cemmap
centre for microdata methods and practice

Partial Identification and Inference in Duration Models with Endogenous Censoring

Shosei Sakaguchi

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP8/20

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Partial Identification and Inference in Duration Models with Endogenous Censoring *

Shosei Sakaguchi †

February 10, 2020

Abstract

This paper studies identification and inference in transformation models with endogenous censoring. Many kinds of duration models, such as the accelerated failure time model, proportional hazard model, and mixed proportional hazard model, can be viewed as transformation models. I allow the censoring of duration outcome to be arbitrarily correlated with observed covariates and unobserved heterogeneity. I impose no parametric restrictions on the transformation function or the distribution function of the unobserved heterogeneity. In this setting, I partially identify the regression parameters and the transformation function, which are characterized by conditional moment inequalities of U-statistics. I provide an inference method for them by constructing an inference approach for the conditional moment inequality models of U-statistics. I apply the proposed inference method to evaluate the effect of heart transplants on patients' survival time using data from the Stanford Heart Transplant Study.

Keywords: Partial identification, duration models, transformation models, censoring, conditional moment inequality.
JEL codes: C14, C24, C41.

*This paper is based on a chapter of my dissertation at Kyoto University. I am grateful to Yoshihiko Nishiyama and Ryo Okui for their guidance and helpful suggestions. Parts of this paper were written while I was visiting Penn State as a visiting student. I would like to thank Keisuke Hirano for his hospitality and helpful advice. I would also like to thank Hidehiko Ichimura, Toru Kitagawa, Francesca Molinari, Daniel Wilhelm, and participants in the seminar at Kyoto University, the Workshop on Advances in Econometrics at Academia Sinica in Taipei, the Workshop on Advances in Econometrics 2018 in Matsuyama, and those in the session at the 2018 Asian Meeting of the Econometric Society in Seoul, for their comments and suggestions. I acknowledge financial support from the Japan Society for the Promotion of Science (JSPS) under KAKENHI Grant Number JP18J00173.

†Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom. Email: s.sakaguchi@ucl.ac.uk.

1 Introduction

Duration models are widely used in various empirical studies in economics and biomedical sciences. This is because the outcomes of interest in empirical studies are often the durations until certain events occur. Durations of interest in economics include unemployment durations, strike durations, insurance claim durations, and durations until the purchase of durable goods.[1]

In practice, duration data are often censored. For example, unemployment durations are likely to be censored due to some individuals dropping out of the survey. Dealing with censoring has been a substantial challenge in duration analysis, and various methods have been proposed. The standard approach is to assume that censoring is independent of unobserved heterogeneity (conditional or unconditional on observed characteristics). Studies employing this approach include Cox (1972), Powell (1984), Ying et al. (1995), Yang (1999), Honoré et al. (2002), Hong and Tamer (2003), and Khan and Tamer (2007), among others. However, in many cases, justifying this independence assumption is difficult. For example, in unemployment duration analysis, unemployed individuals with low motivation to find a job, which is unobserved heterogeneity, may tend to drop out of the survey at an early stage. Szydłowski (2019) presents a number of examples where censoring is correlated with unobserved heterogeneity (i.e., censoring is endogenous).

In this paper, I study identification and inference in transformation models in the presence of endogenous censoring. The transformation model is expressed as

$$T(Y^*) = X'\beta_0 + U, \tag{1.1}$$

where $T(\cdot)$ is a strictly increasing function; Y^* is a dependent variable, which represents a duration outcome in this paper; X is a k-dimensional vector of observed covariates, whose support is denoted by $\mathcal{X}$; β_0 denotes k-dimensional regression parameters; and U is an unobserved random variable that is independent of X. Many kinds of duration models, such as the accelerated failure time model, proportional hazard model, and mixed proportional hazard (MPH) model, can be viewed as transformation models.[2] In this pa-

[1] van den Berg (2001) surveys the many applications of duration models.

[2] Aside from duration models, the class of transformation models contains other important kinds of models, for example, the linear index model and Box-Cox transformation model.

per, I consider nonparametric transformation models in which neither the transformation function nor the distribution function of the unobserved heterogeneity is parametrically specified. One important model represented by the nonparametric transformation model is the nonparametric MPH model in which neither a baseline hazard function nor the distribution function of the unobserved heterogeneity is parametrically specified.

Allowing for endogenous censoring, I partially identify the regression parameters β_0 in the nonparametric transformation model (1.1). The identification is built on the rank property of the nonparametric transformation model proposed by Han (1987). In his work, he shows that if there is no censoring and at least one regressor has full-support on the real line, the regression parameters are point identified up to scale by looking at the rank correlation between the outcomes and regressors. In the presence of endogenous censoring, I partially identify β_0 by supposing in his rank correlation approach that each censored outcome takes an infinitely large value or a value that corresponds to censoring time. This reflects the fact that, concerning each censored outcome, all we know is that it may take any value larger than censoring time. Moreover, unlike Han's (1987) result, the partial identification analysis does not require the full-support condition on the regressors.

The set of the parameters is characterized by conditional moment inequalities whose sample moment is a U-statistic. Based on this, I construct an inference method for the parameters by extending the inference approach for conditional moment inequality models proposed by Andrews and Shi (2013) into the case of U-statistics. The inference method can be applied not only to this work but also to other works involving conditional moment inequalities of U-statistics. In this sense, this paper also contributes to the literature on inference for conditional moment inequality models.[3]

It should be noted here that the set of the parameters discussed above is not necessarily a sharp identified set. On the other hand, using concepts from random set theory (e.g., Beresteanu et al. (2011, 2012)), I characterize the sharp identified set as well. However, constructing a feasible inference method based on it is difficult, whereas the proposed parameters set is tractable to construct a feasible inference method. In the paper, I also discuss the conditions under which the proposed parameters set becomes close to the

[3]Various inference methods for conditional moment inequality models have been proposed, for example, by Andrews and Shi (2013, 2014, 2017), Chernozhukov et al. (2013), Armstrong (2014, 2015), Menzel (2014), and so on. But, none of them can be applied to the sample moment functions of U-statistics.

sharp set.

As an extension, I also study identification and inference for the transformation function, $T(\cdot)$, in the presence of endogenous censoring. In the case of no censoring or covariate dependent censoring, some works have studied identification and inference for the transformation function without parametric specifications (e.g., Horowitz (1996), Ye and Duan (1997), and Chen (2002)). I partially identify the transformation function by incorporating endogenous censoring into Chen's (2002) rank approach as well as provide its inference procedure.

This paper is related to works that study endogenous censoring. Khan and Tamer (2009), Khan et al. (2011, 2016), Li and Oka (2015), and Fan and Liu (2018) study identification and estimation of parameters in quantile regression models under endogenous censoring. For cross-sectional linear quantile regression models, Khan and Tamer (2009) provide a point identification result for the linear coefficients under a certain support condition, and Khan et al. (2011) provide a sharp identification result without the support condition. Under censoring characterized by a certain copula, Fan and Liu (2018) partially identify the linear coefficients of the same model. Li and Oka (2015) and Khan et al. (2016) consider panel quantile regression models with endogenous censoring and provide partial identification results. Differently from these works, the identification result in this paper does not rely on quantile modeling, copula characterization of censoring, or panel data. Aside from quantile models, Szydłowski (2019) considers the parametric MPH model and proposes a sharp identified set and inference for its parameters. While Szydłowski (2019) considers the parametric MPH model, I consider the nonparametric one, which is robust to the misspecification of the hazard function or the distribution function of the unobserved heterogeneity. For competing risks models, Honoré and Lleras-Muney (2006) partially identify the parameters in the accelerated failure time model, and Kim (2018) derives computationally tractable bounds for distributions of latent durations by exploiting the discreteness of observed durations. In this paper, I consider continuous observed durations and do not specify competing risks.

The remainder of this paper is structured as follows. Section 2 describes the setup and identification assumptions and then provides the main identification result for the regression parameters. In this section, I also characterize the sharp identified set and

compare the two parameter sets. Section 3 provides an inference method for the regression parameters and derives its asymptotic properties. Section 4 presents some numerical examples and Monte Carlo simulation results. The numerical examples show how the proposed set varies depending on the degree of censoring and the support of the regressors. The Monte Carlo simulation results show the finite sample properties of the proposed inference method. Section 5 presents an empirical illustration, where I apply the inference method to evaluate the effect of heart transplants on patients' survival duration using data from the Stanford Heart Transplant Study. Section 6 presents an identification result for the transformation function, whose inference procedure is described in Appendix A.3. I conclude this paper with some remarks in Section 7. All the proofs are presented in Appendices A.1 and A.2.

2 Model and Identification

In this section, I first describe the setting of the paper and provide conditions for identification in Section 2.1. Subsequently, in Section 2.2, I present the main identification result for the regression parameters. In Section 2.3, I characterize the sharp identified set using concepts from random set theory and compare the proposed identification set with the sharp identified set.

2.1 Model

We consider the transformation model in the form of (1.1). In the model, we do not specify the transformation function, $T(\cdot)$, or the distribution function of the unobserved heterogeneity, denoted by $F_U(\cdot)$. Because of this, we impose location and scale normalizations. For the location normalization, as in Horowitz (1996), we suppose that the constant term is equal to zero (i.e., X does not contain a constant term) and $T(\tilde{y}) = 0$ for some finite $\tilde{y}$. For the scale normalization, we suppose that the absolute value of the first component of β_0 is equal to one (i.e., $|\beta_{0,1}| = 1$), where $\beta_{0,k}$ denotes the k-th component of β_0. Let the normalized parameter space be denoted by B. Our focus is on the identification and inference of the normalized regression parameters β_0 in B.

The transformation model contains many kinds of duration models as its special cases:

the accelerated failure time model, Cox's proportional hazard model, and MPH model.[4] In particular, the nonparametric MPH model is an important duration model represented by a nonparametric transformation model. The MPH model is an extension of Cox's proportional hazard model in that individual unobserved heterogeneity is incorporated. Since introduced in Lancaster (1979), it has been widely used in various empirical studies in economics. In the nonparametric MPH model, the normalized parameters β_0 can be interpreted as the logs of the scale-normalized hazard ratios (see, e.g., Lancaster (1990)).

When the data are subject to censoring, the duration outcome Y^* cannot always be observed. Instead, for unit $i = 1, \ldots, n$, we observe $W_i = (Y_{0i}, D_i, X_i)$, such that $Y_{0i} = \min\{Y_i^*, C_i\}$ and $D_i = I[Y_i^* \leq C_i]$, where C_i is a random censoring variable and $I[\cdot]$ denotes the indicator function. D_i is a censoring indicator that takes the value zero if Y_i^* is censored and the value one if Y_i^* is observed. Note that we consider right censoring in the paper, but all the results presented below are easily extendable to left and interval censoring. Using D_i, Y_{0i} can be expressed as $Y_{0i} = D_i Y_i^* + (1 - D_i)C_i$. Let P be the distribution function of (Y^*, X, C).

Throughout this paper, we suppose that the following assumptions hold.

