Working papers series

WP ECON 10.11

Unobserved Heterogeneity in Multi-Spell Discrete Time Duration Models

Raquel Carrasco (U. Carlos III de Madrid)

J. Ignacio García Pérez (U. Pablo de Olavide, FEDEA & FCEA)

JEL Classification numbers: J64, J61, C23, C41, J65

Keywords: Duration models; Discrete choice; Multiple spells; Unobserved heterogeneity; Unemployment

Department of Economics

Unobserved Heterogeneity in Multi-Spell Discrete Time Duration Models*

Raquel Carrasco†
Departamento de Economía
Universidad Carlos III de Madrid

J. Ignacio García Pérez
U. Pablo de Olavide, FEDEA & FCEA

October 2010

Abstract

This paper considers the estimation of discrete time duration models. We highlight the enhance identification opportunities embedded in multiple spell data to separately identify the effect of duration dependence and individual time invariant unobserved heterogeneity. We consider two types of models: (i) random effects models specifying a mass point distribution for the unobserved heterogeneity; and (ii) fixed effects models in which the distribution of the effects is left unrestricted. The availability of multiple spell data allows us to consider this type of models, in the spirit of fixed effects discrete choice panel data models. We study the finite sample properties of different estimators for previous models by means of Monte Carlo simulations. Finally, as an empirical illustration, we estimate unemployment duration models using Spanish administrative data with information on the entire labor history of the individuals.
JEL classification: J64, J61, C23, C41, J65

Keywords: Duration models; Discrete choice; Multiple spells; Unobserved heterogeneity; Unemployment.

*We are grateful to Namkee Anh, Manuel Arellano, J.C. Berganza, J. Francisco Jimeno, Sergi Jiménez, Jan C. van Ours, and participants at the XXII ESPE Conference, and U. Malaga and FEDEA workshops for helpful suggestions and comments on this work. All remaining errors are our own. Research funding from the Spanish Ministry of Science and Innovation, Grants No. ECO2009-11165 and SEJ200604803 is acknowledged.
†Corresponding author. Tel.: +34 91 624 9583; *E-mail address:* rcarras@eco.uc3m.es.

1 Introduction

Discrete time duration models have received a great deal of attention in the literature. One of the main areas of research where this type of models has been used is in the econometric analysis of individual unemployment spells. This strand of the literature has focused on the estimation of the duration dependence and on the effect of unemployment benefits over the unemployment hazard rate. Within this context the distinction between what has been called "true" and "spurious" duration dependence is crucial (see Heckman, 1991). It is well known that improper treatment of unmeasured variables is likely to bias the estimated effect of unemployment benefits on the exit rates and to introduce spurious negative duration dependence in the hazard rate.

The basic motivation of this paper is to facilitate the distinction between unobserved heterogeneity and true duration dependence in the estimation of unemployment hazard rates. Microeconometric studies typically analyze this issue by using one single spell of unemployment per individual. This type of data would identify the effect of interest relying on assumptions about the distribution of the unobserved effects, which can be rather restrictive. This paper highlights the enhanced identification opportunities embedded in data with multiple spells for each individual. The key to disentangle unobserved heterogeneity from genuine duration dependence is that one can use information on more than one spell of unemployment for the same individual (see Abbring and Van den Berg, 2003).

We estimate two types of unemployment duration models: (i) random effects models specifying a mass point distribution for the unobserved heterogeneity (see Heckman and Singer, 1984); and (ii) fixed effects models in which the distribution of the effects is left unrestricted. We study the finite sample properties of different estimators for previous models by means of Monte Carlo simulations. We perform several Monte Carlo designs which differ in terms of the true unobserved heterogeneity distribution and of the number of spells available for each individual.

We first estimate random effects models using only individual's unemployment spells and approximating the heterogneity distribution by means of a discrete distribution.[1] Secondly, we allow for a more flexible specification of the unobserved heterogeneity and estimate random effects models using the employment and unemployment spells available in the worker's labor history, assuming a joint discrete distribution for the unobserved heterogeneity in each state.

With multiple spell data, and under the assumption that the unobserved individual component is the same for different spells, one can also estimate a fixed effects model in which the full distribution of the unobserved heterogeneity is left unrestricted and allowed to be dependent of the explanatory variables of the model. Following Frederiksen, Honoré and Hu (2007), we estimate this model, in the spirit of the fixed effects discrete choice panel data models. Our Monte Carlo results suggest that the fixed effects approach may be an useful alternative to estimate discrete time duration models, specially given the computational burden of the random effects approach as the number of support points increases.

Finally, as an empirical illustration, we estimate unemployment duration models using Spanish administrative data with information on the entire labor history of the individuals. Specifically, we use data from the *Muestra Continua de Vidas Laborales* (MCVL), which contains information on the complete labor history of a sample of approximately $1,1$ million workers linked to Social Security within the period 2005-2008.[2] Our results highlight how the estimated effect of Unemployment Benefits over the exit rate from unemployment varies depending on the model used to control for unobserved heterogeneity. We find

[1] The performance of estimators which approximate the distribution of unobserved heterogeneity by means of a discrete distribution is studied by Huh and Sickles (1994), Baker and Melino (2000) and Gaure et al. (2007).

[2] Hansen (2000) and Kalwij (2004) also use multiple spell data to study individuals' unemployment experiences.

that the fixed effects model provides a more accurate estimate of the effect of Unemployment Benefits than the random effects model.

The paper is organized as follows. Section 2 presents the econometric models and estimators. In Section 3 we study the finite sample properties of the estimators by means of Monte Carlo simulations. In Section 4 we present the estimations for the unemployment duration model using Spanish data. Finally, Section 5 concludes.

2 Models and estimators

The starting point is the formulation of a duration model. At any point in time, an individual could be in any of two states: Unemployed or Employed. We estimate the probability that an individual will leave unemployment during next period, given that she has been unemployed for T periods. We treat duration (T) as a discrete variable. For individual i the probability of a spell being completed by time $t+1$ given that it was still continuing at time t is given by the following hazard rate:

$$h_i(t) = \Pr(T_i = t \mid T_i \geq t, b_i(t), x_i(t)) = F(\alpha_0 + \alpha_1(t)b_i(t) + \alpha_2(t)x_i(t) + \gamma(t)). \tag{1}$$

The analysis is conditional on $b_i(t)$, a dummy variable taking the value 1 if the individual receives unemployment benefits in period t, and on a vector of exogenous variables $x_i(t)$, which includes individual, sectorial and aggregate variables. $\gamma(t)$ is a parameter that captures duration dependence and is a function of the number of periods spent in unemployment. $\alpha_1(t)$ and $\alpha_2(t)$ are also functions of t and capture differential effects of the conditioning variables depending on the duration. Finally, $F(\cdot)$ denotes the logistic cumulative distribution function.

2.1 Single-spell duration data

We first consider the estimation of a single spell unemployment duration model which treats different spells for the same individual as independent. This would be a reasonable assumption in the absence of unobserved heterogeneity. Therefore, the number of spells in the sample is equal to the number of individuals times the number of spells available for each individual.

The log likelihood function for all spells of unemployment takes the form

$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{\bar{t}} u_{it} \left\{ (1 - y_{it}) \log(1 - h_i(t)) + y_{it} \log h_i(t) \right\}, \tag{2}$$

where N is the number of unemployment spells in the sample, $\bar{t}$ is the largest observed duration, y_{it} takes the value 1 if an exit from the spell of unemployment is observed in period t and 0 if not, or if the observation is censored at t. The variable u_{it} equals 1 if a spell of unemployment is observed during the period t and zero otherwise.

