Lagged Duration Dependence in Mixed Proportional Hazard Models

M. Picchio

Discussion Paper 2011-37

Institut de Recherches Économiques et Sociales
de l'Université catholique de Louvain

Lagged Duration Dependence in Mixed Proportional Hazard Models*

Matteo Picchio†

October 24, 2011

Abstract

We study the non-parametric identification of a mixed proportional hazard model with lagged duration dependence when data provide multiple outcomes per individual or stratum. We show that the information conveyed by the within strata variation can be exploited to non-parametrically identify lagged duration dependence in more general models than in the literature.

Keywords: lagged duration dependence, mixed proportional hazard models, identification, multiple spells, parallel data.
JEL classification codes: C14, C41.

1 Introduction

The identification of lagged duration dependence in a single-risk mixed proportional hazard (MPH) model is shown in Honoré (1993). Regressor variation and independence between regressors and individual heterogeneity (along with mixed proportionality) are assumptions required for identification. Frijters (2002) sheds further light on this issue by proving identification without exploiting regressor variation and without imposing restrictions on the tail of the unobserved heterogeneity distribution. The price to pay is that the same baseline hazard function

*The author acknowledges financial support by Stichting Instituut GAK, through ReflecT, the Research Institute for Flexicurity, Labor Market Dynamics and Social Cohesion at Tilburg University and by Fonds Wetenschappelijk Onderzoek (FWO).

†Department of Economics, CentER, and ReflecT, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands; tel: +31 (0)134662534; fax: +31 (0)134663042; e-mail: m.picchio@uvt.nl. Sherppa, Department of Social Economics, Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium; e-mail: matteo.picchio@ugent.be. IZA, Germany.

must be imposed in the initial spell and in the subsequent spell. This assumption is too restrictive in many empirical studies, for instance when analysing the impact of unemployment duration on subsequent employment stability, the time until development of HIV infection and outbreak of AIDS, or mother's age until childbirth and child survival.

In this paper, we show that if parallel data (Hougaard, 2000) are available, i.e. multiple realizations of subjects belonging to the same stratum related through unobserved components,[1] the assumption on spell-constant baseline hazard functions can be relaxed without needing regressor variation and restrictions on the mixing distribution. Parallel durations are observed for example when focusing on siblings as they share genetic material and family background (family fixed effect), employees at the same workplace as they share the same environment (firm fixed-effect), and individuals experiencing repeated unemployment and employment spells as their labour market career is affected by time-invarying personal characteristics (individual fixed effect).

There is increasing data availability and interest to apply survival analysis with parallel durations. As a matter of fact, the identification result in this paper can be useful in many empirical frameworks with lagged duration dependence, as far as the grouping results from multiple realizations for the same individual, household, firm, geographical area, school, etc. In economics, applications that exploit multiple durations generated by a single unit are, for example, Doiron and Gørgens (2008) and Cockx and Picchio (2011a,b) who analyse employment stability after unemployment events, Bonnal et al. (1997) who study the impact of training programs on subsequent labour market performance, and Lindeboom and Kerkhofs (2000) and Frederiksen et al. (2007) who focus on, respectively, sickness absenteeism and job tenure with strata at firm level. Further examples can be found in biostatistics: Guo and Rodríguez (1992), Sastry (1997), and Ridder and Tunalı (1999) study child mortality with strata at household level; Therneau and Hamilton (1997) investigate several approaches to recurrent events data, such as recurrent infections in AIDS patients or multiple infarcts in coronary study.

[1] As in Abbring and van den Berg (2003a,b), by stratum we mean a single individual for whom the outcome process is observed (at least) twice or a group of individuals sharing similar unobserved components.

2 MPH Models with Lagged Duration Dependence

The MPH model with lagged duration dependence is characterized in Honoré (1993) by the following hazard functions

$$\theta_1(t_1|x,v) = z_1'(t_1)\phi_1(x)v_1, \tag{1}$$
$$\theta_2(t_2|t_1,x,v) = z_2'(t_2)\phi_2(x)h(t_1)v_2, \tag{2}$$

where $t_s \in \Re_+$ is the duration of the sth spell, z_s' is the baseline hazard, a function of the elapsed duration in spell s, $\phi_s(x)$ is the systematic part, a function of a set of observed characteristics x. $h(t_1)$ is the lagged duration dependence function, i.e. the effect of the duration of the first spell on the hazard function of the subsequent spell. Finally, v_1 and v_2 are non-negative time-constant terms, with distribution function $G(v_1, v_2)$. They capture unobserved heterogeneity that could determine the duration of the first and second spells, respectively. This model could be used, for instance, to investigate the impact of unemployment duration on future (un)employment durations.[2]

Honoré (1993) proves the non-parametric identification of model (1)-(2) assuming that the unobserved heterogeneity is orthogonal to x and under a finite moment condition on the unobserved heterogeneity distribution G. However, the orthogonality condition might be too stringent and not plausible in many applications.

