Propensity Score Weighting with Mismeasured Covariates: An Application to Two Financial Literacy Interventions

Hao Dong*
Southern Methodist University
Daniel L. Millimet
Southern Methodist University & IZA

November 20, 2020

Abstract

Estimation of the causal effect of a binary treatment on outcomes often requires conditioning on covariates to address selection on observed variables. This is not straightforward when one or more of the covariates are measured with error. Here, we present a new semi-parametric estimator that addresses this issue. In particular, we focus on inverse propensity score weighting estimators when the propensity score is of an unknown functional form and some covariates are subject to classical measurement error. Our proposed solution involves deconvolution kernel estimators of the propensity score and the regression function weighted by a deconvolution kernel density estimator. Simulations and replication of a study examining the impact of two financial literacy interventions on the business practices of entrepreneurs show our estimator to be valuable to empirical researchers.

JEL: C18, C21, G21, G53
Keywords: Program Evaluation, Measurement Error, Propensity Score, Unconfoundedness, Financial Literacy

*Corresponding author: Hao Dong, Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496. Tel: (214) 768-3288. E-mail: haod@smu.edu.

1 Introduction

Empirical researchers in economics, finance, management, and other disciplines are often interested in the causal effect of a binary treatment on outcomes. In some cases, randomization is used to ensure comparability across the treatment and control groups. However, researchers must rely on observational data when randomization is not feasible. With observational data, concern over the non-random selection of subjects into the treatment group becomes well-founded. Addressing the possibility of non-random selection requires much of the data at hand. Moreover, even with randomization, demands on the data may be non-trivial since randomization only balances covariates across the treatment and control groups in expectation.

In this paper, we consider the case where adjustment for observed covariates is performed to recover an unbiased estimate of the effect of a treatment. Thus, we are restricting ourselves to the case of *selection on observed variables*. The econometric and statistics literature on the estimation of causal effects in the case of selection on observed variables has grown tremendously of late.[1] This has led to the proliferation of statistical methods designed to estimate the causal effect(s) of the treatment, including parametric regression methods, semi- or non-parametric methods based on the propensity score, and combinations thereof.

Despite the growing number of estimation methods, there are only a few that take into account measurement errors in the data. Here, we present a new semi-parametric estimator that partially fills this gap. In particular, we focus on the case when the propensity score is of an unknown functional form and some covariates are subject to classical measurement error. There are two issues to be dealt with to estimate the treatment effect in such a situation: first, we need to estimate the functional form of the propensity score; second, we need to estimate the moment of a known (or estimated) function of mismeasured covariates. The first issue is solved by using deconvolution kernel regression. For the second issue, as the sample analogue is no longer feasible due to the unobservability of the error-free covariates, we consider the integration weighted by deconvolution kernel density estimator.

We illustrate our estimator both via simulation and by revisiting the randomized control trial (RCT) on financial literacy examined in Drexler et al. (2014). In the experiment, micro-entrepreneurs taking out a loan from ADOPEM, a microfinance institution in the Dominican Republic, are randomly assigned to one of three treatment arms to assess the causal effect of financial literacy programs on a firm's financial practices, objective reporting quality, and business performance. The first treatment provided subjects with standard accounting training. The second treatment provided rule-of-thumb training that covered basic financial heuristics. The final group received neither training and serves as the control group. The

[1] See Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018) for excellent surveys.

authors find significant beneficial effects of the rule-of-thumb training, but not the standard accounting training.

We revisit this study for three reasons. First, proper evaluation of financial literacy interventions is critical. As documented in Lusardi and Mitchell (2014), financial literacy in the US and elsewhere seems woefully inadequate for individuals and small business owners to navigate complex financial matters. McKenzie and Woodruff (2013, p. 48-49) offer the following vivid description:

> "Walk into a typical micro or small business in a developing country and spend a few minutes talking with the owner, and it often becomes clear that owners are not implementing many of the business practices that are standard in most small businesses in developed countries. Formal records are not kept, and household and business finances are combined. Marketing efforts are sporadic and rudimentary. Some inventory sits on shelves for years at a time, whereas more popular items are frequently out of stock. Few owners have financial targets or goals that they regularly monitor and act to achieve."

As evidenced in this quote, the lack of financial literacy among micro-entrepreneurs has real consequences. Lusardi and Mitchell (2014) discuss the wider impacts of a lack of financial literacy, such as lower participation in financial markets, poor investment decisions, susceptibility to financial scams, inadequate retirement planning, increased credit card and mortgage debt, etc. See Morgan and Trinh (2019) for a recent example.

While the impacts are well-documented, knowledge of the efficacy of various programs aiming to teach financial literacy is inadequate. Specifically, the causal effect of specific types of financial literacy training interventions is relatively unexplored. Existing research typically lumps all financial literacy programs together, potentially masking insights into what works and what does not. For example, Fernandes et al. (2015) perform a meta-analysis of 201 studies assessing the impact of financial literacy and education programs on financial behaviors, finding that interventions explain only 0.1% of the variance in financial behaviors. By comparing two different training programs, Drexler et al. (2014) represent an important contribution to the literature.

Second, better understanding the determinants of successful microenterprises is critical in lesser developed countries. Berge et al. (2015, p. 707) state: "Microenterprises are an important source of employment, and developing such enterprises is a key policy concern in most countries, and in particular in developing countries where they employ more than half of the labor force." However, the viability of microenterprises has been found to be heterogenous, as the authors further note that "a growing literature shows that success cannot be taken for granted" (p. 707). Recent research has focused on sources of this heterogeneity, finding that it is not explained fully by variation in capital (Bruhn et al. 2018). The study

by Drexler et al. (2014) addresses this issue by exploring the impact of different types of financial literacy training on firm success.

Finally, our proposed estimator is well-suited to the application. To start, despite training being randomly assigned, the authors control (via regression) for several covariates to increase the precision of the treatment effect estimates. Moreover, one covariate is continuous and potentially suffers from classical measurement error. This covariate reflects the size of the loan received by the entrepreneur. While this variable is unlikely to be mismeasured as it is obtained from bank records, arguably the 'true' covariate of interest is a measure of capital investment in the firm by the entrepreneur. This could be below the official size of the loan due to some funds being diverted to non-business use, or above the official size of the loan due to other funds being used to supplement the loan. As Drexler et al. (2014, p. 2) note, "for microenterprises the boundary between business and personal financial decisions is often blurred."

Applying our proposed estmator, we find the results in Drexler et al. (2014) to be generally robust to 'modest' amounts of measurement error. However, for a few outcomes, the magnitude of the estimated treatment effect changes. With greater amounts of measurement error, the results are not surprisingly less robust. Typically in such cases we find larger point estimates once measurement error is addressed.

The remainder of the paper is organized as follow. Section 2 provides a brief overview of the literature on measurement error in covariates. Section 3 provides an overview of the potential outcomes framework, discusses identification with and without measurement error in covariates, and presents our proposed estimator. Section 4 studies the small sample properties of the proposed estimators by simulation. Section 5 contains our application to assessment of two financial literacy interventions. Section 6 concludes.

2 Measurement Error in Covariates

A small literature has considered measurement error in an observed covariate when estimating the causal effect of a treatment in the case of selection on observed variables. In a regression context with classical measurement error, it is well known that the Ordinary Least Squares (OLS) estimate of the coefficient on the mismeasured regressor suffers from attenuation bias (see, e.g., Frisch 1934; Koopmans 1937; Reiersøl 1950). However, bias will also impact the estimated treatment effect if treatment assignment is correlated with the true value of the mismeasured covariate (Bound et al. 2001). The sign of this covariance determines the sign of the bias. If the measurement error is correlated with treatment assignment (i.e., it is nonclassical), then the direction of the bias depends on whether the partial correlation between the measurement error and treatment assignment is positive or negative (Bound et al. 2001). Finally, if multiple covariates suffer from measurement error, then one is typically unable to sign the bias even under classical measurement error (Bound et al. 2001).

With classical measurement error, a consistent estimate of the treatment effect can be recovered using Instrumental Variable (IV) estimation, where the mismeasured covariate(s) are instrumented for using valid exclusion restrictions. However, this solution places further demands on the data as valid instruments must be available. As an aside, it is also important to realize that the estimated treatment effect will still be inconsistent if treatment assignment is correlated with the measurement error (Bound et al. 2001).

Beyond the regression context, several recent papers consider the effect of measurement error in one or more covariates when relying on semi- or non-parametric estimators of the treatment effect. Battistin and Chesher (2014), extending early work in Cochran and Rubin (1973), focus on the bias of treatment effect parameters estimated using semiparametric (propensity score) methods. The bias, which may be positive or negative, is a function of the measurement error variance. The authors consider bias-corrected estimators where the bias is estimated under different assumptions concerning the reliability of the data.

McCaffrey et al. (2013) develops a consistent inverse propensity score weighted estimator for the case when covariates are mismeasured. In particular, the authors consider a weight function of mismeasured covariates whose conditional expectation given the correctly measured covariates equals the error-free inverse propensity score. Their estimator is then constructed based on approximating the weight function by projecting the inverse of the estimated propensity score onto a set of basis functions. To estimate the propensity score with mismeasured covariates, knowledge of the measurement error distribution is generally needed. It is worth noting that the measurement error considered in this paper could be non-classical, as only conditional independence between the measurement error and the outcome and the treatment given the correctly measured covariates is required. As a cost to this extra flexibility, the authors only establish consistency; further characterization of the asymptotic properties are left as a gap to be filled.

Jakubowski (2015) assesses the performance of propensity score matching when an unobserved covariate is proxied by several variables. The author considers two estimation methods. The first is a propensity score matching estimator where the propensity score model includes the proxy variables. The second is also a propensity score matching model except now the propensity score model includes an estimate of the unobserved covariate obtained via a factor analysis approach.

Webb-Vargas et al. (2017) examines the performance of inverse propensity score weighting with a mismeasured covariate. The authors then consider an inverse propensity score weighting estimator that replaces the mismeasured covariate with multiple imputations. The imputations make use of an auxiliary data source that contains both the true covariate and the mismeasured covariate. Each imputation leads to a unique propensity score model and hence a distinct estimate of the treatment effect. These multiple estimates are then combined into a final estimate.

Rudolph and Stuart (2018) assess the performance of three approaches to deal with measurement

error in covariates when applying propensity score estimators. The first approach is propensity score calibration which, similar to Webb-Vargas et al. (2017), relies on an auxialiary data source that contains both the true covariate and the mismeasured covariate. The second approach is a bias-corrected technique based on the fomulas derived in VanderWeele and Arah (2011). As these bias formulas depend on various unknown sensitivity parameters, this technique relies on either external data to make educated guesses concerning the values of these parameters or sensitivity analysis using a grid of plausible values. The final approach is similar and relies on the sensitivity (to unobserved confounders) approach of Rosenbaum (2010) for matched pairs.

Hong et al. (2019) perform an extensive simulation exercise to explore the impact of multiple mismeasured covariates with and without correlated measurement errors. The authors find that correlation in the measurement errors can exacerbate the bias and that including auxiliary variables that are correlated with the true values of the mismeasured covariates may help mitigate the bias.