Maximum Likelihood (ML) estimates of previous model may be biased by the presence of unobserved heterogeneity. In that case, the duration dependence in the observed hazard function is more negative than otherwise since the individuals with the highest hazards on average leave unemployment quickest. A version of the model allowing for unobserved heterogeneity, η_i, is given by

$$\begin{aligned} h_i(t, \eta_i) &= \Pr(T_i = t \mid T_i \geq t, b_i(t), x_i(t), \eta_i) = \\ &= F(\alpha_0 + \alpha_1(t) b(t) + \alpha_2(t) x_i(t) + \gamma(t) + \eta_i), \end{aligned} \tag{3}$$

Again, assuming independence over the individual spells, the log-likelihood function is

$$\log L = \sum_{i=1}^{N} \int \sum_{t=1}^{\bar{t}} \left[m_{it} \left\{ (1 - y_{it}) \log(1 - h_i(t)) + y_{it} \log h_i(t) \right\} \right] d\mu(\eta), \tag{4}$$

where $\mu(\eta)$ is the unknown distribution of the unobserved heterogeneity.

As it is usual in this type of models, the initial time does not correspond to the date of entry into the labor market for all the individuals in the sample and it is possibly correlated with η_i. Consequently we have to consider the problem of initial conditions. In our application (see Section 4) we follow the approach proposed by Wooldridge (2005) which consists in modeling the unobserved heterogeneity conditional on the initial condition (and on the exogenous variables in all time periods) and to specify the unconditional distribution of unobserved factors.

The problem of how to control for the unobserved mixing distribution $\mu(\eta)$ in the likelihood function given in (4) has been addressed extensively in the literature (see Van den Berg, 2001). Standard approaches require making strong and arbitrary assumptions about distribution functions for population heterogeneity, η. A popular choice is the family of Gamma distributions. This stems from analytic tractability[3] although it suffers from the typical estimation bias due to an incorrect parametrization of $\mu(\eta)$.

Heckman and Singer (1984) proposed a semi-parametric approach to identify the unobserved distribution from a mixed distribution assuming that η_i is a random effect independent of the conditioning variables. Assuming that the random variable η_i is discrete with finite support given by r mass points $s_1, ..., s_r$, and the corresponding probability mass $\Pr(\eta_i = s_\ell) = P_\ell$, the likelihood in this case is

$$\log L = \sum_{i=1}^{N} \sum_{\ell=1}^{r} \sum_{t=1}^{\bar{t}} \left[m_{it} \left\{ (1 - y_{it}) \log(1 - h_i(t, s_\ell)) + y_{it} \log h_i(t, s_\ell) \right\} \right] \Pr(\eta_i = s_\ell), \tag{5}$$

where

$$h_i(t, s_\ell) = F(\alpha_0 + \alpha_1(t) b_i(t) + \alpha_2(t) x_i(t) + \gamma(t) + s_\ell). \tag{6}$$

The idea is that if the number of support points increases, then any true underlying distribution for the unobserved heterogeneity can be approximated well. Nonetheless, in practice it is often difficult to find more than a few different mass points. This fact reflects a lack of informativeness on the distribution of

[3] See Abbring and Van den Berg (2007) for a justification for the choice of the family of Gamma distributions.

the unobserved heterogeneity in the data, especially when only single spell data on durations are available.[4]

The availability of multiple spells for the same individual would enhance the identification of the parameters of interest within the random effects framework. Moreover, with multiple spells the individual unobserved heterogeneity can be ruled out, in the spirit of the fixed effects discrete choice panel data models. Next subsection outlines both methods for multiple spell data.

2.2 Multi-spell duration data

2.2.1 Random effects model

When several spells are observed for each individual, it is possible to allow for dependence across different types of spells for the same individual. Specifically, we can estimate jointly unemployment and employment durations assuming a joint distribution for the unobserved heterogeneity in each state. Therefore, accounting for the two states, unemployment (u) and employment (e), the model is defined by

$$h_i^k(t,\eta_i^k) = F(\alpha_0^k + \alpha_1^k(t)b_i(t)u_{it} + \alpha_2^k(t)x_i(t) + \gamma^k(t) + \eta_i^k), \qquad k = u, e, \quad (7)$$

where $u_{it} = 1$ if during the period t a spell of unemployment is observed and zero otherwise, η_i^u and η_i^e are discrete variables with finite support given by r location points each.

Assuming two location points for each state, (s_1^u,s_2^u) and (s_1^e,s_2^e), with a joint probability distribution, one has to estimate the location points and the corresponding joint probabilities: $P_{11} = \Pr(\eta_i^u = s_1^u, \eta_i^e = s_1^e)$, $P_{12} = \Pr(\eta_i^u = s_1^u, \eta_i^e = s_2^e)$, $P_{21} = \Pr(\eta_i^u = s_2^u, \eta_i^e = s_1^e)$, and $P_{22} = \Pr(\eta_i^u = s_2^u, \eta_i^e = s_2^e)$. Thus, the likelihood function is given by

$$\log L = \sum_{i=1}^{N}\sum_{l=1}^{2}\sum_{m=1}^{2} \log L_i(s_l^u, s_m^e) \Pr(\eta_i^u = s_l^u, \eta_i^e = s_m^e),$$

[4] See Gaure et al. (2007) for a deep insight to the usage of the support point approach.

where $\log L_i(s_l^u, s_m^e)$ takes the following form:

$$\log L_i(s_l^u, s_m^e) = \sum_{t=1}^{\bar{t}} \left\{ \begin{array}{c} [u_{it} \{(1 - y_{it}^u) \log(1 - h_i^u(t, s_l^u)) + y_{it}^u \log h_i^u(t, s_l^u)\}] + \\ [(1 - u_{it}) \{(1 - y_{it}^e) \log(1 - h_i^e(t, s_m^e)) + y_{it}^e \log h_i^e(t, s_m^e)\}] \end{array} \right\}. \tag{8}$$

2.2.2 Fixed-effects model

Multiple spell data allow to identify the model without imposing untestable assumptions of the unobserved heterogeneity distribution. In this case, the duration analysis becomes similar to the dynamic panel data analysis, where one can get rid of the so called "fixed-effects" which can be correlated with the explanatory variables. This is attractive since it ensures that the distribution of the individual effects does not play any role in identifying the parameters of interest. Moreover, within this approach we can obtain consistent estimates without making assumptions on the initial conditions since it is possible to find an objective function that eliminates the unobserved effects.

The fixed effects approach has been scarcely used in duration analysis. Frederiksen *et al.* (2007) proposed a method to estimate discrete time duration models allowing for group level heterogeneity in models for single and multiple spells.[5] We follow this approach and, as in previous sections, we assume a conditional logistic distribution.

To see how the approach works, it is useful to formulate the model as a discrete choice model. We use $y_{ijt}^k = 1$ to denote that the individual i during the spell j leaves the state k in period t. The model is

$$y_{ijt}^k = 1(\alpha_0^k + \alpha_1^k(t) b_{ij}(t) u_{it} + \alpha_2^k(t) x_{ij}(t) + \gamma_j^k(t) + \eta_i^k + \varepsilon_{ijt}^k \geq 0), \qquad k = u, e \tag{9}$$

In the spirit of panel data models, the proposed estimation procedure is based on the observations for which the number of spells per individual, J_i, is larger than 1.[6] It is possible to construct conditional statements and to get

[5] Ridder and Tunali (1999) also follow a fixed effects approach but it only works when durations are continuous.

[6] One could think that this could give rise to an endogenous self-selection problem. In order

rid of the unobserved heterogeneity by using only the spells of unemployment or only the spells of employment. Given that our main interest is the process for unemployment, we drop out the unobserved heterogeneity by using only the spells of unemployment available for each individual.

For simplicity let's assume that the number of spells for all individuals is $J = 2$ and that $\alpha_1^u(t) = \alpha_1^u$ and $\alpha_2^u(t) = \alpha_2^u$. To eliminate the unobserved heterogeneity we compare first to second spells for each individual and each period, t. That is, we compare y_{i2t}^u to y_{i1t}^u assuming that the individual specific effect, η_i, does not depend on the spell number. Therefore, only variables which depend on the spell number are identified, and those variables which only vary with the duration but are constant across spells for the same individual are dropped out. Specifically, within this framework the additive duration dependence, $\gamma_j^k(t)$, is not identified, although interactions between the explanatory variables and the duration dependence can be identified.

