ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Estimation and inference for the counterfactual distribution and quantile functions in continuous treatment models

Chunrong Ai [a], Oliver Linton [b], Zheng Zhang [c,*]

[a] School of Management and Economics, Chinese University of Hong Kong, Shenzhen, China
[b] Faculty of Economics, University of Cambridge, United Kingdom
[c] Center for Applied Statistics, Institute of Statistics & Big Data, Renmin University of China, China

ARTICLE INFO

Article history:
Received 13 July 2019
Received in revised form 28 July 2020
Accepted 13 December 2020
Available online xxxx

JEL classification:
C01
C12
C21

Keywords:
Continuous treatment
Counterfactual distribution
Hypothesis testing
Quantile function

ABSTRACT

Donald and Hsu (2014) studied the estimation and inference for the counterfactual distribution and quantile functions in a binary treatment model. We extend their work to the continuous treatment model. Specifically, we propose a weighted regression estimator for the counterfactual distribution but we estimate the weighting function from a covariate balancing equation by maximizing a globally concave criterion function. We estimate the quantile function by inverting the estimated counterfactual distribution. To test the distributional effect, we consider the (uniform) confidence bands, the sup and L_2 distance, and the Mann–Whitney test. We also consider the stochastic dominance test for the distributional effect and the L_2 test for constant quantiles. A simulation study reveals that our tests exhibit a satisfactory finite-sample performance, and an application shows their practical value.

1. Introduction

Structural economic models have played an important role in empirical analyses and policy evaluation. For example, in the context of labor economics, Keane and Wolpin (1997) and Low et al. (2010) employed a structural model to study labor supply; Blundell et al. (2016) modeled how tax credit reform affects life-cycle female labor supply and human capital accumulation. Further, Low and Pistaferri (2015) studied the long-run effects of reforms to disability insurance. Structural models are also employed in studying market competition, firm and consumer behavior, and industry dynamics in the context of the new empirical industrial organization (e.g., Berry et al., 1995; Goldberg, 1995). For a recent survey on the use of structural economic models, see Low and Meghir (2017). The advantage of a structural model is its ability to distinguish between an economic agent's preferences and the economic environment. It relates outcomes to preferences and relevant factors in the economic environment and identifies mechanisms for outcomes' determination, allowing the researcher to analyze counterfactual policies and to quantify their impacts on specific outcomes (see Heckman et al., 1998b; Lee and Wolpin, 2006; Abbott et al., 2019 for an example on changes in the labor market equilibrium; Chiappori et al., 2018 on the marriage market equilibrium; and, Burdett and Mortensen, 1998 on wages in the frictional labor market equilibrium). The disadvantage of the structural model is that the model is highly nonlinear with no analytical solution; we cannot estimate the model through the traditional maximum likelihood or linear regression methods. In applications where the researcher is only interested in a partially specified structural model such as one with conditional moment

* Corresponding author.
E-mail addresses: chunrongai@cuhk.edu.cn (C. Ai), obl20@cam.ac.uk (O. Linton), zhengzhang@ruc.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.jeconom.2020.12.009
0304-4076/

restrictions, the researcher can estimate the model through the generalized method of moments (GMM; e.g., Gallant et al., 1990). In applications where the researcher is interested in a fully specified structural model, the researcher must fit the structural model numerically and estimate the model by matching the model predicted moments with the sample moments (e.g., Gallant and Tauchen, 1996; Low et al., 2010; Low and Pistaferri, 2015).

Treatment effect models, however, identify the specific causal effect of a policy intervention without referring to a specific economic model. In contrast with structural models, treatment effect models are easy to estimate but they cannot distinguish between preferences and the environment and they cannot identify the mechanisms that determine the outcomes; thereby, their external validity is limited. One proposal is to exploit the advantage of both models. For example, we can fit a structural model to the control group and use the fitted model to predict the outcomes for the treatment group. We then validate the structural model by comparing the predicted outcomes with the observed outcomes. This idea works only for randomized experiments, where the control group and the treatment group are randomly specified. For observational data, however, the two groups are not randomly specified. Fitting the structural model to the control group in this case will give biased estimates and a false validation. Thus, developing a method to estimate the counterfactual distributions or their moments consistently from observational data is the first step for the estimation and validation of a structural model. This study should be viewed as contributing an approach to the aforementioned first step.

There is a large and growing literature on treatment effect models. This literature is mostly concerned with the estimation of moments such as the average treatment effect or the quantile treatment effect (e.g., Heckman et al., 1998a; Hahn, 1998; Imbens, 2000; Hirano et al., 2003; Heckman and Vytlacil, 2005; Firpo, 2007; Florens et al., 2008; Cattaneo, 2010; Chan et al., 2016; Belloni et al., 2017; Abadie and Cattaneo, 2018). For standard quantile regression allowing for continuous covariates subject to parametric restrictions, see Koenker and Bassett (1978) and Firpo et al. (2009). However, since the potential outcomes in these studies are random, the average treatment effect and the quantile treatment effect provide only a partial view of the distributional effects. A complete evaluation should compare how the potential outcomes are distributed. This is exactly what Chernozhukov et al. (2013) and Donald and Hsu (2014) did for the binary treatment model. This study extends their work to the continuous treatment model.

Specifically, we first show that, in the continuous treatment model, the counterfactual distribution is a weighted expectation conditional on the treatment variable, where the weighting function is the ratio of the probability density of the treatment variable and the probability density of the treatment variable conditional on a set of covariates. We then estimate the weighting function (but not the two densities in the ratio separately) from a covariate balancing equation by maximizing a globally concave criterion function; the procedure is computationally fast and stable. The proposed estimation procedure generalizes the covariate balancing approach for the binary treatment (Imai and Ratkovic, 2014) to the continuous treatment model. Finally, we estimate the counterfactual distribution by plugging the estimated weighting function in a kernel regression. We estimate the quantiles by inverting the estimated distribution function, and we estimate the dose response function by using the estimated counterfactual distribution. We show that our estimated distribution and quantile function weakly converge to a Gaussian process at a rate of $\sqrt{Nh}$, where h is a shrinking bandwidth. The slower convergence rate is due to the continuous treatment variable. Despite the slower convergence rate, our estimated distribution and quantile function are more efficient than the ones obtained with the true weighting function. A similar observation is also reported by Hirano et al. (2003) for the average treatment effect in the binary treatment model. Here, we show that their observation holds more generally. We also show that our estimated counterfactual distribution is generally more efficient than the one obtained when the weighting function is obtained by estimating the numerator and denominator densities separately.

To test the distributional effects, we consider three null hypotheses: A. No distributional difference between two levels of treatment. B. Negative distributional difference between two levels of treatment. C. No quantile difference for any levels of treatment. Hypotheses A and B are pair-wise comparisons of the treatments, while hypothesis C is a uniform comparison of all the treatments. We present three classes of tests for hypothesis A: (1) point-wise and uniform confidence bands, (2) a sup and L_2 distance test, and (3) the Mann–Whitney test. For hypotheses B and C, we present a stochastic dominance test and an L_2 distance test, respectively. The test statistics for the pair-wise confidence band, the Mann–Whitney test, and the L_2 distance test for hypothesis C all have a familiar asymptotic distribution so critical values are easy to compute. The statistics for the uniform band, the distance tests for hypothesis B, and the stochastic dominance test all have a nonstandard asymptotic distribution so we compute the critical values through a bootstrap method. Each of these tests has its own merits and weaknesses. In applications, one may apply some or all of them to obtain more precise inference on the distributional effects.

We devote the rest of the paper to the estimation and test procedures outlined above. Specifically, Section 2 sets up the basic framework, Section 3 presents the estimation procedure, Section 4 derives the large-sample properties of the proposed estimators, and Section 6 presents the statistical tests of the distributional and quantile effects. Subsequently, Section 7 reports the results of a simulation study and Section 8 applies the proposed estimation and test procedure to analyze the effect of non-labor income on labor supply. Finally, Section 9 provides some concluding remarks. All the technical proofs appear in the supplemental material (Ai et al., 2020).

2. Basic framework and notation

Let T denote the observed treatment variable with probability density function $f_T(t)$ and support $\mathcal{T} \subset \mathbb{R}$. Let $Y(t)$ denote the potential outcome for treatment t and let $F_t(y)$ denote its cumulative distribution function. We have observations on T,

$Y = Y(T)$, and a vector of covariates $\boldsymbol{X}$. Let $\{T_i, \boldsymbol{X}_i, Y_i\}_{i=1}^N$ denote an independently and identically distributed (i.i.d.) sample of observations drawn from the joint distribution of $(T, \mathbf{X}, Y)$. The goals of this study are (1) to estimate the counterfactual distribution function $F_t(\cdot)$ and its quantile function $q_t(\tau) = \inf\{z : F_t(z) \geq \tau\}$ for $\tau \in [0, 1]$ for all treatment statuses $t \in \mathcal{T}$ and (2) to propose statistical tests for the distributional and quantile effects.

Since the potential outcomes $\{Y(t)\}_{t\in\mathcal{T}}$ are not simultaneously observed, the counterfactual distribution functions $\{F_t(y)\}_{t\in\mathcal{T}}$ cannot be identified without a restriction. Following Rosenbaum and Rubin (1983), we impose the following un-confounded assignment condition:

Assumption 1. For any $t \in \mathcal{T}$, $T \perp Y(t)|\boldsymbol{X}$.

Let $f_{T|X}(t|x)$ denote the conditional probability density function of T given $\boldsymbol{X}$. Hirano and Imbens (2004) and Imai and van Dyk (2004) called $f_{T|X}(t|x)$ the *generalized propensity score*. Let

$$\pi_0(T, \boldsymbol{X}) = \frac{f_T(T)}{f_{T|X}(T|\boldsymbol{X})}.$$

Under Assumption 1, we show in Appendix A that

$$F_t(y) = \mathbb{E}\left\{\pi_0(T, \boldsymbol{X})I(Y \leq y)|T = t\right\}. \tag{2.1}$$

Noticing that $\mathbb{E}\{\pi_0(T, \boldsymbol{X})|T = t\} = 1$, we can write

$$F_t(y) = \frac{\mathbb{E}\{\pi_0(T, \boldsymbol{X})I(Y \leq y)|T = t\}}{\mathbb{E}\{\pi_0(T, \boldsymbol{X})|T = t\}}. \tag{2.2}$$

If $\pi_0(T, \boldsymbol{X})$ were known, we would estimate both the numerator and the denominator by kernel regression and estimate $F_t(y)$ by the ratio:

$$\overline{F}_{t,h}(y) = \frac{\sum_{i=1}^N \pi_0(T_i, \boldsymbol{X}_i)I(Y_i \leq y)K\left(\frac{T_i - t}{h}\right)}{\sum_{i=1}^N \pi_0(T_i, \boldsymbol{X}_i)K\left(\frac{T_i - t}{h}\right)}, \tag{2.3}$$

where $K(\cdot)$ is a known univariate kernel function and h is a bandwidth. It is easy to see that the value of $\overline{F}_{t,h}(y)$ always lies in [0, 1]. Unfortunately, the estimator $\overline{F}_{t,h}(y)$ is infeasible because the weighting function $\pi_0(T, \boldsymbol{X})$ is unknown. Next, we turn our attention to the estimation of the weighting function and the counterfactual distribution function.

3. Estimation procedure

One obvious approach for estimating the weighting function is to estimate $f_T(T)$ and $f_{T|X}(T|\boldsymbol{X})$. The estimators are denoted by $\widehat{f}_T(T)$ and $\widehat{f}_{T|X}(T|\boldsymbol{X})$ respectively. We denote this ratio estimator by $\widetilde{\pi}(T, \boldsymbol{X}) = \widehat{f}_T(T)/\widehat{f}_{T|X}(T|\boldsymbol{X})$ and the resulting estimator of the counterfactual distribution by

$$\widetilde{F}_{t,h}(y) = \frac{\sum_{i=1}^N \widetilde{\pi}(T_i, \boldsymbol{X}_i)I(Y_i \leq y)K\left(\frac{T_i - t}{h}\right)}{\sum_{i=1}^N \widetilde{\pi}(T_i, \boldsymbol{X}_i)K\left(\frac{T_i - t}{h}\right)}.$$

There are three drawbacks to this approach. First, $\widetilde{\pi}(T, \boldsymbol{X})$ does not increase the efficiency of $\widetilde{F}_{t,h}(y)$. In Theorem 4, we derive the asymptotic properties of $\widetilde{F}_{t,h}(y)$ and show that our estimator $\widehat{F}_{t,h}(y)$ (defined below) is more efficient than $\widetilde{F}_{t,h}(y)$, except when the density estimates are carefully under-smoothed. In the latter case, $\widehat{F}_{t,h}(y)$ is as efficient as $\widetilde{F}_{t,h}(y)$. Second, $\widetilde{\pi}(T, \boldsymbol{X})$ is very sensitive to small values of $f_{T|X}(T|\boldsymbol{X})$ since small estimation errors result in large estimation errors in $\widetilde{\pi}(T, \boldsymbol{X})$. Third, $\widetilde{F}_{t,h}(y)$ is not guaranteed to lie in [0, 1] and can even be negative since we must use higher-order kernels to remove the biases. To avoid or mitigate these problems, we estimate the weighting function $\pi_0(T, \boldsymbol{X})$ directly. We note that the weighting function satisfies

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X})u(T)v(\boldsymbol{X})\right] = \mathbb{E}[u(T)] \cdot \mathbb{E}[v(\boldsymbol{X})] \tag{3.1}$$

for any suitable functions $u(t)$ and $v(\boldsymbol{x})$. The following theorem shows that this restriction identifies the weighting function.

Theorem 1. $\mathbb{E}\left[\pi(T, \boldsymbol{X})u(T)v(\boldsymbol{X})\right] = \mathbb{E}[u(T)] \cdot \mathbb{E}[v(\boldsymbol{X})]$ *holds for all integrable functions* $u(T)$ *and* $v(\boldsymbol{X})$ *if and only if* $\pi(T, \boldsymbol{X}) = \pi_0(T, \boldsymbol{X})$ *a.s.*

This result suggests a possible way to estimate the weighting function. The challenge is that (3.1) implies an infinite number of equations. With a finite sample of observations, it is impossible to solve this infinite number of equations. To overcome this difficulty, we approximate the infinite-dimensional function space by a sequence of finite-dimensional sieve spaces. Specifically, let $u_{K_1}(T) = (u_{K_1,1}(T), \ldots, u_{K_1,K_1}(T))^\top$ and $v_{K_2}(\boldsymbol{X}) = \left(v_{K_2,1}(\boldsymbol{X}), \ldots, v_{K_2,K_2}(\boldsymbol{X})\right)^\top$ denote known basis functions with dimensions $K_1 \in \mathbb{N}$ and $K_2 \in \mathbb{N}$, respectively, and let $K = K_1 \cdot K_2$. The functions $u_{K_1}(t)$ and $v_{K_2}(\boldsymbol{x})$ are

approximation sieves that can approximate any suitable functions $u(t)$ and $v(\boldsymbol{x})$ arbitrarily well (see Elbadawi et al., 1983; Gallant and Nychka, 1987; Gallant and Tauchen, 1989, 1996; Coppejans and Gallant, 2002; Ai and Chen, 2003; Chen, 2007 for discussions on the sieve approximation). Since the sieve approximating space is a subspace of the original function space, $\pi_0(T, \boldsymbol{X})$ also satisfies

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right] = \mathbb{E}[u_{K_1}(T)] \cdot \mathbb{E}[v_{K_2}(\boldsymbol{X})]^\top. \tag{3.2}$$

Unfortunately, $\pi_0(T, \boldsymbol{X})$ is not the only solution to (3.2). Indeed, for any increasing and globally concave function $\rho(v)$, with

$$\Lambda^*_{K_1 \times K_2} = \arg \max_{\Lambda \in \mathbb{R}^{K_1 \times K_2}} \mathbb{E}\left[\rho(u_{K_1}(T)^\top \Lambda v_{K_2}(\boldsymbol{X}))\right] - \mathbb{E}[u_{K_1}(T)]^\top \Lambda \mathbb{E}[v_{K_2}(\boldsymbol{X})], \tag{3.3}$$

$\pi^*_K(T, \mathbf{X}) = \rho'\left(u_{K_1}(T)^\top \Lambda^*_{K_1 \times K_2} v_{K_2}(\boldsymbol{X})\right)$ also solves (3.2), where $\rho'(v)$ denotes the first derivative. Since there are many $\rho(\cdot)$ functions, there are many solutions. We shall choose a $\rho(\cdot)$ that has an intuitive interpretation and aids the asymptotic distribution derivations.

