Testing for Independence in Multivariate Duration Models.*

Hans Söderberg† and Johan Lyhagen‡

Working Paper Series in Economics and Finance, No. 302
Stockholm School of Economics

February, 1999

Abstract

Modelling multivariate failure times in a competing risks setting is often performed by assuming independence between risks. However, by wrongly assuming independence, seriously biased parameter estimates may result. The aim of this paper is to evaluate a test for independence previously proposed in the literature. The test is an information matrix test which is evaluated by Monte Carlo methods. The basic duration model used in the simulations is of the mixed proportional hazards type with Weibull distributed latent failure times.

We find the sampling distribution of the test statistic to deviate from the anticipated asymptotic one. However, by applying bootstrap methods it seems that proper critical values can be obtained. Further, the power of the test is found to be highly asymmetrical with respect to the sign of the correlation between risks.

---

*Reinhold Bergström provided helpful comments on an earlier draft. Financial support to Lyhagen from the Tore Browaldhs Foundation for Scientific Research and Education is gratefully acknowledged.

†Department of Information Science, Division of Statistics, Uppsala University, P.O. Box 513, SE-751 20 Uppsala, Sweden. Tel.: +46 18 471 1149; fax: +46 8 18 55 44 22; E-mail: Hans.Soderberg@statistik.uu.se

‡Department of Economic Statistics, Stockholm School of Economics, P.O. Box 6501, S-113 83 Stockholm, Sweden.Tel.:+46 8 736 9232; fax: E-mail: Stjl@hhs.se.

1. Introduction

Duration models are often used in micro economic studies of unemployment spells. The data usually includes information on different exit destinations which constitute possibly dependent competing risks. Nevertheless, many applied studies assume independence between such risks, see for instance Han and Hausman (1990), Narendranathan and Stewart (1993) and Carling et al. (1996). It seems likely, though, that the assumption of independence stems more often from convenience rather than solid knowledge about the true dependence structure. The consequence of not properly account for the dependence being that heavily biased parameter estimates may result. However, some studies do explicitly specify models to capture possible dependence between risks, e.g. Heckman and Walker (1990) and Van den Berg et al. (1994).

Dependence between competing risks can be modelled in several ways (Hougaard, 1987). One such approach, usually applied by econometricians, is to assume that only a subset of relevant explanatory variables is observed. The unobserved set of covariates are in the econometric literature mostly referred to as unobserved heterogeneity while biometricians more often use the term frailty. Dependence between risks is introduced by assuming that the unobserved heterogeneity components are dependent across risks. Consequently, it is possible to construct a test for dependence between risks by examine the dependence structure between unmeasured heterogeneity components.

Lancaster (1990) suggests one such test where the resulting test statistic turns out to be an Information Matrix Test (IMT) statistic (The IMT test procedure was developed by White, 1982). The test statistic is a function of generalized residuals defined in the Cox and Snell (1968, 1971) sense. However, as noted by Lancaster (1990), the properties of such a test based on generalized residuals is not well understood.

The aim of this is paper is to evaluates the performance of the test proposed by Lancaster (1990). Our method for evaluation is based on Monte Carlo simulations. The basic duration model used in the simulations is of the mixed proportional hazards type, with cause specific failure times specified as Weibull distributed.

We find that the sampling distribution of the test statistic heavily departs from the anticipated chi-square distribution. Nevertheless, by applying bootstrap methods we are able to correct the reference distribution and thereby obtain a test size close the nominal one. Our main finding is that the test has a highly asymmetrical power. This is in accordance with a related study in the area performed

by Carling (1996).

The paper is organized as follows. In the next section the duration model is introduced while the third section presents the IMT and its version applied. The fourth section present the results from the Monte Carlo study. Finally, the fifth section closes the paper by a discussion.

