The Proportional Hazard Model in Economics

JERRY A. HAUSMAN AND TIEMEN M. WOUTERSEN[*][†]
MIT AND JOHNS HOPKINS UNIVERSITY

February 2005

ABSTRACT. This paper reviews proportional hazard models and how the thinking about identification and estimation of these models has evolved in the last 30 years.

KEYWORDS: Duration Model, Hazard Model, Mixed Proportional Hazard Model, Time-varying regressors, Heterogeneity.

THE ESTIMATION OF DURATION MODELS has been the subject of significant research in econometrics since the late 1970s. Cox (1972) proposed the use of Proportional Hazard models in biostatistics and they were soon adopted for use in economics. Since Lancaster (1979), it has been recognized among economists that it is important to account for unobserved heterogeneity in models for duration data. Failure to account for unobserved heterogeneity causes the estimated hazard rate to decrease more with the duration than the hazard rate of a randomly selected member of the population. Moreover, the estimated proportional effect of explanatory variables on the population hazard rate is smaller in absolute value than that on the hazard rate of the average population member and decreases with the duration. To account for unobserved heterogeneity Lancaster proposed a parametric Mixed Proportional Hazard (MPH) model, a partial generalization of Cox's Proportional Hazard model, that specifies the hazard rate as the product of a regression function that captures the effect of observed explanatory variables, a base-line hazard that captures variation in the hazard over the spell, and a random variable that accounts for the omitted heterogeneity. In particular, Lancaster (1979) introduced the

[*] Comments are welcome, jhausman@mit.edu and woutersen@jhu.edu.

[†] We thank Su-Hsin Chang for helpful comments.

mixed proportional hazard model in which the hazard is a function of a regressor X, unobserved heterogeneity v, and a function of time $\lambda(t)$,

$$\theta(t \mid X, v) = ve^{X\beta_0}\lambda(t). \tag{1}$$

The function $\lambda(t)$ is often referred to as the baseline hazard and $v|X$ has a gamma distribution. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard $\theta(t \mid X)$ to be decreasing with time. In particular, estimating the mixed proportional hazard model gives the relative importance of the heterogeneity, v, and genuine duration dependence, $\lambda(t)$, see Lancaster (1990) and Van den Berg (2001) for overviews. Lancaster (1979) uses functional form assumptions on $\lambda(t)$, which were not required by the Cox model, and distributional assumptions on v to identify the model. Examples by Lancaster and Nickell (1980) and Heckman and Singer (1984), however, show the sensitivity to these functional form and distributional assumptions. Thus, Lancaster's MPH model is fully parametric and from the outset questions were raised on the role of functional form and parametric assumptions in the distinction between unobserved heterogeneity and duration dependence[1]. This question was resolved by Elbers and Ridder (1982) who showed that the MPH model is semi-parametrically identified if there is minimal variation in the regression function. A single indicator variable in the regression function suffices to recover the regression function, the base-line hazard, and the distribution of the unobserved component, provided that this distribution does not depend on the explanatory variables. Semi-parametric identification means that semi-parametric estimation is feasible, and a number of semi-parametric estimators for the MPH model have been proposed that progressively relaxed the parametric restrictions.

Nielsen et al. (1992) showed that the Partial Likelihood estimator of Cox (1972) can be generalized to the MPH model with Gamma distributed unobserved heterogeneity. Their estimator is semi-parametric because it uses parametric specifications of the regression function and the distribution of the unobserved heterogeneity. The estimator requires numerical integration of the order of the sample size as originally discussed by Han and

[1] Heckman (1991) gives an overview of attempts to make this distinction in duration and dynamic panel data models.

Hausman (1990), which further limits its usefulness and makes it impractical for most situation in econometrics. Heckman and Singer (1984) considered the non-parametric maximum likelihood estimator of the MPH model with a parametric baseline hazard and regression function. Using results of Kiefer and Wolfowitz (1956), they approximate the unobserved heterogeneity with a discrete mixture. The rate of convergence and the asymptotic distribution of this estimator are not known. As a result, these estimators that use discrete mixture with an increasing number of support points cannot be used to test hypotheses. Another estimator that does not require the specification of the unobserved heterogeneity distribution was suggested by Honoré (1990). This estimator assumes a Weibull baseline hazard and only uses very short durations to estimate the Weibull parameter.

Han and Hausman (1990) and Meyer (1990) proposes an estimator that assumes that the baseline hazard is piecewise-constant, to permit flexibility, and that the heterogeneity has a gamma distribution. Both papers find that the hazard rate, conditional on heterogeneity, is non-monotonic so that the Weibull model cannot hold. Hausman and Woutersen (2005) present simulations and a theoretical result that show that using a nonparametric estimator of the baseline hazard with gamma heterogeneity yields inconsistent estimates for all parameters and functions if the true mixing distribution is not a gamma, which limits the usefulness of the Han-Hausman-Meyer approach. Thus, Hausman and Woutersen (2005) find it important to specify a model that does not require a parametric specification of the unobserved heterogeneity.

