A segmented regression model for event history data: An application to the fertility patterns in Italy

Article *in* Journal of Applied Statistics · September 2009

DOI: 10.1080/02664760802552994 · Source: RePEc

CITATIONS
10

READS
303

3 authors:

Vito Muggeo
Università degli Studi di Palermo
97 PUBLICATIONS **5,704** CITATIONS
SEE PROFILE

Massimo Attanasio
Università degli Studi di Palermo
57 PUBLICATIONS **1,346** CITATIONS
SEE PROFILE

Mariano Porcu
Università degli studi di Cagliari
59 PUBLICATIONS **595** CITATIONS
SEE PROFILE

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713428038

A segmented regression model for event history data: an application to the fertility patterns in Italy

Vito M. R. Muggeo [a]; Massimo Attanasio [a]; Mariano Porcu [b]
[a] Dipartimento Scienze Statistiche e Matematiche 'Vianelli', Università di Palermo, Italy [b] Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Italy

First Published:September2009

To cite this Article Muggeo, Vito M. R., Attanasio, Massimo and Porcu, Mariano(2009)'A segmented regression model for event history data: an application to the fertility patterns in Italy',Journal of Applied Statistics,36:9,973 — 988
To link to this Article: DOI: 10.1080/02664760802552994
URL: http://dx.doi.org/10.1080/02664760802552994

Journal of Applied Statistics
Vol. 36, No. 9, September 2009, 973–988

Taylor & Francis
Taylor & Francis Group

A segmented regression model for event history data: an application to the fertility patterns in Italy

Vito M.R. Muggeo[a]*, Massimo Attanasio[a] and Mariano Porcu[b]

[a]*Dipartimento Scienze Statistiche e Matematiche 'Vianelli', Università di Palermo, Italy;* [b]*Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Italy*

(*Received 31 January 2008; final version received 14 October 2008*)

We propose a segmented discrete-time model for the analysis of event history data in demographic research. Through a unified regression framework, the model provides estimates of the effects of explanatory variables and jointly accommodates flexibly non-proportional differences via segmented relationships. The main appeal relies on ready availability of parameters, changepoints, and slopes, which may provide meaningful and intuitive information on the topic. Furthermore, specific linear constraints on the slopes may also be set to investigate particular patterns. We investigate the intervals between cohabitation and first childbirth and from first to second childbirth using individual data for Italian women from the Second National Survey on Fertility. The model provides insights into dramatic decrease of fertility experienced in Italy, in that it detects a 'common' tendency in delaying the onset of childbearing for the more recent cohorts and a 'specific' postponement strictly depending on the educational level and age at cohabitation.

Keywords: segmented regression; discrete-time hazard models; changepoints; parity progression; event occurrence data

1. Introduction

The sharp decline of the fertility pattern experienced by most industrialized countries over the last few decades is one of the main features of the so-called second demographic transition [23,38,42]. This decline is related to a complex set of socioeconomic and demographic factors that have determined important changes over time: a higher mean age of women at first childbearing, a different conception of marriage and motherhood, and a widespread acceptance and practice of divorce and extramarital cohabitation. However, some authors have pointed out that such decline may be explained, at least partially, by the so-called *postponement effect*, i.e. the tendency to delay the onset of childbearing [28]. This postponement effect naturally leads to higher mean ages at the beginning of cohabitation and consequently of the first birth. The issues of fertility

*Corresponding author. Email: vmuggeo@dssm.unipa.it

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760802552994
http://www.informaworld.com

decline, postponement, and related quantifications are crucial in the demographic research and have given rise to several hypotheses and theories on fertility patterns and reproductive behaviours of women [10,34,37].

Important findings coming from a statistical analysis of fertility data should be concerned with the identification of the effect of the socioeconomic factors, a quantification of the natural propensity in childbearing, and an assessment of possible variations over time [22,28]. While several interesting studies estimate the postponement effect and/or the recuperation effect using ecological data [6,37], few papers investigate postponement and socioeconomic factors using individual data [16,34]. Substantially, two different approaches can be recognized: an econometric approach in which birth spacing and timing is accommodated into a female labour supply model [18]; and a demographic approach in which birth spacing and timing is usually investigated through event history models [19,25]. The latter is consolidated in the literature, and typically Cox-type models with proportional hazards are used. However, to the best of our knowledge, simultaneous estimation of two or more birth spacing times with non-proportional hazards effect and explicit modelling of baseline hazards have not been presented for this sort of data.

We propose a discrete time-segmented regression model to investigate, in a novel way, fertility patterns using individual data. After adjusting for some factors in favouring parity, segmented relationships are used to measure the 'background' hazard, i.e. the baseline risk of entering motherhood not dependent on any specific factor. Furthermore, proper interaction terms are included to account for possible postponement and the 'cohort effect'. Roughly speaking, the 'cohort effect' may be considered the sum of the aforementioned time-varying lifestyles and cultural changes which, to some extent, can affect reproductive behaviours over time. We analyse the childbirth intervals studying the first and second parity spacing, namely time since cohabitation (for the first childbirth) and time since the first childbirth (for the second childbirth) for a sample of Italian women. All these facets are analysed in a unique framework.

The goal of this paper is to set up a statistical model suitable to fit parity progression data, to present the relevant estimating algorithm, and to explain how the model parameters may be interpreted. Although some interesting results are briefly discussed, substantive conclusions and their implications will not be emphasized.

The article is structured as follows. Section 2 deals with the statistical methods used, among which the segmented regression represents the novelty of our methodological framework; proper constraints on the slopes are also discussed to model changes over time of the baseline hazards. In Section 3 the data used for application are described in some detail, and the results are presented in Section 4. Finally, Section 4 includes a discussion on substantive implications of the findings.

2. Methodology

In this section, we illustrate the methods of the modelling framework based on discrete recurrent event setting. To facilitate understanding, we shall first present the various aspects separately and then comment on them at the end.

2.1 *The discrete event history model: a short introduction*

This paper considers settings in which the focus is on the intervals between cohabitation, first childbirth, and second childbirth. Each generic woman is characterized by (T_1, T_2) representing time between cohabitation and the first (possible) childbirth (T_1) and between the first childbirth and the second (possible) childbirth (T_2). Dummy variables mean possible censored observations and the vector of covariates denoted by $\boldsymbol{x}$ summarizes the information on the woman's profile; moreover non-informative censoring is assumed. In the present context, however, we are faced with raw measurements which naturally lead to discrete data. The response variable in event

history data for discrete-time processes is constituted by a series of binary outcomes indicating whether or not the event occurred at the observation point. To formalize, let T be the random variable denoting the time of an event occurrence, and let $\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x})$ be the relevant hazard given the covariate vector $\boldsymbol{x}$. A commonly used hazard model is

$$\frac{\lambda(t|\boldsymbol{x})}{1-\lambda(t|\boldsymbol{x})} = \frac{\lambda_0(t)}{1-\lambda_0(t)} \exp\{\boldsymbol{x}^T\boldsymbol{\beta}\}, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard and $\boldsymbol{\beta}$ is the regression parameter relevant to covariate vector $\boldsymbol{x}$. These settings, in both continuous or discrete view, are those met in classical survival/event history analysis, therefore the framework described in this paper may be applied in several fields (see discussion in Section 4).

