Penalized Semiparametric Estimation for Causal Inference with Possibly Invalid Instruments

Yunlong Cao[1], Yuquan Wang[1], Dapeng Shi[2], Dong Chen[1], and Yue-Qing Hu[1,2,*]

[1]Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China

[2]Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China

**email:* yuehu@fudan.edu.cn

SUMMARY: Inferring causal effects with unmeasured confounder is a main challenge in causal inference. Many researchers impose parametric assumptions on the distribution of unmeasured confounder. However, due to the unobservable nature of the unmeasured confounder, it is more reasonable to leave its distribution unrestricted. Another key challenge in causal inference is the involvement of invalid instrumental variables, which may lead to biased inference and possibly misleading scientific conclusions. To this end, we employ a flexible semiparametric model that allows for possibly invalid instruments without specifying the distribution of unmeasured confounder in this work. A penalized semiparametric estimator for causal effects is constructed and its oracle and asymptotic properties are well established for statistical inference. We evaluate the performance of the estimator through simulation studies, revealing that our proposed estimator exhibits asymptotic unbiasedness and robustness in estimating causal effects, along with consistent selection of invalid instruments. We also demonstrate its application using Atherosclerosis Risk in Communities Study data set, which further validates its robustness in the presence of invalid instruments. Additionally, we have implemented the proposed method in R, and the corresponding R code is available for free download.

KEY WORDS: Causal inference; Invalid instruments; Majority rule; Semiparametric estimating equations; SCAD penalty; Unmeasured confounder.

1. Introduction

Causal inference is vital for elucidating cause-and-effect relationships. For observational data, it is rather difficult to make inference due to the existence of unmeasured confounders. Instrumental variables (IV) stand out as a widely used technique for detecting causality and estimating the causal effect of an exposure on an outcome in this situation. A valid IV which is suitable for estimating causal effects, must adhere to three fundamental assumptions (Angrist, 1996; Sargan, 1958), namely:

(1) Relevance: the IV is related to the exposure;

(2) Exchangeability: the IV is independent of unmeasured confounders;

(3) Exclusion Restriction: the IV has no direct effect on the outcome.

The relevance of instruments can be scrutinized through observed data of exposure and instruments. However, checking assumptions (2) and (3) in a data-dependent manner necessitates substantial domain expertise to discern valid IVs.

In certain instances, causal effects can be deduced even with the presence of invalid IVs. In the context of a linear outcome model, Kolesár et al. (2015) and Bowden et al. (2015) provided solutions wherein all candidate IVs may be valid, but the strength of the IV and its direct effect on the outcome are nearly orthogonal. Kang et al. (2016) and Windmeijer et al. (2018) put forth consistent estimators for causal effects, assuming the majority rule that at least 50% of the IVs are valid. Li and Guo (2020) extended the majority rule to nonlinear outcome models, presenting the three-step inference procedure SpotIV for estimating the conditional average treatment effect (CATE).

Semiparametric methodologies are extensively employed in causal inference. Sun et al. (2023) introduced a class of g-estimators guaranteed to maintain consistency and asymptotic normality in estimating the causal effect of interest, even in the presence of invalid instrumental variables. Zhang and Tchetgen Tchetgen (2022) proposed a robust estimator reaching the

efficiency bound for the semiparametric model, without imposing parametric assumptions on the unmeasured confounder. They established the consistency and asymptotic normality of the estimator under appropriate identification and regularity conditions. However, a general identification condition for the semiparametric model is not explicitly stated.

Considering the assumption of majority rule in semiparametric model setting, the penalized semiparametric estimating approaches were developed to estimate causal effects. Diverging from various penalties imposed on the loss function in the conventional parametric models, the semiparametric approach to estimating causal effects does not involve minimizing any objective function. Fu (2003) proposed penalizing the estimating function, instead of the loss function, for generalized linear models with a bridge penalty (Frank and Friedman, 1993; Fu and Knight, 2000). Subsequently, Johnson et al. (2008) presented a comprehensive asymptotic properties of estimators derived from a broad class of penalized estimating functions.

Given semiparametric model setting and majority rule, we explore the penalized semiparametric estimating method to simultaneously estimate causal effects and select invalid instrumental variables in this work. The article is organized as follows. Section Methods serves to introduce our model setting and discuss the identifiability of model. Subsequently, we present the semiparametric estimating equations (SEE) for the model and introduce the penalized semiparametric estimating equations (PSEE). Section Implementation and Results establishes the algorithm and asymptotic theory for PSEE. In Section Simulation Study, numerical results from simulation studies are presented. Moving to Section Real Data Analysis, we apply the PSEE method to the Atherosclerosis Risk in Communities Study (ARIC) dataset. Section Discussions is dedicated to providing some discussions.

2. Methods

2.1 *Semiparametric model setting*

Consider the causal effect of an exposure $D \in \mathbb{R}$ and outcome $Y \in \mathbb{R}$. $\mathbf{Z} \in \mathbb{R}^q$ denote the q-dimensional vector of instrumental variables for inferring the causality, $U \in \mathbb{R}$ is a scalar unmeasured confounder.

We consider the following outcome model,

$$\mathbb{E}(Y \mid D = d, \mathbf{Z} = \mathbf{z}, U = u) = g_1(d\beta + \mathbf{z}^T\alpha + c_1 u). \tag{1}$$

For the exposure D, we consider

$$\mathbb{E}(D \mid \mathbf{Z} = \mathbf{z}, U = u) = g_2(\mathbf{z}^T\gamma + c_2 u), \tag{2}$$

where g_1, g_2 are link functions, $\beta \in \mathbb{R}$ represents the causal parameter of interest, $\alpha \in \mathbb{R}^q$ and $\gamma \in \mathbb{R}^q$ represent the direct effect of the instruments on the outcome and exposure, respectively, c_1, c_2 are fixed sensitivity parameters used to adjust the influence of confounder on the outcome. Let $\theta = (\gamma^T, \beta, \alpha^T)^T \in \mathbb{R}^p$ denote the finite dimensional parameters with $p = 2q + 1$.

In many applications (Harbord et al., 2013), the distribution of U is often assumed to follow a parametric distribution, such as the normal distribution in Shi et al. (2023). Since U is unobservable, it may be more appropriate to refrain from imposing any parametric assumptions on its distribution. Therefore we are exploring a semiparametric model in which both the outcome model and exposure model are accurately specified, as denoted by Equations (1) and (2), respectively. The joint distribution of $(U, \mathbf{Z})$ in this model remains unrestricted.

Note that the model (1)-(2) is highly versatile, encompassing linear and nonlinear outcome as well as exposure as special cases. For example, when g_1 is identity function, it includes

continuous outcome as

$$Y = D\beta + \mathbf{Z}^T\alpha + c_1 U + \epsilon, U \perp \epsilon, \epsilon \sim N(0,1); \tag{3}$$

when g_1 is standard logistic function, it includes binary outcome as

$$Y \mid D = d, \mathbf{Z} = \mathbf{z}, U = u \sim \text{Bernoulli}(\text{logit}^{-1}(d\beta + \mathbf{z}^T\alpha + c_1 u)).$$

The exposure model (2) is similar, it includes continuous and binary exposure depending on g_2.

2.2 *Identifiability of Model*

The presence of direct effects of instruments on the outcome poses a significant challenge for instrumental variable inference. Previous studies have delved into the identifiability conditions within some models, as discussed by Kang et al. (2016) and Li and Guo (2020). This section provides an overview of these works, laying the foundation for the subsequent discussion of our estimating method for the model parameters.

Let $s = \|\alpha\|_0$ denote the number of invalid instruments, i.e. the number of nonzero components of α. For a continuous outcome, Kang et al. (2016) proved the identifiability of the model under the condition of $s < q/2$, known as the *majority rule*. Additionally, they provided a method sisVIVE for estimating the causal effect, which utilizes l_1 penalization on α. Li and Guo (2020) studied the nonlinear causal inference and proposed a method SpotIV to estimate the CATE.

Therefore in this article, we assume that majority rule holds and our model (1)-(2) is identifiable under majority rule. In particular, when outcome Y is continuous and Equation (3) holds, and $\mathbb{E}(U|\mathbf{Z}) = 0$, then our model is actually identifiable under majority rule according to Kang et al. (2016); similarly when exposure D is continuous and $\mathbb{E}(U|\mathbf{Z}) = 0$, our model is identifiable according to Li and Guo (2020).

2.3 *Semiparametric Estimating Equation*

In standard semiparametric theory, we only consider estimators that are regular and asymptotically linear (RAL) (Newey, 1990; Bickel et al., 1993; Van der Vaart, 2000; Tsiatis, 2006). Note that the full data $\mathcal{F} = \{F_i = (Y_i, D_i, \mathbf{Z}_i, U_i), i = 1, 2, \cdots, n\}$, the observed data only consist of $\mathcal{O} = \{O_i = (Y_i, D_i, \mathbf{Z}_i), i = 1, 2, \cdots, n\}$ as U is not observed. An asymptotically linear estimator $\hat{\theta}_n$ of model parameter θ based on the full data satisfies

$$n^{1/2}(\hat{\theta}_n - \theta) = n^{-1/2} \sum_{i=1}^{n} \varphi(F_i) + o_p(1),$$

where the measurable random function $\varphi(F_i)$ is referred to as the i-th influence function of the estimator $\hat{\theta}_n$ and satisfies $E\{\varphi(F)\} = 0$, $E(\varphi\varphi^T)$ is finite and nonsingular. Regularity conditions are imposed to exclude super-efficient estimators, which are unnatural and have undesirable local properties.

Any RAL estimator is asymptotically normally distributed; i.e.,

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, E(\varphi\varphi^T)).$$

We hope to find the efficient influence function $\varphi_{\text{eff}}(F)$, which is the influence function with the smallest variance matrix in the sense that for any influence function $\varphi(F) \neq \varphi_{\text{eff}}(F)$, $\text{var}\{\varphi_{\text{eff}}(\text{F})\} - \text{var}\{\varphi(\text{F})\}$ is negative definite.

By standard semiparametric theory, the efficient influence function based on the full data, $\varphi_{\text{eff}}(F)$, is proportional to the full data efficient score $S_{\text{eff}}(Y, D, \mathbf{Z}, U)$, which can be obtained by projecting the full data score $S_\theta(Y, D, \mathbf{Z}, U)$ onto the orthogonal component of the full data nuisance tangent space Λ^F, which is given by $\Lambda^F = \Lambda_1^F \oplus \Lambda_2^F$, where

$$\Lambda_1^F = \{a_1(\mathbf{Z}) : \mathbb{E}\left[a_1(\mathbf{Z})\right] = 0\}$$

$$\Lambda_2^F = \{a_2(U, \mathbf{Z}) : \mathbb{E}\left[a_2(U, \mathbf{Z}) \mid \mathbf{Z}\right] = 0\}.$$

Accordingly, for observed data efficient score $S_{\text{eff}}(Y, D, \mathbf{Z})$, we need to calculate corresponding

observed data score $S_\theta(Y, D, \mathbf{Z}) = \mathbb{E}[S_\theta(Y, D, \mathbf{Z}, U) \mid Y, D, \mathbf{Z}]$ and the observed data nuisance tangent space $\Lambda = \mathbb{E}[\Lambda^F \mid Y, D, \mathbf{Z}]$.

Zhang and Tchetgen Tchetgen(2022) derived the specific form of the observed data efficient score for the considered model and introduced a *working model* $f^*(U \mid \mathbf{Z}; \xi)$ instead of the unknown $f(U \mid \mathbf{Z})$ to proceed calculation:

$$S^*_{\text{eff}}(Y, D, \mathbf{Z}) = \mathbb{E}_* \left[S^*_\theta(Y, D, \mathbf{Z}, U) \mid Y, D, \mathbf{Z}\right] - \mathbb{E}_*[a(U, \mathbf{Z}) \mid Y, D, \mathbf{Z}],$$

where $a(U, \mathbf{Z})$ satisfies the integral equation:

$$\mathbb{E}_* \left[S^*_\theta(Y, D, \mathbf{Z}) \mid U, \mathbf{Z}\right] = \mathbb{E}_* \left\{\mathbb{E}_*[a(U, \mathbf{Z}) \mid Y, D, \mathbf{Z}] \mid \mathbf{Z}, U\right\}.$$

Note that the true data distribution $\mathcal{P}_F$ can be factored as

$$\mathcal{P}_F = f(Y \mid D, \mathbf{Z}, U) f(D \mid \mathbf{Z}, U) f(U \mid \mathbf{Z}) f(\mathbf{Z}),$$

and the misspecified data distribution with working model is

$$\mathcal{P}^*_F = f(Y \mid D, \mathbf{Z}, U) f(D \mid \mathbf{Z}, U) f^*(U \mid \mathbf{Z}; \xi) f(\mathbf{Z}).$$

$\mathbb{E}[\cdot]$ denotes expectation taken with respect to $\mathcal{P}_F$ and $\mathbb{E}_*[\cdot]$ taken with respect to $\mathcal{P}^*_F$.

The efficient score $S^*_{\text{eff}}(Y, D, \mathbf{Z})$ has an important property:

$$\mathbb{E}[S^*_{\text{eff}}(Y, D, \mathbf{Z})] = 0. \tag{4}$$

Equation (4) yields an estimator for θ that exhibits appealing robustness and efficiency properties. This is achieved by substituting the expectation with its empirical analogue and formulating the following estimating equation:

$$\sum_{i=1}^{n} S^*_{\text{eff}}(Y_i, D_i, \mathbf{Z}_i; \theta) = 0. \tag{5}$$

Under suitable identification and regularity conditions, the solution $\theta = \hat{\theta}$ to the estimating

equation (5) is consistent and asymptotically normal, with variance given by

$$V = \frac{1}{n}\mathbb{E}\left\{\partial S^*_{\text{eff}}(O_i;\theta_0)/\partial\theta\right\}^{-1}\mathbb{E}\left\{S^*_{\text{eff}}(O_i;\theta_0)S^*_{\text{eff}}(O_i;\theta_0)^T\right\} \times \mathbb{E}\left\{\partial S^*_{\text{eff}}(O_i;\theta_0)/\partial\theta^T\right\}^{-1}. \tag{6}$$

If the conditional distribution $f(U|\mathbf{Z})$ is correctly specified, then $\hat{\theta}$ is locally efficient with asymptotic variance $V_{\text{eff}} = \mathbb{E}[S_{\text{eff}}S_{\text{eff}}^T]$.

