Accounting for Unobserved Heterogeneity in Discrete-time, Discrete-choice Dynamic Microsimulation Models.

An application to Labor Supply and Household Formation in Italy.

Ambra Poggi

(University of Milan-Bicocca and LABORatorio R. Revelli)

Matteo Richiardi

(University of Turin and LABORatorio R. Revelli)

May 2011

Abstract

This paper analyzes the implications of unobserved heterogeneity in discrete-time, discrete-choice microsimulation models. We compare the predictions coming from simple pooled probit estimates with those obtained using random effect dynamic probit models, in a dynamic microsimulation of household formation and labor supply in Italy. We show that failing to account for unobserved heterogeneity has important quantitative consequences, which are often neglected in empirical microsimulation work.

Keywords: dynamic microsimulation, unobserved heterogeneity, female labor force participation

JEL Classification: C53, C18, C23, C25, J11, J12, J21

1. Introduction

Dynamic microsimulation models are used for policy analysis and evaluation, and to project into the future trends of economically relevant variables, taking into account the likely evolution of their determinants. The dynamics of each process are generally governed by coefficients that have been estimated on historical data. The choice of the econometric specification is therefore crucial for the quality of the predictions. Ironically microsimulations, which have been explicitly developed to allow distributional analysis and hence a thorough consideration of individual heterogeneity, often fail in modeling the same individual heterogeneity properly, neglecting the role of unobserved heterogeneity. This is especially true for dynamic microsimulations, facing the additional issue of attributing the unobserved individual effects both to the individuals in the initial population and to the new individuals entering the microsimulation in later periods. The goal of the present paper is to investigate whether and to what extent explicitly modeling unobserved heterogeneity makes a difference, in terms of results. As a sample application, we develop a discrete-time dynamic microsimulation of labor supply and household formation tailored to the Italian case, and compare the outcomes in terms of participation rates of different subgroups of the population of two versions of the model: one in which all processes are modeled by means of simple pooled probit specifications with lagged endogenous variables, and the other in which we specify random effect dynamic probit models, where initial conditions are estimated following Heckman (1981a, 1981b). We show that the differences in projected outcomes are large and significant.

For the purpose of our study, Italy is a particularly apt case, as the female participation rates are still very low as compared to other EU states, despite having markedly increased over the past decades (Del Boca et al., 2006). This leaves space for a further increase (Leombruni and Richiardi, 2006), and makes the outcome of the microsimulation highly sensitive to the estimated coefficients.

The paper is structured as follows. Section 2 discusses the methodological problem that motivates our analysis. Section 3 illustrates our strategy to deal with unobserved heterogeneity. Section 4 describes why Italy is an interesting testbed for this exercise, given a long-term trend toward increasing female labor force participation which emphasizes any difference in the estimates. Section 5 discusses the econometric specifications and estimates, while section 6 describes the microsimulation model and section 7 presents the main results of the simulation. Section 8 concludes.

2. Unobserved heterogeneity

In discrete-time microsimulation many processes are dichotomous, e.g. labor market participation, household formation, fertility outcome.[1] They are generally modeled by comparing the value of a latent outcome variable y^*, assumed to be a function of observable characteristics of the individual, with a threshold[2]. In the case of a probit,

$$
\begin{aligned}
& y^*_{i,t} = x_{it}'\beta + \varepsilon_{i,t} \\
& y_{i,t} = 1 \quad \text{if} \quad y^*_{i,t} > 0 \\
& y_{i,t} = 0 \quad \text{otherwise} \\
& \varepsilon_{\mathrm{i,t}} \approx \mathrm{N}(0,1)
\end{aligned} \tag{1}
$$

which gives raise to the standard expression

$$\Pr[y_{i,t} = 1 \mid x_{i,t}] = \Phi(x_{i,t}'\beta) \tag{2}$$

where Φ is the cumulative distribution function of a standard normal distribution, x is a vector of strictly exogenous observed explanatory variables for individual *i* at time *t*, and the vector *β* contains the parameters to be estimated.

However, these processes are often characterized by a high degree of persistence, i.e. the explanatory power of lagged dependent variables is very high. Including the lagged dependent variable in this settings leads to

$$\Pr[y_{i,t} = 1 \mid x_{i,t}, y_{i,t-1}] = \Phi(x_{i,t}'\beta + \gamma y_{i,t-1}) \tag{3}$$

with γ being an additional parameter to be estimated.[3]

---

[1] The same argument applies, *mutatis mutandis,*to a multinomial setting with more than two states.

[2] An alternative, which is however rarely considered, is the linear probability model.

[3] An alternative approach is to use Markovian transitions models, where the processes are split up according to the initial state. This amounts to estimate

$$
\begin{aligned}
& \Pr[y_{i,t} = 1 \mid x_{i,t}, y_{i,t-1} = 0] = \Phi(x_{i,t}'\beta_0) \\
& \Pr[y_{i,t} = 1 \mid x_{i,t}, y_{i,t-1} = 1] = \Phi(x_{i,t}'\beta_1)
\end{aligned} \tag{5}
$$

or, equivalently,

$$\Pr[y_{i,t} = 1 \mid x_{i,t}, y_{i,t-1}] = \Phi(x_{i,t}'\beta_0 + x_{i,t}'\delta\, y_{i,t-1}) \tag{6}$$

The main advantage is the possibility to allow estimated persistence to vary according with the initial state and the observed individual characteristics. The main disadvantage is the high number of parameters to estimate. Since our purpose is to highlight the relevance of unobserved heterogeneity for microsimulation models, we prefer to adopt the simpler approach illustrated by equation (3).

This poses no problems if observed persistence is only due to true state dependence. In this case, being in a certain state (i.e. participate to the labor market) in a specific time period, in itself, increases the probability of being in the same state in subsequent periods. However, observed persistence may also be due to permanent unobserved heterogeneity.[4] In this case, individuals could be heterogeneous with respect to characteristics that are relevant for the chance of being (and remaining) in a certain state. For example, an individual might participate to the labor market despite unfavorable observable characteristics, due to favorable unobservable characteristics like a special taste for work, work ethics, need for money etc. These characteristics are likely to be at least to some extent persistent over time; therefore, they increase the likelihood that the individual will also participate in the future. These unobserved individual effects work in every respect as omitted variables. As Woolridge (2002) puts it, "In nonlinear models, much has been made about the deleterious effects that ignoring heterogeneity can have on the estimation of parameters, even when the heterogeneity is assumed to be indipendent of the observed covariates. A leading case is the probit model with an omitted variable. Yatchew and Griliches (1985) show that when the omitted variable is indipendent of the explanatory variables and normally distribuited, the probit estimators suffer from (asymptotic) attenuation bias. This result is sometimes cited to illustrate how a misspecification that is innocuos in linear models leads to problems in nonlinear models".[5] Moreover, unobserved heterogeneity is not independent of the lagged dependent variable, if included among the covariates, as the latter is correlated by construction with the lagged residual, which includes the unobserved permanent individual effect. As a result, estimates of state dependence that fail to account for permanent unobserved heterogeneity are in general upward biased: the model attributes to true state dependence any individual effect that makes a transition to a different state less likely.

In order to allow for permanent unobserved heterogeneity, we need to write the error term as

(4) $$\varepsilon_{i,t} = \alpha_i + \eta_{i,t}$$

where α_i indicates the individual-specific effect (that is, the unobserved permanent heterogeneity).

In the microsimulation literature, the problem of unobserved heterogeneity has not received homogeneous attention. During the last two decades, since the rise of static microsimulation models with behavioral responses, the community has seen an increasing divide between static and dynamic models, with the former becoming increasingly sophisticated from an econometric perspective.[6] A growing interest in structural modeling has caused more thoughts on the econometric specifications, and the issue of

[4] In facts, we can get significant estimates of state dependence coefficients even when there is no true state dependence and persistence is only due to permanent unobserved heterogeneity (Carro, 2007).
[5] See also Cramer (2005).
[6] See for instance the literature on discrete choice models of labor supply that has followed Aaberge et al. (1995) and van Soest (1995).

unobserved heterogeneity as a separate source of state dependency has gained increasing importance (see for instance Haan, 2006 and Pacifico, 2009).

How is the problem treated, on the other hand, in dynamic microsimulation models? A first answer is that it is often difficult to tell: papers based on dynamic microsimulations generally devote little space to the presentation of the econometric estimates, and sometimes even the list of covariates is missing. This is a well known problem with this approachs: being in general large models, developed over the course of many years often building on pre-existing work, they end up being close to black boxes. Sometimes detailed explanations can be found in technical papers that however remain unpublished or have a limited circulation, while published articles often restrict their attention, given the page constraint, on some specific result / addition to the basic model.

One common way to treat unobserved heterogeneity is to ignore the problem: unobserved heterogeneity is generally (although not explicitly) simply assumed away. This is the case of well established microsimulation models as Corsim, Prism, Dynasim and Dynacan, just to give some examples, as documented in Anderson (2003). But, as we have seen, leaving out unobserved permanent characteristics leads to biases in the estimates since the model compensates for the missing factor by over- or under-estimating the effect of the other explanatory variables. In particular, the coefficient of the lagged dependent variable will be normally over-estimated.

Sometimes practitioners argue that the interest of microsimulation models lies not in interpreting the coefficients, but rather in the predictive power of the whole statistical apparatus. According to this view, neglecting unobserved heterogeneity might lead to biased and inconsistent estimates, but the models are possibly still good for predicting the future evolution of the variables of interest. While the Lucas critique would clearly apply to this line of reasoning, suggesting that the resulting models cannot be used for policy analysis, even the ability to fulfill the more limited goal of projecting trends into the future, assuming the economic environment remains unchanged, needs to be at least substantiated empirically.

