Statistica Sinica Preprint No: SS-2024-0006	
Title	Identifying Causal Effects Using Instrumental Variables from the Auxiliary Dataset
Manuscript ID	SS-2024-0006
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0006
Complete List of Authors	Kang Shuai, Shanshan Luo, Wei Li and Yangbo He
Corresponding Authors	Shanshan Luo
E-mails	shanshanluo@btbu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Identifying causal effects using instrumental variables from the auxiliary dataset

Kang Shuai[a], Shanshan Luo[b*], Wei Li[c] and Yangbo He[a]

[a]School of Mathematical Sciences, Peking University

[b]School of Mathematics and Statistics, Beijing Technology and Business University

[c]Center for Applied Statistics and School of Statistics, Renmin University of China

Abstract: Instrumental variable approaches have gained popularity for estimating causal effects in the presence of unmeasured confounders. However, the availability of instrumental variables in the primary dataset is often challenged due to stringent and untestable assumptions. This paper presents a novel method to identify and estimate causal effects by utilizing instrumental variables from the auxiliary dataset, incorporating a structural equation model, even in scenarios with nonlinear treatment effects. Our approach involves using two datasets: one called the primary dataset with joint observations of treatment and outcome, and another auxiliary dataset providing information about the instrument and treatment. Our strategy differs from most existing methods by not depending on the simultaneous measurements of instrument and outcome. The central idea for identifying causal effects is to establish a valid substitute through the auxiliary dataset, addressing unmeasured confounders. This is achieved by developing a

*Correspondence to: shanshanluo@btbu.edu.cn

control function and projecting it onto the function space spanned by the treatment variable. We then propose a three-step estimator for estimating causal effects and derive its asymptotic results. We illustrate the proposed estimator through simulation studies, and the results demonstrate favorable performance. We also conduct a real data analysis to evaluate the causal effect between vitamin D status and body mass index.

Key words and phrases: Control function, Data fusion, Instrumental variable, Unmeasured confounder.

1. Introduction

Randomized controlled trials are generally considered as the gold standard for evaluating causal effects. However, conducting such trials may not always be feasible due to ethical concerns or practical constraints like cost. In such situations, observational data can be used as an alternative for estimating causal effects. The major challenge in observational studies is the presence of unmeasured confounders, which may often introduce bias and invalidate the conclusions. Instrumental variables have been extensively employed to address such issues [Angrist et al., 1996, Ogburn et al., 2015]. An instrumental variable is a pretreatment variable that satisfies certain criteria. It should be associated with the treatment variable, independent of unmeasured confounders, and only affect the outcome through the treat-

ment variable. It is important to note that the instrumental variable can only be correlated with the treatment while not directly affecting the outcome.

Instrumental variable methods are commonly used to identify and estimate parameters in linear structural equation models (SEMs). When these models are correctly specified, the population average treatment effect corresponds to a specific parameter in the SEMs and can be consistently estimated. The two-stage least squares (2SLS) method and the control function method are two prominent methods in the context of SEMs [Goldberger, 1972, Wooldridge, 2010]. For both methods, the treatment is regressed on the instrument and baseline covariates in the first stage. However, in the second stage, the 2SLS method regresses the outcome on the predicted value of the treatment and baseline covariates, while the control function method regresses the outcome on the treatment, the residual from the first stage regression, and the baseline covariates. The 2SLS method aims to construct a function of the instrument and baseline covariates that is independent of unmeasured confounders, while the control function method incorporates the first stage residual, known as the control function, to control for unmeasured confounders [Imbens and Wooldridge, 2007, Petrin and Train, 2010, Wooldridge, 2015]. Nonparametric or semiparametric identification

of causal effects with an instrumental variable has also been sufficiently investigated in many researches, we refer readers to Angrist et al. [2013], Ogburn et al. [2015], Wang and Tchetgen Tchetgen [2018].

In practice, due to various challenges, it is often difficult to gather complete information on treatments, outcomes, and instrumental variables. Consequently, the utilization of auxiliary data for identifying causal effects is a common strategy in instrumental variable analysis. One of the most widely used method in this context is the two-sample instrumental variable framework [Angrist and Krueger, 1992, Arellano and Meghir, 1992], extensively employed in econometrics, social sciences, and mendelian randomization studies [Inoue and Solon, 2010, Sun and Miao, 2022, Gamazon et al., 2015, Zhao et al., 2019, 2020]. The primary dataset provides data related to instruments and outcomes, whereas the auxiliary dataset provides information about the treatment and instrumental variables. The corresponding two-sample instrumental variable estimators utilize sample moments from both an instrument-treatment sample and an instrument-outcome sample to estimate causal effects. However, stringent assumptions might render a valid instrumental variable unavailable within the primary dataset. A more practical situation is that researchers can only jointly observe the treatment and the outcome that are subject to unmeasured confounders. For example,

when assessing the effect of vitamin D deficiency on body mass index (BMI) in the primary dataset, commonly used instrumental variables like genetic factors might not be available. Meanwhile, there might be an auxiliary dataset that has collected data on both filaggrin mutation and vitamin D for other experimental purposes. The filaggrin mutation plays a significant role in skin barrier function, which suggests a strong association with vitamin D. Moreover, it is commonly acknowledged that the filaggrin mutation does not have a direct effect on BMI [Skaaby et al., 2013]. Consequently, the filaggrin mutation can be considered as a candidate instrumental variable, motivating further exploration of its potential to identify and estimate causal effect within the primary dataset.

In this paper, we introduce a novel approach for identifying and estimating treatment effects with an instrumental variable from the auxiliary dataset. Unlike conventional two-sample instrumental variable methods, we do not require the simultaneous measurements of instrument and outcome in the primary dataset. Instead, for a specific treatment variable that may be confounded in the primary dataset, our approach relies on the presence of an auxiliary dataset that includes a valid instrument and the treatment variable. The presence of unobserved confounders and the absence of instrumental variables commonly occur in the primary dataset, motivating

us to utilize the instrumental variables in the auxiliary dataset for identifying causal effects. This supplementary dataset is typically available from various data sources, containing randomized experiments or observational studies. Although such ideas are straightforward and practical, but as far as we know it has not appeared in the literature. Even within the classical SEM framework, how can we effectively evaluate treatment effects without simultaneous measurements of both instrument and outcome?

The control function approach within the SEMs provides valuable insights into the effective utilization of instrumental variables in the auxiliary dataset. In this paper, we adopt the control function perspective and establish a set of sufficient conditions to guarantee the identification of treatment effects. Specifically, by projecting the control function onto the function space spanned by the treatment variable, we can construct a potentially valid substitute for unmeasured confounders, namely, the control function projection. By incorporating the control function projection variable into the outcome model, we can effectively remove the impact of unmeasured confounders, leading to the identification of the treatment effect. Importantly, our method is also applicable to nonlinear treatment effects, and the conditions considered in this paper are expected to be no stronger than the similar assumptions made by Imbens and Wooldridge [2007] and Guo and

Small [2016]. Based on the identification results, we propose a three-step estimator for estimating the causal effect. We demonstrate its consistency of rate $n^{-1/2}$ under certain regularity conditions, subject to the requirement that the control function projection exhibits a uniform convergence rate of at least $n^{-1/4}$.