Frederiksen *et al.* (2007) assume that the $\varepsilon_{ijt}'s$ are logistically distributed and their framework allows for feedback from the $\varepsilon's$ to future values of the explanatory variables. That is, the explanatory variables can be predetermined. In our application the only explanatory variable which could be considered as predetermined as opposed to strictly exogenous is the indicator of benefits, $b(\cdot)$. Nonetheless, the benefit entitlement is observed in our data, so we can condition on past, current and future values of this variable. In this case, it can be treated as exogenous and therefore we do not need to specify the feedback from ε to future values of $b(\cdot)$ in order to get consistent estimates of the parameters of interest.

Under the previous assumption, Frederiksen *et al.* (2007) show that it is possible to construct conditional statements (see Lemma 1, page 1018) and that

to check for that, in our application we have estimated the model which does not account for unobserved effects only with the individuals with two or more spells. Our results basically hold.

one can estimate the parameters of interest by maximizing

$$\sum_{i=1}^{N}\sum_{t_1=1}^{\bar{t}}\sum_{t_2=1}^{\bar{t}}\{1(T_{1i}=t_1,T_{2i}>t_2)+1(T_{1i}>t_1,T_{2i}=t_2)\} \tag{10}$$
$$\times\log\left(\frac{\exp((b_{i1}(t_1)-b_{i2}(t_2))\alpha_1^u+(x_{i1}(t_1)-x_{i2}(t_2))\alpha_2^u)^{1(T_{1i}=t_1,T_{2i}>t_2)}}{1+\exp((b_{i1}(t_1)-b_{i2}(t_2))\alpha_1^u+(x_{i1}(t_1)-x_{i2}(t_2))\alpha_2^u)}\right).$$

A similar approach can be used when there are more than two spells for each individual and when the α parameters do vary with the duration (see Frederiksen *et al.*, 2007, for details).

3 Experimental evidence

3.1 Experimental design

In this section we study the finite sample properties of the random effects (RE) and fixed effects (FE) estimators described above in an unemployment duration model with unobserved heterogeneity by means of Monte Carlo simulations.

We simulate individual hazard rates using a data generating process based on a standard model of labor flows. The individuals start their labor history at age 16 as unemployed. Then, monthly unemployment and employment spells are generated by assuming that the worker is fired in each period at a rate which depends on age and qualification. Regarding the unemployment hazard, we assume that the worker leaves unemployment in each period at a rate which is function of duration, age, qualification, unemployment benefits receipt and number of months until exhausting unemployment benefits. Our data generating process contains negative duration dependence in the unemployment hazard rate. Specifically, we assume that those unemployed three months or more leave unemployment at a lower rate than those unemployed just one or two months. The unemplopyment hazard rate takes the following form:

$$h_i(t) = F(a_0+a_1 1(T<3)+a_2 UB\times 1(T<3)+a_3 UB\times 1(T\geq 3)+ a_4 1(UBdur \geq 2)+a_5 Young+a_6 Qualification), \tag{11}$$

where $1(T < 3)$ takes the value 1 if the unemployment duration is smaller than 3 months, UB takes the value 1 of the individual receives unemployment benefits, $1(T \geq 3)$ takes the value 1 if the unemployment duration is equal or larger than 3 months, $1(UBdur \geq 2)$ takes the value 1 if the number of months to exhausting unemployment benefits is larger than 2, $Young$ takes the value 1 if the individual is younger than 19 or 21, depending on the number of spells available for each individual.

We assume that the researcher does not observe the worker's level of qualification. Hence, this variable constitutes the unobserved heterogeneity term in the estimates. Various experimental designs are carried out which differ in terms of the process generating the unobserved heterogeneity (qualification) and the number of spells available for each individual. The first design is created by assuming that the variable measuring the worker's qualification level is distributed according to a normal distribution. Then, we modify our data generating process of the unobserved heterogeneity to a discrete distribution with four mass points.

We set the values for the hazards as indicated in Table 1. The baseline hazard determines the probability of leaving unemployment for young and highly qualified individuals with unemployment duration smaller than three months and without unemployment benefits. This is set equal to 60%. Table 1 shows the variation in the hazard when previous characteristics change.

From the previous hazard rates the implied values for the coefficients of the hazard function are derived using the logistic transformation. Table 2 shows the true parameter values obtained for each DGP. As indicated above, in our first DGP (first column in Table 2) we assume that the qualification follows a normal distribution, $N(0.553, 0.137)$. The second DGP (second column in Table 2) assumes that the qualification follows a discrete a distribution with four mass points at $(0, -0.381, -0.784, -1.247)$[7] and associated probabilities $(0.4, 0.2, 0.2, 0.2)$. The hazard rates associated to each of these individual types are $(60\%, 50\%, 40\%, 30\%)$. In the third DGP we assume that the unobserved heterogeneity follows a discrete distribution with four mass points at $(0, -1.013, -3.274, -1.506)$ with probabilities $(0.4, 0.2, 0.2, 0.2)$. In this case, the associated hazard rates for each individual type are $(60\%, 40\%, 10\%, 30\%)$. Notice that in this case the hazard rates are less uniformly distributed than in the previous one.

The model is then estimated by maximum likelihood using the FE approach and the RE approach. In the first case, the unobserved heterogeneity term is dropped out by transforming the model as indicated in previous section. In the second case, since the distribution of the unobserved heterogeneity is assumed unknown to the researcher, it is approximated by a discrete distribution with two points of support.[8] Notice that in the Monte Carlo design the heterogeneity is not generated according to any of the models estimated, so we would only estimate pseudo-true parameters.

In all cases we generate data with different number of spells available for each individual. Increasing the number of spells per individual could be especially relevant for the FE estimates, since one could expect that it will enhance the precision in most parameters estimates. We present estimates with 6 and 18 spells available for each individual.

3.2 Monte Carlo results

For each experiment we generate 100 samples with $N = 3000$.[9] Results from this experiment are presented in Tables 3 and 4, which report mean point estimates, mean estimated standard errors, percentage biases, and mean squared errors

[7] We assume that the unobserved heterogeneity has zero mean. Therefore, our first mass point is equal to zero and the others are differences with respect to the constant.

[8] See Gaure et al. (2007) for a detailed discussion on the optimal number of support points.

[9] Results with $N = 1000$ and also with 12 spells for each individual are available upon request.

(MSE) for the models with different heterogeneity distributions and number of spells available per individual. We first estimate a model which does not account for unobserved heterogeneity. We then present RE estimates using only unemployment spells (labelled RE_U) and using unemployment and employment spells (labelled as RE_UE). Finally, we report the estimates using a FE approach. We show the results regarding the four parameters of main interest.

A first point to note is that, as expected, the bias induced by failing to control for unobserved heterogeneity is large in all models considered. The bias in this coefficient is eliminated by means of accounting for unobserved heterogeneity within a RE approach, but only the model in which the true heterogeneity distribution is discrete with an associated hazard uniformly distributed eliminates the bias almost completely. When the heterogeneity is based on a normal distribution or even on a discrete distribution with hazards non-uniformly distributed, we do not obtain unbiased estimates close to the true parameter values.

Regarding the rest of coefficients, the main result is that FE estimates recovers the parameters irrespective of the way unobserved heterogeneity is distributed in the data. The RE approach tends to do a better job when the true unobserved heterogeneity is discrete and with hazards associated to each individual type uniformly distributed. FE estimates almost always have a smaller mean percentage bias in all experiments, specially as the number of spells available per individual increases.

