Quantile Structural Treatment Effect: Application to Smoking Wage Penalty and its Determinants*

Yu-Chin Hsu

Institute of Economics
Academia Sinica

Kamhon Kan

Institute of Economics
Academia Sinica

Tsung-Chih Lai†

Department of Economics
Feng Chia University

This version: August 19, 2019

* We thank the editor Esfandiar Maasoumi, an anonymous referee, Shakeeb Khan, Chung-Ming Kuan, Tatsushi Oka, and the participants at the 10th International Symposium on Econometric Theory and Applications for helpful comments and suggestions. All errors are our responsibility. Yu-Chin Hsu gratefully acknowledges the research support from Ministry of Science and Technology of Taiwan (MOST103-2628-H-001-001-MY4) and Career Development Award of Academia Sinica, Taiwan.

† Corresponding author. E-mail: tclai@fcu.edu.tw. No. 100, Wenhwa Rd., Taichung, 40724 Taiwan.

Abstract

In this paper, we propose a new parameter called the quantile structural treatment effect (QSTE) to distinguish between observed and unobserved treatment heterogeneity in semiparametric additive treatment effect models. The QSTE is defined as the quantile treatment effect when observed covariates are exogenously set to a fixed value while keeping unobserved heterogeneity unchanged. We show the QSTE is identified under unconfoundedness and propose a semiparametric inverse probability weighted-type estimator that converges weakly to a Gaussian process. A multiplier bootstrap procedure is also proposed to construct uniform confidence bands. Using data from the Panel Study of Income Dynamics and focusing on the female group for the plausibility of the unconfoundedness assumption, we examine observed and unobserved determinants of the adverse effects of smoking on wages known as the smoking wage penalty. Our findings suggest that different levels of observable human capital can partly explain the smoking heterogeneity on wages. However, no evidence is found to support unobservable explanations such as discrimination against smokers, especially in the upper tail of the unobserved heterogeneity distribution.

JEL Classification: C21, I12, J31

Keywords: Heterogeneous treatment effects, Structural functions, Smoking wage penalty, Observed and unobserved determinants.

1 Introduction

In the causal inference literature, various parameters are proposed to characterize different aspects of causal effects under the assumption that selection to treatment is based on observable characteristics. For example, the average treatment effect (ATE) captures the mean impact of a treatment, and the quantile treatment effect (QTE) assesses treatment heterogeneity across the outcome distribution.[1] In this literature, however, the covariates are often used only to achieve the identification of the treatment effect parameters, and researchers do not have a primary interest in how the effects relate to different values of covariates, especially from a distributional perspective.

In this paper, we propose a new parameter called the quantile structural treatment effect (QSTE) to distinguish between observed and unobserved treatment heterogeneity in semiparametric additive treatment effect models. That is, we specify the potential outcomes as (possibly nonlinear) parametrized structural functions of observed covariates plus unobserved errors representing individual heterogeneity. The QSTE is defined as the difference between the quantile structural functions (QSFs, Imbens and Newey, 2009) of the potential outcomes, where the QSF here is the inversion of the distribution function of the potential outcome when covariates are exogenously set to a fixed value while keeping individual heterogeneity unchanged.

The QSTE allows us to answer a number of counterfactual experiments asking "by how much the treatment effect would change if [the covariate] was switched from its value ... to its average value?" (Fortin, Lemieux, and Firpo, 2011, p. 9). This question is of particular interest if one aims to partial out different values of covariates from the total treatment effects. For instance, as will be illustrated in the empirical study, one possible explanation for the adverse wage effects of smoking is that smoking may be correlated with lower investments in human capital, which in turn reduces wages. To understand the heterogeneous effects of observable human capital on the so-called smoking wage penalty after correcting for self-selection bias, we propose to compare the QTE with the QSTE to disentangle observed heterogeneity (due to different levels of human capital) from the total wage differential.

On the other hand, in the QSTE framework the covariates are fixed and the only randomness comes from the unobserved error, the QSTE can therefore be used to examine unobserved treatment heterogeneity across the error distribution. We then apply this idea in our empirical analysis to examine whether unobserved factors (such as reduced productivity and discrimination against smokers) contribute to the smoking wage gap across individual heterogeneity distribution.

To identify the QSTE, we show that the QSF is equivalent to the quantile function of the error with a vertical shift whose magnitude equals to the structural function evaluated

[1] See Imbens and Wooldridge (2009) for a comprehensive review.

at the fixed value, where the parameters within the structural function and the error distribution function are both identifiable under the unconfoundedness assumption. For estimation, we first employ the weighted nonlinear least squares (WNLS) to estimate the structural function parameters, where the weight is estimated nonparametrically in the first stage. We then propose an inverse probability weighted (IPW) estimator of the distribution function of the error. The QSTE estimator can be constructed accordingly and is shown to converge weakly to a zero-mean Gaussian process at the parametric rate. We also propose a multiplier bootstrap to construct uniform confidence bands. Step-by-step implementation of the estimation and inference procedures used in the empirical study is provided as well.

In our empirical application, we revisit the issue of smoking wage penalty and its determinants using 2017 wave of the Panel Study of Income Dynamics (PSID). To our knowledge, this study is the first examining the distributional effects of smoking on wages. One advantage of using the PSID data is that it provides detailed information on the smoking history of respondents, which plays a crucial role in testing the unconfoundedness assumption using the test proposed by Donald, Hsu, and Lieli (2014). After establishing the validity of unconfoundedness for females given a larger and more substantial set of covariates suggested by previous literature, we then apply the proposed methods to yield new insights into the issue. Our findings suggest that different levels of observable characteristics can partly explain the smoking heterogeneity on wages. However, no evidence is found to support unobservable explanations, especially in the upper tail of the unobserved heterogeneity distribution for females.

The QSTE considered in this paper differs from the unconditional QTE proposed by Firpo (2007) in two important respects. First, given the additively separable structure, the QSTE can be decomposed into a structural part and a part consisting of individual heterogeneity as outlined in Section 2. Second, while Firpo (2007) focuses on the pointwise asymptotic results for the QTE, we focus on the QSTE uniformly over a continuum of quantile indices which includes the pointwise QSTE as a special case. Note also that the QSTE is in general different from the conditional QTE considered in Chernozhukov and Hansen (2005), albeit they are the same under an additional independence assumption (see Section 2 for details). The concept of structural functions has a long history in the literature (e.g., Blundell and Powell, 2003, 2004; Imbens and Newey, 2009; Wooldridge, 2015; Chernozhukov et al., 2018). Nevertheless, this paper appears to be the first one applying the structural functions to the potential outcomes.

The remainder of this paper is organized as follows. In Section 2 we provide a formal description of the model and the parameter of interest. The identification and estimation results are given in Section 3. Section 4 covers the asymptotic properties of the estimators and the multiplier bootstrap for uniform inference. The empirical application is illustrated in Section 5. Section 6 concludes. Technical details such as regularity conditions, influence

function estimation, and proofs are delegated to the Appendix.

2 Model and Parameter of Interest

2.1 Model

Following the Rubin causal model (Rubin, 1974), we let D be the binary treatment indicator such that $D = 1$ if the individual receives treatment and $D = 0$ otherwise. Denote Y_1 as the potential outcome for the individual under treatment and Y_0 as that without treatment. Let $Y = DY_1 + (1 - D)Y_0$ be the observed outcome and X be a d_x-dimensional vector of observed characteristics with compact support $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. In this paper, we consider semiparametric additive models for potential outcomes:

$$\begin{aligned} Y_1 &= m_1(X, \beta_1) + \epsilon_1, \quad E(\epsilon_1|X) = 0, \\ Y_0 &= m_0(X, \beta_0) + \epsilon_0, \quad E(\epsilon_0|X) = 0, \end{aligned} \tag{2.1}$$

where for $d = 1$ and 0, $m_d(x, b_d)$ represents the structural function of Y_d given $X = x$ that is known up to a finite-dimensional parameter vector b_d belonging to a compact set $\mathcal{B}_d \subseteq \mathbb{R}^{d_{b_d}}$. We denote β_d as the true parameter vector such that the conditional mean function $E(Y_d|X = x) = m_d(x, \beta_d)$ for all $x \in \mathcal{X}$. That is, we assume the error term ϵ_d, which plays the role of unobserved individual heterogeneity, to be conditional mean zero. Note that (2.1) is semiparametric in that the error distribution is unspecified. In addition, (2.1) is more general than linear models in that it allows for different functional forms of m_1 and m_0 as well as different dimensions of β_1 and β_0. One may extend (2.1) to be nonseparable in covariates and error. However, we follow Brinch, Mogstad, and Wiswall (2017) to impose additive separability for $\sqrt{n}$-consistency and asymptotic normality of our estimator.[2] To ease notation, we write d for 1 and 0 when the arguments or discussions apply to both cases.

2.2 Parameter of Interest

Given (2.1) and a prespecified covariate value x, we follow Imbens and Newey (2009) to define the distribution structural function (DSF) and QSF of the potential outcome Y_d as

$$\begin{aligned} G_d(x, y) &\equiv \int 1\{m_d(x, \beta_d) + e \leq y\} F_{\epsilon_d}(\mathrm{d}e), \\ q_d(x, \tau) &\equiv \inf\{y : G_d(x, y) \geq \tau\}, \end{aligned} \tag{2.2}$$

[2] As pointed out by Brinch, Mogstad, and Wiswall (2017), the additive separability between observed and unobserved components in (2.1) is implied by the linear regression of Y on D and X where the treatment indicator D and covariates X are additively separable.

where $1\{\cdot\}$ denotes the indicator function, $F(\cdot)$ is the distribution function, and $\tau \in [0,1]$. The DSF $G_d(x,y)$ can be interpreted as the distribution function of $m_d(x,\beta_d)+\epsilon_d$, which is the potential outcome Y_d with the structural part $m_d(X,\beta_d)$ being exogenously switched to a fixed value $m_d(x,\beta_d)$ while remaining unobserved individual heterogeneity ϵ_d unchanged. The QSF $q_d(x,\tau)$ is the corresponding quantile function of $G_d(x,y)$. The DSF and QSF allow us to distinguish treatment effects from different values of X across individuals. In fact, in this paper we are interested in the QSTE defined as the difference between QSFs:

$$\delta(x,\tau) \equiv q_1(x,\tau) - q_0(x,\tau), \quad \tau \in [0,1]. \tag{2.3}$$

