The effect of health on retirement decisions in Hungary

Noémi Kreif

Central European University
Department of Economics
In partial fulfillment of the requirements for
the degree of Master of Arts

supervisor: Prof. Gábor Kézdi
submitted: 2008, Budapest, Hungary

Abstract

Using a Hungarian panel dataset, I attempt to estimate the effect of health on the timing of retirement in a discrete duration econometric framework. Making use of a competing risk multinomial logit model, I find that self rated health has a large siginificant effect on the hazard of exiting from the labour force to disability pension, while it does not have any effect on the hazard of old age pension.

Contents

List of Tables

Chapter 1

Introduction

In developed countries, the relation between health and labor force participation has been widely researched over the past 20 years. This relationship is complex, working through various channels and in both directions. Health, as an essential part of human capital, influences productivity, affects the wage offered to the individual, thus works on the demand side. At the same time, health can influence the marginal utility of consumption, leisure, and the disutility of work - making a labor supply effect. In addition, specific labor market situations, like unemployment, hard physical work can affect health. While all of these channels are of great interest, and ideally should be studied together, in this paper I pick one specific issue: the effect of health on the timing of retirement. The objective of this paper is to briefly review the existing empirical and methodological literature and to carry out my own analysis on Hungarian data.

Costs from early retirement - usually categorized as productivity losses - make a substantial share of the societal burden of illnesses and bad health conditions. In order to be able to assess the cost-effectiveness of any public health interventions, it is essential to be able to give a good estimate on the effect of health on the timing of retirement. While the question is relevant in developed and developing countries as well, the emphasis can vary by institutional setting. Hostenkamp and Stolpe (2008) investigate the social cost of health-related early retirement in Germany. Their findings, estimating an enormous loss of early retirement, are highly relevant for the Hungarian setting, since Germany, similarly to Hungary, has a pay-as-you -go pension system. At the same time in the US the existence of private retirement schemes make researchers focus more on disability benefits. Bound, Stinebrickner, and Waidmann (2007) investigate the interplay between financial incentives (like disability benefits), economic resources and health. By using a dynamic modelling framework, they find that a typical individual in poor health is 10 times more likely to retire than a similar person of average health before becoming eligible for retirement benefits.

My analysis makes use of a unique Hungarian panel dataset, the Hungarian Panel Household Survey (1992-1997). Using discrete duration analysis, I find that self assessed health has a significant effect on transiting out of the labor force to disability pension. However, for transition to early retirement health self rated health has no significant effect.

The remaining part of the paper is structured as follows. Section 2 provides a brief review of the literature on the methodology of investigating the effect of health on retirement decisions. Section 4 provides a detailed overview of the method of estimating causal effects in discrete duration data. Section 5 describes the dataset used along with some descriptive statistics. The results and overall conclusions are presented in Section 6. Section 7 is a discussion.

Chapter 2

Review of methodological issues and previous empirical evidence

When choosing the appropriate method of estimating the causal effect of health on the timing of retirement, the classical problem of microeconometric analysis of economic behavior arises. One choice is to build a structural model based on economic theory, than to attempt to estimate and interpret its parameters. On the other hand, it may be sufficient to sketch and estimate a reduced form model based on intuition coming from separate economic models.

Before carefully assessing both approaches, it is worth doing two things. First, I give a definition of retirement and characterize it as rational choice of the individual. Second, I sketch the possible channels through which health may affect the decision to retire.

2.0.1 Retirement as a decision

Following Spataro (2002), retirement can be characterized as:

1. a discrete choice;

2. a typically absorbing state: there is no way back to the labor market. However, there are some instances when the borderline between retirement an unemployment cannot be clearly drawn;

3. a decision which can be made in a certain age interval, determined by the institutional setting of the given country;
4. it requires forward looking behavior. A rational individual assesses her future prospects when making the choice;
5. it mainly depends on individual characteristics, but can also be influenced by characteristics and behaviour of other members of the household;
6. a sequential choice. The decision to retire in one period requires not having retired in a previous period.
7. decided about in an uncertain environment.

   In his article Spataro (2002) assesses the different modeling approaches of retirement by their abilities to capture to complexity resulting from the above mentioned characteristics. Applying it to the present context, the appropriate choice of approach is the one which can capture health the best. While this seems to be an appealing simplification, the endogenous nature of health and the measurement problems attached to it make careful consideration unavoidable.

2.0.2 Health as a determinant of retirement

A good starting point is to regard health as a part of human capital. Grossman (1972) draws an analogy between investment in "health capital" and other forms of human capital. In his model individuals make forward-looking choices regarding the investment into their health capital. The stock of health is a valued for its own sake, while being sick takes time from market and non-market activities. Therefore, labor supply depends on an endogenous health variable. For the present purposes, one lesson of this model is that health must be treated as an endogenous choice.

Currie and Madrian (1999) stress that health is a factor valued by employers but taken by the employee from job to job. Therefore, individuals have to bear the costs of investment in health capital. Good health leads to better expected wages, but better wages also enable the individual to invest more in health.

In view of this, it is surprising that most of the empirical literature on retirement treats health as exogenous. The assumption behind this choice is that in developed countries most of the variation in health is produced by exogenous health shocks (Deschryvere (2004)). If this explanation seems weak – by claiming that he occurrence of bad health events may be associated with economic status – we can get rid off some of the endogeneity by controlling for economic variables in the retirement model.

According to the same author, various channels associated with bad health can be identified as affecting the decision whether to continue work or not. It may lower expected wages for the future, and may increase the disutility from work. Also, it may decrease the utility of consuming other goods, therefore increases the relative valuation of leisure By entitling individuals to non-wage income as disability benefits, it may make retirement attractive. Bad health at the same time decreases expected longevity, influencing the time horizon for decision making. But also, high costs can be associated with poor health therefore requiring higher income.

As the above statements suggest, the theoretical effect of health on the decision to stay in the labor force or not is a priori ambiguous.

2.0.3 Modeling approaches

Structural models

Structural models include the Option Value Model originally set up by Stock and Wise (1990). This is based on an individual assessing the decision of retiring in the current year, by comparing the value of retiring in the given year to the value of retiring in any of the subsequent years. If the difference between the two (the option value of retirement) is positive, the individual will continue to work; if negative, he/she will retire. To incorporate health in an Option Value Model, one has to make assumptions on the role health plays in the value of retiring or staying in the labor force . A second choice for structural modeling is the dynamic programming approach where the behavior of the agent is assumed to be the output of an optimal decision rule, or with other words, " a controlled discrete stochastic process". Spataro (2002). In their recent paper Bound, Stinebrickner, and Waidmann (2007) make use of this methodology, also accounting for the specific problems arising form the measurement of health.

Reduced form models

Reduced form models can be static and dynamic (Spataro (2002)). Static models regard retirement either as a discrete choice in a given period, and apply one of the usual models for binary or multinomial choice to estimate the parameters of interest, or treat the age of retirement as a continuous variable. These models have several problems, referring back to the list characterizing retirement as a decision. First, they cannot incorporate the dynamic nature of the choice -by not being able to handle variables which change over time. The simplest example time varying variable is legal retirement age.

Duration modeling has several advantages over static models. Instead of dealing with a binary left hand side variable, or explaining age of retirement in a static model, it introduces a new dependent variable, "waiting time". It can incorporate dynamic choice by treating the process until the retirement as a sequence of likelihoods. Also, it can incorporate time varying covariates.

