Generalized Residual-Based Specification Testing for Duration Models with Censoring

Yongmiao Hong
Department of Economics
Cornell University
WISE, Xiamen University

Jing Liu
Department of Economics
Cornell University

October 2007

This is Chapter 1 of Jing Liu's doctoral dissertation at Cornell University.

We thank Jerry Hausman, Shakeeb Khan, Nicholas Kiefer and participants at Econometric Society 2006 North America Summer Meeting, Far Eastern Meeting, the 2006 International Symposium on Contemporary Labor Economics, Xiamen, China, University of Connecticut, University of Indiana, Southern Methodist University, Texas A&M University and Cornell University for helpful comments. All errors are our responsibilities. Yongmiao Hong thanks the support from the Cheung Kong Scholarship from Chinese Ministry of Education.

Abstract

We propose a new test for duration models with censoring -- popularly used in economics, finance and other fields -- using a novel computationally simple empirical survival function that utilizes information from censored observations. The impact of parameter estimation uncertainty is properly addressed, ensuring that the proposed test has an asymptotically valid Type I error. A simple resampling method is used to obtain critical values for the proposed test, and extensions to unobserved heterogeneity and competing risks are also considered. A simulation study shows that the proposed test has its accurate sizes and good powers in finite samples.

Key Words: Censoring, Duration analysis, Specification testing,
Parameter Estimation Uncertainty

JEL Classification: C12, C41, C21

1 Introduction

The modeling and analysis of lifetime data is of interest in many fields, such as economics, finance, sociology, biomedicine and engineering (Allison 1984, Lancaster 1992, Lawless 2003).[1] The subject of interest is the time to the occurrence of an event of interest, or equivalently, the instantaneous exit rate from a current state.

Various examples can be found in economics and finance. In labor economics, duration models are regarded as the reduced forms of behavioral models based on job search theory and are widely used to analyze unemployment spells (Kiefer 1988, Lancaster 1992). For example, Kiefer (1985, 1988) introduces a simple job-search model which leads to the exponential distribution model of unemployment duration. In this model, it is assumed that both the instantaneous unemployment utility and job offer arrival rate are exogenous and constant, and the utility of being employed depends on the wage. As a result, the worker's optimal behavior is described by a reservation wage policy, resulting in a constant transition rate to employment. In market microstructure, the asymmetric information model suggests that time between trades contains important information about event uncertainty, thus affecting the behavior of quotes, spreads and transaction prices (Easley and O'Hara, 1992). To accommodate such features, Engle and Russell (1998) and Engle (2000) use the accelerated failure time model, allowing past information to affect trade frequencies. In credit risk analysis, instantaneous probabilities of default for counterparts of banks or credit card companies' portfolios are studied extensively. Representing the state of art in reduced form models, hazard models have showed considerable flexibility to conduct dynamic analysis and bridge the gap between default prediction and default risk pricing (Lando 2004). For example, Shumway (2001) and Chava and Jarrow (2004) adopt discrete logistic hazard models for bankruptcy prediction; Bharath and Shumway (2006) and Duffie *et al.* (2006) use Proportional Hazard models for the same purpose.

Duration models are also widely used in other disciplines in social science. In demographic analysis, men and women enter into or exit from cohabitations or marriages, or enter into parenthood (Hoem 1983; Manning 1995; Michael and Tuma 1985; Monahan 1963). In organizational ecology, firms or organizations (Barnett 1997; Carroll and Hannan 1989; Haveman 1992) are created or ended. In marketing applications, consumers switch from one brand to another or purchase the same brand again (DuWors Jr. and Haines Jr. 1990), to name a few.

Despite the diversity of topics, social science data have several common important features. They are

[1] Terminologies vary across fields. Popular terminologies include duration models, hazard models, lifetime models, failure time models, survival analysis and event history study.

usually nonexperimental, and durations rarely follow the same distribution unconditionally but rather display systematic differences across subgroups. The source of differences is represented by covariates or explanatory variables which could vary across time. More often than not, duration data are censored so that only partial information is available for censored observations. For example, when an individual leaves the sample before completing an unemployment spell, only the lower bound of his or her spell is known. A successful model should accommodate these features and should be consistent with the underlying economic theory as well. Therefore, model specification is always arduous but essential. Misspecification leads to incorrect implications of social agents' behavior and the environment they are assumed to face. The consequences of model misspecification are not trivial; hence model validation has attracted increasing attention from academia, industry and policy community.[2]

Surprisingly however, despite the importance of model specification, relatively little research has been devoted to diagnostic testing of duration models, particularly when dealing with censored observations. There are two categories of diagnostic tests in the existing literature, the informal graphical method and formal statistical method. Lancaster and Chesher (1985) propose a graphical diagnostic based on the integrated hazard function, which is one form of generalized model residuals.[3] Under correct model specification, the integrated hazard has a unit exponential distribution. The informal graphical check investigates on the departure of the integrated hazard from unit exponential by plotting minus the logarithm of the sample survivor function at the point of each exponential residual against the residual itself. The points with uncensored observations should lie approximately on the 45 degree line (subject to sampling variations) if the duration model were correctly specified. Under light censoring, residual plots can often reveal surprising departures from the hypothesized model and suggest directions for potential improvement. This intuitive graphical method can serve as a starting point for diagnostic check and is a useful complement to formal statistical methods. Because only uncensored durations are transformed by the integrated hazard function and used in plotting, the plots do not contain any information on censored

[2] For example, the Payment Cards Center of the Federal Reserve Bank of Philadelphia and the Wharton School's Financial Institutions Center hosted a forum on validation of consumer credit risk models in 2004. This forum brought together experts from industry, academia, and the policy community to discuss challenges surrounding model specification strategies and techniques. Participants agreed that a credit risk model's performance can have important effects on market share, perhaps even creating adverse selection problems due to model misspecification. As competitive pressures and technology advances continue, implementation of new model validation techniques will rise in importance. Another example is well known in labor economics: the failure to account for unobserved heterogeneity usually results in spurious duration dependence. In terms of policy implication, duration dependence may suggest an early installation of reemployment training, whereas heterogeneity suggests the opposite.

[3] Cox and Snell (1968, 1971) give a definition of generalized model residuals. Suppose $\boldsymbol{\beta}$ is an unknown parameter vector and $\{\varepsilon_i\}$ are unobserved i.i.d. random variables. If each observation Y_i depends on only one of the $\{\varepsilon_i\}$, then one could write Y_i as a function of $\boldsymbol{\beta}$ and ε_i, which may have a unique solution such that ε_i can be expressed as a function of Y_i and $\boldsymbol{\beta}$. Substituting $\boldsymbol{\beta}$ with its MLE $\hat{\boldsymbol{\beta}}$ will then yield a generalized model residual e_i.

observations, making this method of limited use with censored data and even inapplicable under heavy censoring. Moreover, the test is based on parameter estimates instead of the true parameter values. This can introduce a nontrivial impact of parameter estimation on the behavior of general residuals.

Formal statistical tests have been contributed by Chesher (1984), Lancaster (1985), Kiefer (1984, 1985, 1988), Sharma (1987), Jaggia (1997) and Prieger (2000). All of them are Lagrange Multiplier (LM) tests based on certain moment restrictions. In duration analysis, tests are often constructed against an arbitrary alternative parametric specification, such as the presence of heterogeneity of unknown parametric form. LM tests are thus preferred to Wald or Likelihood Ratio tests because the alternative model does not have to be estimated when LM tests are used. According to the moment restrictions and their purposes, they can be roughly divided into three categories: raw moment-based (RM) tests, Laguerre Polynomial-based (LGP) tests and LM tests for heterogeneity (Prieger 2000).

The RM test, suggested by Kiefer (1988), gives a simple diagnostic procedure based on the raw moments of the integrated hazard. If the null model were true and the observations are not censored, the exponential residual should have the rth moment equal to $r!$. The RM test checks the validity of these moment restrictions.

Kiefer (1985) also develops an innovative alternative to the RM test– the Laguerre Polynomial (LGP) test. This score test is designed for the null hypothesis of exponentially distributed durations against a general alternative with Laguerre polynomial series expansions. Sharma (1987) generalizes Kiefer's method to test the null hypothesis of Weibull distributed durations. The LGP test essentially checks certain orthogonal polynomial-based moment restrictions implied by the null. This method is appealing for its simplicity and intuitiveness. However as pointed out by Kiefer (1985), "the null that is being tested is tested conditionally on estimated values of the parameters ... . Consequently, the stated asymptotic size of the test reported here is conservative in the sense of leading to more rejections than the unconditional test if the nominal size is strictly interpreted". Meanwhile, similar to the informal graphical method, censored observations are discarded in the LGP test. If censored observations are considered, the generalized residual would behave approximately like a censored sample from standard exponential variables under the null (Lawless 2003, Chapter 6).

Chesher (1984), Kiefer (1984) and Lancaster (1985) develop LM tests for neglected multiplicative heterogeneity. Neglected heterogeneity is of special interest because it can lead to biased predictions and false interpretations. The approaches of Kiefer (1984) and Lancaster (1985) are a bit different, but are both based on approximations to the distribution of the heterogeneous component, leading to essentially the same statistics (Sharma 1987). Chesher (1984) points out that in the uncensored case, their tests are equivalent to White's (1982) Information Matrix (IM) test when the variance of heterogeneity is small. Their tests investigate whether the residual variance is unity, which is the second moment restriction implied by the unit exponential distribution. However the IM tests are notorious for their poor sizes in finite samples (Horowitz and Neumann 1989). In this case Jaggia (1997) shows that the variance calculation of the mean score ignores the covariance between scores (with respect to different parameters), resulting in an underrejection for the null hypothesis. As a remedy, Horowitz (1994) shows that bootstrap can control the size. Nevertheless, its usefulness can be limited under censoring as the interaction of censoring and misspecification complicates the problem (Lancaster 1985).

Prieger (2000) extends all aforementioned tests to censored data. For the RM test, he derives raw moment conditions for censored samples, which are much more tedious than for uncensored samples. For the LGP test, he calculates sample moments for censored observations based on Laguerre polynomials. It turns out that the modified Laguerre polynomials are no longer orthogonal, resulting in the loss of their computational advantage in censored cases. He extends the LM test for heterogeneity to censored data and uses higher-order approximations of the likelihood function in the construction of his test statistic, hence improving power of the test. Prieger's (2000) extension nicely incorporates information of censored observations, but still ignores parameter estimation uncertainty.

In this paper, we develop a new approach to testing the adequacy of duration models with censoring. Essentially, our test inspects misspecification over the whole distribution by using all available information in the complete as well as censored observations. In addition parameter estimation does not affect the asymptotic distribution of our test statistic. Overall our approach has the following advantages over existing tests.

First, our test is based on the conditional duration distribution rather than its moments only. It can detect model misfits in duration distribution even if certain moment conditions hold.

Second, by exploiting the property of observable random censoring, we propose a novel computationally simple empirical survivor function, which efficiently makes use of all available information contained in complete and censored observations. To use this rather than the popular Kaplan-Meier (KM) estimator, we avoid the notoriously difficult asymptotic analysis of KM-based test statistics and the corresponding time expensive computation.

Third, unlike some existing tests, whose applications are limited, our test does not specify any alternative and is generally applicable. On the other hand, the LM test for heterogeneity does not go beyond testing omitted heterogeneity while the LGP test can only handle nested hypotheses,.

Four, our test does not require any particular estimation method; any $\sqrt{n}$-consistent parameter estimators can be used. Thanks to the use of Wooldridge's (1990) device, the asymptotic distribution of our test statistic is not affected by parameter estimation uncertainty, making our test easily implementable.

Last, our test is computationally simple. It is coded easily and takes significantly less time to run than most existing tests. In contrast, the moment derivations of the LGP test and the LM test are tedious especially under censoring, and programming varies according to the number of moments used.

Section 2 introduces the framework and states the hypotheses of interest. Section 3 proposes a new empirical survivor function under censoring, develops our test and derives its asymptotic distribution. This asymptotic distribution is not distribution-free, making the tabulation of critical values impossible. Section 4 introduces a simple resampling method to obtain the critical values of our test statistic and justifies its validity. Section 5 discusses how to extend our test to accommodate unobserved heterogeneity and competing risks. We present Monte Carlo evidence on the finite sample performance of the proposed test in comparison with some existing popular tests in section 6. Section 7 concludes. All mathematical proofs are collected in the appendix. Throughout, we use Δ to denote a generic bounded constant, and $||\cdot||$ the Euclidean norm.

