# Joint and separate score tests for state dependence and unobserved heterogeneity

Sanjiv Jaggia

Pravin K. Trivedi

The paper compares separate, conditional, and joint score tests of duration dependence and unobserved heterogeneity when the null is the exponential model and the alternative is the heterogeneous Weibull model. The score tests based on the conditional score function include the Neyman $C(\alpha)$ test as a special case. An examination of the non-null distribution of the joint test explains when all score tests have low power in the presence of multiple misspecifications. Monte Carlo experiments show that the conditional score tests are superior to the standard separate tests which confound unobserved heterogeneity and duration dependence.

*Key words*: Conditional score; C-alpha test

## 1. Introduction

A number of recent papers consider specification tests for duration models; see especially Lancaster (1985), Kiefer (1985), Burdett et al. (1985), Sharma (1987), Jensen (1987), Horowitz and Neumann (1989). A focus in these papers has been on separate (partial) tests of either unobserved heterogeneity or on tests of duration (state) dependence. In practice an investigator is likely to encounter both neglected heterogeneity and duration dependence simultaneously. It is

often difficult to infer whether the data come from a very heterogeneous population with low duration dependence or a relatively homogeneous population with high duration dependence. This is subsequently referred to as 'confounding'. In this paper we carry out a theoretical analysis of the confounding effects on tests for the two types of misspecifications and suggest ways of avoiding distorted inferences. We also carry out an extensive Monte Carlo investigation of the properties of various test procedures when the true data generation process allows for duration dependence, unobserved heterogeneity, or both.

Our investigation is carried out on the basis of a comparison between three types of score-based tests. The null hypotheses tested are of zero unobserved heterogeneity and zero duration dependence in the context of the heterogeneous Weibull model. The tests used are the separate and joint score tests for two parametric hypotheses and the conditioinal ('adjusted') score test for testing a single separate hypothesis.

It is known that, given duration dependence, neglected heterogeneity leads to inconsistent estimates. Therefore, the motivation for testing for heterogeneity is strong. However, in the presence of multiple misspecifications the use of a separate test, such as the test of unobserved heterogeneity, may not be valid. Due to stochastic dependence of the separate tests a significant test of unobserved heterogeneity may be an indicator of duration dependence. Similarly, an insignificant test may result due to the confounding effects of simultaneous occurrence of duration dependence and unobserved heterogeneity. Consequently, testing separate hypotheses may lead to a model which is either over- or underparameterized, a common presumption being that overparameterization is more likely [Godfrey (1988, p. 80)]. Awareness that the separate tests are not strictly valid in the presence of other possible misspecifications causes some users of separate tests to exercise caution and to claim that the separate test is only a general misspecification ('something is wrong') test and not directed at a particular misspecification. Under that interpretation a significant test cannot suggest a direction for respecification.

A second approach is to test for the presence of duration dependence and heterogeneity jointly and hence deal with the possible correlation between separate tests. A more general model may be inferred if the joint test is significant. However, if only a subset of the joint hypothesis is false, this approach will lead to overparameterization. Further, as shown by an example later in the paper, the joint test may have low power against certain local alternatives.

A third approach tests either of the two separate hypotheses without assuming the truth of the complementary hypotheses. That is, separate tests are constructed which allow for the dependence of the test on certain nuisance parameters. To implement it one either fits a more general model than implied by the joint null as in the Neyman $C(\alpha)$ test or, as in section 4 of the paper,

'adjusts' or 'orthogonalizes' the scores estimated under the joint null with respect to the nuisance parameters. In each case the score test will be based on the conditional score function. This approach is known to be asymptotically equivalent to the separate tests when the joint null is true. Under the alternative hypothesis this approach can be more powerful. When the joint null is untrue, the properties of the conditional score tests depend upon the nature of conditioning.

We explore and elaborate these general themes by a detailed theoretical examination of a model of duration dependence and unobserved heterogeneity for which we develop several specification tests whose properties are subsequently explored in the context of Monte Carlo experiments. This is the main contribution of the paper. The rest of the paper is organized as follows. Though our primary interest is in the specific application, section 2 provides a general exposition of several variants of the conditional score approach and contrasts it with other approaches. Our justification for a general discussion is to exposit the structure and merits of the conditional score approach which is not used widely in the empirical econometric literature. Section 3 discusses the joint score test for duration dependence and neglected heterogeneity, its non-null distribution, and the conditions under which the test has low power. Section 4 specializes the approach of section 2 and derives conditional score tests for heterogeneity and duration dependence based on the maximum likelihood estimator. Section 5 shows how to implement the conditional test for heterogeneity using a variant based on $\sqrt{N}$-consistent least squares estimates of the Weibull regression model. This latter approach is essentially an application of the Neyman $C(\alpha)$ principle and is preferable in principle to the approach of section 4. Sections 6, 7, and 8 report Monte Carlo experiments which compare the three approaches in three different settings. Section 9 concludes.

## 2. Conditional score tests

The common approach of conducting separate score tests of subsets of restrictions, assuming the validity of other untested restrictions, can have misleading consequences if the untested hypotheses are false. For example, in parametric duration models it is common to assume that the functional form of the hazard is correctly specified when testing for neglected heterogeneity. In contrast, a conditional ('adjusted') score test can be constructed which does not make that assumption.

Let $\mathscr{L}(\theta)$ denote the log-likelihood, where $\theta' = [\theta_1' \ \theta_2']$ is the vector of $m$ parameters to be estimated and $\theta_1$ has $r$ elements. Let $s_1 = \partial \mathscr{L}/\partial \theta_1$ and $s_2 = \partial \mathscr{L}/\partial \theta_2$ denote the score vectors. Let $I(\theta) = -\mathbb{E}[\partial^2 \mathscr{L}/\partial \theta_1 \partial \theta_2]$ denote the expected Fisher information matrix and let $\theta^*$ denote the true unknown $\theta$.

Under standard regularity condition [see Holly (1987) or Godfrey (1988)] the score vector has a multivariate normal distribution, that is,

$$N^{-1/2}\begin{bmatrix} s_1(\theta^*) \\ s_2(\theta^*) \end{bmatrix} \sim \mathcal{n}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{N}\begin{pmatrix} I_{11}(\theta^*) & I_{12}(\theta^*) \\ I_{21}(\theta^*) & I_{22}(\theta^*) \end{pmatrix}\right], \tag{2.1}$$

where

$$I(\theta^*) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}. \tag{2.2}$$

Let the joint null hypothesis be $H_0$: $\theta_1 = \theta_{10}$. The score test for $H_0$ is

$$LM[\hat{\theta}] = s_1'(\hat{\theta}_0)[I^{11}(\hat{\theta}_0)]s_1(\hat{\theta}_0), \tag{2.3}$$

where

$$I^{11} = [I_{11} - I_{12}I_{22}^{-1}I_{21}]^{-1}, \tag{2.4}$$

and the carat (^) denotes evaluation at the restricted maximum likelihood estimates of $\theta$. $LM$ is distributed as $\chi^2(r)$ under $H_0$.

