

Causal inference for duration models

An alternative to hazard regression

Stefano Mazzuco
Department of Statistics, University of Padova
Via Cesare Battisti, 241
35121, Padova, Italy
email: mazzuco@stat.unipd.it

October 25, 2005

Abstract

This paper concerns possible approaches to study the impact of an event (treatment) on the duration of an episode. Bivariate regression models with unobserved heterogeneity are usually applied to this problems though they rely on strong assumption which violation can heavily influence the final results. An alternative procedure uses a matching method where survival functions are computed conditional on a function of observed variables thus eliminating the selection based on these variables. However also the matching solution relies on a strong assumption, namely the conditional independence assumption. Therefore we can depict two situations: if we are reasonably confident that the effect of the unobserved factors is negligible, then we can use a matching estimator. If our data is not reach enough and we suspect that the effect of the unobserved factors is relevant, but we observe a valid instrument we can apply the same approach to a IV estimator.

Preliminary version. Please do not quote, nor cite

1 Introduction

Let us consider a sample of n units in an observational framework. We are interested in estimating the impact of one event on the duration of an episode. For instance, we could be interested in estimating the impact of premarital cohabitation on the duration of marriage. This issue has usually been investigated by means of hazard regression models (Lillard et al., 1995), which however usually assumes the time-invariance of impacts. Thus speculations are made on the base on a supposed proportional shift of hazard rate due to the event. In facts, it is likely that the impact of one event is time-varying. Furthermore the proportionality assumption is not the only limitation of hazard regression models:

when the covariate of interest is endogenous with the episode duration, a multivariate model need to be implemented in order to take into account the possible endogeneity (see, for example, for one application, Lillard & Panis, 1998). However, such a model poses new problems in the treatment effect identification. First, since a multivariate random component needs to be specified, it may be the case that results are sensitive of the chosen distribution of the random component (Heckman & Singer, 1982, 1984; Lindeboom & Van Den Berg, 1994). Moreover, in order to identify the model some exclusion-restriction assumption are needed, assumptions that cannot be tested. Most of all, inclusion of time-dependent covariates, can bias estimates of treatment effect since they may cancel the indirect effects.

An alternative approach can be implemented moving to the model suggested by Rubin (1977). This model allows for identification of the survival function (or equivalently the hazard function) of focal episode for those who experienced the event and the survival function for those who do not netting out all spurious effects of other confounding variables. From the Rubin model, analysts derived methods for estimating the impact of one treatment on an outcome, based on matching, instrumental variables, or regression discontinuity designs.

Matching procedures, as well as instrumental variable and regression discontinuity design methods, can provide unbiased estimates of average impacts of “treatment on treated” given certain assumptions are satisfied. However Heckman et al. (1997) showed that with no additional assumption, it is possible to identify the whole distribution of the outcome for treated and control group.

This work aims to take advantage of this feature of matching procedure for deriving a time-varying estimate of the impact of one event on the duration of an episode. This method overcome some limitations of hazard regression: firstly, time pattern of the impact of the event needs not to be specified, secondly the method accounts for confounders only *before* the occurrence of the event. Finally, no exclusion-restriction assumption is needed. Together with these benefits, there is one drawback: matching accounts for selection bias due to observed variables but not on the selection bias due to unobserved variables. To what extent this pitfall affects the results is strongly dependent on richness of data used. The more variables we are able to control for the closer to zero is the bias due to unobserved variables.

The structure of the paper is as follows: section 2 describes the matching procedure comparing it with the hazard regression model, using the application made by Lillard et al. (1995) who aimed to estimate the impact of informal cohabitation on the hazard of marriage dissolution. Section 3 makes a comparison of the results of the two methods using panel data from National Longitudinal Study of Youth 1979 (NLSY79), whereas section 4 concludes.

2 Identification of causal effect: two different strategies

The fundamental problem of causal inference (Holland, 1986) arises whenever we want to estimate the impact of one event on an outcome. In order to formalize the problem let D_i be the treatment variable taking the value 1 if individual i received the treatment and 0 otherwise. Be T the outcome that can be *causally* affected by D . Then let us write T_1 as the outcome of treated individuals and T_0 as the outcome of untreated individuals. One parameter of interests for researcher is the ‘‘Average Treatment Effect’’ (ATE) that is

$$\begin{aligned} E(T_1) - E(T_0) = \\ E(T_1|D = 1) \cdot P(D = 1) + E(T_1|D = 0) \cdot P(D = 0) - \\ E(T_0|D = 1) \cdot P(D = 1) + E(T_0|D = 0) \cdot P(D = 0). \end{aligned} \quad (1)$$

Identification of ATE is not possible as we cannot estimate $E(T_1|D = 0)$ nor $E(T_0|D = 1)$ and the estimate based on the comparison between $E(T_1|D = 1)$ and $E(T_0|D = 0)$ can be seriously biased as it is possible that the population of treated differs from the comparison group not only in terms of treatment status but even in terms of other characteristics. Generally, scholars are more interested in identifying the so-called ‘‘average treatment effect on treated’’ (ATET) that is

$$E(T_1|D = 1) - E(T_0|D = 1) \quad (2)$$

(2) reduces the identification problem since there is only $E(T_0|D = 1)$ that is not identified. Moreover, the ATET is still an interesting parameter giving the average change of T for treated individuals due to the event D only. However identification of $E(T_0|D = 1)$ needs further assumption on the selection process.