2 Framework and Hypotheses of Interest

The focus of duration analysis is the time to the occurrence of event of interest, namely, the lifetime. However, lifetimes are usually incomplete, in which case censoring times are observed instead. Moreover social science data are rarely homogeneous, requiring a careful use of covariates to account for systematic differences across groups (Lancaster 1992). Consistent with these stylized facts, we consider the following data generating process:

Assumption A.1: Available data for duration analysis contain n observations. For the ith observation, $i = 1, ..., n$, the minimum of lifetime $\tilde{T}_i$ and censoring time $\tilde{C}_i$, denoted $\tilde{V}_i = \min(\tilde{T}_i, \tilde{C}_i)$, is observable, together with an indicator $\delta_i \equiv \mathbf{1}(\tilde{T}_i \leq \tilde{C}_i)$ for whether censoring occurs. Moreover, certain individual characteristics denoted by X_i, a $k \times 1$ vector are also observed.

2.1 Survivor Functions and Hazard Functions

In duration analysis, the survivor function gives the upper tail area of the lifetime distribution, i.e., the probability that random variable $\tilde{T}$ is larger than certain value $t \in [0, \infty)$, conditional on $X, S(t|X) = \Pr(\tilde{T} > t|X)$. Let $F(t|X)$ be the lifetime distribution function, conditional on X. Then $S(t|X) = 1 - F(t|X)$.

The hazard function defines the instantaneous exit rate, characterizing the way in which the risk of failure varies with time. In continuous time, it is defined as the probability of exit from a state in the short interval of length Δt after t, conditional on the state still being occupied at t,

$$h(t|X) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t, X)}{\Delta t} = \frac{f(t|X)}{S(t|X)},$$

where $f(t|X) = \frac{d}{dt}F(t|X)$ is the conditional probability density function of $\tilde{T}$ given X.

Mathematically the functions $F(t|X), S(t|X)$ and $h(t|X)$ can be used interchangeably to describe a lifetime distribution. Nevertheless social science theories often suggest direct specification of the hazard function as a result of optimal choices by the agents (structural models) or the relevant regressor variables and the probable directions of their effects (reduced form models). For example, Kiefer (1985) provides a highly stylized two-state job search model. This economic model leads to a constant instantaneous probability of re-employment. Correspondingly, this suggests an exponentially distributed duration model with $h(t|X) = \gamma, S(t|X) = \exp(-\gamma t)$ and $F(t|X) = 1 - \exp(-\gamma t)$.

Another example is Cox's (1972, 1975) proportional hazard model, $h(t|X) = h_0(t)h_1(X)$, where $h_0(t)$ is the baseline hazard function, or the hazard function in absence of covariates. The name "proportional" comes from the fact that the ratio $h(t|X)/h_0(t) = h_1(X)$ is constant over time t. This greatly simplifies inference on duration models, because Cox (1972) suggests an ingenious method to estimate the unknown model parameters of a parametrized $h_1(X)$ without having to specify the form of the common function $h_0(t)$. Although there are few social-scientific justifications of why hazard should be proportional (Lancaster 1992), this specification gains unparalleled popularity because of its analogy to regression models.

By a transformation of time scale, any form of hazard function can be integrated into a constant hazard, known as the integrated hazard, $H(t|X) = \int_0^t h(s|X)ds$ and this facilitates the transformations among $F(t|X), S(t|X)$ and $h(t|X), i.e., S(t|X) = \exp[-H(t|X)]$. Almost all existing tests for durations are based on this property.

2.2 Types of Right Censoring

Lifetime data often come with the property known as right censoring for a variety of reasons that are usually a consequence of a researcher's data collection or observation plan (Lawless 2003). When data are subject to censoring, lifetime $\tilde{T}_i$ is not always observable. It is important to understand the process by which censoring times arise in order to facilitate statistical analysis. Right censoring could come up for different reasons, sometimes planned such as the designed ending of a survey, and sometimes unplanned as in the case when surveyed individuals are lost to follow up. The following three censoring mechanisms are the most common in practice.

Type I Censoring

Type I censoring arises when there is a fixed calendar time censoring for each individual such as the termination of a study. Type I censoring is also called time censoring (Nelson 1982). In this case, $\tilde{T}_i$ is only observed when $\tilde{T}_i \leq \tilde{C}_i$, otherwise only $\tilde{C}_i$ is observed. Thus, $(\delta_i = 1, \tilde{V}_i = \tilde{T}_i)$ denotes a complete observation while $(\delta_i = 0, \tilde{V}_i = \tilde{C}_i)$ denotes a censored observation. However, a fixed calender time censoring does not necessarily imply a uniform censoring time for all individuals. Only under the special case when all individuals start their lifetimes simultaneously would the censoring times be identical (fixed type I censoring). In most cases random samples dictate random entries into the initial state, so individuals have random censoring times (random type I censoring). Type I censoring occurs when a study is conducted over a specific time interval, which often comes up in social science models (Anderson and Portugal 1987, Orbe, Ferreira and Nunez-Anton 2001, Roszbach 2004, Bijwaarda and Ridder 2005).

Type II Censoring

Type II censoring occurs when only the smallest r lifetimes are observed in a random sample: $\tilde{T}_{(1)} \leq \cdots \leq \tilde{T}_{(r)}, 1 \leq r \leq n$. This scheme arises when n individuals start their lifetimes together and the study ends whenever the first r $(r \leq n)$ failures are observed. Type II censoring is thus known as failure censoring (Nelson 1982). The total study time here is *ex-ante* random because $\tilde{T}_{(r)}$ is random. Also the censoring times $\tilde{C} = \tilde{T}_{(r)}$ are the same for the rest $n - r$ individuals. Type II censoring is more common in an experimental engineering environment such as termination of the life test on bulbs when a prespecified number of bulbs fail.

Independent Random Censoring

Another simple yet often realistic random censoring is independent random censoring, where each individual is assumed to have a lifetime $\tilde{T}_i$ and a censoring time $\tilde{C}_i$, and $\tilde{T}_i$, $\tilde{C}_i$ are independent continuous random variables across individuals. Moreover, all lifetimes and censoring times are assumed mutually independent. In this case, $\tilde{C}_i$ may not be observable. This type of censoring usually occurs in medical studies, when competing risks are present, or if individuals drop out of the study or are lost to follow up. It also occurs in social science studies (Hochguertel and Soest 2001, O'Hagan and Stevens 2004).

Other types of censoring could be present as well although less often, such as left censoring or interval censoring, and usually data are subject to more than one type of censoring. For concreteness, we focus on the following censoring scheme in this paper.

Assumption A.2: All censoring times $\{\tilde{C}_i : i = 1, ..., n\}$ are observable and independent across individuals. Moreover, $\{\tilde{C}_i\}$ and $\{\tilde{T}_i\}$ are mutually independent conditional on X_i.

This assumption allows flexible censoring schemes. First of all, (even unconditional) independence between lifetimes and censoring times are usually satisfied in practice. Most of the time, censoring is a consequence of the empirical researcher's observation or data collection plan, normally independent of the sample feature. Secondly, as is almost always the case, random sampling dictates the independence among censoring times. Finally, the assumption of observable censoring times accommodates several types of censoring, random type I censoring and independent random censoring with observable censoring times. Therefore our assumption here actually covers many interesting cases in social science studies. For example, the censoring schemes in credit risks are almost all random type I censoring (Roszbach 2004).

2.3 Hypothesis of Interest

To state the hypotheses of interest, we introduce the following assumption on $\{X_i, \tilde{T}_i\}$:

Assumption A.3: $\{(X_i, \tilde{T}_i) : i \geqslant 1\}$ is an *i.i.d.* sequence with an unknown conditional distribution function $F(\cdot|X_i)$ of $\tilde{T}_i$ given X_i.

Since social science data are usually heterogeneous, lifetimes $\{\tilde{T}_i\}$ seldom follow the same distribution unconditionally. However, conditional on covariates, $\tilde{T}_i|X_i$ often displays the same distribution, so it is appropriate to specify a common conditional distribution function under most scenarios. In general, all regression models (among which proportional hazard is a special case) automatically satisfy Assumption A.3 (Lawless 2003).

Often practitioners specify a parametric model for the hazard function $h\left(t|X_i\right)$, which is equivalent to a parametric specification for the conditional distribution $F(\cdot|X_i)$. For convenience, we state the hypotheses of interest in terms of the conditional distribution here. Let $F_0(\cdot|X_i, \theta)$ be the conditional distribution of $\tilde{T}_i$ given X_i, implied by a hazard model to be tested, where Θ is a finite-dimension parameter space. Then our hypotheses of interest can be stated as follows:

$\mathbb{H}_0 : F(\cdot|X_i) = F_0(\cdot|X_i, \theta_0)$ for some unknown $\theta_0 \in \Theta$ vesus

$\mathbb{H}_A : F(\cdot|X_i) \neq F_0(\cdot|X_i, \theta)$ a.s. for all $\theta \in \Theta$.

3 New Goodness-of-fit Test

We now propose a new approach to testing the parametric adequacy of a duration model. To provide some intuition and insight, we will first discuss the heuristics of our new test, and then introduce it formally.

3.1 Heuristics

Our good-of-fit test is based on the comparison between a simple empirical survivor function and its parametric counterpart under the null $\mathbb{H}_0$, where the empirical survivor function fully makes use of the information from both complete and censored observations. Because our hypotheses of interest are parametric models for the conditional distribution of lifetime $\tilde{T}_i$ given X_i, one might be tempted to use Andrews' (1997) seminal conditional Kolmogorov (CK) test. However, the CK test only applies to the case of no censoring. For lifetime data, censoring is more often than not. Therefore, a new empirical distribution function (or equivalently, survivor function) is needed to take account of censored observations. To address this, we introduce a novel simple empirical survivor function that nicely incorporates all available information.

3.1.1 An Empirical Survivor Function under Censoring

In the absence of censoring, we can easily implement a conditional probability integral transformation. This yields a series of generalized residuals that will be uniformly distributed on $[0,1]$ under $\mathbb{H}_0$, *i.e.*, $\{F_0(\tilde{T}_i|X_i,\theta_0),\ i=1,...,n\}$ is an *i.i.d* $U[0,1]$ sequence under $\mathbb{H}_0$. This suggests that we can construct goodness-of-fit tests by comparing an empirical distribution or survivor function of $F_0(\tilde{T}_i|X_i,\theta_0)$ with $U[0,1]$ distribution. This is the basic idea behind the classical Kolmogorov-Smirnov (KS) test and Cramer-von-Mises (CV) test. Compared to moment-based tests, the use of the distribution function makes it possible to detect a wider range of model misspecifications. However, data incompleteness due to censoring makes the above idea difficult to implement because $F_0(\tilde{T}_i|X_i,\theta_0)$ is no longer uniformly distributed when $\tilde{T}_i$ is censored. Moreover, the true parameter value θ_0 is unknown in practice and has to be replaced by an estimator $\hat{\theta}$ that is consistent for θ_0 under $\mathbb{H}_0$. It is well known that the parameter estimation uncertainty in $\hat{\theta}$ complicates the asymptotic distribution of test statistics such as those of KS test and CV test. In fact Lawless (2003, Chapter 10) suggests the idea of using uniform residuals to form omnibus tests. He cautions, however, "censoring or other forms of incompleteness in the data may make it difficult to find test statistics". Our approach in this paper provides a solution to this difficulty.

To overcome this difficulty, we first introduce a simple empirical survivor function in the presence of censoring, which accommodates most commonly encountered censoring schemes in social science while making tractable test statistics feasible. In particular, we transform both the original lifetimes and censoring times by the null conditional lifetime distribution function $F_0(\tilde{T}_i|X_i,\theta)$. More specifically, we define the following probability integral transforms:

$$\begin{aligned} T_i(\theta) &= F_0(\tilde{T}_i|X_i,\theta), \\ C_i(\theta) &= F_0(\tilde{C}_i|X_i,\theta), \\ V_i(\theta) &= \min\left[T_i(\theta), C_i(\theta)\right]. \end{aligned}$$

Let $\hat{S}_v(t,\theta)$ and $\hat{S}_c\left(t,\theta\right)$ be the empirical survivor functions of $\{V_i(\theta)\}$ and $\{C_i(\theta)\}$ respectively; that is,

$$\begin{aligned} \hat{S}_v(t,\theta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[V_i(\theta) > t\right], \\ \hat{S}_c(t,\theta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[C_i(\theta) > t\right]. \end{aligned}$$

Then we propose the following empirical survivor function for $\{T_i\left(\theta\right)\}$ applicable to both censored and uncensored observations:

$$\hat{S}_T(t,\hat{\theta}) = \frac{\hat{S}_v(t,\hat{\theta})}{\hat{S}_c(t,\hat{\theta})}.$$

Theorem 1 below shows that no matter whether there is censoring, $\hat{S}_T(t,\theta)$ can consistently estimate the population survivor function $S_T\left(t,\theta\right) = E\left\{\mathbf{1}\left[T_i(\theta) > t\right]\right\}$ of $\{T_i(\theta)\}$.

