On two-stage estimation of structural instrumental variable models

By BYEONG YEOP CHOI

*Department of Epidemiology and Biostatistics, University of Texas Health Science Center,
7703 Floyd Curl Drive, San Antonio, Texas 78229, U.S.A.*

choib@uthscsa.edu

JASON P. FINE

*Department of Biostatistics, University of North Carolina, 3103B McGavran-Greenberg Hall,
Chapel Hill, North Carolina 27599, U.S.A.*

jfine@email.unc.edu

AND M. ALAN BROOKHART

*Department of Epidemiology, University of North Carolina, 2105F McGavran-Greenberg Hall,
Chapel Hill, North Carolina 27599, U.S.A.*

mabrook@email.unc.edu

SUMMARY

Two-stage least squares estimation is popular for structural equation models with unmeasured confounders. In such models, both the outcome and the exposure are assumed to follow linear models conditional on the measured confounders and instrumental variable, which is related to the outcome only via its relation with the exposure. We consider data where both the outcome and the exposure may be incompletely observed, with particular attention to the case where both are censored event times. A general class of two-stage minimum distance estimators is proposed that separately fits linear models for the outcome and exposure and then uses a minimum distance criterion based on the reduced-form model for the outcome to estimate the regression parameters of interest. An optimal minimum distance estimator is identified which may be superior to the usual two-stage least squares estimator with fully observed data. Simulation studies demonstrate that the proposed methods perform well with realistic sample sizes. Their practical utility is illustrated in a study of the comparative effectiveness of colon cancer treatments, where the effect of chemotherapy on censored survival times may be confounded with patient status.

Some key words: Censored data; Endogeneity; Instrumental variable; Resampling; Unmeasured confounder.

1. INTRODUCTION

Confounding is a leading cause of bias in observational studies. Regression adjustment or propensity score methods may be used to overcome confounding, but they require that all confounders be observed. Using econometric terminology, the exposure variable is said to be endogenous when the exposure is correlated with an error term due to sharing unmeasured

confounders. Endogeneity may also occur in randomized trials when there is noncompliance, which may be related to unobserved variables that are risk factors for the outcome. In such cases, the usual regression estimators may not be consistent.

An instrumental variable methodology yields an unbiased estimator of the effect of an endogenous exposure. Although the requirements of an instrumental variable depend on the particular analytical method, the following three conditions are commonly used (Brookhart et al., 2010): (i) an instrumental variable V has a causal effect on an exposure Z ; (ii) V affects an outcome Y only through the exposure Z ; (iii) V is unrelated to measured and unmeasured confounders W and U . In randomized trials, a randomization assignment indicator is often used as an instrumental variable (Robins & Tsiatis, 1991; Loeys & Goetghebeur, 2003; Loeys et al., 2005; Nie et al., 2011) to estimate causal treatment effects.

Structural linear equation modelling is popular for estimating causal effects in the instrumental variable setting and provides the foundation for our instrumental variable method. The outcome model of interest relates Y to Z and W via a linear model, and the exposure model relating Z to V and W is also specified via a linear model. The regression parameters in the outcome model are identified by the instrumental variable. In the absence of measured confounders, an instrumental variable estimator may be obtained as the ratio of two covariance estimators. In the case where there are confounders, the generalized method of moments (Hansen, 1982) or two-stage least squares estimation (Anderson & Rubin, 1950) may be used with complete data. Our approach builds upon these earlier methods to accommodate missing data set-ups.

Considerable work has been done in settings where both the outcome and the exposure are fully observed and assumed to satisfy either semiparametric linear or nonlinear or nonparametric structural equation models with unspecified error distributions (Anderson & Rubin, 1950; Amemiya, 1974; Newey, 1990; Newey & Powell, 2003). The popular two-stage least squares estimator has an explicit form, with a well-characterized sampling distribution and plug-in variance estimation, making inference straightforward (Bollen, 1996). However, if either outcome or exposure is incompletely observed, such methods are not applicable. There has been limited work addressing two-stage instrumental variable estimation in such settings.

While the two-stage methods we propose are generally applicable to cases with incomplete data, in this article our focus is on right-censored data. There have been attempts to extend classical two-stage instrumental variable estimators. Robins & Tsiatis (1991) developed instrumental variable methods for correcting noncompliance in randomized trials by using a rank-preserving structural failure time model. This model is an alternative to the usual two-stage model, which is tailored to failure time data. A limitation is that the semiparametric estimation methods require that censoring times always be known, as with fixed follow-up, and hence censoring due to drop-out is not permitted. Brännäs (2000) considered ad hoc two-stage instrumental variable estimators for linear structural equation models, which are adaptations of a symmetric trimmed least squares estimator (Powell, 1986) and a Buckley & James (1979) estimator for right-censored data. However, the theoretical properties of these procedures were not investigated and a rigorous investigation of two-stage instrumental variable estimation in linear models with right censoring does not seem to be available in the literature. Loeys & Goetghebeur (2003) proposed instrumental variable methods for the effect of treatment received in a randomized trial with all-or-nothing compliance based on a proportional hazards model. These methods were extended to permit other covariates in the proportional hazards model (Loeys et al., 2005). Li et al. (2015) and Tchetgen Tchetgen et al. (2015) developed two-stage least squares estimators for an additive hazard model. Nie et al. (2011) proposed an efficient instrumental variable estimator for comparing survival probabilities in randomized trials with noncompliance and administrative censoring, which is an extension of the estimator of Baker (1998).

In § 2, we propose a general framework for two-stage instrumental variable estimation of semiparametric linear structural equation models for outcome and exposure which accommodates incomplete data. The main requirement is that there exist semiparametric methods for fitting linear models to the outcome and exposure. Such methods are well developed for truncated and censored time-to-event outcomes. At stage 1, we fit the exposure model and a reduced form of the outcome model conditional on the instrumental variable. At stage 2, we estimate the regression parameters in the true outcome model using a weighted minimum distance method based on the stage 1 results. This yields a closed-form estimator, for which a particular choice of weight leads to the standard two-stage least squares estimator with fully observed data. For the case of right censoring, the procedure does not require that the censoring time be always observed. We prove that our estimators are consistent and asymptotically normal, and provide a theoretically justified resampling technique for making inferences. The optimal weight is characterized, resulting in a minimum variance estimator which may be superior to the usual two-stage estimator.

In § 3, we discuss details related to the implementation of our semiparametric estimator when either outcome or exposure may be censored, employing existing estimators for accelerated failure time models with right-censored event times. These methods perform well in simulations reported in § 4, where naïve estimation that ignores the unmeasured confounders may produce severely biased estimates of exposure effects. The practical utility of the proposed methods is illustrated by a study of the comparative effectiveness of colon cancer treatments.

2. A GENERAL TWO-STAGE ESTIMATION PROCEDURE

2.1. Model and estimation

For $i = 1, \dots, n$, suppose that Y_i is an outcome, Z_i is an exposure variable, $V_i = (V_{i1}, \dots, V_{ip})^T$ is a $p \times 1$ vector of instrumental variables, $W_i = (W_{i1}, \dots, W_{iq})^T$ is a $q \times 1$ vector of measured confounders, and U_i is an unmeasured confounder.

We consider the linear outcome model

$$Y_i = \alpha_{Y0} + \alpha_0^T X_i + \varepsilon_i, \quad (1)$$

where $\alpha_0 = (\alpha_{YZ}, \alpha_{YW}^T)^T$, $\alpha_{YW} = (\alpha_{YW,1}, \dots, \alpha_{YW,q})^T$ and $X_i = (Z_i, W_i^T)^T$. The unmeasured confounder is captured by ε_i , which is a weighted sum of U_i and a pure error term ε_i^* , i.e., $\varepsilon_i = \alpha_{YU} U_i + \varepsilon_i^*$, with $E(\varepsilon_i^* | X_i, U_i) = 0$. The linear exposure model is

$$Z_i = \beta_{Z0} + \beta_0^T D_i + \delta_i, \quad (2)$$

where $\beta_0 = (\beta_{ZV}^T, \beta_{ZW}^T)^T$, $\beta_{ZV} = (\beta_{ZV,1}, \dots, \beta_{ZV,p})^T$, $\beta_{ZW} = (\beta_{ZW,1}, \dots, \beta_{ZW,q})^T$, $D_i = (V_i^T, W_i^T)^T$ and $\delta_i = \beta_{ZU} U_i + \delta_i^*$ with $E(\delta_i^* | D_i, U_i) = 0$. Models (1) and (2) do not permit interactions between observed covariates and the unmeasured confounder.