Duration modeling can use continuous specification if we can assume that retirement decisions happen in continuous time. In this case one can choose between parametric and semiparametric methods. Because of institutional features it is however much more likely that retirement decisions occur on a yearly basis. This leads to discrete time duration methods. Again, for the hazard rate a parametric functional form can be specified or it can be left nonparametric. I will introduce the discrete time methods in more details in Section 5.

Other issues

Individual heterogeneity can cause different problems depending on the model specified and the estimation method chosen. In any maximum likelihood estimation neglected heterogeneity may bias coefficients. Omitted variables correlated with health will bias the estimated coefficient of health. If the data used have a panel structure, a natural choice is to do discrete duration modeling using time varying covariates. In this case there is a chance to concentrate on transitions between states, by including fixed individual effects. In our context, instead of the level of health, we may focus on the effects of health shocks on the transition from one state to an other, as suggested by Deschryvere (2004).

Couple approaches are emphasizing the correlation between the behaviour of spouses. The behavior of the partner can have an important influence on the timing of retirement, especially for women. This research builds upon family labour supply model: partners are maximizing a single utility function, or have complementarities in their utility coming form leisure (Deschryvere (2004)).

2.0.4 The measurement of health

The parallel between health capital and human capital has already been drawn. One more similarity appears, when trying to empirically capture health. Similarly to the concept of "ability" if we fail to correctly measure health, the problem of biased estimates is unavoidable. With other words, even if the variation in health can be regarded as exogenous in a retirement model, one may not measure this ideal, latent health variable, but something else. This measurement error can have different characteristics for the different health measures, therefore has different consequences on our estimates. Currie and Madrian (1999) divide the usual health measures into the following categories:

1. self-rated health (assessing the own health on a one dimensional scale from very bad to very good)
2. whether having limitations on the ability to work
3. whether having functional limitations with daily living
4. presence of chronic and acute health conditions
5. utilization of medical care

6. clinical assessment of specific conditions

7. nutritional status

8. expected or actual mortality

The above listed measures differ in their ability to capture work capacity and the degree they are prone to reporting bias. While in developed countries the first 5 measures are regarded to be relevant and are usually used, in developing countries the latter four are used mostly.

According to Bound, Stinebrickner, and Waidmann (2007), the global, self rated measures of health contain several problems:

They are discrete, while the latent variable of interest is probably continuous. They presumably contain measurement error since responders are unlikely to use the same "scale" when responding to questions. They are likely to be endogenous to retirement decisions. This latter statement is the so-called "justification hypothesis": someone retiring from the labor market may need to justify his/her choice in the for the society.

Measurement error, if classical, will generally lead to a downward bias of the coefficient of health in a labor force-participation equation. The justification bias, however is likely to exaggerate the role of health in early retirement. Moreover, if one is interested in the effect of economic variables as well (like financial incentives of early retirement), and has the reason to suspect that self-rated health is affected by those incentives, the coefficients of the economic variables are also likely to be biased.

To deal with these issues, several methods have been applied in the literature, some of which is critically reviewed by Bound (1989). One way is to restrict the investigation for the use of the more objective measures. Authors who choose this option usually find smaller effect of health on early retirement than those resulting from the self rated measures. However, Bound (1989) warns that this approach is simplistic: it assumes the bias resulting form the use of these proxies is negligible compared to the bias introduced by the self-reported measures. Some other authors, however, emphasize the validity of self-reported measures from a clinical point of view. Being a valid measure of actual health, the issue of systematic justification bias is still not resolved. If those out of work are substantially more likely to report health problems than those working, the use of self-reported measures would give the wrong answer on why people retire early - even if these measures are highly correlated with actual health. A more sophisticated approach is to instrument self-rated measures with the objective measures. In this case, as Bound (1989) demonstrates, the impact of economic variables is likely to be underestimated. In the above referenced article, the author constructs a statistical model which makes use of the self-assessed measures and the more objective measures as well. In his model, the self reported health is endogenous but is also measured with error. The paper show that without further information, even when using objective and self assessed health measures together, the effect of health on retirement cannot be identified.[1]

Building upon the methodology of Bound (1989) , Rice, Roberts, and Jones (2007) also estimate a latent health stock to purge the health variable from measurement error. They use discrete duration framework, and make use of the availability of panel data to identify the effect of not only health levels, but health shocks by looking at the change in the health variable form one year to an other. This is how they attempt to take account for the endogeneity caused by simultaneity. In my model specification I will strongly rely on their methodology.

2.0.5 Empirical evidence from developed countries

In this section I shortly summarize the results of some of the above referenced articles. For the US, Bound, Stinebrickner, and Waidmann (2007) investigates only single men, to avoid the complication of dealing with the spousal influence on retirement decision. They find that health does play a major role in the timing of retirement. The probability of leaving the labour force before the age of 62 is 5 times higher than the

[1] A recent article of the same authors Bound, Stinebrickner, and Waidmann (2007) follow the same line regarding the health measures, but make us of more ccomputation-intensive techniques. They apply dynamic programming to highlight the interplay between health and financial incentives. and estimate a latent health variable to overcome the bias resulting from the endogeneity of he self reported measure of health.

probability of those in average health. They also compare their results with those estimated the "traditional way" (using only self reported measures), and find that the latter methodology overstates the health effect on retirement. Rice, Roberts, and Jones (2007) use 12 waves of the British Household Panel Survey and also look at the effect of the health status of the spouse on retirement. They find that health shocks are quantitatively more important than pension entitlement both for men and women. They also find that health status of the partner has an influence on the retirement decisions of women, but not on that of men.

Chapter 3

The Hungarian setting

The investigated period is unfortunate from the reasons. First, after the transition retirement was basically a substitute for unemployment. Cseres-Gergely (2007). Also, a major reform in Hungary extending the legal age of retirement only started after 1997. After this period, there was an option called "early retirement" to take the same benefits from old age pension before the actual legal age was reached. While most of the Hungarians go to old age retirement as soon as it becomes legally possible, there is a considerable amount of people only retiring years after the legal age Cseres-Gergely (2007). From this aspect, being in good health could be expected as a predictor for working longer years. Also, since disability pension was used to get rid off unneccessery labour force, we can expect many other effects infuencing disabilty pension than health.

Chapter 4

Methodological framework: duration modelling

In the introduction the research question of this paper was formulated the following way: what is the effect of health on the timing of retirement among Hungarian older workers? Before going into the details of the econometric methodology used, there is a need to formulate this question in a more formalized way. It can be done several ways, as follows. What is the effect of health state on the expected age of retirement? On the probability of being retired before the legal age? On the probability of being in retirement in the given year conditional on not being retired until the present? On the odds of going to retirement compared to staying in the labor force conditional on not being retired. Discrete duration analysis can answer each of these questions with more or less precision. It is able to handle issues which static models (as mentioned above, like estimating ordinary least squares on retirement age, or estimating multiple choice models for a cross section of individuals) could not. These issues are the following: sampling issues leading to left and right censoring, and incorporating information from time varying covariates.

While I do not attempt to introduce here the concepts and methods of duration analysis in general, I have to go briefly through the specific methodology used in for the purposes of the present paper: estimation methods when time spells are discrete, and competing risk duration analysis. I base my discussion - logical steps, concepts and in part the notation - on the work of Jenkins (2004b), who introduces an easy estimation method for discrete time duration data. In some aspects I also reference the broader applied econometric literature on duration analysis.