The comparison between RE_U and RE_UE shows that RE_UE almost always has a smaller MSE than RE_U, being the differences between the two estimators larger when the true unobserved heterogeneity distribution is normal. The RE results may be sensitive to the number of mass points allowed in the distribution for the unobserved heterogeneity. One could argue that increasing the number of support points in the RE approach could give better results specially in the less favorable cases. However, this requires substantial computational resources as the number of support points increases. In many applications it is often specified as just 2 or 3. Therefore, given our simulation results, it may be worthwhile to consider fixed effects estimates which seem to recover reliably the true parameters.

When the number of spells available per individual increases, it turns out that both RE and FE estimates always have a smaller mean bias and also smaller standard deviation, with the reduction becoming wider for the FE estimator. This result shows the importance of having a large panel of spell durations in order to identify the parameters of interest.

In conclusion, the Monte Carlo results for the RE and FE estimates of our models suggest that both RE performs well when the true heterogeneity distribution is similar to the one considered in the estimation approach, but the FE tends to do a better job in all cases. Moreover, we find that in all cases both estimators perform considerably better when the number of spells per individual increases, but FE biases tend to be reduced to a larger extent.

4 Empirical application

4.1 The data

We illustrate previous methods by estimating an unemployment duration model with multiple spells. We use Spanish administrative data from the *Muestra Continua de Vidas Laborales* (MCVL). Administrative data are accessible in many countries, and are likely to play an important role in microeconometric research (see Roed and Raaum, 2003). Our data set is based on a random draw from the Social Security archives and provides a sample of 4% among all the affiliated workers (employed or unemployed) and pensioners. There are four waves available (2005-2008), so that we have information for about 1, 1 million people on their personal characteristics and employment and unemployment spells throughout their entire labor history.[10]

Our sample includes information about 42,396 individuals aged 19 to 62 who were unemployed at some point during the period 2000-2008. We select a sample of male native workers, excluding self-employed and workers in agriculture. Table 5 presents the structure of our data according to the number of spells of unemployment per individual and the duration of each spell. The explanatory variables used are described in the Appendix and summary statistics are presented in Table 6.

To get an idea of the shape of the distribution of durations, we may study the evolution over time of the sample probability of leaving unemployment. That is, we compute the Kaplan-Meier hazard rate which is based on the number of exits from unemployment in each month divided by the population still in unemployment at the beginning of that month. Figure 1 shows a negative duration dependence in the hazard rate. It decreases rapidly up to the twelfth month of unemployment and afterwards it is more or less constant.

Figure 2 represents the effect on the empirical hazard of benefit receipt in a given month. We can see that those individuals not receiving benefits have a higher probability of leaving unemployment than those receiving benefits at all durations, although the difference is larger at the beginning of the spell.

Nonetheless, the observed pattern in the aggregate hazard reflects the fact that different individuals have different exit rates due to differences in observable and unobservable characteristics. The estimation of econometric models allows to disentangle these effects on the hazards and to capture the genuine duration dependence.

[10] To minimize the possible bias due to the selection criteria used to get the sample, which is being linked to Social Security at least one day during the corresponding year, we select all workers appearing in the dataset at least during one sampling year among the four available, 2005-2008.

4.2 Estimation results

Table 7 presents the ML estimates from the different models described in the previous section. First column presents the results from a model which does not account for the effect of unobserved heterogeneity. Second and third columns report the results from the two random effects models estimated. Column 2 reports the results from the RE model estimated using only unemployment spells where we assume a four mass points distribution for the unobserved heterogeneity. Column 3 reports the results for the RE estimated using both unemployment and employment spells where we assume a four mass points distribution for the unobserved heterogeneity as explained in Section 2.2.1.[11] Finally, last column present the results from the FE model.

Duration dependence is captured by a third order polynomial of log duration. We have introduced as regressors interactions of the dummy for the receipt of unemployment benefits, age, qualification, employment growth rate,[12] and also between time to exhausting unemployment benefits and logged duration. We have included also a dummy variable for the year 2008 given the change in economic growth observed since the beginning of that year and to capture a possible different behavior of unemployed workers during the crisis.

The results indicate a non-monotonic duration dependence. As expected, the coefficients for the *log Dur* variable are in general smaller once unobserved heterogeneity is accounted for. The pattern of the predicted hazards are shown in Figure 3 for an individual with the average characteristics of our sample. We can see that the hazard of leaving unemployment decreases with elapsed duration in all models considered, as is usually obtained in previous literature (see for instance Bover *et al.*, 2002). Up to the third month of unemployment, all models predict basically the same hazards. But afterwards, the model which does

[11] In this case, we only report the estimates corresponding to the hazard of leaving unemployment. The estimates on the employment process are available upon request.
[12] This variable captures regional differences across time in the labour market.

not control for unobserved heterogeneity predicts a lower probability of leaving unemployment at all durations. For instance, an individual who remained unemployed for at least 12 months has a probability of leaving unemployment of 10% according to the model which does not account for unobserved heterogeneity and around 15% according to the RE models. Moreover, the predicted hazards by the two RE models are very similar, being the largest differences of around 3 percentage points in the fourth month of unemployment.[13]

On the other hand, Figures 4, 5, and 6, show that the receipt of unemployment benefits reduces the hazard of leaving unemployment, and that the reduction is smaller as duration increases (as indicated by the positive coefficient on the interaction between the dummy for benefit receipt and *log Dur)*. When unobserved heterogeneity is not accounted for, the difference of the effect of receiving unemployment benefits on the hazard of leaving unemployment is smaller than when controlling for unobserved effects (see Figure 7). This result shows that the effect of unemployment benefits is underestimated when unobserved heterogeneity is not accounted for, and this could lead to misleading policy implications. On the other hand, we find that the estimated decrease in the hazard during the beginning of the spell is larger in the RE model which uses employment and unemployment spells than in the one using only unemployment spells.

To asses the effect of allowing for an unrestricted distribution of the individual effects, as in the fixed effects model, Figure 8 displays the odd ratio on the hazard of leaving unemployment for individuals with a benefit entitlement equal to 24 months. The figure shows that the fixed effect estimates provide a negative effect of unemployment benefit much larger than the model without control for unobserved heterogeneity. Moreover, the estimates from the random effects models are in between, although closer to the fixed effects ones.

[13]The FE model is not represented in this figure given the duration depencence parameters are not identified with this approach.

Regarding the effect of the time to exhausting unemployment benefits, estimates from Table 7 shows that there is a negative effect on the hazard rate, but decreasing with duration. Figure 9 depicts the estimated effect in the four models considered for an individuals with a benefit entitlement of 24 months during her first 12 months of unemployment. We can observe that relative to the rest of estimates, the FE estimates predicts a smaller effect on the hazard.[14]

5 Conclusions

This paper considers the estimation of discrete time duration models using multiple spell data. Our basic motivation is to facilitate the distinction between unobserved heterogeneity and true duration dependence in the exit rate from unemployment. We point out that the availability of multiple spell data considerably improves the identification of the parameters of interest.

We present estimates from random effects models assuming that the distribution of the effects is discrete with finite support, using information only on unemployment spells and also on both employment and unemployment spells. On the other hand, since the availability of multiple spells allows us to transform the model to rule out the individual unobserved effects, we also estimate fixed effects models in the spirit of fixed effects discrete choice panel data models.

We report Monte Carlo simulations to asses the finite sample properties of these estimators with several Monte Carlo designs which differ in terms of the true unobserved heterogeneity distribution and of the number of spells available for each individual. Our results show that the fixed effects model gives very reliable estimates of the parameters of interest. We find that random effects estimators perform well when the unknown heterogeneity distribution is assumed to be similar to the true one, but fixed effects tends to do a better job in all cases. Moreover, both estimators perform considerably better when the number

[14] The rest of explanatory variables included in the model have all the expected effect and are not explained in the text.

of spells per individual increases, but fixed effects biases tend to be reduced to a larger extent. These findings suggest that the fixed effects approach may be an useful alternative to estimate discrete time duration models, specially given the computational burden of the random effects approach assuming a discrete distribution for the unobserved heterogeneity as the number of support points increases.