An additional difficulty with dynamic microsimulation models, which might partly explain why researchers are cautious in including unobserved heterogeneity in their specifications, is the need to evolve the initial population into the future. This implies that unobserved individual effects have to be assigned to the individuals in the initial population, if the initial population does not come from the same dataset on which the models were estimated, and also to the new individuals entering the microsimulation at a later stage. Panis (2003) reviews three ways of dealing with the problem: (i) a sample distribution approach - using the same distribution for unobserved heterogeneity as the one estimated in the estimation sample, (ii) a Bayesian approach –deriving the posterior distribution of heterogeneity given the observed past outcomes, and (iii) an optimal assignment approach –assigning to each individual the value for unobserved

heterogeneity that best matches his observed past outcomes. The first approach is feasible only when individuals enter the simulation without a previous history of outcomes, while the two latter approaches are generally quite computationally intensive. For instance, solving the optimal assignment problem involves finding the distribution of individual effects that maximizes the likelihood of observing the true data, which in turns implies in principle to evaluate all *N*! permutations of the individual effects, with *N* being the number of individuals in the simulation. If unobserved heterogeneity is accounted for in the estimation procedure, but is not dealt with in the simulation of the artificial population, individual histories and even the gross totals might be totally inaccurate.

3. Modeling unobserved heterogeneity

In the econometric literature, there are two ways of treating unobserved heterogeneity: random effects or fixed effects models. Interested readers can refer to Honore (2002) for a full discussion on the choice between these two approaches. This paper follows the random effects approach in order to have a fully specified model in which one can estimate all the quantities of interest, including time invariant coefficients (e.g. gender). This allows not only to correctly interpret the coefficients –as in standard microeconometric papers– but, of most importance here, to simulate forward the evolution of the initial population. Note also that the random effects approach usually lead to more efficient estimators of the parameters of the model if the distributional assumptions are satisfied. Moreover, traditional maximum likelihood estimator of non-linear panel data models with fixed effects generally exhibits considerable bias in finite sample when the number of periods is not large.[7]

Finally, in order for the model's results to be fully parameterized the initial conditions also have to be specified since the interpretation of the model depends on the assumption made concerning the initial observation. An initial condition problem arises when the start of the observation period does not coincide with the start of the stochastic process generating individual experiences (i.e. Heckman, 1981a; Arulampalam et el., 2000).[8] Our way to deal with this issue is to use the estimator suggested by Heckman

[7] Fixed effects estimators of nonlinear panel model can be severely biased due to the incidental parameters problem. This problem arises because unobserved individual characteristics are replaced by sample estimates, biasing estimates of model parameters. As far as we know, the solution proposed are not sqrt(N)-consistent (Honore and Kyriazidou, 2000; Hahn, 2001; Honore and Tamer, 2004). Some authors propose modified maximum likelihood estimators that reduces the order of the bias (i.e. Cox and Reid, 1987; Arellano, 2003; Carro, 2007; Arellano and Hahn, 2007; Val, 2009). The latter estimators work only "moderately" well when T is larger than 8, but this is not our case. Very recently, Hoderlein et al. (2011) proposes an nonparametric procedure that generalizes the conditional logit approach leading to an estimators based on nonlinear stochastic integral equations that also seems to works moderately well in finite sample Monte Carlo simulations.

[8] In dynamic panel data models with unobserved effects, the initial condition problem is an important issue. Many authors studied dynamic linear models with an additive unobserved effect with a special focus on the treatment of the initial condition problem (see, for example, Ahn and Schmidt, 1995; Anderson and Hsiao, 1982; Arellano and Bond,

(1981a, 1981b)[9]. His approach involves the specification of an approximation to the reduced form equation for the initial condition and allows for cross-correlation between the dynamic equation and the initial condition:

(7) $$\Pr[y_{i,0} = 1 \mid \alpha_i] = \Phi(z_{i,0}'\lambda + \theta\alpha_i)$$

where z_i is a vector of exogenous covariates (including x_{i0} and, eventually, additional variables that can be viewed as "instruments" such as pre-sample variables).[10] Exogeneity corresponds to $\theta = 0$ and can be tested accordingly. Equations (3), (4) and (7) together specify a complete model for the process that can be estimated by maximum likelihood (for details about the estimation see Arulampalam and Steward, 2007).

In order to distribute the individual effects in the initial population and among the individuals who enter the simulation in later periods, taking into consideration the fact that the random effects and the (ex-post) outcomes are correlated, we follow an optimal assignment approach but use a simplified reverse-engineering algorithm which greatly reduces computing time.:

1. compute the predicted outcome (probability of a positive outcome) using the coefficients of the random effect dynamic model, but imposing a random effect equal to 0 for all *N* individuals to whom random effects need to be assigned;
2. compute the difference between the outcome (0 or 1) and the predicted outcome;
3. order this difference from low to high;
4. extract *N* values for the random effects;
5. order these random effects, from low to high;
6. assign the random effects to the individuals by matching the two rankings described above.

For instance, suppose there are three individuals A, B and C with outcomes Y_A=0, Y_B=1, Y_C=1 and predicted outcomes $\tilde{Y}_A$=.35, $\tilde{Y}_B$=.70, $\tilde{Y}_c$=.85. The differences are Δ_A=-.35, Δ_B=.30, Δ_B=.15. We extract (from a Normal distribution with mean 0 and standard deviation equal to the estimated standard deviation of the random effects, σ_a) three values for the random effects, say α_1=-.12, α_2=-.03, α_3=.05. Individual A has an

1991; Arellano and Bover, 1995; Blundell and Bond, 1998; Hahn, 1999]). The initial condition problem is much more difficult to resolve within non-linear models. Honore (1993) and Honore and Kyriazidou (2000) offer examples of the treatment of the initial condition problem in a semi-parametric context. The interested reader can also refer to Heckman (1981a, 1981b), Hsiao (1986), Orme(1997) and Wooldridge (2005) for a discussion of alternative ways of handling initial conditions in a dynamic non-linear model with unobserved heterogeneity.

[9] Other possible estimators are the ones proposed by Orme (1997) and Wooldridge (2005). However, the three estimators provide similar results (Arulampalam and Stewart, 2007).

[10] If the vector z_i does not include instruments, the model is identified by the functional form.

outcome that is lower than the predicted value, while B and C have an outcome that exceeds their predicted values, with B having a positive outcome despite a relatively low predicted probability. Therefore, A will get α_1, the lowest random effect; B will get α_3, while C will get α_2. B has the highest random effect, coherently with the fact that, abstracting from the random effect, he is the one with a positive outcome that has the lowest predicted probability: his unobserved quality must be high.

4. Female labor supply in Italy

Over the last decades, we observed an increasing long-term trend in female participation rate in most OECD countries. Nevertheless, we also observe persistent differences in levels suggesting that different countries are constrained by country-specific institutional and social factors. Ahn and Mira (2002) and Engelhardt *et al.* (2001) have divided the 21 OECD countries into three groups. The high participation group, in which the participation rate was, at the time of the study, above 60%, includes the U.S., Canada, the U.K., Sweden, Norway, Denmark, Finland and Switzerland. The medium participation group includes countries where the participation rate was in the 50-60% range. Finally, the low participation group includes Italy, Spain and Greece, where the female participation rate was lower than 50%. The latter group was also the target of the the Lisbon 2000 Agenda, which set a goal of (at least) 60% for the female employment rate, to be achieved by 2010. By that year, the female employment rate in Italy was still below 50%, followed only by Malta in the EU27 ranking.

Low female participation has always been a feature of the Italian labor market despite the increase in female employment rate from 35.4% in 1994 to 47.2% in 2008 (Rondinelli and Zizza, 2011). Partecipation rates differ considerable between males and females (see Figure 1). Education level matters in explaining the size of the gap: lower the education levels are associated with larger gender gaps. Even if participation rates of married women increased over the last several decades (Del Boca et al., 2006), employment rates of mothers with children under six in Italy are still very low (Del Boca, 2003): in facts, more than one fourth of women leave the labor market after a birth (Bratti et al. 2005; Casadio et al. 2008).

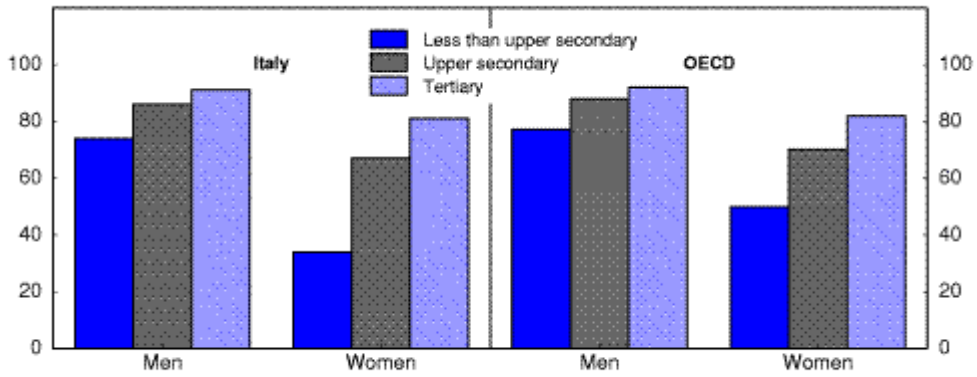

Figure 1. Participation rates by education levels, 2001. Source: OECD, Employment Outlook database; Eurostat, New Chronos

The following factors help explaining the gender participation gap. First, in spite of the recent institutional changes, the Italian labor market still remains highly regulated: strict rules apply to hiring and firing and specify the types of available employment arrangements; these labor market regulations have been largely responsible for the high unemployment rates of women and youth (Bertola et al. 2001). Thus, women have hard time times to enter and re-enter (after breaks during childbearing years) the labor market. This situation affects also participation rates since discouraged women may decide to drop out of the labor force.