The QSTE is different from the unconditional and conditional QTEs defined in, e.g., Firpo (2007) and Chernozhukov and Hansen (2005):

$$\delta_{\text{qte}}(\tau) \equiv Q_{Y_1}(\tau) - Q_{Y_0}(\tau), \qquad \delta_{\text{cqte}}(\tau,x) \equiv Q_{Y_1|X}(\tau|x) - Q_{Y_0|X}(\tau|x),$$

where $Q_{Y_d}(\tau) = \inf\{y : F_{Y_d}(y) \geq \tau\}$ and $Q_{Y_d|X}(\tau|x) = \inf\{y : F_{Y_d|X}(y|x) \geq \tau\}$ are the unconditional and conditional quantile functions of Y_d, respectively. To see the difference between QSTE and unconditional QTE, it is helpful to compare the DSF with the unconditional distribution function of Y_d given (2.1):

$$\begin{aligned} G_d(x,y) &= \int 1\{m_d(x,\beta_d)+e \leq y\} F_{\epsilon_d}(\mathrm{d}e) = E_{\epsilon_d}[1\{m_d(x,\beta_d)+\epsilon_d \leq y\}], \\ F_{Y_d}(y) &= E_{Y_d}[1\{Y_d \leq y\}] = E_{X,\epsilon_d}[1\{m_d(X,\beta_d)+\epsilon_d \leq y\}]. \end{aligned}$$

Clearly, the expectation of $G_d(x,y)$ is taken with respect to the marginal distribution of ϵ_d, while the expectation of $F_{Y_d}(y)$ is taken with respect to the joint distribution of (X,ϵ_d). In other words, $G_d(x,y)$ is the distribution function of $m_d(x,\beta_d)+\epsilon_d$ in which x is fixed over the unconditional distribution of ϵ_d, whereas $F_{Y_d}(y)$ is the distribution function of Y_d or $m_d(X,\beta_d)+\epsilon_d$ in which X is not fixed at one specific point.

We next compare the QSTE with the conditional QTE. The two objects are in principle different for different targeted populations: the QSTE focuses on the entire population (with a manipulated covariate value x) to assess unobserved treatment heterogeneity across individuals, whereas the conditional QTE only focuses on a subpopulation defined by $X = x$. More specifically, note that the conditional distribution function of Y_d given (2.1) can be written as

$$F_{Y_d|X}(y|x) = E_{\epsilon_d|X}[1\{m_d(X,\beta_d)+\epsilon_d \leq y\}|x] = E_{\epsilon_d|X}[1\{m_d(x,\beta_d)+\epsilon_d \leq y\}|x],$$

where the expectation is taken with respect to the conditional distribution of ϵ_d given $X = x$. Clearly, $G_d(x,y)$ and $F_{Y_d|X}(y,x)$ are generally different unless ϵ_d and X are independent. Although we do not impose independence between ϵ_d and X throughout

the paper, our results regarding the QSTE can be readily extended to the conditional QTE provided the independence assumption is satisfied.

It is also worth noting that we do not impose rank invariance assumption on ϵ_1 and ϵ_0 that requires the relative value (rank) of ϵ_1 and ϵ_0 for a given individual to be the same regardless of whether that individual is in the treatment or the control group. Instead, we follow Firpo (2007) to interpret the QSTE from the perspective of a policy-maker who is interested in learning about the marginal distributions of the potential outcomes. For related discussions, please see Firpo (2007).[3]

For identification and estimation, we show the DSF, QSF, and QSTE can be simplified in the following lemma.

Lemma 2.1. *Suppose Y_d is generated according to* (2.1)*. Then the DSF, QSF, and QSTE defined in* (2.2) *and* (2.3) *can be simplified to:*

$$\begin{aligned} G_d(x,y) &= F_{\epsilon_d}(y - m_d(x,\beta_d)), \\ q_d(x,\tau) &= m_d(x,\beta_d) + Q_{\epsilon_d}(\tau), \\ \delta(x,\tau) &= m_1(x,\beta_1) - m_0(x,\beta_0) + Q_{\epsilon_1}(\tau) - Q_{\epsilon_0}(\tau), \end{aligned}$$

where $F_{\epsilon_d}(\cdot)$ and $Q_{\epsilon_d}(\cdot)$ are the distribution and quantile functions of ϵ_d, respectively.

Lemma 2.1 shows that $G_d(x,y)$ corresponds to the distribution function of ϵ_d with a horizontal shift of $m_d(x,\beta_d)$ and $q_d(x,\tau)$ corresponds to the quantile function of ϵ_d with a vertical shift of $m_d(x,\beta_d)$. The QSTE $\delta(x,\tau)$ can thus be decomposed into two parts: (i) the constant part $m_1(x,\beta_1) - m_0(x,\beta_0)$ associated with fixed, observed covariate value x, and (ii) the varying part associated with unobserved treatment heterogeneity $Q_{\epsilon_1}(\tau) - Q_{\epsilon_0}(\tau)$.

3 Identification and Estimation

3.1 Identification

Despite the simplification afforded by Lemma 2.1, one still cannot identify the QSTE since only one of the potential outcomes is observed for each individual. We therefore impose the unconfoundedness assumption introduced by Rosenbaum and Rubin (1983). This assumption requires that treatment assignment is independent of the potential outcomes conditional on observed covariates. In addition, the propensity score $p(x) = P(D = 1|X = x)$ defined as the conditional probability of receiving treatment given $X = x$ is also assumed to be bounded away from 0 and 1 on $\mathcal{X}$. This requirement is known as the

[3]One might modify the tests developed in Frandsen and Lefgren (2018) and Dong and Shen (2018) to test the rank invariance assumption in our framework. However, it requires substantial work to work out the theory which is beyond the scope of this paper.

overlap condition since it equivalently assumes that the treated and untreated supports are completely overlapped.

Assumption 3.1. *Suppose that*

(i) (Unconfoundedness) $D \perp\!\!\!\perp (Y_1, Y_0)|X$.

(ii) (Overlap) $0 < \underline{p} \leq p(x) \leq \overline{p} < 1$ *for all* $x \in \mathcal{X}$.

Note that under (2.1), Assumption 3.1(i) is equivalent to assuming $D \perp\!\!\!\perp (\epsilon_1, \epsilon_0)|X$. However, this assumption is controversial in observational studies. To assess the validity of the unconfoundedness assumption, one can utilize the test proposed by Donald, Hsu, and Lieli (2014) with an additional binary instrumental variable (see Section 5 for more details). On the other hand, if the overlap condition is not met initially, one solution would be trimming $\mathcal{X}$ and redefining the population as advocated in Crump et al. (2009). To identify β_d, we make the following assumptions.

Assumption 3.2. *Suppose that*

(i) $E(Y_d|X = x) = m_d(x, \beta_d)$ *for some* $\beta_d \in \mathcal{B}_d$ *and for all* $x \in \mathcal{X}$.

(ii) $E[m_d(X, \beta_d) - m_d(X, b_d)]^2 > 0$ *for all* $b_d \in \mathcal{B}_d$, $b_d \neq \beta_d$.

Assumption 3.2(i) requires the conditional mean of Y_d to be correctly specified, which is identical to assuming that $E(\epsilon_d|X = x) = 0$ for all $x \in \mathcal{X}$ as mentioned above. Assumption 3.2(ii) requires that β_d is the unique parameter value such that $E(Y_d|X = x) = m_d(x, \beta_d)$ for all $x \in \mathcal{X}$.

Lemma 3.1. *Suppose Assumptions 3.1 and 3.2 hold. Then the QSTE is identified by*

$$\delta(x, \tau) = m_1(x, \beta_1) - m_0(x, \beta_0) + Q_{\epsilon_1}(\tau) - Q_{\epsilon_0}(\tau),$$

where $Q_{\epsilon_d}(\tau) = \inf\{e : F_{\epsilon_d}(e) \geq \tau\}$ *and*

$$\beta_d = \underset{b_d \in \mathcal{B}_d}{\operatorname{argmin}} E\left[\frac{1\{D = d\}[Y - m_d(X, b_d)]^2}{p(X)^d[1 - p(X)]^{1-d}}\right],$$
$$F_{\epsilon_d}(e) = E\left[\frac{1\{D = d\}1\{Y - m_d(X, \beta_d) \leq e\}}{p(X)^d[1 - p(X)]^{1-d}}\right].$$

3.2 Estimation

Given a random sample $\{(D_i, X_i, Y_i) : i = 1, \ldots, n\}$, we propose a two-step semiparametric estimator of the QSTE that is later shown to converge at the parametric $\sqrt{n}$ rate. We first estimate β_d by WNLS:

$$\hat{\beta}_d = \underset{b_d \in \mathcal{B}_d}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1\{D_i = d\}[Y_i - m_d(X_i, b_d)]^2}{\hat{p}(X_i)^d[1 - \hat{p}(X_i)]^{1-d}}, \tag{3.1}$$

where $\hat{p}(x)$ is a nonparametric estimator of $p(x)$.[4,5] Similar to Hirano, Imbens and Ridder (2003), we use the series logit estimator (SLE) to estimate $p(x)$ based on a power series. Other nonparametric estimators can be employed, e.g., local polynomial estimator as in Ichimura and Linton (2005) and kernel estimator as in Abrevaya, Hsu and Lieli (2015). The main advantage of SLE over other nonparametric estimators is that the estimated propensity score is automatically bounded away from 0 and 1 without any trimming. Next, we propose a normalized IPW estimator of F_{ϵ_d}:

$$\hat{F}_{\epsilon_d}(e) = \sum_{i=1}^{n} \frac{1\{D_i = d\}1\{Y_i - m_d(X_i, \hat{\beta}_d) \leq e\}}{\hat{p}(X_i)^d[1-\hat{p}(X_i)]^{1-d}} \Bigg/ \sum_{i=1}^{n} \frac{1\{D_i = d\}}{\hat{p}(X_i)^d[1-\hat{p}(X_i)]^{1-d}}. \quad (3.2)$$

Note that the normalization is necessary for the distribution function estimator $\hat{F}_{\epsilon_d}(e)$ to lie within the unit interval. In addition, since the inverse probability weight in (3.2) is strictly positive due to the SLE, $\hat{F}_{\epsilon_d}(e)$ is a proper distribution function that can be inverted for quantile estimation: $\hat{Q}_{\epsilon_d}(\tau) = \inf\{e : \hat{F}_{\epsilon_d}(e) \geq \tau\}$. Finally, the QSTE estimator is given by

$$\hat{\delta}(x, \tau) = m_1(x, \hat{\beta}_1) - m_0(x, \hat{\beta}_0) + \hat{Q}_{\epsilon_1}(\tau) - \hat{Q}_{\epsilon_0}(\tau). \quad (3.3)$$

4 Asymptotic Properties and Multiplier Bootstrap

In this section, we first investigate the asymptotic properties of the QSTE estimator and then propose a multiplier bootstrap procedure to construct uniform confidence bands. To improve readability, all regularity conditions and the corresponding discussions are postponed to Appendix A.

