Identification Strategy

Tuna Dökmeci

12. Dezember 2019

1 Main Idea

Event history, or survival analysis concerns itself with the occurrence and the timing of events. Event occurrence is an individuals transition from one state to another; from being unemployed to employed, from not being pregnant to being pregnant, from being married to being divorced and so on. As such, it is suitable for analyses that concern themselves with whether and when some people experience the events, why the timing and the occurrence differ from individual to individual. Unlike in other statistical applications, the event history analysis does not take means of outcome and standard deviations as the variable to be explained, but rather probabilities of an individual experiencing a certain event at a given time in their life course.

2 Censoring

Event analysis is especially used to analyze data that suffers from censorship, and especially non-informative and right censorship. Censorship means that we do not have past or future information on an individuals event history. The term left censorship is used to define a situation where for some or all of the units, we do not have information on whether or not the event has occurred in the past. This is a problem much more difficult to deal with than right censorship, and is also more rare (it might however be the case for our data, as I explain below). Right censorship is the opposite case: we do not have the data on whether or not an individual experiences an event after the data collection period has ended. For instance, we do not see in our data whether a family had a child in the following months after the experiment has finished. If non-informative, meaning no household is more likely than the others to suffer from censorship, this can be dealt with event history, although not perfectly (to be explained further).

3 Econometric Model

Event history analysis was first developed for events that are not repeatable. In this context, there are two outcomes of interest: the probability that an individual experiences an event at a given time given that she has never experienced the event before, called the hazard rate, and the survival probability; the probability that an individual has not yet experienced the event by a given time. By their nature, these are probabilities whose ranges are between 0 and 1. The hazard rate can decrease and increase with time; for example the probability of giving birth to first child first increases, then decreases with age. Survival probability, in this context the probability of never having given birth to a child is (not strictly) decreasing with time.

More recently, methods have been developed to analyze events that happen multiple times in a lifetime. Thinking of childbirth, one way to approach this question would be to apply the classic method and only focus on the birth of the first child. This approach however leads to a loss of information.

Extensions of the baseline model exist to accommodate events that can happen multiple times, namely Andersen-Gill (AG), Prentice-Williams and Peterson (PWP), Wei, Lin and Weissfeld (WLW), and frailty models with random effects. Another analysis models the mean number of events or their occurrence rate.

The AG model uses a common baseline hazard function for all events and estimates coefficients for other factors of interest. The model is formulated in terms of time past between events. The data is accordingly structured in events. An individual has one observation if she does not experience the event, two if it happens once and so on. Each observation is thus an episode whose beginning time is the end of one event and end time is the start of the next one. In each row, there are variables indicating the starting time, ending time as well as the time that passes. The model is based on the assumption that the increment between each event is uncorrelated with each other, given covariates. So a person has the same probability of experiencing an event regardless of the past events. This is a rather restrictive assumption, especially in context of childbearing.

The PWP model allows for the effect of co-variates to vary from one event to another. It stratifies events by order. At the beginning of time, age 15 for us, all individuals are in first stratum, at risk to experience the first event. Once an individual experiences an event for the first time, she moves to the second stratum. In other words, an individual can only be in the jth strartum if she has experienced event j-1 times in the past. The data is organized in a similar way as in AG model.

The simplest way to handle this is to treat events separately but this approach has proven

to produce incorrect estimates (Kravdal 2001). Another approach is to treat the events jointly. Because the duration of episodes, the time that passes between two events, is likely to be correlated with each other, this approach includes unit-specific time-invariant random effects. These models are called frailty models, and are often used in fertility analyses.

The frailty model can be written as:

$$log(h_{tij)} = \alpha(t) + \beta * x_{tij} + u_i$$

where $\alpha(t)$ is the estimated baseline hazard rate as a function of age, u_i individual specific random effect that should account for unobserved heterogeneity that can affect fertility decisions as well as the intervals between births for one woman, and x_{tij} time-varying covariates such as income.

(Steele, 2004. http://eprints.ncrm.ac.uk/88/1/MethodsReviewPaperNCRM-004.pdf)

There are further similar analyses that deal with repeated events. Van Hook and Altman (2013) take the example of multiple births and present three models. The baseline model is an additive one, and it is made more flexible with interactions.

Additive model:

$$logit(b_x = 1) = a + j_x + A_x B_1 + CB_2 + C_x B_3$$

b_x is a binary outcome that takes the value 1 when a child is born at a given age, j_x is the birth interval that takes the value 1 if no child has been born, 2 if the first child is born, 3 after the second child is born and so on. A_x is a set of age dummies. Using these dummies in a linear way should be similar to estimating a step hazard function and we allow for non-linearity between fertility and age. B_2 and B_3 are respectively time-fixed and time-varying control variables.

The additive model is restrictive in the assumption that the hazard function is identical across parities. This is a strong assumption that can be relaxed by allowing for interactions between the interval and age dummies. That way, we can have the hazard change depending on the rank of the child that is born (first, second etc.)

