ESTIMATION OF TREATMENT EFFECTS UNDER ENDOGENOUS HETEROSKEDASTICITY*

JASON ABREVAYA† AND HAIQING XU‡

ABSTRACT. This paper considers a treatment effects model in which individual treatment effects may be heterogeneous, even among observationally identical individuals. Specifically, by extending the classical instrumental-variables (IV) model with an endogenous binary treatment, the heteroskedasticity of the error disturbance is allowed to depend upon the treatment variable so that treatment generates both mean and variance effects on the outcome. In this endogenous heteroskedasticity IV (EHIV) model, the standard IV estimator can be inconsistent for the average treatment effects (ATE) and lead to incorrect inference. After nonparametric identification is established, closed-form estimators are provided for the linear EHIV of the mean and variance treatment effects, and the average treatment effect on the treated (ATT). Asymptotic properties of the estimators are derived. A Monte Carlo simulation investigates the performance of the proposed approach. An empirical application regarding the effects of fertility on female labor supply is considered, and the findings demonstrate the importance of accounting for endogenous heteroskedasticity.

Keywords: Endogenous heteroskedasticity, individual treatment effects, average treatment effects, local average treatment effects, instrumental variable

Date: Friday 23rd August, 2019.

*We thank Daniel Ackerberg, Sandra Black, Ivan Canay, Salvador Navarro, Max Stinchcombe, Quang Vuong, and Ed Vytlacil for useful comments. We also thank seminar participants at University of Iowa, University of Hong Kong, McMaster University, Western University, University of Texas at Austin, Xiamen University, Monash University, University of Melbourne, USC, the 2017 Shanghai workshop of econometrics at SUFE, the 2018 Texas Econometrics Camp, and the 2018 CEME conference at Duke University.

†Department of Economics, University of Texas at Austin, Austin, TX, 78712, abrevaya@austin.utexas.edu.

‡Department of Economics, University of Texas at Austin, Austin, TX, 78712, h.xu@austin.utexas.edu.

1. INTRODUCTION

When treatment effects are heterogeneous among observationally identical individuals, the causal inference for policy evaluation is considerably difficult (see e.g. Heckman and Vytlacil, 2005). The seminal paper by Imbens and Angrist (1994) shows that the linear IV estimates should be interpreted as the local average treatment effects (LATE), provided the monotonicity assumption on the selection into treatment. In this paper, we propose a simple but new approach to estimate a heterogeneous treatment effects model without making Imbens and Angrist (1994)'s monotone selection assumption.

Specifically, we extend the classical IV model to include both mean and variance effects rather than just mean effects:

$$Y = \mu(D, X) + \sigma(D, X) \times e(X, \nu), \tag{1}$$

where $Y \in \mathbb{R}$ is the outcome variable of interest, $X \in \mathbb{R}^{d_X}$ is a vector of observed covariates, $D \in \{0, 1\}$ denotes the binary treatment status, and $\nu \in \mathbb{R}^{d_\nu}$ is a vector of latent variables. Moreover, $e : \mathbb{R}^{d_X} \times \mathbb{R}^{d_\nu} \to \mathbb{R}_+$ is an unknown function that describes the essential model disturbance. Under an additional normalization assumption that $e(X, \nu)$ has zero mean and unit variance (given X), the structural functions $\mu(\cdot, X)$ and $\sigma(\cdot, X)$ are the mean and standard deviation of the (potential) outcome, respectively, under different treatment statuses. Hence, $\mu(1, X) - \mu(0, X)$ and $\sigma(1, X) - \sigma(0, X)$ measure the (population-level) *mean effects* and *"variance" effects* of the treatment, respectively.

In the above model, the key feature is to use the simple mean–and–variance–effect structure to parsimoniously characterize heterogeneous treatment effects. See e.g. Chesher (2005); Chernozhukov and Hansen (2005) for a more general characterization using fully nonseparable models. The fact that the heteroskedasticity term $\sigma(\cdot, \cdot)$ depends on the endogenous treatment D implies that treatment effects can differ across individuals even after X has been controlled for. As such, we say that model (1) exhibits *endogenous heteroskedasticity*, and we will call our instrumental-variables method the *endogenous heteroskedasticity IV* (or EHIV) approach. As emphasized in Heckman and Vytlacil (2005), the absence of heterogeneous responses to treatment implies that different treatment effects collapse to the same parameter. If $\sigma(D, X)$ depends upon D in (1), however, heterogeneous treatment effects arise in general, and we show that the standard IV approach is generally inconsistent for estimating the (population) mean effects in the presence of endogenous heteroskedasticity.

On the other hand, if the heteroskedasticity is exogenous, the treatment effects are homogeneous across individuals (after covariates have been controlled for), which can be consistently estimated by the standard IV approach. Therefore, to apply the IV method for the mean effects of the treatment, the exogeneity of heteroskedasticity serves as a key assumption, which should be justified from economic theory and/or statistical tests. By using squared IV estimated residuals, we suggest a Fan and Li (1996) type test statistics for exogenous heteroskedasticity (equivalently, the homogeneous treatment effects). If the heteroskedasticity is not exogenous, the standard IV estimator becomes a mixture of the mean and variance effects, interpreted as LATE under Imbens and Angrist (1994)'s monotonicity condition.

This paper builds upon several strands in the existing literature. The literature on heterogeneous treatment effects (e.g. Imbens and Angrist, 1994; Heckman, Smith, and Clements, 1997; Heckman and Vytlacil, 2005, among many others) is an important antecedent. Within the LATE context, Abadie (2002, 2003) has considered the estimation of the variance and the distribution of treatment effects, but the causal interpretation is limited to compliers. The main difference of our approach from that literature is that we consider additional assumptions on the structural outcome model rather than additional assumptions on a selection equation and/or variation of the instrumental variable. Our approach does not restrict causal interpretation to compliers. As far as we know, the only other paper that explicitly considers a structural treatment-effect model with endogenous heteroskedasticity is Chen and Khan (2014). Under the monotone selection assumption, Chen and Khan (2014) focus on identification and estimation of the ratio of the heteroskedasticity term under different treatment statuses, i.e., $\sigma(1,x)/\sigma(0,x)$.

Another important related literature concerns the identification and estimation of nonseparable models with binary endogeneity (e.g. Chesher, 2005; Chernozhukov and Hansen, 2005; Jun, Pinkse, and Xu, 2011, among many others). In particular, Chernozhukov and Hansen (2005) establish nonparametric (local and global) identification of quantile treatment effects under a rank condition. Extending Chernozhukov and Hansen (2005)'s results, Vuong and Xu (2017) develop a constructive identification strategy for the nonseparable structural model by assuming monotonicity of the selection. This paper also derives closed-form identification for the mean and variance effects of the treatment, but the additional assumptions on the structural outcome equation lead to an estimation strategy that should be considerably simpler for practitioners to use.

While identification does not require additional parametric specification of $\mu(D, X)$, we take a semiparametric approach to estimation that imposes linearity of $\mu(D, X)$, in line with nearly all empirical work, and leaves $\sigma(D, X)$ unspecified. This specification allows for heterogeneous individual treatment effects, but it is quite tractable in the sense that the heterogenous individual treatment effects can be decomposed into mean and variance effects. On the other hand, nonparametric estimation of fully nonseparable models is challenging. See e.g. Chernozhukov and Hansen (2004, 2005) and Feng, Vuong, and Xu (2016), who develop nonparametric estimation of quantiles and density functions of individual treatment effects, respectively, in fully nonseparable frameworks.

The structure of the paper is organized as follows. Section 2 formally introduces the notation and assumptions underlying the endogenous heteroskedasticity model in (1), focusing on the case of a binary instrumental variable. Section 3 provides a constructive approach to nonparametric identification of the mean and variance functions in (1). Section 4 considers a semiparametric version of (1) in which the mean function is a linear index of X and D. An estimator (the EHIV estimator) of the coefficient parameters is proposed, and its asymptotic properties ($\sqrt{n}$-consistency and asymptotic normality) are established. Combining this estimator with a nonparametric estimator of the heteroskedasticity function $\sigma(\cdot, \cdot)$ allows us to consistently estimate the (conditional) distribution of the heterogeneous treatment effects. Section 5 provides Monte Carlo evidence to illustrate the performance of the proposed estimator. Section 6 applies the approach to an empirical application, where the effects of having a third child on female labor supply are estimated (as previously considered by Angrist and Evans, 1998). Section 7 concludes. Proofs are collected in the Appendix.

2. Assumptions and Model

To deal with the endogeneity of treatment status, we consider the canonical case in which a binary instrumental variable $Z \in \{0, 1\}$ exists. The case of binary-valued instruments has been emphasized in the treatment effect literature, in particularly in the applications using natural and social experiments. For each $(x, z) \in \mathscr{S}_{XZ}$, let $p(x, z) = \mathbb{P}(D = 1|X = x, Z = z)$ denote the propensity score. Let further $\epsilon = e(X, \nu)$. The following assumptions are maintained throughout the paper.

Assumption A. *(Normalization) Let* $\mathbb{E}(\epsilon|X) = 0$ *and* $\mathbb{E}(\epsilon^2|X) = 1$.

Assumption B. *(i) (Instrument relevance) For every* $x \in \mathscr{S}_X$, $\mathscr{S}_{Z|X=x} = \{0, 1\}$ *and* $p(x, 0) \neq p(x, 1)$*; (ii) (Instrument exogeneity)* $\mathbb{E}(\epsilon|X, Z) = \mathbb{E}(\epsilon|X)$ *and* $Var(\epsilon|X, Z) = Var(\epsilon|X)$*.*

Assumption A is a normalization on the first two moments of the error term ϵ. Clearly, the scale normalization on $\mathbb{E}(\epsilon^2|X)$ is indispensable for identification of $\sigma(\cdot, \cdot)$. Assumption B contains the instrument relevance and instrument exogeneity conditions. In particular, (ii) is implied by the conditional independence of Z and ϵ given X, i.e., $Z \perp \epsilon|X$, which is usually motivated by the choice of the instrumental variable (see e.g. Angrist and Krueger, 1991). Combining Assumptions A and B(ii), we have $\mathbb{E}(\epsilon|X, Z) = 0$ and $\text{Var}(\epsilon|X, Z) = 1$. For expositional simplicity, we will assume throughout the paper that $p(x, 0) < p(x, 1)$ for all $x \in \mathscr{S}_X$.

Motivated by the fully nonseparable model approach (see e.g. Chesher, 2005; Chernozhukov and Hansen, 2005), our model (1) parsimoniously introduces heterogeneous treatment effects across individuals. In particular, model parameters $\mu(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$, respectively, capture the mean and variance effects of the treatment. Therefore, individual treatment effects can be written as

$$\mu(1, X) - \mu(0, X) + [\sigma(1, X) - \sigma(0, X)] \times \epsilon,$$

which varies across individuals even with the same value of covariates X. Such a semi-nonseparable specification makes our model tractable for estimation and inference.

Vuong and Xu (2017) point out that a key assumption in e.g Chernozhukov and Hansen (2005) is the monotonicity of the outcome equation under which one could define a so-called "counterfactual mapping", i.e.

$$Y_1 = \phi(Y_0; X),$$

where Y_1 and Y_0 are potential outcomes under treatment status $D = 1$ and 0, respectively, and function ϕ is monotone in Y_0 which links the potential outcome Y_0 with its counterfactual Y_1. Vuong and Xu (2017) establish nonparametric identification and estimation of ϕ. In contrast, our model (1) essentially parametrizes such a counterfactual mapping by

$$Y_1 = \phi_0(X) + \phi_1(X) \times Y_0,$$

which significantly simplifies the estimation procedures. More importantly, heterogeneous treatment effects can be interpreted as the mean and variance effects. On the other hand, however, it is certainly an interesting and challenging question to consider a more flexible

specification for counterfactual mapping, e.g.

$$Y_1 = \phi(Y_0, X, \xi)$$

where ξ is an unobserved error term. We leave such a generalization for future research.

2.1. **An economic example for model** (1)**.** For the illustration purpose and also to motivate, we now provide an economic model on the female labor supply and fertility decision for the specification of our econometrics model. In our empirical application, the outcome variable Y of interest is hours worked per week of the mother worked in 1999, the endogenous treatment D is an indicator of having a third child, and the instrument Z is whether the mother's first two children were of the same gender. Exogenous covariates X include mother's education, mother's age at first birth, and age of the first, and second child.

Suppose the mother's utility is given as follows

$$U(C, L, D, Z, X, \eta) = \theta_c \ln C + \theta_\ell \ln L + \theta_d(D, Z, X, \eta)$$

where $C \in \mathbb{R}^+$ is the household consumption, $L \in \mathbb{R}^+$ the leisure measured by hours, $\eta \in \mathbb{R}^{d_\eta}$ is a vector of latent variables that affect mother's utility. Moreover, $\theta_c, \theta_\ell \in \mathbb{R}^+$ and $\theta_d : \{0,1\}^2 \times \mathbb{R}^{d_X} \times \mathbb{R}^{d_\eta} \to \mathbb{R}$ are structural parameters/function. The decision maker will maximize her utility by choosing $C, L, Y \in \mathbb{R}^+$ and $D \in \{0,1\}$, subjecting to both income and time constraints:

$$\begin{aligned} C + q_I(D, X) &\leq w(X, \nu) \times Y; \\ L + Y + q_T(D, X) &\leq T_0, \end{aligned}$$

where $q_I \geq 0$ and $q_T \geq 0$ measures the third-child related money and time expenditure, respectively, T_0 is the total amount of available time; and $w(\cdot, \cdot) > 0$ denotes the hour wage in which ν is the unobserved heterogeneity. To maximize the utility, it is straightforward that both income and time constraints should be binding. Thus, by the first order condition w.r.t. Y, we obtain the following female labor supply function

$$Y = \frac{\theta_c}{\theta_c + \theta_\ell} \times T_0 - \frac{\theta_c}{\theta_c + \theta_\ell} \times q_T(D, X) + \frac{\theta_\ell}{\theta_c + \theta_\ell} \times q_I(D, X) \times w(X, \nu)^{-1}.$$

Note that we assume that the hour wage w should not depend on whether the mother has a third child or not. Furthermore, it can be show that T_0 and q_T cannot be identified separately, and a similar argument also applies to $\frac{\theta_c}{\theta_c+\theta_\ell}$ (resp. $\frac{\theta_\ell}{\theta_c+\theta_\ell}$) and q_T (resp. q_I). Therefore, we specify

the following econometrics model for the female labor supply

$$Y = q_T^*(D, X) + q_I^*(D, X) \times w(X, \nu)^{-1}.$$

In Section 6, we use the 2000 Census data (5-percent public-use microdata sample (PUMS) for an empirical illustration of our method.