2. The Duration Model

The basic duration model used in the simulations is a mixed proportional hazards model, with latent failure times following the Weibull distribution. Several authors, e.g. Van den Berg et al. (1994) and Carling (1996), have previously applied the same model. In this paper we focus on two latent failure times, denoted t_1 and t_2, with associated matrix of observed explanatory variables $\mathbf{x}$. Related to each latent failure time is an unobserved heterogeneity component, assumed independent of $\mathbf{x}$, and symbolized by V_1 and V_2 respectively. V_1 and V_2 are introduced in the duration model as multiplying the hazard function, which imposes a non-negativity restriction upon them. The cause specific conditional hazard functions can be written as

$$\begin{aligned} h\left(t_1|\mathbf{x}, v_1, \theta_1\right) &= \alpha_1 t_1^{\alpha_1 - 1} v_1 e^{\mathbf{x}\beta_1} \\ h\left(t_2|\mathbf{x}, v_2, \theta_2\right) &= \alpha_2 t_2^{\alpha_2 - 1} v_2 e^{\mathbf{x}\beta_2} \end{aligned} \tag{2.1}$$

where $\theta_i = \left(\alpha_i, \beta_i\right)^{'}$ $i = 1, 2$ are vectors of unknown parameters. If $\mathbf{x}$ includes a constant term, the mean of V_1 and V_2 can be normalized to one without loss of generality as any departures from mean one will be captured by the constant. The basic assumption is that $t_1|\mathbf{x}$ and $t_2|\mathbf{x}$ are related only through the dependence structure between V_1 and V_2. The bivariate conditional density of the latent failure times, t_1 and t_2, will in the following be denoted $f\left(t_1, t_2|\mathbf{x}, \theta, v_1, v_2\right)$. This model can also be written in the form of a simultaneous regression model as below

$$\begin{aligned} \alpha_1 \ln\left(t_1\right) &= -\beta_{10} - \mathbf{x}^*\beta_1^* - \ln\left(V_1\right) + \varepsilon_1, \\ \alpha_2 \ln\left(t_2\right) &= -\beta_{20} - \mathbf{x}^*\beta_2^* - \ln\left(V_2\right) + \varepsilon_2. \end{aligned} \tag{2.2}$$

In (2.2) β_{10} and β_{20} represent the intercepts in the regression vectors while ε_1 and ε_2 are independent Extreme Value distributed errors (Kiefer, 1988). Suppose now

that a researcher specify the following (incorrect) model

$$\begin{aligned} \alpha_1 \ln(t_1) &= -\beta_{10} - \mathbf{x}^* \beta_1^* + \varepsilon_1, \\ \alpha_2 \ln(t_2) &= -\beta_{20} - \mathbf{x}^* \beta_2^* + \varepsilon_2 \end{aligned} \tag{2.3}$$

instead of the true model in (2.2). From the representation in (2.2), it can be seen that this misspecification will introduce random variation into the constant terms. That is, the true intercepts consists of the random quantities $\beta_{10}^* = \beta_{10} + \ln(V_1)$ and $\beta_{20}^* = \beta_{20} + \ln(V_2)$ instead of the β_{10} and β_{20}. Moreover, if the variation in β_{10}^* and β_{20}^* is dependent, the latent failure times will be dependent.

It can be noted that by wrongly basing the specification on (2.3) instead of (2.2), two kinds of misspecification can occur. First, as each separate cause specific risk is misspecified, all parameter estimates generally will be biased towards zero (Lancaster, 1990). Secondly, if the unobservables is dependent, additional bias will be inflicted upon parameter estimates. The aim of this paper is only to test for the second type of misspecification. Our general strategy is to wrongly basing the likelihood on (2.3) and then apply the IMT to test for dependence. Generally our specification will suffer from the first type of problem, however, the applied version of the IMT is only intended to detect the second one.

3. The Information Matrix Test

The IMT was first proposed by White (1982) and builds upon a result from standard likelihood theory, namely the information matrix equality. Loosely speaking, the equality states that in a correctly specified model the Hessian of the likelihood function and the corresponding outer product of the gradient vector should asymptotically be equivalent. This is generally not true, however, for a misspecified model. Following the notation of White (1982), define the following matrices:

$$\begin{aligned} A_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta\partial\theta'}, \\ B_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta} \frac{\partial \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta'}, \end{aligned}$$

where n is the sample size and $f(t_1, t_2|\mathbf{x}, \theta)$ the relevant density function depending on the p dimensional parameter vector θ. The IMT investigates the statistical

significance of departures from $A_n(\boldsymbol{\theta}) + B_n(\boldsymbol{\theta}) = 0$. As $A_n(\boldsymbol{\theta}) + B_n(\boldsymbol{\theta})$ is a symmetric matrix it has at most $q = p(p+1)/2$ unique non-zero elements. These unique non-zero elements are, by using White's terminology, called indicators. The most general form of the test is based on all q indicators, but a more directional test for misspecification can be obtained if a specific subset is used.

