Statistica Sinica **35** (2025), 1-22
doi:https://doi.org/10.5705/ss.202022.0288

A DATA FUSION METHOD FOR QUANTILE TREATMENT EFFECTS

Yijiao Zhang and Zhongyi Zhu*

Fudan University

Abstract: With the increasing availability of datasets, developing data fusion methods to leverage the strengths of different datasets to draw causal effects is of great practical importance to many scientific fields. In this paper, we consider estimating the quantile treatment effects using small validation data with fully-observed confounders and large auxiliary data with unmeasured confounders. We propose a Fused Quantile Treatment effects Estimator (FQTE) by integrating the information from two datasets based on doubly robust estimating functions. We allow for the misspecification of the models on the dataset with unmeasured confounders. Under mild conditions, we show that the proposed FQTE is asymptotically normal and more efficient than the initial QTE estimator using the validation data solely. By establishing the asymptotic linear forms of related estimators, convenient methods for covariance estimation are provided. Simulation studies demonstrate the empirical validity and improved efficiency of our fused estimators. We illustrate the proposed method with an application.

Key words and phrases: Calibration, causal inference, double robustness, estimation equation, unmeasured confounder.

1. Introduction

The increasing availability of datasets from multiple sources holds enormous promise for evaluating causal effects. With various datasets at hand, data fusion technology has become more and more important in many medical and biological applications. How to systematically combine multiple datasets sources in an attempt to leverage the strengths of different types of data to improve the estimating efficiency of causal effects is gathering notice from researchers. For example, there are data sources with large sample size, such as electronic health records, claims databases, disease data registries, and census data. However, uncontrolled design mechanisms and limited information on baseline covariates may lead to confounding bias, presenting a major threat to causal inference. In practice, there are also small validation datasets that include all possible confounders and provide detailed information for each individual, especially in some randomized controlled trials (RCTs) in the medical field. A classic example is a two-phase study (Wang et al., 2009), where less expensive covariates are

*Corresponding author. E-mail: zhuzy@fudan.edu.cn