The remaining sections of this paper are organized as follows. In Section 2, we present the notation and outline the proposed model. Section 3 presents a brief review of the control function and introduces sufficient conditions for identifying causal effects. Section 4 provides a three-step estimator and establishes its asymptotic results. To evaluate the empirical performance of the proposed estimator, we conduct a simulation study in Section 5. Furthermore, in Section 6, we apply the proposed estimation procedure to real-world vitamin D datasets. The extension and further discussions are presented in Section 7 and 8, respectively.

2. Notation and Model

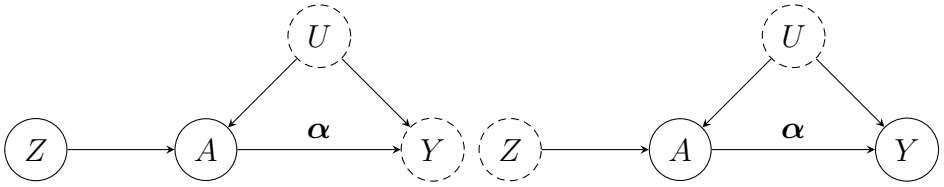
We assume that A denotes a scalar continuous treatment, Y denotes a scalar continuous outcome, $\boldsymbol{U}$ denotes a t-dimensional vector of unmeasured confounders, and Z denotes an instrumental variable that can be either discrete or continuous. For notational simplicity, we condition on

covariates implicitly and firstly omit them below. We adopt the potential outcomes framework to define causal effects and make the stable unit treatment value assumption (SUTVA) throughout the paper, that is, there is only one version of the potential outcomes and there is no interference between units [Rubin, 1980]. The SUTVA allows us to uniquely define the potential outcome Y_a for the outcome if the treatment A is set to be a. Suppose we have two datasets: the primary dataset, denoted as $\mathcal{O}_2 = \{(A_j, Y_j) : j \in \mathcal{S}_2\}$ with $n_2 = |\mathcal{S}_2|$ samples, and the auxiliary dataset, denoted as $\mathcal{O}_1 = \{(Z_i, A_i) : i \in \mathcal{S}_1\}$ with $n_1 = |\mathcal{S}_1|$ samples. For simplicity, we assume that $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, and we denote $r = \lim_{n_2 \to \infty} n_2/n_1 \in (0, \infty)$. We assume that all data are independently and identically distributed for $i \in \mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. Let R_i be an indicator variable, where $R_i = 1$ if the ith unit is from the primary dataset and $R_i = 0$ otherwise. In our analysis, we initially assume that the selection indicator R_i is independent of the variables (Z_i, A_i, Y_i), and later explore the potential relaxation of this assumption in Section 7. Therefore, the combined set of observed data can be represented as $\mathcal{O} = \{(R_i, R_i Y_i, Z_i - Z_i R_i, A_i) : i \in \mathcal{S}\}$, consisting of total $n = n_1 + n_2$ samples. We propose the following model, where $g(\cdot)$ represents

a vector of known linearly independent functions:

$$\begin{aligned} Y &= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{g}(A) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U} + \eta, \\ A &= m(Z) + \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon, \end{aligned} \tag{2.1}$$

where two error terms and unmeasured confounders satisfying $\varepsilon \perp\!\!\!\perp (Z, \boldsymbol{U})$ and $\eta \perp\!\!\!\perp (Z, A, \boldsymbol{U})$, $\mathrm{var}(\boldsymbol{U}) = \boldsymbol{I}_t$ and $\mathrm{var}(\varepsilon) = \sigma^2$. Here, $\eta \perp\!\!\!\perp (Z, A, \boldsymbol{U})$ can be relaxed to $E(\eta \mid Z, A, \boldsymbol{U}) = 0$. However, the independence of ε and $(Z, \boldsymbol{U})$ is important to guarantee identification of $\boldsymbol{\alpha}$ as illustrated by several examples in Section 3.2. Without loss of generality, throughout this paper, we assume that $\{\boldsymbol{U}, \eta, \varepsilon, A, \boldsymbol{g}(A), Y\}$ have zero means. Model (2.1) includes a class of structural equation models (SEMs). For instance, when $\boldsymbol{g}(A) = A$ and $\alpha \in \mathbb{R}^1$, model (2.1) is the widely used linear structural equation model in instrumental variable analysis, where the parameter α represents the causal effect of A on Y for a unit increase of A. Model (2.1) also includes scenarios where the function $\boldsymbol{g}(A)$ may contain nonlinear terms of A, and such nonlinear treatment effects are frequently observed in practice, as discussed by Guo and Small [2016] and Li and Guo [2020]. When Z is binary, the term $m(Z)$ can be characterized by the saturated model, $m(Z) = \gamma_0 + \gamma_1 Z$, and it must be linear with respect to Z. Figure 1 provides a graphical illustration using two observational datasets $\mathcal{O}_1$ and



(a) $\mathcal{O}_1$ with observed instrument and treatment.

(b) $\mathcal{O}_2$ with observed treatment and outcome.

Figure 1: Causal diagrams for two different datasets, one without outcome and the other without instrument. The dashed circles indicate that the corresponding variable is unobserved in the dataset. For simplicity, we have omitted the baseline covariates and selection indicator.

$\mathcal{O}_2$.

Given model (2.1), the causal effect of treatment A on outcome Y can be quantified by comparing the potential outcomes when we change A from a to a', that is, $Y_{a'} - Y_a = \boldsymbol{\alpha}^{\mathrm{T}}\{\boldsymbol{g}(a') - \boldsymbol{g}(a)\}$. Throughout this paper, our primary focus is on identifying and estimating the parameter $\boldsymbol{\alpha}$. As mentioned earlier, in common linear models, $\boldsymbol{\alpha}$ can represent the true causal effect. However, in general nonlinear models, the parameter $\boldsymbol{\alpha}$ represents the change in the potential outcome corresponding to the treatment changing from a to a'. When $\boldsymbol{g}(A)$ is a vector, even with the complete information of (Z, A, Y) available, the two-stage least squares method may not yield a consistent estimator of $\boldsymbol{\alpha}$ without additional conditions [Guo and Small, 2016]. Specifically, within the two-stage least squares framework, it is crucial to ensure that the conditional expectation vector $E\{\boldsymbol{g}(A) \mid Z\}$ is linearly inde-

pendent in order to estimate the parameter $\boldsymbol{\alpha}$. However, this condition can be easily violated when the instrument Z is binary, making direct regression of Y on $E\{\boldsymbol{g}(A) \mid Z\}$ infeasible. In contrast, the control function approach provides an alternative method for estimating $\boldsymbol{\alpha}$ even with a binary instrument Z. By incorporating the vector $\boldsymbol{g}(A)$ into $\{\boldsymbol{g}(A)^{\mathrm{T}}, A - m(Z)\}^{\mathrm{T}}$, we ensure that the resulting vector remains linearly independent, which further leads to more precise estimates of causal parameter $\boldsymbol{\alpha}$ [Imbens and Wooldridge, 2007].