For our choice of $\rho(\cdot)$, we start with the traditional approach, which would estimate the weights by maximizing the log-likelihood (or generalized log-likelihood) function, subject to the sample analog of (3.2). We consider the generalized empirical likelihood (EL) estimation method that solves the following entropy maximization problem:

$$\begin{cases} \{\widehat{\pi}_i\}_{i=1}^N = \arg\max \left\{-N^{-1} \sum_{i=1}^N \pi_i \log \pi_i\right\} \\ \text{subject to } \frac{1}{N} \sum_{i=1}^N \pi_i u_{K_1}(T_i) v_{K_2}(\boldsymbol{X}_i)^\top = \left(\frac{1}{N} \sum_{i=1}^N u_{K_1}(T_i)\right)\left(\frac{1}{N} \sum_{j=1}^N v_{K_2}(\boldsymbol{X}_j)^\top\right). \end{cases} \tag{3.4}$$

Two observations are immediate. First, by including a constant of one in the sieve base functions, (3.4) guarantees that $N^{-1} \sum_{i=1}^N \widehat{\pi}_i = 1$. Second, we notice that

$$\max \left\{-N^{-1} \sum_{i=1}^N \pi_i \log \pi_i\right\} = -\min \left\{\sum_{i=1}^N \{N^{-1}\pi_i\} \cdot \log \frac{N^{-1}\pi_i}{N^{-1}}\right\}.$$

The entropy maximization problem minimizes the Kullback–Leibler divergence between the weights $\{N^{-1}\pi_i\}_{i=1}^N$ and the empirical frequencies $\{N^{-1}\}$, subject to the sample analogue of (3.2). This is similar to the exponential tilting problem considered in Kitamura and Stutzer (1997) and Imbens et al. (1998) . The difference is that their problem is parametric while ours is non-parametric.

The entropy maximization problem is difficult to solve. However, setting $\rho(v) = -\exp(-v - 1)$ for any $v \in \mathbb{R}$, we show in Appendix C that the dual problem to the entropy maximization problem has the solution

$$\hat{\pi}_K(T_i, \boldsymbol{X}_i) = \rho'\left(u_{K_1}(T_i)^\top \hat{\Lambda}_{K_1 \times K_2} v_{K_2}(\boldsymbol{X}_i)\right),$$

where $\hat{\Lambda}_{K_1 \times K_2}$ maximizes the globally concave function

$$\widehat{G}_{K_1 \times K_2}(\Lambda) = \frac{1}{N} \sum_{i=1}^N \rho\left(u_{K_1}(T_i)^\top \Lambda v_{K_2}(\boldsymbol{X}_i)\right) - \left(\frac{1}{N} \sum_{i=1}^N u_{K_1}(T_i)\right)^\top \Lambda \left(\frac{1}{N} \sum_{j=1}^N v_{K_2}(\boldsymbol{X}_j)\right). \tag{3.5}$$

Thus, the choice $\rho(v) = -\exp(-v - 1)$ corresponds to the generalized EL estimation method. Moreover, it satisfies the invariance property $-\rho''(v) = \rho'(v)$, which greatly simplifies the asymptotic distribution derivations.

Having estimated the weighting function, we proceed by estimating the counterfactual distribution function $F_t(y)$ by plugging in the estimated weighting function

$$\widehat{F}_{t,h}(y) = \frac{\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i) I(Y_i \le y) K\left(\frac{T_i - t}{h}\right)}{\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i) K\left(\frac{T_i - t}{h}\right)}.$$

We estimate the quantile function of $Y(t)$ by inverting the estimated distribution function:

$$\widehat{q}_{t,h}(\tau) = \inf\left\{y : \widehat{F}_{t,h}(y) \ge \tau\right\}, \ \tau \in (0, 1).$$

We estimate the dose response function $m(t) = \mathbb{E}[Y(t)] = \int y dF_t(y)$ by plugging in the estimated counterfactual distribution:

$$\widehat{m}_h(t) = \int y d\widehat{F}_{t,h}(y) = \frac{\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i) K\left(\frac{T_i - t}{h}\right) Y_i}{\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i) K\left(\frac{T_i - t}{h}\right)}.$$

Galvao and Wang (2015) and Alejo et al. (2018) also studied the continuous treatment model. They considered a general setting, in which the parameter of interest $\beta(t)$ uniquely solves the moment condition

$$\mathbb{E}[M(Y(t);\beta(t))] = 0 \text{ for all } t \in T, \tag{3.6}$$

where $M(\cdot)$ is a known function. Their setting encompasses our model. For example, $M(Y(t);\beta(t)) = Y(t) - \beta(t)$ gives the dose response function $\beta(t) = m(t)$, $M(Y(t);\beta(t)) = I(Y(t) \geq \beta(t)) - \tau$ gives the $\tau-$quantile $\beta(t) = q_t(\tau)$, and $M(Y(t);\beta(t)) = I(Y(t) \leq y) - \beta(t)$ gives the counterfactual distribution function $\beta(t) = F_t(y)$. Under the un-confounded assignment condition, Galvao and Wang (2015) derived the following moment condition:

$$\mathbb{E}\left[w_0(\boldsymbol{X}, Y; t)M(Y;\beta(t))\right] = 0, \tag{3.7}$$

where $w_0(x, y; t) = f_{T|X,Y}(t|x, y)/f_{T|X}(t|x)$ is their weighting function. With an initial consistent estimate $\widehat{w}(x, y; t)$, they estimated $\beta(t)$ by solving the empirical analogue of (3.7):

$$\frac{1}{N}\sum_{i=1}^{N} \widehat{w}(\boldsymbol{X}_i, Y_i; t)M(Y_i;\widehat{\beta}(t)) = 0.$$

(Galvao and Wang, 2015) derived the asymptotic distribution of $\widehat{\beta}(t)$ uniformly over t under some high-level assumption on $\widehat{w}(x, y; t)$. However, they noted that a non-parametrically estimated $w(x, y; t)$ does not satisfy their high-level assumption but a parametrically estimated $w(x, y; t)$ does. In a sequel study, Alejo et al. (2018) estimated $w(x, y; t)$ by using the Box–Cox model.

There are several differences between their approach and ours. The first difference is the weighting function. Their weighting function depends on all observed variables while our weighting function does not depend on the observed outcome variable. This difference implies that the non-parametric estimate of their weighting function converges to the true function at a rate slower than the non-parametric estimate of our weighting function. Consequently, their estimator of the counterfactual distribution (as well as the quantile and dose response) function converges to the true function at a rate slower than that of our estimator. The second difference is how the weighting function is estimated. We estimate our weighting function by solving the sample analogue of the covariate balancing equation (3.2), which guarantees that the estimated weights are positive and have a sum of one. They placed no corresponding restriction on their weighting function so they had to estimate the weighting function by the ratio of the estimated conditional densities. Their estimated weights could be negative if a higher-order kernel is used. Moreover, if the value of $f_{T|X}(t|x)$ is small, small estimation errors in $\widehat{f}_{T|X}(t|x)$ lead to large estimation errors in the estimated weights, resulting in an unstable estimate of $\beta(t)$.

The covariate balancing equation would improve the finite sample performance of the estimator even if we specified a parametric weighting function. For example, in the binary treatment model, Kang and Schafer (2007) and Smith and Todd (2005) documented that the popular propensity score method could give a substantially biased estimate of the average treatment effect when the propensity score function $P_0(\boldsymbol{X}) = P(T = 1|\boldsymbol{X})$ is slightly mis-specified. Imai and Ratkovic (2014) noticed that the true propensity score function satisfies

$$\mathbb{E}\left[T \cdot P_0(\boldsymbol{X})^{-1} v(\boldsymbol{X})\right] = \mathbb{E}[v(\boldsymbol{X})] \text{ for all integrable functions } v(\boldsymbol{X}).$$

For a parametric propensity score $P(\boldsymbol{X};\gamma)$, they imposed the covariate balancing equation

$$\mathbb{E}\left[T \cdot P(\boldsymbol{X};\gamma)^{-1} v_{K_2}(\boldsymbol{X})\right] = \mathbb{E}[v_{K_2}(\boldsymbol{X})], \tag{3.8}$$

where $v_{K_2}(\boldsymbol{X})$ is a K_2-dimensional basis functions, with K_2 fixed and possibly larger than the dimension of γ. They estimated γ from the covariate balancing equation (3.8) by GMM estimation or EL estimation. Although neither the GMM nor EL enforces the sample analogue of (3.8), they documented that their estimated propensity score substantially improves the finite sample performance of the average treatment effect estimator. Notice that (3.8) is a special case of (3.2), with $u_{K_1}(T) = T$ and $\pi_0(T, \boldsymbol{X}) = T \cdot P(T = 1)/P_0(\boldsymbol{X}) + (1 - T) \cdot P(T = 0)/(1 - P_0(\boldsymbol{X}))$. Evidently, we impose more restrictions (i.e., $K_1 > 1$) and enforce the sample analogue of (3.2). Thus, there is reason to believe our procedure could have a similar finite sample performance.

4. Large sample properties

4.1. Preliminaries

To derive the large-sample properties of the estimated distribution, quantile, and dose response functions, we first show that the estimated weighting function $\widehat{\pi}_K(t, \boldsymbol{x})$ is consistent and compute its convergence rate under the L_2 norm. We impose the following conditions throughout the study:

Assumption 2. (i) The support $\mathcal{X}$ of the control variables $\boldsymbol{X}$ is a compact subset of $\mathbb{R}^r$. The support $\mathcal{T}$ of the treatment variable T is a compact subset of $\mathbb{R}$. (ii) There exist two positive constants η_1 and η_2 such that

$$0 < \eta_1 \leq \pi_0(t, \boldsymbol{x}) \leq \eta_2 < \infty\ ,\ \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}\ .$$

Assumption 3. There exist a $\Lambda_{K_1\times K_2} \in \mathbb{R}^{K_1\times K_2}$ and a constant $\alpha > 0$ such that

$$\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} \left|\rho'^{-1}(\pi_0(t,\boldsymbol{x})) - u_{K_1}(t)^\top \Lambda_{K_1\times K_2} v_{K_2}(\boldsymbol{x})\right| = O(K^{-\alpha}),$$

where $\rho(v) = -\exp(-v-1)$ and $K = K_1 \cdot K_2$.

Assumption 4. (i) The smallest eigenvalues of $\mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]$ and $\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right]$ are bounded away from zero uniformly in K_1 and K_2. (ii) There are two sequences of constants $\zeta_1(K_1)$ and $\zeta_2(K_2)$ satisfying $\sup_{t\in\mathcal{T}}\|u_{K_1}(t)\| \le \zeta_1(K_1)$ and $\sup_{\boldsymbol{x}\in\mathcal{X}}\|v_{K_2}(\boldsymbol{x})\| \le \zeta_2(K_2)$ such that $\sqrt{N}K^{-\alpha} \to 0$ and $\zeta(K)\sqrt{K^2/N} \to 0$ as $N \to \infty$, where $K = K_1 \cdot K_2$ and $\zeta(K) = \zeta_1(K_1)\zeta_2(K_2)$.

Assumption 2(i) requires the covariates and the treatment variable to be bounded. This condition, although restrictive, is commonly imposed in the non-parametric regression literature. However, we can replace it with a restriction on the tail distribution of $(\boldsymbol{X}, T)$. For example, Chen et al. (2008, Assumption 3) assumed that the support of$\boldsymbol{X}$ is the entire Euclidean space but imposed $\int_{\mathbb{R}^r}(1+|x|^2)^\omega f_X(x)dx < \infty$ for some $\omega > 0$. Assumption 2 (ii) requires the weighting function to be bounded and bounded away from zero. We can relax Assumption 2 (ii) by allowing η_1 (η_2) to go to zero (infinity) slowly as $N \to \infty$. Notice that $u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x})$ is a linear sieve approximation for $\rho'^{-1}(\pi_0(t,\boldsymbol{x}))$. Assumption 3 requires the sieve approximation error to shrink to zero at a polynomial rate. A variety of sieve basis functions satisfy this condition. The polynomial rate is positively affected by the smoothness of $\rho'^{-1}(\pi_0(t,\boldsymbol{x}))$ and negatively affected by the number of continuous covariates. Assumption 4 (i) ensures that the sieve estimator is non-degenerate. This condition is common in the sieve regression literature (see Andrews, 1991; Newey, 1997). If the approximation error is nonzero, Assumption 4 (ii) imposes a restriction on the growth rate of the smoothing parameters K_1 and K_2 to ensure under-smoothing. If the approximation error is zero for some fixed K, Assumption 4 (ii) is not needed.

Under these conditions, we prove the following results:

Proposition 1. *Assume that Assumptions 2–4 hold. We have*

$$\int_{\mathcal{T}\times\mathcal{X}} |\widehat{\pi}_K(t,\boldsymbol{x}) - \pi_0(t,\boldsymbol{x})|^2 dF_{T,X}(t,\boldsymbol{x}) = O_p\left(\max\left\{K^{-2\alpha}, \frac{K}{N}\right\}\right),$$

$$\frac{1}{N}\sum_{i=1}^N |\widehat{\pi}_K(T_i,\boldsymbol{X}_i) - \pi_0(T_i,\boldsymbol{X}_i)|^2 = O_p\left(\max\left\{K^{-2\alpha}, \frac{K}{N}\right\}\right).$$

The next proposition provides a representation of the influence function.