To apply the test correctly it is necessary to check that included indicators are not identically equal to zero due to the specification of the likelihood. Such an identity may in some cases be difficult to detect because of algebraic complexity and/or numerical imprecision. Basing the test on indicators that actually are zero would mean that the probability of detecting misspecification decreases. Nor should linearly dependent indicators be used.

The test can be outlined as follows. In the most general form of the test we would use all $q = p(p+1)/2$ indicators and arrange them in the $q \times 1$ vector

$$\begin{aligned} D_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta_j \partial\theta_h} + \frac{\partial \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta_j}\frac{\partial \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta_h}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} d_r(t_{1i}, t_{2i}|\mathbf{x}_i, \theta) \end{aligned}$$

where $r = 1, ..., q$, $j = 1, ..., p$, $i = j, ..., p$. Other matrices to define are the $q \times p$ vector

$$\bigtriangledown D_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial d_r(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta}$$

an the $p \times 1$ gradient vector

$$l(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(t_{1i}, t_{2i}|\mathbf{x}_i, \theta)}{\partial\theta}.$$

In a correctly specified model White showed that the asymptotic sampling distribution of $\sqrt{n}D_n\left(\widehat{\theta}\right)$, where $\widehat{\theta}$ equals any consistent estimate of θ, is multivariate normal with mean zero and asymptotic covariance matrix $V(\theta)$. The asymptotic covariance matrix, $V(\theta)$, may be estimated by several consistent estimators. One such estimator is given by

$$\widehat{V}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[w_i\left(\widehat{\theta}\right)w_i\left(\widehat{\theta}\right)'\right] \tag{3.1}$$

where

$$w_i\left(\widehat{\theta}\right) = d_i\left(\widehat{\theta}\right) - \bigtriangledown D_n\left(\widehat{\theta}\right)\left[A_n\left(\widehat{\theta}\right)\right]^{-1} l_i\left(\widehat{\theta}\right).$$

Equation (3.1) is the original estimator proposed by White, other versions of the test are obtained by replacing $D_n\left(\widehat{\theta}\right)$ and $A_n\left(\widehat{\theta}\right)$ by other asymptotically equivalent estimators. Using the quantities defined above (and supposing q indicators) the test is performed by computing the IMT statistic ϖ as

$$\varpi_q = nD_n\left(\widehat{\theta}\right)'\left[\widehat{V}\left(\widehat{\theta}\right)\right]^{-1} D_n\left(\widehat{\theta}\right)$$

which has an asymptotic chi-square sampling distribution with q degrees of freedom (χ^2_q).

As noted above, unobserved heterogeneity can be interpreted as if the intercepts in the regression vectors, incorrectly, has been specified as fixed instead of random quantities. Moreover, dependence between latent failure times follow if these random quantities are correlated. Consequently, the indicator of interest in a test for independence between risks, is based on the cross-derivative term between the intercepts.

Finally, by noting that only $t = \min(t_1, t_2)$ is observed the relevant log likelihood function, in the absence of unobservables, equals

$$\ln L = \sum_{i=1}^{n} \ln\left[\left(\alpha_1 t_i^{\alpha_1 - 1} e^{\mathbf{x}_i\beta_1}\right)^{c_i}\left(\alpha_2 t_i^{\alpha_2 - 1} e^{\mathbf{x}_i\beta_2}\right)^{(1-c_i)} e^{-t_i^{\alpha_1} e^{\mathbf{x}_i\beta_1} - t_i^{\alpha_2} e^{\mathbf{x}_i\beta_2}}\right]$$

where $c = I(t_1 < t_2)$. The indicator of interest is then calculated as

$$\begin{aligned} D_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} d_{ri}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 \ln L_i}{\partial\beta_{10}\partial\beta_{20}} + \frac{\partial \ln L_i}{\partial\partial\beta_{10}}\frac{\partial \ln L_i}{\partial\beta_{20}}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(t_i^{\alpha_1} e^{\beta_{10} + \mathbf{x}_i\beta_1} - c_i\right)\left(t_i^{\alpha_2} e^{\beta_{02} + \mathbf{x}_i\beta_2} - (1 - c_i)\right) \end{aligned} \tag{3.2}$$

which can be shown to have expectation zero in the absence of unobservables. However, if the specification is contaminated with unobserved heterogeneity the expectation of (3.1) no longer needs to be zero, unless V_1 and V_2 is independent.

