TEST (2024) 33:473–495
https://doi.org/10.1007/s11749-023-00903-9

ORIGINAL PAPER

An instrumental variable approach under dependent censoring

Gilles Crommen[1] · **Jad Beyhum**[2] · **Ingrid Van Keilegom**[1]

Received: 8 January 2023 / Accepted: 4 November 2023 / Published online: 14 December 2023

Abstract
This paper considers the problem of inferring the causal effect of a variable Z on a dependently censored survival time T. We allow for unobserved confounding variables, such that the error term of the regression model for T is dependent on the confounded variable Z. Moreover, T is subject to dependent censoring. This means that T is right censored by a censoring time C, which is dependent on T (even after conditioning out the effects of the measured covariates). A control function approach, relying on an instrumental variable, is leveraged to tackle the confounding issue. Further, it is assumed that T and C follow a joint regression model with bivariate Gaussian error terms and an unspecified covariance matrix, such that the dependent censoring can be handled in a flexible manner. Conditions under which the model is identifiable are given, a two-step estimation procedure is proposed, and it is shown that the resulting estimator is consistent and asymptotically normal. Simulations are used to confirm the validity and finite-sample performance of the estimation procedure. Finally, the proposed method is used to estimate the causal effect of job training programs on unemployment duration.

Keywords Dependent censoring · Causal inference · Instrumental variable · Control function · Survival analysis

Mathematics Subject Classification 62N02 · 62F12 · 62D20

✉ Gilles Crommen
gilles.crommen@kuleuven.be

Jad Beyhum
jad.beyhum@gmail.com

Ingrid Van Keilegom
ingrid.vankeilegom@kuleuven.be

[1] ORSTAT, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

[2] Department of Economics, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

1 Introduction

When estimating the effect of a variable Z on a censored survival time T, unmeasured confounding can be a possible source of bias. A common approach to address this bias without actually having to observe the unmeasured confounders is to use methods based on an instrumental variable (IV). Within survival analysis, IV methods have recently been receiving increased attention to estimate causal effects on censored outcomes. However, almost all of these approaches assume that the censoring time is (conditionally) independent of the survival time. In this work, we propose an IV method that can identify causal effects while allowing for dependent censoring. Precisely, let T depend log-linearly on a vector of observed covariates X, a confounded variable Z and some error term, denoted by u_T, which represents unobserved heterogeneity. There is a confounding issue when Z and u_T are dependent on each other. A common example is when Z is a non-randomized binary treatment variable, even after conditioning on the covariates X. This dependence of Z on the error term implies that the causal effect of Z on T cannot be identified from the conditional distribution of T on (X, Z). Further, we introduce a right censoring mechanism by way of the censoring time C, such that only the minimum of T and C is observed through the follow-up time $Y = \min\{T, C\}$ and the censoring indicator $\Delta = \mathbb{1}(T \leq C)$. We do not assume that T and C are independent, even after conditioning on (X, Z). This possible dependence creates an additional statistical issue since the distribution of T cannot be recovered from that of (Y, Δ) without further assumptions.

1.1 Approach

The confounding issue is tackled by utilizing a control function approach. This method uses an instrumental variable $\tilde{W}$, the observed covariates X and the confounded variable Z to split u_T into two parts: one which is dependent on Z, and one which is not. The exact conditions that $\tilde{W}$ has to satisfy to be a valid instrument are described in Sect. 2.1 of the paper. The part of u_T that is dependent on Z is the control function V, for which it is assumed that u_T is linear in V. This control function (also denoted by g) is a function of $(Z, X, \tilde{W})$ and a parameter γ that captures all unmeasured confounding. Note that the mapping g follows from the reduced form (which is specified by the analyst), but the parameter γ is unknown and will need to be estimated. In this work, we propose two possible control functions for which the expressions depend on the support of Z. Moreover, the control function allows us to estimate the causal effect of Z on T from the conditional distribution of T on (X, Z, V). To allow for dependent censoring, we introduce a joint Gaussian regression model with an unspecified covariance matrix. The specific need for this assumption is explained in Sect. 2.1, where it is formally introduced.

In addition, it is shown that our model is identifiable, which means that we can identify not only the causal effect of Z on T but also the association parameter between T and C. This can be seen as surprising, since we only observe the minimum of T and C through the follow-up time Y and the censoring indicator Δ. In order to estimate the model parameters, a two-step estimation method is proposed. The first step estimates

the parameter γ, which is required to construct the regressor V. Therefore, this control function V can also be seen as a generated regressor. The second step uses maximum likelihood to estimate parameters of interest such as the correlation between T and C and the causal effect of Z on T. Note that the second step uses the generated regressor V, such that a correction for the randomness coming from the first step needs to be applied to get asymptotically valid standard errors. To implement this correction, we treat the two steps as a joint generalized method of moments estimator with their moment conditions stacked in one vector. This allows us to prove consistency and asymptotic normality of the parameter estimates. Using various simulation settings, we show that the estimator demonstrates excellent finite sample performances. We illustrate the procedure by evaluating the effect of federally funded job training services on unemployment duration in the USA.

1.2 Related literature

This paper is firstly related to the literature on dependent censoring. In the survival analysis literature, it is usually assumed that the survival time T is independent of the right censoring time C, which is called independent censoring. However, it is easy to think of situations where this assumption is not a reasonable one to make. A common example of the independent censoring assumption being doubtful can be found in transplant studies. The survival time (time to death) is likely dependent on the censoring time (time to transplant), since selection for transplant is based on the patient's medical condition. In this case we would expect a positive dependence between T and C, as usually the most ill patients are selected for transplant (Staplin et al 2015). In the literature, many methods have been proposed to handle dependent censoring. An important result comes from Tsiatis (1975), who proved that it is impossible to identify the joint distribution of two failure times by their minimum in a fully nonparametric way. Because of this, more information about the dependence and/or marginal distributions of T and C is needed to identify their joint distribution. The most popular approaches are based on copulas, and Zheng and Klein (1995) were the first to apply this idea. Under the assumption of a fully known copula for the joint distribution of T and C, a nonparametric estimator of the marginals was proposed. This estimator is called the copula-graphic estimator, which extends the Kaplan and Meier (1958) estimator to the dependent censoring case. Rivest and Wells (2001) further investigated the copula-graphic estimator for Archimedean copulas. Note that both of these methods rely on a completely known copula. In particular, this means that the association parameter specifying the dependence between T and C is assumed to be known, which is often not the case in practice. The copula methods were extended to include covariates by Braekers and Veraverbeke (2005), Huang and Zhang (2008) and Sujica and Van Keilegom (2018) among others. Nevertheless, these methods still rely on a fully known copula. More recently a new method was proposed by Czado and Van Keilegom (2023), which does not require the association parameter to be known. As a trade-off, this requires the marginals to be fully parametric for the association parameter to be identifiable. Deresa and Van Keilegom (2020c) and Deresa and Van Keilegom (2020a) propose a semiparametric and parametric transformed joint

regression model, respectively, where the transformed variables T and C follow a bivariate normal distribution after adjusting for covariates. Deresa and Van Keilegom (2020b) extends the parametric transformed joint regression model to allow for different types of censoring. The present paper relies on a similar Gaussian model as Deresa and Van Keilegom (2020a), but nevertheless differs from it as we allow for confounding. Therefore, our method can be seen as a generalization of the one proposed by Deresa and Van Keilegom (2020a). The added complication comes from the generated regressor that is introduced by the control function.

Secondly, the present work falls within the instrumental variable and control function literature. A confounding issue could occur due to a multitude of reasons such as noncompliance (Angrist et al 1996), sample selection (Heckman 1979), measurement error or omitting relevant variables. The control function approach used in this work has been discussed extensively in the literature on confounding and endogeneity by Lee (2007), Navarro (2010) and Wooldridge (2015) among others. The idea is that adding an appropriate parametric control function to the regression, which is estimated in the first stage using a valid instrument, solves the confounding issue. The advantages of this approach are that it is computationally simple and that it can handle complicated models that are nonlinear in the confounded variable in a parsimonious manner. It is interesting to note that using the control function method creates a generated regressor problem. See Pagan (1984), Oxley and McAleer (1993) and Sperlich (2009) for an overview of possible methods and issues raised when using generated regressors. Moreover, Escanciano et al (2016) look at a general framework for two-step estimators with a non-parametric first step. In this work, they consider the example of a control function estimator for a binary choice model with an endogenous regressor.