Proposition 2. *Assume that Assumptions 2–4 hold. For any square-integrable random variable $\phi(T,\boldsymbol{X},Y) \in L^2$ such that $\mathbb{E}[\phi(T,\boldsymbol{X},Y)|T=t,\boldsymbol{X}=\boldsymbol{x}]$ is continuously differentiable, we have the following representation:*

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^N \{\widehat{\pi}_K(T_i,\boldsymbol{X}_i)\phi(T_i,\boldsymbol{X}_i,Y_i) - \mathbb{E}[\pi_0(T,\boldsymbol{X})\phi(T,\boldsymbol{X},Y)]\} \\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^N \Big\{\pi_0(T_i,\boldsymbol{X}_i)\phi(T_i,\boldsymbol{X}_i,Y_i) - \pi_0(T_i,\boldsymbol{X}_i)\cdot\mathbb{E}[\phi(T_i,\boldsymbol{X}_i,Y_i)|T_i,\boldsymbol{X}_i] \\
&\quad + \mathbb{E}[\pi_0(T_i,\boldsymbol{X}_i)\phi(T_i,\boldsymbol{X}_i,Y_i)|\boldsymbol{X}_i] - \mathbb{E}[\pi_0(T,\boldsymbol{X})\phi(T,\boldsymbol{X},Y)] \\
&\quad + \mathbb{E}[\pi_0(T_i,\boldsymbol{X}_i)\phi(T_i,\boldsymbol{X}_i,Y_i)|T_i] - \mathbb{E}[\pi_0(T,\boldsymbol{X})\phi(T,\boldsymbol{X},Y)]\Big\} + o_p(1).
\end{aligned}$$

4.2. Asymptotic distribution

To derive the asymptotic properties of the estimated distribution and quantile function, we impose the following additional conditions.

Assumption 5. For each $t \in \mathcal{T}$, $Y(t)$ has a compact support.

Assumption 6. (i) For any given $(t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}$, the conditional distribution function $F_{Y|T,X}(y|t,\boldsymbol{x})$ is continuous in $y \in \mathcal{Y}$. (ii) For any $y \in \mathcal{Y}$, $F_{Y|T,X}(y|t,\boldsymbol{x})$ is continuously differentiable with respect to $(t,\boldsymbol{x}) \in \mathcal{T}\times\mathcal{X}$. (iii) The density function $f_T(t)$ is third-order continuously differentiable; (iv) The function $F_t(y)$ is continuous in y and third-order continuously differentiable with respect to t.

Assumption 7. $K(\cdot)$ is a univariate kernel function, symmetric around the origin, and satisfying (i) $\int K(u)du = 1$; (ii) $\int u^2K(u)du = \kappa_{21} \in (0,\infty)$; (iii) $\int K^2(u)du = \kappa_{02} < \infty$; and, (iv) $\int |K(u)|^{2+\delta}du < \infty$, for some $\delta > 0$.

Assumption 8. As $N \to \infty$, $h \to 0$ and $Nh \to \infty$.

We do not need Assumption 5 to derive the large-sample properties of the estimated distribution function but we need it to derive those of the estimated quantile function (also see Assumption 3.1 of Donald and Hsu (2014)). Assumption 8 is common in the kernel regression literature (see Li and Racine, 2007).

Theorem 2. *Assume that Assumptions 1–8 hold and $Nh^5 \to 0$. We show that, for any fixed $t \in \mathcal{T}$,*

$$\sup_{y\in\mathcal{Y}} \left| \sqrt{Nh}\left\{\widehat{F}_{t,h}(y) - F_t(y)\right\} - \sqrt{\frac{h}{N}} \sum_{i=1}^{N} \psi_{t,h}(Y_i, T_i, \boldsymbol{X}_i; y) \right| = o_P(1),$$

where

$$\psi_{t,h}(Y_i, T_i, \boldsymbol{X}_i; y) = \frac{\pi_0(T_i, \boldsymbol{X}_i)}{p_{t,h}} K\left(\frac{T_i - t}{h}\right)\left\{I(Y_i \le y) - F_{Y|T,X}(y|T_i, \boldsymbol{X}_i)\right\}$$

and $p_{t,h} = \mathbb{E}\left[K(\frac{T-t}{h})\right]$. Furthermore, we have

$$\sqrt{Nh}\left\{\widehat{F}_{t,h}(\cdot) - F_t(\cdot)\right\} \Rightarrow \Psi_t(\cdot).$$

where "$\Rightarrow$" denotes weakly convergence, and $\Psi_t(\cdot)$ is a mean-zero Gaussian process with covariance function

$$\Omega_t(y_1, y_2) = \lim_{h\to 0} h \cdot \mathbb{E}\left[\psi_{t,h}(Y_i, T_i, \boldsymbol{X}_i; y_1)\psi_{t,h}(Y_i, T_i, \boldsymbol{X}_i; y_2)\right] = \frac{\kappa_{02}}{f_T(t)} \times$$
$$\mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)^2 \left\{I(Y_i \le y_1) - F_{Y|T,X}(y_1|T_i, \boldsymbol{X}_i)\right\}\left\{I(Y_i \le y_2) - F_{Y|T,X}(y_2|T_i, \boldsymbol{X}_i)\right\} \middle| T_i = t\right],$$

and $\kappa_{ij} = \int u^i K^j(u)du$.

The following theorem shows that the estimated counterfactual distribution function $\widehat{F}_{t,h}(y)$ is more efficient than the infeasible $\overline{F}_{t,h}(y)$.

Theorem 3. *Assume Assumption 8 holds and $Nh^5 \to 0$. For any fixed $t \in \mathcal{T}$, we have*

$$\sqrt{Nh}\left\{\overline{F}_{t,h}(\cdot) - F_t(\cdot)\right\} \Rightarrow \overline{\Psi}_t(\cdot),$$

where $\overline{\Psi}_t(\cdot)$ is a Gaussian process with covariance function

$$\overline{\Omega}_t(y_1, y_2) = \frac{\kappa_{02}}{f_T(t)} \cdot \mathbb{E}\left[\pi_0(T, \boldsymbol{X})^2 \{I(Y \le y_1) - F_t(y_1)\}\{I(Y \le y_2) - F_t(y_2)\}|T = t\right].$$

Furthermore, we have $\Omega_t(y, y) < \overline{\Omega}_t(y, y)$. Hence, $\widehat{F}_{t,h}(y)$ is more efficient than $\overline{F}_{t,h}(y)$.

The theorem above states that a perfectly estimated weighting function is not as important as the information contained in the covariate-balancing equation. It is not clear if a less accurately estimated weighting function is still not as important. To study this problem, we consider estimating the weighting function by using separate kernel density estimates. Specifically, let $K_1(t)$, $K_2(\boldsymbol{x})$, and $K_3(\boldsymbol{x}, t)$ denote kernel density functions of orders s_1, s_2, and s_3, respectively. Although we can construct a kernel function of any order, the bias of the kernel density estimator is bounded by the smoothness condition of the density function. Since the joint density function is as smooth as the marginal density functions, it is reasonable to assume that $s_3 = \min\{s_1, s_2\}$ and let $K_3(\boldsymbol{x}, t) = K_1(t) \cdot K_2(\boldsymbol{x})$. Let r denote the dimension of $\boldsymbol{X}$. Let $K_{1,h_1}(t) = h_1^{-1}K_1\left(\frac{t}{h_1}\right)$, $K_{2,h_2}(\boldsymbol{x}) = h_2^{-r}K_2\left(\frac{\boldsymbol{x}}{h_2}\right)$, and $K_{3,h_3}(\boldsymbol{x}, t) = K_{1,h_1}(t)K_{2,h_2}(\boldsymbol{x})$. We estimate the density functions as follows:

$$\widehat{f}_T(T_i) = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} K_{1,h_1}(T_j - T_i),\ \widehat{f}_X(\boldsymbol{X}_i) = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} K_{2,h_2}(\boldsymbol{X}_j - \boldsymbol{X}_i),$$

$$\widehat{f}_{T,X}(T_i, \boldsymbol{X}_i) = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} K_{2,h_2}(\boldsymbol{X}_j - \boldsymbol{X}_i)K_{1,h_1}(T_j - T_i),$$

$$\widehat{f}_{T|X}(T_i|\boldsymbol{X}_i) = \widehat{f}_{T,X}(T_i, \boldsymbol{X}_i)/\widehat{f}_X(\boldsymbol{X}_i).$$

We estimate the weighting function as

$$\widetilde{\pi}(T_i, \boldsymbol{X}_i) = \frac{\widehat{f}_T(T_i)}{\widehat{f}_{T|X}(T_i|\boldsymbol{X}_i)}$$

and the counterfactual distribution function as

$$\widetilde{F}_{t,h}(y) = \frac{\sum_{i=1}^N \widetilde{\pi}(T_i, \boldsymbol{X}_i) I(Y_i \le y) K\left(\frac{T_i - t}{h}\right)}{\sum_{i=1}^N \widetilde{\pi}(T_i, \boldsymbol{X}_i) K\left(\frac{T_i - t}{h}\right)}.$$

We prove the following result in the supplemental materials.

Theorem 4. *Assume that* $Nh_2^r h_1 \to \infty$, $Nhh_1^{2s_1} \to 0$, $Nhh_2^{2s_2} \to 0$, $Nh^5 \to 0$, *and* $\frac{\log^2 N}{N} \cdot \frac{h}{h_1^2 h_2^{2r}} \to 0$. *In addition, assume that one of the following conditions holds:*

$$(C1): \ \frac{h}{h_1} \to 0 \text{ and } \frac{Nh^5}{h_1^4} \to 0; \ (C2): \ \frac{h_1}{h} \to 0 \text{ and } \frac{Nh_1^{2s_1}}{h^{2s_1 - 1}} \to 0.$$

Then, for any fixed $t \in \mathcal{T}$, *we have*

$$\sqrt{Nh}\left\{\widetilde{F}_{t,h}(\cdot) - F_t(\cdot)\right\} \Rightarrow \widetilde{\Psi}_t(\cdot),$$

where $\widetilde{\Psi}_t(\cdot)$ *is a Gaussian process with covariance function* $\widetilde{\Omega}_t(y_1, y_2)$ *given by*

$$\widetilde{\Omega}_t(y_1, y_2) = \overline{\Omega}_t(y_1, y_2)$$

if (C_1) *holds, and*

$$\widetilde{\Omega}_t(y_1, y_2) = \Omega_t(y_1, y_2)$$

if (C_2) *holds.*

The theorem above states that, for the estimated distribution function to have a familiar limiting distribution, the bandwidths h and h_1 cannot shrink at the same rate. If h shrinks at a faster rate than h_1, $\widetilde{F}_{t,h}(y)$ is less efficient than $\widehat{F}_{t,h}(y)$. This can happen if h shrinks at a rate slightly faster than $N^{-1/5}$ and h_1 shrinks at a rate slightly faster than $N^{-\frac{1}{2s_1+1}}$ with $s_1 > 2$. Both rates minimize the mean-squared error of the estimated density. On the other hand, if h_1 shrinks at a faster rate than h, $\widetilde{F}_{t,h}(y)$ is as efficient as $\widehat{F}_{t,h}(y)$. However, this would require a careful selection of (h, h_1, h_2), a large degree of under-smoothing in $\widehat{f}_T(T_i)$ regardless of the order of $K_1(\cdot)$, and a very high-order kernel $K_2(\cdot)$. In particular, we cannot choose h_1 to minimize the mean-squared error of $\widehat{f}_T(T_i)$. Given the difficulty in calibrating the bandwidths to satisfy all conditions in the theorem, case (C_2) is less likely to happen in applications. Additionally, we have to contend with the possibility of a negative distribution estimate. In Section 7 of the supplemental material, we compare the performance between the proposed estimator $\widehat{F}_{t,h}$ and the plug-in estimator $\widetilde{F}_{t,h}$ via a simulation study. The simulation results indicate that the proposed estimator dominates the naive estimator.

Next, we derive the asymptotic properties of statistics that are functions of the estimated distribution function. We begin with the quantile function, which is Hadamard differentiable. By applying the delta method and Theorem 20.8 and Lemma 21.3 of van der Vaart (1998), we have

$$\sqrt{Nh}\left\{\widehat{q}_{t,h}(\tau) - q_t(\tau)\right\} = -\frac{1}{f_{Y(t)}(q_t(\tau))} \cdot \sqrt{Nh}\left\{\widehat{F}_{t,h}(q_t(\tau)) - F_t(q_t(\tau))\right\} + o_p(1),$$

where $f_{Y(t)}$ is the density function of $Y(t)$. The following result is an application of Theorem 2.

Corollary 5. *Assume that Assumptions 1–8 hold and* $Nh^5 \to 0$. *We show that, for any fixed* $t \in \mathcal{T}$,

$$\sqrt{Nh}\left\{\widehat{q}_{t,h}(\cdot) - q_t(\cdot)\right\} \Rightarrow Q_t(\cdot),$$

where $Q_t(\cdot)$ *is a mean-zero Gaussian process with covariance function* $\Gamma_t(\tau_1, \tau_2)$ *for all* $(\tau_1, \tau_2) \in [0, 1] \times [0, 1]$, *where*

$$\begin{aligned}\Gamma_t(\tau_1, \tau_2) = &\frac{1}{f_{Y(t)}(q_t(\tau_1)) f_{Y(t)}(q_t(\tau_2))} \times \frac{\kappa_{02}}{f_T(t)} \\ &\times \mathbb{E}\Big[\pi_0(T, \boldsymbol{X})^2 \left\{I(Y \le q_t(\tau_1)) - F_{Y|T,X}(q_t(\tau_1)|T, \boldsymbol{X})\right\} \\ &\quad \times \left\{I(Y \le q_t(\tau_2)) - F_{Y|T,X}(q_t(\tau_2)|T, \boldsymbol{X})\right\} \Big| T = t\Big].\end{aligned}$$

We also apply Theorem 2 to the estimated dose response function and obtain the following corollary:

Corollary 6. *Assume that Assumptions 1–8 hold and* $Nh^5 \to 0$. *For any fixed* $t \in \mathcal{T}$, *we have*

$$\sqrt{Nh}\left\{\widehat{m}_h(t) - m(t)\right\} \xrightarrow{d} N(0, V_t), \text{ with}$$

$$V_t = \frac{\kappa_{02}}{f_T(t)} \cdot \mathbb{E}\Big[\pi_0(T, \boldsymbol{X})^2 \left\{Y - \mathbb{E}[Y|T, \boldsymbol{X}]\right\}^2 \Big| T = t\Big].$$

Kennedy et al. (2017, Theorem 3) estimated the dose response function through a parametric weighting function. Although we use a non-parametric weighting function, our estimated dose response function is as efficient as theirs.