In sum, it is now well known that measurement error in covariates that belong in the propensity score model introduces bias in the estimated treatment effect. While a few solutions have been proposed, these solutions have not completely solved the problem. Some solutions rely on auxiliary data that contain both the true and mismeasured covariates. Other solutions are based upon bias-corrected estimates requiring the specification of parameter values whose true values are typically unknown. Finally, some solutions are based on trying to reduce the bias through the use of multiple proxies or assessing how severe the measurement error would have to be to explain the treatment effect ignoring measurement error.

Compared to most of these existing works, our estimator has the advantage of not requiring a specific functional form of the propensity score so that it can avoid the bias caused by potential model misspecification. McCaffrey et al. (2013) is an exception as it also treat the propensity as a nonparametric object. However, instead of the consistency established in McCaffrey et al. (2013), our estimator allows us to further characterize the differences in the convergence rates when the measurement errors are of different smoothness.

For the technical aspects, this paper contributes to the vast literature on estimating the non-/semi-parametric measurement error models using deconvolution. See books by Meister (2009) and Horowitz (2009) and surveys by Chen, Hong and Nekipelov (2011) and Schennach (2016) for reviews. This literature started with the density estimation; See Carroll and Hall (1988), Stefanski and Carroll (1990), Fan (1991a,b), Bissantz, Dümbgen, Holzmann and Munk (2007), van Es and Gugushvili (2008), Lounici and Nickl (2011) among others. The deconvolution approach used to estimate the density later extend to the estimation of regression function; See Fan and Truong (1993), Fan and Masry (1992), Delaigle and Meister (2007), Delaigle, Fan and Carroll (2009) and Delaigle, Hall and Jamshidi (2015). Like works in other semi-parametric setups (Fan, 1995; Dong, Otsu and Taylor, 2020b), our estimator is constructed using the

deconvolution kernel estimators of both the density and the regression function as the building blocks.

3 Empirics

3.1 Potential Outcomes Framework

Our analysis is couched within the potential outcomes framework (see, e.g., Neyman 1923; Fisher 1935; Roy 1951; Rubin 1974). We consider a random sample of N individuals from a large population, where individuals are indexed by $j = 1, ..., n$. Define $Y_j(T)$ to be the potential outcome of individual j under treatment T, $T \in \mathcal{T}$.[2] In this paper, we limit ourselves to binary treatments: $\mathcal{T} = \{0, 1\}$. The causal effect of the treatment for a given individual is defined as the individual's potential outcome under treatment ($T = 1$) relative to the individual's potential outcome under control ($T = 0$). Formally,

$$\tau_j = Y_j(1) - Y_j(0). \tag{1}$$

In the evaluation literature, several population parameters are of potential interest. Here, attention is given to the average treatment effect (ATE)

$$\tau = E[\tau_j] = E[Y_j(1) - Y_j(0)]$$

and the average treatment effect for the treated (ATT)

$$\tau_{\text{treat}} = E[\tau_j | T = 1] = E[Y_j(1) - Y_j(0) | T = 1].$$

The ATE is the expected treatment effect of an observation chosen at random from the population, whereas the ATT is the expected treatment effect of an observation chosen at random from the treatment group.

Each observation is characterized by the quadruple $\{Y_j, T_j, X_j, Z_j\}$, where Y_j is the observed outcome, T_j is a binary indicator of the treatment received, X_j is a scalar covariate, and Z_j is a d-dimensional vector of covariates. The covariates included in X_j and Z_j must be pre-determined (i.e., they are not affected by T_j) and must not perfectly predict treatment assignment. The observed outcomes is

$$Y_j = T_j Y_j(1) + (1 - T_j) Y_j(0) \tag{2}$$

which makes clear that only one potential outcome is observed for any individual. Absent randomization, τ and τ_{treat} are not identified in general due to the selection problem, that is the distribution of $(Y(0), Y(1))$

[2] We assume that the Stable Unit Treatment Value Assumption (SUTVA), where potential outcomes of individuals do not depend on the treatment assignment of others, to hold (Neyman 1923; Rubin 1986).

may depend on T. Even with randomization, the efficiency of estimates can be improved by incorporating the covariates.

3.2 Strong Ignorability

To overcome the selection problem, or to improve the efficiency of estimates obtained under randomization, a set of fully observed covariates are commonly assumed, conditional on which $(Y(0), Y(1))$ and T are independent. This is referred to as the *conditional independence* or *unconfoundedness* assumption (Rubin 1974; Heckman and Robb 1985). Formally, this assumption is expressed as

Assumption 1. $(Y(0), Y(1)) \perp T|(X, Z)$.

In addition to Assumption 1, the following *overlap* or *common support* assumption concerning the joint distribution of treatment assignment and covariates is also needed. Let $p_{X,Z}(x, z) = P(T = 1|X = z, Z = z)$ denote the propensity score, and $\mathcal{X}$ and $\mathcal{Z}$ denote supports of X and Z respectively.

Assumption 2. $0 < p_{X,Z}(x, z) < 1$ *for all* $(x, z) \in \mathcal{X} \times \mathcal{Z}$.

Assumptions 1 and 2 are jointly referred to as *strong ignorability* in Rosenbaum and Rubin (1983) and lead to the following well known result

$$\tau = E\left[\frac{(T - p_{X,Z}(X, Z))Y}{p_{X,Z}(X, Z)(1 - p_{X,Z}(X, Z))}\right] \tag{3}$$

$$\tau_{\text{treat}} = E\left[\frac{(T - p_{X,Z}(X, Z))Y}{p(1 - p_{X,Z}(X, Z))}\right], \tag{4}$$

where $p = P(T = 1)$ is the probability of getting treated; See Proposition 18.3 of Wooldridge (2010). Thus, strong ignorability is sufficient to identify the estimands, τ and τ_{treat}, when all variables are accurately measured.

3.3 Strong Ignorability with Measurement Error

Consider the case where Assumptions 1 and 2 continue to hold, but the quadruple $\{Y_j, T_j, W_j, Z_j\}$ is observed by the researcher instead of $\{Y_j, T_j, X_j, Z_j\}$. Here, the observed scalar, W_j, is assumed to be a noisy measure of X_j, generated by

$$W_j = X_j + \epsilon_j,$$

where ϵ_j is measurement error. Let f_V denote the density of a random variable V and $f^{\text{ft}}(t) = \int e^{\mathrm{i}tx} f(x)dx$ denote the Fourier transform of a function f with $i = \sqrt{-1}$. To identify τ and τ_{treat} in the presence of contaminated data, we impose the following assumption in addition to strong ignorability.

Assumption 3. $\epsilon \perp (Y, T, X, Z)$, *$f_\epsilon$ is known, and f_ϵ^{ft} vanishes nowhere.*

Assumption 3 requires the measurement error to be classical. Although this is somewhat restrictive, it is worth noting that this setup is consistent with multiplicative measurement error of the form $W = X\epsilon$, as this can be transformed to an additive structure by taking the natural logarithm. In fact, as argued in Schennach (2019), we do not need full independence; only $f_W^{\text{ft}}(t) = f_X^{\text{ft}}(t) f_\epsilon^{\text{ft}}(t)$ for all $t \in \mathbb{R}$ is necessary, which is as equally strong as a conditional mean restriction. The assumption of a known error distribution is unlikely to hold in practice, but is imposed here for simplicity. We discuss the relaxation of this assumption when auxiliary information is available, such as the repeated measurements of X, in Section 3.5.

The identification result in the presence of contaminated data is given in the following theorem.

Theorem 1. *Under Assumptions 1-3, τ and τ_{treat} are identified from $\{Y, T, W, Z\}$.*

The intuition behind Theorem 1 is straightforward. Based on (3) and (4), to identify τ and τ_{treat}, it is sufficient to identify $f_{Y,X,Z|T}$, which follows by implementing the convolution theorem to $f_{Y,W,Z|T}$ under Assumption 3.

Theorem 1 in McCaffrey et al. (2013) provides results that be used to achieve the point identification of τ and τ_{treat} under similar assumptions. In particular, using their Theorem 1, τ and τ_{treat} can be identified by (3) and (4) if the inverse propensity score is replaced by a non-stochastic function A of W and Z whose conditional expectation given X and Z equals $1/p_{X,Z}$, i.e. $E[A(W,Z)|X,Z] = 1/p_{X,Z}(X,Z)$, and A is needed to finally pin down τ and τ_{treat}. McCaffrey et al. (2013), however, do not provide further details on A except in very special cases.

Under Assumptions 1 - 3, where Assumption 3 is slightly stronger than Assumption 1 in McCaffrey et al. (2013), we can derive a general explicit form of their function A. For example, for $E[Y(1)]$, which is needed to construct τ, the function A is given by

$$A(w,z) = \frac{p \int e^{-\mathrm{i}tw} \frac{\{f_{W,Z|T=1}(\cdot,z)\}^{\text{ft}}(t)}{|f_\epsilon^{\text{ft}}(t)|^2} dt}{2\pi f_{W,Z}(w,z)}, \tag{5}$$

where $\{f_{W,Z|T=1}(\cdot,z)\}^{\text{ft}}(t) = \int e^{\mathrm{i}tw} f_{W,Z|T=1}(w,z)dw$. While it is not easy to give an intuitive interpretation of (5), using the law of iterated expectation, the result shown in Appendix A.2 implies that (5) is the equivalent quantity of the inverse propensity score in the contaminated case. As can be seen, the functional form of A depends on $f_{W,Z|T}$ and f_ϵ. The former, $f_{W,Z|T}$, is identified as $\{T, W, Z\}$ are directly observed, but extra knowledge on the latter, f_ϵ, is needed to identify A, which echos the known error distribution part of Assumption 3.

In fact, the functional form of A not only matters to the identification of τ and τ_{treat}, but also matters to the convergence rates of estimators of τ and τ_{treat}. In particular, as will be seen in Section 3.4, varying the smoothness of f_ϵ (implying f_ϵ^{ft} decays to zero at different rates as $t \to \infty$) alters the convergence rate. Intuitively, as f_ϵ^{ft} appears in the denominator of A, even if the same estimator of $f_{W,Z|T}$ is used to construct the estimator of A and then the estimators of τ and τ_{treat}, due to the integration, the resulting estimators of τ and τ_{treat} should converge at different speeds if f_ϵ^{ft} decays to zero at different rates.