Finally, as an empirical illustration, we estimate previous models using a large administrative data set for Spain which contains information on multiple spell data. The results show that lack of control of unobserved heterogeneity leads to underestimating the negative effect of unemployment benefits and that the random effects models correct the bias, although not completely. Specifically, the fixed effect estimates provide a negative effect of unemployment benefit much larger than the models without unobserved heterogeneity. Moreover, the estimates from the random effects models are in between, although closer to the fixed effects ones. We could interpret this result as an indication that random effects models do not eliminate all the unobserved heterogeneity present in the data, as opposed to the fixed effects model. In the same line, lack of control for unobserved heterogeneity leads to an underestimation of the negative effect of the time to exhausting benefits on the hazard rate.

The contrast between these sets of estimates emphasizes the point that different individuals behave differently due to heterogeneous characteristics. Lack of proper control for these effects could lead to the conclusion that unemployment benefits have a smaller effect on the probability of leaving unemployment than the true one.

Appendix

Variables Definition

Variable Name	Definition
Unempl. Benefits	The worker receives unemployment benefits in the current period
Time to exhausting	Number of months until the exhaustion of Unemployment Benefits
Unempl. Subsidy	Unemployment assistance benefits
Industry	Sector of activity in the previous job
Construction	Sector of activity in the previous job
Non-market services	Sector of activity in the previous job
Δ Empl. rate	Annual growth rate of employed population by region and year
High Occupation	Occupation held in the previous job
Intermediate Occupation	Occupation held in the previous job
Age 31-44	The age in the current period belongs to the interval 31-44
Age 45-62	The age in the current period belongs to the interval 45-62
Fired	Non voluntary exit from the previous job
Firm$\geq$250 workers	The previous firm of the worker had more than 250 workers
New Firm	Worker's previous firm was created one year before the worker was hired or less
THA	Coming from a Temporary Help Agency
Permanent contract	The previous job of the worker was under a permanent contract
Part-time job	The previous job of the worker was under a part-time contract
Total empl.	N^o months of employment before the first observation in our sample
Same Employer	Same employer in the following job as in the pervious one
Private firm	The previous firm did not belong to the Public sector

References

[1] Abbring, J.H. and G.J. van den Berg (2003), "The nonparametric identification of treatment effects in duration models", *Econometrica,* 71, 1491-1517.

[2] Abbring, J.H. and G.J. van den Berg (2007), "The unobserved heterogeneity distribution in duration analysis", *Biometrika,* 94, 87-99.

[3] Baker, M. and A. Melino (200), "Duration dependence and nonparametric heterogeneity: a Monte Carlo study", *Journal of Econometrics*, 96, 357-393.

[4] Bover, O., Arellano, M. and S. Bentolila (2002), "Unemployment Duration, Benefit Duration, and the Business Cycle", *The Economic Journal*, 112, 223-265.

[5] Frederiksen, A., Honoré, B.E., and L. Hu (2007), "Discrete Time Duration Models with Group-level Heterogeneity", *Journal of Econometrics*, 141, 1014-1043.

[6] Gaure, S., Roed, and T. Zhang (2007), "Time and causality: A monte Carlo assessment of the timing-of-events approach", *Journal of Econometrics*, 141, 1159-1195.

[7] Heckman, J. (1991), "Identifying the Hand of Past. Distinguishing State Dependence from Heterogeneity", *American Economic Review Papers and Proceedings*, 81, 75-79.

[8] Heckman, J. and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, 52, 271-320.

[9] Huh, K. and R.C. Sickles (1994), "Estimation of the duration model by nonparametric maximum likelihood, maximum penalized likelihood, and probability simulators", *The Review of Economics and Statistics*, 76, 683-694.

[10] Jenkins, S. (1995), "Easy Estimation Methods for Discrete Time Duration Models", *Oxford Bulletin of Economics and Statistics*, 57(1), 120-138.

[11] Kalbfleisch, J. D. and R. L. Prentice (1980), The Statistical Analysis of Failure Time Data, Wiley, New York.

[12] Kalwij, A.S. (2004), "Individuals' Unemployment Experiences: Heterogeneity and Business Cycle Effects", *Oxford Bulletin of Economics and Statistics*, 66 (2), 205-237.

[13] Narendranathan, W., and M.B. Stewart (1993), "Modelling the probability of leaving unemployment: competing risks models with flexible base-line hazards", *Journal of the Royal Statistical Society, Series C,* 42, 68-83.

[14] Ridder, G., and I. Tunali (1999), "Stratified partial likelihood estimation", *Journal of Econometrics,* 92, 193–232.

[15] Roed, K., and O. Raaum (2003), "Administrative registers - unexplored reservoirs of scientific knowledge?", *Economic Journal,* 113, F258-F281.

[16] Van den Berg, G.J. (2001), "Duration models: specification, identification, and multiple durations", in: J.J. Heckman and E. Leamer, eds. Handbook of Econometrics, Volume V (North-Holland, Amsterdam).

[17] Wooldridge J.M. (2005). "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity", *Journal of Applied Econometrics*, 20(1), 39–54.

Table 1: Hazard values for the DGP

Individual type	Hazard rate
Young, highly qualified, $1(T < 3) = 1, UB = 0$	60%
BUT	
$1(T \geq 3) = 1$	55%
$UB = 1$ and	
$1(UBdur \geq 2) = 1$	40%
$1(T < 3) = 1$	50%
$1(T \geq 3) = 1$	55%

Table 2: True parameter values for the DGP

True value	DGP (i)	DGP (ii)	DGP (iii)
a_1	$+0.157$	$+0.229$	$+0.272$
a_2	-0.554	-0.491	-0.600
a_3	-0.384	-0.276	-0.285
a_4	-1.498	-1.152	-1.230
a_0	-4.010	-0.733	-0.881
Unobserved Heterogeneity:			
	$N(0.55, 0.14)$		
η_1		-1.247	-1.013
η_2		-0.784	-3.274
η_3		-0.381	-1.506

Table 3: Monte Carlo Results. 6 spells

Variable	Without control unob. heterog.				RE_U				RE_UE				FE			
	Mean	St.dev.	Mean Bias	MSE	Mean	St.dev.	Mean Bias	MSE	Mean	St.dev.	Mean Bias	MSE	Mean	St.dev.	Mean Bias	MSE
DGP (i)																
$dur(U) < 3$	0,361	0,034	129,40%	4,255	0,182	0,035	16,03%	0,188	0,207	0,035	31,49%	0,366				
$UB \times dur(U) < 3$	-0,365	0,044	-34,10%	3,764	-0,447	0,046	-19,38%	1,366	-0,501	0,047	-9,55%	0,503	-0,623	0,059	12,41%	0,821
$UB \times dur(U) \geq 3$	-0,277	0,062	-27,82%	1,524	-0,460	0,066	19,70%	1,001	-0,434	0,066	12,93%	0,682	-0,417	0,094	8,58%	0,992
$dur(UB) \geq 2$	-1,545	0,095	3,12%	1,114	-1,595	0,096	6,47%	1,851	-1,598	0,099	6,70%	1,992	-1,669	0,104	11,45%	4,018
DGP (ii)																
$dur(U) < 3$	0,402	0,024	75,75%	3,011	0,216	0,025	-5,45%	0,091	0,239	0,024	4,41%	0,073				
$UB \times dur(U) < 3$	-0,337	0,022	-31,39%	2,529	-0,437	0,023	-11,12%	0,395	-0,440	0,023	-10,49%	0,358	-0,470	0,033	-4,32%	0,137
$UB \times dur(U) \geq 3$	0,147	0,058	-153,26%	17,857	-0,097	0,062	-64,86%	3,404	-0,227	0,060	-17,58%	0,491	-0,228	0,086	-17,34%	0,877
$dur(UB) \geq 2$	-0,575	0,034	-50,09%	33,652	-0,776	0,037	-32,58%	14,469	-0,895	0,036	-22,26%	6,881	-0,926	0,043	-19,60%	5,369
DGP (iii)																
$dur(U) < 3$	0,900	0,031	231,37%	39,630	0,366	0,037	34,73%	1,026	0,379	0,037	39,66%	1,298				
$UB \times dur(U) < 3$	-0,120	0,036	-80,01%	23,208	-0,474	0,040	-21,01%	1,754	-0,520	0,041	-13,37%	0,811	-0,627	0,043	4,46%	0,257
$UB \times dur(U) \geq 3$	0,188	0,050	-166,18%	22,613	-0,319	0,055	11,94%	0,412	-0,260	0,057	-8,73%	0,382	-0,321	0,100	12,78%	1,125
$dur(UB) \geq 2$	-1,084	0,041	-11,80%	22,740	-1,388	0,041	12,86%	2,669	-1,340	0,044	9,00%	1,417	-1,272	0,045	3,49%	0,386