4.3. Consistent variance

In this section, we present some consistent estimates of the variance functions $\Omega_t(y,y)$, $\Gamma_t(y,y)$ and V_t. Let

$$\widehat{p}_{t,h} = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{T_i - t}{h}\right), \ \widehat{f}_{Y(t)}(y) = \partial_t \widehat{F}_t(y),$$

$$\widehat{F}_{Y|T,X}(y|t,\boldsymbol{x}) = \frac{\sum_{i=1}^N I(Y_i \le y) K\left(\frac{T_i - t}{h_T}\right)\prod_{j=1}^r K\left(\frac{X_{ji}-x_j}{h_X}\right)}{\sum_{i=1}^N K\left(\frac{T_i - t}{h_T}\right)\prod_{j=1}^r K\left(\frac{X_{ji}-x_j}{h_X}\right)},$$

$$\widehat{E}[Y|T=t,\boldsymbol{X}=\boldsymbol{x}] = \frac{\sum_{i=1}^N Y_i K\left(\frac{T_i - t}{h_T}\right)\prod_{j=1}^r K\left(\frac{X_{ji}-x_j}{h_X}\right)}{\sum_{i=1}^N K\left(\frac{T_i - t}{h_T}\right)\prod_{j=1}^r K\left(\frac{X_{ji}-x_j}{h_X}\right)}$$

and

$$\widehat{\psi}_{t,h}(T_i,\boldsymbol{X}_i,Y_i;y) = \frac{\widehat{\pi}_K(T_i,\boldsymbol{X}_i)}{\widehat{p}_{t,h}} K\left(\frac{T_i - t}{h}\right)\left\{I(Y_i \le y) - \widehat{F}_{Y|T,X}(y|T_i,\boldsymbol{X}_i)\right\}.$$

We estimate the variance functions as follows:

$$\widehat{\Omega}_t(y,y) = \frac{h}{N}\sum_{i=1}^{N}\left\{\widehat{\psi}_{t,h}(T_i,\boldsymbol{X}_i,Y_i;y)\right\}^2,$$

$$\widehat{\Gamma}_t(\tau,\tau) = \frac{1}{\widehat{f}_{Y(t)}(\widehat{q}_t(\tau))^2}\cdot\frac{h}{N}\sum_{i=1}^{N}\left\{\widehat{\psi}_{t,h}(T_i,\boldsymbol{X}_i,Y_i;\widehat{q}_{t,h}(\tau))\right\}^2,$$

$$\widehat{V}_t = \frac{h}{N}\sum_{i=1}^{N}\left\{\frac{\widehat{\pi}_K(T_i,\boldsymbol{X}_i)}{\widehat{p}_{t,h}} K\left(\frac{T_i - t}{h}\right)\left\{Y_i - \widehat{E}[Y_i|T_i,\boldsymbol{X}_i]\right\}\right\}^2.$$

The consistency results $\widehat{\Omega}_t(y,y) \xrightarrow{p} \Omega_t(y,y)$, $\widehat{\Gamma}_t(\tau,\tau) \xrightarrow{p} \Gamma_t(\tau,\tau)$, and $\widehat{V}_t \xrightarrow{p} V_t$ follow from standard argument.

5. Data-driven smoothing parameter

The large-sample properties of the proposed estimator hold for a range of values of K and h. This presents a dilemma for applied researchers, who have only one finite sample. Too little smoothing yields a large variance and too much smoothing yields a large bias. Therefore, applied researchers would like to have some guidance on the choice of K and h. In this section, we propose a cross-validation method for choosing the smoothing parameters K and h.

The proposed method combines several methods proposed in the non-parametric regression literature. We notice that the weighting function satisfies

$$\mathbb{E}[\pi_0(T,\boldsymbol{X})Y|T] = m(T).$$

We follow Härdle et al. (1988) by choosing K and h to minimize the mean-squared error

$$\mathbb{E}\left[(\widehat{\pi}_K(T,\boldsymbol{X})Y - \widehat{m}_h(T))^2\right].$$

To avoid over-fitting, we follow Kennedy et al. (2017, Section 3.5) and Li and Racine (2007, Section 15.2) and adopt the generalized cross-validation function

$$CV(K_1,K_2,h) = \left\{\left(1 - \frac{K_1\cdot K_2}{N}\right)\left(1 - \frac{K(0)}{Nh}\right)\right\}^{-2}\cdot\frac{1}{N}\sum_{i=1}^{N}\{\widehat{\pi}_K(T_i,\boldsymbol{X}_i)Y_i - \widehat{m}_h(T_i)\}^2.$$

Alternatively, we can apply the leave-one-out approach and define

$$CV(K_1,K_2,h) = \left\{\left(1 - \frac{K_1\cdot K_2}{N}\right)\right\}^{-2}\frac{1}{N}\sum_{i=1}^{N}\left\{\widehat{\pi}_K(T_i,\boldsymbol{X}_i)Y_i - \widehat{m}_h^{(-i)}(T_i)\right\}^2,$$

where

$$\widehat{m}_h^{(-i)}(T_i) = \frac{\sum_{j\neq i} Y_j K\left(\frac{T_j - T_i}{h}\right)\widehat{\pi}_K(T_j, \boldsymbol{X}_j)}{\sum_{j\neq i} K\left(\frac{T_j - T_i}{h}\right)\widehat{\pi}_K(T_j, \boldsymbol{X}_j)}.$$

Then, we choose K_1, K_2, and h to minimize $CV(K_1, K_2, h)$.

6. Testing distributional effects

Detecting evidence of a treatment effect is one of the goals of the program evaluation literature. The existing literature is mostly concerned with comparing some moments of counterfactual distributions (e.g., means and quantiles). However, for the best detection of a treatment effect, one should compare the entire distributions. This section considers three hypotheses.

6.1. Distributional differences between two treatments

The first hypothesis compares the counterfactual distributions of two treatment levels. For fixed t_0 and t_1, we consider the following null and alternative hypotheses:

$$\begin{aligned} &H_0 : F_{t_1}(y) = F_{t_0}(y) \text{ for all } y \in \mathcal{Y}; \\ &H_1 : F_{t_1}(y) \neq F_{t_0}(y) \text{ for some } y \in \mathcal{Y}. \end{aligned} \tag{6.1}$$

We present three classes of tests for 6.1.

6.1.1. Confidence bands

The first class of tests is to test the difference $\Delta_{t_1,t_0}(y) = F_{t_1,h}(y) - F_{t_0,h}(y)$. Letting $\widehat{\Delta}_{t_1,t_0}(y) = \widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)$, Theorem 2 shows that $\sqrt{Nh}\left(\widehat{\Delta}_{t_1,t_0}(y) - \Delta_{t_1,t_0}(y)\right)$ converges in distribution to a normal distribution point-wise for each $y \in \mathcal{Y}$, and $\sqrt{Nh}\left(\widehat{\Delta}_{t_1,t_0}(y) - \Delta_{t_1,t_0}(y)\right)$ converges in distribution to a Gaussian process uniformly over $y \in \mathcal{Y}$. Based on these results, we construct point-wise confidence bands and uniform confidence bands. With $\widehat{\psi}_{t,h}(Y_i, T_i, \boldsymbol{X}_i; y)$ defined as in 4.3 , we estimate the variance of $\widehat{\Delta}_{t_1,t_0}(y)$ by

$$\widehat{\Sigma}_{t_1,t_0}(y) = \frac{h}{N}\sum_{i=1}^{N}\left\{\widehat{\psi}_{t_1,h}(Y_i, T_i, \boldsymbol{X}_i; y) - \widehat{\psi}_{t_0,h}(Y_i, T_i, \boldsymbol{X}_i; y)\right\}^2.$$

For some small $\alpha > 0$, let $z_{1-\alpha/2}$ denote the $(1-\alpha/2)$ quantile of the standard normal distribution. The point-wise confidence bands are given by

$$CI_{1-\alpha}(y) = \left[\widehat{\Delta}_{t_1,t_0}(y) - z_{1-\alpha/2}\cdot\widehat{\Sigma}^{1/2}_{t_1,t_0}(y)/\sqrt{Nh}, \widehat{\Delta}_{t_1,t_0}(y) + z_{1-\alpha/2}\cdot\widehat{\Sigma}^{1/2}_{t_1,t_0}(y)/\sqrt{Nh}\right].$$

We reject the null hypothesis if 0 is outside the point-wise confidence bands.

The point-wise confidence bands have the correct coverage probability for each y but not uniformly over all y. To construct the uniform confidence bands, we must find the critical value that yields the correct coverage probability for all y. Let $\widetilde{\Sigma}_{t_1,t_0}(y)$ denote a uniform consistent estimator of the variance of $\widehat{\Delta}_{t_1,t_0}(y)$. Then $\sqrt{Nh}\widetilde{\Sigma}^{-1/2}_{t_1,t_0}(y)\left(\widehat{\Delta}_{t_1,t_0}(y) - \Delta_{t_1,t_0}(y)\right)$ converges in distribution to the standard Gaussian process uniformly over $y \in \mathcal{Y}$. The Kolmogorov–Smirnov (KS) maximal t-statistic is

$$t_{KS} = \sup_{y\in\mathcal{Y}}\sqrt{Nh}\cdot\widetilde{\Sigma}^{-1/2}_{t_1,t_0}(y)\cdot|\widehat{\Delta}_{t_1,t_0}(y) - \Delta_{t_1,t_0}(y)|.$$

Let $\widehat{t}_{1-\alpha}$ denote the critical value satisfying $P\left(t_{KS} > \widehat{t}_{1-\alpha}\right) = \alpha$. The uniform confidence bands are given by

$$\left[\widehat{\Delta}_{t_1,t_0}(y) - \widehat{t}_{1-\alpha}\cdot\widetilde{\Sigma}^{1/2}_{t_1,t_0}(y)/\sqrt{Nh}, \widehat{\Delta}_{t_1,t_0}(y) + \widehat{t}_{1-\alpha}\cdot\widetilde{\Sigma}^{1/2}_{t_1,t_0}(y)/\sqrt{Nh}\right].$$

We reject the null hypothesis if 0 is outside the uniform confidence bands.

The distribution of t_{KS} is unknown. To compute the critical value, we apply the *exchangeable bootstrap* (Praestgaard and Wellner, 1993; Van Der Vaart and Wellner, 1996). The idea is to bootstrap the distribution of t_{KS}. Let $(w_1, \ldots, w_N)$ denote an independent sample drawn from a distribution satisfying Condition EB in Section 5 of Chernozhukov et al. (2013) (e.g., an exponential distribution). Compute

$$\widehat{F}^*_{t,h}(y) = \frac{\sum_{i=1}^N w_i\cdot\widehat{\pi}_K(T_i, \boldsymbol{X}_i)I(Y_i \leq y)K\left(\frac{T_i - t}{h}\right)}{\sum_{i=1}^N w_i\cdot\widehat{\pi}_K(T_i, \boldsymbol{X}_i)K\left(\frac{T_i - t}{h}\right)},$$

$$\widehat{\Delta}^*_{t_1,t_0}(y) = \widehat{F}^*_{t_1,h}(y) - \widehat{F}^*_{t_0,h}(y), \text{ and } \widehat{Z}^*_{t_1,t_0}(y) = \sqrt{Nh}\left\{\widehat{\Delta}^*_{t_1,t_0}(y) - \widehat{\Delta}_{t_1,t_0}(y)\right\}.$$

The bootstrap algorithm involves repeating the calculation above multiple times:

1. Draw B samples $(w_{1b}, \ldots, w_{Nb})$, $b = 1, 2, \ldots, B$.
2. Compute $\{\widehat{\Delta}^*_{t_1,t_0;b}(y) : 1 \le b \le B\}$ and $\{\widehat{Z}^*_{t_1,t_0;b}(y) : 1 \le b \le B\}$.
3. To bootstrap the variance, let $q_p(y)$ denote the pth quantile of $\{\widehat{\Delta}^*_{t_1,t_0;b}(y) : 1 \le b \le B\}$ and let z_p denote the pth quantile of $N(0, 1)$. The bootstrap variance is
$$\widetilde{\Sigma}^{1/2}_{t_1,t_0}(y) = \frac{q_{0.75}(y) - q_{0.25}(y)}{z_{0.75} - z_{0.25}}.$$
4. Compute the maximal t-statistic
$$\widehat{t}_{KS;b} = \sup_{y\in\mathcal{Y}} \widetilde{\Sigma}^{-1/2}_{t_1,t_0}(y) \cdot |\widehat{Z}^*_{t_1,t_0;b}(y)|, \text{ for } 1 \le b \le B.$$
5. $\widehat{t}_{1-\alpha}$ is the $(1-\alpha)$ sample quantile of $\{\widehat{t}_{KS;b} : 1 \le b \le B\}$.

6.1.2. KS and Cramér–von Mises (CvM) test

The second class of tests is distance tests. We consider two distance measures: the sup and the L_2 distance. The sup distance gives the KS test statistic
$$\widehat{I}_{KS} = \sup_{y\in\mathcal{Y}} \left|\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)\right|$$
and the L_2 distance gives the CvM test statistic
$$\widehat{I}_{CvM} = \int_{\mathcal{Y}} \left\{\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)\right\}^2 d\widehat{F}_Y(y) = \frac{1}{N}\sum_{j=1}^{N} \left\{\widehat{F}_{t_1,h}(Y_j) - \widehat{F}_{t_0,h}(Y_j)\right\}^2.$$

Theorem 7. *Assume Assumptions 1–8 hold and $Nh^5 \to 0$. We show that, under H_0,*
$$\sqrt{Nh}\cdot\widehat{I}_{KS} \xrightarrow{d} \sup_{y\in\mathcal{Y}} |G_{t_1,t_0}(y)| \text{ and } Nh\cdot\widehat{I}_{CvM} \xrightarrow{d} \int_{y\in\mathcal{Y}} |G_{t_1,t_0}(y)|^2 dF_Y(y),$$
where $G_{t_1,t_0}(y)$ is a Gaussian process indexed by $y \in \mathcal{Y}$ with covariance function
$$\begin{aligned} V_G(y_1,y_2) =& \frac{\kappa_{02}}{f_T(t_1)} \cdot \mathbb{E}\Big[\pi_0(T_i,\boldsymbol{X}_i)^2 \left\{I(Y_i \le y_1) - F_{Y|T,X}(y_1|T_i,\boldsymbol{X}_i)\right\} \\ &\times \left\{I(Y_i \le y_2) - F_{Y|T,X}(y_2|T_i,\boldsymbol{X}_i)\right\} |T_i = t_1\Big] \\ &+ \frac{\kappa_{02}}{f_T(t_0)} \cdot \mathbb{E}\Big[\pi_0(T_i,\boldsymbol{X}_i)^2 \left\{I(Y_i \le y_1) - F_{Y|T,X}(y_1|T_i,\boldsymbol{X}_i)\right\} \\ &\times \left\{I(Y_i \le y_2) - F_{Y|T,X}(y_2|T_i,\boldsymbol{X}_i)\right\} |T_i = t_0\Big]. \end{aligned}$$

The limiting distributions of $\widehat{I}_{KS}$ and $\widehat{I}_{CvM}$ do not have an analytical form. The critical values can be computed by applying the bootstrap procedure. For more details, see Li et al. (2003).

6.1.3. Mann–Whitney test

The third class of tests is based on the Mann–Whitney indicator $\theta_{t_1,t_0} = \int_{\mathcal{Y}} F_{t_1}(y) dF_{t_0}(y)$, where $\theta_{t_1,t_0} = 1/2$ under the null. $\theta_{t_1,t_0} > 1/2$ if $F_{t_1}(y) > F_{t_0}(y)$ for all y and $\theta_{t_1,t_0} < 1/2$ if $F_{t_1}(y) < F_{t_0}(y)$ for all y. One advantage of this indicator is that it reduces the comparison of two distributions to a single parameter. Another advantage is that the value of θ_{t_1,t_0} may reveal the stochastic dominance of potential outcomes. By plotting θ_{t_1,t_0} against (t_1, t_0) we could infer the range of treatments and associated stochastically dominant outcomes. The third advantage is that the asymptotic distribution of the Mann–Whitney test statistic
$$\widehat{\theta}_{t_1,t_0,h} = \int_{\mathcal{Y}} \widehat{F}_{t_1,h}(y) d\widehat{F}_{t_0,h}(y)$$
is normal.