3.4 Estimation

If we directly observe X, τ and τ_{treat} can be estimated by

$$\begin{aligned} \check{\tau} &= \frac{1}{n}\sum_{j=1}^{n} \frac{(T_j - \check{p}_{X,Z}(X_j, Z_j))Y_j}{(1-\check{p}_{X,Z}(X_j, Z_j))\check{p}_{X,Z}(X_j, Z_j)} \\ \check{\tau}_{\text{treat}} &= \frac{1}{n}\sum_{j=1}^{n} \frac{(T_j - \check{p}_{X,Z}(X_j, Z_j))Y_j}{\hat{p}\check{p}_{X,Z}(X_j, Z_j)}, \end{aligned}$$

where $\hat{p} = \frac{1}{n}\sum_{j=1}^{n} T_j$ and $\check{p}_{X,Z}$ is a nonparametric estimator of the propensity score, $p_{X,Z}$. These are known as the inverse propensity score weighting (IPW) estimators; see Horvitz and Thompson (1952). However, this estimator is no longer feasible when X is unobserved due to measurement error. To overcome this, note that we can alternatively express τ and τ_{treat} as

$$\tau = \iiint \left\{ \frac{p f_{Y,X,Z|T=1}(y,x,z)}{p_{X,Z}(x,z)} - \frac{(1-p) f_{Y,X,Z|T=0}(y,x,z)}{1-p_{X,Z}(x,z)} \right\} y dy dx dz, \tag{6}$$

$$\tau_{\text{treat}} = \iiint \left\{ f_{Y,X,Z|T=1}(y,x,z) - \frac{(1-p) p_{X,Z}(x,z) f_{Y,X,Z|T=0}(y,x,z)}{p(1-p_{X,Z}(x,z))} \right\} y dy dx dz, \tag{7}$$

Derivation of (6) and (7) are discussed in Appendix A.1. To keep the notation simple, we will focus on the case when both X and Z are scalar for the rest of this section. By applying the deconvolution method with f_ϵ known and given the i.i.d. sample $\{Y_j, T_j, W_j, Z_j\}_{j=1}^n$ of (Y, T, W, Z), the conditional densities $f_{Y,X,Z|T=1}(y,x,z)$ and $f_{Y,X,Z|T=0}(y,x,z)$ can be estimated by

$$\tilde{f}_{Y,X,Z|T=1}(x,y,z) = \frac{b_n^{-3}\sum_{j=1}^{n} T_j K\left(\frac{y-Y_j}{b_n}\right) \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right)}{\sum_{j=1}^{n} T_j}, \tag{8}$$

$$\tilde{f}_{Y,X,Z|T=0}(x,y,z) = \frac{b_n^{-3}\sum_{j=1}^{n} (1-T_j) K\left(\frac{y-Y_j}{b_n}\right) \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right)}{n - \sum_{j=1}^{n} T_j}, \tag{9}$$

and the propensity score $p_{X,Z}(x,z)$ can be estimated by

$$\tilde{p}_{X,Z}(x,z) = \frac{\sum_{j=1}^{n} T_j \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right)}{\sum_{j=1}^{n} \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right)}, \tag{10}$$

where b_n is a bandwidth, K is a (ordinary) kernel function, and $\mathbb{K}$ is a deconvolution kernel function defined as

$$\mathbb{K}(x) = \frac{1}{2\pi} \int e^{-itx} \frac{K^{\text{ft}}(t)}{f_\epsilon^{\text{ft}}(t/b_n)} dt.$$

Plugging (8), (9), and (10) into (6) and (7), we obtain estimators of τ and τ_{treat} as follow.

$$\tilde{\tau} = \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \frac{\tilde{q}_{1,1}(x,z)}{\tilde{q}_{0,1}(x,z)} - \frac{\tilde{q}_{1,0}(x,z)}{\tilde{q}_{0,0}(x,z)} \right\} \tilde{q}(x,z) dx dz, \tag{11}$$

$$\tilde{\tau}_{\text{treat}} = \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \tilde{q}_{1,1}(x,z) - \frac{\tilde{q}_{1,0}(x,z)\tilde{q}_{0,1}(x,z)}{\tilde{q}_{0,0}(x,z)} \right\} \frac{1}{\hat{p}} dx dz \tag{12}$$

where $\mathcal{X}$ and $\mathcal{Z}$ separately denote the support of X and Z, and

$$\tilde{q}(x,z) = \frac{1}{nb_n^2} \sum_{j=1}^{n} \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right),$$

$$\tilde{q}_{k,s}(x,z) = \frac{1}{nb_n^2} \sum_{j=1}^{n} Y_j^k T_j^s (1-T_j)^{1-s} \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right) \text{ for } k,s = 0,1.$$

Derivation of (11) and (12) are left to Appendix A.3.

Remark 1 (Case of vector X and Z)**.** *The proposed method can be easily generalized to case when X and Z are vector even though it is constructed based on the case when both X and Z are scalar. In particular, let $X = (X_1, \ldots, X_{d_x})$ and $Z = (Z_1, \ldots, Z_{d_z})$ be d_x- and d_z-dimensional vectors, respectively, and $W = (W_1, \ldots, W_{d_x})$ a d_x-dimensional vector of noisy measures of X generated by $W_d = X_d + \epsilon_d$ for $d = 1, \ldots, d_x$. Following a simiar route as when X and Z are both scalar, we can estimate τ and τ_{treat} by*

$$\tilde{\tau}' = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{dx}} \int_{\mathcal{Z}_1} \cdots \int_{\mathcal{Z}_{dz}} \left\{ \frac{\tilde{q}'_{1,1}(x,z)}{\tilde{q}'_{0,1}(x,z)} - \frac{\tilde{q}'_{1,0}(x,z)}{\tilde{q}'_{0,0}(x,z)} \right\} \tilde{q}'(x,z) dx dz,$$

$$\tilde{\tau}'_{\text{treat}} = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{dx}} \int_{\mathcal{Z}_1} \cdots \int_{\mathcal{Z}_{dz}} \left\{ \tilde{q}'_{1,1}(x,z) - \frac{\tilde{q}'_{1,0}(x,z)\tilde{q}'_{0,1}(x,z)}{\tilde{q}'_{0,0}(x,z)} \right\} \frac{1}{\hat{p}} dx dz,$$

where $\mathcal{X}_{d_1}$ and $\mathcal{Z}_{d_2}$ separately denote the support of X_{d_1} and Z_{d_2} for $d_1 = 1, \ldots, d_x$ and $d_2 = 1, \ldots, d_z$, $x =$

$(x_1, \ldots, x_{d_x})$, $z = (z_1, \ldots, z_{d_z})$, *and*

$$\tilde{q}'(x,z) = \frac{1}{nb_n^{d_x+d_z}} \sum_{j=1}^{n} \prod_{d=1}^{d_x} \mathbb{K}_d \left(\frac{x_d - W_{d,j}}{b_n} \right) \prod_{j=1}^{d_z} K \left(\frac{z_d - Z_{d,j}}{b_n} \right),$$

$$\tilde{q}'_{k,s}(x,z) = \frac{1}{nb_n^{d_x+d_z}} \sum_{j=1}^{n} Y_j^k T_j^s (1-T_j)^{1-s} \prod_{d=1}^{d_x} \mathbb{K}_d \left(\frac{x_d - W_{d,j}}{b_n} \right) \prod_{j=1}^{d_z} K \left(\frac{z_d - Z_{d,j}}{b_n} \right) \text{ for } k,s = 0,1,$$

with

$$\mathbb{K}_d(x) = \frac{1}{2\pi} \int e^{-\mathrm{i}tx} \frac{K^{\mathrm{ft}}(t)}{f^{\mathrm{ft}}_{\epsilon_d}(t/b_n)} dt \text{ for } d = 1, \ldots, d_x.$$

We conjecture that analogous results to our main theorems can be established for the multivariate case.

To derive the convergence rates of $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$, we need following conditions.

Assumption 4.

(i) $\{Y_j, T_j, W_j, Z_j\}_{j=1}^n$ *is an i.i.d. sample of* (Y, T, W, Z). $f_{X,Z}$ *and* $E[Y(s)|X,Z]$ *are bounded away from zero, and* $f_{X,Z}$ *and* $E[Y^2(s)|X,Z]$ *are bounded for* $s = 0,1$ *over compact support* $\mathcal{X} \times \mathcal{Z}$.

(ii) $f_{X,Z}$, $p_{X,Z}$, *and* $E[Y(s)|X,Z]$ *for* $s = 0,1$ *are* γ*-times continuously differentiable with bounded and integrable derivatives for some positive integer* γ.

(iii) K *is differentiable to order* γ *and satisfies*

$$\int K(u)du = 1, \quad \int u^{p+1} K(u)du \neq 0, \quad \int u^l K(u)du = 0 \text{ for } l = 1, 2, \ldots, \gamma.$$

Also, K^{ft} *is compactly supported on* $[-1,1]$*, symmetric around zero, and bounded.*

(iv) $b_n \to 0$ *and* $nb_n \left(\inf_{|t| \leq b_n^{-1}} |f^{\mathrm{ft}}_{\epsilon}(t)| \right)^2 \to \infty$ *as* $n \to \infty$.

Assumption 4 (i) requires the random sampling and the regularity of densities and conditional moments. Assumption 4 (ii) imposes smoothness restrictions on the densities and conditional moments, which are needed to control the magnitude of bias in the estimation together with the properties of kernel function K as imposed in Assumption 4 (iii). In addition to the standard properties of a high-order kernel function, Assumption 4 (iii) also requires K^{ft} to be compactly supported, which is commonly used in deconvolution problems to truncate the ill-behaved tails of the integrand for regularization purposes. Meister (2009) discusses how kernels of any order can be constructed quite simply. Assumption 4 (iv) imposes mild bandwidth restrictions. In particular, it simply requires that the bandwidth must decay to zero as the sample size grows, but should not decay too fast. The second part of Assumption 4 (iv) is needed so that the higher order components of the estimation error are asymptotically negligible.

Theorem 2. *Under Assumptions 1 - 4, it holds*

$$|\tilde{\tau} - \tau| = O_p\left(n^{-1/2} b_n^{-3/2} \left(\inf_{|t|\leq b_n^{-1}} |f_\epsilon^{\text{ft}}(t)|\right)^{-1} + b_n^\gamma\right),$$
$$|\tilde{\tau}_{\text{treat}} - \tau_{\text{treat}}| = O_p\left(n^{-1/2} b_n^{-3/2} \left(\inf_{|t|\leq b_n^{-1}} |f_\epsilon^{\text{ft}}(t)|\right)^{-1} + b_n^\gamma\right).$$

Theorem 2 presents the convergence rates of $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$. The second term b_n^γ in the convergence rate characterizes the magnitude of the estimation bias, which is identical to the error-free case. The first term characterizes the magnitude of the estimation variance. Compared to the error-free case, the estimation variance of $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$ decays slower due to the extra term $b_n^{-1/2}\left(\inf_{|t|\leq b_n^{-1}} |f_\epsilon^{\text{ft}}(t)|\right)^{-1}$. In particular, the smoother is the error distribution, the larger will be the estimation variance and, hence, the slower will be the convergence rate.

As is typical in the nonparametric measurement literature, to further specify the convergence rates of $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$, we consider two separate cases characterized by different smoothness of the measurement error: the ordinary smooth case and the supersmooth case. For the ordinary smooth case, the error characteristic function decays at a polynomial rate. In particular, we impose the following condition.

Assumption 5. *There exist positive constants α and $c_0^{\text{os}} \leq c_1^{\text{os}}$ such that*

$$c_0^{\text{os}}(1+|t|)^{-\alpha} \leq |f_\epsilon^{\text{ft}}(t)| \leq c_1^{\text{os}}(1+|t|)^{-\alpha} \text{ for all } t \in \mathbb{R}.$$

If f_ϵ satisfies Assumption 5, we say that it is ordinary smooth of order α. Popular examples of ordinary smooth densities include the Laplace and gamma density. The convergence rates of $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$ in the presence of ordinary smooth error of order α are specified as follow.