In some cases variables T_1 and T_0 may represent duration of an episode. For instance, T_1 could be the marriage duration in case marriage was preceded by an informal cohabitation ($D = 1$) and T_0 the marriage duration without premarital cohabitation ($D = 0$). In this case one may be interested in identifying the effect of D not directly on T but on its survival function $S_T(t)$ or its hazard function $h_T(t)$. Many researchers (see, for instance, Teachman & Polonko, 1990; DeMaris & MacDonald, 1992) tried to single out the effect of cohabitation on marriage stability but once again we face the selectivity issues. Bennett et al. (1988) argue that couples who cohabited before marriage are selected in terms of propensity to divorce. In this case the difference between the survival function of cohabitators $S_{T_1}(t|D = 1)$ and the survival function of non-cohabitators $S_{T_0}(t|D = 0)$ is biased by this form of selection.

One common solution invoked in these cases is the use of simultaneous equations with bivariate normally distributed heterogeneity components to control for the unobserved variables (Lillard, 1993). In this way one should account for former unobserved propensity to divorce and identify the true effect of cohabitation on marriage stability.

This is the route followed by Lillard et al. (1995), who find that, after

having controlled for observed covariates and unobserved heterogeneity, the effect of cohabitation on the risk of marriage dissolution is not significant. Next I give details of this approach outlining the assumption on which it relies and its possible drawbacks: firstly, in simultaneous hazard regression it is often required to make some likely exclusion-restriction hypothesis for identification purposes, which are sometimes difficult to find. Moreover the unobserved component is assumed to be drawn from a bivariate normal distribution, violation of such an assumption can influence the estimate of treatment effect. Besides in the hazard regression model it is useful to use some time varying covariates but in this way one can net out even the indirect effect of treatment. Finally, the treatment effect is assumed to be time-invariant.

2.1 The regression solution

The practical problem is estimating the effect of premarital cohabitation on the risk of union dissolution. The treatment variable is premarital cohabitation and it is modeled via a probit function

$$\Phi((2D - 1)(\alpha_0 + \alpha_1 X_m^c + \xi)) \quad (3)$$

where D is the treatment status (equal to one if the couple cohabited before marriage and zero otherwise), X_m^c is a set of covariates, and ξ is a individual-specific error term, representing the unobserved heterogeneity in the propensity to cohabit prior marriage. Φ represents the normal distribution function.

Differently, marital dissolution is modeled with a regression function on the log-hazard, assuming that the logarithm of the hazard function of marriage dissolution can be defined as

$$\log h_m^d(t) = \log h_{0m} + \delta_1 X_m^d + \delta_2 D + \varepsilon \quad (4)$$

where h_{0m} is the baseline hazard function (estimated by means of spline functions), X_m^d is a set of time-varying covariates (not coincident with X_m^c), and ε is a normally distributed heterogeneity component.

Then, it assumed that (ξ, ε) are drawn from a bivariate normal distribution. Estimates of parameters α and δ are based on marginal joint probabilities of premarital cohabitation and marital dissolution, maximizing the joint marginal likelihood

$$L = \int_{\varepsilon} \int_{\xi} \phi \left(\frac{\frac{\varepsilon}{\sigma_{\varepsilon}}, \frac{\xi}{\sigma_{\xi}} | \rho_{\varepsilon\xi}}{\sigma_{\varepsilon}\sigma_{\xi}} \right) \prod_{m=1}^M S_m^d(t_m, X_m^d(t), \varepsilon) \cdot [h_m^d(t_m, \varepsilon)]^{C_m} \cdot \Phi((D - 1)(\alpha_0 + \alpha_1 X_m^c + \xi)) d\xi d\varepsilon \quad (5)$$

where $S_m^d(t_m, X_m^d(t), \varepsilon)$ is the survival function of marriage m , depending on covariates $X_m^d(t)$ and on the unobserved heterogeneity ε , and C_m is a dummy variable for right censorship (i.e. it is equal to one if the episode censored and 0 otherwise). The estimate of δ_2 in the (5) gives the proportional effect of premarital cohabitation on the risk of marriage dissolution, netted out of former attitude toward divorce.

The approach described above suffers, anyway, from several limitations

that can severely bias parameter estimation, then even δ_2 that is the parameter of interest in this study. The first drawback is the need for assumptions for model identification, i.e. we need to assume that exists at least one variable significantly correlated with cohabitation but not correlated with hazard of divorce. Some authors suggest that parental characteristics, such as educational level can be valid instruments. In their work Lillard et al. claim that they do not need any exclusion-restriction assumption as the model is identified by means of multiple individual records for those who married more than once. However, such a solution hide the assumption that unobserved heterogeneity is not varying from one marriage episode to the other. This is in facts a strong assumption given that much of unobserved heterogeneity is likely to depend on partner's characteristics, which change from one marriage to the other. Thus model identification remains a problem.

Another limitation is that the effect of focal variable δ_2 (premarital cohabitation in this case) on the log hazard is proportional, i.e. it is considered as time fixed. In facts the effect of cohabitation may taper off or increase over time. This is not an unsolvable issue¹ but there is no way to determine the exact time pattern of the effect.