Assumption 2.1. *The vectors (Y_i^*, C_i, X_i), $i = 1, \ldots, n$, are independent and identically distributed (i.i.d) from the latent transformation model (1.1) with the distribution function P, and B is a compact subset of $\mathbb{R}^k$.*

Assumption 2.2. *U is distributed independently of X and has a continuous distribution.*

Assumption 2.3. *Let $D = I[Y^* \leq C]$ and $\mathcal{X}_{uc} = \{x \in \mathcal{X} : P(D = 1 \mid X = x) > 0\}$. Then, $P(\mathcal{X}_{uc}) > 0$.*

Assumption 2.4. *$\mathcal{X}_{uc}$ contains at least two distinct values.*

Assumption 2.2 requires that U is independent of X. However, it does not restrict the relationship between U and C, as we consider endogenous censoring. Assumption 2.3

[4] If $T(Y^*) = \log Y^*$, the transformation model corresponds to the accelerated failure time model; if $T(Y^*) = \log \Delta(Y^*)$,where $\Delta(\cdot)$ is the integrated baseline hazard function, and U has the CDF $F(u) = 1 - \exp(-e^u)$, the transformation model corresponds to Cox's proportional hazard model; if $T(Y^*) = \log \Delta(Y^*)$ and $U = \epsilon + \nu$ where ν is unobserved heterogeneity and ϵ has the CDF $F(\epsilon) = 1 - \exp(-e^\epsilon)$, the transformation model corresponds to the MPH model. For more details, see Horowitz (2009, Ch. 6).

requires that the probability of censoring is not equal to one for all x. Under Assumption 2.4, in contrast to many works in the semiparametric literature, we do not impose a full-support or full-rank condition on the regressors, in line with Magnac and Maurin (2008), Blevins (2011), and Komarova (2013). By not doing so, we allow all the regressors to be discrete random variables, to not have large support, or to be arbitrarily correlated with each other. Magnac and Maurin (2008), Blevins (2011), and Komarova (2013) discuss the difficulties of justifying these conditions in a number of cases, and provide partial identification results for some other semiparametric models in the absence of these conditions.

2.2 Identification for the Regression Parameters

This section presents the partial identification result for the regression parameters. Because the identification result I propose in this section is based on Han (1987), I first briefly introduce his identification result in the absence of censoring.

Suppose now that there is no censoring (i.e., Y^* is always observed). In this case, under Assumptions 2.1–2.3 and the full-support and full-rank conditions on the regressors, Han (1987) shows that β_0 uniquely satisfies the following rank property,

$$x_i'\beta_0 \geq x_j'\beta_0 \Leftrightarrow P(Y_i^* \geq Y_j^* \mid x_i, x_j) \geq P(Y_j^* \geq Y_i^* \mid x_i, x_j)$$

for all $(x_i, x_j) \in \mathcal{X}^2$. Heuristically, β_0 uniquely satisfies

$$x_i'\beta_0 \geq x_j'\beta_0 \Leftrightarrow P(Y_i^* \geq Y_j^* \mid x_i, x_j) \geq \frac{1}{2} \tag{2.1}$$

for all $(x_i, x_j) \in \mathcal{X}^2$. This rank property means that, for any given pair of (x_i, x_j), the probability that Y_i^* is larger than or equal to Y_j^* is greater than or equal to $1/2$ if and only if $x_i'\beta_0$ is larger than or equal to $x_j'\beta_0$. Then, β_0 is the unique value in B that satisfies this rank relationship for any pair of (x_i, x_j). In other words, for any $\beta \neq \beta_0$, there exist at least one pair of (x_i, x_j) that violates the rank property (2.1).

In this paper, we actually suppose that censoring exists and it may be endogenous. We cannot always observe Y_i^* and do not have any information about the mechanism of censoring. The censoring variable C_i may be arbitrarily correlated with the observed

covariates X_i and the unobserved heterogeneity U_i.

In this situation, I derive a set of parameters that contains the true parameters β_0. Let $Y_{1i} = D_i Y_i^* + (1 - D_i)(+\infty)$, which is an outcome variable that takes an arbitrary large value when the primary outcome is censored, and recall that $Y_{0i} = D_i Y_i + (1 - D_i)C_i$. Then, because $P(Y_{1i} \geq Y_{0j} \mid x_i, x_j) \geq P(Y_i^* \geq Y_j^* \mid x_i, x_j)$ holds for all (x_i, x_j), the following rank property holds from (2.1),

$$x_i'\beta_0 \geq x_j'\beta_0 \Rightarrow P(Y_{1i} \geq Y_{0j} \mid x_i, x_j) \geq \frac{1}{2} \tag{2.2}$$

for all $(x_i, x_j) \in \mathcal{X}^2$. Therefore, defining

$$B_I \equiv \{\beta \in B \mid \ x_i'\beta \geq x_j'\beta \Rightarrow P\left(Y_{1i} \geq Y_{0i} \mid x_i, x_j\right) \geq \tfrac{1}{2} \text{ for all } (x_i, x_j) \in \mathcal{X}^2\},$$

β_0 is contained in B_I. This parameters set is derived from a worst-case analysis where we suppose that censored outcomes may take extreme values, C or $+\infty$, for any given value of x. This reflects the fact that concerning each censored outcome, all we know is that it may take any value at least larger than its censored time. The following theorem summarizes this identification result.

Theorem 2.1. *Under Assumptions 2.1–2.4, $\beta_0 \in B_I$.*

I provide the proof in Appendix A.1. There are some notes on this theorem. First, in this theorem, we do not impose the full-support or full-rank condition on the regressors, in contrast to many works in the semiparametric literature. Second, B_I is not necessarily sharp regardless of whether the full-support and full-rank conditions are imposed. Although it is not necessarily sharp, it is easy to compute and tractable to construct an inference method based on it, as we will see in Section 3. The following section shows how the sharp identified set can be characterized, although it is difficult to construct an inference method based on it.

2.3 Characterization of the Sharp Identified Set

In this section, I illustrate how the sharp identified set can be characterized using concepts from random set theory. Subsequently, I compare the sharp identified set with B_I. This

comparison clarifies why B_I is not sharp and in which situations it approaches the sharp set. For the definitions and notations for random set theory used in this section, see, for example, Molchanov (2005) or Beresteanu et al. (2012, Appendix A). Throughout this section, for any variable A, we denote by $\tilde{A}$ an independent copy of A.

Using concepts from random set theory, we can characterize the incomplete information for the latent outcome variable Y^*. For any random variable A, let A_x be the random variable that has the conditional distribution of A given $X = x$. Then, for a given $x \in \mathcal{X}$, what we observe for the latent outcome variable in the presence of endogenous censoring can be expressed as the random set $\mathcal{Y}_x$ defined by

$$\mathcal{Y}_x = \begin{cases} \{Y_x^*\} & \text{if } D_x = 1 \\ (C_x, +\infty) & \text{otherwise} \end{cases},$$

where Y_x^* and C_x are, respectively, a latent outcome and a censoring variable given $X = x$, and $D_x = I[Y_x^* \leq C_x]$ is a censoring indicator given $X = x$. Hence, all the information for the latent outcome variable can be expressed by stating that $Y_x^* \in \text{Sel}(\mathcal{Y}_x)$.[5] Let B_0 denote the sharp identified set of β_0. Throughout this section, we suppose that the full-support and full-rank conditions on the regressors hold to ensure the sharp identification result.[6]

Combining the above random set representation with Han's (1987) identification result (2.1), B_0 is characterized as the set of β such that there exists a family of pairs of selections $(Y_{x_i}, \tilde{Y}_{x_j}) \in \text{Sel}(\mathcal{Y}_{x_i}) \times \text{Sel}(\tilde{\mathcal{Y}}_{x_j})$ over $(x_i, x_j) \in \mathcal{X}^2$ that satisfy the following:

$$x_i'\beta \geq x_j'\beta \Leftrightarrow P(Y_{x_i} \geq \tilde{Y}_{x_j}) \geq \frac{1}{2} \tag{2.3}$$

for all $(x_i, x_j) \in \mathcal{X}^2$, where $\tilde{\mathcal{Y}}_x$ is the random set of $\tilde{Y}_x$. Therefore, B_0 is equivalent to

$$B_0 = \left\{ \beta \in B \mid \exists \left\{ Y_x \in \text{Sel}(\mathcal{Y}_{x_i}), \tilde{Y}_{x_j} \in \text{Sel}(\tilde{\mathcal{Y}}_{x_j}) \right\}_{(x_i,x_j)\in\mathcal{X}^2}, \forall (x_i, x_j) \in \mathcal{X}^2, (2.3) \text{ holds} \right\}. \tag{2.4}$$

[5] For any random set $\mathcal{Y}$, a random variable Y is called a measurable selection of $\mathcal{Y}$ if $Y \in \mathcal{Y}$ a.s., and $\text{Sel}(\mathcal{Y})$ is defined to be the set of all measurable selections of $\mathcal{Y}$. See, for example, Molchanov (2005, Ch. 1) or Beresteanu et al. (2012, Appendix A).

[6] These conditions might not be needed to obtain the sharp identification result. But, to my knowledge, there is no work that derives the sharp identification result for the nonparametric transformation model in the absence of these conditions.

Next, we look at the proposed set B_I. From the definitions, $Y_{1,x}$ and $Y_{0,x}$ satisfy (i) $Y_{1,x}, Y_{0,x} \in \text{Sel}(\mathcal{Y}_x)$ for any $x \in \mathcal{X}$ and (ii) $Y_{0,x} \leq Y_x \leq Y_{1,x}$ for any $Y_x \in \text{Sel}(\mathcal{Y}_x)$ and $x \in \mathcal{X}$. Thus, for some given pair (x_i, x_j), the parameters set

$$\{\beta \in B \mid x_i'\beta \geq x_j'\beta \Rightarrow P\left(Y_{1i} \geq Y_{0j} \mid x_i, x_j\right) \geq \tfrac{1}{2}\}$$

is equivalent to

$$\left\{\beta \in B \mid \exists \left(Y_{x_i}, \tilde{Y}_{x_j}\right) \in \text{Sel}(\mathcal{Y}_{x_i}) \times \text{Sel}(\tilde{\mathcal{Y}}_{x_j}), (2.3) \text{ holds}\right\},$$

which is the set of β such that, for the given (x_i, x_j), there exists a pair of selections $\left(Y_{x_i}, \tilde{Y}_{x_j}\right) \in \text{Sel}(\mathcal{Y}_{x_i}) \times \text{Sel}(\tilde{\mathcal{Y}}_{x_j})$ that satisfies the inequality (2.3). Therefore, from the definition of B_I, B_I is characterized as a set of β such that, for any pair $(x_i, x_j) \in \mathcal{X}^2$, there exists a pair of selections $\left(Y_{x_i}, \tilde{Y}_{x_j}\right) \in \text{Sel}(\mathcal{Y}_{x_i}) \times \text{Sel}(\tilde{\mathcal{Y}}_{x_j})$ that satisfy the inequality (2.3). Formally, B_I is characterized as

$$B_I = \left\{\beta \in B \mid \forall\, (x_i, x_j) \in \mathcal{X}^2, \exists \left(Y_{x_i}, \tilde{Y}_{x_j}\right) \in \text{Sel}(\mathcal{Y}_{x_i}) \times \text{Sel}(\tilde{\mathcal{Y}}_{x_j}), (2.3) \text{ holds}\right\}. \quad (2.5)$$

The difference in the statements between (2.4) and (2.5) shows that B_0 is contained in B_I, and hence B_I is not necessarily sharp. An intuition for the non-sharpness of B_I is as follows. Suppose now a triple of realized values of X, (x_i, x_j, x_k), and some $\beta \in B$ such that $x_i'\beta \leq x_j'\beta \leq x_k'\beta$. In the construction of B_I, when we compare x_j with x_k, we suppose that the latent outcome variable $Y^*_{x_j}$ takes its smallest value, C_{x_j}, whereas when we compare x_j with x_i, we suppose that $Y^*_{x_j}$ takes its largest value, $+\infty$. However, when characterizing the sharp identified set as (2.4), we compare among fixed selections $Y_x \in \text{Sel}(\mathcal{Y}_x)$ over all $x \in \mathcal{X}$; that is, Y_{x_i} is not changeable when compared with different $\tilde{Y}_{x_j}$ over $x_j \in \mathcal{X}$. This difference explains why B_I is not smaller than B_0.