Corollary 1. *Under Assumptions 1 - 4, if Assumption 5 holds true, we have*

$$|\tilde{\tau} - \tau| = O_p\left(n^{-1/2} b_n^{-(3/2+\alpha)} + b_n^\gamma\right),$$
$$|\tilde{\tau}_{\text{treat}} - \tau_{\text{treat}}| = O_p\left(n^{-1/2} b_n^{-(3/2+\alpha)} + b_n^\gamma\right).$$

Corollary 1 shows that $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$ converge in a polynomial rate n^{-r} for some constant $r > 0$. The value of r depends on the choice of the bandwidth b_n, which will be discussed in Section 4.

For the supersmooth case, the error characteristic function decays at an exponential rate. In particular, we impose the following condition.

Assumption 6. *There exist positive constants β, β_0, and $c_0^{\text{ss}} \leq c_1^{\text{ss}}$ such that*

$$c_0^{\text{ss}} e^{-\beta_0 |t|^\beta} \leq |f_\epsilon^{\text{ft}}(t)| \leq c_1^{\text{ss}} e^{-\beta_0 |t|^\beta} \text{ for all } t \in \mathbb{R}.$$

If f_ϵ satisfies Assumption 6, we say that it is supersmooth of order β. Popular examples of supersmooth densities include the Cauchy and Gaussian density. The convergence rates of $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$ in the presence of supersmooth error of order β are specified as follow.

Corollary 2. *Under Assumptions 1 - 4, if Assumption 6 holds true, we have*

$$|\tilde{\tau} - \tau| = O_p\left(n^{-1/2} b_n^{-3/2} e^{\beta_0 b_n^{-\beta}} + b_n^\gamma\right),$$
$$|\tilde{\tau}_{\text{treat}} - \tau_{\text{treat}}| = O_p\left(n^{-1/2} b_n^{-3/2} e^{\beta_0 b_n^{-\beta}} + b_n^\gamma\right).$$

Corollary 2 shows that $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$ can only converge at a logarithm rate, which is much slower than the polynomial rate that has been seen in the ordinary smooth case. In particular, a normal error would make the estimator much more data-demanding than the case with a Laplace error. Again, the specific rate will depends on the choice of the bandwidth b_n, which will be discussed in Section 4.

3.5 Case of Unknown Measurement Error Distribution

Assuming the measurement error distribution to be fully known is usually unrealistic in practice. Auxiliary information, such as repeated measurements of X, can be used to relax the assumption of a known error distribution imposed in Assumption 3.

Suppose we have two independent noisy measures of X, W and W^r, determined as follows

$$\begin{aligned} W_j &= X_j + \epsilon_j \\ W_j^r &= X_j + \epsilon_j^r, \end{aligned}$$

for $j = 1, \ldots, n$. To identify the distribution of ϵ, we impose the following assumption.

Assumption 7. *(ϵ, ϵ^r) are mutually independent and independent of (X, Y, Z, T), the distributions of ϵ and ϵ^r are identical, and f_ϵ is symmetric around zero.*

The error characteristic function, f_ϵ^{ft}, can be estimated by

$$\hat{f}_\epsilon^{\text{ft}}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} \cos\{t(W_j - W_j^r)\} \right|^{1/2}$$

under Assumption 7 (Delaigle et al. 2008)[3].

When the measurement error distribution is unknown, we can estimate τ and τ_{treat} by plugging in this estimator, yielding

$$\hat{\tau} = \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \frac{\hat{q}_{1,1}(x,z)}{\hat{q}_{0,1}(x,z)} - \frac{\hat{q}_{1,0}(x,z)}{\hat{q}_{0,0}(x,z)} \right\} \hat{q}(x,z) dx dz, \tag{13}$$

$$\hat{\tau}_{\text{treat}} = \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \hat{q}_{1,1}(x,z) - \frac{\hat{q}_{1,0}(x,z)\hat{q}_{0,1}(x,z)}{\hat{q}_{0,0}(x,z)} \right\} \frac{1}{\hat{p}} dx dz, \tag{14}$$

where

$$\hat{q}(x,z) = \frac{1}{nb_n^2} \sum_{j=1}^{n} \hat{\mathbb{K}} \left(\frac{x - W_j}{b_n} \right) K \left(\frac{z - Z_j}{b_n} \right),$$

$$\hat{q}_{k,s}(x,z) = \frac{1}{nb_n^2} \sum_{j=1}^{n} Y_j^k T_j^s (1 - T_j)^{1-s} \hat{\mathbb{K}} \left(\frac{x - W_j}{b_n} \right) K \left(\frac{z - Z_j}{b_n} \right) \text{ for } k, s = 0, 1,$$

and the deconvolution kernel function based on estimated error characteristic function defined by

$$\hat{\mathbb{K}}(x) = \frac{1}{2\pi} \int e^{-\mathrm{i}tx} \frac{K^{\text{ft}}(t)}{\hat{f}_{\epsilon}^{\text{ft}}(t/b_n)} dt.$$

3.6 Inference

The proposed estimators, both in the case where the error distribution is known and the case where it is not, are constructed based on deconvolution methods. Unfortunately, deconvolution-based inference is known to be extremely challenging. Most of the existing work on the deconvolution-based inference focuses on nonparametric objects, such as the density and regression function. In particular, Bissantz et al. (2007) develops uniform confidence bands for the density of a mismeasured variable when the error distribution is known. Lounici and Nickl (2011) derive upper bounds for the sup-norm risk of a wavelet deconvolution estimator of the density of a mismeasured variable when the error distribution is unknown but repeated measurements are available. The authors construct uniform confidence bands for the density using these bounds. The resulting confidence bands, however, could be conservative as the upper bound of the coverage probability is derived using the concentration inequalities. Kato and Sasaki (2018, 2019) establish uniform confidence bands with asymptotic validity for the density of a mismeasured variable and the regression function of a mismeasured covariate, respectively, in the case when the error distribution is unknown but repeated measurements are available. Inference concerning a *function* of densities

[3]As ϵ and ϵ^r are independent and identically distributed under Assumption 7, we have $E\left[e^{\mathrm{i}t(W-W^r)}\right] = E\left[e^{\mathrm{i}t(\epsilon-\epsilon^r)}\right] = |f_{\epsilon}^{\text{ft}}(t)|^2$. As f_{ϵ} is symmetric around zero under Assumption 7, $f_{\epsilon}^{\text{ft}}(t) > 0$, which implies $f_{\epsilon}^{\text{ft}}(t) = \left|E\left[e^{\mathrm{i}t(W-W^r)}\right]\right|^{1/2} = |E\left[\cos\{t(W-W^r)\}\right]|^{1/2}$. Thus, $\hat{f}_{\epsilon}^{\text{ft}}(t)$ is obtained by plugging in the sample analogue of $E\left[\cos\{t(W-W^r)\}\right]$.

and regression functions under the deconvolution problem is remains an open question in the literature. We leave for future work the examination of bootstrap methods for the construction of confidence intervals of our proposed estimators. In particular, a non-parametric bootstrap as in Bissantz et al. (2007) could be considered for the case when the error distribution is known, and a wild bootstrap as in Kato and Sasaki (2019) could be considered for the case when the error distribution is unknown but repeated measurements are available.

4 Simulation

In this section, we evaluate the finite sample performance of the proposed estimators using Monte Carlo simulation. In particular, we focus on the case with a single covariate for which we can only observe its noisy measurement, and the following data generating process is considered

$$Y(T) = g(T, X) + U,$$

where covariate X is drawn from $U[0.5, 1.5]$ and is independent of U, the error term U is drawn from $N(0, 1)$ and is independent of (T, X), the treatment is assigned according to $P(T = 1|X) = \exp(0.5 - X)$, and three specifications of g are considered

$$\begin{aligned}
&\text{DGP1 }: g(t, x) = t + x, \\
&\text{DGP2 }: g(t, x) = t + x + x^2 - x^3, \\
&\text{DGP3 }: g(t, x) = t + x - \sin(x).
\end{aligned}$$

While X is assumed unobserved, we suppose $W = X + \epsilon$ and $W^r = X + \epsilon^r$ are observed, where (ϵ, ϵ^r) is mutually independent and independent of (T, X, U). For the distributions of ϵ and ϵ^r, we consider two cases. First, as an example of the case of ordinary smooth errors, we consider the case when (ϵ, ϵ^r) have a zero mean Laplace distribution with standard deviation $1/3$. Second, as an example of the case of supersmooth errors, we consider the case when (ϵ, ϵ^r) have a normal distribution with zero mean and standard deviation of $1/3$.

Throughout the simulation study, we use the kernel function whose Fourier transform is

$$K^{\text{ft}}(t) = \begin{cases} 1 & \text{if } |t| \leq 0.05, \\ \exp\left\{\frac{-\exp(-(|t|-0.05)^2)}{(|t|-1)^2}\right\} & \text{if } 0.05 < |t| < 1, \\ 0 & \text{if } |t| \geq 1. \end{cases}$$

This is the infinite-order flat-top kernel proposed by McMurry and Politis (2004), whose Fourier transform is compactly supported and can be used for the regularization purpose in the deconvolution estimation.

A trimming term is used in the denominators $\hat{q}_{0,0}$ and $\hat{q}_{0,1}$ to ensure the stability. Specifically, all values of the denominators below 0.01 are set to 0.01. Two sample sizes are considered, n=250 and 500, and all results are based on 500 Monte Carlo replications. For the choice of the bandwidth, to reduce the computation cost, we apply Delaigle and Gijbels (2004) to the first experiment, and use it for all subsequent simulations. To increase the robustness, we use 2 times the optimal bandwidth suggested by Delaigle and Gijbels (2004) as the deconvolution based estimator is more sensitive to smaller bandwidth; see Dong et al. (2020a).

Table 1: DGP1

Known Error Distribution								
Estimator	$\tilde{\tau}$				$\tilde{\tau}_{\text{treat}}$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
Bias	0.056	-0.217	-0.316	-0.292	0.131	-0.226	-0.322	-0.297
SD	0.371	0.104	0.126	0.096	0.559	0.109	0.128	0.098
RMSE	0.375	0.241	0.340	0.307	0.575	0.251	0.346	0.313
Unknown Error Distribution								
Estimator	$\hat{\tau}$				$\hat{\tau}_{\text{treat}}$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
Bias	0.094	-0.216	-0.315	-0.292	0.185	-0.224	-0.321	-0.297
SD	0.451	0.103	0.127	0.096	0.690	0.108	0.129	0.098
RMSE	0.461	0.239	0.340	0.307	0.714	0.249	0.346	0.313

Table 2: DGP2

Known Error Distribution								
Estimator	$\tilde{\tau}$				$\tilde{\tau}_{\text{treat}}$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
Bias	-0.082	-0.121	-0.170	-0.152	-0.086	-0.138	-0.178	-0.160
SD	0.273	0.103	0.125	0.096	0.400	0.107	0.127	0.098
RMSE	0.285	0.159	0.211	0.180	0.409	0.175	0.219	0.188
Unknown Error Distribution								
Estimator	$\hat{\tau}$				$\hat{\tau}_{\text{treat}}$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
Bias	-0.077	-0.120	-0.169	-0.152	-0.064	-0.137	-0.177	-0.160
SD	0.333	0.103	0.126	0.095	0.523	0.108	0.128	0.098
RMSE	0.342	0.158	0.211	0.180	0.527	0.174	0.219	0.188

Table 3: DGP3

Known Error Distribution								
Estimator	$\tilde{\tau}$				$\tilde{\tau}_{\text{treat}}$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
Bias	-0.023	-0.159	-0.240	-0.218	0.002	-0.165	-0.244	-0.221
SD	0.287	0.102	0.124	0.095	0.416	0.107	0.126	0.097
RMSE	0.288	0.189	0.270	0.237	0.417	0.197	0.275	0.242
Unknown Error Distribution								
Estimator	$\hat{\tau}$				$\hat{\tau}_{\text{treat}}$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
Bias	-0.008	-0.158	-0.239	-0.217	0.021	-0.163	-0.243	-0.221
SD	0.337	0.102	0.125	0.094	0.538	0.106	0.127	0.097
RMSE	0.338	0.188	0.269	0.237	0.539	0.195	0.274	0.241

In Tables 1-3, results are reported for both $\tilde{\tau}$ and $\tilde{\tau}_{\text{treat}}$, the proposed estimators for the case when the

error distribution is known and only W is observed, and $\hat{\tau}$ and $\hat{\tau}_{\text{treat}}$, the proposed estimators for the case when the error distribution is unknown and both W and W^r are observed. Results are given for the bias (Bias), standard deviation (SD), and the rooted mean squared error (RMSE) of each estimator in different settings.