As mentioned above the asymptotic sampling distribution of the IMT statistic, ϖ, is chi-square with degrees of freedom equal to the number of indicators. However, several papers have indicated that the IMT may have a size problem as it tends to reject a true null hypothesis too often, unless the sample size is large. See Kennan and Neuman (1987); Orme (1987); Chesher and Spady (1988). Horowitz (1994), suggests the use of bootstrap based critical values instead of asymptotic

ones as a solution to the small sample size problem. In agreement with the authors above, we also find the IMT test to reject the true null hypothesis too often. However, this effect seems not to be associated with the small sample problem. Nevertheless, our solution is, in accordance with Horowitz, to rely on bootstrap based critical values.

4. Experimental design

As the aim of this paper is to evaluate the performance of the IMT, the focus naturally is on the size and the power of the test. To judge the power of the test it is desirable to have a benchmark to which the IMT could be compared. Furthermore, a comparable test should ideally be of equal generality as the IMT. A previous study performed by Carling (1996) can be used for this purpose. Carling (1996) applies the mass-point technique to test for independence between two latent failure times. His primary target is to study the asymptotic behavior of the non parametric Maximum Likelihood estimator (Heckman and Singer, 1984). In addition, he also provides size corrected power estimates which we use as a benchmark for the IMT. As Carling (1996) is chosen as a benchmark for power comparisons, the experimental design in this study is made to closely follow the one used in his paper. However, the studies differ somewhat in the sample sizes applied.

The data generating process can be described as follows. The matrix of explanatory variables, $\mathbf{x}$, is defined as $n \times 3$ where the first column consists of a vector of ones. The remaining columns contains two variables drawn from the bivariate Normal distribution with variance equal to five, mean equal to zero and the correlation coefficient, ρ, set equal to 0.3. Referring to (2.2) the heterogeneity components $W_i = \ln(V_i)$, $i = 1, 2$, are drawn from the bivariate standard Normal distribution, $N(0, 1)$, with correlation coefficient chosen from the set $\rho^* \in (-0.6, -0.2, 0, 0.2, 0.6)$. New heterogeneity components, explanatory variables and errors, ε, are generated in each replicate. The parameter vectors are set to $\theta_1 = (\alpha_1, \beta_{01}, \beta_{11}, \beta_{12})' = (1.1, 0, 0.5, 0.5)'$ for the first failure time and to $\theta_2 = (\alpha_2, \beta_{20}, \beta_{21}, \beta_{22}) = (0.9, 0, 0.8, -0.3)$ for the second failure time. This setup yields $prob(t_1 < t_2)$ to be approximately equal to 0.5. By using (2.2) the latent failure times are generated as $t_{ij} = \exp((-\mathbf{x}_j\beta_i - w_{ij} + \varepsilon_{ij})/\alpha_i)$, $i = 1, 2$, $j = 1, .., n$. Finally, the observed failure time, t, and the censoring indicator, c, is obtained as $t = \min(t_1, t_2)$ and $c = I(t_1 < t_2)$. The sample sizes considered are 100, 500, 1000 and 2000. The experiments cover all combinations of these sample

sizes and ρ^*. Throughout, the number of replicates used are 1000, except otherwise stated. Finally, when critical values are obtained by bootstrap methods we use 500 bootstrap samples from each replicate data set to estimate the 90:th, 95:th and 99:te percentile of the empirical distribution. Estimates of the percentiles are obtained by ordering the bootstrap statistics as $\widehat{\varpi}_{boot_1} < ... < \widehat{\varpi}_{boot_{500}}$ and using $\widehat{\varpi}_{boot_{451}}$, $\widehat{\varpi}_{boot_{476}}$ and $\widehat{\varpi}_{boot_{496}}$ as percentile estimates respectively.