On the other hand, if there is unobserved heterogeneity in the money and time expenditures, i.e. $q_I = q_I(D, X, \nu)$ and/or $q_T = q_T(D, X, \nu)$, then the solution of female labor supply would violate our mean-and–variance effect specification. Instead, a fully non-separable model might be more adequate.

2.2. **Inconsistency of IV estimation.** With non-degenerate variance effects, the standard IV estimator is generally inconsistent for estimating the model parameter μ. In particular, a closed-form expression for the bias of the IV estimator can be derived under our model specification. For expositional simplicity, the covariates X are suppressed in the following discussion. Under Assumption B(i), define the quantities r_0 and r_1 as follows

$$r_0 = \mu(0) + [\sigma(1) - \sigma(0)] \times \frac{\mathbb{E}(D\epsilon|Z=0)p(1) - \mathbb{E}(D\epsilon|Z=1)p(0)}{p(1) - p(0)},$$
$$r_1 = \mu(1) - \mu(0) + [\sigma(1) - \sigma(0)] \times \frac{\mathbb{E}(D\epsilon|Z=1) - \mathbb{E}\left(D\epsilon|Z=0\right)}{p(1) - p(0)}.$$

Then, model (1) can be represented by the following linear IV projection:

$$Y = r_0 + r_1 D + \tilde{\epsilon},$$

where $\tilde{\epsilon} \equiv \mu(D) + \sigma(D)\epsilon - r_0 - r_1 D$. By definition, $\tilde{\epsilon}$ measures the discrepancy between the structural model and its linear IV projection, which satisfies $\mathbb{E}(\tilde{\epsilon}|Z) = 0$ under Assumptions A and B. Therefore, the standard IV regression would estimate the coefficient r_1, which is a linear mixture of the mean effect, $\mu(1) - \mu(0)$, and the variance effect, $\sigma(1) - \sigma(0)$.

The seminal paper by Imbens and Angrist (1994) show that the coefficient r_1 from the above linear IV projection has a LATE interpretation. Specifically, suppose that the selection to treatment satisfies the monotonicity condition, e.g.,

$$D = \mathbb{1}[\eta \leq m(Z)], \tag{2}$$

where $\eta \in \mathbb{R}$ is a scalar-valued latent variable and $m(\cdot)$ is a real-valued function with $m(0) < m(1)$.[1] Under this selection assumption, the LATE can be written as

$$r_1 = \mu(1) - \mu(0) + [\sigma(1) - \sigma(0)] \times \mathbb{E}[\epsilon | m(0) < \eta \leq m(1)].$$

The bias term of the LATE, i.e. $[\sigma(1) - \sigma(0)] \times \mathbb{E}[\epsilon | m(0) < \eta \leq m(1)]$, depends on the degree to which heteroskedasticity depends upon treatment, as well as the average error disturbance for the compliers.

2.3. **Test for endogenous heteroskedasticity.** When treatment effects are homogeneous, i.e. the heteroskedasticity is exogenous, the ATE can be estimated by the LATE. Therefore, it can be worthwhile to test the homogeneous treatment effects hypothesis via testing for exogenous heteroskedasticity, i.e. $\mathbb{H}_0 : \sigma(0, \cdot) = \sigma(1, \cdot) = \tilde{\sigma}(\cdot)$ for some $\tilde{\sigma} : \mathbb{R}^{d_X} \to \mathbb{R}_+$. Because the standard IV approach consistently estimates homogeneous treatment effects under the null hypothesis, a direct test can be conducted by determining whether the squared IV estimated residuals depend upon the instrumental variable Z or not. Although the IV estimator may be inconsistent under the alternative hypothesis, we show that such a test is surprisingly consistent.

Lemma 1. *Suppose* (1) *and Assumptions A to C hold. Then* $\sigma(X, 0) = \sigma(X, 1)$ *if and only if*

$$\mathbb{E}(\tilde{\epsilon}^2 | X, Z = 0) = \mathbb{E}(\tilde{\epsilon}^2 | X, Z = 1) \tag{3}$$

where $\tilde{\epsilon} = Y - \tilde{r}_0(X) - \tilde{r}_1(X)D$*, in which* $\tilde{r}_1(X) = Cov(Y, Z|X)/Cov(D, Z|X)$ *and* $\tilde{r}_0(X) = [Cov(YZ, D|X) - Cov(Y, DZ|X)]/Cov(D, Z|X)$*. In addition, suppose the semiparametric model* (9) *holds. Then* $\sigma(X, 0) = \sigma(X, 1)$ *if and only if*

$$\mathbb{E}\big[(Y - X'\tilde{\beta}_1 - \tilde{\beta}_2 D)^2 | X, Z = 0\big] = \mathbb{E}\big[(Y - X'\tilde{\beta}_1 - \tilde{\beta}_2 D)^2 | X, Z = 1\big]$$

where $\tilde{\beta} = (\tilde{\beta}_1', \tilde{\beta}_2)'$ *satisfies* $\mathbb{E}(Y - X'\tilde{\beta}_1 - \tilde{\beta}_2 D | X, Z) = 0$*.*

In Lemma 1, note that $\tilde{\epsilon}$ is the residual from the nonparametric IV regression, and $\tilde{\beta}$ could be estimated by the usual IV approach.

The model restriction (3) can be equivalently written as

$$\mathbb{E}[\Psi(Y, X, D, Z) | X] = 0,$$

[1] See Vytlacil (2002) for a proof of the observational equivalence between (2) and the monotone selection condition.

where $\Psi(Y,X,D,Z) \equiv \tilde{\epsilon}^2 \times [\mathbb{P}(Z=1|X) - Z] \times f(X)$. Under $\mathbb{H}_0$, suppose we obtain ($\sqrt{n}$)–consistent residuals $\{\tilde{\epsilon}_i : i = 1, \cdots, n\}$. Following Fan and Li (1996), we suggest the following test statistics:

$$T_n = \frac{1}{n(n-1)(n-2)h^{3d_X}} \sum_{i \neq j \neq k_1 \neq k_2} \tilde{\epsilon}_i^2 \tilde{\epsilon}_j^2 (Z_{k_1} - Z_i)(Z_{k_2} - Z_j) K(\frac{X_{k_1} - X_i}{h}) K(\frac{X_{k_2} - X_j}{h}) K(\frac{X_i - X_j}{h}).$$

Under regularity conditions and proper assumptions on the bandwidth h and kernel function K, the asymptotic behavior of T_n has been established in e.g. Fan and Li (1996) and Zheng (1996). Namely, under H_0, higher order terms in the Hoeffding decomposition of the above U-statistics determine its limiting distribution:

$$nh^{d_X/2} \times T_n \to N(0, V)$$

where $V = 2\mathbb{E}[\Psi^2(Y,D,X,Z) f_X(X)] \times \int^{\mathbb{R}^{d_X}} K^2(u) du$ can be consistently estimated by its sample analog. Under alternative hypothesis, it can be shown that T_n a $\sqrt{n}$–consistent estimator of $\mathbb{E}\{\Psi(Y,D,X,Z) \times \mathbb{E}[\Psi(Y,D,X,Z)|X] \times f_X(X)\}$, a non-zero constant under misspecification.

3. Nonparametric Identification

In this section, we provide a constructive identification that involves two steps. First, we identify $\sigma(\cdot, X)$ up-to-scale. Second, we transform (1) into a model with exogenous heteroskedasticity, from which both $\mu(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ are identified.

Some additional notation is required. For $d = 0, 1$, let

$$\delta_d(X) = \frac{\mathbb{E}[Y \times \mathbb{1}(D=d)|X, Z=1] - \mathbb{E}[Y \times \mathbb{1}(D=d)|X, Z=0]}{\mathbb{P}(D=d|X,Z=1) - \mathbb{P}(D=d|X,Z=0)}; \tag{4}$$

$$V_d(X) = \frac{\mathbb{E}[Y^2 \times \mathbb{1}(D=d)|X, Z=1] - \mathbb{E}[Y^2 \times \mathbb{1}(D=d)|X, Z=0]}{\mathbb{P}(D=d|X,Z=1) - \mathbb{P}(D=d|X,Z=0)} - \delta_d^2(X). \tag{5}$$

Under Assumption B(i), both $\delta_d(X)$ and $V_d(X)$ are well defined. Similarly to Imbens and Angrist (1994), $\delta_d(X)$ and $V_d(X)$ can be written in terms of covariances of the observables:

$$\delta_d(X) = \frac{\mathrm{Cov}\big(Y \times \mathbb{1}(D=d), Z|X\big)}{\mathrm{Cov}(\mathbb{1}(D=d), Z|X)};$$

$$V_d(X) = \frac{\mathrm{Cov}\big(Y^2 \times \mathbb{1}(D=d), Z|X\big)}{\mathrm{Cov}(\mathbb{1}(D=d), Z|X)} - \delta_d^2(X).$$

Note that both $\delta(\cdot)$ and $V_d(\cdot)$ are identified from the data.

Moreover, for $\ell = 1, 2$, denote

$$\xi_\ell(x) = \frac{\mathbb{E}(\epsilon^\ell \times D|X = x, Z = 1) - \mathbb{E}(\epsilon^\ell \times D|X = x, Z = 0)}{p(x, 1) - p(x, 0)}.$$

By definition, $\xi_\ell(x)$ depends on the (unknown) distribution of $F_{\epsilon D|XZ}$. Then, model (1) and Assumption A imply

$$\delta_d(X) = \mu(d, X) + \sigma(d, X) \times \xi_1(X),$$
$$V_d(X) = \sigma^2(d, X) \times \left[\xi_2(X) - \xi_1^2(X)\right].$$

Let $C(X) = \xi_2(X) - \xi_1^2(X)$. Thus, the vector $(V_0(X), V_1(X))'$ identifies the heterogeneity component $\sigma(\cdot, X)$ up to the scale $C(X)$. The above discussion is summarized by the following lemma.

Lemma 2. *Suppose Assumptions A and B hold. Then*

$$V_d(X) = \sigma^2(d, X) \times C(X), \text{ for } d = 0, 1.$$

Lemma 2 implies that $\text{sign}(V_0(X)) = \text{sign}(V_1(X))$, which is a testable model restriction. As a matter of fact, Lemma 2 provides a basis for the identification of our model. Before proceeding, however, an assumption ruling out zero-valued variances is needed:

Assumption C. $C(X) \neq 0$ *almost surely.*

Assumption C is verifiable since $C(X) \neq 0$ if and only if $V_d(X) \neq 0$. Moreover, note that if (2) holds, $C(X)$ is interpreted as the (conditional) variance of ϵ given X and the "complier group". In this case, $C(X) > 0$ if and only if the (conditional) distribution of ϵ is non-degenerate.

Model (1) can now be transformed to deal with the issue of endogenous heteroskedasticity. Defining

$$S = |V_0(X)|^{\frac{1}{2}} \times (1 - D) + |V_1(X)|^{\frac{1}{2}} \times D,$$

one can show that $S = \sigma(D, X) \times |C(X)|^{\frac{1}{2}}$ by Lemma 2. Dividing the original model (1) by S yields the transformed model

$$\frac{Y}{S} = \frac{\mu(D, X)}{S} + \frac{\epsilon}{|C(X)|^{\frac{1}{2}}}, \tag{6}$$

for which Z satisfies the instrument exogeneity condition with the (transformed) error disturbance $\epsilon/|C(X)|^{\frac{1}{2}}$.

Closed-form expressions for $\mu(\cdot, x)$ and $\sigma(\cdot, x)$ are now provided. Fixing $x \in \mathscr{S}_X$, note that

$$\mathbb{E}\Big(\frac{Y}{S}\Big|X = x, Z = z\Big) = \frac{\mu(1,x)}{|V_1(x)|^{\frac{1}{2}}} \times p(x,z) + \frac{\mu(0,x)}{|V_0(x)|^{\frac{1}{2}}} \times [1 - p(x,z)], \text{ for } z = 0, 1,$$

which is a linear equation system in $\mu(0, x)$ and $\mu(1, x)$. Assumption B implies

$$\mu(1,x) = \frac{\mathbb{E}\big(\frac{Y}{S}\big|X = x, Z = 1\big)[1 - p(x,0)] - \mathbb{E}\big(\frac{Y}{S}\big|X = x, Z = 0\big)[1 - p(x,1)]}{p(x,1) - p(x,0)} \times |V_1(x)|^{\frac{1}{2}}; \quad (7)$$

$$\mu(0,x) = \frac{\mathbb{E}\big(\frac{Y}{S}\big|X = x, Z = 1\big)p(x,0) - \mathbb{E}\big(\frac{Y}{S}\big|X = x, Z = 0\big)p(x,1)}{p(x,0) - p(x,1)} \times |V_0(x)|^{\frac{1}{2}}. \quad (8)$$

Moreover, it is straightforward to show that

$$\sigma^2(d,x) = |V_d(x)| \times \mathbb{E}\left\{\left[\frac{Y - \mu(D,X)}{S}\right]^2 \big|X = x\right\}.$$

which can be equivalently rewritten as

$$\sigma^2(d,x) = \left|\frac{V_d(x)}{V_1(x)}\right| \times \mathbb{E}\big[D(Y - \mu(D,X))^2\big|X = x\big] + \left|\frac{V_d(x)}{V_0(x)}\right| \times \mathbb{E}\big[(1-D)(Y - \mu(D,X))^2\big|X = x\big].$$

It should also be noted that one could further obtain identification of the *average treatment effect on the treated* (ATT, see e.g. Heckman and Vytlacil, 2005). Specifically,

$$\begin{aligned}
\text{ATT} &= \mathbb{E}\left[\mu(1,X) - \mu(0,X)|D = 1\right] + \mathbb{E}\big\{[\sigma(1,X) - \sigma(0,X)] \times \epsilon|D = 1\big\} \\
&= \mathbb{E}\left[\mu(1,X) - \mu(0,X)|D = 1\right] + \mathbb{E}\left\{\left[1 - \frac{|V_0(X)|^{\frac{1}{2}}}{|V_1(X)|^{\frac{1}{2}}}\right] \times \mathbb{E}[Y - \mu(1,X)|X, D = 1]\right\}.
\end{aligned}$$

Interestingly, once $\mu(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ have been identified, Vuong and Xu (2017)'s *counterfactual mapping* approach can be used to identify counterfactual outcomes for each individual. Let $Y_d \equiv \mu(d,X) + \sigma(d,X) \times \epsilon$ be the "potential outcome" under the treatment status d. By definition, Y_d is observed in the data if and only if $D = d$. The endogeneity issue arises due to the missing observations of Y_{1-d} when $D = d$. Given model (1), the unobserved potential outcomes (counterfactuals) can be explicitly constructed by the distribution of the observables: Suppose w.l.o.g. $D = 1$. Then, $Y_1 = Y$, and by Lemma 2,

$$Y_0 = \mu(0,X) + [Y - \mu(1,X)] \times \frac{\sigma(0,X)}{\sigma(1,X)} = \delta_0(X) + [Y - \delta_1(X)] \times \frac{|V_0(X)|^{\frac{1}{2}}}{|V_1(X)|^{\frac{1}{2}}},$$

which is constructively identified from the data. This also suggests an alternative expression for ATT:

$$\text{ATT} = \mathbb{E}\left\{ Y - \delta_0(X) - [Y - \delta_1(X)] \times \frac{|V_0(X)|^{\frac{1}{2}}}{|V_1(X)|^{\frac{1}{2}}} \Big| D = 1 \right\}.$$

3.1. **Interpretations under monotone selection and misspecification.** If the linear outcome equation is misspecified, Imbens and Angrist (1994) point out that the usual IV estimator should be interpreted as LATE (under an additional monotone selection assumption) rather than ATE. Though our model is less restrictive, it is still useful to interpret the EHIV estimators when the underlying structure for the data generating process is fully nonseparable.