The results appear encouraging and display several interesting features. First, the estimators have better performance with a larger sample size, and the performance of estimators is better with ordinary smooth error compared to the supersmooth error case. We also note that the performance of the estimators for the case when the error distribution is unknown is close to that of the estimators for the case when the error distribution is known. Using the estimated error distribution generally adds extra noise to the estimation. There are cases when the symmetry of the error distribution may allow the performance of the estimator to be independent of whether the error distribution is estimated or not; see Dong et al. (2020b). Finally, as to be expected, the proposed estimators have similar performance across different data generating processes, which implies that they are robust to unobserved nonlinearity in the conditional expectation function.

5 Application

To illustrate our estimator in practice, we revisit the analysis in Drexler et al. (2014).[4] Drexler et al. (2014) examine a randomized control trial (RCT) in which micro-entrepreneurs taking out a loan from ADOPEM, a microfinance institution in the Dominican Republic, are randomly assigned to one of three treatment arms to assess the causal effect of financial literacy programs on a firm's financial practices, objective reporting quality, and business performance. The first treatment provided subjects with standard accounting training. The second treatment provided rule-of-thumb training that covered basic financial heuristics. The final group received neither training and serves as the control group. The authors find significant beneficial effects of the rule-of-thumb training, but not the standard accounting training.

Our analysis deviates from Drexler et al. (2014) in one main respect. Whereas Drexler et al. (2014) examine both treatments simultaneously using a single regression model estimated by OLS, we do not. As our estimator is based on the IPW estimator, we examine each treatment separately. To do so, we restrict the sample to a single treatment arm along with the control group. Thus, our sample sizes diverge from the original study. Nonetheless, we present OLS estimates for comparison to Drexler et al. (2014) and they are essentially identical.

Our results are presented in Tables 4 and 5. The only difference across the two tables is the set of

[4]Data are available at `https://www.aeaweb.org/articles?id=10.1257/app.6.2.1`.

outcomes being analyzed.[5] In each table, Columns 2 and 5 report the OLS estimates of the treatment effect. These are most directly comparable to Drexler et al. (2014), subject to the caveat mentioned above that we assess each treatment separately. Columns 3 and 6 report IPW estimates of the ATE treating all covariates as correctly measured and estimating the propensity score via logit. Finally, Columns 4 and 7 report the results of our estimator for the ATE, treating the size of the loan as potentially mismeasured. Because the application lacks any auxiliary information on possible measurement error in this covariate, we assume the measurement error is normally distributed with mean zero and three different variance, corresponding to increasing levels of measurement error. Specifically, we set the standard deviation of the measurement error to be 1/6, 1/3, and 2/3 of the standard deviation of the observed loan values. The bandwidth for the observed loan values, as in the simulation experiment, is chosen as two times the optimal bandwidth suggested by Delaigle and Gijbels (2004), and bandwidths for other covariates, which are supposed to be error-free, are chosen based on Li and Racine (2003).

The results are interesting. In terms of the standard accounting treatment, the results appear robust to modest measurement in loan size. The only outcome for which the treatment effect is statistically significant ignoring measurement error is "setting aside cash for business purposes". Here, the OLS and IPW estimates are both 0.07. With modest measurement error, our estimator yields a point estimate of 0.08. It is noteworthy, however, that we find even stronger effects as we increase the variance of the measurement error.

In terms of the the rule-of-thumb treatment, the results appear predominantly robust to modest measurement in loan size as well. In Table 4, all OLS and IPW estimates are statistically significant at conventional levels. With modest measurement error, our estimator yields point estimates are qualitatively unchanged; sometimes slightly larger and sometimes slightly smaller in absolute value. As we increase the variance of the measurement error, however, the point estimates generally increase in magnitude. Thus, the economic magnitudes of the treatment effects are sensitive to the degree of measurement error. For example, increasing the standard deviation of the measurement error from 1/6 to 2/3 of the standard deviation of the observed loan values at least doubles the magnitude of the ATE for the outcomes "setting aside cash for business purposes," "keep accounting records," "separate business and personal accounting," and "calculate revenues formally".

In Table 5, the only outcomes for which the treatment effect is statistically significant ignoring measurement error are "any reporting errors" and "revenue index". For the former, our estimator suggests, if anything, a larger ATE in absolute value once measurement error is addressed. For the latter, our estimator suggests a smaller ATE once measurement error is addressed.

[5]Note, we do not analyze one outcome included in Drexler et al. (2014). Savings amount is excluded from our analysis as the first-stage propensity score ignoring measurement error did not converge.

Table 4: Impact of Training on Business Practices and Performance

Dependent Variable	**Standard Accounting**			**Rule-of-Thumb**		
	OLS	**IPW**	**IPW-ME**	**OLS**	**IPW**	**IPW-ME**
Business and Personal Financial Practices						
Separate Business and	0.00	0.00	-0.05	0.08	0.08	0.08
Personal Cash	(0.03)	(0.03)	{0.02}	(0.03)	(0.03)	{0.10}
			{{0.14}}			{{0.24}}
	524			532		
Keep Accounting Records	0.04	0.04	0.05	0.12	0.12	0.08
	(0.05)	(0.05)	{0.10}	(0.03)	(0.03)	{0.09}
			{{0.25}}			{{0.21}}
	524			533		
Separate Business and	0.04	0.04	0.00	0.12	0.12	0.11
Personal Accounting	(0.05)	(0.05)	{0.08}	(0.03)	(0.03)	{0.14}
			{{0.24}}			{{0.25}}
	521			532		
Set Aside Cash for	0.07	0.07	0.08	0.12	0.12	0.13
Business Purposes	(0.03)	(0.03)	{0.14}	(0.04)	(0.04)	{0.14}
			{{0.19}}			{{0.23}}
	524			532		
Calculate Revenues	0.01	0.01	-0.03	0.06	0.06	0.07
Formally	(0.04)	(0.04)	{0.04}	(0.03)	(0.03)	{0.15}
			{{0.16}}			{{0.23}}
	524			533		
Business Practices	0.07	0.07	-0.15	0.14	0.14	0.13
Index	(0.06)	(0.06)	{-0.17}	(0.04)	(0.04)	{0.18}
			{{-0.14}}			{{0.15}}
	525			534		
Any Savings	0.01	0.01	-0.03	0.08	0.08	0.05
	(0.05)	(0.05)	{0.01}	(0.04)	(0.04)	{0.04}
			{{0.19}}			{{0.10}}
	529			540		

Notes: Sample includes only those individuals with own business and either exposed to the treatment in the column heading or neither treatment. Number of observations beneath results for each model. Standard errors for OLS and IPW are in parentheses and are clustered at the barrio-level. IPW-ME reports only point estimates. The IPW-ME estimates in first row for each outcome assume the variance of the measurement is 1/6 of the variance of the observed covariate; estimates in {} assume the variance of the measurement is 1/3 of the variance of the observed covariate; estimates in {{}} assume the variance of the measurement is 2/3 of the variance of the observed covariate. IPW and IPW-ME estimates are of the average treatment effect.

Table 5: Impact of Training on Business Practices and Performance

Dependent Variable	**Standard Accounting**			**Rule-of-Thumb**		
	OLS	**IPW**	**IPW-ME**	**OLS**	**IPW**	**IPW-ME**
Objective Reporting Quality						
Any Reporting Errors	-0.04	-0.04	-0.09	-0.09	-0.09	-0.15
	(0.04)	(0.04)	{-0.07}	(0.03)	(0.03)	{-0.15}
			{{-0.04}}			{{-0.21}}
	496			508		
Raw Profit	918	914	1123	1094	1086	857
Calculation Difference	(746)	(726)	{1158}	(551)	(538)	{925}
(RD$), weekly			{{262}}			{{690}}
	273			289		
Absolute Value Profit	-324	-368	-633	-642	-641	-840
Calculation Difference	(643)	(622)	{-595}	(471)	(460)	{-803}
(RD$), weekly			{{-98}}			{{-919}}
	273			289		
Business Performance						
Total Number	0.08	0.08	0.37	-0.04	-0.04	0.11
of Employees	(0.09)	(0.09)	{-0.02}	(0.09)	(0.09)	{-0.01}
			{{0.61}}			{{0.98}}
	523			533		
Revenue Index	-0.02	-0.02	-0.02	0.10	0.10	0.04
	(0.05)	(0.05)	{-0.03}	(0.05)	(0.05)	{0.04}
			{{-0.09}}			{{0.03}}
	511			518		
Sales,	-649	-686	-543	604	665	220
Average Week (RD$)	(810)	(791)	{-619}	(942)	(941)	{-480}
			{{1663}}			{{130}}
	367			386		
Sales,	-672	-696	-386	1168	1111	389
Bad Week (RD$)	(513)	(497)	{-291}	(538)	(533)	{641}
			{{116}}			{{-35}}
	359			373		

Notes: Sample includes only those individuals with own business and either exposed to the treatment in the column heading or neither treatment. Number of observations beneath results for each model. Standard errors for OLS and IPW are in parentheses and are clustered at the barrio-level. IPW-ME reports only point estimates. The IPW-ME estimates in first row for each outcome assume the variance of the measurement is 1/6 of the variance of the observed covariate; estimates in {} assume the variance of the measurement is 1/3 of the variance of the observed covariate; estimates in {{}} assume the variance of the measurement is 2/3 of the variance of the observed covariate. IPW and IPW-ME estimates are of the average treatment effect.

6 Conclusion

Estimation of the causal effect of a binary treatment on outcomes, even in the case of selection on observed covariates, can be complicated when one or more of the covariates are measured with error. In this paper, we present a new semi-parametric estimator that addresses this issue. In particular, we focus on the case when the propensity score is of an unknown functional form and some covariates are subject to classical measurement error. Allowing the functional form of the propensity score to be unknown as well as a function of unobserved, error-free covariates, we consider an integration weighted by deconvolution kernel density estimator. Our simulations and replication exercise show our estimator to be valuable to empirical researchers. However, future work is needed to conduct inference with this estimator.