Second, part-time employment is still not common in Italy: it is less than 30% for women, while it is above 75% in the Netherlands (Eurostat data for 2010); moreover, this is an important factor in accounting for low participation rates of married women, particularly those with children (Del Boca, 2002).

Third, women do not have the same opportunities as men within the family (in terms of division of household labor and child bearing); in facts, the reconciliation of roles within and outside the family is more difficult for a working mother than for a working father, and often the strategies adopted are completely different: men typically increase the time devoted to paid work and women decrease their working time or even exit the labor market (Anxo et al., 2007, Mencarini and Tanturri, 2004, Lo Conte and Prati, 2003).

Fourth, the public childcare system does not provide services which are of much assistance to married women in terms of reducing the direct costs of participation; in particular, there are a limited number of slots available (especially in some regions in the South) and the hours of childcare is typically non compatible with full-time jobs hours; also having school age children does not necessarily increase the participation rates since school days often end in mid-afternoon much early than the end of full-time work days (Del Boca, 2002; Del Boca and Vuri, 2007).

5. Data, econometric specifications and estimation results

5.1 The data

The initial population is a random extraction (with replacement) from the 2005 Italian Statistics on Income and Living conditions (IT-SILC) data. IT-SILC is the Italian component of the European Community Statistics on Income and Living conditions (EU-SILC), a data source provides cross-sectional and longitudinal micro data mainly referred to objective living and employment conditions. Weights are provided in order to achieve representativeness of the total population. The 2005 data contain observations on 56,105 individuals. The population covered by EU-SILC is composed by individuals aged 16+. Individuals reporting missing observations in the variables of interest are excluded from the sample.

Input estimations are run using Italian micro-data from all eight waves 1994 to 2001 of the European Community Household Panel (ECHP), the precursor of EU-SILC. ECHP is a household panel survey conducted annually by following the same sample of households and persons (aged 16+). The main advantage of ECHP is that it permits us to analyze participation in the labor market, schooling and household formation from a dynamic point of view. Unfortunately, the exact moment of graduation is often only poorly observed in ECHP. For this reason, we use the same coefficients for graduation as in Leombruni and Richiardi (2006), estimated on the 1993-2003 Italian Labor Force Surveys (LFS) data.

5.2 Initial status

ECHP and EU-SILC data cover individuals aged 16+. Since we need information on lagged status, individuals enter the simulation at age 17. Assignment to the initial status (in education, activity, employment) is random, based on observed probabilities (**Errore. L'origine riferimento non è stata trovata.**). As a benchmark, we assume these probabilities to remain constant over the whole simulation period.

	In education	Active	Unemployed
		Males	
North	93.8%	6.2%	21.6%
Centre	95.3%	4.7%	90.9%
South	83.3%	16.4%	88.5%
		Females	
North	87.5%	12.3%	34.7%
Centre	98.8%	2.0%	60.0%
South	86.8%	8.5%	73.0%

Table 1. Status at age 17. Source: IT-SILC (2005)

5.3 Education

In Italy school attendance is compulsory until age 16, while it is illegal to work under 15.[11] Primary school has 8 grades, and students should normally complete it at 13-14. Then, they compulsorily enrol in secondary school, which should last for 5 years. According to Sistan (2006, 2007), secondary school attendance from age 16 to age 18 is above 80% (2005 data), while the probability of achieving a diploma is slightly above 70% (67% for males and 78% for females). Early school leavers (individuals aged 18-24 that achieved at least an education level equal to ISCED 2) are about 22%, higher than the EU-25 average (15%).

Among those who achieved a diploma, three in four enrol at university (79% females; 66% males), while less than one in two of those enrolled actually graduate (51% females; 37% males). Among those who make it, 52% graduates before age 25 and the 80% before age 29. Overall, university enrolment rate is about 56%, about the OCSE average (54%). Enrolment rates have increased from 1998 to 2005 (+16%). Dropouts after the first year of university are about 20%, and they remain significant even later. Graduation rates have remained fairly stable over the years. Very few people come back to formal education, once left.

Coherently with this picture, we estimate two separate equations, one for secondary school attendance and one for university attendance. In both cases the probability of being a student at time *t* is modeled as a function of sex, age, age-squared, year of birth and area of residence. Moreover, the model estimation is conditioned on not having entered the labor market. Estimates are reported in **Errore.**

[11] Illegal dropouts are approximately zero in primary school and about 1.5% in secondary school (Sistan, 2006).

L'origine riferimento non è stata trovata.. The probability of being enrolled in secondary school decreases with age, as individuals are supposed take a diploma at about 18-19 years old.

	Secondary school			University		
	Coef.		SE	Coef.		SE
Female	-0.008		0.069	0.113	*	0.055
Age	-11.285	**	2.665	0.691	**	0.214
Age-squared	0.287	**	0.070	-0.016	**	0.004
Centre	0.090		0.098	0.010		0.070
South	0.087		0.081	0.110	**	0.033
Year of birth	0.073	**	0.021	0.036	**	0.014
Constant	-33.024		50.76	-77.856	**	27.901

Table 2. Enrolment

Graduation is modeled by means of a constant probability in the relevant age brackets for secondary school, and with a linear term in age for university. We use the same coefficients for graduation as in Leombruni and Richiardi (2006), estimated on LFS 1993-2003 data (Figure 2).

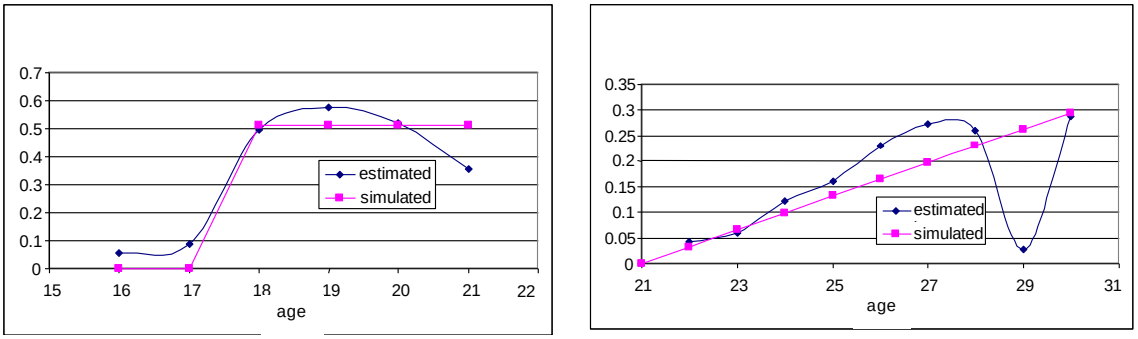

Figure 2. Probability of graduating, high school (left panel) and university (right panel). Source: Leombruni and Richiardi (2006)

5.4 Unemployment and male participation

The unemployment status at time *t* is modeled as function of lagged unemployment status, age, educational level, area of residence, the status of student at time *t*-1 and the overall unemployment rate.

Male participation at time *t* is modeled as a function of lagged participation, age, year of birth, educational level, area of residence and the status of student at time *t*-1. In both cases, the model estimation is conditioned on not being retired or student. We first estimate standard probit models. To account for unobserved heterogeneity and solve the initial conditions problem, we then estimate dynamic

random effects models (see section 2 for details). The estimated coefficients and standard errors are shown in **Errore. L'origine riferimento non è stata trovata.**. To compare the probit coefficients with those from the random effects estimators, the latter need to be multiplied by an estimate of $1/\sqrt{(1+\sigma_a^2)}$, where σ_a represents the size of the unobserved heterogeneity (see Arulampalam, 1999). Allowing for the different normalizations, the scaled estimate of the coefficient on lagged participation (unemployment status) is 0.5 (0.7), really less than the pooled probit estimate. In both cases, participation today increases the probability of participate tomorrow in the labor market. The relationship between male participation and age is an inverse U: participation initially increases with age, than it slows down and starts to decrease close to the retirement age. A higher education increases male participation. Living in the South decreases male participation.

	Unemployment						**Male Participation**					
	Probit			Dynamic r.e. model			Probit			Dynamic r.e. model		
	Coef.		Robust SE	Coef.		SE	Coef.		Robust SE	Coef.		SE
Lparticipate	---		---	---		---	1.766	**	0.052	0.799	**	0.077
Lag(unemployment)	1.816	**	0.026	1.133	**	0.038	---		---	---		---
Female	0.343	**	0.021	0.551	**	0.047	---		---	---		---
Age	-0.133	**	0.006	-0.245	**	0.015	0.130	**	0.011	0.326	**	0.026
Age2	0.001	**	0.000	0.002	**	0.000	-0.002	**	0.000	-0.004	**	0.000
High education	-0.266	**	0.038	-0.600	**	0.080	0.287	**	0.066	0.341	**	0.115
Medium education	-0.254	**	0.022	-0.395	**	0.044	0.110	**	0.032	0.170	**	0.058
Centre	0.288	**	0.035	0.491	**	0.075	-0.054		0.046	-0.104		0.088
South	0.764	**	0.028	1.478	**	0.063	-0.155	**	0.035	-0.313	**	0.069
Year of birth	---		---	---		---	-0.006		0.007	0.010		0.014
Lag(student)	1.165	**	0.043	0.504	**	0.116	0.205	**	0.061	-0.026		0.135
Unemployment rate	14.118	**	1.842	13.234	**	3.308	---		---	---		---
_cons	-0.253		0.185	1.334	**	0.389	9.376		13.075	-24.420		27.245
σ_a	---		---	1.263			---		---	1.071		

Table 3.Unemployment and male participation estimates

5.5 Female labor market participation and household formation

We estimate two separate equations, one for the choice of participation in the labor market and one for the choice of living in consensual union.[12] Reflecting our interest in uncovering the presence of dynamic

[12] Female labor force participation and the choice of living in consensual union may be correlated. Since the aim of this paper is to show that ignoring unobserved heterogeneity leads to wrong results in the microsimulation model, for simplicity we do not consider the issue of the joint determination of female participation and marital status. However,

spillover effects from participation to marriage, and from marriage to participation, the equations for female labor market participation and household formation also include cross-effect lagged variables: lagged marriage dummy is included in the female participation equation and lagged female participation is included in the estimation of the probability of being married. Cross-effect lagged variables are assumed weakly exogenous. Therefore, both the participation and the consensual union status at time *t* are modeled as a function of lagged participation, lagged consensual union status, the existence of children aged under three at *t*-1, age, year of birth, educational level, area of residence and the status of student at time *t*-1. Moreover, the model estimation is conditioned on not being retired or student. Dummies for the area of residence capture also regional differences in the availability of childcare and other (local) institutional factors. In order to simplify the model and keep a dichotomous participation outcome variable we do not explicitly model work hours. This implies that we do not consider the availability of part-time. Since the share of female part-time employment in the total employment has increased almost linearly since 1990, from about 10% to about 30% with few regional differences (the share of female part-time employment is 29.0% in the North, 29.4% in the Centre and 25.3% in the South), this increase is catched by the cohort effect.