Results from $N = 3000$ and 100 replications based on the model with different heterogeneity distributions and 6 spells for each individual. MSE units: 1E-4

Table 4: Monte Carlo Results. 18 spells

Variable	Without control unob. heterog.				RE_U				RE_UE				FE			
	Mean	St.dev.	Mean Bias	MSE	Mean	St.dev.	Mean Bias	MSE	Mean	St.dev.	Mean Bias	MSE	Mean	St.dev.	Mean Bias	MSE
DGP(i)																
$dur(U) < 3$	0,378	0,029	140,75%	4,984	0,192	0,031	22,44%	0,222	0,204	0,0318	29,90%	0,321				
$UB \times dur(U) < 3$	-0,391	0,037	-29,49%	2,809	-0,480	0,039	-13,31%	0,692	-0,5077	0,040	-8,40%	0,376	-0,592	0,050	6,75%	0,385
$UB \times dur(U) \geq 3$	-0,281	0,025	-26,88%	1,133	-0,424	0,029	10,30%	0,242	-0,397	0,030	3,36%	0,109	-0,364	0,040	-5,30%	0,198
$dur(UB) \geq 2$	-1,520	0,082	1,49%	0,716	-1,537	0,086	2,59%	0,889	-1,521	0,090	1,52%	0,847	-1,516	0,089	1,23%	0,811
DGP(ii)																
$dur(U) < 3$	0,348	0,015	52,20%	1,453	0,229	0,015	0,16%	0,022	0,249	0,015	8,72%	0,062				
$UB \times dur(U) < 3$	-0,383	0,016	-22,09%	1,205	-0,476	0,016	-3,13%	0,051	-0,477	0,016	-2,82%	0,046	-0,514	0,021	4,69%	0,097
$UB \times dur(U) \geq 3$	-0,020	0,027	-92,60%	6,604	-0,221	0,027	-20,00%	0,379	-0,244	0,028	-11,57%	0,177	-0,255	0,031	-7,75%	0,143
$dur(UB) \geq 2$	-0,882	0,021	-23,41%	7,315	-1,075	0,021	-6,67%	0,634	-1,116	0,022	-3,09%	0,173	-1,183	0,026	2,74%	0,167
DGP(iii)																
$dur(U) < 3$	0,846	0,021	211,46%	33,070	0,417	0,022	53,41%	2,155	0,385	0,023	41,84%	1,346				
$UB \times dur(U) < 3$	-0,183	0,022	-69,54%	17,479	-0,521	0,022	-13,28%	0,683	-0,532	0,022	-11,33%	0,510	-0,642	0,028	7,00%	0,257
$UB \times dur(U) \geq 3$	0,205	0,036	-171,89%	24,063	-0,209	0,041	-26,49%	0,735	-0,129	0,045	-54,49%	2,604	-0,290	0,043	1,93%	0,183
$dur(UB) \geq 2$	-1,003	0,025	-18,43%	5,201	-1,287	0,035	4,68%	0,456	-1,155	0,055	-6,10%	0,860	-1,233	0,030	0,24%	0,089

Results from $N = 3000$ and 100 replications based on the model with different heterogeneity distributions and 18 spells for each individual.

Table 5: Unemployment Spells and Unemployment Duration

N^o of Unempl. spells per individual	*1*	*2-4*	*5-10*	*+10*
(%)	31.47	38.78	23.62	6.13
Unempl. Dur. in months. All spells	*1-3*	*3-6*	*6-12*	*+12*
(%)	64.68	14.40	12.20	8.71
Unempl. Dur. in months. Completed spells	*1-3*	*3-6*	*6-12*	*+12*
(%)	69.98	13.71	11.48	4.84

Table 6: Descriptive Statistics

	Completed spells (%)	Censored spells (%)
With Unemployment Benefits	33.34	43.41
With Contributive Unempl. Benefits	84.39	84.89
Sector: Industry	12.72	14.46
Construction	30.18	29.16
Non-market services	14.15	14.11
Market services	42.83	41.87
High Occupation	15.78	19.78
Intermediate Occupation	37.14	38.36
Low Occupation	47.08	41.86
Age 19-30	55.30	45.23
Age 31-44	31.41	32.78
Age 45-62	13.29	21.99
Non voluntary exit from previous job	83.27	81.31
Permanent contract	10.71	23.35
Part-time job	13.19	13.79
No. of Spells	113,997	22,234

Table 7: ML Estimates

	Without control for unob. het.	RE_U	RE_UE	FE
log Dur	-1,9291	-1,7610	-1,8283	-
	(0,0282)	(0,0301)	(0,0296)	
(log Dur)^2	0,9658	0,9971	1,0223	-
	(0,0260)	(0,0274)	(0,0270)	
(log Dur)^3	-0,1835	-0,1850	-0,1912	-
	(0,0063)	(0,0066)	(0,0065)	
U. Benefits	-0,9216	-1,3608	-1,2189	-1,2692
	(0,0174)	(0,0206)	(0,0192)	(0,0271)
U. Benefits$x$ logDur	0,0794	0,1284	0,0936	0,1735
	(0,0134)	(0,0148)	(0,0142)	(0,0249)
Time to exhausting	-0,0101	-0,0019	-0,0043	-0,0290
	(0,0014)	(0,0016)	(0,0015)	(0,0023)
Time to exh.x logDur	0,0178	0,0127	0,0139	0,0204
	(0,0012)	(0,0013)	(0,0013)	(0,0024)
U. Assitance	-0,2350	-0,2015	-0,2243	-0,2790
	(0,0168)	(0,0207)	(0,0200)	(0,0278)
Δ Empl. rate	4,2710	5,1595	4,8970	4,5675
	(0,2374)	(0,2689)	(0,2599)	(0,3357)
Δ Empl. rate$x$ logDur	-2,2752	-2,2783	-2,2293	-1,9831
	(0,1611)	(0,1753)	(0,1713)	(0,2794)
Age 31-44	0,0045	0,0659	0,0449	0,2548
	(0,0135)	(0,0185)	(0,0167)	(0,0367)
Age 45-64	-0,4377	-0,4774	-0,4239	0,2244
	(0,0210)	(0,0301)	(0,0259)	(0,0670)
Age 31-44x logDur	-0,0707	-0,0733	-0,0718	-0,0376
	(0,0091)	(0,0102)	(0,0097)	(0,0339)
Age 45-64x logDur	-0,2025	-0,1987	-0,2056	-0,0199
	(0,0117)	(0,0131)	(0,0125)	(0,0593)
High qualification	0,0379	0,0144	0,0368	0,1060
	(0,0139)	(0,0201)	(0,0188)	(0,0303)
Interm. qualification	0,1378	0,1066	0,1301	0,0371
	(0,0119)	(0,0145)	(0,0136)	(0,0199)
High qualifi.x logDur	-0,0080	0,0156	0,0022	0,0391
	(0,0112)	(0,0126)	(0,0121)	(0,0271)
Interm. qualif.x logDur	-0,0469	-0,0243	-0,0357	0,0371
	(0,0087)	(0,0097)	(0,0094)	(0,0184)

Note: Numbers in brackets are st.errors. Firm's characteristics and seasonal and sectorial dummies included.

