Discussion Papers No. 600, December 2009
Statistics Norway, Research Department

Christian N. Brinch

Non-parametric identification of the mixed proportional hazards model with interval-censored durations

Abstract:
This note presents identification results for the mixed proportional hazards model when duration data are interval-censored. Earlier positive results on identification under intervalcensoring require both parametric specification on how covariates enter the hazard functions and assumptions of unbounded support for covariates. New results provided here show how one can dispense with both of these assumptions. The mixed proportional hazards model is non-parametrically identified with interval-censored duration data, provided covariates have support on an open set and the hazard function is a non-constant continuous function of covariates.

Keywords: duration analysis, interval-censoring, non-parametric identification

JEL classification: C41

Acknowledgement: Thanks to Rolf Aaberge and Magne Mogstad for helpful comments on this paper.

Address: Christian Brinch, Statistics Norway, Research Department and Center for Ecological and
Evolutionary Synthesis, Department of Biology, University of Oslo,
e-mail: cnb@ssb.no

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

The Mixed Proportional Hazards (MPH) model is the main workhorse in econometric duration analysis with a focus on separating heterogeneity from structural duration dependence. A large and growing literature has been concerned with identification of the MPH model under different assumptions, see e.g. van den Berg (2001) for a survey. This note describes identification results for MPH models when durations are not observed exactly, but are interval-censored. Existing identification results for the MPH model with interval-censored data require both parametric specification of how covariates enter the model and assumptions of unbounded support for regressors. I here demonstrate that it is possible to dispense with both assumptions: The MPH model is non-parametrically identified under interval-censoring provided covariates have support on an open set and the hazard function is a continuous non-constant function of the covariates.

Non-parametric identification is an important issue for duration models with unobserved heterogeneity. The inherent non-linearities in commonly applied models ensure parametric identication of structural duration dependence. In the absence of non-parametric identification, estimation results depend crucially on parametric specifications - which may often be ad hoc. With non-parametric identification results to fall back on, one can at least hope for results that do not depend crucially on parametric specifications, even if full non-parametric estimation is often not feasible, and parametric or semi-parametric estimators are applied.

Elbers and Ridder (1982) prove identification of the MPH model with minimal requirements on variation in covariates, under an assumption of finite mean for the heterogeneity distribution, while Heckman and Singer (1984) prove a similar result with alternative tail assumptions for the heterogeneity distribution. Ridder (1990) clarifies the differences within the GAFT class that generalizes the MPH model. Heckman and Honoré (1989) and Abbring and van den Berg (2003) generalize these results to dependent competing risks models.

The above results rely crucially on exact observation of durations. In practice, duration data should usually be considered interval-censored or discrete. That is, durations are not observed exactly, but only observed to lie within some interval, e.g. one observes spell lengths that are less than one month, between one and two months etc. The combination of continuous time hazard rate models and interval-censored duration data is common enough to have generated a voluminous literature. There are basically three approaches to estimation of models in this setting. The first is to simply assume away the interval-censoring in the sense that data are treated as if they were not censored. Not surprisingly, this may lead to problems, see e.g. Bergstrøm and Edin (1992) or Røed and Zhang (2002). The second approach is to derive the likelihood of the interval-censored observations from a continuous time model and use this likelihood as a basis for estimation. Flinn and Heckman (1982) give an early discussion of this. The third approach is to specify the model as a discrete duration model. A discrete duration model may or may not be consistent with a hazard rate model. For cases where the discrete time models are consistent with such underlying continuous time models, the second and third approaches are equivalent. Han and Hausman (1990) and Sueyoshi (1995) estimate discrete duration models that are consistent with hazard rate models, while e.g. van den Berg and van Ours (1994) estimate discrete duration models that are not consistent with hazard rate models, but on the other hand allow for simplification of some estimation procedures.