Finally, the last string of research linked to this paper is that of instrumental variable methods for right censored data. We first discuss methods assuming that the censoring mechanism is independent. Some papers follow a nonparametric approach assuming that both Z and $\tilde{W}$ are categorical: Frandsen (2015), Sant'Anna (2016) and Beyhum et al (2022a). Other approaches are semiparametric, such as Bijwaard and Ridder (2005), Li et al (2015), Tchetgen Tchetgen et al (2015), Chernozhukov et al (2015) and Beyhum et al (2022b) among others. Note that Tchetgen Tchetgen et al (2015) also propose a control function approach. Centorrino and Florens (2021) study nonparametric estimation with continuous regressors. Confounding has also been discussed in a competing risks framework by Richardson et al (2017), Zheng et al (2017), Martinussen and Vansteelandt (2020) and Beyhum et al (2023). However, research on confounding within a dependent censoring framework is sparse. Firstly, Robins and Finkelstein (2000) look at a correction for noncompliance and dependent censoring. However, they make the strong assumption that conditional on the treatment arm and the recorded history of six time-dependent covariates, C does not further depend on T. It is clear that this assumption is violated if there is a variable affecting both T and C that is not observed. Secondly, Khan and Tamer (2009) discuss an endogenously censored regression model, but they make a strong assumption (IV2, page 110 in Khan and Tamer (2009)) regarding the relationship between the instruments and the covariates. An example of this assumption being violated is when the support of the natural logarithm of C given Z and X is the whole real line, which is allowed for in our model. Finally, Blanco et al (2020) look at treatment effects on duration outcomes

under censoring, selection, and noncompliance. However, they derive bounds on the causal effect instead of point estimates.

1.3 Outline

In Sect. 2, we specify the model to be studied and describe some distributions such that the expected conditional log-likelihood can be defined. Section 3.1 derives the identification results and Sect. 3.2 outlines the estimation procedure. Section 3.3 shows consistency and asymptotic normality for the estimator described in Sect. 3.2. Section 3.4 describes how the asymptotic variance can be estimated. The technical details for the three theorems outlined in Sect. 3 can be found in Sections B and C of the supplementary information. Simulation results and an empirical application regarding the impact of Job Training Partnership Act (JTPA) programs on time until employment are described in Sects. 4 and 5, respectively. The R code used for both of these sections can be found on https://github.com/GillesCrommen.

2 The model

2.1 Model specification

Let T and C be the natural logarithm of the survival and censoring time, respectively. Because T and C censor each other, only one of them is observed through the follow-up time $Y = \min\{T, C\}$ and the censoring indicator $\Delta = \mathbb{1}(T \leq C)$. The measured covariates that have a direct effect on both T and C are given by $X = (1, \tilde{X}^\top)^\top$ and Z, where $\tilde{X}$ and Z are of dimension m and 1, respectively. More precisely, suppose we have the following structural equation system:

$$\begin{cases} T = X^\top \beta_T + Z\alpha_T + V\lambda_T + \epsilon_T \\ C = X^\top \beta_C + Z\alpha_C + V\lambda_C + \epsilon_C \end{cases}, \tag{1}$$

where (ϵ_T, ϵ_C) are unobserved error terms and V an unobserved confounder of Z. Note that if V were to be observed, we could directly estimate the causal effect α_T by applying the method of Deresa and Van Keilegom (2020a) to model (1). However, since V is not observed, using this method would lead to biased estimates of α_T. To resolve this issue, we use an instrumental variable $\tilde{W}$ that is sufficiently dependent on Z (conditionally on X). In addition, if we were to construct V such that $\mathbb{E}[V \mid W] = \mathbb{E}[V]$, we can think of V as the part of Z that does not depend on W, where $W = (X^\top, \tilde{W})^\top$. Note that all the V's proposed in Sect. 2.2 satisfy this mean independence property.

More precisely, we let $V = g_\gamma(Z, W)$ for which the control function g is known up to the parameter γ. Note that this control function follows from the reduced form, which is specified by the analyst. Further, it is assumed that:

(A1) $\begin{pmatrix}\epsilon_T\\ \epsilon_C\end{pmatrix} \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \quad \Sigma_\epsilon = \begin{pmatrix}\sigma_T^2 & \rho\sigma_T\sigma_C\\ \rho\sigma_T\sigma_C & \sigma_C^2\end{pmatrix}\right)$, with Σ_ϵ positive definite ($\sigma_T, \sigma_C > 0$ and $|\rho| < 1$).
(A2) $(\epsilon_T, \epsilon_C) \perp\!\!\!\perp (W, Z)$, where $\perp\!\!\!\perp$ denotes statistical independence.
(A3) The covariance matrix of $(\tilde{X}^\top, Z, V)$ is full rank and $\text{Var}(\tilde{W}) > 0$.
(A4) The probabilities $\mathbb{P}(Y = T \mid W, Z)$ and $\mathbb{P}(Y = C \mid W, Z)$ are both strictly positive almost surely.

Assumption (A1) implies that, conditional on (W, Z), both T and C are normally distributed and allowed to be dependent on each other due to the correlation parameter ρ. As mentioned in the introduction, Tsiatis (1975) showed that it is impossible to identify the joint distribution of two failure times by their minimum in a fully nonparametric way. Because of this result, we need to make some assumptions regarding the dependence and/or marginal distributions of T and C in order to identify their joint distribution. When dependent censoring is still present after conditioning on the covariates, there are two common approaches that can be considered. The first one consists of specifying a fully known copula for the joint distribution of T and C, while leaving the marginals unspecified (see Emura and Chen (2018) for more details). This means that the association parameter, which describes the dependence between T and C, is assumed to be known. As this is often not the case in practice, we opt to use a different method. At the cost of using the fully parametric model that follows from Assumption (A1), it will later be shown by Theorem 1 that we can actually identify the association parameter ρ. We deem this to be an acceptable price to pay, as there is no good way of choosing the association parameter in practice. The possible relaxation of this assumption is discussed in the Future research section at the end of the paper. Secondly, Assumption (A2) tells us that V is the only unobserved confounder of Z. Lastly, Assumptions (A3) and (A4) are commonly made in a survival analysis context, except for the second part of Assumption (A3), which can be interpreted as a nontrivial assignment assumption when $\tilde{W}$ is binary.

2.2 The control function

In the literature, different control functions have been proposed. Following Wooldridge (2010) and Navarro (2010), we give two examples of possible control functions that will be used throughout the paper. Consider first the case where Z is a continuous random variable and the relation between Z and W follows a linear model, that is

$$Z = W^\top\gamma + \nu \qquad \text{with} \qquad \mathbb{E}[\nu \mid W] = 0, \tag{2}$$

where ν is an unobserved error term and $\gamma \in \mathbb{R}^{m+2}$. In this setting it is natural to set $V = g_\gamma(Z, W) = Z - W^\top\gamma$ such that V is the confounded part of Z, that is, the part that does not depend on W. Another, more involved, example follows from Z being a binary random variable where the relation between Z, W and ν is specified as

$$Z = \mathbb{1}(W^\top\gamma - \nu > 0) \qquad \text{with} \qquad \nu \perp\!\!\!\perp W. \tag{3}$$

Since we cannot directly separate ν from Z and W, we let

$$V = g_\gamma(Z, W) = Z\,\mathbb{E}[\nu \mid W^\top\gamma > \nu] + (1 - Z)\,\mathbb{E}[\nu \mid W^\top\gamma < \nu]. \tag{4}$$