Third, it has been shown (Heckman & Singer, 1982, 1984; Lindeboom & Van Den Berg, 1994) that the results of duration models are quite sensitive of the functional form of unobserved heterogeneity. Usually it is assumed that the unobserved component is normally distributed mainly for ease of manipulation and not on the base of prior information. Anyway, it emerges that if this assumption does not hold the parameter estimates (and in our case even δ_2) can be severely biased.

The fourth problem concerns time-varying covariates: in the work of Lillard et al. the number of children is controlled for using a time-dependent variable. It is generally acknowledged from a causal point of view it is convenient to account for time variability of certain covariates, such as the number of children (see, for example, Blossfeld & Rowher, 1995). This means that the model accounts for births occurred even after cohabitation and after marriage. However, controlling for children born *after* marriage can net out indirect effects of cohabitation. Let us suppose that premarital cohabitation has a negative effect on the risk of having a child, i.e. cohabiting couples tend to have children later with respect to those who marry without cohabiting. Moreover, let us assume that childbearing has a negative effect on divorce risk, i.e. having a child makes marital disruption less likely. Then, we find that cohabitation makes *indirectly* more likely a marital dissolution since it delays childbearing. If we control for children born after cohabitation we also control for this indirect effect (see Heckman et al., 1999).

2.2 The matching solution

The matching approach to identification of treatment effect is pretty different from the previous one. It moves from the seminal paper by Rosenbaum

¹One may, for instance, create a linear spline variable defining some knots at which the effect changes its slope. In this case δ_2 changes from one knot to the other.

& Rubin (1983). This method invokes the *strong ignorability* assumption: given a vector of covariates X , if $T_0 \perp D|X$ then the ATET is identified, in fact since conditional on X $E(T_0|D = 1, X) = E(T_0|D = 0, X)$ we can compute ATET by

$$E[T_1 - T_0|D = 1] = E_X \{[E(T_1|D = 1, X) - E(T_0|D = 0, X)]|D = 1\} \quad (6)$$

This approach is referred to as “matching” since treated individuals are matched to untreated individuals on the base of X . Though theoretically appealing, the matching approach is in practice difficult to apply when the dimension of X is high because of the difficulties in calculating the conditional expectations in the (2). Rosenbaum & Rubin (1983) showed that instead of matching on the base of X one can *equivalently* match treated and comparison units on the base of any balancing score, and in particular on the base of the “propensity score” that is the conditional probability of receiving the treatment given the values of X , formally

$$p(X) = Pr(D = 1|X) \quad (7)$$

This result reduces the dimensionality problem in computing the conditional expectation and the ATET can be unbiasedly estimated with

$$\begin{aligned} E[T_1 - T_0|D = 1] &= E_{p(X)} \{[E(T_1|D = 1, p(X)) - E(T_0|D = 0, p(X))]|D = 1\} \quad (8) \\ &= \int_{p(X)} \{[E(T_1|D = 1, p(X)) - E(T_0|D = 0, p(X))]|D = 1\} \cdot dF[p(X)] \end{aligned}$$

The wide debate raised around propensity score matching (PSM) has shown that one can go further and identify not only the ATET but the entire marginal distribution of variable $T_0|D = 1$. Heckman et al. (1997) and Imbens (2004) show that with no other assumption than strong ignorability, $F(t_1|D = 1)$ and $F(t_0|D = 1)$ (i.e. the distribution function of $T_1|D = 1$ and $T_0|D = 1$) are identified but not the joint distribution of $(T_1, T_0|D = 1)$, which could be useful if one is interested to the distribution of the treatment effect on treated ($T_1 - T_0|D = 1$).

In duration models anyway, studying the impact of D on the marginal distributions of $T_1|D = 1$ and $T_0|D = 1$ can be very useful, indeed. The matching approach allows to identify, for example, the function

$$\phi(t|D = 1) = S(t_1|D = 1) - S(t_0|D = 1) \quad (9)$$

where $S(t_1|D = 1)$ is the survival function of treated and $S(t_0|D = 1)$ the survival function of the same treated individuals had they not received the treatment. To see this, note that it possible to identify the ATET on any function of T $Y = g(T)$ so that the ATET on Y is

$$E[Y_1 - Y_0|D = 1] = E[g(T_1) - g(T_0)|D = 1]$$

Hence, in order to identify $\phi(t)$ for a fixed t it suffices computing the ATET on the variable $g(t) = 1\{T > t\}$. Firpo (2002) shows that in this way one can identify the “quantile treatment effects”.

