are also increasingly considered undesirable, either for their lack of a longitudinal dimension or for their lack of certain variables needed for specific applications. As a result, an increasing reliance on data from small, specialist surveys has been seen in demography. Small survey data, however, have major disadvantages with respect to statistical efficiency. They may also be subject to bias due to attrition and other forms of non-response. These are the concerns that have led to the development of methods for combining population or large-scale data with small-sample survey data in economics (Hellerstein and Imbens 1999; Ridder and Moffitt 2007).

In previous applications to fertility estimation, Handcock et al. (2000, 2005) introduced and implemented a constrained maximum likelihood estimator (MLE) in a logistic regression model. They demonstrated large efficiency gains first in estimating the intercept parameter by constraining survey estimates to an overall fertility rate (Handcock et al. 2000), and second in estimating coefficient parameters by constraining to the fertility rates of population subgroups (Handcock et al. 2005). In the first case, the reduction in the variance about the intercept parameter resulted in a 50% reduction in the variance about the predicted birth probabilities. In the second case, even larger reductions in standard errors about the parameter estimates for population subgroup coefficients were achieved. They referred to these coefficients as being "directly constrained" by the population data. Consistent with Imbens and Lancaster's (1994) simulation results, however, Handcock et al. (2005) found that no more than trivial gains in efficiency may be expected for regression parameters that are *not* directly constrained by population data.

The present study builds on those earlier studies by addressing the problem of how to improve efficiency of estimation of regression parameters that are not directly constrained by population data. It does so by pooling data across surveys while still

constraining to population data. In an application to first births by education in Italy, observations from a larger, general-purpose survey dataset (the 1998 Multiscopo survey) are pooled with observations from a smaller, specialist dataset (the 1995/1996 Fertility and Family Survey, or FFS). We consider only first childbearing after age 25 to focus the analysis on the process of entry to motherhood after completion of studies, and to illustrate the utility of population constraints for ages at which survey observations of women who have not yet given birth are relatively few. Women born in the early 1950s are compared to women born ten years later in the early 1960s, thus providing examples of estimation respectively for complete and censored hazards.

Even though the two surveys are conducted three years apart, their retrospective fertility histories overlap for all years up to the survey year of the FFS. This allows for the potential to realize gains in statistical efficiency by simply pooling sample observations across the two surveys. We first derive a basic theoretical result on the relationship between survey sample sizes and the variance-reducing effect of inclusion of population constraints: that the *proportionate* reduction in variance from the inclusion of population constraints is independent of the size of the survey sample. This implies that pooling observations across sample surveys will not alter the relative efficiency gains achieved through applying population constraints. This result is confirmed empirically by comparing the gains between unconstrained and constrained estimation when using the smaller survey dataset only with the gains when pooling the larger survey observations with those of the smaller survey.

The population data, however, directly constrain survey estimation only of the relationship of age and cohort to first birth. The relationship of education to first births is not directly constrained, and so no significant improvements in its estimation are achieved by adding population constraints. The pooling of surveys in a constrained MLE, however, achieves substantial increases in the efficiency of the estimates of the relationship of education to first births.

The remainder of the article is organized as follows. In Sect. 2, we first describe the sample survey data and evaluate the comparability of the two survey datasets against population data. We then describe the method of constrained MLE as applied to the problem of estimating first birth probabilities using these survey and population data sources. We also derive the main analytical result relevant to pooling the survey datasets with constrained estimation: that the proportionate reduction in variance of the regression parameter estimates is independent of the total sample size. In Sect. 3, we compare the results obtained under constrained MLE on the pooled survey datasets with results from estimators that either ignore the population data or that forego the opportunity to pool the survey data. Both non-parametric and parametric specifications of the relationship of first births to age are used in the alternatives that ignore the population data. Finally, Sect. 4 follows.

2 Data and method

In this section, we first describe the three separate data sources to be eventually combined in the estimation. We note here their comparability in terms of universes and topics

covered. Second, we compare the three data sources empirically. We note the comparability of their estimates of first birth probabilities by cohort across all three sources and of the estimates of the distribution of women by education and cohort across the two survey sources. Third, we describe a logistic regression model that may be estimated with survey data alone or with combinations of the survey and population data.

2.1 The sample and population data

Italy has two survey datasets that collected women's fertility histories in the 1990s: the smaller, 1995/1996 Italian Fertility and Family Survey ("FFS", De Sandre et al. 2000); and the larger, 1998 Italian Multipurpose Survey ("Multiscopo", ISTAT 2000). As its name implies, the FFS was designed explicitly for fertility analysis and for other subjects related to family formation and change. The FFS included approximately 4,800 female sample members aged 20–49 at survey date. From the fertility history asked of all sample members, we use here only the year of birth of a woman's first live-born child, if any, born up to the end of the year (1994) before the survey year. The FFS also recorded highest educational qualification obtained, coded to ISCED77 (International Standard Classification of Education, OECD 2003) categories. We coded "high education" for women with any tertiary education qualification (ISCED77 codes 5 and above).

The 1998 Multiscopo is a large, general purpose survey. Its sample included more than 20,000 households with approximately 54,000 individuals. A fertility history was collected for all female sample members aged 15 and over. We use here only the year of birth of a woman's first live-born child, if any, born up to the end of the year before the survey year (1997). The Multiscopo also included a question on highest educational qualification obtained, from which we were able to code "high education" in the same way as for the FFS.

From both the FFS and Multiscopo, we use data from female respondents born in the years 1951–1955 and 1961–1965, and create person-years of exposure to first birth from age 25 and above. We define age throughout the analysis using the "generation" definition of number of years attained this calendar year. On average this is half a year younger than the "age at last birthday" definition. The women born in the 1950s have only just completed their childbearing years by survey date, assuming age 44 to be the oldest age of childbearing. The FFS data, collected in 1995/1996, allow for exposure to childbearing only to age 42. The Multiscopo data, collected in 1998, allow for exposure up to age 44. For the 1960s cohort, the FFS data allow for exposure to childbearing to age 32. We use the Multiscopo data for exposure to childbearing up to age 34.

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 probabilities, subsequently updated by Giorgi to 1997, as our population-level estimates of first-birth probabilities by single-year age. We calculated the geometric mean of individual birth-year-specific probabilities to convert them into five-year birth-cohort averages.

2.2 Population representativeness of the two survey datasets

Handcock et al. (2005) showed that even when the sample survey data deviate from being exactly representative of the population for which the constraint data are obtained, the constrained MLE will improve estimation compared to using an unconstrained alternative estimator. Bias in the survey data in this case will also be reduced by incorporating the exact population constraints, but will not be eliminated (see also Hellerstein and Imbens 1999). We now show that in the present application, the two survey datasets sample in an approximately unbiased way from the same population, and therefore that the issue of estimation from non-representative survey data will not play a major role in the analysis.