References

[1] Abadie, Alberto, and Matias D. Cattaneo. 2018. Econometric Methods for Program Evaluation. *Annual Review of Economics* 10: 465-503.

[2] Battistin, Erich, and Andrew Chesher. 2014. Treatment Effect Estimation with Covariate Measurement Error. *Journal of Econometrics* 178: 707-715.

[3] Berge, Lars Ivar Oppedal, Kjetil Bjorvatn, and Bertil Tungodden. 2015. Human and Financial Capital for Microenterprise Development: Evidence from a Field and Lab Experiment. *Management Science* 61(4): 707-722.

[4] Bissantz, Nicolai, Lutz Dümbgen, Hajo Holzmann, and Axel Munk. 2007. Non-parametric confidence bands in deconvolution density estimation. *Journal of the Royal Statistical Society, Series B* 69(3): 483-506.

[5] Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. Measurement Error in Survey Data. In J.J. Heckman and E.E. Leamer (eds.) *Handbook of Econometrics* Vol. 5, 3705-3843.

[6] Bruhn, Miriam, Dean Karlan, and Antoinette Schoar. 2018. The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico. *Journal of Political Economy* 126(2): 635-687.

[7] Carroll, Raymond J., and Peter Hall. 1988. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association* 83(404): 1184-1186.

[8] Chen, Xiaohong, Han Hong, and Denis Nekipelov. 2011. Nonlinear models of measurement errors. *Journal of Economic Literature* 49(4): 901-937.

[9] Cochran, William G. and Donald B. Rubin. 1973. Controlling Bias in Observational Studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* 35(4): 417-446.

[10] Delaigle, Aurore, and Irene Gijbels. 2004. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics* 56(1): 19-47.

[11] Delaigle, Aurore, and Alexander Meister. 2007. Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association* 102(480): 1416-1426.

[12] Delaigle, Aurore, Peter Hall, and Alexander Meister. 2008. On Deconvolution with Repeated Measurements. *The Annals of Statistics* 36: 665–685.

[13] Delaigle, Aurore, Jianqing Fan, and Raymond J. Carroll. 2009. A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association* 104(485): 348-359.

[14] Delaigle, Aurore, Peter Hall, and Farshid Jamshidi. 2015. Confidence bands in non-parametric errors-in-variables regression. *Journal of the Royal Statistical Society, Series B* 77(1): 149-169.

[15] Dong, Hao, Taisuke Otsu, and Luke N. Taylor. 2020a. Estimation of Varying Coefficient Models with Measurement Error. Working paper.

[16] Dong, Hao, Taisuke Otsu, and Luke N. Taylor. 2020b. Average Derivative Estimation under Measurement Error. Forthcoming in *Econometric Theory*.

[17] Drexler, Alejandro, Greg Fischer, and Antoinette Schoar. 2014. Keeping It Simple: Financial Literacy and Rules of Thumb. *American Economic Journal: Applied Economics* 6(2): 1-31.

[18] Fan, Jianqing. 1991. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* 19(3): 1257-1272.

[19] Fan, Jianqing. 1991. Asymptotic normality for deconvolution kernel density estimators. *Sankhyā: The Indian Journal of Statistics, Series A* 53: 97-110.

[20] Fan, Jianqing, and Elias Masry. Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes. *Journal of multivariate analysis* 43(2): 237-271.

[21] Fan, Jianqing, and Young K. Truong. 1993. Nonparametric regression with errors in variables. *The Annals of Statistics* 21(4): 1900-1925.

[22] Fan, Yanqin. 1995. Average derivative estimation with errors-in-variables. *Journaltitle of Nonparametric Statistics* 4(4): 395-407.

[23] Fernandes, Daniel, John G. Lynch Jr., Richard G. Netemeyer. 2014. Financial Literacy, Financial Education, and Downstream Financial Behaviors. *Management Science* 60(8): 1861-1883.

[24] Fisher, Ronald A. 1935. The logic of inductive inference. *Journal of the Royal Statistical Society* 98: 39–82.

[25] Frisch, Ragnar. 1934. *Statistical Confluence Analysis By Means of Complete Regression Systems*. Oslo, University Institute for Economics.

[26] Heckman, James J., and Richard Robb Jr. 1985. Alternative Methods for Evaluating the Impact of Interventions: An overview. *Journal of Econometrics* 30: 239–267.

[27] Hong, Hwanhee, David A. Aaby, Juned Siddique, and Elizabeth A.Stuart. 2019. Propensity Score–Based Estimators With Multiple Error-Prone Covariates. *American Journal of Epidemiology* 188(1): 222-230.

[28] Horowitz, Joel L. 2009. Semiparametric and nonparametric methods in econometrics. Springer.

[29] Horvitz, D.G. and D.J. Thompson. 1952. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* 47(260): 663–685.

[30] Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* 47: 5-86.

[31] Jakubowski, Maciej. 2010. Latent Variables and Propensity Score Matching: A Simulation Study with Application to Data from the Programme for International Student Assessment in Poland. *Empirical Economics* 48: 1287-1325.

[32] Kato, Kengo, and Yuya Sasaki. 2018. Uniform confidence bands in deconvolution with unknown error distribution. *Journal of Econometrics* 207(1): 129-161.

[33] Kato, Kengo, and Yuya Sasaki. 2019. Uniform confidence bands for nonparametric errors-in-variables regression. *Journal of Econometrics* 213(2): 516-555.

[34] Koopmans, Tjalling. 1937. *Linear Regression Analysis of Economic Time Series*. Amsterdam, Netherlands Econometric Institute, Harrlem-de Erwen F Bohn N.V.

[35] Li, Qi, and Jeff Racine. 2003. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* 86(2): 266-292.

[36] Lounici, Karim, and Richard Nickl. 2011. Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics* 39(1): 201-231.

[37] Lusardi, Annamaria, and Olivia S. Mitchell. 2014. The Economic Importance of Financial Literacy. *Journal of Economic Literature* 52: 5-44.

[38] McCaffrey, Daniel F., J.R. Lockwood, and Claude M. Setodji. 2013. Inverse Probability Weighting with Error-Prone Covariates. *Biometrika* 100(3): 671-680.

[39] McKenzie, David, and Chrisopher Woodruff. 2013. What Are We Learning from Business Training and Entrepreneurship Evaluations around the Developing World? *World Bank Research Observer* 29: 48-82.

[40] McMurry, Timothy L., and Dimitris N. Politis. 2004. Nonparametric Regression with Infinite Order Flat-Top Kernels. *Journal of Nonparametric Statistics* 16: 549-562.

[41] Meister, Alexander. 2009. Deconvolution problems in nonparametric statistics. Springer.

[42] Morgan, Peter J., and Long Q. Trih. 2019. Determinants and Impacts of Financial Literacy in Cambodia and Viet Nam. *Journal of Risk and Financial Management* 12: 1-24.

[43] Neyman, Jerzy S. 1923. On the Application of Probability Theory to Agricultural Experiments. Essays on Principles. Section 9. *Annals of Agricultural Sciences*, 10: 1—51.

[44] Reiersøl, Olav. 1950. Identifiability of a Linear Relation Between Variables Which Are Subject to Error. *Econometrica* 18(4): 375-389.

[45] Rosenbaum, Paul R., and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1): 41-55.

[46] Rosenbaum, Paul R. 2010. *Design of Observational Studies* (Springer Series in Statistics). New York, NY: Springer Publishing Company.

[47] Roy, A. D. 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers* 3: 135–146.

[48] Rubin, Donald B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66: 688–701.

[49] Rubin, Donald B. 1986. Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics* 4: 87–94.

[50] Rudolph, Kara E., and Elizabeth A. Stuart. 2018. Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods. *American Journal of Epidemiology* 187(3): 604-613.

[51] Schennach, Susanne M. 2016. Recent advances in the measurement error literature. *Annual Review of Economics* 8: 314-377.

[52] Schennach, Susanne M. 2019. Convolution without independence. *Journal of Econometrics* 211: 308–318.

[53] Stefanski, Leonard A., and Raymond J. Carroll. 1990. Deconvolving kernel density estimators. *Statistics* 21: 169–184.

[54] VanderWeele, Tyler J., and Onyebuchi A. Arah. 2011. Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders. *Epidemiology* 22(1): 42-52.

[55] Van Es, Bert, Shota Gugushvili, and Peter Spreij. 2008. Deconvolution for an atomic distribution. *Electronic Journal of Statistics* 2: 265-297.

[56] Webb-Vargas, Yenny, Kara E. Rudoplh, David Lenis, Peter Murkami, and Elizabeth A. Stuart. 2017. An Imputation-Based Solution to Using Mismeasured Covariates in Propensity Score Analysis. *Statistical Methods in Medical Research* 26(4): 1824-1837.

[57] Wooldridge, Jeffrey M. 2010. Econometric analysis of cross section and panel data. MIT press.

A Derivation of Equations

A.1 Derivation of (6) and (7)

(6) follows by $\tau = E[Y(1) - Y(0)]$ and for $t = 0, 1$,

$$\begin{aligned}
E[Y(t)] &= E[E[Y(t)|X,Z]] = E[E[Y(t)|T=t,X,Z]] = E[E[Y|T=t,X,Z]] \\
&= \iiint y f_{Y|T=t,X=x,Z=z}(y)dy f_{X,Z}(x,z)dxdz \\
&= P(T=t)\int y \left\{ \iint \frac{f_{Y,X,Z|T=t}(y,x,z)f_{X,Z}(x,z)}{f_{X,Z|T=t}(x,z)P(T=t)}dxdz \right\} dy \\
&= P(T=t)\iiint \frac{y f_{Y,X,Z|T=t}(y,x,z)}{P(T=t|X=x,Z=z)}dydxdz,
\end{aligned}$$

where the second step follows by Assumption 1 and the last step requires Assumption 2. (7) follows by $\tau_{\text{treat}} = E[Y(1) - Y(0)|T=1]$, $E[Y(1)|T=1] = E[Y|T=1] = E[TY]/p$, and

$$\begin{aligned}
E[Y(0)|T=1] &= E[E[Y(0)|T=1,X,Z]|T=1] \\
&= E[E[Y(0)|T=0,X,Z]|T=1] \\
&= E[E[Y|T=0,X,Z]|T=1] \\
&= \iint \left\{ \int y f_{Y|T=0,X=x,Z=z}(y)dy \right\} f_{X,Z|T=1}(x,z)dxdz \\
&= \iiint \frac{y f_{X,Z|T=1}(x,z)}{f_{X,Z|T=0}(x,z)} f_{Y,X,Z|T=0}(y,x,z)dydxdz \\
&= E\left[\frac{Y f_{X,Z|T=1}(X,Z)}{f_{X,Z|T=0}(X,Z)} \middle| T=0 \right] \\
&= \frac{1-p}{p} E\left[\frac{Y p_{X,Z}(X,Z)}{1-p_{X,Z}(X,Z)} | T=0 \right],
\end{aligned}$$

where the first step follows by Assumption 1 and the last step requires Assumption 2.