4.1 Asymptotic Properties

To study the asymptotic properties of the QSTE estimator, we first provide influence function representations of $\hat{\beta}_d$ and $\hat{F}_{\epsilon_d}(e)$. Let $\nabla m_d(x, b_d)$ denote the $d_{b_d} \times 1$ gradient of $m_d(x, b_d)$ with respect to b_d.

Lemma 4.1. *Suppose Assumptions 3.1, 3.2, and A.1–A.4 in Appendix A hold. Let $\hat{\beta}_d$ and*

[4] Alternatively, one can estimate $p(x)$ parametrically by logit or probit regression similar to Wooldridge (2007).

[5] Although the WNLS is not the most efficient estimator among the class of estimators under the conditional moment restriction or Assumption 3.2(i), the focus of this paper is the estimation and inference of the QSTE so we use WNLS here for its simplicity.

$\hat{F}_{\epsilon_d}(e)$ *be the estimators defined in* (3.1) *and* (3.2)*, respectively. Then for* $W = (X, Y, D)$*,*

$$\sqrt{n}(\hat{\beta}_d - \beta_d) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\beta_d}(W_i, \beta_d) + o_p(1),$$
$$\sqrt{n}(\hat{F}_{\epsilon_d}(e) - F_{\epsilon_d}(e)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\epsilon_d}(W_i, e) + o_p(1),$$

where the influence functions $\psi_{\beta_d}(W, \beta_d)$ *and* $\psi_{\epsilon_d}(W, e)$ *are given by*

$$\psi_{\beta_d}(W, \beta_d) = E\left[\frac{1\{D = d\}\nabla m_d(X, \beta_d)\nabla m_d(X, \beta_d)'}{p(X)^d[1 - p(X)]^{1-d}}\right]^{-1} \cdot \frac{1\{D = d\}\nabla m_d(X, \beta_d)[Y - m_d(X, \beta_d)]}{p(X)^d[1 - p(X)]^{1-d}},$$
$$\psi_{\epsilon_d}(W, e) = \frac{1\{D = d\}1\{Y - m_d(X, \beta_d) \le e\}}{p(X)^d[1 - p(X)]^{1-d}} - F_{\epsilon_d}(e) + \left\{1 - \frac{1\{D = d\}}{p(X)^d[1 - p(X)]^{1-d}}\right\} F_{\epsilon_d|X}(e|X) + f_{\epsilon_d}(e)E[\nabla m_d(X, \beta_d)]'\psi_{\beta_d}(W, \beta_d),$$

where $F_{\epsilon_d|X}(\cdot|x)$ *is the conditional distribution function of* ϵ_d *given* $X = x$ *and* $f_{\epsilon_d}(\cdot)$ *is the density function of* ϵ_d*.*

We make several remarks concerning Lemma 4.1. First, the estimation effect of $\hat{p}(x)$ on $\hat{\beta}_d$ is asymptotically negligible in that the asymptotic variance of $\hat{\beta}_d$ is equivalent to the variance of $\psi_{\beta_d}(W, \beta_d)$ despite the fact that $\hat{p}(x)$ is estimated nonparametrically. Second, the influence function $\psi_{\epsilon_d}(W, e)$ consists of three components: (i) the first line of $\psi_{\epsilon_d}(W, e)$ corresponds to the influence function if the true propensity score $p(x)$ is known and need not be estimated, (ii) the second line is the contribution of estimating $p(x)$ to the asymptotic process of $\hat{F}_{\epsilon_d}(e)$, and (iii) the last line comes from the estimation error associated with $\hat{\beta}_d$.

By (3.3) and Lemma 4.1, the following theorem states that the QSTE estimator will converge weakly to a Gaussian process at the parametric rate.

Theorem 4.1. *Suppose Assumptions 3.1, 3.2, and A.1–A.4 in Appendix A hold. Let* $\hat{\delta}(x, \tau)$ *be the estimator defined in* (3.3)*. Then*

$$\sqrt{n}(\hat{\delta}(x, \tau) - \delta(x, \tau)) \Rightarrow \Delta(x, \tau),$$

where $\Rightarrow$ *denotes weak convergence and* $\Delta(x, \tau)$ *is a zero-mean Gaussian process with the*

covariance kernel being generated by the influence function

$$\psi_\delta(W,x,\tau) = \nabla m_1(x,\beta_1)'\psi_{\beta_1}(W,\beta_1) - \nabla m_0(x,\beta_0)'\psi_{\beta_0}(W,\beta_0) - \left[\frac{\psi_{\epsilon_1}(W,Q_{\epsilon_1}(\tau))}{f_{\epsilon_1}(Q_{\epsilon_1}(\tau))} - \frac{\psi_{\epsilon_0}(W,Q_{\epsilon_0}(\tau))}{f_{\epsilon_0}(Q_{\epsilon_0}(\tau))}\right],$$

where $\psi_{\beta_d}(W,\beta_d)$ and $\psi_{\epsilon_d}(W,e)$ are defined in Lemma 4.1.

From (3.3), it is not surprising that the influence function $\psi_\delta(W,x,\tau)$ has a similar two-component decomposition: (i) the first line of $\psi_\delta(W,x,\tau)$ is associated with the estimation effect of $\hat{\beta}_d$ on $m_d(x,\hat{\beta}_d)$ and (ii) the second line is associated with the influence function of $Q_{\epsilon_d}(\tau)$ obtained by the functional delta method given the Hadamard differentiability of the quantile map. Since $\{\psi_\delta(W,x,\tau) : \tau \in [0,1]\}$ belongs to some Donsker class under regularity conditions, Theorem 4.1 follows immediately from the functional central limit theorem.

4.2 Multiplier Bootstrap

Based on the asymptotic properties in Theorem 4.1, one can easily conduct pointwise inference on $\delta(x,\tau)$ at a fixed point τ provided the influence function $\psi_\delta(W,x,\tau)$ is consistently estimated. However, conducting uniform inference is not a straightforward extension since it involves a continuum of quantile indices rather than a single point. One popular alternative is the nonparametric bootstrap, but doing this could be time consuming because $p(x)$, β_d, and $Q_{\epsilon_d}(\tau)$ have to be estimated in each replication. Hence, we propose a simulation-based method known as the multiplier bootstrap to approximate the limiting process $\Delta(x,\tau)$. The multiplier bootstrap is computationally more attractive than the nonparametric bootstrap since its source of randomness comes from the exogenously generated multipliers instead of resampling. However, it comes at a cost that the estimator of $\psi_\delta(W,x,\tau)$ must be uniformly consistent. We provide such an estimator in Appendix B. The explicit form of the estimated influence function can also be found there.

Given the estimated influence function $\hat{\psi}_\delta(W_i,x,\tau)$, we let the multipliers $\{U_i : i = 1,\ldots,n\}$ be i.i.d. pseudo random variables with mean 0 and variance 1 (e.g., standard normal random variables) that are independent of the sample. The simulated process for $\Delta(x,\tau)$ is given by

$$\Delta^u(x,\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \cdot \hat{\psi}_\delta(W_i,x,\tau). \tag{4.1}$$

In fact, one can show that under regularity conditions, $\Delta^u(x,\tau)$ approximates the limiting process well in that $\Delta^u(x,\tau)$ converges weakly to the limiting process $\Delta(x,\tau)$ conditional on the sample path with probability approaching 1. That is, by generating multipliers

$\{U_i : i = 1, \ldots, n\}$ many times (say B times), the simulated processes $\{\Delta_b^u(x,\tau) : b = 1, \ldots, B\}$ allow us to approximate the whole limiting process as well as critical values over a continuum of quantile indices.

To illustrate the usefulness of the above result, we briefly demonstrate the construction of one-sided and two-sided uniform confidence bands here. For a confidence level α and for $[\eta, 1-\eta]$ where $\eta \in [0, 1/2)$, let $c_{\alpha,\eta}^{\text{1-sided}}$ and $c_{\alpha,\eta}^{\text{2-sided}}$ be the one- and two-sided critical values for $\{\delta(x,\tau) : \tau \in [\eta, 1-\eta]\}$ that satisfy

$$c_{\alpha,\eta}^{\text{1-sided}} = \inf\left\{ y : \Pr\left(\sup_{\tau\in[\eta,1-\eta]} \frac{\Delta^u(x,\tau)}{\hat{\sigma}(x,\tau)} \le y \right) \ge \alpha \right\},$$
$$c_{\alpha,\eta}^{\text{2-sided}} = \inf\left\{ y : \Pr\left(\sup_{\tau\in[\eta,1-\eta]} \frac{|\Delta^u(x,\tau)|}{\hat{\sigma}(x,\tau)} \le y \right) \ge \alpha \right\},$$

where $\Delta^u(x,\tau)$ is from (4.1) and

$$\hat{\sigma}(x,\tau) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_\delta^2(W_i, x, \tau)} \tag{4.2}$$

is the pointwise standard deviation of the QSTE. That is, $c_{\alpha,\eta}^{\text{1-sided}}$ and $c_{\alpha,\eta}^{\text{2-sided}}$ are the α-th quantiles of $\Delta^u(x,\tau)/\hat{\sigma}(x,\tau)$ and $|\Delta^u(x,\tau)|/\hat{\sigma}(x,\tau)$, respectively. The lower and upper one-sided $(1-\alpha)$ uniform confidence bands for the QSTE are then given by

$$\hat{\delta}(x,\tau) - c_{\alpha,\eta}^{\text{1-sided}} \frac{\hat{\sigma}(x,\tau)}{\sqrt{n}} \quad \text{and} \quad \hat{\delta}(x,\tau) + c_{\alpha,\eta}^{\text{1-sided}} \frac{\hat{\sigma}(x,\tau)}{\sqrt{n}}, \tag{4.3}$$

and the two-sided $(1-\alpha)$ uniform confidence band for the QSTE is

$$\left[\hat{\delta}(x,\tau) - c_{\alpha,\eta}^{\text{2-sided}} \frac{\hat{\sigma}(x,\tau)}{\sqrt{n}},\ \hat{\delta}(x,\tau) + c_{\alpha,\eta}^{\text{2-sided}} \frac{\hat{\sigma}(x,\tau)}{\sqrt{n}} \right]. \tag{4.4}$$

5 Empirical Study

In this section, we apply the proposed methods to investigate the wage penalty of smoking and the underlying determinants. It is well established in the literature that smokers on average earn 4–24% less than non-smokers after controlling for observed and unobserved characteristics that might be correlated with smoking and wages (Levine, Gustafson, and Velenchik, 1997; van Ours, 2004; Auld, 2005; Grafova and Stafford, 2009; Lång and Nystedt, 2018). To examine potential explanations and address some issues, we briefly discuss these studies below.