4.0.6 The base case: single destination states

Basic concepts

The object of analysis is the time until retirement occurs. What can be observed, is individuals with different *spells* of labour force participation. The spell ends when the individual retires. When the spell begins, is less clear - unlike in the case of unemployment spells, for example However, without loss of generality, some age between 45 and 50 can be regarded as the beginning of the spell. Throughout being in the spell, the person is under risk. Ideally we would be able to observe a complete spell of the individual. The ending of the spell is called *failure*. Failures can happen in continuous time or in discrete time. With other words, time until failure can be a discrete or continuous random variable Discrete time in social sciences usually denotes period of observation, like a month, quarter or a year. Discrete time can be only seemingly discrete: while the event can happen in any point of time, we can only observe it at the end of the period whether the event has occurred or not. Time can also be intrinsically discrete. In our case, nothing is lost by assuming intrinsically discrete time. In this case survival time, T is a discrete random variable. Time is now measured in cycles, which is, in our case years.

The probability of surviving exactly up to t cycles - staying in the labour force up to t years - is:

$$f(t) = f_t \equiv \Pr(T = t) \tag{4.1}$$

The discrete time survival function –the probability of not being retired up to t cycles – is:

$$S(t) = \Pr(T \geqq t) = \sum_{k=t}^{\infty} f(k) \tag{4.2}$$

The discrete time hazard in the t-th year - the probability of retirement in the year t conditioning on being in the labour force up to the year before - is

$$h(t) = h_t = \Pr(T = t | T \geqq t) = \frac{f(t)}{S(t-1)} \tag{4.3}$$

An other useful formulation of the discrete density function is:

$$f(t) = h_t S(t-1) \tag{4.4}$$

The link between the discrete time survival function and hazard function can also be formulated as follows:

$$S(t) = \prod_{k=1}^{t} (1 - h_k) \tag{4.5}$$

Sampling

Ideally we would observe all the individuals from their –say – 45th age until their year of retirement. This way we would observe the actual spell length for everybody. This cannot hold for two reasons. Starting with the less problematic, our observation of the surveyed individuals ends in 1997. Some individuals are still not retired at that time, leading to the case of right censoring. Right censoring will be easily dealt with when calculating the likelihood contribution of a censored individual. In the case of this specific dataset, there is attrition in the panel from year to year. Therefore, censoring can not only occur if somebody has not retired until the last survey year, 1997, but also if somebody is lost to follow up in the meantime. Left censoring is a different issue: again, coming from the nature of household panel data, we have a random sample of households from 1992, the first wawe of the survey. This sample includes individuals of all age groups and labor force status. Naturally, to be able to observe spells from the beginning to the end (or until they are right censored) we need individuals still in the labor force in 1992. Therefore, our originally random sample should be restricted, by sampling only those who are still in the labour force in 1992. Because we are only interested in individuals at risk of retirement, those above 45 and working in 1992 are chosen.[1] By this procedure, however, I exclude those who for some reason retired before 1992. This should be also dealt with when calculating the individual likelihood contribution.

Time indexing may differ from calendar time by individuals, the following way. $\tau_i = 0$ when the the person is 45 years old. Δt_i is the part of the spell spent until the calendar time 1992, when the first wave of the survey actually happens. Δt_i +5 equals the part of the spell until 1997, the last year of the survey when the individual is possibly still under risk. Δt_i.Summarizing, an individual contributes t_i periods of spell under risk, where

$$t_{i=} \left\{ \begin{array}{ll} \Delta t_i + z_i \text{ if } z_i \,\epsilon (1,2,3,4,5) \\ \Delta t_i + h_i & \text{if } h_i < z_i \\ \Delta t_i + 5 & \text{if } z_i > 5 \end{array} \right\} \tag{4.6}$$

In the expression, h_i is the number of period the individual spends in the sample, and z_i is the period after entering into the sample that the individual retires. Therefore, an individual contributes totally $\Delta t_i + z_i$ spells if he/she retires until 1997, $\Delta t_i + h_i$ if does not retire before he/she

[1] A further restricton is given by the attrition of the panel dataset. To be able to calculate the hazard in the first period under risk, 1993, those who are already lost to follow up after the year 1992 are excluded. The reason is that there is no information in these spells which could contribute the sample likelihood.

is lost to follow up, and $\Delta t_i + 5$ if he/she stays in the sample until 1997, but also does not retire until 1997. From this point on the third case can be regarded as the special case of the second one, where $h_i = 5$.

The sample likelihood

The sample likelihood of a discrete time survival sample is derived in the Appendix.

$$\log L = \sum_{i=1}^{N} \sum_{k=u_i+1}^{t} y_{ik} \log h_{ik} + (1 - y_{ik}) \log(1 - h_{ik}) \qquad (4.7)$$

The new indicator variable y_{ik} takes the value of one if an individual experiences failure, and takes the value of zero in all other cases (when the individual is censored or remains in the state). This sample log likelihood has exactly the form of the log likelihood of a binary response model, where for each individual (i) there are more observations (k).The corresponding data structure can be regarded as an unbalanced panel. For each individual we have exactly as many observations as long as that individual is under risk. The fact the the individual likelihoods are summed up from $k = u_u + 1$ to $k = t$ deals with the stock sampling (left truncation).

Specifying the hazard function

Now having the sample likelihood the next step is to find the link between or variables of interest - in our case, especially health - and the discrete hazard function.

There are two main approaches used in the literature: the proportional odds model and the proportional hazard model.(Jenkins (2004b)).

The proportional odds model

The proportional odds model assumes that the relative odds of making a transition in year t , given survival up to that period is proportional to the a baseline relative odds. The baseline relative odds is the function of the baseline hazard functions, where the values of all explanatory variables are set to zero. The factor of proportionality is given by some function of the observed variables, for example an exponential of a linear link function. In mathematical form it can be expressed as follows:

$$\frac{h_i(t, x_{it})}{1 - h_i(t, x_{it})} = \frac{h_0(t)}{1 - h_0(t)} \exp(\beta' x_{it}) \tag{4.8}$$

If we denote $\ln \frac{h_0(t)}{1-h_0(t)} = \alpha_t$,and take the logarithm of the above equation, we get

$$\ln \left(\frac{h(t, x_{it})}{1 - h(t, x_{it})} \right) = \alpha_t \ + \beta' x_{it} \tag{4.9}$$

After some manipulation, we get the following:

$$h(t, x_{it}) = \frac{1}{1 + \exp(-a_t - \beta' x_{it})} \tag{4.10}$$

If we interpret the hazard function as a probability conditional on several explanatory variables, we can easily discover the usual logit functional form in the last expression. The linearized version of it, in addition, tells us about the interpretation of the coefficients of the explanatory variables: they influence the log of the relative odds of making a transition to retirement or not. Therefore, if we think that this assumption of the role of the explanatory variables in our hazard function is a valid one, we can simply maximize the above sample likelihood function as a likelihood function of a logit model. The exponential of a coefficient estimated form a proportional odds model can be interpreted the follwoing way: one unit increase in the explanatory variable causes and $\exp(\beta)$ incrase in the relative odds of teh event, compared to staying in the state.

