Journal of Econometrics 157 (2010) 458–480

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Testing for unobserved heterogeneity in exponential and Weibull duration models

Jin Seo Cho [a], Halbert White [b,*]

[a] School of Economics, Yonsei University, Republic of Korea

[b] Department of Economics, University of California, San Diego, 92093-0508, United States

ARTICLE INFO

Article history:
Received 1 September 2008
Received in revised form
6 March 2010
Accepted 28 March 2010
Available online 8 April 2010

JEL classification:
C12
C22
C24
C41
C80
J22
J64

Keywords:
Unobserved heterogeneity
Mixture models
Likelihood ratio test
Search theory
Interarrival times

ABSTRACT

We examine the use of the likelihood ratio (LR) statistic to test for unobserved heterogeneity in duration models, based on mixtures of exponential or Weibull distributions. We consider both the uncensored and censored duration cases. The asymptotic null distribution of the LR test statistic is not the standard chi-square, as the standard regularity conditions do not hold. Instead, there is a nuisance parameter identified only under the alternative, and a null parameter value on the boundary of the parameter space, as in Cho and White (2007a). We accommodate these and provide methods delivering consistent asymptotic critical values. We conduct a number of Monte Carlo simulations, comparing the level and power of the LR test statistic to an information matrix (IM) test due to Chesher (1984) and Lagrange multiplier (LM) tests of Kiefer (1985) and Sharma (1987). Our simulations show that the LR test statistic generally outperforms the IM and LM tests. We also revisit the work of van den Berg and Ridder (1998) on unemployment durations and of Ghysels et al. (2004) on interarrival times between stock trades, and, as it turns out, affirm their original informal inferences.

1. Introduction

Econometric specifications for duration data are often based on exponential or Weibull distributions. In labor economics, Lancaster (1979) exploits these distributions to analyze unemployment spells. In financial econometrics, Engle and Russell (1998) and Engle (2000) exploit exponential and Weibull distributions to model interarrival times of stock transactions based upon market microstructure theory of Easley and O'Hara (1992) and O'Hara (1995). The properties and analyses of various duration models in economics are well reviewed in Kiefer (1988), Lancaster (1992) and Hong and Liu (2007). The popularity of the exponential and Weibull distributions is not restricted solely to economics. They are also widely applied in epidemiology, and especially clinical trials, to model the time to occurrence of significant milestones, such as death or recovery.

The presence of unobserved heterogeneity creates serious challenges for duration models. As Heckman and Singer (1984) point out, estimated parameters can be quite sensitive to the presence of unobserved heterogeneity. Thus, testing for unobserved heterogeneity often accompanies parameter estimation. For this, Lancaster (1979) and Kalbfleisch and Prentice (1980) assume a conventional gamma distribution for the unobserved heterogeneity and test for its presence by measuring the variance of the gamma distribution. Chesher (1984) and Lancaster (1985) propose an information matrix (IM) test (White, 1982, 1994, ch. 11), as the information matrix equality holds in the absence of unobserved heterogeneity. Kiefer (1985), Sharma (1987) and Prieger (2000, 2003) propose Lagrange multiplier (LM) tests exploiting the fact that the exponential and Weibull distributions can be represented using Laguerre polynomials.

In order to obtain estimates less sensitive to unobserved heterogeneity, researchers have developed a variety of flexible specifications. Heckman and Singer (1984) exploit a discrete mixture distribution for heterogeneity and estimate parameters using nonparametric maximum-likelihood estimation. Honoré (1990) assumes a Weibull distribution for the durations but does not impose a specific distribution on the heterogeneity, obtaining consistent parameter estimates. Meyer (1990) develops an estimation theory without specifying the baseline hazard, but, for convenience,

* Corresponding author. Tel.: +1 858 534 3502.
E-mail address: drhalwhite@yahoo.com (H. White).

0304-4076/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2010.03.046

retains the gamma distribution for heterogeneity. For the most part, however, the theories in this literature specify hazard or conditional mean functions in which the unknown coefficients multiply the explanatory variables, implementing a form of linearity. As yet, it is unknown how these theories may need to be modified when specifying general nonlinear models for the conditional mean. As we see in our empirical time-series application, a conditional mean equation embodying linearity can easily be misspecified. On the other hand, Horowitz (1999) provides a nonparametric estimation procedure under the condition that the distribution of heterogeneity is highly smooth; but if the associated heterogeneity is a discrete mixture, then this procedure may not work. There thus remains a need for tests of unobserved heterogeneity, whether continuous or discrete, applicable when the hazard function or conditional mean does not necessarily embody linearity.