Table 7(Cont.): ML Estimates. Males

	Without control for unob. het.	RE_U	RE_UE*	FE
Total empl.	0,0292	0,0300	0,0293	-0,0318
	(0,0008)	(0,0012)	(0,0010)	(0,0045)
Year 2008	-0,2134	-0,2424	-0,2379	-0,2380
	(0,0162)	(0,0186)	(0,0181)	(0,0262)
2008xU.Benefits	0,0752	0,0550	0,0574	0,0427
	(0,0322)	(0,0350)	(0,0339)	(0,0449)
2008xTime to exh.	-0,0217	-0,0226	-0,0209	-0,0197
	(0,0025)	(0,0028)	(0,0027)	(0,0043)
Constant	-1,3193	-	-	-
	(0,0291)			
s_1^u	-	0,3044	-2,0708	-
		(0,1416)	(0,0367)	
s_2^u	-	-0,3214	-0,7267	-
		(0,1395)	(0,0364)	
s_3^u		-2,6056	-	-
		(0,0663)		
s_4^u		-0,7520	-	-
		(0,1146)		
P_1		-1,7336	-	-
		(0,3268)		
P_2		-0,4343	-	-
		(0,0888)		
P_3		-0,3214	-	-
		(0,1395)		
P_{11}		-	0,8878	-
			(0,0357)	
P_{12}		-	-0,1286	-
			(0,0407)	
P_{21}		-	0,4112	-
			(0,0377)	
N^o Obs.	587.998	587.998	2.007.629**	198.852
Log Lik.	-241.136	-235.759	-687.446	-71.134

*Only results for the unemployment hazard are reported.
**Total number of observations, including those tho exit to employment.

Figure 1: Empirical Hazard. Kaplan-Meier estimates

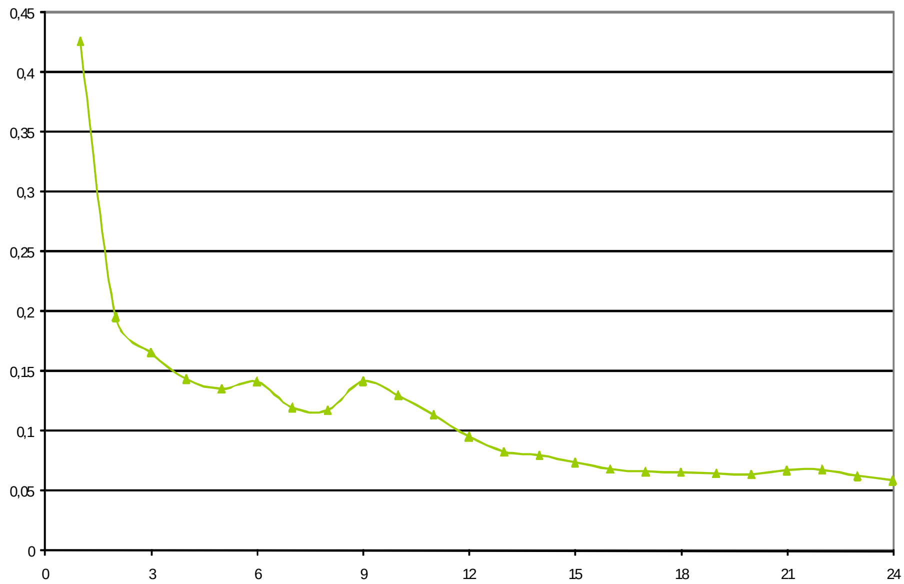
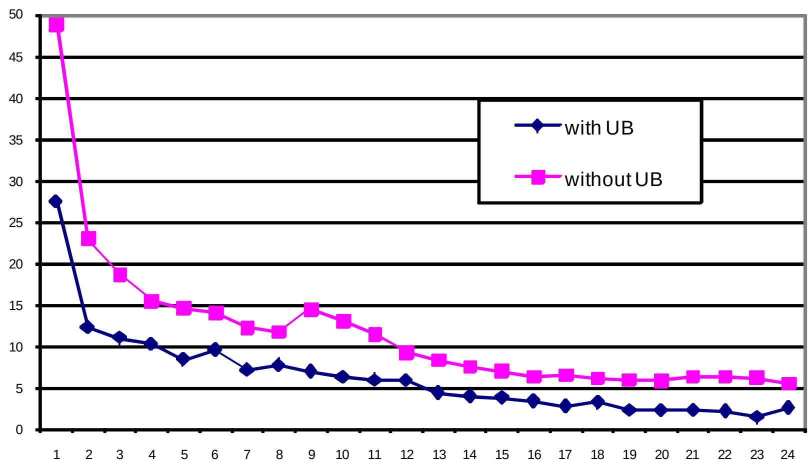


Figure 2: Empirical Hazard and Unemployment Benefit



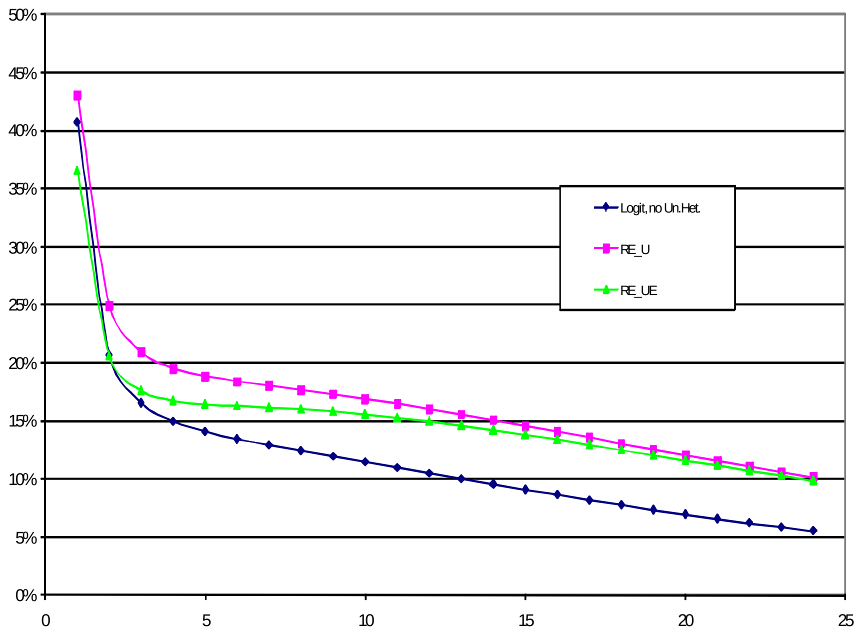


Figure 3: Predicted Hazards

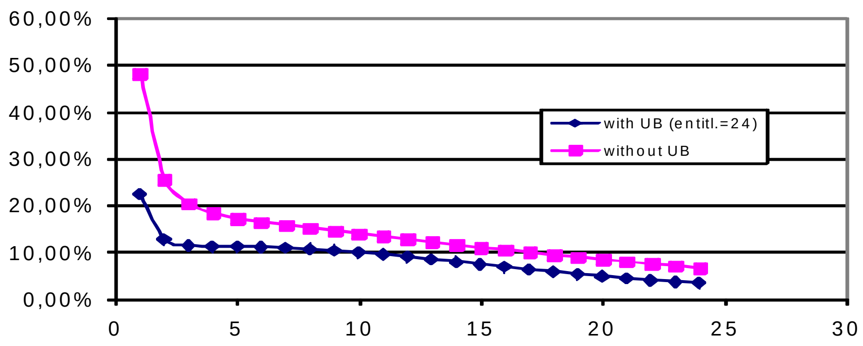


Figure 4: Predicted Hazards by UB receipt. Model without Unob. Het.