Theorem 8. *Assume that Assumptions 1–8 hold and $Nh^5 \to 0$. We have*
$$\sqrt{Nh}\{\widehat{\theta}_{t_1,t_0,h} - \theta_{t_1,t_0}\} \xrightarrow{d} \mathcal{N}(0, V_{t_1,t_0}),$$

where

$$V_{t_1,t_0} = \frac{\kappa_{02}}{f_T(t_1)} \cdot \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)^2 \left\{\int_{\mathcal{Y}} \{I(Y_i \le y) - F_{Y|T,X}(y|T_i, \boldsymbol{X}_i)\} dF_{t_0}(y)\right\}^2 \middle| T = t_1\right] + \frac{\kappa_{02}}{f_T(t_0)} \cdot \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i)^2 \left\{\int_{\mathcal{Y}} \{I(Y_i \le y) - F_{Y|T,X}(y|T_i, \boldsymbol{X}_i)\} dF_{t_1}(y)\right\}^2 \middle| T = t_0\right].$$

The asymptotic variance in Theorem 8 is estimated as follows:

$$\widehat{V}_{t_1,t_0} = \frac{\kappa_{02}}{\widehat{f}_T(t_1)} \cdot \frac{\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i)^2 \left\{\int_{\mathcal{Y}} \{I(Y_i \le y) - \widehat{F}_{Y|T,X}(y|T_i, \boldsymbol{X}_i)\} d\widehat{F}_{t_0,h}(y)\right\}^2 K\left(\frac{T_i - t_1}{h_T}\right)}{\sum_{i=1}^N K\left(\frac{T_i - t_1}{h_T}\right)} + \frac{\kappa_{02}}{\widehat{f}_T(t_0)} \cdot \frac{\sum_{i=1}^N \widehat{\pi}_K(T_i, \boldsymbol{X}_i)^2 \left\{\int_{\mathcal{Y}} \{I(Y_i \le y) - \widehat{F}_{Y|T,X}(y|T_i, \boldsymbol{X}_i)\} d\widehat{F}_{t_1,h}(y)\right\}^2 K\left(\frac{T_i - t_0}{h_T}\right)}{\sum_{i=1}^N K\left(\frac{T_i - t_0}{h_T}\right)}.$$

We reject the null hypothesis H_0 if

$$\sqrt{\frac{Nh}{\widehat{V}_{t_1,t_0}}} \left|\widehat{\theta}_{t_1,t_0,h} - 1/2\right| > z_{1-\alpha/2}.$$

6.2. Stochastic dominance test

The Mann–Whitney test is effective for detecting a distributional difference. However, we cannot be certain of stochastic dominance if $\theta_{t_1,t_0} = 1/2$ is rejected. This is because $\theta_{t_1,t_0} > 1/2$ does not necessarily mean that $F_{t_1}(y)$ dominates $F_{t_0}(y)$. To test stochastic dominance, we consider the following null and alternative hypotheses, which rank the distribution functions:

$$H_0 : F_{t_1}(y) \le F_{t_0}(y) \text{ for all } y \in \mathcal{Y};$$
$$H_1 : F_{t_1}(y) > F_{t_0}(y) \text{ for some } y \in \mathcal{Y}.$$

We apply the stochastic dominance test, which was first introduced in econometrics by McFadden (1989) and further studied by Anderson (1996), Davidson and Duclos (2000), Barrett and Donald (2003), Linton et al. (2005), Linton et al. (2010), and Donald and Hsu (2014). The KS statistic in this case is

$$\widehat{S}_h = \sqrt{Nh} \sup_{y \in \mathcal{Y}} \left(\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)\right).$$

Clearly, this statistic should not take a large positive value if the null hypothesis is true. Thus, for a small positive value c, the decision rule is

$$\text{Reject } H_0 \text{ if } \widehat{S}_h > c.$$

To find the critical value c, we need the asymptotic distribution of $\widehat{S}_h$, which, unfortunately, depends on the unknown true distributions under the null hypothesis. One proposed solution is the least favorable configuration (LFC). The LFC finds an upper bound that is equal to the KS statistic when the two distributions are the same. The asymptotic distribution of the upper bound is known so that we can use it to find the critical value. Applying this idea, we find that under the null hypothesis,

$$\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y) = \left\{(\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)) - (F_{t_1}(y) - F_{t_0}(y))\right\} + (F_{t_1}(y) - F_{t_0}(y)) \le (\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)) - (F_{t_1}(y) - F_{t_0}(y)).$$

Hence,

$$\widehat{S}_h = \sqrt{Nh} \sup_{y \in \mathcal{Y}} \left\{\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)\right\} \le \sqrt{Nh} \sup_{y \in \mathcal{Y}} \left\{(\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)) - (F_{t_1}(y) - F_{t_0}(y))\right\}.$$

Applying Theorem 2, the upper bound $\sqrt{Nh} \sup_{y \in \mathcal{Y}} \left\{(\widehat{F}_{t_1,h}(y) - \widehat{F}_{t_0,h}(y)) - (F_{t_1}(y) - F_{t_0}(y))\right\}$ is asymptotically equal to $\sup_{y \in \mathcal{Y}} \left(\Psi_{t_1}(y) - \Psi_{t_0}(y)\right)$.

The distribution of $\sup_{y\in\mathcal{Y}}\left(\Psi_{t_1}(y)-\Psi_{t_0}(y)\right)$ is complex. To approximate it, we apply the wild bootstrap approach of Donald and Hsu (2014). By Theorem 2, for any fixed $t\in\mathcal{T}$, we have

$$\sup_{y\in\mathcal{Y}}\left|\sqrt{Nh}\left\{\widehat{F}_{t,h}(y)-F_t(y)\right\}-\sqrt{\frac{h}{N}}\cdot\sum_{i=1}^{N}\psi_{t,h}(Y_i,T_i,\boldsymbol{X}_i;y)\right|=o_P(1).$$

Let $\{w_i\}_{i=1}^N$ be i.i.d. random variables with mean zero and variance one, independent of the sample $\{T_j,\boldsymbol{X}_j,Y_j\}_{j=1}^N$. We bootstrap $\sup_{y\in\mathcal{Y}}\left(\Psi_{t_1}(y)-\Psi_{t_0}(y)\right)$ by using

$$\overline{S}_w=\sup_{y\in\mathcal{Y}}\left(\Psi^w_{t_1,h}(y)-\Psi^w_{t_0,h}(y)\right),\ \text{where}\ \Psi^w_{t,h}(y)=\sqrt{\frac{h}{N}}\cdot\sum_{j=1}^{N}w_j\widehat{\psi}_{t,h}(Y_j,T_j,\boldsymbol{X}_j;y).$$

Employing the same arguments as those in Donald and Hsu (2014), we show that $\Psi^w_{t,h}(\cdot)\Rightarrow\Psi_t(\cdot)$, conditional on the sample $\{T_i,\boldsymbol{X}_i,Y_i:i=1,\ldots,N\}$ with probability approaching to one. For a significance level α, the simulated critical value $\widehat{c}$ is the $(1-\alpha)$-th quantile of $\overline{S}_u$:

$$\widehat{c}=\sup\{q:P_w\left(\overline{S}_w\le q\right)\le 1-\alpha\}.$$

Evidently, the critical value determined by the upper bound is larger than necessary and so the KS test is conservative in terms of Type I error.

Theorem 9. *Assume that Assumptions 1–8 hold. Consider the decision rule "reject H_0 when $\widehat{S}_h>\widehat{c}$". We show that*

1. *If H_0 is true, $\limsup P(\text{reject } H_0)=\limsup P(\widehat{S}_h>\widehat{c})\le\alpha_0$, where the equality holds when $F_{t_0}(y)=F_{t_1}(y)$ for all $y\in\mathcal{Y}$.*
2. *If H_0 is false, $\lim_N P(\text{reject } H_0)=1$.*

6.3. Quantile treatment effect

While the two types of hypotheses considered above compare the distribution functions, the third type of hypothesis compares the quantiles of different treatments. Specifically, for some fixed t_0, the null and alternative hypotheses are:

$$H_0: q_t(\tau)=q_{t_0}(\tau)\ \text{for all}\ t\in\mathcal{T};$$
$$H_1: q_t(\tau)\neq q_{t_0}(\tau)\ \text{for some}\ t\in\mathcal{T}.$$

Assume that $F_t(y)$ is strictly monotone in y for all $t\in\mathcal{T}$. Then, $q_t(\tau)=q_{t_0}(\tau)$ for all $t\in\mathcal{T}$ is equivalent to $F_t(q_{t_0}(\tau))=F_t(q_t(\tau))=\tau$ for all $t\in\mathcal{T}$. The test statistic is

$$\widehat{I}_{t_0}=\int_{\mathcal{T}}\left\{\widehat{F}_{t,h}(\widehat{q}_{t_0,h}(\tau))-\tau\right\}^2w(t)dt,$$

where the weighting function $w(t)$ is given by

$$w(t)=\left\{\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)K_h\left(T_i-t\right)\right\}^2.$$

After some manipulation, we have

$$\begin{aligned}\widehat{I}_{t_0}&=\int_{\mathcal{T}}\left\{\frac{\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)I(Y_i\le\widehat{q}_{t_0,h}(\tau))K_h\left(T_i-t\right)}{\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)K_h\left(T_i-t\right)}-\tau\right\}^2w(t)dt\\&=\int_{\mathcal{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)\left\{I(Y_i\le\widehat{q}_{t_0,h}(\tau))-\tau\right\}K_h\left(T_i-t\right)\right\}^2dt\\&=\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\widehat{\pi}_K(T_i,\boldsymbol{X}_i)\left\{I(Y_i\le\widehat{q}_{t_0,h}(\tau))-\tau\right\}\widehat{\pi}_K(T_j,\boldsymbol{X}_j)\left\{I(Y_j\le\widehat{q}_{t_0,h}(\tau))-\tau\right\}\overline{K}_h(T_i,T_j),\end{aligned}$$

where $\overline{K}_h(T_i,T_j)=h^{-1}\cdot\overline{K}(\{T_i-T_j\}/h)$ and $\overline{K}(v)=\int K(u)K(v-u)du$ is the convolution kernel derived from $K(\cdot)$.

Theorem 10. *Assume that $F_t(y)$ is strictly monotone in y, Assumptions 1–8 hold, $\zeta^2(K)\sqrt{K^2/N}\to 0$, and $Nh^5\to 0$. Under the null hypothesis, we have*

$$\frac{Nh\left\{\widehat{I}_{t_0}-\widehat{b}_{t_0}\right\}}{\widehat{\sigma}_{t_0}}\xrightarrow{d}\chi^2_1,$$

where χ_1^2 denotes the chi-square distribution with one degree of freedom, and

$$\widehat{b}_{t_0} = \frac{\overline{K}(0)}{N^2 h} \sum_{i=1}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i)^2 \left\{ I(Y_i \le \widehat{q}_{t_0,h}(\tau)) - \tau \right\}^2,$$

$$\widehat{\sigma}_{t_0} = \left\{ \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{i=1, i\neq j}^{N} \overline{K}_h(T_i, T_j) \times \widehat{\pi}_K(T_i, \boldsymbol{X}_i) \widehat{f}_{Y(t)}(\widehat{q}_{t_0,h}(\tau)) \Big|_{t=T_i} \right.$$
$$\left. \times \widehat{\pi}_K(T_j, \boldsymbol{X}_j) \widehat{f}_{Y(t)}(\widehat{q}_{t_0,h}(\tau)) \Big|_{t=T_j} \right\} \times \widehat{\Gamma}_{t_0}(\tau, \tau),$$

$$\widehat{f}_{Y(t)}(y) = \frac{\sum_{i=1}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) \widehat{f}_{Y|T,X}(y|T_i, \boldsymbol{X}_i) K\left(\frac{T_i - t}{h}\right)}{\sum_{i=1}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) K\left(\frac{T_i - t}{h}\right)},$$

where $\widehat{f}_{Y|T,X}$ is a kernel estimator for the conditional density $f_{Y|T,X}$.

We use a heuristic argument to derive the result above. Under the null hypothesis, $\tau = F_t(q_{t_0}(\tau)) = F_t(\widehat{q}_{t_0,h}(\tau)) - f_{Y(t)}(\widetilde{q}_{t_0,h}(\tau)) \cdot \{\widehat{q}_{t_0,h}(\tau) - q_{t_0}(\tau)\}$ for all $t \in \mathcal{T}$, where $\widetilde{q}_{t_0,h}(\tau)$ is between $\widehat{q}_{t_0,h}(\tau)$ and $q_{t_0,h}(\tau)$. We can write

$$\widehat{I}_{t_0} - \widehat{b}_{t_0} = \frac{1}{N^2} \sum_{j\neq i} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) \left\{ I(Y_i \le \widehat{q}_{t_0,h}(\tau)) - F_t(\widehat{q}_{t_0,h}(\tau))|_{t=T_i} \right\} \tag{6.2}$$
$$\times \widehat{\pi}_K(T_j, \boldsymbol{X}_j) \left\{ I(Y_j \le \widehat{q}_{t_0,h}(\tau)) - F_t(\widehat{q}_{t_0,h}(\tau))|_{t=T_j} \right\} \overline{K}_h(T_i, T_j)$$
$$+ \frac{2}{N^2} \sum_{j\neq i} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) \left\{ I(Y_i \le \widehat{q}_{t_0,h}(\tau)) - F_t(\widehat{q}_{t_0,h}(\tau))|_{t=T_i} \right\} \tag{6.3}$$
$$\times \widehat{\pi}_K(T_j, \boldsymbol{X}_j) f_{Y(t)}(\widetilde{q}_{t_0,h}(\tau)) \Big|_{t=T_j} \overline{K}_h(T_i, T_j) \left\{ \widehat{q}_{t_0,h}(\tau) - q_{t_0}(\tau) \right\}$$
$$+ \frac{1}{N^2} \sum_{j\neq i} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) f_{Y(t)}(\widetilde{q}_{t_0,h}(\tau)) \Big|_{t=T_i} \tag{6.4}$$
$$\times \widehat{\pi}_K(T_j, \boldsymbol{X}_j) f_{Y(t)}(\widetilde{q}_{t_0,h}(\tau)) \Big|_{t=T_j} \overline{K}_h(T_i, T_j) \left\{ \widehat{q}_{t_0,h}(\tau) - q_{t_0}(\tau) \right\}^2.$$

Employing arguments similar to those used to prove Lemma 3.3 (a) and 3.3(b) in Zheng (1996), we show that the right hand side of (6.2) and the term (6.3) are $O_P(1/Nh^{1/2})$. By Corollary 5, the term (6.4) multiplied by $Nh/\widehat{\sigma}_{t_0}$ is asymptotically χ_1^2 distributed.