5.1 Literature Review

As a pioneer study in this field, Levine, Gustafson, and Velenchick (1997) discovered that smoking reduces wages by roughly 4–8% using panel and sibling data to control for unobservable characteristics that are constant over time and are constant within a family. They also offered several explanations which can be broadly classified into observable and unobservable ones. For the former, smoking might indicate a high rate of time preference which would be associated with fewer investments in observable human capital and hence lower wages. For the latter, reduced productivity (due to ill-health and absenteeism) and discrimination against smokers are two leading unobservable explanations for this wage penalty.

However, their investigation on potential explanations yields no conclusive results partly due to the lack of attention to the heterogeneous nature of smokers. In recognizing the differences among individuals for smoking behavior, Grafova and Stafford (2009), using PSID data, constructed a retrospective sample by smoking history and found adverse wage effects of 7–12% for persistent smokers versus never or former smokers. They also suggest that the wage differential may be driven by a non-causal explanation rather than by smoking per se, indicating the importance of controlling for self-selection when evaluating the consequences of smoking.

Taking self-selection into account, van Ours (2004) and Auld (2005) studied the wage effects of the use of alcohol and tobacco based on instrumental variable estimation. Using early exposure to alcohol and tobacco as instruments, van Ours (2004) found that drinking generates a wage premium for males of about 10% while smoking reduces wages by about 10%. Auld (2005) further showed that the smoking wage penalty rises from 8% to 24% after correcting for endogeneity by assuming that the substance prices and religious status only affect drinking and smoking but not wages directly. Using twin data to neutralize the endogeneity arising from the twins' unobserved common background, Lång and Nystedt (2018) analyzed the smoking-related earnings penalty in two different social contexts 30 years apart. While the earnings penalty were insignificant or in the opposite direction in the 1970s for men and women, smokers were estimated to earn 6–10% less than non-smokers in the 2000s by different genders.

Concluding from these studies, to understand the mechanisms underlying the smoking wage penalty, one needs to address at least three issues: (i) self-selection into smoking; (ii) heterogeneous wage effects; (iii) observed and unobserved determinants. For the first issue, we assess the validity of the unconfoundedness assumption which requires that smoking is as good as randomly assigned conditional on a sufficiently large set of observed covariates. To deal with the second one, we make use of the QTE to uncover the heterogeneous smoking effects across wage quantiles. Finally, we utilize the QSTE to set observed covariates at a fixed value but keeping unobserved smoking heterogeneity unaltered. The

difference between QTE and QSTE thus contributes to the observed heterogeneity due to different levels of human capital, whereas the QSTE along can be used to examine unobserved smoking heterogeneity across individual heterogeneity distribution.

5.2 Data and Unconfoundedness Test

We use data from the 2017 wave of the PSID consisting of over 9,600 household heads (respondents). Began in 1968, the PSID is the longest-running longitudinal household survey containing rich information on socio-economic and demographic characteristics. In particular, the PSID collects data on smoking behavior in the 1986 wave and every wave since the 1999's. This panel structure allows us to track respondents' smoking history that plays a vital role in testing the unconfoundedness assumption. The outcome variable is the log of the average hourly wage rate. The treatment variable is the current smoking status. Baseline covariates include age, years of education (highest grade completed), a race dummy, two occupational dummies (white-collar and blue-collar), and union status.[6]

In addition to the baseline covariates, we also control for the state cigarette price and religious sentiment to account for possible correlations between smoking and unobserved regional and individual characteristics that might affect wages directly.[7] For the state cigarette price, we match the average cost per pack of cigarettes from the Tax Burden on Tobacco (Orzechowski and Walker, 2019) to PSID respondents' state of residence. For the religious sentiment, we use the frequency of attending religious services to measure religiosity. Following Grafova and Stafford (2009), the sample is restricted to individuals between the ages of 25 and 65 who worked at least 1,500 hours a year. Descriptive statistics by gender and smoking status are summarized in Table 1.

To assess the validity of the unconfoundedness assumption, we utilize the test proposed by Donald, Hsu, and Lieli (2014) and the early-smoking status (whether started smoking before age 25) as a binary instrument similar to van Ours (2004).[8] Donald, Hsu, and Lieli (2014) showed that, given the binary instrument satisfying one-sided non-compliance, the local average treatment effect on the treated and the average treatment effect on the treated would be identical under the null hypothesis of unconfoundedness. They then proposed a Durbin-Wu-Hausman-type test based on this observation. The one-sided non-compliance condition in our case is equivalent to requiring that for all individuals,

[6]We set the white-collar dummy equal to one if the respondent's occupation code belongs to management, business, science, and arts occupations as classified by the 2010 Census occupation codes. The blue-collar dummy equals one if the code belongs to natural resources, construction, maintenance, production, transportation, and material moving occupations.

[7]Although prices and religiosity were treated as instruments in Auld (2005), Some authors (e.g., French and Popovici, 2011) argue that religiosity might be correlated with unobserved personal characteristics that directly affect labor market outcomes.

[8]Van Ours (2004) used whether started smoking before age 16 as an instrument because his sample includes individuals aged 16 to over 65. We use age 25 as a cutoff because we focus on individuals between the ages of 25 and 65.

should they not smoke before age 25, would not smoke currently as well. Although this condition is not testable for the early-smokers, it can be verified for the remaining and larger group—those who never smoked before age 25—thanks to the smoking history data collected in the PSID. In our sample, only a small fraction (69 out of 3,297, around 2%) of individuals who did not smoke before the age of 25 became smokers afterward. We then exclude these individuals to (partially) justify the one-sided non-compliance of our instrument.[9]

The results for the unconfoundedness test are presented in Table 2. For male and female groups, we report p-values of the test and consider several implementations of the SLE in estimating the instrument propensity scores. Specifically, we start with a constant model and then add linear, interaction, and quadratic terms of the covariates in the power series. From the upper panel of Table 2, it can be seen that the unconfoundedness assumption is rejected at 10% significance level in most cases when the conditional set only includes baseline covariates. However, if we additionally condition on the cigarette price and religiosity, the unconfoundedness cannot be rejected for females at 10% level. This finding suggests that a comprehensive set of observable characteristics can plausibly explain the self-selection into smoking status for females. As a result, we focus on the female respondents for the remainder of our analysis.

5.3 Implementation Details

Here we outline the implementation of the estimation and inference procedures used in the empirical study.

1. (SLE for Propensity Score). We introduce a power series $R^K(X)$ including all polynomials of X up to order K, where K satisfies Assumption A.4(ii). For example, if X only includes two variables $X = (X_1, X_2)'$ and $K = 2$, then $R^K(X) = (1, X_1, X_2, X_1 \cdot X_2, X_1^2, X_2^2)'$. The SLE $\hat{p}(X_i)$ is defined as the fitted value from the logistic regression of D_i on $R^K(X_i)$. One can also trim $\hat{p}(X_i)$ for a better finite sample performance following Crump et al. (2009).

2. (QTE). Given the estimated propensity score $\hat{p}(X_i)$ and a grid of quantiles $\mathcal{T}$, say $\mathcal{T} = \{0.05, 0.06, \ldots, 0.95\}$, we estimate the quantile functions of the potential outcomes and the QTE based on the IPW method proposed by Firpo (2007). One can also conduct uniform inference on QTE following Donald and Hsu (2014).

3. (QSTE). Given $\hat{p}(X_i)$, we estimate β_d and $F_{\epsilon_d}(e)$ according to (3.1) and (3.2). In particular, if the structural function is of linear form, i.e., $m_d(X, \beta_d) = X'\beta_d$, the

[9]For the early-smokers, we assume that they do not smoke currently had they not smoked before age 25.

WNLS estimator $\hat{\beta}_d$ can be simplified to a closed-form expression:

$$\hat{\beta}_d = \left\{ \sum_{i=1}^{n} \frac{1\{D_i = d\} X_i X_i'}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}} \right\}^{-1} \sum_{i=1}^{n} \frac{1\{D_i = d\} X_i Y_i}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}}.$$

Hence, we consider linear specifications of the structural functions m_1 and m_0 throughout the empirical study. Moreover, we estimate $F_{\epsilon_d}(e)$ at $e = y - \overline{y}$, where y consists of all unique values of Y_i's and $\overline{y}$ is the average of y. The QSTE can be obtained by plugging $\hat{\beta}_d$ and $\hat{F}_{\epsilon_d}(e)$ in (3.3) with a prespecified covariate value x.

4. (Fixed covariate value). The fixed value x we consider is given in Table 3. Specifically, for dummy variables (race, occupational status, and union status), the fixed values are assigned to the most frequently occurring combination in the dataset. For continuous variables (age, education, cigarette price, and religiosity), the fixed values are assigned to the means of the corresponding variables in the most frequently occurring combination group.[10]

5. (Influence Function). To construct uniform confidence bands of the QSTE via multiplier bootstrap, we first estimate the corresponding influence function based on uniformly consistent estimators $\hat{F}_{\epsilon_d|X}(e|x)$ and $\hat{f}_{\epsilon_d}(e)$ given in Appendix B. We use the same power series as in step 1 for $\hat{F}_{\epsilon_d|X}(e|x)$. For $\hat{f}_{\epsilon_d}(e)$, we use the Gaussian kernel with normal reference bandwidth $h = 1.06\hat{\sigma}_d n^{-1/5}$, where $\hat{\sigma}_d$ is the standard deviation of the residual $Y_i - X_i'\hat{\beta}_d$.