The implied model for X_i is

$$\begin{aligned} X_i &= \begin{pmatrix} \beta_{Z0} \\ 0_{q \times 1} \end{pmatrix} + \begin{pmatrix} \beta_{ZV}^T & \beta_{ZW}^T \\ 0_{q \times p} & I_q \end{pmatrix} \begin{pmatrix} V_i \\ W_i \end{pmatrix} + \begin{pmatrix} \delta_i \\ 0_{q \times p} \end{pmatrix} \\ &= \beta_{Z0}^l + B_0^T D_i + \delta_i^l, \end{aligned} \quad (3)$$

where B_0 is a $(p+q) \times (1+q)$ parameter matrix, $0_{q \times p}$ is the $q \times p$ zero matrix and I_q is the q -dimensional identity matrix. Substituting (3) into (1) gives the reduced form of the outcome

model conditional on the instrumental variable and measured confounders but not the exposure:

$$Y_i = \gamma_{Y0} + \gamma_0^T D_i + \tau_i, \quad (4)$$

where $\gamma_{Y0} = \alpha_{Y0} + \alpha_{YZ}\beta_{Z0}$, $\gamma_0 = (\gamma_{YV}^T, \gamma_{YW}^T)^T = B_0\alpha_0$ is a $(p+q) \times 1$ parameter vector and $\tau_i = \varepsilon_i + \alpha_{YZ}\delta_i$.

The key to our two-stage estimation procedure is that the reduced-form model (4) is not subject to unmeasured confounding. This assumption is closely connected to the usual exclusion restriction for instrumental variable estimation (Angrist et al., 1996), which implies that $E(\varepsilon_i^* | Z_i, V_i, W_i, U_i) = 0$ and hence $E(\varepsilon_i^* | D_i) = 0$. Clearly, $E(\delta_i^* | D_i) = 0$. Since ε_i and δ_i are sums of certain functions of U_i and the pure error terms, the condition $E(U_i | D_i) = 0$ is necessary for $E(\tau_i | D_i) = E(\delta_i | D_i) = 0$.

A two-stage estimator will be developed under the assumption that conditional on D_i , (τ_i, δ_i) are independent and identically distributed with mean zero and covariance matrix Σ_e . A sufficient condition for this mean independence assumption on (τ_i, δ_i) is that $E(U_i | D_i) = 0$, which we call the instrumental variable independence condition. This implies that the mean of the unmeasured confounder is the same across the categories generated by the instrumental variable and the measured confounders.

Naïve estimation of model (1) may be inconsistent because X_i and ε_i are correlated through U_i , so $E(\varepsilon_i | X_i)$ is not equal to zero in general unless $E(U_i | Z_i, W_i^T) = 0$. However, since $E(\tau_i | D_i) = 0$ in model (4), γ_0 can be consistently estimated using $\{(Y_1, D_1^T), \dots, (Y_n, D_n^T)\}$. The proposed instrumental variable estimators require the assumptions below that consistent and asymptotically normal estimators of $\theta_0 = (\gamma_0^T, \beta_0^T)^T$ exist.

Assumption 1. The estimator $\hat{\theta} = (\hat{\gamma}^T, \hat{\beta}^T)^T$ converges in probability to $\theta_0 = (\gamma_0^T, \beta_0^T)^T$ as $n \rightarrow \infty$.

Assumption 2. The random quantity $n^{1/2}(\hat{\theta} - \theta_0)$ has a limiting normal distribution with mean zero and covariance matrix Σ_{θ_0} .

Consistent and asymptotically normal estimators may be obtained by a least squares method for fully observed data or by a Buckley–James method (Buckley & James, 1979) and a rank-based method (Prentice, 1978) for right-censored data. The asymptotic properties of the Buckley–James and rank estimators were studied by Tsiatis (1990), Lai & Ying (1991), Ying (1993) and Jin et al. (2006b). Other incomplete data settings may also be of interest, as discussed in § 6.

Given consistent estimators $\hat{\gamma}$ and $\hat{B}$, a consistent estimator for α_0 can be obtained by minimizing a weighted quadratic distance criterion

$$(\hat{\gamma} - \hat{B}\alpha_0)^T A_n (\hat{\gamma} - \hat{B}\alpha_0),$$

where A_n is a nonnegative-definite symmetric weight matrix which may be data-dependent, with $A_n = A + o_p(1)$. The minimum distance estimator is

$$\hat{\alpha} = (\hat{B}^T A_n \hat{B})^{-1} \hat{B}^T A_n \hat{\gamma}.$$

For complete data, the two-stage least squares estimator is obtained by replacing the exposure with its predicted value calculated from the least squares fit of the exposure model. Let X_c and D_c

denote matrices with i th rows X_{ci}^T and D_{ci}^T , where $X_{ci} = X_i - \bar{X}$, $D_{ci} = D_i - \bar{D}$, $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $\bar{D} = n^{-1} \sum_{i=1}^n D_i$. The two-stage least squares estimator can be written as

$$\hat{\alpha}_{2\text{sls}} = (\hat{X}_c^T \hat{X}_c)^{-1} \hat{X}_c^T Y,$$

where $Y = (Y_1, \dots, Y_n)^T$ and $\hat{X}_c = D_c \hat{B}$. It follows that

$$\hat{\alpha}_{2\text{sls}} = \{\hat{B}^T (D_c^T D_c) \hat{B}\}^{-1} \hat{B}^T D_c^T Y,$$

which is equivalent to $\hat{\alpha}$ with $A_n = D_c^T D_c$ and $\hat{\gamma} = (D_c^T D_c)^{-1} D_c^T Y$.

Next, we present the major theoretical results for our general two-stage estimator.

THEOREM 1. *Under Assumption 1, $\hat{\alpha}$ converges in probability to α_0 as $n \rightarrow \infty$.*

Proof. It follows from the continuous mapping theorem that

$$\hat{\alpha} = \{\hat{B}^T (A_n/n) \hat{B}\}^{-1} \hat{B}^T (A_n/n) \hat{\gamma} = (B_0^T A B_0)^{-1} B_0^T A \gamma_0 + o_p(1) = \alpha_0 + o_p(1),$$

which establishes the assertion. $\square$

THEOREM 2. *Under Assumptions 1 and 2, $n^{1/2}(\hat{\alpha} - \alpha_0)$ has a normal limiting distribution with mean zero and covariance matrix $(B_0^T A B_0)^{-1} B_0^T A \Omega(\alpha_0) A B_0 (B_0^T A B_0)^{-1}$, where $\Omega(\alpha_0) = \text{var}\{n^{1/2}(\hat{\gamma} - \hat{B}\alpha_0)\}$.*

Proof. Note that $(\hat{B}^T A_n \hat{B})^{-1} \hat{B}^T A_n = (B_0^T A B_0)^{-1} B_0^T A + o_p(1)$ and that $n^{1/2}(\hat{\gamma} - \hat{B}\alpha_0)$ converges to a zero-mean multivariate normal distribution with covariance matrix $\Omega(\alpha_0)$. Therefore, from a multivariate Slutsky's theorem, with

$$\begin{aligned} n^{1/2}(\hat{\alpha} - \alpha_0) &= n^{1/2}\{(\hat{B}^T A_n \hat{B})^{-1} \hat{B}^T A_n \hat{\gamma} - (\hat{B}^T A_n \hat{B})^{-1} \hat{B}^T A_n \hat{B}\alpha_0\} \\ &= (\hat{B}^T A_n \hat{B})^{-1} \hat{B}^T A_n n^{1/2}(\hat{\gamma} - \hat{B}\alpha_0), \end{aligned}$$

the result follows easily. $\square$

Although Theorems 1 and 2 may appear straightforward, their generality is useful in converting the problem of finding consistent and asymptotically normal instrumental variable estimators to that of finding well-established estimators for the exposure and reduced outcome linear models. While our focus here is on time-to-event applications with censored outcome or exposure, these theorems are broadly applicable. They accommodate, for example, more complicated time-to-event observation schemes such as left truncation and interval censoring of the outcome as well as missing confounders in either the exposure or the outcome model. In § 3 we present corollaries which give the asymptotic properties of the proposed instrumental variable estimators for right-censored data, based on the properties of existing estimation procedures for right-censored data. Additional corollaries could be established on a case-by-case basis for other missing-data scenarios where estimation methodology is available.

Following well-known results for generalized method of moments estimators, a lower bound on the covariance matrix of $n^{1/2}(\hat{\alpha} - \alpha_0)$ is $\{B_0^T \Omega(\alpha_0)^{-1} B_0\}^{-1}$. This is obtained by taking $A = \Omega(\alpha_0)^{-1}$. The corresponding $\hat{\alpha}$ is obtained by using the weight $A_n = \hat{\Omega}(\hat{\alpha})^{-1}$, which

is a consistent estimator for $\Omega(\alpha_0)^{-1}$ if $\hat{\alpha}$ is consistent for α_0 . In order to compute $A_n = \hat{\Omega}(\hat{\alpha})^{-1}$, we need an initial estimator that is consistent for α_0 . In practice, we could use the initial estimator $\hat{\alpha}_I = (\hat{B}^\top \hat{B})^{-1} \hat{B}^\top \hat{\gamma}$ with an identity weight matrix $A_n = I$. We remark that this estimator is optimal only within the proposed class of estimators and may not be fully efficient.