Sample sizes and comparisons of the variables of interest between the sample and birth-registration data, and between sample and Labour Force Survey (LFS) estimates, are presented in Table 1. FFS and Multiscopo sample sizes are of female respondents born in the years 1951–1955 and 1961–1965, respectively. There are approximately three times as many women from both cohorts in the Multiscopo (2,100 and 2,690, respectively) as in the FFS (760 and 840, respectively). The three extra years of observation per woman in the Multiscopo as compared to the FFS raise the ratio to approximately four times as many person-years of observation in the Multiscopo (see below). For the LFS, we use published reports and special tabulations that are not accompanied by confidence intervals or sample sizes (ISTAT 1996, 2005), and therefore treat them as if they are from population data. The effect is to make it more likely to reject the null hypothesis of no difference between the FFS and Multiscopo estimates and those of the LFS. Given the very large overall sample size of the LFS (320,000 individuals each trimester in 1985 and 200,000 in 1995, ISTAT 1996), this bias is likely to be small.

Comparisons of education at survey date indicate small deviations only of the survey estimates from population data, and between the surveys. Compared to the LFS of 1995, both the FFS and Multiscopo have significantly higher proportions of women with higher-education qualifications, at around 11%, but differences between the FFS and Multiscopo are small and not significant. Surprisingly, given international trends towards increased female participation in higher education, no statistically significant change is seen across the two Italian cohorts born ten years apart in either the FFS or Multiscopo surveys (statistical test results not shown). To check whether this lack of observed change is due to the different ages of the women from the two cohorts at survey date (early-to-mid 30s for the 1961–1965 cohort versus early-to-mid 1940s for the 1951–1955 cohort), we compared also the 1951–1955 cohort's proportion with higher qualifications 10 years before, in the 1985 LFS. While the 1995 LFS recorded almost identical percentages of women with a higher education between the 1951–1955 cohort (9.4%) and the 1961–65 cohort (9.3%), only 8.0% of women from the 1951–1955 cohort had a higher qualification in 1985. The real growth in higher education across cohorts implied by the LFS, however, is still small: from 8.0% of the 1951–1955 cohort to 9.3% of the 1961–1965 cohort.

The survey data on births are also similar to estimates from population data, and the FFS and Multiscopo data are similar to each other. Compared to the birth-registration data, both the FFS and Multiscopo have similar proportions still childless at age 24

Table 1 First birth timing and education: sample survey and population data for Italian women born 1951–1955 and 1961–1965

	Fertility and Family Survey (FFS)		Multiscopo survey		Birth registration data (population)		Labour force survey		
Birth cohort:	1951–1955	1961–1965	1951–1955	1961–1965	1951–1955	1961–1965	1951–1955		1961–1965
Year of survey:	1995	1995	1998	1998			1985	1995	1995
Percentage with high education qualification	10.7	11.4*	11.0*	11.2**	–	–	8.0	9.4	9.3
Childless percent of all 24-year-old women	45.2	60.6**,+	49.0	65.3	48.3	65.5	–	–	–
Childless percent of 24-year-old women by education									
High education qualification	82.2	96.5	86.5	94.0	–	–	–	–	–
No high education qualification	40.8	56.0++	44.3	61.7	–	–	–	–	–
Sample size (persons)	760	844	2,098	2,693					

* Statistically different from the population value (Birth Registration data or Labour Force Survey) $p < 0.05$
** Statistically different from the population value (Birth Registration data or Labour Force Survey) $p < 0.01$
+ Statistically different from the large survey value (Multiscopo Survey) $p < 0.05$
++ Statistically different from the large survey value (Multiscopo Survey) $p < 0.01$

(the beginning of the year the woman attained age 25). The FFS proportions appear slightly lower, than either the Multiscopo or birth registration data, and the deficit is statistically significant compared to both the Multiscopo and birth registration estimates for the 1961–1965 cohort. The FFS and Multiscopo exhibit similar differentials by education in proportions childlessness at age 24 (much higher among "high education" women) and by cohort (substantially higher for the 1960s cohort than for the 1950s cohort). The lower overall childlessness in the FFS' 1961–1965 cohort is seen to be due to the "no high education" group.

The first-birth probabilities by single-year age over all the observed ages of the study (from age 25 to the oldest age available for each cohort in the respective surveys) are compared between the two survey data sources and the population data source in Fig. 1a and b. From the population data, it is clear that the true pattern of age-specific first-birth hazard is smooth. It is also clear that a major change in the pattern occurred between the 1950s and 1960s cohorts. The hazard is generally lower, and the age pattern later, for the 1960s cohort. The hazard slopes upward until age 30 for the 1960s cohort, while it descends for the 1950s cohort already from age 26. At age 25, the first-birth probability was 0.139 for the 1950s cohort, but only 0.085 for the 1960s cohort. By age 30, the probabilities were similar: 0.111 for the 1950s cohort and 0.104 for the 1960s cohort. By age 34, the probability for the 1960s cohort (0.078) exceeded slightly that for the 1960s cohort (0.069).

From Fig. 1a and b, both sample surveys appear to be approximately representative of the population with respect to both levels and cross-cohort changes. Statistical tests of differences between the surveys for the full age, cohort, and education relationship to first birth were conducted by adding a full set of interactions for survey (Multiscopo against a reference FFS), using a polynomial specification for the relationship of age to the first-birth probability. The addition of the Multiscopo dummy and interactions of age, cohort, and education with this dummy resulted in an improvement in model fit that was statistically significant at the $p = 0.05$ level, but with none of the individual coefficients added for "Multiscopo" and interactions with "Multiscopo" being statistically significant (results available from the first author on request). This indicates again that the two surveys are sampling from approximately the same population process.

Sampling fluctuations appear to be substantially greater in the smaller FFS estimates than in the larger Multiscopo survey estimates, as would be expected given their respective sample sizes. Fluctuations are especially large towards the oldest ages observed for the 1960s cohort (see Fig. 1b). This is due to fewer single-year age birth cohorts contributing exposed years just before survey date. For example, only the 1961 and 1962 cohorts attain age 32 in the FFS observation period. Thus the population pattern of increasing first birth probabilities to age 30 followed by decreases thereafter is not evident in the sample series.

2.3 Constrained maximum likelihood estimation and unconstrained alternatives

We specify a regression model whose dependent variable Y takes the value of 1 in the year that a woman has her first live birth, and 0 in every year that she remains childless. Let X be a vector of length p representing the values of the regressors.