The proportional hazard model

The second choice is perhaps more intuitive, starting from a proportional hazard assumption:

$$h_i(t, x_{it}) = h_0(t) \exp(\beta' x_{it}) \tag{4.11}$$

, where the individual specific hazard should be proportional to the baseline hazard. (Here, to the hazard of somebody for whom all the characterisics take a value of zero). Any parameter derived from such an assumption has a nice interpretation:

$$\frac{h_i(t, x_{it})}{h_0(t)} = \exp(\beta' x_{it}) \tag{4.12}$$

$$\frac{d}{dx_k} \log \left[\frac{h_i(t, x_{it})}{h_0(t)} \right] = \beta_k \tag{4.13}$$

Verbally, the exponentiated coefficient can be interpreted as follows: a one unit increase in teh explanatory variable causes an $\exp(\beta)$ increase in the hazard of the event. Therefore, here I slightly modify the derivation of Jenkins (2004b), while he derives the form of the hazard for interval censored, continuous time, I do it for intrinsically discrete time. The derivation can be found in the Appendix.

$$h(t, x_{it}) = 1 - \exp(\exp(\beta' x_{it} + \gamma_t). \tag{4.14}$$

γ_t is denoting the baseline hazard.

In both the logit and the complementary log-log specification, the baseline hazard can be treated in a flexible way, by including dummy variables for each period under risk. One period should be omitted if we also want to specify an intercept term in the vector β.

Identification of the effect of health

Self rated health is the main variable of interest. The question is the the following: the health of which period does effect the hazard (or the odds) of entering retirement in period t.If we assume that retirement is a decision to be made which takes some time, we can rightfully include the $t-1$ of health in the model of $h(t, X)$. We can also assume that health shocks may matter: a suddenly deteriorating health may effect $h(t, X)$.

Therefore, the vector of explanatory variables looks the following way for the logistic model:

$$h(t, x_{it}) = \frac{1}{1 + \exp(-a_t - \beta' x_{it})}, \tag{4.15}$$

where

$$\beta' X_{it} = \beta_0 + \beta_1 H\_1_{it} + \beta_2 dH_{it} + \delta' Z_{it} \tag{4.16}$$

$$a_t = \rho_1 D93 + \rho_2 D94 + \rho_3 D95 + \rho_4 D96 \tag{4.17}$$

Z is a vector of socioeconomic and demographic variables, $D93 - D96$ are a series of dummy variables denoting the subsequent years from 1993 to 1996. 1997 is chosen as the base category.

The fact that lagged health is controlled for mitigates the fear from one source of endogeneity, justification bias. The ceteris paribus effect of lagged health is coming from the variation of previous period health among those having the same amount of change in sef rated health.

4.0.7 The role of unobserved heterogeneity

So far it was implicitly assumed that unobserved heterogeneity is not present in the discrete hazard function. Not dealing with unobserved heterogeneity would come at a price. According to Van den Berg (2001), omitted unobserved heterogeneity may over-estimate negative duration dependence in the hazard. The reason is that those people with a relatively large (and positive) unobserved component in their hazard will leave the state soon, so those who stay in the sample will have lower values of the unobserved heterogeneity, therefore smaller hazard. The same authors show that in a continuos time proportional hazard model, omitting unobserved heterogeneity may cause the underestimation of the effects of covariates on the hazard. Nicolett and Rondellini (2006) assesses the effects of ignoring unobserved heterogeneity in a single spell discrete duration framework, using Monte Carlo experiments. They find that neglecting unobserved heterogeneity causes a bias in the duration dependence estimates. In the coefficients of the covariates, it leads to a constant rescaling factor, which is smaller than one if variables are iid across individuals but not iid across time.

There are several ways to incorporate unobserved heterogeneity in a discrete duration model. Following Jenkins (2004b), one starting point is the complementary log-log model,

which, as already described has the proportional hazard characteristics:

$$h_i(t, x_{it}) = h_0(t) \exp(\beta' x_{it}) \tag{4.18}$$

The unobserved heterogeneity can enter in a multiplicative form:

$$h_i(t, x_{it}) = h_0(t) \exp(\beta' x_{it}) \varepsilon_i \tag{4.19}$$

where ε is a continous or discrete random variable. In applied work often ε is assumed to follow a continous Gamma distribution, with mean zero, and variance of $\sigma^2 = v$.

As we saw in the previous section, the discrete time hazard function of the proportional hazard model looks as follows:

$$h(t, x_{it}) = 1 - \exp(\exp(\beta' x_{it} + \gamma_t). \tag{4.20}$$

therefore, the unobserved heterogeneity enters the model expression the following way:

$$h(t, x_{it}) = 1 - \exp(\exp(\beta' x_{it} + \gamma_t + \log \varepsilon_i). \tag{4.21}$$

The survival function of this model has a closed form. Jenkins (2004a). The log likelihood function of this model looks as follows:

$$\log L = \sum_{i=1}^{N} \log \{(1 - c_i) A_i + c_i B_i\} \tag{4.22}$$

,where

$$A_i = [1 + v\sum_{j=1}^{t_i} \exp[\beta' x_{it} + \gamma_t]]^{-1/v} \tag{4.23}$$

and

$$B_i = \begin{Bmatrix} [1 + v\sum_{j=1}^{t_i - 1} \exp[\beta' x_{it} + \gamma_t]]^{-1/v} - A_i \text{ if } t_i > 1 \\ 1 - A_i \quad \text{if } t_i = 1 \end{Bmatrix} \tag{4.24}$$

, where c_i is the usual censoring indicator, with $c_i = 1$ if the observation is censored and $c_i = 0$ if the observation is not censored.

In the case of the proportional odds specification, unobserved heterogeneity has to be incorporated in the logit framework. A natural generalization is the random effects logit model. The main assumption here is that unobserved heterogeneity has to be uncorrelated with the time varying covariates, and has to be constant over time for an individual. The conditional hazard for period t, individual i can be than written as follows:

$$h_i(t, x|a_i) = \frac{1}{1 + \exp(-\alpha D_{it} - \beta' X_{it} - a_i)} \tag{4.25}$$

Since a_i is unobserved, the unconditional hazards of retirement also have to be derived. The likelihood contribution of the individual can be obtained by numerically integrating out the unobserved heterogeneity.

4.0.8 Multiple destinations: competing risk models

The above described concepts and procedures have to be somewhat modified if we have the reason to think that failure is not a single event. Competing risk situation arises when two or more hazards exist, which may cause the failure. In the Hungarian retirement setting, there are two well separable - both absorbing - ways out of the labor force: disability pension as old age pension.

In this section, following Jenkins (2004b) it will be shown why the multinomial logit setup is a natural extension of the previously discussed logit model in the presence of multiple destination states.

Uncorrelated risks

Let $j = (1, 2)$ be the type of the destination state, 1=disability pension, 2=retirement. The censored state can be denoted as j=0. The discrete time hazard of exit at any period is now the sum of the hazards of the two specific exit routes:

$$h_i(t) = h_{1_i}(t) + h_{2_i}(t) \tag{4.26}$$

The likelihood contribution of the individual exiting to disability pension is (assuming stock sampling as before):

$$L_i^1 = \frac{h_{1_i}(t)S_i(t-1)}{S_i(u_i)} = \frac{\frac{h_{1_i}(t)}{1-h_i(t)}S_i(t)}{S_i(u_i)} \tag{4.27}$$

The likelihood contribution of the individual exiting to retirement is:

$$L^2 = \frac{h_2(t)S(t-1)}{S_i(u_i)} = \frac{\frac{h_{2_i}(t)}{1-h_i(t)}S_i(t)}{S_i(u_i)} \tag{4.28}$$

And finally, the likelihood contribution of being censored, therefore staying in the state until the end of the period is:

$$L_i^0 = S_i(t) = \prod_{k=u_i+1}^{t} (1 - h_i(k) = \prod_{k=u_i+1}^{t} (1 - h_{1_i}(k) - h_{2i}(k)) \tag{4.29}$$

Incorporating these three likelihoods in the overall likelihood contribution of the individual, the following can be written:

$$L_i = \left(L_i^1\right)^{\delta_1} \left(L_i^2\right)^{\delta_2} \left(L_1^0\right)^{1-\delta_1-\delta_2} \tag{4.30}$$

where $\delta_1 = 1$ if the person enters into disability pension, 0 otherwise. Similarly, $\delta_2 = 1$ if the person enters into old age pension, 0 otherwise.