Theorem 1. *Suppose Assumption A.1-A.3 hold. Then under the null hypothesis* $\mathbb{H}_0$,

$$\frac{\hat{S}_v(t,\theta_0)}{\hat{S}_c(t,\theta_0)} \xrightarrow{p} 1 - t.$$

Obviously when data are complete, *i.e.*,when $\tilde{T}_i \leq \tilde{C}_i$ for all i, we have $V_i(\theta_0) = T_i(\theta_0)$ and $\hat{S}_c(t,\theta_0) = 1$. In this case, the function $\hat{S}_T\left(t,\theta_0\right)$ simplifies to the conventional empirical survivor function, $\hat{S}_T(t,\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[T_i(\theta_0) > t]$. Andrews' (1997) CK test applies to this uncensored case, but is still more computationally burdensome than our test to be proposed below. To see that, let us briefly review several properties of the CK test. Firstly, to circumvent the problem that the parametric model does not specify the distribution function of X_i, the CK test compares the empirical distribution function with the semi-parametric/semi-empirical distribution function. Secondly, the CK statistic is defined by taking supremum over the sample $\{X_i,\tilde{T}_i\}_{i=1}^{n}$. Consequently, the CK test depends on signs of the elements $(X_i,\tilde{T}_i)$ in the random sample. To obtain a sign invariant CK test statistic, one has to explore all possible sign permutations and define a resultant CK test statistic to be the maximum of these statistics, which is undoubtedly time consuming. In contrast, we transform the original data by the null conditional parametric distribution function, which obviates the difficulty of defining a semi-parametric distribution function. The computational advantage is phenomenal, because we "reduce" the dimensionality from $\mathbb{R}^{k+1}$ to $\mathbb{R}$ and we do not have to worry about the problem of sign dependence of test statistics.

In duration analysis, the Kaplan-Meier empirical survivor function is generally applicable to various random censoring schemes, including the random censoring scheme we are considering. Unfortunately it leads to formidable statistical inference procedures. When the Kaplan-Meier estimator is used, a KS or CV test statistic can only be derived under the framework of counting processes (Lawless 2003), and its asymptotic analysis is notoriously difficult. As Fleming and Harrington (1991) and Andersen *et al.*(1992) point out, elegant mathematical derivations fail to generate easily usable tests. Sun (1997) gives some results. Even more complicated, with heterogeneous data which are normally encountered in social science, a conditional Kaplan-Meier estimator (Beran 1981) has to be used, where survivor functions are estimated locally for different $X_i's$. Thus, although we could use the conditional Kaplan-Meier estimator to form a test, the asymptotic distribution may not be tractable and is computationally expensive. Meanwhile in economic settings, available samples are usually small after conditioning on covariates, therefore the Kaplan-Meier estimator is “unlikely to prove successful in econometrics because the available samples are small especially after cross-classification by regressor variables” (Heckman and Singer 1984). The Kaplan-Meier estimator covers all random censoring schemes,[4] while in social science studies certain types of censoring as described in Assumption A.2 are predominantly common. Our simple empirical survivor function $\hat{S}_T(t,\theta)$ thus exploits the characteristic of this specific but commonly encountered type of censoring. The simplicity of this estimator transmits to the manageability of the asymptotic theory associated with the proposed test statistic. It also simplifies a great deal the implementation of the proposed test.

3.1.2 Impact of Parameter Estimation Uncertainty

Under $\mathbb{H}_0$ and no censoring, the probability integral transforms $\{T_i(\theta_0)\}$ is $i.i.d.U[0,1]$. Therefore the population survivor function $S_T(t,\theta_0) = 1-t$ under $\mathbb{H}_0$. Intuitively we can test $\mathbb{H}_0$ against $\mathbb{H}_A$ by comparing $\hat{S}_T(t,\hat{\theta})$ with $1-t$. Any significant difference between them is evidence of model misspecification. However, we cannot proceed with this intuition without scrutiny, because $\{T_i(\hat{\theta})\}$ are not exactly $i.i.d.U[0,1]$ (Lawless 2003).[5] In another word, the estimated parameter $\hat{\theta}$ in some sense "contaminates"

[4] The Kaplan-Meier estimator gives the nonparametric maximum likelihood estimator of the survivor function of randomly censored lifetime data (Fleming and Harrington, 1991).

[5] Lawless (2003) warns that, when using estimated parameters to implement probability integral transformations, the estimated residual $T_i(\hat{\theta}) = F_0(\tilde{T}_i|X_i,\theta)$ is only approximately, but not exactly *i.i.d.* $U[0,1]$, so care must be given to the distribution of any such statistic.

the asymptotic distribution of our test statistic. In fact one common caveat for the existing tests in the literature is that the impact of parameter estimation uncertainty in $\hat{\theta}$ is not taken into account. When tests are constructed using the estimated parameter rather than the true parameter θ_0, nontrivial uncertainty is introduced into the test statistic even asymptotically. Kiefer (1985) notes that such uncertainty normally generates a poor (indeed asymptotically invalid) size of the test. To gain insight into the impact of parameter estimation uncertainty and how we remove it, we define

$$U_i(t,\theta) = \frac{\mathbf{1}[V_i(\theta) > t] - (1-t)\mathbf{1}[C_i(\theta) > t]}{\hat{S}_c(t,\theta)},\ t \in [0,1],$$

then $\hat{S}_T(t,\theta) - (1-t) = \frac{1}{n}\sum_{i=1}^{n} U_i(t,\theta) \equiv \frac{1}{\sqrt{n}}\hat{M}_n(t,\theta)$, say. Under $\mathbb{H}_0$, we have $\hat{M}_n(t,\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(t,\theta_0) \xrightarrow{p} \sqrt{n}[S_T(t,\theta_0) - (1-t)] = 0$ for all $t \in [0,1]$. This forms the basis of our test.

Under regularity conditions (see Assumption A.4 and A.5 below), we can show (see proof of Theorem 2) that, $\hat{M}_n(t,\theta)$ has the following asymptotic representation:

$$\hat{M}_n(t,\theta) = \hat{M}_n(t,\theta_0) - \frac{1}{S_c(t,\theta)}\bar{g}(t,\theta,\theta_0)'\sqrt{n}(\theta - \theta_0) + o_p(1), \tag{3.1.1}$$

where $\bar{g}(t,\theta,\theta_0) = p\lim \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{1}[C_i(\theta) > t]\frac{\partial}{\partial\theta}F_0[F_0^{-1}(t,\theta),\theta_0]|_{\theta=\theta_0}\right\}$ and $o_p(1)$ is uniform in $t \in [0,1]$.[6] Clearly the asymptotic distribution of $\hat{M}_n(t,\hat{\theta})$ depends on the limiting distribution of $\hat{M}_n(t,\theta_0)$ and the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$.[7] Consequently test statistics based on $\hat{M}_n(t,\hat{\theta})$ will not be asymptotically free of the impact from the parameter estimation $\hat{\theta}$. Deriving the limiting distribution of $\hat{M}_n(t,\hat{\theta})$ normally entails finding the asymptotic variance of $\hat{\theta}$ and the covariance matrices between $\hat{M}_n(t,\theta_0)$ and $\sqrt{n}(\hat{\theta} - \theta_0)$, but the resulting test statistics can be hard to compute (Wooldridge 1990). This is particularly relevant to the present context, because of the involvement of nuisance parameters. To take into account such impact, one either needs to make additional assumption about the expansionary form of $\hat{\theta}$ (Andrews 1997), or needs to rely on certain conditions derived in parameter estimation (for example, the score functions of MLE), which in turn ties the method to one specific estimation procedure. The convenient "purging" technique introduced by Wooldridge (1990), on the other hand, requires neither of these, making it especially flexible and attractive. With the adoption of Wooldridge's (1990) device, the asymptotic distribution of our test statistic is free of the impact of parameter estimation. One

[6] $F_{X_i}^{-1}(t,\theta)$ is the inverse function of $F_{X_i}(t,\theta)$.

[7] Actually it depends on how θ_0 is estimated.

can treat the test statistic as if it were calculated at the true parameter value θ_0, and this saves all the trouble of calculating corresponding covariance matrices similar as the ones between the first term and the second term in (3.1.1). Moreover, this device is computationally simple, only requiring the running of an OLS regression. We will use Wooldridge's (1990) idea to purge parameter estimation impact in our test statistic.

Wooldridge's (1990) idea is based on the fact that *OLS* residuals are orthogonal to explanatory variables. Suppose $E\left[\tau\left(T_i, X_i, \theta_0\right)|X_i\right] = 0$ is the hypothesis of interest, where function $\tau\left(T_i, X_i, \theta_0\right)$ is differentiable with respect to θ. Then the validity of such hypothesis can be tested by choosing some misspecification indicator function $\Lambda\left(X_i, \theta_0\right)$ of the explanatory variables X_i and checking whether the sample covariance between $\tau\left(T_i, X_i, \theta_0\right)$ and $\Lambda\left(X_i, \theta_0\right)$ is significantly different from zero. To derive the asymptotic distribution of this sample covariance, we employ a Taylor series expansion:

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda(X_i,\hat{\theta})\tau(T_i,X_i,\hat{\theta}) \\
= \quad & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda\left(X_i,\theta_0\right)\tau\left(T_i,X_i,\theta_0\right) + \sqrt{n}(\hat{\theta}-\theta_0)'\left[\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(X_i,\theta_0\right)\frac{\partial}{\partial\theta}\tau\left(T_i,X_i,\theta_0\right)\right] + o_p(1).
\end{aligned}$$

The second term is the uncertainty impact of parameter estimation and it affects the asymptotic distribution of the sample covariance. To purge this, Wooldridge (1990) proposes the modified sample moment

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\Lambda(X_i,\hat{\theta}) - G(X_i,\hat{\theta})'\hat{\beta}(\hat{\theta})\right]\tau(T_i,X_i,\hat{\theta}), \tag{3.1.2}$$

where $G(X_i,\theta) = E\left[\frac{\partial}{\partial\theta}\tau\left(T_i,X_i,\theta\right)|X_i\right]$ and $\hat{\beta}\left(\theta\right) = \left[\sum_{i=1}^{n} G\left(X_i,\theta\right)' G\left(X_i,\theta\right)\right]^{-1}\sum_{i=1}^{n} G\left(X_i,\theta\right)'\Lambda\left(X_i,\theta\right)$. Note that $\Lambda(X_i,\hat{\theta}) - G(X_i,\hat{\theta})'\hat{\beta}(\hat{\theta})$ is the *OLS* residual of regressing $\Lambda(X_i,\hat{\theta})$ on the gradient $G(X_i,\hat{\theta})$.

Since $\hat{\beta}(\theta) = \beta(\theta_0) + o_p(1)$, where $\beta\left(\theta_0\right) = \{E\left[G\left(X_i,\theta_0\right)G\left(X_i,\theta_0\right)'\right]\}^{-1}E\left[G\left(X_i,\theta_0\right)\Lambda\left(X_i,\theta_0\right)\right]$, we have the modified sample variance

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\Lambda(X_i,\hat{\theta})-G(X_i,\hat{\theta})'\hat{\beta}(\hat{\theta})\right]\tau(T_i,X_i,\hat{\theta}) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\Lambda(X_i,\hat{\theta})-G(X_i,\hat{\theta})'\beta(\theta_0)\right]\tau(T_i,X_i,\hat{\theta})+o_p(1) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\Lambda\left(X_i,\theta_0\right)-G(X_i,\theta_0)'\beta(\theta_0)]\tau\left(T_i,X_i,\theta_0\right) \\
& +\sqrt{n}(\theta-\theta_0)'\frac{1}{n}\sum_{i=1}^{n}[\Lambda\left(X_i,\theta_0\right)-G(X_i,\theta_0)'\beta(\theta_0)]\frac{\partial}{\partial\theta}\tau\left(T_i,X_i,\theta_0\right) \\
& +\sqrt{n}(\theta-\theta_0)'\frac{1}{n}\sum_{i=1}^{n}[\frac{\partial}{\partial\theta}\Lambda\left(X_i,\theta_0\right)-\frac{\partial}{\partial\theta}G(X_i,\theta_0)\beta(\theta_0)]\tau\left(T_i,X_i,\theta_0\right)+o_p(1) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\Lambda\left(X_i,\theta_0\right)-G(X_i,\theta_0)'\beta(\theta_0)]\tau\left(T_i,X_i,\theta_0\right)+o_p(1).
\end{aligned}$$

Under certain regularity conditions, the Uniform Law of Large Numbers ($ULLN$) holds for all the average terms above, then by Law of iterated expectation and the definition of $\beta(\theta_0)$, the second summand above is $o_p(1)$. Thus the modified sample covariance is free of the impact of parameter estimation uncertainty, because the replacement of θ_0 by $\hat{\theta}$ does not alter its asymptotic distribution. This feature is used by Wooldridge (1990) to develop moment based asymptotic χ^2 tests that are not subject to the impact of parameter estimation uncertainty.