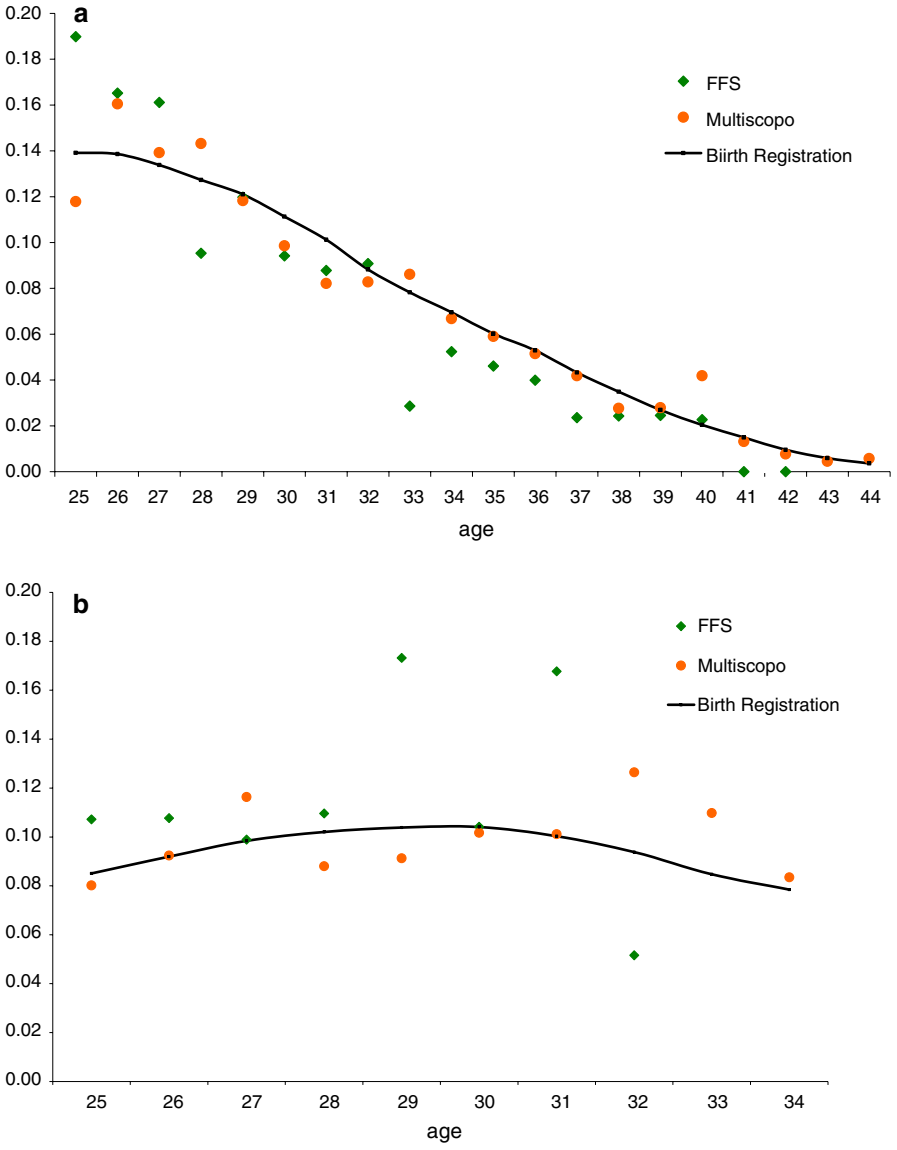

Fig. 1 **a** Italy 1951–1955 cohort first birth probabilities by source of data. **b** Italy 1961–1965 cohort first birth probabilities by source of data

These may be fixed or time-varying. Let $\beta = (\beta_o, \beta_1, \ldots, \beta_p)$ be a parameter vector consisting of an intercept parameter β_0 plus the p coefficient parameters corresponding to each of the regressors in X. A discrete-time first-birth hazard function, where age is the "duration" variable of the hazard, may then be specified as a binomial logistic regression model (e.g., Maddala 1983). Under the logistic model, the birth probability

$P(Y = 1|X = x, \beta)$ is transformed through a log odds function that is linear in x:

$$\log\left[\frac{P(Y = 1|X = x, \beta)}{P(Y = 0|X = x, \beta)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

Denote the survey data by $D = (y_i, x_i)$, $i = 1, \ldots, n$. These are person-year observations of women at ages at which they had not yet had a first birth at the beginning of the year. The value y_i is the realization of Y indicating a birth ($y_i = 1$) or no birth ($y_i = 0$) and x_i is the vector of length p of regressors for the ith person-year in the dataset. The distribution of X may depend on some design parameter γ. We will assume that the parameter space of γ and the parameter space of β are disjoint. The likelihood function for the person-year data given the model of Eq. (1) can be written as:

$$\begin{aligned} L(\beta, \gamma; y, x) &= \prod_{i=1}^{n} P(Y = y_i, X = x_i|\beta, \gamma) \\ &= \prod_{i=1}^{n} P(Y = y_i|X = x_i, \beta) P(X = x_i|\gamma) \end{aligned} \tag{2}$$

Because the parameter spaces of γ and β are disjoint, maximizing (2) with respect to β is equivalent to maximizing $L(\beta; y|x) = \prod_{i=1}^{n} P(Y = y_i|X = x_i, \beta)$. Let $I_n(\beta)$ be the expected Fisher information matrix for the parameter β. It is the $(p + 1) \times (p + 1)$matrix with $(j + 1, k + 1)$th element $-E_\beta\left[\frac{\partial^2 \log[L(\beta; y|x)]}{\partial\beta_j \partial\beta_k}\right]$, $j = 0, 1, 2, \ldots, p$ and $k = 0, 1, 2, \ldots, p$. If the survey data are all the information we have, under standard regularity conditions, the estimated value $\hat{\beta}$ that maximizes the above likelihood is asymptotically unbiased and Gaussian with asymptotic variance V_S, where V_S is the inverse of $I_n(\beta)$ (Casella and Berger 2002). We refer to $\hat{\beta}$ and V_S as the *unconstrained model* estimates.

To introduce the *constrained* model, we first introduce notation for the specific set of variables in our application, age a, birth cohort c, and education e. Education takes the value of $e = 1$ for women with a higher education qualification and e=0 otherwise. Recall that age and cohort are available in both the survey and population data, while education is available in the survey data only. The probability of a first birth for each age and cohort, independent of education, $\phi_{a,c} \equiv P(Y = 1|a, c)$, is known for each observed combination of age and cohort from the population data $\{\phi_{a,c}\}$. These data are exactly those plotted in Fig. 1a and b (the "Birth registration" lines). In this constrained regression model, they are the constraint *values*.

The constraint *function* additionally includes the education variable. Let $P(Y = 1|a, c, e)$ be the probability of a first birth for a woman of age a, from cohort c and with education e. The probability $P(Y = 1|a, c)$ can be expressed as the weighted sum of the probability of a first birth for a woman with a higher education qualification $P(Y = 1|a, c, 1)$ and the probability of a first birth for a woman with no higher education qualification $P(Y = 1|a, c, 0)$, where the weights are given by the proportions of women of that age and cohort with a higher qualification, $\pi(e = 1|a, c)$, and without one, $1 - \pi(e = 1|a, c)$. Both $P(Y = 1|a, c, 1)$ and $P(Y = 1|a, c, 0)$ are derivable as

predicted values from the logistic regression model (1). As the model depends on β, this induces a dependence of these probabilities on β. For given values of $\pi(e = 1|a, c)$, the following constraint function C(β) may therefore be defined:

$$\begin{aligned}\phi_{a,c} = C(\beta) = {} & P(Y = 1|a, c, 1; \beta)\pi(e = 1|a, c) \\ & + P(Y = 1|a, c, 0; \beta)[1 - \pi(e = 1|a, c)]\end{aligned} \tag{3}$$

The constrained maximum likelihood estimator $\hat{\beta}_{\text{con}}$ that solves Eq. (2) subject to constraint function (3) is still asymptotically efficient, unbiased and Gaussian. However, while the asymptotic variance in the unconstrained version is given by V_S, in the constrained version the asymptotic variance of $\hat{\beta}_{\text{con}}$ is:

$$V_{\text{con}} = V_S - V_S H^T [H V_S H^T]^{-1} H V_S \tag{4}$$

where H is the gradient matrix of C(β) with respect to β. H is the $m \times (p + 1)$ matrix with $(l, j + 1)$th element $\frac{\partial C_l(\beta)}{\partial \beta_j}$, $j = 0, 1, 2, \ldots, p$ parameters and $l = 1, 2, \ldots, m$ constraints. Since the second term in expression (4) is positive definite, the inclusion of the population information always leads to an improvement in the estimation of β. That is, the constrained estimator $\hat{\beta}_{\text{con}}$ is, on average, closer to the true value of β than is the unconstrained estimator $\hat{\beta}$. In particular, the standard error of the estimator in the version using the population information will always be less than the unconstrained estimator that ignores it. This is the key result of the constrained maximum likelihood model.