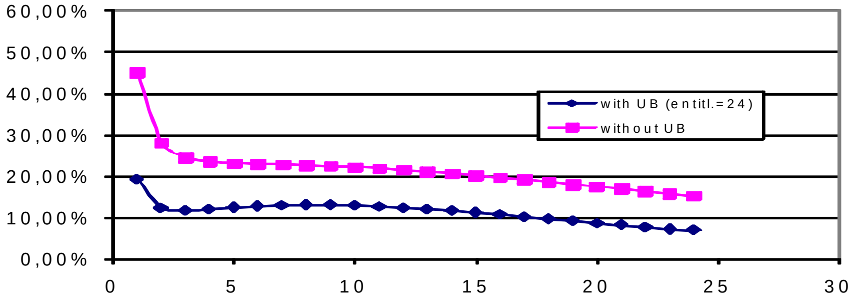


Figure 5: Predicted Hazards by UB. RE_U model

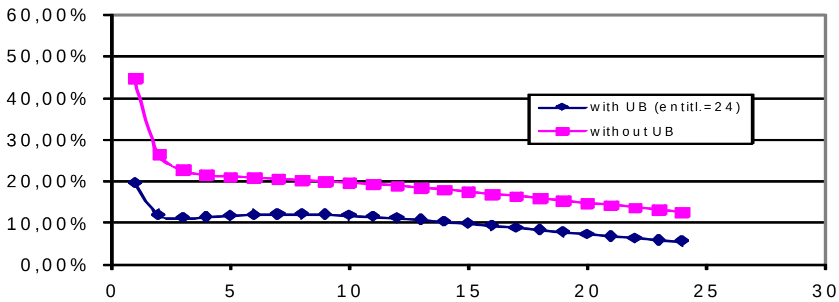


Figure 6: Predicted Hazards by UB receipt. RE_UE model.

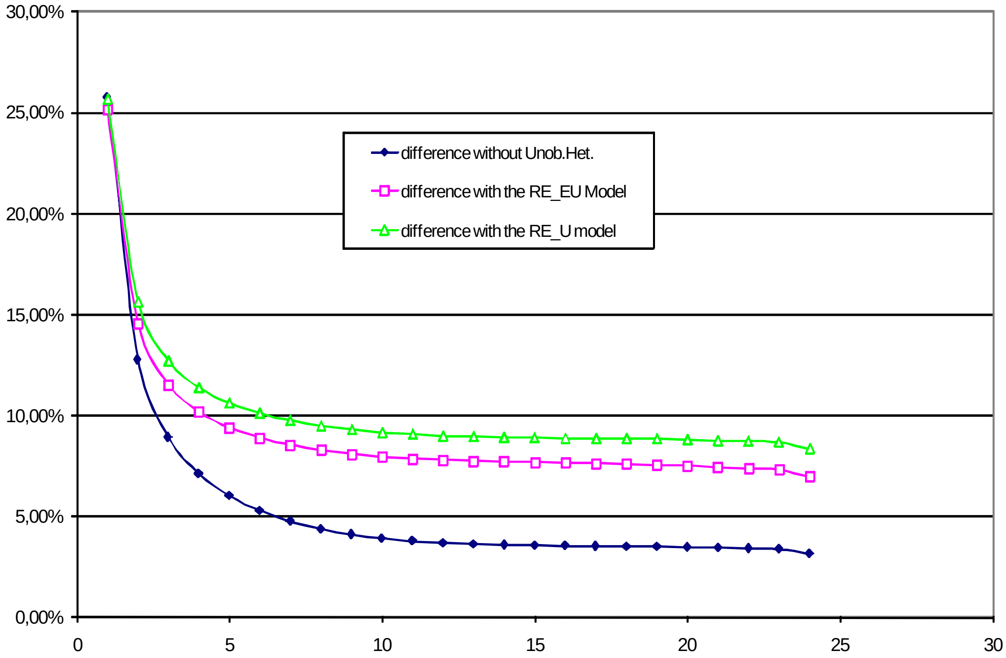


Figure 7: Difference in predicted hazards by UB receipt

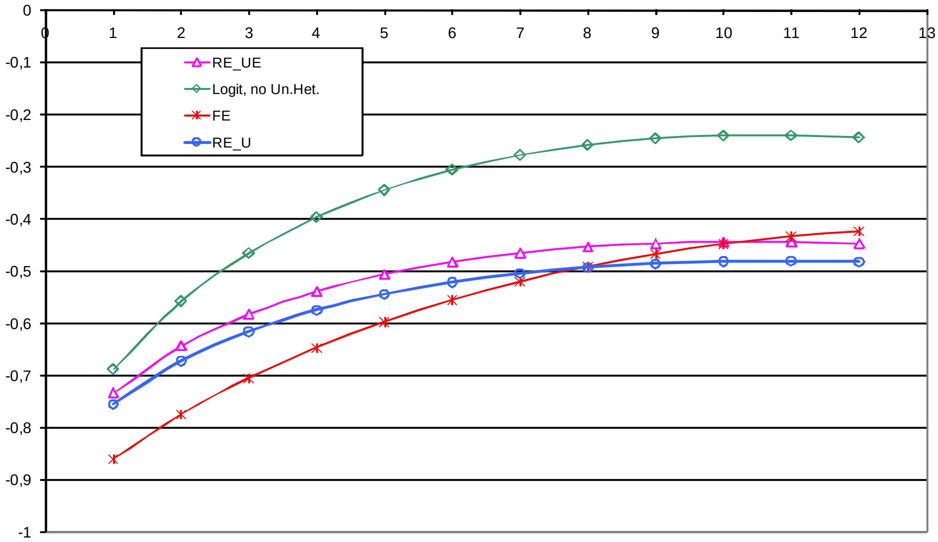


Figure 8: Effect of UB. Odd ratio. Entitlement 24 months

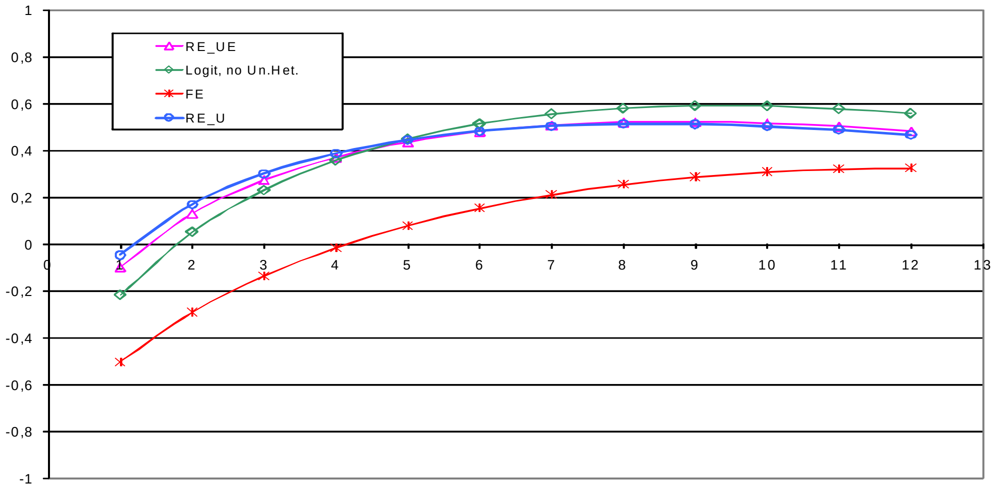


Figure 9: Effect of exhausting UB. Entitlement 24 months