Therefore we can apply the propensity score matching to survival functions $S(t_1|D = 1)$ and $S(t_0|D = 0)$

$$\phi(t|D = 1) = \int_{p(X)} \{[S(t_1|D = 1, p(X)) - S(t_0|D = 0, p(X))]|D = 1\} \cdot dF[p(X)]. \quad (10)$$

More than one matching method is available (see, for instance, Becker & Ichino, 2002; Smith & Todd, 2005), here we use a stratification method: the sample is divided into q blocks then the treatment effect is estimated in every block through

$$\phi_q(t|D = 1) = S(t_1|D = 1, p(X), Q = q) - S(t_0|D = 0, p(X), Q = q) \quad (11)$$

where $S(t_1|D = 1, p(X), Q = q)$ and $S(t_0|D = 0, p(X), Q = q)$ are product-limit estimate of survival function at time t of treated and control units in block q . Blocks are determined so that within every block treated and control groups have the same distribution for X and the same average value of propensity score. The overall treatment effect is then computed using the formula

$$\phi(t|D = 1) = \sum_{q=1}^Q \phi_q(t|D = 1) \frac{\sum_{i \in I(q)} D_i}{\sum_{i=1}^n D_i}. \quad (12)$$

$\phi_q(t|D = 1)$ is then weighted with the corresponding fraction of treated units in block q . The variance estimate is as well a weighted sum of variances within every block. Let $V_q^T(t)$ the variance of $S(t_1|D = 1, p(X), Q = q)$ and $V_q^C(t)$ the variance of $S(t_0|D = 0, p(X), Q = q)$, then

$$Var[\phi(t|D = 1)] = \sum_{q=1}^Q \left[\frac{V_q^T(t)}{N_q^T} + \frac{V_q^C(t)}{N_q^C} \right] \cdot \left(\frac{\sum_{i \in I(q)} D_i}{\sum_{i=1}^n D_i} \right)^2 \quad (13)$$

This estimate is unbiased provided that the study is free from *hidden bias* (Rosenbaum, 1995), i. e. self selection occurs only on the base of the observed variables X .

This approach to identification of the effect of a treatment on a episode duration has some benefits in respect to ordinary hazard regression: first of all, in this approach we only need to account for characteristics of treated and controls *prior* the treatment and not after. Then we are not running the risk of removing the indirect effects of treatment from our estimate as we do in a hazard regression model with time-varying covariates. Secondly, we do not impose any time pattern to the treatment effect. In a hazard regression model it is usually assumed that treatment effect is either time-fixed or with a specified time pattern. The matching approach allows to identify the true time pattern of treatment effect without any assumption on it. Of course, we have to take in mind that together with these benefits, there is one drawback of the matching approach, which is that the *strong ignorability assumption* may not hold. Anyway this assumption does not seem stronger than assumptions made in the hazard regression approach on the functional form of the unobserved component and on the exclusion-restriction rules. From the practical point of view,

the choice of matching variables is the most important part of this strategy, since the better is specified the propensity score the lower is the hidden bias, and the more likely is the strong ignorability assumption.

3 Comparison of methods

The next step is to compare hazard regression and matching approaches in estimating the impact of premarital cohabitation on marriage dissolution risk using the same source of data, which is the “National Longitudinal Survey of Youth, 1979” (NLSY79). The NLSY79 is a nationally representative sample of 12,686 young men and women who were 14-22 years old when they were first surveyed in 1979. These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis.

3.1 Description of Data

The NLSY79 is a nationally representative sample of 12,686 young men and women born between 1 January 1957 and 31 December 1964 who resided in the United States in 1979. These individuals are now in their late thirties and forties, and have been personally interviewed for over 2 decades. Data were collected yearly from 1979 to 1994, and biennially from 1996 to the present (see Bureau of Labor Statistics, 2003). The primary purpose of the NLSY79 is the collection of data on each respondent’s labor force experiences, labor market attachment, and investments in education and training. However, the actual content of the NLSY79 is much broader due to the interests of governmental agencies besides the Department of Labor. Examples of other topical areas include marriage and fertility histories, family background and demographic characteristics, high school performance, time use, and alcohol abuse. One main issue of longitudinal prospective data as NLSY79 is attrition of the sample. The retention rate at every round remained close to 90 percent during the first 16 interview rounds and were approximately 85 percent for rounds 17 and 18. In the round 19 (2000) survey, 8,033 civilian and military respondents out of the 9,964 eligible were interviewed, for an overall retention rate of 80.6 percent (Bureau of Labor Statistics, 2003). Table 1 shows the descriptive statistics for the sample of individuals. The sample is voluntarily made up of only ever married individuals (females and males), as we are interested in union dissolutions. Since information on premarital cohabitation is available only for marriages occurred after 1990, those celebrated before that date are dropped. Then every record corresponds to a marriage and for every marriage we have some information on spouses and on the marriage itself. Sampled individuals belong to cohorts from 1957 to 1964, almost 36% of them got a high school diploma, whereas about 20% of their mothers have a high school diploma. Catholics constitute one third of the sample and protestants are even more (44%). As for marriage characteristics we see that the mean age at marriage is 24 and 37 out of 100 marriages were preceded by informal cohabitation and the marriage dissolution rate is 27%.

Table 2 show the results from a probit model on the probability of en-

Table 1: Descriptives statistics of sample

Variable	Mean
cohort range	1957-1964
% female	50.3%
% ever married	100%
Age at marriage	24.05%
% cohabited before marriage	37.4%
% with multiple marriages	7.8%
% Marriages ended with a divorce	27.5%
% graduated or higher	35.5%
% premarital conceptions	9.30%
% Blacks	19.4%
% Catholics	34.3%
% Protestants	44.2%
% Atheists	8.7%
% Mother graduated or more	19.39%
% Rural	3.11 %
% Intact family	75.1 %

tering in cohabitation; the results are in line with literature. Blacks and Hispanics are less likely to cohabit before marriage as well as catholics, protestants and those belonging to other religions (those without any religious belief are the reference group). Having lived in an intact family makes less likely the pre-marital cohabitation. Conversely having had an early (before 18) sexual experience, a previous marriage or a child makes cohabitation more probable.