For the dose response function $m(t)$, we consider the following null and alternative hypotheses:

$$H_0 : m(t) = m(t_0) \text{ for all } t \in \mathcal{T};$$
$$H_1 : m(t) \neq m(t_0) \text{ for some } t \in \mathcal{T}.$$

The test statistic is

$$\widehat{I}_m = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) \{Y_i - \widehat{m}_h(t_0)\} \widehat{\pi}_K(T_j, \boldsymbol{X}_j) \left\{ Y_j - \widehat{m}_h(t_0) \right\} \overline{K}_h(T_i, T_j).$$

Corollary 11. *Assume that Assumptions 1–8 hold, $\zeta^2(K)\sqrt{K^2/N} \to 0$, and $Nh^5 \to 0$. Under the null hypothesis, we have*

$$\frac{Nh\left\{\widehat{I}_m - \widehat{b}_m\right\}}{\widehat{\sigma}_m} \xrightarrow{d} \chi_1^2,$$

where

$$\widehat{b}_m = \frac{\overline{K}(0)}{N^2 h} \sum_{i=1}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i)^2 \{Y_i - \widehat{m}_h(t_0)\}^2,$$

$$\widehat{\sigma}_m = \left\{ \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{i=1, i\neq j}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) \widehat{\pi}_K(T_j, \boldsymbol{X}_j) \overline{K}_h(T_i, T_j) \right\} \cdot \widehat{V}_{t_0}.$$

7. Monte Carlo simulation

To assess the finite sample performance of the proposed tests, we conduct a small-scale simulation study. We consider three scenarios. In all scenarios, we assume there is one covariate $X = 0.3 + 0.4U_x$ with U_x drawn from the uniform distribution over [0, 1]. We generate the continuous treatment variable as $T = X + \varepsilon$, with ε drawn from the standard normal distribution with mean zero and variance one. We specify the potential outcome $Y(t)$ as

- Scenario I: $Y(t) = I(t \le 0.5)\left(I(U_y \le X)\frac{U_y^2}{X} + I(U_y > X)U_y\right) + I(t > 0.5) \times (I(U_y \le 1 - X)\frac{U_y^2}{1-X} + I(U_y > 1 - X)U_y)$
- Scenario II: $Y(t) = -t + X + U_y$
- Scenario III: $Y(t) = |t - 0.25| + X + U_y$

where U_y is drawn from the uniform distribution over [0, 1], and U_x, U_y, and ε are independent. The observed outcome is $Y = Y(T)$. For simplicity, we set $t_0 = 0$ and let t_1 vary over the interval (0, 1]. Specifically, we set t_1 equal to the grid points $\{0.1, 0.2, \ldots, 0.9, 1\}$.

In Scenario I, by design, $F_{t_1}(\cdot) = F_{t_0}(\cdot)$, for all t_1. This scenario examines the size properties of the tests. In Scenario II, $F_{t_1}(\cdot) > F_{t_0}(\cdot)$, and $F_{t_1}(y) - F_{t_0}(y)$ is increasing in t_1. Therefore, the Mann–Whitney indicator $\theta_{t_1,0}$ exceeds 1/2 and is increasing in t_1. This scenario examines the power properties of the tests. In Scenario III, the relation between $F_{t_1}(\cdot)$ and $F_{t_0}(\cdot)$ varies with the treatment $t_1 \in (0, 1]$. For $t_1 \in (0, 0.25]$, $F_{t_1}(\cdot) > F_{t_0}(\cdot)$ and $F_{t_1}(y) - F_{t_0}(y)$ is increasing in t_1 so $\theta_{t_1,0}$ exceeds 1/2 and is increasing in t_1. For $t_1 \in (0.25, 0.5)$, $F_{t_1}(\cdot) > F_{t_0}(\cdot)$ and $F_{t_1}(y) - F_{t_0}(y)$ is decreasing in t_1 so $\theta_{t_1,0}$ exceeds 1/2 and is decreasing in t_1. For $t_1 = 0.5$, $F_{t_1}(y) = F_{t_0}(y)$ and $\theta_{t_1,0} = 1/2$. For $t_1 \in (0.5, 1]$, $F_{t_1}(\cdot) < F_{t_0}(\cdot)$ and $F_{t_1}(y) - F_{t_0}(y)$ is decreasing in t_1 so $\theta_{t_1,0}$ is less than 1/2 and is decreasing in t_1. This scenario examines the performance of the Mann–Whitney statistic.

In all scenarios, we set the sample size to $N = 100, 200$, and 400. We apply the data-driven approach to choose the smoothing parameters (K_1, K_2, h). For the stochastic dominance test, we use a standard normal random variable as perturbation to compute the critical value. We use 1000 perturbations. The significance level is $\alpha = 5\%$ and in all designs, the number of Monte Carlo simulations is 500.

We report the simulation results in Tables 1 and 2 and Figs. 1 and 2. Fig. 1 graphs the confidence bands for $F_{t_1} - F_{t_0}$. Due to space limitations, we only report the graphs for the case with $N = 400$ and $t_1 \in \{0.25, 0.75\}$. Table 1 reports the rejection rates of the Mann–Whitney test under Scenarios I, II, and III. Fig. 2 plots the average Mann–Whitney statistic $\widehat{\theta}_{t_1,0,h}$ against t_1 under Scenarios I, II, and III. Table 2 reports the rejection rates of the stochastic dominance test under the three scenarios.

Given these Figures and Tables, we make the following observations:

1. In Scenario I, we expect all three tests to accept the null hypothesis and the Mann–Whitney indicator to be a constant. Fig. 1 shows that zero is inside the confidence band and the difference test accepts the null hypothesis at the 95% confidence level. Table 1 reveals that, except in a few cases, the rejection rates of the Mann–Whitney test are close to 0.05, and are closer to 0.05 in larger samples, implying that the Mann–Whitney test accepts the null hypothesis at a level close to 0.05. The average Mann–Whitney statistic $\widehat{\theta}_{t_1,0,h}$ is roughly constant at 0.5 for all t_1, consistent with the model used. Table 2 reveals that the rejection rates of the KS test are higher than 0.05 in small samples. In larger samples, although the rejection rates are still higher than 0.05, they are closer to 0.05, implying that the KS test rejects a true null hypothesis at a higher frequency than what we would like.
2. In Scenario II, we expect all three tests to reject the null hypothesis and the Mann–Whitney indicator to be a monotone curve above 0.5. Fig. 1 shows that zero is outside the confidence bands and the difference test rejects the null hypothesis with the expected frequency. Both Tables 1 and 2 show that, except for lower levels of treatment, both the Mann–Whitney test and the KS test reject the null at a proportion close to one. Moreover, $\widehat{\theta}_{t_1,0,h}$ is always above 0.5 and continuously increases with t_1. These results are consistent with our expectations.
3. In Scenario III, we expect the difference test and the Mann–Whitney test to reject the null hypothesis for all t_1 and the KS test to reject the null hypothesis if $t_1 < 0.5$. We expect the Mann–Whitney indicator to decline and cross the horizontal line at 0.5. Fig. 1 shows that zero is outside the confidence band, implying the difference test rejects the null hypothesis. Table 1 reveals that the rejection rates of the Mann–Whitney test vary and seem to increase with the treatment level. Nevertheless, the rejection rates converge to one as the sample size increases. Fig. 2 shows that the values of $\widehat{\theta}_{t_1,0,h}$ are consistent with our expectations. Table 2 shows that the rejection rates of the KS test are closer to one when $t_1 < 0.5$ and closer to or smaller than 0.05 when $t_1 \ge 0.5$. These results are consistent with our predictions.

8. Empirical study

Economic theory suggests that labor supply decreases with non-labor income and the rate of decrease depends on the total non-labor income. The higher the non-labor income, the larger the decrease in labor supply. To test this theory, one would need an experiment that gives money to people randomly. The lottery is such an experiment as those participating

Table 1
Rejection rates of Mann–Whitney test.

N = 100										
t_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Scenario I	0.012	0.044	0.086	0.106	0.082	0.080	0.068	0.058	0.076	0.120
Scenario II	0.152	0.568	0.788	0.922	0.988	0.994	1.000	1.000	1.000	1.000
Scenario III	0.104	0.262	0.260	0.140	0.102	0.234	0.522	0.802	0.914	0.962
N = 200										
t_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Scenario I	0.000	0.030	0.070	0.082	0.078	0.066	0.072	0.078	0.084	0.082
Scenario II	0.206	0.738	0.942	0.988	0.998	1.000	1.000	1.000	1.000	1.000
Scenario III	0.158	0.528	0.490	0.220	0.116	0.302	0.650	0.940	0.984	1.000
N = 400										
t_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Scenario I	0.000	0.020	0.048	0.060	0.064	0.066	0.086	0.080	0.058	0.054
Scenario II	0.364	0.942	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000
Scenario III	0.276	0.768	0.710	0.304	0.098	0.422	0.908	1.000	1.000	1.000

Table 2
Rejection rates of KS test.

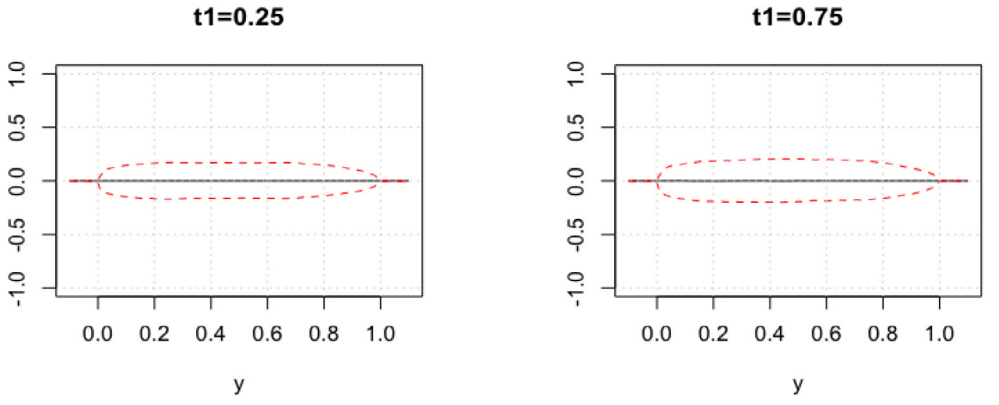
N = 100										
t_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Scenario I	0.124	0.130	0.104	0.100	0.118	0.106	0.100	0.068	0.080	0.088
Scenario II	0.656	0.712	0.834	0.930	0.988	0.996	1.000	1.000	1.000	1.000
Scenario III	0.586	0.522	0.422	0.210	0.078	0.026	0.004	0.000	0.000	0.000
N = 200										
t_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Scenario I	0.082	0.086	0.094	0.094	0.100	0.078	0.070	0.084	0.092	0.100
Scenario II	0.710	0.822	0.944	0.986	0.998	1.000	1.000	1.000	1.000	1.000
Scenario III	0.668	0.672	0.550	0.258	0.084	0.018	0.002	0.002	0.000	0.000
N = 400										
t_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Scenario I	0.074	0.084	0.086	0.078	0.086	0.088	0.086	0.078	0.062	0.054
Scenario II	0.892	0.968	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Scenario III	0.858	0.856	0.740	0.338	0.072	0.002	0.000	0.000	0.000	0.000

in it have an equal chance of winning regardless of their backgrounds. Moreover, the winning prizes vary across winners, providing variations to detect the impact on labor supply. It is well documented that participation in the lottery is not random. People with a low income are more likely to take part in it than those with a high income. Unfortunately, non-labor income is usually unobservable to researchers but it is highly correlated with personal background. This is a setting where the treatment (i.e., winning prize money) is continuous and the potential outcomes (i.e., post lottery earnings) are independent of the treatment given the proxies for non-labor income (i.e., personal characteristics). To study this setting, Imbens et al. (2001) obtained a dataset from a survey of Massachusetts lottery winners and labor earnings (a proxy for labor supply) from US Social Security records. We apply our procedure to this dataset to estimate and test the treatment effect of prize money on labor supply.

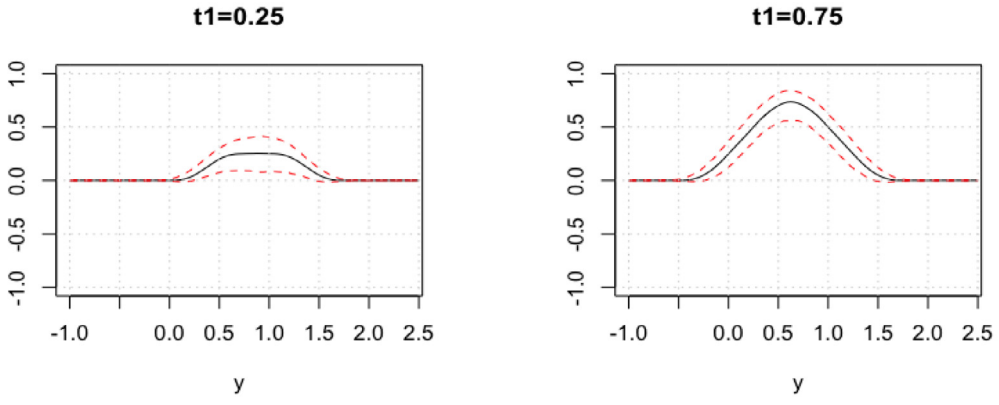
There are 237 lottery winners in the dataset. For each winner, we have data on the winning prize, age, gender, years of high school, years of college, winning year, number of tickets bought, work status after winning, and social security earnings $s \in \{1, 2, \ldots, 6\}$ years before winning the lottery and six years after. We denote the earnings six years after winning the lottery by Y, the logarithm of lottery prize by T, and the variables representing the other characteristics by $\boldsymbol{X}$. The value of Y is available only for 202 of the 237 winners. Fifty-two percent of the 202 winners exhibit no earnings, that is, $Y = 0$. Forty-seven percent of those with no earnings are male. Detailed descriptive statistics can be found in Imbens et al. (2001) and Hirano and Imbens (2004).

As a treatment variable, we take the logarithm of lottery prize instead of the lottery prize itself because the distribution of the latter is severely right-skewed (see Fig. 3(a)) while that of the logarithm of lottery prize is similar to a normal distribution (see Fig. 3(b)). Since the range of the treatment levels is between 0.76 and 6.1, we set the benchmark treatment level at $t_0 = 0.7$. To detect the evidence of the distributional treatment effect, we apply the proposed Mann–Whitney test for the null hypothesis $H_0 : F_{t_0}(\cdot) = F_{t_1}(\cdot)$ for $t_1 \in \{1, 2, \ldots, 6\}$. We use the leave-one-out cross-validation to choose smoothing parameters. Fig. 4 reports the estimated Mann–Whitney statistic $\widehat{\theta}_{t_1,t_0,h}$, the 95% point-wise confidence bands, and the corresponding P-values. From Fig. 4, we find that: (i) $\widehat{\theta}_{t_1,t_0,h} > 0.5$ for all $t_1 \in [1, 6]$; (ii) $\widehat{\theta}_{t_1,t_0,h}$ increases as t_1 increases; and (iii) the null hypothesis is rejected for $t_1 > 1.5$.