3. Identification

3.1 Review of control function approach

In this section, we will give a brief review of the control function approach, which provides a different perspective for estimating possibly nonlinear treatment effects compared to the two-stage least squares method. The control function approach is a widely adopted method for identifying and estimating linear or nonlinear treatment effects with instrumental variables [Heckman, 1976, Rivers and Vuong, 1988, Imbens and Wooldridge, 2007, Petrin, 2011, Cai et al., 2011, Wooldridge, 2015, Guo and Small, 2016, Li and Guo, 2020]. In this section, we start by considering a straightforward scenario involving a complete dataset, where measurements for the instru-

mental variable, treatment variable, and outcome variable are all collected simultaneously. We will begin by reviewing the control function approach within the context of a linear outcome model, and then proceed to extend it to the nonlinear case. Specifically, in the simplest case where $\boldsymbol{g}(A) = A$, we define $\boldsymbol{U}_{\text{proj}}$ to signify the linear projection of $\boldsymbol{U}$ on $A - m(Z)$:

$$\begin{aligned}\boldsymbol{U}_{\text{proj}} &= \text{cov}\{\boldsymbol{U}, A - m(Z)\} \cdot \text{cov}^{-1}\{A - m(Z)\} \cdot \{A - m(Z)\} \\ &= \frac{\boldsymbol{l}}{\boldsymbol{l}^{\text{T}}\boldsymbol{l} + \sigma^2}\{A - m(Z)\}.\end{aligned}$$

The outcome model can be rewritten as follows:

$$\begin{aligned}Y &= \alpha A + \boldsymbol{\beta}^{\text{T}}\boldsymbol{U}_{\text{proj}} + \boldsymbol{\beta}^{\text{T}}(\boldsymbol{U} - \boldsymbol{U}_{\text{proj}}) + \eta \\ &= \alpha A + \frac{\boldsymbol{\beta}^{\text{T}}\boldsymbol{l}}{\boldsymbol{l}^{\text{T}}\boldsymbol{l} + \sigma^2}\{A - m(Z)\} + \boldsymbol{\beta}^{\text{T}}(\boldsymbol{U} - \boldsymbol{U}_{\text{proj}}) + \eta.\end{aligned} \tag{3.1}$$

By the definition of linear projection and $Z \perp\!\!\!\perp (\boldsymbol{U}, \varepsilon)$, we have that $\text{cov}\{A - m(Z), \boldsymbol{U} - \boldsymbol{U}_{\text{proj}}\} = 0$ and

$$\text{cov}\left\{m(Z), \boldsymbol{U} - \boldsymbol{U}_{\text{proj}}\right\} = \text{cov}\left\{m(Z), \boldsymbol{U} - \frac{\boldsymbol{l}}{\boldsymbol{l}^{\text{T}}\boldsymbol{l} + \sigma^2}(\boldsymbol{l}^{\text{T}}\boldsymbol{U} + \varepsilon)\right\} = 0.$$

The above two conditions also imply the zero covariance between A and $\boldsymbol{U} - \boldsymbol{U}_{\text{proj}}$, that is, $\text{cov}\{A, \boldsymbol{U} - \boldsymbol{U}_{\text{proj}}\} = 0$. Therefore, $\boldsymbol{\beta}^{\text{T}}(\boldsymbol{U} - \boldsymbol{U}_{\text{proj}}) + \eta$

in (3.1) can be regarded as a new error term, which is uncorrelated with A and $A - m(Z)$. Hence, if we regress Y on A and $A - m(Z)$, then the coefficient of A will be equal to the true causal effect α. This method is the so-called control function approach, which has been previously discussed in Heckman [1976].

However, the outcome model may also be nonlinear with respect to the treatment, the direct regression procedure may not be sufficient for identifying causal parameter $\boldsymbol{\alpha}$ due to the potential non-zero covariance between $\boldsymbol{g}(A)$ and $\boldsymbol{U} - \boldsymbol{U}_{\text{proj}}$. In such cases, the identifiability condition outlined by Imbens and Wooldridge [2007], and further discussed by Guo and Small [2016] and Li and Guo [2020], becomes essential for identifying and estimating the parameter $\boldsymbol{\alpha}$. We summarize the condition as follows:

Condition 1. $E(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}+\eta \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U}+\varepsilon) = \rho(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{U}+\varepsilon)$, where ρ is some constant.

A sufficient condition for Condition 1 is that the conditional expectation $E(\boldsymbol{U} \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon)$ is linear in $\boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon$. Condition 1 implies that regressing Y on $\boldsymbol{g}(A)$ and the residual $A - m(Z)$ will give the consistent estimator of $\boldsymbol{\alpha}$,

as shown below:

$$\begin{aligned} E(Y \mid Z, A) &= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{g}(A) + E(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U} + \eta \mid Z, A) \\ &= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{g}(A) + E(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U} + \eta \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) \\ &= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{g}(A) + \rho\{A - m(Z)\}, \end{aligned} \tag{3.2}$$

where the last equality holds due to model (2.1). However, equation (3.2) still requires the joint measurements of the instrument and outcome, and it does not directly address the specific problem encountered in Figure 1. In the upcoming section, we present a series of novel conditions to identify the causal parameter $\boldsymbol{\alpha}$ using the combined dataset $\mathcal{O}$ without directly imposing Condition 1.

3.2 Identification using auxiliary dataset

To identify the causal effect $\boldsymbol{\alpha}$, we introduce a conditional expectation $C(A)$, defined as $C(A) = E\{A - m(Z) \mid A\} = E(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon \mid A)$. Obviously, the conditional expectation $C(A) = A - E\{E(A \mid Z, R = 0) \mid A, R = 0\}$ can be identified using the auxiliary dataset $\mathcal{O}_1$. The term $A - m(Z) = \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon$ is known as the control function [Heckman, 1976], and $C(A)$ represents the projection of the control function onto the function space spanned by the treatment variable A. Intuitively, the control function projection variable

$C(A)$ captures the information of unmeasured confounders $\boldsymbol{U}$ entailed by treatment A, and can be considered as a potentially valid substitute for unmeasured confounders $\boldsymbol{U}$. We consider the following regular condition:

Assumption 1. The vector $\boldsymbol{g}(A)$ and $C(A)$ are linearly independent.

The plausibility of Assumption 1 can be assessed using the observed data, since the conditional expectation $C(A)$ is identifiable and the vector $\boldsymbol{g}(A)$ is known. When $\boldsymbol{g}(A) = A$, this assumption implies that the term $E\{m(Z) \mid A\}$ is nonlinear with respect to A, which is often easily satisfied when Z is binary.