As far as marriage duration is concerned, in figure 1 we can see the Kaplan-Meier curves for marriage preceded by cohabitation and marriage that were not. From the figure it is evident that *unconditional* on spurious effects of confounding variables, the risk of dissolution is higher for marriages preceded by informal cohabitation.

Figure 2 maps the log negative log of survival functions versus the log of survival time. This is to show that the issue of proportionality cannot be neglected, in fact the curves are not parallel as they should be, in the case of proportional effects. However, even looking only at figure 1 it turns out that survival functions are not parallel and that the effect of informal cohabitation (unconditional on X) is not time-independent.

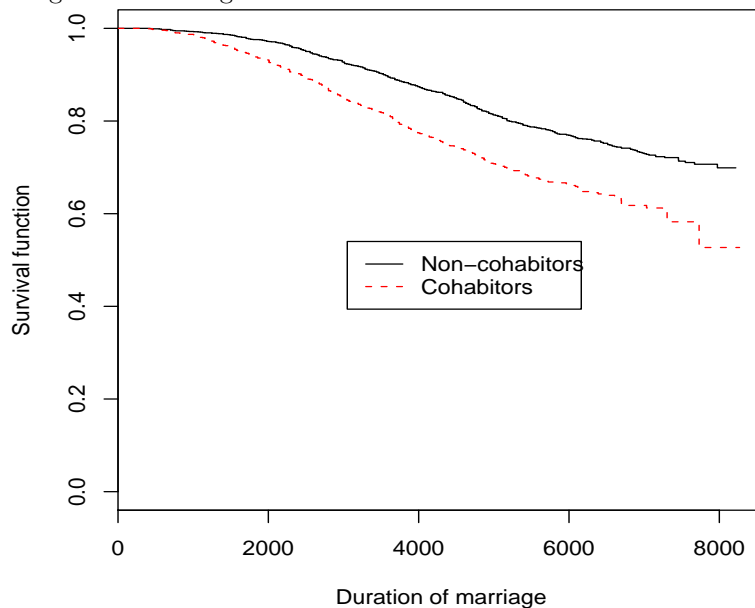
3.2 Results from hazard regression

Basically I replicate the analysis made by Lillard et al. (1995) using different specification of the hazard model. For instance, I show results from model using time-varying covariates (number of children) and the same model not using such variables. The aim is to show that inclusion of time dependent variables can bias the estimate of the parameter of interest. In this section only the value of parameters in the union dissolution equation

Table 2: Probit model on cohabitation.

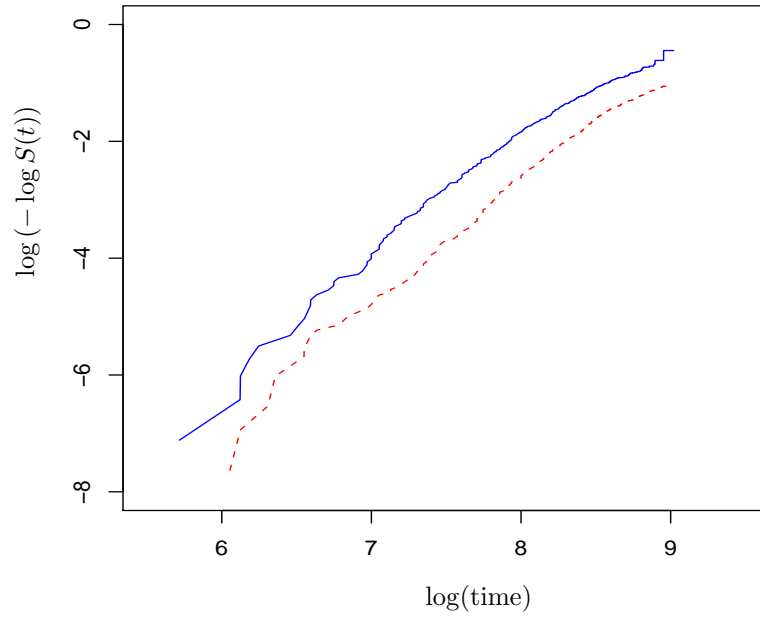
	Coeff.	St. Err.	t-value
Constant	-0,623	1,886	-0,331
Year of birth	0,015	0,028	0,543
Female	0,019	0,058	0,330
<i>Reference: White</i>			
Black	-0,165	0,080	-2,064
Hispanic	-0,299	0,094	-3,173
Mother graduated or more	0,091	0,082	1,113
Father graduated or more	0,059	0,075	0,779
<i>Reference: No religion</i>			
Catholic	-0,380	0,111	-3,438
Protest	-0,376	0,105	-3,578
Other religion	-0,525	0,126	-4,153
Intact family	-0,161	0,068	-2,376
Graduated or more	-0,024	0,027	-0,911
Enrolled in education	-0,161	0,083	-1,929
Rural area	-0,056	0,069	-0,807
Poverty status (in 1979)	0,001	0,001	0,860
Health limitation	0,170	0,135	1,265
Sex int. before 18	0,199	0,095	2,089
Multiple marr.	0,449	0,094	4,795
Children before marr.	0,682	0,091	7,467
Age difference	0,003	0,007	0,496

Figure 1: Marriage duration for cohabitators and non-cohabitators



are shown while the coefficients from the cohabitation equation are not shown given they are quite similar to the coefficient reported in table 2.