2.4 *Penalized Semiparametric Estimating Equation*

Given estimating equation (5) and assumed identification condition, majority rule, we consider the penalized semiparametric estimating equation for simultaneous estimation and invalid instrumental variable selection. Specifically, the penalized semiparametric estimating functions are defined as

$$S^P(\theta) = S(\theta) - nq_\lambda(|\theta|)\,\text{sgn}(\theta),$$

where $S(\theta) = \sum_{i=1}^{n} S^*_{\text{eff}}(Y_i, D_i, \mathbf{Z}_i;\theta)$, $q_\lambda(|\theta|) = (q_{\lambda,1}(|\theta_1|), \cdots, q_{\lambda,p}(|\theta_p|))^T$ and the second term is the componentwise product of q_λ and $\text{sgn}(\theta)$. To select invalid instrumental variables, we design $q_\lambda(|\theta|)$ as follows: (i) for $j = 1, 2, \cdots, q+1$, set $q_{\lambda,j}(|\theta_j|) = 0$; (ii) for $j = q+2, \cdots, p$, set $q_{\lambda,j}(|\theta_j|)$ as SCAD penalty:

$$q_\lambda(|\theta_j|) = \lambda\left\{I(\theta_j \leqslant \lambda) + \frac{(a\lambda - \theta_j)_+}{(a-1)\lambda}I(\theta_j > \lambda)\right\}$$

with $a > 2$.

Here we adopt the nonconvex SCAD penalty proposed by Fan and Li (2001), which results in an estimator with oracle property: that is, the estimator has the same limiting distribution as an estimator that konws the true model a priori.

Note that we only penalize the latter q component of $S(\theta)$ which corresponding to parameter α, therefore γ and β will not be shrunken. Intuitively, $q_\lambda(|\alpha_j|)$ is zero for a large value of $|\alpha_j|$, while it increases significantly for a small value of $|\alpha_j|$. Consequently, the jth component of the semiparametric estimating function $S(\theta)$, denoted as $S_j(\theta)$, is not penalized

when $|\alpha_j|$ is large. Conversely, $S_j(\theta)$ is heavily penalized if $|\alpha_j|$ is close (but not equal) to zero, compelling the estimator of α_j to shrink to zero. When α_j is shrunk to zero, it implies that the jth instrumental variable is deemed valid and is consequently excluded from the outcome model.

3. Implementation and Results

3.1 *Algorithm*

To solve the penalized semiparametric estimating equation, we employ an iterative algorithm similar to that utilized in Johnson et al. (2008), Wang et al. (2012, 2013). This algorithm combines the minorization-maximization algorithm for the nonconvex penalty introduced by Hunter and Li (2005) with the Newton-Raphson algorithm:

$$\hat{\theta}^k = \hat{\theta}^{k-1} + \left[H(\hat{\theta}^{k-1}) + nE(\hat{\theta}^{k-1})\right]^{-1} S^P(\hat{\theta}^{k-1}).$$

where

$$H(\theta) = -\frac{1}{n}\frac{\partial S(\theta)}{\partial \theta},$$
$$E(\theta) = \mathrm{diag}\{\underbrace{0, \cdots 0}_{q+1}, \frac{q_{\lambda_n}(|\alpha_1|)}{\varepsilon + |\alpha_1|}, \cdots, \frac{q_{\lambda_n}(|\alpha_q|)}{\varepsilon + |\alpha_q|}\},$$

the constant ϵ represents a small perturbation, set to 10^{-6} in our simulation studies. We initialize the algorithm with the adaptive lasso estimator $\hat{\theta}^0$. For a chosen tuning parameter, the algorithm iterates until the convergence criterion $\|\hat{\theta}^{k+1} - \hat{\theta}^k\| \leqslant 10^{-4}$ is satisfied. Typically, this criterion is met within 20 iterations in our simulation studies. Additionally, any coefficient that becomes sufficiently small is constrained to zero; specifically, if $|\hat{\theta}_j| \leqslant 10^{-4}$ upon convergence, then the estimator for this coefficient is set to exactly zero.

We need to select (a, λ) for the SCAD penalty. Fan and Li (2001, 2002) demonstrated that the choice $a \equiv 3.7$ performs well across various scenarios, and we adopt this recommendation for our numerical analyses. In practice, cross-validation is a widely used data-driven method

for choosing λ. In the same vein, we employ k-fold cross-validation, minimizing the L_2 norm of the estimating equation $S(\theta)$. This approach aligns with the fact that the parameter of interest is θ, which sets the expected value of the estimating equation to zero (see Equation (4)).

We derive the following sandwich formula from the algorithm to estimate the asymptotic covariance matrix of $\hat{\theta}$:

$$\mathrm{Cov}(\hat{\theta}) \approx [H(\hat{\theta}) + nE(\hat{\theta})]^{-1} M(\hat{\theta}) [H(\hat{\theta}) + nE(\hat{\theta})]^{-1}$$

where $M(\hat{\theta}) = \sum_{i=1}^{n} \mathrm{S}^*_{\mathrm{eff}}(O_i; \hat{\theta})\, \mathrm{S}^*_{\mathrm{eff}}(O_i; \hat{\theta})^T$.

3.2 *Asymptotic Theory for Penalized Semiparametric Estimator*

Let $\theta_0 = (\gamma_0^T, \beta_0, \alpha_0^T)^T = (\gamma_0^T, \beta_0, \alpha_{10}^T, \alpha_{20}^T)^T := (\theta_{10}^T, \theta_{20}^T)^T$ denote the true value of θ, where $\alpha_{10} \in \mathbb{R}^s, \theta_{10} = (\gamma_0^T, \beta_0, \alpha_{10}^T)^T, \theta_{20} = \alpha_{20}$ and $\alpha_0 = (\alpha_{01}, \cdots, \alpha_{0q})^T$. Without loss of generality, suppose that $\theta_{0j} \neq 0$ for $j \leqslant q + 1 + s$ and $\theta_{0j} = 0$ for $j > q + 1 + s$. For the asymptotic theory, we require the following regularity conditions.

(a) $n^{-1} \sum_{i=1}^{n} \partial S^*_{\mathrm{eff}}(Y_i, D_i, \mathbf{Z}_i; \theta)/\partial\theta$ exists and is continuous in an open neighborhood of β_0;

(b) $n^{-1} \sum_{i=1}^{n} \partial S^*_{\mathrm{eff}}(Y_i, D_i, \mathbf{Z}_i; \theta)/\partial\theta$ converges uniformly to its limit in a neighborhood of θ_0;

(c) $\mathbb{E}\left\{\partial S^*_{\mathrm{eff}}(Y, D, \mathbf{Z}; \theta)/\partial\theta\right\}|_{\theta=\theta_0}$ is invertible;

(d) $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$;

Remark 1: Conditions (a)-(c) are imposed by Zhang and Tchetgen Tchetgen (2022) to ensure the consistency and asymptotic normality of the estimator derived from the semi-parametric estimating equation $S(\theta) = 0$. Condition (d) represents a standard requirement concerning the rate of the tuning parameter to attain the oracle property (Fan and Li, 2001).

Theorem 1: Assuming conditions (a)-(d), the following results hold:

a. There exists a root-n-consistent approximate solution of $S^P(\theta)$, $\hat{\theta} = \theta_0 + O_p(n^{-\frac{1}{2}})$, in the

sense that $n^{-\frac{1}{2}}S^P(\hat{\theta}) = O_p(1)$.

b. (Oracle Property). For any root-n-consistent approximate solution $\hat{\theta} = \theta_0 + O_p(n^{-\frac{1}{2}})$, we have that $P(\hat{\theta}_j = 0) = 1$ for $j > q + 1 + s$. Furthermore, if $n^{-\frac{1}{2}}S^P(\hat{\theta}) = o_p(1)$, then $\hat{\theta}_1$ has the asymptotic normality

$$n^{\frac{1}{2}}(A_{11} + \Sigma_{11})\left\{\hat{\theta}_1 - \theta_{10} + (A_{11} + \Sigma_{11})^{-1}b_n\right\} \xrightarrow{d} N(0, V_{11}),$$

where A_{11}, V_{11} are the first $(q + 1 + s) \times (q + 1 + s)$ submatrices of $A = \frac{1}{n}\frac{\partial S(\theta_0)}{\partial \theta}$ and V (as defined by Equation (6)), $\Sigma_{11} = \text{diag}\{\underbrace{0, \cdots, 0}_{q+1}, -q'_{\lambda n}(|\alpha_{10}|)\,\text{sgn}(\alpha_{10})\}$, and

$$b_n = -(\underbrace{0, \cdots, 0}_{q+1}, q_{\lambda n}(|\alpha_{01}|)\,\text{sgn}(\alpha_{01}), \cdots, q_{\lambda n}(|\alpha_{0s}|)\,\text{sgn}(\alpha_{0s}))^T.$$

c. Let $S_1^P(\theta)$ denote the first $(q+1+s)$-components of $S^P(\theta)$, then there exists $\hat{\theta}_1$ such that

$$S_1^P((\hat{\theta}_1^T, \mathbf{0}^T)^T) = 0;$$

that is, the solution is exact.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

4. Simulation Study

We consider a binary response Y and continuous exposure D in this section. Assume there are $n = 1000$ individuals and $q = 10$ candidate instruments. The observations $(Y_i, D_i, \mathbf{Z}_i), i = 1, \cdots, n$ are generated by

$$D_i = \mathbf{Z}_{i\cdot}^T\gamma + c_1U_i + \epsilon_i, \epsilon_i \sim N(0, 1)$$

$$\text{logit}\left[\mathbb{E}(Y_i \mid \mathbf{Z}_{i\cdot}, D_i, U_i)\right] = \mathbf{Z}_{i\cdot}^T\alpha + D_i\beta + c_2U_i,$$

where $\mathbf{Z}_{i\cdot}$ is drawn from a multivariate normal with zero mean and identity covariance matrix. We set $\beta = 2$, the fixed effect γ are drawn from $N(0, 1)$. We vary (i) the direct

effect parameter $\alpha = (1, 1, \cdots, 0, 0)$ where we change s in $\|\alpha\|_0 = s$, (ii) the distribution of unmeasured confounder U to test the robustness of our PSEE estimator.

Under each simulation scenario, we conduct 1000 replications. Our evaluation involves comparing the proposed PSEE method for estimating β with the original SEE method developed by Zhang and Tchetgen Tchetgen (2022) . Additionally, we compute estimates from the "naive" Two-Stage Least Squares (TSLS) method under the assumption that all instruments are valid, and the "oracle" TSLS method, assuming perfect knowledge of which instruments are valid. Our focus is on the estimation accuracy and invalid instrument selection properties of these methods, assessed through bias and root mean square error (RMSE), along with the average number of correct (C) and incorrect (I) zero estimates, respectively. Additionally, we calculate the sample standard deviation of $\hat{\beta}$ and the mean of the estimated standard deviation using the sandwich variance, denoted as SD_1 and SD_2. We also employ the sandwich variance formula to construct approximate 95% confidence intervals, relying on asymptotic normality theory, and report the corresponding empirical coverage probabilities.

The results of Table 1 summarize the performance of the naive TSLS, oracle TSLS, PSEE, and SEE for different number of invalid instruments s. The true distribution of unmeasured confounder U is Bernoulli(0.2) and $c_1 = c_2 = 1$. The PSEE (correct) and SEE (correct) denote estimators with correctly specified working model $U \sim$ Bernoulli(0.2), the PSEE (incorrect) and SEE (incorrect) denote estimators with incorrectly specified working model $U \sim$ Bernoulli(0.5). We observe that when majority rule holds, PSEE performs close to oracle TSLS in terms of instrument selection properties, and the estimated standard deviation closely approximates the empirical standard deviation, and the empirical coverage probability closely approaches 95%. These numerical results suggest the effective performance of the sandwich variance formula. Additionally, it is evident that PSEE outperforms naive TSLS

and SEE in bias and RMSE, even when the majority rule is violated. We also observe that PSEE performs well with incorrectly specified working model, these indicate Equation (4).

Table 2 summarizes the performance when the unmeasured confounder is continuous, $U \sim N(0,1)$, and $c_1 = c_2 = 0.5$. Our working model for U is a discrete uniform distribution on the interval $[-0.5, 0.5]$ with mesh size h. For computationally efficiency, we take $h = 0.5$. We observe that when majority rule holds, PSEE performs close to oracle TSLS in terms of bias, RMSE and instrument selection properties, and coverage approaches 95%. When majority rule does not hold, PSEE also has good performance in bias and RMSE. Table 3 presents the results for an alternative continuous unmeasured confounder setting, $U \sim t(3)$, with $c_1 = c_2 = 0.25$. The working model is consistent with that of Table 2. It is concluded from Table 3 again that PSEE exhibits robust performance across various evaluation metrics in this scenario as well.

5. Real Data Analysis

We demonstrate the potential advantages of our method in Mendelian randomization (MR) by analyzing the effect of BMI on suffering stroke. For this analysis, we leverage data from the Atherosclerosis Risk in Communities Study (ARIC), which is a prospective longitudinal epidemiological study conducted in four U.S. communities in North Carolina, Massachusetts, Maryland, and Minnesota.

Similar to another analysis with the ARIC data, we include individuals of white origin and extract European ancestry individuals and impute the data set on Michigan Imputation Center with EUR population from 1000 Genomes Phase 3 v5 reference panel (Shi et al., 2023). We remove 0.82% missing data in the following analysis. Finally, 8739 individuals are selected.

We consider potential candidate instruments for our MR analysis using the following SNPs in the ARIC data that have been previously associated with BMI: rs725959, rs1147199,

rs3817334, and rs6477694. While we have no specific reason to believe any of these SNPs are invalid IVs, uncertainty arises due to incomplete knowledge about their biological functions. Additionally, the inability to control all confounders precisely is a common scenario in MR studies.

Under the assumption that all instruments are valid, the TSLS method estimates a causal effect of 0.0805 (OR = 1.0838). In contrast, PSEE (with working model $U \sim$ Bernoulli(0.5)) estimates a causal effect of 0.1107 (SE: 0.0265, OR = 1.1171) with a 95% confidence interval [0.0588, 0.1626], excluding 0. The difference between TSLS and PSEE may stem from the underlying distribution of unmeasured confounder, as demonstrated in our simulations. Importantly, PSEE does not identify any SNPs as invalid IVs.