There are some identification results for MPH models with interval-censoring in the literature. Clearly, it is not possible to recover hazard function behavior within intervals (Sueyoshi, 1995). Ridder (1990) shows that the GAFT class is not identified under assumptions corresponding to the classical results for uncensored data, but that the model is identified in a corresponding way if covariates are assumed to enter the log structural hazard function linearly and covariates have support on the full real line. McCall (1994) shows that the model is still identified when the coefficients associated with the linear function of covariates are interval specific. Meyer (1995) contains an identification result for the MPH model similar to the positive result in Ridder (1990) and also comments that the result also holds in the more general case where the structural hazard function is a known function of the linear function of covariates. Bierens (2008) proves identification of the same model, while also relaxing the assumption on the support of covariates some-what and in addition providing alternative conditions on the heterogeneity distribution. All identification results for MPH models with interval-censored durations in the literature are semi-parametric, in that they require a known function of the structural hazard function to be linear in covariates. All results also rely on unbounded covariate support.

In the next section, I first show how the unbounded support assumption may be relaxed within the semi-parametric framework. Secondly, I show that full non-parametric identification can be achieved, regardless of the negative identification result in Ridder (1990).

2 Identification results

The MPH model describes the family of distributions of a positive random variable T, the duration, conditional on covariates $x \in \mathcal{X}$. Assuming continuous distribution functions for T, these are fully described through hazard functions. The MPH model is specified in terms of an independent random variable V with support on R_+, representing unobserved heterogeneity, and a hazard function, conditional on both covariates x and $V = v$ specified as $vf(t)g(x)$. The survival function of the MPH model, after integrating out V, follows as

$$G(t,x) = \mathrm{E}(\exp(-VF(t)g(x))) = \mathcal{L}(F(t)g(x)), \tag{1}$$

where E denotes expectation with respect to V, $F(t) = \int_0^t f(r)dr$, and $\mathcal{L}$ is the Laplace transform of the random variable V, see e.g. Feller (1971).

In addition, I will discuss the Generalized Accelerated Failure Time (GAFT) class introduced by Ridder (1990), a generalization of the MPH model. Define the GAFT class directly by

$$G(t,x) = L(F(t)g(x)), \tag{2}$$

where L is a continuously differentiable, strictly decreasing, positive function defined on $\mathbb{R}_+$ with $L(0) = 1$. L corresponds to $\mathcal{L}$ in equation (1), which satifies the restrictions on L. $\mathcal{L}$ has more properties. The essential extra property in our context is that $\mathcal{L}$ is analytic and hence uniquely determined by its values on an open set.

Here identification of the model will be studied under discrete or more precisely interval censored duration data. With interval-censoring, durations are not observed exactly, but only observed to fall within a certain interval. Equivalently, whether or not durations "have ended" is only observed at a finite number of points in time.

In the literature, identification under interval censoring has been studied in models with parametric functional form restrictions and unbounded support assumptions on covariates. Let us first see how we can dispense with the latter assumption.

Assumption 1 *The random variable V has finite mean, normalized to unity.*

Thus, $\mathcal{L}'(0) = -1$.

Assumption 2 *We impose the parametric restriction $g(x) = \exp(x\beta)$, with $\beta \neq 0$.*

I assume scalar covariates. It is straightforward to extend results to the case with vector valued covariates.

Assumption 3 *x takes on values on an open set $\mathcal{X} \subset \mathbb{R}$.*

Assumption 4 *G(t,x) is only known at $t = t_a$, with $G(t_a, x) < 1$ for some $x \in \mathcal{X}$.*

This corresponds to an observation plan where it is only observed whether durations have ended at one point in time.

A *structure* of the MPH model is a set $\{\mathcal{L}, f, g\}$ that conforms to the definitions above. We say that the MPH model is *identified* if the structure of the model is uniquely determined from the unconditional survival function. Under Assumptions 3 and 4, the starting point is what one can identify from $G(t, x) = \mathcal{L}(F(t)g(x))$ for $t = t_a$ and $x \in \mathcal{X}$. Clearly, it is then impossible to identify $F(t)$ for $t \neq t_a$. Use the notation $F_a = F(t_a)$. Under assumptions 2-4, a structure of the MPH model can now be represented by the set $\{\mathcal{L}, F_a, \beta\}$.