Then, the function g is known, up to γ, when the distribution of ν is known. This specification of the control function is discussed and justified in Section 19.6.1 and Section 21.4.2 by Wooldridge (2010). If $\nu \sim N(0, 1)$ or ν follows a standard logistic distribution, we have a probit or logit model for Z, respectively. Specific expressions of g for the probit and logit model can be found in Section A of the supplementary material. Moreover, when Z is binary, Tchetgen Tchetgen et al (2015) give another example of a possible control function:

$$V = Z - \mathbb{P}(Z = 1 \mid W).$$

2.3 Useful distributions and definitions

Using the assumptions that have been made so far, some conditional distributions and densities are derived. They are useful in proving the identification theorem and to define the estimator in Sect. 3. The expected log-likelihood function is also defined. For a given $\theta = (\beta_T, \alpha_T, \lambda_T, \beta_C, \alpha_C, \lambda_C, \sigma_T, \sigma_C, \rho)^\top$ and γ, we define $F_{T|W,Z}(\cdot \mid w, z, \gamma; \theta)$ and $F_{C|W,Z}(\cdot \mid w, z, \gamma; \theta)$ as the conditional distribution function of T and C given $W = w = (x^\top, \tilde{w})^\top$ and $Z = z$, respectively. Thanks to Assumptions (A1) and (A2), we have that:

$$F_{T|W,Z}(t \mid w, z, \gamma; \theta) = \Phi\left(\frac{t - x^\top\beta_T - z\alpha_T - g_\gamma(z, w)\lambda_T}{\sigma_T}\right),$$
$$F_{C|W,Z}(c \mid w, z, \gamma; \theta) = \Phi\left(\frac{c - x^\top\beta_C - z\alpha_C - g_\gamma(z, w)\lambda_C}{\sigma_C}\right),$$

with Φ the cumulative distribution function of a standard normal variable. It follows that for a given γ and θ, the conditional density functions of T and C given $W = w$ and $Z = z$ are, respectively:

$$f_{T|W,Z}(t \mid w, z, \gamma; \theta) = {\sigma_T}^{-1}\phi\left(\frac{t - x^\top\beta_T - z\alpha_T - g_\gamma(z, w)\lambda_T}{\sigma_T}\right),$$
$$f_{C|W,Z}(c \mid w, z, \gamma; \theta) = {\sigma_C}^{-1}\phi\left(\frac{c - x^\top\beta_C - z\alpha_C - g_\gamma(z, w)\lambda_C}{\sigma_C}\right),$$

where ϕ is the density function of a standard normal variable. For ease of notation, define $b_C = y - x^\top\beta_C - z\alpha_C - g_\gamma(z, w)\lambda_C$ and $b_T = y - x^\top\beta_T - z\alpha_T - g_\gamma(z, w)\lambda_T$. The sub-distribution function $F_{Y,\Delta|W,Z}(\cdot, 1 \mid w, z, \gamma; \theta)$ of (Y, Δ) given (W, Z) and

(γ, θ) can be derived as follows:

$$\begin{aligned} F_{Y,\Delta|W,Z}(y, 1 \mid w, z, \gamma; \theta) &= \mathbb{P}(Y \leq y, \Delta = 1 \mid W = w, Z = z) \\ &= \mathbb{P}(Y \leq y, T \leq C \mid W = w, Z = z) \\ &= \mathbb{P}(\epsilon_T \leq b_T, b_C - b_T + \epsilon_T \leq \epsilon_C). \end{aligned}$$

This expression is equivalent to

$$\int_{-\infty}^{b_T} \mathbb{P}(\epsilon_C \geq b_C - b_T + e \mid \epsilon_T = e) f_{\epsilon_T}(e) de.$$

Since $(\epsilon_C \mid \epsilon_T = e) \sim N\left(\rho \frac{\sigma_C}{\sigma_T} e, \sigma_C^2(1-\rho^2)\right)$ and $\epsilon_T \sim N(0, \sigma_T^2)$, it follows that

$$f_{Y,\Delta|W,Z}(y, 1 \mid w, z, \gamma; \theta) = \frac{1}{\sigma_T}\left[1 - \Phi\left(\frac{b_C - \rho \frac{\sigma_C}{\sigma_T} b_T}{\sigma_C (1-\rho^2)^{\frac{1}{2}}}\right)\right] \phi\left(\frac{b_T}{\sigma_T}\right).$$

Using the same arguments, it can be shown that

$$f_{Y,\Delta|W,Z}(y, 0 \mid w, z, \gamma; \theta) = \frac{1}{\sigma_C}\left[1 - \Phi\left(\frac{b_T - \rho \frac{\sigma_T}{\sigma_C} b_C}{\sigma_T (1-\rho^2)^{\frac{1}{2}}}\right)\right] \phi\left(\frac{b_C}{\sigma_C}\right).$$

Since $\mathbb{P}(Y \leq y) = \mathbb{P}(T \leq y) + \mathbb{P}(C \leq y) - \mathbb{P}(T \leq y, C \leq y)$, we have that:

$$F_{Y|W,Z}(y \mid w, z, \gamma; \theta) = \Phi\left(\frac{b_T}{\sigma_T}\right) + \Phi\left(\frac{b_C}{\sigma_C}\right) - \Phi\left(\frac{b_T}{\sigma_T}, \frac{b_C}{\sigma_C}; \rho\right), \tag{5}$$

where $\Phi(\cdot, \cdot, \rho)$ is the distribution function of a bivariate normal distribution with covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Further, let $S = (Y, \Delta, \tilde{X}, \tilde{W}, Z)$ with distribution function G on $\mathcal{G} = \mathbb{R} \times \{0, 1\} \times \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}$ and $\ell : \mathcal{G} \times \Gamma \times \Theta \to \mathbb{R} : (s, \gamma, \theta) \mapsto \ell(s, \gamma, \theta) = \log f_{Y,\Delta|W,Z}(y, \delta \mid w, z, \gamma; \theta)$, where $\Theta \subset \{\theta : (\beta_T, \alpha_T, \lambda_T, \beta_C, \alpha_C, \lambda_C) \in \mathbb{R}^{2m+6}, (\sigma_T, \sigma_C) \in \mathbb{R}^2_{>0}, \rho \in (-1, 1)\}$ is the parameter space of θ and Γ the parameter space of γ (usually $\Gamma \subset \mathbb{R}^{m+2}$). The expected conditional log-likelihood (given W, Z) can be defined as follows:

$$L(\gamma, \theta) = \mathbb{E}\left[\ell(S, \gamma, \theta)\right] = \int_{\mathcal{G}} \ell(s, \gamma, \theta) dG(s).$$

3 Model identification and estimation

3.1 Identification of the model

We will start by showing that model (1) is identifiable in the sense that two different values of the parameter vector (γ, θ) result in two different distributions of S. Let

(γ^*, θ^*) denote the true parameter vector. In order to prove the identifiability of the model, it will be assumed that:

(A5) γ^* is identified.

Considering again the examples from Sect. 2.2, when Z is a continuous random variable for which (2) holds, it is well known that the assumption that the covariance matrix of $(\tilde{X}, \tilde{W})$ is full rank implies Assumption (A5). When Z is a binary random variable for which (3) holds, the assumption that the covariance matrix of $(\tilde{X}, \tilde{W})$ is full rank together with a known distributional assumption on ν (e.g., $\nu \sim N(0, 1)$ or $\nu \sim$ Logistic$(0, 1)$) implies Assumption (A5) as shown by Manski (1988).

Theorem 1 *Under Assumptions (A1)–(A5), suppose that* (T_1, C_1) *and* (T_2, C_2) *satisfy model* (1) *with* (γ, θ_1) *and* (γ, θ_2) *as parameter vectors, respectively. If* $f_{Y_1,\Delta_1|W,Z}(\cdot, k \mid w, z, \gamma; \theta_1) \equiv f_{Y_2,\Delta_2|W,Z}(\cdot, k \mid w, z, \gamma; \theta_2)$ *for almost every* (w, z)*, then*

$$\theta_1 = \theta_2.$$

The proof of the theorem can be found in Section C.1 of the supplementary material, and is based on the proof of Theorem 1 by Deresa and Van Keilegom (2020a). The fact that the proposed joint regression model is identifiable can be seen as surprising, since this means that we can identify the relationship between T and C while only observing their minimum through the follow-up time Y and the censoring indicator Δ.

3.2 Estimation of the model parameters

We consider estimation when the data consist of an i.i.d. sample $\{Y_i, \Delta_i, W_i, Z_i\}_{i=1,\ldots,n}$. Further, it is assumed that:

(A6) There exists a known function $m : (w, z, \gamma) \in \mathbb{R}^{m+2} \times \mathbb{R} \times \Gamma \mapsto m(w, z, \gamma)$ twice continuously differentiable with respect to γ such that the estimator

$$\hat{\gamma} \in \arg\max_{\gamma \in \Gamma} n^{-1} \sum_{i=1}^{n} m(W_i, Z_i, \gamma), \tag{6}$$

is consistent for the true parameter γ^*.