The following form for the destination specific hazard can be assumed:

$$h_j(t) = \frac{\exp(\beta_j X)}{1 + \exp(\beta_1 X) + \exp(\beta_2 X)} \tag{4.31}$$

$$j = 1, 2 \tag{4.32}$$

, where the vector of X-es again includes dummy variables for the different time periods, to capture duration dependence. And the hazard of the third event, staying employed is:

$$1 - h_1(t) - h_2(t) = \frac{1}{1 + \exp(\beta_1 X) + \exp(\beta_2 X)} \tag{4.33}$$

Substituting the hazard rates back to the likelihood contribution of the individual:

$$L_i = \left[\frac{\exp(\beta_1 X)}{1 + \exp(\beta_1 X) + \exp(\beta_2 X)}\right]^{\delta^1} \left[\frac{\exp(\beta_1 X)}{1 + \exp(\beta_1 X) + \exp(\beta_2 X)}\right]^{\delta^2} \left[\frac{1}{1 + \exp(\beta_1 X) + \exp(\beta_2 X)}\right]^{1-\delta^1-\delta^2} \prod_{k=u+1}^{t-1} \left[\frac{1}{1 + \exp(\beta_1 X) + \exp(\beta_2 X)}\right] \tag{4.34}$$

In the person-period dataset, described above, this is the likelihood contribution of all the periods at risk experienced by the person. The sample likelihood is the product of all these individual likelihoods, and leads to the sample likelihood of a multinomial logit. It should be noted that in the present form, unobserved individual-specific heterogeneity is not included in the model.

The parameters estimated from this model have a very similar interpretation to those from the logit model: they have a proportional effect on the relative odds of the given event, compared to the base event (staying in the state).

Correlated risks

A multinomial logit specification carries the strong assumption of the independence of irrelevant alternatives (IIA).Verbally IIA requires that the predicted odds of two alternatives stays the same if any of the other alternatives is removed from the set of alternatives. Formally it means the following:

$$\frac{h_{ij}(t)}{h_{i0(t)}} = \exp(\beta_1 X_{ijt}) \tag{4.35}$$

Following the logic of the famous red bus -blue bus example of McFadden (1974) (Cited byWooldridge (2002)) this can be verbalized as follows. Suppose for some people for some reason only the two choices exist, continuing to work and old age pension,and these people, predicted from their observed characteristics (like age, health, etc.) would choose old age pension with a three times higher probability than work (this means, would distribute themselves between old age pension and working as 75% and 25%. By adding the third possibility of disability pension, we would expect this relative probability to go down: several people who would have chosen old age pension in the first instance would choose now disability pension, but only a few people who would be predicted to choose, leading to a new distribution, say, 40% for old age pension, 40% for disability pension, and 20% working. But now, the relative odds for old age pension and working is 40% divided by 20%, ie. two, in contrast with the original three. In order to the IIA assumption to hold, the new distribution of people among the three states should be, for example, 30%, 30% and 10, therefore, the predicted probability of old age pension and continuing working should fall with the same proportion. If we can assume that those unobserved factors leading to old age pension are similar to those leading to disability pension are similar, the IIA assumption cannot hold. It is possible to test for the validity of the IIA assumption with a Hausman's specification test Green (2008). The logic of the test is the following: if one subset (in our case, only one) of the choices is truly irrelevant, omitting it will not lead to different parameter estimates. The IIA assumption can be relaxed in multiple choice modelling, by specifying a multinomial probit, a mixed logit, or a nested logit model.

A mixed logit specification

As the test results described below no evidence against the IIA assumption was found. This is surprising, because it would be expected the individuals who choose one of the types of retirement are similar in unobserved characteristics. Still, unobserved heterogeneity raises the same problems in competing risk framework as in the single destination situation. A mixed logit (random effects multinomial logit could offer a solution for this problem.[2]

[2]The challenge those who try to estimate a random effect multinomial logit is well signallad by the fact that for the software STATA three independent routines exist for the estimation of mixed logit models: GLLAMM, Mixlogit, and XTMelogit, which differ in the type of computation involved.

Chapter 5

Data

5.0.9 Variables

Both time varying and time constant variables are included as regressors. The analysis is carried out separately for men and women. The self rated health variable is an answer to a single survey question: How would you rate your satisfaction with your health (scale incrasing by one, from 0: not at all to 10 "perfect").

The laboour force participation variable is derived from a question asking for the current labour force status. Among the possible answers there is employed, unemployed, working while retired, receiving disability benefits, receiving old age pension ,self-emplyed, entrepreneur, on maternity leave, housewife, other non-working household member. I used the self-reported disability pension or old age pension status as my retirement variables. All other states of labour force participation I grouped in one "in the labour force" group. Also including here unemployment may seem to be somewhat brave, however it can be justified by the present focus on retirement as an absorbing state. Lagged unemployment status may. however influence retirement, therefore this variable is included among the regressors,

Health is measured with the lagged self rated health variable, and also as the change in health state from the previous period. Health was not measured in 1995, for this year the average of the 1994 and the 1996 variables was imputed. Because of attrition, many individuals present in the 1995 wave are missing from the 1996 wave. For these people, the 1994 values was imputed. Other, occasionally missing values in the health measure were also imputed using this procedure. Wage was measured as the yearly income from the main occupation, udjusted to 1992, using the consumer price index Hivatal (2008). The lagged wage is included among the regressors, since it can be rightfully assumed that adjustment to a changed wage offer – in this case by transition out of the labour market – takes time. Missing values of the wage variable were imputed using the "last observation carried forward" method, by inserting the last observed wage instead of the missing wage variable. Three dummies are used for the highest educational attainment: vocational, secondary and higher education, while primary school or less is the baseline category.

Duration depedence is captured by the inclusion of dummy variables for the years 1993–1996, taking 1997 as baseline. Beside period dummies, age is also included in the regressor. This raisis a question if identification: it is hard to make a difference between the effect of being one year older, or experiencing a new survey year. This way, ceteris paribus, the year dummies are there to capture macro-influences on the probability of retirement for people of the same age across time periods. Since legal age is expected to be a very powerful predictor of retirement, I also include a dummy variables for legal retirement age, which is calculated separately for each individual and each time period, taking into account the year of birth ad the gender of the individual.

5.0.10 Descriptive statistics

The two tables below represent the labour market status of men and women separately, and their mean age for each subgroups. Since the dataset used is an unbalanced panel dataset, keeping each individual only until retired or censored, the frequencies at both types of retirement represent the actual transition frequencies. Similarly, the number of those staying in the labor force (either employed or unemployed) and those who are lost to follow up ("Attr".) for the next period are also listed. It can be seen that a large fraction of the initial sample was lost to follow up due to attrition: 43% of men, and 48% of women.