A further result of (4) of particular importance for the present study is that the asymptotic ratio of the variances of the constrained to unconstrained parameters is independent of the survey sample size. The relevance of this derives from the increase in total sample size achieved by pooling observations across the two surveys. Because individuals in the surveys are sampled at random, independence holds. Consequently, when X is fixed and known, the expected Fisher information for the sample, $I_n(\beta)$, can be represented as $nI(\beta)$, where $I(\beta)$ is the expected Fisher information for a given individual. When X is random, the expected Fisher information for (β, γ) is $I_n(\beta, \gamma)$, a block diagonal matrix because γ and β are disjoint. Therefore $I_n(\beta)$ and the expected information for γ, $I_n(\gamma)$, are independent and can be easily extracted from $I_n(\beta, \gamma)$. In both the fixed and random X cases, $V_s = I_n^{-1}(\beta) = [nI(\beta)]^{-1} = \frac{1}{n}V$, where $V = I^{-1}(\beta)$. As a result, the asymptotic variance matrix of the constrained and unconstrained parameters can be represented as $\frac{1}{n}(V – V H^T [H V H^T]^{-1} H V)$ and $\frac{1}{n}V$, respectively. The ratio of the constrained to unconstrained variances is therefore asymptotically independent of sample size. Hence, the *percentage* reduction in the asymptotic standard errors of the regression parameters will be the same for all sample sizes.

The form of constraint equation (3) is very general in demographic applications. It expresses an overall rate P as a weighted sum of covariate-dependent ("specific") rates $P(0)$ and $P(1)$. The weights are given by the population distribution of the covariate $\{\pi, 1 - \pi\}$. This population distribution may be approximated by the sample distribution with the loss of some efficiency, the analytical result for this loss being

derived in Hellerstein and Imbens (1999). In the present application, while sample survey data are used to approximate the population distribution of the regressors, we treat this distribution as if it were calculated from population data. This allows us to apply a pre-written constrained maximization routine (the SAS PROC NLP, SAS Institute 1997) to the likelihood (2) and constraint function (3), and thereby obtain the constrained version of the variance–covariance matrix (4).

Finally, supplied with the sample survey data are sample weights to account for differential probability of selection. We use these weights throughout the analysis. Before pooling the two surveys for the regression estimation, we normalize to a mean of 1 the sample weights of each of the two surveys separately. These normalized weights then form part of the likelihood function (2).

3 Results

The results are divided into two subsections. In the first subsection, we present parameter estimates and predicted birth probabilities under identical regression specifications between the unconstrained and constrained models. This allows us to compare standard errors for coefficients with and without constraints. We compare unconstrained and constrained models estimated with data from only one survey with models estimated with data that are pooled across the two surveys. This shows the incremental benefits respectively from pooling surveys and from constraining survey estimation through the inclusion of population data.

Following this, we present a second unconstrained regression specification that parameterizes the age function as polynomial, allowing for a smoothing of the first birth relationship with age. The parametric approach to hazard estimation is a common solution to the problem of high sampling variability with survey data. We show here that this parametric approach is nevertheless inferior to the approach that smoothes the first birth relationship to age by using the population data as formal constraints on the regression estimation.

3.1 Constrained and unconstrained model estimates under identical specifications

In Table 2, constrained and unconstrained parameter estimates and standard errors are presented for the logistic regression of first birth on age, cohort, and education. Separate results are reported using the small (FFS) survey only, the large (Multiscopo) survey only, and the FFS and Multiscopo surveys with their observations pooled. The function of age and cohort to first birth is specified using single-year ages (that is, completely non-parametric), while we parameterize (with a second-order polynomial) the education by age interaction. This is because we have exact population information about the age and cohort relationships, but must rely on survey data for information about the education relationship.

Consistent with Eq. (4) in the statistical theory presented above, all standard errors in the constrained version are as low as, or lower than, the corresponding standard errors of the unconstrained version. The standard errors of the age parameters are seen to be reduced by very large amounts by constraining survey-based estimates to the

Table 2 Unconstrained and constrained logisitic regressions using the FFS, Multiscopo, and pooled FFS and Multiscopo

	Unconstrained						Constrained					
	FFS		Multiscopo		FFS and Multiscopo		FFS		Multiscopo		FFS and Multiscopo	
	Param.	SE	Param.	SE	Param.	SE	Param.	SE	Param.	SE	Param.	SE
Intercept	−1.282**	0.139	−1.868**	0.097	−1.700**	0.079	−1.661**	0.028	−1.676**	0.017	−1.674**	0.014
High education qualification	−1.215**	0.338	−1.180**	0.213	−1.164**	0.178	−1.203**	0.337	−1.177**	0.213	−1.162**	0.178
Age (ref. age = 25)												
26	−0.184	0.209	0.344**	0.131	0.187	0.110	−0.016	0.011	−0.023**	0.004	−0.021**	0.004
27	−0.258	0.222	0.148	0.141	0.023	0.118	−0.098**	0.024	−0.091	0.010	−0.092**	0.010
28	−0.912**	0.277	0.140	0.147	−0.123	0.127	−0.207**	0.043	−0.188	0.018	−0.191**	0.017
29	−0.719**	0.271	−0.126	0.162	−0.294*	0.138	−0.328**	0.060	−0.292	0.027	−0.297**	0.024
30	−1.048**	0.310	−0.377*	0.180	−0.562**	0.155	−0.482**	0.076	−0.434*	0.036	−0.440**	0.032
31	−1.185**	0.335	−0.623**	0.200	−0.780**	0.171	−0.647**	0.092	−0.585**	0.044	−0.593**	0.039
32	−1.202**	0.346	−0.655**	0.207	−0.806**	0.177	−0.851**	0.103	−0.777**	0.051	−0.786**	0.045
33	−2.473**	0.572	−0.636**	0.212	−1.006**	0.194	−1.023**	0.111	−0.932**	0.053	−0.944**	0.048
34	−1.873**	0.452	−0.929**	0.240	−1.171**	0.211	−1.185**	0.119	−1.079**	0.055	−1.094**	0.050
35	−2.041**	0.491	−1.077**	0.260	−1.325**	0.228	−1.373**	0.136	−1.248**	0.059	−1.267**	0.054
36	−2.194**	0.537	−1.216**	0.281	−1.467**	0.247	−1.514**	0.154	−1.380**	0.060	−1.402**	0.056
37	−2.733**	0.691	−1.420**	0.313	−1.737**	0.283	−1.719**	0.186	−1.578**	0.064	−1.602**	0.061
38	−2.666**	0.697	−1.833**	0.382	−2.056**	0.333	−1.910**	0.214	−1.785**	0.070	−1.807**	0.067
39	−2.589**	0.785	−1.794**	0.386	−2.005**	0.345	−2.112**	0.215	−2.022**	0.075	−2.039**	0.070
40	−2.659**	0.976	−1.347**	0.330	−1.616**	0.311	−2.385**	0.277	−2.280**	0.084	−2.300**	0.081
41	−14.870	523.834	−2.504**	0.564	−2.814**	0.560	−2.650**	0.270	−2.571**	0.091	−2.589**	0.088
42	−14.510	554.688	−3.012**	0.732	−3.280**	0.730	−3.077**	0.329	−2.992**	0.096	−3.011**	0.095
43			−3.521**	1.090	−3.710**	1.089			−3.433**	0.116	−3.454**	0.115
44			−3.228**	1.110	−3.416**	1.109			−3.873**	0.123	−3.893**	0.122