(a) Scenario I



(b) Scenario II



(c) Scenario III

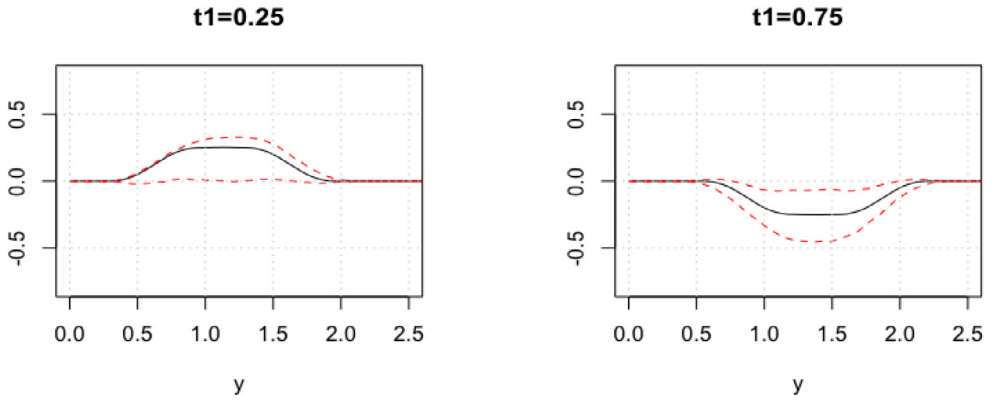


Fig. 1. 95% Confidence Bands for $F_{t_1} - F_{t_0}$ ($N = 400$).
The horizontal axis denotes the value of the outcome, and the vertical axis denotes the distributional effect. The black line represents the true curve of $F_{t_1} - F_{t_0}$. The red dotted lines represent the 95% confidence bands.

We also apply our method to estimate the average treatment effect $m(t) - m(t_0)$ and the quantile treatment effect $\{q_t(\tau) - q_{t_0}(\tau) : \tau \in [0, 1]\}$ for $t \in \{0.7, 0.8, 0.9, \ldots, 7.0\}$. We compute the 95% point-wise confidence bands through the

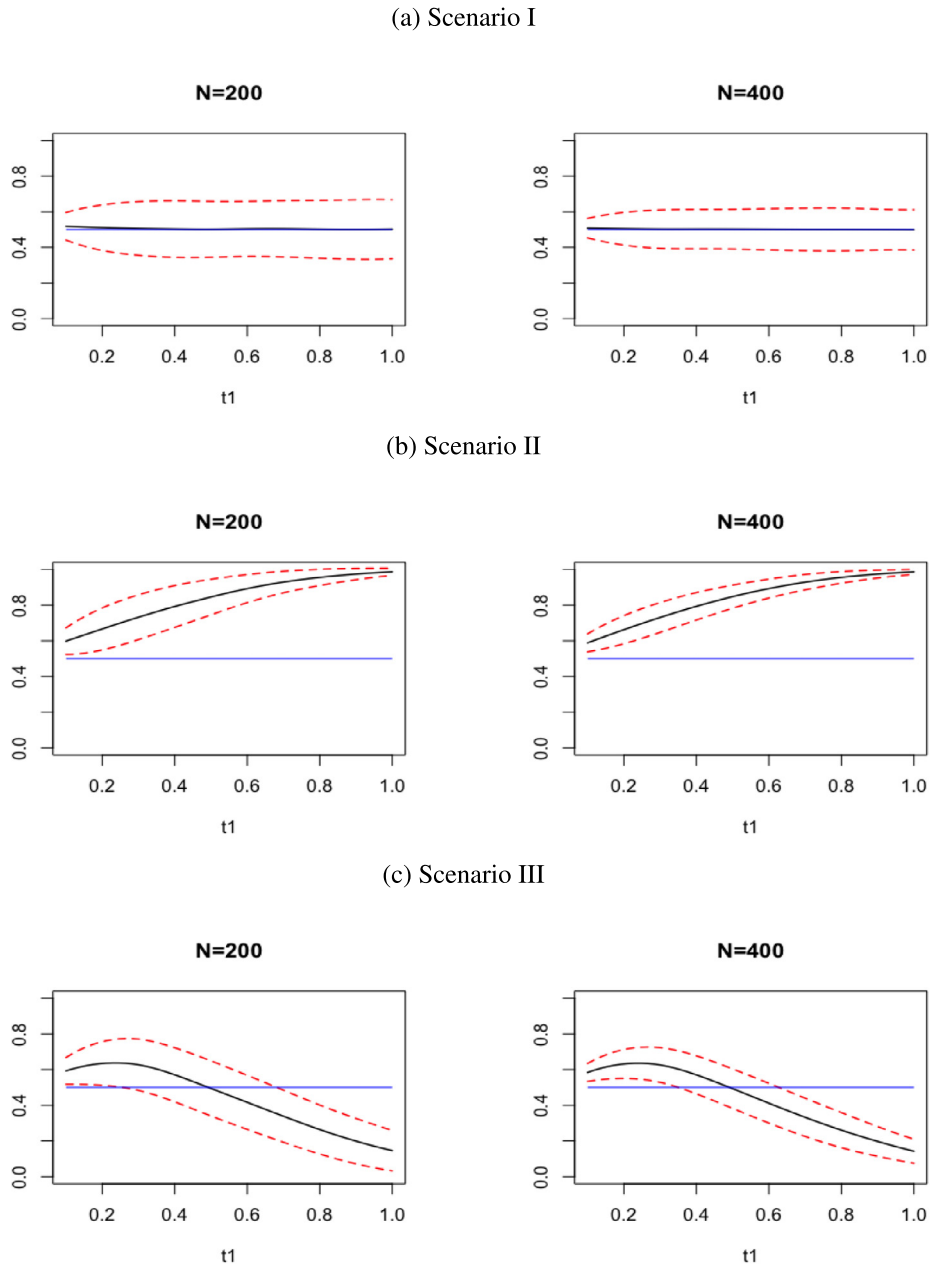


Fig. 2. The Mann–Whitney Statistic.
The horizontal axis denotes the treatment level and the vertical axis denotes the outcome of Mann–Whitney statistic. The black line is the average of the estimated Mann–Whitney statistic $\hat{\theta}_{t_1,0,h}$ based on 500 Monte Carlo, the red dotted lines are the confidence bands of $\hat{\theta}_{t_1,0,h}$, the blue line denotes $\theta = 0.5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

quantile-based non-parametric bootstrap method, with 500 samples generated by empirical bootstrap. For comparison, we also apply the naïve method, which sets $\pi_0(T, \boldsymbol{X}) \equiv 1$.

We report the estimated average treatment effect curve $\widehat{m}_h(t) - \widehat{m}_h(t_0)$ and its 95% point-wise confidence bands in Fig. 5(a). We notice that for low quantiles, that is, $\tau < 0.5$, $\widehat{q}_{t,h}(\tau) = 0$ for all t. Hence, we only plot the quantile treatment effect curve $\widehat{q}_{t,h}(\tau) - \widehat{q}_{t_0,h}(\tau)$ and the 95% confidence bands for $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ in Figs. 5(b)–5(f).

Based on figs. 5(a)–5(f), we observe the following: In Fig. 5(a), we see that the average treatment effect estimated by both the proposed and the naïve method declines as the lottery prize increases. The proposed estimate declines faster than the naïve estimate. We conclude that there is a negative lottery effect on labor earnings, which is consistent with the finding in Hirano and Imbens (2004). In figs. 5(b)–5(f), we see that in most cases, the estimated quantile

(a) Histogram for lottery prize (b) Histogram for log(lottery prize)

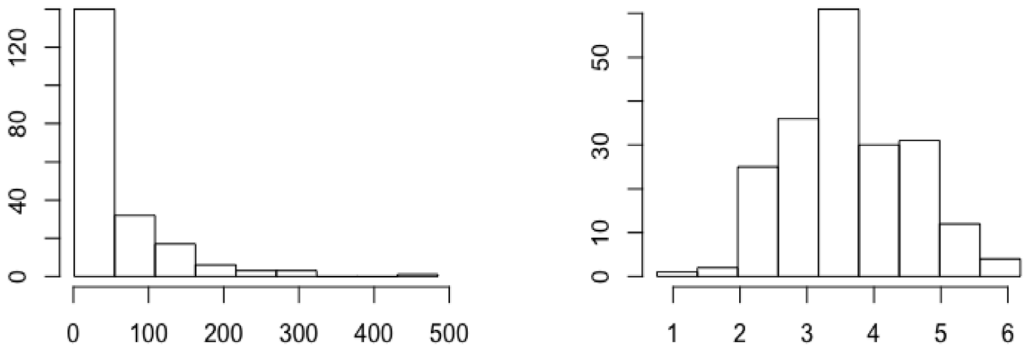
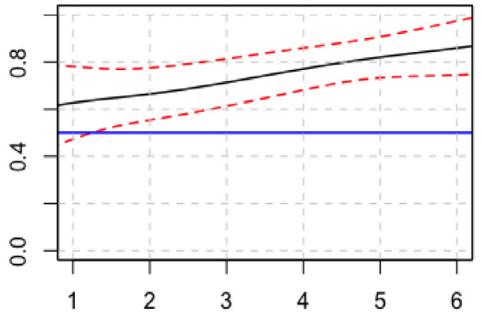


Fig. 3. Histograms for Lottery Prize and log(Lottery Prize).

(a) The Plot of Mann-Whitney Statistic (b) The P-Values of Mann-Whitney Test



t_1	1	2	3	4	5	6
P-value	0.106	0.004	0	0	0	0

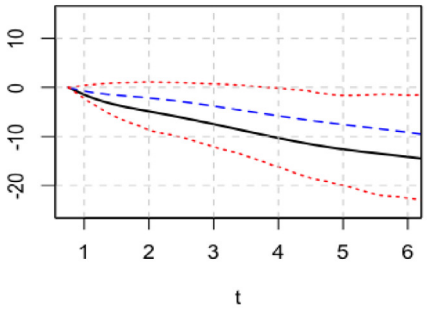
Fig. 4. The Result of Mann–Whitney Test.
Fig. 4(a) is the plot of $\widehat{\theta}_{t_1,t_0,h}$ for $t_1 \in [1, 6]$. The horizontal axis denotes the treatment level, and the vertical axis denotes the outcome of Mann–Whitney statistic. The black line is the estimated Mann–Whitney statistic $\widehat{\theta}_{t_1,t_0,h}$, the red dotted lines are the confidence bands of $\hat{\theta}_{t_1,t_0,h}$, and the blue line denotes $\theta = 0.5$. Fig. 4(b) reports the P-values of the Mann–Whitney tests for the hypothesizes $H_0 : F_{t_0}(\cdot) = F_{t_1}(\cdot)$ for $t_1 \in \{1, 2, \ldots, 6\}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

treatment effect declines as the lottery prize increases, and the declining rate decreases as the quantile level increases. For $\tau \in \{0.5, 0.6, 0.7\}$, the estimated quantile treatment effect curve $\widehat{q}_{t,h}(\tau) - \widehat{q}_{t_0,h}(\tau)$ decreases up to a certain treatment level and then remains constant after that. This implies that there is a prize threshold value that makes people with low earnings stop working. For $\tau = 0.8$, the estimated quantile treatment effect $\widehat{q}_{t,h}(\tau) - \widehat{q}_{t_0,h}(\tau)$ decreases slightly, for large enough t. For $\tau = 0.9$, $\widehat{q}_{t,h}(\tau) - \widehat{q}_{t_0,h}(\tau)$ exhibits no significant decrease across all treatment levels. The 95% point-wise confidence bands contain zero for all treatment levels. This implies that the lottery prize has no significant effect for people with high earnings.

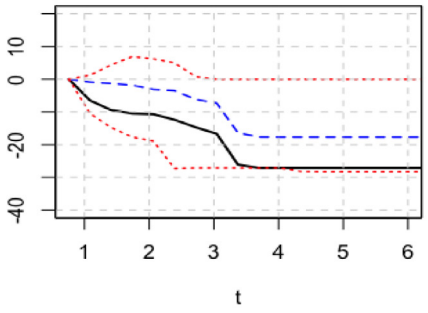
9. Conclusions

Counterfactual analysis is the primary focus of the program evaluation literature. This study extends the existing literature on discrete treatment (e.g., binary treatment) models to continuous treatment models. Specifically, we propose estimating the weighting function from a finite and expanding number of equations by maximizing a globally concave function and estimating the counterfactual distribution by plugging the estimated weighting function in a kernel regression. The estimated counterfactual distribution is then inverted to estimate quantiles. To test the distributional

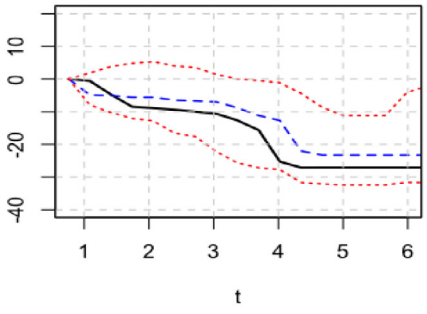
(a) $m(t) - m(t_0)$



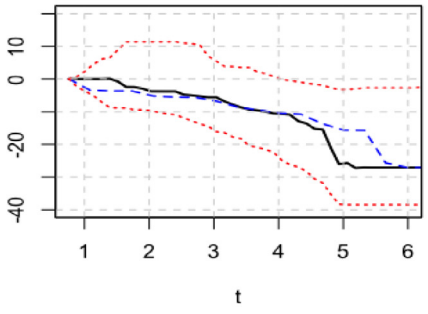
(b) $q_t(0.5) - q_{t_0}(0.5)$



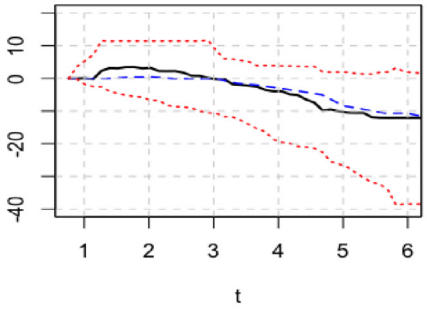
(c) $q_t(0.6) - q_{t_0}(0.6)$



(d) $q_t(0.7) - q_{t_0}(0.7)$



(e) $q_t(0.8) - q_{t_0}(0.8)$



(f) $q_t(0.9) - q_{t_0}(0.9)$

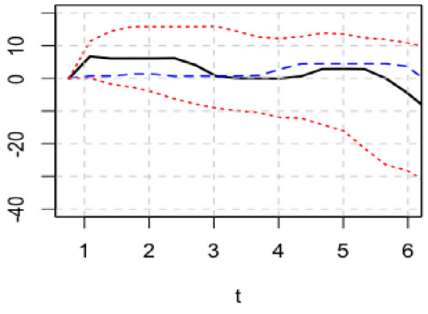


Fig. 5. Average Treatment Effect and Quantile Treatment Effect Curves.
(a): Estimated average treatment effect curve; (b)–(f): Estimated quantile treatment effect curves. The horizontal axis is treatment level, and the vertical axis is the treatment effect. The black solid line represents the proposed estimator. The red dashed lines are the 95% point-wise confidence bands obtained by empirical bootstrap with 500 repetitions. The blue dotted line is the naive estimator. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and quantile effects, we consider three types of null hypotheses: A. No distributional difference between two levels of treatment, B. Negative distributional difference between two levels of treatment, and C. No quantile difference for any levels of treatment. We propose three classes of tests for hypothesis A, a stochastic dominance test for hypothesis B, and

an L_2- distance test for hypothesis C. Each of these tests has its own merits and weaknesses. In applications, one may use them all to obtain more precise inference on the distributional effect.