To further validate our method, we introduce another instrument, rs42039, associated with both BMI and stroke. Under the assumption that all four instruments are valid, TSLS estimates an effect of -0.0331 (OR = 0.9674). Conversely, PSEE (with working model $U \sim$ Bernoulli(0.5)) estimates a causal effect of 0.1108 (SE: 0.0265, OR = 1.1171), similar to the estimates when using four instruments. PSEE also excludes rs42039, suspected to be invalid.

In the real data analysis, PSEE provides similar estimates and consistently excludes the suspected invalid instrument (rs42039) when additional instrument is introduced, this indicates its robustness to possibly invalid instruments compared to TSLS. Furthermore, Harshfield et al. (2021) found that heightened obesity will increase the risk of ischemic, large artery, and small vessel stroke, which supports our PSEE results. Therefore, it is recommended to focus on interventions that reduce obesity to mitigate the risk of stroke.

6. Discussions

In this paper, we consider a flexible semiparametric instrumental variable model accommodating for continuous/binary exposure/outcome. We assume identifiability of this model under majority rule and propose a penalized semiparametric approach to estimate causal

effect. The asymptotic and oracle properties are outlined in Theorem 1 and further illustrated in comprehensive simulation studies. Specifically, our proposed method, PSEE, demonstrates performance close to that of the oracle TSLS regarding bias, RMSE, and instrument selection properties for both binary and continuous unmeasured confounder scenarios when majority rule holds. Even when majority rule does not hold, PSEE performs well in terms of bias and RMSE. Additionally, the robustness of our estimator is evident in both simulation experiments and real data analysis. We emphasize that PSEE does not impose specific requirements on whether the outcome and exposure are continuous or binary; it only requires the ability to specify the conditional probability distributions of the outcome and exposure. However, in our simulations, we specifically consider scenarios with binary outcome and continuous exposure. Nevertheless, it is worth noting that PSEE exhibits a lower coverage for increased values of sensitivity parameters c_1 and c_2, as illustrated in the Appendix C.

Further work could involve extending the considered model to the presence of additional complexities, such as nonlinear (Staley and Burgess, 2017) or time-varying exposure (Labrecque and Swanson, 2019), and vector-valued confounder. One could also consider different penalty functions, such as MCP (Zhang, 2010) and ALASSO (Zou, 2006) penalty, to compare their estimation accuracy and instrument selection properties. Moreover, the computational challenges posed by a considerable number of instrumental variables need to be tackled in future research. In our simulation experiments, involving a sample size of 1000 and 10 instrumental variables, the task of conducting 1000 replications on an 80-node cluster required an average of 30 hours. This highlights the necessity of developing efficient algorithms or optimization techniques, which would significantly enhance the scalability and practical applicability of our proposed method.

Acknowledgements

We express our gratitude to ARIC for generously providing the data. This research was supported partially by the National Key R&D Program of China [2023YFF1205101].

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444–455.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology* **44,** 512–525.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30,** 74–99.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35,** 109–135.

Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28,** 1356–1378.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59,** 126–132.

Harbord, R. M., Didelez, V., Palmer, T. M., Meng, S., Sterne, J. A., and Sheehan, N. A. (2013). Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. *Statistics in Medicine* **32,** 1246–1258.

Harshfield, E. L., Georgakis, M. K., Malik, R., Dichgans, M., and Markus, H. S. (2021).

Modifiable lifestyle factors and risk of stroke: A Mendelian randomization analysis. *Stroke* **52,** 931–936.

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33,** 1617–1642.

Johnson, B. A., Lin, D., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103,** 672–680.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association* **111,** 132–144.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business and Economic Statistics* **33,** 474–484.

Labrecque, J. A. and Swanson, S. A. (2019). Interpretation and potential biases of Mendelian randomization estimates with time-varying exposures. *American Journal of Epidemiology* **188,** 231–238.

Li, S. and Guo, Z. (2020). Causal inference for nonlinear outcome models with possibly invalid instrumental variables. *arXiv preprint arXiv:2010.09922* .

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5,** 99–135.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* **26,** 393–415.

Shi, D., Wang, Y., Zhang, Z., Cao, Y., and Hu, Y.-Q. (2023). MR-BOIL: Causal inference in one-sample Mendelian randomization for binary outcome with integrated likelihood method. *Genetic Epidemiology* **47,** 332–357.

Staley, J. R. and Burgess, S. (2017). Semiparametric methods for estimation of a non-linear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genetic Epidemiology* **41,** 341–352.

Sun, B., Liu, Z., and Tchetgen Tchetgen, E. (2023). Semiparametric efficient G-estimation with invalid instrumental variables. *Biometrika* **110,** 953–971.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Wang, L., Kai, B., Heuchenne, C., and Tsai, C.-L. (2013). Penalized profiled semiparametric estimating functions. *Electronic Journal of Statistics* **7,** 2656–2682.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68,** 353–360.

Windmeijer, F., Farbmacher, H., Davies, N., and Smith, G. D. (2018). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association* **114,** 1339–1350.

Zhang, B. and Tchetgen Tchetgen, E. J. (2022). A semi-parametric approach to model-based sensitivity analysis in observational studies. *Journal of the Royal Statistical Society Series A: Statistics in Society* **185,** S668–S691.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38,** 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Appendix

Appendix A: Proof of Theorem 1

PROOF: To prove part a, we consider $\hat{\theta} = (\hat{\theta}_1^T, \mathbf{0}^T)^T$, where $\hat{\theta}_1 = \theta_{10} + O_p(n^{-\frac{1}{2}})$. For $j = 1, 2, \cdots, p$, we have

$$\begin{aligned} n^{-\frac{1}{2}} S_j^P(\hat{\theta}) &= n^{-\frac{1}{2}} S_j(\theta_0) + n^{\frac{1}{2}} A_j(\theta - \theta_0) + o_p(1) - n^{\frac{1}{2}} q_{\lambda n}(|\hat{\theta}_j|)\,\mathrm{sgn}(\theta_j) \\ &= O_p(1) + O_p(1) + o_p(1) + o_p(1) \\ &= O_p(1). \end{aligned}$$

where A_j is the j^{th} row of A.

To prove part b, we consider the sets in the probability $C_j = \left\{\hat{\theta}_j \neq 0\right\}, j = q + 1 + s + 1, \cdots, p$. We show that for any $\varepsilon > 0$, when n is sufficiently large, $P(C_j) < \varepsilon$. Because $\hat{\theta}_j = O_p(n^{-\frac{1}{2}})$, there exists some M such that when n is large enough,

$$\begin{aligned} P(C_j) &= P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| \geqslant Mn^{-\frac{1}{2}}) + P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| < Mn^{-\frac{1}{2}}) \\ &< \frac{\varepsilon}{2} + P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| < Mn^{-\frac{1}{2}}) \end{aligned}$$

Using the jth component of the penalized estimating function and the definition of the approximate solution, we obtain that on the set of $\left\{\hat{\theta}_j \neq 0, |\hat{\theta}_j| < Mn^{-\frac{1}{2}}\right\}$,

$$\begin{aligned} O_p(1) &= n^{-\frac{1}{2}} S_j^P(\hat{\theta}) \\ &= n^{-\frac{1}{2}} S_j(\theta_0) + n^{\frac{1}{2}} A_j(\hat{\theta} - \theta_0) + o_p(1) - n^{\frac{1}{2}} q_{\lambda_n}(|\hat{\theta}_j|)\,\mathrm{sgn}(\hat{\theta}_j). \end{aligned}$$

The first three terms on the right side are of order $O_p(1)$. As a result, there exists some M' such that for large n,

$$P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| < Mn^{-\frac{1}{2}}, n^{\frac{1}{2}} q_{\lambda n}(|\hat{\theta}_j|)\,\mathrm{sgn}(\hat{\theta}_j) > M') < \frac{\varepsilon}{2}.$$

Because $\lim_{n\to\infty} \sqrt{n} \inf_{|\theta| \leqslant Mn^{-\frac{1}{2}}} q_{\lambda n}(|\theta|) \to \infty$ by condition (d), $\hat{\theta}_j \neq 0$ and $|\hat{\theta}_j| < Mn^{-\frac{1}{2}}$ imply that $n^{\frac{1}{2}} q_{\lambda n}(|\hat{\theta}_j|) > M'$ for large n. Thus $P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| < Mn^{-\frac{1}{2}}) = P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| <$

$Mn^{-\frac{1}{2}}, n^{\frac{1}{2}} q_{\lambda_n}(|\hat{\theta}_j|)\,\mathrm{sgn}(\hat{\theta}_j) > M')$. Therefore,

$$P(C_j) < \frac{\varepsilon}{2} + P(\hat{\theta}_j \neq 0, |\hat{\theta}_j| < Mn^{-\frac{1}{2}}, n^{\frac{1}{2}} q_{\lambda_n}(|\hat{\theta}_j|)\,\mathrm{sgn}(\hat{\theta}_j) > M') < \varepsilon.$$

We next show the asymptotic normality of $\hat{\theta}_1$ when the order of $n^{-\frac{1}{2}} S^P(\hat{\theta})$ is $o_p(1)$. We have

$$\begin{aligned} o_p(1) &= n^{-\frac{1}{2}} S_1^P(\hat{\theta}) \\ &= n^{-\frac{1}{2}} S_1(\theta_0) + n^{\frac{1}{2}} A_{11}(\hat{\theta}_1 - \theta_{10}) + o_p(1) - n^{\frac{1}{2}} q_{\lambda_n}(|\hat{\theta}_1|)\,\mathrm{sgn}(\hat{\theta}_1). \end{aligned}$$

Employing Taylor expansions of $q_{\lambda n}(|\hat{\theta}_1|)\,\mathrm{sgn}(\hat{\theta}_1)$ at θ_{10} yields

$$n^{\frac{1}{2}}(\hat{\theta}_1 - \theta_{10}) = -n^{\frac{1}{2}}(A_{11} + \Sigma_{11})^{-1}\left[S_1(\theta_{10}) + b_n\right] + o_p(1).$$

where

$$\Sigma_{11} = \mathrm{diag}(\underbrace{0, \cdots, 0}_{q+1}, -q'_{\lambda_n}(|\alpha_{01}|), \cdots, -q'_{\lambda_n}(|\alpha_{0s}|))$$

and

$$b_n = -(\underbrace{0, \cdots, 0}_{q+1}, q_{\lambda_n}(|\alpha_{01}|)\,\mathrm{sgn}(\alpha_{01}), \cdots, q_{\lambda_n}(|\alpha_{0s}|)\,\mathrm{sgn}(\alpha_{0s}))^\top.$$

we then obtain that

$$\sqrt{n}(A_{11} + \Sigma_{11})\left[(\hat{\theta}_1 - \theta_{10}) + (A_{11} + \Sigma_{11})^{-1} b_n\right] \xrightarrow{d} N(0, V_{11})$$

where V_{11} is the $(q + 1 + s) \times (q + 1 + s)$ submatrix in the upper-left corner of V. This completes the proof.

To establish part c, we examine $\theta_1 \in \mathbb{R}^{q+1+s}$ situated on the boundary of a ball centered around θ_{10}, defined as $\theta_1 = \theta_{10} + n^{-1/2}u$ with $|u| = r$ for a constant r. Leveraging the

penalized estimating function S_1^P, we derive the following expression:

$$\begin{aligned}
&n^{-1/2}(\theta_1-\theta_{10})^T A_{11}^T S_1^P(\theta)\\
&=(\theta_1-\theta_{10})^T A_{11}^T\left\{n^{-1/2}S_1(\theta)-n^{1/2}q_{\lambda_n}(|\theta_1|)\,\mathrm{sgn}(\theta_1)\right\}\\
&=O_p(|\theta_1-\theta_{10}|)+n^{1/2}(\theta_1-\theta_{10})^T A_{11}^T A_{11}(\theta_1-\theta_{10})\\
&\quad-n^{1/2}(\theta_1-\theta_{10})A_{11}^T\,\mathrm{diag}\left\{q'_{\lambda_n}(\left|\theta_j^*\right|)\,\mathrm{sgn}(\theta_{0j})\right\}(\theta_1-\theta_{10}),
\end{aligned}$$

where θ_j^* lies between θ_j and θ_{0j} for $j=1,\ldots,s$. As A_{11} is nonsingular, the second term on the right side exceeds $a_0 r^2 n^{-1/2}$, where a_0 denotes the smallest eigenvalue of $A_{11}^T A_{11}$. The first term is of order $rO_p(n^{-1/2})$. Due to the convergence of $\max_j q'_{\lambda_n}(\left|\theta_j^*\right|)$ to 0, the third term is dominated by the second term. Therefore, by selecting r adequately large such that, for large n, the probability that the absolute value of the first term surpasses the second term is less than ϵ, we obtain

$$P\left[\min_{|\theta_1-\theta_{10}|=n^{-1/2}r}(\theta_1-\theta_{10})^T A_{11}^T S_1^P((\theta_1^T,\mathbf{0}^T)^T)>0\right]>1-\epsilon.$$

Applying the Brouwer fixed-point theorem to the continuous function $S_1^P((\theta_1^T,\mathbf{0}^T)^T)$, we see that $\min_{|\theta_1-\theta_{10}|=n^{-1/2}r}(\theta_1-\theta_{10})^T A_{11}^T\times\ S_1^P((\theta_1^T,\mathbf{0}^T)^T)>0$ implies that $A_{11}^T S_1^P((\theta_1^T,\mathbf{0}^T)^T)$ has a solution within this ball. In other words, $A_{11}^T S_1^P((\theta_1^T,\mathbf{0}^T)^T)$ has a solution within this ball, or equivalently, $S_1^P((\theta_1^T,\mathbf{0}^T)^T)$ has a solution within this ball. Thus, we can select an exact solution $\widehat{\theta}=(\widehat{\theta}_1^T,\mathbf{0}^T)^T$ to $S_1^P(\theta)=\mathbf{0}$ with $\widehat{\theta}=\theta_0+O_p(n^{-1/2})$. □

Appendix B: Boxplots and histograms for simulation studies

$U\sim$ Bernoulli(0.2).

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

$U \sim N(0,1)$.

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

$U \sim t(3)$.

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

Appendix C: Simulation results for increased values of sensitivity parameters

$U \sim N(0,1)$ *and* $c_1 = c_2 = 0.5$.