Discrete-time models of type (1) were first proposed by Cox [9], and further useful references include [1,7,12,26,36]. Aranda-Ordaz [2] motivates discrete models by grouping and also presents a test to discriminate between additive and multiplicative models with complementary log–log link. If the complementary log–log link is used, as an alternative to the logit link, the regression parameters become log hazard rates rather than log odds ratios. In each case, however, the parameter estimates measure the effect sizes and, moreover, the differences between the two link functions are expected to be negligible, especially when the hazard is small [12].

Our framework extends model (1) to consider complexities in fertility studies; the regression equation accounting for the specific features of the analysis may be written as

$$\log\left\{\frac{\lambda_k(t|\boldsymbol{x}_{tkc})}{1-\lambda_k(t|\boldsymbol{x}_{tkc})}\right\} = f_{kc}(t) + \boldsymbol{x}_{tkc}^T\boldsymbol{\beta}, \tag{2}$$

where the subscripts t and c refer to possible time- and cohort-varying effects, and $k = 1, 2$ identifies the different strata, namely first and second childbirth. For instance, $f_{kc}(\cdot)$ is the logit of the stratum-specific segmented baseline hazard for cohort c. The subscript k emphasizes that baseline hazards and covariate effects are allowed to vary between first and the second childbirth. Model (2) is similar to Guo and Lin's marginal model [17], which assumes different baseline hazards; the Guo and Lin model may be considered a discrete version of the model proposed by Wei *et al.* [43] which, however, is not appropriate for recurrent events. On the contrary, our model can be considered as a discrete data version of Prentice–Williams–Peterson's continuous proportional hazard model [32] designed for recurrent events settings, where baseline hazards and covariates effect are allowed to vary among the strata. In the following sections, we shall explicitly discuss the different aspects of the model embedded in Equation (2).

2.2 *Baseline hazards via segmented linear parameterizations*

To illustrate the idea, we consider the simplified basic model (1) with a single covariate x and write it as

$$\text{logit}\,\lambda(t|x) = f(t) + \beta x,$$

where $f(\cdot) = \log\lambda_0(\cdot)/(1-\lambda_0(\cdot))$ is the baseline hazard whose shape may give indications on the fertility pattern variations with respect to time (time since cohabitation or since the first childbirth in our context). Note that $f(t)$ measures what has been termed 'natural' or 'background' fertility at time point t, because it is not affected by the effect of x which is accounted for by its parameter β.

Typically, the baseline hazards are non-linear, and this is particularly true in our context (see Figure 1). While several approaches could be used to account for nonlinearity, for instance, non-parametric smoothing [41] or regression splines [34], a function with meaningful parameters is desirable. We use a *segmented* (also called *piecewise-linear* or *broken-line*) parameterization.

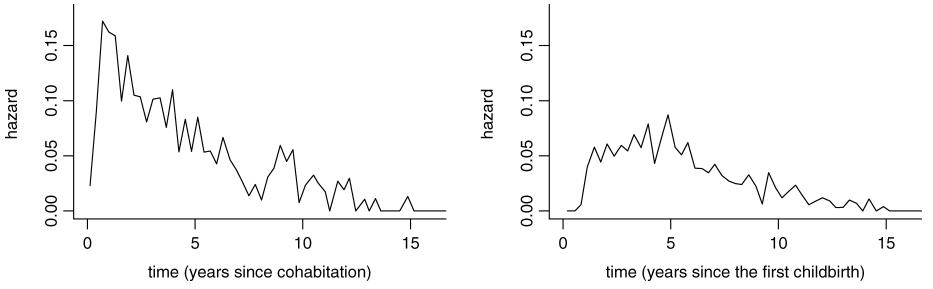

Figure 1. Raw estimates of the hazard functions for the events first and second childbirth.

The rationale is to approximate the non-linear function via two or more straight lines connected at unknown time-points, usually referred to as joinpoints or changepoints, where the slope changes [35, Chapter 9]. Thus the segmented baseline hazard $f(t)$ with s joinpoints can be expressed as

$$f(t) = \alpha_0 + \alpha_1 t + \delta_1(t - \psi_1)_+ + \cdots + \delta_s(t - \psi_s)_+, \tag{3}$$

where $(t - \psi_1)_+ = (t - \psi_1)$ if $(t - \psi_1) > 0$ and zero otherwise. This segmented function says that the time axis of event occurrence (in our case, time from cohabitation to the first childbirth or from the first childbirth to the second childbirth) is split into $s + 1$ intervals with different hazard patterns whose slopes change at the break-points. On the logit scale, the slope is given by $\{\alpha_1 + \delta_1 + \cdots + \delta_{s-1}\}$ at time $\psi_{s-1} < t \leq \psi_s$ and $\{\alpha_1 + \delta_1 + \cdots + \delta_{s-1} + \delta_s\}$ for $t > \psi_s$. Each slope measures the risk increase of having a child for unit time rise on the logit scale.

However, alternative techniques to model baseline hazards are available. A widespread approach uses the piece-wise constant model in which the time axis is split into a set of time intervals (for instance, months 0–12, 13–24, ...) with the hazard constant over each interval [25,36]: overall, the hazard is a step function. The possible disadvantages lie in the shape of the hazard which is sometimes unreasonable and to the potentially large number of parameters which have to be employed to ensure sufficient flexibility. Another approach relies on fully parametric assumptions [16] (for instance, Weibull or Gamma); however, parametric functions are often too inflexible or involve complicated formulas, and in the context of discrete data such methods are not tenable. Yet another approach is the non-parametric one, where the hazard is modelled by an unspecified smooth function [41]. While segmented modelling may lack some of the flexibility of smooth modelling, there is a benefit in the ease of interpretation: in fact the straightforwardness of the parameters used in Equation (3) makes it a convenient and attractive choice in that it allows the quantification of hazard functions in a very concise and objective way. On the other hand, the detection of changepoints and differences in slopes using non-parametric techniques is very subjective, difficult, and sometimes even impossible. It should also be noticed that while Equation (3) may resemble a spline formulation based on truncated power functions with several ψ's *arbitrarily* fixed *a priori* [34], our approach is completely different since both slopes (the α's and the δ's parameters) and the joinpoints (the ψ's) are parameters which will be estimated. Typically, unless the relationship is strongly non-linear, the number of changepoints involved rarely exceeds 2 and hence such an approach turns out to be parsimonious in practice. Moreover, likelihood-based criteria may be employed to select the appropriate number of changepoints (see Section 3.2).

In the next section, we shall discuss how such a small number of parameters may be employed to investigate possible variations over the whole time period in order to summarize the main features. Unfortunately, in spite of its apparent simplicity, function (3) is quite awkward since classical regularity conditions are not met in a likelihood-based approach [35, Chapter 9]. We shall present our estimation method in the next section.

