Behaviormetrika (2023) 50:241–262
https://doi.org/10.1007/s41237-022-00177-9

ORIGINAL PAPER

Instrumental variable estimation of causal effects with applying some model selection procedures under binary outcomes

Shunichiro Orihara[1] · Atsushi Goto[1] · Masataka Taguri[1]

Received: 12 March 2022 / Accepted: 10 June 2022 / Published online: 9 July 2022

Abstract

In observational studies, unmeasured covariates are an important problem. In the presence of some unmeasured covariates, some instrumental variable methods, such as the two-stage residual inclusion (2SRI) estimator or limited-information maximum likelihood (LIML) estimator, can still obtain an unbiased estimate for causal effects despite the existence of nonlinear models, such as logistic regression and probit models. However, not only a correct outcome model but also a correct treatment model needs to be specified. Therefore, it is important to identify the correct models. In this paper, we consider model selection procedures for 2SRI and LIML, and confirm their properties through simulation and real datasets. Specifically, we confirm the model selection procedures can detect the correct treatment and outcome models, and unbiased causal effects can be estimated. The model selection properties are confirmed through simulation datasets and GENEVA Diabetes Study datasets. From the simulation and data analysis results, we recommend that LIML with any model selection procedures is a good choice when there are binary outcomes and any concerns about unmeasured covariates.

Keywords Instrumental variable · Limited-information maximum likelihood · Mendelian randomization · Model selection · Two-stage residual inclusion · Unmeasured covariates

Mathematics Subject Classification 62F12 · 62P10

Communicated by Maomi Ueno.

✉ Shunichiro Orihara
w205702a@yokohama-cu.ac.jp

[1] 22-2 Seto, Kanazawa-ku, Yokohama City, Kanagawa, Japan

1 Introduction

Observational studies are usually interested in estimating causal effects between treatments and outcomes. When all covariates or confounders (hereafter, referred to as "covariates") are observed, the covariates can be adjusted and an unbiased estimator for causal effects can be obtained, as in the case of "no unmeasured confounding" (c.f. Hernán and Robins 2020). No unmeasured confounding is a sufficient assumption for the estimation of an unbiased estimator of causal effects. However, there are serious risks in estimating biased causal effects unless the covariates are adjusted appropriately. When some covariates are not observed, usually, an unbiased estimator cannot be obtained, as in the case of some unmeasured covariates. Unmeasured covariates constitute an important problem in causal inference, since no unmeasured confounding is no longer applied. Therefore, different estimation methods should be applied.

In this study, the focus is on instrumental variable (IV) methods. A two-stage least squares (2SLS) estimator is one of the most important two-step procedures in the estimation of IV causal effects when there are unmeasured covariates (Wooldridge 2010). 2SLS is useful, but it requires the key assumption that there is a linear relationship between the treatment variable and outcome variable. If this assumption is violated, a biased estimate of the causal effect of interest may be obtained. Terza et al. (2008) introduced a two-stage residual inclusion (2SRI) estimator similar to the control function approach (Wooldridge 2010). 2SRI is another two-step procedure expanded to include nonlinear models, such as logistic regression and probit models, whereby an unbiased estimate of the causal effect can be obtained even when there are nonlinear models. Although 2SRI overcomes the problem of 2SLS, it may derive biased causal effects, as mentioned in Basu et al. (2017) and Wan et al. (2018). According to the simulation results of Basu et al. (2017), a full-likelihood approach derives a more accurate estimate than 2SRI (see also Section 5 of Burgess et al. 2017). Therefore, in this manuscript, a limited-information maximum likelihood (LIML) estimator (Wooldridge 2014) is also considered. LIML estimator uses a full-likelihood approach, but has features similar to those of 2SRI and the control function approach. Both 2SRI and LIML can be used for nonlinear models; however, not only the correct outcome model but also the correct treatment model needs to be specified (Basu et al. 2017). Therefore, detecting the correct models is an important process when using 2SRI and LIML.

In model selection, information criteria are commonly used to select the "correct" model. The Akaike information criterion (AIC; Akaike 1974) is the best-known information criterion for selecting the best model in the prediction of future outcomes. The Bayesian information criterion (BIC) proposed by Schwarz (1978) is another well-known information criterion with model selection consistency (i.e., it selects the correct model with probability 1) under certain assumptions (Nishii 1984, Shao 1997). In the field of causal inference, some previous studies exist (Brookhart and van der Laan 2006; Vansteelandt et al. 2012) and Taguri et al. (2014). Although the considered procedures varied among the studies, motivation is the same: to estimate unbiased causal effects, a valid model needs to be considered. However, to best of our knowledge, there are no previous reports related to any model selection procedures for 2SRI and LIML.

In this paper, we consider model selection procedures for 2SRI and LIML, and confirm their properties through simulation and real datasets. Specifically, we confirm the model selection procedures can detect the correct treatment and outcome model, and unbiased causal effects can be estimated. Since previous studies have considered model selection procedures neither for 2SRI (the control function approach) nor for LIML in this context, the contribution of this study may be considered significant for these estimation procedures. In Sect. 2, a motivational example is introduced and the model considered in this study is presented. Two situations are considered: continuous and dichotomous treatments. In addition, we introduce AIC-type and BIC-type information criteria. In Sect. 3, the properties of 2SRI and LIML with model selection are confirmed using simulation datasets. In the simulation, we consider a case in which the distribution of unmeasured covariates is correctly specified. In Sect. 4, data analysis is performed using the GENEVA Diabetes Study dataset. Supplementary information on simulations and the GENEVA Diabetes Study datasets are found in the Appendix. Some calculations and supplemental simulations are found in the Web Appendix.

2 Motivation example and IV methods

First, a motivational example, the GENEVA Diabetes Study datasets which store subjects' demographic information (phenotype), genetic information (genotype), and outcomes (presence or absence of diabetes), is introduced. In this study, the causal effect of body mass index (BMI) on the incidence of diabetes is investigated. As is well known, diabetes affects some parts of the body, such as eyes, kidney, and heart. There are more than 400 million diabetic patients worldwide (Cheng et al. 2019). In addition, BMI and the incidence of diabetes have a positive relationship such that high BMI increases the likelihood of developing diabetes. To estimate the causal effect correctly, the covariates, regardless of whether they are observed or not, need to be adjusted when the datasets are derived from observational studies. Cheng et al. (2019) and Richardson et al. (2020) used an instrumental variable approach with the genetic information constituting the instrumental variables. This analysis strategy is called "Mendelian randomization" (Burgess et al. 2017). In this study, Mendelian randomization was also conducted using the genetic information included in the GENEVA Diabetes Study datasets.