Using the first-order conditions of program (6), we obtain that $n^{-1} \sum_{i=1}^{n} \nabla_\gamma m(W_i, Z_i, \hat{\gamma}) = 0$. Hence, Assumption (A6) implies that we possess a consistent Z-estimator of γ. The theory on M-estimators (Newey and McFadden 1994) allows us to find sufficient conditions for the assumption that $\hat{\gamma}$ is consistent. Assumption (A6) will hold when (i) the true parameter γ^* belongs to the interior of Γ, which is compact, (ii) $\mathcal{L}(\gamma) = \mathbb{E}\left[m(W, Z, \gamma)\right]$ is continuous and uniquely maximized at γ^* and (iii) $\hat{\mathcal{L}}(\gamma) = n^{-1} \sum_{i=1}^{n} m(W_i, Z_i, \gamma)$ converges uniformly (in $\gamma \in \Gamma$) in probability to $\mathcal{L}(\gamma)$. In the case where $\hat{\mathcal{L}}(\cdot)$ is concave, (i) can be weakened to γ^* being an

element of the interior of a convex set Γ, while (iii) is only required to hold pointwise rather than uniformly. Returning again to the examples given in Sect. 2.2, when Z is a continuous random variable for which (2) holds, it is well known that ordinary least squares is an extremum estimation method that consistently estimates γ under the assumption that the covariance matrix of $(\tilde{X}, \tilde{W})$ is full rank. In this case, we can define $m(W, Z, \gamma) = -(Z - W^\top \gamma)^2$. When Z is a binary random variable for which (3) holds and the distribution of ν is known, maximum likelihood estimation can be used to consistently estimate γ under weak regularity conditions that can be found in Aldrich and Nelson (1991). In this case, we can define $m(W, Z, \gamma) = Z \log \mathbb{P}(W^\top \gamma > \nu \mid W) + (1 - Z) \log \mathbb{P}(W^\top \gamma < \nu \mid W)$.

After obtaining $\hat{\gamma}$ from (6), the parameters from model (1) can be estimated using maximum likelihood with the estimates given by the second-step estimator:

$$\hat{\theta} = (\hat{\beta}_T, \hat{\alpha}_T, \hat{\lambda}_T, \hat{\beta}_C, \hat{\alpha}_C, \hat{\lambda}_C, \hat{\sigma}_T, \hat{\sigma}_C, \hat{\rho}) = \underset{\theta \in \Theta}{\arg\max}\, \hat{L}(\hat{\gamma}, \theta), \tag{7}$$

with Θ the parameter space as defined before and

$$\begin{aligned}
\hat{L}(\hat{\gamma}, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \log f_{Y,\Delta|W,Z}(Y_i, \Delta_i \mid W_i, Z_i, \hat{\gamma}; \theta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \Delta_i \left(-\log(\sigma_T) + \log \left[1 - \Phi \left(\frac{b_{C_i} - \rho \frac{\sigma_C}{\sigma_T} b_{T_i}}{\sigma_C (1 - \rho^2)^{\frac{1}{2}}} \right) \right] \right. \right. \\
&\quad \left. + \log \left[\phi \left(\frac{b_{T_i}}{\sigma_T} \right) \right] \right) \\
&\quad + (1 - \Delta_i) \left(-\log(\sigma_C) + \log \left[1 - \Phi \left(\frac{b_{T_i} - \rho \frac{\sigma_T}{\sigma_C} b_{C_i}}{\sigma_T (1 - \rho^2)^{\frac{1}{2}}} \right) \right] \right. \\
&\quad \left. \left. + \log \left[\phi \left(\frac{b_{C_i}}{\sigma_C} \right) \right] \right) \right\},
\end{aligned}$$

with $b_{C_i} = Y_i - X_i^\top \beta_C - Z_i \alpha_C - g_{\hat{\gamma}}(W_i, Z_i) \lambda_C$ and $b_{T_i} = Y_i - X_i^\top \beta_T - Z_i \alpha_T - g_{\hat{\gamma}}(W_i, Z_i) \lambda_T$.

3.3 Consistency and asymptotic normality

In this section, it will be shown that the parameter estimates $\hat{\theta}$, as defined in (7), are consistent and asymptotically normal. Theorems 2 and 3 show consistency and asymptotic normality, respectively. The proofs can be found in Section C of the supplementary material. We start by providing some definitions and assumptions that will be useful in stating these theorems. Let

$$h_\ell(S, \gamma^*, \theta^*) = \nabla_\theta \ell(S, \gamma^*, \theta^*), \qquad H_\theta = \mathbb{E}\left[\nabla_\theta h_\ell(S, \gamma^*, \theta^*)\right],$$
$$h_m(W, Z, \gamma^*) = \nabla_\gamma m(W, Z, \gamma^*), \qquad H_\gamma = \mathbb{E}\left[\nabla_\gamma h_\ell(S, \gamma^*, \theta^*)\right],$$

$$M = \mathbb{E}\left[\nabla_\gamma h_m(W, Z, \gamma^*)\right], \qquad \Psi = -M^{-1} h_m(W, Z, \gamma^*),$$
$$\tilde{h}(S, \gamma^*, \theta^*) = \left(h_m(W, Z, \gamma^*)^\top, h_\ell(S, \gamma^*, \theta^*)^\top\right)^\top, \qquad H = \mathbb{E}\left[\nabla_{\gamma,\theta} \tilde{h}(S, \gamma^*, \theta^*)\right].$$

The following assumptions will be used in the proofs of Theorems 2 and 3:

(A7) The parameter space Θ is compact and θ^* belongs to the interior of Θ.

(A8) There exists a function $\mathcal{D}(s)$ integrable with respect to G and a compact neighborhood $\mathcal{N}_\gamma \subseteq \Gamma$ of γ^* such that $|\ell(s, \gamma, \theta)| \leq \mathcal{D}(s)$ for all $\gamma \in \mathcal{N}_\gamma$ and $\theta \in \Theta$.

(A9) $\mathbb{E}\left[\|\tilde{h}(S, \gamma^*, \theta^*)\|^2\right] < \infty$ and $\mathbb{E}\left[\sup_{(\gamma,\theta)\in\mathcal{N}_{\gamma,\theta}} \|\nabla_{\gamma,\theta} \tilde{h}(S, \gamma, \theta)\|\right] < \infty$, with $\mathcal{N}_{\gamma,\theta}$ a neighborhood of (γ^*, θ^*) in $\Gamma \times \Theta$.

(A10) $H^\top H$ is nonsingular.

Note that $\|\cdot\|$ represents the Euclidean norm. Assumption (A8) is necessary to show the consistency and asymptotic normality of the parameter estimates. Sufficient conditions for this assumption are that the support of S is bounded, Γ being compact and Assumption (A7). Assumptions (A7), (A9) and (A10) are regularity conditions that are commonly made in a maximum likelihood context. We have the following consistency theorem.

Theorem 2 *Under Assumptions (A1)–(A8), suppose that $\hat{\theta}$ is a parameter estimate as described in (7), then*

$$\hat{\theta} \xrightarrow{p} \theta^*.$$

The challenge in proving this theorem comes from the fact that we are using a two-step estimation method, meaning that the results from the first step are used in the second step. To ensure consistency of $\hat{\theta}$, in the proofs, we show uniform convergence (in $\theta \in \Theta$) in probability of the empirical likelihood function $\hat{L}(\hat{\gamma}, \theta)$ in (7) to the true likelihood of the model at γ^*. We also have the following asymptotic normality result:

Theorem 3 *Under Assumptions (A1)–(A10), suppose that $\hat{\theta}$ is a parameter estimate as described in (7), then*

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \Sigma_\theta),$$

with

$$\Sigma_\theta = H_\theta^{-1}\, \mathbb{E}\left[\{h_\ell(S, \gamma^*, \theta^*) + H_\gamma \Psi\}\{h_\ell(S, \gamma^*, \theta^*) + H_\gamma \Psi\}^\top\right]\left(H_\theta^{-1}\right)^\top.$$

The difficulty in proving this theorem is related to the fact that the randomness coming from the first step inflates the asymptotic variance of $\hat{\theta}$. Hence, ignoring the first step would lead to inconsistent standard errors and confidence intervals that are not asymptotically valid. To obtain correct standard errors, we treat the two steps as a

joint generalized method of moments (GMM) estimator with their moment conditions stacked in one vector (Newey and McFadden 1994). Indeed, given that $\hat{\gamma}$ and $\hat{\theta}$ are consistent by Theorem 2, they are the unique solutions to the first-order conditions of their respective objective functions in a neighborhood of γ^* and θ^* (with probability going to 1). Therefore, the two-step estimator is asymptotically equivalent to the GMM estimator corresponding to the following moments:

$$\mathbb{E}[h_m(W, Z, \gamma)] = 0 \text{ and } \mathbb{E}[h_\ell(S, \gamma, \theta)] = 0, \tag{8}$$

for the first and second step, respectively. Because of this theoretical equivalence, we could also jointly minimize some norm of the sample version of (8), with respect to (γ, θ), in a single step. However, we opt for the two-step estimation procedure as it is most natural in this context and is less computational complex as the joint minimization. This is because the joint GMM estimator would require solving a system of $3m + 11$ equations of first-order derivatives that would have to be derived analytically or approximated numerically. Moreover, it is important to note that the results from Theorem 2 and 3 also hold for the maximum likelihood estimation of alternative models, as long as they are identified. This implies that Assumption (A1) is a sufficient, but not a necessary condition for both of these theorems to hold. Other possible models could include different parametric copulas for the dependence structure of the error terms (e.g., Frank, Clayton or Joe) and different marginal distributions for each of the error terms (e.g. Gumbel, exponential or logistic distribution). The possible identification of these models is further discussed in the Future research section at the end of the paper. As a last remark, if we were to remove the correction $H_\gamma \Psi$ for the first step, the covariance matrix simplifies to the inverse of Fisher's information matrix (assuming the model is correctly specified).