Figure 2: Log negative log survival function for cohabitators and non-cohabitators



From table 3 show the first set of results:

Table 3: Simultaneous Model on Marriage duration and cohabitation.

	Model 1		Model 2		Model 3		Model 4	
Constant	-22,150	***	-19,538	***	-21,922	***	-19,387	***
<i>Marriage Duration</i>								
0-1 years	1,620	***	1,538	***	1,614	***	1,524	***
1-7 years	0,360	***	0,267	***	0,361	***	0,269	***
7-15 years	0,091	***	0,049	***	0,091	***	0,050	***
more than 15 years	0,007		-0,010		0,004		-0,013	
Cohort (Base: 1957)	0,230	***	0,189	***	0,227	***	0,186	***
Marriage 2+	0,540	***	0,604	***	0,513	***	0,565	***
Cohabited	-0,297		-0,208		-0,149		-0,015	
Duration of cohab.	0,099	**	0,115	***				
Premarital child	-0,310	**	0,119		-0,291	**	0,153	
<i>Children born in current marriage</i>								
1 child (time-varying)	-0,615	***			-0,625	***		
2 children (time-varying)	-1,251	***			-1,265	***		
Black	0,142		0,182		0,141		0,182	
Hispanic	-0,108		-0,134		-0,109		-0,133	
Female	-0,244	***	-0,180	**	-0,232	**	-0,167	*
Catholic	-0,111		-0,159		-0,110		-0,158	
Protest	-0,169		-0,145		-0,173		-0,152	
Other rel.	-0,028		-0,041		-0,029		-0,038	
Intact family	-0,214	**	-0,179	*	-0,222	**	-0,187	**
Graduated or more	0,013		-0,002		0,011		-0,004	
Enrolled to educ.	-0,358	***	-0,303	**	-0,356	***	-0,302	**
Rural area	0,003		0,025		0,004		0,027	
Poverty	-0,001		-0,001		-0,001		-0,001	
Sex before 18	0,274	**	0,227	*	0,285	**	0,239	*
Health limitations	0,332		0,273		0,326		0,270	
Age difference	0,002		0,000		0,003		0,001	
σ_{ξ}	1,185	***	0,991	***	1,196	***	1,017	***
σ_{ϵ}	0,759	***	0,762	***	0,759	***	0,761	***
$\rho_{\xi,\epsilon}$	0,601	***	0,652	***	0,595	***	0,615	***

***=p-value < 0.01, **=p-value < 0.05, *=p-value < 0.1.

when cohabitation and marital dissolution are considered simultaneously, the positive effect of cohabitation becomes negative and not significant². In our case, differently from Lillard et al. (1995) there is an effect of duration of cohabitation even in the simultaneous model. The correlation is positive and significant meaning that those who are more prone to cohabitation are even more prone to marriage dissolution. The second column of table 3 reports the results of the same model without the time-varying covariates on the number of children born during marriage. As remarked before, the use of these variables can bias the effect of cohabitation. If the number of children born during marriage has a negative effect on marital dissolution, which is the case given the results in the first column, and cohabitation has a negative on the hazard of having children during marriage then controlling for the number of children means also controlling for an indirect effect of cohabitation. This seems to be the case, indeed the

²The results of single equation model on marital dissolution are not shown for the sake of brevity. What we need to know is that in this model the effect of cohabitation is positive and significant.

effect of cohabitation on marital dissolution is -0.297 when the number of children is controlled for and -0.208 when the number of children is not controlled for. Similarly the effect of duration of cohabitation pass from 0.099 to 0.115. In the third and the fourth column of table 3 the same models are replicated except for variable “duration of cohabitation” that was dropped. Thus the change in the value of the cohabitation parameter is more evident passing from -0.149 to -0.015. In our case this change is not relevant as the value of δ_2 is not significant this example show that using a time-varying covariate can bias the estimate of treatment effect and in other cases this bias can be more relevant.

3.3 Cohabitation effect: estimation with the matching procedure

The first step for estimating the effect of pre-marital cohabitation on the risk of divorce is to estimate the propensity score via a logit model. The main issue in this case is not the correct specification of the logit model but the inclusion of all the variables that are suspected of confounding the effect of cohabitation on the divorce risk. For example, it is necessary to include a dummy variable for second or higher order marriages. Those who already experienced a divorce are more likely to start a new marriage with a cohabitation and on the other hand are probably more prone to marriage dissolution. Other variables are likely to be correlated with the propensity of experiencing a divorce: religious affiliation, for example, has to be controlled for, as well as ethnicity: catholics are less likely to cohabit before marriage and less inclined to divorce (considering the Catholic precepts about marriage) than non-catholics. After having estimated the propensity score, The sample is divided into blocks based on percentiles of propensity score. Thus the first block includes all individual with the lowest value of propensity score up to percentile p_1 , the second block all the individuals whose propensity score is between p_1 and p_2 , and so on³. Figure 3 shows the estimate of $\phi(t|D = 1)$ as described in section 2.2. The second graph shows the reconstruction of survival curves after matching. The survival function of treated is estimated through ordinal product-limit method, as $S(t_1|D = 1)$ is identified. The survival function for control units is reconstructed as a difference between $S(t_1|D = 1)$ and $\phi(t|D = 1) = S(t_1|D = 1) - S(t_0|D = 1)$.