The conditional distribution of $s_1(\theta)$, given $s_2(\theta) = c_2$, where $c_2$ is the realized value of $s_1$, is given by

$$(s_1 | s_2 = c_2) \sim \mathcal{n}[I_{12}I_{22}^{-1}c_2, (I^{11})^{-1}]. \tag{2.5}$$

Define the *conditional score function* as

$$s_1^c(\theta) = (s_1(\theta) | s_2 = c_2) - I_{12}(\theta)I_{22}^{-1}(\theta)c_2, \tag{2.6}$$

where the second term on the right-hand side is the 'adjustment' whose effect may be interpreted as purging $s_1$ of the correlation with $s_2$. An alternative interpretation is in terms of orthogonalization of $s_1^c$ and $s_2$ in the sense of zero asymptotic covariance. The adjusted score $s_1^c = s_1$ if $I_{12} = 0$. Notice that no conditioning is necessary if all the parameters that are not being tested are estimated unrestricted using maximum likelihood. Here $c_2 = 0$ by the first-order condition $\partial\mathcal{L}/\partial\theta_2 = 0$, and $s_1^c = s_1$. The adjusted score may be interpreted also as the residual from the regression of $s_1$ on $s_2$.

Once again, under standard regularity conditions,

$$N^{-1/2}\begin{bmatrix} s_1^c(\theta^*) \\ s_2(\theta^*) \end{bmatrix} \sim \mathcal{n}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{N}\begin{pmatrix} [I^{11}(\theta^*)]^{-1} & 0 \\ 0 & I_{22}(\theta^*) \end{pmatrix}\right], \tag{2.7}$$

which establishes the orthogonality (zero asymptotic covariance) between $s_1^c(\theta^*)$ and $s_2(\theta^*)$. The conditional score test statistic is given by

$$LM^c[\theta^*] = s_1^{c\prime}(\theta^*)\, I^{11}(\theta^*)\, s_1^c(\theta^*)\,. \tag{2.8}$$

Let $\theta$ be further partitioned as $\theta' = (\theta'_{11}\ \theta'_{12}\ \theta'_{22})$. This redefinition of $\theta$ facilitates the implementation of the two types of conditional score tests that we consider in this paper. The following are the two seperate hypotheses we need to test: first, for $H_{01}$: $\theta_{11} = \theta_{110}$, $\theta'_2 = (\theta'_{12}\ \theta'_{22})$. Conversely, $\theta'_2 = (\theta'_{11}\ \theta'_{22})$, when testing $H_{02}$: $\theta_{12} = \theta_{120}$. Consider $H_{01}$: $\theta_{11} = \theta_{110}$. The first type of conditional score test consists of using the $\sqrt{N}$-consistent estimate of $(\theta'_{12}\ \theta'_{22})$, denoted by a tilde ($\tilde{}$). The second type of test uses the restricted maximum likelihood estimate of only $\theta_{22}$, denoted by a carat ($\hat{}$). For testing $H_{01}$, where $(\tilde{\theta}_{12}\ \tilde{\theta}'_{22})$ is a $\sqrt{N}$-consistent estimator of $(\theta^*_{12}\ \theta^{*\prime}_{22})$. Then under $H_{01}$, $s_1^c(\tilde{\theta}) \xrightarrow{d} s_1(\theta^*)$ and the conditional score test, given by

$$LM^c[\tilde{\theta}] = s_1^{c\prime}(\tilde{\theta})\, I^{11}(\tilde{\theta})\, s_1^c(\tilde{\theta})\,, \tag{2.9}$$

has the same asymptotic distribution as $LM^c[\theta^*]$.

Now consider the conditional score based on the restricted maximum likelihood estimator. The restriction is placed on both elements of $\theta_{110}$ and $\theta_{120}$, however, only $\theta_{110}$ is being tested. Let $\hat{\theta}' = (\theta'_{110}\ \theta'_{120}\ \hat{\theta}'_{22})$. Define the new conditional score $s_1^c(\hat{\theta})$ analogously to (2.6) with $\theta$ replaced by $\hat{\theta}$. Therefore, another conditional score test of $H_{0j}$: $\theta_{1j} = \theta_{1j0}$ $(j = 1, 2)$ is given by

$$LM_j^c(\hat{\theta}) = s_j^{c\prime}(\hat{\theta})\, I^{jj}(\hat{\theta})\, s_j^c(\hat{\theta})\,. \tag{2.10}$$

Since the score test based on $s_j^c(\tilde{\theta})$ does not impose $H_{0k}$ $(k \neq j)$ and is asymptotically equivalent to the score test based on $s_1^c(\theta^*)$, its asymptotic distribution is independent of $(\theta^{*\prime}_{1j}\ \theta^{*\prime}_{22})$, $j = 1, 2$. We shall refer to this as the *proper conditioning* case. In contrast the score test based on $s_j^c(\hat{\theta})$ imposes $H_{01}$ and $H_{02}$ simultaneously. If the untested auxillary hypothesis is false and $I(\theta^*)$ is not block-diagonal in the elements under test, the distribution of $s_j^c(\hat{\theta})$ will not converge to that of $s_j^c(\theta^*)$ and hence will not, in general, be independent of the nuisance parameter. We shall refer to this as the case of *improper conditioning*.

The above argument explains why in general $LM_j^c[\hat{\theta}]$ and $LM_j^c[\tilde{\theta}]$ will have different properties. Although $LM_j^c[\tilde{\theta}]$ is preferred to $LM_j^c[\hat{\theta}]$, the latter may still be preferred to $LM_j[\hat{\theta}]$ which uses restricted maximum likelihood estimates but no conditioning. $LM_j^c[\hat{\theta}]$ is a potentially useful test statistic if $\hat{\theta}$ is not readily computable.

The conditional score test based on $\tilde{\theta}$ and defined by (2.9) is known as the Neyman $C(\alpha)$ test in the statistical literature [Neyman (1959), Moran (1970)]

and has been discussed in the econometrics literature by Breusch and Pagan (1980), Engle (1984), and Holly (1987). Traditionally the test has been motivated by computational considerations since the $\sqrt{N}$-consistent estimate of $\theta$ under the null is often easier to obtain than the maximum likelihood estimate required for the score test. Hence the test is sometimes dubbed 'pseudo-score test'.