Specifically, suppose the outcome equation is given as follows:

$$Y = h(D, X, \epsilon)$$

where h is nonseparable in the error term ϵ, and in addition equation (2) holds with $m(X,0) < m(X,1)$. First, we argue that $V_d(X)$ can be interpreted as the (conditional) variance of the corresponding potential outcome given the "compliers group". To fix ideas, define

$$\text{Complier}(X) \equiv \{\eta \in \mathbb{R} : m(X,0) < \eta \leq m(X,1)\}$$

as the group of compliers who switch their treatment participation decision with the realization of Z. Specifically, a complier chooses $D = 0$ if and only if $Z = 0$. Moreover, define

$$\text{Always-Taker}(X) \equiv \{\eta \in \mathbb{R} : \eta \leq m(X,0)\};$$
$$\text{Never-Taker}(X) \equiv \{\eta \in \mathbb{R} : \eta > m(X,1)\},$$

as the group of individuals who always participate in the treatment and the group of individuals who never participate, respectively, regardless the realization of Z; see Imbens and Angrist (1994) for a detailed discussion on these three groups. By a similar argument to Imbens and Angrist (1994), one can show that $\delta_d(X)$ can be interpreted as the (conditional) mean of the potential outcome Y_d given X and the group of compliers:

$$\delta_d(X) = \mathbb{E}(Y_d | X, \text{Complier}(X)).$$

In addition, $V_d(X)$ is the (conditional) variance of potential outcome Y_d given X and the group of compliers:

$$V_d(X) = \text{Var}(Y_d | X, \text{Complier}(X)).$$

It is worth pointing out that such a "local variance" interpretation does not depend on the functional form specification in model (1).

Furthermore, denote $R(X) = \sqrt{V_0(X)/V_1(X)}$. Let further $Q_1(X) = 1 - p(X,0) + R(X)p(X,0)$ and $Q_0(X) = p(X,1) + R^{-1}(X)[1 - p(X,1)]$. By definition, $Q_1(X) = R(X)Q_0(X) + [1 - R(X)] \times [p(X,1) - p(X,0)]$, and both $Q_0(X)$ and $Q_1(X)$ are positive. Using eqs. (7) and (8), we have

$$\begin{aligned} & \mu(1,X) - \mu(0,X) \\ = \; & \mathbb{E}[h(1,X,\epsilon)|X, \text{Complier}(X)] \times Q_1(X) + \mathbb{E}[h(1,X,\epsilon)|X, \text{Always-Taker}(X)] \times [1 - Q_1(X)] \\ - \; & \mathbb{E}[h(0,X,\epsilon)|X, \text{Complier}(X)] \times Q_0(X) - \mathbb{E}[h(0,X,\epsilon)|X, \text{Never-Taker}(X)] \times [1 - Q_0(X)], \end{aligned}$$

which we call the "adjusted" LATE if model (1) is indeed misspecified. Note that the LATE uses information contained only in the complier group. The "adjusted" LATE, however, depends upon information contained in all three groups. Moreover, if $V_0(X) = V_1(X)$, i.e. the case of exogenous heteroskedasticity, we have $Q_0(X) = Q_1(X) = 1$, then $\mu(1,X) - \mu(0,X)$ becomes the (conditional) LATE. Alternatively, suppose $p(X,0) = 0$ and $p(X,1) = 1$. Then we also have $Q_0(X) = Q_1(X) = 1$. Our "adjusted" LATE extrapolates information from the three groups to the whole population, depending on the relative variance of potential outcomes in the complier groups as well as the probability masses of the three groups. It should also be noted that under misspecification, our model can provide a "better" approximation to the underlying data generating structure than the standard IV model with exogenous heteroskedasticity since the latter is nested in our model.

4. Semiparametric Estimation

For ease of implementation and in line with empirical practice, a linear specification for the $\mu(\cdot,\cdot)$ is considered here. Specifically, the following model with $\mu(D,X) = X'\beta_1 + \beta_2 D$ is considered:

$$Y = X'\beta_1 + \beta_2 D + \sigma(D,X) \times \epsilon \tag{9}$$

where $\beta_1 \in \mathbb{R}^{d_X}$ and $\beta_2 \in \mathbb{R}$. Such a specification is parsimonious, with the average treatment effects measured by the scalar parameter β_2. This semiparametric model is a natural extension of the standard linear IV model with (exogenous) heteroskedasticity. While it is possible to estimate $\mu(\cdot,\cdot)$ in model (1) nonparametrically, such an approach would suffer from the curse of dimensionality.

For notational simplicity, let $W = (X', Z)' \in \mathbb{R}^{d_X} \times \{0, 1\}$ and $\beta = (\beta_1', \beta_2)' \in \mathbb{R}^{d_X+1}$. Let $\{(Y_i, D_i, W_i')' : i \leq n\}$ be an i.i.d. random sample of $(Y, D, W')'$ generated from (9), where $n \in \mathbb{N}$ is the sample size. To simplify the theoretical development, all the components of X are assumed to be continuously distributed, with $f_X(\cdot)$ denoting the density function. In practice, if X contains discrete variables which are ordered with rich support, then the discrete components can be simply treated as continuous random variables or a smoothing method (see e.g. Racine and Li, 2004) can be applied. Denote $\Delta_\sigma(X) \equiv \sigma(1, X) - \sigma(0, X)$ and $\Delta_p(X) \equiv p(X, 1) - p(X, 0)$.

First, we nonparametrically estimate $\delta_d(X_i)$ and $V_d(X_i)$ for each $i \leq n$. Let $K : \mathbb{R}^{d_X} \to \mathbb{R}$ and h be a Nadaraya-Watson kernel and bandwidth, respectively. Conditions on K and h will be formally introduced in the asymptotic analysis below. For a generic random variable $A \in \mathbb{R}$, denote $\phi_A(X_i) \equiv f_X(X_i) \times \mathbb{E}(A_i|X_i)$. Following the standard kernel estimation literature, $\phi_A(X_i)$ is estimated by

$$\hat{\phi}_A(X_i) = \frac{1}{(n-1)h^{d_X}} \sum_{j \neq i} A_j K\big(\frac{X_j - X_i}{h}\big).$$

In particular, when A is a constant, e.g. $A = 1$, we have

$$\hat{\phi}_1(X_i) = \frac{1}{(n-1)h^{d_X}} \sum_{j \neq i} K\big(\frac{X_j - X_i}{h}\big),$$

which is a kernel density estimator of $f_X(X_i)$. Note that the estimation of $\phi_A(X_i)$ leaves the i-th observation out to improve its finite sample performance. Moreover, for $d = 0, 1$, let

$$\hat{\delta}_d(X_i) = (-1)^{1+d} \times \frac{\hat{\phi}_1(X_i)\hat{\phi}_{Y\mathbb{1}(D=d)Z}(X_i) - \hat{\phi}_{Y\mathbb{1}(D=d)}(X_i)\hat{\phi}_Z(X_i)}{\hat{\phi}_1(X_i)\hat{\phi}_{DZ}(X_i) - \hat{\phi}_D(X_i)\hat{\phi}_Z(X_i)},$$

$$\hat{V}_d(X_i) = (-1)^{1+d} \times \frac{\hat{\phi}_1(X_i)\hat{\phi}_{Y^2\mathbb{1}(D=d)Z}(X_i) - \hat{\phi}_{Y^2\mathbb{1}(D=d)}(X_i)\hat{\phi}_Z(X_i)}{\hat{\phi}_1(X_i)\hat{\phi}_{DZ}(X_i) - \hat{\phi}_D(X_i)\hat{\phi}_Z(X_i)} - \hat{\delta}_d^2(X_i).$$

In the above expressions, the term $(-1)^{1+d}$ is introduced due to the fact that

$$\text{Cov}(\mathbb{1}(D = d), Z|X) = (-1)^{1+d} \times \text{Cov}(D, Z|X), \text{ for } d = 0, 1.$$

Thereafter, we estimate S_i by the plug-in method:

$$\hat{S}_i \equiv |\hat{V}_0(X_i)|^{\frac{1}{2}} \times (1 - D_i) + |\hat{V}_1(X_i)|^{\frac{1}{2}} \times D_i.$$

Let $\varphi_{ni} = \hat{\phi}_1(X_i)\hat{\phi}_{DZ}(X_i) - \hat{\phi}_D(X_i)\hat{\phi}_Z(X_i)$ be the denominator from the estimators above. Clearly, small values of φ_{ni} could lead to a denominator issue. Moreover, it is well known that

the above kernel estimators will be biased at the boundaries of the support. Therefore, attention is restricted to nonparametric estimation on an inner support $\mathscr{X}_n \equiv \{x \in \mathscr{S}_X : \mathcal{B}_x(h) \subseteq \mathscr{S}_X\}$, where $\mathcal{B}_x(h) \equiv \{\tilde{x} \in \mathbb{R}^{d_X} : \|\tilde{x} - x\| \leq h\}$.

In the second step of estimation, β is estimated. Note that the conventional IV regression model with exogenous heteroskedasticity is a special case of (9). When $\sigma(1,\cdot) \neq \sigma(0,\cdot)$, however, the standard IV estimator of β is inconsistent:

$$\hat{\beta}_{IV} = \Big[\sum_{i=1}^n W_i(X_i', D_i)\Big]^{-1} \sum_{i=1}^n W_i Y_i = \beta + \Big[\sum_{i=1}^n W_i(X_i', D_i)\Big]^{-1} \sum_{i=1}^n W_i \sigma(D_i, X_i)\epsilon_i$$
$$\xrightarrow{p} \beta + \mathbb{E}^{-1}[W(X', D)] \times \mathbb{E}\left[W \Delta_\sigma(X) D \epsilon\right]$$

under standard conditions for applying the WLLN in the last step. Clearly, the bias term is equal to zero if and only if $\mathbb{E}\left[W \Delta_\sigma(X) D \epsilon\right] = 0$. (The Monte Carlo experiments of Section 5 provide empirical evidence of the inconsistency of $\hat{\beta}_{IV}$). The proposed *endogenous heteroskedasticity IV (EHIV)* estimator is defined as follows:

$$\hat{\beta} = \left[\frac{1}{n}\sum_{i=1}^n \frac{T_{ni} W_i(X_i', D_i)}{\hat{S}_i}\right]^{-1} \times \frac{1}{n}\sum_{i=1}^n \frac{T_{ni} W_i Y_i}{\hat{S}_i},$$

where $\{T_{ni} : i \leq n\}$ is a trimming sequence for dealing with the denominator issue and the boundary issue in the nonparametric estimation. Specifically,

$$T_{ni} = \mathbb{1}\big(|\varphi_{ni}| \geq \tau_n;\ |\hat{V}_0(X_i)| \geq \kappa_{0n};\ |\hat{V}_1(X_i)| \geq \kappa_{1n};\ X_i \in \mathscr{X}_n\big)$$

for positive deterministic sequences $\tau_n \downarrow 0$, $\kappa_{0n} \downarrow 0$, and $\kappa_{1n} \downarrow 0$ as $n \to \infty$. Conditions on τ_n, κ_{0n}, and κ_{1n} will be introduced later for the asymptotics properties of $\hat{\beta}$. Note that it is possible to apply more sophisticated trimming mechanisms used in the nonparametric regression literature (see, e.g., Klein and Spady, 1993).

Next, the heteroskedasticity function $\sigma(\cdot,\cdot)$ is estimated, which immediately leads to estimates of the variance effects of the treatment. Fix $x \in \mathscr{X}_n$. For $d = 0, 1$, let $d' = 1 - d$, and then define

$$\hat{\sigma}^2(d, x) = \frac{|\hat{V}_d(x)|}{|\hat{V}_1(x)|} \times \frac{\sum_{i=1}^n D_i \hat{u}_i \times K\big(\frac{X_i - x}{h}\big)}{\sum_{i=1}^n K\big(\frac{X_i - x}{h}\big)} + \frac{|\hat{V}_d(x)|}{|\hat{V}_0(x)|} \times \frac{\sum_{i=1}^n (1 - D_i)\hat{u}_i^2 \times K\big(\frac{X_i - x}{h}\big)}{\sum_{i=1}^n K\big(\frac{X_i - x}{h}\big)}$$

where $\hat{u}_i = Y_i - X_i'\hat{\beta}_1 - \hat{\beta}_2 D_i$. Under additional conditions, it is shown below that $\hat{\beta}$ converges to β at the parametric rate, and therefore $\hat{u}_i$ converges to $u_i \equiv \sigma(D_i, X_i) \times \epsilon_i$ at the same rate. Therefore, the estimation errors associated with $\hat{u}_i$ are asymptotically negligible in the estimation

of $\sigma^2(d,x)$ under some regularity conditions. The variance effects of the treatment are estimated by $\hat{\sigma}(1,x) - \hat{\sigma}(0,x)$ for all $x \in \mathscr{X}_n$, and also the median of the variance effects, denoted as MVE, is estimated by Median $\{T_{ni}\,[\hat{\sigma}(1,X_i) - \hat{\sigma}(0,X_i)]\}$. Note that the MVE differs from the variance of the treatment effects.