Theorem 1 *Under Assumptions 2, 3 and 4, observationally equivalent structures $\{\mathcal{L}_1, F_{a1}, \beta_1\}$ and $\{\mathcal{L}_2, F_{a2}, \beta_2\}$ of the MPH model must satisfy*

$$F_{a2} = AF_{a1}^{b}, \tag{3}$$

$$\beta_2 = \beta_1 b, \tag{4}$$

and

$$\mathcal{L}_2(As^b) = \mathcal{L}_1(s), \tag{5}$$

for positive constants A and b.

Under Assumptions 1, 2, 3 and 4, the MPH model is identified.

Proof. Assume that two structures $\{\mathcal{L}_1, F_{a1}, \beta_1\}$ and $\{\mathcal{L}_2, F_{a2}, \beta_2\}$ are observationally equivalent. That is,

$$\mathcal{L}_1(F_{a1}\exp(\beta_1 x)) = \mathcal{L}_2(F_{a2}\exp(\beta_2 x)), \text{ for all } x \in \mathcal{X}. \tag{6}$$

Equivalently

$$\beta_2 x + \log F_{a2} = h_1(\beta_1 x + \log F_{a1})), \text{ for all } x \in \mathcal{X}. \tag{7}$$

where $h_1 = \log \circ \mathcal{L}_2^{-1} \circ \mathcal{L}_1 \circ \exp$, where $\circ$ denotes composition of functions. Then h_1 must be a linear function for $x \in \mathcal{X}$. Let $h_1(z) = \log(A) + bz$, with two arbitrary constants $A > 0$ and $b > 0$. (h_1 is increasing, by the properties of the component functions.) Next, let $h_2 = \mathcal{L}_1^{-1} \circ \mathcal{L}_2$. Then

$$h_2(z) = Az^b. \tag{8}$$

Thus, for all s on some open set,

$$\mathcal{L}_2(As^b) = \mathcal{L}_1(s). \tag{9}$$

When this equation holds for all s on an open set, it holds for all $s > 0$ through the analyticity of Laplace transforms.

Substituting for $\mathcal{L}_1$ in equation (6), we find

$$\mathcal{L}_2(AF_{a1}^b \exp(b\beta_1 x)) = \mathcal{L}_2(F_{a2}\exp(\beta_2 x)), \text{ for all } x \in \mathcal{X}, \tag{10}$$

hence

$$AF_{a1}^b \exp(b\beta_1 x)) = F_{a2} \exp(\beta_2 x)), \tag{11}$$

leading to equations (3) and (4).

Differentiation of both sides of equation (9) with respect to s gives

$$\mathcal{L}_2'(As^b)Abs^{b-1} = \mathcal{L}_1'(s), s \in \mathbb{R}_+. \tag{12}$$

Under Assumption 1, both $\mathcal{L}_1'(s)$ and $\mathcal{L}_2'(As^b)$ are required to approach -1 as $s \to 0$, which again requires that $Abs^{b-1} \to 1$ as $s \to 0$, giving $A = b = 1$. ■

Theorem 1 is very similar to Theorem 2 in Ridder (1990) with an identification result for the corresponding GAFT class - where Assumption 1 is not invoked. The main difference from the first part of Theorem 1 is that analytical continuation can not be applied for the GAFT class and that Assumption 3 must therefore be strengthened such that x takes on values on $\mathbb{R}$. This is precisely the point with Theorem 1, to demonstrate that for the MPH model, the unbounded covariate support assumption is not necessary for identification. Theorem 1 contains the main identification result for interval-censored durations in Meyer (1995) and Theorem 5 in Bierens (2008) as special cases. These apply stronger conditions - either that x takes on values on $\mathbb{R}$ - or in the case of Bierens (2008), that $x\beta$ has no lower bound.

