

Testing for Heterogeneity in Duration Models: A Semiparametric Test

TIEMEN WOUTERSEN[†]
JOHNS HOPKINS UNIVERSITY

April 2007

ABSTRACT. This paper presents a new test for heterogeneity in duration models. The test allows for a nonparametric baseline hazard and does not impose any distributional assumptions on the unobserved heterogeneity. Moreover, test allows the durations to be censored.

KEYWORDS: Mixed Proportional Hazard Model, Heterogeneity.

1. INTRODUCTION

THE ESTIMATION OF DURATION MODELS has been the subject of significant research in econometrics since the late 1970s. Since Lancaster (1979), it has been recognized 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. To account for unobserved heterogeneity Lancaster proposed a parametric Mixed Proportional Hazard (MPH) model, a generalization of Cox's (1972) 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.

Lancaster's MPH model was fully parametric, as opposed to Cox's semi-parametric approach, and from the outset questions were raised on the role of functional form and

*This paper was presented at a conference in honour of Tony Lancaster, April 2007, Brown University.

[†]Comments are welcome and woutersen@jhu.edu.

parametric assumptions in the distinction between unobserved heterogeneity and duration dependence¹. 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.

Hausman (1990) and Meyer (1990) proposes an estimator that assumes that the base-line hazard is piecewise-constant, to permit flexibility, and that the heterogeneity has a gamma distribution. Hausman and Woutersen (2006) 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 (2006) find it important to specify a model that does not require a parametric specification of the unobserved heterogeneity. In applied work, one may wonder whether to estimate Cox's (1972) model or whether to allow for unobserved heterogeneity, as Lancaster (1979) does. It may be noted that some popular test may not even work for parametric heterogeneity. In particular we discuss how one cannot use the likelihood ratio, Lagrange multiplier or Wald test to distinguish one unobserved type from two unobserved types. Baker and Melino (2000) show that trying to estimate the heterogeneity and the duration dependence simultaneously is problematic if one wants to detect heterogeneity. Our test avoids the problems of several unobserved types and also avoids estimating the duration dependence.

In this paper, we propose a new test to test for unobserved heterogeneity in the mixed proportional hazard model of Lancaster (1979). This paper is organized as follows. Section 2 discusses the mixed proportional hazard model (with heterogeneity) and presents our

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

test. Section 3 gives simulation results and section 4 concludes.

2. MIXED PROPORTIONAL HAZARD MODEL

2.1 The semi-parametric MPH model

Lancaster (1979) introduced the mixed proportional hazard model in which the hazard is a function of a regressor X , unobserved heterogeneity v , and a function of time $\lambda(t)$,

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

The function $\lambda(t)$ is often referred to as the baseline hazard. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard $\theta(t | X)$ 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 $\lambda(t)$ 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. Our test avoids such functional form and distributional assumptions. In particular, our test relies on the empirical distribution function and the inverse of the empirical distribution function.

The mixed proportional hazard model of equation (1) implies the following survival probabilities,

$$\begin{aligned} P(T \geq t|x, v) &= \bar{F}(t|x, v) = \exp(-v \int_0^t e^{x\beta}\lambda(s)ds) \text{ and} \\ \bar{F}(t|x) &= P(T \geq t|x) = E_v\{\bar{F}(t|x(t), v)\} = E_v\{\exp(-v \int_0^t e^{x\beta}\lambda(s)ds)\}. \end{aligned} \tag{2}$$

The unconditional (on v) integrated hazard at the population values of the parameters is defined as

$$R = \Lambda(T, \alpha)e^{\beta'X} \tag{3}$$

with $\Lambda(t, \alpha) = \int_0^t \lambda(s, \alpha)ds$. Ridder and Woutersen (2003) we show that

$$R \stackrel{d}{=} \frac{W}{e^U} \tag{4}$$

with W a standard exponential random variable that is independent of U, X and $\stackrel{d}{=}$ means that the random variables on both sides have the same distribution.

2.2 Semi-parametric identification

Elbers and Ridder (1982) show that this MPH model is semi-parametrically identified if the following assumptions hold.

- (A1) $\Lambda(t_0, \alpha_0) = 1$ for some $t_0 > 0$, and $\Lambda(\infty, \alpha_0) = \infty$.
- (A2) $E(e^U) < \infty$.
- (A3) There are x_1, x_2 in the support of X with $\beta'_0 x_1 \neq \beta'_0 x_2$ and there is no constant in X ; U and X are independent.
- (A4) If $\lambda(t, \alpha_0) = \lambda(t, \tilde{\alpha}_0)$ for all $t > 0$, then $\alpha_0 = \tilde{\alpha}_0$, and if $\beta'_0 x = \tilde{\beta}'_0 x$ for all x in the support of X , then $\beta_0 = \tilde{\beta}_0$.