### 3. Score tests of heterogeneity and duration dependence

Following several recent papers [for example, Lancaster (1985)], we consider a locally heterogeneous Weibull model with possibly censored observations on durations $t_i$, $i = 1, \ldots, N$, whose joint log-likelihood is given by

$$\mathscr{L} = \sum \ln f(t_i), \tag{3.1}$$

where

$$\ln f(t_i) = C_i\left[\ln\alpha + (\alpha - 1)\ln t_i + \ln\mu_i + \ln\left(1 + \frac{\sigma^2}{2}(\varepsilon_i^2 - 2\varepsilon_i)\right)\right]$$

$$- \varepsilon_i + (1 - C_i)\ln\left(1 + \frac{\sigma^2}{2}\varepsilon_i^2\right),$$

$$\mu_i = \exp(x_i'\beta) = \exp(\beta_0 + x_{1i}'\beta_1), \tag{3.3}$$

$$\varepsilon_i = \mu_i t_i^{\alpha}. \tag{3.4}$$

Here $x_i$ is a $(K \times 1)$ vector of exogenous variables and $\varepsilon_i$ is a *generalized error* in the sense of Cox and Snell [see Lancaster (1985)]. $C_i$ is the censoring indicator which takes the value unity for complete durations and zero for right-censored durations. The duration dependence parameter is $\alpha$, $\alpha < 1$ implying negative duration dependence and $\alpha > 1$ implying positive duration dependence. The heterogeneity parameter is $\sigma^2$.

With $\theta' = (\theta_1' \ \theta_2')$, where $\theta_1' = (\sigma^2 \ \alpha)$, $\theta_2' = (\beta_0 \ \beta_1')$, and let $\theta_0' = (0 \ 1 \ \beta_0 \ \beta_1')$ denote the restricted vector and $\hat{\theta}_0'$ its maximum likelihood estimate. The joint null hypothesis of zero heterogeneity and no duration dependence is

$$\mathrm{H}_0: \quad \sigma^2 = 0 \quad \text{and} \quad \alpha = 1. \tag{3.5}$$

The scores with respect to $\sigma^2$ and $\alpha$ are as follows:

$$\left.\frac{\partial \mathscr{L}}{\partial \sigma^2}\right|_{H_0} = \tfrac{1}{2}\sum_i (\varepsilon_i^2 - 2C_i\varepsilon_i) \equiv s_{11}(\theta)\,, \tag{3.6}$$

$$\left.\frac{\partial \mathscr{L}}{\partial \alpha}\right|_{H_0} = \sum_i (C_i + (C_i - \varepsilon_i)\ln t_i) \equiv s_{12}(\theta)\,. \tag{3.7}$$

However, most of the theoretical part of the paper up to section 7 will deal only with the special case of no censoring. We have derived the joint score test for $H_0$ for the uncensored case. The result is now summarized.

*Proposition 1. Since asymptotically* $[I(\theta)]^{-1/2} s(\theta) \sim \mathscr{N}(0, I)$, *the joint score test statistic for heterogeneity and duration dependence is*

$$LM_{HD}[\hat{\theta}] = s_1'(\hat{\theta}_0)\, I^{11}(\hat{\theta}_0)\, s_1(\hat{\theta}_0)\,, \tag{3.8}$$

*where*

$$I^{11}(\theta_0) = \frac{(\Psi'(1) - 1)^{-1}}{N}\begin{bmatrix} \Psi'(1) & 1 \\ 1 & 1 \end{bmatrix}, \tag{3.9}$$

*where* $\Psi(r)$ *is the digamma function,* $\mathrm{d}\log\Gamma(r)/\mathrm{d}r$, *and* $\Psi'(r)$ *is the trigamma function,* $\mathrm{d}^2\log\Gamma(r)/\mathrm{d}r^2$. ■

*Proposition 2. Let* $\delta' = (\delta_1\ \delta_2)$, *where* $\delta_1$ *and* $\delta_2$ *are scalar constants. Let the sequence of alternative hypotheses* $H_A$ *be given by*

$$\theta_1 = \theta_0 + N^{-1/2}\delta\,. \tag{3.10}$$

*Under certain regularity conditions, asymptotically,*

$$LM_{HD} \sim \chi^2(2, \lambda)\,, \tag{3.11}$$

*where the noncentrality parameter* $\lambda$ *is defined as*

$$\lambda = [\delta_1\ \delta_2]\begin{bmatrix} 1 & -1 \\ -1 & \Psi'(1) \end{bmatrix}\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \delta_1^2 - 2\delta_1\delta_2 + \delta_2^2\,\Psi'(1). \quad \blacksquare \tag{3.12}$$

We note that

$$\lambda|_{\alpha=1} = N\sigma^4 > 0\,, \tag{3.13}$$

$$\lambda|_{\sigma^2=1} = N(\alpha-1)^2\,\Psi'(1) > 0\,. \tag{3.14}$$

The function $\lambda$ is biquadratic in $(\delta_1, \delta_2)$, where $\delta_1, \delta_2 > 0$. We see from the sign of the middle term in (3.12) that certain configurations of values of $\sigma^2$ and $\alpha$ will make $\lambda$ 'small'. This happens when $\alpha > 1$, i.e., $\delta_2 > 0$. In words, the joint presence of heterogeneity and *positive* duration dependence induces a small value of the noncentrality parameter, thereby reducing the power of the test against local alternatives. It is readily seen that for a given $\delta_1$ the minimum value of $\lambda$, which is given by

$$\min_{\{\delta_2>0\}} \lambda = \lambda|_{\delta_2=\delta_1/\Psi'(1)} = \delta_1^2(1 - 1/\Psi'(1))\,, \tag{3.15}$$

may still be high if $\delta_1^2$ is large. On the other hand, when $\alpha < 1$, implying negative duration dependence, the power of the joint score test is increased. The simultaneous presence of positive duration dependence and heterogeneity poses a problem for both joint and separate score tests, reducing the power of the former and invalidating the latter. It may also lead the researcher to underparameterize the model.