Frijters (2002) attempts to avoid the orthogonality condition by considering the following MPH model

$$\theta_1(t_1|v) = z'(t_1)v, \tag{3}$$
$$\theta_2(t_2|t_1,v) = z'(t_2)h(t_1)v. \tag{4}$$

He shows that z, h, and G are non-parametrically identified without requiring regressor variation and any assumption on the mixing distribution. Nevertheless, a critical price is paid in terms of model flexibility: in model (3)-(4), the baseline hazards and the unobserved heterogeneity terms have to be the same in the initial spell and in the subsequent spell. This type of model has been estimated since the 1980s to understand whether unemployment duration affects the duration of

[2]The outcome variable t_2 might not necessarily be a duration outcome, but it might represent any other non-negative outcome variable, such as starting wages (Cockx and Picchio, 2011b), earnings (Arni et al., 2009), working hours, prices, etc. The hazard function in (2) would then fully characterize the corresponding distribution function.

subsequent unemployment spells (e.g., Heckman and Borjas, 1980). Nonetheless, this restriction might not be natural for a lot of applications; for example, investigations into the way unemployment duration t_1 affects subsequent job or employment stability t_2. This arises the question of under which conditions we can allow for different baseline hazards and different unobserved heterogeneity components in each spell without requiring the variation of orthogonal regressors.

In what follows we show that if data provide information on (t_1, t_2) at least twice in a stratum which is characterized by a single realization of (v_1, v_2), it is possible to allow the baseline hazards and the fixed-effects to be spell-specific without needing variation of exogenous regressors and assumptions about the mixing distribution.

The MPH model with lagged duration dependence that we study is

$$\theta_1^k(t_1^k|v) = z_1^{k\prime}(t_1^k)v_1, \tag{5}$$

$$\theta_2(t_2^k|t_1^k, v) = z_2^{k\prime}(t_2^k)h^k(t_1^k)v_2, \qquad (k = 1, 2) \tag{6}$$

where the superscript $k = 1, 2$ identifies the recurrence of the outcome variables within a stratum. In a panel data framework k would be the kth time we observe the outcome variables (t_1, t_2) for a given individual. In twins or matched pairs studies (Holt and Prentice, 1974), k is sibling's identifier. Note that model (5)-(6) is more flexible than the one in (3)-(4), in that the baseline hazards are allowed to be different over-spells and within-stratum. Furthermore, differently from model (1)-(2), the regressors x do not enter the model specification, so that the analysis can be thought of as conditional on x.

We assume that conditional on (v_1, v_2), (t_1^1, t_2^1) and (t_1^2, t_2^2) are independent. With hazard functions specified as in (5) and (6), the joint survivor function of $(t_1^1, t_2^1, t_1^2, t_2^2)$ is

$$\begin{aligned} S(t_1^1, t_2^1, t_1^2, t_2^2) &= \int_{\Re_+^2} \exp\left\{-v_1\Big[\sum_{k=1,2} z_1^k(t_1^k)\Big] - v_2\Big[\sum_{k=1,2} z_2^k(t_2^k)h^k(t_1^k)\Big]\right\} dG(v_1, v_2) \\ &= \mathscr{L}_G\Big[\sum_{k=1,2} z_1^k(t_1^k), \sum_{k=1,2} z_2^k(t_2^k)h^k(t_1^k)\Big], \end{aligned} \tag{7}$$

where $\mathscr{L}_G(s_1, s_2)$ is the Laplace transform of G. In the identification analysis that follows, $S(t_1^1, t_2^1, t_1^2, t_2^2)$ is observed and taken to be known, as well as the subsurvival function $\partial S(t_1^1, t_2^1, t_1^2, t_2^2)/\partial t_2^k$, for $k = 1, 2$.