[Table 4 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

$U \sim t(3)$ *and* $c_1 = c_2 = 0.5$.

[Table 5 about here.]

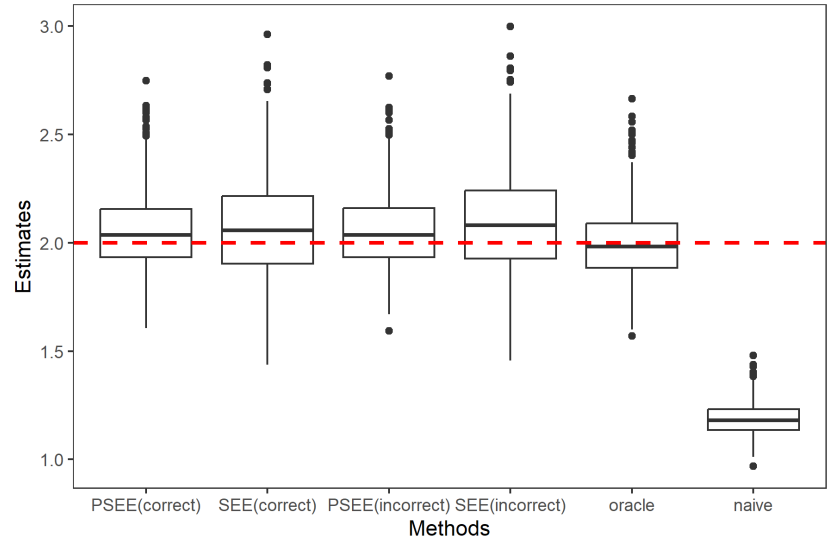
[Figure 17 about here.]

[Figure 18 about here.]

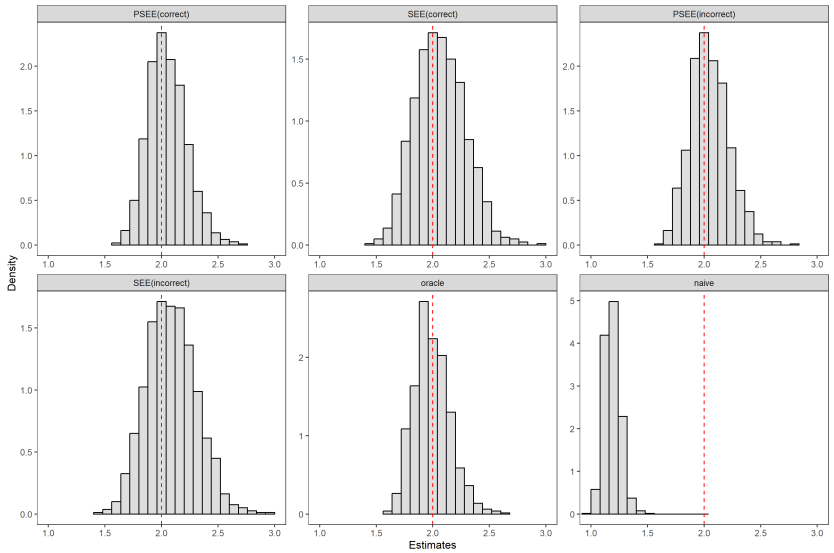
[Figure 19 about here.]

[Figure 20 about here.]

Figure A1: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim$ Bernoulli$(0.2), c_1 = c_2 = 1$, 1 invalid IV, 1000 replications.

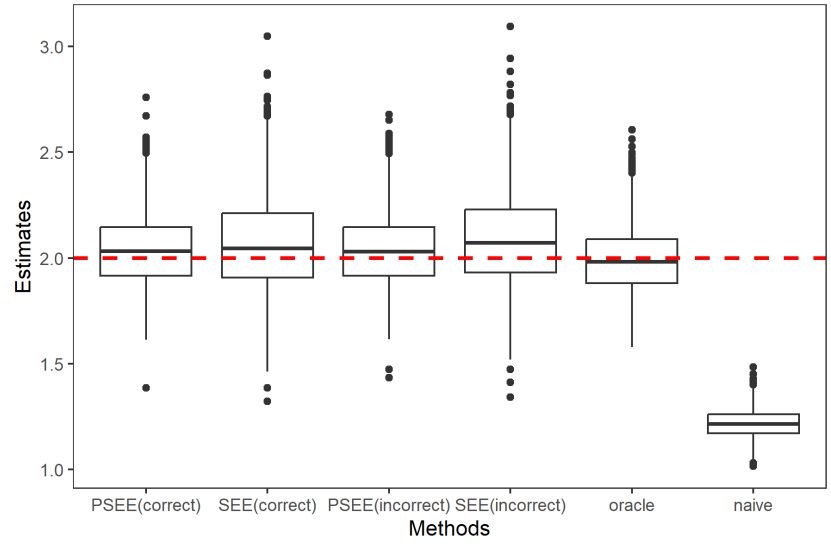


(a) Boxplot: true value of β is represented by a horizontal red dashed line

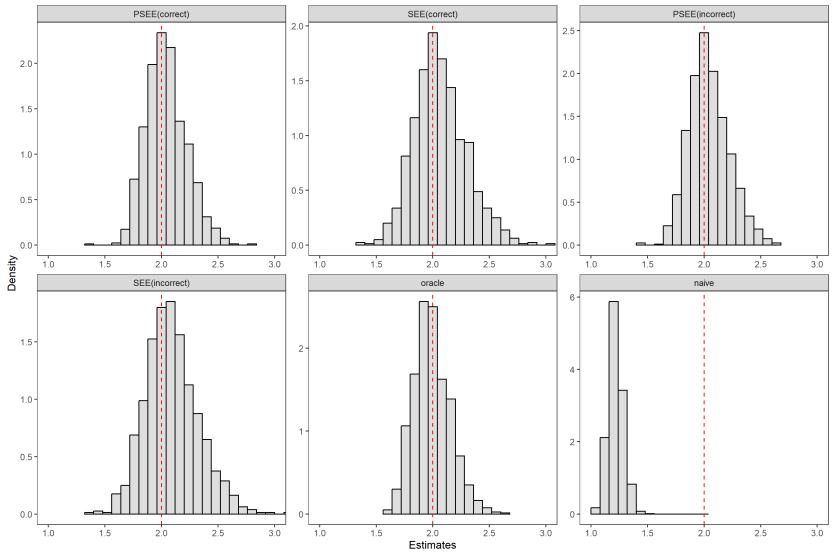


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A2: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim$ Bernoulli$(0.2), c_1 = c_2 = 1$, 3 invalid IVs, 1000 replications.

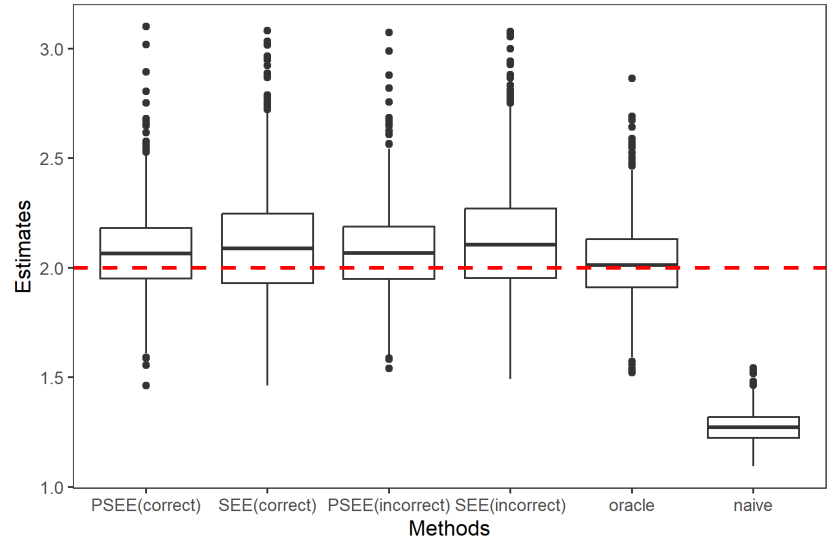


(a) Boxplot: true value of β is represented by a horizontal red dashed line

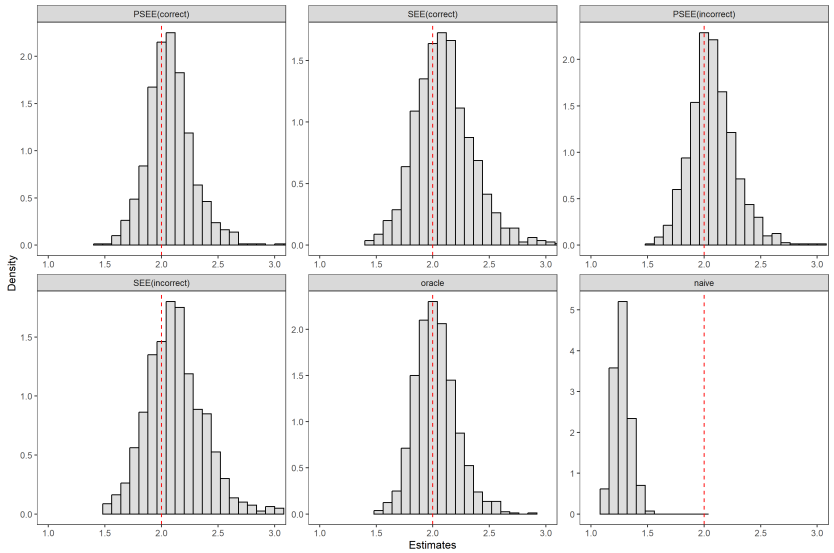


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A3: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim$ Bernoulli$(0.2), c_1 = c_2 = 1$, 5 invalid IVs, 1000 replications.

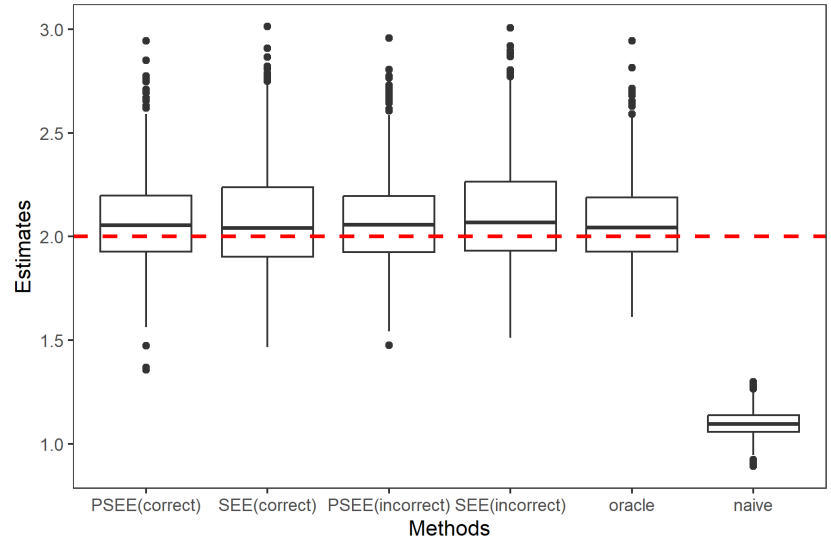


(a) Boxplot: true value of β is represented by a horizontal red dashed line

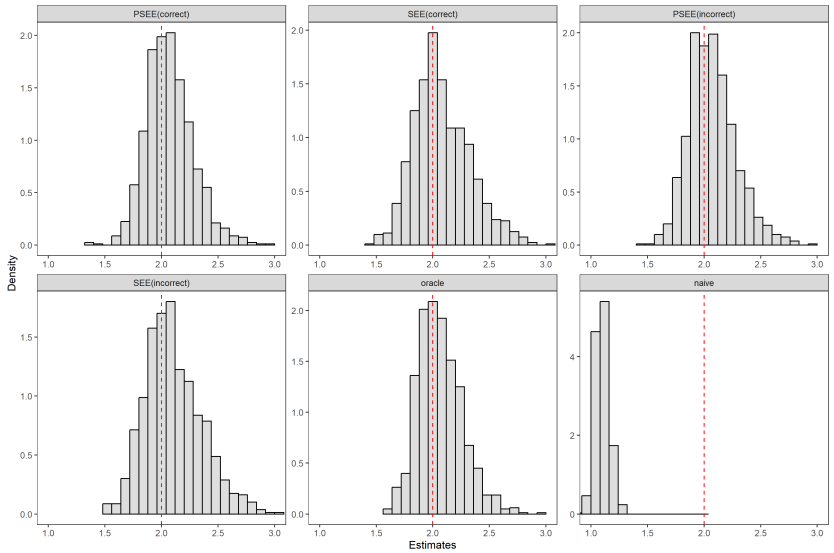


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A4: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim$ Bernoulli$(0.2), c_1 = c_2 = 1$, 7 invalid IVs, 1000 replications.