It follows from the discussion in Ridder (1990) that, if $\mathcal{L}_1$ is the unique structure conforming to Assumption 1, then the constant b characterizing observationally equivalent structures must be larger than one. Other values of b lead to $\mathcal{L}_2$ that do not conform to the requirements of Laplace transforms.

Assumption 1 should be seen as a necessary assumption for identification in the context of separating heterogenity and structural duration dependence. There are observationally equivalent structures without Assumption 1 that imply qualitatively different structural duration dependence. There are however alternative necessary assumptions. Bierens (2008) discusses two such alternative assumptions. As should be clear from the GAFT definition above, $1 - \mathcal{L}$ can be interpreted as a cumulative distribution function, say for a random variable Y, with support on $\mathbb{R}_+$. Bierens (2008) considers the distribution of $Z = \exp(-Y)$, which has support on the unit interval. The first of the alternative identification conditions in Bierens (2008) is to pre-specify two quantiles of Z or equivalently of Y. Clearly, setting two quantiles is sufficient for pinning down the constants A and b above. The second alternative identification condition is to pre-specify the first two moments of Z - which again suffices for determining A and b. Thus, the alternative conditions in Bierens (2008) can be substituted for Assumption 1 in Theorem 1, although Assumption 3 is weaker than the corresponding assumption in Theorems 6 and 7 in Bierens (2008). Similarly, identification based on Theorem 1 could use the tail assumptions from Heckman and Singer (1984) in place of Assumption 1. These different ways of achieving identification pins down different combinations of the constants A and b and potentially leads to qualitatively different structural duration dependence.

At first glance, one can hardly claim to be identifying structural duration dependence through Theorem 1, as the integrated structural hazard rate is only identified at one point. However, identification of structural duration dependence is trivial when the heterogeneity distribution is identified:

Corollary 2 *Given Assumptions 1-4, if G(t,x) is also observed for some $t_b > t_a$, then the integrated structural hazard function is also identified over the interval from t_a to t_b without Assumption 2 or indeed any proportional hazards structure in the period beyond t_a.*

Proof.

$$G(t_b, x) = \mathcal{L}(F_a \exp(x\beta) + \Lambda(t_a, t_b, x)), \tag{13}$$

where the integrated structural hazard function $\Lambda(t_a, t_b, x) = \int_{t_a}^{t_b} \lambda(s, x)ds$, with $\lambda(t, x)$ a general structural hazard function (corresponding to $f(t)g(x)$ in the MPH setup) specified as a function of elapsed duration t and covariates x. Straightforwardly,

$$\Lambda(t_a, t_b, x) = \mathcal{L}^{-1}(G(t_b, x)) - F_a \exp(x\beta). \tag{14}$$

The unknown functions on the right hand side are identified through Theorem 1. ■

McCall (1994) studies a model where the survival probability is observed at more than one point of time and where the function $g(x)$ may differ over intervals, while retaing the exponential structure from Assumption 2. Corollary 2 generalizes the result in McCall (1994).

It is now clear that the parametric restriction in Assumption 2 suffices for identification of the MPH model without assuming unbounded support of covariates. Let us now see where we get without imposing parametric restrictions.

Assumption 5 *g is a continuous, non-constant function of x.*

Assumption 5 is strictly weaker than Assumption 2. It follows from Assumptions 3 and 5 that $g(x)$ takes on values on an open set.

Ridder (1990) contains a demonstration that even unbounded support is not sufficient for identification without parametric restrictions in the GAFT class. Since this non-identification result may not hold in the specialized MPH model, we provide the simple theorem below.