2.3 *Overview of the estimation procedure*

To estimate model (2) with $f_{kc}(\cdot)$ representing segmented baseline hazards, standard likelihood methods cannot be applied; the changepoints ψ are non-linear and non-regular parameters and the log-likelihood is only piecewise differentiable [35, Chapter 9]. In a frequentist perspective, several authors have used grid-search algorithms to find the changepoint estimates and have performed estimation for the remaining model parameters given the selected joinpoint values [27]. However, grid-search type algorithms become prohibitive with several changepoints and, furthermore, standard errors for joinpoints are not generally available. To overcome such difficulties, we suggest a method which, in a joint framework, provides estimates and relevant standard errors for all the model parameters, including the joinpoints [29]. The method is iterative and relies on linear approximation of the generic 'hockey-stick' function $\delta(t-\psi)_+$ on the right-hand side of Equation (3). Omitting the indices and using a Taylor expansion, this non-linear function is replaced by two linear terms, i.e.

$$\delta(t-\psi)_+ \approx \delta(t-\tilde{\psi})_+ + \gamma I(t > \tilde{\psi})^-,$$

where $I(t > \tilde{\psi})^- = -1$ if $t > \tilde{\psi}$ and zero otherwise and $\tilde{\psi}$ is an approximate value of the joinpoint, which may be a starting guess or a value derived from the previous iteration. The variable $I(t > \tilde{\psi})^-$, through the associated parameter γ, accounts for estimation of ψ. The estimation procedure starts from an initial guess for the joinpoint and iterates towards a standard 'working' linear model including the two constructed terms $(t-\tilde{\psi})_+$ and $I(t > \tilde{\psi})^-$. At each iteration, a new estimate of the join-point is computed via $\hat{\psi} = \tilde{\psi} + \hat{\gamma}/\hat{\delta}$. Though we are skipping over some other details, it should be noticed that the algorithm does not depend on the number of joinpoints to be estimated and therefore it allows the estimation of several joinpoints without an excessive computational burden. This is not a negligible advantage in our context, because we are concerned with the estimation of segmented relationships having multiple joinpoints: grid-search methods based on the profile likelihoods become quite prohibitive when the number of change points increases.

Using linear approximations, the working linear terms reflecting a segmented baseline hazard in each of the different cohorts $c = 0, 1, 2, \ldots, C$ may be written as

$$\sum_{c=0}^{C} \delta_c (t-\tilde{\psi}_c)_+ + \sum_{c=0}^{C} \gamma_c I(t > \tilde{\psi}_c)^- \tag{4}$$

and the improvements of the cohort-specific joinpoint values are simply given by $\hat{\psi}_c = \tilde{\psi}_c + \hat{\gamma}_c/\hat{\delta}_c$.

In our case, we are interested in cohort-varying coefficients and therefore the above estimating algorithm has to be modified. By accounting for the linear cohort-varying effects, the generic difference-in-slopes parameter may be expressed as $\delta_c = \omega_0 + \omega_1 c$, where c represents the score for cohort and ω_1 quantifies how δ varies among the cohorts. The next section includes discussion about modelling cohort-varying parameters. Hence Equation (4) becomes

$$\begin{aligned}
&\sum_{c=0}^{C} [\omega_0 + \omega_1 c](t-\tilde{\psi}_c)_+ + \sum_{c=0}^{C} \gamma_c I(t > \tilde{\psi}_c)^- \\
&\quad = \omega_0 \left[\sum_{c=0}^{C} (t-\tilde{\psi}_c)_+ \right] + \omega_1 \left[\sum_{c=0}^{C} c(t-\tilde{\psi}_c)_+ \right] + \sum_{c=0}^{C} \gamma_c I(t > \tilde{\psi}_c)^-.
\end{aligned} \tag{5}$$

It is thus seen that estimation is performed in terms of (few) ω-parameters rather than the (several) δ-ones: in particular, at each step, the two different synthetic variables relevant to ω_0

and ω_1 are computed, the model is fitted, and the estimates of the difference-in-slopes parameters and jointpoints are obtained accordingly (see the Appendix for details).

2.4 *Explaining the model parameters*

Interpretation of the regression parameters for models like Equations (1) and (2) is generally straightforward and well established (see, for instance [7,36]). However, it is probably worth stressing the rationale of our modelling framework by focusing on the meaning and interpretability of the parameters and their connection with important demographic issues. As discussed previously, our modelling framework is aimed at answering to substantive demographic research questions,

(i) What is the effect of socioeconomic factors on fertility?
(ii) What is the 'natural' or background risk of childbirth? More specifically, what is the risk after adjusting for the effect of the explanatory variables?
(iii) Is there a postponement tendency?
(iv) Are there differences among cohorts, namely differences among women with the same covariates but born in different periods? From a substantive standpoint, recall the 'cohort' comprises all plain cultural changes considered as a whole, and may therefore reflect other non-specified variations in socio-demographic factors and life-style habits in general.

The answer to (i) is easily given by the regression coefficient β, which provides the (log) odds ratio of having a child at every time t. Of course, $\beta > 0$ suggests a higher risk of having a child.

To answer to question (ii), we remark that the baseline hazard $f(\cdot)$ accounts for the natural or background fertility versus the elapsed time (since cohabitation or since first childbirth). A segmented relationship is $f(t) = \alpha_0 + \alpha_1 t + \delta_1(t - \psi)_+$, with α_1 and $(\alpha_1 + \delta_1)$ measuring the change in risk of having a child for one-year increase when $t \leq \psi$ and $t > \psi$, respectively.

Question (iii) is answered by introducing a time-varying regression coefficient in the right hand side of the model, i.e. $\text{logit}\lambda(t|x) = f(t) + \beta(t)x$, where for instance $\beta(t) = \beta_0 + \beta_1 t$, say. If $\beta_0 < 0$ and $\beta_1 > 0$ this time-varying structure would suggest that the risk of entering motherhood at $t = 0$ (for instance at the beginning of cohabitation) is lower but over time the risk increases and eventually becomes positive. The aforementioned postponement tendency *due to the covariate*, is expressed by the coefficient $\beta(t) = \beta_0 + \beta_1 t$: $\beta_0 < 0$ refers to an initial low tendency of childbearing, and $\beta_1 > 0$ represents a subsequent recuperation effect . Whether such recuperation is complete or not will depend on the values of the estimated coefficients β_0 and β_1. It is clear that such a time-varying structure is fitted by including the interaction terms $\beta_0 x + \beta_1 tx$ in the model.

To answer question (iv) we must investigate whether some differences occurred across the cohorts, that is to study possible cohort-varying effects of covariates and of baseline hazards. A natural approach would be to construct product terms among the explanatory variables and a suitable 'cohort' variable. Although a categorical variable might seem a natural choice, we have chosen to use equally spaced zero-origin scores $c = 0, 1, 2, \ldots$ to identify the different cohorts. In the previous section, we presented a linear cohort-varying pattern for the difference-in-slopes parameter of the baseline hazard $\delta_c = \omega_0 + \omega_1 c$; likewise, we assume a similar linear-varying effect for the generic regression coefficient β and the left slope of the baseline hazards α_1. Overall, the cohort-varying structure for the model parameters can be expressed by

$$\beta_c = \theta_0 + \theta_1 c \tag{6}$$

for the generic regression parameter and

$$\alpha_{1c} = \tau_0 + \tau_1 c, \quad (\alpha_{1c} + \delta_{1c}) = (\tau_0 + \omega_0) + (\tau_1 + \omega_1)c, \tag{7}$$

for the left and right slope, respectively.