As a benchmark, the estimates of the standard pooled probit models are reported in **Errore. L'origine riferimento non è stata trovata.**, columns 1 and 2. Then, column 3 and 4 report the coefficients and standard errors of the dynamic probit model with random effects. As explained above, the random effects probit and pooled probit models involve different normalizations. The scaled estimate of the coefficient on lagged participation (cohabitation) is 0.9 (2.2), really less than the pooled probit estimate. This indicates that omitting permanent unobserved heterogeneity leads to overestimation of state dependence. There is a lot of heterogeneity that cannot be accounted for by the explanatory variables: the estimated σ_a is equal to 1.2 (0.89). Instead, the estimated coefficients on the other covariate are similar (sometime larger) than the pooled estimates.

Focusing on the participation equation, we find that the coefficient on the lagged participation status in positive and statistical significant: participation today increases the probability of participate tomorrow in the labor market. Instead, the coefficients on the lagged cohabitation and on the lagged presence of children aged under three are negative: either living in consensual union or having small children reduce the future probability of participating in the labor market. The level of education (high and medium) seems to significantly increase the probability of participating in the labor market. Living in the Centre and South of Italy is associated with lower activity rates.

---

this issue needs to be discussed in future research and it can be treated extending the dynamic random effects model allowing for correlation in the error terms (see Alessie et al., 2004; Devicienti and Poggi, forthcoming).

Quite obviously, we find that living in consensual union in the previous year strongly increases the probability of living in consensual union in the current year. The chances to living in consensual union increase when the woman has young children. The relationship between living in consensual union and age takes the form of an inverse U. The level of education and the area of residence do not significantly affect the probability of living in consensual union. Instead, being a student the year before reduce the probability of living in consensual union in the current year.

Females participation	Probit			Random effects model		
(females)	Coef.		Robust SE	Coef.		SE
Lag(participation)	2.387	**	0.027	1.417	**	0.038
Lag(union)	-0.417	**	0.027	-0.595	**	0.049
Lag(children under 3)	-0.159	**	0.032	-0.197	**	0.050
Age	0.038	**	0.007	0.087	**	0.015
Age2	-0.001	**	0.000	-0.001	**	0.000
High education	0.808	**	0.047	1.665	**	0.093
Medium education	0.371	**	0.021	0.775	**	0.044
Centre	-0.116	**	0.027	-0.355	**	0.064
South	-0.270	**	0.021	-0.738	**	0.052
Year of birth	0.000		0.004	0.009		0.008
Lag(student)	0.549		0.056	0.272	*	0.121
Constant	-1.326		7.845	-19.025		16.684
σ_a				1.218		
Union	Probit			Random effects model		
females)	Coef.		Robust SE	Coef.		SE
Lag(participation)	-0.049		0.033	-0.093		0.052
Lag(union)	3.795	**	0.043	3.010	**	0.075
Lag(children under 3)	0.347	**	0.098	0.516	**	0.122
Age	0.074	**	0.010	0.263	**	0.036
age2	-0.001	**	0.000	-0.003	**	0.000
High education	-0.001		0.060	-0.170		0.093
Medium education	-0.029		0.032	-0.135	*	0.057
Centre	0.052		0.042	0.145	*	0.073
South	0.015		0.031	0.003		0.053
Year of birth	0.002		0.007	-0.026		0.014
Lag(student)	-0.530	**	0.066	-0.706	**	0.090
Constant	-7.667		13.146	44.404		28.285
σ_a				0.890		

Table 4. Females participation and household formation estimates

Finally, we estimate the probability of having a child at time *t*. The latter is modeled as a function of the existence of children aged under three at time *t*-1, the number of children aged under 16 at time *t*-1,

dummies about the labor market status at time *t*-1 (participation, unemployment, being student), age, educational level, area of residence and the fertility rate. Moreover, the model estimation is conditioned on living in consensual union. Estimates are reported in **Errore. L'origine riferimento non è stata trovata.**. The probability decreases if in the household there are already children under three. The larger is the number of children under 16 in the household, the lower is the probability of having a newborn. The latter initially increases and then decreases with age. Having high education increases the probability of having a child. No significant geographical differences are found.

Born	Estimates ECHP		
	Coef.		SE
Lag (children uneder 3)	-0.403	**	0.057
Lag (No. Children under 18)	-0.301	**	0.041
Lag (participation)	-0.069		0.046
Lag (unemployment)	-0.095		0.074
Lag (student)	-0.276		0.155
Age	0.357	**	0.044
Age2	-0.007	**	0.001
High education	0.197	**	0.074
Medium education	0.031		0.043
Centre	0.065		0.061
South	0.079		0.073
Fertility rate	12.499	*	5.105
Constant	-6.061	**	0.746

Note: the sample includes females aged 17-45.

Table 5. Birth probability estimates

6. The microsimulation model

Our model is a discrete-time dynamic population-based microsimulation of labor supply, with an open population: flags are switched on and off for partners and children for the female population, but no simulated individuals are actually matched.

The microsimulation is comprised of four modules: Demography, Education, Household formation and Employment. The overall structure of the microsimulation is depicted in Figure 3.

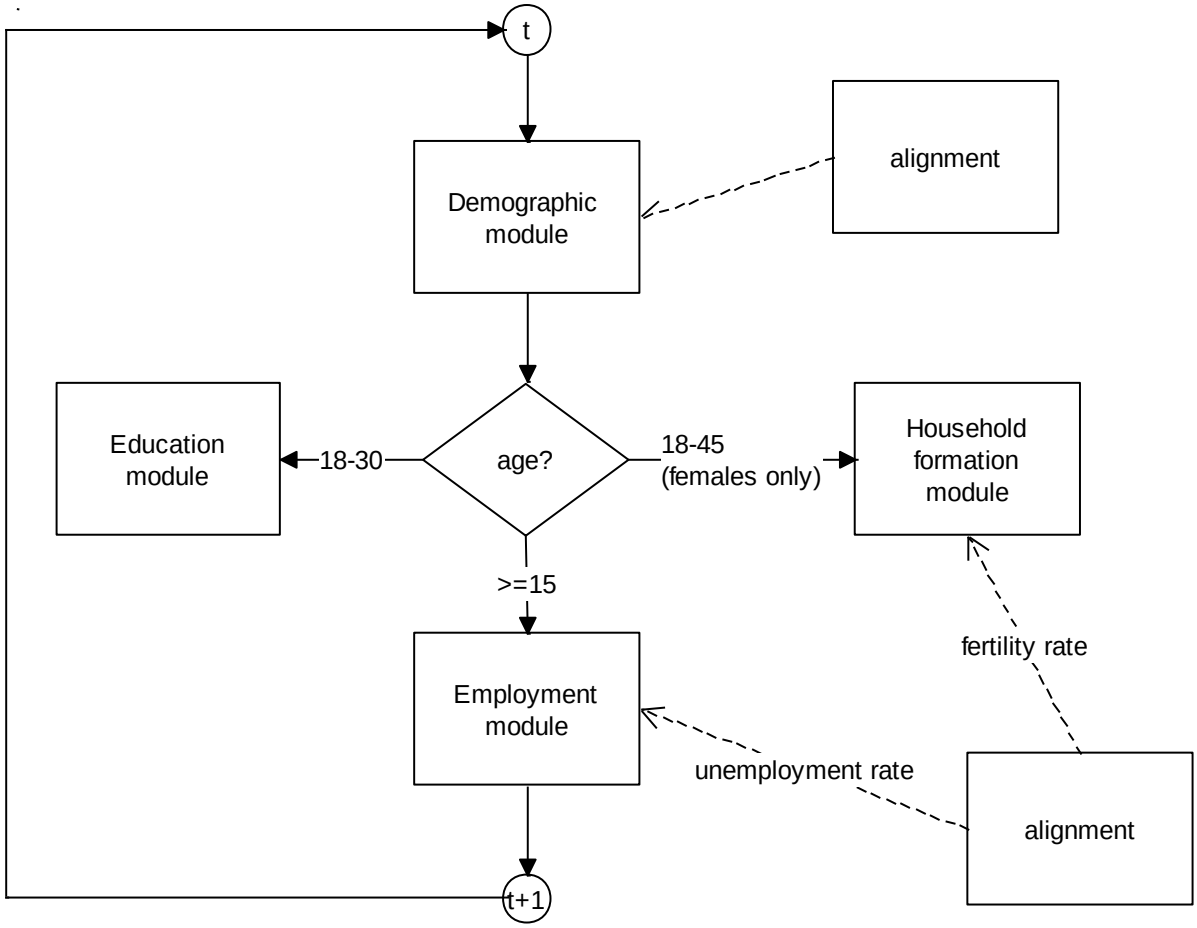

Figure 3. Structure of the microsimulation model

6.1 Demographic module

Population is aligned to official demographic projections by year, age, gender and macro-area of living (north, centre and south of Italy). Whenever the population is over-represented in a given age, gender and area cell, simulated individuals are killed at random. Whenever the population is under-represented, new individuals are created by cloning at random existing individuals in the same cell. We do not model (internal nor external) migration.