Remark 2.1. *Although we could characterize the sharp set B_0 as (2.4), it is hard to compute. When examining whether a certain value of β is contained in B_0, one would have to search for the existence of selections $Y_x \in Sel(\mathcal{Y}_x)$, for all $x \in \mathcal{X}$, that satisfy the rank inequality (2.3) for all pairs $(x_i, x_j) \in \mathcal{X}^2$. By contrast, the proposed set B_I is easy to compute. For this reason, I focus on B_I in this paper rather than the sharp set.*

Beresteanu et al. (2011, 2012) suggest using the support function and Aumann expectation to easily compute a sharp identified set. However, using this approach, one still has to search for the selections $Y_x \in Sel(\mathcal{Y}_x)$, for all $x \in \mathcal{X}$, to satisfy a certain equality. Thus this approach does not so much ease the computation.

Remark 2.2. *There are some situations when B_I is close to B_0. Comparing (2.4) with (2.5) suggests that if the random set $\mathcal{Y}_x$ does not widely vary, then B_I is close to B_0. There are some such cases. First, with a lower amount of censoring, B_I is closer to B_0. This is because when the censoring is unlikely to occur given any value of $x \in \mathcal{X}$, the measurable selections of $\mathcal{Y}_x$, in (2.4) and (2.5), take the single value Y_x^* with high probabilities. The second case is when Y_x^* is not censored at small values; that is, C_x takes a large value when Y_x^* is censored. In this case, measurable selections in $\mathcal{Y}_x$ do not widely vary, under which the difference between (2.4) and (2.5) does not make much difference between B_I and B_0. In the empirical example of the heart transportation study in Section 5, this case corresponds to the case when each patient is unlikely to drop out of the study at an early stage.*

3 Inference

This section provides a statistical inference approach for the regression parameters in model (1.1) based on the identification result presented in Section 2.2. I suggest a method to construct a confidence set that covers the true parameter value β_0 with a probability greater than or equal to $1-\alpha$ for $\alpha \in (0,1)$. Because B_I is characterized by conditional moment inequalities involving U-statistics, I construct the inference method by extending the inference approach for conditional moment inequality models proposed by Andrews and Shi (2013) (hereafter AS) into the U-statistics case. The approach transforms conditional moment inequalities into an infinite number of unconditional ones, without information loss, to construct a test statistic, and a confidence set is constructed by inverting the test statistic and using critical values obtained via moment selection. We consider continuous regressors in this section; but, if all regressors are discrete, we can apply inference methods for unconditional moment inequality models.[7] The inference method I propose below

[7] Various inference methods for unconditional moment inequality models have been proposed by Imbens and Manski (2004), Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Romano and Shaikh

is for U-statistics of order two, but it can be applied to U-statistics of greater order with some obvious modifications.

3.1 Test Statistic and Critical Value

In this subsection, I construct the test statistic and then describe the inference procedure. Let

$$m\left(W_i, W_j, \beta\right) = -\frac{1}{2} + I[Y_{1i} \geq Y_{0j}] \cdot I[X_i'\beta \geq X_j'\beta] + I[Y_{1j} > Y_{0i}] \cdot I[X_j'\beta > X_i'\beta].$$

Then, B_I is a set of parameters that satisfy the following conditional moment inequalities,

$$E_P\left[m\left(W_i, W_j, \beta\right) \mid x_i, x_j\right] \geq 0 \text{ for all } \left(x_i, x_j\right) \in \mathcal{X}^2. \tag{3.1}$$

To transform all the information from the conditional moment inequalities (3.1) into unconditional ones, I adopt AS's instrumental functions approach. From here, we suppose, without loss of generality, that X_i is transformed via a one-to-one mapping so that each of its elements lies in $[0,1]$ (i.e., $\mathcal{X} = [0,1]^k$).[8] The set of instrumental functions that we consider is of the following form:

$$\mathcal{G} = \left\{g\left(x_i, x_j\right) = I\left[x_i \in C_1, x_j \in C_2\right] \text{ for } \left(C_1, C_2\right) \in \mathcal{C}\right\},$$

where

$$\begin{aligned}\mathcal{C} = \Big\{C_{a,\tilde{a},r} = \bigtimes_{u=1}^{k}\left(\left(a_u - 1\right)/\left(2r\right), a_u/\left(2r\right)\right] \times \bigtimes_{u=1}^{k}\left(\left(\tilde{a}_u - 1\right)/\left(2r\right), \tilde{a}_u/\left(2r\right)\right] : \\ a = \left(a_1, \ldots, a_k\right)',\ \tilde{a} = \left(\tilde{a}_1, \ldots, \tilde{a}_k\right)',\ \left(a_u, \tilde{a}_u\right) \in \left\{1, 2, \ldots, 2r\right\}^2 \\ \text{for } u = 1, \ldots, k \text{ and } r = 1, 2, \ldots\Big\}.\end{aligned}$$

This set of instrumental functions transforms the conditional moment inequalities (3.1) into infinitely many unconditional ones without loss of information. Accordingly, under

(2008, 2010), Stoye (2009), Andrews and Soares (2010), Bugni (2010), and so on.

[8] For example, following AS, the transformed regressors may be $X_i^o = \Phi\left(\hat{\Sigma}_{X,n}^{-1/2}\left(X_i - \bar{X}_n\right)\right)$ where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $\hat{\Sigma}_{X,n} = n^{-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)\left(X_i - \bar{X}_n\right)'$, and $\Phi\left(x\right) = \left(\Phi\left(x_1\right), \ldots, \Phi\left(x_K\right)\right)'$, where $\Phi\left(\cdot\right)$ denotes the standard normal cumulative distribution function and $x = \left(x_1, \ldots, x_k\right)'$.

Assumptions 2.1–2.4, B_I is equivalent to

$$\{\beta \in B : \ E_P [m (W_i, W_j, \beta, g)] \geq 0 \text{ for all } g \in \mathcal{G}\},$$

where $m (W_i, W_j, \beta, g) = m (W_i, W_j, \beta) \cdot g (X_i, X_j)$ for $g \in \mathcal{G}$. I formalize this result as Lemma A.2 in Appendix A.2 with a proof. Other kinds of instrumental functions introduced in AS could be applicable with modifications.

Define the sample moment function and sample variance function of $m (W_i, W_j, \beta, g)$, respectively, by

$$\bar{m}_n (\beta, g) = \frac{1}{n(n-1)} \sum_{i \neq j} m (W_i, W_j, \beta, g)$$

and

$$\hat{\sigma}_n^2 (\beta, g) = \left\{ \frac{1}{n (n-1) (n-2)} \sum_{i \neq j \neq k} m (W_i, W_j, \beta, g) m (W_i, W_k, \beta, g) - \left(\frac{1}{n(n-1)} \sum_{i \neq j} m (W_i, W_j, \beta, g) \right)^2 \right\}.$$

Note that $\bar{m}_n (\beta, g)$ and $\hat{\sigma}_n^2 (\beta, g)$ are U-statistics of orders two and three, respectively. Because $\bar{m}_n (\beta, g)$ is a non-degenerate U-statistic of order two, the asymptotic variance of $\sqrt{n}\bar{m}_n (\beta, g)$ is $\mathrm{Var}_P (E_P [m (W_i, W_j, \beta, g) \mid W_i])$, which is equivalent to[9]

$$E_P [m (W_i, W_j, \beta, g) m (W_i, W_k, \beta, g)] - (E_P [m (W_i, W_j, \beta, g)])^2.$$

Thus, $\hat{\sigma}_n^2 (\beta, g)$ is a consistent estimator of the asymptotic variance of $\sqrt{n}\bar{m}_n (\beta, g)$. However, in practice, $\hat{\sigma}_n^2 (\beta, g)$ could be zero for some $g \in \mathcal{G}$; so we use the modification proposed by AS for $\hat{\sigma}_n^2 (\beta, g)$. The modified version of $\hat{\sigma}_n^2 (\beta, g)$ is

$$\bar{\sigma}_n^2 (\beta, g) = \hat{\sigma}_n^2 (\beta, g) + \epsilon \hat{\sigma}_n^2,$$

[9]For the variance of U-statistics, see, for example, van der Vaart (1998, Ch. 12).

where $\hat{\sigma}_n^2 = \hat{\sigma}_n^2(\beta, 1)$, which is a consistent estimator of

$$\sigma_P^2(\beta) = E_P\left[m(W_i, W_j, \beta)\, m(W_i, W_k, \beta)\right] - \left(E_P\left[m(W_i, W_j, \beta)\right]\right)^2,$$

and ϵ is a regularization parameter that takes some fixed positive value. Based on some simulation experiments, I recommend taking $\epsilon = 0.0001$.[10]

Then, with $g_{a,\tilde{a},r}(x_i, x_j) = 1\left[(x_i, x_j) \in \left(\frac{a_u - 1}{2r}, \frac{a_u}{2r}\right] \times \left(\frac{\tilde{a}_u - 1}{2r}, \frac{\tilde{a}_u}{2r}\right]\right]$, the test statistic at β takes the form

$$T_n(\beta) = \sum_{r=1}^{\infty} \left(r^2 + 100\right)^{-1} \sum_{(a,\tilde{a}) \in \{1, \cdots, 2r\}^2} (2r)^{-2K} \left[\frac{n^{\frac{1}{2}} \bar{m}_n(\beta, g_{a,\tilde{a},r})}{\bar{\sigma}_n(\beta, g_{a,\tilde{a},r})}\right]_-^2,$$

where $[x]_- = -x$ if $x < 0$ and $[x]_- = 0$ if $x \geq 0$. This test statistic is a version of AS's test statistic that is extended to the U-statistics of order two. Here, the inner summation is taken over two indices, a and $\tilde{a}$. In the implementation, we instead use an approximate test statistic at β:

$$T_{n,R}(\beta) = \sum_{r=1}^{R} \left(r^2 + 100\right)^{-1} \sum_{(a,\tilde{a}) \in \{1, \cdots, 2r\}^2} (2r)^{-2K} \left[\frac{n^{\frac{1}{2}} \bar{m}_n(\beta, g_{a,\tilde{a},r})}{\bar{\sigma}_n(\beta, g_{a,\tilde{a},r})}\right]_-^2,$$

where R is some truncation integer chosen by the researcher.

To compute a critical value for $T_{n,R}(\beta)$, I propose using an asymptotic approximation version of the critical value. This is a simulated quantile of

$$T_{n,R}^{Asy}(\beta) = \sum_{r=1}^{R} \left(r^2 + 100\right)^{-1} \sum_{(a,\tilde{a}) \in [1, \cdots, 2r]^2} (2r)^{-2K} \left[\frac{v_n(\beta, g_{a,\tilde{a},r}) + \varphi_n(\beta, g_{a,\tilde{a},r})}{\bar{\sigma}_n(\beta, g_{a,\tilde{a},r})}\right]_-^2,$$

where $(v_n(\beta, g))_{g \in \mathcal{G}}$ is a zero mean Gaussian process with a covariance kernel evaluated by

$$\hat{h}_2(\beta, g, g^*) = \left\{ \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} m(W_i, W_j, \beta, g)\, m(W_i, W_k, \beta, g^*) \right.$$
$$\left. - \left[\frac{1}{n(n-1)} \sum_{i \neq j} m(W_i, W_j, \beta, g)\right] \cdot \left[\frac{1}{n(n-1)} \sum_{i \neq j} m(W_i, W_j, \beta, g^*)\right] \right\}.$$

[10] This value is different from the value recommended by AS, which is $\epsilon = 0.05$.