In conducting program evaluation, decision-makers might also be interested in the distributional effects of the treatment (see e.g. Heckman and Vytlacil, 2007). From the model in (9), the *individual treatment effect* (ITE) is given by

$$\text{ITE} = \beta_2 + \Delta_\sigma(X) \times \epsilon,$$

which takes a non-degenerate probability distribution as long as $\Delta_\sigma(X) \neq 0$ with strict positive probability. By Lemma 2 and $S = \sigma(D,X) \times |C(X)|^{\frac{1}{2}}$, the ITE can be re-written as

$$ITE = \beta_2 + \Delta_\sigma(X) \times \epsilon = \beta_2 + \frac{|V_1(X)|^{\frac{1}{2}} - |V_0(X)|^{\frac{1}{2}}}{S} \times [Y - (X', D)\beta].$$

Based upon this expression, we estimate the ITE for observation i (if $T_{ni} \neq 0$) by

$$\widehat{\text{ITE}}_i = \hat{\beta}_2 + \frac{|\hat{V}_1(X_i)|^{\frac{1}{2}} - |\hat{V}_0(X_i)|^{\frac{1}{2}}}{\hat{S}_i} \times \hat{u}_i.$$

Then, to estimate the distribution of ITE (conditional on covariates), we follow Guerre, Perrigne, and Vuong (2000) by using the pseudo-sample of $\widehat{\text{ITE}}_i$'s estimated above:

$$\hat{f}_{\text{ITE}|X}(e|x) = \frac{h_f^{-(d_X+1)} \sum_{i=1}^n T_{ni} K_f\left(\frac{X_i - x}{h_f}, \frac{\widehat{\text{ITE}}_i - e}{h_f}\right)}{h_X^{-d_X} \sum_{i=1}^n T_{ni} K_X\left(\frac{X_i - x}{h_X}\right)}, \quad \forall\, e \in \mathbb{R},$$

where $K_f : \mathbb{R}^{d_X+1} \to \mathbb{R}$ and $K_X : \mathbb{R}^2 \to \mathbb{R}$ are Nadaraya-Watson kernels; $h_f \in \mathbb{R}^+$ and $h_X \in \mathbb{R}^+$ are bandwidths. By a similar argument to Guerre, Perrigne, and Vuong (2000), conditions for the choice of h_f (see below) imply oversmoothing due to the fact that the ITE is estimated rather than directly observed.

4.1. **Discussion.** It is worth noting that our model (9) fits Ai and Chen (2003)'s general framework of *sieve minimum distance (SMD)* estimation. Therefore, given the identification of structural functions established in Section 3, Ai and Chen (2003)'s SMD approach could apply here to construct a $\sqrt{n}$-consistent estimator for β. The SMD approach would estimate the finite-dimensional parameter β and nonparametric functions $\sigma(\cdot,\cdot)$ simultaneously from the following conditional

moments:

$$\mathbb{E}\Big[\frac{Y - X'\beta_1 - \beta_2 D}{\sigma(D,X)}\Big|W\Big] = 0,$$
$$\mathbb{E}\Big[\frac{(Y - X'\beta_1 - \beta_2 D)^2}{\sigma^2(D,X)}\Big|W\Big] = 1.$$

In contrast to SMD, the EHIV approach described above leads to closed-form expressions for all of the estimators of interest.

In addition, suppose one assumes the following parametric variance model:

$$\sigma(D,X) = \exp\left[(1,X') \times \pi_1 + \pi_2 D\right],$$

where $\pi_1 \in \mathbb{R}^{d_X+1}$ and $\pi_2 \in \mathbb{R}$ are coefficients. In particular, π_2 characterizes the endogenous heteroskedasticity. Thus, we can estimate β_1, β_2 and π_2 from the following moment equations:

$$\mathbb{E}\Big[\frac{Y - X'\beta_1 - \beta_2 D}{\exp(\pi_2 D)}\Big|W\Big] = 0,$$
$$\mathbb{E}\Big[\frac{(Y - X'\beta_1 - \beta_2 D)^2}{\exp(2\pi_2 D)}\Big|W\Big] = \mathbb{E}\Big[\frac{(Y - X'\beta_1 - \beta_2 D)^2}{\exp(2\pi_2 D)}\Big|X\Big].$$

A standard GMM approach applies. Note that the first moment equation provide a closed-form solution of β_1 and β_2 depending on the scalar parameter π_2.

4.2. **Asymptotic properties.** In this subsection, we establish asymptotic properties for the EHIV estimator by following the semiparametric two-step estimation literature (e.g. Bierens, 1983; Powell, Stock, and Stoker, 1989; Andrews, 1994; Newey and McFadden, 1994, among many others). Before we proceed, it is worth pointing out that the EHIV estimator $\hat{\beta}$ is $\sqrt{n}$-consistent if the heteroskedasticity is exogenous, i.e., $\sigma(d,\cdot) = \tilde{\sigma}(\cdot)$ for some $\tilde{\sigma}$, without additional conditions on the first-stage estimation. In the presence of endogeneity, however, the following consistency (resp. $\sqrt{n}$-consistency) argument of $\hat{\beta}$ requires that the first-stage estimation error, i.e. $\hat{V}_d(X_i) - V_d(X_i)$, uniformly converges to zero (resp. uniformly converges to zero faster than $n^{-1/4}$).

To begin with, we make the following assumptions. Most of them are weak and standard in the literature.

Assumption D. *(i) Eq.* (9) *holds; (ii) The data* $\{(Y_i, D_i, W_i')' : i \leq n\}$ *is an i.i.d. random sample; (iii) The support* $\mathscr{S}_X$ *is compact with nonempty interior; (iv) The density of* X *is bounded and bounded away*

from zero on $\mathscr{S}_X$; (v) The function $\mathbb{P}(Z=0|X=x)$ is bounded away from 0 and 1 on $\mathscr{S}_X$; (vi) The parameter space $\mathbb{B} \subseteq \mathbb{R}^{d_X+1}$ of β is compact.

Assumption E. *For each $x \in \mathscr{S}_X$, $|\Delta_p(x)| \geq C_0$ for some $C_0 \in \mathbb{R}_+$.*

Assumption F. *For some integer $R \geq 2$, the functions $\sigma(d,\cdot)$, $p(\cdot,z)$, $f_{XZ}(\cdot,z)$, $\mathbb{E}(\epsilon|D=d,X=\cdot,Z=z)$ and $\mathbb{E}(\epsilon^2|D=d,X=\cdot,Z=z)$ are R-times continuously differentiable on $\mathscr{S}_X$.*

Assumption G. *Let $K:\mathbb{R}^{d_X}\to\mathbb{R}$ be a kernel function satisfying: (i) $K(\cdot)$ has bounded support; (ii); $\int k(u)du=1$; (iii) $K(\cdot)$ is an R-th order kernel, i.e.,*

$$\begin{aligned}\int u_1^{r_1}\cdots u_{d_X}^{r_{d_X}}K(u)du &= 0, \quad \text{if } 1\leq \textstyle\sum_{\ell=1}^{d_X} r_\ell \leq R-1;\\ &< \infty, \quad \text{if } \textstyle\sum_{\ell=1}^{d_X} r_\ell = R,\end{aligned}$$

where $(r_1,\cdots,r_{d_X})\in\mathbb{N}^{d_X}$; (iv) $K(\cdot)$ is differentiable with bounded first derivatives on $\mathbb{R}^{d_X}$.

Assumption H. *As $n\to\infty$, (i) $h\to 0$; (ii) $nh^{d_X}/\ln n\to\infty$.*

Assumption D can be relaxed to some extent: Assumption D-(ii) could be extended to allow for weak time/spatial dependence across observations. Regarding Assumption D-(iii) , unbounded regressors can be accommodated by using high order moment restrictions on the tail distribution of X at the expense of longer proofs. Assumption E is introduced for expositional simplicity. Assumptions F to H are standard in the kernel regression literature. See e.g. Pagan and Ullah (1999). In particular, Assumption F is a smoothness condition that can be further relaxed by a Lipschitz condition. Assumptions E and F imply that for $d,z=0,1$, the functions $\delta_d(\cdot)$, $V_d(\cdot)$, $\mathbb{E}[Y\mathbb{1}(D=d)|X=\cdot,Z=z]$ and $\mathbb{E}[Y^2\mathbb{1}(D=d)|X=\cdot,Z=z]$ are R-times continuously differentiable on $\mathscr{S}_X$ with bounded R-th partial derivatives. In Assumption H, the $\ln n$ arises because we drive uniform consistency for the first-stage nonparametric estimation.

Lemma 3. *Under Assumptions D to H, we have*

$$\sup_{x\in\mathscr{S}_X}\left|\hat{V}_d(x)-V_d(x)\right| = O_p\Big(h^R+\sqrt{\frac{\ln n}{nh^{d_X}}}\Big).$$

The uniform convergence result in Lemma 3 is standard in the kernel estimation literature (see e.g. Andrews, 1995), and therefore proofs are omitted. In particular, the choice of h should balance the bias and variance in the nonparametric estimation. Suppose $h=\lambda_0(n/\ln n)^{-\gamma}$ for some $\lambda_0>0$ and $\gamma\in(0,1/d_X)$. Note that such a choice of h satisfies Assumption H. Then, the convergence rate in Lemma 3 becomes $\big(n/\ln n\big)^{-(R\gamma\wedge\frac{1-\gamma d_X}{2})}$.

Assumption I. *The random matrix* $\frac{W(X',D)}{S}$ *has finite second moments and* $\frac{W\epsilon}{\sqrt{|C(X)|}}$ *has finite forth moments, i.e.,*

$$\mathbb{E}\Big\|\frac{W(X',D)}{S}\Big\|^2 < +\infty; \text{ and } \mathbb{E}\Big\|\frac{W\epsilon}{\sqrt{|C(X)|}}\Big\|^4 < +\infty.$$

Assumption J. *The matrix* $\mathbb{E}\big[\frac{W(X',D)}{S}\big]$ *is invertible.*

Assumption K. *For each* $x \in \mathscr{S}_X$ *and* $d = 0, 1$, *let* $|V_d(x)| \geq C_1$ *for some* $C_1 \in \mathbb{R}_+$.

Assumption L. *As* $n \to +\infty$, *the trimming parameters satsify (i)* $\tau_n \downarrow 0$, $\kappa_{0n} \downarrow 0$, *and* $\kappa_{1n} \downarrow 0$; *(ii)* $\tau_n^{-1}\big(h^{2R} + \frac{\ln n}{nh^{d_X}}\big) \downarrow 0$, $\kappa_{01}^{-1}\big(h^{2R} + \frac{\ln n}{nh^{d_X}}\big) \downarrow 0$, *and* $\kappa_{1n}^{-1}\big(h^{2R} + \frac{\ln n}{nh^{d_X}}\big) \downarrow 0$.

Assumption I is standard, allowing us to apply the WLLN and CLT. Assumption J is a testable rank condition, given that S_i can be consistently estimated. Similar to Assumption E, Assumption K is introduced for expositional simplicity, dealing with the denominator issue. Such a condition can be relaxed at the expense of a longer proof and exposition. Assumption L imposes mild restrictions on the choice of the trimming parameters.

Theorem 1. *Suppose all the assumptions in Lemma 3 and Assumptions I to L hold. Then,* $\hat{\beta} \xrightarrow{p} \beta$.

Theorem 1 shows that if the first-stage nonparametric estimation is uniformly consistent, then the EHIV converges to the true parameter in probability.

With consistency, we are now ready to establish the limiting distribution of $\hat{\beta}$. Following Powell, Stock, and Stoker (1989), we impose conditions on the kernel function and the bandwidth such that the first-stage estimation bias vanishes faster than $\sqrt{n}$. It is worth pointing out that our model fits the general framework in the semiparametric two-step estimation literature (e.g. Andrews, 1994, 1995). Thus, the $\sqrt{n}$-consistency of $\hat{\beta}$ requires that the first-stage estimator $\hat{V}_d(\cdot)$ converges to $V_d(\cdot)$ faster than $n^{-1/4}$.

Assumption M. *As* $n \to +\infty$, *(i)* $n^{\frac{1}{2}}h^R \to 0$; *(ii)* $n^{\frac{1}{4}}\sqrt{\frac{\ln n}{nh^{d_X}}} \to 0$.

Assumption M strengthens Assumption H by requiring that both the first-stage estimation bias $\mathbb{E}[\hat{V}_d(\cdot)] - V_d(\cdot)$ and variance of $\hat{V}_d(\cdot)$ vanish faster than $n^{-1/2}$. Note that this assumption implies that $R \geq d_X$. For instance, one could choose e.g. $h = \lambda \times (n/\ln n)^{1/(2R-\iota)}$ for some positive constants λ and ι to satisfy Assumption M, as long as $d_X - R + \frac{1}{2}\iota > 0$ and $\iota < 2R$.

To derive $\hat{\beta}$'s limiting distribution, we plug (9) into the expression of $\hat{\beta}$, which gives us

$$\hat{\beta} = \beta + \Big[\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i', D_i)}{\hat{S}_i}\Big]^{-1} \times \frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i\epsilon_i}{\sqrt{|C(X_i)|}} + \Big[\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i', D_i)}{\hat{S}_i}\Big]^{-1} \times \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_{ni}W_i\epsilon_i}{\sqrt{|C(X_i)|}}\Big(\frac{S_i}{\hat{S}_i} - 1\Big)\right].$$

Note that the last term on the right-hand side comes from the first-stage estimation error. Unlike the semiparametric weighted least squares estimator (see e.g. Andrews, 1994), the last term on the right hand side converges in distribution to a limiting normal distribution under additional assumptions, instead of being $o_p(n^{-1/2})$. This is because the weighting function used for transformation (i.e. $1/S_i$) depends on the endogenous variable D_i.

Define

$$\psi(Y, D, X) = \frac{D[Y - \delta_1(X)]^2}{V_1(X)} + \frac{(1-D)[Y - \delta_0(X)]^2}{V_0(X)}$$

and let $\Psi = \psi(Y, D, X)$ be a random variable. By Lemma 2, we have $\Psi = [\epsilon - \xi_1(X)]^2/C(X)$, which is uncorrelated with Z conditional on X, i.e., $\text{Cov}(\Psi, Z|X) = 0$. Thus, $\mathbb{E}(\Psi|W) = \mathbb{E}(\Psi|X)$. Let further

$$\zeta = \frac{[\Psi - \mathbb{E}(\Psi|X)] \times [Z - \mathbb{E}(Z|X)]}{2\text{Cov}(D, Z|X)} \times \left[\frac{\mathbb{E}(D\epsilon|X)}{|C(X)|^{1/2}}X', \frac{\mathbb{E}(ZD\epsilon|X)}{|C(X)|^{1/2}}\right]'.$$

By definition, ζ is a random vector of $d_X + 1$-dimensions and $\mathbb{E}(\zeta|W) = 0$.