A.2 Derivation of (5)

$$
\begin{aligned}
E[Y(1)] &= E\left[\frac{TY}{p_{X,Z}(X,Z)}\right] = p\iint E[TY|X=x,Z=z]f_{X,Z|T=1}(x,z)dxdz \\
&= p\iint\left\{\frac{1}{2\pi}\int e^{-\mathrm{i}sx}\frac{\{E[TY|W=\cdot,Z=z]\}^{\mathrm{ft}}(s)}{f_\epsilon^{\mathrm{ft}}(s)}ds\right\}\left\{\frac{1}{2\pi}\int e^{-\mathrm{i}tx}\frac{\{f_{W,Z|T=1}(\cdot,z)\}^{\mathrm{ft}}(t)}{f_\epsilon^{\mathrm{ft}}(t)}dt\right\}dxdz \\
&= \frac{p}{2\pi}\iiint\left\{\frac{1}{2\pi}\int e^{-\mathrm{i}(s+t)x}dx\right\}\frac{\{E[TY|W=\cdot,Z=z]\}^{\mathrm{ft}}(s)\{f_{W,Z|T=1}(\cdot,z)\}^{\mathrm{ft}}(t)}{f_\epsilon^{\mathrm{ft}}(s)f_\epsilon^{\mathrm{ft}}(t)}dsdtdz \\
&= \frac{p}{2\pi}\iint\frac{\{E[TY|W=\cdot,Z=z]\}^{\mathrm{ft}}(-t)\{f_{W,Z|T=1}(\cdot,z)\}^{\mathrm{ft}}(t)}{|f_\epsilon^{\mathrm{ft}}(t)|^2}dtdz \\
&= \frac{p}{2\pi}\iint\left\{\int E[TY|W=w,Z=z]e^{-\mathrm{i}tw}dw\right\}\frac{\{f_{W,Z|T=1}(\cdot,z)\}^{\mathrm{ft}}(t)}{|f_\epsilon^{\mathrm{ft}}(t)|^2}dtdz \\
&= \iint E[TY|W=w,Z=z]\left\{\frac{p\int e^{-\mathrm{i}tw}\frac{\{f_{W,Z|T=1}(\cdot,z)\}^{\mathrm{ft}}(t)}{|f_\epsilon^{\mathrm{ft}}(t)|^2}dt}{2\pi f_{W,Z}(w,z)}\right\}f_{W,Z}(w,z)dwdz \\
&= E[TYA(W,Z)],
\end{aligned}
$$

where the sixth equality follows by $\int\delta(x-b)f(x)dx = f(b)$ with Dirac delta function $\delta(x) = \frac{1}{2\pi}\int e^{-\mathrm{i}tx}dx$.

A.3 Derivation of (11) and (12)

$$
\begin{aligned}
\tilde{\tau} &= \iiint\left\{\frac{\hat{p}\tilde{f}_{Y,X,Z|T=1}(y,x,z)}{\tilde{p}_{X,Z}(x,z)} - \frac{(1-\hat{p})\tilde{f}_{Y,X,Z|T=0}(y,x,z)}{1-\tilde{p}_{X,Z}(x,z)}\right\}ydydxdz \\
&= \frac{1}{nb_n^3}\sum_{j=1}^n\iint\left\{\frac{T_j\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)}{\frac{\sum_{l=1}^n T_l\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^n\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}} - \frac{(1-T_j)\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)}{\frac{\sum_{l=1}^n(1-T_l)\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^n\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}}\right\}\left\{\int K\left(\frac{y-Y_j}{b_n}\right)ydy\right\}dxdz \\
&=_{(1)} \frac{1}{nb_n^2}\sum_{j=1}^n\iint\left\{\frac{T_j\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)}{\frac{\sum_{l=1}^n T_l\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^n\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}} - \frac{(1-T_j)\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)}{\frac{\sum_{l=1}^n(1-T_l)\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^n\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}}\right\}\left\{\int K(\tilde{y})(Y_j+b_n\tilde{y})d\tilde{y}\right\}dxdz \\
&=_{(2)} \frac{1}{nb_n^2}\sum_{j=1}^n Y_j\iint\left\{\frac{T_j\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)}{\frac{\sum_{l=1}^n T_l\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^n\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}} - \frac{(1-T_j)\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)}{\frac{\sum_{l=1}^n(1-T_l)\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^n\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}}\right\}dxdz \\
&= \iint\left\{\frac{\hat{q}_{1,1}(x,z)}{\hat{q}_{0,1}(x,z)} - \frac{\hat{q}_{1,0}(x,z)}{\hat{q}_{0,0}(x,z)}\right\}\hat{q}(x,z)dxdz,
\end{aligned}
$$

where (1) follows by change of variables $\tilde{y} = \frac{y-Y_j}{b_n}$ and (2) follows by $\int K(x)dx = 0$ and $\int K(x)xdx = 0$.

$$
\begin{aligned}
\tilde{\tau}_{\text{treat}} &= \frac{1}{\hat{p}} \iiint \left\{ \hat{p}\tilde{f}_{X,Y,Z|T=1}(x,y,z) - \frac{\tilde{p}_{X,Z}(x,z)(1-\hat{p})\tilde{f}_{X,Y,Z|T=0}(x,y,z)}{1-\tilde{p}_{X,Z}(x,z)} \right\} y dx dy dz \\
&= \frac{1}{nb_n^3\hat{p}} \sum_{j=1}^{n} \iint \left\{ \begin{array}{c} T_j \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right) \\ -\frac{(1-T_j)\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)\sum_{l=1}^{n} T_l \mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^{n}(1-T_l)\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)} \end{array} \right\} \left\{ \int K\left(\frac{y-Y_j}{b_n}\right) y dy \right\} dx dz \\
&=_{(1)} \frac{1}{nb_n^2\hat{p}} \sum_{j=1}^{n} \iint \left\{ \begin{array}{c} T_j \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right) \\ -\frac{(1-T_j)\mathbb{K}\left(\frac{x-W_j}{b_n}\right)K\left(\frac{z-Z_j}{b_n}\right)\sum_{l=1}^{n} T_l \mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^{n}(1-T_l)\mathbb{K}\left(\frac{x-W_l}{b_n}\right)K\left(\frac{z-Z_l}{b_n}\right)} \end{array} \right\} \left\{ \int K(\tilde{y})\,(Y_j + b_n\tilde{y}) d\tilde{y} \right\} dx dz \\
&=_{(2)} \frac{1}{nb_n^2\hat{p}} \sum_{j=1}^{n} Y_j \iint \left\{ T_j - \frac{(1-T_j)\sum_{l=1}^{n} T_l \mathbb{K}\left(\frac{x-W_l}{b_n}\right) K\left(\frac{z-Z_l}{b_n}\right)}{\sum_{l=1}^{n}(1-T_l)\mathbb{K}\left(\frac{x-W_l}{b_n}\right) K\left(\frac{z-Z_l}{b_n}\right)} \right\} \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right) dx dz \\
&= \iint \left\{ \hat{q}_{1,1}(x,z) - \frac{\hat{q}_{1,0}(x,z)\hat{q}_{0,1}(x,z)}{\hat{q}_{0,0}(x,z)} \right\} \frac{1}{\hat{p}} dx dz,
\end{aligned}
$$

where (1) follows by change of variables $\tilde{y} = \frac{y-Y_j}{b_n}$ and (2) follows by $\int K(x)dx = 0$ and $\int K(x)x dx = 0$. The integrations in (11) and (12) are restricted on $\mathcal{X}$ and $\mathcal{Z}$ to emphasize that it is sufficient to integrate over the supports of X and Z.

B Proof of Theorems

B.1 Proof of Theorem 1

To identify τ and τ_{treat}, it is sufficient to identify $f_{Y,X,Z|T}$ as $p(x,z) = \frac{p f_{X,Z|T=1}(x,z)}{p f_{X,Z|T=1}(x,z) + (1-p) f_{X,Z|T=0}(x,z)}$ and $f_{X,Z|T}(x,z) = \int f_{Y,X,Z|T}(y,x,z)dy$. Let $\{f_{Y,W,Z|T}(y,\cdot,z)\}^{\text{ft}}(t) = \int e^{\text{i}tw} f_{Y,W,Z|T}(y,w,z)dw$. The identification of $f_{Y,X,Z|T}$ follows by

$$
f_{Y,X,Z|T}(y,x,z) = \frac{1}{2\pi} \int e^{-\text{i}tx} \frac{\{f_{Y,W,Z}(y,\cdot,z)\}^{\text{ft}}(t)}{f_{\epsilon}^{\text{ft}}(t)} dt,
$$

for which we use the convolution theorem under Assumption 3.

B.2 Proof of Theorem 2

Define $q(x,z) = f_{X,Z}(x,z)$ and $q_{k,s}(x,z) = \{m_{X,Z,s}^k p_{X,Z,s} f_{X,Z}\}(x,z)$ for $k,s = 0,1$, where $m_{X,Z,s}(x,z) = E[Y(s)|X=x,Z=z]$ and $p_{X,Z,s}(x,z) = p_{X,Z}^s(x,z)(1-p_{X,Z}(x,z))^{1-s}$. Then, we have

$$
\tau = \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \frac{q_{1,1}(x,z)}{q_{0,1}(x,z)} - \frac{q_{1,0}(x,z)}{q_{0,0}(x,z)} \right\} q(x,z) dx dz, \tag{15}
$$

$$\tau_{\text{treat}} = \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ q_{1,1}(x,z) - \frac{q_{1,0}(x,z) q_{0,1}(x,z)}{q_{0,0}(x,z)} \right\} \frac{1}{p} dx dz. \tag{16}$$

First, using $\hat{u}\hat{v} - uv = (\hat{u}-u)v + (\hat{v}-v)u + (\hat{u}-u)(\hat{v}-v)$ and $\hat{v}^{-1} - v^{-1} = -(\hat{v}-v)\{v(\hat{v}-v) + v^2\}^{-1}$, we have

$$\frac{\tilde{q}_{1,s}\tilde{q}}{\tilde{q}_{0,s}} - \frac{q_{1,s}q}{q_{0,s}} = \frac{\{(\tilde{q}_{1,s} - q_{1,s})q + (\tilde{q} - q)q_{1,s} + (\tilde{q}_{1,s} - q_{1,s})(\tilde{q} - q)\}q_{0,s} - (\tilde{q}_{0,s} - q_{0,s})q_{1,s}q}{q_{0,s}(\tilde{q}_{0,s} - q_{0,s}) + q_{0,s}^2}, \tag{17}$$

$$\frac{\tilde{q}_{1,1}}{\hat{p}} - \frac{q_{1,1}}{p} = \frac{(\tilde{q}_{1,1} - q_{1,1})p - (\hat{p} - p)q_{1,1}}{p(\hat{p} - p) + p^2}, \tag{18}$$

$$\frac{\tilde{q}_{1,0}\tilde{q}_{0,1}}{\tilde{q}_{0,0}\hat{p}} - \frac{q_{1,0}q_{0,1}}{q_{0,0}p} = \frac{\left\{ \begin{array}{c} \{(\tilde{q}_{1,0} - q_{1,0})q_{0,1} + (\tilde{q}_{0,1} - q_{0,1})q_{1,0} + (\tilde{q}_{1,0} - q_{1,0})(\tilde{q}_{0,1} - q_{0,1})\}pq_{0,0} \\ -\{(\hat{p} - p)q_{0,0} + (\tilde{q}_{0,0} - q_{0,0})p + (\hat{p} - p)(\tilde{q}_{0,0} - q_{0,0})\}q_{1,0}q_{0,1} \end{array} \right\}}{pq_{0,0}\{(\hat{p} - p)q_{0,0} + (\tilde{q}_{0,0} - q_{0,0})p + (\hat{p} - p)(\tilde{q}_{0,0} - q_{0,0})\} + p^2 q_{0,0}^2}, \tag{19}$$

where we intentionally suppress the dependence of $\hat{q}$, $\hat{q}_{k,s}$, q, and $q_{k,s}$ on x and z for $k, s = 0, 1$ to keep the notation simple.