In the present context, our interest lies in testing whether $\frac{1}{n}\sum_{i=1}^{n}U_i(t,\theta_0)\xrightarrow{p}0$ for all t. We choose $\tau\left(T_i,X_i,\theta\right)=U_i(t,\theta)$ and $\Lambda\left(X_i,\theta\right)=1$. However, there are two key differences from Wooldridge's (1990) original device. First, we need to check for all $t\in[0,1]$ rather than a single or finitely many points of t. Second, Wooldridge (1990) assumes that $\tau\left(T_i,X_i,\theta\right)$ is differentiable with respect to θ, whereas our generalized residual $U_i\left(t,\theta\right)$ is not differentiable with respect to θ because it involves the indicator function. Fortunately, using the expansion in (3.1.1), we can define $G_i(t,\theta)=\frac{\partial}{\partial\theta}F_0[F_0^{-1}(t|X_i,\theta)|X_i]$ analogously, and

$$\hat{\beta}(t,\theta)=\left[\sum_{i=1}^{n}\mathbf{1}[C_i(\theta)>t]G_i(t,\theta)G_i(t,\theta)'\right]^{-1}\sum_{i=1}^{n}\mathbf{1}[C_i(\theta)>t]G_i(t,\theta)$$

is the OLS estimator of regressing unity on $\mathbf{1}[C_i(\theta)>t]G_i(t,\theta)$. Now instead of using $\hat{M}_n(t,\theta)$, we consider the modified empirical process

$$\hat{J}_n(t,\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\mathbf{1}[V_i(\theta) > t] - (1-t)\mathbf{1}[C_i(\theta) > t]}{\hat{S}_c(t,\theta)} \right\} \left\{ 1 - \mathbf{1}[C_i(\theta) > t] G_i(t,\theta)' \hat{\beta}(t,\theta) \right\}, \; t \in [0,1]. \tag{3.1.3}$$

Unlike $\hat{M}_n(t,\hat{\theta})$, $\hat{J}_n(t,\hat{\theta})$ is free of uncertainty impact from parameter estimation asymptotically in the sense that $\hat{J}_n(t,\hat{\theta}) = \hat{J}_n(t,\theta_0) + o_p(1)$. Of course our test statistic is different from Wooldridge's χ^2 tests because we have to take care of the "nuisance parameter" t properly to ensure the global power of our test.

3.2 Test Statistic

To derive the null asymptotic distribution of our test statistic, we impose the following conditions.

Assumption A.4: (a) The conditional cdf $F_0(t|X_i,\theta)$ is continuous and strictly increasing in t, so its inverse function $F_0^{-1}(\cdot,|X_i,\theta)$ exists and is well defined; (b) $F_0(t|X_i,\theta)$ is twice continuously differentiable with respect to $\theta \in \Theta$ with $E \sup_{t\in[0,1]} \sup_{\theta\in\Theta_0} \left\| \frac{\partial}{\partial\theta} F_0(t|X_i,\theta) \right\|^2 \leq \Delta$ for some constant $\Delta < \infty$, where $\Theta_0 \equiv \{\theta \in \Theta : \sqrt{n}\,\|\theta - \theta_0\| \leq \Delta_0\}$ and Δ_0 is a bounded constant; (c) The cdf of censoring variable $\tilde{C}_i$, say $F_{\tilde{C}}(\cdot)$, is continuous.

Assumption A.5: $\hat{\theta}$ is an estimator of θ_0 such that $\sqrt{n}(\hat{\theta} - \theta^*) = O_p(1)$, where $\theta^* \equiv p\lim \hat{\theta}$, and $\theta^* = \theta_0$ under $\mathbb{H}_0$.

Assumption A.4 provides regular smoothness and moment conditions on the conditional lifetime distribution model $F_0(t|X_i,\theta)$ of T_i on X_i. Assumption A.5 does not require any specific estimation method: any $\sqrt{n}$ consistent estimator of θ_0 applies. Examples include MLE, approximate MLE, QMLE and GMM. In particular, we allow but do not require any asymptotically most efficient estimator. Moreover, we need not know the asymptotic expansion structure of $\hat{\theta}$ because $\hat{\theta}$ does not affect asymptotic distribution of the test statistic. These properties greatly simplify the construction and implementation of our test based on $\hat{J}_n(t,\theta)$.

Given Assumption A.5, for any given constant $\varepsilon > 0$, there exists $\Delta_0 \equiv \Delta_0(\varepsilon) < \infty$ such that $P(\sqrt{n}\,\|\theta - \theta^*\| > \Delta_0) < \varepsilon$ for n sufficiently large. Putting $\Theta_0 \equiv \{\theta \in \Theta : \sqrt{n}\,\|\theta - \theta^*\| \leq \Delta_0\}$, we have $\hat{\theta} \in \Theta_0$ with probability approaching 1 as $n \to \infty$.

Theorem 2 considers the impact of parameter estimation uncertainty on empirical processes $\hat{M}_n(t,\hat{\theta})$ and $\hat{J}_n(t,\hat{\theta})$ respectively.

Theorem 2. *Suppose Assumptions A.1 -A.5 hold. Then under H_0, for all $t \in [0,1]$ and $\theta \in \Theta_0$,* *where* $\Theta_0 = \{\theta \in \Theta : \sqrt{n}\,\|\theta - \theta^*\| \leq \Delta_0\}$, *for some constant* Δ_0, *we have*

$$\begin{aligned}\hat{M}_n(t,\theta) &= \hat{M}_n(t,\theta_0) - \frac{1}{S_c(t,\theta)}\bar{g}(t,\theta,\theta_0)'\sqrt{n}(\theta-\theta_0) + o_p(1), \\ \hat{J}_n(t,\theta) &= \hat{J}_n(t,\theta_0) + o_p(1),\end{aligned}$$

where $\bar{g}(t,\theta,\theta_0) = E\left\{\mathbf{1}[C_i(\theta) > t]\frac{\partial}{\partial\theta}F_0[F_0^{-1}(t|X_i,\theta)|X_i]|_{\theta=\theta_0}\right\}$.

Theorem 2 implies that unlike $\hat{M}_n(t,\hat{\theta})$, parameter estimation uncertainty has no impact on the asymptotic distribution of $\hat{J}_n(t,\hat{\theta})$. Asymptotically the factor $1 - \mathbf{1}[C_i(\theta) > t]G_i\,(t,\theta)'\,\hat{\beta}\,(t,\theta)$ removes the nontrivial uncertainty impact of parameter estimation. Now the derivation of limit distribution of our test statistic under $\mathbb{H}_0$ is no longer complicated by the substitution of parameter estimator $\hat{\theta}$ for the unknown true parameter value θ_0. Specifically one can proceed as if θ_0 were known and equal to $\hat{\theta}$. This greatly simplifies the construction and implementation of our test statistic because we need not know the asymptotic expansion of $\hat{\theta}$ and can choose any convenient estimation method that yields a $\sqrt{n}$-consistent parameter estimator.

We can derive the asymptotic distribution of the modified empirical process $\hat{J}_n(t,\hat{\theta})$, as stated in Theorem 3 below.

Theorem 3. Suppose Assumption A.1 - A.5 hold. *Then under* H_0

$$\hat{J}_n(t,\theta_0) \Rightarrow W(t),$$

where $\Rightarrow$ *denotes weak convergence, and* $W(t)$ *is a zero mean Gaussian process with covariance kernel*

$$\begin{aligned}&Cov[W(t),W(s)] \\ =\; &\frac{(1-t\vee s)-(1-t)(1-s)}{S_c(t,\theta_0)S_c(s,\theta_0)}E\{\mathbf{1}[C_i\,(\theta_0) > t\vee s][1-G_i(t,\theta_0)'\beta(t,\theta_0)][1-G_i(s,\theta_0)'\beta(s,\theta_0)]\},\end{aligned}$$

where $\beta(t,\theta_0) = \{E\,(\mathbf{1}[C_i\,(\theta_0) > t]G_i(t,\theta_0)G_i(t,\theta_0)')\}^{-1}E\{\mathbf{1}[C_i\,(\theta_0) > t]G_i(t,\theta_0)\}$.

With Theorem 2 and 3, and the continuous mapping theorem (e.g., Billingsley 1995), we can construct many test statistics based on $\hat{J}_n(t,\hat{\theta})$. Our primary test statistic is defined as follows:

$$GCV = \int_0^1 \hat{J}_n^2(t,\hat{\theta})dt. \qquad (3.2.1)$$

This can be viewed as a *Generalized Cramer-von-Mises (GCV) test.*

We can also define a *Generalized Kolmogorov-Smirnov (GKS) test:*

$$GKS = \sup_{0\leq t\leq 1}\left|\hat{J}_n(t,\hat{\theta})\right|.$$

However, our simulation studies show that GKS has poor size in finite sample. For this reason, we focus on the GCV test in this paper.

The following corollary gives the asymptotic distribution of GCV.

Corollary 1. *Suppose Assumptions A.1 -A.5 hold. Then under H_0, we have*

$$GCV \xrightarrow{d} \int_0^1 W^2(t)\,dt,$$

where $\xrightarrow{d}$ denotes convergence in distribution.

Our GCV reduces to the conventional Cramer-von-Mises statistic (but with the impact of parameter estimation uncertainty properly removed) when there is no censoring. In this special case we have $\mathbf{1}[C_i(\hat{\theta}) > t] = 1$ and $\mathbf{1}[V_i(\hat{\theta}) > t] = \mathbf{1}[T_i(\hat{\theta}) > t]$. It follows that GCV becomes the following form:

$$GCV = \int_0^1 \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\mathbf{1}[T_i(\hat{\theta}) > t] - (1-t)\right\}\left[1 - G_i(t,\hat{\theta})'\hat{\beta}(t,\hat{\theta})\right]\right)^2 dt,$$

where $\hat{\beta}(t,\hat{\theta}) = \left(\sum_{i=1}^n G_i(t,\hat{\theta})G_i(t,\hat{\theta})'\right)\sum_{i=1}^n G_i(t,\hat{\theta})$.

However, it worths noting that the free-of-parameter-impact property does not come "freely". Specifically, $\hat{M}_n(t,\hat{\theta})$ and $\hat{J}_n(t,\hat{\theta})$ are not always asymptotically equivalent in the sense that $\hat{M}_n(t,\hat{\theta}) - \hat{J}_n(t,\hat{\theta}) \xrightarrow{p} 0$ under the null. The asymptotic equivalence between $\hat{M}_n(t,\hat{\theta})$ and $\hat{J}_n(t,\hat{\theta})$ occurs when

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\frac{\mathbf{1}[V_i(\hat{\theta}) > t] - (1-t)\mathbf{1}[C_i(\hat{\theta}) > t]}{\hat{S}_c(t,\hat{\theta})}\right\}\mathbf{1}[C_i(\hat{\theta}) > t]G_i(t,\hat{\theta})'\hat{\beta}(t,\hat{\theta}) = o_p(1).$$

When this condition fails, the tests based on $\hat{M}_n(t,\hat{\theta})$ and $\hat{J}_n(t,\hat{\theta})$ may test misspecification in different directions. This is the price we have to pay by using $\hat{J}_n(t,\hat{\theta})$.

Theorem 3 implies that our test statistic GCV is not asymptotic distribution free (ADF). Before we move on to discuss the resampling method we use for critical values, we first consider how we can potentially get an ADF test in this setting. To derive an ADF test in this setting, we can use the so-called Khmaladze transformation on the appropriate empirical process (Khmaladze 1981, 1993). This transformation has been used by Bai (2003) and Koenker and Xiao (2002) in econometrics. To illustrate the essence of Khmaladzation, we consider the simple case without censoring. Define $\vartheta_n(t,\theta) = \sqrt{n}\left\{\frac{1}{n}\sum_i \mathbf{1}[T_i(\theta) \leq t] - t\right\} = -\hat{M}_n(t,\theta)$. The limiting distribution of $\vartheta_n(t,\hat{\theta})$ is some zero mean Gaussian process $\hat{v}$. Khmaladze's transformation (Khmaladzation hereafter) is performed through three steps: first, we need to transform process $\hat{v}$ to its innovation martingale $\hat{w}$ through the Doob-Meyer transformation (see for example, Fleming and Harrington, 1991);[8] second, $\hat{w}$ is then transformed to a standard Wiener process w (a much easier step than the first one); finally, in the resulting transformation from $\hat{v}$ to w, substitute $\vartheta_n(t,\theta)$ for $\hat{v}$. In the uncensored case, define $g'(t,\theta) = \frac{\partial}{\partial t}\bar{g}(t,\theta)$, then Khmaladzation generates a process

$$w_n(t,\theta) = \vartheta_n(t,\theta) - \int_0^t \left(g'(s,\theta)^T \left[\int_s^1 g'(\tau,\theta)\, g'(\tau,\theta)^T d\tau\right]^{-1} \int_s^1 g'(\tau,\theta) d\vartheta_n(\tau,\theta)\right) ds$$

which has a standard Wiener process limiting distribution. Intuitively, Wooldridge's transformation is a point transformation or reweighting of each observation to purge parameter estimation uncertainty impact, while Khmaladzation is the infinite dimension transformation.[9] As a result, even for this simplest case, we can see that Khmaladzation requires the calculation of stochastic integral, which inevitably imposes much heavier computational burden. When censoring is present, the transformation would involve a composition of two transformations, because $\vartheta_n(t,\theta_0)$ is not the familiar Brownian bridge to start with. Nikabadze and Stute (1997) derive the Khmaladzation formula in the situation when lifetimes follow the same unconditional distribution and random censoring is present. As expected, Khmaladzation in this case is much more intricate, and this mathematical elegancy does not generate easily computable test statistics. Moreover, a necessary and sufficient condition for the existence of innovation process $\hat{w}$ in step 1, is that the functions $1, g_1'(t,\theta), g_2'(t,\theta), ..., g_k'(t,\theta)$ are linearly independent in the neighborhood of 1.[10] Although Tsigroshvili (1998) shows that this condition can be relaxed, a generalized inverse is inevitable whenever this condition fails. On the contrary, to compute our test, we only need to perform the convenient OLS regression of 1 on $\mathbf{1}[C_i(\hat{\theta}) > t]G_i(t,\hat{\theta})$.