As shown in (3.2), the control function approach for a linear treatment effect with the full dataset $\mathcal{F} = \{(Z_i, A_i, Y_i) : i = 1, ..., n\}$ does not require any additional assumptions. However, when $\boldsymbol{g}(A)$ contains a nonlinear term of A, even with the full dataset $\mathcal{F}$, additional assumptions must be introduced, such as Condition 1 [Guo and Small, 2016]. Nevertheless, Condition 1 does not address the challenge of lacking the joint observation of instrumental and outcome variables. To address this issue, we propose the following assumption:

Assumption 2. $E(\boldsymbol{U} \mid A) = \boldsymbol{v}E(\varepsilon \mid A)$, where $\boldsymbol{v}$ is a constant vector.

It is important to mention that the constant vector $\boldsymbol{v}$ can be calcu-

lated as $\boldsymbol{v} = \boldsymbol{l}/\sigma^2$, based on the fact $E(A\boldsymbol{U}) = \boldsymbol{v}E(A\varepsilon)$. Assumption 2 plays a pivotal role in our subsequent discussions on identifiability. This assumption is crucial not only for ensuring the identification of the causal parameter $\boldsymbol{\alpha}$, but also for addressing the data fusion problem in cases where the instrument and outcome variables are not observed jointly. In the context of model (2.1) with a scalar unmeasured confounder, Assumption 2 is indeed weaker compared to Condition 1, as illustrated in the following Example 1. We also provide a mathematical relaxation of this assumption in the supplementary material. Additionally, Assumption 2 imposes constraints on the joint distribution of the unmeasured confounder $\boldsymbol{U}$ and the noise ε, requiring the two corresponding conditional expectations with respect to A to be proportional. A similar idea has also been previously considered by Shuai et al. [2023] for identifying causal effects in the presence of unmeasured confounders. It is important to note that the conditional expectation $E(\boldsymbol{U} \mid A) = E_{Z|A}\{E(\boldsymbol{U} \mid Z, \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U}+\varepsilon)\} = E_{Z|A}\{E(\boldsymbol{U} \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U}+\varepsilon)\}$ and $E(\varepsilon \mid A) = E_{Z|A}\{E(\varepsilon \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon)\}$, which implies that we only need to derive the proportional equality as a sufficient condition when conditional on $\boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon$. When this holds, $E(\boldsymbol{U} \mid A)$ will be proportional to $E(\varepsilon \mid A)$. We now provide additional examples for illustrating the plausibility and applicability of Assumption 2.

Example 1. Under our setting, Assumption 4 in Guo and Small [2016] can be interpreted as:

$$E(\beta U + \eta \mid lU + \varepsilon) = \rho(lU + \varepsilon),$$

where $\rho \in \mathbb{R}^1 \backslash \{0\}$ is a constant. This equality is essentially Condition 1 stated in Section 3.1. This entails that $E(U \mid lU+\varepsilon)$ is linear in $lU+\varepsilon$. Since $E(lU + \varepsilon \mid lU + \varepsilon)$ is obviously linear in $lU + \varepsilon$, we thus have $E(\varepsilon \mid lU + \varepsilon)$ is also linear in $lU + \varepsilon$. Given $Z \perp\!\!\!\perp (U, \varepsilon)$, we immediately have:

$$E\{U \mid m(Z), A\} = E\{U \mid m(Z), lU + \varepsilon\} = E(U \mid lU + \varepsilon),$$
$$E\{\varepsilon \mid m(Z), A\} = E\{\varepsilon \mid m(Z), lU + \varepsilon\} = E(\varepsilon \mid lU + \varepsilon).$$

Therefore, two conditional expectations $E\{U \mid m(Z), A\}$ and $E\{\varepsilon \mid m(Z), A\}$ should also be linear in $lU + \varepsilon$. Then there must exist some constant v satisfying

$$E\{U \mid m(Z), A\} = vE\{\varepsilon \mid m(Z), A\}.$$

This implies Assumption 2.

Example 1 illustrates that the conditions considered in this paper are expected to be no stricter than Condition 1 within the context of a scalar

unmeasured confounder [Imbens and Wooldridge, 2007, Guo and Small, 2016]. We next provide specific examples of the possible joint distribution of $(\boldsymbol{U}^{\mathrm{T}}, \varepsilon)^{\mathrm{T}}$ based on Assumption 2 in a more general case.

Example 2 (Elliptical Distribution)**.** Suppose the joint probability density function of $\boldsymbol{\zeta} = (\boldsymbol{U}^{\mathrm{T}}, \varepsilon)^{\mathrm{T}}$ has the following elliptical distributional form:

$$p_{\boldsymbol{\zeta}}(x) = k \cdot \varphi\{(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\},$$

where k is the normalizing constant, $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix, and $\varphi(\cdot)$ is some known function. By Theorem 2.18 in Fang et al. [2018], we have:

$$E(\boldsymbol{\zeta} \mid \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\zeta}) = \boldsymbol{a}_{\boldsymbol{B}} + \boldsymbol{M}_{\boldsymbol{B}} \boldsymbol{B}^{\mathrm{T}} \zeta,$$

where $\boldsymbol{a}_{\boldsymbol{B}} \in \mathbb{R}^{t+1}$ is some constant vector and $\boldsymbol{B} \in \mathbb{R}^{(t+1)\times r}$ is some constant matrix with full column rank. Because $E(\boldsymbol{\zeta}) = \boldsymbol{0}$, we must have $\boldsymbol{a}_{\boldsymbol{B}} = \boldsymbol{0}$. Specifically, let $\boldsymbol{B} = (\boldsymbol{l}^{\mathrm{T}}, 1)^{\mathrm{T}}$, this implies:

$$E(\boldsymbol{\zeta} \mid \boldsymbol{l}^{\mathrm{T}} \boldsymbol{U} + \varepsilon) = \boldsymbol{M}_{\boldsymbol{B}} (\boldsymbol{l}^{\mathrm{T}} \boldsymbol{U} + \varepsilon).$$

Thus, there must exist some constant vector $\boldsymbol{v} \in \mathbb{R}^t$ satisfying

$$E\{\boldsymbol{U} \mid m(Z), A\} = E(\boldsymbol{U} \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) = vE(\varepsilon \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) = vE\{\varepsilon \mid m(Z), A\}.$$

This implies Assumption 2.

Elliptical distributions play an important role in various fields, including portfolio theory [Owen and Rabinovitch, 1983]. Example 2 covers a broad class of distributions satisfying Assumption 2, such as the multivariate normal distribution, multivariate t-distribution, multivariate Laplace distribution, and so on [Fang et al., 2018]. The elliptical distribution can be either bounded or unbounded, depending on the choice of the function $\varphi(\cdot)$. Additionally, the Pearson system of distributions can also fulfill Assumption 2 [Kotz et al., 2004]. The following example provides another case that guarantees Assumption 2 without requiring a specific distributional form.