Theorem 3 *The MPH model is not identified under Assumptions 1, 3, 4 and 5, in fact there exists an observationally equivalent structure for every heterogeneity distribution.*

Proof. Normalize F_a to one. (Any other value can be captured by g.) Let $\{\mathcal{L}, g\}$ denote a structure of the model. Let $\mathcal{L}_0$ denote the Laplace transform of an arbitrary distribution function with support on $\mathbb{R}_+$.

$$G(t,x) = \mathcal{L}(g(x)) = \mathcal{L}_0(\mathcal{L}_0^{-1}(\mathcal{L}(g(x)))) \tag{15}$$

The observationally equivalent structure can now be specified as $\{\mathcal{L}_0, g_0\}$ with $g_0 = \mathcal{L}_0^{-1} \circ \mathcal{L} \circ g$. ■

In view of the negative result on non-parametric identification in Ridder (1990) for the GAFT class, and the straightforward extension of this result to the MPH model in Theorem 3, it is not surprising that positive identification results for this case has not been searched for. These negative results do however depend critically on the extreme interval-censoring as implemented in Assumption 4. To get positive results, we will instead use the alternative

Assumption 6 *G(t,x) is known at t_a and $t_b > t_a$, with $G(t_a, x) < 1$ and $G(t_b, x) < G(t_a, x)$ for some $x \in \mathcal{X}$.*

Thus, whether durations have ended is observed at two points in time, and some durations end between these points of time.

The following key result shows that neither the parametric assumptions nor the unbounded support assumptions are necessary for identification.

Theorem 4 *The MPH model is identified under Assumptions 1, 3, 5 and 6.*

First note that, if we strengthen Assumption 5 to ensure that $g(x)$ contains points arbitrarily close to zero, we can directly apply the proof of the classical result in Elbers and Ridder (1982). They prove the sufficiency of two different values of $g(x)$ and t that varies such that $F(t) \to 0$. Here, we have two different values of F(t) and x that varies such that $g(x) \to 0$. The proof in Elbers and Ridder (1982) can however straightforwardly be extended to prove Theorem 4. I provide the necessary extension here.

Proof. Specify two equivalent structures by $\{\mathcal{L}_1, g_1, F_1\}$ and $\{\mathcal{L}_2, g_2, F_2\}$, where

$$\mathcal{L}_1(g_1(x)) = \mathcal{L}_2(g_2(x)), \text{ for all } x \in \mathcal{X} \tag{16}$$

specify survival to t_b and survival to t_a is specified as

$$\mathcal{L}_1(F_1 g_1(x)) = \mathcal{L}_2(F_2 g_2(x)), \text{ for all } x \in \mathcal{X}. \tag{17}$$

Thus $F_1 < 1$ and $F_2 < 1$, by Assumption 6.

Let $z(w) = \mathcal{L}_2^{-1}(\mathcal{L}_1(w))$. Now, $g_2(x) = z(g_1(x))$ and

$$z(F_1 g_1(x)) = F_2 z(g_1(x)), x \in \mathcal{X}. \tag{18}$$

For this functional equation to hold for all $g_1(x)$ on an open set, it must, due to the analyticity of z, also hold for all $g_1(x) \in \mathbb{R}_+$.

It is shown in Elbers and Ridder (1982), page 409, that the only solution to this functional equation when $\mathcal{L}_1$ and $\mathcal{L}_2$ are Laplace transforms of finite (and normalized) mean distribution functions is that z is the identity function - and that the observationally equivalent structures must therefore be identical. ■

It is pointed out in Meyer (1995) that the proof of his main identification result (which is covered by Theorem 1 above) also applies beyond the parametric specification in Assumption 2 above. Specifically, in the notation applied here, it is required that g is a known, strictly monotone, continuously differentiable function in a linear function of x. Clearly, Theorem 4 goes beyond this result from Meyer (1995), as it is here not required that the function g is known.

3 Discussion

The results provided here close the gap between identification results for the Mixed Proportional Hazards model with exact and interval censored duration data. The model is non-parametrically identified under interval-censoring, assuming the structural hazard function is a continuous non-constant function of covariates with support on an open set.

It is clearly not possible to straightforwardly extend the results to the case with covariates with finite support. The combination of interval censored duration and covariates with finite support gives us only a finite number of cell probabilities as empirical predictions - hardly enough for full identification of infinite-dimensional models. Still, sets of observationally equivalent models may be sufficiently similar for identification in the intuitive sense to hold in practice. See Bierens (2008) or Honoré and Lleras-Muney (2006) for related discussions.