[8] In this case, the innovation martingale is some Gaussian process with independent increments (Khmaladze 1981).

[9] We want to point out that, Khmaladzation also incurs some loss of asymptotic power since the transformed process is not always asymptotically equivalent to the original process. The cost is in some sense unavoidable in order to derive a test statistic free of parameter estimation uncertainty impact.

[10] This integer k is the dimension of vector $g'(t,\theta)$.

4 Resampling Method For Critical Values

The asymptotic distribution of GCV is not distribution-free, since it depends on θ_0 and $F_0(\cdot|\cdot,\cdot)$. As a result, asymptotic critical values for GCV cannot be tabulated. We now propose a simple resampling method that can easily generate asymptotically valid critical values for the proposed test statistic.

We first describe our resampling procedure:

(i) Simulate B *i.i.d.* $U[0,1]$ samples, each with size n. The *bth* *i.i.d.* $U[0,1]$ sample is denoted as $\{T_{ib}^*\}_{i=1}^n$ for $b = 1,..,B$.

(ii) Compute the *bth* resample test statistic for GCV, using $\{T_{ib}^*\}_{i=1}^n$ and the original observed data $\{X_i, \tilde{C}_i\}_{i=1}^n$. This resample test statistic is defined as follows:

$$GCV_b^* = \int_0^1 \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\mathbf{1}[C_i(\hat{\theta}) > t]\{\mathbf{1}\left[T_{ib}^* > t\right] - (1-t)\}}{\hat{S}_c(t,\hat{\theta})} \right\} \left[1 - \mathbf{1}[C_i(\hat{\theta}) > t] G_i(t,\hat{\theta})' \hat{\beta}(t,\hat{\theta})\right] \right)^2 dt. \tag{4.1}$$

(iii) Repeat steps (i) and (ii) for $b = 1,...,B$, and obtain a collection of resample test statistics $\{GCV_b^*\}_{b=1}^B$.

(iv) The sample of $\{GCV_b^*\}_{b=1}^B$ mimics random draws from the distribution of GCV under the null hypothesis $\mathbb{H}_0$. Hence, its $(1-\alpha)\,th$ sample percentile yields the critical value for GCV at a prespecified significance level $\alpha \in (0,1)$. This is asymptotically valid if $B \to \infty$ and $n \to \infty$, as is justified in Theorem 4 below.

Theorem 4. *Suppose Assumption A.1-A.5 and H_0 hold. Then for any $b \in \{1,2,...,B\}$, $GCV_b^* \xrightarrow{d} \int_0^1 W^2(t)\,dt$, where $W(t)$ is defined in Theorem 3.*

Note that in resampling, the covariates $\{X_i\}$ and censoring times $\{\tilde{C}_i\}$ are the same as in the observed sample. This is similar to Andrews' (1997) parametric bootstrap. But our resampling method is much more computationally simpler for reasons stated in section 3.1.1. Moreover, since Andrews' (1997) CK statistic is based on the difference between an empirical distribution function and a semiparametric distribution function, his parametric bootstrapping procedure simulates the original dependent variable $\tilde{T}_i^*$ using a parametric conditional distribution function $F_0(t|X_i,\hat{\theta})$ and model re-estimation is needed for each resample data to account for the impact of parameter estimation uncertainty. In contrast, thanks to the use of the probability integral transform, we simply generate the transformed lifetimes T_i^* from a $U[0,1]$ distribution, which is model-free. In addition, we need not re-estimate model parameters in any iteration. As a result, our resampling method is computationally simple.

5 Extensions

Now we will discuss the extensions of our test to several interesting and important scenarios.

5.1 Extension to Duration Models with Unobserved Heterogeneity

Since Lancaster (1979), it has been recognized in the literature that it is often necessary to account for population variations in both observed and unobserved variables (Heckman and Singer 1984b), the latter known as unobserved heterogeneity in duration analysis. Failure to adequately control for population heterogeneity (observed and unobserved) can produce severe bias in structural estimates as well as inferences of duration models. Existence of unobserved heterogeneity is a special case of general model misspecification, and our proposed test developed earlier can detect it. Our interest here is the particular parametric form of the lifetime distribution conditional on both observed and unobserved covariates. Heckman and Singer (1984b) show that empirical parameter estimates of the lifetime duration model conditional on all covariates (both observed and unobserved) are rather sensitive to the distribution specification of the unobservable. However, economic theories rarely suggest a concrete functional form for the unobserved heterogeneity distribution. Estimation methods not specifying the distribution of unobserved heterogeneity have been proposed in the literature (Chesher 1984, Kiefer 1984, Lancaster 1985, and recently, Hausman and Woutersen 2005). Similarly, it will be highly desirable to develop a test for duration models with unobserved heterogeneity that does not assume an unobserved heterogeneity distribution or is robust to any possible misspecification of an unobserved heterogeneity. We now propose such a test.

Assumption A.3*: $\{(X_i, \nu_i, \tilde{T}_i) : i \geqslant 1\}$ is an *i.i.d* sequence with unknown conditional distribution function $F(\cdot|X_i, \nu_i)$ of $\tilde{T}_i$ given X_i and ν_i, where the $\{X_i\}$ are observable covariates while the $\{\nu_i\}$ are unobservable random heterogeneities.

Assumption A.4*: $H(\nu)$ is a prespecified *cdf*.

In practice, the popular choice of the Gamma distribution, or more generally, the exponential family distribution is mainly based on tractability and computational efficiency (Heckman and Singer 1984a, 1984b), since all functions of interest have simple explicit expressions in this case (Lancaster, 1992). Recently Abbring and Van Den Berg (2007) prove that in a large class of hazard models with proportional unobserved heterogeneity, the distribution of the heterogeneity converges to a gamma distribution often at a rapid rate. However, it should be emphasized that the prespecified cdf $H(\nu)$ does not have to be the true distribution function of ν, so our test below is robust to misspecification of the distribution of omitted heterogeneity ν_i.

Our hypotheses of interest are:

$\mathbb{H}_0^*: F(\cdot|X_i,\nu_i) = F_0(\cdot|X_i,\nu_i,\theta_0)$ for some unknown $\theta_0 \in \Theta$ vesus

$\mathbb{H}_A^*: \mathbb{H}_0^*$ is not true.

To extend the test developed earlier, we define

$$\begin{aligned} T_i(\theta|\nu) &= F_0(\tilde{T}_i|X_i,\nu,\theta), \\ C_i(\theta|\nu) &= F_0(\tilde{C}_i|X_i,\nu,\theta), \\ V_i(\theta|\nu) &= F_0(\tilde{V}_i|X_i,\nu,\theta). \end{aligned}$$

Under $\mathbb{H}_0^*$, we have $T_i(\theta_0|\nu_i) \sim i.i.d. U[0,1]$, which implies that $E\{\mathbf{1}[T_i(\theta_0|\nu_i) > t|X_i,\nu_i] = 1-t$. Correspondingly, we have

$$\int_\nu E\{\mathbf{1}[T_i(\theta_0|\nu) > t]|X_i,\nu\}dH(\nu) = 1-t.$$

We can exchange the order of integral and expectation and obtain

$$E\int_\nu \mathbf{1}[T_i(\theta_0|\nu) > t]dH(\nu) = 1-t.$$

This suggests the format of our empirical survivor function with complete observations as follows:

$$\hat{S}_T(t,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \int_\nu \mathbf{1}[T_i(\hat{\theta}|\nu) > t]dH(\nu).$$

When there are censored observations, the survivor function becomes

$$\hat{S}_T(t,\hat{\theta}) = \frac{\frac{1}{n}\sum_{i=1}^n \int_\nu \mathbf{1}[V_i(\hat{\theta}|\nu) > t]dH(\nu)}{\frac{1}{n}\sum_{i=1}^n \int_\nu \mathbf{1}[C_i(\hat{\theta}|\nu) > t]dH(\nu)}.$$

Then our extended GCV test can be defined as follows:

$$GCV^* = \int_0^1 \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\int_\nu \{\mathbf{1}[V_i(\hat{\theta}|\nu) > t] - (1-t)\mathbf{1}[C_i(\hat{\theta}|\nu) > t]\}dH(\nu)}{\frac{1}{n}\sum_{i=1}^n \int_\nu \mathbf{1}[C_i(\hat{\theta}|\nu) > t]dH(\nu)} \right\} \left\{1 - G_i^*(t,\hat{\theta})'\hat{\beta}^*(t,\hat{\theta})\right\} \right|^2 dt, \tag{5.1}$$

where

$$
\begin{aligned}
G_i^*(t,\theta) &= \int_\nu \mathbf{1}[C_i(\hat{\theta}|\nu) > t] \frac{\partial}{\partial\theta} F_0[F_0^{-1}(t|X_i,\nu,\theta)|X_i,\nu]dH(\nu), \\
\hat{\beta}^*(t,\theta) &= \left[\sum_{i=1}^n G_i^*(t,\theta)G_i^*(t,\theta)'\right]^{-1} \sum_{i=1}^n G_i^*(t,\theta).
\end{aligned}
$$

In practice, given the result of Abbring and Van Den Berg (2007), one can use Gamma distribution for omitted heterogeneity in the estimation procedure without worrying that the parameter estimates are too sensitive to the specification of unobservable heterogeneity But the choice of $H(\cdot)$ in the testing procedure is rather flexible due to Assumption A.4* and how we construct GCV^*. Such flexibility makes our test robust to misspecified omitted heterogeneity and easily extends its applicability to this often encountered scenario.

5.2 Extension to Duration Models with Competing Risks

Competing risk models arise when failure arises in different ways or for different reasons. For example, an unemployment spell can end with a new job, or a recall from previous job, or withdrawal from the labor force (Kiefer 1988, Lancaster 1992). In statistics, there is a well known nonidentification theorem proved by Cox and Tsiatis (Kalbfleisch and Prentice 1980, Lawless 2003), which states that "for any joint distribution of the latent failure times there exists a joint distribution with independent failure times which gives the same distribution of the identified minimum", and it has "led much empirical work on multistate duration models to be conducted within an independent risks paradigm" (Heckman and Honore 1989). However, this theorem applies to settings where covariates are absent. In social science settings where covariates are more common than not, some researchers have shown identifiable results under certain conditions (Han and Hausman 1986, Heckman and Honore 1989). Nevertheless, independent risks remain popular in empirical studies (e.g., Katz 1986, Katz and Meyer 1990, Idson and Valletta 1996, Wheelock and Wilson 2000). This is because on one hand, even though interdependent risks are more plausible, there are some evidences that the independence hypothesis cannot be rejected by data (Han and Hausman 1990, Fallick 1993); on the other hand, identification might be demanding in terms of the amount of data required. Given this, we restrict our attention to independent competing risks in this section.

Assumption A.3:** There are M types of causes for a failure on individual i. Let $\tilde{T}_{qi}$ be the type q latent failure time and let X_i be observable covariates for individual i. The sample $\{(X_i, \tilde{T}_{1i}, ..., \tilde{T}_{Mi}) :$

$i \geqslant 1\}$ is an *i.i.d* sequence with unknown conditional distribution functions $F^q(\cdot|X_i)$ of $\tilde{T}_{qi}$ given X_i, for $q = 1, ..., M$. The failure times $\{\tilde{T}_{qi}, q = 1, ..., M\}$ and censoring times $\{\tilde{C}_i\}$ are mutually independent of each other given X_i.

Under this conditional independent competing risk framework, one may be interested in testing parametric specification for one specific type of failure, say type $m \in \{1, 2, ..., M\}$. That is, whether the failure of type m follows a parametric conditional distribution specification $F^m(\cdot|X_i, \theta)$. Formally, our hypotheses of interest are:

$\mathbb{H}_0^{**} : F^m(\cdot|X_i) = F_0^m(\cdot|X_i, \theta_0)$ for some unknown $\theta_0 \in \Theta$ vesus

$\mathbb{H}_A^{**} : \mathbb{H}_0^{**}$ is not true.