Table 2 continued

	Unconstrained						Constrained					
	FFS		Multiscopo		FFS and Multiscopo		FFS		Multiscopo		FFS and Multiscopo	
	Param.	SE	Param.	SE	Param.	SE	Param.	SE	Param.	SE	Param.	SE
Cohort 61–65	−0.718**	0.199	−0.433**	0.132	−0.519**	0.110	−0.598**	0.039	−0.560**	0.017	−0.566**	0.016
Cohort 61–65 and high education	0.339	0.445	−0.395	0.273	−0.199	0.226	0.330	0.444	−0.395	0.273	−0.200	0.226
Age for Cohort 61–65 (ref. age = 25)												
26	0.175	0.291	−0.198	0.182	−0.086	0.153	0.089**	0.011	0.098	0.004	0.095**	0.004
27	0.136	0.309	0.244	0.187	0.237	0.159	0.229**	0.025	0.229	0.010	0.225**	0.009
28	0.882*	0.351	−0.080	0.199	0.150	0.171	0.355**	0.047	0.347	0.016	0.340**	0.014
29	1.210**	0.351	0.195	0.212	0.435*	0.180	0.483**	0.073	0.439	0.023	0.437**	0.021
30	0.946*	0.441	0.531*	0.226	0.630**	0.197	0.636**	0.117	0.548*	0.033	0.550**	0.031
31	1.620**	0.475	0.735**	0.245	0.869**	0.213	0.746**	0.154	0.623**	0.045	0.630**	0.042
32	0.309	0.905	0.984**	0.250	1.031**	0.220	0.859**	0.144	0.706**	0.058	0.722**	0.054
33			0.779**	0.270	1.075**	0.251			0.723**	0.070	0.748**	0.065
34			0.739*	0.324	0.917**	0.298			0.756*	0.086	0.792**	0.079
Interaction												
Age and high education	0.377**	0.139	0.341**	0.073	0.339**	0.063	0.372**	0.138	0.340**	0.073	0.338**	0.063
Age, high education and cohort 61–65	−0.153	0.127	0.019	0.052	−0.013	0.045	−0.150	0.126	0.019	0.052	−0.013	0.045
Age squared and high education	−0.016	0.011	−0.016**	0.005	−0.015**	0.004	−0.016	0.011	−0.016**	0.005	−0.015**	0.004
$-2\log L$ intercept and covariates	3,237.987		11,937.004		15,230.601		3,281.535		11,979.917		15,267.134	
Person years	4,945		19,596		24,541		4,945		19,596		24,541	

* $p < 0.05$
** $p < 0.01$

overall population values, generally by 75% or more as compared to the unconstrained version, and sometimes by as much as 90 percent. Only for the age parameters, cohort-by-age parameters, and intercept, however, are the reductions in standard errors other than of negligible magnitudes. That is, for none of the parameters for education and its interaction with age and cohort is there a non-negligible reduction in the standard error. This makes intuitive sense, as the constraints offer exact information about the relationship of age to first childbearing, but no information about how this relationship differs by education.

A further result of Eq. (4) noted in the statistical theory description above is confirmed empirically in Table 2: the ratio of the variances of the constrained to unconstrained parameters is independent of the survey sample size. The asymptotic result is that the *percentage* reduction in the standard errors of the regression parameters from the unconstrained to the constrained versions will be equal. This is seen to be closely approximated in practice for the FFS and the Multiscopo. Thus even while the sample size of the Multiscopo is approximately four times as high as the sample size of the FFS, there is no difference in the proportionate reduction of the standard error about the first-birth model coefficient estimates. Importantly, the standard errors for the pooled sample are reduced by similar amounts in percentage terms as are the standard errors for either of the two surveys alone. For example, for the age-40 coefficient, the standard error for estimation with the FFS is reduced from an unconstrained-model 0.976 to a constrained-model 0.277, an approximately 75% reduction. When estimating the unconstrained and constrained models with the pooled FFS and Multiscopo, the standard error falls from 0.311 to 0.081, again an approximately 75% reduction.

While the population constraints have a negligible effect on the standard errors of the coefficients for education, and for the interaction of education with age and cohort, pooling the two samples results in substantial reductions in these standard errors. These reductions are seen equally in the constrained and unconstrained estimates, although we focus on the constrained estimates. Compared with using the FFS alone, the standard error for the parameter for the main effect (at age 25 for the 1951–1955 cohort) of having a higher education qualification is halved (from 0.337 to 0.178). Compared with using the Multiscopo alone, the standard error for the same parameter is reduced from 0.213 to 0.178. Similarly large reductions by adding the Multiscopo data to the FFS data, and much smaller but still substantial reductions by adding the FFS data to the Multiscopo data, are seen in the standard errors for the parameters for higher-education interactions with cohort and age.

The practical advantages of pooling survey data under population-constrained estimation are best seen by graphing the predicted first birth probabilities by age, cohort, and education. These predicted probabilities for the estimation that uses the pooled survey data with the population information as constraints to the survey estimation are first presented in Fig. 2a and b. We consider these our best estimates of the relationship of age, cohort, and education to first-birth, since they take into account all available survey and population data. Confidence intervals for these estimates, as for all the predicted probabilities presented in this article, were generated using a bootstrap procedure (Efron and Tibshirani 1994) with 1,000 iterations. The 95% confidence interval shown in the graphs consists of the 5th percentile and 95th percentile of the bootstrapped estimates.