6. (Multiplier Bootstrap). Given the estimated influence function, our multiplier bootstrap procedure is as follows: (i) generate random variables $\{U_i : i = 1, \ldots, n\}$ from the standard normal distribution B times, say $B = 1000$; (ii) for each replication $b = 1, \ldots, B$, obtain the simulated process $\Delta_b^u(x, \tau)$ according to (4.1). (iii) store $M_b = \max_{\tau \in \mathcal{T}} |\Delta_b^u(x, \tau)| / \hat{\sigma}(x, \tau)$ for the critical value, where $\hat{\sigma}(x, \tau)$ is the standard deviation given in (4.2).

7. (Uniform Confidence Bands). Rank M_b in an ascending order such that $M_{(1)} \leq \cdots \leq M_{(B)}$. Given confidence level α, define the critical value $c_\alpha = M_{(\lfloor \alpha B \rfloor)}$, where $\lfloor c \rfloor$ is the floor function returning the largest integer not greater than c. In our empirical study, we let $\alpha = 0.90$ and $B = 1000$ so that the critical value is given by $M_{(900)}$ or the 900th largest value of $\{M_1, \ldots, M_{1000}\}$. The two-sided uniform confidence bands can then be constructed according to (4.4).

[10] We also try different fixed values for continuous variables such as the overall means. The results are similar and are available upon request.

5.4 Results

Our empirical results are two-fold. First, Figure 1 depicts the quantile functions of (i) potential wages if all individuals were smoking (thick solid line); (ii) potential wages if all individuals were not smoking (thin solid line); (iii) observed wages for smokers (thick dashed line); (iv) observed wages for non-smokers (thin dashed line). Not surprisingly, the potential and observed wages in the smoking case are lower than their non-smoking counterparts for all quantiles considered. However, it is more interesting that the gap between potential wages remains relatively wide (albeit narrower than the observed one), indicating the presence of smoking wage penalty for females even after correcting for self-selection bias. This argument is reinforced by the QTE and the corresponding 90% uniform confidence band shown in Figure 2. It follows that smoking causes significant adverse effects on wages at the middle quantiles, where the magnitudes (around 25%) are similar to the average effect (24%) reported in Auld (2005) after correcting for endogeneity.

Second, and more importantly, we investigate the observed and unobserved determinants of the smoking wage differential via QSTE in Figure 3. For comparison, we also depict the QTE and the observed wage gap. As seen in Figure 3, the gap between QSTE and QTE is considerable and is more pronounced in the upper quantiles. Since this difference accounts for the observed smoking heterogeneity as mentioned earlier, we argue that different levels of observable human capital characteristics could be one of the driving forces contributing to the smoking wage penalty for females. On the other hand, it can also be seen that the QSTE is statistically insignificant across all quantiles and is closer to zero in the upper quantiles. This finding suggests that, given the observed covariates fixed, women with high individual heterogeneity suffer less from the smoking-induced wage losses. Hence, we interpret this as supporting evidence that the smoking wage penalty is not driven by unobserved factors such as reduced productivity or discrimination against smokers, at least for females in the upper end of the individual heterogeneity distribution. For robustness checks, we also try different, representative fixed values for observable characteristics. The results are similar and are available upon request.

6 Conclusion

In this paper, we propose the QSTE to distinguish between observed and unobserved treatment heterogeneity by partialling out different values of covariates while remaining unobserved heterogeneity unchanged. We provide the identification, estimation, and uniform inference for the QSTE in semiparametric additive treatment effect models under unconfoundedness. Using PSID data, we investigate the observed and unobserved determinants of the smoking-induced wage differential after establishing the validity of unconfoundedness for the female group. Our findings suggest that, while the smoking wage

penalty is significant for females after accounting for selection bias, this wage differential is mainly driven by different levels of observable characteristics instead of unobserved factors, especially for females in the upper end of the individual heterogeneity distribution.

APPENDIX

A Regularity Conditions

We impose the following regularity conditions for the asymptotic properties of the QSTE estimator. Let $\nabla m_d(x, b_d)$ denote the $d_{b_d} \times 1$ gradient of $m_d(x, b_d)$ with respect to b_d and $\nabla^2 m_d(x, b_d)$ denote the $d_{b_d} \times d_{b_d}$ Hessian of $m_d(x, b_d)$. Let $\|\cdot\|$ denote the Euclidean norm of a matrix.

Assumption A.1 (Support of X)**.**

(i) The support of X is a Cartesian product of compact intervals, $\mathcal{X} = \prod_{i=1}^{d_x}[x_{i\ell}, x_{iu}]$.

(ii) The density function of X is bounded away from 0 on $\mathcal{X}$.

Assumption A.2 (Structural Functions)**.**

(i) β_d is in the interior of $\mathcal{B}_d$ which is a compact subset of $\mathbb{R}^{d_{b_d}}$.

(ii) For each $b_d \in \mathcal{B}_d$, $m_d(\cdot, b_d)$ is Borel measurable on $\mathcal{X}$.

(iii) For each $x \in \mathcal{X}$, $m_d(x, \cdot)$ is continuously differentiable of order 2 in $b_d \in \mathcal{B}_d$.

(iv) $E[\sup_{b_d \in \mathcal{B}_d} \|\nabla m_d(X, b_d)\|^2] < \infty$ and $E[\sup_{b_d \in \mathcal{B}_d} \|\nabla^2 m_d(X, b_d)\|^2] < \infty$.

(v) $E[\nabla m_d(X, \beta_d) \nabla m_d(X, \beta_d)']$ is positive definite.

Assumption A.3 (Unobservables)**.**

(i) ϵ_d has a compact support $[e_{d\ell}, e_{du}]$ and denote $\mathcal{E} = [e_\ell, e_u]$ where $e_\ell = \min\{e_{0\ell}, e_{1\ell}\}$ and $e_u = \max\{e_{0u}, e_{1u}\}$.

(ii) The density function $f_{\epsilon_d}(e)$ is bounded away from 0 and continuously differentiable of order 2 on $[e_{d\ell}, e_{du}]$.

Assumption A.4 (Series Logit Estimator)**.**

(i) $p(x)$ is continuously differentiable of order $s \geq 7d_x$.

(ii) $\hat{p}(x)$ uses a power series with order $K = a \cdot n^\nu$ for some $a > 0$ and $d_x/4(s - d_x) < \nu < 1/9$.

Assumption A.2 introduces standard conditions for the asymptotic properties of the WNLS estimator, see Wooldridge (2010) for details. Assumption A.3(i) requires ϵ_d to be supported on a compact interval, which is not restrictive in that the theory regarding the estimator of $F_{\epsilon_d}(e)$ remains the same when the support of ϵ_d is the whole real line. If this is the case, however, we need an additional assumption that $\text{Var}(\epsilon_d) < \infty$ (which holds automatically when ϵ_d has a compact support) so that the asymptotic results of the WNLS estimator would hold. Assumption A.3(ii) implies that $F_{\epsilon_d}(e)$ is strictly increasing on $[e_{d\ell}, e_{du}]$. Assumption A.4(i) requires that all of the covariates are continuous. It is not restrictive since one can easily deal with the case

where X has both continuous and discrete components by sample splitting as in Donald and Hsu (2014). Assumption A.4(ii) regulates the rate at which additional terms are added to the series depending on the dimension of X and the number of derivatives of $p(x)$. See Hirano, Imbens, and Ridder (2003) for more details.

B Influence Function Estimation

The plug-in estimator of the influence function in Theorem 4.1 is given by

$$\begin{aligned}\hat{\psi}_\delta(W_i, x, \tau) = \nabla m_1(x, \hat{\beta}_1)' \hat{\psi}_{\beta_1}(W_i, \hat{\beta}_1) - \nabla m_0(x, \hat{\beta}_0)' \hat{\psi}_{\beta_0}(W_i, \hat{\beta}_0) \\ - \left[\frac{\hat{\psi}_{\epsilon_1}(W_i, \hat{Q}_{\epsilon_1}(\tau))}{\hat{f}_{\epsilon_1}(\hat{Q}_{\epsilon_1}(\tau))} - \frac{\hat{\psi}_{\epsilon_0}(W_i, \hat{Q}_{\epsilon_0}(\tau))}{\hat{f}_{\epsilon_0}(\hat{Q}_{\epsilon_0}(\tau))} \right],\end{aligned}$$

where $\hat{\beta}_d$ is from (3.1), $\hat{Q}_{\epsilon_d}(\tau) = \inf\{e : \hat{F}_{\epsilon_d}(e) \geq \tau\}$ with $\hat{F}_{\epsilon_d}(e)$ from (3.2), and

$$\begin{aligned}\hat{\psi}_{\beta_d}(W_i, \beta_d) = \left[\frac{1}{n} \sum_{i=1}^n \frac{1\{D_i = d\} \nabla m_d(X_i, \beta_d) \nabla m_d(X_i, \beta_d)'}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}} \right]^{-1} \cdot \\ \frac{1\{D_i = d\} \nabla m_d(X_i, \beta_d)[Y_i - m_d(X_i, \beta_d)]}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}},\end{aligned}$$

$$\begin{aligned}\hat{\psi}_{\epsilon_d}(W_i, e) = \frac{1\{D_i = d\} 1\{Y_i - m_d(X_i, \hat{\beta}_d) \leq e\}}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}} - \hat{F}_{\epsilon_d}(e) \\ + \left\{ 1 - \frac{1\{D_i = d\}}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}} \right\} \hat{F}_{\epsilon_d|X}(e|X_i) \\ + \hat{f}_{\epsilon_d}(e) \left[\frac{1}{n} \sum_{i=1}^n \nabla m_d(X_i, \hat{\beta}_d) \right]' \hat{\psi}_{\beta_d}(W_i, \hat{\beta}_d),\end{aligned}$$

where $\hat{F}_{\epsilon_d|X}(e|x)$ and $\hat{f}_{\epsilon_d}(e)$ are uniformly consistent estimators given below.