The interaction parameters θ_1, τ_1, and $(\tau_1 + \omega_1)$ may be employed to answer specifically to question (iv). For instance, θ_1 models the change in the effect of the factor x in moving from cohort c to cohort $c + 1$. Similarly for the baseline hazards: due to the parametrization (3) it turns out that the cohort-varying effects are modelled by τ_1 for the left slopes and by $(\tau_1 + \omega_1)$ for the right slopes. Inference on such quantities will provide evidence on changes of baseline hazards among cohorts. In particular, as right slopes are aimed at measuring risk of late childbirth, comparing them throughout the cohort may be useful to investigate on tendencies in postponing childbearing. It is important to emphasize that the postponement effect here considered, unlike the one expressed by the covariate–time interactions, is not ascribable to any particular factor: it can therefore be termed as 'pure' or 'common' postponement since it simply reflects women's behaviour unaffected by any other factor. Finally it is worth noting that the use of scores $c = 0, 1, \ldots$ provides meaningful monotone patterns and enables us to perform a test for trend throughout the cohorts. In this respect, the use of scores turns out to be quite an attractive choice and may possibly be supported by a sensitivity analysis to assess misfitting with respect to the unconstrained dummy variables.

In practice, linear cohort-varying parameters are estimated by the relevant products of covariates and scores (such as the products of covariates and time for the time-varying effects discussed above), while interaction terms for the parameters of the segmented baseline hazards are included through the algorithm presented in the previous section.

3. Application to the Inf2 survey data

3.1 *Data*

The data of the present work come from the Second National Survey on Fertility (*Inf2*), a retrospective study carried out on a sample of 4824 women and 1206 men. The statistical unit is an Italian resident woman, aged between 20 and 49, i.e. born between 1946 and 1975. Some further eligibility criteria have been set out for our study, i.e. women with twin births, those who had experienced their first childbirth before the beginning of cohabitation, and women who submitted incomplete and incorrect questionnaires have all been deleted from this study. Thus the final sample size is $n = 3164$, and only first and second childbirth have been considered, neglecting higher-order births which involve only 16% of our sample. Even if the data are quite old, we believe that the present analysis is noteworthy in that they refer to a decade when Italy experienced social and cultural changes which led the total fertility rate below 1.30, a fact that earned Italy the definition of 'lowest-low fertility country' [24]. Furthermore, the *Inf2* data is a well-known data set in the demographic literature with comprehensive information on the woman's life. Complete details on the sample design used for the survey can be found in [11]. The covariates considered in the present study are those typically investigated in most demographic studies. However, we are aware that further important information, not available in the survey, as the accessibility to child care facilities, should be included.

3.1.1 *Cohort of birth* (COHORT)

Women are classified into five cohorts following the usual 5-year grouping (see Table 1); due to their size, the last two cohorts have been collapsed into a single one. As previously argued,

Table 1. Descriptive statistics by parity: number of observations and median interval length in years (in parentheses).

Covariate	Parity 0	Parity 1	Parity 2+	Total
AGECO				
<24	171	1823 (1.17)	1292 (3.33)	1994
≥24	273	897 (1.42)	475 (3.33)	1170
SIBL				
1	52	182 (1.75)	80 (3.33)	234
2	170	733 (1.58)	405 (3.92)	903
≥3	222	1798 (1.17)	1276 (3.25)	2020
AREA				
North	257	1178 (1.75)	648 (3.83)	1435
Centre	94	596 (1.25)	385 (3.83)	690
South	93	946 (1.00)	734 (3.00)	1039
EDU				
First stage basic	32	623 (1.00)	519 (3.17)	655
Second stage basic	117	941 (1.25)	606 (3.50)	1058
Upper secondary	238	926 (1.58)	515 (3.42)	1164
Degree	57	229 (2.00)	126 (3.67)	286
OCC				
Worker	297	1225 (1.58)	681 (3.67)	1522
Non-worker	146	1495 (1.17)	1086 (3.25)	1641
COHORT				
1946–1950	41	135 (1.08)	454 (3.17)	630
1951–1955	36	150 (1.25)	474 (3.63)	660
1956–1960	63	178 (1.50)	410 (3.58)	651
1961–1965	92	264 (1.58)	311 (3.33)	667
1966–1975	212	226 (1.25)	118 (2.75)	556
Total	444	2720 (1.25)	1767 (3.33)	3164

the cohort may be thought to encompass changes in lifestyle, habits, and generic behaviours of society, namely it accounts for the transition of Italy from the rural and non-urban way of life to a new urban post-industrial society.

3.1.2 *Age at the beginning of the cohabitation* (AGECO)

This variable is obviously interrelated with the woman's reproductive period and to her future birth rate. To obtain a more readable estimate of the parameter, we have dichotomized it into 'less than 24 years' and '24 years and above'. This cut-off was set to correspond with the highest value of the hazard calculated for the 'beginning of cohabitation' event, as discussed in Ref. [40].

3.1.3 *Siblings in the family of origin of the woman* (SIBL)

It is natural to believe that reproductive strategies could be influenced by the nature of the mother's own family of origin. We have classified the woman's siblings into three categories: one, two, and more than two; the number of siblings of the man of the couple was not available for all the records.

3.1.4 *Geographic area* (AREA)

In Italy, residence is a proxy covering for the social and cultural status. Reproductive patterns have also been historically different between the South, Centre, and North. The two biggest islands

(Sardinia and Sicily) have always demonstrated different reproductive behaviours, so they cannot be included into a unique category. We have therefore chosen to aggregate Sardinia with the Centre and Sicily with the South, in consideration of historical and cultural issues.

3.1.5 *Educational attainment* (EDU)

We have set four levels, namely 'first stage basic', 'second stage basic', 'upper secondary', and 'degree' according to the International Standard Classification of Education of 1997.

3.1.6 *Occupational status* (OCC)

Two categories have been considered: students have been included in the category 'workers', while the unemployed and housekeepers form the the category of 'non-workers'. It should be recognized that this variable is actually a time-dependent variable which may affect the risk of having a child. However, the percentage of women who have changed their occupational status after the beginning of cohabitation is only 14.6%.

3.1.7 *Calendar time (year) of childbirth* (YEAR)

We have considered this 'contextual' variable as a proxy suitable to enclose some of the long-term changes that have affected fertility over time and which have not already been encompassed by the other covariates.

Table 1 shows some descriptive statistics of the analysed data: the number of observations by covariates and parity order. Figure 1 highlights the 'hazards' of having a child through the two-parity order. The raw hazard of first birth, as expected, rises sharply in the first year and then shows a descending trend till the tenth year of cohabitation. The raw hazard of the second birth is constantly lower and it apparently can be roughly split into three intervals. Similar plots per cohort (here not shown) maintain the same features described above, although there are differences in terms of intensity.

3.2 *Results*

The model has been fitted using the iterative procedure illustrated above and detailed in the Appendix. The robust Wald test based on the 'corrected' covariance matrix has been used to test for significance of coefficients and only statistically significant terms have been retained in the final model.

The cohort-specific baseline hazards were fitted via piecewise linear parametrizations with one and two joinpoints for the first and second childbirth, respectively. The number of changepoints has been first suggested by a visual inspection of the raw estimates reported in Figure 1, and then the appropriateness of the our choice was checked by comparing models with different numbers of joinpoints by means of the Bayesian information criterion (BIC) [39]. In the formula of the BIC, we have penalized for the number of events rather than for the number of the observations as recommended by Raftery [33] in the context of event history analysis. The BIC value for the selected model was always better (i.e. lower) than the ones from possible competitors; for instance, the difference in the BIC of the selected model with respect to the model with one changepoint for the baseline hazard of the second childbirth was -75.3.