3 Identification Result

Theorem 1 *Functions G, z_1^k, z_2^k, and h^k, with $k = 1, 2$, in (7) are uniquely identified from the distribution of $\cap_{k=1,2}(T_1^k, T_2^k)$ under the following assumptions:*

A1 *$z_1^k(t)$ and $z_2^k(t)$, for $k = 1, 2$, are non-negative, differentiable, and strictly increasing $\forall t \in \Re_+$. $z_1^1(t^0) = z_2^1(t^0) = 1$ for some fixed $t^0 \in \Re_+$.*

A2 *h^1 and h^2 are non-negative on $\Re_+$. $h^1(t^*) = h^2(t^*) = 1$ for some fixed $t^* \in \Re_+$.*

Proof. Under Assumption A1, from the marginal distribution of (T_1^1, T_1^2) we can identify z_1^1 and z_1^2 by invoking Theorem 1 in Honoré (1993). Identification of the remaining functions is shows in steps. In step **(a)**, identification of lagged duration functions h^1 and h^2 is proven. Step **(b)** concerns identification of the integrated baseline hazards z_2^1 and z_2^2. Finally, step **(c)** deals with identification of the individual heterogeneity distribution G.

(a) From a large data set we can compute the subsurvival function

$$\frac{\partial S(t_1^1, t_2^1, t_1^2, t_2^2)}{\partial t_2^1} = z_2^{1\prime}(t_2^1) h^1(t_1^1) D_{s_2} \mathscr{L}_G \Big[\sum_{k=1,2} z_1^k(t_1^k), \sum_{k=1,2} z_2^k(t_2^k) h^k(t_1^k) \Big], \tag{8}$$

where $D_{s_2} \mathscr{L}_G(s_1, s_2) \equiv \partial \mathscr{L}_G(s_1, s_2)/\partial s_2$. We can also compute the subsurvival function

$$\frac{\partial S(t_1^1, t_2^1, t_1^2, t_2^2)}{\partial t_2^2} = z_2^{2\prime}(t_2^2) h^2(t_1^2) D_{s_2} \mathscr{L}_G \Big[\sum_{k=1,2} z_1^k(t_1^k), \sum_{k=1,2} z_2^k(t_2^k) h^k(t_1^k) \Big]. \tag{9}$$

If we divide the subsurvival function in (8) by the one in (9), the component related to the first derivative of the Laplace transform drops out. This is the advantage of having variation within strata. Consider an arbitrary $(t_2^1, t_1^2, t_2^2) \in \Re_+^3$ and pick $(t_1^1, t^*) \in \Re_+^2$. From the the ratio

$$\frac{\frac{\partial S(t_1^1, t_2^1, t_1^2, t_2^2)/\partial t_2^1}{\partial S(t_1^1, t_2^1, t_1^2, t_2^2)/\partial t_2^2}}{\frac{\partial S(t^*, t_2^1, t_1^2, t_2^2)/\partial t_2^1}{\partial S(t^*, t_2^1, t_1^2, t_2^2)/\partial t_2^2}} = \frac{\frac{z_2^{1\prime}(t_2^1) h^1(t_1^1)}{z_2^{2\prime}(t_2^2) h^2(t_1^2)}}{\frac{z_2^{1\prime}(t_2^1) h^1(t^*)}{z_2^{2\prime}(t_2^2) h^2(t_1^2)}} = \frac{h^1(t_1^1)}{h^1(t^*)}$$

we get identification of h^1 up to a constant. Identification of h^2 is similarly yielded considering an arbitrary $(t_1^1, t_2^1, t_2^2) \in \Re_+^3$ and picking $(t_1^2, t^*) \in \Re_+^2$.

(b) Taking the ratio of (8) over (9) and solving with respect to z_2^2 with the normalization $z_2^1(t^0) = 1$ yield

$$z_2^2(t_2^2) = \frac{h^1(t_1^1)}{h^2(t_1^2)} \int_0^{t_2^2} \left[\int_0^{t_0} \frac{\partial S(t_1^1, \tau^1, t_1^2, \tau^2)/\partial \tau^1}{\partial S(t_1^1, \tau^1, t_1^2, \tau^2)/\partial \tau^2} d\tau^1 \right]^{-1} d\tau^2. \tag{10}$$

Since h^1 and h^2 have already been identified, fixing $(t_1^1, t_1^2, t^0) \in \Re_+^3$ and letting t_2^2 vary over $\Re_+$ give identification of z_2^2.[3] Similar computations yield

$$z_2^1(t_2^1) = z_2^2(t_2^2) \frac{h^2(t_1^2)}{h^1(t_1^1)} \int_0^{t_2^1} \left[\int_0^{t_2^2} \frac{\partial S(t_1^1, \tau^1, t_1^2, \tau^2)/\partial \tau^2}{\partial S(t_1^1, \tau^1, t_1^2, \tau^2)/\partial \tau^1} d\tau^2 \right]^{-1} d\tau^1,$$

which identifies z_2^1 for arbitrary $(t_1^1, t_1^2, t_2^2) \in \Re_+^3$.