Herein, a more general formulation is considered. Let n be the sample size and assume that $i = 1, 2, \dots, n$ are i.i.d. samples. $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Z} \in \mathbb{R}^K$ denote vectors of covariates and IVs, respectively. The following relationship is assumed for the unmeasured variables:

$$\begin{pmatrix} V \\ U \end{pmatrix} \sim F(v, u; \xi), \quad \begin{pmatrix} V \\ U \end{pmatrix} \perp\!\!\!\perp \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Z} \end{pmatrix}, \tag{2.1}$$

where ξ is a parameter of the joint distribution $(V, U) \in \mathbb{R}^2$, referred to as "unmeasured covariates" in this study. These assumptions are similar to those of Wooldridge

(2014); (2.1) suggests a LIML estimation procedure. Next, the models considered in this paper are introduced.

- Treatment model (continuous treatment)

$$W = \varphi_1(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\alpha}) + V \tag{2.2}$$

- Treatment model (dichotomous treatment)

$$W = \mathbf{1}\left\{\varphi_1(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\alpha}) + V \geq 0\right\}$$

- Outcome model

$$Y = \mathbf{1}\left\{\varphi_2(W, \boldsymbol{X}; \boldsymbol{\beta}) + U \geq 0\right\}, \tag{2.3}$$

where φ_1 and φ_2 are twice differentiable predictors with respect to parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. For instance, φ_1 and φ_2 can be selected as a linear model:

$$\varphi_1(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\alpha}) = Z_i^\top \boldsymbol{\alpha}_z + X_i^\top \boldsymbol{\alpha}_x, \quad \varphi_2(W, \boldsymbol{X}; \boldsymbol{\beta}) = W_i^\top \beta_w + X_i^\top \boldsymbol{\beta}_x.$$

In addition, the parameter spaces of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are denoted by $\Theta_{\boldsymbol{\alpha}}$ and $\Theta_{\boldsymbol{\beta}}$, respectively.

The explanation of the above variables by DAGs (c.f. Hernán and Robins 2020) is shown in the Web Appendix A.

The pair of models (2.2) and (2.3) is called the Rivers-Vuong model (RV model; Rivers and Vuong (1988)), where

$$F(v, u; \xi) = N_2\left(\mathbf{0}_2, \begin{pmatrix} \sigma_v^2 & \rho\sigma_v \\ \rho\sigma_v & 1 \end{pmatrix}\right), \ \ \xi = (\sigma_v, \rho)^\top, \ \sigma_v > 0, \ \rho \neq 0.$$

Under the RV model, (2.3) becomes a probit model. Note that the following discussions are not limited to the above treatment / outcome models but apply to other parametric models, as well. Note also that the IVs $\boldsymbol{Z}$ follow three IV features (see Baiocchi et al. 2014): (1) causal association with the treatment variable T, (2) no association with the unobserved variables (V, U), and (3) no direct causal association with the outcome variable Y. The first and third features are explained using the above treatment and outcome models, respectively. The second feature is explained by (2.1).

To estimate the parameters $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \xi^\top\right)^\top \in \Theta = \Theta_{\boldsymbol{\alpha}} \times \Theta_{\boldsymbol{\beta}} \times \Theta_{\xi}$, two IV estimators are introduced: a 2SRI estimator and a LIML estimator.

2.1 Two-stage residual inclusion

The 2SRI estimator estimates the causal effects in two steps. In the first step, the treatment variable is regressed onto the instrumental variables to construct the residuals of the treatment variable. Specifically, (2.2) and (2.3) are considered. In particular, consider the ordinary least squares estimator of $\boldsymbol{\alpha}$:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(w_i - \varphi_1(z_i, \boldsymbol{x}_i; \boldsymbol{\alpha})\right)^2.$$

For each predictor $\varphi_1(\boldsymbol{Z}_i, \boldsymbol{X}_i; \hat{\boldsymbol{\alpha}})$, the residuals are derived:

$$v_i(\hat{\boldsymbol{\alpha}}) = w_i - \varphi_1(z_i, \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}). \tag{2.4}$$

In the second step, the outcome variable is regressed not only onto the treatment variables but also onto the residuals of the treatment variables. In the following model, the residuals are plugged into (2.3). For instance, when U is a logistic distribution, the outcome model becomes a logistic regression model:

$$p_i(\boldsymbol{\beta}, \gamma) = expit\left\{\varphi_2(w_i, \boldsymbol{x}_i; \boldsymbol{\beta}) + v_i(\hat{\boldsymbol{\alpha}})\gamma\right\}. \tag{2.5}$$

In contrast, when U is a normal distribution, the above outcome model becomes a probit model:

$$p_i(\boldsymbol{\beta}, \gamma) = \Phi\left(\varphi_2(w_i, \boldsymbol{x}_i; \boldsymbol{\beta}) + v_i(\hat{\boldsymbol{\alpha}})\gamma\right). \tag{2.6}$$

Under (2.5) or (2.6), the maximum likelihood estimator of $(\boldsymbol{\beta}, \gamma)$ is considered:

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\gamma} \end{pmatrix} &= \arg\max_{\boldsymbol{\beta}, \gamma} \ \log \left[\prod_{i=1}^{n} p_i(\boldsymbol{\beta}, \gamma)^{y_i} \left(1 - p_i(\boldsymbol{\beta}, \gamma)\right)^{1-y_i}\right] \\ &= \arg\max_{\boldsymbol{\beta}, \gamma} \ \ell_{2SRI}(\boldsymbol{\beta}, \gamma). \end{aligned} \tag{2.7}$$

Through the above procedures, a 2SRI estimator $\hat{\boldsymbol{\beta}}$ is obtained. Note that the residuals are more complicated than (2.4) for dichotomous treatment (Tchetgen et al. 2015); this is drawback of 2SRI. As mentioned below, LIML need not be considered the residual; a full-likelihood is only necessary.

In Section 3, to consider performance of model selection procedures, the following AIC and BIC are considered:

$$\begin{aligned} AIC_{2SRI} &= -2\ell_{2SRI}(\hat{\boldsymbol{\beta}}, \hat{\gamma}) + 2(|\hat{\boldsymbol{\beta}}| + 1) \\ BIC_{2SRI} &= -2\ell_{2SRI}(\hat{\boldsymbol{\beta}}, \hat{\gamma}) + (|\hat{\boldsymbol{\beta}}| + 1)\log(n), \end{aligned}$$

where $|\cdot|$ is the number of elements. Note that these are applied to select an outcome model without considering v as the predicted residuals (i.e. the same handling as the other covariates).