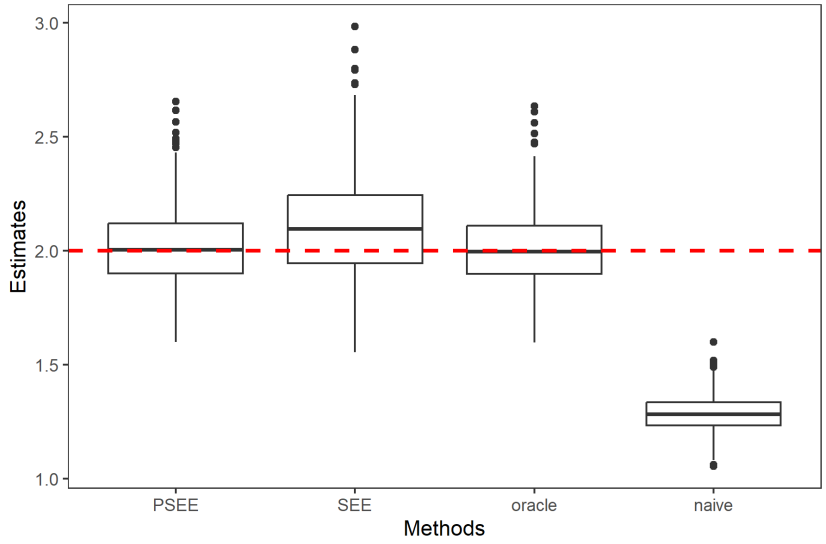


(a) Boxplot: true value of β is represented by a horizontal red dashed line

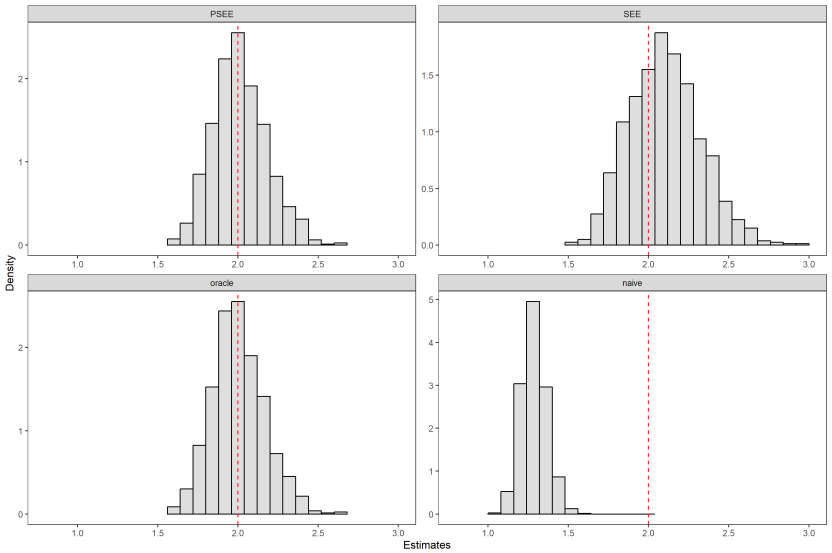


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A5: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0, 1), c_1 = c_2 = 0.25$, 1 invalid IV, 1000 replications.

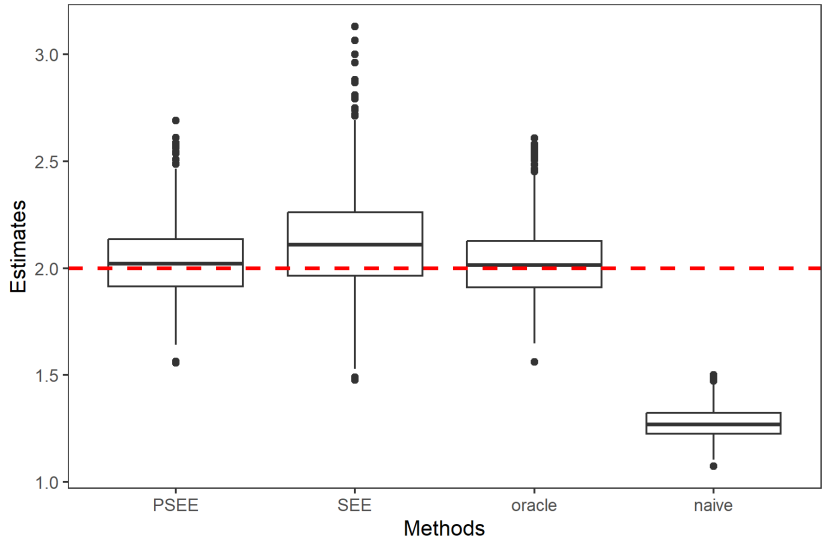


(a) Boxplot: true value of β is represented by a horizontal red dashed line

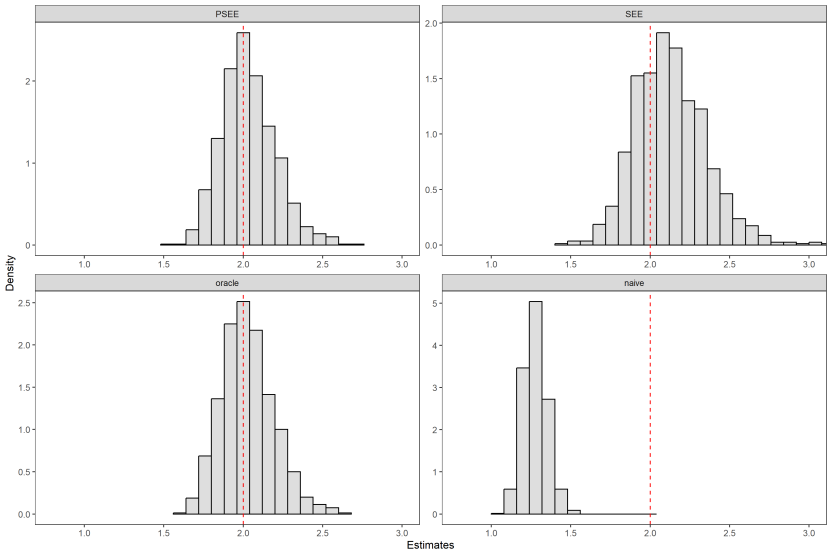


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A6: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0, 1), c_1 = c_2 = 0.25$, 3 invalid IVs, 1000 replications.

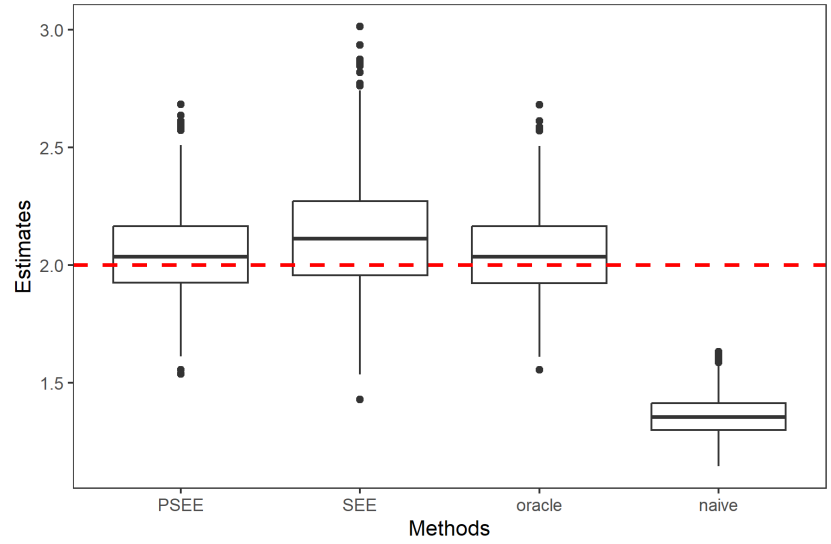


(a) Boxplot: true value of β is represented by a horizontal red dashed line

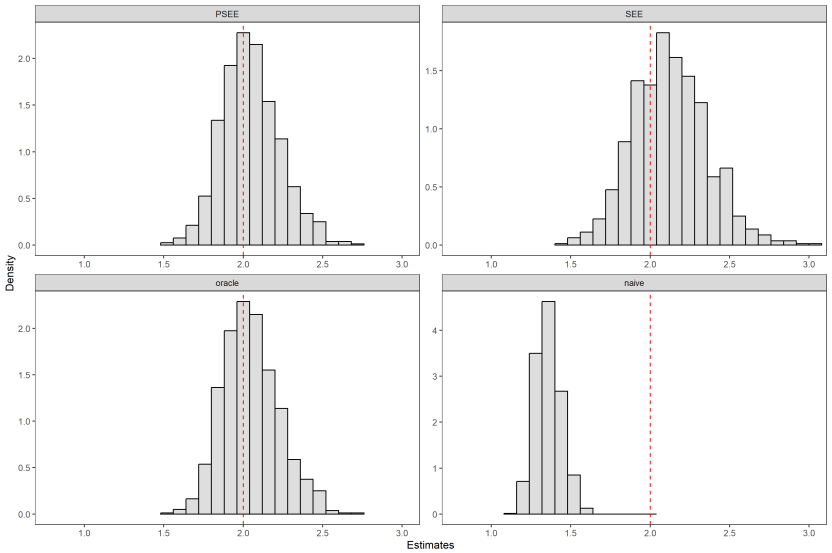


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A7: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0, 1), c_1 = c_2 = 0.25$, 5 invalid IVs, 1000 replications.

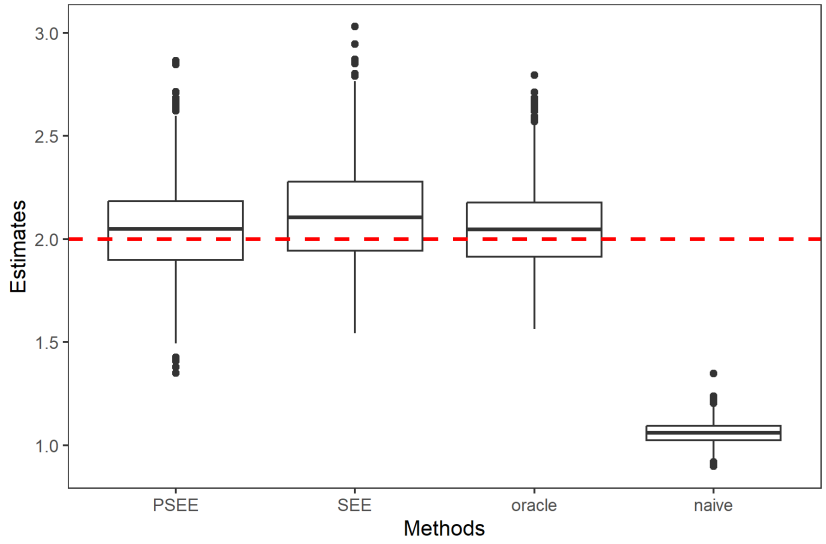


(a) Boxplot: true value of β is represented by a horizontal red dashed line

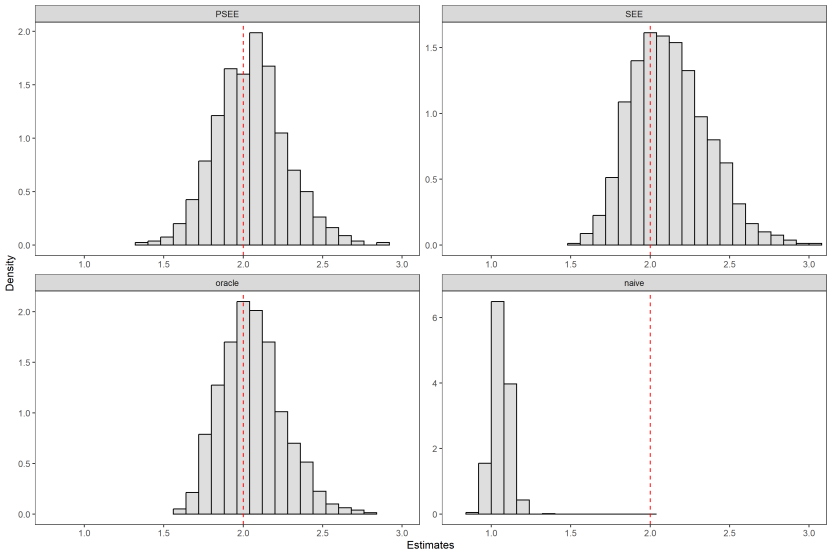


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A8: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.25$, 7 invalid IVs, 1000 replications.

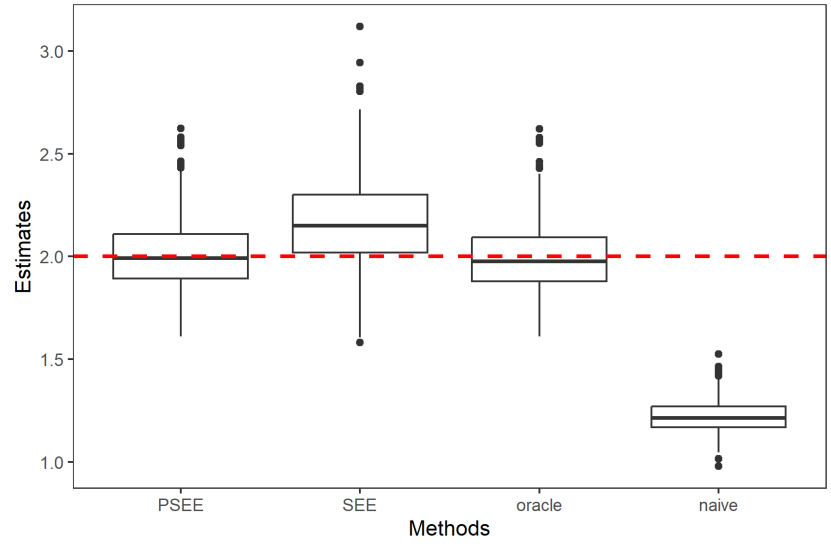


(a) Boxplot: true value of β is represented by a horizontal red dashed line

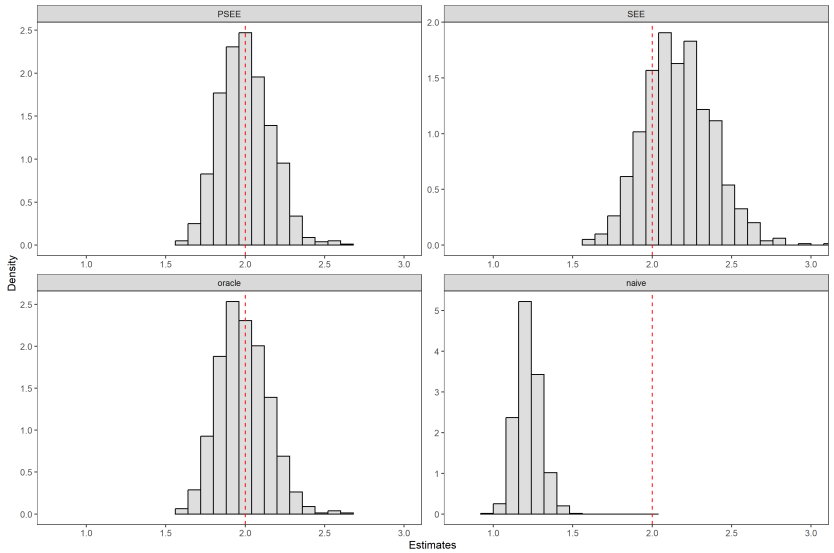


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A9: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.25$, 1 invalid IV, 1000 replications.

