A period TFR with covariates

Gustavo De Santis (Dept. of Statistics, University of Florence, desantis@ds.unifi.it)
Sven Drefahl (Stockholm University, Demography Unit, sven.drefahl@sociology.su.se)
Daniele Vignoli (Dept. of Statistics, University of Florence, vignoli@ds.unifi.it)

Introduction and purpose

In this paper we suggest a (relatively) simple and (partly) original way of studying fertility with cross-sectional survey data, calculating the state-specific contributions to Total Fertility. This avoids a few of the selection problems that frequently affect other applications (specific by birth order), and explicitly aims at reconciling micro- and macro analysis, because the group-specific fertility estimates that we obtain are consistent with the observed (or estimated) period TFR for the entire population.

Logistic regression and EHA

Logistic regression is typically applied to fertility micro-data within an EHA (Event History Analysis) framework. A panel of subjects (normally women, or sometimes couples) are observed over a long time span, with the purpose of estimating the likelihood of a birth - in each subgroup, relative to a reference group. The analysis, which is normally conducted separately by birth order, need not cover the entire reproductive age span, or all the women. The covariates are observed at the beginning of each period (before fertility manifests itself) and they can be interpreted as determinants of higher or lower fertility with respect to a reference group. The interest is indeed typically not in the absolute level of fertility, but in the relative distance between an arbitrarily-chosen reference group (e.g. married women, aged 25-29, with medium education, employed, etc.) and the others, differing by one characteristic at a time (the "cause" under scrutiny), with everything else equal.
However, when the observation period is short, life courses cannot be followed, which inherently contradicts the requirements and the logic of EHA. Occasionally, even the mere analysis of order-specific probabilities of birth may be impossible (parity, or previous number of births, may not be asked of women) or lead to distorted results, because of selection. Imagine being interested in the probability of first birth: obviously, all women who have already become mothers are excluded from the analysis, which leaves only a selected sample of women under observation (those who delayed first birth). The effect of this selection is a priori unpredictable.
Other types of selection, too, may be at work, and occasionally go unnoticed. Imagine that one is interested in the probability of first birth, and includes among the explanatory variables employment and marital state (e.g. married, cohabiting, single). Being unemployed may results in a low probability of marriage, and in a few contexts, including Italy, this depresses fertility. But it may well happen that, *ceteris paribus* (that is: for every given marital state of the woman), the probabilities of first birth do not differ much between the employed and the unemployed, which may lead to the erroneous conclusion that unemployment has little or no impact on fertility. With a short period of observation, still another selection may operate. Recently-formed couples are more fertile than others, especially in terms of first births. But recently-formed couples may also be selected by certain characteristics - e.g. they more frequently live in rented, instead of owned, apartments, and have precarious instead of stable jobs. If the length of the union is not known, the wrong conclusion that rented apartments or precarious jobs stimulate fertility may come to one's mind.
Beside standard applications of EHA and logistic regression, researchers have moved a step forward in handling selection/endogeneity issues using multiprocess modelling. This set of methods allows

one to identify selection into specific careers (for instance, fertility or employment) and among parallel careers (e.g., Matysiak, 2009). Similarly, fixed-effect models account for selection on time-invariant (for instance, maternal) characteristics (e.g., Steele et al., 2009). Such approaches, however, tend to complicate both the theoretical framework and the empirical set of results. The approach proposed here is rather simple and, in addition, makes it possible to reconcile individual-level findings and aggregate measures of fertility. Incidentally, the latter advantage is crucial because it facilitates to bridge the increasing gap between macro- and micro-level research. Imagine for instance that the data tell us that being unemployed lowers the probability of first birth by 10%. The overall impact of this finding on the TFR will differ sharply if unemployment is widespread or rare - but this is something that odds ratios alone cannot reveal.