6.2 Education module

We separately model enrolment and graduation, for both secondary school and university. Individuals enter the simulation at 17, after completion of compulsory education. Dropouts at 16 are modelled by aligning the initial status (in education, in the labor force, in employment) to the observed frequencies (see section 6.2 below). Dropouts from school exit the education module and enter the labor market module. Students can graduate from secondary school starting at age 18. Those who fail to

graduate before age 22 exit the education module and enter the labor market module. Secondary school graduates can enter university. University participation is allowed until age 30, while graduation can take place beginning at age 21. University dropouts leave the education module and enter the labor market module. We make the simplifying assumption that people never go back to education, once they have left. This is justified, as already discussed, by the very small number of students of older ages.

The detected (linear) trends toward higher high school participation are stopped for individuals born in 1990 or later, while those toward higher university participation are stopped for individuals born in 1985 or later (we prudentially assume all trends have already come to an end in the base year).

The flowchart of the Education module is represented in Figure 4.

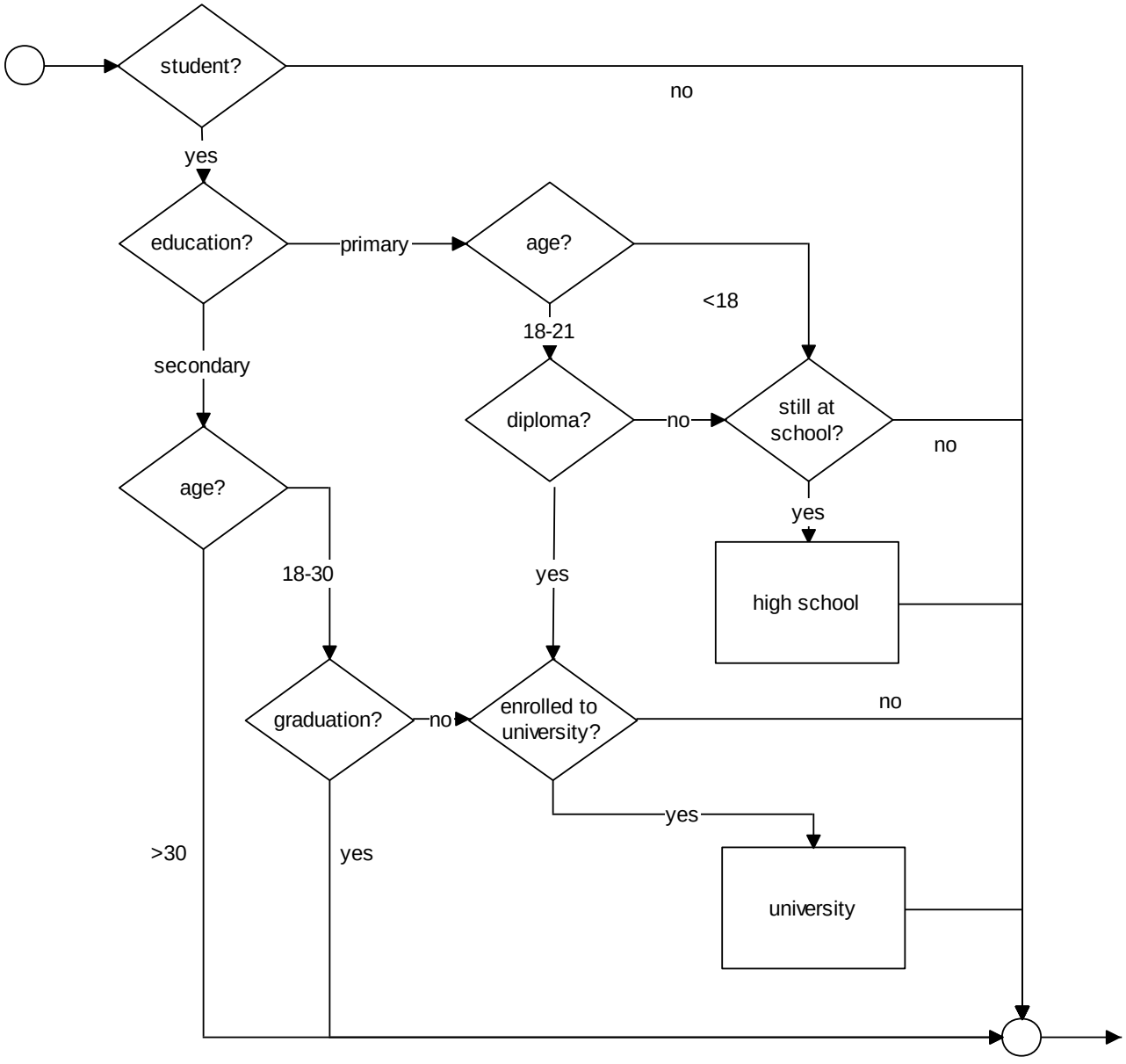

Figure 4. The Education module.

6.3 Household formation module

Given that the presence of a partner is not relevant, at a first order approximation, for male labor market participation, the household formation module is only applied to female.[13] It is comprised of two sub-modules: Living in consensual union and Maternity. Women aged 18 or older and who are not student can enter a consensual union. Note that at young ages living in consensual union is likely to be a choice, while at older ages it might reflect a partner's death, hence a state of widowhood. The (linear) cohort effect in the equation for living in consensual union is stopped for individuals born in 1990 or later. Women aged 18-45 who live in a consensual union can have children. The total amount of births in each year is aligned with demographic data. The flowchart of the Household formation module is represented in Figure 5.

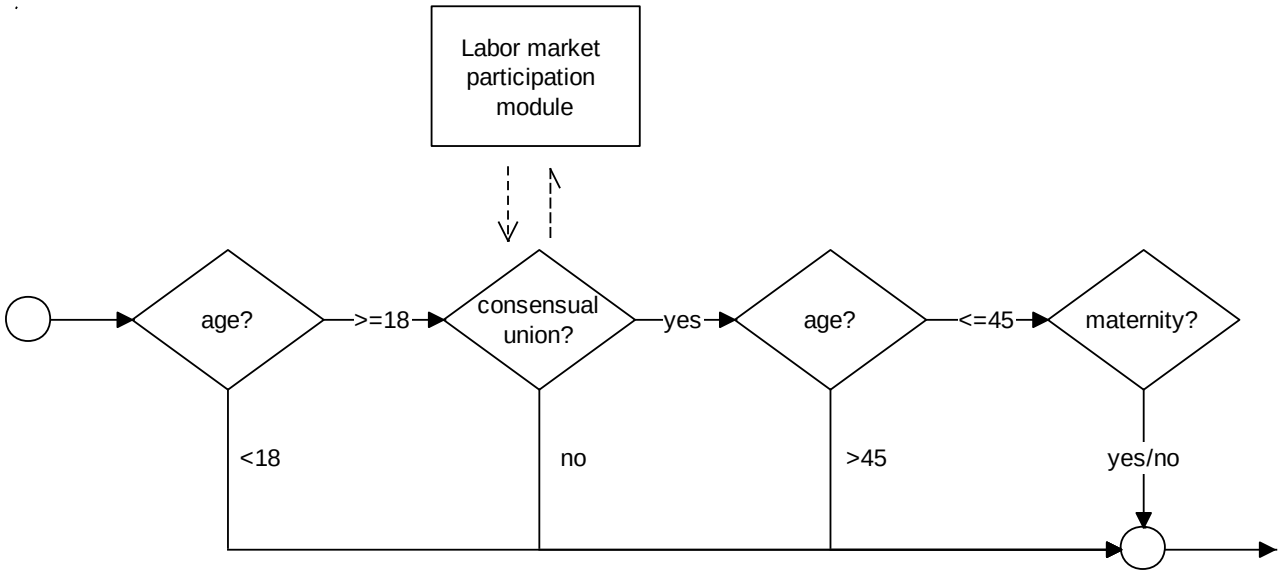

Figure 5. The Household formation module (females only).

6.4 Employment module

The labor market module is applied to all individuals who are not in education or retired (remember individuals enter the simulation above the minimum working age). It is composed of two sub-modules: Labor market participation and Unemployment.

Labor market participation is modeled separately by gender, and the model for females is conditional on household composition. The (linear) cohort effect in the equations for labor market participation are stopped for individuals born in 1990 or later. Alignment with an exogenous trend is

[13] Living in union might well affect the decision about how much to work, and possibly wages; however, we do not model work hours, nor wages.

performed for the overall unemployment rate. Consequently, the Unemployment module is to be regarded as a model of unemployment rate differentials, rather than unemployment rate levels. This is justified by the fact that we do not model the demand side: the overall unemployment rate is therefore considered as an exogenous parameter of the simulation. To get away with business cycle considerations, this exogenous unemployment rate is kept constant at 7.5%. The flowchart of the Employment module is represented in Figure 6.