The first part of assumption A1 and the absence of a constant in X are normalizations. Assumption A4 ensures parametric identification of α_0, β_0 .

Ridder and Woutersen (2003) propose an alternative for assumption A2.

- (A2*) $0 < \lim_{t \downarrow 0} \lambda(t, \alpha_0) = \lambda(0, \alpha_0) < \infty$.

For the remainder of the paper assume that the mixed proportional hazard model is identified. That is, we assume that either assumptions A1-A4 holds or that A1, A2*, A3 and A4 hold. Ridder and Woutersen (2003) discuss that, although both sets of conditions ensure that the semi-parametric MPH model is identified, they have different implications for the information bound of this model. In particular, with the finite mean assumption the information matrix can be singular, while with assumption A2* this cannot be the case.

Examples of parametric models where assumption A2* holds for all parameter values are the Gompertz baseline hazard, the rational log specification (Lancaster (1990)), and the normal hazard. Examples of models in which assumption A2* is a parametric restriction are the piecewise-constant baseline hazard and the Box-Cox baseline hazard. Finally, the lognormal hazard does not satisfy A2* for all parameter values.

2.2 Test for unobserved heterogeneity

In applied work, one may wonder whether to estimate Cox’s (1972) model or whether to allow for unobserved heterogeneity, as Lancaster (1979) does and Lancaster (1990) and Van der Berg (2001) discuss. It may be noted that some popular tests may not even work for parametric heterogeneity. In particular, one cannot use the likelihood ratio, Lagrange multiplier or Wald test to distinguish one unobserved type from two unobserved types. That is, one cannot use these test if

$$\begin{aligned}
 H_0 & : P(v = v_1) = 1 \text{ for some } v_1 \\
 H_1 & : P(v = v_1) = p, P(v = v_2) = 1 - p \text{ for some } v_1, v_2.
 \end{aligned}$$

The reason that one cannot use these test is v_2 is not defined under the null hypothesis². Another drawback is that H_1 is rather restrictive and does not capture the notion of ‘unobserved heterogeneity’. Baker and Melino (2000) show that trying to estimate the heterogeneity and $\Lambda(t)$, the duration dependence, simultaneously is problematic if one wants to detect heterogeneity.. Our test avoids the problems of several unobserved types and also avoids estimating $\Lambda(t)$, the duration dependence.

Suppose one can estimate the cumulative distribution function of t at two points. For example, using nonparametric local averaging or by reducing the dimension of the regressors using the maximum rank correlation estimator of Han (1987). In particular suppose that the regressor x has two values, $x \in \{0, 1\}$ and let $\phi = e^{\beta_0}$. We use the following notation. Let $\bar{F}_0(t) = \bar{F}(t|x = 0)$, $\bar{F}_1(t) = \bar{F}(t|x = 1)$ and let $\bar{F}_0^{-1}(q)$ denote the inverse of $\bar{F}_0(t)$. Moreover, let $G_v(s) = E_v e^{-vs}$. Define $M(\kappa_1, \kappa_2)$ as follows,

$$M(\kappa_1, \kappa_2) = \frac{\ln[\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\}]}{\ln[\bar{F}_1\{\bar{F}_0^{-1}(\kappa_2)\}]}$$

Note that $\bar{F}_0(t) = E_v e^{-v\Lambda(t)} = G_v\{\Lambda(t)\}$ and that $\bar{F}_0^{-1}(\kappa_1) = \Lambda^{-1}(G_v^{-1}(\kappa_1))$. Moreover,

$$\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\} = G_v(\phi G_v^{-1}(\kappa_1)).$$

²In particular, the likelihood ratio test relies on a well behaved estimator for v_2 , which does not exist under H_0 . The Wald test relies on the difference of the well behaved estimate for v_2 and the value of v_2 under H_0 , and neither of these exist. Finally, the Lagrange multiplier relies on v_2 under H_0 , which does not exist.