Table 5.1: **Age and transition of men, by waves**

Men										
	Emp.		**Unemp.**		**Disab. pens.**		**Old. pens.**		$\sum$ **N**	**Attr.**
Wave	*N*	*age (sd)*	*N*	*age (sd)*	*N*	*age (sd)*	*N*	*age (sd)*		
1992	416	49.28 (5.85)	50	48.4 (5.25)	-	-	-	-	466	-
1993	364	49.67 (5.50)	66	49.34 (5.37)	15	53.13 (6.25)	21	59.57 (1.98)	466	53
1994	296	50.05 (5.02)	45	50.33 (5.51)	12	50.83 (5.44)	24	59.75 (2.23)	377	54
1995	241	50.46 (4.89)	12	50 (5.09)	14	53 (3.82)	20	58.3 (2.00)	287	35
1996	188	51.07 (4.70)	11	49.72 (4.85)	6	50.5 (4.59)	13	58.61 (4.21)	218	50
1997	131	51.82 (4.97)	7	49.14 (2.60)	2	53 (7.07)	5	58.4 (1.81)	149	

Table 5.2: **Age and transition of women, by waves**

Wom.										
	Emp.		Unemp.		Dis. pens.		Old pens.		$\sum$ **N**	**Attr.**
Wave	N	age	N	age	N	age	N	age		
1992	417	48.08	32	47.62	-	-	-	-	449	-
		(4.88)		(5.15)						
1993	382	48.58	27	47.44	9	50.4	31	55.87	449	57
		(4.62)		(3.55)		(3.64)		(4.28)		
1994	297	49.16	13	49.23	21	49.42	21	54.57	352	49
		(4.483)		(3.05)		(3.00)		(2.08)		
1995	228	49.58	8	48	5	51.6	20	55.45	261	31
		(4.10)		(2.67)		(1.51)		(2.85)		
1996	173	50.09	6	48.5	3	49.3	23	54.95	205	78
		(4.05)		(2.50)		(4.16)		(2.05)		
1997	103	50.79	7	49.14	10	50.1	16	55.75	136	
		(4.01)		(2.47)		(2.07)		(3.82)		

Chapter 6

Results

6.0.11 Pooled destination states

The result of four models are compared here. The proportional odds specification, estimated both with a plain logit model (not accounting for unobserved heterogeneity), and with a random effects logit model.[1] The proportional hazards specification complementary log-log model is also estimated both ways, using the *pgmhaz* procedure of Jenkins (2004a) and the complementary log-log model using the cloglog procedure of STATA. There is a negative significant effect of health on the hazard of retirement for men, both in the gamma mixture and in the complementary log-log specification. While in the table below the original coefficients are reported, it is worth to look at the exponentiated coefficients for some of the variables of interest. First, the cloglog model for men suggests that the hazard ratio for having a one unit higher self-rater health 0.79. This means, for those having a one unit higher self rated health in the previous period the hazard of retirement is 79% of those having a lower self rated health. For women, this effect is somewhat smaller; 81%. The results from the gamma mixture model and from the cloglog model are very similar, although the significance level of the constant term from the gamma distribution suggests that there is neglected heterogeneity in the model.

[1]STATA estimates random effect logit by numerical integration. (Gauss hermit Quadratures, see http://www.stata.com/help.cgi?xtlogit)

It was expected that the dummy variable indicating the legal age of retirement will be very powerful, the value of the coefficient is around one, which indicates a hazard ratio of 2.5 for those close to the legal retirement age. We can check our expectations on the effects of neglected heterogeneity on duration dependence as well.

Having a look at the logit and random effects logit estimates, we find very similar magnitudes of coefficients. For men, there is almost no difference between the logit and random effects estimates. For women, however, there is a very obvious difference. Now the interpretation is also different: for those women having one unit higher self rated health, the odds of being retired versus staying in the labour force is 70% of those with a one unit lower self rated health.

Table 6.1: **Complementary log-log coefficients**

parameter	mix., men	cloglog., men	mix, women	cloglog, wom
health_d	-0.077	-0.077	-0.126*	-0.063
	(0.044)	(0.044)	(0.058)	(0.039)
health_1	-0.230***	-0.230***	-0.346***	-0.202***
	(0.040)	(0.039)	(0.074)	(0.037)
wage	-0.000**	-0.000**	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
age	0.200***	0.200***	0.388***	0.149***
	(0.041)	(0.022)	(0.087)	(0.016)
dum93	0.52	0.52	-2.001***	-1.026***
	(0.437)	(0.431)	(0.505)	(0.285)
dum94	0.635	0.635	-1.068**	-0.533*
	(0.436)	(0.430)	(0.400)	(0.269)
dum95	0.974*	0.975*	-1.164**	-0.854**
	(0.435)	(0.428)	(0.390)	(0.299)
dum96	0.762	0.762	-0.671	-0.45
	(0.457)	(0.450)	(0.349)	(0.287)
secondary	0.071	0.071	0.257	0.092
	(0.296)	(0.289)	(0.363)	(0.206)
vocational	0.339	0.339	0.003	-0.157
	(0.228)	(0.222)	(0.431)	(0.248)
higher	-0.506	-0.506	-0.646	-0.37
	(0.380)	(0.359)	(0.515)	(0.336)
legal	0.943***	0.943***	1.039**	1.215***
	(0.282)	(0.229)	(0.325)	(0.198)
unemp	0.503	0.503	1.046**	1.092***
	(0.264)	(0.263)	(0.399)	(0.287)
Constant	-12.119***	-12.119***	-17.993***	-7.793***
	(2.347)	(1.322)	(3.904)	(0.894)
Gamm const	-12.726		0.722*	
	(195.617)		(0.360)	
N	1497	1497	1403	1403

p<0.05, ** p<0.01, *** p<0.001

6.0.12 Multiple destination states

To test for the independence of irrelevant alternatives assumption, Hausman-type tests were carried out fro men and women separately, fro both choices. The procedure was the following: the coefficients from the multinomial logit model were compared with the coefficients of a multinomial logit model restricicited for only one of the choices (practically running a multinomial logit on a restricted sample.) None of the Hausman tests has shown evidence against the IIA assumption, therefore the multinomial logit can be accepted as an appropriate choice for modelling competing risks in this situation. In Table 5 below the marginal effects are reported, ie. the partial effects of each variable on unconditional probability of retirement (in contrast to the hazard!). It is not very surprisig, that the coefficient estimates for the two ways of retirement are very different. While the lagged health has a significant effect on the probability of disability pension, it does not have a significant effect on the probability of old age pension. For men, the only significant predictor of old age pension is age (a positive effect, as expected), and higher education (those with higher education have a lower hazard of retiremnt. Naturally, the exponentiated coeffiecients have a meaning in themselves. Here, the coefficient of lagged health is -0.59 in the disability pension equation for men, which equals a very small odds ratio of 55%. Just to emphasize the interpretation of this coefficient, somebody who has a one unit higher self rated health ceteris paribus has a 55% smaller odds of going in disability pension the next year. It should be noted that the effect of wage is significant negatively in the disability equations for both genders, while its coefficient is practically small.