In the form of $T_{n,R}^{Asy}(\beta)$, $\left(v_n\left(\beta, g_{a,\tilde{a},r}\right)\right)_{a,\tilde{a},r}$ approximates the asymptotic distribution of

$$\left(n^{\frac{1}{2}}\left[\bar{m}_n\left(\beta, g_{a,\tilde{a},r}\right) - E_P\left[m\left(W_i, W_j, \beta, g_{a,\tilde{a},r}\right)\right]\right]\right)_{a,\tilde{a},r}.$$

$\varphi_n\left(\beta, g_{a,\tilde{a},r}\right)$ is a generalized moment selection (GMS) function to select binding moment restrictions, which is given by

$$\varphi_n\left(\beta, g_{a,\tilde{a},r}\right) = \hat{\sigma}_n^2 B_n I\left[\kappa_n^{-1} n^{\frac{1}{2}} \bar{m}_n\left(\beta, g_{a,\tilde{a},r}\right) / \bar{\sigma}_n\left(\beta, g_{a,\tilde{a},r}\right) > 1\right],$$

where B_n and κ_n are two tuning parameters that should satisfy $\kappa_n \to \infty$, $\kappa_n / n^{1/2} \to 0$, and $B_n \to \infty$ as $n \to \infty$ a.s. In this paper, based on some simulation experiments, I recommend using $\kappa_n = \left(\left(1 - \hat{p}_{1-D}^{1/3}\right)^{2/5} \times 0.6 \ln(n)\right)^{\frac{1}{2}}$ and $B_n = \left(0.8 \ln(n) / \ln\ln(n)\right)^{\frac{1}{2}}$, where $\hat{p}_{1-D} = \frac{1}{n}\sum_i^n (1 - D_i)$ is the sample censoring rate.[11] The recommended value of κ_n decreases with the sample censoring rate. The following assumption summarizes the requirements for the tuning parameters in the GMS function.

Assumption 3.1. *The tuning parameters (κ_n, B_n) satisfy $\kappa_n \to \infty$, $\kappa_n / n^{1/2} \to 0$, and $B_n \to \infty$ as $n \to \infty$ a.s.*

For a significance level of $\alpha < 1/2$, the critical value is set to be the $1 - \alpha + \eta$ simulated quantile of $T_{n,R}^{Asy}(\beta)$, where η is an arbitrarily small positive number (e.g., 10^{-6} following AS). Letting $\hat{c}_{n,\eta,1-\alpha}(\beta)$ be the $1 - \alpha + \eta$ quantile of $T_{n,R}^{Asy}(\beta)$, a nominal level $1 - \alpha$ confidence set is computed by

$$\widehat{CS}_{n,\eta,1-\alpha} = \left\{\beta \in B : T_{n,R}(\beta) \leq \hat{c}_{n,\eta,1-\alpha}(\beta)\right\}.$$

Note again that the inference approach presented above is for the U-statistics of order two, but it can be applied to U-statistics of a greater order with some modifications. In these modifications, the class of instrumental functions, the moment function, and its sample variance function need to be changed for greater order; the inner double summation in the test statistic needs to be replaced with more summation; and the tuning parameters probably should be customized.

[11] The values for κ_n and B_n are different from the values recommend by AS, which are $\kappa_n = (0.3 \ln(n))^{\frac{1}{2}}$ and $B_n = \left(0.4 \ln(n) / \ln\ln(n)\right)^{\frac{1}{2}}$, respectively.

3.2 Asymptotic Size and Power Properties

This subsection provides uniform asymptotic size and power properties of the inference method. Let $\mathcal{Q}$ be the collection of all pairs of the regression parameters and distribution, (β, P), that satisfy (3.1) and Assumptions 2.1–2.4. Define

$$\begin{aligned} h_{2,P}(\beta, g, g^*) =& E_P\left[E_P\left[m\left(W_i, W_j, \beta, g\right) m\left(W_i, W_k, \beta, g^*\right) \mid W_i\right]\right] \\ & - E_P\left[m\left(W_i, W_j, \beta, g\right)\right] E_P\left[m\left(W_i, W_j, \beta, g^*\right)\right], \end{aligned} \tag{3.2}$$

which is the covariance kernel between $m\left(W_i, W_j, \beta, g\right)$ and $m\left(W_i, W_j, \beta, g^*\right)$ under distribution P. Let $\mathcal{H}_2$ be the collection of all possible covariance kernel functions on $\mathcal{G} \times \mathcal{G}$. The following theorem presents the uniform size and power properties of the proposed inference method.

Theorem 3.1. *Suppose that Assumptions 2.1–2.4 and 3.1 hold, $R = \infty$, and $\alpha < 1/2$.*
(a) For every compact subset $\mathcal{H}_{2,cpt}$ of $\mathcal{H}_2$, the confidence set $\widehat{CS}_{n,\eta,1-\alpha}$ satisfies

$$\lim_{\eta \to 0} \liminf_{n \to \infty} \inf_{\{(\beta,P) \in \mathcal{Q}: h_{2,P} \in \mathcal{H}_{2,cpt}\}} P\left(\beta \in \widehat{CS}_{n,\eta,1-\alpha}\right) = 1 - \alpha.$$

(b) Let $\tilde{\beta} \in B$ be a vector of parameters such that (3.1) is violated for some $(x_i, x_j) \in \mathcal{X}_{uc}^2$. Then, $\lim_{n \to \infty} P(\tilde{\beta} \in \widehat{CS}_{n,\eta,1-\alpha}) = 0$.

The proof is provided in Appendix A.2. Theorem 3.1 (a) states that the proposed confidence set is asymptotically conservative, which corresponds to Theorem 2(b) of AS. The uniformity in the statement enables the asymptotic result to provide a good finite sample approximation, which is well discussed in AS. Theorem 3.1 (b) states that the test is consistent against a fixed alternative.

4 Simulation Studies

This section presents numerical examples and Monte Carlo simulation results. The numerical examples show how the set B_I varies with the degree of censoring and the support of the regressors. The Monte Carlo simulations show the finite sample performance of

the proposed inference method and demonstrate how it varies with various choices of the tuning parameters.

4.1 Numerical Examples

This section provides some examples to show how B_I varies depending on the degree of censoring and the support of the covariates. We consider three MPH models (Models 1–3) with endogenous censoring that have the following form:

$$\begin{aligned} \log Y &= \beta_1 X_1 + \beta_2 X_2 + \log U + \log V, \\ \log C &= \alpha_0 + (\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2) \times \log U + \log W. \end{aligned}$$

In all the models, I set $(\beta_1, \beta_2) = (0.5, 1.5)$ and $(\gamma_0, \gamma_1, \gamma_2) = (-0.5, 0.5, -1)$. In Models 1–3, I set α_0 equal to $+\infty$, 3, and 1.6, respectively. In Model 1, there is no censoring; in Models 2 and 3, there is censoring, which is correlated with the covariates and unobserved heterogeneity. The outcome is likely to be more censored in Model 3 than in Model 2. In all the models, U, V, and W have unit exponential distributions, and X_2 takes values in $\{0, 1\}$. As for X_1, we consider three cases; X_1 takes values in (i) $\{-2.5, -2.0, \ldots, 2.5\}$, (ii) $\{-5, -2.5, \ldots, 5\}$, or (iii) $\{-5, -4.8, \ldots, 5\}$. Support (ii) is wider than support (i), while they have the same fineness; support (iii) is finer than support (ii), while they have the same width. In these data generating processes (DGPs), the censoring is endogenous because Y is correlated with C even conditional on X_1 and X_2, which occurs due to the presence of U in both equations for $\log Y$ and $\log C$. The parameter of interest here is the scale normalized value of β_2, whose true value is $\beta_2 / |\beta_1| = 3$.

Given the DGPs described above, B_I is a set of parameter values that satisfy the conditional moment inequality (3.1) for each pair of values of (X_1, X_2). I numerically obtain B_I by simulating the distributions of $\log Y$ and $\log C$ for each pair of values of (X_1, X_2), from 5,000 random draws from each of the DGPs given the parameter space $B = \{-1, 1\} \times [0, 9]$.

Table 1 presents the numerical results. In Table 1, each cell provides the numerically computed B_I for each model and each support of X_1. As expected, B_I shrinks as the censoring rate decreases or the support becomes wider or finer. In particular, in Model 1,

in which there is no censoring, B_I shrinks to a singleton at the true parameter value when X_1 takes values from support (iii). In the same model, the computed sets from supports (i) and (ii) are the same, which implies that the difference in widths between supports (i) and (ii) does not affect the width of B_I in Model 1.

Table 1: Computed B_I for Models 1–3 and Supports (i)–(iii)

Model / Support of X_1	Support (i)	Support (ii)	Support (iii)
Model 1	$[2.51, 3.49]$	$[2.51, 3.49]$	$\{3\}$
Model 2	$[2.01, 3.50]$	$[2.01, 3.49]$	$[2.41, 3.39]$
Model 3	$[1.50, 4.99]$	$[1.51, 4.00]$	$[1.81, 3.80]$

4.2 Monte Carlo Experiments

I also conduct Monte Carlo experiments to evaluate the size and power properties of the proposed inference method. I use two DGPs (DGP1 and DGP2). In DGP1 and DGP2, the data are derived from Model 2 and Model 3, respectively, where the distribution of X_1 is replaced with a normal distribution with a mean of zero and a standard deviation of two; X_2 takes values in $\{0, 1\}$ with probability $1/2$ for each; U, V, and W have unit exponential distributions. The censoring rates in DGP1 and DGP2 are about 16% and 30%, respectively.

For the Monte Carlo experiments, 500 samples are drawn with sample sizes of 250 and 500. The critical values are simulated using 1,000 repetitions for the significance level $\alpha = 0.05$. Based on the inference method, I conduct a test of H_0 : (3.1) holds against H_1 : (3.1) is violated at each value of $(\beta_1, \beta_2) \in \{1\} \times \{0, 0.5, \ldots, 9\}$, where I set β_1 equal to one for scale normalization. The true value of the normalized parameter $\beta_2/|\beta_1|$ is 3. As a base case, I set the tuning parameters in the GMS function κ_n, B_n, and ϵ to the values recommended in Section 3.1, and set $R = 5$. I also compare the results from the base case with the results from various choices of the tuning parameters. I do not assume that the researcher knows the exact distribution of X_1; hence, I transform X_1 into X_1^o described in Section 3.1 and use X_1^o instead of X_1. As for X_2, I assume that the researcher knows its exact distribution. Since X_2 takes the two discrete values (0 and 1), I set the instrumental

function as

$$g_{a,\tilde{a},d,\tilde{d},,r}(x_i, x_j) = I\left[(x_{1i}, x_{1j}) \in \left(\frac{a-1}{2r}, \frac{a}{2r}\right] \times \left(\frac{\tilde{a}-1}{2r}, \frac{\tilde{a}}{2r}\right],\ (x_{2i}, x_{2j}) \in \left(d, \tilde{d}\right)\right]$$

for all $(a, \tilde{a}) \in \{1, 2, \ldots, 2r\}^2$ and $\left(d, \tilde{d}\right) \in \{0, 1\}^2$, and use the following test statistic,

$$T_{n,R}(\beta) = \sum_{r=1}^{R} \left(r^2 + 100\right)^{-1} \sum_{\left(d,\tilde{d}\right)\in\{0,1\}^2} \sum_{(a,\tilde{a})\in\{1,\cdots,2r\}^2} (2r \times 2)^{-2} \left[\frac{n^{\frac{1}{2}} \bar{m}_n\left(\beta, g_{a,\tilde{a},d,\tilde{d},r}\right)}{\bar{\sigma}_n\left(\beta, g_{a,\tilde{a},d,\tilde{d},r}\right)}\right]_-^2.$$

The critical value is taken as a simulated quantile of

$$T_{n,R}^{Asy}(\beta) = \sum_{r=1}^{R} \left(r^2 + 100\right)^{-1} \sum_{\left(d,\tilde{d}\right)\in\{0,1\}^2} \sum_{(a,\tilde{a})\in\{1,\cdots,2r\}^2} (2r \times 2)^{-2} \times \left[\frac{n^{\frac{1}{2}}\left(v_n\left(\beta, g_{a,\tilde{a},d,\tilde{d},r}\right) + \varphi_n\left(\beta, g_{a,\tilde{a},d,\tilde{d},r}\right)\right)}{\bar{\sigma}_n\left(\beta, g_{a,\tilde{a}d,\tilde{d},r}\right)}\right]_-^2.$$

Figure 1 shows the graphs of rejection frequencies for the base case in DGP1 and DGP2. The solid horizontal line in each figure indicates a rejection frequency of 0.05. The dashed curve and dotted curve indicate the rejection frequencies for sample sizes of 250 and 500, respectively. As expected, all the rejection frequencies at the true point are close to the nominal size $\alpha = 0.05$. Furthermore, the rejection frequencies are close to 0.05 not only at the true parameter value but also in the intervals that contain it. The 95% confidence interval in DGP2 is wider than that in DGP1 for each sample size. It is also wider with a sample size of 500 than with a sample size of 250 in each DGP. But the rejection frequencies at extreme points, such as 0 and 8, are larger with a sample size of 500 than with a sample size of 250 in each DGP. All the intervals are stretched more toward positive values than toward negative values.