## 4. Conditional score tests for heterogeneity and duration dependence

To derive exact expressions for the conditional score tests for the heterogeneous Weibull model consider only one restriction at a time. For the test of heterogeneity this is $H_{01}$: $\sigma^2 = 0$. The derivation of the test uses the following expressions:

$$\frac{1}{N} I_{11} = \begin{bmatrix} 2 & -(m+1) \\ -(m+1) & 1 + \Psi'(2) + \beta_1' \bar{\Omega} \beta_1 + m^2 \end{bmatrix},$$

$$\frac{1}{N} I_{12} = \begin{bmatrix} -1 & 0 \\ m & -\beta_1' \bar{\Omega} \end{bmatrix},$$

$$\frac{1}{N} I_{22} = \begin{bmatrix} -1 & 0 \\ 0 & \bar{\Omega} \end{bmatrix},$$

where $m = \Psi(2) - \beta_0$, $x' = (1\ x_1')$, and $\mathbb{E}(x_1') = 0$, $\mathbb{E}[x_1 x_1'] = \bar{\Omega}$ (nonsingular), $\mathbb{E}[xx'] = \Omega$. The test statistic is easily shown to be

$$LM_{\mathrm{H}}^{\mathrm{c}}(\hat{\theta}) = \frac{[s_{11}(\hat{\theta}) + (1/\Psi'(1))c_{12}(\hat{\theta})]^2}{N(1 - 1/\Psi'(1))}, \tag{4.1}$$

where $s_{11} = \frac{1}{2}\sum_i (\varepsilon_i^2 - 2\varepsilon_i)$ and $c_{12} = \sum_i (1 + (1 - \varepsilon_i)\ln t_i)$ are, respectively, the score with respect to $\sigma^2$ and the score with respect to $\alpha$, evaluated at $\alpha = 1$. Further, $\partial\mathscr{L}/\partial\beta = c_{22} = 0$, since the maximum likelihood estimator of $\beta$ is used. Analogously the conditional score test for zero duration dependence, viz., $\mathrm{H}_{02}$: $\alpha = 1$, is

$$LM_{\mathrm{D}}^{\mathrm{c}}(\hat{\theta}) = \frac{[s_{12}(\hat{\theta}) + c_{11}(\hat{\theta})]^2}{N(\Psi'(1) - 1)}. \tag{4.2}$$

The reader is again reminded that in general (4.1) and (4.2) involve improper conditioning. The case of proper conditioning is dealt with in section 5.

*Relationship between the conditional test and some separate tests*

Given $\mathscr{L}(\theta_{110}, \theta_{120}, \hat{\theta}_{22})$, the likelihood evaluated at the jointly restricted maximum likelihood parameter values, a separate or partial score test for (say) the first restriction is based on the efficient score for that parameter ignoring its possible correlation with the second score. For example, assuming $\alpha = 1$ and testing $\mathrm{H}_{01}$: $\sigma^2 = 0$ yields the partial or separate test for heterogeneity [Lancaster (1985)], viz.,

$$LM_{\mathrm{H}} = \frac{1}{N} s_{11}^2(\hat{\theta}_0). \tag{4.3}$$

Analogously, the partial test of duration dependence, $\mathrm{H}_{02}$: $\alpha = 1$, assuming $\sigma^2 = 0$, is

$$LM_{\mathrm{D}} = \frac{1}{N\Psi'(1)} s_{12}^2(\hat{\theta}_0). \tag{4.4}$$

Both $LM_{\mathrm{H}}$ and $LM_{\mathrm{D}}$ are $\chi^2(1)$ tests, and though they are quite appropriate under the joint null, the actual test size will differ from the nominal significance level under the alternative since in that case the covariance between them is nonzero. Block diagonality of $I^{11}$ with respect to $\theta_{11}$ and $\theta_{12}$ is a necessary and

sufficient condition for the joint test to be additive in $LM_D$ and $LM_H$ [Bera and McKenzie (1987)]. However, $I^{11}(\hat{\theta}_0)$ is not block-diagonal. Tests which ignore this may be misleading; see Jaggia (1991) for an empirical example.

Usually the joint presence of heterogeneity and duration dependence cannot be ruled out *a priori*. Hence the joint score test has obvious attractions over the partial and conditional tests. On the other hand, if the joint null is rejected, one may wish to test the component hypotheses. Further, for some parameter configurations the joint test is likely to have low power.

## 5. OLS-based $C(\alpha)$ test of heterogeneity

The distribution of the test $LM^c[\hat{\theta}]$ developed above will depend in general upon unknown nuisance parameters. We desire a properly conditioned test, viz. the $C(\alpha)$ test, which is asymptotically equivalent to that based on the maximum likelihood estimates of the Weibull model. Appropriate expressions for the score and information matrix when the data are uncensored and a procedure for obtaining $\sqrt{N}$-consistent estimates are required.

To construct a $C(\alpha)$ test of $H_0$: $\sigma^2 = 0$, $\sqrt{N}$-consistent estimates of the Weibull model are estimated as follows: if $t_i$ $(i = 1, \ldots, N)$ are Weibull distributed, then using $y_i = \ln(t_i)$ we can write

$$y = -X\beta/\alpha + u/\alpha \tag{5.1}$$

$$= X\omega + U$$

$$= \omega_0 + X_1\omega_1 + U, \tag{5.2}$$

where $u$ has an extreme value distribution with $E(u) = \Psi(1) = -0.5772$ and $\text{var}(u) = \Psi'(1) = 1.6449$, $\omega' = \beta'/\alpha$, and $U = u/\alpha$. Hence

$$\alpha = [1.6449/\text{var}(U)]^{1/2}, \tag{5.3}$$

$$\beta_0 = -\alpha\omega_0 - 0.5772, \tag{5.4}$$

$$\beta_1 = -\alpha\omega_1. \tag{5.5}$$

As $\omega$ and $\text{var}(U)$ can be consistently estimated by ordinary least squares, all parameters of the Weibull model can be estimated. To construct the Neyman $C(\alpha)$ version of the conditional moment test we use the expression (2.9), $LM_H^c[\tilde{\theta}] = s_1^{c\prime}(\tilde{\theta}) I^{11}(\tilde{\theta}) s_1^c(\tilde{\theta})$.

## 6. Monte Carlo experiments: Uncensored observations

*Statistics compared*: The Monte Carlo experiments are designed to throw light on the agreement between asymptotic and finite sample distributions when $H_0$ holds, and to allow comparison between the power properties of six tests viz., $LM_H$, $LM_D$, $LM_{HD}$, $LM_H^c[\hat{\theta}]$, $LM_D^c[\hat{\theta}]$, and $LM_H^c[\tilde{\theta}]$, defined as follows:

$LM_H$: separate test of heterogeneity,

$LM_D$: separate test of duration dependence,

$LM_{HD}$: joint test of heterogeneity and duration dependence,

$LM_H^c[\hat{\theta}]$: (improperly) conditioned test of heterogeneity

$LM_D^c[\hat{\theta}]$: (improperly) conditioned test of duration dependence,

$LM_H^c[\tilde{\theta}]$: $C(\alpha)$ or properly conditioned test of heterogeneity.