Table 6.2: **Logit and RE logit estimated coefficients**

Variable	logit, men	RE logit, men	logit, wom.	RE logit, wom.
health_d	-0.055	-0.055	-0.085	-0.117
	(0.055)	(0.055)	(0.050)	(0.064)
health_1	-0.268***	-0.268***	-0.241***	-0.346***
	(0.046)	(0.046)	(0.043)	(0.086)
wage	-0.000**	-0.000**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
age	0.220***	0.220***	0.190***	0.315***
	(0.026)	(0.026)	(0.023)	(0.085)
dum93	0.557	0.557	-1.158***	-2.095**
	(0.476)	(0.476)	(0.336)	(0.737)
dum94	0.703	0.703	-0.553	-1.061*
	(0.475)	(0.475)	(0.318)	(0.497)
dum95	1.092*	1.092*	-0.947**	-1.291**
	(0.474)	(0.474)	(0.353)	(0.465)
dum96	0.828	0.828	-0.475	-0.685
	(0.499)	(0.499)	(0.342)	(0.407)
secondary	0.05	0.05	0.08	0.151
	(0.337)	(0.337)	(0.244)	(0.360)
vocational	0.421	0.421	-0.169	-0.126
	(0.269)	(0.269)	(0.292)	(0.434)
higher	-0.488	-0.488	-0.439	-0.696
	(0.399)	(0.399)	(0.376)	(0.567)
legal	1.150***	1.151***	1.397***	1.661***
	(0.291)	(0.291)	(0.259)	(0.376)
unemp	0.488	0.488	1.157***	1.317**
	(0.312)	(0.312)	(0.346)	(0.472)
Constant	-12.947***	-12.948***	-9.460***	-15.050***
	(1.483)	(1.484)	(1.222)	(3.927)
lnsig2u				
Constant		-10.429		0.843
		(21.913)		(0.806)
N	1497	1497	1403	1403

p<0.05, ** p<0.01, *** p<0.001

Table 6.3: **Multinomial logit marginal effects at mean**

Variable	men, disab.	men, old age	wom., disab	wom, old age.
health_d	-0.001	0	-0.001*	0.001
	(0.001)	(0.001)	(0.001)	(0.002)
health_lag	-0.004***	0	-0.002**	-0.002
	(0.001)	(0.000)	(0.001)	(0.001)
wage	-0.000**	0	-0.000**	0
	(0.000)	(0.000)	(0.000)	(0.000)
age	0.001*	0.003***	0	0.007***
	(0.000)	(0.001)	(0.000)	(0.001)
secondary (d)	0.003	-0.001	0.003	-0.008
	(0.004)	(0.003)	(0.002)	(0.008)
vocational (d)	0.006	0.003	-0.002	-0.005
	(0.004)	(0.003)	(0.001)	(0.009)
higher (d)	0.003	-0.009**	0.006	-0.025**
	(0.005)	(0.003)	(0.006)	(0.008)
dum93 (d)	0.011	0.001	-0.003	-0.033***
	(0.010)	(0.005)	(0.002)	(0.009)
dum94 (d)	0.012	0.005	0.001	-0.027***
	(0.011)	(0.007)	(0.002)	(0.008)
dum95 (d)	0.017	0.009	-0.003	-0.022**
	(0.015)	(0.009)	(0.002)	(0.007)
dum96 (d)	0.01	0.007	-0.004*	-0.008
	(0.012)	(0.009)	(0.002)	(0.009)
legal (d)	-0.003	0.021	0	0.089***
	(0.003)	(0.011)	(0.003)	(0.025)
unemp (d)	-0.002	0.02	-0.001	0.110*
	(0.002)	(0.011)	(0.002)	(0.045)
N	1497	1497	1403	1403

(d) for discrete change of dummy variable from 0 to 1
p<0.05, ** p<0.01, *** p<0.001

Chapter 7

Conclusion and discussion

While the above analysis was a challenging technical exercise to deal with micro panel data in discrete duration setting, in contrast to the international literature, no other effect of health on early retirement was found than that working through the disability pension. While it could have been expected that those in better health may actually retire later than the legal age, no such effect was found. One explanation for this could be that in Hungary in the investigated period retirement was not so much a choice but rather something happening to those becoming unneeded in the labour market.

For further research it would be interesting to investigate newer Hungarian micro datasets, to compare the effects found here with those affecting retirement decisons nowadays. A structural approach to modellig retirement is probably also unavoidable.

For the other focal point of the topic, the suspected measurement error in the self rated health variable coold not have been dealt with in this analysis. It would be challenging to look at more detailed health meaures in Hungarian setting, to gain more insights in the actual interplay between health and retirement.

Bibliography

BOUND, J. (1989): "Self-Reported vs. Objective Measures of Health in Retirement Models," NBER Working Papers 2997, National Bureau of Economic Research, Inc.

BOUND, J., T. STINEBRICKNER, AND T. WAIDMANN (2007): "Health, Economic Resources and the Work Decisions of Older Men," NBER Working Papers 13657, National Bureau of Economic Research, Inc, `http://ideas.repec.org/p/nbr/nberwo/13657.html`.

CSERES-GERGELY, Z. (2007): "Inactivity in Hungary-the persistent effect of the pension system," Budapest Working Papers on the Labour Market BWP-2007/1, Institut of Economics. Hungarian Academy of Sciences. Corvinus University, Budapest (Hungary).

CURRIE, J., AND B. C. MADRIAN (1999): *Handbook of Labor Economics v 3c*chap. Health, Health Insurance and the Labour Market, pp. 3309–3416. New York: Elsevier.

DESCHRYVERE, M. (2004): "Health and Retirement. An Update of the Literature," Discussion Papers 932, The Research Institute of the Finnish Economy, `http://ideas.repec.org/p/rif/dpaper/932.html`.

GREEN, W. H. (2008): *Econometric Analysis*. New York University.

GROSSMAN, M. (1972): "On the Concept of Health Capital and the Demand for Health," *The Journal of Political Economy*, 80(2), 223–255, `http://www.jstor.org/stable/1830580`,.

Hivatal, K. S. (2008): "Website,"
http://portal.ksh.hu/portal/page?_pageid=37,447530&_dad=portal&_schema=PORTAL.

Hostenkamp, G., and M. Stolpe (2008): "The Social Costs of Health-related Early Retirement in Germany: Evidence from the German Socio-economic Panel," Kiel Working Papers 1415, Kiel Institute for the World Economy,
http://ideas.repec.org/p/kie/kieliw/1415.html.

Jenkins, S. P. (2004a): "PGMHAZ8: Stata module to estimate discrete time (grouped data) proportional hazards models," Statistical Software Components, Boston College Department of Economics,
http://ideas.repec.org/c/boc/bocode/s438501.html.

——— (2004b): "Survival analysis," Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK. Downloadable from
http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/pdfs/ec968lnotesv6.pdf.

McFadden, D. L. (1974): *Frontiers in Econometrics*chap. Conditional Logit Analysis of Qualitative Choice Analysis. New York: Academic Press.

Nicolett, C., and C. Rondellini (2006): "The (Mis)specification of Discrete Time Duration Models with Unobserved Heterogeneity: a Monte Carlo Study," ISER Working Paper 53, Colchester: Univesity of Essex,
http://www.iser.essex.ac.uk/pubs/workpaps/.

Rice, N., J. Roberts, and A. M. Jones (2007): "Sick of work or too sick to work? Evidence on health shocks and early retirement from the BHPS," Working Papers 2007002, The University of Sheffield, Department of Economics,
http://ideas.repec.org/p/shf/wpaper/2007002.html,.

Spataro, L. (2002): "New Tools in Micromodeling Retirement Decisions: Overview and Applications to the Italian Case," CeRP Working Papers 28, Center for Research on Pensions and Welfare Policies, Turin (Italy),
http://ideas.repec.org/p/crp/wpaper/28.html.

Stock, J. H., and D. A. Wise (1990): "Pensions, the Option Value of Work, and Retirement," *Econometrica*, 58(5), 1151–80,
http://ideas.repec.org/a/ecm/emetrp/v58y1990i5p1151-80.html.

Van den Berg, G. J. (2001): *Handbook of Econometrics*vol. 5, chap. Duration Models: Specification, Identification and Multiple Durations, pp. 3381–3460. Elsevier.