Example 3. Let $\boldsymbol{l} = (1, ..., 1)^{\mathrm{T}}$, Assumption 2 can also be satisfied when $U_1, ..., U_t, \varepsilon$ are independent and identically distributed. Specifically, we have

$$E(U_1 \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) = \cdots = E(U_t \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) = E(\varepsilon \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon),$$

where $\boldsymbol{U} = (U_1, ..., U_t)^{\mathrm{T}}$. This means

$$E(\boldsymbol{U} \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) = \boldsymbol{v}E(\varepsilon \mid \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon) \text{ and } \boldsymbol{v} = \boldsymbol{l}/\sigma^2.$$

This implies Assumption 2.

The example above illustrates that if the underlying causal mechanisms of unmeasured variables $U_1, U_2, ..., U_t$, and ε are identical with respect to the treatment variable A, then Assumption 2 also hold. As a result, the causal parameter $\boldsymbol{\alpha}$ can be identified without imposing any restrictions on their joint distribution.

Theorem 1. *Under Assumptions 1-2, the causal parameter* $\boldsymbol{\alpha}$ *is identified by*

$$\boldsymbol{\alpha} = \boldsymbol{D}_{11}E\{\boldsymbol{g}(A)Y \mid R = 1\} + \boldsymbol{D}_{12}E\{C(A)Y \mid R = 1\}, \tag{3.3}$$

where $\boldsymbol{D}_{11}$ *and* $\boldsymbol{D}_{12}$ *are the corresponding block matrices of the variance matrix* $\boldsymbol{D}$*:*

$$\boldsymbol{D} = \left[E\left\{\boldsymbol{h}(A)\boldsymbol{h}(A)^{\mathrm{T}}\right\}\right]^{-1} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{pmatrix}, \quad \boldsymbol{h}(A) = \{\boldsymbol{g}(A)^{\mathrm{T}}, C(A)\}^{\mathrm{T}}.$$

As discussed below Assumption 1, the control function projection $C(A)$ can be obtained using the auxiliary dataset $\mathcal{O}_1$. Once we have derived the projection $C(A)$, we can directly perform a regression of Y on $h(A)$ in the dataset $\mathcal{O}_2$. The coefficient of $\boldsymbol{g}(A)$ can then be used to identify the causal parameter $\boldsymbol{\alpha}$. The theorem essentially requires that either the instrument or the outcome variable is missing completely at random (MCAR), namely, $R \perp\!\!\!\perp (Z, A, Y)$. However, our situation is more complex compared to the traditional MCAR framework. Whether restricted to the subpopulation with $R = 1$ or $R = 0$, we cannot obtain a complete observed dataset, making it challenging to directly apply standard missing data analysis methods [Rubin, 1976]. To partially relax this MCAR assumption, we will address the same identification issue even when the selection indicator R depends on certain observed variables, as discussed in Section 7. By the way, we provide the following Corollary to include baseline covariates W into analysis.

Corollary 1. *Let W represent the baseline covariates, replace Assumption 2 by $E(U \mid A, W) = vE(\varepsilon \mid A, W)$ and assume $g(A), E\{m(Z, W) \mid A, W\}$ are linearly independent and the following model holds*

$$\begin{aligned} Y &= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{g}(A) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U} + h(W) + \eta, \\ A &= m(Z, W) + \boldsymbol{l}^{\mathrm{T}}\boldsymbol{U} + \varepsilon, \end{aligned} \tag{3.4}$$

where $\boldsymbol{g}(\cdot)$ *is known,* $h(\cdot)$ *is unknown and* $(\boldsymbol{U}, \varepsilon) \perp\!\!\!\perp (Z, W)$*. Then we can conclude that* $\boldsymbol{\alpha}$ *is identifiable.*

4. Estimation and Inference

In this section, we omit the covariates and first describe an estimation procedure in full generality.

Step 1. Obtain a treatment model $\hat{m}(Z)$ for $m(Z)$ based on the auxiliary dataset $\mathcal{O}_1$.

Step 2. Project the residue $A - \hat{m}(Z)$ onto the function space of A to derive the control function projection estimator $\hat{C}(A)$ for $C(A)$ based on the auxiliary dataset $\mathcal{O}_1$.

Step 3. Plug the estimate $\hat{C}(A)$ into (3.3) to estimate $\boldsymbol{\alpha}$ based on the primary dataset $\mathcal{O}_2$.

In Step 1, estimating the treatment model $m(Z) = E(A \mid Z)$ can be achieved parametrically or nonparametrically using standard density estimation techniques. This step is relatively straightforward. However, Step 2 poses a more significant challenge as it involves solving a reverse estimation problem for $C(A) = A - E\{m(Z) \mid A, R = 0\}$. In most literature, the primary focus remains on estimating $m(Z)$, representing the forward

estimation problem in Step 1, which also aligns with the data generation mechanism. Nevertheless, even if a parametric model for $m(Z)$ is known, or in the case where Z is binary, obtaining the true parametric model for $C(A)$ can still be challenging. In practice, we recommend utilizing non-parametric estimation techniques. Estimating Step 3 is standard in causal inference problems. We will outline some regularity conditions required for Step 2, which are sufficient to ensure the asymptotic results for the causal parameter $\boldsymbol{\alpha}$.

Assumption 3. (i) $\hat{C}(A)$ is uniformly consistent, namely, $\sup_{A\in\mathcal{A}}|\hat{C}(A) - C(A)| = o_p(1)$; and (ii) $\sup_{A\in\mathcal{A}}|\hat{C}(A) - C(A)| = o_p(n_1^{-1/4})$.

The conditions imposed in Assumption 3 are standard regularity conditions used to establish the asymptotic results. Assumption 3(ii) is commonly imposed in the causal inference literature to derive the asymptotic distribution of the estimators when the nuisance functions are estimated with certain convergence rates. For example, if $\hat{C}(A)$ is estimated based on a correctly specified parametric model, $\hat{C}(A)$ is $\sqrt{n_1}$-consistent by using maximum likelihood estimation or generalized moments of estimation [Hansen, 1982]. When $m(Z)$ and $\hat{C}(A)$ are estimated nonparametrically, it is often expected that the estimators can achieve the rate $o_p(n_1^{-1/4})$. Specifically, we can adopt the classical linear sieve methods with certain linearly in-

dependent basis functions. Except for the sieve method, popular machine learning approaches, including kernel regression or random forest, can also be easily implemented to achieve the desired square root consistency. Furthermore, while sample splitting or cross-fitting techniques are not required for our estimation, they provide a valuable perspective for alleviating some technical conditions in future work [Chernozhukov et al., 2018]. Although Assumption 3(ii) related to the sample size n_1, it is important to note that we assume the ratio n_2/n_1 of the two sample sizes tends to a constant r. This implies that condition $\sup_{A\in\mathcal{A}}|\hat{C}(A) - C(A)| = o_p(n_2^{-1/4})$ also holds.