Horowitz (1999) was the first to propose an estimator that estimates both the baseline hazard and the distribution of the unobserved heterogeneity nonparametrically. His estimator is an adaptation of the semi-parametric estimator for a transformation model that he introduced in Horowitz (1996). In particular, if the regressors are constant over the duration then the MPH model has a transformation model representation with the logarithm of the integrated baseline hazard as the dependent variable and a random error that is equal to the logarithm of a log standard exponential minus the logarithm of a positive random variable. In the transformation model the regression coefficients are identified only up to scale. As shown by Ridder (1990) the scale parameter is identified in the MPH model if the unobserved heterogeneity has a finite mean. Horowitz (1999) suggests an estimator of the scale parameter that is similar to Honoré's (1990) estimator of the Weibull parameter and is consistent if the finite mean assumption holds so that his approach allows estimation of the regression coefficients (not just up to scale). However, the Horowitz approach only permits estimation of the regression coefficients at a slow rate of convergence and it is not $N^{-1/2}$ consistent, where N is the sample size. The reason for the slower than $N^{-1/2}$ convergence is that the information matrix of the MPH model is singular under Horowitz assumptions[2]. In particular, Horowitz (1999) assumes that the first three moments of the heterogeneity distribution exist and Ishwaran (1996b) shows that the fastest possible rate of convergence is $N^{-2/5}$ for that case and Horowitz' (1999) estimator converges arbitrarily close to that rate. In other words, the slow rate of convergence is implied by the assumptions and is not a peculiarity of the estimator.

Subsequent research has focused on strengthening the assumptions of the MPH model so that $N^{-1/2}$ convergence is possible. Ridder and Woutersen (2003) derive a $N^{-1/2}$ consistent estimator for the MPH model by assuming that the baseline hazard rate is constant over a small interval, $\lambda(t) = \lambda$ for $0 \leq t \leq \varepsilon$ for any $\varepsilon > 0$ while allowing for a nonparametric baseline hazard function for $t > \varepsilon$. For parametric baseline hazards, Ridder and Woutersen (2003) assume that $\underset{t\downarrow 0}{lim}\, \lambda(t) = \lambda$ for $0 < \lambda < \infty$ and derive another $N^{-1/2}$ consistent estimator. Hausman and Woutersen (2005) derive an estimator for the mixed proportional hazard model (with heterogeneity) that allows for a nonparametric baseline hazard and uses time-varying regressors. No parametric specification of the heterogeneity distribution nor nonparametric estimation of the heterogeneity distribution is necessary. Intuitively, Hausman and Woutersen (2005) condition out the heterogeneity distribution, which makes it unnecessary to estimate it. Thus, Hausman and Woutersen (2005) eliminate the problems that arise with the Lancaster (1979) approach to MPH models. In this model the baseline hazard rate is nonparametric and the estimator of the integrated baseline hazard rate converges at the regular rate, $N^{-1/2}$, where N is the sample size. This convergence rate is the same rate as for a duration model without

[2] See Hahn (1994) and Ishwaran (1996a).

heterogeneity. The regressor parameters also converge at the regular rate. A nice feature of the estimator is that it allows the durations to be measured on a finite set of points. Such discrete measurement of durations is important in economics; for example, unemployment is often measured in weeks. In the case of discrete duration measurements, the estimator of the integrated baseline hazard only converges at this set of points, as would be expected.

It may be argued that the bias in the estimates of the regression coefficients is small, if the estimates of the MPH model indicate that there is no significant unobserved heterogeneity. The problem with this argument is that estimates of the heterogeneity distribution are usually not very accurate. Given the results in Horowitz (1999) this finding should not come as a surprise. The simulation results in Baker and Melino (2000) show that it is empirically difficult to find evidence of unobserved heterogeneity, in particular if one chooses a flexible parametric representation of the baseline hazard. However, Han and Hausman (1990) and applications of their approach have found significant heterogeneity using a flexible approach to the baseline hazard. Bijwaard and Ridder (2002) find that the bias in the regression parameters is largely independent of the specification of the baseline hazard. Hence, failure to find significant unobserved heterogeneity should not lead to the conclusion that the bias due to correlation of the regressors and the unobservables that affect the hazard is small.

Because it is empirically difficult to recover the distribution of the unobserved heterogeneity, estimators that rely on estimation of this distribution may be unreliable. Therefore, it may be advisable to avoid estimating the unobserved heterogeneity distribution and the remainder of the MPH model simultaneously. Nevertheless, after estimating the baseline hazard and regression function, one can usually identify the mixing distribution. In particular, Horowitz (1999) uses the following equation to estimate the mixing distribution,

$$\ln\{\Lambda(T)\} + X\beta - \ln(Z) = -\ln(v)$$

where $\Lambda(T)$ and β can be estimated and the unobserved Z has an exponential distribution with mean one. Thus, Horowitz (1999) solves a deconvolution problem and the speed of convergence depends on the assumptions on the distribution of v.