The main goal of this paper is therefore to develop convenient test statistics having power comparable to or better than standard test statistics for unobserved heterogeneity and applicable to flexibly specified models. To achieve our goal, we develop log-likelihood ratio (*LR*) test statistics based upon a discrete mixture of exponential or Weibull distributions. We consider both the uncensored and censored duration cases. To develop our tests, we apply results of Cho and White (2007a), who build on work of Andrews (1999, 2001). Cho and White (2007a) use discrete mixtures to develop a test for regime-switching in a time-series context and derive the asymptotic distribution of the *LR* statistic under the null hypothesis of a single regime. Here, we obtain the asymptotic null distributions of our *LR* statistics under the hypothesis of no heterogeneity. As pointed out in the literature, the null distribution of the *LR* statistic in such situations is model dependent (e.g., Hartigan, 1985; Chernoff and Lander, 1995; Cho and White, 2007a). Thus, the null distributions derived for the discrete mixtures of binomials in Chernoff and Lander (1995) or normals in Cho and White (2007a) cannot be applied either to the discrete mixture of exponentials or the mixture of Weibulls. We separately derive the null distributions for these two cases. In particular, we find that the censored case differs substantially from the uncensored case. We provide procedures to obtain consistent asymptotic critical values for our *LR* test statistics, and we conduct large scale Monte Carlo simulations under various heterogeneity assumptions, including continuous distributions for heterogeneity. As we see, our *LR* test statistics have well behaved levels, and they appear to be consistent not only for discrete forms of heterogeneity, but also for many continuous heterogeneous alternatives.

Another goal of this paper is to revisit the empirical analyses of van den Berg and Ridder (1998) and Ghysels et al. (2004). Search theory predicts that unemployment durations follow an exponential distribution for each segmented labor market (see Yoon, 1981; van den Berg and Ridder, 1998; and the references therein). Nevertheless, most empirical papers in the literature admit that, due to unobserved heterogeneity, exponential distributions are hard to verify empirically. van den Berg and Ridder (1998) estimate reduced form equations for labor markets in the Netherlands and identify unobserved heterogeneity using *LR* statistics based on mixtures of exponential and Weibull distributions. Nevertheless, van den Berg and Ridder (1998) use an informal procedure to test for unobserved heterogeneity. Here we provide a formal testing procedure valid under their assumptions; as it turns out, we affirm their original inferences. Ghysels et al. (2004) note that an accurate analysis of financial market liquidity needs to accommodate both conditional mean and variance at the same time; for this they propose the stochastic volatility duration (SVD) model, which extends the exponential duration model with gamma heterogeneity to the time-series context. We examine their data using the methods proposed here, and affirm the presence of unobserved heterogeneity, motivating the use of the SVD model. As we discuss, correct specification of the conditional mean equation plays a key role for inference in this context.

The plan of this paper is as follows. In Section 2, we derive asymptotic null distributions for our *LR* statistics under the null hypothesis of no unobserved heterogeneity. We first treat the uncensored case. As we show, discrete mixtures of the exponential or Weibull distributions have different asymptotic null distributions. In particular, we represent the limiting behaviors of the *LR* statistic as functions of different Gaussian processes. We provide alternate representations of these processes; these yield consistent asymptotic critical values. We next discuss the censored case. As we see, the censored case differs substantially from the uncensored case. In particular, more involved methods based on those of Hansen (1996) are required to obtain consistent asymptotic critical values. In Section 3, we conduct Monte Carlo simulations comparing the performance of *LR*-based tests with IM and LM tests for both uncensored and censored cases. These experiments corroborate the results of Section 2 and provide useful insight into specifying key aspects of the parameter space. Section 4 contains our analysis of van den Berg and Ridder's (1998) and of Ghysels et al.'s (2004) data on Netherlands unemployment durations and interarrival durations of stock transactions, respectively. We provide a summary and conclusions in Section 5. The Appendix contains formal statements of the relevant assumptions and a link to proofs of our results.

2. Modeling durations with Weibull distributions

2.1. The data generating process

Let $\{(Y_t, X_t')'\}$ be a strictly stationary geometric β-mixing stochastic process, where Y_t is scalar-valued and X_t is $\mathbb{R}^k$-valued, $k \in \mathbb{N}$. Y_t represents a duration and must thus be non-negative. In the time-series context, for example, in the study of interarrival times for stock transactions, X_t can contain lagged values of Y_t. The β-mixing condition permits X_t to be a function of the entire history of Y_t, as is true for autoregressive conditional duration (ACD) models. (See Carrasco and Chen (2002).) As a special case, $\{(Y_t, X_t')'\}$ can be an independent identically distributed (IID) process, as is suitable for cross-section data. In this case, t indexes individuals in the sample, and the elements of X_t are duration-invariant. For either time-series or cross-section data, we assume without loss of generality that X_t does not contain a constant term, as explained below.

Throughout, we assume that our interest focuses on the conditional distribution of Y_t given X_t. For the uncensored case and in the absence of unobserved heterogeneity, we suppose that the conditional probability density function (PDF) of the durations is defined by the Weibull density

$$f(y \mid X_t; \delta^*, \beta^*, \gamma^*) = \delta^* \gamma^* g(X_t; \beta^*) y^{\gamma^*-1} \exp(-\delta^* g(X_t; \beta^*) y^{\gamma^*}),$$

for some $(\pi^*, \delta^*, \beta^{*\prime}, \gamma^*) \in [0, 1] \times D \times B \times \Gamma \subset [0, 1] \times \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^+$, where $g(X_t; \cdot)$ satisfies the differentiability condition given in the Appendix. We leave the functional form of $g(X_t; \beta^*)$ unspecified, as this form differs from application to application. In labor economics, $g(X_t; \beta^*)$ is often Cox's (1972) proportional hazard function, and in many cases, $g(X_t; \beta^*) = \exp(X_t'\beta^*)$ is the chosen specification. In finance, $g(X_t; \beta^*)$ is often associated with the conditional mean of $Y_t \mid X_t$. In particular, ACD models specify autoregressive functions for the conditional mean of