(c) All the functions entering the Laplace transform $\mathscr{L}_G$ have already been identified. Thereby, we can trace $\mathscr{L}_G$ on a non-empty open set by appropriately varying $(t_1^1, t_2^1, t_1^2, t_2^2)$. As $\mathscr{L}_G$ is real analytic, it is uniquely determined on $\Re_+^4$. Uniqueness of the Laplace transform implies identification of G and concludes the proof. ∎

This theorem states that assumptions on regressor orthogonality and on the moments (or the tail) of the mixing distribution are not necessary for model identification with within strata variation. Finally, the result can be extended to cover the identification of a model where the baseline hazards and lagged duration functions are conditional on x and the mixing distribution G depends on x.

References

Abbring, J.H. and G.J. van den Berg, "The Identifiability of the Mixed Proportional Hazards Competing Risks Model," *Journal of the Royal Statistical Society Series B*, 2003, *65* (3), 701–710.

__ **and** __ , "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 2003, *71* (5), 1491–1517.

Arni, P., R. Lalive, and J.C. van Ours, "How Effective Are Unemployment Benefit Sanctions? Looking Beyond Unemployment Exit," 2009. IZA Discussion Paper No. 4509.

[3] If $t_1^1 = t_1^2 = t^*$, this step of the proof is like Honoré's (1993) Theorem 1. Note however that it can be repeated for all $(t_1^1, t_1^2) \in \Re_+^2$. Since all the solutions to equations (10) should be the same, potentially testable overidentifying restrictions arise, similarly to Abbring and van den Berg (2003a) and Melino and Sueyoshi (1990).

Bonnal, L., D. Fougere, and A. Serandon, "Evaluating the Impact of French Employment Policies on Individual Labour Market Hystories," *Review of Economic Studies*, 1997, *64* (4), 683–713.

Cockx, B. and M. Picchio, "Are Short-Lived Jobs Stepping Stones to Long-Lasting Jobs?," *Oxford Bulletin of Economics and Statistics,*, 2011, forthcoming.

__ **and** __ , "Scarring Effects of Remaining Unemployed for Long-Term Unemployed School-Leavers," 2011. IZA Discussion Paper No. 5937, Bonn.

Doiron, D. and T. Gørgens, "State Dependence in Youth Labor Market Experiences, and the Evaluation of Policy Interventions," *Journal of Econometrics*, 2008, *145* (1), 81–97.

Frederiksen, A., B.E. Honoré, and L. Hu, "Discrete Time Duration Models with Group-Level Heterogeneity," *Journal of Econometrics*, 2007, *141* (2), 1014–1043.

Frijters, P., "The Non-Parametric Identification of Lagged Duration Dependence," *Economics Letters*, 2002, *75* (3), 289–292.

Guo, G. and G. Rodríguez, "Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, with an Application to Child Survival in Guatemala," *Journal of the American Statistical Association*, 1992, *87* (420), 969–976.

Heckman, J.J. and G.J. Borjas, "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence," *Economica*, 1980, *47* (187), 247–283.

Holt, J.D. and R.L. Prentice, "Survival Analyses in Twin Studies and Matched Pair Experiments," *Biometrika*, 1974, *61* (1), 17–30.

Honoré, B.E., "Identification Results for Duration Models with Multiple Spells," *Review of Economic Studies*, 1993, *60* (1), 241–246.

Hougaard, P., *Analysis of Multivariate Survival Data*, New York: Springer-Verlag, 2000.

Lindeboom, M. and M. Kerkhofs, "Multistate Models for Clustered Duration Data – An Application to Workplace Effects on Individual Sickness Absenteeism," *Review of Economics and Statistics*, 2000, *82* (4), 668–684.

Melino, A. and G.T. Sueyoshi, "A Simple Approach to the Identifiability of the Proportional Hazards Model," *Economics Letters*, 1990, *33* (1), 63–68.

Ridder, G. and İ Tunalı, "Stratified Partial Likelihood Estimation," *Journal of Econometrics*, 1999, *92* (2), 193–232.

Sastry, N., "A Nested Frailty Model for Survival Data, With an Application to the Study of Child Survival in Northeast Brazil," *Journal of the American Statistical Association*, 1997, *92* (438), 426–435.

Therneau, T.M. and S.A. Hamilton, "rhDNase as an Example of Recurrent Event Analysis," *Statistics in Medicine*, 1997, *16* (18), 2029–2047.

Institut de Recherches Économiques et Sociales
Université catholique de Louvain

Place Montesquieu, 3
1348 Louvain-la-Neuve, Belgique

ISSN 1379-244X D/2011/3082/037