2.2 Limited-information maximum likelihood

Let us consider the likelihood function $L_{LIML}(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_{LIML,i}(\boldsymbol{\theta})$ conditioning on $\boldsymbol{z}$ and $\boldsymbol{x}$:

$$L_{LIML}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, w_i | \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \mathrm{P}(y_i | w_i, \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) f(w_i | \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\alpha}). \tag{2.8}$$

In the following, the specific form of the likelihood for the two cases is explicitly defined. In the case of the Rivers-Vuong model, (2.8) becomes

$$L_{LIML}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \Phi\left(\frac{\varphi_{i2}(\boldsymbol{\beta}) + \rho v_i(\boldsymbol{\alpha})}{\sqrt{1-\rho^2}} \right)^{y_i} \left(1 - \Phi\left(\frac{\varphi_{i2}(\boldsymbol{\beta}) + \rho v_i(\boldsymbol{\alpha})}{\sqrt{1-\rho^2}} \right) \right)^{1-y_i} \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left\{ -\frac{v_i^2(\boldsymbol{\alpha})}{2\sigma_v^2} \right\} \tag{2.9}$$

(see Web Appendix B.1). Therefore, the log-likelihood $\ell_{LIML}(\boldsymbol{\theta}) = \log L_{LIML}(\boldsymbol{\theta})$ becomes:

$$\begin{aligned} \ell_{LIML}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Bigg\{ & y_i \log \Phi\left(\frac{\varphi_{i2}(\boldsymbol{\beta}) + \rho v_i(\boldsymbol{\alpha})}{\sqrt{1-\rho^2}} \right) + (1-y_i) \log \left(1 - \Phi\left(\frac{\varphi_{i2}(\boldsymbol{\beta}) + \rho v_i(\boldsymbol{\alpha})}{\sqrt{1-\rho^2}} \right) \right) \\ & - \frac{v_i^2(\boldsymbol{\alpha})}{2\sigma_v^2} - \log\left(\sqrt{2\pi\sigma_v^2} \right) \Bigg\}. \end{aligned}$$

For dichotomous treatment, (2.8) becomes

$$\begin{aligned} L_{LIML}(\boldsymbol{\theta}) = \prod_{i=1}^{n} & \mathrm{P}\big(y_i = 1, w_i = 1 | \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}\big)^{y_i w_i} \mathrm{P}\big(y_i = 0, w_i = 1 | \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}\big)^{(1-y_i)w_i} \\ & \times \mathrm{P}\big(y_i = 1, w_i = 0 | \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}\big)^{y_i(1-w_i)} \mathrm{P}\big(y_i = 0, w_i = 0 | \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}\big)^{(1-y_i)(1-w_i)}. \end{aligned}$$

Therefore, the log-likelihood becomes

$$\begin{aligned} \ell_{LIML}(\boldsymbol{\theta}) = \sum_{i=1}^{n} & \big\{ y_i w_i \log \big\{ 1 - F(\infty, -\varphi_{i2}(\boldsymbol{\beta}); \xi) - F(-\varphi_{i1}(\boldsymbol{\alpha}), \infty; \xi) + F(-\varphi_{i1}(\boldsymbol{\alpha}), -\varphi_{i2}(\boldsymbol{\beta}); \xi) \big\} \\ & + (1-y_i) w_i \log \big\{ F(\infty, -\varphi_{i2}(\boldsymbol{\beta}); \xi) - F(-\varphi_{i1}(\boldsymbol{\alpha}), -\varphi_{i2}(\boldsymbol{\beta}); \xi) \big\} \\ & + y_i (1-w_i) \log \big\{ F(-\varphi_{i1}(\boldsymbol{\alpha}), \infty); \xi) - F(-\varphi_{i1}(\boldsymbol{\alpha}), -\varphi_{i2}(\boldsymbol{\beta}); \xi) \big\} \\ & + (1-y_i)(1-w_i) \log \big\{ F(-\varphi_{i1}(\boldsymbol{\alpha}), -\varphi_{i2}(\boldsymbol{\beta}); \xi) \big\} \big\} \end{aligned} \tag{2.10}$$

(see Web Appendix B.2), where

$$F(v, \infty; \xi) = \lim_{u \to \infty} F(v, u; \xi), \quad F(\infty, u; \xi) = \lim_{v \to \infty} F(v, u; \xi).$$

By maximizing the likelihood (2.8), a limited-information maximum likelihood estimator can be derived as

$$\hat{\boldsymbol{\theta}} = \mathop{\arg\max}_{\theta \in \Theta} \ \ell_{LIML}(\boldsymbol{\theta}). \tag{2.11}$$

Note that the joint distribution $F(v, u; \xi)$ has to be specified when using LIML. However, the distribution is somewhat flexible; for instance, some parametric copulas can be selected (e.g.,Biller and Corlu 2012; Fantazzini 2009). When the marginal distributions of V and U are assumed to be the logistic distributions $F_V^{logis}(v)$ and

$F_U^{logis}(u)$, respectively, and some parametric copulas, such as the t-copula or Clayton copula $C(\cdot,\cdot;\xi)$ are assumed, the joint distribution becomes

$$F(v,u;\xi) = C(F_V^{logis}(v), F_U^{logis}(u);\xi).$$

In Section 3, to consider performance of model selection procedures, the following AIC and BIC are considered:

$$AIC_{LIML} = -2\ell_{LIML}(\hat{\boldsymbol{\beta}},\hat{\gamma}) + 2|\hat{\boldsymbol{\theta}}|$$
$$BIC_{LIML} = -2\ell_{LIML}(\hat{\boldsymbol{\beta}},\hat{\gamma}) + |\hat{\boldsymbol{\theta}}|\log(n)$$

2.3 Interpretation under potential outcomes

The 2SRI and LIML can estimate the average treatment effects (ATE) (Rosenbaum and Rubin 1983). To estimate ATE by these methods, G-computation (e.g., Hernán and Robins, 2020)) can be applied:

1. By solving (2.7) and (2.11), the 2SRI estimates or LIML estimates of $\boldsymbol{\beta}$ can be obtained.
2. To estimate a probability under the particular treatment value (written as w'), the average is calculated over all populations; for instance, U under the normal distribution and the probit model:

$$\hat{\mathrm{P}}\left(Y_{w'} = 1\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathrm{P}}\left(Y = 1|w',\boldsymbol{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\varphi_2\left(w',\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}\right)\right),$$

   where $Y_{w'}$ corresponds to the potential outcome under treatment w', and $\boldsymbol{x}_i$ are the observed covariates. Regarding 2SRI, $\boldsymbol{x}_i$ also includes the residual term of the 1st stage model.

From the above steps, ATE is estimated:

$$\hat{\mathrm{E}}[Y_{w'}] - \hat{\mathrm{E}}[Y_{w''}] = \hat{\mathrm{P}}\left(Y_{w'} = 1\right) - \hat{\mathrm{P}}\left(Y_{w''} = 1\right).$$

3 Simulations

In this section, the properties of model selection procedures and parameter estimates of 2SRI and LIML are confirmed. Because no previous studies have considered model selection procedures for 2SRI (or the control function approach) and LIML, our simulation results may provide some guidance for using these estimation procedures. To confirm these properties, (1) the number of times the true model was selected for each procedure and the corresponding proportions were determined, and

(2) descriptive statistics of estimates for each procedure were calculated. The number of iterations for the simulations was 1000.