Table 2 shows the rejection frequencies at the true parameter value and $(\beta_1, \beta_2) = (1, 0)$ for several choices of the tuning parameters in DGP1 and DGP2. The point $(\beta_1, \beta_2) = (1, 0)$ is not contained in B_I, as seen from the results of the numerical examples. Table 2 shows the degree of sensitivity of the inference to variation in sample size n, the choice of the truncation integer R in the approximate test statistic, the value of ϵ for the modified variance estimator $\bar{\sigma}_n^2(\beta, g)$, and the choice of (κ_n, B_n) in the GMS function. The base

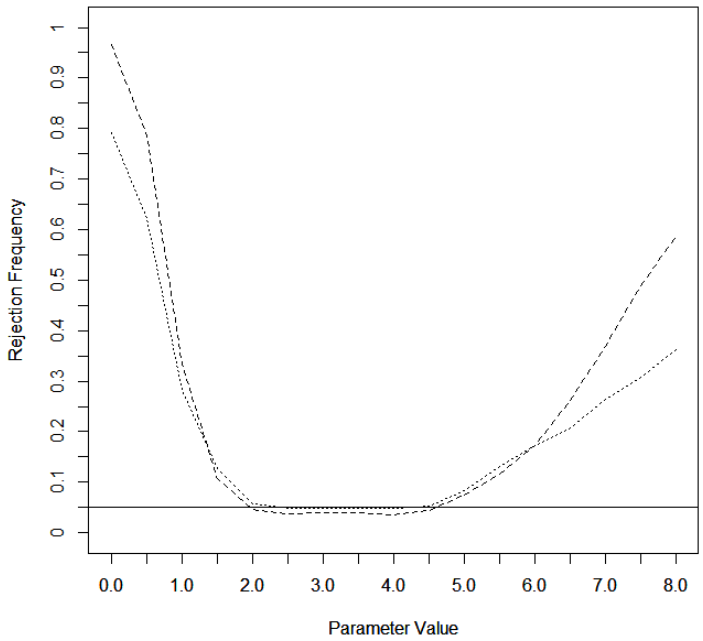
case in Table 2 uses $n = 250$, $R = 5$, and the values of $(\kappa_n, B_n, \epsilon)$ recommended in Section 3.1. In the case of the last row of Table 2, the tuning parameter κ_n does not depend on the sample censoring rate. The results in Table 2 shows that there is some sensitivity to the sample size, the choice of (κ_n, B_n), R, and the value of ϵ. In particular, the sensitivity to the choice of (κ_n, B_n) is high. A small value of ϵ leads to a high power of the test, whereas it increases the size of the test.

Table 2: Rejection Frequencies for the Inference Method: Variation in Sample Size and Choice of the Tuning Parameters

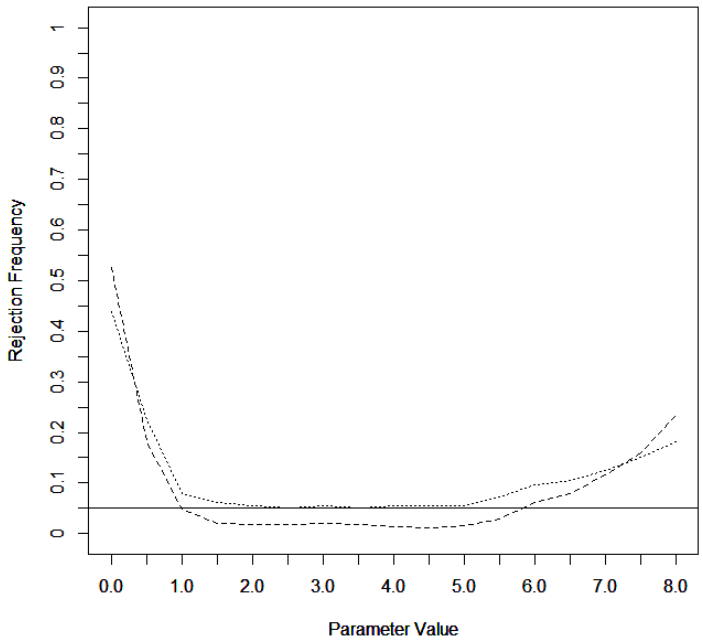
Case	DGP1		DGP2	
	$(\beta_1, \beta_2) = (1, 3)$	$(\beta_1, \beta_2) = (1, 0)$	$(\beta_1, \beta_2) = (1, 3)$	$(\beta_1, \beta_2) = (1, 0)$
Base Case $(n = 250,\ R = 5,\ \epsilon = 0.0001)$	0.046	0.79	0.032	0.362
$\epsilon = 0.001$	0.046	0.716	0.028	0.29
$\epsilon = 0.00001$	0.056	0.802	0.036	0.372
$R = 3$	0.040	0.590	0.038	0.180
$R = 7$	0.036	0.804	0.014	0.348
$n = 100, \epsilon = 0.001$	0.050	0.328	0.044	0.152
$n = 100, \epsilon = 0.0001$	0.068	0.438	0.052	0.216
$n = 100, \epsilon = 0.00001$	0.082	0.460	0.050	0.232
$n = 500, \epsilon = 0.001$	0.028	0.940	0.014	0.364
$n = 500, \epsilon = 0.0001$	0.040	0.962	0.020	0.456
$n = 500, \epsilon = 0.00001$	0.046	0.972	0.022	0.472
$n = 1000, \epsilon = 0.001$	0.032	1.000	0.008	0.492
$n = 1000, \epsilon = 0.0001$	0.034	1.000	0.018	0.644
$n = 1000, \epsilon = 0.00001$	0.040	1.000	0.020	0.666
$(\kappa_n, B_n) = 1/2\,(\kappa_n, B_n)$	0.302	0.950	0.144	0.568
$(\kappa_n, B_n) = 2\,(\kappa_n, B_n)$	0.00	0.256	0.00	0.04
$\kappa_n = (0.6 \ln(n))^{\frac{1}{2}}$	0.022	0.694	0.016	0.24

Figure 1: Rejection Frequencies in DGP1 and DGP2

(a) DGP1



(b) DGP2



Notes: In each line, the horizontal line indicates parameter values at $0.0, 0.5, \ldots, 8.0$ and, at each of these, the dashed curve and dotted curve indicate rejection frequencies for a sample size of 250 and 500, respectively. The solid horizontal line in each graph indicates a rejection frequency of 0.05. For each graph and curve, the set of parameter values at which the curve is below the solid horizontal line is a computed 95% confidence set.

5 Empirical Illustration

I apply the proposed inference method to evaluate the effect of heart transplants on patients' survival duration using the Stanford Heart Transplant Data taken from Kalbfleisch and Prentice (1980). This data set consists of survival times (in days) of 103 patients; an indicator of censoring, which takes the value one if the patient was dead (uncensored) or zero if the patient was censored; an indicator of receiving a heart transplant, which takes the value one if the patient received a heart transplant or zero otherwise; and ages (in years) of patients at the time of acceptance into the program. Among the 103 patients, 27% (28 patients) are censored due to attrition or administrative censoring. The censoring rates for the treated (transplanted) and untreated (not transplanted) groups are 35% and 22%, respectively.

We consider the following censored transformation model,

$$T\left(Y_{0i}\right) = \min\left\{X_{i,age}\beta_{age} + X_{i,treat}\beta_{treat} + U_i, T\left(C_i\right)\right\},$$

where Y_{0i} is the observed survival time, $X_{i,age}$ is the age, $X_{i,treat}$ is the transplant indicator, U_i is the patient's unobserved heterogeneity, and C_i is the censoring time of patient i. Applying the proposed method, we allow the censoring to be arbitrarily correlated with the patient's age and unobserved heterogeneity. Further, we do not specify the transformation function or the distribution function of the patient's unobserved heterogeneity. For scale normalization, I set $|\beta_{age}| = 1$. Our interest is then on the scale-normalized regression parameter of $X_{i,treat}$ (i.e., $\beta_{treat}/|\beta_{age}|$), which can be interpreted as the log of the scale-normalized hazard ratio. We compare the proposed method with the partial rank estimator (PRE) proposed by Khan and Tamer (2007). This is robust up to covariate-dependent censoring and consistently estimates the regression parameters in a nonparametric transformation model.

Table 3 shows the inference results. It presents the point estimate obtained from the PRE and 95% confidence intervals obtained from the PRE and the proposed method. The confidence interval obtained from the PRE is computed based on 1,000 bootstrap pseudo samples from the data. For the proposed method, I set $R = 5$ and use the values of the tuning parameters recommended in Section 3.1 but $\epsilon = 0.001$ as this seems more

conservative for the small sample according to the Monte Carlo simulation results in the previous section. The confidence interval obtained from the proposed method does not have a finite upper bound. This may be because the age does not have sufficiently large support to derive a finite upper bound of the identified set. The estimate obtained from the PRE is positive and is significantly different from zero. The 95% confidence interval obtained from the proposed method is entirely positive and covers the confidence interval obtained from the PRE. The inference result from the proposed method shows that even if the censoring is arbitrarily correlated with the patient's age or unobserved heterogeneity, the heart transplant has a positive effect on the patient's survival time.

Table 3: Empirical Illustration: Inference Results

	PRE	Proposed Method
Estimate	41.4	-
95% Confidence Interval	[17.3, 57.3]	[10.6, $+\infty$]

6 Identification of the Transformation Function

In this section, I propose a partial identification result for the transformation function $T(\cdot)$ in the presence of endogenous censoring. We focus on the transformation function at a particular value of $y \in \mathbb{R}$, $T(y)$. Although we do not identify the distribution function of the unobserved heterogeneity in this paper, knowing about $T(\cdot)$ and β_0 enables us to predict some useful parameters (e.g., average partial effect) given the distribution of the unobserved heterogeneity. Further, the shape of $T(\cdot)$ is informative to infer the type of duration model.

The identification is built on the rank approach proposed by Chen (2002). Let $T_0(\cdot)$ be the true transformation function. In the case with no censoring, provided that Assumptions 2.1–2.3 and the full-rank and full-support conditions on the regressors hold and that the true regression parameters β_0 are given,[12] Chen (2002) shows that $T_0(y)$

[12]Under the supposed conditions, β_0 can be point identified by, for example, applying Han's (1987) maximum rank correlation approach.

uniquely satisfies the following rank property:

$$P\left(Y_i^* \geq y \mid x_i\right) \geq P\left(Y_j^* \geq \tilde{y} \mid x_i\right) \text{ whenever } x_i'\beta_0 - x_j'\beta_0 \geq T_0(y) \tag{6.1}$$

for all $(x_i, x_j) \in \mathcal{X}^2$, where recall that $\tilde{y}$ is such that $T(\tilde{y}) = 0$ for the location normalization. Thus, $T_0(y)$ is point identified under the supposed conditions. He also provides an inference method based on this identification result.