In the final part of this section, we will outline simpler conditions to ensure that Assumption 3 is satisfied within the context of a linear model framework for $m(Z)$. These conditions provide insights into solving the classical linear structural equation model more effectively. Without the linear structure of $m(Z)$, we refer readers to the two-step sieve M estimator previously discussed by Hahn et al. [2018].

Theorem 2. *Under Assumptions 3(i), we have that* $\hat{\boldsymbol{\alpha}}$ *is consistent for* $\boldsymbol{\alpha}$ *as* $n_2 \to \infty$. *Additionally, suppose Assumption 3(ii) holds, then* $\hat{\boldsymbol{\alpha}}$ *is asymptotically normal, namely,* $\sqrt{n_2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} N(0, \boldsymbol{V})$ *as* $n_2 \to \infty$, *where*

$$\boldsymbol{V} = \mathrm{var}\left\{\boldsymbol{D}_{11}\boldsymbol{g}(A)Y + \boldsymbol{D}_{12}C(A)Y + \frac{\partial\boldsymbol{\alpha}(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}X\right\},$$

$$\boldsymbol{\alpha}(\boldsymbol{\mu}) = \boldsymbol{D}_{11}E\{\boldsymbol{g}(A)Y\} + \boldsymbol{D}_{12}E\{C(A)Y\},$$

$$\boldsymbol{X} = \left[\text{vec}\{\boldsymbol{g}(A)\boldsymbol{g}(A)^{\mathrm{T}}\}^{\mathrm{T}}, \boldsymbol{g}(A)^{\mathrm{T}}C(A), C(A)^2\right]^{\mathrm{T}}, \quad \boldsymbol{\mu} = E(\boldsymbol{X}).$$

The above theorem demonstrates that the consistency of $\hat{\boldsymbol{\alpha}}$ can be achieved when a uniformly consistent control function projection $\hat{C}(A)$ is available. Furthermore, the $\sqrt{n_2}$-consistency of $\hat{\boldsymbol{\alpha}}$ requires the uniform convergence rate of $\hat{C}(A)$ to be at least faster than $n_2^{-1/4}$. In this scenario, estimating $C(A)$ has a negligible impact on the asymptotic variance of $\sqrt{n_2}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})$. Theorem 2 also implies the asymptotic variance of $\sqrt{n_2}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})$ will be the same even if we know the true control function projection $C(A)$.

At the end of this subsection, we provide further discussion about Assumption 3 by specifically considering the case where $m(Z)$ follows a linear model. This choice is particularly useful as it covers all cases where Z is binary. The decision to impose a linear restriction on $m(Z)$ is beneficial because it simplifies the estimation process in Step 1. Without loss of generality, we assume that $\hat{m}(Z) = \hat{\gamma}_0 + \hat{\gamma}_1 Z$, where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ can be obtained through linear regression. However, despite this simplification, there are still numerous challenges that need to be addressed in the following step.

In Step 2 of our analysis, given the linear model for $m(Z)$, we can derive

the following expression using straightforward algebra:

$$\hat{C}(A) = A - \hat{\gamma}_0 - \hat{\gamma}_1 \hat{E}(Z \mid A). \tag{4.1}$$

Estimating $E(Z \mid A)$ will be somewhat simpler than directly estimating $C(A)$, given its more direct expression. To derive $\hat{E}(Z \mid A)$ in (4.1), we suggest utilizing nonparametric techniques or other machine learning approaches, such as kernel estimation [Gasser and Müller, 1979] and random forests [Breiman, 2001]. Similar estimation approaches can be employed for further estimating the $\boldsymbol{\alpha}$ in Step 3. The asymptotic results for $\hat{\boldsymbol{\alpha}}$ are straightforward, as Assumption 3 is satisfied by the estimator $\hat{C}(A)$ constructed in (4.1). To summarize:

Corollary 2. *If $\hat{E}(Z \mid A)$ is uniformly consistent and $E(Z \mid A)$ is bounded, then Assumption 3(i) holds. Also, when $\sup_{A \in \mathcal{A}} |\hat{E}(Z \mid A) - E(Z \mid A)| = o_p(n_1^{-1/4})$, Assumption 3(ii) holds.*

5. Simulation Studies

To evaluate the finite sample performance of the proposed three-step estimator, we firstly generate data according to the following model:

$$\text{Scenario 1}: A = \gamma Z + lU + \varepsilon, \quad Y = \alpha A + \beta U + \eta,$$

$$\text{Scenario 2}: A = \gamma Z + lU + \varepsilon, \quad Y = \alpha A^2 + \beta U + \eta.$$

In Scenario 1, we consider the linear treatment effect setting, while in Scenario 2, we examine the nonlinear treatment effect setting. The goal is to demonstrate the consistency of our proposed estimator in various situations involving different distributional combinations of U and ε, as well as different sample sizes. We have n_2 samples for (A, Y) from the primary dataset and another n_1 samples for (Z, A) from the auxiliary dataset. For both scenarios, we consider different combinations of n_1 and n_2, where $n_1, n_2 \in \{5000, 10000\}$, and we use either a binary or continuous instrument variable Z. We first set $\alpha = \gamma = \beta = 1$ and assume $(U, \varepsilon) \sim \mathrm{N}(0, I_2)$ with $l = 0.5$. In this case, (U, ε) belongs to an elliptically contoured distribution, satisfying Assumption 2 from the previous Example 2. Next, we fix $l = 1$ and consider $U, \varepsilon \sim \exp(1)$ or $U(-1, 1)$. These distributions also satisfy Assumption 2 from Example 3. Additional simulation settings, in-

cluding larger sample sizes or other forms of distribution combinations, are provided in the supplementary material. We also display the simulation results for control function approach with complete dataset at the end of the supplement.

For both scenarios, we employ the estimation method proposed in Section 4 for analysis. Intuitively, the estimation performance will be affected by the choice of the nonparametric estimation basis function, the distributional form of (Z, A), and the sample sizes n_1 and n_2. Table 1 presents the bias, mean squared error (MSE), and 95% coverage probability of our estimator with 500 bootstrap resampling iterations and 500 Monte Carlo runs. Across all scenarios, the bias for our estimator is found to be very small, indicating that it provides unbiased estimates of the causal effect $\boldsymbol{\alpha}$. Additionally, the coverage probability of our estimator improves as n_1 and n_2 increase, particularly when n_2 is increased. This observation aligns with the convergence performance of $\hat{\boldsymbol{\alpha}}$ with respect to the sample size n_2, as demonstrated in Theorem 2. The MSE varies under different distributional combinations. For example, when (U, ε) are normally distributed and Z follows a Bernoulli or uniform distribution, the MSE is relatively larger compared to other cases. This difference in MSE might be attributed to the chosen estimation method for $C(A)$. There is also an interesting observa-

Table 1: Bias, mean square error (MSE), and 95% coverage probability (95% CP) of $\hat{\boldsymbol{\alpha}}$ based on 500 repetitions. The bias and MSE have been multiplied by 100.