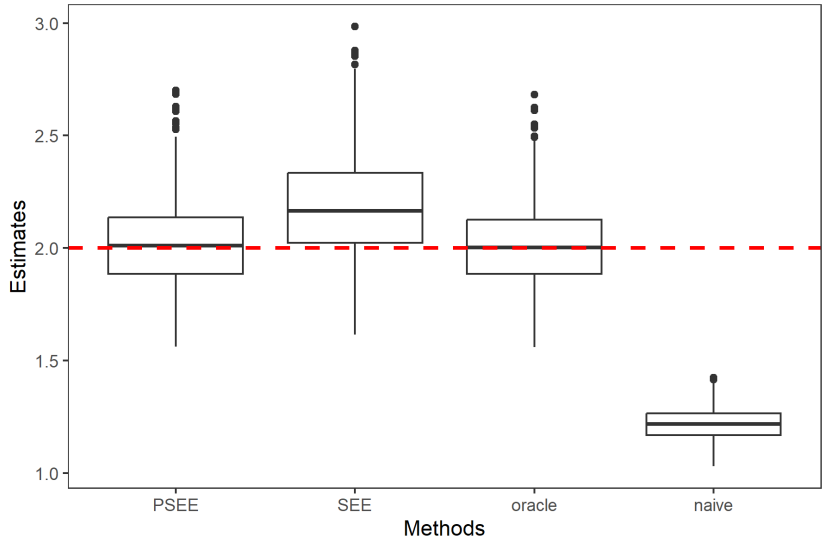


(a) Boxplot: true value of β is represented by a horizontal red dashed line

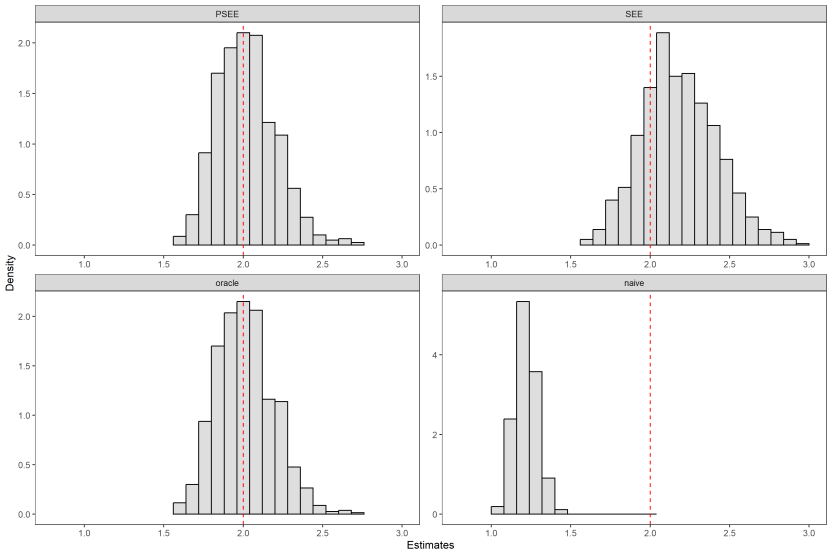


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A10: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.25$, 3 invalid IVs, 1000 replications.

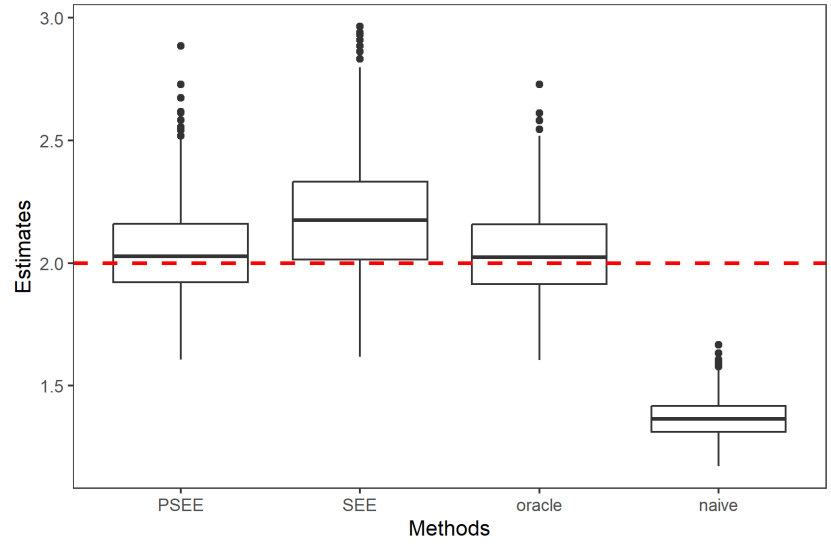


(a) Boxplot: true value of β is represented by a horizontal red dashed line

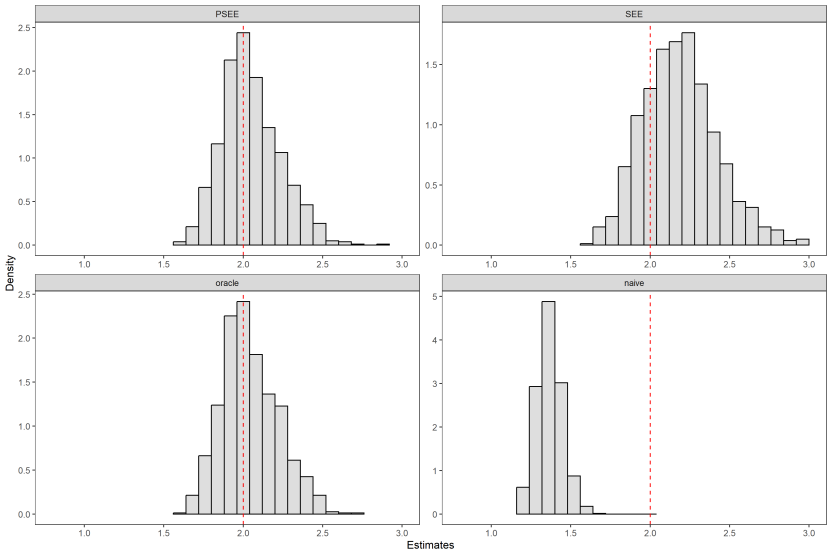


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A11: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.25$, 5 invalid IVs, 1000 replications.

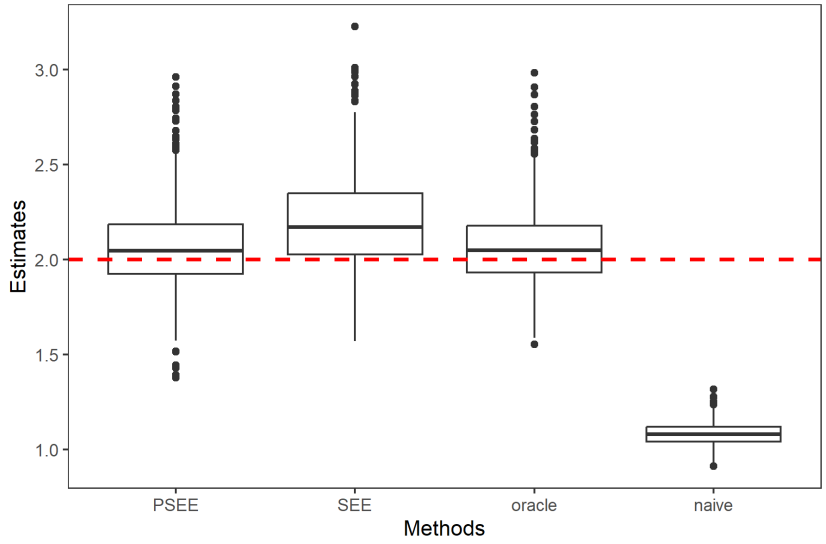


(a) Boxplot: true value of β is represented by a horizontal red dashed line

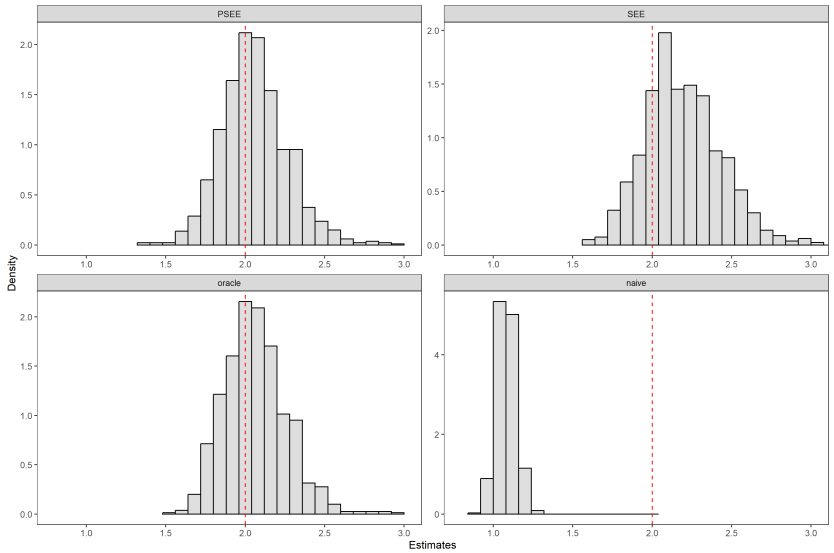


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A12: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.25$, 7 invalid IVs, 1000 replications.

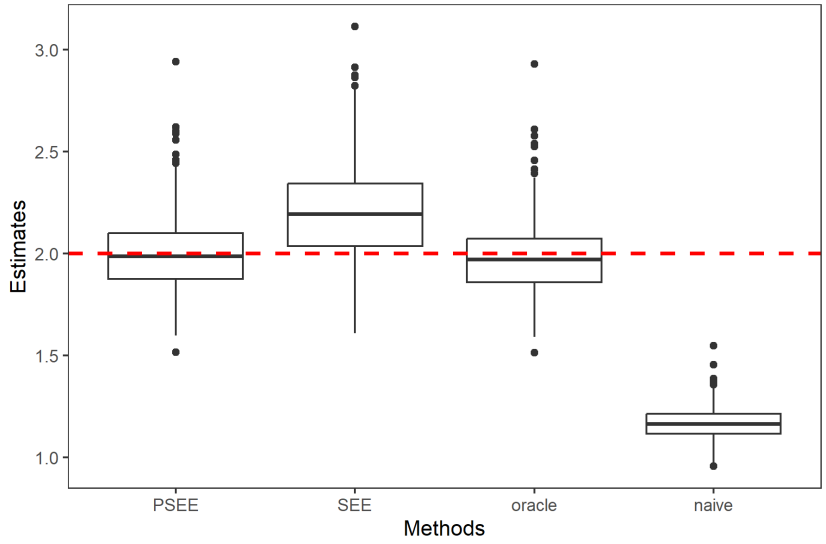


(a) Boxplot: true value of β is represented by a horizontal red dashed line

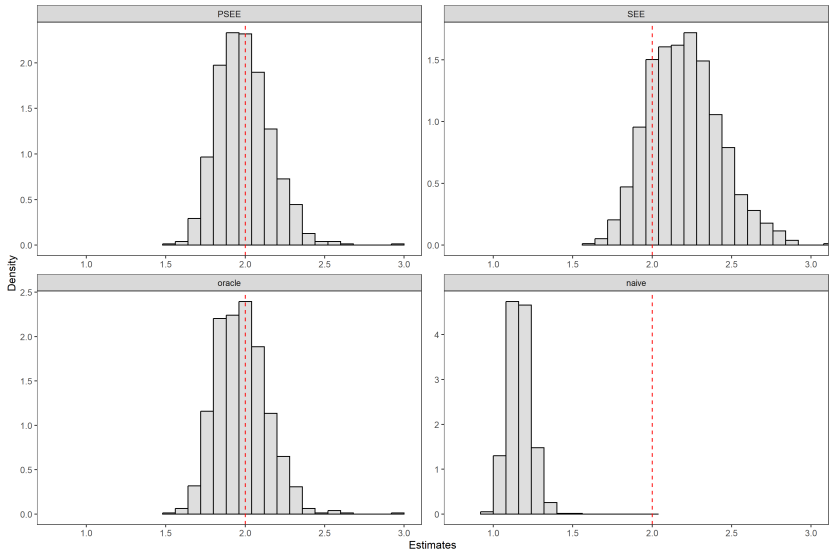


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A13: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.5$, 1 invalid IV, 1000 replications.

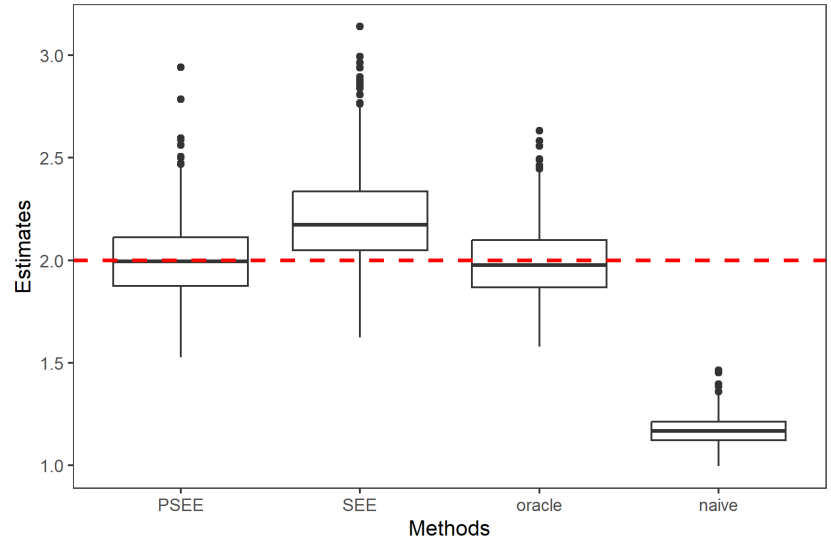


(a) Boxplot: true value of β is represented by a horizontal red dashed line

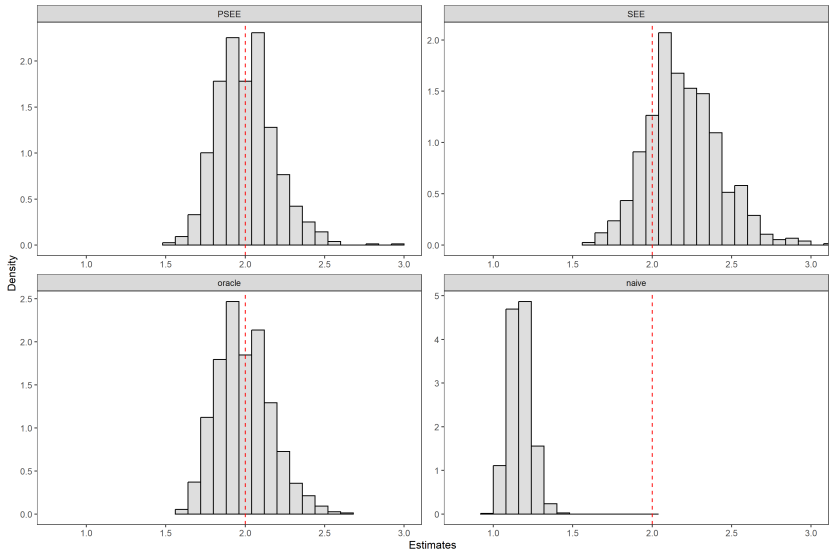


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A14: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.5$, 3 invalid IVs, 1000 replications.