B.1 Sieve Estimator of $F_{\epsilon_d|X}(e|x)$

We construct the sieve estimator of $F_{\epsilon_d|X}(e|x)$ similar to Donald and Hsu (2014). Note that the estimator must satisfy three requirements: (i) bounded between 0 and 1; (ii) monotonically increasing in e for any given x; (iii) converging uniformly in probability to $F_{\epsilon_d}(e|x)$ in both e and x. The estimator plays an important role in the multiplier bootstrap method. We first denote $\tilde{F}_{\epsilon_d|X}(e|x)$ as

$$\left\{ \sum_{i=1}^n \frac{1\{D_i = d\} 1\{Y_i - m_d(X_i, \hat{\beta}_d) \leq e\}}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}} R^K(X_i) \right\}' \left\{ \sum_{i=1}^n R^K(X_i) R^K(X_i)' \right\}^{-1} R^K(x),$$

where $R^K(x)$ is the same power series used in the SLE estimator of $p(x)$. Note that $\tilde{F}_{\epsilon_d|X}(e|x)$ is a step function in e with jumps at $Y_i - m_d(X_i, \hat{\beta}_d)$'s for any given x, and will converge in probability to $F_{\epsilon_d|X}(e|x)$ in both arguments e and x. However, $\tilde{F}_{\epsilon_d|X}(e|x)$ is not necessarily

bounded between 0 and 1 nor monotonically increasing in e for any given x. Hence we introduce a modification as follows.

Let $\varepsilon_i = Y_i - m_d(X_i, \hat{\beta}_d)$. Without loss of generality assume that $\mathcal{E} = [0, \overline{e}]$ and there are no ties between ε_i's. We add $\varepsilon_{(0)} = 0$ and $\varepsilon_{(n+1)} = \overline{e}$. Let $\varepsilon_{(i)}$ denote the i-th smallest element among the ε_i's so that we have $0 = \varepsilon_{(0)} < \varepsilon_{(1)} < \cdots < \varepsilon_{(n)} < \varepsilon_{(n+1)} = \overline{e}$. We then define $\hat{F}_{\epsilon_d|X}(e|x)$ by induction. First, define $\hat{F}_{\epsilon_d|X}(e|x) = \tilde{F}_{\epsilon_d|X}(e|x) = 0$ for $\varepsilon_{(0)} \leq e < \varepsilon_{(1)}$ and $\hat{F}_{\epsilon_d|X}(\varepsilon_{(n+1)}|x) = 1$. Next, suppose $\hat{F}_{\epsilon_d|X}(e|x) = 0$ is already defined for $\varepsilon_{(0)} \leq e < \varepsilon_{(i)}$, we then define for $\varepsilon_{(i)} \leq e < \varepsilon_{(i+1)}$,

$$\begin{aligned}\hat{F}_{\epsilon_d|X}(e|x) = \hat{F}_{\epsilon_d|X}(\varepsilon_{(i-1)}|x) \cdot 1\{0 \leq \tilde{F}_{\epsilon_d|X}(\varepsilon_{(i)}|x) \leq \hat{F}_{\epsilon_d|X}(\varepsilon_{(i-1)}|x)\} \\ + \tilde{F}_{\epsilon_d|X}(\varepsilon_{(i)}|x) \cdot 1\{\hat{F}_{\epsilon_d|X}(\varepsilon_{(i-1)}|x) < \tilde{F}_{\epsilon_d|X}(\varepsilon_{(i)}|x) \leq 1\} + 1\{\tilde{F}_{\epsilon_d|X}(\varepsilon_{(i)}|x) > 1\}.\end{aligned}$$

The idea is that if $\tilde{F}_{\epsilon_d|X}(e|x)$ jumps down at $\varepsilon_{(i)}$, then we set $\hat{F}_{\epsilon_d|X}(e|x) = \hat{F}_{\epsilon_d}(\varepsilon_{(i-1)}|x)$ for $\varepsilon_{(i)} \leq e < \varepsilon_{(i+1)}$. At the same time, we force $\hat{F}_{\epsilon_d|X}(e|x)$ to lie between 0 and 1 by defining $\hat{F}_{\epsilon_d|X}(e|x) = 0$ when $\tilde{F}_{\epsilon_d|X}(e|x) < 0$ and defining $\hat{F}_{\epsilon_d|X}(e|x) = 1$ when $\tilde{F}_{\epsilon_d|X}(e|x) > 1$. It is easy to see that for any given x, $\hat{F}_{\epsilon_d|X}(e|x)$ is bounded between 0 and 1 and is monotonically increasing in e. More importantly, one can show that under regularity conditions, $\hat{F}_{\epsilon_d|X}(e|x)$ is uniformly consistent over $\mathcal{E}$ and $\mathcal{X}$:

$$\sup_{e \in \mathcal{E}, x \in \mathcal{X}} \left| \hat{F}_{\epsilon_d|X}(e|x) - F_{\epsilon_d|X}(e|x) \right| = o_p(1).$$

The above result follows from the fact that $\sup_{e \in \mathcal{E}, x \in \mathcal{X}} |\tilde{F}_{\epsilon_d|X}(e|x) - F_{\epsilon_d|X}(e|x)| = o_p(1)$ and $\sup_{x \in \mathcal{X}} |\hat{F}_{\epsilon_d|X}(e|x) - F_{\epsilon_d|X}(e|x)| \leq \sup_{x \in \mathcal{X}} |\tilde{F}_{\epsilon_d|X}(e|x) - F_{\epsilon_d|X}(e|x)| = o_p(1)$ for all $x \in \mathcal{X}$. Note that the compactness of $\mathcal{X}$ in Assumption A.1 is needed to obtain the uniform result. Alternatively, one can use the kernel method to estimate $F_{\epsilon_d|X}(e|x)$ instead of the sieve estimator. The multiplier bootstrap remains valid provided that the kernel estimator satisfies the three requirements mentioned above.

B.2 Kernel-Based Estimator of $f_{\epsilon_d}(e)$

In addition to $F_{\epsilon_d|X}(e|x)$, we need to estimate $f_{\epsilon_d}(e)$ before approximating the limiting process $\Delta(x, \tau)$. We introduce the IPW kernel estimator of $f_{\epsilon_d}(e)$ which is uniformly consistent over $\mathcal{E}$. Let $h = h_n$ denote a bandwidth which depends on the sample size n and $K(\cdot)$ a kernel function. For $e \in [e_\ell + h, e_u - h]$, define $\tilde{f}_{\epsilon_d}(e)$ as

$$\tilde{f}_{\epsilon_d}(e) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1\{D_i = d\}}{\hat{p}(X_i)^d [1 - \hat{p}(X_i)]^{1-d}} K\left(\frac{Y_i - m_d(X_i, \hat{\beta}_d) - e}{h} \right).$$

For all $e \in \mathcal{E}$, the estimator of $f_{\epsilon_d}(e)$ is given by

$$\hat{f}_{\epsilon_d}(e) = \begin{cases} \tilde{f}_{\epsilon_d}(e_\ell + h) & \text{if} \quad e \in [e_\ell, e_\ell + h), \\ \tilde{f}_{\epsilon_d}(e) & \text{if} \quad e \in [e_\ell + h, e_u - h], \\ \tilde{f}_{\epsilon_d}(e_u - h) & \text{if} \quad e \in (e_u - h, e_u]. \end{cases}$$

The reason to use $\hat{f}_{\epsilon_d}(e)$ instead of $\tilde{f}_{\epsilon_d}(e)$ is because $\tilde{f}_{\epsilon_d}(e)$ is in general inconsistent around the boundary points e_ℓ and e_u. Therefore, we modify $\tilde{f}_{\epsilon_d}(e)$ around the boundary to obtain uniform consistency. This method is also used in Donald, Hsu, and Barrett (2012) and Donald and Hsu (2014). Under suitable conditions, it can be shown that $\hat{f}_{\epsilon_d}(e)$ is uniformly consistent over $\mathcal{E}$:

$$\sup_{e \in \mathcal{E}} \left| \hat{f}_{\epsilon_d}(e) - f_{\epsilon_d}(e) \right| = o_p(1).$$

C Proofs

Proof of Lemma 2.1

For the DSF,

$$G_d(x, y) = E_{\epsilon_d}[1\{m_d(x, \beta_d) + \epsilon_d \leq y\}] = E_{\epsilon_d}[1\{\epsilon_d \leq y - m_d(x, \beta_d)\}] = F_{\epsilon_d}(y - m_d(x, \beta_d)),$$

where the expectation is taken with respect to the unconditional distribution of ϵ_d. The QSF follows by

$$\begin{aligned} q_d(x, \tau) &= \inf\{y : G_d(x, y) \geq \tau\} \\ &= \inf\{y : F_{\epsilon_d}(y - m_d(x, \beta_d)) \geq \tau\} \\ &= \inf\{m_d(x, \beta_d) + y - m_d(x, \beta_d) : F_{\epsilon_d}(y - m_d(x, \beta_d)) \geq \tau\} \\ &= \inf\{m_d(x, \beta_d) + z : F_{\epsilon_d}(z) \geq \tau\} \\ &= m_d(x, \beta_d) + \inf\{z : F_{\epsilon_d}(z) \geq \tau\} \\ &= m_d(x, \beta_d) + Q_{\epsilon_d}(\tau), \end{aligned}$$

where the first equality holds by definition, the second holds by the fact that $G_d(x, y) = F_{\epsilon_d}(y - m_d(x, \beta_d))$, and the third holds by rewriting y. The fourth equality holds by changing $y - m_d(x, \beta_d)$ to z and the fifth equality holds because $m_d(x, \beta_d)$ is a constant in the infimum operator. Finally, the last equality holds again by definition. □