3.4 Estimation of the asymptotic variance

Using the result from Theorem 3, we can construct a consistent estimator $\hat{\Sigma}_\theta$ for the covariance matrix of the parameters in θ in the following way:

$$\hat{\Sigma}_\theta = \hat{H}_\theta^{-1}\left[n^{-1}\sum_{i=1}^{n}\{h_\ell(S_i, \hat{\gamma}, \hat{\theta}) + \hat{H}_\gamma \hat{\Psi}_i\}\{h_\ell(S_i, \hat{\gamma}, \hat{\theta}) + \hat{H}_\gamma \hat{\Psi}_i\}^\top\right]\left(\hat{H}_\theta^{-1}\right)^\top,$$

where $S_i = \left(Y_i, \Delta_i, \tilde{X}_i, \tilde{W}_i, Z_i\right)$ and

$$h_\ell(S_i, \hat{\gamma}, \hat{\theta}) = \nabla_\theta \ell(S_i, \hat{\gamma}, \hat{\theta}), \qquad h_m(W_i, Z_i, \hat{\gamma}) = \nabla_\gamma m(W_i, Z_i, \hat{\gamma}),$$

$$\hat{H}_\theta = n^{-1}\sum_{i=1}^{n}\nabla_\theta h_\ell(S_i, \hat{\gamma}, \hat{\theta}), \qquad \hat{H}_\gamma = n^{-1}\sum_{i=1}^{n}\nabla_\gamma h_\ell(S_i, \hat{\gamma}, \hat{\theta}),$$

$$\hat{M} = n^{-1}\sum_{i=1}^{n}\nabla_\gamma h_m(W_i, Z_i, \hat{\gamma}), \qquad \hat{\Psi}_i = -\hat{M}^{-1} h_m(W_i, Z_i, \hat{\gamma}).$$

Thanks to the asymptotic normality and the consistent estimator for the variance of the estimators, confidence intervals can easily be constructed. Note that since $\sigma_T, \sigma_C > 0$ and $\rho \in (-1, 1)$, their confidence intervals will be constructed using a logarithm and a Fisher's z-transformation, respectively. These transformations project the estimates on the real line, after which the delta method can be used to obtain their standard errors. The confidence intervals can then be constructed and transformed back to the original scale. This procedure makes sure that our confidence intervals are reasonable (e.g. no negative values for the confidence limits of the standard deviation estimates). Also note that instead of calculating $h_\ell(S_i, \hat{\gamma}, \hat{\theta})$, $\hat{H}_\theta$ and $\hat{H}_\gamma$ using their analytical expressions, they are approximated. This is due to the complexity of these expressions and the amount of them that would have to be derived. For example, $\hat{H}_\theta$ is already a $(2m + 9) \times (2m + 9)$ matrix of derivatives where m is the dimension of $\tilde{X}$. The calculation of these approximations is done by making use of Richardson's extrapolation (Richardson 1911), resulting in more accurate estimates. A general description of the method to approximate the Jacobian matrix can be given as repeated calculations of the central difference approximation of the first derivative with respect to each component of θ, using a successively smaller step size. Richardson's extrapolation uses this information to estimate what happens when the step size goes to zero. A similar description can be given for the approximation of the Hessian matrices. Note that these calculations can be quite time consuming depending on the required level of accuracy.

4 Simulation study

In this section, a simulation study is performed to investigate the finite sample performance of the proposed two-step estimator. Further, we look at the impact of model misspecification. In particular, we investigate what happens when Assumption (A1) does not hold and when the control function is misspecified.

4.1 Comparison of estimators

We consider the four possible combinations of the cases where Z and $\tilde{W}$ are continuous or binary random variables. It is assumed that when Z is binary, it follows a logit model. The proposed estimator is compared to three other estimators: one which does not account for the confounding issue, one which assumes T and C are independent and one which uses the proposed method but treats V as observed. The parameters are estimated for samples of 250, 500 and 1000 observations. The first step of the data generating process (DGP) is as follows:

$$\tilde{X} \sim N(0, 1),$$
$$\begin{pmatrix} \epsilon_T \\ \epsilon_C \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1.1^2 & 0.75 \cdot 1.1 \cdot 1.4 \\ 0.75 \cdot 1.1 \cdot 1.4 & 1.4^2 \end{pmatrix}\right).$$

We have 4 different designs depending on whether Z and $\tilde{W}$ are assumed to be a continuous or binary random variable:

Design 1: $\tilde{W} \sim U[0,2], \quad \nu \sim N(0,2)$	$Z = W^\top \gamma_{cont} + \nu,$
Design 2: $\tilde{W} \sim \text{Bernoulli}(0.5), \quad \nu \sim N(0,2)$	
Design 3: $\tilde{W} \sim U[0,2], \quad \nu \sim \text{Logistic}(0,1)$	$Z = \mathbb{1}(W^\top \gamma_{bin} - \nu > 0),$
Design 4: $\tilde{W} \sim \text{Bernoulli}(0.5), \quad \nu \sim \text{Logistic}(0,1)$	

with $W = (1, \tilde{X}, \tilde{W})^\top$, $\gamma_{cont} = (0.5, -0.4, 1)^\top$ and $\gamma_{bin} = (-1, 0.6, 2.3)^\top$. Further, it is assumed that

$$(\epsilon_T, \epsilon_C) \perp\!\!\!\perp (\tilde{X}, \tilde{W}, \nu), \qquad \tilde{W} \perp\!\!\!\perp (\tilde{X}, \nu) \quad \text{and} \quad \tilde{X} \perp\!\!\!\perp \nu.$$

From this, we have that:

Design 1-2:	$V = Z - W^\top \gamma_{cont}.$
Design 3-4:	$V = (1-Z)\Big[\big(1+\exp\{W^\top\gamma_{bin}\}\big)\log\big(1+\exp\{W^\top\gamma_{bin}\}\big) - W^\top\gamma_{bin}\exp\{W^\top\gamma_{bin}\}\Big]$ $-Z\Big[\big(1+\exp\{-W^\top\gamma_{bin}\}\big)\log\big(1+\exp\{-W^\top\gamma_{bin}\}\big) + W^\top\gamma_{bin}\exp\{-W^\top\gamma_{bin}\}\Big].$

Finally, T and C can be constructed for each design in the following way:

$$\begin{cases} T = \beta_{T,0} + \tilde{X}\beta_{T,1} + Z\alpha_T + V\lambda_T + \epsilon_T \\ C = \beta_{C,0} + \tilde{X}\beta_{C,1} + Z\alpha_C + V\lambda_C + \epsilon_C \end{cases},$$

where $(\beta_{T,0}, \beta_{T,1}, \alpha_T, \lambda_T) = (1.5, 0.6, 0.4, 0.3)$ and $(\beta_{C,0}, \beta_{C,1}, \alpha_C, \lambda_C) = (1.6, 0.4, -0.3, -0.2)$. It follows that $Y = \min\{T, C\}$ and $\Delta = \mathbb{1}(T \leq C)$. It is important to remember that V is not observed and therefore can only be used as a benchmark to compare our estimation results to, which we will introduce in the next paragraph as the oracle estimator.

This data generating process was repeated 1000 times for the four possible designs. The parameter values were chosen such that there is between 55% and 60% censoring for each design. For each sample size, there are four different estimators. The first, which we call the naive estimator, ignores the confounding issue and therefore does not include estimates for λ_T and λ_C. The second, which we call the independent estimator, assumes that T and C are independent from each other (no estimates for ρ as it is assumed to be zero). The third, which we call the oracle estimator, uses the control function approach to handle the confounding issue but treats V as if it were observed. The fourth and last, which we call the two-step estimator, uses the two-step estimation method proposed in this article. This means that V is estimated using $\hat{\gamma}$ from the first step. The estimation is performed in *R* and uses the package *nloptr* to maximize certain functions and the package *numDeriv* for computing the necessary Hessian and Jacobian matrices. The package *MASS* is used to generate the bivariate normal variables.