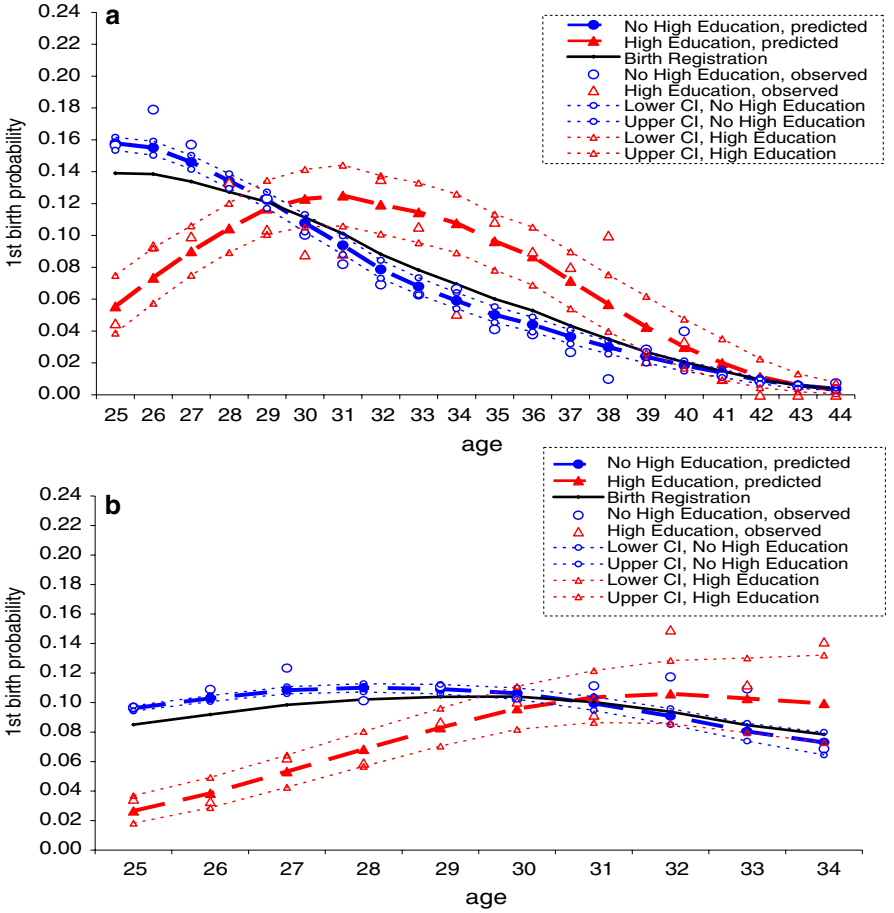

Fig. 2 **a** 1951–1955 cohort constrained estimation with pooled (FFS and Multiscopo) survey data. **b** 1961–1965 cohort constrained estimation with pooled (FFS and Multiscopo) survey data. **c** 1951–1955 cohort constrained estimation with small (FFS) survey. **d** 1961–1965 cohort constrained estimation with small (FFS) survey

The 1950s cohort's predicted first-birth probabilities show highly differentiated patterns by education (see Fig. 2a). The downward-sloping profile from age 26 seen in the birth registration data is modeled for women without a high education, while the pattern for women with a high education is modeled as sloping steeply upwards to a peak first-birth probability at age 31. The modeled pattern follows the observed probabilities closely for women with no high education. The observed probabilities for women with high education qualifications, however, fluctuate much more around the predicted line. This is expected given that relatively few women in the cohort, and therefore also in the sample, have a higher qualification.

Some similar remarks may be made about the 1960s cohort's constrained estimates versus the observed data and overall first-birth probabilities in the population data (see Fig. 2b). Up to about age 30, the fit of the lines to the observed data appears as

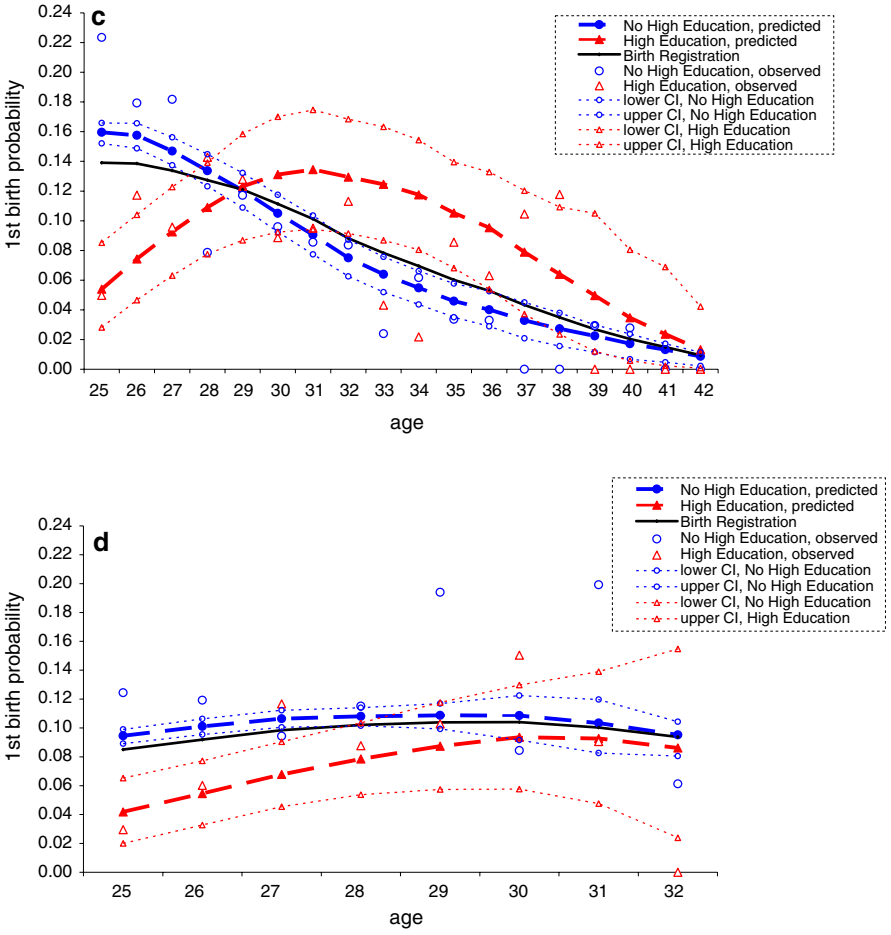

Fig. 2 continued

if it were a simple smoothing of the sample data. After age 30, however, the effect of the constraint is clearly much stronger than seen either before age 30 or in the case of the 1950s cohort. The constraint pulls both education-specific lines downwards so that they are on average much lower than their observed sample points. For the higher-education women, for example, little evidence of a downward slope emerging by age 34 is seen in the sample points. The implication of the predicted education-specific lines after age 30 is that the observed sample points may be biased upwards. This may be because, for example, non-response is differentially low for women who had children in the year before survey date. The population data, however, are not subject to response differentials, and therefore are expected to be unbiased. Using them in the constrained estimation therefore will correct for bias in the survey data.

We present in Fig. 2c and d the predicted values for the constrained estimator using only the smaller, FFS dataset. The main objective here is to show, by contrast with

Fig. 2a and b, how pooling survey data may lead to substantial improvements especially in estimating those parts of the relationship for which population information is not available. While a similar relationship of education to first birth is seen under constrained estimation using the FFS only, the confidence intervals around the predicted probabilities are much wider. For example, while the confidence intervals for "High Education" and "No High Education" women over 30 in the 1950s cohort are non-overlapping only between the ages 32 and 35 for the FFS, they are non-overlapping from ages 31 to 38 with the combined FFS and Multiscopo surveys.

The largest improvements achieved by using all of the available data are again seen for the 1960s cohort. Here, the constrained estimator with the FFS data results in the higher educated women's first birth hazard approaching but never exceeding the hazard for women without a higher-education qualification (see Fig. 2d). This contrasts with the cross-over at about age 31 seen for the constrained estimator that pools the FFS and Multiscopo data (Fig. 2b). The failure of the FFS constrained estimator to model the education cross-over is due to a combination of its observations going only up to age 32 and to its much smaller sample size. Note that at age 32, no first births were observed in the FFS sample (see the "High Education, observed" points on the plot).