Note that $\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\}$ depends on the heterogeneity distribution³ but that $\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\}$ does *not* depend on the nonparametric baseline hazard rate. Therefore, the transformation $\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\}$ implies that we only need to compare exponentials to a mixture of exponentials. This is not difficult. In particular, consider $M(\kappa_1, \kappa_2)$ if the data generating process has no heterogeneity, i.e. $P(v = v^*) = 1$ for some v^* . Then⁴

$$M(\kappa_1, \kappa_2) = \frac{\ln[\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\}]}{\ln[\bar{F}_1\{\bar{F}_0^{-1}(\kappa_2)\}]} = \frac{\ln(\kappa_1)}{\ln(\kappa_2)}.$$

This suggests to estimate $M(\kappa_1, \kappa_2)$ and use the null hypothesis of no heterogeneity, $M(\kappa_1, \kappa_2) = \frac{\ln(\kappa_1)}{\ln(\kappa_2)}$. In particular, define

$$\hat{M}(\kappa_1, \kappa_2) = \frac{\ln[\hat{S}_1\{\hat{S}_0^{-1}(\kappa_1)\}]}{\ln[\hat{S}_1\{\hat{S}_0^{-1}(\kappa_2)\}]}$$

where $\hat{S}_1(t)$ is the empirical survivor function, $\hat{S}_1(t) = \frac{1}{N} \sum_i 1(T_i \geq t)$ and $\hat{S}_0^{-1}(q)$ is the inverse of the empirical survivor function, $\hat{S}_0^{-1}(q)$ is the largest value of t for which $\hat{S}_1(t) \leq q$. The properties of $\hat{S}_1(t)$ and $\hat{S}_0^{-1}(q)$ follow immediately from the properties of the empirical distribution function and its inverse, see Athey and Imbens (2006). In particular, joint normality of $\hat{S}_1(t) - S_1(t)$ and $\hat{S}_0^{-1}(q) - S_0^{-1}(q)$ imply that $\hat{M}(\kappa_1, \kappa_2) - M(\kappa_1, \kappa_2)$ is also normally distributed around zero. The regularity conditions for the bootstrap, see Horowitz (2001) are satisfied and we bootstrap the test statistic $\hat{M}(\kappa_1, \kappa_2)$.

Consider the following example.

Example: Let

$$\theta(t \mid X, v) = v\phi^x \lambda(t) \tag{5}$$

where $x \in \{0, 1\}$ and $v|x \sim \text{Gamma}(\gamma, \delta)$.

Then

$$\bar{F}(t|x) = \frac{1}{(1 + \frac{\phi^x \Lambda(t)}{\delta})^\gamma}$$

and

$$\begin{aligned} G_v(s) &= E_v e^{-vs} = \frac{1}{(1 + \frac{s}{\delta})^\gamma} \\ G_v(\kappa) &= \delta(-1 + \kappa^{-1/\gamma}). \end{aligned}$$

³through the functions $G_v(\cdot)$ and $G_v^{-1}(\cdot)$.

⁴using $\bar{F}_1(t) = \exp(-v^* \phi \Lambda(t))$, $\bar{F}_0^{-1}(\kappa_1) = \Lambda^{-1}\{-\frac{1}{v^*} \ln(\kappa_1)\}$, and $\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\} = q^\phi$.

Thus,

$$\begin{aligned} \bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\} &= G_v(\phi G_v^{-1}(\kappa_1)) \\ &= \frac{1}{(1 + \phi(-1 + \kappa^{-1/\gamma}))^\gamma}. \end{aligned}$$

$$M(\kappa_1, \kappa_2) = \frac{\ln[\bar{F}_1\{\bar{F}_0^{-1}(\kappa_1)\}]}{\ln[\bar{F}_1\{\bar{F}_0^{-1}(\kappa_2)\}]} = \frac{\ln(1 + \phi(-1 + \kappa_1^{-1/\gamma}))}{\ln(1 + \phi(-1 + \kappa_2^{-1/\gamma}))}.$$

This example illustrated that the statistic $M(\kappa_1, \kappa_2)$ does not depend on the nonparametric baseline hazard $\Lambda(t)$. We use the empirical survival function and its inverse as estimates for $\bar{F}_1(\cdot)$ and $\bar{F}_0^{-1}(\cdot)$.

3. SIMULATION

Let $\kappa_1 = 0.25$ and $\kappa_2 = 0.75$ so that $H_0 : M(0.25, 0.75) = 4.8188$.

Model 1a: $\theta(t | x, v) = v \cdot 2^x$, $N = 800$

v	Mean M	Mean $t - statistic$	Rejection frequency
$v = 1$	5.0172	-0.2051	0.0385
$Exp(1)$	3.9169	-1.4437	0.2932
χ_1^2	3.7540	-1.6841	0.3631
$v = \begin{cases} 0.5 & w.p. 0.5 \\ 1.5 & w.p. 0.5 \end{cases}$	4.4486	-0.7180	0.1168