The bootstrap samples are generated by first computing residuals in the Cox and Snell (1968, 1971) sense from the replicate data set in question as

$$\widehat{e}_{ij} = 1 - \exp(-t_{ij}^{\widehat{\alpha}_i} e^{\mathbf{x}_j \widehat{\beta}_i}) \tag{4.1}$$

where $i = 1, 2$ defines each cause specific risk and $j = 1, ..., n$. By solving (4.1) for t_{ij} and inserting a resampled residual coming from the relevant cause specific risk we construct a bootstrap data set. The observed failure time is again obtained by $\min(t_1, t_2)$ and the censoring indicator, c, is defined in the same manner as when the replicate data set was created. The basic idea behind this resampling procedure is that if the latent failure times are dependent, it is mirrored in the residuals. By randomly pairing resampled residuals from the two cause specific risk the dependence structure will be eliminated.

Details of the bootstrap method applied are now given. The cause specific residuals e_1 and e_2 are assumed to be *iid* random variables having cumulative distribution functions F_1 end F_2 respectively. As we have an incomplete data set due to censoring the resampling procedure only resample uncensored residuals. The problem here is to assign the proper resampling probabilities to the uncensored residuals. One possible way to solve this problem is to estimate F_1 and F_2 by $\widehat{F}_i = 1 - \widehat{S}_i$, $i = 1, 2$, where $\widehat{S}_i$ equals some estimate of the associated survival function. $\widehat{S}_i$ are in this paper obtained by the Nelson-Aalen estimator (for a discussion of this estimator, see for instance Andersen et al., 1993)[1].

To be able to estimate F_i we need to distinguish between complete and censored observations. In the following we denote each uncensored residual by e_{ij}^*, where each marginal has r_i such residuals. It follows that $n = r_1 + r_2$. We define $Y_i\left(e_{ij}^*\right)$ as the number of observations larger than or equal to e_{ij}^*, i.e. the size of

[1] Another possible estimator is the Kaplan-Meier estimator (Kaplan and Meier, 1958). However, some informal experiments showed that the Nelson-Aalen estimator had a superior performance.

the risk set at e_{ij}^*. $\widehat{F}_i$ is based on the Nelson-Aalen estimator and is computed as

$$\widehat{F}_i(u) = 1 - \exp\left(- \sum_{e_{ij}^* \leq u} \frac{1}{Y_i\left(e_{ij}^*\right)} \right) \quad i = 1, 2 \tag{4.2}$$

In (4.2) we use the variable u instead of t to emphasize that the time scale has been changed due to the transformation in (4.1). To each uncensored residual a resample probability equal to $P_i\left(e_{ij}^*\right) = \widehat{F}_i\left(e_{ij}^*\right) - \widehat{F}_i\left(e_{ij-1}^*\right)$ is calculated from the ordered residuals, $0 < e_{i1}^* < ... < e_{ij-1}^* < e_{ij}^* < ... < e_{ir_i}^*$. In summary our resample procedure consists of the following steps.

1. For each cause specific risk calculate residuals according to (4.1).
2. For each cause specific risk estimate F_i by (4.2).
3. To each uncensored residual attach a resample probability $P_i\left(e_{ij}^*\right)$.
4. Draw one uncensored residual from each cause specific risk according to the probabilities given by P_1 and P_2. Denote this first pair of resample residuals e_{11B}^* and e_{21B}^*.
5. Repeat step four n times, where n is the required sample size. Finally, holding $\mathbf{x}$ fixed, new latent failure times and the censoring indicator are generated as $t_{ij} = \exp\left(\frac{\left(-\mathbf{x}_j\widehat{\beta}_i + \varepsilon_{ijB}^*\right)}{\widehat{\alpha}_i}\right)$ $i = 1, 2$ $j = 1, ..., n$. Observed failure times are then obtained as $t_j = \min\left(t_{1j}, t_{2j}\right)$ and $c_j = I\left(t_{1j} < t_{2j}\right)$.

For a general discussion of resampling censored data see, e.g. Efron (1981), Reid (1981) and Akritas (1986).

The computer programs employed in the study are written in GAUSS code (Gauss, 1992). Maximization is done by the algorithm in Berndt et. al. (1974) and iterations are terminated when the relative gradients of estimated parameters are less than 0.0001. Finally, all calculations are performed by means of analytical expressions.