(n_1, n_2)	Metrics	Scenario 1					
		Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6
$(5000, 5000)$	Bias	0.123	0.264	0.607	0.650	0.199	0.771
	MSE	1.820	46.101	0.125	0.177	1.130	48.648
	95% CP	95.4	97.2	94.8	93.8	96.2	98.2
$(5000, 10000)$	Bias	0.930	0.865	0.856	0.130	0.361	2.686
	MSE	0.971	22.769	0.084	0.110	0.622	25.968
	95% CP	96.0	96.6	93.4	94.4	95.4	98.0
$(10000, 5000)$	Bias	1.763	3.105	0.267	0.125	1.320	3.589
	MSE	1.781	45.471	0.083	0.167	1.079	44.511
	95% CP	94.4	96.0	94.2	94.8	96.4	96.6
$(10000, 10000)$	Bias	0.497	1.482	0.511	0.145	0.002	1.589
	MSE	0.909	25.123	0.060	0.086	0.585	24.015
	95% CP	96.0	94.4	95.0	95.2	95.6	96.4
(n_1, n_2)	Metrics	Scenario 2					
		Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6
$(5000, 5000)$	Bias	0.045	0.066	0.072	0.021	0.027	0.070
	MSE	0.013	0.008	0.002	0.002	0.012	0.007
	95% CP	92.4	96.0	94.2	96.6	92.2	96.0
$(5000, 10000)$	Bias	0.005	0.019	0.084	0.034	0.017	0.024
	MSE	0.007	0.004	0.001	0.001	0.006	0.004
	95% CP	93.2	94.8	94.4	94.8	93.4	94.8
$(10000, 5000)$	Bias	0.019	0.000	0.025	0.051	0.015	0.014
	MSE	0.012	0.009	0.001	0.002	0.012	0.008
	95% CP	93.0	95.0	95.8	93.0	93.2	93.6
$(10000, 10000)$	Bias	0.063	0.020	0.056	0.035	0.060	0.034
	MSE	0.005	0.004	0.001	0.001	0.005	0.004
	95% CP	95.6	94.2	93.4	94.4	95.0	93.6

Setting 1 & 2: $l = 1$, $Z \sim B(0.5)$, $U \sim \exp(1)$; $l = 0.5$, $Z \sim B(0.5)$, $U \sim N(0, 1)$;
Setting 3 & 4: $l = 1$, $Z \sim \exp(1)$, $U \sim U(-1, 1)$; $l = 0.5$, $Z \sim \exp(1)$, $U \sim N(0, 1)$;
Setting 5 & 6: $l = 1$, $Z \sim U(-1, 1)$, $U \sim \exp(1)$; $l = 0.5$, $Z \sim U(-1, 1)$, $U \sim N(0, 1)$;

tion in Table 1: Scenario 2, which incorporates nonlinearity, demonstrates superior performance compared to Scenario 1, which adopts a linear treatment model. Overall, as sample size increases, the performance results demonstrate the theoretic results discussed in Section 4.

6. Application

We illustrate the proposed method using two different datasets. The primary dataset $\mathcal{O}_2$ is derived from the National Health and Nutrition Examination Survey (NHANES) program conducted by the Centers for Disease Control and Prevention (CDC) in the United States during 2011-2012. Vitamin D deficiency has been proven to be closely associated with many common diseases, such as diabetes or cancer. In this study, our main focus is on estimating the treatment effect of vitamin D (the treatment A) status on BMI (the outcome Y), aiming to explore the potential causal relationship between vitamin D deficiency and obesity. However, some crucial confounders, such as gene expressions, which could affect both the vitamin D status and BMI, have not been included in the dataset $\mathcal{O}_2$. This implies that the treatment effect estimate may suffer from severe bias if these confounders are not considered. After removing missing values and outliers, our analysis includes 7539 individual samples.

We consider another dataset from the population-based study Monica10 as the auxiliary dataset $\mathcal{O}_1$. The Monica10 study was previously conducted in 1982-1984 and contained examinations of 3785 individuals of Danish origin, recruited from the Danish Central Personal Register. Specifically, this dataset contains five important background variables, including vitamin D

status, filaggrin genotype, mortality, age, and follow-up time. The mutations in the filaggrin gene are associated with a higher vitamin D status through an increased UV sensitivity of keratinocytes [Skaaby et al., 2013]. Therefore, we treat the indicator of filaggrin mutation as the instrumental variable (Z) since filaggrin plays a crucial role in the skin barrier function but does not seem to affect BMI directly. After removing individuals with missing information and outliers, we finally include 2571 individuals in our analysis [Martinussen et al., 2019].

In this section, we apply our method to evaluate the effect of vitamin D status on BMI. The binary instrumental variable Z takes two values, 0 and 1, respectively representing the two most common null mutations of the filaggrin gene, including R501X and 2282del4. We employ the estimation method proposed in Section 5 for our analysis. The corresponding assumptions can be empirically verified by the implementation process of our method. For example, Assumption 1 can be directly satisfied when the control function projection $C(A)$ is nonlinear in A for linear treatment effect cases. The results are shown in Table 2, where we provide the point estimate of α, along with the bootstrap standard errors and 95% confidence interval calculated using the corresponding z-score. The quantiles and histogram of the corresponding bootstrap results are also provided in the supplementary

Table 2: Point estimate, bootstrap standard error (SE), and 95% confidence interval for the causal effect of vitamin D status on BMI.

	Point estimate	SE	95% confidence interval
The proposed method	-0.6386	14.2409	(-28.5508, 27.2736)
Naive method	-0.0418	0.0034	(-0.0484, -0.0352)

material. For comparison, we have also included the naive results obtained through the least square in the second row of Table 2. We note a significant negative effect of vitamin D on BMI using naive estimates. However, vitamin D is widely used in practice, and no studies suggest that vitamin D has significant side effects. Our proposed method indicates that vitamin D status has no significant effect on BMI, which is consistent with previous findings reported by Duan et al. [2020]. However, possibly due to the lack of important pre-treatment covariates related to distinct timeframes and locations, the primary and auxiliary datasets may not be completely independent and identically distributed. Thus, we suggest practitioners collect more covariates information and further consider Model (3.4) and the methods proposed in Section 7 for analysis.

7. Extension

In previous sections, we assumed that the selection indicator variable $R \perp\!\!\!\perp (Z, A, Y)$, implying that the corresponding missing mechanism is missing completely at random [Rubin, 1976]. However, even in cases where the

selection indicator R depends on some observed variables, it remains feasible to attain the identification of causal parameter $\boldsymbol{\alpha}$. We summarize this in the following assumption.

Assumption 4. $R \perp\!\!\!\perp (Z, Y) \mid A$.

Assumption 4 can be viewed as a form of missing at random (MAR) assumption, where the selection indicator R is allowed to depend on treatment variable A. This assumption covers scenarios in which the two data sources $\mathcal{O}_1$ and $\mathcal{O}_2$ share overlapping samples that provide treatment information, a situation commonly encountered in practice. Intuitively, since the treatment variable A is allowed to be observed in both datasets, Assumption 4 can still sufficiently establish identification results similar to Theorem 1.