The result confirms that the effect of cohabitation is not time-fixed: during the first 3 years the hazard of marriage dissolution of cohabitators is at the same level of non-cohabitators. After this period the survival curve of cohabitators decline more rapidly then again after 4000 days of marriage (almost 10 years) $\phi(t|D = 1)$ becomes flat again. At the end, we see that after 6000 days of marriage the “treated” units experienced a higher rate of marriage dissolution than the control units did.

³The number of blocks is determined by an algorithm testing within every block that average propensity score of treated and control units does not differ. If the test fail in one block the latter is split into two blocks and repeat the test (Becker & Ichino, 2002, see).

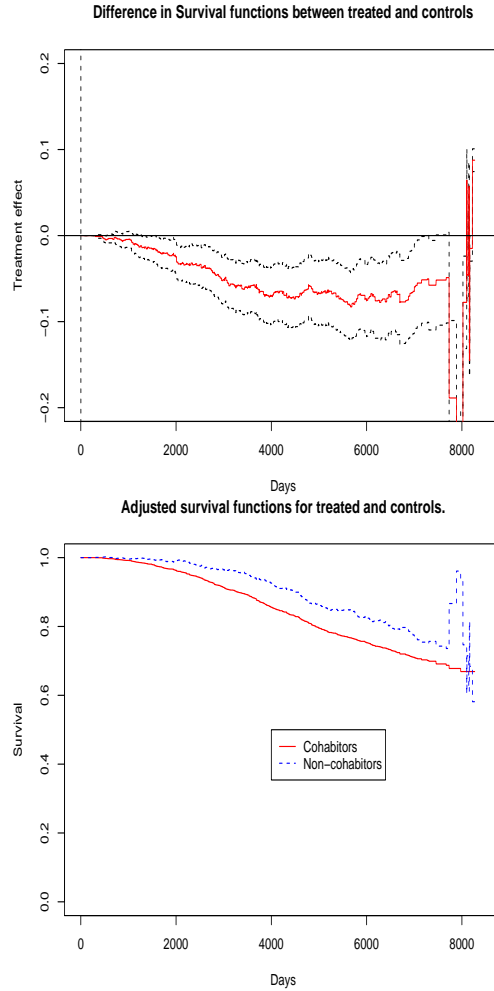
Table 4: logit model for pre-marital cohabitation, NLSY79.

	Estimate	Std. Err.	t-value
(Intercept)	0,111	0,511	0,218
<i>Reference: 1957</i>			
Year of birth 1958	0,093	0,198	0,472
Year of birth 1959	0,184	0,195	0,941
Year of birth 1960	0,153	0,195	0,788
Year of birth 1961	0,211	0,239	0,886
Year of birth 1962	0,217	0,249	0,874
Year of birth 1963	0,418	0,254	1,645
Year of birth 1964	0,289	0,263	1,099
Female	0,053	0,119	0,448
Graduated or more	0,018	0,181	0,097
Enrolled in education	-0,194	0,125	-1,546
Female * Graduated	-0,296	0,180	-1,643
Rural area	-0,079	0,108	-0,731
Health limitations	0,320	0,202	1,582
<i>Reference: No religion</i>			
Catholic	-0,657	0,160	-4,100
Protestant	-0,573	0,155	-3,699
Other religion	-0,849	0,188	-4,511
Black	-0,020	0,124	-0,160
<i>Original family characteristics</i>			
Family size (before union)	-0,094	0,024	-3,919
Poverty status (in 1979)	-0,016	0,120	-0,133
Intact family	-0,123	0,105	-1,171
Mother graduated or more	0,052	0,127	0,407
Father graduated or more	0,079	0,117	0,675
Siblings	0,017	0,020	0,822
<i>Sentimental history</i>			
Age at union	0,003	0,016	0,164
Multiple marriage	0,642	0,166	3,877
Children before union	0,654	0,115	5,677
Sex int. before 18	0,294	0,141	2,087
Age difference	0,000	0,000	-1,172
Female * Age difference	0,000	0,000	2,272

4 Conclusion

The paper shows two different approaches used in determining the effect of pre-marital cohabitation on marriage duration. The first approach uses a bivariate model of both processes taking into account the effect of unobserved heterogeneity by means of a bivariate random component with normal distribution. The second approach assumes “strong ignorability”

Figure 3: Treatment effect on duration of marriage and survival functions after matching



or “conditional independence”, i.e. the effect of unobserved heterogeneity is assumed to be irrelevant in the determination of the cohabitation effect. In this paper I criticize the first approach because of some pitfalls: the functional form of the bivariate random component can heavily influence the estimates of the cohabitation effect, and given there is no theoretical justification for imposing one particular distribution to the unobserved heterogeneity this makes the results unreliable. Another pitfall is provided by the time-varying covariates which use can have undesired effect on the the treatment effect estimates, canceling out a possible indirect effect of the treatment. Moreover the treatment effect is often assumed to be time-invariant, alternatively a set of knots has to be specified in order

to allow for time variability of the effect. Finally, estimates rely on a set of exclusion-restriction hypothesis. In the example provided identification is based on multiple individual records provided by multiple marriages of individuals but this solution implies that the unobserved heterogeneity does not change across marriage episodes.