Here, the failure times of types $1, 2, ..., m-1, m+1, ..., M$ are treated as the censoring times for type m. To apply our method developed earlier, we need to transform the original data by the conditional parametric probability distribution model $F_0^m(\cdot|X_i, \theta)$ of type m under $\mathbb{H}_0^{**}$. Define

$$\begin{aligned} V_i^m(\theta) &= \min\left[T_{1i}^m(\theta), ..., T_{Mi}^m(\theta), C_i(\theta)\right], \\ C_i^m(\theta) &= \min[T_{1i}^m(\theta), ..., T_{(m-1)i}^m(\theta), T_{(m+1)i}^m(\theta), ..., T_{Mi}^m(\theta), C_i(\theta)], \end{aligned}$$

where $T_{qi}^m(\theta) = F_0^m(\tilde{T}_{qi}|X_i, \theta)$, for $q = 1, ..., M$ and $C_i(\theta) = F_0^m(\tilde{C}_i|X_i, \theta)$. Also define the sample survivor function $\hat{S}_{c^m}(t, \theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[C_i^m(\theta) > t]$. Then our *GCV* test can be defined as follows:

$$GCV^m = \int_0^1 \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\mathbf{1}[V_i^m(\hat{\theta}) > t] - (1-t)\mathbf{1}[C_i^m(\hat{\theta}) > t]}{\hat{S}_{c^m}(t, \hat{\theta})} \right\} \left[1 - \mathbf{1}[C_i^m(\hat{\theta}) > t] G_i^m(t, \hat{\theta})' \hat{\beta}^m(t, \hat{\theta})\right] \right|^2 dt, \tag{5.2}$$

where

$$\begin{aligned} G_i^m(t, \theta) &= \frac{\partial}{\partial \theta} F_0^m[F_0^{m^{-1}}(t|X_i, \theta)|X_i, \theta], \\ \hat{\beta}^m(t, \theta) &= \left[\sum_{i=1}^n \mathbf{1}[C_i^m(\theta) > t] G_i^m(t, \theta) G_i^m(t, \theta)'\right]^{-1} \sum_{i=1}^n \mathbf{1}[C_i^m(\theta) > t] G_i^m(t, \theta). \end{aligned}$$

6 Finite Sample Performance

We now investigate the finite sample performance of the *GCV* test, in comparison with three existing popular tests, namely *RM*, *LGP*, and the *LM* test for heterogeneity, with application to testing the null hypothesis of a conditional exponential distributed duration with censored observations. Prieger (2000) calculates the explicit forms of *RM*, *LGP* and *LM* tests for this case, which we follow here. Since the first few moments are usually of special interest, we choose the second and third moment conditions to perform the moment tests.[11]

6.1 Simulation Design

6.1.1 Size

To investigate sizes of tests, we consider the following Data Generating Processes (hereafter DGP):

· *DGP1:*

$$\tilde{T}_i|X_i \sim \textit{Exponential distribution} \text{ with pdf } f(t|X_i) = \mu_i \exp(-\mu_i t),$$

where $\mu_i = \exp[-(X_{1i} + 2X_{2i})], X_i = (X_{1i,}X_{2i})', X_{1i} \sim i.i.d.N(0,1), X_{2i} \sim i.i.d.N(0,1)$, and X_{1i} and X_{2i} are mutually independent.

We evaluate the sizes of tests under different degrees of random censoring, checking whether censoring distorts sizes, and if so, to what extent. For simplicity, we use an independent random censoring scheme, $C_i \sim \textit{Exponential Distribution}$ with different means to generate desirable censoring percentages: 0%, around 10% and around 20% respectively. Under the null hypothesis of a conditional exponential distribution, we estimate a null conditional exponential distributed duration model via MLE and calculate test statistics with the parameter estimates. For GCV, we use the resampling method described in Section 4 to obtain critical values, with the resampling iteration number $B = 100$. For RM, LGP and LM tests, we use both asymptotic critical values and bootstrap critical values. The bootstrap for the latter is conducted as follows. First, we simulate B $(B = 100)$ bootstrap samples, each of size n. In the bth sample, covariates and censoring time are the same as in the real data, *i.e.*, $(X_{ib}, \tilde{C}_{ib}) = (X_i, \tilde{C}_i)$; lifetime $\tilde{T}_{ib}$ is simulated using the null distribution $F_0(\cdot \mid X_i, \hat{\theta})$, where $\hat{\theta}$ is the parameter estimate based on the real data. Then we estimate the null model using the bootstrap sample $(\tilde{T}_{ib}, X_i, \tilde{C}_i)$ and compute test statistics for the bth bootstrap sample. The sample $\{RM_b\}, \{LGP_b\}, \{LM_b\}$ mimic random draws from the distributions of RM, LGP and LM under the null hypothesis. Hence their $(1-\alpha)th$ sample percentiles yield the critical values of RM, LGP and LM respectively at significance level α. Because the bootstrap takes into account the impact of parameter estimate, moment tests using bootstrap critical values can compare fairly with our test. Five different sample sizes are considered: $n = 100, 200, 300, 400, 500$. The number of Monte Carlo trials in all cases is 1000.

[11] The first moment restriction is automatically satisfied from the likelihood equations (Kiefer 1988, Lancaster 1992, Prieger 2000).

6.1.2 Power

We also examine power of tests for neglected heterogeneity and misspecification of duration distribution respectively. The DGPs are as follows:

· *DGP2* (Omitted Heterogeneity):

$$\tilde{T}_i|X_i \sim \text{exponential distribution with pdf } f(t|X_i) = \mu_i \exp(-\mu_i t),$$

where $\mu_i = \exp[-(X_{1i} + 2X_{2i} + X_{1i}^2)], X_i = (X_{1i}, X_{2i})', X_{1i} \sim i.i.d.U[0,1], X_{2i} \sim i.i.d.U[0,1]$ and X_{1i} and X_{2i} are mutually independent.

· *DGP3* (Misspecification of Duration Density):

$$\tilde{T}_i|X_i \sim \text{lognormal distribution with pdf } f(t|X_i) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\left[-\frac{1}{2}\left(\frac{\log t - \mu_i}{\sigma}\right)^2\right],$$

where $\mu_i = \exp[-(X_{1i} + 2X_{2i})], X_i = (X_{1i}, X_{2i}), X_{1i} \sim i.i.d.N(0,1), X_{2i} \sim i.i.d.N(0,1), \sigma = 0.8$, and X_{1i} and X_{2i} are mutually independent.

In both cases, we use MLE to estimate the null conditional exponential distributed duration model:

$$\tilde{T}_i|X_i \sim \text{exponential distribution with pdf } f(t|X_i, \theta) = \lambda_i \exp(-\lambda_i t),]$$

where $\lambda_i = \exp[-(\alpha_1 X_{1i} + \alpha_2 X_{2i})]$ and $\theta = (\alpha_1, \alpha_2)'$.

$DGP2$ is designed for power comparison among all tests against omitted heterogeneity. In this scenario, hypotheses are nested, so all tests are applicable. We use both empirical critical values and bootstrap/resampling critical values. To obtain the empirical critical values, we first generate $\{X_i\}, \{\tilde{C}_i\}$ according to the design in $DGP2$, and $\{\tilde{T}_i\}$ according to the null model, then we use this data to estimate the null duration model, and use the parameter estimates and the data to compute test statistics. After we repeat the above procedure for 1000 times, we can rank these 1000 test statistics, and the $1000(1-\alpha)$ percentile gives the corresponding Empirical Critical Value (ECV) at significance level α. We then generate $\{X_i\}, \{\tilde{C}_i\}$ and $\{\tilde{T}_i\}$ under $DGP2$, estimate the null model with the data, and compute test statistics with the parameter estimates. The decision rule is to compare these test statistics with the corresponding ECV. Empirical critical values provide a fair comparison of powers among different tests. However, empirical critical values are not applicable in practice, because the $DGPs$ for $\{X_i, \tilde{C}_i\}$ are unknown. Therefore we also conduct power studies using bootstrap critical values for RM, LGP and LM tests and resampling critical values for GCV, which are always feasible in practice.

Under $DGP3$, there exists misspecification in the conditional duration density. In this case, the LM test for heterogeneity is no longer applicable because the design only accommodates the omitted heterogeneity cases. Therefore we only compare the power of LGP and RM tests with that of GCV, using both ECV and bootstrap/resampling critical values.

For both $DGP2$ and $DGP3$, we are interested in studying the impact of censoring on power of tests. Different degrees of censoring are generated by the same method as in the size study. In all cases, the bootstrap and resampling iteration numbers $B = 100$, and the number of Monte Carlo trials is equal to 500. Since $DGP2$ is designed as a close alternative to the null hypothesis (with the omitted squared term of X_{1i}^2, where $X_{1i} \sim i.i.d.U[0,1]$), we consider the sample size $n = 100, 200, 300, 500, 2000$. For $DGP3$, we use the sample size $n = 100, 200, 300, 400, 500$.

6.2 Monte Carlo Evidences

Table I reports the empirical rejection rates of the tests under $\mathbb{H}_0$ at the 0.05 and 0.10 significance levels. For the GCV test, its empirical size is close to its nominal level even for the sample size n as small as 100. It is also robust to different degrees of censoring. On the other hand, none of the moment tests give reasonable sizes when asymptotic critical values are used. Specifically, the RM test excessively overrejects at both levels, although there is some tendency that its empirical null rejection probabilities get closer to its nominal levels gradually as the sample size n increases. This is due to the fact that RM converges very slowly (Prieger 2000). Not surprisingly, LM underrejects in all cases since the theoretical information matrices are not available (Jaggia 1997); When there is no censoring, LGP underrejects, although not dramatically for all sample sizes, but it seems that its empirical sizes converge to its nominal levels as the sample size increases. However, censoring vastly distorts the sizes of the LGP test: in fact, the empirical null rejection probabilities are 0 everywhere, implying invalid sizes. This is because whenever there is censoring, the modified Laguerre polynomials are no longer orthogonal with respect to the censored exponential distribution (Prieger 2000), discounting the validity of the test. The last six columns of Table I report the empirical sizes of RM, LGP and LM tests using bootstrap critical values. Once bootstrap critical values are adopted, sizes are noticeably improved for all moment tests at all sample sizes and censoring percentages, as explained by Horowitz (1994). In particular, for LGP and LM tests, empirical sizes are close to nominal levels, while the RM test still shows under rejection in most cases. However, this is achieved at the price of computation burden. On average, it takes at least 3 times longer to run a bootstrap moment test than to run our test through resampling.

Table II reports the powers of all tests at the 0.05 and 0.10 significance levels under $DGP2$, using empirical critical values. Apparently the LGP and LM tests for heterogeneity have little power detecting this close alternative and large sample sizes do not boost their powers.[12] The RM test demonstrates a slow power improvement with increasing sample sizes. At the largest sample size $n = 2000$ we consider, the empirical power for the RM test is roughly around 0.3 at the 0.05 level and 0.5 at the 0.10 level . Our GCV test is the most powerful for detecting this omitted heterogeneity. Its power improves significantly as the sample size increases. For example, for $n = 2000$, and under the uncensored case, the rejection rates for GCV are 0.696 and 0.796 at the 0.05 and 0.10 levels respectively.

Table III reports the powers of all tests at the 0.05 and 0.10 levels under $DGP2$, using bootstrap/resampling critical values. The power pattern is similar to the one based on the empirical critical values in Table II.

Table IV reports the empirical powers of GCV, LGP and RM tests at the 0.05 and 0.10 levels under $DGP3$, using empirical critical values. In this scenario, the LM test for heterogeneity is not applicable. GCV again has the highest power for all sample sizes and censoring levels. GCV achieves unit power when $n \geq 400$ at all censoring levels. The LGP test also has good power. As the sample size n increases, its power approaches unit gradually. In comparison, the RM test is the least powerful for detecting this density misspecification. In all cases, its power never exceeds 0.14 and there is no evidence that increasing the sample size improves its power.

Table V reports the powers under $DGP3$ using bootstrap/resampling critical values. Again, the power pattern is similar to that in Table IV.