3.2 Parametric and non-parametric specifications of age in unconstrained estimation

The researcher who uses sample data only is unlikely to specify the non-parametric, single-year age dummy model used in constrained estimation. Instead, a smooth relationship of the first-birth probability with age is likely to be imposed parametrically. We now illustrate graphically that both the non-parametric and parametric approaches will be inferior to the approach that uses the population data as formal constraints to the estimation. For the parametric version, a polynomial age specification regression with linear, squared, and cubed terms for the reference, 1950s cohort, and with linear and squared interaction terms for the 1960s cohort is estimated (parameter estimates available from the first author on request). The non-parametric version uses the specification from Table 2 above. The two versions are intended to give the range of likely alternative estimation strategies (from completely non-parametric to the simplest parametric specification) in the case that no statistical method for the incorporation of known population information is available to the researcher. The predicted values for the non-parametric and parametric unconstrained specifications, in all cases using the pooled survey data, are shown in Fig. 3a–d.

The population line is included in the graphs to show how estimates from the sample data, whether using non-parametric or parametric specifications, may be inconsistent with the overall population values. This contrasts with Fig. 2a and b, where such inconsistency is prevented by the method of constraining to the population values. The predicted values for the unconstrained non-parametric, single-year age dummy specifications shown in Fig. 3a and b generate jagged lines for both the education-specific probability series. False local peaks in the hazard, for example, occur at age 40 for the 1950s cohort and at age 27 for the 1960s cohort. This is clearly attributable to sampling error, as the population function is known from population data to be smooth across these ages.

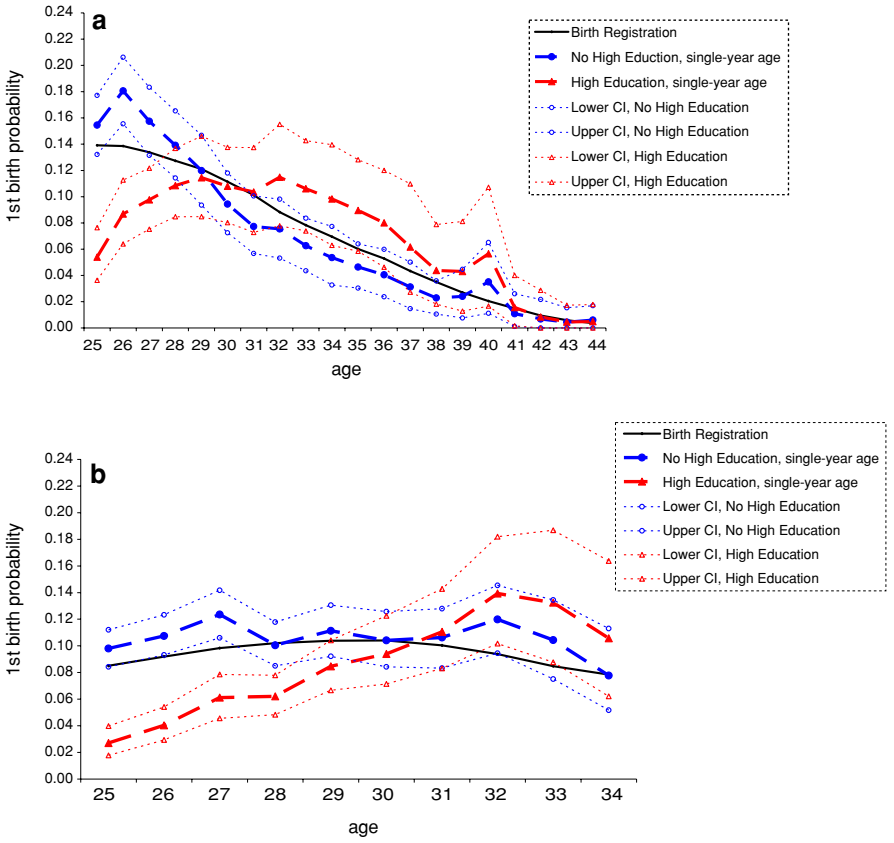

Fig. 3 **a** 1951–1955 cohort unconstrained estimation with pooled (FFS and Multiscopo) survey data, single year age dummy ("Non-parametric") specification. **b** 1961–1965 cohort unconstrained estimation with pooled (FFS and Multiscopo) survey data, single age dummy ("Non-parametric") specification. **c** 1951–1955 cohort unconstrained estimation with pooled (FFS and Multiscopo) survey data, polynomial ("Parametric") age specification. **d** 1961–1965 cohort unconstrained estimation with pooled (FFS and Multiscopo) survey data, polynomial ("Parametric") age specification

Predicted values are presented in Fig. 3c and d for the parametric version. For the 1950s cohort (see Fig. 3c), the unconstrained polynomial age specification lines are very similar in pattern to those seen for the constrained estimate of Fig. 2a. There is a similar cross-over point, at about age 29, between the higher-qualified and not-higher-qualified women. This parametric specification appears to model reasonably well the relationship seen in the sample data. For the 1960s cohort, however, it produces predicted values that exceed the population values for *both* higher-qualified and not-higher-qualified women after age 30 (see Fig. 3d). Such deviations from a known population relationship are possible because the parametric smoothing has no effect on the overall level of the hazard.

The effect of the population constraint in Fig. 2a through d is now clearer when contrasted with Fig. 3a through d. While the patterns of first-birth probabilities in Fig. 2a through d appear to be similar to those that would emerge from a parametric