Theorem 2. *Suppose Assumptions A to M hold. Then we have* $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\big(0, \Omega\big)$, *where* $\Omega \equiv \mathbb{E}^{-1}\big[\frac{(X',D)'W'}{S}\big] \times Var\big[\frac{W\epsilon}{\sqrt{|C(X)|}} - \zeta\big] \times \mathbb{E}^{-1}\big[\frac{W(X',D)}{S}\big]$.

In the asymptotic variance matrix Ω, the term ζ accounts for the first-stage estimation error.

For inference based on Theorem 2, it's necessary to estimate the variance matrix Ω. First, we estimate $\mathbb{E}\big[\frac{(X',D)'W'}{S}\big]$ by

$$\mathbb{E}_n\Big[\frac{(X', D)'W'}{S}\Big] = \frac{1}{\sum_{i=1}^{n} T_{ni}} \times \sum_{i=1}^{n} T_{ni}\frac{(X_i', D_i)'W_i'}{\hat{S}_i}.$$

Next, we construct a pseudo sample of $\{\zeta_i : i \leq n; T_{ni} = 1\}$. Let

$$\mathbb{E}_n\big(\frac{X_i D_i \epsilon_i}{\sqrt{|C(X_i)|}}\big|X_i\big) = \frac{X_i}{\sqrt{|\hat{V}_1(X_i)|}} \times \frac{\sum_{j\neq i} D_j \hat{u}_j K\big(\frac{X_j - X_i}{h}\big)}{\sum_{j\neq i} K\big(\frac{X_j - X_i}{h}\big)},$$

$$\mathbb{E}_n\big(\frac{Z_i D_i \epsilon_i}{\sqrt{|C(X_i)|}}\big|X_i\big) = \frac{1}{\sqrt{|\hat{V}_1(X_i)|}} \times \frac{\sum_{j\neq i} Z_j D_j \hat{u}_j K\big(\frac{X_j - X_i}{h}\big)}{\sum_{j\neq i} K\big(\frac{X_j - X_i}{h}\big)},$$

be estimators of $\mathbb{E}\big(\frac{X_i D_i \epsilon_i}{\sqrt{|C(X_i)|}}\big|X_i\big)$ and $\mathbb{E}\big(\frac{Z_i D_i \epsilon_i}{\sqrt{|C(X_i)|}}\big|X_i\big)$, respectively. For all $i, j \leq n$ satisfying $T_{ni} = 1$, let further

$$\hat{\Psi}_{ji} = \frac{D_j[Y_j - \hat{\delta}_1(X_i)]^2}{\hat{V}_1(X_i)} + \frac{(1 - D_j)[Y_j - \hat{\delta}_0(X_i)]^2}{\hat{V}_0(X_i)}$$

and $\hat{\Psi}_i = \hat{\Psi}_{ii}$. Thus, we construct ζ_i by

$$\hat{\zeta}_i = \frac{\frac{1}{n-1}\sum_{j\neq i}(\hat{\Psi}_i - \hat{\Psi}_{ji}) K_h\big(X_j - X_i\big) \times \frac{1}{n-1}\sum_{j\neq i}(Z_i - Z_j) K_h\big(X_j - X_i\big)}{2\big[\hat{\phi}_1(X_i)\hat{\phi}_{DZ}(X_i) - \hat{\phi}_D(X_i)\hat{\phi}_Z(X_i)\big]} \times \Big\{\mathbb{E}_n\Big[\frac{X_i' D_i \epsilon_i}{\sqrt{|C(X_i)|}}\big|X_i\Big], \mathbb{E}_n\Big[\frac{Z_i D_i \epsilon}{\sqrt{|C(X_i)|}}\big|X_i\Big]\Big\}',$$

where $K_h(\cdot) = K(\cdot/h)/h^{d_X}$. Hence, we obtain a pseudo sample $\{\hat{\zeta}_i : i \leq n, T_{ni} = 1\}$ of ζ. Furthermore, because $\frac{W\epsilon}{\sqrt{|C(X)|}} = \frac{Wu}{S}$, we estimate $\mathrm{Var}(\frac{W\epsilon}{\sqrt{|C(X)|}} - \zeta)$ by the sample variance of $\Big\{\frac{W_i \hat{u}_i}{\hat{S}_i} - \hat{\zeta}_i : i \leq n, T_{ni} = 1\Big\}$, denoted as $\hat{\mathrm{Var}}(\frac{W\epsilon}{\sqrt{|C(X)|}} - \zeta)$.

We are now ready to define an estimator of Ω as follows:

$$\hat{\Omega} \equiv \mathbb{E}_n^{-1}\left[\frac{(X', D)'W'}{S}\right] \times \hat{\mathrm{Var}}\left(\frac{W\epsilon}{\sqrt{|C(X)|}} - \zeta\right) \times \mathbb{E}_n^{-1}\left[\frac{W(X', D)}{S}\right].$$

The consistency is given by a similar argument to Theorem 1. In practice, one could also obtain the standard errors of $\hat{\beta}$ by the bootstrap (see e.g. Abadie, 2002) and/or by simulation methods (see e.g. Barrett and Donald, 2003).

Finally, we provide the asymptotic properties of $\hat{\sigma}(\cdot,\cdot)$. Note that $\hat{u}_i = u_i - (X_i', D_i)(\hat{\beta} - \beta) = u_i + O_p(n^{-1/2})$, where the $O_p(n^{-1/2})$ holds uniformly. Therefore, we have

$$\hat{\sigma}^2(d,x) = \frac{|\hat{V}_d(x)|}{|\hat{V}_1(x)|} \times \frac{\sum_{i=1}^n D_i u_i^2 \times K\big(\frac{X_i - x}{h}\big)}{\sum_{i=1}^n K\big(\frac{X_i - x}{h}\big)} + \frac{|\hat{V}_d(x)|}{|\hat{V}_0(x)|} \times \frac{\sum_{i=1}^n (1 - D_i) u_i^2 \times K\big(\frac{X_i - x}{h}\big)}{\sum_{i=1}^n K\big(\frac{X_i - x}{h}\big)} + O_p(n^{-1/2}),$$

provided that the conditions in Theorem 2 hold. Following the standard nonparametric literature (e.g. Pagan and Ullah, 1999), we obtain the asymptotic properties of $\hat{\sigma}(\cdot,\cdot)$.

Theorem 3. *Suppose all the assumptions in Theorem 2 hold. Then for any compact subset $\mathbb{C}$ of $\mathbb{R}^{d_X}$,*

$$\sup_{x \in \mathbb{C}} |\hat{\sigma}(d,x) - \sigma(d,x)| = O_p\Big(\sqrt{\frac{\ln n}{n h^{d_X}}}\Big), \text{ for } d = 0, 1.$$

Theorem 3 establishes the uniform convergence of $\hat{\sigma}(d,\cdot)$ on any compact subset $\mathbb{C}$. Note that Assumption M implies that the bias in the estimation of $\sigma(d,\cdot)$ vanishes faster than $\sqrt{n}$. Therefore, the convergence rate of $\hat{\sigma}(d,\cdot)$ is fully determined by the asymptotic variance of the nonparametric estimator $\hat{V}_d(x)$.

By a similar argument to Guerre, Perrigne, and Vuong (2000), one can also establish the uniform convergence of $\hat{f}_{\text{ITE}|X}(\cdot|\cdot)$ to $f_{\text{ITE}|X}(\cdot|\cdot)$ under their conditions.

5. Monte Carlo Evidence

To illustrate our two-step semiparametric procedure, we conduct a Monte Carlo study. In particular, we consider the following triangular model as the data generating process:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 D + (0.1 + 0.25|X| + \lambda_0 D)\epsilon, \\ D &= \mathbb{1}\big[\Phi(\eta) \geq 0.2|X| + r_0 Z\big] \end{aligned}$$

where $X \sim N(0,1)$, $Z \sim Bernoulli(0.5)$, (ϵ, η) has a bivariate normal distribution with unit variance and correlation coefficient $\rho_0 \in (-1,1)$, and $\Phi(\cdot)$ denotes the CDF of the standard normal distribution. Moreover, $\lambda_0 \in \mathbb{R}_+$ and $r_0 \in \mathbb{R}_+$ are two positive constants to be specified, with the former measuring the level of endogenous heteroskedasticity and the latter capturing the size of the "complier group". Let $(X, Z) \perp (\epsilon, \eta)$ to satisfy Assumptions A and B. For simplicity, let further $X \perp Z$. Assumption C holds trivially. Regarding conditions for asymptotics,

Assumptions D-(iv) and E are not satisfied in our setting, but note that these conditions are imposed for the simplicity of proofs and expositions.

For each replication, we draw an i.i.d. random sample $\{(W_i, \epsilon_i, \eta_i) : i \leq n\}$ and then generate a random sample $\{(Y_i, D_i, W_i) : i \leq n\}$ of size $n = 1000, 2000, 4000$ from the data generating process. Next, we apply our estimation procedure for each replication. All reported results are based on 500 replications.

To assess the finite sample behavior of the estimators, we set $\beta = (0, 1, 1)'$ and $(\lambda_0, r_0, \rho_0) = (0.5, 0.5, 0.5)$ and then compare EHIV's performance with the standard IV estimator. For the first stage estimation of $V_d(\cdot)$, we consider two kernel functions of order $R = 4$, i.e., the Gaussian kernel and the Epanechnikov kernel:

$$K_G(u) = \frac{1}{2}(3 - u^2) \times \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2});$$
$$K_E(u) = \frac{15}{8}(1 - \frac{7}{3}u^2) \times \frac{3}{4}(1 - u^2) \times \mathbb{1}(|u| \leq 1).$$

Note that the bounded support condition in Assumption G-(i) is satisfied by $K_E(\cdot)$, but not by $K_G(\cdot)$. Moreover, we follow Silverman's rule of thumb to choose the bandwidth, i.e., $h = 1.06 \times n^{-1/5}$. Clearly, Assumption M is satisfied. For the trimming sequence T_{ni}, we choose $\tau_n = \kappa_{0n} = \kappa_{1n} = 0.1$. We also considered other values for the trimming parameters (e.g., $\tau_n = \kappa_{0n} = \kappa_{1n} = 0.05$ and 0.01), for which the results are qualitatively similar.

Table 4 in the Appendix reports the finite performance of the EHIV estimator in terms of the Mean Bias (MB), Median Bias (MEDB), Standard Deviation (SD), and Root Mean Square Error (RMSE). For comparison, we also provide summary statistics of the IV estimates. In particular, the MB and MEDB of the IV estimates of β_2 do not shrink with the sample size, which provides evidence for inconsistency of the IV estimation. In contrast, both the bias (MB, MEDB) and the variance (SD) of the EHIV estimator decrease at the expected $\sqrt{n}$-rate. Moreover, the summary statistics show that the EHIV behaves similarly for the difference choices of kernel functions.

Figure 6 in the Appendix illustrates the performance of the nonparametric estimates of the endogenous heteroskedasticity $\sigma(\cdot, \cdot)$. The figures on the left side display the true functions $\sigma(d, \cdot)$ and the averages of $\hat{\sigma}(d, \cdot)$ over 500 replications for different sample sizes. As sample size increases, the bias of $\hat{\sigma}(d, \cdot)$ converges to zero quickly. Note that there is a positive finite-sample bias, in particular when the endogenous heteroskedasticity is small. The figures on the right side of Figure 6 provide 95% confidence intervals for $\sigma(d, x)$ for a sample size of 4000.

Next, we estimate $f_{\text{ITE}|X}(\cdot|x)$ at $x = -0.6745$, 0, and 0.6745, which are the first, second, and third quartiles of the distribution of X, respectively. Note that our specification implies that the conditional ITE follows a normal distribution with mean β_0 and variance λ_0^2, regardless of the value of x. Figure 7 in the Appendix shows that $\hat{f}_{\text{ITE}|X}(\cdot|x)$ behaves well for all sample sizes.

As a robustness check, we also consider different sizes of the compliers group (varying r_0), degrees of endogeneity (varying ρ_0), and levels of heteroskedasticity (varying λ_0). For different values of r_0, we use $\tau_n = 0.2 \times r_0$ for the trimming mechanism; otherwise, more observations would be trimmed out as r_0 decreases. Table 5 in the Appendix reports the summary statistics for $n = 4000$. The results are qualitatively similar across different settings. The EHIV performs worse as r_0 decreases to zero, in line with the asymptotic results in Theorem 3.

6. EMPIRICAL APPLICATION

In this section, we apply the EHIV estimation approach to an empirical application, specifically studying the causal effects of fertility on female labor supply. Motivated by Angrist and Evans (1998), we investigate the effects of having a third child on hours worked per week. Having a third child might be expected to affect a mother's labor supply heterogeneously, given that fertility and labor supply are determined simultaneously and some latent variables may interact with the presence of a third child. Following Angrist and Evans (1998), we use the gender mix of the first two children to instrument for the decision of having a third child.[2] There is a strong argument for the validity of this instrument since child gender is randomly assigned and families with first two children of the same gender are significantly more likely to have a third child. Given households' (heterogenous) preferences over consumption, leisure and childrearing, female labor supply is mainly determined by financial and time constraints. Having a third child might cause time constraints to become more stringent and therefore reduce the role of preference heterogeneity, which implies variance effects in the labor supply model.

For our application, the sample is drawn from the 2000 Census data (5-percent public-use microdata sample (PUMS)). The outcome of interest (Y) is hours worked per week of the mother worked in 1999, the binary endogenous explanatory variable (D) is the presence of a third child, and the instrument (Z) is whether the mother's first two children were of the same gender. The

[2]There is also a sizable literature that use twins at first birth as an IV to estimate the relationship between childbearing and female labor supply; see e.g. Rosenzweig and Wolpin (1980a,b), Bronars and Grogger (1994), and Gangadharan, Rosenbloom, Jacobson, and Pearre III (1996), and references therein. Relatedly, Maurin and Moschion (2009) consider the peer mechanism and suggest neighbors' children sex mix as an IV to identify peer effects in female labor market participation.

specifications considered below include mother's education, mother's age at first birth, and age of first child as exogenous covariates (X). To have the units of education in years, we recode some of the Census education classifications as detailed in Table 1. Table 2 provides descriptive statistics for the observable realizations of (Y, D, Z, X) in our sample.

