A Test of Heterogeneity in Constant Hazard Models Using Least Squares*

Bong Joon Yoon
Dongguk University and SUNY-Binghamton

The presence of unmeasured heterogeneity can seriously bias inference in economic duration models. To detect the presence of heterogeneity in the hazard models of duration, the tests proposed in the past utilize cumbersome maximum likelihood procedures. This paper presents an alternative test assuming a constant hazard. Our diagnostic test is based on the least squares regression, and hence it is simple to implement.

I. Introduction

It has been widely recognized in recent literature that the presence of unobserved or unmeasured heterogeneity can seriously bias inference in economic duration models. To detect the presence of this heterogeneity bias in duration model estimation, namely, spurious duration dependence, several tests have been proposed in the past. However, these tests are based on the maximum likelihood technique, which require cumbersome estimation procedures.

This paper presents an alternative method to test for the presence of unmeasured heterogeneity when the hazard of a duration variable is invariant over time. Our diagnostic test is based on the least squares regression, hence it is simple to implement.

II. The Model

Let T denote a random economic duration variable, e.g., unemployment duration, time till the next child birth, and so on. A characterization of the distribution of the duration variable T is given by the

*I appreciate comments from an anonymous referee which has significantly improved the quality of the paper. Any remained errors are my responsibility.

hazard $h(t)$, which is the instantaneous probability that the current stage (e.g., unemployment) terminates at t. Assume a constant hazard so that $h(t) = h$ for all t. Then, the probability density function of T, $f(t)$, is given by the following exponential density:

$$f(t) = h \exp(-ht), \tag{1}$$

following the relation $h(t) = f(t)/[1 - F(t)]$ where $F(\cdot)$ is the cumulative density. Assume that h depends on regressors denoted by a vector $x = (x_1, x_2, \ldots, x_k)$. As an illustration, consider the specification $h = \exp(\beta' x)$ where β is a vector of coefficients.

The model (1) can be estimated using the maximum likelihood technique.[1] The data may come in the form of i) completed duration, ii) censored duration, or iii) a combination of both. As long as the null hypothesis of the exponential density is maintained, (1) can be used to specify the likelihood for all the three cases.[2]

Now suppose that there exists unmeasured heterogeneity in the constant hazard duration model (1). This may arise from either omitted variables in the list of the regressors x or truly unmeasurable variables, e.g., ability. Let the unmeasured heterogeneity component be denoted by a random variable u. Then the duration model under heterogeneity is described by the density

$$f^*(t) = h^* \exp(-h^* t) \tag{2}$$
$$\text{where } h^* = \exp(\beta' x + u).$$

Model (2) also exhibits a constant hazard. However, ignoring the presence of the unmeasured heterogeneity term u will seriously bias the coefficient estimates, resulting in spurious duration dependence of the estimated hazard function.[3] Diagnostic tests for the presence of the heterogeneity component have been suggested, for example, by Chesher (1984) in a general setting and Kiefer (1984) for duration models. These past approaches use maximum likelihood estimation results to form a score test. This paper provides an alternative diagnostic test based on the least squares regression results.

[1]For estimation of exponential duration models, see Lancaster (1979) and Yoon (1981) among others.

[2]See Feller (1971) and Heckman and Singer (1984).

[3]See Lancaster (1979), Kiefer (1984), and Heckman and Singer (1984).

III. The Test

Define a new random variable W such that

$$T = (1/h)\ \exp(-W) \tag{3}$$

Then using the change of variable rule one can show that the pdf of W is given by

$$g(w) = \exp(-w - \exp(-w)). \tag{4}$$

That is, the random error term W has an extreme value (or Fisher–Tippett Type I) distribution with parameters $\delta_1 = 0$ and $\delta_2 = 1$.[4] Thus

$$\begin{aligned} &\mathrm{E}(W) = \nu \\ \text{and } &\mathrm{Var}(W) = \pi^2/6 \\ \text{where } &\nu = \text{Euler's constant} = 0.577\ldots. \end{aligned}$$

Assume absence of nonmeasured heterogeneity. Then, from (3) it follows that

$$\ln T = -\beta' x - W, \tag{5}$$

where x is a $k \times 1$ vector of explanatory variables including a constant term. Now denote a coefficient vector by $\alpha = (\alpha_1,\ldots, \alpha_k)$ such that $\alpha_1 = -\beta_1 - \nu$ and $\alpha_j = -\beta_j$ for $j = 2,\ldots, k$. Defining a new random variable $\varepsilon = \nu - W$, (5) can be rewritten as

$$\ln T = \alpha' x + \varepsilon . \tag{6}$$

The error term ε has mean 0 and a constant (homoscedastic) variance $\pi^2/6$.[5] The regression model (6) corresponds to the density f in (1) based on the assumption of no unmeasured heterogeneity component in the hazard function. Now assume that the heterogeneity component u is present in the hazard so that the density of T is f^* as given in (2). Then, following the same steps as above, we can specify the regression analogue of the model f^* as

$$\ln T = \alpha' x + \varepsilon^*, \tag{7}$$

where $\varepsilon^* = -u + \varepsilon$ and ε is as defined earlier. Since x includes

[4]See Abramowitz and Stegun (1972, p. 930).