$LM_D^c[\tilde{\theta}]$ is not included because the expected information matrix required for computing the variance of the test was difficult to evaluate in this case.

*Design of sampling experiments*: Twelve models are used and each simulation experiment is based on 500 replications. In all experiments the parameters $(\beta_0, \beta_1)$ are fixed at $(-5.0, 1.0)$. The variable $x_1$ is taken as a random draw from a uniform [0, 1] distribution and held fixed for all experiments.

Weibull distributed $t_i$ are generated using the relation $\ln t_i = -M_i/\alpha + w_i/\alpha$, where $M_i = x_i'\beta$ and $w_i$ are i.i.d. random variables with $\text{pdf}(w) = \exp(w - e^w)$: in case of heterogeneity, $M_i = x_i'\beta + v_i$, and $v_i$ is a random draw from $\mathcal{N}(0, \sigma^2)$ distribution. To condition on a given vector $v$, the lognormal heterogeneity term should be held fixed across all replications of a given sample size. This means that one is conditioning the test statistic on exogenous variables and given heterogeneity; such conditioning also reduces sampling variability between alternatives [Jensen (1987)]. However, in a Monte Carlo this may lead to correlation between tests in different experiments. Hence, following the suggestion of an anonymous referee, for each replication we make independent draws of $v$.

The convolution of a Weibull with a lognormal heterogeneity distribution does not lead to a closed form. However, to give the reader some feel for the data-generating mechanism being used, it is convenient to replace the lognormal by gamma heterogeneity because a closed form is available for this case. (For some parameter values lognormal will be a good approximation to the gamma in any case.) For the Weibull–gamma mixture it is known that

Table 1

Actual rejection rate of $H_0$ at 5% nominal significance level for Models 1–1C.

| | $N = 50$ | $N = 100$ | $N = 200$ | $N = 500$ | $N = 200$ (censored) |
|---|---|---|---|---|---|
| $LM_H$ | 2.2 | 6.8 | 3.8 | 3.0 | 6.8 |
| $LM_D$ | 5.6 | 4.6 | 5.0 | 4.4 | 6.2 |
| $LM_{HD}$ | 3.4 | 4.8 | 4.4 | 3.4 | 7.8 |
| $LM_H^*[\hat{\theta}]$ | 3.2 | 3.4 | 3.0 | 2.6 | 7.2 |
| $LM_D^*[\hat{\theta}]$ | 4.2 | 5.0 | 4.2 | 5.2 | 6.2 |
| $LM_H^*[\tilde{\theta}]$ | 8.4 | 7.8 | 7.0 | 5.7 | NA |

$\mathbb{E}(t_i \mid x_i) = \exp(x_1'\beta/\alpha) \cdot B[\alpha^{-1}, \sigma^{-2} - \alpha^{-1}]/\alpha^{-1}\sigma^{-2\alpha}$, where $B[\ldots]$ is the beta function [see Lanacaster (1979, p. 952)]. It is easy to verify that $\mathbb{E}(t_i)$ is decreasing in $\alpha$ (given $\sigma^2$) and increasing in $\sigma^2$ (given $\alpha$). The difficulty of inferring whether the data are characterized by small $\alpha$ and large $\sigma^2$, or large $\alpha$ and small $\sigma^2$, is referred to as 'confounding'.

The twelve combinations of $(\sigma^2, \alpha)$ are used for the heterogeneous Weibull model. For Model 1 it is (0, 1.0); for the remaining models the combinations can be read off from the table 2. The data generation process for Model 1 is exponential and for Models 2–12 it is either Weibull ($\alpha \neq 1$) or exponential ($\alpha = 1$), with heterogeneity ($\sigma^2 > 0$) or without heterogeneity ($\sigma^2 = 0$). The same twelve experiments are repeated with Type I censoring, these being referred to as Models 1C–12C. For the uncensored exponential case $\mathbb{E}(t_i \mid x_i = \bar{x})$ is about 90 with the same standard deviation; for $\alpha = 1$, $\sigma^2 = 0.6$, it is 225, for $\alpha = 1$, $\sigma^2 = 0.8$, it is 450. For $\sigma^2 = 0$, the mean duration falls from 480 when $\alpha = 0.75$ to about 20 when $\alpha = 1.45$.

*The results*: For the correctly specified Model 1, the actual rejection frequencies based on the nominal 5% significance level are shown below in table 1 for $N = 50, 100, 200, 500$. In the absence of size distortion, this should be around 5% (taking account of Monte Carlo error) and it is close to that value. This was also found to be the case for 10% and 1% significance levels. The latter results are omitted from here to save space. Thus, all tests are satisfactory from a size consideration.

The upper half of table 2 contains the results for uncensored Models 2–12 and the lower half for censored Models 2C–12C. Models 2, 3, and 4 incorporate duration dependence (Weibull model) but not unobserved heterogeneity. Model 2 data are subject to negative duration dependence ($\alpha = 0.75$) and Model 3 and 4 data are subject to positive duration dependence ($\alpha = 1.3, 1.45$). Since there is zero heterogeneity in this case, the separate heterogeneity test should, ideally, show a rejection rate of about 5%. This, however, is not the case: $LM_H$ is unable to distinguish between duration dependence and neglected heterogeneity. The rejection rate of $LM_H$ is between 95% and 100% for all

Table 2

Percentage rejections of $H_0$ at 5% significance level for Models 2–12, 2C–12C; $N = 200$.