Proof of Lemma 3.1

We show the identification of β_d and $F_{\epsilon_d}(e)$ which is sufficient for the parameter of interest. For β_d, we know from Assumption 3.2(ii) that β_d uniquely solves the population problem

$$\min_{b_d \in \mathcal{B}_d} E[Y_d - m_d(X, b_d)]^2.$$

By law of iterated expectations,

$$\begin{aligned}
&E\left[\frac{1\{D=d\}[Y - m_d(X,b_d)]^2}{p(X)^d[1-p(X)]^{1-d}}\right] \\
&= E\left\{E\left[\frac{1\{D=d\}[Y - m_d(X,b_d)]^2}{p(X)^d[1-p(X)]^{1-d}}\middle| X\right]\right\} \\
&= E\left\{\frac{1}{p(X)^d[1-p(X)]^{1-d}} E\left[1\{D=d\}[Y - m_d(X,b_d)]^2 | X, D=d\right] \cdot \Pr(D=d|X)\right\} \\
&= E\{E\{[Y_d - m_d(X,b_d)]^2 | X, D=d\}\} \\
&= E\{E\{[Y_d - m_d(X,b_d)]^2 | X\}\} \\
&= E\{[Y_d - m_d(X,b_d)]^2\},
\end{aligned}$$

where the second equality holds by expanding the conditional expectation according to D, the third equality comes from $\Pr(D=d|X) = p(X)^d[1-p(X)]^{1-d}$ and $Y = Y_d$ when $D = d$. By Assumption 3.1(i), the fourth equality holds. Then by law of iterated expectations again the fifth equality holds. Since D, X, Y are all observable, β_d is then identified. For $F_{\epsilon_d}(e)$, the result follows immediately by replacing $[Y - m_d(X, b_d)]^2$ above with $1\{Y - m_d(X, \beta_d) \leq e\}$. □

Proof of Lemma 4.1

We show the influence function representation of $\hat{\beta}_d$ here. The proof for that of $\hat{F}_{\epsilon_d}(e)$ is a combination of Theorem 3.3 of Donald, Hsu, and Barrett (2012) and Theorem 3.6 of Donald and Hsu (2014) so is omitted. By a mean-value expansion about β_d in the first-order condition for $\hat{\beta}_d$ in (3.1), it is true that

$$\sqrt{n}(\hat{\beta}_d - \beta_d) = -E[H(\beta_d, p(X))]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\beta_d, \hat{p}(X_i))\right] + o_p(1),$$

where the score $s(\beta_d, p(X))$ and Hessian $H(\beta_d, p(X))$ are defined as

$$\begin{aligned}
s(\beta_d, p(X)) &= -\frac{1\{D=d\}\nabla m_d(X,\beta_d)[Y - m_d(X,\beta_d)]}{p(X)^d[1-p(X)]^{1-d}}, \\
H(\beta_d, p(X)) &= -\frac{1\{D=d\}\nabla^2 m_d(X,\beta_d)[Y - m_d(X,\beta_d)]}{p(X)^d[1-p(X)]^{1-d}} + \frac{1\{D=d\}\nabla m_d(X,\beta_d)\nabla m_d(X,\beta_d)'}{p(X)^d[1-p(X)]^{1-d}}.
\end{aligned}$$

Similar to the proof of Lemma 3.1, one can show that $E[s(\beta_d, p(X))] = 0$ and

$$E[H(\beta_d, p(X))] = E\left[\frac{1\{D=d\}\nabla m_d(X,\beta_d)\nabla m_d(X,\beta_d)'}{p(X)^d[1-p(X)]^{1-d}}\right].$$

Next, by replacing Y_i's with $-\nabla m_d(X_i,\beta_d)\big[Y_i - m_d(X_i,\beta_d)\big]$'s in the addendum of Hirano, Imbens, and Ridder (2003), it is true that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n s(\beta_d,\hat{p}(X_i)) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{-\frac{1\{D_i=d\}\nabla m_d(X_i,\beta_d)[Y_i - m_d(X_i,\beta_d)]}{p(X_i)^d[1-p(X_i)]^{1-d}} + \left[1 - \frac{1\{D_i=d\}}{p(X_i)^d[1-p(X_i)]^{1-d}}\right] E[\nabla m_d(X,\beta_d)[Y_d - m_d(X,\beta_d)]|X=X_i]\right\}\right| = o_p(1),$$

where the last term of the left-hand side is zero by Assumption 3.2(i). Thus,

$$\begin{aligned}&\left|\sqrt{n}(\hat{\beta}_d - \beta_d) + \frac{1}{\sqrt{n}}\sum_{i=1}^n E[H(\beta_d,p(X))]^{-1}s(\beta_d,\hat{p}(X_i))\right| \\ &= \left|\sqrt{n}(\hat{\beta}_d - \beta_d) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{E\left[\frac{1\{D=d\}\nabla m_d(X,\beta_d)\nabla m_d(X,\beta_d)'}{p(X)^d[1-p(X)]^{1-d}}\right]^{-1}\cdot \right.\right. \\ &\qquad\left.\left.\frac{1\{D_i=d\}\nabla m_d(X_i,\beta_d)[Y_i - m_d(X_i,\beta_d)]}{p(X_i)^d[1-p(X_i)]^{1-d}}\right\}\right| = o_p(1).\end{aligned}$$

That is, $\hat{\beta}_d$ is asymptotically linear with influence function

$$E\left[\frac{1\{D=d\}\nabla m_d(X,\beta_d)\nabla m_d(X,\beta_d)'}{p(X)^d[1-p(X)]^{1-d}}\right]^{-1}\frac{1\{D=d\}\nabla m_d(X,\beta_d)[Y - m_d(X,\beta_d)]}{p(X)^d[1-p(X)]^{1-d}}. \qquad \square$$

Proof of Theorem 4.1

Given (3.3) and Lemma 4.1, the first part of the influence function $\psi_\delta(W,x,\tau)$,

$$\nabla m_1(x,\beta_1)'\psi_{\beta_1}(W,\beta_1) - \nabla m_0(x,\beta_0)'\psi_{\beta_0}(W,\beta_0),$$

can be derived directly from the delta method. The second part of $\psi_\delta(W,x,\tau)$,

$$-\left[\frac{\psi_{\epsilon_1}(W,Q_{\epsilon_1}(\tau))}{f_{\epsilon_1}(Q_{\epsilon_1}(\tau))} - \frac{\psi_{\epsilon_0}(W,Q_{\epsilon_0}(\tau))}{f_{\epsilon_0}(Q_{\epsilon_0}(\tau))}\right],$$

can also be obtained by the functional delta method since the quantile map is Hadamard differentiable. Since $p(x)$ is bounded away from 0 and 1 by Assumption 3.1(ii), $\nabla m_d(X,\beta_d)$ is square integrable by Assumption 3.2(iv), and $f_{\epsilon_d}(e)$ is bounded away from 0 by Assumption A.3(i), $\{\psi_\delta(W,x,\tau) : \tau \in [0,1]\}$ is a combination of Type I and II functions defined by Andrews (1994) which forms a Donsker class of functions. Thus, Theorem 4.1 follows immediately from the functional central limit theorem. A similar proof can be found in Theorem 3.6 of Donald and Hsu (2014). $\square$

References

Abrevaya, J., Y.-C. Hsu and R. P. Lieli (2015): "Estimating Conditional Average Treatment Effects," *Journal of Business and Economic Statistics*, **33**, 485–505.

Andrews, D. W. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, **4**, 2247–2294.

Auld, M. C. (2005): "Smoking, Drinking, and Income," *Journal of Human Resources*, **40**, 505–518.

Blundell, R. and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," *Advances in Economics and Econometrics*, **2**, 312–357.

Blundell, R. and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *The Review of Economic Studies*, **71**, 655-679.

Brinch, C. N., M. Mogstad and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, **125**, 985–1039.

Chernozhukov, V., I. Fernández-Val and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, **81**, 2205–2268.

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli and F. Vella (2018): "Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models," *Working Paper*.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, **73**, 245–261.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, **96**, 187–199.

Donald, S. G. and Y.-C. Hsu (2014): "Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models," *Journal of Econometrics*, **178**, 383–397.

Donald, S. G., Y.-C. Hsu and G. F. Barrett (2012): "Incorporating Covariates in the Measurement of Welfare and Inequality: Methods and Applications," *The Econometrics Journal*, **15**, C1–C30.

Donald, S. G., Y.-C. Hsu and R. P. Lieli (2014): "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business and Economic Statistics*, **32**, 395–415.

Dong, Y. and S. Shen (2018): "Testing for Rank Invariance or Similarity in Program Evaluation," *Review of Economics and Statistics*, **100**, 78–85.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, **75**, 259–276.

Fortin, N., T. Lemieux and S. Firpo (2011): "Decomposition Methods in Economics," in D. Card and O. Ashenfelter (Eds.), *Handbook of Labor Economics*, **4**, 1–102. Elsevier.

Frandsen, B. R. and L. J. Lefgren (2018): "Testing Rank Similarity," *Review of Economics and Statistics*, **100**, 86–91.

French, M. T. and I. Popovici (2011): "That Instrument is Lousy! In Search of Agreement When Using Instrumental Variables Estimation in Substance Use Research," *Health Economics*, **20**, 127–146.

Grafova, I. B. and F. P. Stafford (2009): "The Wage Effects of Personal Smoking History," *Industrial and Labor Relations Review*, **62**, 381–393.

Hirano, K., G. Imbens and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, **71**, 1161–1189.

Ichimura, H. and O. Linton (2005): "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," in D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models*, 149–170. Cambridge University Press.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," *Econometrica*, **77**, 1481–1512.

Imbens, G. W. and J. W. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, **47**, 5–86.

Levine, P. B., T. A. Gustafson and A. D. Velenchik (1997): "More Bad News for Smokers? The Effects of Cigarette Smoking on Wages," *Industrial and Labor Relations Review*, **50**, 493–509.

Orzechowski, W. and R. Walker (2019): *The Tax Burden on Tobacco, 1970-2018*.

Rosenbaum, P. and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, **70**, 41–55.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, **66**, 688.

Van Ours, J. C. (2004): "A Pint a Day Raises a Man's Pay; But Smoking Blows that Gain Away," *Journal of Health Economics*, **23**, 863–886.

Wooldridge, J. M. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, **141**, 1281–1301.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition. The MIT Press.

Wooldridge, J. M. (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, **50**, 420–445.

Table 1: Descriptive statistics.