A hazard model is a natural framework for time-varying regressors if a flow or a transition probability depends on a regressor that changes with time since a hazard model avoids the curse of dimensionality that would arise from interacting the regressors at each point in time with one another. A nonconstructive identification proof[3] for the duration model with time-varying regressors can be produced using techniques similar to Honoré (1993b) and Honoré (1993a) gives such a proof. In particular, Honoré (1993a) does not assume that the mean of the heterogeneity distribution is finite[4]. Ridder and Woutersen (2003) argue that it is precisely the finite mean assumption that makes the identification of Elbers and Ridder (1982) 'weak' in the sense that the model of Elbers and Ridder (1982) cannot be estimated at rate $N^{-1/2}$. As in Honoré (1993a), Hausman and Woutersen (2005) do not need the finite mean assumption which gives an intuitive explanation why Hausman and Woutersen (2005) can estimate the model at rate $N^{-1/2}$.

References

[1] Baker, M. and A. Melino (2000): "Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study," *Journal of Econometrics*, 96, 357-93.

[2] Bijwaard, G. and G. Ridder (2002): "Efficient Estimation of the Semi-parametric Mixed Proportional Hazard Model", in preparation.

[3] Cox, D. R. (1972): "Regression models and life tables (with discussion)", *Journal of the Royal Statistical Society* B, 34: 187-220.

[4] Elbers, C. and G. Ridder (1982): "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model," *Review of Economic Studies*, 49, 402-409.

[5] Hahn, J. (1994): "The Efficiency Bound of the Mixed Proportional Hazard Model, " *Review of Economic Studies*, 61, 607-629.

[6] Han, A. K. and J. A. Hausman (1990): "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics.*

[3] A nonconstructive identification proof is an identification proof that does not suggest an estimator.

[4] nor does Honoré (1993a) assume a tail condition as in Heckman and Singer (1985).

[7] Hausman, J. A. and T. M. Woutersen (2005): “Estimating a Semi-Parametric Duration Model without Specifying Heterogeneity”, UCL CenMap working paper.

[8] Heckman, J. J. (1991): “Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity,” *American Economic Review*, 81, 75-79.

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

[10] Honoré, B. E. (1990): “Simple Estimation of a Duration Model with Unobserved Heterogeneity,” *Econometrica*, 58, 453-473.

[11] Honoré, B. E. (1993a): “Identification Results for Duration Models with Multiple Spells or Time-Varying Regressors,” Northwestern working paper.

[12] Honoré, B. E. (1993b): “Identification Results for Duration Models with Multiple Spells,” *Review of Economic Studies*, 60, 241-246.

[13] Horowitz, J. L. (1996): “Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable,” *Econometrica*, 64, 103-107.

[14] Horowitz, J. L. (1999): “Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity” *Econometrica*, 67, 1001-1028.

[15] Ishwaran, H. (1996a): “Identifiability and Rates of Estimation for Scale Parameters in Location Mixture Models,” *The Annals of Statistics*, 24, 1560-1571.

[16] Ishwaran, H. (1996b): “Uniform Rates of Estimation in the Semiparametric Weibull Mixture Model,” *The Annals of Statistics*, 24, 1572-1585.

[17] Kiefer, J. and J. Wolfowitz (1956): "Consistency of Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", *Annals of Mathematical Statistics,* 27, 887-906.

[18] Lancaster, T. (1979): “Econometric Methods for the Duration of Unemployment,” *Econometrica*, 47, 939-956.

[19] Lancaster, T. (1990): *The Econometric Analysis of Transition Data.* Cambridge: Cambridge University Press.

[20] Lancaster, T. and S. J. Nickell, (1980): "The Analysis of Re-employment Probabilities for the Unemployed", *Journal of the Royal Statistical Society*, A, 143, 141-165.

[21] Meyer, B. D. (1990): "Unemployment Insurance and Unemployment Spells," *Econometrica*, 58, 757-782.

[22] Nielsen, G.G., Gill, R.D., Andersen, P.K. & Sørensen, T.I.A. (1992): "A counting-process approach to maximum likelihood estimation in frailty models", *Scandanavian Journal of Statistics.* 19, 25-43

[23] Ridder, G. (1990): "The Non-Parametric Identification of Generalized Accelerated Failure Time Models, *Review of Economic Studies*, 57, 167-182.

[24] Ridder, G. and T. M. Woutersen (2003): "The Singularity of the Information Matrix of the Mixed Proportional Hazard Model" *Econometrica*, 71, 1579-1589.

[25] Van den Berg, G. J. (2001): "Duration Models: Specification, Identification, and Multiple Duration," in *Handbook of Econometrics,* Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: North-Holland.