For each estimator, the bias of each parameter estimate is given together with the empirical standard deviation (ESD) and the root mean squared error (RMSE). Note

that, as the bias decreases, these last 2 statistics should converge to the same value. To better explain how these statistics are calculated, we give the formulas for α_T as an example. Let N represent the total amount of simulations with $j = 1, ..., N$ and $(\hat{\alpha}_T)_j$ the estimate of α_T for the j'th simulation. The ESD and RMSE for α_T are given as follows:

$$\text{ESD} = \sqrt{(N-1)^{-1} \sum_{j=1}^{N} \left[(\hat{\alpha}_T)_j - \bar{\alpha}_T\right]^2}, \quad \text{with } \bar{\alpha}_T = N^{-1} \sum_{j=1}^{N} (\hat{\alpha}_T)_j.$$

$$\text{RMSE} = \sqrt{N^{-1} \sum_{j=1}^{N} \left[(\hat{\alpha}_T)_j - \alpha_T^*\right]^2}, \quad \text{with } \alpha_T^* \text{ the true parameter value.}$$

Lastly, the coverage rate (CR) shows in which percentage of the simulations the true parameter value is included in the estimated 95% confidence interval that follows from Theorem 3 and the estimator $\hat{\Sigma}_\theta$ given in Sect. 3.4.

Table 1 shows the results for design 4, meaning that both Z and W are binary. The naive estimates show a very noticeable bias for almost each parameter, especially α_T. Note that this bias remains the same as the sample size increases. The table also shows that the estimated standard errors are not asymptotically valid as the CR is inconsistent and does not converge to the expected 95%. Moreover, we find similar results when looking at the independent estimator. As with the naive estimator, the bias does not seem to decrease when the sample size increases. Next, the bias for the proposed two-step estimation method is very close to 0 and is clearly an improvement over the naive and independent estimator. It is also very close to that of the oracle estimator, which treats V as observed. This implies that the error from estimating V is negligible compared to the one from the second step. This makes sense, as V is much simpler to estimate. The RMSE decreases when the sample size increases and the ESD and RMSE converge to the same value, which also decreases as the sample size increases. The CR is mostly around 95%, meaning that we have asymptotically valid standard errors and confidence intervals. Nevertheless, the CR seems to behave particularly poorly for α_C. However, when we increased the number of simulations, the CR converged to the nominal level such that this can be attributed to random noise. From these results, it is clear that the two-step estimator performs well, even for small sample sizes. The results for the other designs can be found in Section D of the supplementary material.

4.2 Misspecification of the model

In this subsection, we consider four different types of misspecification. More specifically, we investigate what happens for design 4 (Z and $\tilde{W}$ both binary). For each scenario, we generated 1000 data sets with a sample size of 500 each. The exact results of these simulations can be found in Tables 7–0 from Section D of the supplementary material. To fix ideas, let $\mathbb{P}(\varepsilon_T \leq u, \varepsilon_C \leq v) = \mathcal{C}\left(F_{\varepsilon_T}(u), F_{\varepsilon_C}(v)\right)$, where

Table 1 Estimation results for design 4 with 54% censoring and 1000 simulations.

	$n = 250$				$n = 500$				$n = 1000$			
	Bias	ESD	RMSE	CR	Bias	ESD	RMSE	CR	Bias	ESD	RMSE	CR
Naive estimator												
$\beta_{T,0}$	0.377	0.467	0.600	0.855	0.369	0.343	0.504	0.812	0.373	0.247	0.447	0.659
$\beta_{T,1}$	0.082	0.147	0.168	0.888	0.073	0.101	0.125	0.888	0.073	0.075	0.105	0.829
α_T	−0.672	0.222	0.708	0.128	−0.673	0.161	0.692	0.024	−0.671	0.118	0.682	0.000
$\beta_{C,0}$	−0.222	0.194	0.294	0.731	−0.227	0.133	0.263	0.596	−0.229	0.098	0.249	0.349
$\beta_{C,1}$	−0.043	0.127	0.134	0.928	−0.049	0.092	0.105	0.916	−0.050	0.064	0.081	0.890
α_C	0.444	0.217	0.494	0.421	0.443	0.154	0.469	0.161	0.451	0.111	0.465	0.019
σ_T	0.054	0.094	0.108	0.952	0.036	0.059	0.069	0.954	0.027	0.038	0.046	0.936
σ_C	0.005	0.138	0.138	0.918	0.008	0.098	0.098	0.935	0.011	0.073	0.074	0.934
ρ	−0.120	0.361	0.380	0.907	−0.087	0.263	0.277	0.908	−0.071	0.191	0.204	0.917
Independent estimator												
$\beta_{T,0}$	0.456	0.258	0.524	0.556	0.448	0.186	0.485	0.310	0.455	0.134	0.474	0.086
$\beta_{T,1}$	0.118	0.126	0.173	0.811	0.107	0.082	0.135	0.753	0.107	0.062	0.124	0.577
α_T	0.371	0.460	0.591	0.906	0.373	0.332	0.499	0.811	0.364	0.231	0.431	0.663
λ_T	0.265	0.184	0.322	0.703	0.265	0.131	0.296	0.417	0.261	0.091	0.276	0.142
$\beta_{C,0}$	0.463	0.283	0.542	0.667	0.466	0.205	0.509	0.357	0.454	0.144	0.476	0.097
$\beta_{C,1}$	−0.070	0.135	0.152	0.891	−0.075	0.094	0.120	0.842	−0.076	0.067	0.101	0.760
α_C	−0.266	0.466	0.537	0.912	−0.275	0.335	0.433	0.878	−0.256	0.240	0.351	0.826
λ_C	−0.187	0.178	0.258	0.844	−0.189	0.129	0.229	0.722	−0.184	0.093	0.206	0.482
σ_T	0.095	0.083	0.126	0.770	0.099	0.059	0.116	0.571	0.105	0.042	0.113	0.255
σ_C	0.186	0.091	0.207	0.444	0.196	0.064	0.206	0.104	0.199	0.048	0.205	0.005
Oracle estimator												
$\beta_{T,0}$	0.025	0.272	0.272	0.935	0.006	0.192	0.192	0.939	0.004	0.137	0.136	0.942

Table 1 continued

	$n = 250$				$n = 500$				$n = 1000$			
	Bias	ESD	RMSE	CR	Bias	ESD	RMSE	CR	Bias	ESD	RMSE	CR
$\beta_{T,1}$	0.011	0.110	0.111	0.942	−0.001	0.074	0.074	0.950	−0.003	0.053	0.053	0.942
α_T	−0.007	0.405	0.405	0.908	−0.004	0.285	0.285	0.914	0.001	0.198	0.198	0.942
λ_T	0.000	0.206	0.206	0.911	−0.000	0.143	0.143	0.928	0.000	0.102	0.102	0.930
$\beta_{C,0}$	0.018	0.309	0.309	0.929	0.009	0.226	0.226	0.932	0.004	0.165	0.165	0.921
$\beta_{C,1}$	0.005	0.117	0.117	0.937	−0.000	0.082	0.082	0.933	0.000	0.056	0.056	0.940
α_C	−0.013	0.384	0.385	0.919	−0.013	0.283	0.284	0.907	−0.004	0.207	0.206	0.900
λ_C	−0.007	0.169	0.169	0.925	−0.004	0.126	0.126	0.912	−0.004	0.092	0.092	0.924
σ_T	0.018	0.080	0.082	0.968	0.008	0.050	0.051	0.961	0.004	0.034	0.034	0.962
σ_C	−0.008	0.119	0.120	0.922	−0.005	0.082	0.082	0.935	−0.002	0.061	0.061	0.929
ρ	−0.069	0.296	0.304	0.934	−0.034	0.196	0.199	0.957	−0.018	0.140	0.141	0.959
Two-step estimator												
$\beta_{T,0}$	0.020	0.282	0.282	0.938	0.003	0.197	0.197	0.954	0.003	0.141	0.141	0.944
$\beta_{T,1}$	0.009	0.115	0.115	0.942	−0.002	0.074	0.074	0.963	−0.003	0.055	0.056	0.952
α_T	−0.000	0.428	0.427	0.912	0.002	0.300	0.300	0.909	0.003	0.206	0.206	0.945
λ_T	0.002	0.212	0.212	0.903	0.002	0.148	0.148	0.927	0.001	0.105	0.105	0.938
$\beta_{C,0}$	0.017	0.315	0.315	0.938	0.012	0.229	0.229	0.938	0.004	0.166	0.166	0.932
$\beta_{C,1}$	0.006	0.118	0.119	0.941	0.000	0.083	0.083	0.943	0.000	0.057	0.057	0.943
α_C	−0.013	0.395	0.395	0.921	−0.017	0.289	0.289	0.911	−0.005	0.209	0.209	0.900
λ_C	−0.006	0.171	0.172	0.929	−0.006	0.128	0.128	0.917	−0.004	0.093	0.093	0.926
σ_T	0.019	0.080	0.082	0.964	0.008	0.050	0.051	0.966	0.004	0.034	0.034	0.959
σ_C	−0.008	0.120	0.120	0.918	−0.005	0.082	0.082	0.938	−0.002	0.061	0.061	0.938
ρ	−0.068	0.296	0.303	0.923	−0.034	0.196	0.199	0.955	−0.018	0.140	0.142	0.955