One can write

$$\Sigma_{\theta_0} = \begin{pmatrix} \Sigma_{\gamma_0} & \Sigma_{\gamma_0, \beta_0} \\ \Sigma_{\beta_0, \gamma_0} & \Sigma_{\beta_0} \end{pmatrix}.$$

Then $\Omega(\alpha_0) = \Sigma_{\gamma_0} - \alpha_{YZ}(\Sigma_{\gamma_0, \beta_0} + \Sigma_{\beta_0, \gamma_0}) + \alpha_{YZ}^2 \Sigma_{\beta_0}$. For the two-stage least squares estimator with complete data, $\Omega(\alpha_0) = \text{var}(\varepsilon_i - \alpha_{YZ} \delta_i) \bar{M}^{-1}$ where $\bar{M} = \lim_{n \rightarrow \infty} D_c^\top D_c / n$.

If the instrumental variable is univariate ($p = 1$) and B_0 is nonsingular, then $\alpha_0 = B_0^{-1} \gamma_0$ and $\hat{\alpha}$ does not depend on A_n . The covariance matrix of $n^{1/2}(\hat{\alpha} - \alpha_0)$ with univariate instrumental variable is $\{B_0^\top \Omega(\alpha_0)^{-1} B_0\}^{-1}$, which matches the lower bound in the general case. If there are no confounders, then $\hat{\alpha}$ reduces to $\hat{\gamma}_{yv} / \hat{\beta}_{zv}$, which is the standard instrumental variable estimator (Angrist et al., 1996).

2.2. A resampling method for variance estimation

Variance estimation for $\hat{\alpha}$ is of practical importance. Computational difficulties may arise if the estimation procedures for the exposure and reduced-form models do not yield simple closed-form estimators for Σ_{θ_0} . For example, with right-censored data, complicated nonparametric function estimation may be needed if the estimating equation for the regression parameter is not smooth enough. This occurs with rank-based estimators for the accelerated lifetime model, where the variance involves the derivative of the hazard function of the error term in the linear model. To avoid such difficulties, resampling methods may be used which require only that the estimating equations for stage 1 estimation be consistent and asymptotically normal. These methods are particularly useful when the stage 1 estimators can easily be computed, as is the case for right-censored accelerated failure time models.

We propose a general resampling scheme for $\hat{\alpha}$ which adapts the work of Jin et al. (2001). The main idea is to repeatedly perturb and optimize the objective function used for estimation. If the objective function has a first derivative, then this approach is equivalent to perturbing the corresponding estimating equation. In Jin et al. (2001), the empirical variance of the bootstrap estimators is consistent for the true variance if both the estimating equation and the perturbed estimating equation have good quadratic approximations around the true parameter values. This assumption holds quite generally under mild regularity conditions, e.g., complete data estimation under the L_p -norm and rank regression. The approach has been carefully studied in the context of the accelerated lifetime model with right-censored data, including rank estimation (Jin et al., 2001, 2006a), Buckley–James estimation (Jin et al., 2006b), and local Buckley–James estimation (Pang et al., 2015). Details for the right-censored instrumental variable estimators are given in the next section.

3. INFERENCE

3.1. Estimating equation framework

We start by sketching our two-stage instrumental variable method, which involves solving two separate estimating equations. To obtain $\hat{\gamma}$ and $\hat{\beta}$, we find the roots of the estimating functions,

i.e., the solutions to $U_1(\gamma) = 0$ and $U_2(\beta) = 0$ where

$$U_1(\gamma) = \sum_{i=1}^n U_{1i}(\gamma), \quad U_2(\beta) = \sum_{i=1}^n U_{2i}(\beta)$$

are the estimating equations for models (4) and (2), respectively.

The joint distribution of $\hat{\gamma}$ and $\hat{\beta}$ can be generated by iteratively solving the perturbed estimating equations using the same positive multipliers, whose mean and variance are 1 and which are independent of the data. Then the asymptotic covariance matrix of $\hat{\gamma}$ and $\hat{\beta}$ can be estimated by the sample covariance matrix of the simulated estimators. Let $R = (R_1, \dots, R_n)^T$ denote a vector of random variables used for perturbation. The perturbed estimating equations are

$$U_1^*(\gamma) = \sum_{i=1}^n U_{1i}^*(\gamma) = \sum_{i=1}^n U_{1i}(\gamma)R_i, \quad (5)$$

$$U_2^*(\beta) = \sum_{i=1}^n U_{2i}^*(\beta) = \sum_{i=1}^n U_{2i}(\beta)R_i. \quad (6)$$

We perturb the estimating equations by multiplying the summands in the original estimating equations by the same $(R_1, \dots, R_n)$, which ensures that the covariance of the estimating equations is correctly accounted for in the resampling. For $l = 1, 2$, under mild conditions, the conditional covariance matrix of $n^{-1/2}U_1^*(\gamma)$ and $n^{-1/2}U_2^*(\beta)$ given the data converges to the asymptotic covariance matrix of $n^{-1/2}U_1(\gamma)$ and $n^{-1/2}U_2(\beta)$ (Jin et al., 2001). For the accelerated failure time model, the resampling method used in (5) is sufficient to generate the marginal distribution of $\hat{\gamma}$ or $\hat{\beta}$ (Jin et al., 2003). However, to generate the joint distribution of the estimators, we need to modify (5), as discussed later. The resampling of the local Buckley–James estimator is similar to that of the rank estimator, but is more complex because perturbing the Kaplan–Meier estimator of an error distribution is needed.

Suppose that we repeatedly perturb the estimating equations a large number of times, say K , while fixing the data. Let $R^k = (R_1^k, \dots, R_n^k)^T$ denote the variables used for the k th perturbation, where $k = 1, \dots, K$. Denote by $\hat{\gamma}^k$ and $\hat{\beta}^k$ the corresponding estimators. Then the covariance matrix of $\hat{\gamma}$ and $\hat{\beta}$ can be estimated by the sample covariance matrix of $\hat{\gamma}^k$ and $\hat{\beta}^k$. The optimal two-stage instrumental variable estimator is obtained with the optimal weight $\Omega(\alpha_0)^{-1}$, which can be estimated by taking the inverse of the sample variance of $n^{1/2}(\gamma^k - \hat{B}^k \hat{\alpha}_I)$, where $\hat{\alpha}_I$ is considered to be fixed. The asymptotic variance of the optimal two-stage instrumental variable estimator is estimated by $n^{-1}\{\hat{B}^T \hat{\Omega}(\hat{\alpha}_I)^{-1} \hat{B}\}^{-1}$ or the empirical variance of $(\hat{\alpha}^1, \dots, \hat{\alpha}^K)$, where $\hat{\alpha}^k$ ($k = 1, \dots, K$) is the k th generated $\hat{\alpha}$.

In what follows, we assume that the outcome is right-censored, that is, Y is the log survival time. We consider several scenarios of right-censored outcomes, which are distinguished by the type of exposure. Case 1 involves a continuous exposure and Case 2 a binary exposure. Each main case is divided into two subcases based on whether the exposure is observed via coarsening or not. Case 1A is when the exposure is fully observed and Case 1B is when the exposure is censored. Case 2A is when the observed exposure is modelled using a continuous latent variable model, and Case 2B is when the observed exposure is modelled using a linear model. Case 2B is important in applications, since the interpretation of the linear model is more direct than for the latent variable model.

3.2. Case 1A: fully observed continuous exposure

Model (4) is the accelerated failure time model under the assumption that $(\tau_1, \dots, \tau_n)$ are independent error terms with a common but unspecified distribution. Let $C^Y = (C_1^Y, \dots, C_n^Y)^T$ be the vector of log censoring times. The data consist of $(\tilde{Y}_i, \Delta_i^Y, D_i)$ ($i = 1, \dots, n$), where $\tilde{Y}_i = \min(Y_i, C_i^Y)$ and $\Delta_i^Y = I(Y_i \leq C_i^Y)$. Here, $I(Q)$ takes the value 1 when Q is true and the value 0 otherwise. The usual censoring assumption is that Y_i and C_i^Y are independent conditionally on Z_i , W_i and U_i . For instrumental variable estimation, under the exclusion restriction assumption it is necessary to assume that Y_i and C_i^Y are independent conditionally on D_i .

Define $e_i(\gamma) = \tilde{Y}_i - \gamma^T D_i$, $N_i(\gamma; t) = \Delta_i^Y I\{e_i(\gamma) \leq t\}$ and $Y_i(\gamma; t) = I\{e_i(\gamma) \geq t\}$. Note that $N_i(\gamma; t)$ and $Y_i(\gamma; t)$ are the counting process and at-risk process on the residual time scale. The Gehan-type rank estimator $\hat{\gamma}_G$ is a root of $U_{1,G}(\gamma) = 0$, where

$$U_{1,G}(\gamma) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \Delta_i^Y (D_i - D_j) I\{e_i(\gamma) \leq e_j(\gamma)\}. \quad (7)$$

The perturbed version of (7) is

$$U_{1,G}^*(\gamma) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \Delta_i^Y (D_i - D_j) I\{e_i(\gamma) \leq e_j(\gamma)\} R_i R_j, \quad (8)$$

where $(R_1, \dots, R_n)$ are positive random variables with $E(R_i) = \text{var}(R_i) = 1$ which are independent of the data. The perturbation in (8) is more complex than in the usual approach, where each term in the estimating equation is multiplied by a single R_i . Jin et al. (2006a) showed that the resampling technique with (8) is valid for correlated failure time data.