Table 5.1: Empirical size computed by asymptotic critical values

		Empirical size of nominal		
	Sample size	0.10-level test	0.05-level test	0.01-level test
$\rho^* = 0$				
	100	0.153	0.084	0.019
	500	0.159	0.085	0.023
	1000	0.218	0.124	0.034
	2000	0.821	0.706	0.425
	5000	0.992	0.981	0.909

5. Results

We start by evaluating the size properties of the test using asymptotic critical values. Table (5.1) presents the result obtained by letting the correlation coefficient, ρ^*, of V_1 and V_2 be equal to zero, that is a true null hypothesis. The number of replicates for each sample size is 10000. From the table it can be seen that asymptotic critical values provide a test size larger than the nominal one for all sample sizes. Moreover, as the sample size increases, the empirical test size apporaches one. This is not in lines with previous findings, e.g. Kennan and Neuman (1987); Orme (1987); Chesher and Spady (1988), which indicates that the IMT test has a size problem at small or medium sample sizes. On the contrary, in the present case the size problem of the test gets worse as the sample size increases.

To further check the asymptotic properties of the test, some simulations, not reported though, were performed without unobservables in the specification. In these simulations, the test statistic did converge to the correct asymptotic sampling distribution, however, slowly.

Consequently, given that the suggested test statistic is to be maintained, critical values must be obtained otherwise. A natural and relatively simple way is to apply bootstrap methods.

To evaluate our suggested bootstrap method we now perform two experiments. In the first experiment we evaluate the size properties of the test under the null hypothesis, that is, the case of independence ($\rho^* = 0$). Table (5.2) present the results from the first experiment. As can be seen in the table the bootstrap based critical values seem to provide the correct size of the test. In none of the cases can it be concluded that the empirical size significantly deviates from the nominal test size. However, at the sample size 100 there is a slight indication of a less satisfying

Table 5.2: Empirical size computed by bootstrap-based critical values

		Empirical size of nominal		
	Sample size	0.10-level test	0.05-level test	0.01-level test
$\rho^* = 0$				
	100	0.113	0.069	0.016
	500	0.101	0.055	0.011
	1000	0.091	0.044	0.011
	2000	0.093	0.051	0.009
	5000	0.088	0.044	0.015

Note: A test for equivalence between empirical and nominal size was performed. In none of the cases the test rejected the hypothesis of equality at the 0.01 level. Statistical significance was determined using the binomial distribution.

performance. Probably this a consequence of somewhat poorer estimates of F_1 and F_2 due to the small sample size.

The second experiment concerns the power of the test.

In table (5.3) power estimates of all considered sample sizes and correlations are presented. Due to the heavy computational burden we restrict the sample sizes to 100, 500, 1000 and 2000 in this experiment. A striking feature is the extremely asymmetrical power. This effect was also reported by Carling (1996). We agree with his conclusion that a plausible explanation for this result is that the variance of $\min(t_1, t_2)$ becomes much smaller when the correlation between heterogeneity components is negative. Surprisingly, though, this asymmetrical effect is reversed for the smallest sample size 100. Further, comparing the empirical power estimates in table (5.3) with the ones reported by Carling (1996) we find that for negative correlations the IMT test provides lower power throughout. Surprisingly, the power is reduced as the sample size increases.

Turning to the results for positive correlations the IMT test seems to provide higher power at the sample size 1000 and lower power at the sample size 2000. For sample sizes below 1000 no comparison is possible as Carling (1996) used 1000 as the minimum sample size.

6. Discussion

In the present paper we have examined a test of dependence between competing risks that utilizes the correlation between cause specific residuals. We find that

Table 5.3: Empirical power computed by bootstrap-based critical values

		Empirical power		
	Sample size	0.10-level test	0.05-level test	0.01-level test
$\rho^* = 0.6$	100	0.246	0.147	0.041
	500	0.479	0.340	0.127
	1000	0.661	0.505	0.250
	2000	0.775	0.664	0.432
$\rho^* = 0.2$	100	0.158	0.081	0.022
	500	0.188	0.077	0.021
	1000	0.228	0.117	0.031
	2000	0.267	0.161	0.044
$\rho^* = -0.2$	100	0.106	0.052	0.010
	500	0.065	0.031	0.005
	1000	0.039	0.016	0.005
	2000	0.016	0.009	0.002
$\rho^* = -0.6$	100	0.336	0.193	0.051
	500	0.020	0.011	0.005
	1000	0.008	0.005	0.000
	2000	0.002	0.000	0.000

the sample distribution of the test statistic deviates from the anticipated asymptotic chi-square distribution. The exact reason for this, however, is unknown. Nevertheless, by constructing a proper bootstrap method we were able to obtain a correct test size from the sample. The power of the test was found to be highly asymmetrical, a phenomenon also found in a related study performed by Carling (1996). For a positive correlation structure and at the sample sizes 100, 500, and 1000 there is some evidence that the IMT has higher power than the masspoint technique examined by Carling (1996). On the other hand, at larger sample sizes, it seems that the test provides a somewhat lower empirical power. A negative correlation structure between latent failure times produced very low empirical power estimates at all sample sizes applied, with one surprising exception. The power structure was almost equal for positive and negative correlations at the sample size 100.