3.1 Continuous treatment and normal unmeasured covariates

The Rivers–Vuong model was considered. In this setting, it was confirmed that LIML and 2SRI estimator with model selection perform well. Since we can apply 2SLS under this situation, the results of the 2SLS are summarized as well for reference. The simulation settings were as follows:

Covariates: $X_1 \sim N(0,1)$, $X_2 \sim Ber(0.5)$, $X_3 \sim N(0,1)$
An instrumental variable: $Z \sim Ber(0.5)$
Unmeasured covariates: $\begin{pmatrix} V \\ U \end{pmatrix} \sim N\left(\mathbf{0}_2, \begin{pmatrix} 1 & \rho \\ & 1 \end{pmatrix}\right)$

- Weak correlation: $\rho = 0.3$
- Strong correlation: $\rho = 0.6$

A treatment model: $W = 1 + \alpha_z Z + X_2 + X_3 + V$

- Weak instrumental variable: $\alpha_z = 0.2 \Rightarrow$ The correlation between a treatment and IV is approximately 0.06.
- Strong instrumental variable: $\alpha_z = 1 \Rightarrow$ The correlation between a treatment and an IV is approximately 0.3.

An outcome model: $Y = \mathbf{1}\{0.5 + 0.6W + 0.5X_1 + 0.5X_2 + U \geq 0\}$

To select a treatment model and outcome model, candidate models were prepared. The supplemental information is provided in Appendix A.

The simulation results are summarized in tables and supplemental figures. The results of model selection are summarized in Table 1, where **2SRI: AIC** and **2SRI: BIC** are 2SRI with each model selection procedures, **LIML: AIC** and **LIML: BIC** are LIML with each model selection procedures. The column "True model" shows the number of times the selected method was the true model (i.e., the pair of models a4 and b2; see Appendix). The column "Including true model" shows the number of times each selected method was the true or larger models (i.e., not misspecified models; see Appendix). The column "Both true model" shows the number of times the 2SRI estimator selected the true model in the first step and the second step. Throughout the simulations, "(1) Weak correlation and Strong IV" were used as reference settings. In terms of selection probabilities of the "True model," BIC did not display high probability for small samples ($N = 100$); however, it was the best out of the three selection procedures in all cases of large samples ($N = 300$). This result is the same as the previous theoretical results (the model selection consistency; see Nishii 1984 and Shao 1997). For selection probabilities of "Including true model," both 2SRI and LIML displayed high probability even for small samples. This is also the same feature as that of AIC. Regarding 2SRI, the selection probabilities of "Both true models" were also

Table 1 Summary of the results of model selection for each estimator (continuous treatment and normal unmeasured covariates)

Samplesize	**Situation**	**Method**	Step	True model n (%)	Including true model n (%)	Both true model n (%)
$N = 100$	(1) Weak correlation and Strong IV	**2SRI: AIC**	1st	871 (87.1)	1000 (100)	840 (84.0)
			2nd	450 (45.0)	840 (84.0)	
		2SRI: BIC	1st	989 (98.9)	996 (99.6)	438 (43.8)
			2nd	355 (35.5)	440 (44.0)	
		LIML: AIC	–	390 (39.0)	837 (83.7)	–
		LIML: BIC	–	347 (34.7)	430 (43.0)	–
	(2) Weak correlation and Weak IV	**2SRI: AIC**	1st	871 (87.1)	1000 (100)	883 (88.3)
			2nd	486 (48.6)	883 (88.3)	
		2SRI: BIC	1st	989 (98.9)	996 (99.6)	492 (49.2)
			2nd	424 (42.4)	493 (49.3)	
		LIML: AIC	–	422 (42.2)	737 (73.7)	–
		LIML: BIC	–	331 (33.1)	366 (36.6)	–
	(3) Strong correlation and Strong IV	**2SRI: AIC**	1st	881 (88.1)	1000 (100)	894 (89.4)
			2nd	493 (49.3)	894 (89.4)	
		2SRI: BIC	1st	992 (99.2)	997 (99.7)	532 (53.2)
			2nd	444 (44.4)	533 (53.3)	
		LIML: AIC	–	452 (45.2)	887 (88.7)	–
		LIML: BIC	–	451 (45.1)	518 (51.8)	–
Samplesize	**Situation**	**Method**	Step	True model n (%)	Including true model n (%)	Both true model n (%)
$N = 300$	(1) Weak correlation and Strong IV	**2SRI: AIC**	1st	872 (87.2)	1000 (100)	1000 (100)
			2nd	672 (67.2)	1000 (100)	
		2SRI: BIC	1st	1000 (100)	1000 (100)	881 (88.1)
			2nd	846 (84.6)	881 (88.1)	
		LIML: AIC	–	588 (58.8)	1000 (100)	–
		LIML: BIC	–	846 (84.6)	880 (88.0)	–
	(2) Weak correlation and Weak IV	**2SRI: AIC**	1st	872 (87.2)	1000 (100)	998 (99.8)
			2nd	675(67.5)	998 (99.8)	
		2SRI: BIC	1st	1000 (100)	1000 (100)	935 (93.5)
			2nd	897 (89.7)	935 (93.5)	
		LIML: AIC	–	582 (58.2)	987 (98.7)	–
		LIML: BIC	–	881 (88.1)	921 (92.1)	–
	(3) Strong correlation and Strong IV	**2SRI: AIC**	1st	876 (87.6)	1000 (100)	999 (99.9)
			2nd	617 (61.7)	999 (99.9)	
		2SRI: BIC	1st	997 (99.7)	1000 (100)	953 (95.3)
			2nd	894 (89.4)	953 (95.3)	
		LIML: AIC	–	553 (55.3)	999 (99.9)	–
		LIML: BIC	–	900 (90.0)	951 (95.1)	–

high. In (2) and (3), these correspond to the weak IV and the strongly correlated unmeasured covariate situations, and labelled as “(2) Weak correlation and Weak IV” and “(3) Strong correlation and Strong IV” respectively. The selection probabilities of both “True model” and “Including true model” are somewhat different from (1); however, these are no remarkable difference from the results of model selection only. Therefore, it can be seen that model selections have stable results regardless of the situation and model selection procedure.

The estimated coefficients of treatment W in the outcome model are summarized in Table 2, Figs. 1, and 2, where **2SLS** is 2SLS (without model selection), **2SRI: Full model** is 2SRI with the largest model among the candidates, and **LIML: Full model** is LIML with the largest model among the candidates in the table; the red line denotes the true value in the figure.