Each cohort baseline hazard refers to an employed woman, cohabiting before her 24th birthday, with no siblings, living in the North of Italy, with a first stage basic level education.

The estimated segmented hazards are displayed in Figures 2 and 3, and estimates of these parameters are reported in Table 2.

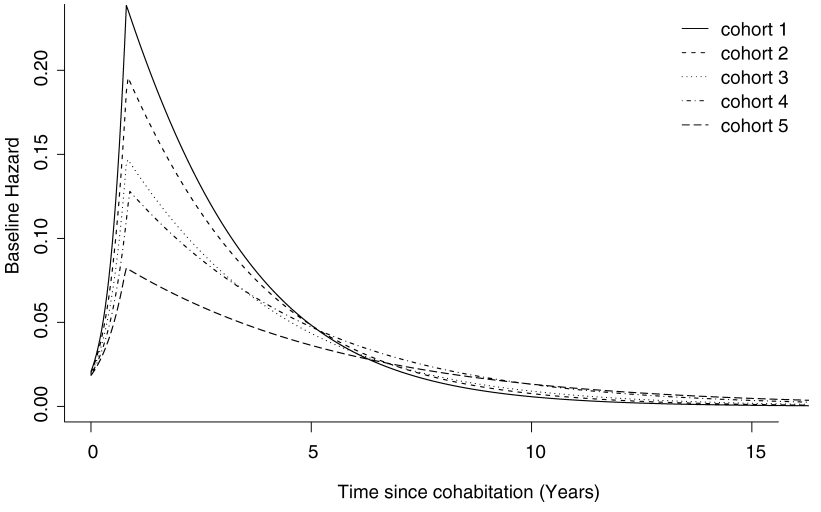

Figure 2. Estimated baseline hazards by cohort for the first childbirth.

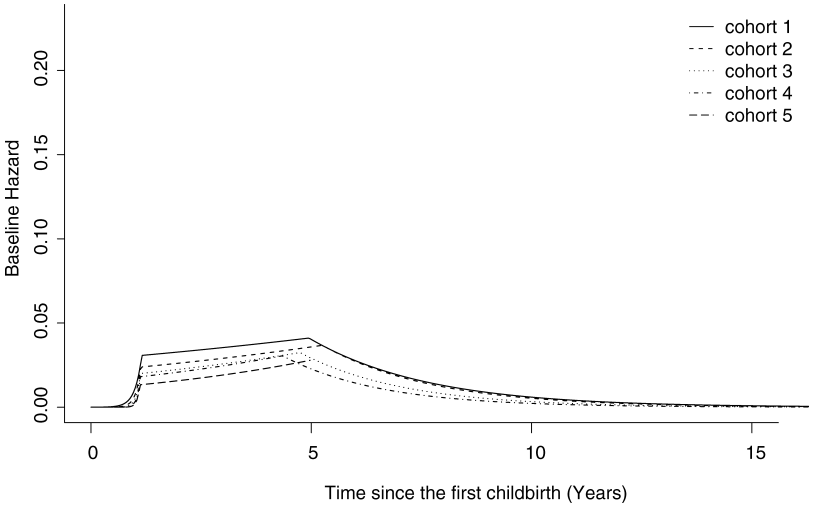

Figure 3. Estimated baseline hazards by cohort for the second childbirth.

For the first childbirth, the baseline hazards (BH1) show similar shapes with close changepoints (about 0.8 years) and sharply different slopes between the cohorts. In each cohort, the probability of having a child reaches its maximum at approximately 10 months after cohabitation with no notable difference among the cohorts. As expected, after the first year BH1 decreases for each cohort, but this reduction is less for the more recent cohorts. In fact for the first childbirth the annual baseline risk after the changepoint is (on the log scale) -0.427 for the first cohort, while it rises to -0.238 for the last one (see the 'right slope' row for the first childbirth in Table 2). Table 3 reports the estimated changes per cohort of the annual baseline risks, i.e. the point estimates with corresponding standard errors for parameters τ_1 and $\tau_1 + \omega_1$ as discussed in formula (7); the z-values in the last column, representing the empirical values of the Wald test, highlight the significant change of the right slopes for first childbirth.

Table 2. Parameter estimates (standard errors in parentheses) of the baseline hazards for the first and the second childbirth: slopes (on the log odds scale) and changepoints (in years).

	Cohort				
	1	2	3	4	5
First childbirth					
Left slope	3.372 (0.239)	2.999 (0.170)	2.627 (0.151)	2.254 (0.197)	1.882 (0.279)
Right slope	−0.427 (0.033)	−0.380 (0.029)	−0.332 (0.030)	−0.285 (0.035)	−0.238 (0.042)
ψ	0.806 (0.029)	0.819 (0.026)	0.796 (0.028)	0.851 (0.041)	0.773 (0.061)
Second childbirth					
Left slope	6.755 (1.280)	9.210 (1.408)	11.666 (2.522)	14.121 (3.843)	16.576 (5.217)
Middle slope	0.072 (0.043)	0.100 (0.031)	0.128 (0.027)	0.157 (0.034)	0.185 (0.048)
Right slope	−0.389 (0.034)	−0.309 (0.022)	−0.229 (0.025)	−0.149 (0.040)	−0.069 (0.058)
ψ_1	1.154 (0.037)	1.129 (0.019)	1.119 (0.017)	1.124 (0.021)	1.112 (0.023)
ψ_2	4.981 (0.361)	5.327 (0.332)	5.071 (0.371)	4.494 (0.525)	–

Table 3. Assessing cohort-varying slopes in the baseline hazards: estimated changes per cohort with corresponding robust standard errors and z-values.

			Wald test	
Baseline hazard slopes	Est.	SE	$\|z\|$-value	p-value
First childbirth				
Left slope	−0.374	0.106	3.53	0.0004
Right slope	0.045	0.012	3.75	0.0002
Second childbirth				
Left slope	2.447	1.419	1.72	0.085
Mid-slope	0.023	0.018	1.28	0.200
Right slope	0.074	0.021	3.52	0.0004

The baseline hazards for the second childbirth (BH2) are obviously much lower, with two estimated time-points where the risk values change abruptly. As for the first childbirth, the general shape appears substantially similar for all the cohorts, with a peak around the second changepoint (about 5 years). Once again, in the more recent cohorts the annual risk decreases more slowly after the second join-point. Across the cohorts, the annual risk of having the second child afterwards rises significatively from −0.389 in the first cohort up to −0.069 in the last cohort (see Tables 2 and 3). Note that for the last cohort, only the first joinpoint has been estimated because of lack of sufficient information necessary to estimate the segmented relationship in that time period.

Overall, segmented modelling emphasizes shapes of baseline hazards which are similar for all the cohorts. In each cohort, two and three different patterns of annual risk of motherhood have been detected for the first and the second childbirth, respectively. However, the annual risk of having a child (i.e. the slopes) varies among the cohorts, in particular, the risk of late births is higher in the more recent cohorts. This suggests that, controlling for the covariates included in the model, women in the more recent cohorts have delayed the onset of childbearing: this feature may be considered evidence of the postponement effect and subsequent partial recuperation. However, as shown by cumulative hazards (plots not shown), at the end of the reproductive period (15 years after cohabitation, say) the 'overall' fertility of the last cohorts is greatly reduced with respect to the older ones. This highlights how complete recuperation does not take place for 'baseline' women.