| | $\sigma^2 = 0.0$ | | | $\sigma^2 = 0.6$ | | | | $\sigma^2 = 0.8$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| $\alpha$ | 0.75 | 1.3 | 1.45 | 1.00 | 0.75 | 1.3 | 1.45 | 1.0 | 0.75 | 1.3 | 1.45 |
| $LM_H$ | 96.0 | 95.0 | 100.0 | 99.0 | 100.0 | 33.2 | 15.2 | 99.6 | 100.0 | 63.6 | 22.2 |
| $LM_D$ | 100.0 | 99.6 | 100.0 | 99.2 | 100.0 | 7.4 | 54.2 | 99.8 | 100.0 | 23.8 | 23.4 |
| $LM_{HD}$ | 99.8 | 98.8 | 100.0 | 99.0 | 100.0 | 49.6 | 78.2 | 99.8 | 100.0 | 67.0 | 70.2 |
| $LM_H^*[\hat{\theta}]$ | 35.6 | 0.0 | 0.0 | 85.8 | 95.8 | 55.6 | 41.2 | 93.4 | 98.8 | 72.0 | 60.0 |
| $LM_D^*[\tilde{\theta}]$ | 83.2 | 87.8 | 99.6 | 46.8 | 80.6 | 48.2 | 83.2 | 58.4 | 87.6 | 47.6 | 79.0 |
| $LM_H^*[\tilde{\theta}]$ | 8.6 | 8.8 | 6.8 | 78.6 | 76.2 | 79.6 | 80.0 | 86.2 | 86.2 | 84.8 | 87.0 |
| Model | 2C | 3C | 4C | 5C | 6C | 7C | 8C | 9C | 10C | 11C | 12C |
| $\alpha$ | 0.75 | 1.3 | 1.45 | 1.00 | 0.75 | 1.3 | 1.45 | 1.0 | 0.75 | 1.3 | 1.45 |
| $LM_H$ | 90.6 | 92.8 | 99.6 | 76.4 | 100.0 | 10.0 | 51.2 | 89.6 | 100.0 | 6.4 | 25.8 |
| $LM_D$ | 98.0 | 98.2 | 100.0 | 69.8 | 100.0 | 39.4 | 91.0 | 89.4 | 100.0 | 19.2 | 76.6 |
| $LM_{HD}$ | 93.8 | 95.8 | 99.6 | 67.4 | 100.0 | 52.2 | 90.6 | 84.4 | 100.0 | 42.8 | 86.2 |
| $LM_H^*[\hat{\theta}]$ | 4.4 | 27.4 | 42.4 | 24.2 | 38.0 | 37.8 | 57.6 | 28.4 | 54.4 | 45.8 | 62.8 |
| $LM_D^*[\hat{\theta}]$ | 44.8 | 70.4 | 90.4 | 2.6 | 46.0 | 58.2 | 87.4 | 5.2 | 52.2 | 50.2 | 86.6 |

values of $\alpha$. Thus the separate test for zero unobserved heterogeneity is very misleading. For the Weibull model the conditional mean of $t$ depends upon $\alpha$, and hence the imposition of the restriction $\alpha = 1$ leads to a model with misspecified first conditional moment. Consequently, if one tests $H_0$: $\sigma^2 = 0$, the resulting separate test cannot have the correct asymptotic size. The conditional score test based on $\hat{\theta}$, that are intended to make at least a partial adjustment for the misspecified first moment, are somewhat more informative. $LM_H^c[\hat{\theta}]$ should show a 5% rejection rate, but it is 35.6% for $\alpha = 0.75$ and 0% for $\alpha = 1.30$ and 1.45. Similarly, even though one would ideally expect a rejection rate close to 95% for $LM_H^c[\hat{\theta}]$, it actually equals 83.7%, 87.8%, and 99.6% for $\alpha = 0.75$, 1.30, and 1.45, respectively. The properly conditioned test, $LM_H^c[\tilde{\theta}]$, should perform better than $LM_H^c[\hat{\theta}]$, and it does. Though the conditional test based on $\hat{\theta}$ is theoretically incorect, in particular cases it performs similarly to the test based on $\tilde{\theta}$ and almost always outperforms the incorrect separate test $LM_H$.

Now consider the case of Models 5 and 9, where there is unobserved heterogeneity but no duration dependence, though $LM_H^c[\hat{\theta}]$ still has a high rejection probability, unfortunately $LM_D^c[\hat{\theta}]$ incorrectly identifies duration dependence in 46.8% and 58.4% of the cases, reflecting the high correlation between the test procedures.

Models 7, 8, 11, and 12 have heterogeneity and positive duration dependence. Here $LM_H$ has low rejection probability [Jensen (1987) reports a similar result]. The joint test $LM_{HD}$ has relatively higher power which increases with the magnitude of $\alpha$. The conditional tests based on $\hat{\theta}$ also suffer a reduction in power for Models 5 and 8 compared with Models 9 and 12, respectively. By contrast, $LM_H^c[\tilde{\theta}]$ is robust and has higher power in all cases. For all values of $\alpha$, power increases when $\sigma^2$ is raised from 0.6 to 0.8. Further, even in situations where a test such as $LM_H$ has low power, one of the two conditional tests based on $\hat{\theta}$ can have quite high power (Models 8 and 12). Nevertheless, it is problematic that there are parameter configurations in which all score tests based on the restricted maximum likelihood estimator have rather low rejection probability.

*To summarize*: Separate tests of heterogeneity and duration dependence, being correlated to an extent that is model-dependent, are potentially very misleading. The same is true to a much lesser extent for the conditional score tests based on restricted maximum likelihood. The joint test is less misleading than the separate test but its performance is also model-dependent, being most unsatisfactory when positive duration dependence and unobserved heterogeneity occur simultaneously. The best test is the $C(\alpha)$ test. For the practitioner, conditioning on restricted maximum likelihood is convenient; the Monte Carlo results suggest that this is generally preferrable to the separate tests.

## 7. Monte Carlo experiments: Censored observations

Since in practice most samples include (right-)censored observations of $t_i$, we investigate the performance of the above test procedures to censoring. Currently no clear conclusions on the sensitivity of conventional diagnostic tests to censoring are available in the literature [Neumann and Horowitz (1989)]. Tests developed for uncensored data can be expected to retain their properties for low degrees of (Type I) censoring. An important practical issue is whether the performance rankings for the uncensored case carry over to the censored case.

A difficulty in the application of score tests to censored data arises from the fact that the expected Fisher information matrix depends upon the censoring mechanism with unknown parameters, thereby requiring additional assumptions if it is to be used. One approach is to use the OPG form of the information matrix, or the sample Hessian of the log-likelihood. Efron and Hinkley (1978) advocate the use of the sample information matrix as an estimate of the unknown expected Fisher information matrix. We use the OPG estimator.

*Design of sampling experiments*: We use Type I censoring in our experiments, which can be induced in practice by a finite observation period. In this case $t_i = \min(T_i, L)$, where $T_i$ are completed durations and $L$ is the censored duration, $T_i \leq L$. The same twelve experiments reported in the previous section are run with a fixed censoring point $L$, chosen by trial and error, to yield about 20–22% censored observations. Other features of the sampling design are left unchanged to facilitate comparisons with the uncensored case. Since we do not have the regression equivalent of the censored Weibull model, only five tests are compared; the $C(\alpha)$ test is omitted. The experiments 2–12 summarized in the upper half of table 2 are renumbered 2C to 12C in the lower half, where the suffix C indicates censoring.