95% confidence level, # simulations = 1000

Model 1b: $\theta(t | x, v) = v \cdot 2^x$, $N = 1600$

v	Mean M	Mean $t - statistic$	Rejection frequency
$v = 1$	4.9488	-0.1008	0.0440
$Exp(1)$	3.9166	-1.9072	0.4340
χ_1^2	3.7329	-2.4308	0.5610
$v = \begin{cases} 0.5 & w.p. 0.5 \\ 1.5 & w.p. 0.5 \end{cases}$	4.3867	-0.9583	0.1760

95% confidence level, # simulations = 1000

Model 2a: $\theta(t | x, v) = vt^{-\frac{1}{2}}2^{x-1}$, $N = 800$

v	Mean M	Mean $t - statistic$	Rejection frequency
$v = 1$	5.0221	-0.2085	0.0589
$v \sim Exp(1)$	3.9491	-1.4062	0.2944
$v \sim \chi_1^2$	3.7557	-1.6989	0.3541
$v = \begin{cases} 0.5 & w.p. 0.5 \\ 1.5 & w.p. 0.5 \end{cases}$	4.4783	-0.7092	0.1149

95% confidence level, # simulations = 1000

Model 2a: $\theta(t | x, v) = vt^{-\frac{1}{2}}2^{x-1}$, $N = 1600$

v	Mean M	Mean $t - statistic$	Rejection frequency
$v = 1$	4.9429	-0.1096	0.0390
$v \sim Exp(1)$	3.8874	-1.9863	0.4360
$v \sim \chi_1^2$	3.6880	-2.5181	0.5940
$v = \begin{cases} 0.5 & w.p. 0.5 \\ 1.5 & w.p. 0.5 \end{cases}$	4.3994	-0.9139	0.1650

95% confidence level, # simulations = 1000

Model 3a: $\theta(t | x, v) = v \cdot 4^x$, $N = 1600$

v	Mean M	Mean $t - statistic$	Rejection frequency
$v = 1$	4.5692	-0.5318	0.0455
$v \sim Exp(1)$	3.0686	-5.2806	0.9570
$v \sim \chi_1^2$	2.9457	-6.1709	0.9790
$v = \begin{cases} 0.5 & w.p. 0.5 \\ 1.5 & w.p. 0.5 \end{cases}$	4.0619	-1.5585	0.3285

95% confidence level, # simulations = 1000

Model 3b: $\theta(t | x, v) = v \cdot v^{2^x - 1} t^{-\frac{1}{2}}$, $N = 1600$

v	Mean M	Mean $t - statistic$	Rejection frequency
$v = 1$	4.3215	-0.8800	0.0833
$v \sim Exp(1)$	3.0749	-5.2441	0.9570
$v \sim \chi_1^2$	2.9566	-6.0516	0.9830
$v = \begin{cases} 0.5 & w.p. 0.5 \\ 1.5 & w.p. 0.5 \end{cases}$	4.0436	-1.6070	0.3717

95% confidence level, # simulations = 1000

4. CONCLUSION

Since Lancaster (1979), it has been recognized that it is important to account for unobserved heterogeneity in models for duration data. Failure to account for unobserved heterogeneity makes the estimated hazard rate decreases more with the duration than the hazard rate of a randomly selected member of the population. In this paper, we derive a new test for unobserved heterogeneity. This test allows for a nonparametric baseline hazard and does not impose any distributional assumptions on the unobserved heterogeneity. Moreover, the test allows the durations to be censored.

REFERENCES

- [1] Athey, S. and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-in-Differences Models", 74, 431 - 497.
- [2] Baker, M. and A. Melino (2000): "Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study," *Journal of Econometrics*, 96, 357-93.
- [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. (1987): "Non-parametric Analysis of a Generalized Regression Model, the Maximum Rank Correlation Estimator", *Journal of Econometrics*, 35, 303-316.
- [7] Han, A. K. and J. A. Hausman (1990): "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*.
- [8] Hausman, J. A. and T. Woutersen (2006): Estimating a Semi-Parametric Duration Model without Specifying Heterogeneity, MIT working paper.
- [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] Heckman, J. J. (1991): "Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity," *American Economic Review*, 81, 75-79.
- [11] Honoré, B. E. (1990): "Simple Estimation of a Duration Model with Unobserved Heterogeneity," *Econometrica*, 58, 453-473.

- [12] Horowitz, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103-107.
- [13] Horowitz, J. L. (1999): "Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity" *Econometrica*, 67, 1001-1028.
- [14] Horowitz, J. L. (2001): "The Bootstrap" in *Handbook of Econometrics*, Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: North-Holland.
- [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] Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.
- [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] Sherman, R. P. (1993): “The Limiting Distribution of the Maximum Rank Correlation Estimator”, *Econometrica*, 61, 123-137.
- [26] 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.