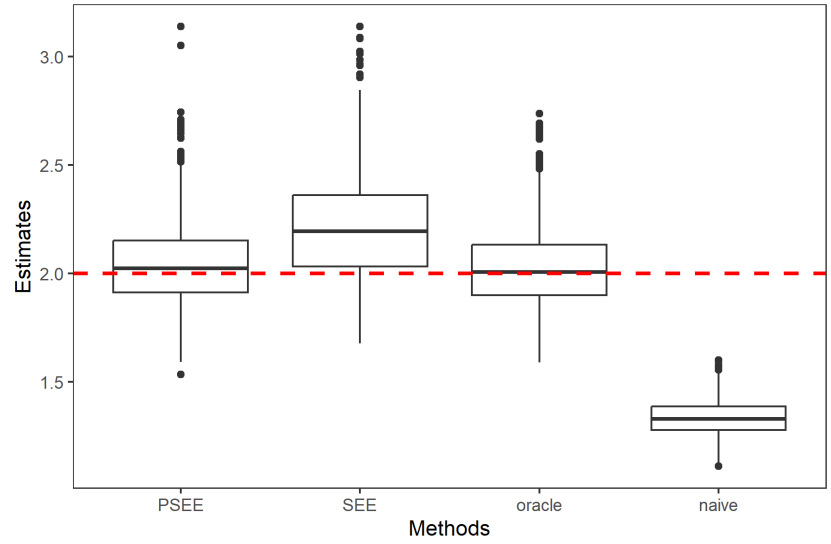


(a) Boxplot: true value of β is represented by a horizontal red dashed line

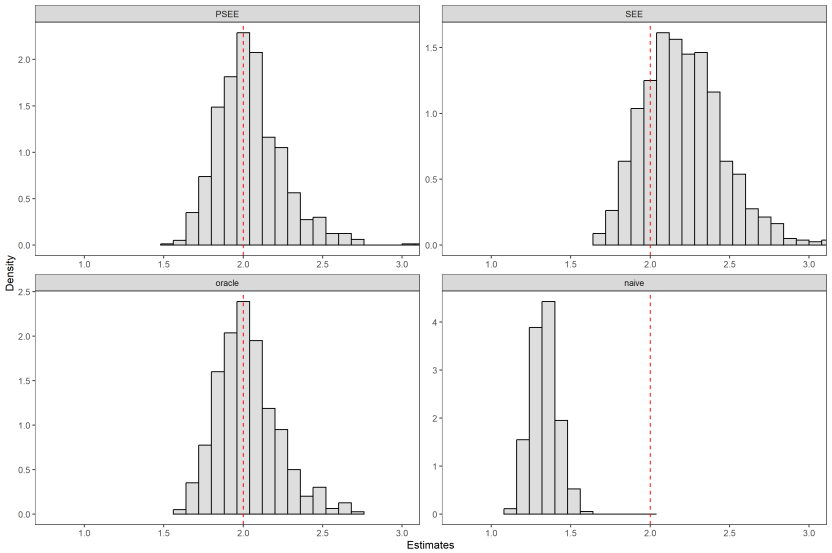


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A15: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.5$, 5 invalid IVs, 1000 replications.

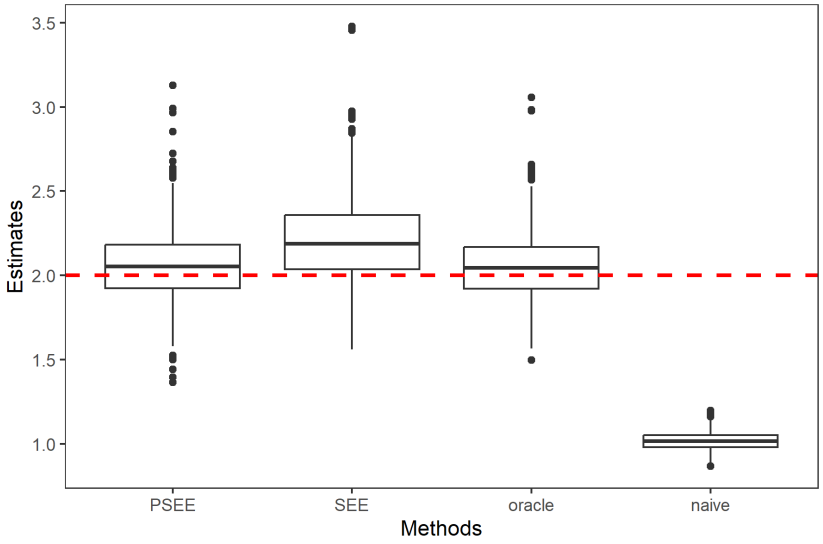


(a) Boxplot: true value of β is represented by a horizontal red dashed line

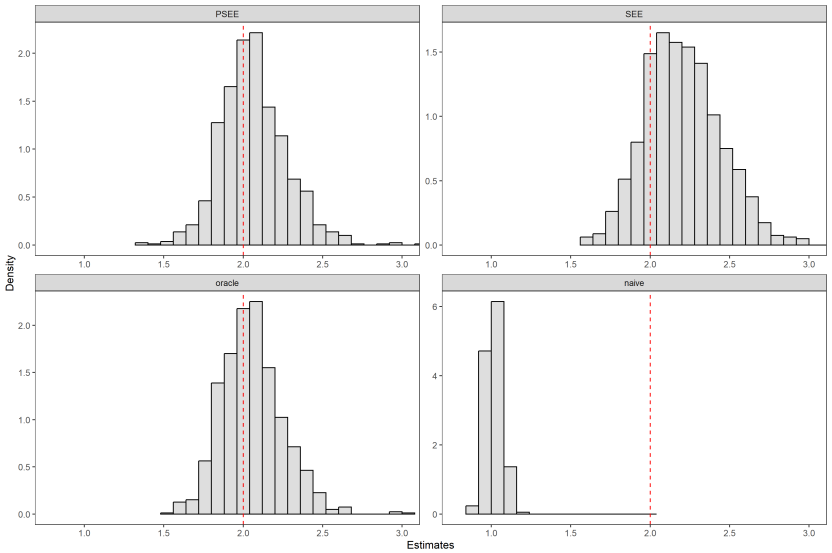


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A16: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.5$, 7 invalid IVs, 1000 replications.

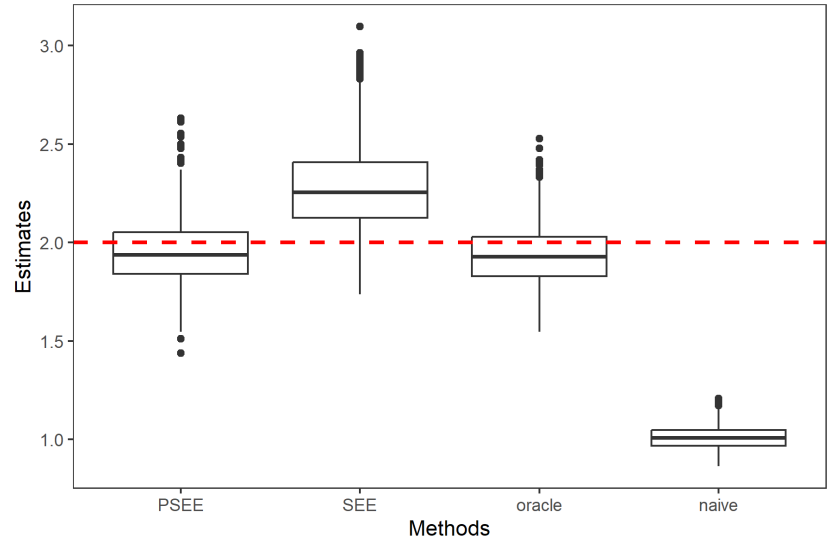


(a) Boxplot: true value of β is represented by a horizontal red dashed line

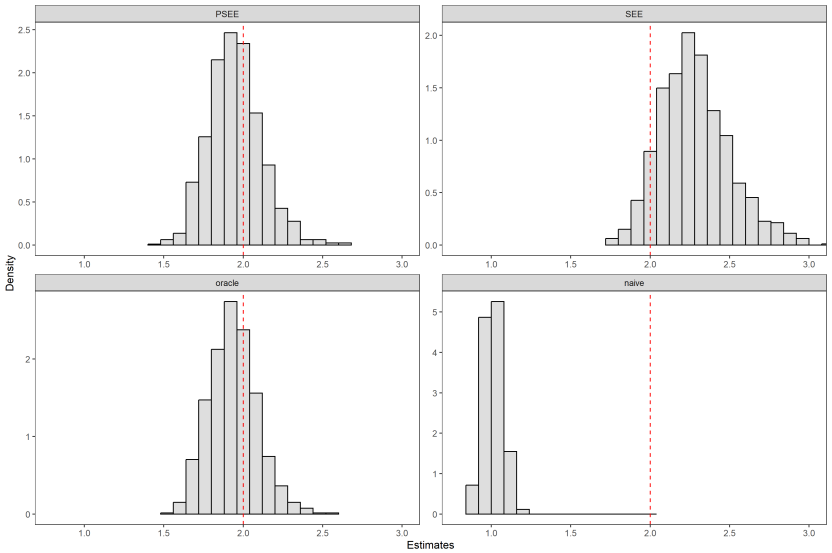


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A17: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.5$, 1 invalid IV, 1000 replications.

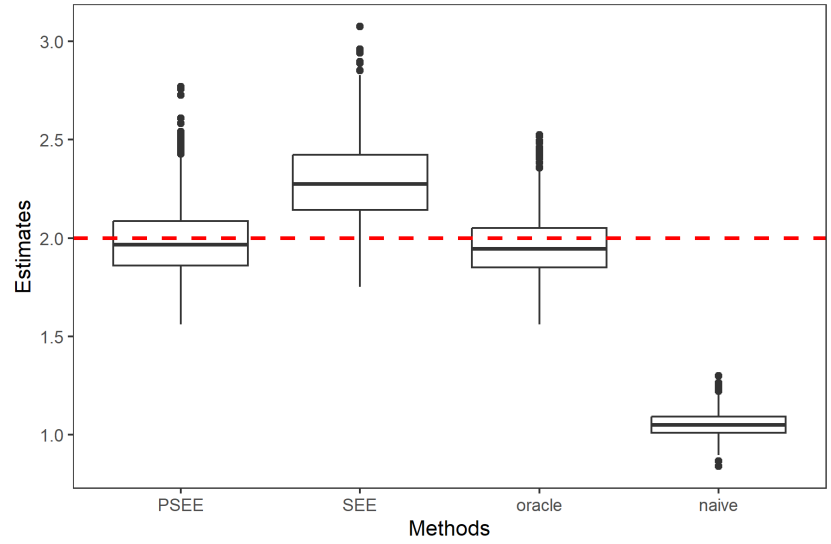


(a) Boxplot: true value of β is represented by a horizontal red dashed line

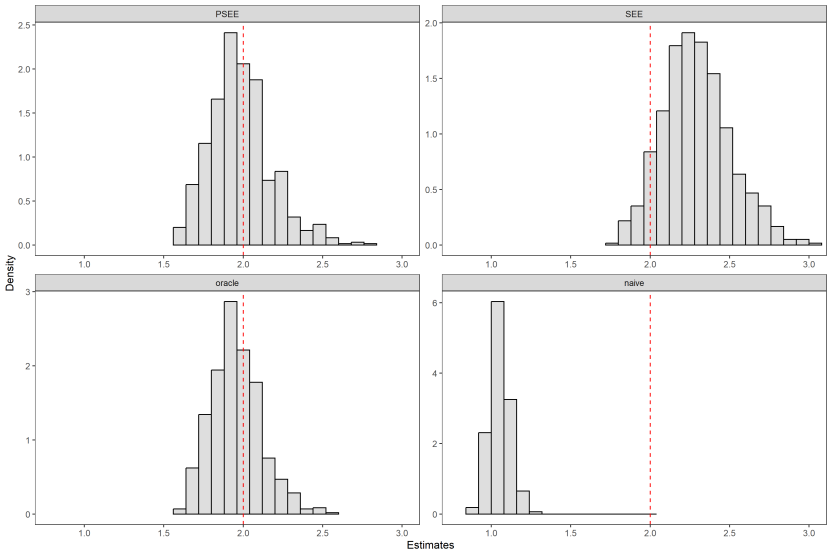


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A18: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.5$, 3 invalid IVs, 1000 replications.