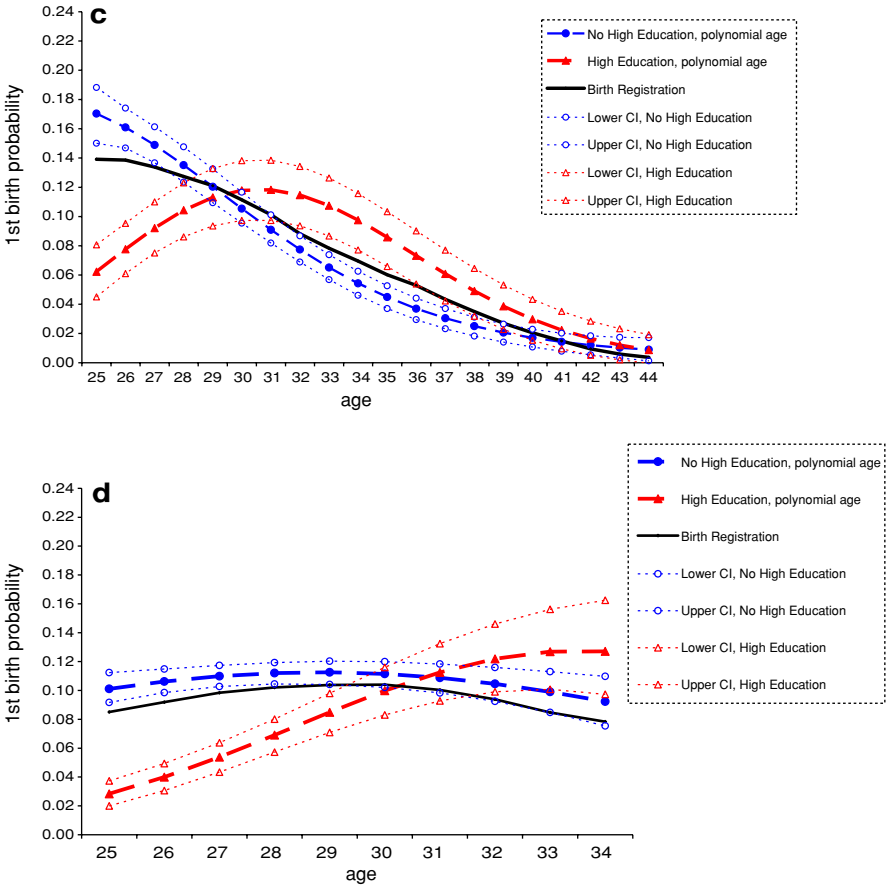

Fig. 3 continued

or non-parametric smoothing of the two education-specific series, the two education-specific lines always surround the population constraint line. This is a result of the population line's being a weighted sum of the two education-specific lines at each single-year age. This is most obvious at the point at which the education-specific lines cross, which is forced to be the point at which they are equal to the known overall first-birth probability in the population (the constraint line). Both the parametric and non-parametric versions of the unconstrained estimation, in contrast, allow drift in the two education-specific hazards from the known overall population hazard of first birth by age and cohort.

4 Summary and conclusions

Previous demographic and economic studies have demonstrated large efficiency gains through combining population data with survey data in regression estimation. These

gains, however, have been limited to the intercept parameter and the coefficients for variables for which population data are also available. The present study demonstrated how this limitation can be overcome by pooling data from more than one survey sample and constraining estimates from the pooled surveys to population data.

Full use of available population data was achieved by imposing population constraints by single-year age, parity, and cohort. This introduced an exact, baseline relationship of age to first childbearing separately for two five-year birth cohorts. Observations from a second, large-scale survey (the 1998 Multiscopo) pooled with observations from a specialist demographic survey (the 1995/1996 FFS) allowed for much greater efficiency in the estimation of the relationship of a key socio-economic variable (educational attainment) to first birth by age. As expected, however, negligible reductions in the standard errors for the parameters for education and its interaction with age and cohort were achieved by the imposing of population constraints. The intuition for this is that the constraints offer exact information about the relationship of age and cohort to first birth, but no information about how this relationship differs by the education levels of cohort members.

Additional information about how first birth differs by education was instead obtained by pooling the data from the small survey with observations on women from the same cohorts in a larger survey in which the education variable and fertility histories were also present. Here, the efficiency gains over using the smaller survey alone are equivalent to increasing the latter's sample by the number of observations in the larger survey. Because the larger, Multiscopo survey has approximately four times the person-year sample size of the smaller, FFS, the standard errors about the education coefficients were approximately half those estimated using the FFS data alone. Pooling the survey data, moreover, does nothing to reduce the effectiveness of using population constraints. Both theoretical and empirical results were presented showing that the percentage reduction in the standard errors achieved by applying population constraints is independent of the survey sample size, and therefore equally effective when surveys are pooled.

The structures of the survey datasets and population data used in the present study have permitted a largely straightforward statistical treatment. The two survey datasets used here have been treated as though they sample from the same population, and contain the same variables needed to estimate the relationships of interest. In one practically important way, the larger dataset also contributed variables not present in the smaller survey. These were from observations of women at ages 43 and 44 in the 1950s cohort and at ages 33 and 34 in the 1960s cohort. Their practical significance is that they complete the ages of reproduction for the 1950s cohort, and extend predictions over ages at which first birth hazards are high, especially for women with higher education, in the 1960s cohort. This presents no statistical complication for hazard modeling, since adding ages of observation does no more than relax the degree of right-censoring of first-birth exposure. Pooling data from surveys with more general differences in their regressor variables is also possible, but involves greater statistical challenges (see Ridder and Moffitt 2007).

The population data used here were treated as exact, in the senses both of being unbiased and having negligible sampling error. This assumption will not hold for all population data collections. The Italian statistical system for the collection of births

data was itself overhauled in 1999, such that information on mother's age and parity is no longer available in a single, complete-enumeration source (LoConte et al. 2003). This means that only by using data collections that include sampling error will it be possible to construct age- and parity-specific population constraints from 1999 onwards. This complicates, but does not eliminate, the possibilities for improving survey estimates. Hellerstein and Imbens (1999) show this by deriving a variance estimator that adjusts for sampling error in "population" constraints from large-scale sample survey data.

Acknowledgments We are very grateful to Piero Giorgi for providing us with first-birth probabilities by age and cohort calculated from Italian birth-registration data, and for comments received at presentations of earlier versions at the August 2004 Meeting of the Logic and Methodology section of the International Sociological Association, and at the June 2006 Meeting of the Italian Statistical Society. This work was funded by the National Institute of Child Health and Human Development under investigator grant R01-HD04347201, and under center grants to the RAND Labor and Population program (R24-HD050906), and to the University of Washington Center for Studies in Demography and Ecology (R24-HD41025); and by a grant to Alessandra DeRose from the University of Rome (Inter-faculties Research on "Integrating current data and survey samples for the analysis of family behaviors" 2000/02).

References

Casella G, Berger RL (2002) Statistical inference, 2nd edn. Duxbury Press, Pacific Grove

Deming WE, Stephan FF (1942) On the least squares adjustment of a sampled frequency table when the expected marginal tables are known. Ann Math Stat 11:427–424

De Sandre P et al (2000) Fertility and Family Surveys in the ECE Region Standard Country Report: Italy Geneva: United Nations Economic Commission for Europe, Population Activities Unit

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman and Hall, New York

Giorgi P (1993) Una rilettura della fecondità del momento per ordine di nascita in Italia nel periodo 1950–1990 considerando la struttura per parità. Genus 40(3–4):177–204

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

Handcock MS, Rendall MS, Cheadle JE (2005) Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. Sociol Methodol 35(1):291–334

Hellerstein J, Imbens GW (1999) Imposing moment restrictions from auxiliary data by weighting. Rev Econ Stat 81(1):1–14

Imbens GW, Lancaster T (1994) Combining micro and macro data in microeconometric models. Rev Econ Stud 61:655–680

Ireland CT, Kullback S (1968) Contingency tables with given marginals. Biometrika 55:179–188

ISTAT (2000) Indagine Statisitca Multiscopo sulla Famiglia, 1998. Istituto Nazionale di Statistica, Rome

ISTAT (1996) Forze di Lavoro, Media 1995 Serie Annuari. Istituto Nazionale di Statistica, Rome

ISTAT (2005) Elaborazioni Istat su dati ricostruiti progetto MARSS. Istituto Nazionale di Statistica, Rome

LoConte M, Castagnaro C, Talucci V, Prati S (2003) The first sample survey on births in Italy: purposes and results. Paper presented at the 2003 European Population Conference, Warsaw, Poland

Maddala GS (1983) Limited-dependent and qualitative variables in econometrics. Cambridge University Press, New York

OECD (2003) Education statistics and indicators, education at a glance, 2002 edition. http://www.oecd.org

Rice JA (1995) Mathematical statistics and data analysis. Wadsworth, Pacific Grove

Ridder G, Moffitt RA (2007) The econometrics of data combination. In: Heckman JJ, Leamer EE (eds) Handbook of econometrics. North Holland, Amsterdam

SAS Institute (1997) SAS/OR Technical Report: The NLP Procedure. SAS Institute Inc., Cary, NC