In the presence of endogenous censoring, I partially identify $T_0(y)$ using a similar idea to that presented in Section 2.2. If β_0 was given, since $P\left(Y_{1i} \geq y \mid x_i\right) \geq P\left(Y_i^* \geq y \mid x_i\right)$ and $P\left(Y_j^* \geq y \mid x_j\right) \geq P\left(Y_{0j} \geq y \mid x_j\right)$ hold for all (x_i, x_j), it follows from (6.1) that $T_0(y)$ is contained in

$$\left\{t \in \mathbb{R} : X_i'\beta_0 - X_j'\beta_0 \geq t \Rightarrow P\left(Y_{1i} \geq y \mid x_i\right) \geq P\left(Y_{0j} \geq \tilde{y} \mid x_i\right) \text{ for all } (x_i, x_j) \in \mathcal{X}^2\right\}. \tag{6.2}$$

However, in the presence of endogenous censoring, we cannot point identify β_0; instead, we can obtain the set B_I, as described in Section 2.2, which contains β_0. Thus, letting $T_{I,\beta}(y)$ equal

$$\left\{t \in \mathbb{R} : X_i'\beta - X_j'\beta \geq t \Rightarrow P\left(Y_{1i} \geq y \mid x_i\right) \geq P\left(Y_{0j} \geq \tilde{y} \mid x_i\right) \text{ for all } (x_i, x_j) \in \mathcal{X}^2\right\}$$

and $T_B(y) = \{T_{I,\beta}(y) \mid \beta \in B\}$ for any parameter set B, we have that $T_0(y) \in T_{B_I}(y)$. Note that due to the similar reason discussed in Section 2.3, $T_{I,\beta_0}(y)$ is not a sharp identified set of $T_0(y)$ even if β_0 is known. Hence, T_{B_0} is not a sharp identified set even if we obtain B_0. The following theorem formalizes the identification result.

Theorem 6.1. *Under Assumptions 2.1–2.4, $T_0(y) \in T_{B_I}(y)$ for any $y \in \mathbb{R}$.*

The proof is provided in Appendix A.1. This identification result also does not depend on the full-rank or full-support condition on the regressors. The identified set shrinks as censoring is less likely to occur. In Appendix A.3, I present a joint inference procedure for β_0 and $T_0(y)$.

7 Concluding Remarks

In this paper, I propose a partial identification and inference approach for a nonparametric transformation model in the presence of endogenous censoring. I partially identify the regression parameters and the transformation function, each of which is characterized by conditional moment inequalities involving a U-statistic. I also characterize the sharp identified set of the regression parameters, using the concepts from random set theory, though it is hard to compute. Comparison between the proposed set and sharp identified set makes clear when the proposed set approaches to the sharp set. Based on the identification result, I propose an inference method by extending the inference approach for conditional moment inequality models, proposed by Andrews and Shi (2013), into the U-statistics case, and I derive its asymptotic properties. Numerical examples illustrate the characteristics of the proposed set, and the results of Monte Carlo experiments show the size and power properties of the proposed inference method. As an empirical application, I apply the inference method to evaluate the effect of heart transplants on patients' survival duration by using data from the Stanford Heart Transplant Study, from which I find that heart transplants have a positive effect on patients' survival duration regardless of the structure of censoring.

A Appendix

In this appendix, Section A.1 provides proofs of Theorems 2.1 and 6.1. Section A.2 provides a proof of Theorem 3.1 with some auxiliary lemmas. Section A.3 provides a joint inference procedure for the regression parameters and the transformation function based on the identification results presented in Sections 2.2 and 6.

A.1 Proofs of Theorems 2.1 and 6.1

This section provides proofs of Theorems 2.1 and 6.1.

Proof of Theorem 2.1. From the definitions of Y_{1i} and Y_{0i}, the following holds for all $(x_i, x_j) \in \mathcal{X}^2$,

$$P(Y_{1i} \geq Y_{0j} \mid x_i, x_j) - P(Y_{0j} \geq Y_{1i} \mid x_i, x_j) \geq P(Y_i^* \geq Y_j^* \mid x_i, x_j) - P(Y_j^* \geq Y_i^* \mid x_i, x_j).$$

For the conditional multinomial distribution $P(Y_i^*, Y_j^* \mid x_i, x_j)$, it holds for all $(x_i, x_j) \in \mathcal{X}^2$ that

$$\begin{aligned}
&P\left(Y_i^* \geq Y_j^* \mid x_i, x_j\right) - P\left(Y_j^* \geq Y_i^* \mid x_i, x_j\right) \\
=&P\left(T^{-1}\left(x_i'\beta_0 + U_i\right) \geq T^{-1}\left(x_j'\beta_0 + U_j\right) \mid x_i, x_j\right) - P\left(T^{-1}\left(x_j'\beta_0 + U_j\right) \geq T^{-1}\left(x_i'\beta_0 + U_i\right) \mid x_i, x_j\right) \\
=&P\left(x_i'\beta_0 + U_i \geq x_j'\beta_0 + U_j \mid x_i, x_j\right) - P\left(x_j'\beta_0 + U_j \geq x_i'\beta_0 + U_i \mid x_i, x_j\right) \\
=&P\left(U_i - U_j \geq -\Delta x'\beta_0 \mid x_i, x_j\right) - P\left(U_j - U_i \geq \Delta x'\beta_0 \mid x_i, x_j\right),
\end{aligned}$$

where $\Delta x \equiv x_i - x_j$. The first and second equality holds because $T(\cdot)$ and, automatically, its inversion $T^{-1}(\cdot)$ are strictly monotonic increasing functions under Assumption 2.1.

Under Assumptions 2.3 and 2.4, X takes at least two distinct values on $\mathcal{X}_{uc}$. Then, because the above difference has the same sign as $\Delta x'\beta_0 = x_i'\beta_0 - x_j'\beta_0$ under Assumptions 2.1 and 2.2, it follows that

$$\begin{aligned}
x_i'\beta_0 \geq x_j'\beta_0 &\Rightarrow P\left(Y_i^* \geq Y_j^* \mid x_i, x_j\right) \geq \left(Y_j^* \geq Y_i^* \mid x_i, x_j\right) \\
&\Rightarrow P(Y_{1i} \geq Y_{0j} \mid x_i, x_j) \geq P(Y_{0j} \geq Y_{1i} \mid x_i, x_j) \\
&\Leftrightarrow P(Y_{1i} \geq Y_{0j} \mid x_i, x_j) \geq \frac{1}{2}
\end{aligned}$$

for all $(x_i, x_j) \in \mathcal{X}^2$. This implies that $\beta_0 \in B_I$. □

Proof of Theorem 6.1. Under Assumptions 2.1–2.4, by Theorem 2.1, $\beta_0 \in B_I$. Then, it suffices to show that $T_0(y) \in T_{I,\beta_0}(y)$ for any $y \in \mathbb{R}$.

Note that, from the definitions of Y_{1i} and Y_{0i}, the following holds for all $(x_i, x_j) \in \mathcal{X}^2$,

$$P(Y_{1i} \geq y \mid x_i) - P(Y_{0j} \geq \tilde{y} \mid x_j) \geq P(Y_i^* \geq y \mid x_i) - P(Y_j^* \geq \tilde{y} \mid x_j).$$

For the conditional multinomial distribution $P(Y_i^*, Y_j^* \mid x_i, x_j)$, it holds that

$$\begin{aligned}
&P(Y_i^* \geq y \mid x_i) - P(Y_j^* \geq \tilde{y} \mid x_j) \\
=&P\left(x_i'\beta_0 + U_i \geq T_0(y) \mid x_i\right) - P\left(x_j'\beta_0 + U_j \geq 0 \mid x_j\right) \\
=&F_U\left(-x_j'\beta_0\right) - F_U\left(T_0(y) - x_i'\beta_0\right)
\end{aligned}$$

for all $(x_i, x_j) \in \mathcal{X}^2$, where $F_U(\cdot)$ is the distribution function of U and $T(\tilde{y}) = 0$ holds from the location normalization. Under Assumptions 2.3 and 2.4, X takes at least two distinct values on $\mathcal{X}_{uc}$. Then, since the above difference has the same sign as $\Delta x'\beta_0 - T_0(y)$ under Assumptions 2.1 and 2.2, we have

$$\begin{aligned}
x_i'\beta_0 - x_j'\beta_0 \geq T_0(y) &\Rightarrow P(Y_i^* \geq y \mid x_i) - P(Y_j^* \geq \tilde{y} \mid x_j) \geq 0 \\
&\Rightarrow P(Y_{1i} \geq y \mid x_i) - P(Y_{0j} \geq \tilde{y} \mid x_j) \geq 0
\end{aligned}$$

for all $(x_i, x_j) \in \mathcal{X}^2$. This implies that $T_0(y) \in T_{I,\beta_0}(y)$. □

A.2 Proof of Theorem 3.1

This section provides a proof of the uniform asymptotic probability results for the inference method presented in Section 3. The outline of the proof is same as that of the proofs of Theorems 2 (b) and 3 in AS, but I modify them for the case of U-statistics. Let $\rightsquigarrow$ denote weak convergence of a stochastic process in the sense of Pollard (1990). The following

notations are similar to the notations introduced in AS,

$$v_{n,P}(\beta, g) = n^{\frac{1}{2}} \left(\bar{m}_n(\beta, g) - E_P[m(W_i, W_j, \beta, g)]\right) / \sigma_p(\beta)$$

and

$$\hat{h}_{2,n,P}(\beta, g, g^*) = \left\{ \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} m(W_i, W_j, \beta, g)\, m(W_i, W_k, \beta, g^*) \right.$$
$$\left. - E_P[m(W_i, W_j, \beta, g)] \cdot E_P[m(W_i, W_j, \beta, g^*)] \right\} / \sigma_P^2(\beta).$$

Let $\{v_{h_2}(g) : g \in \mathcal{G}\}$ be a mean zero Gaussian process with some covariance kernel $h_2(\cdot, \cdot)$ on $\mathcal{G} \times \mathcal{G}$.

To prove Theorem 3.1, I first prove that the following two lemmas hold. Lemma A.1 implies that Assumption EP in AS holds. Lemmas A.2 implies that a version of Assumption CI in AS, which is modified for the case of U-statistics, holds.

Lemma A.1. *Suppose that Assumptions 2.1–2.4 hold. For any subsequence $\{(\beta_{a_n}, P_{a_n}) \in \mathcal{Q} : n \geq 1\}$ such that*

$$\lim_{n \to \infty} \sup_{g, g^* \in \mathcal{G}} \left\| h_{2,P_{a_n}}(\beta_{a_n}, g, g^*) - h_2(g, g^*) \right\| = 0$$

for some covariance kernel $h_2(\cdot, \cdot)$ on $\mathcal{G} \times \mathcal{G}$, we have
(a) $\sqrt{a_n} v_{a_n, P_{a_n}}(\beta_{a_n}, \cdot) \rightsquigarrow v_{h_2}(\cdot)$ as $n \to \infty$, and
(b) $\sup_{(g,g^*) \in \mathcal{G} \times \mathcal{G}} \left\| \hat{h}_{2,a_n,P_{a_n}}(\beta_{a_n}, g, g^*) - h_2(g, g^*) \right\| \underset{p}{\to} 0$ as $n \to \infty$.

Lemma A.2. *Suppose that Assumptions 2.1–2.4 hold. For any $\beta \in B$ and any distribution function P, let*

$$\mathcal{X}_P(\beta) = \left\{ (x_i, x_j) \in \mathcal{X}^2 : E_P[m(W_i, W_j, \beta) \mid x_i, x_j] < 0 \right\}.$$

Then, for any $\beta \in B$ and P for which $P((X_i, X_j) \in \mathcal{X}_P(\beta)) > 0$, there exists some $g \in \mathcal{G}$ such that

$$E_P[m(W_i, W_j, \beta, g)] < 0.$$

The next lemma with a proof is auxiliary to Lemma A.1.