TABLE 1. Re-coding of mother's education based upon Census classifications

Education level	Coded value	Recoded value
No schooling completed	1	0
Nursery school to 4th grade	2	2
5th grade or 6th grade	3	5.5
7th grade or 8th grade	4	7.5
9th grade	5	9
10th grade	6	10
11th grade	7	11
12th grade, No diploma	8	11.5
High school graduate	9	12
Some college credit, but less than 1 year	10	12.5
1 or more years of college, no degree	11	14
Associate degree	12	14
Bachelor's degree	13	16
Master's degree	14	18
Professional degree	15	18
Doctorate degree	16	21

TABLE 2. Descriptive statistics

Variable	Description	Mean	Median	SD
Hours	Hours worked per week in 1999	23.291	25	18.755
Had third child	1 if had third child, 0 otherwise	0.257	0	0.437
Same-sex	1 if first two children are same gender, 0 otherwise	0.502	1	0.500
Education	Mother's education level (in years)	13.951	14	2.228
Age at first birth	Mother's age when first child was born	26.364	26	5.034
1st child's age	Age of first child in 2000	7.550	8	3.032
2nd child's age	Age of second child in 2000	4.548	4	3.061
Sample Size	293,771			

In our estimation, we assume $R = 6$ for Assumption F and use the 6th order Gaussian kernel, i.e.,

$$k_j(u) = \frac{1}{8}(15 - 10u^2 + u^4) \times \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}), \quad \forall u \in \mathbb{R},$$

and $K(u) = k_1(u)k_2(u)k_3(u)k_4(u)$. The bandwidth is chosen by

$$h_z = 1.06 \times \hat{\sigma}_X \times (\hat{c}_z \times n)^{-1/9},$$

where $\hat{\sigma}_X$ is the sample standard deviation of the covariates and $\hat{c}_z = n^{-1} \sum_{i=1}^{n} \mathbb{1}(Z_i = z)$. With these choices, one can verify that Assumptions H and M are satisfied. Moreover, to specify our trimming sequence T_{ni}, we set $\tau_n = 10^{-10}$ and $\kappa_{0n} = \kappa_{1n} = 10^{-2}$. For this trimming sequence, 75,654 observations (roughly 26% of the whole sample) are "trimmed away."

TABLE 3. Estimation Results

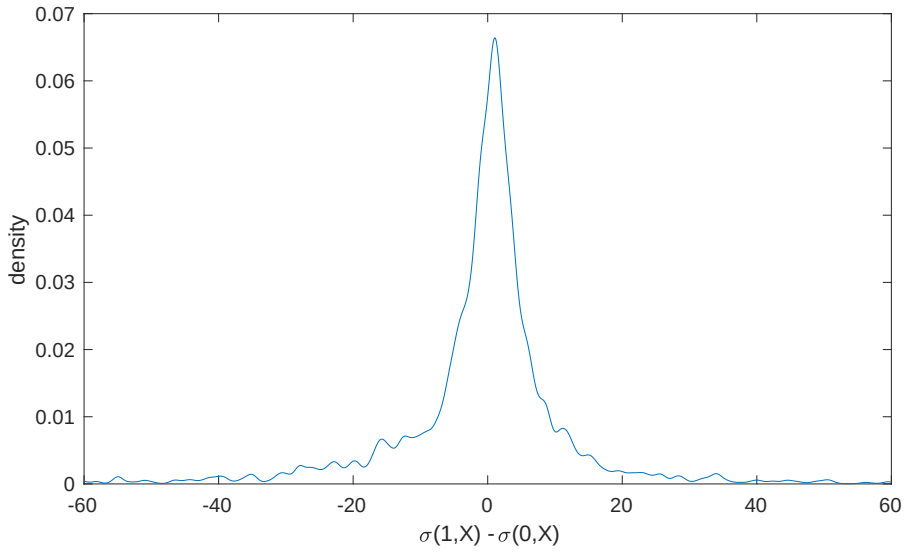
Hours worked per week	OLS	IV	EHIV
Has a third child	-7.597**	-4.226**	-5.343**
	(0.084)	(1.123)	(1.401)
Education	1.046**	1.005**	0.685**
	(0.017)	(0.023)	(0.033)
Age at first birth	-0.341**	-0.282**	-0.368**
	(0.007)	(0.023)	(0.010)
1st child's age	0.635**	0.740**	0.731**
	(0.022)	(0.044)	(0.043)
2nd child's age	0.022	-0.225**	-0.045
	(0.022)	(0.093)	(0.047)
Constant	14.761**	13.219**	19.163**
	(0.271)	(0.625)	(0.830)
ATT			-4.861
			(2.980)
Endogenous heterogeneity test:			
t statistic		10.114	
p-value		0.000	

Table 3 reports the main results from EHIV estimation along with the results obtained from OLS and IV. Across the three methods, there is consistently a negative relationship between having a third child and labor supply. In looking at the OLS and IV results, a similar finding to that in Angrist and Evans (1998) is obtained, with the LATE effect of a third child being considerably lower in magnitude (4.226 hour reduction) than the OLS estimate (7.597 hour reduction). As we've shown previously, the IV estimate of -4.226 may be an inconsistent estimate of the ATE in the presence of endogeneous heteroskedasticity. The EHIV, in contrast, is consistent for the ATE under our model of endogenous heteroskedasticity. In this application, the EHIV estimate is more negative (-5.343) than the IV estimate, although it is still within a standard deviation of the latter. It is interesting to note that, despite the non-parametric estimates that play a role in EHIV estimation, the EHIV standard error is less than 30% larger

than the IV estimator, and this difference is likely to be largely driven by the trimming described above. For the exogenous covariates, EHIV estimates are all of the same sign as the IV estimates, with the largest difference in magnitudes seen for the education and age-at-first-birth covariates. Moreover, the estimate of ATT is -4.861, though this estimate is not significant at a 5% level.

Next, we estimate $\sigma(1, X_i)$ and $\sigma(0, X_i)$ for each observation in the sample. Using the kernel approach, we show the density function of variance effects (i.e., $\sigma(1, X) - \sigma(0, X)$) in Figure 1. Overall, variance effects are distributed around zero. This means, having a third child could either increase or decrease the standard deviation of the mother's labor supply, depending on the value of covariates.

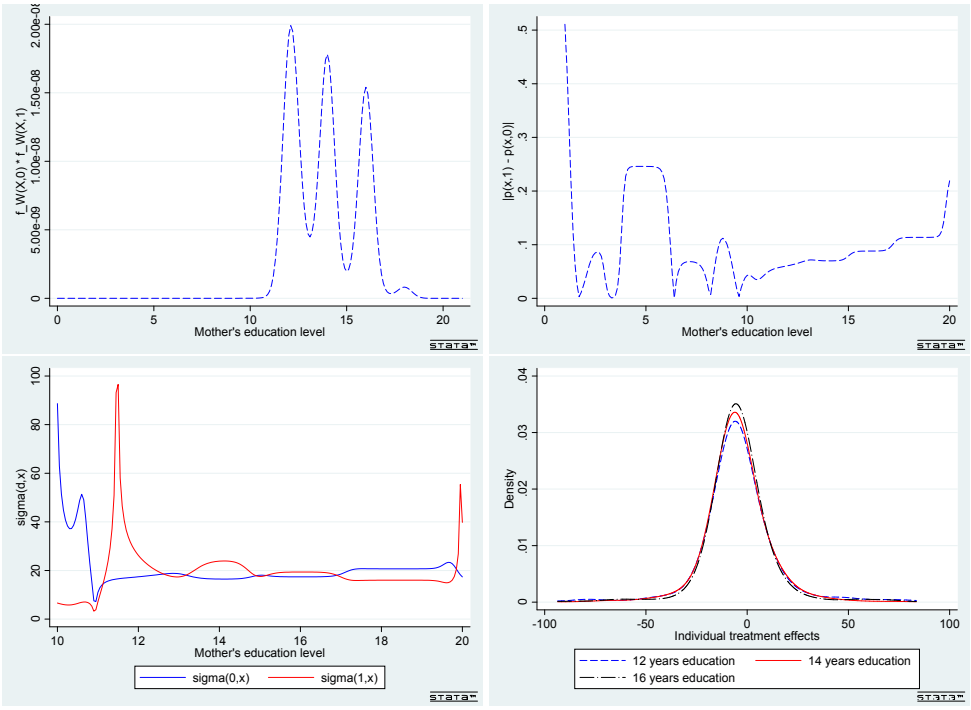
FIGURE 1. Density of EHIV variance effects



We also plot $\sigma(d, x)$ at different values of x. Fixing age at first birth, 1st child's age, and 2nd child's age at their median values, we first estimate $\sigma(d, x)$ as a function of the treatment variable and the mother's education level. The top-left figure in Figure 2 shows the density of the education variable, which leads us to focus our estimation of $\sigma(d, x)$ on the range between 10 and 20 years of education. The estimated $\sigma(d, 0)$ and $\sigma(d, 1)$ functions (i.e., as a function of education) are shown in the bottom-left figure of Figure 2. The top-right figure of Figure 2 gives a sense of the size of the complier group, as it shows $|\hat{p}(x, 1) - \hat{p}(x, 0)|$ as a function of education (again fixing other covariates at their median). Finally, we provide the estimated ITE distributions for three different levels of education (12 years, 14 years, 16 years) in the

bottom-right figure of Figure 2. The most notable feature of the ITE distributions is the large amount of heterogeneity in the ITE's. Although the center of these ITE distributions lines up with the EHIV coefficient estimate (-5.343) from Table 3, the region of non-negligible positive weight includes positive ITE's of up to 20 hours and negative ITE's as low as -30 hours.

FIGURE 2. EHIV variance effects and ITE distributions (education)



Figures 3 to 5 are similar to Figure 2, except that they consider the other three exogenous variables (age at first birth, 1st child's age, and 2nd child's age, respectively). For example, Figure 3 provides estimates of $\sigma(d, x)$ and the ITE distributions as functions of age at first birth, with the other exogenous covariates fixed at their median values. Not surprisingly, the large heterogeneity found in the ITE distributions (each in the lower-right of the corresponding figure) is similar to that seen in Figure 2. In terms of how these distributions vary for different covariate values, it appears that the largest differences are found for age at first birth (Figure 3) and 2nd child's age (Figure 5).

FIGURE 3. EHIV variance effects and ITE distributions (age at first birth)

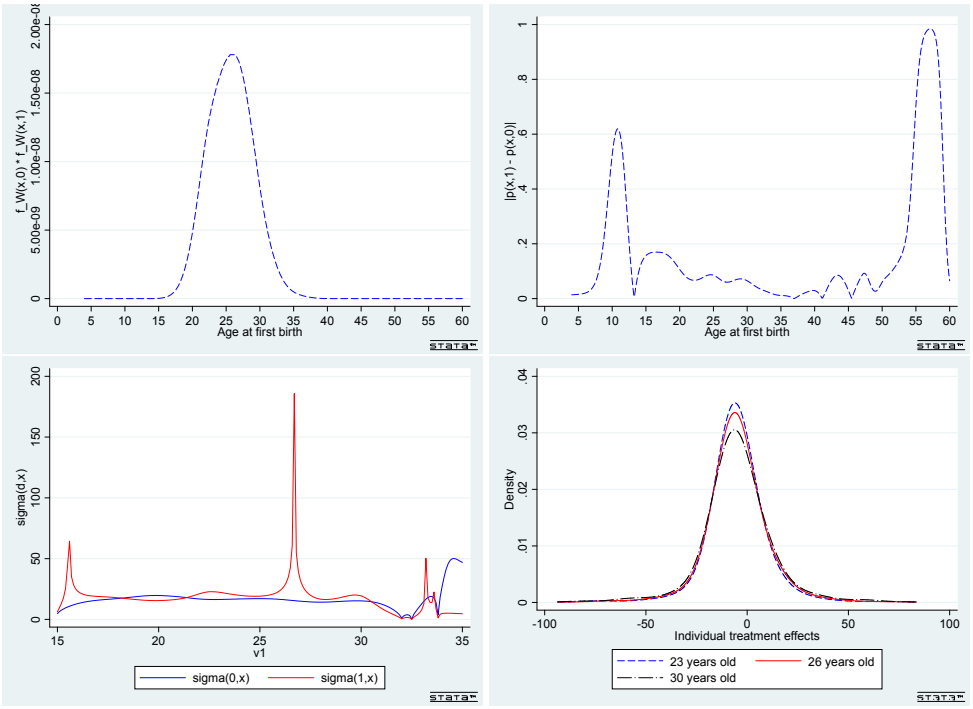


FIGURE 4. EHIV variance effects and ITE distributions (1st child's age)