For fully observed exposure, one may use the least squares estimator. For simplicity, we assume that $(\delta_1, \dots, \delta_n)$ in model (2) are independent with a common unspecified distribution. The least squares estimator of β_0 , denoted by $\hat{\beta}_L$, is obtained by solving $U_{2,L}(\beta) = 0$, which is the normal equation, with

$$U_{2,L}(\beta) = \sum_{i=1}^n (D_i - \bar{D})(Z_i - D_i^T \beta). \quad (9)$$

The perturbed estimating equation is

$$U_{2,L}^*(\beta) = \sum_{i=1}^n (D_i - \bar{D})(Z_i - D_i^T \beta) R_i, \quad (10)$$

where $(R_1, \dots, R_n)$ are the same random variables as in (8). Employing the same perturbations is essential to generating the joint distribution of $(\hat{\gamma}_G, \hat{\beta}_L)$.

Below we present a corollary on the asymptotic properties of the two-stage instrumental variable estimator using estimating equations (7) and (9) and a theorem on the validity of the resampling method in (8) and (10) for approximating the asymptotic distribution of the estimator.

COROLLARY 1. *For Case 1A, the Gehan-type rank estimator from (7) for γ_0 , denoted by $\hat{\gamma}_G$, and the least squares estimator in (9) for β_0 , denoted by $\hat{\beta}_L$, satisfy Assumptions 1 and 2 under*

Conditions A1–A4 in the Supplementary Material. Therefore, the two-stage estimator $\hat{\alpha}$ with $\hat{\gamma}_G$ and $\hat{\beta}_L$ converges in probability to α_0 and is asymptotically normal by Theorems 1 and 2.

THEOREM 3. *For Case 1A, under Conditions A1–A4 in the Supplementary Material, the asymptotic distribution of $\hat{\alpha}$ can be approximated by the empirical distribution of $(\hat{\alpha}^1, \dots, \hat{\alpha}^K)$ from (8) and (10) conditionally on the data.*

3.3. Case 1B: censored exposure

Left censoring often occurs in the measurement of biomarkers where assays may have lower limits of detection. A naïve approach to evaluating the association of these left-censored biomarker measurements with an outcome of interest which accounts for unmeasured confounding is to employ instrumental variable models discarding left-censored subjects (Smith et al., 2005). Such an approach was adopted in analysing the relationship between left-censored C-reactive protein levels and blood pressure, using Mendelian randomization as an instrumental variable. We propose a rigorous semiparametric analysis using rank-based methods that permit censoring of both exposure and outcomes, with perturbation resampling used for inference. The key point is that with left-censored data, one can use methods of estimation for accelerated failure time models with right-censored data, after reversing the signs of the event times (Klein & Moeschberger, 2003). In the Supplementary Material, a detailed construction of Gehan-type estimating equations is described, along with a proof of their theoretical validity, which follows Jin et al. (2006a, § 2). The results are stated in the following corollary and theorem.

COROLLARY 2. *For Case 1B, the Gehan-type rank estimators for γ_0 and β_0 , denoted by $\hat{\gamma}_G$ and $\hat{\beta}_G$, satisfy Assumptions 1 and 2 under Conditions A1–A4 in the Supplementary Material. Therefore, the two-stage estimator $\hat{\alpha}$ converges in probability to α_0 and is asymptotically normal by Theorems 1 and 2.*

THEOREM 4. *For Case 1B, under Conditions A1–A4 in the Supplementary Material, the asymptotic distribution of $\hat{\alpha}$ can be estimated by the empirical distribution of $(\hat{\alpha}^1, \dots, \hat{\alpha}^K)$ conditionally on the data.*

3.4. Case 2A: dichotomous exposure with a latent variable model

In Case 2A, we assume that Z_i in model (1) is a latent continuous variable that is not directly observed and determines receipt of treatment via a binary choice model, $\tilde{Z}_i = I(Z_i > 0)$, where $\tilde{Z}_i$ is the observed treatment variable. This modelling strategy measures the effect of the latent variable on the outcome. Latent variable models are often employed for dummy variables in instrumental variable methods with complete data (Heckman, 1978). Using this approach, we cannot make direct inference on the effect of $\tilde{Z}_i$, unlike with the method of Case 2B discussed in the next subsection. After estimating the effect of the latent variable, it may be of interest to determine the implied effect of the observed binary exposure. In the Supplementary Material, we investigate a method to determine the effect of the observed binary exposure from the latent variable model. However, this method involves strong assumptions, which may not be valid and which are challenging to verify in practice. Thus, latent variable methods are most useful in settings where the effect of the underlying latent variable is of interest. As an example, in psychological research, participants are often asked a binary question (Bollen, 2002); the participant answers yes if the underlying latent variable exceeds a certain threshold and no otherwise. It is common for researchers to be interested in the underlying latent variable in addition to the observed yes/no variable.

A probit regression model is considered for the observed binary exposure. For the identification of the model, we assume that $(\delta_1, \dots, \delta_n)$ are independent standard normal random variates. The probit model is

$$\text{pr}(\tilde{Z}_i = 1 \mid D_i) = \text{pr}(Z_i > 0 \mid D_i) = \Phi(\beta_{ZO} + \beta_0^\top D_i),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The maximum likelihood estimator for β_0 , $\hat{\beta}_M$, is obtained by solving the likelihood score equation

$$U_{2,M}(\beta_o, \beta) = \sum_{i=1}^n \frac{\{\tilde{Z}_i - \Phi(\beta_o + \beta^\top D_i)\} \phi(\beta_o + \beta^\top D_i)}{\Phi(\beta_o + \beta^\top D_i) \{1 - \Phi(\beta_o + \beta^\top D_i)\}} D_i,$$

where β_o is an index for an intercept, β is an index for the parameter β_0 , and $\phi(\cdot)$ is the standard normal density function. To generate the resampled maximum likelihood estimator for β_0 , we solve the perturbed score equation with the R used for $U_{1,G}(\gamma)$,

$$U_{2,M}^*(\beta_o, \beta) = \sum_{i=1}^n \frac{\{\tilde{Z}_i - \Phi(\beta_o + \beta^\top D_i)\} \phi(\beta_o + \beta^\top D_i)}{\Phi(\beta_o + \beta^\top D_i) \{1 - \Phi(\beta_o + \beta^\top D_i)\}} D_i R_i.$$

COROLLARY 3. *For Case 2A, the Gehan-type rank estimators for γ_0 , denoted by $\hat{\gamma}_G$, and the maximum likelihood estimator for β_0 , denoted by $\hat{\beta}_M$, satisfy Assumptions 1 and 2 under Conditions A1–A4 in the Supplementary Material. Therefore, the two-stage estimator $\hat{\alpha}$ with $\hat{\gamma}_G$ and $\hat{\beta}_M$ converges in probability to α_0 and is asymptotically normal by Theorems 1 and 2.*

THEOREM 5. *For Case 2A, under Conditions A1–A4 in the Supplementary Material, the asymptotic distribution of $\hat{\alpha}$ can be approximated by the empirical distribution of $(\hat{\alpha}^1, \dots, \hat{\alpha}^K)$ conditionally on the data.*

3.5. Case 2B: binary exposure

In Case 2B, Z_i in models (1) and (2) is binary. Hence, we estimate the effect of the binary exposure on the outcome. In this case, model (2) becomes a linear probability model and the variance of error terms depends on the covariates. That is, by construction, $E(\delta_i^* \mid D_i, U_i) = 0$ and $\text{var}(\delta_i^* \mid D_i, U_i) = \mu_z(D_i, U_i) \{1 - \mu_z(D_i, U_i)\}$, where $\mu_z(D_i, U_i) = E(Z_i \mid D_i, U_i)$. Since we fit the exposure model without U_i , a question arises as to whether the model is still the linear probability model, i.e., whether $\text{var}(\delta_i \mid D_i) = \mu_z(D_i) \{1 - \mu_z(D_i)\}$ is true, where $\mu_z(D_i) = E(Z_i \mid D_i)$. This is addressed in the following remark.