Table 4 reports results concerning the covariate effects: estimated log odds ratios and corresponding standard errors.

Table 4. Parameter estimates (on the log odds scale) of the effect of covariates on the first and the second childbirth.

	First childbirth		Second childbirth	
Terms	Est.	SE	Est.	SE
AGECO (ref: <24)				
≥24	0.043	0.155	−0.100	0.091
≥24 × c	−0.162	0.068	0.127	0.046
≥24 × log(time + 1)	−0.057	0.141	–	–
≥24 × log(time + 1) × c	0.158	0.065	–	–
SIBL (ref: 1)				
2	0.117	0.086	0.369	0.132
≥3	0.203	0.084	0.568	0.128
AREA (ref: North)				
Centre	0.239	0.056	0.189	0.066
South	0.563	0.055	0.759	0.062
EDU (ref: First stage basic)				
Second stage basic	−0.084	0.089	−0.144	0.065
Second stage basic × time	0.057	0.034	–	–
Upper secondary	−0.394	0.092	−0.168	0.071
Upper secondary × time	0.114	0.033	–	–
Degree	−0.602	0.135	0.014	0.113
Degree × time	0.160	0.038	–	–
OCC (ref: Worker)				
Non-worker	−0.084	0.075	0.289	0.055
Non-worker × c	0.151	0.033	–	–
Year (continuous)	−0.038	0.007	−0.038	0.007

Estimates for the siblings and area variables are undoubtedly according to expectations and they will not be discussed in detail: a higher number of siblings and living in South increase the risk of childbirth, as elsewhere reported [3]. Their effect is time-independent and does not vary among cohorts.

As argued in Section 2.4, possible postponement is modelled by interactions with time. Two factors, 'Age at cohabitation' (AGECO) and 'educational attainment' (EDU), show significant interactions with log(time + 1) and 'time', respectively.

AGECO is a variable that, in Italy, can be usually replaced by 'age at marriage', especially at the time of the survey. For the first birth, it appears that the effect of AGECO varies with the time since cohabitation and also differs among cohorts. In the first cohorts (first and second), the effect of AGECO is substantially time-independent and equal to zero. Thus, for women belonging to the first two cohorts (i.e. born in the interval 1946–1954) 'reproductive planning' appears to be already set as the time elapsed since cohabitation does not affect decision to have a first child. On the other hand, the most recent cohorts experience a negative risk at the beginning of cohabitation [i.e. when log(time + 1) = 0] followed by an increase as time since cohabitation elapses: in particular, the more recent the cohort, the steeper the 'slope'. Figure 4a displays graphically the time-cohort-varying effect of the AGECO variable. As discussed previously, this shape is coherent with the postponement effect which appears quite clear in the last cohorts. It, therefore, emerges that in the more recent cohorts, women beginning their cohabitation late delay the onset of childbearing.

With regard to the variable EDU, it appears that at the beginning of cohabitation (i.e. time = 0) the higher the educational attainment, the lower the hazard of having a first child: in particular, women with the highest level of education exhibit around half the risk compared to the lowest level ($\hat{\beta} = -0.602$). However, the risk itself is time-varying as it increases as time increases,

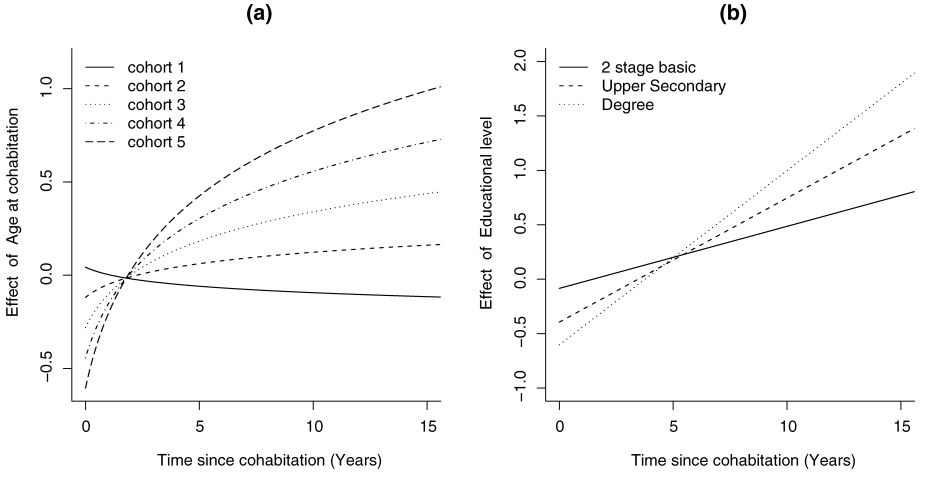

Figure 4. The effect of age at cohabitation (a), and of the educational level (b) on the first childbirth.

especially for the most educated groups (see Figure 4b). Like AGECO, this factor identifies a group of women characterized by the postponement effect which is even more enhanced for the more educated group. A possible reason for this is that women with high level of education may postpone their childbirth to try to consolidate their work position first.

As regards occupational status (OCC), the corresponding estimates underline that the effect of this factor on childbearing differs from cohort to cohort. This is also because a non-worker condition appears to be negligible in the older cohorts and increasingly more important in the later ones.

For the second childbirth event, no time-varying effect has been revealed, although the findings about the various factors are essentially unchanged: we only noted the effect of having a 'degree' as against a 'first stage basic' level of education is estimated with high uncertainty and is hence meaningless, while the role of AGECO suggests that women in the more recent cohorts beginning cohabitation late have a higher propensity to have a second child.

Finally, the model also includes the continuous variable year-of-birth (YEAR): our guess is that this variable covers long-term fertility trends: its estimated negative value accounts for the decline of natural fertility regarding first and second childbirth.

4. Discussion and conclusions

This paper presents a unified framework to model and quantify determinants of fertility patterns: through a discrete-time regression model, we have succeeded in quantifying the role of some demographic and socioeconomic factors and in obtaining estimates for the baseline hazards which can be considered the natural risk of having a child. Covariate effects and baseline hazards for both first and second childbirth are estimated simultaneously in the same framework, thus allowing the separation and identification of the contribution of specific variables and to obtain deeper insight into the reproductive behaviour of women. We also assess cohort- and time-varying effects via appropriate interaction terms in the linear predictor. The cohort-specific estimates of the baseline hazards have been obtained via segmented parameterizations. Segmented modelling may provide a sound model-based basis for making accurate and convincing assertions on the background hazards using a small number of meaningful parameters, slopes, and changepoints. Piecewise linear parameterizations have been successfully employed in different fields, including epidemiology [4,20], toxicology, and biology [5], where the term 'threshold' is often used instead

of changepoint or joinpoint. More specifically and in context of event history data, the framework of our model may be potentially useful to analyse any time-to-event data when the interest is not confined to the covariate effect, but also lies in the baseline hazard whose shape changes at some point in time [12,15,31]; time-to-divorce and time-to-development of depression in clinical psychology are further possible examples. Instead of treating the baseline hazard as nuisance in the spirit of Cox model [9], or instead of explicitly modelling it via parametric function or spline [12,13], a segmented parametrization could be both helpful and effective [12,31]. As pointed out by a referee, segmented modelling is under-used in practice, and a possible reason for this is the lack of specialized software. This is probably true in general although some software is now available, including the joinpoint software from the National Cancer Institute designed for the analysis of trends in cancer incidence data [20] at http://www.srab.cancer.gov/joinpoint/, and the more general R package segmented [30] at http://cran.r-project.org/web/packages/segmented/ which may be used within the framework of generalized linear models. Although the model presented in this paper requires some task, it may be easily implemented within any statistical environment with programming facilities, for instance, R, S-Plus, or SAS; R code is available on request from the first author.