4 Data Structure

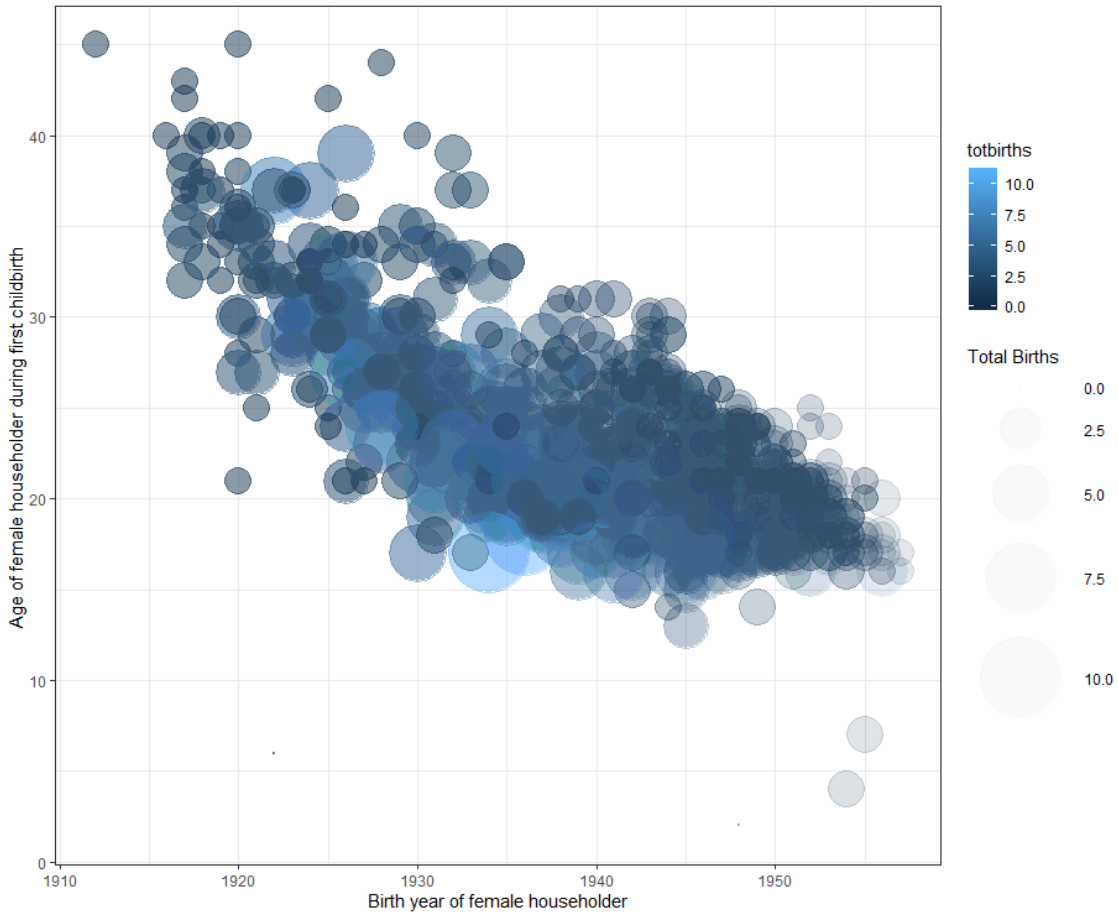
Frailty models where a random effect is incorporated require a person-period level data. For each individual, we have a number of observations equal to the number of periods we observe the person. In event history data, we start with a common "beginning of time" where no individual has ever experienced the event before. In studies analyzing fertility, this can be the age where a woman is considered to be fertile. Because this age is usually 15, and because in our sample earliest recorded birth age is 16, I organized the data with each woman having 36 observations from the age 15 to 50. Women who are censored, so who were younger than 50 in 1977 have observations until the age they are in in 1977. They are the right censored observations. The data we have at the end is organized in years, and has observation for each female householder from the age 15 to the end of 1977. You can look at file datapersonperiod.rds, which is on Github and our server under data to see the data structure.

Using the data on birth dates of all individuals in the households, we can then reconstruct the birth history for each female in our sample. The variable event takes value of 1 if in a given year, the female householder has given birth and 0 if not. Variable j represents the episode: it is 1 if the individual has never had a child, and 2 between the first and second children, 3 between second and third children and so on. Variable spell-time counts the years until the next event happens so that we can observe the different lengths of event occurrence. Furthermore, for each age, we have a dummy so that we can estimate the hazard rate as a nonlinear function of age. Finally, we have the variables treated and experiment, treated taking the value of 1 if the household received payments and the experiment 1 if the year is 1975, 1976 or 1977. The coefficient of interest is that of the interaction of these two variables.

The second dataset is adapted to AG and PWP models, and can be found under the name dataag.rds.

It is likely that our sample suffers also from left censorship. The following plot shows the age of the female householder when she had her first child by her own birth year. The size and the color of bubbles indicate how many children the female householder has in total. We see that females born between 1910 and 1930 have their children at a much later age and have less children in total compared to younger female householders. It is very unlikely that this is the case. More likely is that the children of the older householders have already moved out and are not counted in the household. This would create a problem regardless of if we are using an event history analysis or the logit regression with cross section as previously, as also the number of children in some households is not accurately measured.

Tabelle 1: Age at the First Birth by Birth Years



In the family composition data, there are two variables that could help us. V7862 gives information on how many children of the householder are not living at home and V7962 asks the age of the oldest child not at home. By using this information, we can correct the number of children the householder has as well as the age she had the first child, but we would not be able to construct a full event history with exact ages where the individual had each child. (Should we then use the total number of children as a control instead of a random effect, or what else could we do? Or else Andersen-Gill could be used maybe)

Literatur

[1] Leila DAF Amorim and Jianwen Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, 44(1):324–333, 2015.

[2] Nancy Brandon Tuma, Michael T. Hannan, and Lyle P. Groeneveld. Dynamic analysis of event histories. *American Journal of Sociology*, 84(4):820–854, 1979.

[3] Jennifer Van Hook and Claire E Altman. Using discrete-time event history fertility models to simulate total fertility rates and other fertility measures. *Population research and policy review*, 32(4):585–610, 2013.