Remark 1. By a simple probability argument,

$$\text{var}(\delta_i \mid D_i) = E\{\text{var}(\delta_i \mid D_i, U_i) \mid D_i\} + \text{var}\{E(\delta_i \mid D_i, U_i) \mid D_i\}.$$

First we can show that $\text{var}\{E(\delta_i \mid D_i, U_i) \mid D_i\} = \beta_{ZU}^2 \text{var}(U_i \mid D_i)$. From $\text{var}(\delta_i \mid D_i, U_i) = \text{var}(\delta_i^* \mid D_i, U_i)$ and $\mu_z(D_i, U_i) = \mu_z(D_i) + \beta_{ZU} U_i$, it follows that $E\{\text{var}(\delta_i \mid D_i, U_i) \mid D_i\} = E\{\text{var}(\delta_i^* \mid D_i, U_i) \mid D_i\}$ is equal to

$$\mu_z(D_i) \{1 - \mu_z(D_i)\} - \beta_{ZU}^2 E(U_i^2 \mid D_i) + \beta_{ZU} E(U_i \mid D_i) - 2\beta_{ZU} \mu_z(D_i) E(U_i \mid D_i).$$

Since $E(U_i | D_i) = 0$, we have $\beta_{ZU}^2 E(U_i^2 | D_i) = \beta_{ZU}^2 \text{var}(U_i | D_i)$. Based on these results, we obtain that $\text{var}(\delta_i | D_i) = \mu_z(D_i)\{1 - \mu_z(D_i)\}$.

The error term of the reduced outcome model is $\tau_i = \varepsilon_i + \alpha_{YZ}\delta_i$. This implies that the variance of τ_i also depends on D_i , as does that of δ_i . Since the rank and Buckley–James methods assume equal variance, we adapt a recently developed method, the local Buckley–James method (Pang et al., 2015), to estimate the heteroscedastic accelerated failure time model. Pang et al. (2015) investigated models which assume that the residual variance is a nonparametric function of the mean. In our setting, the model can be written as

$$Y_i = \gamma_{YO} + \gamma_0^T D_i + \sigma(\gamma_0^T D_i)\omega_i,$$

where $(\omega_1, \dots, \omega_n)$ are independent and identically distributed random variables with mean zero and variance one, and $\sigma(\gamma_0^T D_i)$ is a nonparametric function of $\gamma_0^T D_i$. Since the heteroscedastic variance of the reduced outcome model arises from that of the linear exposure model, instead of $\gamma_0^T D_i$, the variance of the reduced outcome model will depend on $\beta_0^T D_i$. In other words, we have

$$Y_i = \gamma_{YO} + \gamma_0^T D_i + \sigma(\beta_0^T D_i)\omega_i, \quad (11)$$

where $\sigma(\beta_0^T D_i)$ is a nonparametric function of $\beta_0^T D_i$. Local Buckley–James estimation is directly applicable to model (11) because β_0 can be consistently estimated by least squares. In the following remark, we discuss simple sufficient conditions for $\tau_i = \sigma(\beta_0^T D_i)\omega_i$.

Remark 2. The conditional variance of τ_i given D_i is

$$\text{var}(\tau_i | D_i) = \text{var}(\varepsilon_i | D_i) + 2\alpha_{YZ} \text{cov}(\varepsilon_i, \delta_i | D_i) + \alpha_{YZ}^2 \text{var}(\delta_i | D_i).$$

If $\text{var}(\varepsilon_i | D_i)$ and $\text{cov}(\varepsilon_i, \delta_i | D_i)$ are constants, which may be unknown, then one can write $\tau_i = \sigma(\beta_0^T D_i)\omega_i$. This happens when $\text{var}(\varepsilon_i^* | D_i)$, $\text{var}(U_i | D_i)$, $\text{cov}(\varepsilon_i^*, U_i | D_i)$, $\text{cov}(\varepsilon_i^*, \delta_i^* | D_i)$ and $\text{cov}(\delta_i^*, U_i | D_i)$ are constants.

We now describe the local Buckley–James estimation procedure for model (11). The conditional mean of Y_i is

$$\begin{aligned} E(Y_i | Y_i \geq C_i^Y, \tilde{Y}_i, D_i) &= E(\tau_i | Y_i \geq C_i^Y, \tilde{Y}_i, \gamma_0^T D_i, \beta_0^T D_i) + \gamma_0^T D_i \\ &= \frac{\int_{\tilde{Y}_i - \gamma_0^T D_i}^{\infty} u \, dF_{\theta_0}(u | \beta_0^T D_i)}{1 - F_{\theta_0}(\tilde{Y}_i - \gamma_0^T D_i | \beta_0^T D_i)} + \gamma_0^T D_i, \end{aligned}$$

where $F_{\theta_0}(u | v)$ is an unknown cumulative distribution function of τ_i conditional on $\beta_0^T D_i = v$. Since $F_{\theta_0}(u | \beta_0^T D_i)$ depends on $\beta_0^T D_i$, it cannot be consistently estimated by the Kaplan–Meier method. Instead, a local Kaplan–Meier estimator (Dabrowska, 1987) will be used. The local Buckley–James estimation of model (11) can be implemented as follows.

Step 1. Obtain an initial estimator for γ_0 with the Buckley–James estimator or the rank estimator, and obtain an estimator for β_0 with the least squares estimator $\hat{\beta}_L$, as in Case 1A.

Step 2. At the a th iteration, compute the imputed Y_i by

$$\hat{Y}_i(\hat{\gamma}_a) = \Delta_i^Y \tilde{Y}_i + (1 - \Delta_i^Y) \hat{E}(Y_i | Y_i \geq C_i^Y, \tilde{Y}_i, \hat{\gamma}_a^T D_i, \hat{\beta}_L^T D_i) \quad (i = 1, \dots, n),$$

where

$$\hat{E}(Y_i | Y_i \geq C_i^Y, \tilde{Y}_i, \hat{\gamma}_a^T D_i, \hat{\beta}_L^T D_i) = \hat{\gamma}_a^T D_i + \frac{\int_{e_i(\hat{\gamma}_a)}^{\infty} u d\hat{F}_{\hat{\theta}_a}(u | \hat{\beta}_L^T D_i)}{1 - \hat{F}_{\hat{\theta}_a}\{e_i(\hat{\gamma}_a) | \hat{\beta}_L^T D_i\}}$$

with $e(\hat{\gamma}_a) = \tilde{Y}_i - \hat{\gamma}_a^T D_i$ and $\hat{\theta}_a = (\hat{\gamma}_a^T, \hat{\beta}_L^T)^T$. The local Kaplan–Meier estimate of $F_{\theta}(t | \beta^T D_i)$ is

$$\hat{F}_{\theta}(t | \beta^T D_i) = 1 - \prod_{j: e_j(\gamma) < t} \left\{ 1 - \frac{B_{nj}(\beta^T D_j) \Delta_i^Y}{\sum_{k=1}^n I\{e_k(\gamma) \geq e_j(\gamma)\} B_{nk}(\beta^T D_i)} \right\},$$

where $B_{nk}(\cdot)$ ($k = 1, \dots, n$) is a sequence of nonnegative weights with $\sum_{k=1}^n B_{nk}(\cdot) = 1$. For $B_{nk}(\beta^T D_i)$, the Nadaraya–Watson-type weight may be used:

$$B_{nk}(\beta^T D_i) = \frac{K\{(\beta^T D_i - \beta^T D_k)/h_n\}}{\sum_{l=1}^n K\{(\beta^T D_i - \beta^T D_l)/h_n\}},$$

where h_n is a bandwidth satisfying $h_n \rightarrow 0$ as $n \rightarrow \infty$ and $K(\cdot)$ is a symmetric kernel function.

Step 3. Apply least squares estimation to the imputed Y_i ($i = 1, \dots, n$) to obtain an updated estimator $\hat{\gamma}_{a+1}$:

$$\hat{\gamma}_{a+1} = \left\{ \sum_{i=1}^n (D_i - \bar{D}_n)^{\otimes 2} \right\}^{-1} \sum_{i=1}^n (D_i - \bar{D}_n) \{ \hat{Y}_i(\hat{\gamma}_a) - \bar{Y}_n(\hat{\gamma}_a) \},$$

where $\bar{Y}_n(\hat{\gamma}_a) = n^{-1} \sum_{i=1}^n \hat{Y}_i(\hat{\gamma}_a)$.

Step 4. Repeat Steps 2 and 3 until convergence is achieved. We denote by $\hat{\gamma}_B$ the converged estimator for γ_0 .