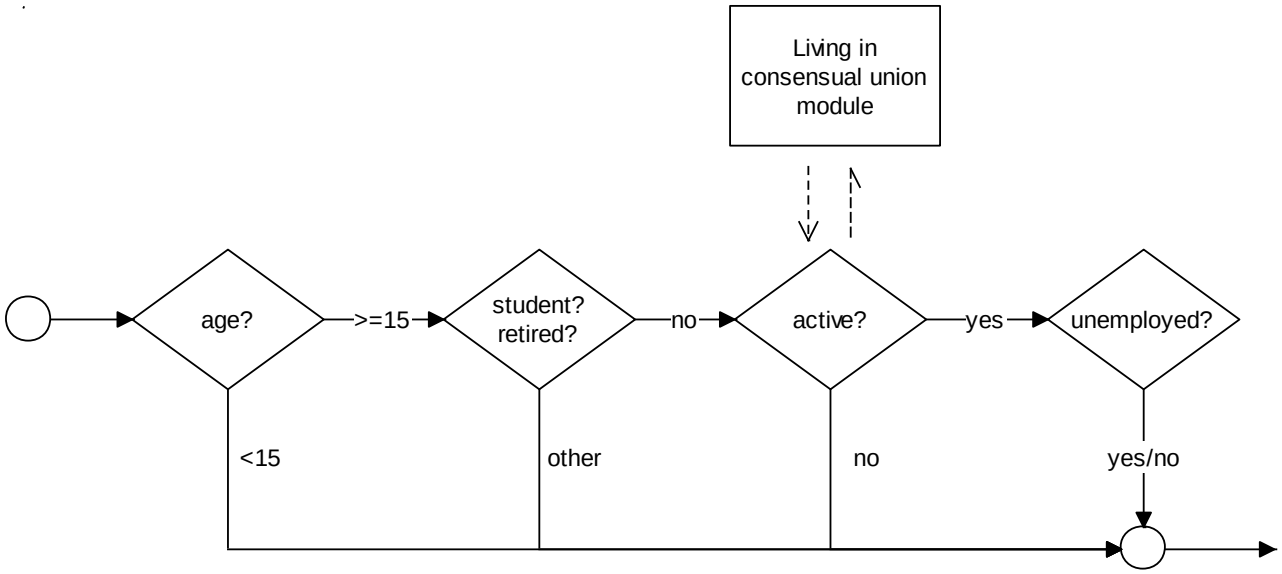

Figure 6. The Employment module.

7. Microsimulation results

Figure *7* points out how differences in the estimates used as inputs may lead to large differences in the findings. Depending on the estimation method producing these inputs (standard probit model versus dynamic probit model with random effects), we find significant differences in the female participation rates. The latter differences become larger over time: at the end of the simulation process, female participation rates result more than 5 points percent higher if robust estimates are used as input. This results can be explained observing that omitting permanent unobserved heterogeneity leads to an overestimation of true state dependence and, therefore, to lower participation rates stronger constrained by past experiences.

In the following, we illustrate shortly the simulated evolution of female participation in Italy obtained after controlling for unobserved heterogeneity. In recent decades women have been given the same opportunities as men in education, and to some extent, in the labor market. However, female participation

rate is still low in Italy: it is currently below 60% and it is expected to remain under 70% for many decades (Figure 7, which shows the difference in the projected activity rates in the age groups 17-65 and 17-45 with the two estimation strategies).

From Figure 8 onwards we focus exclusively on the results obtained by using dynamic probit models with random effects estimates, our preferred specification. Most dynamics will take place in the two decades to 2025, when the baby boom generation will have moved to retirement age. Figure 8 compares the activity rate for young adult men with that for young adult women, with and without children (in the age group 17-45, having completed education). Activity rates in these groups are higher, even if the female participation rate is still currently below 75%, as compared to more than 95% for young adult men. The gender gap between young adult men and women is actually larger than 20 points percent and it will only partially decrease in the medium run. Figure 8 also shows that having children lowers the participation rate for young adult women; however, the penalization for children is expected to decrease over time, as the activity rate for women without children is projected to remain roughly constant at around 85%, while the activity rate for women with children will grow from 60% to 80% by 2025. The penalization for the number of children is also expected to shrink considerably (

Figure *9*).

Figure 10 shows that the differentials in female participation rate by educational level are expected, if any, to widen further, suggesting that a main driver in the increase in the female participation rate is the increase in the fraction of the female population with a high education.

Finally, Figure 11 shows the differences in the projected participation rates for young adult women by region and household composition. Activity rates for women without children are projected to remain constant, while the already documented increase in the participation rates of women with children will be stronger in the South.

From the factors illustrated above emerges a complex picture which points to some convergence in the activity rates of men and women, and in the activity rates of women with and without children, though the transition will be incomplete and slow, with respect to the one needed to meet the Lisbon 2010 (not to speak about Europe 2020) targets. The changes are mainly due to an increase in the educational levels of women with children. This is coherent with a literature suggesting that education is able to break traditional roles and increase female participation. For example, couples with higher education have a more even division of household labor compared to those with lower levels of education (Gershuny and Robinson, 1988; Mencarini and Tanturri, 2004). Moreover, highly educated women have higher opportunity costs or “more to lose” if they do not participate in the labor market and are able to outsource domestic tasks easily.

Is it possible to speed up the convergence process? Institutions and related policies that support women and men to achieve work-life balance can also help in promoting female participation in the labor market. [14] Appropriate policy may modify cultural norms surrounding working women and help women in combining work and family commitment, promoting participation. Also, homogeneity in the availability, affordability and use of childcare over all Italian regions, which we keep constant throughout the simulation, will likely end up in higher female participation rates.

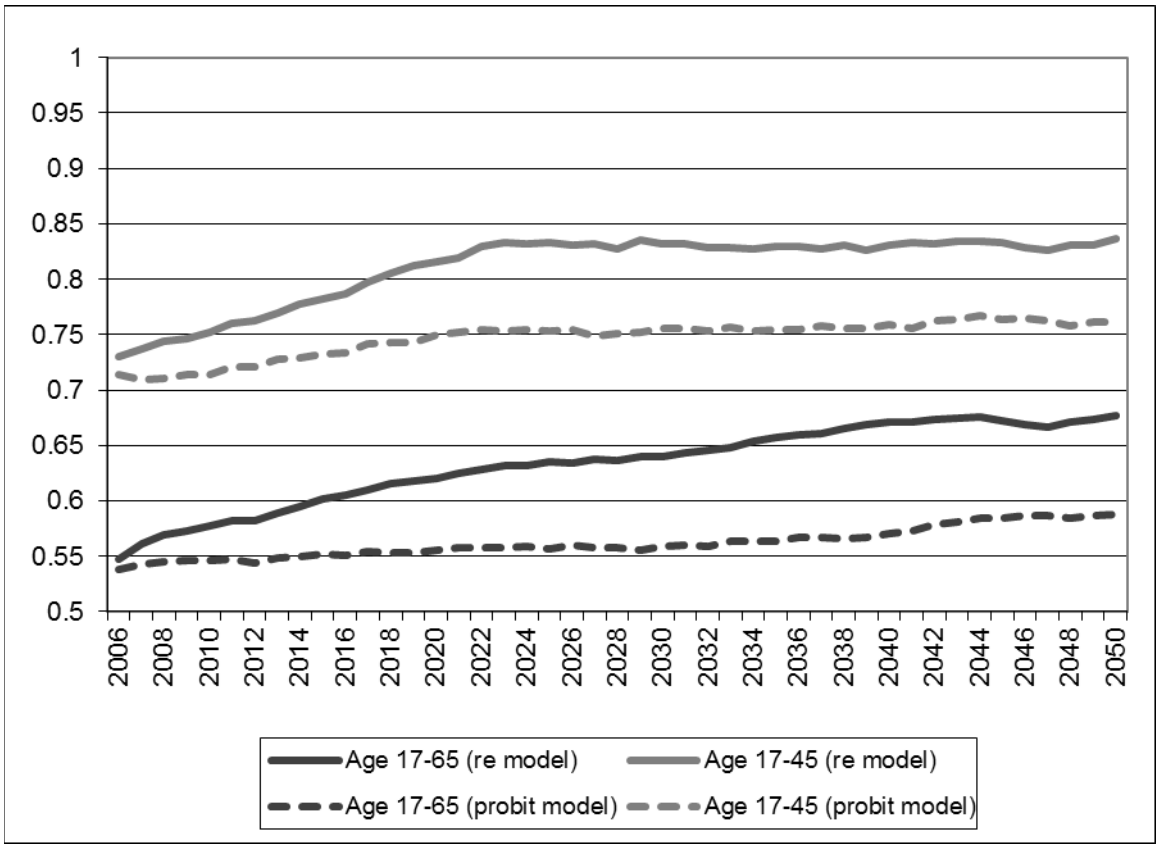

Figure 7. Female participation rates (students excluded)

[14] Institutions and related policies relevant for this topic include: tax systems and regulations, employment regulations and specifically in relation to flexible or part-time work, level and acceptability of working women and mothers, contraceptive availability and acceptance, and childcare legislation, affordability and availability.