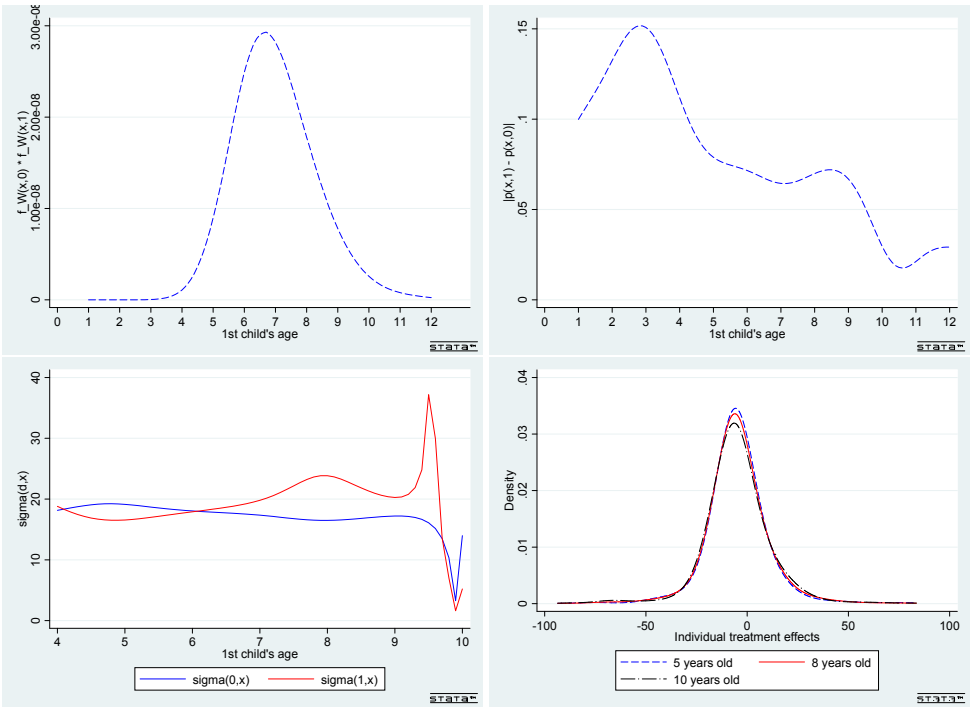
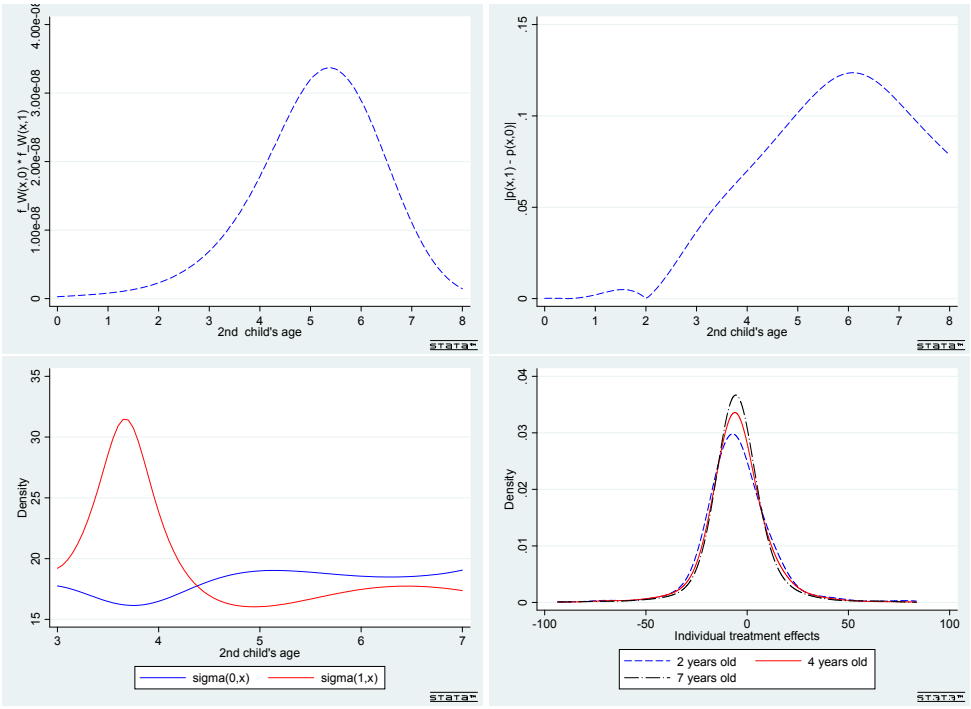


FIGURE 5. EHIV variance effects and ITE distributions (2nd child's age)



7. Conclusion

This paper has considered identification and estimation of a linear model with endogenous heteroskedasticity. Our model assumes that the treatment variable has both mean and variance effects on the outcome variable, which implies heterogenous treatment effects even among observationally identical individuals. Because of the endogenous heteroskedasticity, the standard IV estimator is inconsistent. We then propose a consistent estimation procedure, modified from the IV approach, which has a closed-form expression and is simple to implement. Under appropriate conditions, we establish the $\sqrt{n}$-consistency and the limiting normal distribution for the proposed estimator. Monte Carlo simulations show that the EHIV estimator works well even in moderately sized samples.

References

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American statistical Association*, 97(457), 284–292.

Ai, C., and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

Andrews, D. W. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, pp. 43–72.

——— (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11(03), 560–586.

Angrist, J. D., and W. N. Evans (1998): "Children and their parents' labor supply: Evidence from exogenous variation in family size," *The American Economic Review*, 88(3), 450.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Barrett, G. F., and S. G. Donald (2003): "Consistent tests for stochastic dominance," *Econometrica*, 71(1), 71–104.

Bierens, H. J. (1983): "Uniform consistency of kernel estimators of a regression function under generalized conditions," *Journal of the American Statistical Association*, 78(383), 699–707.

Bronars, S. G., and J. Grogger (1994): "The economic consequences of unwed motherhood: Using twin births as a natural experiment," *The American Economic Review*, pp. 1141–1156.

Chen, S. H., and S. Khan (2014): "Semi-parametric estimation of program impacts on dispersion of potential wages," *Journal of Applied Econometrics*, 29(6), 901–919.

CHERNOZHUKOV, V., AND C. HANSEN (2004): "The effects of 401(K) participation on the wealth distribution: an instrumental quantile regression analysis," *The Review of Economics and Statistics*, 86(3), 735–751.

——— (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

CHESHER, A. (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73(5), 1525–1550.

FAN, Y., AND Q. LI (1996): "Consistent model specification tests: omitted variables and semiparametric functional forms," *Econometrica: Journal of the econometric society*, pp. 865–890.

FENG, Q., Q. VUONG, AND H. XU (2016): "Nonparametric estimation of heterogeneous individual treatment effects with endogenous treatments," *arXiv preprint arXiv:1610.08899*.

GANGADHARAN, J., J. ROSENBLOOM, J. JACOBSON, AND J. W. PEARRE III (1996): "The effects of child-bearing on married women's labor supply and earnings: Using twin births as a natural experiment," Discussion paper, National Bureau of Economic Research.

GUERRE, E., I. PERRIGNE, AND Q. VUONG (2000): "Optimal nonparametric estimation of first–price auctions," *Econometrica*, 68(3), 525–574.

HECKMAN, J. J., J. SMITH, AND N. CLEMENTS (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *The Review of Economic Studies*, 64(4), 487–535.

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

HECKMAN, J. J., AND E. J. VYTLACIL (2007): "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," *Handbook of econometrics*, 6, 4779–4874.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–475.

JUN, S. J., J. PINKSE, AND H. XU (2011): "Tighter bounds in triangular systems," *Journal of Econometrics*, 161(2), 122–128.

KLEIN, R. W., AND R. H. SPADY (1993): "An efficient semiparametric estimator for binary response models," *Econometrica: Journal of the Econometric Society*, pp. 387–421.

MAURIN, E., AND J. MOSCHION (2009): "The social multiplier and labor market participation of mothers," *American Economic Journal: Applied Economics*, 1(1), 251–72.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press.

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

RACINE, J., AND Q. LI (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119(1), 99–130.

ROSENZWEIG, M. R., AND K. I. WOLPIN (1980a): "Life-cycle labor supply and fertility: Causal inferences from household models," *Journal of Political economy*, 88(2), 328–348.

——— (1980b): "Testing the quantity-quality fertility model: The use of twins as a natural experiment," *Econometrica: journal of the Econometric Society*, pp. 227–240.

VUONG, Q., AND H. XU (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with discrete endogeneity," *Quantitative Economics*.

VYTLACIL, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70(1), 331–341.

WAN, Y., AND H. XU (2015): "Inference in semiparametric binary response models with interval data," *Journal of Econometrics*, 184(2), 347–360.

ZHENG, J. X. (1996): "A consistent test of functional form via nonparametric estimation techniques," *Journal of Econometrics*, 75(2), 263–289.

APPENDIX A. PROOFS

A.1. **Proof of Lemma 1.**

Proof. We first show the first half. It suffices to show the if part. By definition,

$$\tilde{r}_1(X) = \mu(1,X) - \mu(0,X) + [\sigma(1,X) - \sigma(0,X)] \times \frac{\mathbb{E}(D\epsilon|Z=1) - \mathbb{E}\left(D\epsilon|Z=0\right)}{p(X,1) - p(X,0)};$$
$$\tilde{r}_0(X) = \mu(0,X) - [\sigma(1,X) - \sigma(0,X)] \times \frac{\mathbb{E}(\epsilon D|X,Z=1)p(X,0) - \mathbb{E}(\epsilon D|X,Z=0)p(X,1)}{p(X,1) - p(X,0)}.$$

Under the condition $\mathbb{E}(\tilde{\epsilon}^2|X,Z=1) = \mathbb{E}(\tilde{\epsilon}^2|X,Z=0)$, we have

$$\frac{\mathbb{E}\left\{\left[Y - \tilde{r}_0(X) - \tilde{r}_1(X)D\right]^2 \middle| X, Z=1\right\} - \mathbb{E}\left\{\left[Y - \tilde{r}_0(X) - \tilde{r}_1(X)D\right]^2 \middle| X, Z=0\right\}}{p(X,1) - p(X,0)} = 0.$$

Plug (1) into the above equation, so that

$$\begin{aligned}
0 &= [\mu(1,X) - \mu(0,X) - \tilde{r}_1(X)]^2 + [\sigma(1,X) - \sigma(0,X)]^2 \times \xi_2(X) \\
&+ 2[\mu(0,X) - \tilde{r}_0(X)] \times [\mu(1,X) - \mu(0,X) - \tilde{r}_1(X)] \\
&+ 2[\mu(0,X) - \tilde{r}_0(X)] \times [\sigma(1,X) - \sigma(0,X)] \times \xi_1(X) \\
&+ 2\sigma(0,X) \times [\mu(1,X) - \mu(0,X) - \tilde{r}_1(X)] \times \xi_1(X) \\
&+ 2[\mu(1,X) - \mu(0,X) - \tilde{r}_1(X)] \times [\sigma(1,X) - \sigma(0,X)] \times \xi_1(X) \\
&+ 2\sigma(0,X) \times [\sigma(1,X) - \sigma(0,X)] \times \xi_2(X) \\
&= [\sigma(1,X) - \sigma(0,X)]^2 \times \xi_1^2(X) \\
&- 2[\sigma(1,X) - \sigma(0,X)]^2 \times \frac{\mathbb{E}(\epsilon D|X,Z=1)p(X,0) - \mathbb{E}(\epsilon D|X,Z=0)p(X,1)}{p(X,1) - p(X,0)} \times \xi_1(X) \\
&+ 2[\sigma(1,X) - \sigma(0,X)]^2 \times \frac{\mathbb{E}(\epsilon D|X,Z=1)p(X,0) - \mathbb{E}(\epsilon D|X,Z=0)p(X,1)}{p(X,1) - p(X,0)} \times \xi_1(X) \\
&- 2[\sigma(1,X) - \sigma(0,X)]^2 \times \xi_1^2(X) + [\sigma^2(1,X) - \sigma^2(0,X)] \times \xi_2(X) \\
&= [\sigma^2(1,X) - \sigma^2(0,X)] \times C(X).
\end{aligned}$$

Under Assumption C, it follows that $\sigma(0,X) = \sigma(1,X)$.

We now show the second half. Again, the only if part is straightforward and it suffices to show the if part. Suppose $\mathbb{E}\big[(Y - X'\tilde{\beta}_1 - \tilde{\beta}_2 D)^2|X,Z=0\big] = \mathbb{E}\big[(Y - X'\tilde{\beta}_1 - \tilde{\beta}_2 D)^2|X,Z=1\big]$ holds for $\tilde{\beta}$ satisfying $\mathbb{E}(Y - X'\tilde{\beta}_1 - \tilde{\beta}_2 D|X,Z) = 0$. Then, it follows that

$$\begin{aligned}
0 &= \mathbb{E}\left\{\left[X'(\beta_1 - \tilde{\beta}_1) + (\beta_2 - \tilde{\beta}_2)D + \sigma(0,X)\epsilon + (\sigma(1,X) - \sigma(0,X))D\epsilon\right]^2 \middle| X, Z=1\right\} \\
&- \mathbb{E}\left\{\left[X'(\beta_1 - \tilde{\beta}_1) + (\beta_2 - \tilde{\beta}_2)D + \sigma(0,X)\epsilon + (\sigma(1,X) - \sigma(0,X))D\epsilon\right]^2 \middle| X, Z=0\right\}.
\end{aligned}$$

Dividing both sides by $p(X,1)-p(X,0)$, we have

$$\begin{aligned} 0 &= (\beta_2-\tilde{\beta}_2)^2+[\sigma(1,X)-\sigma(0,X)]^2\xi_2(X) \\ &+ 2X'(\beta_1-\tilde{\beta}_1)(\beta_2-\tilde{\beta}_2)+2\big[X'(\beta_1-\tilde{\beta}_1)+(\beta_2-\tilde{\beta}_2)\big][\sigma(1,X)-\sigma(0,X)]\xi_1(X) \\ &+ 2\sigma(0,X)(\beta_2-\tilde{\beta}_2)\xi_1(X)+2\sigma(X,0)[\sigma(1,X)-\sigma(0,X)]\xi_2(X). \end{aligned}$$

Since $\mathbb{E}(Y-X'\tilde{\beta}_1-\tilde{\beta}_2 D|X,Z)=0$,

$$\begin{aligned} &\mathbb{E}\big[X'(\beta_1-\tilde{\beta}_1)+(\beta_2-\tilde{\beta}_2)D+(\sigma(1,X)-\sigma(0,X))D\epsilon|X,Z=1\big] \\ &\qquad -\mathbb{E}\big[X'(\beta_1-\tilde{\beta}_1)+(\beta_2-\tilde{\beta}_2)D+(\sigma(1,X)-\sigma(0,X))D\epsilon|X,Z=0\big]=0. \end{aligned}$$

Therefore, $\beta_2-\tilde{\beta}_2=-[\sigma(1,X)-\sigma(0,X)]\times\xi_1(X)$. It follows that

$$\begin{aligned} 0 &= [\sigma(1,X)-\sigma(0,X)]^2\xi_1^2(X)+[\sigma(1,X)-\sigma(0,X)]^2\xi_2(X)-2[\sigma(1,X)-\sigma(0,X)]^2\xi_1^2(X) \\ &- 2\sigma(X,0)[\sigma(1,X)-\sigma(0,X)]\xi_1(X)+2\sigma(X,0)[\sigma(1,X)-\sigma(0,X)]\xi_2(X) \\ &= [\sigma^2(1,X)-\sigma^2(0,X)]C(X). \end{aligned}$$

Under Assumption C, we have $\sigma(0,X)=\sigma(1,X)$. □

A.2. **Proof of Theorem 1.**

Proof. By the definition of $\hat{\beta}$ and (9),

$$\hat{\beta}-\beta=\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i',D_i)}{\hat{S}_i}\Big)\Big]^{-1}\times\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i,X_i)\epsilon_i}{\hat{S}_i}$$

By Lemmas 4 and 5, $\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i'(X_i',D_i)}{\hat{S}_i}$ converges in probability to $\mathbb{E}\big[\frac{W(X',D)}{S}\big]$ and $\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i,X_i)\epsilon_i}{\hat{S}_i}$ converges in probability to zero. By Assumption J and Slutsky's Theorem, $\hat{\beta}-\beta\xrightarrow{p}0$. □

APPENDIX B. PROOF OF THEOREM 2

Proof. By definition of $\hat{\beta}$ and (9), we have

$$\sqrt{n}(\hat{\beta}-\beta)=\left[\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i',D_i)}{\hat{S}_i}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i,X_i)\epsilon_i}{\hat{S}_i}.$$

First, note that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i'(X_i',D_i)}{\hat{S}_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i',D_i)}{S_i}+\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{S_i}{\hat{S}_i}-1\Big)\frac{T_{ni}W_i(X_i',D_i)}{S_i}.$$

By Lemmas 4 and 5,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i'(X_i',D_i)}{\hat{S}_i}\xrightarrow{p}\mathbb{E}[W(X',D)/S].$$

Hence, it suffices to derive the limiting distribution of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i,X_i)\epsilon_i}{\hat{S}_i}$.