The second approach is not free from drawbacks. Basically, the estimates relies on the conditional independence assumption, an assumption that is often overly strong (Heckman et al., 1998). However is possible to apply the same approach to Instrumental Variable estimator (Angrist & Kruger, 2001; Imbens & Angrist, 1994). This approach does not need the conditional independence assumption but it needs a valid instrument, without any other assumption on functional form of unobservables. Thus we can outline two possible situations: in the first one we can have rich information on the process we are studying. In this case we may be confident that the effect of the unobserved factors is negligible, then we can use a matching estimator. In the second case we have not rich data or we suspect that the effect of the unobserved factors is relevant, but we observe a valid instrument. Then we can apply the approach illustrated in section 2.2 to a IV estimator.

This work is still at a preliminary stage: basically I aim to run some sensitivity analysis both on regression and matching estimates. The hazard regression will be estimated with different specification of unobserved heterogeneity term distribution in order to see how estimate of treatment effect is sensitive to random effect functional form. Furthermore I intend to check how the matching estimate is robust to different specification of the set of matching variables.

References

- ANGRIST, J. D. & KRUGER, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiment. *Journal of Economic Perspectives* **15**:4, 69–85.
- BECKER, S. & ICHINO, A. (2002). Estimation of Average Treatment Effects Based on Propensity Scores. *The STATA journal* **2**(4), 358–377.
- BENNETT, N. G., BLANC, A. K. & BLOOM, D. E. (1988). Commitment and the Modern Union: Assessing the Link Between Premarital Cohabitation and Subsequent Marital Stability. *American Sociological Review* **53**(2), 127–138.
- BLOSSFELD, H.-P. & ROWHER, G. (1995). *Techniques of Event History Modeling. New Approaches to Causal Analysis*. Mahwah, New Jersey.
- BUREAU OF LABOR STATISTICS (2003). *NLS Handbook 2003*. Washington, DC.
- DEMARIS, A. & MACDONALD, W. (1992). Premarital Cohabitation and Subsequent Marital Stability in the United States: A Reassessment. *Journal of Marriage and the Family* **54**, 178–190.

- FIRPO, S. (2002). *Efficient Semiparametric Estimation of Quantile Treatment Effect*. Ph.D. thesis, Department of Economics, University of California, Berkeley.
- HECKMAN, J. J., ICHIMURA, H., TODD, P. E. & SMITH, J. (1998). Matching As an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies* **64**, 605–654.
- HECKMAN, J. J., LALONDE, R. J. & SMITH, J. A. (1999). The Economics and Econometrics of Active Labor Market Programs. In *Handbook of Labor Economics*, O. Ashenfelter & D. Card, eds. Elsevier Science, pp. 1865–2097.
- HECKMAN, J. J. & SINGER, B. (1982). The Identification Problem in Econometric Models for Duration Data. In *Advances in Econometrics*, W. Hildenbrand, ed. Cambridge: Cambridge University Press. Proceedings of the Fourth World Congress of the Econometric Society.
- HECKMAN, J. J. & SINGER, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica* **52**, 271–320.
- HECKMAN, J. J., SMITH, J. & CLEMENTS, N. (1997). Making the Most Out of Programme Evaluation and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *The Review of Economic Studies* **64**, 487–535.
- HOLLAND, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81**, 945–970.
- IMBENS, G. & ANGRIST, J. D. (1994). Identification and Estimation of Local Average Treatment Effect. *Econometrica* **62:2**, 467–475.
- IMBENS, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics* **86**, 4–29.
- LILLARD, L. A. (1993). Simultaneous Equations for Hazards. Marriage Duration and Fertility Timing. *Journal of Econometrics* **56**, 189–217.
- LILLARD, L. A., BRIEN, M. J. & WAITE, L. J. (1995). Premarital Cohabitation and Subsequent Marital Dissolution: A Matter of Self-Selection? *Demography* **32(3)**, 437–457.
- LILLARD, L. A. & PANIS, C. W. A. (1998). Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status, and Mortality. *The Journal of Human Resources* **33(2)**, 437–458.
- LINDEBOOM, M. & VAN DEN BERG, G. J. (1994). Heterogeneity in Models for Bivariate Survival: The Importance of Mixing Distribution. *Journal of the Royal Statistical Society. Series B* **56**, 49–60.

- ROSENBAUM, P. R. (1995). *Observational Studies*. Leipzig, Germany: Springer-Verlag.
- ROSENBAUM, P. R. & RUBIN, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41–55.
- RUBIN, D. B. (1977). Assignment to on the Basis of a Covariate. *Journal of Educational Statistics* **2**, 1–26.
- SMITH, J. A. & TODD, P. E. (2005). Does Matching Overcome LaLonde Critique of Nonexperimental Estimators? *Journal of Econometrics* **125**, 305–353.
- TEACHMAN, J. D. & POLONKO, K. A. (1990). Premarital Cohabitation and Subsequent Marital Stability in the United States. *Social Forces* **69(1)**, 207–220.