The last column in table gives the empirical size of the five tests at 5% nominal significance level. In comparing it to the corresponding figure in column 4, the reader is reminded that the differences reflect partly the pure effect of loss of information from censoring and partly the effect of using the OPG estimator of the information matrix. A difference from the uncensored case is that there is a slight tendency towards overrejection of the true null in the censored case.

In Models 2C, 3C, and 4C, where the data are generated by models with duration dependence but *no* heterogeneity, there is still a tendency for $LM_H$ to overwhelmingly reject the true null, as was the case for (uncensored) Models 2, 3, and 4 earlier. Though, as before, the conditional tests fare relatively better than the corresponding separate tests, the performance of the $LM_H^c$ deteriorates as $\alpha$ increases. The rejection rate for the false null goes from 4.4% at $\alpha = 0.75$ to 42.4% for $\alpha = 1.45$. The corresponding figures for $LM_H$ are 90.6% and 99.6%,

respectively. A comparison with table 2, upper half, shows that there is a marked deterioration of performance of $LM_D^c[\hat{\theta}]$ in the censored case.

For the heterogeneous Models 7C and 8C, where $\sigma^2 = 0.6$, and Models 11C and 12C, where $\sigma^2 = 0.8$, we again observe the underrejection of the false null of $\sigma^2 = 0$ (the 'cancellation phenomenon') which was discussed earlier with Proposition 2. However, for $\alpha = 1.45$ the tendency towards underrejection is *less* marked than in the uncensored case. Further, the conditional tests are not unambiguously better than the separate tests. The confounding effect due to the joint presence of heterogeneity and duration dependence is present, but in the censored case it affects separate and conditional tests in different ways for different parameter values.

*To summarize*: Standard specification tests applied to censored data, after approximating the expected Fisher information matrix by the OPG matrix, will be frequently misleading. However, for the separate as well as the joint tests, the performance rankings of the tests are comparable to the uncensored case. The fact that (improperly) conditioned tests show a sharper deterioration in performance suggests that they may be relatively more sensitive to the use of the OPG estimator. Better estimators of the information matrix are required to reveal the pure effect of censoring on the performance of the tests.

## 8. Monte Carlo experiments: The case of lognormal and gamma hazards

It is useful to have tests with good power properties against a variety of distributional alternatives. For example, from an empirical perspectice it is desirable that specification tests for a heterogeneous Weibull model work well even if the data come from (say) heterogeneous lognormal or heterogeneous gamma populations. In this section additional Monte Carlo experiments are reported which help to evaluate whether this holds for the tests developed in earlier sections.

We evaluate the power of the joint, conditional, and separate tests developed for the heterogeneous Weibull alternative when the true alternative is either the heterogeneous lognormal or the heterogeneous gamma. Unlike the Weibull model, the lognormal model has a nonmonotone hazard in general; the gamma model has a monotone hazard function.

To generate the data for the lognormal model we use the relation $\ln t_i = -M_i + \rho W_i$, where $W_i$ is an independent draw from a $\mathcal{N}(0, 1)$ and $M_i = x_i'\beta$. To allow for unobserved heterogeneity we add the term $v_i$ to $M_i$, where $v_i$ is a random draw from $\mathcal{N}(0, \sigma^2)$ distribution. The shape of the hazard function for the lognormal depends on $\rho$, larger values yielding a more clearly defined nonmonotone form.

Table 3

Percentage rejections of $H_0$ at 5% significance level for Models 1HL–8HL; $N = 200$; true data-generating process: heterogeneous lognormal.

| | $\sigma^2 = 0.0$ | | | | $\sigma^2 = 0.6$ | | | |
|---|---|---|---|---|---|---|---|---|
| Model | 1HL | 2HL | 3HL | 4HL | 5HL | 6HL | 7HL | 8HL |
| $\rho$ | 1.00 | 0.90 | 0.80 | 0.70 | 1.00 | 0.90 | 0.80 | 0.70 |
| $LM_{\mathrm{H}}$ | 66.2 | 30.0 | 34.2 | 83.2 | 99.6 | 98.2 | 93.4 | 82.0 |
| $LM_{\mathrm{D}}$ | 15.0 | 64.6 | 98.4 | 100.0 | 97.2 | 86.6 | 50.6 | 23.8 |
| $LM_{\mathrm{HD}}$ | 98.6 | 99.8 | 100.0 | 100.0 | 99.8 | 98.8 | 97.2 | 99.0 |
| $LM_{\mathrm{H}}^{\mathrm{c}}[\hat{\theta}]$ | 97.2 | 94.6 | 90.4 | 86.2 | 99.8 | 99.4 | 97.8 | 98.8 |
| $LM_{\mathrm{D}}^{\mathrm{c}}[\hat{\theta}]$ | 99.4 | 100.0 | 100.0 | 100.0 | 84.2 | 89.2 | 92.0 | 98.0 |

Table 4

Percentage rejections of $H_0$ at 5% significance level for Models 1HG–6HG; $N = 200$; true data-generating process: heterogeneous gamma.

| | $\sigma^2 = 0.0$ | | | $\sigma^2 = 0.6$ | | |
|---|---|---|---|---|---|---|
| Model | 1HG | 2HG | 3HG | 4HG | 5HG | 6HG |
| $\kappa$ | 0.60 | 2.00 | 3.00 | 0.60 | 2.00 | 3.00 |
| $LM_{\mathrm{H}}$ | 93.8 | 100.0 | 100.0 | 100.0 | 75.8 | 46.2 |
| $LM_{\mathrm{D}}$ | 100.0 | 100.0 | 100.0 | 100.0 | 22.4 | 24.6 |
| $LM_{\mathrm{HD}}$ | 100.0 | 100.0 | 100.0 | 100.0 | 87.6 | 98.2 |
| $LM_{\mathrm{H}}^{\mathrm{c}}[\hat{\theta}]$ | 56.2 | 2.0 | 2.4 | 79.6 | 91.8 | 92.2 |
| $LM_{\mathrm{D}}^{\mathrm{c}}[\hat{\theta}]$ | 99.2 | 100.0 | 100.0 | 79.4 | 83.2 | 99.6 |

To generate the gamma distributed durations with heterogeneity we use the relation $\ln t_i = -M_i + W_i$, where $W_i$ are i.i.d. random variables with $\mathrm{pdf}(W) = \exp(\kappa w - \mathrm{e}^w)/\Gamma(\kappa)$, $\kappa > 0$. To introduce unobserved lognormal heterogeneity we proceed as before. The gamma hazard function does not have a closed form but is known to be montone decreasing for $\kappa < 1$ and monotone increasing for $\kappa > 1$ [Kalbfleisch and Prentice (1980, ch. 2)].