From a substantive point of view, one of the most noticeable results concerns the simplicity of describing baseline hazards via segmented parametrizations. With respect to the first cohorts, the more recent ones exhibit a *common* tendency to delay the onset of childbearing for both children: in fact the 'natural' risk of (first and second) childbirth increases as we move from older to more recent cohorts after the estimated changepoints which appear to be quite similar among the cohorts. Apart from a common tendency towards postponement independent of the effect of any measured factors, more educated women emerge to have a higher propensity in delaying the onset of childbearing. The reason is probably due to the fact that for an educated woman the choice to have children is likely to follow the completion of her education and the attempt to establish herself on the labour market, something that inevitably delays the beginning of childbearing. This result is consistent with recent findings for other European countries [21], and it is in agreement with results from Rondinelli [34]. Although in the earlier cohorts, as expected, late cohabitation is associated with a lower risk of entering motherhood, in the more recent cohorts we found that women starting their cohabitation late also have an initial period of low fertility, but later exhibit a recuperation effect, at least for the first childbirth. For second childbirth, in fact, the effect of age-at-cohabitation does not vary with time, probably due to the fact that most women have already established their professional position before starting their childbearing career. Interestingly, it appears that postponement does not concern employed women: working class women reduce their fecundity but do not delay it, the reason being that only high wage levels (strongly dependent on education and age) are associated with delayed childbearing [34], while for a generic working woman a low wage might be an obstacle to childbearing. Although we have applied our model to the non-updated *Inf2* survey data, we believe that they include a mine of information on potential determinants of fertility which are still topical nowadays.

Acknowledgements

The authors would like to thank the two anonymous referees for their valuable comments which improved the quality and the clarity of the manuscript. The research of M. Attanasio has been funded by grants 'Fondi di Ateneo (ex 60%)' 2004-ORPA049434 and 2005-ORPA053442.

References

[1] P.D. Allison, *Discrete-time methods for the analysis of event history*, in *Sociological Methodology*, S. Leinhardt, ed., Jossey-Bass, San Francisco, 1982, pp. 61–98.

[2] F.J. Aranda-Ordaz, *An extension of the proportional hazards model for grouped data*, Biometrics 39 (1983), pp. 109–117.
[3] P. Astolfi, L. Liuzzi, and L. Zonta, *Trends in chlidbearing and stillbirth risk: heterogeneity among Italian regions*, Hum. Biol. 74 (2002), pp. 185–196.
[4] H. Bang, M. Mazumdara, and D. Spence, *Tutorial in biostatistics: analyzing associations between total plasma homocysteine and b vitamins using optimal categorization and segmented regression*, Neuroepidemiology 27 (2006), pp. 188–200.
[5] M. Betts, G. Forbes, and A. Diamond, *Thresholds in songbird occurrence in relation to landscape structure*, Conserv. Biol. 21 (2007), pp. 1046–1058.
[6] J. Bongaarts, *The end of the fertility transition in the developed world*, Popul. Dev. Rev. 28 (2002), pp. 419–443.
[7] J.M. Box-Steffensmeier and B.S. Jones, *Event History Modelling: A Guide for Social Scientists*, Cambridge University Press, Cambridge, UK, 2004.
[8] C. Brown, *On the use of indicator variables for studying the time dependence of parameters in a response time models*, Biometrics 31 (1975), pp. 863–872.
[9] D. Cox, *Regression models and life tables* (*with discussion*), J. Roy. Statist. Soc., Ser. B 34 (1972), pp. 187–220.
[10] G. Dalla Zuanna, *The banquet of æolus: a familistic interpretation of italy's lowest low fertility*, Demogr. Res. 4 (2001), pp. 131–162.
[11] P. De Sandre *et al.*, *Matrimonio e figli: tra rinvio e rinuncia*, il Mulino, Bologna, 1997.
[12] B. Efron, *Logistic regression, survival analysis, and the Kaplan-Meier curve*, J. Amer. Statist. Assoc. 83 (1988), pp. 414–425.
[13] J. Etezadi-Amoli and A. Ciampi, *Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function*, Biometrics 43 (1987), pp. 181–192.
[14] C.P. Farrington, *Interval censored survival data: a generalized linear modelling approach*, Stat. Med. 15 (1996), pp. 283–292.
[15] F. Gao, A.K. Manatunga, and S. Chen, *Non-parametric estimation for baseline hazards function and covariate effects with time-dependent covariates*, Stat. Med. 26 (2007), pp. 857–868.
[16] G. Ghilagaber *et al.*, *The use of flexible parametric duration functions in modelling the tempo of fertility: applications to the analysis of birth intervals in rural China*, Popul. Rev. 44 (2005), pp. 11–35.
[17] S.W. Guo and D.Y. Lin, *Regression analysis of multivariate grouped survival data*, Biometrics 50 (1994), pp. 632–639.
[18] J.J. Heckman and J.R. Walker, *The relationship between wages and income and the timing and spacing of births: Evidence from Swedish longitudinal data*, Econometrica 58 (1990), pp. 1411–1441.
[19] A.S. Kalwij, *The effects of female employment status on the presence and number of children*, J. Popul. Econ. 13 (2000), pp. 221–239.
[20] H. Kim *et al.*, *Permutation tests for joinpoint regression with applications to cancer rates*, Stat. Med. 19 (2000), pp. 335–351.
[21] S. Klasen and A. Launov, *Analysis of the determinants of fertility decline in the Czech Republic*, J. Popul. Econ. 19 (2006), pp. 25–54.
[22] T. Kögel, *Did the association between fertility and female employment within OECD countries really change its sign*? J. Popul. Econ. 17 (2004), pp. 45–65.
[23] H. Kohler, J. Ortega, and F. Billari, *Towards a theory of lowest-low fertility*, MPIDR Working Paper S WP-2001-032, Max Planck Institute for Demographic Research, Rostock, Germany 2001, pp. 1–57.
[24] H. Kohler, J. Ortega, and F. Billari, *The emergence of lowest-low fertility in Europe during the 1990s*, Popul. Dev. Rev. 28 (2002), pp. 641–680.
[25] K. Köppen, *Second births in western Germany and France*, Demogr. Res. 14 (2006), pp. 295–330.
[26] J. Lawless, *Statistical Models and Methods for Lifetime Data*, 2nd ed., Wiley, New York, 2002.
[27] P.M. Lerman, *Fitting segmented regression models by grid search*, Appl. Statist. 29 (1980), pp. 77–84.
[28] S.P. Martin, *Diverging fertility among U.S. women who delay childbearing past age 30*, Demography 37 (2000), pp. 523–533.
[29] V.M.R. Muggeo, *Estimating regression models with unknown break-points*, Stat. Med. 22 (2003), pp. 3055–3071.
[30] V.M.R. Muggeo, *Segmented: an R software to fit regression models with broken-line relationships*, R News 8(1) (2008), pp. 20–25.
[31] A.A. Noura and K.L.Q. Read, *Proportional hazards changepoint models in survival analysis*, Appl. Statist. 39 (1990), pp. 241–253.
[32] R.L. Prentice, B.J. Williams, and A.V. Peterson, *On the regression analysis of multivariate failure time data*, Biometrics 68 (1981), pp. 373–379.
[33] A.E. Raftery, *Bayesian model selection in social research*, Soc. Methodol. 25 (1995), pp. 111–163.
[34] C. Rondinelli, A. Aassve, and F. Billari, *Socio-economic differences in postponement and recuperation of fertility in Italy: results from a multi-spell random effect model*, ISER Working Paper, Institute for Social and Economic Research, University of Essex, Colchester, UK, 2006.