	Male			Female		
	Smoker	Non-smoker	Difference	Smoker	Non-smoker	Difference
Average hourly	19.575	32.044	−12.469	14.865	20.800	−5.935
wage	(13.838)	(35.321)	[−14.964]	(9.755)	(14.122)	[−7.590]
Ever smoked	0.937	0.260	0.677	0.872	0.201	0.671
before 25	(0.244)	(0.438)	[54.731]	(0.335)	(0.401)	[26.181]
Age	39.438	42.258	−2.819	41.115	41.081	0.034
	(10.328)	(10.563)	[−6.300]	(10.085)	(10.999)	[0.045]
Years of	12.446	14.043	−1.597	12.775	14.098	−1.323
education	(2.348)	(2.507)	[−15.579]	(1.806)	(2.277)	[−9.472]
Race (white=1)	0.585	0.658	−0.073	0.431	0.362	0.069
	(0.493)	(0.475)	[−3.439]	(0.496)	(0.481)	[1.896]
White-collar	0.157	0.377	−0.220	0.179	0.372	−0.193
	(0.364)	(0.485)	[−13.116]	(0.384)	(0.483)	[−6.490]
Blue-collar	0.540	0.355	0.185	0.183	0.098	0.085
	(0.499)	(0.479)	[8.650]	(0.388)	(0.298)	[3.079]
Union	0.110	0.142	−0.032	0.096	0.153	−0.057
	(0.313)	(0.349)	[−2.350]	(0.296)	(0.361)	[−2.514]
Cigarette price	6.104	6.235	−0.131	5.978	6.169	−0.191
	(1.213)	(1.323)	[−2.462]	(1.229)	(1.338)	[−2.068]
Religiosity	2.676	2.716	−0.04	2.945	3.181	−0.236
	(8.051)	(7.533)	[−0.116]	(6.756)	(8.078)	[−0.458]
Sample size	648	3,128	3,776	218	1,141	1,359

Notes: The table reports means and standard deviations (in parentheses) for respondents between 25 and 65 years old, working at least 1,500 hours a year. The columns showing differences in means (by smoking status) report the t-statistics (in brackets) for the null hypothesis of equality in means. See text for the definition of the variables.

Table 2: Results for the unconfoundedness test.

		Specification			
Group	Sample size	Constant	Linear	Interaction	Quadratic
$X = \{$age, education, race, occupational status, union$\}$					
Male	3,735	0.000	0.003	0.006	0.417
Female	1,331	0.067	0.140	0.081	0.091
$X = \{$age, education, race, occupational status, union, price, religiosity$\}$					
Male	3,735	0.000	0.002	0.034	0.146
Female	1,331	0.067	0.184	0.592	0.563

Notes: The table reports p-values for the null hypothesis of unconfoundedness using the test proposed by Donald, Hsu, and Lieli (2014) and whether smoked before 25 as a binary instrument following van Ours (2004). Different specifications indicate the inclusion of constant, linear, interaction, and quadratic terms of covariates as power series in estimating the instrument propensity score. See Donald, Hsu, and Lieli (2014) for more details.

Table 3: Fixed covariate values.

Dummy variable	Fixed value (most common combination)	Frequency (%)
Race (white=1)	0	434/1331 (32.52%)
White-collar	0	
Blue-collar	0	
Union	0	
Continuous variable	Fixed value (group mean)	Overall mean
Age	40.673	40.979
Years of education	13.180	13.903
Cigarette price	5.891	6.141
Religiosity	3.293	3.134

Notes: For dummy variables, the fixed values are assigned to the most frequently occurring combination in the dataset. For continuous variables, the fixed values are assigned to the means of the corresponding variables in the most frequently occurring combination group.

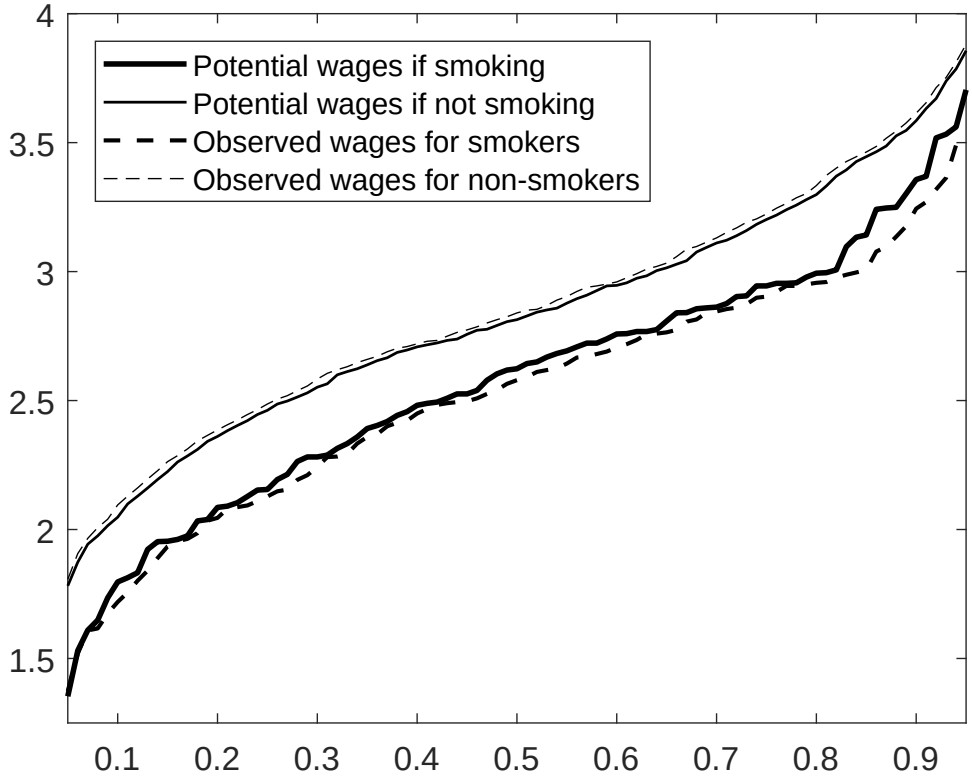

Figure 1: Quantile functions of potential and observed wages for females.

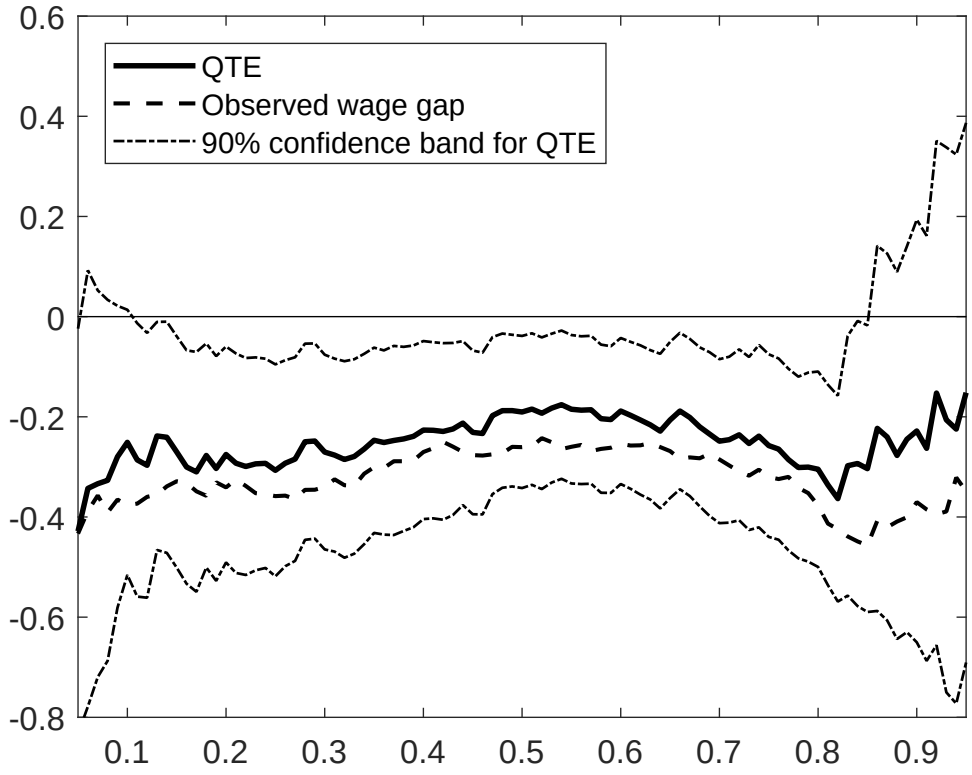

Figure 2: QTE of smoking on wages for females.

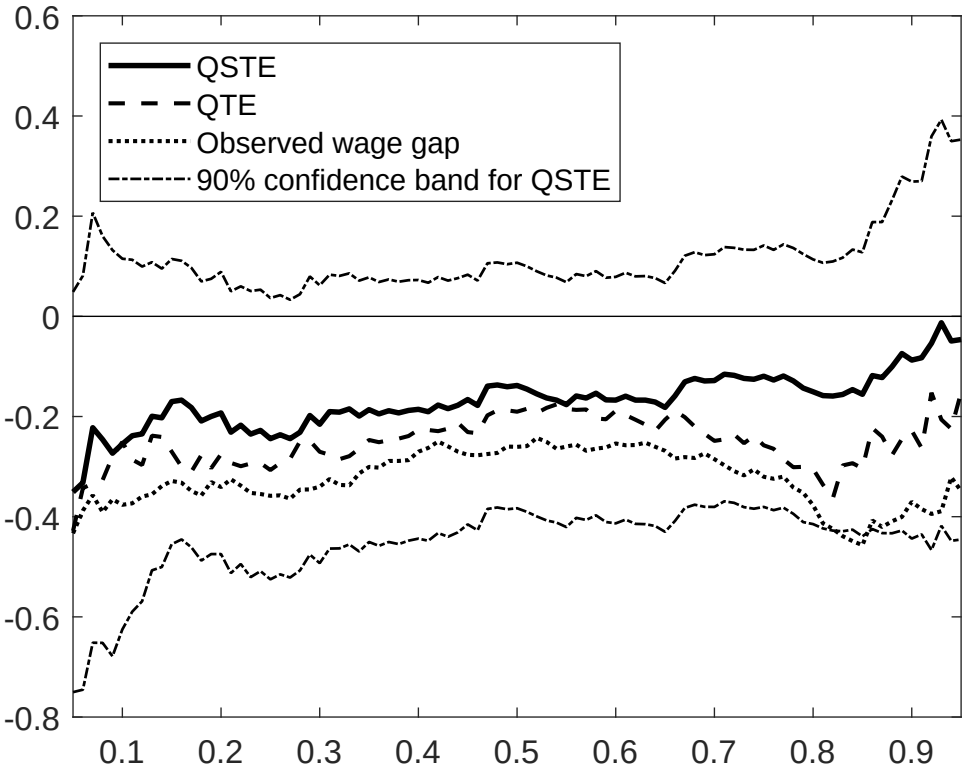

Figure 3: QSTE of smoking on wages for females.