Pang et al. (2015) adapted the resampling technique of Jin et al. (2006b) to make inference using $\hat{\gamma}_B$. Let R_i ($i = 1, \dots, n$) be positive and have unit mean and variance. One may use $\hat{\gamma}_B$ as an initial estimator for this resampling procedure. Let $\hat{\gamma}_a^*$ denote the resampled estimator at the a th iteration. Define

$$L^*(\hat{\gamma}_a^*) = \left\{ \sum_{i=1}^n R_i (D_i - \bar{D}_n)^{\otimes 2} \right\}^{-1} \left[\sum_{i=1}^n R_i (D_i - \bar{D}_n) \{ \hat{Y}_i^*(\hat{\gamma}_a^*) - \bar{Y}_n^*(\hat{\gamma}_a^*) \} \right],$$

where

$$\begin{aligned} \hat{Y}_i^*(\hat{\gamma}_a^*) &= \Delta_i^Y \tilde{Y}_i + (1 - \Delta_i^Y) \left[\frac{\int_{e_i(\hat{\gamma}_a^*)}^{\infty} u d\hat{F}_{\hat{\theta}_a^*}^*(u | \hat{\beta}_L^T D_i)}{1 - \hat{F}_{\hat{\theta}_a^*}^*\{e_i(\hat{\gamma}_a^*) | \hat{\beta}_L^T D_i\}} + \hat{\gamma}_a^{*T} D_i \right], \\ \hat{F}_{\theta}^*(t | \beta^T D_i) &= 1 - \prod_{j: e_j(\gamma) < t} \left\{ 1 - \frac{R_j B_{nj}(\beta^T D_j) \Delta_i^Y}{\sum_{k=1}^n R_k I\{e_k(\gamma) \geq e_j(\gamma)\} B_{nk}(\beta^T D_i)} \right\}, \end{aligned}$$

$\bar{Y}_n^*(\gamma) = n^{-1} \sum_{i=1}^n \hat{Y}_i^*(\gamma)$ and $\hat{\theta}_a^* = (\hat{\gamma}_a^{*T}, \hat{\beta}_L^T)^T$. We denote the converged estimator by $\hat{\gamma}_B^*$.

COROLLARY 4. For Case 2B, the local Buckley–James estimator for γ_0 , denoted by $\hat{\gamma}_B$, and the least squares estimator for β_0 , denoted by $\hat{\beta}_L$, satisfy Assumptions 1 and 2 under Conditions B1–B9 in the Supplementary Material. Therefore, the two-stage estimator $\hat{\alpha}$ with $\hat{\gamma}_B$ and $\hat{\beta}_L$ converges in probability to α_0 and is asymptotically normal by Theorems 1 and 2.

THEOREM 6. For Case 2B, under Conditions B1–B9 in Supplementary Material, the asymptotic distribution of $\hat{\alpha}$ can be approximated by the empirical distribution of $(\hat{\alpha}^1, \dots, \hat{\alpha}^K)$ conditionally on the data.

4. SIMULATION STUDY

In this section, we report the results of numerical studies on the four data set-ups described in § 3. In Case 1A, we investigate how the skewness of the error distributions affects the performance of the estimators, including the proposed instrumental variable estimator and the naïve rank estimator based on fitting the accelerated failure time model directly to the exposure and measured confounders. In Case 1B, we consider a left-censored exposure and compare the instrumental variable method with the complete case analysis, which excludes the observations with left censoring of covariates and requires that the censoring be exogenous. In Case 2A, we examine the performance of the instrumental variable method with the latent variable. In Case 2B, we compare two instrumental variable methods: the method proposed in § 3.5 and the method proposed in § 3.2 that ignores heteroscedasticity.

The simulation models were

$$Y_i = \alpha_{YO} + \alpha_{YZ}Z_i + \alpha_{YW}^T W_i + \alpha_{YU}U_i + \varepsilon_i, \quad (12)$$

$$Z_i = \beta_{ZO} + \beta_{ZV}^T V_i + \beta_{ZW}^T W_i + \beta_{ZU}U_i + \delta_i. \quad (13)$$

Each V_i and W_i was two-dimensional in Cases 1A and 1B, and one-dimensional in Cases 2A and 2B. The parameter values in model (12) were set equal to 1 for all cases considered. The parameter values in model (13) were equal to 1 for Cases 1A and 1B, $(0, 1, 1, 1)$ for Case 2A, and $(0, b, 0.1, 0.2)$ for Case 2B where $b \in \{0.2, 0.4\}$. In Case 2A, Z_i is a latent variable, with the observed exposure defined as $\tilde{Z}_i = I(Z_i > 0)$. In Case 2B, $\delta_i = 0$ and the binary exposure was generated from $\text{Ber}\{P(Z_i = 1 \mid D_i, U_i)\}$.

We assumed that covariates $(V_i^T, W_i^T, U_i)^T$ were standard normal variates truncated at ± 2 . The censoring times for Y_i were generated from $\text{Un}(0, c_y)$, where c_y was chosen to yield a desired right-censoring rate of 20%. In Case 1A, we used standard normal and standard Gumbel distributions to generate ε_i and δ_i independently, and the standard Gumbel variate was standardized to have mean zero and variance one. For the distribution of $(\varepsilon_i, \delta_i)$, four combinations were considered: (N, N) , (N, G) , (G, N) and (G, G) , where N and G refer to $N(0, 1)$ and $G(0, 1)$ distributions, respectively. Only the results for (N, N) are presented here; results for the other cases are given in the Supplementary Material. In Case 1B, $\varepsilon_i \sim N(0, 1)$ and $\delta_i \sim N(0, 1)$ independently. The censoring times for Z_i were generated from $\text{Un}(-7, c_z)$, where c_z was chosen to yield a desired left-censoring rate of 20%. In Case 2A, for the identification of the probit model, we assumed that $\varepsilon_i \sim N(0, 1)$, $U_i \sim N(0, 0.5)$ and $\delta_i \sim N(0, 0.5)$, and this gave $\beta_{ZU}U_i + \delta_i \sim N(0, 1)$. In Case 2B, $\varepsilon_i \sim N(0, 1)$.

In the perturbation resampling, we generated R_i from the unit exponential distribution, with 500 resampled analyses conducted. We used the R (R Development Core Team, 2017) package `lss` (Huang & Jin, 2006) to implement the methods. We compared the proposed instrumental variable methods with the naïve rank method. For each estimator, Tables 1–3 show the average

Table 1. *Simulation results for Case 1A with $\varepsilon \sim N(0, 1)$ and $\delta \sim N(0, 1)$ and for Case 1B: the two-stage instrumental variable estimators with an identity matrix weight or an optimal weight and the naïve rank estimator; all values have been multiplied by 100*

n	Method	Parameter	Case 1A				Case 1B			
			Bias	ESE	ASE	ECR	Bias	ESE	ASE	ECR
100	Identity matrix	α_{YZ}	-0.2	14.0	14.7	95.6	0.0	15.7	15.4	93.6
		$\alpha_{YW,1}$	-0.1	23.1	25.4	95.0	1.2	25.5	25.5	93.6
		$\alpha_{YW,2}$	1.7	24.3	25.2	96.2	-1.7	27.1	25.5	93.6
	Optimal weight	α_{YZ}	-0.4	14.2	14.5	96.0	-0.1	15.8	15.2	93.7
		$\alpha_{YW,1}$	-0.2	23.5	24.9	94.4	1.2	25.6	25.0	93.7
		$\alpha_{YW,2}$	1.5	24.7	24.8	94.8	-1.8	27.2	25.1	93.7
	Naïve	α_{YZ}	23.7	8.1	7.9	17.2	23.6	10.1	9.6	29.8
		$\alpha_{YW,1}$	-24.6	17.7	18.0	70.6	-22.7	20.4	20.4	76.6
		$\alpha_{YW,2}$	-22.9	18.3	17.8	71.2	-24.6	20.4	20.6	74.4
200	Identity matrix	α_{YZ}	-0.2	9.7	9.9	95.4	0.1	9.9	10.4	95.1
		$\alpha_{YW,1}$	0.8	16.4	17.0	95.0	-0.4	17.2	17.5	95.1
		$\alpha_{YW,2}$	0.7	15.9	17.0	95.6	-0.3	18.2	17.5	95.1
	Optimal weight	α_{YZ}	-0.3	9.8	9.8	95.4	0.1	9.9	10.3	94.5
		$\alpha_{YW,1}$	0.8	16.5	16.8	95.0	-0.5	17.1	17.3	94.5
		$\alpha_{YW,2}$	0.8	16.0	16.8	95.2	-0.4	18.2	17.3	94.5
	Naïve	α_{YZ}	23.2	5.6	5.5	1.8	23.3	6.3	6.6	4.6
		$\alpha_{YW,1}$	-23.0	12.7	12.5	54.2	-24.1	13.4	14.2	61.6
		$\alpha_{YW,2}$	-22.8	12.1	12.5	57.4	-23.2	15.5	14.2	61.0

Bias, average bias; ESE, empirical standard error; ASE, average of the estimated standard errors; ECR, empirical coverage rate of the 95% Wald confidence intervals.

Table 2. *Simulation results for Case 2A: the two-stage instrumental variable estimator and the naïve rank estimator; all values have been multiplied by 100*

n	Method	Parameter	Bias	ESE	ASE	ECR
100	Two-stage	α_{YZ}	-2.5	24.5	25.4	94.6
		α_{YW}	0.2	30.4	29.8	95.4
	Naïve	α_{YZ}	28.9	10.1	10.1	18.6
		α_{YW}	-28.8	18.1	17.6	62.2
200	Two-stage	α_{YZ}	-2.5	17.7	17.4	91.7
		α_{YW}	0.6	20.8	20.3	93.3
	Naïve	α_{YZ}	28.5	6.8	7.1	1.2
		α_{YW}	-28.1	12.3	12.3	37.8
400	Two-stage	α_{YZ}	-0.3	11.7	12.3	95.8
		α_{YW}	-0.5	13.7	14.2	97.0
	Naïve	α_{YZ}	27.9	4.7	5.0	0.0
		α_{YW}	-27.7	8.8	8.6	13.0

Bias, average bias; ESE, empirical standard error; ASE, average of the estimated standard errors; ECR, empirical coverage rate of the 95% Wald confidence intervals.

bias, empirical standard error, average of the estimated standard errors, and empirical coverage rate of the 95% Wald confidence intervals from 500 samples.