Our approach

Cross sectional data are very common in social sciences and it seems therefore convenient to find a way to exploit them as much as possible - even in the absence of complete fertility histories. Let us imagine that we are interested in women and in their fertility in the course of the next short interval of time (let us say, 12 months). At the beginning of the period these women are aged x (where x covers the whole reproductive period, possibly by age groups: 15-19; 20-24; ...; 45-49), and have certain other observed characteristics (all discrete) in terms of, for instance, employment (yes/no); living condition (single; with an unmarried partner; with a husband); education (low, medium, high), etc. Let us cumulatively refer to this set of characteristics with c (1, ..., c, ..., C): the notation (x,c) will thus characterize women with age x and a given and known set of other characteristics (e.g. employed, single, with high education).
As customary, a specific group (x_s,c_s) will act as a standard of reference against which, with logistic regression, we will estimate the relative probability of all other groups to give birth to a child in the next (short) period. Let us call this estimated relative probability $p_{x,c}$. From this, we can calculate the age-specific probability across all groups, as a weighted average of group-specific probabilities, at age x

1) $p_x = \Sigma_c p_{x,c} w_{x,c}$

We can also obtain

2) $TPR = \Sigma_x p_x$

which is a "total probability rate" - an odd definition, admittedly, but its use (as a proxy of the *TFR*) will be clearer shortly.
Let us now assume that $p_{x,c}$ is sufficiently close to a linear transformation of the age- and group-specific fertility rate $f_{x,c}$. This is indeed normally the case (see e.g., Caselli, Vallin, Wunsch, 2006 and Appendix A). Besides, the *TFR* of the period under consideration is normally known (from official data), or can be estimated (from the sample of data itself, or from some other source).
We can thus compute the ratio k

3) $k = TFR/TPR$

and use this k as a correction factor. Assuming that k is constant (and not group- or age- specific), we can now obtain

4) $f_{x,c} = k\, p_{x,c}$

and calculate group- or path-specific *TFRs*. For variables that can be assumed to remain constant over time (e.g. religious affiliation), we simply get

5) $TFR_c = \Sigma_x f_{x,c}$

In other cases, e.g. marital status, it may make more sense to imagine a life-path, by which a woman is first single (e.g. at ages 15-19 and 20-24), then cohabiting (e.g. at age 25-29), then married (between 30 and, say, 39), then again single. Of course, one may study any trajectory, but privilege will normally be given to the trajectories that are most frequent in any given community.
Finally, since

6) $TFR = k\ TPR = k\ \Sigma_x \Sigma_c p_{x,c}\ w_{x,c}$

the relative contribution of each state *c* to the overall fertility level of the period can be determined, in terms of both fertility ($f_{x,c} = k\ p_{x,c}$) and relative frequency ($w_{x,c}$). Note, however, that the passage from group-specific to the general *TFR* is not as straightforward as it may appear at first sight, as discussed in appendix B.

Data

For our empirical application, we use four waves of the Italian section of the EU-SILC (Community Statistics on Income and Living Condition), 2004-2007. The EU-SILC survey is the statistical data reference source for comparative statistics on income for the European Union and is conducted in each member state: it collects detailed longitudinal information on the social and economic characteristics of individuals and households. The Italian EU-SILC, launched in 2004, follows the rotational design proposed by Eurostat (European Commission 2010). Each year a new sample is drawn, and it is then followed for 4 years. Each sample is representative of the whole Italian population.
In our data set we include individuals who were first interviewed in 2004, 2005, or 2006, and re-interviewed at least once in subsequent years, and were thus observed for 1 to 3 consecutive years. Weights are provided by the Italian National Institute of Statistics to correct the biases that may derive from this complex sampling scheme and from non-response. We consider all women, aged 16-49 at each EU-SILC wave, who can be observed for a whole year. In total, we include about 39 000 women and about 1 600 births. Each additional year is considered as an independent observation, adjusting standard errors to allow for possible intra-group correlations.
We rely on the set of covariates proposed in Vignoli et al. (2011). The marital condition of the woman distinguishes three cases: no partner, married to, or simply living together with a partner. The respondent's education is grouped into three categories (low, medium and high), consistent with the International Standard Classification of Education (ISCED)[1].The region of residence is broken down into three categories: Northern, Central, and Southern Italy (including the Islands). The main activity status indicates whether the woman is employed or not. Among those who do not work we distinguish between the unemployed and those who are not active or still in education. This variable is self-declared and it thereby reflects the respondents' perception of their current main activity status.