The results from both experiments show that a bootstrap procedure can be used to obtain a correct test size. The power in the presence of a positive correlation structure between risks is reasonable good, especially at the smallest sample sizes applied. On the other hand, the power of the test is poor for negative correlations.

References

[1] Akritas, M. G. , (1986), 'Bootstrapping the Kaplan-Meier estimator', *Journal of American Statistical Association*, 81, 1032-1038.

[2] Andersen, P. K. , Borgan, O. , Gill, R. D. , Keiding, N. , (1993), 'Statistical Models Based on Counting Processes', Springer-Verlag, New York, Inc.

[3] Carling, K. , (1996), 'Testing for Independence in a Competing Risks Model', *Computational Statistics and Data Analysis*, vol. 22, 527-535.

[4] Carling, K. , Edin, P.A. , Harkman, P.A. , Holmlund, B.A. , (1996), 'Unemployment Duration, Unemployment Benefits and Labor Market Programs in Sweden', *Journal of Public Economics*, vol. 59, 313-334.

[5] Cox, D.R. , Snell, E.J. , (1968), 'A general Definition of Residuals (With discussion)', *Journal of the Royal Statistical Society*, B30, 248-275.

[6] Cox, D.R. , Snell, E.J. , (1971), 'On Test Statistics Calculated from Residuals', *Biometrika*, vol. 58 589-594.

[7] Efron, B. , (1981), 'Censored Data and the Bootstrap', *Journal of American Statistical Association*, 76, 312-319.

[8] GAUSS 3.0 Applications, (1992), MAXLIK, Aptech Systems, Inc., Maple Valley, WA.

[9] Han, A. , Hausman, A.J. , (1990), 'Flexible Parametric Estimation of Duration and Competing Risks Models', *Journal of Applied Statistics*, 5, 1-28.

[10] Heckman, J.J. , Honoré, B.E. , (1989), 'The Identifiability of the Competing Risks Model', *Biometrika*, vol. 76:2, 325-330.

[11] Heckman, J.J. , Walker, J.R. , (1990), 'The Relationship Between Wages and Income and the Timing and Spacing of Births: Evidence from Swedish Longitudinal Data', *Econometrica*, vol. 58, 1411-1441.

[12] Horowitz, J.L. , (1994), 'Bootstrapped-based Critical Values for the Information Matrix Test', *Journal of Econometrics*, vol. 61, 395-411.

[13] Kaplan, E.L. , Meier, P. , (1958), 'Nonparametric Estimation from Incomplete Observations', *Journal of American Statistical Association*, 53, 457-481.

[14] Kiefer, J.R. , (1988), K.M. 'Econometric Duration Data and Hazard Functions', *Journal of Economic Literature*, vol. XXVI, 646-679.

[15] Lancaster, T. , '*The Econometric Analysis of Transition Data*', Cambridge University Press: Cambridge, 1990.

[16] Narendranathan, W. , Stewart, M.B. , (1993), 'Modelling the Probability of leaving Unemployment: Competing Risks Models with Flexible Base-line Hazards', *Journal of Applied Statistics*, vol. 42, 66-83.

[17] Orme, C. , (1990), 'The Small-sample Performance of the Information-matrix test', *Journal of Econometrics*, vol. 46, 309-331.

[18] Reid, N. , (1981), 'Estimating the Median Survival Time', *Biometrika*, 68, 601-608.

[19] Van den Berg, G.J. , Lindeboom, M. , Ridder G. , (1994), 'Attrition in Panel Data and the Empirical Analysis of Dynamic Labor Market Behavior', *Journal of Applied Econometrics*, vol. 9, 421-435.

[20] White, H. , (1982), 'Maximum Likelihood Estimation of Misspecified Models', *Econometrica*, vol. 50, 1-25.