- With model selection vs. without model selection Comparing the results with and without model selection procedures, it is appeared that the estimates with model selection procedures are more efficient and more unbiased; especially when there are large samples. In particular, the results of 2SRI without model selection is unstable. From the results, using model selection procedure is important to estimate the causal effects correctly.
- 2SRI vs. LIML In (1), for both the small sample and large sample cases, the LIML estimator with BIC was the most efficient result among the three results with model selection procedures. The LIML estimator with AIC also worked well in the sense of the unbiased result. The 2SRI estimator displayed a large bias and low efficiency; however, both results improved for large samples. In (2), surprisingly the LIML estimator yielded more accurate and unbiased results than (1) when there are only small samples. As mentioned in Burgess et al. (2017), the LIML estimator is more robust than any other well-known IV methods under weak IV situations. However, the simulation result is notable since the weak IV results are more accurate than (1). Table 1 shows that the model selection probabilities of both “True model” and “Including true model” are smaller than the other situations; more simple and accurate models tend to be selected over the true model. Therefore, the LIML estimator can derive the causal effects correctly using model selection procedures. Whereas, the 2SRI estimator suffers from the weak IV problems (c.f. Burgess et al. 2017). In (3), the LIML estimator yielded results similar to (1). However, the 2SRI estimator with and without model selection displayed a large bias and low efficiency for small samples. In the large samples, the efficiency somewhat improved; however, one important point is that 2SRI is still biased. These results are similar to those reported by Basu et al. (2017) and Wan et al. (2018). It is derived from the model construction of 2SRI that the residual term included in 2nd step is assumed as fixed covariate (see section 2.2 also). Actually, the bias is included in the all results of 2SRI. In particular, the situation where there are strongly correlated unmeasured covariates derives large bias (c.f. Wan et al. 2018).
- 2SLS vs. 2SRI & LIML The 2SLS method had some bias and instability compared with other methods. As there is no linear relationship between treatment

Table 2 Summary of descriptive statistics for each estimator (coefficients of W, continuous treatment and normal unmeasured covariates)

Sample size	Situation	Method	Mean (SD)	Median (Range)	Bias	RMSE
$N = 100$	(1) Weak	2SLS	0.379 (3.358)	0.375 (-93.56,23.35)	− 0.221	3.366
	correlation	2SRI: AIC	4.022 (48.711)	0.751 (-174.43, 1411.34)	3.422	48.831
	and	2SRI: BIC	1.212 (11.163)	0.700 (−174.43, 216.76)	0.612	11.180
	Strong IV	**LIML: AIC**	1.409 (9.654)	0.652 (−0.97, 228.93)	0.809	9.688
		LIML: BIC	0.899 (4.402)	0.634 (−0.72, 131.56)	0.299	4.412
		2SRI:Full model	≫10000 (≫10000)	0.873 (−91.93, ≫10000)	≫10000	≫10000
		LIML:Full model	0.943 (1.893)	0.739 (−0.94, 30.89)	0.343	1.924
	(2) Weak	**2SLS**	1916.803 (≫10000)	0.573 (≪−10000, ≫10000)	1916.203	≫10000
	correlation	**2SRI: AIC**	3.628 (102.496)	0.751 (−1717.59, 2514.4)	3.028	102.541
	and	**2SRI: BIC**	2.369 (30.028)	0.703 (−258.51, 616.05)	1.769	30.080
	Weak IV	**LIML: AIC**	0.655 (0.467)	0.662 (−0.80, 1.94)	0.055	0.470
		LIML: BIC	0.646 (0.316)	0.631 (−0.75, 1.91)	0.046	0.319
		2SRI:Full model	11.960 (270.608)	1.189 (−2228.26, 7771.30)	11.360	270.846
		LIML:Full model	0.775 (3.332)	0.875 (−0.80, 95.64)	0.175	3.337
	(3) Strong	**2SLS**	0.431 (1.263)	0.328 (−7.72, 24.63)	−0.169	1.274
	correlation	**2SRI: AIC**	6.968 (48.441)	0.971 (−191.22, 1154.13)	6.368	48.858
	and	**2SRI: BIC**	3.989 (26.372)	0.891 (−85.8, 591.93)	3.389	26.589
	Strong IV	**LIML: AIC**	5.588 (46.558)	0.692 (−0.60, 914.37)	4.988	46.824
		LIML: BIC	2.146 (23.893)	0.669 (−0.6, 672.81)	1.546	23.943
		2SRI:Full model	≫10000 (≫10000)	1.126 (−191.22, ≫10000)	≫10000	≫10000
		LIMLE:Full model	3.043 (16.018)	0.779 (−0.54, 289.51)	2.443	16.203
Samplesize	Situation	Method	Mean (SD)	Median (Range)	Bias	RMSE
$N = 300$	(1) Weak	**2SLS**	0.396 (0.228)	0.376 (−0.50, 1.74)	−0.204	0.306
	correlation	**2SRI: AIC**	0.673 (0.268)	0.666 (−0.28, 1.83)	0.073	0.278
	ans	**2SRI: BIC**	0.668 (0.183)	0.662 (−0.33, 1.83)	0.068	0.195
	Strong IV	**LIML: AIC**	0.625 (0.264)	0.618 (−0.21, 1.59)	0.025	0.265

Table 2 (continued)

Samplesize	Situation	Method	Mean (SD)	Median (Range)	Bias	RMSE
		LIML: BIC	0.626 (0.178)	0.617 (−0.21, 1.58)	0.026	0.180
		2SRI:Full model	0.703 (0.322)	0.704 (−0.32, 1.98)	0.103	0.338
		LIMLE:Full model	0.651 (0.326)	0.658 (−0.22, 1.62)	0.051	0.330
	(2) Weak correlation and Weak IV	**2SLS**	670.576 (≫10000)	1.317 (≪ −10000, ≫10000)	669.976	≫10000
		2SRI: AIC	1.298 (18.004)	0.656 (−60.84, 554.97)	0.698	18.010
		2SRI: BIC	0.500 (4.532)	0.646 (−134.15, 34.32)	−0.100	4.533
		LIML: AIC	0.592 (0.433)	0.611 (−0.71, 1.65)	−0.008	0.434
		LIML: BIC	0.616 (0.196)	0.605 (−0.67, 1.48)	0.016	0.197
		2SRI: Full model	2.431 (38.012)	0.851 (−60.89, 1046.45)	1.831	38.056
		LIMLE:Full model	0.530 (0.788)	0.691 (−0.76, 2.20)	−0.070	0.791
	(3) Strong correlation and Strong IV	**2SLS**	0.357 (0.211)	0.348 (−0.90, 1.38)	−0.243	0.322
		2SRI: AIC	0.824 (0.334)	0.807 (−0.41, 2.80)	0.224	0.402
		2SRI: BIC	0.809 (0.246)	0.788 (−0.41,2.07)	0.209	0.323
		LIML: AIC	0.650 (0.315)	0.618 (−0.41, 2.57)	0.050	0.319
		LIML: BIC	0.634 (0.216)	0.617 (−0.19, 1.91)	0.034	0.219
		2SRI: Full model	0.848 (0.391)	0.853 (−0.53, 2.80)	0.248	0.463
		LIMLE:Full model	0.675 (0.384)	0.654 (−0.39, 2.57)	0.075	0.391

and outcome, the results are natural. Therefore, we need to pay attention carefully when using 2SLS.

Thus, a good choice is to consider using LIML with model selection procedures; however, the best model selection procedure depends on the specific case, as mentioned in the Introduction. Whereas, overcoming weak IV problems is an advantage of LIML with model selection because other well-known IV methods do not have this feature. 2SRI with model selection can be selected for “valid" causal relationships, however, biased or unstable estimates may be obtained when there are only weak IVs, or there are strong unobserved relationships between treatments and outcomes.

4 Data analysis

In this section, the real data analysis is performed using 2SRI and LIML with model selection procedures. From the simulation results, the difference between the AIC and BIC is a little under large samples. Therefore, the AIC is used to select the valid model.