The results are given in table 3 for the heterogeneous lognormal, identified by the suffix HL, and in table 4 for the heterogeneous gamma, identified by the suffix HG. They show that the confounding of heterogeneity and duration dependence remains a serious problem. From table 3, where $\sigma^2 = 0$, it is seen that both $LM_{\mathrm{H}}$ and $LM_{\mathrm{H}}^{\mathrm{c}}[\hat{\theta}]$, especially the latter, have a high rejection rate of the true null when $\rho = 1$. Tests of heterogeneity developed for an alternative with monotone hazards overrejects the true null when the correct alternative has nonmonotone hazards; the confounding effect is even worse in this case.

For testing duration dependence, however, the conditional test is better. The separate test $LM_{\mathrm{D}}$ rejects the false null of zero duration dependence in only 15% of the cases when $\rho = 1$, whereas $LM_{\mathrm{D}}^{\mathrm{c}}[\hat{\theta}]$ does so in over 99% of the cases.

Considering $\rho = 0.9, 0.8$, and 0.7, it is seen that the tests give a better indication of duration dependence for smaller values of $\rho$, but a generally poor indication of unobserved heterogeneity. When $\rho = 0.7$, the hazard function is closer to being monotone increasing, whereas it is more like an inverted 'U' shape when $\rho = 0.9$ or 0.8. When we have both heterogeneity and duration dependence, the conditional tests again improve significantly on the unconditional ones; the superiority of $LM_{\mathfrak{H}}$ over $LM_{\mathrm{D}}$ is especially marked.

The general pattern of the results for the gamma alternative given in table 4 is similar to that obtained when the data were indeed generated by the heterogeneous Weibull; compare tables 2 and 4. The conditional test generally performs better than the unconditional test and the joint test retains high power. The similarity of the hazard function in the two cases is the likely reason for this.

## 9. Summary and conclusion

This paper motivates and exposits the conditional score, including $C(\alpha)$, tests as useful alternatives to several separate and joint score tests, by reference to a model of heterogeneity and duration dependence. By a detailed theoretical and Monte Carlo investigation of the non-null distribution of a test of duration dependence and unobserved heterogeneity, we show that when different tests are asymptotically correlated there is a serious problem of size distortion and of confounding of the source of misspecification. Hence, the separate tests are unreliable and potentially misleading.

The tests investigated are developed for the case of uncensored duration data with an exponential null and a heterogeneous Weibull alternative. However, the Monte Carlo results show that our general results remain valid even when durations are censored and alternatives are not Weibull.

As a result of confounding, the heterogeneity tests based on the exponential null and the Weibull alternative appear to overreject the null. This highlights the known difficulty of selecting between a model with a more flexible hazard function and no heterogeneity and one in which the hazard function specification is less flexible but the model incorporates heterogeneity. An example is the exponential model with exponential heterogeneity, which generates the same reduced form as the log-logistic model with no heterogeneity. Good *a priori* information about the form of the hazard would help to solve the identification problem.

Monte Carlo results suggest that in general the conditional tests have greater power than separate tests, but their performance is case-dependent, especially with censored data. Further, a ranking of their relative power properties is ambiguous when the true hazard function is not monotone, but tests are developed for a monotone alternative. In contrast, the joint test has excellent

power properties against non-Weibull alternatives. Thus, the evidence suggests that conditional and joint score tests are to be preferred to the separate tests.

For the empirical researcher we recommend the strategy of specifying the most general data-coherent functional form for the hazard function (about which economic theory is in any case likely to be more informative) that is computationally feasible, before testing for heterogeneity. If several misspecification tests are to be applied, we recommend conditional tests in preference to separate tests. The paper reinforces the argument that the rejection of the null does not imply the acceptance of the alternative.

## References


Bera, A.K. and C.R. McKenzie, 1987, Additivity and separability of the Lagrange multiplier, likelihood ratio, and Wald tests, Journal of Quantitative Economics 3, 55–63.

Breusch, T.S. and A.R. Pagan, 1980, Lagrange multiplier test and its applications to model specification in econometrics, Review of Economic Studies 47, 239–253.

Burdett, K., N.M. Kiefer, and S. Sharma, 1985, Layoffs and duration dependence in a model of turnover, Journal of Econometrics 28, 51–62.

Efron, B. and D.V. Hinkley, 1978, Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, Biometrika 65, 457–482.

Engle, R.F., 1984, Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, Ch. 13 in: Z. Griliches and M. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam).

Godfrey, L.G., 1988, Specification testing in econometrics (Cambridge University Press, Cambridge).

Holly, A., 1987, Specification tests: An overview, Ch. 2 in: T. Bewley, ed., Advances in econometrics – Fifth World Congress, Vol. I (Cambridge University Press, Cambridge).

Horowitz, J.L. and G.R. Neumann, 1989, Specification testing in censored regression models: Parametric and nonparametric methods, Journal of Applied Econometrics 4, S61–S86.

Jaggia, S., 1991, Specification tests based on the heterogeneous generalized gamma model of duration: With an application to Kennan's strike data, Journal of Applied Econometrics 6, 169–180.

Jensen, P., 1987, Testing for unobserved heterogeneity and duration dependence in econometric duration models, Unpublished paper (University of Aarhus, Aarhus).

Kalbfleisch, J.D. and R.L. Prentice, 1980, The statistical analysis of failure time data (Wiley, New York, NY).

Kiefer, N.M., 1985, Specification diagnostics based on Laguerre alternatives for econometric models of duration, Journal of Econometrics Annals 28, 135–154.

Lancaster, T., 1979, Econometric methods for the duration of unemployment, Econometrica 47, 939–956.

Lancaster, T., 1985, Generalized residuals and heterogeneous duration models: With applications to the Weibull model, Journal of Econometrics 28, 155–169.

Moran, P.A.P., 1970, On asymptotically optimal tests of composite hypothesis, Biometrika 57, 47–55.

Neyman, J., 1959, Optimal asymptotic tests of composite statistical hypothesis, in: U. Grenander, ed., Probability and statistics (Wiley, New York, NY) 213–234.

Sharma, S., 1987 Specification diagnostics for econometric models of duration, Unpublished working paper (University of California, Los Angeles, CA).