Lemma A.3. *Let $\beta \in B$ and P be the limit distribution of P_n. Define $\mathcal{F}_1 = \{f_1(w_i, w_j, \beta, g) : g \in \mathcal{G}\}$ and $\mathcal{F}_2 = \{f_2(w_i, w_j, w_k, \beta, g, g^*) : (g, g^*) \in \mathcal{G} \times \mathcal{G}\}$, where*

$$f_1(w_i, w_j, \beta, g) = m(w_i, w_j, \beta, g) - E_P[m(w_i, w_j, \beta, g)]$$

and

$$\begin{aligned} f_2(w_i, w_j, w_k, \beta, g, g^*) =& m(w_i, w_j, \beta, g)\, m(w_i, w_k, \beta, g^*) \\ & - E_P[m(W_i, W_j, \beta, g)] \cdot E_P[m(W_i, W_j, \beta, g^*)]. \end{aligned}$$

Then, $\mathcal{F}_1$ and $\mathcal{F}_2$ are Euclidean classes of functions for constant envelopes 1 and 1/2, respectively.

Proof of Lemma A.3. We first consider the class of function $\mathcal{G}$ defined in Section 3.1. This class of functions is represented as

$$\begin{aligned} \mathcal{G} = \{ & I[(a_u - 1)/(2r) < x_i \leq a_u/(2r)] \cdot I[(\tilde{a}_u - 1)/(2r) < x_j \leq \tilde{a}_u/(2r)] : \\ & a = (a_1, \ldots, a_k)', \tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_k)', \ (a_u, \tilde{a}_u) \in \{1, 2, \ldots, 2r\}^2 \\ & \text{for } u = 1, \ldots, k \text{ and } r = 1, 2, \ldots \}. \end{aligned}$$

Because the collection of cells on the real line is a Vapnik-Chervonenkis (VC) class of sets (see van der Vaart and Wellner (1996, Example 2.6.1), the collection of all subgraphs, $\{(x_i, x_j, t) : t < g(x_i, x_j)\}$, of the function in $\mathcal{G}$ forms a VC class of sets in $\mathcal{X}^2 \times \mathbb{R}$. Hence, $\mathcal{G}$ is a VC class of functions. Combining this result with Lemma 2.6.18 in van der Vaart and Wellner (1996), $\mathcal{F}_1$ and $\mathcal{F}_2$ are VC-classes of functions. Thus, from Corollary 19 in Nolan and Pollard (1987), $\mathcal{F}_1$ and $\mathcal{F}_2$ are Euclidean classes of functions. $\mathcal{F}_1$ and $\mathcal{F}_2$ obviously have the constant envelopes 1 and 1/2, respectively, from their definitions. □

I provide proofs of Lemmas A.1 and A.2 and Theorem 3.1 below.

Proof of Lemma A.1.(a). While Lemma A.1 is stated in terms of a subsequence $\{a_n\}$, for notational simplicity, I prove it for the sequence $\{n\}$. All of the arguments in this and the next proofs proceed with $\{a_n\}$ instead of $\{n\}$.

I use Theorem 5 in Nolan and Pollard (1988) to show that the weak convergence result in Lemma A.1.(a) holds. Let $N_p(\epsilon, R, \mathcal{F}, F)$ denote the $L_p(Q)$-covering number of radius ϵ for the functional space $\mathcal{F}$ with envelope function F where Q is some probability measure. Denote the class of functions $Pf_1(x, \cdot, \beta, g)$ on $\mathcal{X}$ by $P\mathcal{F}_1$, where f_1 and $\mathcal{F}_1$ are defined in Lemma A.3. Let $\beta \in B$ and P be the limit distribution of P_n.

To apply Theorem 5 in Nolan and Pollard (1988), it suffices to show that the following conditions hold:

(i) $\sup_Q \int_0^1 \log N_2(\epsilon, Q, \mathcal{F}_1, F)\, d\epsilon < \infty$, $\sup_Q P\left[\int_0^1 \log N_2(\epsilon, Q, \mathcal{F}_1, F)\, d\epsilon\right]^2 < \infty$, and $\sup_Q P\left[\int_0^1 \log N_2(\epsilon, Q, P\mathcal{F}_1, PF)\, d\epsilon\right]^2 < \infty$;

(ii) for each $\eta > 0$ and $\epsilon > 0$, there exists a $\gamma > 0$ such that

$$\lim_{n\to\infty} \sup_Q P\left(\int_0^\gamma \log N_2(\epsilon, Q, P\mathcal{F}_1, PF)\, d\epsilon > \eta\right) < \epsilon,$$

where Q is any probability measure. I show below that these two conditions are satisfied.

I first consider condition (i). From Lemma A.3, the class of functions $\mathcal{F}_1$ is Euclidean for the constant envelope $F = 1$. Then, from Corollaries 21 in Nolan and Pollard (1987), the class of functions $P\mathcal{F}_1$ is also a Euclidean class for the constant envelope 1. From page 789 in **?**, if a Euclidean class has a constant envelope function, then the upper bound of the $L_p(Q)$-covering number of radius ϵ for it is uniform in any probability measure Q. Therefore, since $\mathcal{F}_1$ and $P\mathcal{F}_1$ are Euclidean classes with constant envelopes, for any $0 < \epsilon \leq 1$, there exist some constants K_2, K_2^*, V_2, and V_2^* such that $N_2(\epsilon, R, \mathcal{F}_1, F) \leq K_2\epsilon^{-2V_2}$ and $N_2(\epsilon, R, P\mathcal{F}_1, PF) \leq K_2^*\epsilon^{-2V_2^*}$, for any probability measure Q. Then, it follows that

$$\sup_Q \int_0^1 \log N_2(\epsilon, Q, \mathcal{F}_1, F)\, d\epsilon \leq \int_0^1 \left(\log K_2\epsilon^{-2V_2}\right) d\epsilon < \infty,$$

$$\sup_Q \left[\int_0^1 \log N_2(\epsilon, Q, \mathcal{F}_1, F)\, d\epsilon\right]^2 \leq \left[\int_0^1 \left(\log K_2\epsilon^{-2V_2}\right) d\epsilon\right]^2 < \infty,$$

and

$$\sup_{Q}\left[\int_0^1 \log N_2\left(\epsilon, Q, P\mathcal{F}_1, PF\right) d\epsilon\right]^2 \leq \left[\int_0^1 \left(\log K_2^* \epsilon^{-2V_2^*}\right) d\epsilon\right]^2 < \infty.$$

These imply that condition (i) is satisfied.

Next, as $\gamma \searrow 0$,

$$\begin{aligned}\sup_{R}\int_0^\gamma \log N_2\left(\epsilon, R, P\mathcal{F}_1, PF\right) d\epsilon &\leq \int_0^\gamma \log K_2^* \epsilon^{-2V_2^*} d\epsilon \\ &= \gamma \log K_2^* - 2V_2^* \gamma \log \gamma \\ &\to 0.\end{aligned}$$

This implies that condition (ii) is satisfied. Therefore, from Theorem 5 in Nolan and Pollard (1988) and $\sigma_{P_n}(\beta) \underset{p}{\to} \sigma_P(\beta)$, Lemma A.1.(a) holds. □

Proof of Lemma A.1.(b). Let $\beta \in B$ and P be the limit distribution of P_n. Since $\mathcal{F}_1$ and $\mathcal{F}_2$ are Euclidean classes with constant envelopes from Lemma A.3, by applying Corollary 7 in Sherman (1994), it follows that

$$\sup_{\mathcal{G}}\left\|\frac{1}{n(n-1)}\sum_{i\neq j} f_1(w_i, w_j, \beta, g)\right\| \underset{p}{\to} 0$$

and

$$\sup_{\mathcal{G}^2}\left\|\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k} f_2(w_i, w_j, w_k, \beta, g, g^*)\right\| \underset{p}{\to} 0.$$

Therefore, letting $h_2(\beta, g, g^*)$ be given by (3.2) and further divided by $\sigma_P^2(\beta)$, as $\sigma_{P_n}(\beta) \underset{p}{\to} \sigma_P(\beta)$, we have

$$\begin{aligned}&\sup_{g,g^*\in\mathcal{G}^2}\left\|\hat{h}_{2,a_n,P_n}\left(\beta_{a_n}, g, g^*\right) - h_2\left(\beta, g, g^*\right)\right\| \\ \leq\ &\sup_{\mathcal{G}^2}\left\|\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k} f_2(w_i, w_j, w_k, \beta, g, g^*)/\sigma_P(\beta)\right\|\end{aligned}$$

$$+\left\{\sup_{\mathcal{G}}\left\|\frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_{i\neq j}f_1(w_i,w_j,\beta,g)/\sigma_P\left(\beta\right)\right\|\right\}^2+o_p\left(1\right)$$

$$\underset{p}{\to}0.$$

□

Proof of Lemma A.2. It suffices to show that

$$\begin{aligned} &E_P\left[m\left(W_i,W_j,\beta,g\right)\right]\geq 0 \quad \forall g\in\mathcal{G} \\ \Rightarrow\ &E_P\left[m\left(W_i,W_j,\beta\right)\mid X_i,X_j\right]\geq 0 \quad \text{a.s.} \end{aligned} \tag{A.1}$$

I invoke Lemma C1 in AS. Let $\mathcal{R}$ be a semiring of subsets of $\mathbb{R}^{2k}$ and

$$\mu\left(C\right)=E_P\left[m\left(W_i,W_j,\beta\right)I\left[\left(X_i,X_j\right)\in C\right]\right]$$

for $C\in\sigma\left(\mathcal{C}\right)=\mathcal{B}\left(\mathbb{R}^{2k}\right)$, where $\sigma\left(\mathcal{C}\right)$ denotes the σ-field generated by $\mathcal{C}$ and $\mathcal{B}\left(\mathbb{R}^{2k}\right)$ is the Borel σ-field on $\mathbb{R}^{2k}$. $\sigma\left(\mathcal{C}\right)=\mathcal{B}\left(\mathbb{R}^{2k}\right)$ is a well known result. I show that all conditions of Lemma C1 in AS are satisfied. Then, condition (A.1) holds from Lemma C1 in AS.

First, $\mathcal{C}$ is a semiring of subsets of $\mathbb{R}^{2k}$. Since $m\left(W_i,W_j,\beta\right)$ and $I\left[\left(X_i,X_j\right)\in C\right]$ are bounded functions, $\mu(\cdot)$ satisfies the boundedness condition of Lemma C1 in AS. The other conditions of Lemma C1 in AS also hold by the same argument of Lemma 3 in AS. Thus, by applying Lemma C1 in AS, $\mu\left(C\right)\geq 0$ for all $C\in\mathcal{C}$ implies that $E_P\left[m\left(W_i,W_j,\beta\right)I\left[\left(X_i,X_j\right)\in C\right]\right]\geq 0$ for all $\sigma\left(\mathcal{C}\right)$, which is equivalent to $\mathcal{B}\left(\mathbb{R}^{2k}\right)$. This implies that the result of Lemma A.2 in this appendix holds. □

Proof of Theorem 3.1. To show that Theorem 3.1 holds, it suffices to show that all required conditions in Theorems 2.1.(b) and 3 in AS are satisfied. Since we use the modified method moments function proposed by AS (Equation 3.8 in AS), Lemma 1 in AS guarantees that our definition of $T_{n,R}\left(\beta\right)$ satisfies Assumptions S1–4 in AS. Lemma A.1 in this appendix implies that Assumption EP in AS is satisfied. Assumption 2.2 guarantees that the continuity condition of Assumption GMS (a) in AS is satisfied. Assumption 3.1 implies that the tuning parameters in the GMS function satisfy Assumptions GMS1 and

GMS2 (b) and (c) in AS. Lemma A.2 in this appendix has the same role as Assumption CI in AS. Therefore, by the same arguments in Sections 12 and 14.1–2 in AS, Theorem 3.1 in this paper holds. □

A.3 Inference for the Transformation Function

This section provides a joint inference procedure for β_0 and $T_0(y)$, for a particular value of $y \in \mathbb{R}$, based on the identification results presented in Sections 2 and 6. The inference procedure is constructed by applying the conditional moment inequality inference approach presented in Section 3 to both the conditional moment inequality (3.1) and the one in (6.2).