4.1 Analysis plan

The analysis follows the flow outlined below:

1. Detecting the genetic information used as instrumental variables. According to Cheng et al. (2019), there are 52 SNPs related to BMI; however, only 19 SNPs are included in the GENEVA Diabetes Study datasets. Since SNP is a weak instrument variable (weak correlation with a treatment variable), as many SNPs as possible should be used to increase the efficiency of the estimation.
2. Detecting the risk factors related to incidence of diabetes. According to Chen et al. (2018) and Narayan et al. (2007), age and sex are two important risk factors. In addition, both factors may have interaction effects on the incidence of diabetes. Therefore, a candidate model with interaction terms must be included.
3. Detecting the “valid” model using the AIC and estimating the causal effect. To select a treatment model and an outcome model, candidate models were prepared and the AIC was used to select the model. The supplemental information is provided in the Appendix B.

Age categorization was considered as follows:

- Age categories If $Age < 50$, then age was coded as $\}\}0''$; otherwise, if $50 \leq Age < 60$, then age was coded as $\}\}1''$; otherwise, if $60 \leq Age < 70$, then age was coded as “2"; otherwise, age was coded as $\}\}3''$.

There were 5481 subjects with either demographic or genetic data. In this study, a complete case analysis was conducted; subjects who had no missing data in the sex, age, BMI, and genetic data categories, were included in the analysis. Consequently, 5,036 subjects ($100 \times 5036/5481 = 91.9\%$) were included in the analysis. Note that BMI and SNPs are treated as continuous variables in the following analysis.

4.2 Analysis results

First, the participants' demographic data were confirmed. Table 3 summarizes the mean (SD) for continuous parameters and the number of subjects (%) for categorical parameters.

Regarding demographic data, there were some differences between the BMI categories. Therefore, there are concerns regarding the confounding effects of age and sex. In addition, the incidence of diabetes is different.

Table 4 summarizes the correlations between the two parameters. The correlation of SNP (instrumental variable) with BMI (a treatment variable) was quite small, raising a concern about the weak IV problem, as expected.

The estimated causal effects are summarized in Table 5. The result of the logistic regression (“Naive,” without using SNPs) has some obvious biases. The 2SRI estimation may provide a somewhat questionable result compared with the results of Hu et al. (2001) (risk ratio: 2.67). This is from the results that the 2SRI estimates are unstable similar to the simulation results. Regarding LIML, however, this result may be more

Table 3 Demographic data

Parameters	Category	Group		Total
		BMI < 30	BMI ≥ 30	
		N = 3909	N = 1127	N = 5036
BMI (continuous)	–	–	–	26.98 (4.88)
BMI (category)	BMI < 30	–	–	3909 (77.6)
	BMI ≥ 30	–	–	1127 (22.4)
Age (continuous)	–	57.85 (7.83)	55.78(7.28)	57.39 (7.76)
Age (category)	Age < 50	776 (19.9)	289 (25.6)	1065 (21.1)
	50 ≤ Age < 60	1454 (37.2)	476 (42.2)	1930 (38.3)
	60 ≤ Age < 70	1464 (37.5)	338 (30.0)	1802 (35.8)
	70 ≤Age	215 (5.5)	24 (2.1)	239 (4.7)
Sex	Male	2007 (51.3)	359 (31.9)	2366 (47.0)
	Female	1902 (48.7)	768 (68.1)	2670(53.0)
Diabetes	Yes	1496 (38.3)	833 (73.9)	2329(46.2)
	No	2413 (61.7)	294 (26.1)	2707 (53.8)

Table 4 Association of genetic variants with BMI and diabetes

SNP	Chr	Gene	BMI beta (SE)	Diabetesbeta(SE)	Effect/Other allele	Effect allele frequency (%)	*p*-value of HW-test
rs543874	1	*SEC16B, LINC01741*	0.129 (0.123)	0.002 (0.013)	G/A	18.9	0.698
rs2820292	1	*NAV1, IPO9-AS1*	−0.030 (0.098)	0.000 (0.010)	C/A	53.9	0.319
rs10182181	2	*ADCY3, DNAJC27*	0.072 (0.098)	0.003 (0.010)	G/A	46.7	0.424
rs2121279	2	*TMEM163*	0.147 (0.148)	0.018 (0.015)	T/C	12.5	0.750
rs7599312	2	*ERBB4*	0.131 (0.110)	0.013 (0.011)	G/A	73.9	0.740
rs492400	2	*USP37*	−0.054 (0.098)	−0.011 (0.010)	C/T	43.0	0.498
rs13107325	4	*SLC39A8*	−0.038 (0.186)	0.006 (0.019)	T/C	7.6	0.963
rs11727676	4	*HHIP*	0.028 (0.167)	−0.013 (0.017)	T/C	90.4	0.895
rs13191362	6	*PRKN*	0.052 (0.149)	0.009 (0.015)	A/G	88.3	0.976
rs1167827	7	*HIP1*	0.028 (0.097)	−0.007 (0.010)	G/A	57.5	0.830
rs6477694	9	*EPB41L4B, FRRS1L*	0.133 (0.103)	−0.010 (0.010)	C/T	34.4	0.797
rs1928295	9	*TRPL35AP22*	0.171 (0.097)	0.006 (0.010)	T/C	54.5	0.276
rs10733682	9	*LMX1B*	0.119 (0.096)	0.011 (0.010)	A/G	49.0	0.865
rs7899106	10	*GRID1*	0.254 (0.223)	−0.017 (0.023)	G/A	5.0	0.189
rs11030104	11	*BDNF, BDNF-AS*	0.292 (0.120)	−0.005 (0.012)	A/G	79.1	0.133
rs12286929	11	*CADM1*	0.185 (0.097)	0.014 (0.010)	G/A	53.7	0.769
rs3736485	15	*DMXL2*	0.169 (0.098)	0.016 (0.010)	A/G	46.4	0.389
rs7239883	18	*LINC00907*	−0.061 (0.100)	0.011 (0.010)	G/A	39.6	0.237
rs2836754	21	*LINC01700, RPSAP64*	0.146 (0.102)	−0.008 (0.010)	C/T	62.8	0.975

Note 1: Relationship between BMI/Diabetes and SNPs is summarized as regression coefficients (SE).

Note 2: For BMI, the ordinary linear model is applied. For Diabetes, the logistic regression model is applied

Table 5 Summary of estimates of the causal effects (point estimates (95%CI))

	2SRI	LIML	Naive
Risk ratio	0.989 (0.907, 8.486)	1.896 (1.027, 2.407)	2.634 (1.948, 3.907)

Note1: Causal effects are estimated by G-computation (e.g., Hernán and Robins 2020).