For $\tilde{\tau}$, note that

$$\begin{aligned}
|\tilde{\tau} - \tau| &= \left| \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \begin{array}{c} \left\{ \frac{\tilde{q}_{1,1}(x,z)\tilde{q}(x,z)}{\tilde{q}_{0,1}(x,z)} - \frac{q_{1,1}(x,z)q(x,z)}{q_{0,1}(x,z)} \right\} \\ -\left\{ \frac{\tilde{q}_{1,0}(x,z)\tilde{q}(x,z)}{\tilde{q}_{0,0}(x,z)} - \frac{q_{1,0}(x,z)q(x,z)}{q_{0,0}(x,z)} \right\} \end{array} \right\} dx dz \right| \\
&= O\left(\max_{s\in\{0,1\}} \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} \left| \frac{\tilde{q}_{1,s}(x,z)\tilde{q}(x,z)}{\tilde{q}_{0,s}(x,z)} - \frac{q_{1,s}(x,z)q(x,z)}{q_{0,s}(x,z)} \right| \right) \\
&= O\left(\max_{s\in\{0,1\}} \left\{ \begin{array}{c} \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{1,s}(x,z) - q_{1,s}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,s}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}(x,z) - q(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,s}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,s}(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{1,s}(x,z) - q_{1,s}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}(x,z) - q(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,s}(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{0,s}(x,z) - q_{0,s}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,s}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q(x,z)| \end{array} \right\} \right) \\
&= O\left(\max_{k,s\in\{0,1\}} \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{k,s}(x,z) - q_{k,s}(x,z)| + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}(x,z) - q(x,z)| \right),
\end{aligned}$$

where the first step follows by (6) and (15), the second step follows by the compactness of $\mathcal{X}$ and $\mathcal{Z}$ (Assumption 4 (i)), the third step follows by (17), $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\hat{q}_{0,s}(x,z) - q_{0,s}(x,z)| = o_p(1)$ (Lemma 1 and Assumption 4 (iv)), and $\inf_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,s}(x,z)| > 0$ (Assumption 4 (i)) for $s = 0, 1$, and the last step follows by $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q(x,z)| < \infty$ (Assumption 4 (i)), $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{k,s}(x,z)| < \infty$ (Assumption 4 (i)), and $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\hat{q}_{k,s}(x,z) - q_{k,s}(x,z)| = o_p(1)$ (Lemma 1 and Assumption 4 (iv)) for any $k, s = 0, 1$.

For $\tilde{\tau}_{\text{treat}}$, note that

$$
\begin{aligned}
|\tilde{\tau}_{\text{treat}} - \tau_{\text{treat}}| &= \left| \int_{\mathcal{X}} \int_{\mathcal{Z}} \left\{ \begin{array}{c} \left\{ \frac{\tilde{q}_{1,1}(x,z)}{\hat{p}} - \frac{q_{1,1}(x,z)}{p} \right\} \\ - \left\{ \frac{\tilde{q}_{1,0}(x,z)\tilde{q}_{0,1}(x,z)}{\hat{p}\tilde{q}_{0,0}(x,z)} - \frac{q_{1,0}(x,z)q_{0,1}(x,z)}{pq_{0,0}(x,z)} \right\} \end{array} \right\} dxdz \right| \\
&= O\left(\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} \left| \frac{\tilde{q}_{1,1}(x,z)}{\hat{p}} - \frac{q_{1,1}(x,z)}{p} \right| + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} \left| \frac{\tilde{q}_{1,0}(x,z)\tilde{q}_{0,1}(x,z)}{\hat{p}\tilde{q}_{0,0}(x,z)} - \frac{q_{1,0}(x,z)q_{0,1}(x,z)}{pq_{0,0}(x,z)} \right| \right) \\
&= O\left(\begin{array}{c} \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{1,1}(x,z) - q_{1,1}(x,z)| + |\hat{p} - p| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,1}(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{1,0}(x,z) - q_{1,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,1}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,0}(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{0,1}(x,z) - q_{0,1}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,0}(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{1,0}(x,z) - q_{1,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{0,1}(x,z) - q_{0,1}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,0}(x,z)| \\ + |\hat{p} - p| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,1}(x,z)| \\ + \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{0,0}(x,z) - q_{0,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,1}(x,z)| \\ + |\hat{p} - p| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{0,0}(x,z) - q_{0,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{1,0}(x,z)| \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,1}(x,z)| \end{array} \right) \\
&= O\left(\max_{k,s\in\{0,1\}} \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{k,s}(x,z) - q_{k,s}(x,z)| + |\hat{p} - p| \right),
\end{aligned}
$$

where the first step follows by (7) and (16), the second step follows by the compactness of $\mathcal{X}$ and $\mathcal{Z}$ (Assumption 4 (i)), the third step follows by (18), (19), $|\hat{p} - p| = o_p(1)$ (Lemma 1 and Assumption 4 (i)), $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{0,0}(x,z) - q_{0,0}(x,z)| = o_p(1)$ (Lemma 1 and Assumption 4 (iv)), $0 < p < 1$ (Assumption 2), $\inf_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{0,0}(x,z)| > 0$ (Assumption 4 (i)), and the last step follows by $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |q_{k,s}(x,z)| < \infty$ (Assumption 4 (i)) and $\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\tilde{q}_{k,s}(x,z) - q_{k,s}(x,z)| = o_p(1)$ (Lemma 1 and Assumption 4 (iv)) for any $k, s = 0, 1$. The conclusion then follows by implementing Lemma 1.

C Lemmas

To facilitate the proof of Theorem 4, we introduce following Lemma, where q, $\hat{q}$, $q_{k,s}$, and $\hat{q}_{k,s}$ for $k, s = 0, 1$ are the same as defined in Appendix B.2.

Lemma 1. *Under Assumptions 3 and 4 (i) - (iii), it holds* $|\hat{p} - p| = O_p\left(n^{-1/2}\right)$ *and*

$$
\begin{aligned}
\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\hat{q}(x,z) - q(x,z)| &= O_p\left(n^{-1/2} b_n^{-3/2} \left(\inf_{|t|\le b_n^{-1}} |f_\epsilon^{\text{ft}}(t)| \right)^{-1} + b_n^{\gamma} \right), \\
\max_{k,s\in\{0,1\}} \sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} |\hat{q}_{k,s}(x,z) - q_{k,s}(x,z)| &= O_p\left(n^{-1/2} b_n^{-3/2} \left(\inf_{|t|\le b_n^{-1}} |f_\epsilon^{\text{ft}}(t)| \right)^{-1} + b_n^{\gamma} \right).
\end{aligned}
$$

Proof. The first statement follows by $E|\hat{p} - p|^2 = \frac{E|T|^2}{n} \le n^{-1}$. For the rest two statements, we focus on

the last one as the second statement can be shown in a similar way.

First, note that

$$
\begin{aligned}
E[\hat{q}_{k,s}(x,z)] &= b_n^{-2} E\left[Y_j^k T_j^s (1-T_j)^{1-s} \left\{ \frac{1}{2\pi} \int e^{-\mathrm{i}t\left(\frac{x-W_j}{b_n}\right)} \frac{K^{\mathrm{ft}}(t)}{f_\epsilon^{\mathrm{ft}}(t/b_n)} dt \right\} K\left(\frac{z-Z_j}{b_n}\right) \right] \\
&= b_n^{-2} E\left[Y_j^k T_j^s (1-T_j)^{1-s} \left\{ \frac{1}{2\pi} \int e^{-\mathrm{i}t\left(\frac{x-X_j}{b_n}\right)} K^{\mathrm{ft}}(t) dt \right\} K\left(\frac{z-Z_j}{b_n}\right) \right] \\
&= b_n^{-2} E\left[Y_j^k T_j^s (1-T_j)^{1-s} K\left(\frac{x-X_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right) \right] \\
&= \int q_{k,s}(x-b_n\tilde{u}, z-b_n\tilde{v}) K(\tilde{u}) K(\tilde{v}) d\tilde{u} d\tilde{v} \\
&= q_{k,s}(x,z) + O\left(b_n^{\gamma}\right),
\end{aligned}
$$

where the first step follows by the definition of the deconvolution kernel $\mathbb{K}$, the second step follows by the independence between ϵ and (Y,T,X,Z) (Assumption 3), the third step follows by $K(x) = \frac{1}{2\pi}\int e^{-\mathrm{i}tx} K^{\mathrm{ft}}(t)dt$, the fourth step follows by the change of variables $\tilde{u} = \frac{x-u}{b_n}$ and $\tilde{v} = \frac{z-v}{b_n}$, and the last step follows by the smoothness of $f_{X,Z}$, $p_{X,Z}$, and $E[Y(s)|X,Z]$ (Assumption 4 (ii)) and properties of the γ-th order kernel function K (Assumption 4 (iii)).

Also note that

$$
\begin{aligned}
Var(\hat{q}_{k,s}(x,z)) &\leq \frac{1}{nb_n^4} E\left| Y_j^k T_j^s (1-T_j)^{1-s} \mathbb{K}\left(\frac{x-W_j}{b_n}\right) K\left(\frac{z-Z_j}{b_n}\right) \right|^2 \\
&= O\left(\frac{1}{nb_n^4} \left(\inf_{|t|\leq b_n^{-1}} |f_\epsilon^{\mathrm{ft}}(t)| \right)^{-2} \iiint \left| K\left(\frac{z-v}{b_n}\right) \right|^2 \{E[Y_j^{2k} T_j^s (1-T_j)^{1-s} | X,Z] f_{X,Z}\}(u,v) f_\epsilon(\eta) du dv d\eta \right) \\
&= O\left(\frac{1}{nb_n^3} \left(\inf_{|t|\leq b_n^{-1}} |f_\epsilon^{\mathrm{ft}}(t)| \right)^{-2} \iint |K(\tilde{v})|^2 \{E[Y_j^{2k} T_j^s (1-T_j)^{1-s} | X,Z] f_{X,Z}\}(u, z-b_n\tilde{v}) du d\tilde{v} \right) \\
&= O\left(\frac{1}{nb_n^3} \left(\inf_{|t|\leq b_n^{-1}} |f_\epsilon^{\mathrm{ft}}(t)| \right)^{-2} \right),
\end{aligned}
$$

where the first step follows by the random sampling (Assumption 4 (i)), the second step follows by the fact that K^{ft} is supported on $[-1,1]$ (Assumption 4 (iii)), the third step uses the change of variables $\tilde{v} = \frac{z-v}{b_n}$, and the last step follows by the boundedness of $E[Y_j^2(s)|X,Z]$ for $s = 0,1$ (Assumption 4 (i)), the smoothness of $f_{X,Z}$ (Assumption 4 (ii)), and properties of the γ-th order kernel function K (Assumption 4 (iii)). ∎