Let

$$\begin{aligned} m_y^\dagger(W_i, W_j, \beta, t) = &(I[Y_{1i} \geq y] - I[Y_{0j} \geq \tilde{y}]) \cdot I[X_i'\beta - X_j'\beta \geq t] \\ &+ (I[Y_{1j} \geq y] - I[Y_{0i} \geq \tilde{y}]) \cdot I[X_j'\beta - X_i'\beta \geq t] \end{aligned}$$

and $m_y^\dagger(W_i, W_j, \beta, t, g) = m_y^\dagger(W_i, W_j, \beta, t) \cdot g(X_i, X_j)$ for any $g \in \mathcal{G}$. Then, $T_{I,\beta}(y)$ is equivalent to

$$\left\{t \in \mathbb{R} : \; E_P\left[m_y^\dagger(W_i, W_j, \beta, t, g)\right] \geq 0 \text{ for all } g \in \mathcal{G}\right\}.$$

Define the sample moment function and sample variance function, respectively, by

$$\bar{m}_{y,n}^\dagger(\beta, t, g) = \frac{1}{n(n-1)} \sum_{i \neq j} m_y^\dagger(W_i, W_j, \beta, t, g)$$

and

$$\begin{aligned} \hat{\sigma}_{y,n}^{\dagger 2}(\beta, t, g) = \Bigg\{ &\frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} m_y^\dagger(W_i, W_j, \beta, t, g) \, m_y^\dagger(W_i, W_k, \beta, t, g) \\ &- \left(\frac{1}{n(n-1)} \sum_{i \neq j} m_y^\dagger(W_i, W_j, \beta, t, g)\right)^2 \Bigg\}. \end{aligned}$$

Since the function $\bar{m}_{y,n}^\dagger(\beta, t, g)$ is a U-statistic of order two, the estimator of its asymptotic variance, $\hat{\sigma}_{y,n}^{\dagger 2}(\beta, t, g)$, is constructed by a similar form to $\hat{\sigma}_n^2(\beta, g)$ in Section 3.1. In

practice, we use the modified sample variance function:

$$\bar{\sigma}_{y,n}^{\dagger 2}(\beta,t,g)=\hat{\sigma}_{y,n}^{\dagger 2}(\beta,t,g)+\epsilon\hat{\sigma}_{y,n}^{\dagger 2},$$

where $\hat{\sigma}_{y,n}^{\dagger 2}=\hat{\sigma}_{y,n}^{\dagger 2}(\beta,t,1)$ and ϵ is the regularization parameter (e.g., $\epsilon=0.0001$). Then, an approximate test statistic at β and t is constructed as

$$\begin{aligned}T_{y,n,R}(\beta,t)=&\sum_{r=1}^{R}\left(r^2+100\right)^{-1}\sum_{(a,\tilde{a})\in\{1,\cdots,2r\}^2}(2r)^{-2K}\left(\left[\frac{n^{\frac{1}{2}}\bar{m}_n(\beta,g_{a,\tilde{a},r})}{\bar{\sigma}_n(\beta,g_{a,\tilde{a},r})}\right]_-^2\right.\\&\left.+\left[\frac{n^{\frac{1}{2}}\bar{m}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})}{\bar{\sigma}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})}\right]_-^2\right)\end{aligned}$$

for some truncation integer R chosen by the researcher. Note that this test statistic comprises two normalized sample moments, $\bar{m}_n(\beta,g_{a,\tilde{a},r})/\bar{\sigma}_n(\beta,g_{a,\tilde{a},r})$ and $\bar{m}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})/\bar{\sigma}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})$, for β_0 and $T_0(y)$, respectively.

We can compute the critical value for $T_{y,n,R}(\beta,t)$ as a simulated quantile of

$$\begin{aligned}T_{y,n,R}^{Asy}(\beta,t)&=\sum_{r=1}^{R}\left(r^2+100\right)^{-1}\sum_{(a,\tilde{a})\in[1,\cdots,2r]^2}(2r)^{-2K}\left(\left[\frac{v_n(\beta,g_{a,\tilde{a},r})+\varphi_n(\beta,g_{a,\tilde{a},r})}{\bar{\sigma}_n(\beta,g_{a,\tilde{a},r})}\right]_-^2\right.\\&\left.+\left[\frac{v_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})+\varphi_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})}{\bar{\sigma}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})}\right]_-^2\right),\end{aligned}$$

where $\left(v_{y,n}^{\dagger}(\beta,t,g)\right)_{g\in\mathcal{G}}$ is a zero mean Gaussian process with a covariance kernel evaluated by

$$\begin{aligned}\hat{h}_{y,2}^{\dagger}(\beta,t,g,g^*)&=\left\{\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k}m_y^{\dagger}(W_i,W_j,\beta,t,g)\,m_y^{\dagger}(W_i,W_k,\beta,t,g^*)\right.\\&\left.-\left[\frac{1}{n(n-1)}\sum_{i\neq j}m_y^{\dagger}(W_i,W_j,\beta,t,g)\right]\cdot\left[\frac{1}{n(n-1)}\sum_{i\neq j}m_y^{\dagger}(W_i,W_j,\beta,t,g^*)\right]\right\},\end{aligned}$$

and $\varphi_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})$ is a GMS function given by

$$\varphi_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})=\hat{\sigma}_{y,n}^{\dagger 2}B_nI\left[\kappa_n^{-1}n^{\frac{1}{2}}\bar{m}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})/\bar{\sigma}_{y,n}^{\dagger}(\beta,t,g_{a,\tilde{a},r})>1\right].$$

Here B_n and κ_n are two tuning parameters that should satisfy Assumption 3.1. (e.g., $\kappa_n = \left(\left(1-\hat{p}_{1-D}^{1/3}\right)^{2/5} \times 0.6\ln(n)\right)^{\frac{1}{2}}$ and $B_n = (0.8\ln(n)/\ln\ln(n))^{\frac{1}{2}}$). For a significance level of $\alpha < 1/2$, let $\hat{c}_{y,\eta,1-\alpha}(\beta,t)$ be the $1-\alpha+\eta$ sample quantile of $T^{Asy}_{y,n,R}(\beta,t)$ with an arbitrarily small number η (e.g., 10^{-6}). Then, the $(1-\alpha)$-level confidence set for $(\beta_0, T_0(y))$ is computed by

$$\{(\beta,t) \in B \times \mathbb{R} : T_{y,n,R}(\beta,t) \leq \hat{c}_{y,\eta,1-\alpha}(\beta,t)\}.$$

The size and power properties stated in Theorem 3.1 could apply to this confidence set.

References

Andrews, D. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666.

——— (2014): "Nonparametric inference based on conditional moment inequalities," *Journal of Econometrics*, 179, 31–45.

——— (2017): "Inference based on many conditional moment inequalities," *Journal of Econometrics*, 196, 275–287.

Andrews, D. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157.

Armstrong, T. (2014): "Weighted KS statistics for inference on conditional moment inequalities." *Journal of Econometrics*, 181, 92–116.

——— (2015): "Asymptotically exact inference in conditional moment inequality models," *Journal of Econometrics*, 186, 51–65.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821.

——— (2012): "Partial identification using random set theory," *Journal of Econometrics*, 166, 17–92.

Beresteanu, A. and I. Molinari (2008): "Asymptotic properties for a class of partially identified models," *Econometrica*, 76, 763–814.

Blevins, J. (2011): "Partial identification and inference in binary choice and duration panel data models," Working Paper.

Bugni, F. (2010): "Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set," *Econometrica*, 78, 735–753.

Chen, S. (2002): "Rank estimator of transformation models," *Econometrica*, 70, 1683–1697.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. Rosen (2013): "Inference with intersection bounds," *Econometrica*, 81, 667–737.

Cox, D. (1972): "Regression models and life tables," *Journal of the Royal Statistical Society (Series B)*, 34, 187–220.

Fan, Y. and R. Liu (2018): "Partial identification and inference in censored quantile regression," *Journal of Econometrics*, 206, 1–38.

Han, A. (1987): "Non-parametric analysis of a generalized regression model," *Journal of Econometrics*, 35, 303–316.

Hong, H. and E. Tamer (2003): "Inference in censored models with endogenous regressors," *Econometrica*, 71, 905–932.

Honoré, B., S. Khan, and J. Powell (2002): "Quantile regression under random censoring," *Journal of Econometrics*, 109, 67–105.

Honoré, B. and A. Lleras-Muney (2006): "Bounds in competing risks models and the war on cancer," *Econometrica*, 74, 1675–1698.

Horowitz, J. (1996): "Semiparametric estimation of a regression model with an unknown transformation of the dependent variable," *Econometrica*, 64, 103–137.

——— (2009): *Semiparametric and Nonparametric Methods in Econometrics*, New York: Springer-Verlag.

Imbens, G. and C. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

Kalbfleisch, J. and R. Prentice (1980): *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Khan, S., M. Ponomareva, and E. Tamer (2011): "Sharpness in randomly censored linear models," *Economics Letters*, 113, 23–25.

——— (2016): "Identification of panel data models with endogeneous censoring," *Journal of Econometrics*, 194, 57–75.

Khan, S. and E. Tamer (2007): "Partial rank estimation of duration models with general forms of censoring," *Journal of Econometrics*, 136, 251–280.

——— (2009): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152, 104–119.

Kim, D. (2018): "Partially identifying competing risks models: Applications to the war on cancer and unemployment spells," Job Market Paper.

Komarova, T. (2013): "Binary choice models with discrete regressors: Identification and misspecification," *Journal of Econometrics*, 177, 14–33.

Lancaster, T. (1979): "Econometric methods for the duration of unemployment," *Econometrica*, 47, 939–957.

——— (1990): *The Econometric Analysis of Transition Data*, Cambridge: Cambridge University Press.

Li, T. and T. Oka (2015): "Set identification of the censored quantile regression model for short panels with fixed effects," *Journal of Econometrics*, 188, 363–377.

Magnac, T. and E. Maurin (2008): "Partial identification in monotone binary models: Discrete regressors and interval data," *Review of Economic Studies*, 75, 835–864.

Menzel, K. (2014): "Consistent estimation with many moment inequalities," *Journal of Econometrics*, 182, 329–350.

Molchanov, I. (2005): *Theory of Random Sets*, London: Springer-Verlag.

Nolan, D. and D. Pollard (1987): "U-processes: Rates of convergence," *Annals of Statistics*, 15, 780–799.

——— (1988): "Functional limit theorems for U-processes," *Annals of Probability*, 16, 1291–1298.

Pollard, D. (1990): "Empirical Process Theory and Application," in *NSF-CBMS Regional Conference Series in Probability and Statistics*, Hayward: Institute of Mathematical Statistics, vol. II.

Powell, J. (1984): "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, 25, 303–325.

Romano, J. and A. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planing and Inference*, 138, 2786–2807.

——— (2010): "Inference for the identified set in partially identified econometric models," *Econometrica*, 78, 169–211.

Sherman, R. (1994): "Maximal inequalities for degenerate U-processes with applications to optimization estimators," *The Annals of Statistics*, 22, 439–459.

Stoye, J. (2009): "More on confidence regions for partially identified parameters," *Econometrica*, 77, 1299–1315.

Szydłowski, A. (2019): "Endogenous censoring in the mixed proportional hazard model with an application to optimal unemployment insurance," *Journal of Applied Econometrics*, 34, 1086–1011.

van den Berg, G. (2001): "Duration models: Specification, identification and multiple durations," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Amsterdam: Elsevier, vol. 5, 3381–3460.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes*, New York: Springer.

Yang, S. (1999): "Censored median regression using weighted empirical survival and hazard functions," *Journal of the American Statistical Association*, 94, 137–145.

Ye, J. and N. Duan (1997): "Nonparametric $n^{-1/2}$-consistent estimation for the general transformation models," *Annals of Statistics*, 6, 2682–2717.

Ying, Z., S. Jung, and L. Wei (1995): "Survival analysis with median regression models," *Journal of the American Statistical Association*, 90, 178–184.