The identification results provided here do not generalize directly to the case with dependent competing risks. Dependent competing risks model with interval-censoring are difficult to work with. State dependent integrated structural hazard functions are not even directly identifiable when the unobserved heterogeneity distribution is known. Within interval behavior of transition rates to one state may affect the population at risk for transitions to other states. Honoré and Lleras-Muney (2006) show how bounds may still be achieved on interesting parameters in a closely related model.

The identification results provided here rely crucially on the parametric hazards assumption and the finite mean assumption on the heterogeneity distribution. Brinch (2008) provide results that show we can dispense with these assumption if covariates vary over time as well as across observations, corresponding to results in Brinch (2007) for models without interval-censoring.

References

[1] Abbring, J. H. and G. J. van den Berg (2003), “The identifiability of the mixed proportional hazards competing risks model”, Journal of the Royal Statistical Society Series B, 65(3), 701-710.

[2] Bergström, R. and P. A. Edin (1992), “Time aggregation and the distribution shape of unemployment duration”, Journal of Applied Econometrics, 7(1), 5-30.

[3] Bierens, H. (2008), “Semi-Nonparametric Interval-Censored Mixed Proportional Hazard Models: Identification and Consistency Results”, Econometric Theory, 24, 749-794.

[4] Brinch, C. N. (2007), “Nonparametric Identification of the Mixed Hazards Model with Time-varying Covariates”, Econometric Theory, 23, 349-354.

[5] Brinch, C. N. (2008), “Nonparametric Identification of the Mixed Hazards Model with Interval-Censored Durations”, Statistics Norway Discussion Paper 539.

[6] Elbers, C. and G. Ridder (1982), “True and Spurious Duration Dependence: The Identifiablity of the Proportional Hazards Model”, Review of Economic Studies, 49, 403-409.

[7] Feller, W. (1971), An Introduction to Probability Theory and Its Applications, Vol. II, New York: John Wiley.

[8] Flinn, C. and Heckman, J. J. (1982), "Models for the analysis of labor force dynamics", Advances in Econometrics, vol. 1, eds. R. Bassman and G. Rhodes. Greenwich, Conn.: JAI Press, 35-95.

[9] Han, A. and J. A. Hausman (1990), "Flexible parametric estimation of duration and competing risk models", Journal of Applied Econometrics, 5(1), 1-28.

[10] Heckman, J. J. and B. Honoré (1989), "The identifiability of the competing risks models", Biometrika, 76(2), 325-330.

[11] Heckman, J. J. and B. Singer (1984), "The Identifiability of the Proportional Hazard Model", Review of Economic Studies, 51(2), 231-241.

[12] Honoré, B. E. and A. Lleras-Muney (2006), "Bounds in Competing Risks Models and the War on Cancer", Econometrica, 74(6), 1675-1698.

[13] McCall, B. P. (1994), "Testing the Proportional Hazards Assumption in the Presence of Unmeasured Heterogeneity", Journal of Applied Econometrics, 9, 321-334.

[14] Meyer, B. D. (1995), "Semiparametric Estimation of Hazard Models", Unpublished manuscript, Northwestern University.

[15] Ridder, G. (1990), "The Non-Parametric Identification of Generalized Accelerated Failure Time Models", Review of Economic Studies, 57, 167-181.

[16] Røed, K. and T. Zhang (2002), "A Note on the Weibull Distribution and Time Aggregation Bias", Applied Economics Letters, 9(7), 469-472.

[17] Sueyoshi, G. T. (1995), "A Class of Binary Response Models for Grouped Duration Data", Journal of Applied Econometrics, 10, 411-431.

[18] van den Berg, G. J. (2001), "Duration models: specification, identification, and multiple durations", Handbook of Econometrics, Vol. 5, Amsterdam: North-Holland.

[19] van den Berg, G. J. and J. C. van Ours (1994), "Unemployment Dynamics and Duration Dependence in France, the Netherlands and the United Kingdom", The Economic Journal, 104, 432-443.