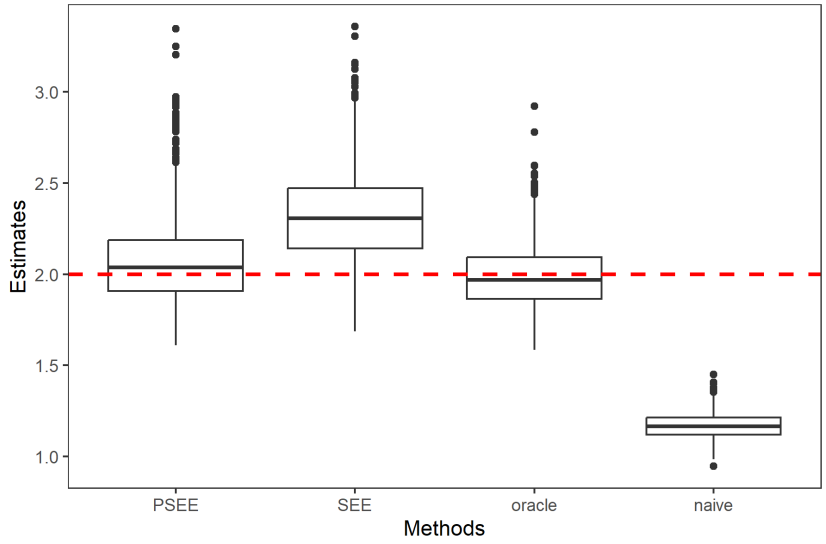


(a) Boxplot: true value of β is represented by a horizontal red dashed line

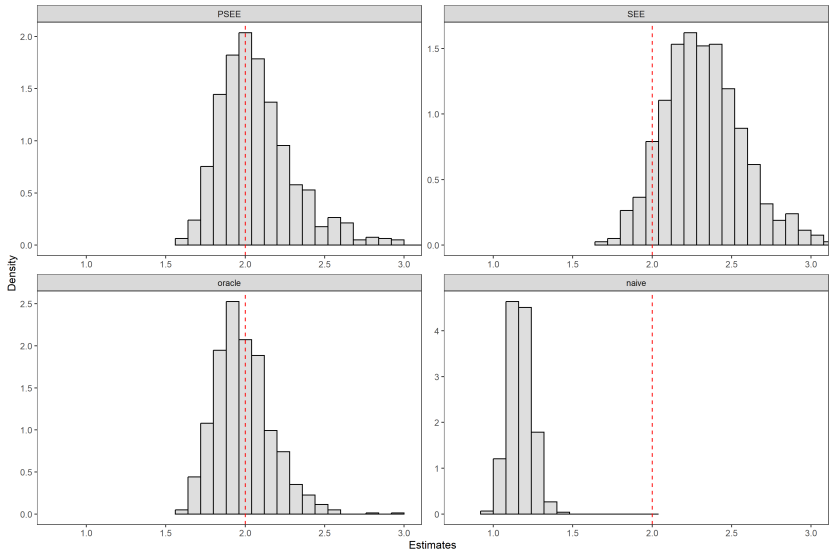


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A19: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.5$, 5 invalid IVs, 1000 replications.

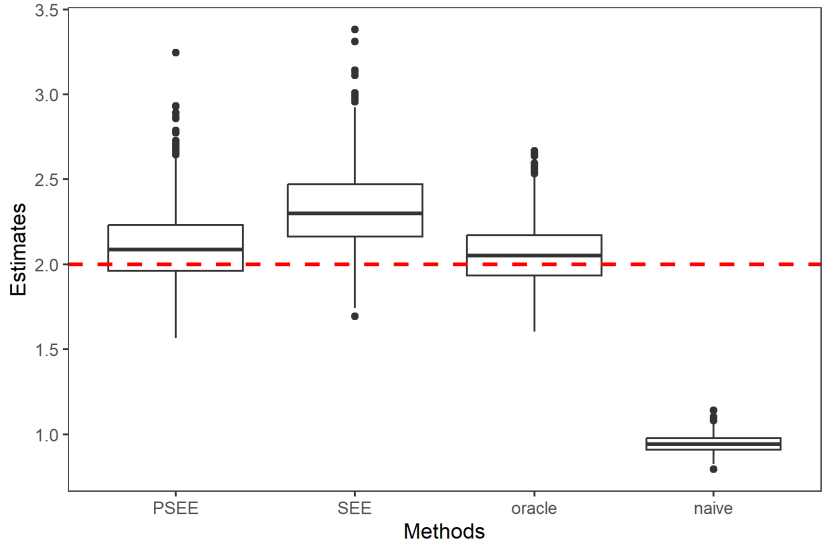


(a) Boxplot: true value of β is represented by a horizontal red dashed line

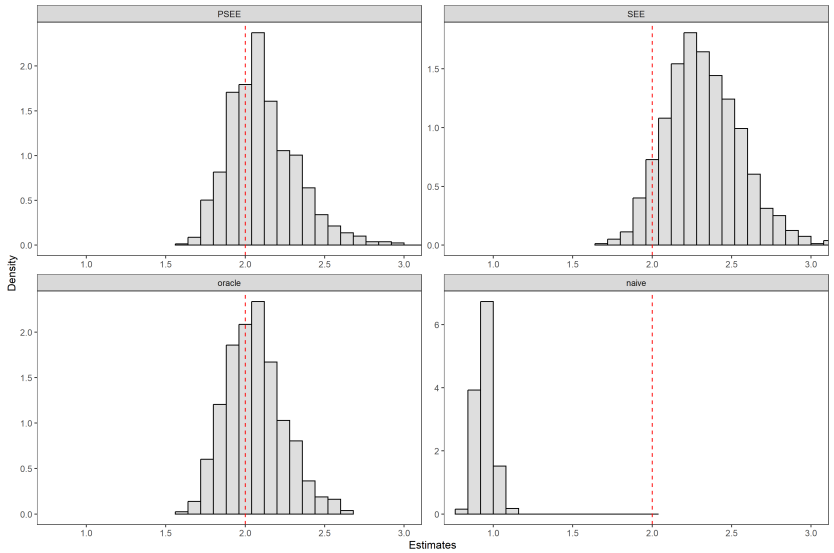


(b) Histogram: true value of β is represented by a vertical red dashed line

Figure A20: Boxplot and histogram of estimates of the causal effect $\beta, n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.5$, 7 invalid IVs, 1000 replications.



(a) Boxplot: true value of β is represented by a horizontal red dashed line



(b) Histogram: true value of β is represented by a vertical red dashed line

Table 1

Estimation Results for β ($n = 1000, q = 10, U \sim \text{Bernoulli}(0.2), c_1 = c_2 = 1$): comparison of naive TSLS, oracle TSLS, PSEE (correct), SEE (correct), PSEE (incorrect) and SEE (incorrect). Here "correct" denotes estimator with correctly specified working model, i.e., $f^*(U \mid \mathbf{Z}; \xi) = \text{Bernoulli}(0.2)$; "incorrect" denotes estimator with incorrectly specified working model, i.e., $f^*(U \mid \mathbf{Z}; \xi) = \text{Bernoulli}(0.5)$. SD_1 and SD_2 denote the sample standard deviation and the mean of the estimated standard deviation using the sandwich variance. C and I denote the average number of correct and incorrect zero estimates, respectively. True β equals 2.0.

	Bias	RMSE	SD_1	SD_2	Coverage	C	I
			Majority rule holds				
1 invalid IV							
PSEE(correct)	0.053	0.180	0.162	0.172	93.4%	8.644	0.026
SEE(correct)	0.065	0.231	0.216	0.222	95.1%		
PSEE(incorrect)	0.051	0.180	0.162	0.172	93.4%	8.652	0.021
SEE(incorrect)	0.089	0.239	0.216	0.222	95.1%		
Oracle	-0.008	0.161				9	0
Naive	-0.814	0.817					
3 invalid IVs							
PSEE(correct)	0.043	0.183	0.170	0.178	94.8%	6.877	0.097
SEE(correct)	0.067	0.242	0.227	0.233	94.1%		
PSEE(incorrect)	0.042	0.183	0.171	0.178	94.4%	6.868	0.078
SEE(incorrect)	0.090	0.250	0.227	0.233	93.6%		
Oracle	-0.007	0.164				7	0
Naive	-0.781	0.784					
			Majority rule is violated				
5 invalid IVs							
PSEE(correct)	0.079	0.217	0.181	0.202	91.8%	4.881	0.235
SEE(correct)	0.099	0.271	0.235	0.252	93.6%		
PSEE(incorrect)	0.078	0.216	0.180	0.202	91.8%	4.883	0.268
SEE(incorrect)	0.122	0.281	0.235	0.253	93.3%		
Oracle	0.026	0.187				5	0
Naive	-0.724	0.728					
7 invalid IVs							
PSEE(correct)	0.072	0.221	0.181	0.208	91.7%	2.951	0.756
SEE(correct)	0.081	0.263	0.230	0.251	93.6%		
PSEE(incorrect)	0.074	0.222	0.181	0.210	91.5%	2.945	0.753
SEE(incorrect)	0.106	0.272	0.230	0.251	93.0%		
Oracle	0.066	0.208				3	0
Naive	-0.901	0.903					

Table 2

Estimation Results for β ($n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.25$): comparison of naive TSLS, oracle TSLS, PSEE, SEE. Working model for U is a discrete uniform distribution on the interval $[-0.5, 0.5]$ with mesh size h, here we take $h = 0.5$. SD_1 and SD_2 denote the sample standard deviation and the mean of the estimated standard deviation using the sandwich variance. C and I denote the average number of correct and incorrect zero estimates, respectively. True β equals 2.0.

	Bias	RMSE	SD_1	SD_2	Coverage	C	I
			Majority rule holds				
1 invalid IV							
PSEE	0.015	0.168	0.159	0.168	93.9%	8.694	0.005
SEE	0.105	0.245	0.215	0.221	94.4%		
Oracle	0.006	0.162				9	0
Naive	-0.715	0.719					
3 invalid IVs							
PSEE	0.033	0.173	0.167	0.170	95.0%	6.942	0.018
SEE	0.128	0.259	0.221	0.225	94.1%		
Oracle	0.027	0.167				7	0
Naive	-0.724	0.728					
			Majority rule is violated				
5 invalid IVs							
PSEE	0.047	0.187	0.177	0.181	94.1%	4.976	0.076
SEE	0.124	0.265	0.226	0.234	93.8%		
Oracle	0.048	0.183				5	0
Naive	-0.641	0.646					
7 invalid IVs							
PSEE	0.050	0.227	0.181	0.222	89.4%	2.943	0.438
SEE	0.124	0.270	0.229	0.240	93.4%		
Oracle	0.059	0.204				3	0
Naive	-0.940	0.942					

Table 3

Estimation Results for β ($n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.25$): comparison of naive TSLS, oracle TSLS, PSEE, SEE. Working model for U is a discrete uniform distribution on the interval $[-0.5, 0.5]$ with mesh size h, here we take $h = 0.5$. SD_1 and SD_2 denote the sample standard deviation and the mean of the estimated standard deviation using the sandwich variance. C and I denote the average number of correct and incorrect zero estimates, respectively. True β equals 2.0.

	Bias	RMSE	SD_1	SD_2	Coverage	C	I
			Majority rule holds				
1 invalid IV							
PSEE	0.004	0.160	0.155	0.160	94.1%	8.633	0.014
SEE	0.163	0.268	0.208	0.213	90.4%		
Oracle	-0.010	0.155				9	0
Naive	-0.780	0.783					
3 invalid IVs							
PSEE	0.023	0.186	0.168	0.185	94.0%	6.886	0.040
SEE	0.182	0.297	0.224	0.235	89.6%		
Oracle	0.014	0.179				7	0
Naive	-0.782	0.785					
			Majority rule is violated				
5 invalid IVs							
PSEE	0.049	0.191	0.175	0.185	94.5%	4.927	0.119
SEE	0.187	0.299	0.226	0.233	91.5%		
Oracle	0.049	0.191				5	0
Naive	-0.633	0.638					
7 invalid IVs							
PSEE	0.061	0.220	0.179	0.212	91.5%	2.905	0.599
SEE	0.195	0.310	0.231	0.241	89.5%		
Oracle	0.065	0.206				3	0
Naive	-0.917	0.919					

Table A1

Estimation Results for β ($n = 1000, q = 10, U \sim N(0,1), c_1 = c_2 = 0.5$): comparison of naive TSLS, oracle TSLS, PSEE, SEE. Working model for U is a discrete uniform distribution on the interval $[-0.5, 0.5]$ with mesh size h, here we take $h = 0.5$. SD_1 and SD_2 denote the sample standard deviation and the mean of the estimated standard deviation using the sandwich variance. C and I denote the average number of correct and incorrect zero estimates, respectively. True β equals 2.0.

	Bias	RMSE	SD_1	SD_2	Coverage	C	I
			Majority rule holds				
1 invalid IV							
PSEE	-0.005	0.165	0.158	0.165	94.9%	8.683	0.024
SEE	0.201	0.300	0.217	0.223	87.9%		
Oracle	-0.023	0.161				9	0
Naive	-0.834	0.837					
3 invalid IVs							
PSEE	0.005	0.180	0.163	0.180	92.4%	6.878	0.079
SEE	0.194	0.299	0.219	0.227	88.1%		
Oracle	-0.006	0.169				7	0
Naive	-0.830	0.833					
			Majority rule is violated				
5 invalid IVs							
PSEE	0.045	0.210	0.180	0.206	92.2%	4.858	0.265
SEE	0.208	0.322	0.234	0.246	89.1%		
Oracle	0.029	0.194				5	0
Naive	-0.667	0.672					
7 invalid IVs							
PSEE	0.065	0.217	0.176	0.207	91.4%	2.908	0.973
SEE	0.207	0.320	0.227	0.244	86.9%		
Oracle	0.057	0.202				3	0
Naive	-0.983	0.984					

Table A2

Estimation Results for β ($n = 1000, q = 10, U \sim t(3), c_1 = c_2 = 0.5$): comparison of naive TSLS, oracle TSLS, PSEE, SEE. Working model for U is a discrete uniform distribution on the interval $[-0.5, 0.5]$ with mesh size h, here we take $h = 0.5$. SD_1 and SD_2 denote the sample standard deviation and the mean of the estimated standard deviation using the sandwich variance. C and I denote the average number of correct and incorrect zero estimates, respectively. True β equals 2.0.

	Bias	RMSE	SD_1	SD_2	Coverage	C	I
			Majority rule holds				
1 invalid IV							
PSEE	-0.049	0.174	0.146	0.167	88.0%	8.628	0.142
SEE	0.281	0.355	0.204	0.218	76.7%		
Oracle	-0.062	0.163				9	0
Naive	-0.990	0.991					
3 invalid IVs							
PSEE	-0.018	0.194	0.159	0.194	87.2%	6.481	0.347
SEE	0.291	0.365	0.217	0.220	77.4%		
Oracle	-0.049	0.168				7	0
Naive	-0.949	0.951					
			Majority rule is violated				
5 invalid IVs							
PSEE	0.075	0.254	0.177	0.243	89.1%	4.268	0.815
SEE	0.323	0.410	0.233	0.253	76.2%		
Oracle	-0.010	0.177				5	0
Naive	-0.831	0.835					
7 invalid IVs							
PSEE	0.113	0.242	0.172	0.215	89.1%	2.715	2.188
SEE	0.323	0.400	0.228	0.236	75%		
Oracle	0.061	0.191				3	0
Naive	-1.056	1.057					