The results demonstrate that the proposed instrumental variable estimators are unbiased and the proposed variance estimators perform well. The proposed estimators with the identity matrix and with the optimal weight performed similarly in our simulation settings. The naïve method gave biased estimators and their empirical coverage rates were far below the target coverage rate

Table 3. *Simulation results for Case 2B: the instrumental variable estimators of Cases 2B and 1A and the naïve rank estimator; all values have been multiplied by 100*

<i>n</i>	Method	Parameter	<i>b</i> = 0.2				<i>b</i> = 0.4			
			Bias	ESE	ASE	ECR	Bias	ESE	ASE	ECR
800	Case 2B IV	α_{YZ}	-3.5	25.7	25.2	96.2	-1.3	12.4	12.1	94.8
		α_{YW}	0.2	5.4	5.5	96.2	-0.2	4.8	4.9	95.2
	Case 1A IV	α_{YZ}	-15.2	26.4	26.1	92.0	-2.4	14.3	13.8	92.6
		α_{YW}	0.3	5.6	5.6	96.6	-0.1	5.1	5.3	95.8
	Naïve	α_{YZ}	28.8	5.7	5.9	0.2	23.4	5.3	5.3	0.6
		α_{YW}	-2.7	4.4	4.6	90.6	-2.3	4.4	4.6	92.6
1600	Case 2B IV	α_{YZ}	-1.0	17.7	17.3	94.9	-0.4	8.7	8.5	93.6
		α_{YW}	0.1	3.8	3.8	94.7	0.0	3.4	3.5	95.9
	Case 1A IV	α_{YZ}	-13.4	18.4	18.2	89.6	-1.4	9.8	9.7	94.2
		α_{YW}	0.2	3.8	3.9	95.2	0.0	3.6	3.7	96.4
	Naïve	α_{YZ}	29.2	3.9	4.2	0.0	23.3	3.7	3.7	0.0
		α_{YW}	-2.8	3.2	3.2	85.2	-2.1	3.1	3.2	90.2

Bias, average bias; ESE, empirical standard error; ASE, average of the estimated standard errors; ECR, empirical coverage rate of the 95% Wald confidence intervals; Case 2B IV, instrumental variable method of Case 2B; Case 1A IV, instrumental variable method of Case 1A.

of 95%. In Case 1A, the naïve estimator had much greater bias when $\delta_i \sim G(0, 1)$ than when $\delta_i \sim N(0, 1)$. This suggests that skewness of the exposure distribution may yield larger biases with unmeasured confounding. The proposed estimators performed well across the range of error distributions we considered. In Case 1B, the complete case estimator was biased because the censoring of the exposure is not exogenous due to unmeasured confounding. Thus the method of Wang & Feng (2012), which requires exogenous censoring, will not be applicable. The method in Case 2B gave unbiased estimators when $b = 0.2$ and 0.4 because it accounts for heteroscedasticity correctly. The method in Case 1A gave biased estimators when $b = 0.2$ but unbiased estimators when $b = 0.4$. This suggests that a strong instrumental variable may reduce the bias due to heteroscedasticity.

5. COLON CANCER DATA

We applied the proposed method to the Surveillance, Epidemiology and End Results data for elderly stage III colon cancer patients (Warren et al., 2002). Oxaliplatin is a chemotherapeutic agent that is used as part of a multi-agent adjuvant chemotherapy regimen for stage III colon cancer patients. Based on efficacy results from the MOSAIC trial in 2003 (Andre et al., 2004), the U.S. Food and Drug Administration approved oxaliplatin for use in treatment of stage III colon cancer. After the Food and Drug Administration's approval for this new indication, it was disseminated rapidly among stage III colon cancer patients to replace 5-fluorouracil, 5-FU, monotherapy as the standard of care. The objective of our analysis is to determine whether oxaliplatin, compared with 5-FU alone, improves survival in an older patient population, a question that was not addressed in the MOSAIC trial.

The cohort included individuals aged 65 and over who had been diagnosed with primary stage III colon cancer between 2003 and 2007, with follow-up through April 2010. Included patients were those who received surgical resection within 90 days of diagnosis, survived longer than 30 days, and initiated treatment with either oxaliplatin or 5-FU/capecitabine without oxaliplatin within 110 days of surgery and 120 days of diagnosis. Patients who received radiation, were diagnosed at autopsy, or had Health Maintenance Organization coverage or incomplete Medicare claims during the 12 months pre- and post-diagnosis or until death were excluded.

Table 4. *Results for the colon cancer data analysis*

Method	Parameter	Est	SE	<i>p</i> -value	LB	UB
Case 1A IV (R)	Oxaliplatin	0.368	0.142	0.010	0.089	0.647
	Age ₁	−0.140	0.102	0.171	−0.340	0.060
	Age ₂	−0.270	0.115	0.018	−0.495	−0.045
	Age ₃	−0.650	0.108	0.000	−0.863	−0.438
	Income	0.151	0.072	0.035	0.011	0.292
	Diabetes	−0.210	0.087	0.015	−0.379	−0.040
Case 1A IV (BJ)	Oxaliplatin	0.392	0.147	0.008	0.105	0.679
	Age ₁	−0.122	0.106	0.250	−0.329	0.086
	Age ₂	−0.310	0.121	0.010	−0.547	−0.073
	Age ₃	−0.656	0.117	0.000	−0.886	−0.426
	Income	0.152	0.076	0.045	0.003	0.300
	Diabetes	−0.245	0.093	0.009	−0.427	−0.062
Case 2B IV	Oxaliplatin	0.398	0.144	0.006	0.116	0.681
	Age ₁	−0.115	0.102	0.258	−0.314	0.084
	Age ₂	−0.312	0.121	0.010	−0.549	−0.075
	Age ₃	−0.637	0.117	0.000	−0.866	−0.408
	Income	0.158	0.075	0.034	0.012	0.304
	Diabetes	−0.244	0.093	0.009	−0.425	−0.062
Naïve	Oxaliplatin	0.149	0.079	0.059	−0.006	0.304
	Age ₁	−0.162	0.102	0.112	−0.362	0.038
	Age ₂	−0.294	0.114	0.010	−0.517	−0.071
	Age ₃	−0.709	0.104	0.000	−0.913	−0.505
	Income	0.167	0.070	0.017	0.030	0.304
	Diabetes	−0.220	0.085	0.010	−0.387	−0.053

Case 1A IV (R), Case 1A instrumental variable method with γ_0 estimated by the rank method; Case 1A IV (BJ), Case 1A instrumental variable method with γ_0 estimated by the Buckley–James method; Est, estimated coefficient; SE, standard error; LB, lower bound of the 95% Wald confidence interval; UB, upper bound of the 95% Wald confidence interval.

The outcome, Y , is log survival time in years. The binary exposure variable, Z , was coded as 1 if the patient was treated with oxaliplatin and 0 if treated with 5-FU. The instrumental variable was coded as 1 if the patient was treated after the FDA's approval of oxaliplatin for use in stage III colon cancer and 0 otherwise (Mack et al., 2015). Further details of the instrumental variable construction are provided in Mack et al. (2015). Three confounders were used: age in years, an indicator of whether household median income in 2000 was greater than \$50 000, and an indicator for diabetes. To account for a possible nonlinear effect of age on survival, we generated four groups based on quartiles of the age distribution and the corresponding three dummy variables: Age _{k} compares group k with the baseline group for $k = 1, 2, 3$, where increasing k indicates older groups. The sample size is 2879, with the resampling size equal to 200 when computing the standard errors of the parameter estimates with unit exponential perturbations.

We applied the method of Case 2B. The exogenous covariates are $D = (\text{Time}, \text{Age}_1, \text{Age}_2, \text{Age}_3, \text{Income}, \text{Diabetes})$, where Time is the instrumental variable. It was assumed that δ_i has a Bernoulli distribution with mean zero and variance $E(Z | D)\{1 - E(Z | D)\}$. Although the method of Case 1A may not be theoretically justified, for comparison we applied the homoscedastic instrumental variable methods using either rank or Buckley–James methodology. We also applied the naïve method based on fitting the outcome model to Y directly.