Under some sufficient conditions, we show that the estimated counterfactual Distributions and quantiles, and all the test statistics are $\sqrt{Nh}$ consistent. Compared with estimation methods in the literature, our estimation and testing procedure have several advantages. First, our weighting function has an empirical likelihood interpretation but it is much easier to compute than the standard EL estimator. Second, our estimated weighting function improves the efficiency of the estimated counterfactual distributions and the quantile functions, and thereby the power of the tests. Third, we estimate the weighting function as a whole (as opposed to the ratio of two separately estimated densities), avoiding large extreme weights and unstable distribution estimates.

Acknowledgments

The authors would like to sincerely thank the Guest Editor—George Tauchen, the Associate Editor, and the two referees for their constructive suggestions and comments. The authors thank Lukang Huang for her assistance. Chunrong Ai acknowledges the financial support from the National Natural Science Foundation of China through the project 71873138. Oliver Linton acknowledges the financial support provided by Cambridge INET, United Kingdom. Zheng Zhang acknowledges the financial support from the National Natural Science Foundation of China through the project 12001535, and the fund for building world-class universities (disciplines) from the Renmin University of China. Zheng Zhang wishes to dedicate this paper to his beloved uncle, Mr Chuanbin Zhang, who lost his battle with cancer.

Appendix A. Proof of (2.1)

Using the tower property of conditional probability, we obtain:

$$\begin{aligned}
&F_t(y) = P(Y(t) \le y) \\
=&\int_{\mathcal{X}} P(Y(t) \le y | \boldsymbol{X} = x) f_X(x) dx \\
=&\int_{\mathcal{X}} P(Y(t) \le y | T = t, \boldsymbol{X} = x) f_X(x) dx \quad \text{(by Assumption 1)} \\
=&\int_{\mathcal{X}} P(Y \le y | T = t, \boldsymbol{X} = x) f_X(x) dx \\
=&\int_{\mathcal{X}} P(Y \le y | T = t, \boldsymbol{X} = x) \frac{f_X(x)}{f_{X|T}(x|t)} f_{X|T}(x|t) dx \\
=&\int_{\mathcal{X}} P(Y \le y | T = t, \boldsymbol{X} = x) \frac{f_T(t)}{f_{T|X}(t|x)} f_{X|T}(x|t) dx \\
=&\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) I(Y \le y) | T = t\right].
\end{aligned}$$

Appendix B. Proof of Theorem 1

The sufficiency part is obvious. We prove the necessity part. Let $u(T) = \exp(a \cdot T)$ and $v(\boldsymbol{X}) = \exp(b^\top \boldsymbol{X})$ be the test functions, where $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\{\pi(T, \boldsymbol{X}) - \pi_0(T, \boldsymbol{X})\} \exp\left\{a \cdot T + b^\top \boldsymbol{X}\right\}\right] + \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) \exp\left\{a \cdot T + b^\top \boldsymbol{X}\right\}\right] \\
=&\mathbb{E}\left[\exp(a \cdot T)\right] \cdot \mathbb{E}\left[\exp(b^\top \boldsymbol{X})\right],
\end{aligned}$$

which, in turn, implies $\mathbb{E}\left[\{\pi(T, \boldsymbol{X}) - \pi_0(T, \boldsymbol{X})\} \exp\left\{a \cdot T + b^\top \boldsymbol{X}\right\}\right] = 0$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$. By the uniqueness of the Laplace transform, we obtain $\pi(T, \boldsymbol{X}) = \pi_0(T, \boldsymbol{X})$ a.s.

Appendix C. Duality of primal problem (3.4)

We first introduce some notations.

- Let $m_K(T, \boldsymbol{X}) = \text{vec}\left(u_{K_1}(T) v_{K_2}^\top(\boldsymbol{X})\right)$ denote the K -dimensional column vector formed by the elements of the matrix $u_{K_1}(T) v_{K_2}^\top(\boldsymbol{X})$ and let $M_{K \times N} = (m_K(T_1, \boldsymbol{X}_1), \ldots, m_K(T_N, \boldsymbol{X}_N))$ be the $K \times N$ matrix.
- Let $u_{K_1,k}(T)$ (resp. $v_{K_2,k'}(\boldsymbol{X})$) denote the kth (resp. k'^{th}) component of $u_{K_1}(T)$ (resp. $v_{K_2}(\boldsymbol{X})$), and let

$$\bar{u}_{K_1,k} = \frac{1}{N} \sum_{j=1}^N u_{K_1,k}(T_j), \quad \bar{v}_{K_2,k'} = \frac{1}{N} \sum_{j=1}^N v_{K_2,k'}(\boldsymbol{X}_j).$$

Let b_K be the K-dimensional vector whose elements are formed by $\{\bar{u}_{K_1,k} \bar{v}_{K_2,k'}; k = 1, \ldots, K_1, k' = 1, \ldots, K_2\}$.

- Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ and $F(\boldsymbol{\pi}) = \sum_{i=1}^N \pi_i \log \pi_i$.

The primal optimization problem (3.4) can be written as

$$\begin{cases} \min_{\boldsymbol{\pi}} F(\boldsymbol{\pi}) \\ \text{subject to} \quad M_{K\times N} \cdot \boldsymbol{\pi} = N \cdot b_K \end{cases} \tag{C.1}$$

Based on Tseng and Bertsekas (1991), the conjugate convex function of $F(\cdot)$ is

$$F^*(\mathbf{z}) = \sup_{\boldsymbol{\pi}} \sum_{i=1}^N \{z_i\pi_i - \pi_i \log \pi_i\} = \sum_{i=1}^N \left\{z_i\pi_i^* - \pi_i^* \log \pi_i^*\right\} ,$$

where π_j^* satisfies the first order condition

$$z_j = \log \pi_j^* + 1 \Rightarrow \pi_j^* = e^{z_j - 1} = \rho'(z_i),$$

where $\rho\left(z_j\right) = -e^{-z_j-1}$. The conjugate convex function is simplified as follows:

$$F^*(\mathbf{z}) = \sum_{i=1}^N \left\{z_i e^{z_i-1} - e^{z_i-1}(z_i - 1)\right\} = \sum_{i=1}^N e^{z_i-1} = \sum_{i=1}^N -\rho(-z_i).$$

Based on Tseng and Bertsekas (1991), the dual problem (C.1) is

$$\begin{aligned} &\max_{\lambda\in\mathbb{R}^K} \left\{\lambda^\top (N \cdot b_K) - F^*\left(\lambda^\top M_{K\times N}\right)\right\} \\ &= \max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}} \sum_{j=1}^N \left\{\bar{u}_{K_1}^\top \Lambda \bar{v}_{K_2} + \rho\left(-u_{K_1}(T_j)^\top \Lambda v_{K_2}(\boldsymbol{X}_j)\right)\right\} \\ &= \max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}} \sum_{j=1}^N \left\{\rho\left(u_{K_1}(T_j)^\top \Lambda v_K(\boldsymbol{X}_j)\right) - \bar{u}_{K_1}^\top \Lambda \bar{v}_{K_2}\right\} \\ &= \max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}} \widehat{G}_{K_1\times K_2}(\Lambda), \end{aligned}$$

where

$$\widehat{G}_{K_1\times K_2}(\Lambda) = \frac{1}{N}\sum_{j=1}^N \rho(u_{K_1}(T_j)^\top \Lambda v_K(\boldsymbol{X}_j)) - \bar{u}_{K_1}^\top \Lambda \bar{v}_{K_2}.$$

Hence, the dual solution of (3.4) is

$$\widehat{\pi}_K(T_i, \boldsymbol{X}_i) = \rho'\left(u_{K_1}(T_i)^\top \hat{\Lambda}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i)\right) ,$$

where $\widehat{\Lambda}_{K_1\times K_2}$ is the maximizer of the strictly concave objective function $\widehat{G}_{K_1\times K_2}$.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.12.009.

References

Abadie, A., Cattaneo, M.D., 2018. Econometric methods for program evaluation. Annu. Rev. Econ. 10, 465–503.
Abbott, B., Gallipoli, G., Meghir, C., Violante, G.L., 2019. Education policy and intergenerational transfers in equilibrium. J. Polit. Econ. 127 (6), 2569–2624.
Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71 (6), 1795–1843.
Ai, C., Linton, O., Zhang, Z., 2020. Supplemental material for 'Estimation and inference for the counterfactual distribution and quantile functions in continuous treatment models', Discussion paper.
Alejo, J., Galvao, A.F., Montes-Rojas, G., 2018. Quantile continuous treatment effects. Econom. Stat. 8, 13–36.
Anderson, G., 1996. Nonparametric tests of stochastic dominance in income distributions. Econometrica 64 (5), 1183–1193.
Andrews, D.W.K., 1991. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Econometrica 59 (2), 307–345.
Barrett, G.F., Donald, S.G., 2003. Consistent tests for stochastic dominance. Econometrica 71 (1), 71–104.
Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.
Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. Econometrica 63 (4), 841–890.
Blundell, R., Costa Dias, M., Meghir, C., Shaw, J., 2016. Female labor supply, human capital, and welfare reform. Econometrica 84 (5), 1705–1753.

Burdett, K., Mortensen, D.T., 1998. Wage differentials, employer size, and unemployment. Internat. Econom. Rev. 39 (2), 257–273.
Cattaneo, M.D., 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. J. Econometrics 155 (2), 138–154.
Chan, K.C.G., Yam, S.C.P., Zhang, Z., 2016. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (3), 673–700.
Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. Handb. Econom. 6 (B), 5549–5632.
Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Statist. 36 (2), 808–843.
Chernozhukov, V., Fernández-Val, I., Melly, B., 2013. Inference on counterfactual distributions. Econometrica 81 (6), 2205–2268.
Chiappori, P.-A., Dias, M.C., Meghir, C., 2018. The marriage market, labor supply, and education choice. J. Polit. Econ. 126 (S1), S26–S72.
Coppejans, M., Gallant, A.R., 2002. Cross-validated SNP density estimates. J. Econometrics 110 (1), 27–65.
Davidson, R., Duclos, J.-Y., 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. Econometrica 68 (6), 1435–1464.
Donald, S.G., Hsu, Y.-C., 2014. Estimation and inference for distribution functions and quantile functions in treatment effect models. J. Econometrics 178 (3), 383–397.
Elbadawi, I., Gallant, A.R., Souza, G., 1983. An elasticity can be estimated consistently without a priori knowledge of functional form. Econometrica 51 (6), 1731–1751.
Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75 (1), 259–276.
Firpo, S., Fortin, N.M., Lemieux, T., 2009. Unconditional quantile regressions. Econometrica 77 (3), 953–973.
Florens, J.P., Heckman, J.J., Meghir, C., Vytlacil, E., 2008. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Econometrica 76 (5), 1191–1206.
Gallant, A.R., Hansen, L.P., Tauchen, G., 1990. Using conditional moments of asset payoffs to infer the volatility of intertemporal marginal rates of substitution. J. Econometrics 45 (1–2), 141–179.
Gallant, A.R., Nychka, D.W., 1987. Semi-nonparametric maximum likelihood estimation. Econometrica 55 (2), 363–390.
Gallant, A.R., Tauchen, G., 1989. Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. Econometrica 57 (5), 1091–1120.
Gallant, A.R., Tauchen, G., 1996. Which moments to match? Econometric Theory 12 (4), 657–681.
Galvao, A.F., Wang, L., 2015. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. J. Amer. Statist. Assoc. 110 (512), 1528–1542.
Goldberg, P.K., 1995. Product differentiation and oligopoly in international markets: The case of the US automobile industry. Econometrica 63 (4), 891–951.
Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66 (2), 315–331.
Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? J. Amer. Statist. Assoc. 83 (401), 86–95.
Heckman, J.J., Ichimura, H., Todd, P., 1998a. Matching as an econometric evaluation estimator. Rev. Econom. Stud. 65 (2), 261–294.
Heckman, J.J., Lochner, L., Taber, C., 1998b. Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Rev. Econ. Dyn. 1 (1), 1–58.
Heckman, J.J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73 (3), 669–738.
Hirano, K., Imbens, G.W., 2004. The propensity score with continuous treatments. In: Gelman, A., Meng, X.-L. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. John Wiley & Sons Ltd., pp. 73–84, chap. 7.
Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.
Imai, K., van Dyk, D.A., 2004. Causal inference with general treatment regimes: Generalizing the propensity score. J. Amer. Statist. Assoc. 99 (467), 854–866.
Imai, K., Ratkovic, M., 2014. Covariate balancing propensity score. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 243–263.
Imbens, G.W., 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87 (3), 706–710.
Imbens, G.W., Rubin, D.B., Sacerdote, B.I., 2001. Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. Amer. Econ. Rev. 91 (4), 778–794.
Imbens, G., Spady, R., Johnson, P., 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66 (2), 333–357.
Kang, J., Schafer, J., 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22 (4), 523–539.
Keane, M.P., Wolpin, K.I., 1997. The career decisions of young men. J. Polit. Econ. 105 (3), 473–522.
Kennedy, E.H., Ma, Z., McHugh, M.D., Small, D.S., 2017. Non-parametric methods for doubly robust estimation of continuous treatment effects. J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (4), 1229–1245.
Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65 (4), 861–874.
Koenker, R., Bassett, Jr., G., 1978. Regression quantiles. Econometrica 33–50.
Lee, D., Wolpin, K.I., 2006. Intersectoral labor mobility and the growth of the service sector. Econometrica 74 (1), 1–46.
Li, Q., Hsiao, C., Zinn, J., 2003. Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. J. Econometrics 112 (2), 295–325.
Li, Q., Racine, J.S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press.
Linton, O., Maasoumi, E., Whang, Y.-J., 2005. Consistent testing for stochastic dominance under general sampling schemes. Rev. Econom. Stud. 72 (3), 735–765.
Linton, O., Song, K., Whang, Y.-J., 2010. An improved bootstrap test of stochastic dominance. J. Econometrics 154 (2), 186–202.
Low, H., Meghir, C., 2017. The use of structural models in econometrics. J. Econ. Perspect. 31 (2), 33–58.
Low, H., Meghir, C., Pistaferri, L., 2010. Wage risk and employment risk over the life cycle. Amer. Econ. Rev. 100 (4), 1432–1467.
Low, H., Pistaferri, L., 2015. Disability insurance and the dynamics of the incentive insurance trade-off. Amer. Econ. Rev. 105 (10), 2986–3029.
McFadden, D., 1989. Testing for stochastic dominance. In: Studies in the Economics of Uncertainty. Springer, pp. 113–134.
Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79 (1), 147–168.
Praestgaard, J., Wellner, J.A., 1993. Exchangeably weighted bootstraps of the general empirical process. Ann. Probab. 21 (4), 2053–2086.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.
Smith, J.A., Todd, P.E., 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? J. Econometrics 125 (1–2), 305–353.
Tseng, P., Bertsekas, D.P., 1991. Relaxation methods for problems with strictly convex costs and linear constraints. Math. Oper. Res. 16 (3), 462–481.
van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.
Van Der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes with Applications to Statistics. Springer.
Zheng, J.X., 1996. A consistent test of functional form via nonparametric estimation techniques. J. Econometrics 75 (2), 263–289.