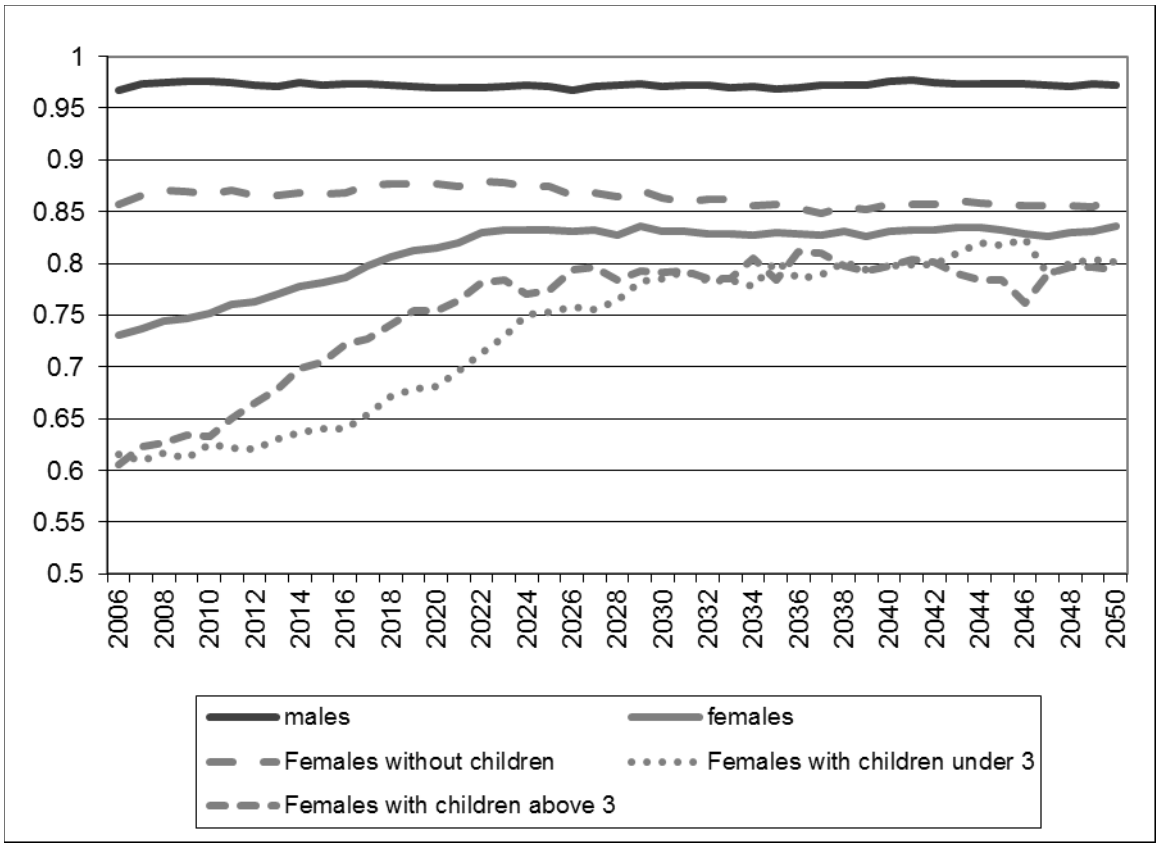

Figure 8. Participation rates by gender (individuals aged 17-45) excluding students

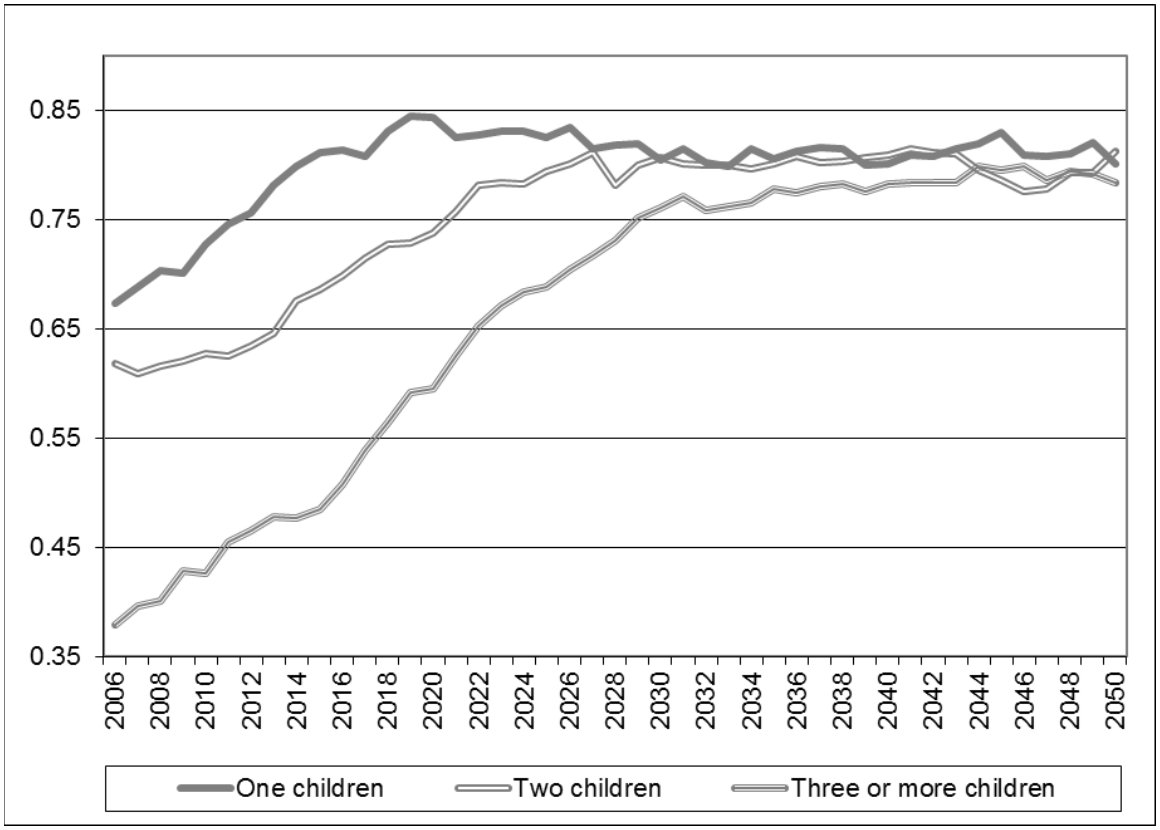

Figure 9. Participation rates: females (aged 17-45) with children under 18 - excluding students

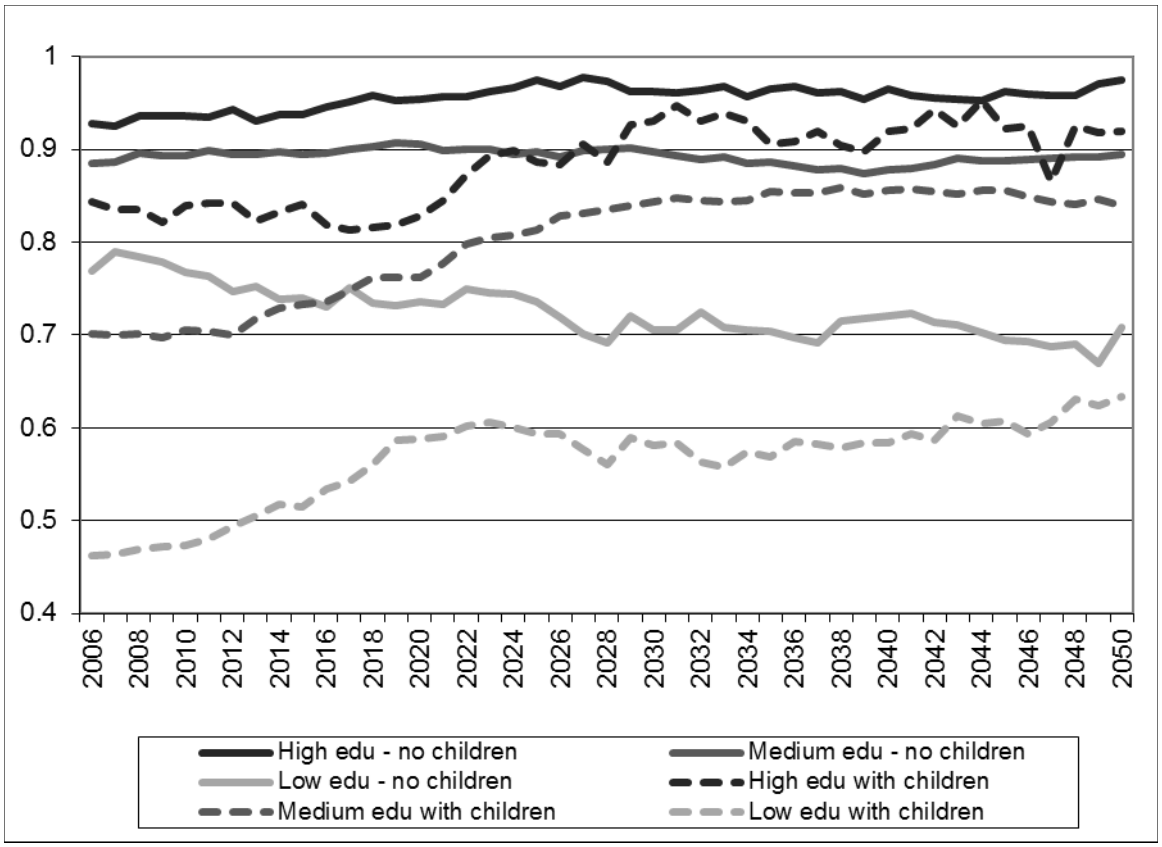

Figure 10. Females participation rates by education: individuals aged 17-45 excluding students