Note2: The results are derived as difference from $BMI = 18.5$ to $BMI = 25$.**Note3:** 95%CI is derived from a bootstrap method. The sample size of bootstrap sample is 500, and the number of iteration is 1,000

Table 6 Summary of estimates of the causal effects by each sex (point estimates (95%CI))

	Male		Female	
	2SRI	LIML	2SRI	LIML
Risk ratio	2.952 (0.062, 139.543)	2.969 (2.921, 3.017)	2.397 (0.239, 24.006)	1.574 (1.552, 1.596)

Note1: The results are derived as difference from $BMI = 18.5$ to $BMI = 25$

Note2: 95%CI is derived from a normal approximation

plausible since it is less sensitive to weak IV problems as shown in the simulation results. The estimated causal effects for each sex are also summarized in Table 5. Since the cohorts of males and females are different, a supplemental analysis was planned. Unfortunately, the estimates become more unstable since there are only small sample size (male: 2366 and female: 2670). Regarding 2SRI, the results are also unstable. Whereas, the results are the same direction of the causal effect as the main result. From the results in Table 6, the main result (Table 5) is quite reasonable.

From the viewpoint of model selection, "Model 51" was selected for 2SRI and LIML (see also Appendix). According to Chen (2018), there are interactions between age and BMI; however, the interaction model was not selected. From the results of Chen (2018), younger subjects may display stronger interaction effects than older subjects, whereas our data included only subjects aged 40 years or older. Therefore, the interaction term may not be selected.

The above analyses have some limitations. First, as mentioned previously, only 19 SNPs were used in our data analysis; thus, there may be some concerns about the weak IV problem. Cheng et al. (2019) used 52 SNPs; however, a critical limitation of this study is that only 19 SNPs were used in the analysis. Second, the sample size was limited for the dbGaP data. To overcome the weak IV problem, a large sample size is necessary for Mendelian randomization (Burgess et al. 2017). Therefore, the derived result may be inefficient and requires care when interpreting the results.

5 Conclusion and future work

In this study, a binary outcome model with unmeasured covariates was considered. Two-stage residual inclusion (2SRI) is applied in this situation; however, some biased estimates may be derived (Basu et al. 2017). Therefore, limited-information maximum likelihood (LIML), which has features similar to those of 2SRI, was also considered in this study. Since model selections are important to estimate unbiased causal effects, the AIC and BIC for 2SRI and LIML are considered in this study. From the simulation results, LIML with the AIC or BIC works well compared with using full models when an unmeasured covariate distribution is specified correctly, especially, overcoming the weak IV problem is an advantage of LIML. In contrast, 2SRI may derive biased or unstable estimates when there are only weak IVs or strong unobserved relationships between treatments and outcomes. 2SRI and LIML with the AIC were applied to the GENEVA Diabetes Study as Mendelian randomization. The results show that the causal effects are similar to those of previous research; however, there may be some concern about weak IV problems. From the above, we recommend that LIML with any model selection procedures is a good choice when there are binary outcomes and any concerns about unmeasured covariates.

As mentioned, the results are significant contributions in cases of unmeasured covariates and nonlinear outcomes because there has been no research on model selection procedures when both the true treatment model and the true outcome model need to be specified. However, several future studies should be conducted. First, only a binary outcome was considered in this study. Because 2SRI considers a likelihood in the 2nd step and LIML considers a full-likelihood, the method can be expanded to more complex models, for instance, a more general outcome of an exponential family or a time-to-event outcome (Kianian et al. 2019 and Martínez-Camblor et al. 2019). In particular, LIML needs to consider likelihoods of both the outcome and treatment variables; however, the other restrictions are limited. For instance, LIML is not restricted to binary instrumental variables (Wang and Tchetgen 2018 and Kianian et al. 2019) or continuous treatment (Martínez-Camblor et al. 2019). Therefore, LIML with any model selection procedures has great potential expandability. Next, the impact of the misspecification of an unmeasured covariate distribution needs to be carefully confirmed. As the simulation results in Web Appendix C show, the impact may be limited; however, the estimation behavior in other cases is not clear. Therefore, it is necessary to continue with simulations to consider more varied situations.

A Supplementary information for simulations

Data generating programs, simulation datasets, simulation programs, simulation results, and programs for deriving tables and figures are available at the following URL:

- https://drive.google.com/file/d/17nZla3cQDYTvka-260Ib2Qc9XUjn3B5o/view?usp=sharing

Table 7 Settings of candidates for treatment and outcome models

Candidate model	α_0	α_1	α_2	α_3	α_4	α_5	α_6
Modela1	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$	$= 0$	$= 0$
Modela2	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$	$= 0$
Modela3	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Modela4	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Modela5	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$
Modela6	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$
Modela7	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$

Candidate model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Model b1	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$	$= 0$	$= 0$	$= 0$
Model b2	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$	$= 0$
Model b3	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Model b4	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$
Model b5	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$

Note that Model a4 and Model b2 represent the true models. Models a1-a3 a5 and a6, and Model b1 are misspecified models. For 2SRI, a residual term is also included in the 2nd model

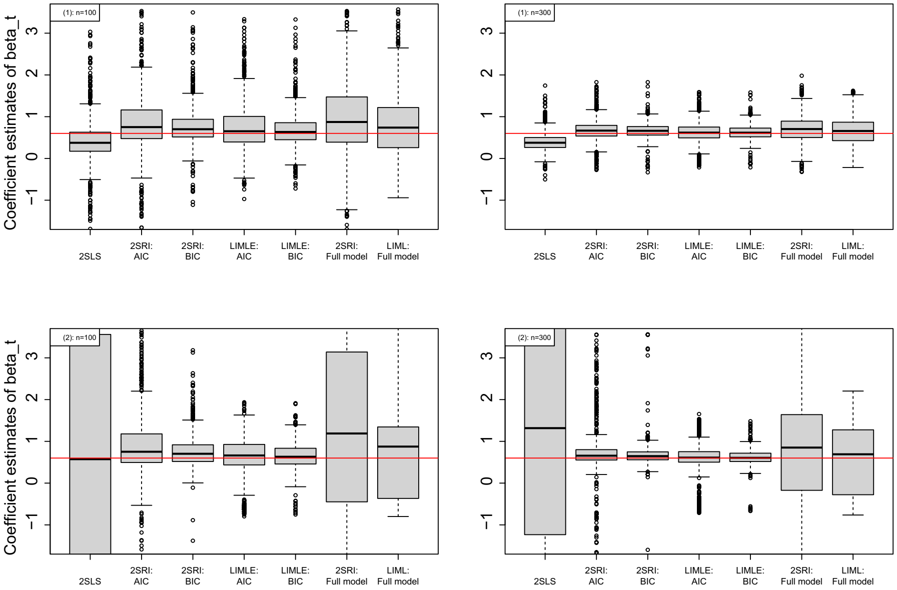

Fig. 1 Boxplots of descriptive statistics for each estimator (continuous treatment and normal unmeasured covariates) 1/2