[35] G.A.F. Seber and C.J. Wild, *Nonlinear regression*, Wiley, New York, 1989.
[36] J.D. Singer and J.B. Willett, *Applied Longitudinal Data Analysis*, Oxford University Press, New York, 2003.
[37] T. Sobotka, *Is lowest-low fertility in Europe explained by the postponement of childbearing?* Popul. Dev. Rev. 30 (2004), pp. 195–220.
[38] J. Surkyn and R. Lesthaeghe, *Value orientations and the second demographic transition (SDT) in Northern, Western and Southern Europe: an update*, Demogr. Res. Special Collection 3 (2004), pp. 45–86.
[39] R. Tiwari *et al., Bayesian model selection for join point regression with application to age-adjusted cancer rates*, Appl. Statist. 54 (2005), pp. 919–939.
[40] P.M. Todd, F.C. Billari, and J. Simão, *Aggregate age-at-marriage patterns from individual mate- search heuristics*, Demography 42 (2005), pp. 559–574.
[41] G. Tutz and L. Pritscher, *Nonparametric estimation of discrete hazard functions*, Lifetime Data Anal. 2 (1996), pp. 291–308.
[42] D. van de Kaa, *Europe's second demographic transition*, Popul. Bull. 42 (1987), pp. 1–57.
[43] L.J. Wei, D.Y. Lin, and L. Weissfeld, *Regression analysis of multivariate incomplete failure time data by modelling marginal distributions*, J. Amer. Statist. Assoc. 84 (1989), pp. 1065–1073.

Appendix: The estimating algorithm

Here we detail the estimating algorithmic steps used to fit the regression model applied in this work. We use superscripts to distinguish between quantities corresponding to first and second childbirth. Thus the joinpoints of the baseline hazards are $\tilde{\psi}_c^1$ for the first event and $\tilde{\psi}_{1c}^2$ and $\tilde{\psi}_{2c}^2$ for the second one; the 'time since cohabitation' and 'time since first childbirth' are written as t^1 and t^2, and similarly for the other variables and parameters.

1. Fix starting values for the joinpoints: $\tilde{\psi}_c^1$, $\tilde{\psi}_{1c}^2$, and $\tilde{\psi}_{2c}^2$ for $c = 0, \ldots, 4$.
2. Compute the following constructed variables:
   - $\mathcal{V}_c^1 = I(t^1 > \tilde{\psi}_c^1)^-$, $\mathcal{V}_{1c}^2 = I(t^2 > \tilde{\psi}_{1c}^2)^-$, and $\mathcal{V}_{2c}^2 = I(t^2 > \tilde{\psi}_{2c}^2)^-$ where $c = 0, \ldots, 4$,
   - $\mathcal{U}^1 = \sum_c (t^1 - \tilde{\psi}_c^1)_+$ and $\mathcal{W}^1 = \sum_c c(t^1 - \tilde{\psi}_c^1)_+$, $\mathcal{U}_1^2 = \sum_c (t^2 - \tilde{\psi}_{1c}^2)_+$ and $\mathcal{W}_1^2 = \sum_c c(t^2 - \tilde{\psi}_{1c}^2)_+, \mathcal{U}_2^2 = \sum_c (t^2 - \tilde{\psi}_{2c}^2)_+$ and $\mathcal{W}_2^2 = \sum_c c(t^2 - \tilde{\psi}_{2c}^2)_+$,

   and fit the model with linear predictor including the additional terms:

$$\omega_0^1 \mathcal{U}^1 + \omega_1^1 \mathcal{W}^1 + \omega_{10}^2 \mathcal{U}_1^2 + \omega_{11}^2 \mathcal{W}_2^2 + \omega_{20}^2 \mathcal{U}_2^2 + \omega_{21}^2 \mathcal{W}_2^2 + \sum_c \gamma_c^1 \mathcal{V}_c^1 + \sum_c \gamma_{1c}^2 \mathcal{V}_{1c}^2 + \sum_c \gamma_{2c}^2 \mathcal{V}_{2c}^2.$$

3. Compute the difference-in-slope parameter estimates

$$\hat{\delta}_c^1 = \hat{\omega}_0^1 + \hat{\omega}_1^1 c, \quad \hat{\delta}_{1c}^2 = \hat{\omega}_{10}^2 + \hat{\omega}_{11}^2 c, \quad \hat{\delta}_{2c}^2 = \hat{\omega}_{20}^2 + \hat{\omega}_{21}^2 c$$

   and update the estimates of the joinpoints via

$$\hat{\psi}_c^1 = \tilde{\psi}_c^1 + \frac{\hat{\gamma}_c^1}{\hat{\delta}_c^1}, \quad \hat{\psi}_{1c}^2 = \tilde{\psi}_{1c}^2 + \frac{\hat{\gamma}_{1c}^2}{\hat{\delta}_{1c}^2}, \quad \hat{\psi}_{2c}^2 = \tilde{\psi}_{2c}^2 + \frac{\hat{\gamma}_{2c}^2}{\hat{\delta}_{2c}^2},$$

   for $c = 0, \ldots, 4$.
4. Repeat steps (2) and (3). up to some convergence criterion is met.

In this way, at each step, a 'working' model is fitted by maximizing the Bernoulli log-likelihood $\log \prod_i^n \prod_j^{J_i} \lambda(t_j|\boldsymbol{x}_i)^{y_{ij}} (1 - \lambda(t_j|\boldsymbol{x}_i))^{y_{ij}}$, where $\lambda(t_j|\boldsymbol{x}_i)$ is given by step (1), and $(y_{i1}, y_{i2}, \ldots, y_{iJ_i})^{\mathrm{T}}$ is the J_i-dimensional binary response vector for woman i [8,12,14].

We have applied this algorithm to fit the *Inf2* data, and convergence was achieved in reasonable time after a few iterations. As starting values we set $\tilde{\psi}_c^1 = 0.8$, $\tilde{\psi}_{1c}^2 = 1.5$, and $\tilde{\psi}_{2c}^2 = 4$ for $c = 0, \ldots, 4$. However, due to the rather clear-cut segmented relationships, different initial guesses with reasonable values did not affect the final estimates.