Theorem 3. *Under Assumptions 1, 2 and 4, the causal parameter* $\boldsymbol{\alpha}$ *can be identified as follows*

$$\boldsymbol{\alpha} = \boldsymbol{D}_{11} E\{\boldsymbol{g}(A) E(Y \mid A, R = 1)\} + \boldsymbol{D}_{12} E\{C(A) E(Y \mid A, R = 1)\},$$

where the conditional expectation $C(A)$ *can be identified through* $C(A) = A - E\{E(A \mid Z) \mid A, R = 0\}$, *and the conditional expectation* $E(A \mid Z)$ *can*

be identified through $E(A \mid Z = z) = \int a f(a \mid z) da$ *with*

$$f(a \mid z) = \frac{f(z \mid a, r = 0) f(a)}{f(z)}, \quad f(z) = \int f(z \mid a, r = 0) f(a) da.$$

Based on the identification results presented in Theorem 3, we can develop an estimation procedure for causal parameter $\boldsymbol{\alpha}$ and provide its asymptotic properties using a similar approach as described in Section 4. For the sake of simplicity, we omit the detailed methodology in this section.

8. Discussion

In this article, we consider identifying and estimating the causal effect using an instrumental variable from auxiliary dataset. Existing researches often rely on joint observations of the instrumental variable and outcome, such as two-sample instrumental estimators, to identify treatment effects. However, this poses challenges when the instrumental variable is not available in the primary dataset. To address this issue, we propose a novel identification strategy from the control function perspective. We further consider the estimation and asymptotic theory for the proposed estimator.

There are several potential extensions that can be explored in future researches. Firstly, if additional information, such as proxy variables or

multiple candidate instrumental variables, becomes available, the identification conditions may be further relaxed or extended to allow for nonparametric identification [Kang et al., 2016, Miao et al., 2018]. Secondly, the scenario with binary treatment variables or non-continuous outcome variables is pretty common in practice, and it would be interesting to extend our model to accommodate such cases [Wang and Tchetgen Tchetgen, 2018]. The difficulty is about how to propose a similar condition like the proportional conditional expectation equality under linear additive model. Finally, investigating causal inference in settings with multiple or high-dimensional treatments holds practical and theoretical importance. These are all left for future work because they are beyond the scope of this paper.

Supplementary Material

The supplementary material available online includes additional technical proofs and simulation results.

Acknowledgements

We sincerely thank the editor, associate editor, and reviewers for their insightful and helpful comments, which have significantly improved our paper. Kang Shuai and Yangbo He are supported by the National Key R&D

Program of China (2022ZD0160300). Wei Li is supported by the Beijing Natural Science Foundation (1232008), the National Natural Science Foundation of China (12101607, 12071015), the National Key R&D Program of China (2022YFA1008100), and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD910001). Shanshan Luo is supported by the Disciplinary funding of Beijing Technology and Business University and the Research Foundation for Youth Scholars of Beijing Technology and Business University.

References

Joshua Angrist, Ivan Fernandez-Val, Daron Acemoglu, Manuel Arellano, and D Eddie. *Advances in Economics and Econometrics*. Cambridge University Press, 2013.

Joshua D Angrist and Alan B Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87:328–336, 1992.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.

Manuel Arellano and Costas Meghir. Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. *The Review of Economic Studies*, 59: 537–559, 1992.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Bing Cai, Dylan S Small, and Thomas R Ten Have. Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. *Statistics in Medicine*, 30:1809–1824, 2011.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21:C1 –

C68, 2018.

Leizhen Duan, Ling Han, Qin Liu, Yili Zhao, Lei Wang, and Yan Wang. Effects of vitamin d supplementation on general and central obesity: results from 20 randomized controlled trials involving apparently healthy populations. *Annals of Nutrition and Metabolism*, 76: 153–164, 2020.

Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 2018.

Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47:1091–1098, 2015.

Theo Gasser and Hans-Georg Müller. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop Held in Heidelberg*, pages 23–68. Springer, 1979.

Arthur S Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001, 1972.

Zijian Guo and Dylan S Small. Control function instrumental variable estimation of nonlinear causal effect models. *The Journal of Machine Learning Research*, 17:3448–3482, 2016.

Jinyong Hahn, Zhipeng Liao, and Geert Ridder. Nonparametric two-step sieve m estimation and inference. *Econometric Theory*, 34:1281–1324, 2018.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

Guido Imbens and Jeffrey Wooldridge. Control function and related methods. *What's new in Econometrics*, 2007.

Atsushi Inoue and Gary Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92:557–561, 2010.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables es-

timation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111:132–144, 2016.

Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004.

Sai Li and Zijian Guo. Causal inference for nonlinear outcome models with possibly invalid instrumental variables. *arXiv preprint arXiv:2010.09922*, 2020.

Torben Martinussen, Ditte Nørbo Sørensen, and Stijn Vansteelandt. Instrumental variables estimation under a structural cox model. *Biostatistics*, 20:65–79, 2019.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105:987–993, 2018.

Elizabeth L Ogburn, Andrea Rotnitzky, and James M Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:373–396, 2015.

Joel Owen and Ramon Rabinovitch. On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance*, 38:745–752, 1983.

Amil Petrin. Revisiting instrumental variables and the classic control function approach, with implications for parametric and non-parametric regressions kyoo il kim. Working Paper, 2011.

Amil Petrin and Kenneth Train. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47:3–13, 2010.

Douglas Rivers and Quang H Vuong. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39:347–366, 1988.

Donald B Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

Donald B Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75:591–593, 1980.

Kang Shuai, Shanshan Luo, Yue Zhang, Feng Xie, and Yangbo He. Identification and estimation of causal effects using non-gaussianity and auxiliary covariates. *arXiv preprint arXiv:2304.14895*, 2023.

Tea Skaaby, Lise Lotte Nystrup Husemoen, Torben Martinussen, Jacob P Thyssen, Michael Melgaard, Betina Heinsbæk Thuesen, Charlotta Pisinger, Torben Jørgensen, Jeanne D

Johansen, Torkil Menné, et al. Vitamin d status, filaggrin genotype, and cardiovascular risk factors: a mendelian randomization approach. *PloS one*, 8:e57647, 2013.

BaoLuo Sun and Wang Miao. On semiparametric instrumental variable estimation of average treatment effects through data fusion. *Statistica Sinica*, 32:569–590, 2022.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:531–550, 2018.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

Jeffrey M Wooldridge. Control function methods in applied econometrics. *Journal of Human Resources*, 50:420–445, 2015.

Qingyuan Zhao, Jingshu Wang, Wes Spiller, Jack Bowden, and Dylan S Small. Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science*, 34:317–333, 2019.

Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S Small. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *Annals of Statistics*, 48:1742–1769, 2020.