Overall, our GCV test has a great finite sample performance. The empirical sizes of GCV are close to its nominal levels, and it has the highest power for two alternatives considered. In comparison, the popular moment tests, LGP, LM for heterogeneity and RM tests have invalid sizes when asymptotic critical values are used. Their sizes are corrected and become reasonable when bootstrap critical values are used, but the corresponding computing programs take at least 3 times longer to run. In terms of power studies, the LGP test has good power against the misspecified density, while no power against the neglected heterogeneity. In addition, once censoring is involved, the modified polynomials are no

[12] Moreover, the calculation of moment conditions for LGP and LM is very tedious. For example, the 2nd and 3rd moments for LGP test are:
$\lambda_2 = \frac{1}{n}\sum_i 0.5[\hat{\varepsilon}_i^2 - 4\hat{\varepsilon}_i + 2 + 2(1-\delta_i)(\hat{\varepsilon}_i - 1)]$
$\lambda_3 = \frac{1}{n}\sum_i \frac{1}{6}[-\hat{\varepsilon}_i^3 + 9\hat{\varepsilon}_i^2 - 18\hat{\varepsilon}_i + 6 + 3(1-\delta_i)(-\hat{\varepsilon}_i^2 + 4\hat{\varepsilon}_i - 2)]$;
And the 2nd and 3rd moments for LM test are:
$s_2 = \frac{1}{n}\sum_i 0.5(\hat{\varepsilon}_i^2 - 2\delta_i\hat{\varepsilon}_i)$
$s_3 = \frac{1}{n}\sum_i -\frac{1}{6}(\hat{\varepsilon}_i^3 - 3\delta_i\hat{\varepsilon}_i^2)$;
where $\hat{\varepsilon}_i = \tilde{V}_i \exp(-X_i'\hat{\beta})$ for our null conditional exponential duration model.

longer orthogonal with respect to null weighting function (Prieger 2000), implying that the advantage of easy computation for LGP is lost in censoring cases. The LM test for heterogeneity fails to detect the omitted heterogeneity. Moreover, the applicability of both LGP and LM tests is limited: the LM for heterogeneity does not go beyond omitted heterogeneity testing, and the LGP test does not apply to any non-nested hypotheses. The RM test, on the other hand, has the applicability as wide as our GCV test. However, not surprisingly, it has limited power for certain alternatives, like the misspecified density in $DGP3$. This illustrates the limit of moment based tests.

7 Conclusion

Duration models with censoring have continuously attracted attentions in economics, finance and other fields. Their increasing popularity has been accompanied by the potential increase in "model risk", which can only be reduced by model checking. In this paper, we have proposed a generalized residual based goodness-of-fit test, which has a number of significant advantages over the existing approaches. Firstly, the existing duration literature only develops moment-based tests, while our approach takes a more comprehensive approach by inspecting the conditional duration distribution. Secondly, some existing methods fail to incorporate the censoring information, while our test incorporates all available information from complete and censored observations. This is achieved through a novel simple survivor function that is applicable to both censored and uncensored observations. Lastly, our approach does not require any specific estimation method and parameter estimation uncertainty does not affect the asymptotic distribution of the proposed test statistic, thanks to adopting a purging device of Wooldridge's (1990) type. We propose a simple resampling method to obtain the critical values of our test, and discuss the extension to accommodate unobserved heterogeneity and competing risks. The finite sample performance of the proposed test is assessed via simulation studies in comparison with a number of popular existing tests, and our test has a nice finite sample performance (both in terms of size and power) and computational advantage.

References

Abbring, J.H. and G. Van Den Berg. The Unobserved Heterogeneity Distribution in Duration Analysis. *Biometrika 94(2007)*: 87-99.

Addison, J.T. and P. Portugal. On the Distribution Shape of Unemployment Duration. *The Review of Economics and Statistics 69(1987)*: 520-526.

Allison, P. Event History Analysis : Regression for Longitudinal Event Data (Quantitative Applications in the Social Sciences). *Sage Publication, Inc. (1984).*

Anderson, P.K., O. Borgan, R. D. Gill and N. Keiding. Statistical Models Based on Counting Processes (Springer Series in Statistics). *Springer; 1 edition (1996).*

Andrews, D. W.K. A Conditional Kolmogorov Test. *Econometrica 65 (1997)*: 1097-1128.

Bai, J. Testing Parametric Conditional Distributions of Dynamic Models. *Review of Economics and Statistics 85 (2003)*: 531-549.

Barnett, W.P. The Dynamics of Competitive Intensity. *Administrative Science Quarterly 42 (1997)*: 128-160.

Beran, R. Nonparametric Regression with Randomly Censored Survival Data. *University of California, Berkeley working paper (1981).*

Bijwaarda, G.E. and G. Ridder. Correcting for selective compliance in a re-employment bonus experiment. *Journal of Econometrics 125 (2005)*: 77-111.

Billingsley, P. Convergence of Probability Measures (Wiley Series in Probability and Statistics). *Wiley-Interscience; 2 edition (1999).*

Billingsley, P. Probability and Measure (Wiley Series in Probability and Mathematical Statistics). *Wiley-Interscience; 3 edition (1995).*

Blossfeld, H.P. and G. Rohwer. Techniques of event history modeling: New approaches to causal analysis. *Lawrence Erlbaum Associates, 2nd edition* (2001).

Bharath, Sreedhar and T. Shumway. Forecasting Default with KMV-Merton Model. *Working Paper, University of Michigan* (2004)

Carroll, G.R. and M.T. Hannan. Density dependence in the evolution of populations of newspaper organizations. *American Sociological Review 54 (1989)*: 524-541.

Chava, S. and R. Jarrow. Bankruptcy Prediction with Industry Effects. *Review of Finance (2004)*: 537-569.

Chesher, A.D. Testing for Neglected Heterogeneity. *Econometrica 52 (1984):* 865-872.

Cox, D.R. and E.J. Snell. A General Definition of Residuals. *Journal of Royal Statistical Society*

B. 2 (1968): 248-275.

Cox, D.R. and E.J. Snell. On Test Statistics Calculated from Residuals. *Biometrica 58 (1971):* 589-594.

Cramer, H. Mathematical Methods of Statistics. *Princeton University Press (1964).*

Duffie, Darrell, L. Saita and K. Wang. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics (2006)*: forthcoming.

Easley, D., and M. O'Hara. Time and the Process of Security Price Adjustment, *Journal of Finance 47 (1992)*, 577-605.

Engle, R., and J. R. Russell. Autoregressive Conditional Duration: A new Model for Irregularly Spaced Data, *Econometrica 66 (1998)*, 1127-1162.

Engle, R. The Econometrics of High Frequency Data, *Econometrica 68 (2000)*, 1-22.

Fallick, B.C. The Industrial Mobility of Displaced Workers. *Journal of Labor Economics 11 (1993):* 302-323.

Fleming, T.R. and D.P. Harrington. Counting Processes and Survival Analysis. *Wiley-Interscience, 1 edition (1991).*

DuWors Jr., R.E. and G.H. Haines Jr. Event history analysis measures of brand loyalty. *Journal of Marketing Research 27 (1990)*: 485-493

Gray, R.J. and D.A. Pierce. Goodness-of-Fit Tests for Censored Survival Data. *The Annals of Statistics 13 (1985)*: 552-563.

Han, A. and J. Hausman. Identification of Continuous and Discrete Competing Risk Models. *MIT Working Paper (1986).*

Han, A. and J. Hausman. Flexible Parametric Estimation of Duration and Competing Risk Models. *Journal of Applied Econometrics 5 (1990)*: 1-28.

Hausman, J. and T. Woutersen. Estimating a Semi-Parametric Duration Model without Specifying Heterogeneity. *MIT Economics Working Papers (2005).*

Haveman, H.A. Between a rock and a hard place: Organizational change and performance under conditions of fundamental environmental transformation. *Administrative Science Quarterly 37 (1992)*: 48-75.

Heckman, J.J. and B.E. Honore. The Identifiability of the Competing Risks Model. *Biometrika 76 (1989):* 325-330.

Heckman, J.J. and B. Singer. Econometric Duration Analysis. *Journal of Econometrics 24 (1984a)*: 63-132.

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

Hochguertel, S. and A. van Soest. The Relation between Financial and Housing Wealth: Evidence from Dutch Households. *Journal of Urban Economics 49(2001)*: 374-403.

Hoem, J.M. Distortions Caused by Nonobservation of Periods of Cohabitation Before the Latest. *Demography 20 (1983)*: 491-506.

Horowitz, J. L. Bootstrap-based Critical Values for the Information Matrix Test. *Journal of Econometrics 61 (1994):* 395-411.

Horowitz, J. L. and G. Neumann. Specification Testing in Censored Regression Models: Parametric and Semiparametric Methods. *Journal of Applied Econometrics 4 (1989):* s61-s89.

Idson, T.L. and R.G. Valletta. Seniority, Sectoral Decline, and Employment Retension: An Analysis of Layoff Unemployment Spells. *Journal of Labor Economics 14 (1996):* 654-676.

Jaggia, S. Alternative Forms of the Score Test for Heterogeneity in a Censored Exponential Model. *The Review of Economics and Statistics 79 (1997):* 113-119.

Kalbfleisch, J.D. and R.L. Prentice. Statistical Analysis of Failure Time Data. *John Wiley & Sons, Inc. (1980).*

Katz, L.F. Layoffs, Recall and the Duration of Unemployment. *NBER Working Paper No. 1825 (1986).*

Katz, L.F. and B.D. Meyer. Unemployment Insurance, Recall Expectation and Unemployment Outcomes. *The Quarterly Journal of Economics 105 (1990):* 973-1002.

Khmaladze, E.V. Martingale Approach in the Theory of Goodness-of-fit Tests. *Theory of Probability and its Applications 26 (1981):* 240-257.

Khmaladze, E.V. Goodness of Fit Problem and Scanning Innovation Martingales. *Annals of Statistics (1993):* 798-829.

Kiefer, N.M. A Simple Test for Heterogeneity in Exponential Models of Duration. *Journal of Labor Economics 2 (1984):* 539-549.

Kiefer, N.M. Specification Diagnostics based on Laguerre Alternatives for Econometric models of Duration. *Journal of Econometrics 28 (1985):* 135-154.

Kiefer, N.M. Economic Duration Data and Hazard Functions. *Journal of Economic Literature 26 (1988):* 646-679.

Koenker, R. and Z. Xiao. Inference on the Quantile Regression Process. *Econometrica 70(2002):* 1583-1612.

Lancaster, T. Econometric Methods for the Duration of Unemployment. *Econometrica 47(1979):* 939-956.

Lancaster, T. Generalised Residuals and Heterogeneous Duration Models with Applications to the Weibull Model. *Journal of Econometrics 28(1985):* 155-169.

Lancaster, T. The Econometric Analysis of Transition Data (Econometric Society Monographs). *Cambridge University Press, New Ed edition (1992).*

Lancaster, T. and A.D. Chesher. Residual Analysis for Censored Duration Data. *Economic Letters 12(1985):* 723-725.

Lawless, J. F. Statistical models and methods for lifetime data (Wiley Series in Probability and Statistics). *Wiley-Interscience; 2 edition (2003).*

Loynes, R.M. The Empirical Distribution Function of Residuals from Generalised Regression. *The Annals of Statistics 8(1980)*: 285-198.

Manning, W.D. Cohabitation, Marriage, and Entry into Motherhood. *Journal of Marriage and the Family 57 (1995)*: 191-200.

Michael, R.T. and N.B. Tuma. Entry Into Marriage and Parenthood by Young Men and Women: The Influence of Family Background. *Demography 22 (1985)*: 515-544.

Manton, K.G., E. Stallard and J.W. Vaupel. Alternative Models for the Heterogeneity of Mortality Risks among the Aged. *Journal of American Statistical Society 81 (1986): 635-644.*

Monahan, T.P. When Married Couples Part: Statistical Trends and Relationships in Divorce. *American Sociological Review 27 (1963)*: 625-633.

Nelson, W.B. Applied Life Data Analysis (Wiley Series in Probability and Statistics). *Wiley; 1 edition (1982).*

Nikabadze, A and W. Stute. Model Checks Under Random Censorship. *Statistics and Probability Letters 32 (1997):* 249-259

O'Hagan, A and J.W. Stevens. On estimators of medical costs with censored data. *Journal of Health Economics 23(2004)*: 615-625.

Orbe, J., E. Ferreira and V. Nunez-Anton. Modelling the duration of firms in Chapter 11 bankruptcy using a flexible model. *Economic Letters 71(2001)*: 35-42.

Prieger, J.E. Conditional Moment Tests for Parametric Duration Models. *U.C. Davis Working Paper (2000).*

Rao,J.S. and J. Sethuraman. Weak Convergence of Empirical Distribution Functions of Random Variables Subject to Perturbations and Scale Factors. *The Annals of Statistics 3 (1975):* 299-313.

Rosenblatt, M. Remarks on a Multivariate Transformation. *Annals of Mathematical Statistics 23 (1952):* 470-472.

Roszbach, Kasper. Bank Lending Policy, Credit Scoring, and the Survival of Loans. *The Review of Economics and Statistics 84(2004)*: 946-958.

Sharma, S. Specification Diagnostics for Econometric Models of Duration. *UCLA Economics Working Papers 440 (1987).*

Shumway, T. Forecasting Bankruptcy more accurately: a Simple Hazard Model. *Journal of Business (2001)*: 101-124.

Sun, Y. Weak Convergence of the Generalized Parametric Empirical Processes and Goodness-of-fit tests for Parametric Models. *Communications in Statistics: Theory and Methods 26 (1997):* 2393-2413.

