Statistica Sinica **36** (2026), 1-20
doi:https://doi.org/10.5705/ss.202024.0006

IDENTIFYING CAUSAL EFFECTS USING INSTRUMENTAL VARIABLES FROM THE AUXILIARY DATASET

Kang Shuai[1], Shanshan Luo*[2], Wei Li[3] and Yangbo He[1]

[1]*Peking University,* [2]*Beijing Technology and Business University and* [3]*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 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, Imbens and Rubin, 1996; Ogburn, Rotnitzky and Robins, 2015). An instrumental variable is a pretreatment variable that satisfies certain

*Corresponding author. E-mail: shanshanluo@btbu.edu.cn