The results are given in Table 4. Since there is a single instrumental variable, the two-stage estimators do not depend on A_n . The naïve estimate of the treatment effect is 0.149, with p -value 0.059, which is not significant at level 0.05. In contrast, the estimated effect using the instrumental

variable methods is between 0.35 and 0.40, notably larger than the naïve estimate and statistically significant. To summarize the effect of oxaliplatin, we use the fact that the treatment effect parameter may be interpreted in terms of the differences in median survival times for the two treatments. Specifically, there is a $100\{\exp(\alpha_{YZ}) - 1\}\%$ increase for oxaliplatin over 5-FU when α_{YZ} is positive and a corresponding decrease when α_{YZ} is negative. The Case 2B method gives a 49% increase with a 95% confidence interval ranging from 12% to 98% in median survival with oxaliplatin, while the naïve rank method yields a 16% increase with a 95% confidence interval ranging from -1% to 36%.

Based on the instrumental variable analysis, we conclude that oxaliplatin is more beneficial than 5-FU in treating colon cancer patients. The differences between the naïve and instrumental variable results suggest that there may exist unmeasured confounders. The partial F-test statistic for the instrumental variable in the exposure model is 1324.64, which is much larger than the rule of thumb of 10 (Staiger & Stock, 1997), implying that the variable Time is a strong instrument. The estimated treatment effects from the two Case 1A methods are very similar to that from the Case 2B method. This similarity could be due to the instrumental variable being strong so that the effect of heteroscedascity may not be great, as evidenced in the simulations in § 4.

6. DISCUSSION

While we have focused on censored outcomes and exposures in this paper, the instrumental variable methods in § 2 are generally applicable to scenarios that involve incomplete observation of either the outcome, the exposure, the instrumental variable, or the measured confounders. The setting of Case 2A, where the exposure in the outcome model is a latent variable defining the observed binary exposure, only partially illustrates this broad applicability. Other scenarios involving either missing or mismeasured variables can be handled by the proposed framework, as long as there exist estimation procedures for the linear exposure model and the reduced-form outcome model which accommodate the incompletely observed data. Additional applications are currently under investigation.

For the binary exposure, Case 2B, we used the linear probability model. One may be tempted to use other models such as logistic regression for the binary exposure. However, doing so would lead to the reduced models being nonlinear. For such nonlinear cases, it is not straightforward to derive minimum distance estimators without strong model assumptions. To obtain instrumental variable estimators under the logistic exposure model, one might consider alternative approaches, such as the two-stage predictor substitution method (Terza et al., 2008), where the binary exposure is replaced by its predicted value. This is beyond the scope of the present work.

The linear probability model can have fitted values very close to or outside $[0, 1]$, which may lead to unreliable instrumental variable estimators. A possible remedy is to discretize continuous covariates or reduce the number of measured confounders in the model. This approach is valid unless the independence assumption does not hold.

There are two ways in which the censored exposure might occur: one where the exposure is the time to some event, which could be right-censored due to drop-out or loss to follow-up, and the other in which the censored exposure is a measured variable that is subject to a limit of detection. The latter is a good fit for the Case 1B methods. The former situation may involve both time-varying exposure and time-varying confounding, where it may not be straightforward to construct valid structural models using our approach. The development of such models and associated inferential procedures is complicated and merits further investigation.

An application of our method to longitudinal data is straightforward when responses are recorded at common time-points. Hogan & Lancaster (2004) developed instrumental variable

methods in such a setting, but only considered completely observed data. Complications may arise either with missing data or with time-dependent measured and unmeasured confounding. Under the Markov independence assumption in [Hogan & Lancaster \(2004\)](#), our models and methods of estimation can be extended to such settings.

ACKNOWLEDGEMENT

We thank William R. Carpenter, Christina D. Mack, and Anne-Marie Meyer for assistance in obtaining the data and for valuable insight based on their research studies examining innovation diffusion, which were supported by the National Cancer Institute. The current analysis is based on the data developed through their work.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs and additional simulation results.

REFERENCES

- AMEMIYA, T. (1974). The nonlinear two-stage least-squares estimator. *J. Economet.* **2**, 105–10.
- ANDERSON, T. W. & RUBIN, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.* **21**, 570–82.
- ANDRE, T., BONI, C., MOUNEDJI-BOUDIAF, L., NAVARRO, M., TABERNERO, J., HICKISH, T., TOPHAM, C., ZANINELLI, M., CLINGAN, P., BRIDGEWATER, J. et al. (2004). Oxaliplatin, uorouracil, and leucovorin as adjuvant treatment for colon cancer. *New Engl. J. Med.* **350**, 2343–51.
- ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–55.
- BAKER, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *J. Am. Statist. Assoc.* **93**, 929–34.
- BOLLEN, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika* **61**, 109–21.
- BOLLEN, K. A. (2002). Latent variables in psychology and the social sciences. *Ann. Rev. Psychol.* **53**, 605–34.
- BRÄNNÄS, K. (2000). Estimation in a duration model for evaluating educational programs. *IZA Discussion Paper* no. 103, Institute for the Study of Labor (IZA), Bonn, Germany.
- BROOKHART, M. A., RASSEN, J. A. & SCHNEEWEISS, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol. Drug Safety* **19**, 537–54.
- BUCKLEY, J. & JAMES, I. (1979). Linear regression with censored data. *Biometrika* **65**, 429–36.
- DABROWSKA, D. M. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.* **14**, 181–97.
- HANSEN, L. B. (1982). Large sample properties of generalized method of moment estimators. *Econometrica* **50**, 1029–54.
- HECKMAN, J. J. (1978). Dummy endogeneous variables in a simultaneous equation system. *Econometrica* **46**, 931–59.
- HOGAN, J. W. & LANCASTER, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statist. Meth. Med. Res.* **13**, 17–48.
- HUANG, L. & JIN, Z. (2006). *lss: The accelerated failure time model to right censored data based on least-squares principle*. R package version 0.52.
- JIN, Z., YING, Z. & WEI, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–90.
- JIN, Z., LIN, D. Y., WEI, L. J. & YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–53.
- JIN, Z., LIN, D. Y. & YING, Z. (2006a). Rank regression analysis of multivariate failure time data based on marginal linear models. *Scand. J. Statist.* **33**, 1–23.
- JIN, Z., LIN, D. Y. & YING, Z. (2006b). On least-squares regression with censored data. *Biometrika* **93**, 147–61.
- KLEIN, J. P. & MOESCHBERGER, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.
- LAI, T. L. & YING, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist.* **19**, 1370–402.
- LI, J., FINE, J. P. & BROOKHART, M. A. (2015). Instrumental variable additive hazards models. *Biometrics* **71**, 122–30.

- LOEYS, T. & GOETGHEBEUR, E. (2003). A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics* **59**, 100–5.
- LOEYS, T., GOETGHEBEUR, E. & VANDEBOSCH, A. (2005). Causal proportional hazards models and time-constant exposure in randomized clinical trials. *Lifetime Data Anal.* **11**, 435–49.
- MACK, C. D., BROOKHART, M., GLYNN, R., CARPENTER, W. & STURMER, T. (2015) Comparative effectiveness of oxaliplatin versus 5-fluorouracil in older adults: An instrumental variable analysis. *Epidemiology* **26**, 690–9.
- NIE, H., CHENG, J. & SMALL, D. S. (2011). Inference for the effect of treatment on survival probability in randomized trials with noncompliance and administrative censoring. *Biometrics* **67**, 1397–405.
- NEWBY, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica* **58**, 809–37.
- NEWBY, W. K. & POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–78.
- PANG, L., LU, W. & WANG, H. J. (2015). Local Buckley-James estimation for heteroscedastic accelerated failure time model. *Statist. Sinica* **25**, 863–77.
- POWELL, J. L. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica* **54**, 1435–60.
- PRENTICE, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167–79.
- R DEVELOPMENT CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- ROBINS, J. M. & TSIATIS, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun. Statist. A* **20**, 2609–31.
- SMITH, G. D., LAWLOR, D. A., HARBORD, R., TIMPSON, N., RUMLEY, A., LOWE, G. D. O., DAY, I. N. M. & EBRAHIM, S. (2005). Association of C-reactive protein with blood pressure and hypertension: Life course confounding and Mendelian randomization tests of causality. *Arterioscler. Thromb. Vasc. Biol.* **25**, 1051–6.
- STAIGER, D. & STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–86.
- TCHETGEN TCHETGEN, E. J., WALTER, S., VANSTEELENDT, S., MARTINUSSEN, T. & GLYMOUR, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology* **26**, 402–10.
- TERZA, J. V., BASU, A. & RATHOUZ, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *J. Health Econ.* **27**, 531–43.
- TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354–72.
- WANG, H. J. & FENG, X. (2012). Multiple imputation for M-regression with censored covariates. *J. Am. Statist. Assoc.* **107**, 194–204.
- WARREN, J. L., KLABUNDE, C. N., SCHRAG D., BACH P. B. & RILEY, G. F. (2002). Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med. Care* **40**(8 Suppl), IV-3-18.
- YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76–99.

[Received on 6 January 2014. Editorial decision on 13 July 2017]