Tsigroshvili, Z. Some Notes on Goodness-of-fit Tests and Innovation Martingales. *Proceedings of A. Razmadze Mathematical Institute117 (1998):* 89-102

White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica 50 (1982):* 1-25.

Wheelock, D.C. and P.W. Wilson. Why do Banks Disappear? The Determinants of U.S. Bank Failures and Acquisitions. *The Review of Economics and Statistics 82 (2000):* 127-138.

Wooldridge, J.M. A Unified Approach to Robust, Regression-based Specification Tests. *Econometric Theory 6 (1990):* 17-43.

Table I: Empirical Size

	GCV		LM2, acv		LGP2, acv		RM2, acv		LM2, bp		LGP2, bp		RM2, bp	
	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
n=100														
Censor=0%	0.058	0.106	0.022	0.034	0.025	0.052	0.407	0.459	0.052	0.11	0.057	0.112	0.07	0.124
Censor=11.49%	0.054	0.112	0.013	0.02	0.00	0.00	0.424	0.474	0.038	0.087	0.053	0.096	0.038	0.098
Censor=20.74%	0.058	0.109	0.017	0.02	0.00	0.00	0.388	0.441	0.059	0.085	0.073	0.126	0.05	0.096
n=200														
Censor=0%	0.06	0.127	0.02	0.037	0.041	0.07	0.298	0.351	0.053	0.108	0.057	0.117	0.03	0.064
Censor=11.61%	0.063	0.116	0.022	0.03	0.00	0.00	0.287	0.358	0.048	0.092	0.054	0.109	0.06	0.082
Censor=20.85%	0.059	0.104	0.018	0.026	0.00	0.00	0.273	0.332	0.046	0.07	0.054	0.101	0.042	0.088
n=300														
Censor=0%	0.066	0.122	0.028	0.04	0.043	0.076	0.239	0.296	0.063	0.122	0.064	0.113	0.04	0.082
Censor=11.61%	0.054	0.10	0.017	0.03	0.00	0.00	0.255	0.309	0.062	0.105	0.07	0.12	0.056	0.104
Censor=20.81%	0.053	0.106	0.019	0.028	0.00	0.00	0.245	0.307	0.04	0.066	0.065	0.117	0.036	0.07
n=400														
Censor=0%	0.069	0.121	0.027	0.035	0.042	0.064	0.227	0.282	0.069	0.115	0.067	0.115	0.016	0.05
Censor=11.56%	0.044	0.098	0.023	0.033	0.00	0.00	0.211	0.273	0.048	0.086	0.069	0.124	0.06	0.102
Censor=20.78%	0.058	0.102	0.022	0.028	0.00	0.00	0.221	0.271	0.032	0.069	0.058	0.116	0.042	0.082
n=500														
Censor=0%	0.065	0.106	0.022	0.037	0.06	0.086	0.195	0.26	0.062	0.106	0.053	0.105	0.008	0.044
Censor=11.57%	0.057	0.107	0.022	0.034	0.00	0.00	0.189	0.25	0.057	0.107	0.062	0.106	0.04	0.068
Censor=20.86%	0.055	0.096	0.015	0.025	0.00	0.00	0.207	0.268	0.034	0.067	0.054	0.105	0.03	0.068

Note: Iteration number k=1000, Bootstrap Iteration number B=100, Sample size n=100, 200, 300, 400, 500
GCV—Generalized Cramer-von Mise Statistic
LM 2 – LM test for Heterogeneity, 2nd and 3rd moments used
LGP 2 – Laguerre-based test, 2nd and 3rd moments used
RM2 – Raw Moment test, 2nd and 3rd moments used
acv: asymptotic critical value; bp: bootstrap critical values

DGP: X1～N(0,1); X2～N(0,1)
μ_i=exp[-(x_{1i}+2x_{2i})]
Exponential pdf for lifetime T_i: $\mu_i exp(-\mu_i t)$.

Table II: Empirical Power over Neglected Heterogeneity (ECV)

	GCV, ecv		LM 2, ecv		LGP 2, ecv		RM 2, ecv	
	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
n=100								
Censor=0%	0.078	0.126	0	0	0.048	0.084	0.07	0.126
Censor=10.43%	0.06	0.124	0.036	0.072	0.044	0.086	0.052	0.086
Censor=20.10%	0.094	0.15	0.03	0.054	0.028	0.072	0.038	0.1
n=200								
Censor=0%	0.082	0.166	0	0	0.052	0.09	0.04	0.092
Censor=10.31%	0.134	0.24	0.018	0.032	0.042	0.094	0.064	0.11
Censor=20.14%	0.108	0.164	0.022	0.044	0.036	0.068	0.056	0.106
n=300								
Censor=0%	0.2	0.286	0	0	0.044	0.1	0.036	0.092
Censor=10.4%	0.134	0.232	0.014	0.046	0.034	0.09	0.052	0.132
Censor=20.02%	0.14	0.238	0.036	0.066	0.032	0.078	0.056	0.106
n=500								
Censor=0%	0.214	0.322	0	0	0.08	0.116	0.058	0.146
Censor=10.35%	0.178	0.31	0.018	0.062	0.038	0.1	0.054	0.104
Censor=20.09%	0.186	0.31	0.022	0.062	0.032	0.05	0.082	0.174
n=2000								
Censor=0%	0.696	0.796	0	0	0.066	0.118	0.292	0.478
Censor=10.33%	0.608	0.732	0.02	0.06	0.032	0.06	0.316	0.486
Censor=20.02%	0.532	0.662	0.034	0.078	0.032	0.05	0.298	0.486

Note: Iteration number k=500, ECV Replication number B=1000, Sample size n=100, 200, 300, 500, 2000. ecv: empirical critical value.

Simulation DGP:

X1～U[0,1]; X2～U[0,1]

$\mu_i=\exp[-(x_{1i}+2x_{2i}+x_{1i}^2)]$

Exponential pdf for lifetime T_i: $\mu_i exp(-\mu_i t)$.

Model (H$_0$):

$\lambda_i=\exp[-(\alpha_1 x_{1i}+\alpha_2 x_{2i})]$

Exponential pdf for lifetime y_i: $\lambda_i exp(-\lambda_i t)$.

Table III: Empirical Power over Neglected Heterogeneity (Bootstrap)

	GCV		LM 2, bp		LGP 2, bp		RM 2, bp	
	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
n=100								
Censor=0%	0.08	0.164	0.056	0.102	0.058	0.102	0.062	0.114
Censor=10.24%	0.084	0.136	0.052	0.098	0.058	0.108	0.068	0.1
Censor=20.08%	0.07	0.138	0.048	0.08	0.062	0.088	0.048	0.108
n=200								
Censor=0%	0.136	0.22	0.052	0.114	0.05	0.12	0.072	0.112
Censor=10.41%	0.1	0.158	0.042	0.106	0.03	0.078	0.056	0.112
Censor=20.24%	0.1	0.14	0.024	0.076	0.032	0.062	0.078	0.126
n=300								
Censor=0%	0.194	0.274	0.044	0.088	0.048	0.086	0.076	0.15
Censor=10.2%	0.158	0.246	0.05	0.088	0.056	0.09	0.066	0.116
Censor=20.16%	0.142	0.2	0.036	0.084	0.024	0.064	0.054	0.118
n=500								
Censor=0%	0.254	0.368	0.074	0.108	0.08	0.124	0.064	0.144
Censor=10.35%	0.186	0.274	0.06	0.116	0.04	0.09	0.08	0.164
Censor=19.96%	0.166	0.256	0.034	0.08	0.026	0.068	0.064	0.136
n=2000								
Censor=0%	0.658	0.758	0.078	0.126	0.096	0.138	0.314	0.468
Censor=10.43%	0.594	0.72	0.072	0.134	0.038	0.064	0.276	0.452
Censor=20.08%	0.588	0.696	0.056	0.1	0.026	0.05	0.218	0.386

Note: Iteration number k=500, Bootstrap Iteration number B=100, Sample size n=100, 200, 300, 400, 500

Simulation DGP:

X1～U[0,1]; X2～U[0,1]

$\mu_i = \exp[-(x_{1i}+2x_{2i}+x_{1i}^2)]$

Exponential pdf for lifetime T_i: $\mu_i exp(-\mu_i t)$.

Model (H_0):

$\lambda_i = \exp[-(\alpha_1 x_{1i}+\alpha_2 x_{2i})]$

Exponential pdf for lifetime y_i: $\lambda_i exp(-\lambda_i t)$.

Table IV: Empirical Power over Misspecification in Density (ECV)

	GCV, ecv		LM 2, ecv		LGP 2, ecv		RM 2, ecv	
	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
n=100								
Censor=0%	0.522	0.772	N/A	N/A	0.44	0.574	0.1	0.162
Censor=10.27%	0.406	0.73	N/A	N/A	0.452	0.622	0.114	0.162
Censor=%	0.452	0.71	N/A	N/A	0.54	0.668	0.084	0.136
n=200								
Censor=0%	0.972	0.996	N/A	N/A	0.686	0.828	0.07	0.1
Censor=10.15%	0.952	0.994	N/A	N/A	0.752	0.852	0.07	0.114
Censor=%	0.94	0.984	N/A	N/A	0.752	0.852	0.088	0.132
n=300								
Censor=0%	1	1	N/A	N/A	0.858	0.952	0.042	0.07
Censor=10.14%	0.996	0.996	N/A	N/A	0.88	0.958	0.056	0.088
Censor=%	0.996	1	N/A	N/A	0.902	0.944	0.068	0.11
n=400								
Censor=0%	1	1	N/A	N/A	0.952	0.99	0.046	0.126
Censor=10.19%	1	1	N/A	N/A	0.98	0.992	0.036	0.094
Censor=%	1	1	N/A	N/A	0.98	0.996	0.062	0.096
n=500								
Censor=0%	1	1	N/A	N/A	0.994	0.998	0.056	0.136
Censor=10.1%	1	1	N/A	N/A	0.984	0.996	0.062	0.134
Censor=%	1	1	N/A	N/A	0.988	1	0.03	0.09

Note: Iteration number k=500, ECV Replication number B=1000, Sample size n=100, 200, 300, 400, 500

Simulation DGP:

X1～N(0,1); X2～N(0,1)

$\mu_i=x_{1i}+2x_{2i}$

Lognormal pdf for lifetime T_i: $(2\pi)^{-0.5}(\sigma t)^{-1}\exp[-(\log t-\mu_i)^2/(2\sigma^2)]$, $\sigma=0.8$.

Model (H_0):

$\lambda_i=\exp[-(\alpha_1 x_{1i}+\alpha_2 x_{2i})]$

Exponential pdf for lifetime y_i: $\lambda_i exp(-\lambda_i t)$.

Table V: Empirical Power over Misspecification in Density (Bootstrap)

	GCV		LM 2, bp		LGP 2, bp		RM 2, bp	
	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
n=100								
Censor=0%	0.568	0.812	N/A	N/A	0.448	0.602	0.118	0.176
Censor=10.13%	0.52	0.756	N/A	N/A	0.462	0.604	0.124	0.172
Censor=21.69%	0.44	0.668	N/A	N/A	0.506	0.648	0.088	0.146
n=200								
Censor=0%	0.97	1	N/A	N/A	0.698	0.846	0.068	0.112
Censor=10.14%	0.964	0.992	N/A	N/A	0.738	0.856	0.072	0.116
Censor=21.94%	0.982	0.986	N/A	N/A	0.784	0.89	0.058	0.082
n=300								
Censor=0%	1	1	N/A	N/A	0.85	0.938	0.056	0.094
Censor=10.32%	0.998	1	N/A	N/A	0.908	0.956	0.042	0.08
Censor=21.7%	0.996	0.998	N/A	N/A	0.95	0.98	0.058	0.086
n=400								
Censor=0%	1	1	N/A	N/A	0.95	0.988	0.042	0.11
Censor=10.16%	1	1	N/A	N/A	0.96	0.986	0.042	0.098
Censor=21.22%	1	1	N/A	N/A	0.982	0.996	0.024	0.062
n=500								
Censor=0%	1	1	N/A	N/A	0.98	0.998	0.04	0.124
Censor=10.20%	1	1	N/A	N/A	0.992	1	0.05	0.11
Censor=21.8%	1	1	N/A	N/A	0.988	0.996	0.02	0.052

Note: Iteration number k=500, Bootstrap Iteration number B=100, Sample size n=100, 200, 300, 400, 500

Simulation DGP:

$X1 \sim N(0,1)$; $X2 \sim N(0,1)$

$\mu_i = x_{1i} + 2x_{2i}$

Lognormal pdf for lifetime T_i: $(2\pi)^{-0.5}(\sigma t)^{-1}\exp[-(\log t-\mu_i)^2/(2\sigma^2)]$, $\sigma = 0.8$.

Model (H_0):

$\lambda_i = \exp[-(\alpha_1 x_{1i} + \alpha_2 x_{2i})]$

Exponential pdf for lifetime y_i: $\lambda_i exp(-\lambda_i t)$.