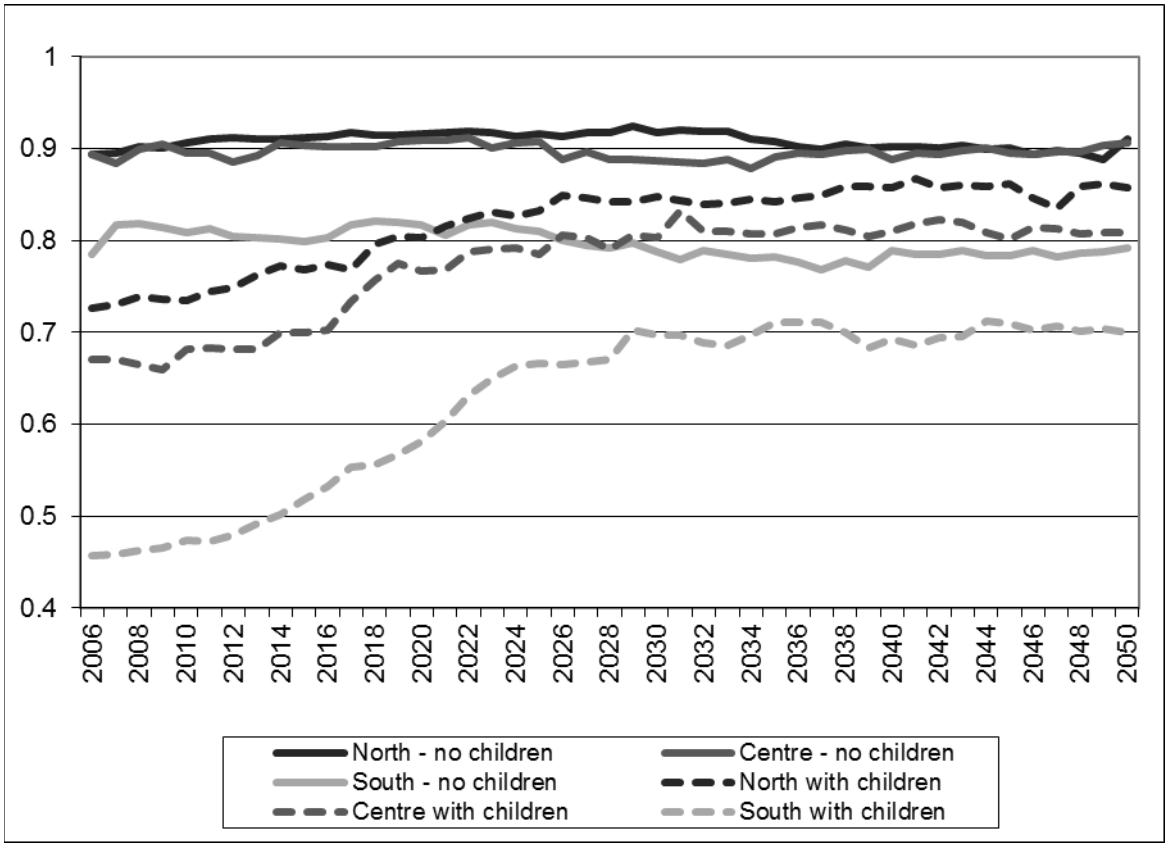

Figure 11. Females participation rates by area: individuals aged 17-45 excluding students

8. **Conclusions**

In this paper we have documented the existence of a marked though slow process of convergence in the activity rates of different subgroups of the Italian population: between men and women, between women with children and women without children, between the North and the South of Italy. This process would have been significantly underestimated should we have not accounted for unobserved heterogeneity in labor market participation and household formation, by means of dynamic probit models with random effects. This is because omitting permanent unobserved heterogeneity leads to an overestimation of true state dependence and, therefore, to lower participation rates stronger constrained by low past participation rates.

The results of this paper strongly suggest that misspecification can be an important source of bias in dynamic microsimulation, even when the model is used only for predicting future trends and not for policy evaluation and "what-if" scenario analysis.

References

Aaberge, R., J.K. Dagsvik and S. Strøm (1995): "Labor Supply Responses and Welfare Effects of Tax Reforms", *Scandinavian Journal of Economics*, 97(4), pp. 635-659.

Ahn, S.C., Schmidt, P. (1995): Efficient estimation of models for dynamic panel data. Journal of Econometrics 68, pp. 5–28 .

Anderson, J.M. (2003): *Models for Retirement Policy Analysis*, report prepared for the Society of Actuaries

Anderson, T.W., Hsiao, C. (1982): "Formulation and estimation of dynamic models using panel data". *Journal of Econometrics* 68, pp. 5–27.

Anxo, D., L. Flood, L. Mencarini, A. Pailhé, A. Solaz, and M.L. Tanturri (2007): "Time Allocation between Work and Family Over the Life-Cycle: A Comparative Gender Analysis of Italy, France, Sweden and the United States". *IZA Discussion Paper* No. 3193 (November)

Arellano, M. (2003): "Discrete choice with panel data", *Investigaciones Económicas* XVII (3), pp. 423–458

Arellano, M., Bond, S.R. (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". *Review of Economic Stud*ies 58, pp. 277–297 .

Arellano, M., Bover, O. (1995): Another look at the instrumental variables estimation of errorcomponent models. Journal of Econometrics 68, pp. 29–51.

M. Arellano and J. Hahn (2007): "Understading bias in nonlinear panel models: Some recent developments". In: R. Blundell, W.K. Newey and T. Persson, Editors, *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Vol. 3*, Cambridge University Press, Cambridge .

Arulampalam, W. (1999): "A Note on estimated effects in random effect probit models", *Oxford Bulletin of Economics and Statistics*, 61(4), pp. 597-602.

Arulampalam, W. and Steward, M.B. (2007): "Simplified Implementation of the Heckman Estimator of the Dynamic Probit Model and a Comparison with Alternative Estimators", *IZA discussion paper* No. 3039, Institute for the Study of Labor, Bonn

Arulampalam, W., Booth A.L., Taylor, M.P.: Unemployment persistence. *Oxf. Econ. Pap*. 52, pp. 24–50 (2000)

Bertola, G., Jimeno, J.F., Marimon, R., Pissarides, C. (2001): "Welfare Systems and Labor Markets in Europe: what convergence before and after EMU?" in in Bertola, G., Boeri, T. and Nicoletti, G. (edited by) Welfare and Employment in a United Europe, MIT Press.

Blundell, R., Bond, S. (1998): "Initial conditions and moment restrictions in dynamic panel data models". *Journal of Econometrics* 87, pp. 115–143

Bratti et al. (2005):

Carro, J.M. (2007): "Estimating dynamic panel data discrete choice models with fixed effects", *Journal of Econometrics*, Volume 140, Issues 2, pp.503-528.

Casadio et al. (2008):

Cox, D.R. and N. Reid (1987): "Parameter orthogonality and approximate conditional inference", *Journal of the Royal Statistical Society, Series B* 49, pp. 1–39

Cramer, J.S: (2005): "Omitted Variables and Misspecified Disturbances in the Logit Model", *Tinbergen Institute Discussion Paper* TI 2005-084/4

Del Boca, D. (2002): The effect of child care and part time opportunities on participation and fertility decisions in Italy, *Journal of Population Economics*, 15(3), pp. 1432-1475.

Del Boca (2003):

Del Boca, D., Aaberge, R., Colombino U., Ermish, J., Francesconi, M., Pasqua, S. and Strom, S. (2006): "Labour market participation of women and fertility: the effect of social policies", mimeo

Del Boca, D. and D. Vuri, (2007): The mismatch between employment and child care in Italy: the impact of rationing, *Journal of Population Economics*, 20 (4), pp. 805-832.

Gershuny, J. and Robinson, J.P. (1988): Historical Changes in the Household Division of Labor, *Demography*, 25(4), pp. 537-552.

Haan, P. (2006): "Slowly, But Changing: How Does Genuine State Dependency Affect Female Labor Supply On The Extensive and Intensive Margin", *JEPS Working Papers* No. 06-002, JEPS

Hahn, J. (2001): The information bound of a dynamic panel logit model with fixed effects, *Econometric Theory* 17, pp. 913–932

Hahn, J. (1999): "How informative is the initial condition in the dynamic panel data model with fixed effects?" *Journal of Econometrics* 93, pp. 309–326.

Heckman, J. J. (1981a): “Heterogeneity and state dependence”, in S. Rose (ed.), *Studies inLabor Markets*, Chicago Press, Chicago, IL.

Heckman, J. J. (1981b): “The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process”, in C. F. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with EconometricApplications*, MIT Press, Cambridge, MA, pp. 114-178.

Hoderlein, S., Mammen, E. and Yu, K. (2011): “Nonparametric models in Binary choice fixed effects panel data”, *Econometrics Journal*, forthcoming

Honoré, B., and Tamer, E. (2004): “Bounds on parameters in dynamic discrete choice models”. *CAM Working paper* 2004-23, University of Copenhagen

Honore, B.E., Kyriazidou, E. (2010): “Panel data discrete choice models with lagged dependent variables”. *Econometrica* 68, pp. 839–874.

Honore, B.E. (2002): “Non-linear models with panel data. Centre for Microdata Methods and Practice”, *Working paper CWP* 13/02, the Institute for Fiscal Studies, London .

Honore, B.E. (1993): “Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables.” *Journal of Econometrics* 59(1–2), pp.35–61.

Hsiao, C. (1986): *Analysis of Panel Data*. Cambridge University Press, Cambridge.

Lo Conte, M. and S. Prati. (2003): “Maternità e partecipazione femminile al mercato del lavoro. Un’analisi della situazione professionale delle neo-madri” paper presented at Workshop Cnel-Istat on *Motherhood and female participation to job market, between constraints and re-conciliation strategies*, Rome

Mencarini, L. and M.L. Tanturri (2006): “High Fertility or Childlessness: Micro-Level Determinants of Reproductive Behaviour in Italy”, *Population*, 61 (4), pp. 389-416.

Orme, C. D. (2001): “Two-step inference in dynamic non-linear panel data models”, mimeo, University of Manchester.

Pacifico, D. (2009): "On the role of unobserved preference heterogeneity in discrete choice models of labor supply," *MPRA Paper* 19030, University Library of Munich, Germany

Rondinelli and Zizza (2011):

Sistan – Sistema Statistico Nazionale (2008): “L’Università in cifre 2007”, Ministero dell’Università e della Ricerca, Roma: Roroform s.r.l

Sistan – Sistema Statistico Nazionale (2006): “La dispersione scolastica: indicatori di base per l’analisi del fenomeno. Anno Scolastico 2004-2005”, Ministero dell’Università e della Ricerca

Val, F. (2009): “Fixed effects estimation of structural parameters and marginal effects in panel probit modles”, *Journal of Econometrics*, 150(1), pp. 71-85.

Van Soest, A. (1995): “Structural Models of Family Labor Supply: A Discrete Choice Approach”, *Journal of Human Resources*, 30, pp. 63-88.

Woolridge, J. (2002):

Wooldridge, J. (2005): “Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity”, *Journal of Applied Econometrics*, 20, pp. 39-54.

Yatchew and Griliches (1985):