Given are the bias, the empirical standard deviation (ESD), the root mean squared error (RMSE) and the confidence rate (CR)

F_{ε_T} and F_{ε_C} are the cumulative distribution functions of ε_T and ε_C, respectively, and $\mathcal{C}$ a parametric copula.

Scenario 1 In this first scenario, let ε_T and ε_C be Gumbel distributed random variables (instead of them being normally distributed). The location and scale parameters are chosen such that both random variables have a mean of 0 and a standard deviation of 1.1 and 1.4, respectively. Moreover, $\mathcal{C}$ is still a Gaussian copula with correlation parameter $\rho = 0.75$.

Scenario 2 Next, let ε_T and ε_C be normally distributed with location and scale parameters as before. However, $\mathcal{C}$ is now a Frank copula. The dependence parameter for the Frank copula is chosen such that it is equal to the same Kendall's tau as a correlation parameter $\rho = 0.75$ ($\tau \approx 0.54$) for the Gaussian copula.

Scenario 3 Thirdly, we look at a misspecification of the control function V. More precisely, instead of $\nu \sim \text{Logistic}(0, 1)$, we let $\nu \sim N(0, 1)$.

Scenario 4 The final scenario looks at the case where V is identically equal to zero, such that there is no unmeasured confounding.

Overall, misspecifying the model seems to have little impact on the bias of the main parameter of interest α_T, which remains small. However, it is to be noticed that the ESD of α_T more than doubles in both Scenario 1 and 2. Moreover, the coverage rates behave poorly for all scenarios, but this is to be expected when misspecifying the model. For most of the other parameters, the first and second scenario give rise to a lot of bias. However, both the third and fourth scenario (where the control function is misspecified) seem to have little influence on the proposed estimator for all parameters except λ_T and λ_C. From these results, we can conclude that for estimating the main parameter of interest, α_T, our proposed method seems to be fairly robust to these types of misspecification.

5 Data application

In this section, we apply the outlined methodology to estimate the effect of Job Training Partnership Act (JTPA) services on time until employment. The data come from a large-scale randomized experiment known as the National JTPA Study and have been analyzed extensively by Bloom et al (1997), Abadie et al (2002) and Frandsen (2015) among others. The data and problem investigated is the same as in Frandsen (2015), but the method used nevertheless differs as we allow for dependent censoring. Later in this section, we give our reasoning as to why there could be dependent censoring present in the data.

This study was performed to evaluate the effectiveness of more than 600 federally funded services, established by the Job Training Partnership Act of 1982, that were intended to increase the employability of eligible adults and out-of-school youths. These services included classroom training, on-the-job training and job search assistance. The JTPA started to fund these programs in October of 1983 and continued funding up until the late 1990's. Between 1987 and 1989, a little over 20,000 adults and out-of-school youths who applied for JTPA were randomly assigned to be in either a treatment or a control group. Treatment group members were eligible to receive JTPA services, while control group members were not eligible for 18 months. However, due

to local program staff not always following the randomization rules closely, about 3% of the control group members were able to participate in JTPA services. It is important to note that we are not comparing JTPA services to no services but rather JTPA services versus no and other services, since control group members were still eligible for non-JTPA training. Between 12 and 36 months after randomization, with an average of 21 months, the participants were surveyed by data collection officers. Next, a subset of 5468 subjects participated in a second follow-up survey, which focused on the period between the two surveys. The second survey took place between 23 and 48 months after randomization. See Figure 1 in Section D of the supplementary information for a graphical representation of the interview process.

In this application, we will focus our attention on the effect of JTPA programs on the sample of 1298 fathers who reported having no job at the time of randomization, for which participation data is available. The outcome of interest is the time between randomization and employment. For the individuals that were only invited to the first interview, the outcome is measured completely if an individual is employed by the time of the survey and censored at the time of the interview otherwise. For the fathers that were invited to the second follow-up interview and participated, the outcome is measured completely if an individual is employed by the time of the second follow-up interview, but is otherwise censored at the second interview date. If the individual does not participate in the second survey after being invited, they will be censored at the time of the first interview. It follows that there could be some dependence between T and C when this decision to go to the second follow-up interview is influenced by them having found a job between the two interview dates. This possible dependence combined with the fact that the data suffer from two-sided noncompliance makes it an appropriate application of the proposed methodology.

The instrument $\tilde{W}$ will be a binary variable indicating whether an individual is in the control or treatment group (0 and 1, respectively). The confounded variable Z indicates whether they actually participated in a JTPA program (0 for no participation and 1 otherwise). This participation variable is confounded due to individuals moving themselves between the treatment and control group in a non-random way. The covariates include the participant's age, race (white or non-white), marital status and whether they have a high school diploma or GED. We expect $\tilde{W}$ to be a valid instrument because it is randomly assigned, correlated with JTPA participation and should have no impact on time until employment other than through participation in a JTPA funded program. Rows 2 through 5 from Table 3 in Section D of the supplementary information show that the individual characteristics are balanced across the control and treatment group. This indicates satisfactory random assignment. The first row shows that about 31% of the total sample was assigned to the control group. The last 3 rows show summary statistics for variables observed after randomization. It is interesting to note that 13% of the fathers in the control group were nevertheless able to participate in JTPA services compared to 3% for the entire control group. The mean time to employment also seems to be about 30 days shorter for the individuals assigned to the treatment group compared to the control group. The censoring rate is similar for both groups. Figure 2 in Section D of the supplementary information plots a histogram of the observed follow-up time Y, where darker shading indicates a higher censoring rate. A lot of the censored observations are around the 600 days mark, at

which time most of the first follow-up interviews took place. Since everyone in the sample participated in the first follow-up interview, an observation before the date of the first follow-up survey cannot be censored.

The results of applying the two-step estimator (using a logit model for Z), compared to other estimators, can be found in Table 2. The naive estimator, which does not treat Z as a confounded variable, seems to underestimate the effect of JTPA services on time until employment compared to the proposed two-step estimator. At a 5% significance level, both of these estimators find a significant effect of JTPA training reducing time until employment. However, the two-step estimate is almost twice the naive estimate which indicates that the individuals participating in the treatment are those with a lower ability to find employment. The independent estimator, which assumes independent censoring, seems to only slightly overestimate α_T compared to the proposed two-step estimator. However, looking at the p-values, we notice that the effect estimated by the independent estimator is not significantly different from zero at a 5% significance level. On the contrary, the estimated effect is deemed to be significant for the proposed two-step estimator. Age seems to be (borderline) significant across the estimators as does marriage status and having a high school diploma or GED. Being older seems to increase time until employment, while being married and having a diploma reduces it. Both the naive and the two-step estimator seem to agree that there is a quite strong negative correlation of about -0.43 between T and C.