Wooldridge, J. M. (2002): *Econometric Analysis of Cros Section and Pane Data*. MIT Press.

Appendix A

Deriving the sample likelihood for the single destination case

The sample likelihood of on individual whose spell would be completely observed - which, to emphasize, is not possible with this dataset - and has retired in year $1992 + z_i$ is

$$L_i^{flow,uncens,} = \Pr(T_i = t_i) = f_{it}$$
$$= h_{it} S_i(t-1) = \frac{h_{it}}{1-h_{it}} \prod_{k=1}^{t} (1 - h_{ik}) \quad \text{(A.1)}$$

with other words, the probability that the individual retires exactly after t_i periods, counting from the beginning of his/her spell.

The likelihood contribution of someone, whose spell was observed from the beginning, but was right censored at at the t_i-th period (which may differ by individual because of the attrition):

$$L_{.i}^{flow,cens} = \Pr(T_i > t_i) = S_i(t) = \prod_{k=1}^{t} (1 - h_{ik}) \quad \text{(A.2)}$$

Now we have to take into account that we have a sock sample: we have a sample of those individuals with an ongoing spell in 1992, which means, at the Δt_i-th period, counting from $\tau_i = 0$, where the individual was 45 years old. This is basically conditioning on the survival up to this period. We have to inflate the likelihood function in both of the above cases by the survival function evaluated at the Δt_i-th period.

Therefore for the uncensored observations

$$L_i^{stock,uncens,} = \frac{f_{it}}{S(\Delta t_i)} \tag{A.3}$$

Similarly, for the censored observations

$$L_i^{stock,cens,} = \frac{S_i(t)}{S(\Delta t_i)} \tag{A.4}$$

Now we can easily constrict the sample likelihood, where there are K uncensored and $N-K$ censored observations in the sample of N.

$$L = \prod_{i=1}^{K} \frac{f_{it}}{S(\Delta t_i)} \prod_{i=K+1}^{N} \frac{S_i(t)}{S(\Delta t_i)} \tag{A.5}$$

By introducing an indicator variable $c_i = 1$ for completed spells and $c_i = 0$ for censored observations, we can rewrite this as

$$L = \prod_{i=1}^{N} \left[\frac{f_{it}}{S(\Delta t_i)}\right]^{c_i} \left[\frac{S_i(t)}{S(\Delta t_i)}\right]^{1-c_i} \tag{A.6}$$

Substituting the appropriate formulas to the density and survival functions we get

$$L = \prod_{i=1}^{N} \left[\frac{\frac{h_{it}}{1-h_{it}} \prod_{k=1}^{t} (1-h_{ik})}{S(\Delta t_i)} \right]^{c_i} \left[\frac{\prod_{k=1}^{t} (1-h_{ik})}{S(\Delta t_i)} \right]^{1-c_i} \tag{A.7}$$

Now denote the period of the sampling $\Delta t_i = u_i$.

We can write the discrete survival function until period u_i as

$$S(u_i) = \prod_{k=1}^{u} (1-h_{ik}) \tag{A.8}$$

.Plugging this back to the sample likelihood we get

$$\begin{aligned} L = & \prod_{i=1}^{N} \left[\frac{\frac{h_{it}}{1-h_{it}} \prod_{k=1}^{t} (1-h_{ik})}{\prod_{k=1}^{u} (1-h_{ik})} \right]^{c_i} \left[\frac{\prod_{k=1}^{t} (1-h_{ik})}{\prod_{k=1}^{u} (1-h_{ik})} \right]^{1-c_i} && \text{(A.9)} \\ = & \prod_{i=1}^{N} \left[\frac{h_{it}}{1-h_{it}} \prod_{k=u_i+1}^{t} (1-h_{ik}) \right]^{c_i} \left[\prod_{k=u_i+1}^{t} (1-h_{ik}) \right]^{1-c_i} && \text{(A.10)} \end{aligned}$$

Now if we only concentrate on the individual likelihood:

$$\begin{aligned} L_i = & \left[\frac{h_{it}}{1-h_{it}} \prod_{k=u_i+1}^{t} (1-h_{ik}) \right]^{c_i} \left[\prod_{k=u_i+1}^{t} (1-h_{ik}) \right]^{1-c_i} && \text{(A.11)} \\ = & \left[\frac{h_{it}}{1-h_{it}} \right]^{c_i} \prod_{k=u_i+1}^{t} (1-h_{ik}) && \text{(A.12)} \end{aligned}$$

We can create a new indicator variable

$$y_{it} = \begin{Bmatrix} 1 \text{ if } c_i = 1 \text{ and } k = t \\ 0 \quad\quad \text{otherwise} \end{Bmatrix} \tag{A.13}$$

Verbally, we create a binary variable which takes the value of 1 for each individual only for the period when failure (transition to retirement) happens. In all other states (for censored individuals for each period, and for individuals with completed spells before the period of failure) this variable takes the value of 0.

Now we can write up the individual likelihood as follows:

$$L_i = \prod_{k=u_i+1}^{t} (1 - h_{ik})^{(1-y_{ik})} h_{ik}^{y_{ik}} \tag{A.14}$$

It is more convenient to write up the individual likelihood

$$\log L_i = \sum_{k=u_i+1}^{t} y_{ik} \log h_{ik} + (1 - y_{ik}) \log(1 - h_{ik}) \tag{A.15}$$

Appendix A

Deriving the complementary log-log hazard function from the proportional hazard model

The main assumption is the following: the hazard and the baseline hazard are proportional to each other, and the factor of proportionality is again $\exp(\beta' x_i) = \lambda_i$.Therefore,

$$h_i(k, x) = h_0(k)\lambda_{it} \quad \text{(A.1)}$$

$$S_i(t, x_i) = \prod_{k=1}^{t}(1 - h(k, x_{it})) = \prod_{k=1}^{t}(1 - h_0(k)\lambda_{it}) \quad \text{(A.2)}$$

$$\log S(t, x_{it}) = \sum_{k=1}^{t} \log(1 - h_0(k)\lambda_{it}) \quad \text{(A.3)}$$

if $h_0(k)\lambda_i$ is small enough, we can write the approximation:

$$\log S_i(t, x_{it}) = -\sum_{k=1}^{t} h_0(k)\lambda_{it} = -\lambda_i \sum_{k=1}^{t} h_0(k) \quad \text{(A.4)}$$

denoting

$$\sum_{k=1}^{t} h_0(k) = H(t) \tag{A.5}$$

$$S(t, x_{it}) = \exp(-\lambda_{it} H(t)) \tag{A.6}$$

$$h_i(t, x_i) = \frac{S_i(t, x_{it}) - S(t-1, x_{it})}{S(t-1, x_{it})} = \tag{A.7}$$

$$= 1 - \frac{S(t, x_{it})}{S(t-1, x_{it})} = \exp(\lambda_i H(t) - H(t-1)) \tag{A.8}$$

This implies that

$$\log(1 - h_i(t, x_{it})) = \lambda_{it}(H(t-1) - H(t)) \tag{A.9}$$

Taking one more logarithm,

$$\log(-\log(1 - h_i(t, x_{it})) = \beta' X_{it} + \log(H(t) - H(t-1) \tag{A.10}$$

In our discrete time specification,

$$H(t) - H(t-1) = h_0(t), \tag{A.11}$$

which is the baseline hazard. Let's denote if for now as $h_0(t) = \gamma_t$.

After some more manipulation, we get the following:

$$h_{it}(t, x_{it}) = 1 - \exp(\exp(\beta' X_{it} + \gamma_t). \tag{A.12}$$