Results

(** to be written**)

Discussion

(** to be written**)

[1] The lowest category corresponds to lower secondary school, primary school, or lower education. In the intermediate level we find people with upper secondary education or post-secondary, but non-tertiary, education. Individuals with tertiary education are assigned to the highest category.

Appendix A. From probabilities to rates

Let us first start with absolute, not relative probabilities. Probabilities are virtually indistinguishable from rates when: 1) the event does not eliminate the subject from observation; and 2) it can happen at most once in the observed period.

Condition (1) is perfectly fulfilled with fertility. Condition (2) is not, or better, not perfectly because of short spacing (a woman can, in principle, have two births in less than 12 months, e.g., in January and then again in December) and because of twins. Both case, however, are rare (less than 2%, all together). Besides, if the occurrence of both phenomena (short spacing and twins) is not correlated with the classification adopted (our c groups: living condition, employment, education, etc.), the comparative analysis will not be distorted. The only consequence is that *TPR* will be lower than *TFR* (by about 2%), which will lead to a correction factor k (slightly) higher than 1 in equation (3).

If, as in equation (3), one works with relative probabilities, each of it will not be too far from 1 (1 meaning that a birth is as likely in group c as in the reference group), and the *TPR* in equation (3) will therefore be much greater than the *TFR*. In this case, the correction factor k will be (much) lower than 1. But, once again, this will not affect the comparison between sub-groups.

Appendix B. From group-specific to the general TFR

Assuming that the underestimation factor k is constant, and that the group or path-specific TFR's are $f_{x,c} = k\, p_{x,c}$, eq. (6) becomes

B.1) $$TFR = \Sigma_c \, \Sigma_x f_{x,c} \, w_{x,c}$$

Note, first, that

B.2) $$\Sigma_c \, \Sigma_x f_{x,c} \, w_{x,c} = A$$

where A is the number of age classes considered. This could be, for instance, 35 with yearly age classes (15 to 49). In our case we have A=7, because we work with 7 age classes (16-19; 20-24; ...; 45-49 - these are 5-year age classes, except for the first). Now, consider a category of women whose relative frequency in the population, over the whole reproductive period is w_c ($=W_{c,16\text{-}49}/W_{16\text{-}49}$). If the relative frequency of these women is (at least approximately) constant in each of the age classes considered, and therefore if $w_{x,c} \approx w_c$, then eq. B.1 becomes

B.3) $$TFR = \Sigma_c \, \Sigma_x f_{x,c} \, w_{x,c} \approx \Sigma_c \, \Sigma_x f_{x,c} \, w_c = \Sigma_c \, w_c \, \Sigma_x f_{x,c} = \Sigma_c \, w_c \, TFR_c$$

which means that the general *TFR* can, at least approximately, be thought of as a weighted average of the c group-specific *TFR*'s, each with weight w_c. If, on the contrary, $w_{x,c} \neq w_c$, then the relative impact of group c on the global *TFR* will have to be derived step by step, that is by considering separately what happens within each age class, as in eq. B.1.

Preliminary references

Caselli G., Vallin J., Wunsch G. (2006), *Demography: Analysis and Synthesis*, Four Volume Set: *A Treatise in Population*. Academic Press.

Matysiak, A. (2009), Employment first, then childbearing: Women's strategy in post-socialist Poland, *Population Studies* 63(3): 253-276

Steele F., Sigle-Rushton W., Kravdal Ø. (2009), Consequences of Family Disruption on Children's Educational Outcomes in Norway, *Demography* 46(3): 553–574.

Vignoli D., Drefahl S., De Santis G. (2011), Whose job instability affects the likelihood of becoming a parent in Italy? A tale of two partners, *Demographic Research* (forthcoming).