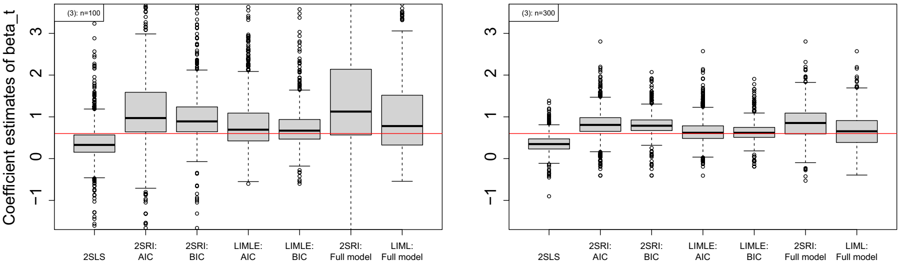

Fig. 2 Boxplots of descriptive statistics for each estimator (continuous treatment and normal unmeasured covariates) 2/2

A.1 Candidates of models

To select a treatment model and outcome model, the following candidate models are presented (Table 7):

Candidates for treatment models

$$W = \alpha_0 + z\alpha_1 + x_2\alpha_2 + x_3\alpha_3 + \left(z \times x_2\right)\alpha_4 + \left(z \times x_3\right)\alpha_5 + \left(x_2 \times x_3\right)\alpha_6 + V$$

Candidates for outcome models

$$Y = \mathbf{1}\left\{\beta_0 + w\beta_1 + x_1\beta_2 + x_2\beta_3 + x_3\beta_4 + \left(x_1 \times x_2\right)\beta_5 + \left(x_1 \times x_3\right)\beta_6 + \left(x_2 \times x_3\right)\beta_7 + U \geq 0\right\}$$

A.2 Supplemental figures for continuous treatment and normal unmeasured covariates

B. Supplementary information for data analysis

B.1 Descriptions of GENEVA diabetes study datasets

The details of the GENEVA Diabetes Study are available at the following URL:

– https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1

There are two main genotype datasets and one phenotype dataset, named “phg000036v1,” “phg000048v1,” and “phenotype,” respectively. Note that the genotype datasets constitute one dataset per subject, and the data and subject IDs are connected through annotation files. The GENEVA Diabetes Study datasets are encrypted using the NCBI data encryption algorithm. To decode the encryption, the “SRA Toolkit” is required. The details of the data encryption are found at the following URL:

– https://www.ncbi.nlm.nih.gov/books/NBK570250/

B.2 Candidate models

See Table 8.

Candidates for treatment models

$$BMI = \alpha_0 + SNPs\,\alpha_1 + age\,\alpha_2 + sex\,\alpha_3 + (age \times sex)\alpha_4 + V$$

Candidates for outcome models

$$\begin{aligned} Diabetes = \mathbf{1}\big\{&\beta_0 + BMI\beta_{11} + BMI^2\beta_{12} + age\,\beta_2 + sex\,\beta_3 + (BMI \times age)\beta_4 \\ &+(BMI \times sex)\beta_5 + (age \times sex)\beta_6 + U \geq 0\big\} \end{aligned}$$

Table 8 Candidate settings for analysis models

Candidate model	α_0	α_1	α_2	α_3	α_4	β_0	β_{11}	β_{12}	β_2	β_3	β_4	β_5	β_6
Model 11	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$	$= 0$	$= 0$
Model 12	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$	$= 0$	$= 0$
Model 21	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Model 22	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Model 31	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Model 32	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$= 0$
Model 41	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$
Model 42	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	$= 0$
Model 51	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$
Model 52	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$	$\neq 0$	$= 0$
Model 61	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$
Model 62	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$
Model 71	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$
Model 72	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41237-022-00177-9.

Acknowledgements We would like to express our grateful thanks to editors for their useful comments. This manuscript is sophisticated by their comments, and becomes more useful. This work was supported by JSPS KAKENHI Grant Number JP21K10500. Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract "High-throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). The datasets used for the analyses described in this manuscript were obtained from dbGaP from http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v1.p1 through dbGaP accession number phs000091.v1.p.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Baiocchi M, Cheng J, Small DS (2014) Instrumental variable methods for causal inference. Stat Med 33(13):2297–2340

Basu A, Coe N, Chapman CG (2017) Comparing 2SLS VS 2SRI for binary outcomes and binary exposures (No. w23840). National Bureau of Economic Research

Biller B, Corlu CG (2012) Copula-based multivariate input modeling. Surv Oper Res Manag Sci 17(2):69–84

Brookhart MA, van der Laan MJ (2006) A semiparametric model selection criterion with applications to the marginal structural model. Comput Stat Data Anal 50(2):475–498

Burgess S, Small DS, Thompson SG (2017) A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res 26(5):2333–2355

Chen Y et al (2018) Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study. BMJ Open 8(9):e021768

Cheng L, Zhuang H, Ju H, Yang S, Han J, Tan R, Hu Y (2019) Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study. Front Genet 10:94

Fantazzini D (2009) The effects of misspecified marginals and copulas on computing the value at risk: a Monte Carlo study. Comput Stat Data Anal 53(6):2168–2188

Hernán MA, Robins JM (2020) Causal inference: what if. Chapman & Hill/CRC, New York

Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med 345(11):790–797

Kianian B, Kim JI, Fine JP, Peng L (2019) Causal proportional hazards estimation with a binary instrumental variable. arXiv:1901.11050

Martínez-Camblor P, Mackenzie T, Staiger DO, Goodney PP, O'Malley AJ (2019) Adjusting for bias introduced by instrumental variable estimation in the Cox proportional hazards model. Biostatistics 20(1):80–96

Narayan KV, Boyle JP, Thompson TJ, Gregg EW, Williamson DF (2007) Effect of BMI on lifetime risk for diabetes in the US. Diabetes Care 30(6):1562–1566

Nishii R (1984) Asymptotic properties of criteria for selection of variables in multiple regression. Ann Stat, 758–765

Richardson TG, Sanderson E, Elsworth B, Tilling K, Smith GD (2020) Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. bmj, 369

Rivers D, Vuong QH (1988) Limited information estimators and exogeneity tests for simultaneous probit models. J Econ 39(3):347–366

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Shao J (1997) An asymptotic theory for linear model selection. Stat Sin 221–242

Taguri M, Matsuyama Y, Ohashi Y (2014) Model selection criterion for causal parameters in structural mean models based on a quasi-likelihood. Biometrics 70(3):721–730

Tchetgen EJT, Walter S, Vansteelandt S, Martinussen T, Glymour M (2015) Instrumental variable estimation in a survival context. Epidemiology (Camb, MA) **26**(3):402

Terza JV, Basu A, Rathouz PJ (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ 27(3):531–543

Vansteelandt S, Bekaert M, Claeskens G (2012) On model selection and model misspecification in causal inference. Stat Methods Med Res 21(1):7–30

Wan F, Small D, Mitra N (2018) A general approach to evaluating the bias of 2-stage instrumental variable estimators. Stat Med 37(12):1997–2015

Wang L, Tchetgen ET (2018) Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. J R Stat Soc Ser B Stat Methodol **80**(3):531

Wooldridge JM (2010) Econometric analysis of cross section and panel data. MIT Press, New York

Wooldridge JM (2014) Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. J Econ 182(1):226–234

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.