Future research

It is important to note that this work is only a first step toward a set of models that will allow for the estimation of causal effects under dependent censoring and unobserved confounding. A first extension could be to select different parametric marginals and copulas by making use of an information criterion. Up until now, the association parameter has been shown to be identified only for certain combinations of parametric copulas and marginals without confounding or covariates (see Czado and Van Keilegom (2023)). Implementing this would therefore complicate the identifiability proof. However, it is to be noted that the consistency and asymptotic normality results would still hold for the alternative maximum likelihood estimator, provided that the model is identified. Another line of research is to allow for semi-parametric marginals (see Deresa and Van Keilegom (2020c) and Deresa and Van Keilegom (2023) for examples without unmeasured confounding). These lines of research would greatly increase the flexibility of the model and are currently being investigated. Lastly, a goodness-of-fit test could be developed. Under the null hypothesis, we would have that the distribution function of Y is equal to the marginal version of (5) for some (γ, θ). An Anderson-Darling type test statistic could be developed, where we look at the distance between the empirical distribution function (EDF) of Y and the proposed parametric estimate of the distribution of Y. Note that we use the variable Y, as the EDF cannot be computed for T or C (only one of them is observed). Large values of this test statistic would therefore indicate a possible misspecification. The distribution of this test statistic

Table 2 Estimation results for the naive, independent and two-step estimator. Given are the parameter estimate, standard error (SE) and the p-value

	Naive estimator			Independent estimator			Two-step estimator		
	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Survival time (T)									
Intercept	4.753	0.223	0.000	4.949	0.318	0.000	4.866	0.370	0.000
Age	0.015	0.006	0.007	0.013	0.006	0.025	0.015	0.008	0.056
White	−0.197	0.108	0.068	−0.197	0.118	0.096	−0.187	0.622	0.764
Married	−0.331	0.123	0.007	−0.330	0.133	0.013	−0.331	0.132	0.012
GED	−0.166	0.102	0.104	−0.179	0.112	0.109	−0.172	0.217	0.430
α_T	−0.218	0.102	0.033	−0.483	0.260	0.063	−0.428	0.210	0.041
λ_T				−0.113	0.116	0.330	−0.097	0.087	0.268
Censoring time (C)									
Intercept	6.866	0.134	0.000	6.672	0.164	0.000	6.848	0.403	0.000
Age	−0.001	0.002	0.436	−0.001	0.002	0.605	−0.001	0.003	0.605
White	0.012	0.033	0.713	−0.005	0.050	0.913	0.010	1.206	0.993
Married	0.010	0.032	0.746	−0.006	0.075	0.933	0.013	0.661	0.984
GED	−0.063	0.042	0.137	−0.069	0.041	0.093	−0.062	0.177	0.728
α_C	−0.062	0.042	0.138	−0.052	0.117	0.655	−0.028	1.335	0.983
λ_C				0.006	0.042	0.876	0.015	0.511	0.976
σ_T	1.817	0.040	0.000	1.804	0.038	0.000	1.816	0.038	0.000
σ_C	0.323	0.037	0.000	0.285	0.013	0.000	0.323	0.029	0.000
ρ	−0.430	0.196	0.028				−0.432	0.176	0.014

could be approximated by using parametric bootstrap or one could try to obtain an expression for the limiting distribution of the test statistic.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-023-00903-9.

Acknowledgements Financial support from the European Research Council (2016-2022, Horizon 2020 / ERC grant agreement No. 694409) is gratefully acknowledged. The authors are grateful for the valuable suggestions and feedback from the reviewers. Moreover, the authors would like to thank Sara Rutten and Ilias Willems for their useful comments and remarks.

Declarations

Conflict of interest The authors have declared no conflict of interest

References

Abadie A, Angrist J, Imbens G (2002) Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70(1):91–117
Aldrich JH, Nelson FD (1991) Linear probability, logit, and probit models, 10th edn. Quantitative applications in the social sciences 45, Sage, Newbury Park
Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. J Am Stat Assoc 91(434):444–455
Beyhum J, Florens JP, Van Keilegom I (2023) A nonparametric instrumental approach to confounding in competing risks models. Lifetime Data Anal 1–26
Beyhum J, Florens JP, Van Keilegom I (2022) Nonparametric instrumental regression with right censored duration outcomes. J Business Econ Stat 40(3):1034–1045
Beyhum J, Tedesco L, Van Keilegom I (2023) Instrumental variable quantile regression under random right censoring. Economet J utad015
Bijwaard GE, Ridder G (2005) Correcting for selective compliance in a re-employment bonus experiment. J Econ 125(1):77–111
Blanco G, Chen X, Flores CA et al (2020) Bounds on average and quantile treatment effects on duration outcomes under censoring, selection, and noncompliance. J Business Econ Stat 38(4):901–920
Bloom HS, Orr LL, Bell SH et al (1997) The benefits and costs of jtpa title ii-a programs: key findings from the national job training partnership act study. J Hum Resour 32(3):549–576
Braekers R, Veraverbeke N (2005) A copula-graphic estimator for the conditional survival function under dependent censoring. Can J Stat 33(3):429–447
Centorrino S, Florens JP (2021) Nonparametric estimation of accelerated failure-time models with unobservable confounders and random censoring. Electron J Stat 15(2):5333–5379
Chernozhukov V, Fernández-Val I, Kowalski AE (2015) Quantile regression with censoring and endogeneity. J Econ 186(1):201–221
Czado C, Van Keilegom I (2023) Dependent censoring based on parametric copulas. Biometrika 110(3):721–738
Deresa NW, Van Keilegom I (2020) Flexible parametric model for survival data subject to dependent censoring. Biom J 62(1):136–156
Deresa NW, Van Keilegom I (2020) A multivariate normal regression model for survival data subject to different types of dependent censoring. Comput Stat Data Anal 144(106):879
Deresa NW, Van Keilegom I (2020) On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. Biometrika 108(4):965–979
Deresa NW, Van Keilegom I (2023) Copula based Cox proportional hazards models for dependent censoring. J Am Stat Assoc (just-accepted):1–23
Emura T, Chen YH (2018) Analysis of survival data with dependent censoring: copula-Based Approaches. Springer

Escanciano JC, Jacho-Chávez D, Lewbel A (2016) Identification and estimation of semiparametric two-step models. Quant Econ 7(2):561–589
Frandsen BR (2015) Treatment effects with censoring and endogeneity. J Am Stat Assoc 110(512):1745–1752
Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47(1):153–161
Huang X, Zhang N (2008) Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. Biometrics 64(4):1090–1099
Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282):457–481
Khan S, Tamer E (2009) Inference on endogenously censored regression models using conditional moment inequalities. J Econ 152(2):104–119
Lee S (2007) Endogeneity in quantile regression models: a control function approach. J Econ 141(2):1131–1158
Li J, Fine J, Brookhart A (2015) Instrumental variable additive hazards models. Biometrics 71(1):122–130
Manski CF (1988) Identification of binary response models. J Am Stat Assoc 83(403):729–738
Martinussen T, Vansteelandt S (2020) Instrumental variables estimation with competing risk data. Biostatistics 21(1):158–171
Navarro S (2010) Control functions. Palgrave Macmillan UK, London, pp 20–28
Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. Handb Econ 4:2111–2245
Oxley L, McAleer M (1993) Econometric issues in macroeconomic models with generated regressors. J Econ Surv 7(1):1–40
Pagan A (1984) Econometric issues in the analysis of regressions with generated regressors. Int Econ Rev 25(1):221–247
Richardson A, Hudgens MG, Fine JP et al (2017) Nonparametric binary instrumental variable analysis of competing risks data. Biostatistics 18(1):48–61
Richardson LF (1911) Ix. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. Philos Transact Royal Soc London Ser A, Contain Papers Math Phys Character 210(459–470):307–357
Rivest LP, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. J Multivar Anal 79(1):138–155
Robins JM, Finkelstein DM (2000) Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. Biometrics 56(3):779–788
Sant'Anna PHC (2016) Program evaluation with right-censored data. arXiv preprint arXiv:1604.02642
Sperlich S (2009) A note on non-parametric estimation with predicted variables. Economet J 12(2):382–395
Staplin N, Kimber A, Collett D et al (2015) Dependent censoring in piecewise exponential survival models. Stat Methods Med Res 24(3):325–341
Sujica A, Van Keilegom I (2018) The copula-graphic estimator in censored nonparametric location-scale regression models. Econom Stat 7:89–114
Tchetgen Tchetgen EJ, Walter S, Vansteelandt S et al (2015) Instrumental variable estimation in a survival context. Epidemiology 26(3):402–410
Tsiatis A (1975) A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci–PNAS 72(1):20–22
Wooldridge JM (2010) Econometric Analysis of Cross Section and Panel Data. MIT press
Wooldridge JM (2015) Control function methods in applied econometrics. J Hum Resour 50(2):420–445
Zheng C, Dai R, Hari PN et al (2017) Instrumental variable with competing risk model. Stat Med 36(8):1240–1255
Zheng M, Klein JP (1995) Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika 82(1):127–138