[5]Note that in (1) $\mathrm{Var}(t) = h^{-2} = \exp(-2\beta' x)$ varies with x. This heteroscedasticity disappears in model (6).

a constant term, it is innocuous to assume that $\mathrm{E}(u) = 0$ and $\mathrm{E}(\varepsilon^*) = 0$. However,

$$\mathrm{Var}(\varepsilon^*) = \mathrm{Var}(u) + \mathrm{Var}(\varepsilon) > \mathrm{Var}(\varepsilon) = \pi^2/6, \tag{8}$$

assuming that u and ε are uncorrelated.[6]

In discriminating between the constant hazard model with and without the heterogeneity component u, i.e., between the models (1) and (2), this paper proposes to utilize the fact that $\mathrm{Var}(\varepsilon^*) > \mathrm{Var}(\varepsilon)$. Let the null hypothesis be absence of heterogeneity. Under this null hypothesis, $\mathrm{Var}(\varepsilon)$ is consistently estimated by the least squares variance estimator $S^2 = Y'(I - X(X'X)^{-1}X')Y/(n-k)$ where X is an $n \times k$ matrix formed by $(x_1', \ldots, x_i', \ldots, x_n')$ and Y is an $n \times 1$ vector of the dependent variable denoted by $(\ln t_1, \ldots, \ln t_i, \ldots, \ln t_n)$ with i representing the i-th observation and n the sample size. The present strategy of diagnostic testing for the presence of heterogeneity is to check whether S^2 is significantly greater than $\pi^2/6$.

Under the null hypothesis of no heterogeneity, it can be shown that $n^{1/2}(S^2 - \mathrm{Var}(\varepsilon)) = n^{1/2}(S^2 - \pi^2/6)$ is asymptotically distributed as $N(0, \mu_4 - \mathrm{E}(\varepsilon^4))$ where μ_4 is the fourth central moment of ε.[7] Following the density (4) of w, it can be shown that

$$\mu_4 = \mathrm{E}[\varepsilon - \mathrm{E}(\varepsilon)]^4 = \mathrm{E}[w - \mathrm{E}(w)]^4 = 11.429.$$

If the observed value of the statistic $n^{1/2}(S^2 - \pi^2/6)$ turns out to be significantly larger than 0, one may reject the null hypothesis and conclude that there exists unmeasured heterogeneity component in the duration model.

An even simpler test is obtained from the observation that the statistic $(n-k)S^2/\sigma^2$ has a χ^2 distribution with $(n-k)$ degree of freedom, where σ^2 is $\pi^2/6$ under the maintained hypothesis.[8] Thus, if the observed value of $(n-k)S^2/\sigma^2$ with the maintained value of σ^2 exceeds the upper critical value at a pre-specified significance level, then the hypothesis of no unmeasured heterogeneity is rejected.[9]

[6]This paper focuses on correcting the spurious duration dependence due to the presence of unmeasured heterogeneity per se. An interesting new problem is to analyze the effect of correlation between heterogeneity component and the usual disturbance term, that is, between u and ε.

[7]See Schmidt (1976).

[8]See Theil (1971, p. 130).

[9]Our model assumes a time-invariant hazard. A more general one is obtained by assuming a Weibull hazard which allows for time dependency of the $p-1$ hazard. Namely, $h(t) = \lambda p \exp(\lambda p)$. Hence our constant hazard model is limited in that $p = 1$. With p not

IV. Conclusion

This paper has proposed a test of unmeasured heterogeneity in economic duration models in a constant hazard context. This test is based on the least squares regression in constrast to the previous tests using the cumbersome maximum likelihood estimation procedures. That is, our test is simpler to implement. Specifically, the test is offered as a useful alternative to the ones given by Chesher (1984) and Kiefer (1984), for example. A complete evaluation of the performance of the test, however, requires considerations of its small sample properties and power. This awaits further study.

References

Abramowitz, M., and Stegun, I. *Handbook of Mathematical Functions.* New York: Dover Publications, 1972.

Chesher, A. "A Test of Neglected Heterogeneity." *Econometrica* 52 (1984): 865-72.

Feller, W. *An Introduction to Probability Theory and Its Applications*, Vol. II, New York: John Wiley and Sons, 1971.

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

Kiefer, N. "Simple Test for Heterogeneity in Exponential Models of Duration." *Journal of Labor Economics* (1984).

Lancaster, T. "Econometric Models for the Duration of Unemployment." *Econometrica* 47 (1979): 989-56.

Schmidt, P. *Econometrics.* New York: Marcel Dekker, 1976.

Theil, H. *Principles of Econometrics.* New York: John Wiley and Sons, 1971.

Yoon, B. "A Model of Unemployment Duration with Variable Search Intensity." *Review of Economics and Statistics* 63 (1981): 599-609.

being fixed at 1, detecting unmeasured heterogeneity requires estimation of the parameter p. A possible approach is to estimate p first and then investigate the variance of ε / p accounting for the variance of the first round estimator of p. To devise a simple test for the presence of unmeasured heterogeneity with this time-varying hazard would be a valuable generalization of the current study. This requires further study.