Next, note that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i,X_i)\epsilon_i}{\hat{S}_i}&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{ni}W_i\epsilon_i}{\sqrt{|C(X_i)|}}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T_{ni}\Big(\frac{S_i}{\hat{S}_i}-1\Big)\frac{W_i\epsilon_i}{\sqrt{|C(X_i)|}}\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{W_i\epsilon_i}{\sqrt{|C(X_i)|}}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T_{ni}\left[\frac{\sqrt{|V_1(X)|}}{\sqrt{|\hat{V}_1(X)|}}-\frac{\sqrt{|V_0(X)|}}{\sqrt{|\hat{V}_0(X)|}}\right]\frac{W_iD_i\epsilon_i}{\sqrt{|C(X_i)|}}+o_p(1),\end{aligned}$$

where the last step comes from Lemma 7 and the fact that $\frac{S}{\hat{S}}-1=\frac{\sqrt{|V_0(X)|}}{\sqrt{|\hat{V}_0(X)|}}-1+\left[\frac{\sqrt{|V_1(X)|}}{\sqrt{|\hat{V}_1(X)|}}-\frac{\sqrt{|V_0(X)|}}{\sqrt{|\hat{V}_0(X)|}}\right]\times D$. Applying a Taylor expansion, we have

$$T_{ni}\frac{\sqrt{|V_d(X_i)|}}{\sqrt{|\hat{V}_d(X_i)|}}=T_{ni}\left\{1-\frac{1}{2V_d(X_i)}\big[\hat{V}_d(X_i)-V_d(X_i)\big]\right\}+o_p(n^{-1/2})$$

where the o_p term holds uniformly over i by Theorem 1. Hence, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i,X_i)\epsilon_i}{\hat{S}_i}&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{W_i\epsilon_i}{\sqrt{|C(X_i)|}}\\&\quad+\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{W_iD_i\epsilon_i}{\sqrt{|C(X_i)|}}\times T_{ni}\Big[\frac{\hat{V}_0(X_i)-V_0(X_i)}{V_0(X_i)}-\frac{\hat{V}_1(X_i)-V_1(X_i)}{V_1(X_i)}\Big]\right\}+o_p(1).\end{aligned}\tag{10}$$

Let $\tilde{T}_{ni}=\mathbb{1}\left(|\varphi_{ni}|\geq\tau_n;\ |V_0(X_i)|\geq\kappa_{0n};\ |V_1(X_i)|\geq\kappa_{1n};\ X_i\in\mathscr{X}_n\right)$. By a similar argument to Wan and Xu (2015, Lemma B.7) and Bernstein's tail inequality, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{W_iD_i\epsilon_i}{\sqrt{|C(X_i)|}}\times T_{ni}\Big[\frac{\hat{V}_0(X_i)-V_0(X_i)}{V_0(X_i)}-\frac{\hat{V}_1(X_i)-V_1(X_i)}{V_1(X_i)}\Big]\right\}\\&\qquad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{W_iD_i\epsilon_i}{\sqrt{|C(X_i)|}}\times\tilde{T}_{ni}\Big[\frac{\hat{V}_0(X_i)-V_0(X_i)}{V_0(X_i)}-\frac{\hat{V}_1(X_i)-V_1(X_i)}{V_1(X_i)}\Big]\right\}+o_p(1).\end{aligned}\tag{11}$$

Let $A(X_i)=f_X(X_i)\mathrm{Cov}(D_i,Z_i|X_i)$. By Lemma 6, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{W_iD_i\epsilon_i}{\sqrt{|C(X_i)|}}\times\tilde{T}_{ni}\Big[\frac{\hat{V}_0(X_i)-V_0(X_i)}{V_0(X_i)}-\frac{\hat{V}_1(X_i)-V_1(X_i)}{V_1(X_i)}\Big]\right\}\\&=-\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\tilde{T}_{ni}W_iD_i\epsilon_i}{A(X_i)}\big[\Psi_{ji}-\mathbb{E}(\Psi_i|X_i)\big]\big[Z_j-\mathbb{E}(Z_i|X_i)\big]K_h\big(X_j-X_i\big)+o_p(1).\end{aligned}$$

Let further $T^*_{ni} = \mathbb{1}\left(|\varphi_i| \geq \tau_n;\ |V_0(X_i)| \geq \kappa_{0n};\ |V_1(X_i)| \geq \kappa_{1n};\ X_i \in \mathscr{X}_n\right)$, where $\varphi_i = \phi_1(X_i)\phi_{DZ}(X_i) - \phi_D(X_i)\phi_Z(X_i)$. By Assumption D-(ii) and Assumption E, $T^*_{ni} = \mathbb{1}(X_i \in \mathscr{X}_n)$ for sufficiently large n. Thus,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{W_iD_i\epsilon_i}{\sqrt{|C(X_i)|}} \times \tilde{T}_{ni}\Big[\frac{\hat{V}_0(X_i) - V_0(X_i)}{V_0(X_i)} - \frac{\hat{V}_1(X_i) - V_1(X_i)}{V_1(X_i)}\Big]\right\} \\
= \quad & -\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{T^*_{ni}W_iD_i\epsilon_i}{A(X_i)}\big[\Psi_{ji} - \mathbb{E}(\Psi_i|X_i)\big]\big[Z_j - \mathbb{E}(Z_i|X_i)\big]K_h\big(X_j - X_i\big) + o_p(1).
\end{aligned}$$

Following the Hoeffding's Decomposition in Powell, Stock, and Stoker (1989), we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{T^*_{ni}W_iD_i\epsilon_i}{A(X_i)}\big[\Psi_{ji} - \mathbb{E}(\Psi_i|X_i)\big]\big[Z_j - \mathbb{E}(Z_i|X_i)\big]K_h\big(X_j - X_i\big) \\
= \quad & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbb{E}\left\{\frac{T^*_{ni}W_iD_i\epsilon_i}{A(X_i)}\big[\Psi_{ji} - \mathbb{E}(\Psi_i|X_i)\big]\big[Z_j - \mathbb{E}(Z_i|X_i)\big]K_h\big(X_j - X_i\big)\,\middle|\,\mathscr{F}_j\right\} + o_p(1) \\
= \quad & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\mathbb{E}(W_jD_j\epsilon_j|X_j)}{\mathrm{Cov}(D_j, Z_j|X_j)}\big[\Psi_j - \mathbb{E}(\Psi_j|X_j)\big]\big[Z_j - \mathbb{E}(Z_j|X_j)\big] + o_p(1).
\end{aligned}$$

where the last step uses a similar argument to Lemma 7.

Thus, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{ni}W_i\sigma(D_i, X_i)\epsilon_i}{\hat{S}_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{W_i\epsilon_i}{\sqrt{|C(X_i)|}} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i + o_p(1).$$

The results then simply follow from the CLT and Slutsky's Theorem. □

APPENDIX C. TECHNICAL LEMMAS

Lemma 4. *Suppose the assumptions in Theorem 1 hold. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i', D_i)}{S_i} = \mathbb{E}\Big[\frac{W(X', D)}{S}\Big] + o_p(1)$$

Proof. Because

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{T_{ni}W_i(X_i', D_i)}{S_i} &= \frac{1}{n}\sum_{i=1}^{n}\frac{W_i(X_i', D_i)}{S_i} + \frac{1}{n}\sum_{i=1}^{n}(T_{ni} - 1)\frac{W_i(X_i', D_i)}{S_i} \\
&= \mathbb{E}\left[\frac{W(X', D)}{S}\right] + \frac{1}{n}\sum_{i=1}^{n}(T_{ni} - 1)\frac{W_i(X_i', D_i)}{S_i} + o_p(1)
\end{aligned}$$

where the last step comes from the WLLN. By the Cauchy-Schwarz inequality,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(T_{ni}-1)\frac{W_i(X_i',D_i)}{S_i}\right\| = \mathbb{E}\left\|(T_{ni}-1)\frac{W_i(X_i',D_i)}{S_i}\right\| \leq \left\{\mathbb{E}\left[\frac{\|W_i(X_i',D_i)\|^2}{S_i^2}\right]\times\mathbb{E}(T_{ni}-1)^2\right\}^{1/2}.$$

Because of Assumptions E, K and L and $\mathscr{X}_n \to \mathscr{S}_X$, we have

$$\mathbb{E}(T_{ni}-1)^2 \leq \mathbb{P}(|\phi_D(X_i)\phi_Z(X_i)|<\tau_n)+\mathbb{P}(|\hat{V}_0(X_i)|\geq\kappa_{0n})+\mathbb{P}(|\hat{V}_1(X_i)|\geq\kappa_{1n})+\mathbb{1}(X_i\in\mathscr{X}_n^c)\to 0.$$

By Assumption I,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(T_{ni}-1)\frac{W_i(X_i',D_i)}{S_i}\right\|\to 0. \quad \square$$

Lemma 5. *Suppose the assumptions in Theorem 1 hold. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}T_{ni}\Big(\frac{S_i}{\hat{S}_i}-1\Big)\frac{W_i(X_i',D_i)}{S_i}=o_p(1)$$

Proof. By Cauchy Schwarz inequality,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}T_{ni}\Big(\frac{S_i}{\hat{S}_i}-1\Big)\frac{W_i(X_i',D_i)}{S_i}\right\| \leq \left\{\mathbb{E}\left[T_{ni}\Big(\frac{S_i}{\hat{S}_i}-1\Big)^2\right]\times\mathbb{E}\left\|\frac{W_i(X_i',D_i)}{S_i}\right\|^2\right\}^{-1/2}.$$

By Lemma 3 and assumption L-(ii), $\mathbb{E}\left[T_{ni}\big(\frac{S_i}{\hat{S}_i}-1\big)^2\right]\to 0$. $\square$

Lemma 6. *Suppose all the assumptions in Lemma 3 and Assumption M hold. Then,*

$$\begin{aligned}&\frac{\hat{V}_0(X_i)-V_0(X_i)}{V_0(X_i)}-\frac{\hat{V}_1(X_i)-V_1(X_i)}{V_1(X_i)}\\ &=\frac{1}{A(X_i)}\times\frac{1}{n-1}\sum_{j\neq i}\big[\big(\Psi_{ji}-\mathbb{E}(\Psi_i|X_i)\big)(Z_j-\mathbb{E}(Z_i|X_i))K_h\big(X_j-X_i\big)-Cov(\Psi_i,Z_i|X_i)f_X(X_i)\big]+o_p(n^{-1/2})\end{aligned}$$

where the $o_p(\cdot)$ term holds uniformly over i, and $A(X_i)\equiv f_X(X_i)Cov(D_i,Z_i|X_i)$.

Proof. Let $A(X_i) = f_X(X_i)\text{Cov}(D_i, Z_i|X_i)$. By Taylor expansion, we have

$$
\begin{aligned}
&\frac{\hat{\phi}_1(X_i)\hat{\phi}_{Y^2DZ}(X_i) - \hat{\phi}_{Y^2D}(X_i)\hat{\phi}_Z(X_i)}{\hat{\phi}_1(X_i)\hat{\phi}_{DZ}(X_i) - \hat{\phi}_D(X_i)\hat{\phi}_Z(X_i)} - \frac{\phi_1(X_i)\phi_{Y^2DZ}(X_i) - \phi_{Y^2D}(X_i)\phi_Z(X_i)}{\phi_1(X_i)\phi_{DZ}(X_i) - \phi_D(X_i)\phi_Z(X_i)} \\
= \;& \frac{1}{A(X_i)} \times \frac{1}{n-1}\sum_{j\neq i}\big[Y_j^2 D_j Z_j K_h\big(X_j - X_i\big) - \mathbb{E}(Y_i^2 D_i Z_i|X_i) f_X(X_i)\big] \\
+ \;& \frac{1}{A(X_i)} \times \frac{\mathbb{E}(Y_i^2 D_i Z_i|X_i)}{n-1}\sum_{j\neq i}\big[K_h\big(X_j - X_i\big) - f_X(X_i)\big] \\
- \;& \frac{1}{A(X_i)} \times \frac{\mathbb{E}(Z_i|X_i)}{n-1}\sum_{j\neq i}\big[Y_j^2 D_j K_h\big(X_j - X_i\big) - \mathbb{E}(Y_i^2 D_i|X_i) f_X(X_i)\big] \\
- \;& \frac{1}{A(X_i)} \times \frac{\mathbb{E}(Y_i^2 D_i|X_i)}{n-1}\sum_{j\neq i}\big[Z_j K_h\big(X_j - X_i\big) - \mathbb{E}(Z_i|X_i) f_X(X_i)\big] \\
- \;& \frac{V_1(X_i) + \delta_1^2(X_i)}{A(X_i)} \times \frac{1}{n-1}\sum_{j\neq i}\big[D_j Z_j K_h\big(X_j - X_i\big) - \mathbb{E}(D_i Z_i|X_i) f_X(X_i)\big] \\
- \;& \frac{V_1(X_i) + \delta_1^2(X_i)}{A(X_i)} \times \frac{\mathbb{E}(D_i Z_i|X_i)}{n-1}\sum_{j\neq i}\big[K_h\big(X_j - X_i\big) - f_X(X_i)\big] \\
+ \;& \frac{V_1(X_i) + \delta_1^2(X_i)}{A(X_i)} \times \frac{\mathbb{E}(Z_i|X_i)}{n-1}\sum_{j\neq i}\big[D_j K_h\big(X_j - X_i\big) - \mathbb{E}(D_i|X_i) f_X(X_i)\big] \\
+ \;& \frac{V_1(X_i) + \delta_1^2(X_i)}{A(X_i)} \times \frac{\mathbb{E}(D_i|X_i)}{n-1}\sum_{j\neq i}\big[Z_j K_h\big(X_j - X_i\big) - \mathbb{E}(Z_i|X_i) f_X(X_i)\big] + o_p(n^{-1/2}),
\end{aligned}
$$

where all higher order terms are of $o_p(n^{-1/2})$ uniformly over i due to a similar argument to Lemma 3 and Assumption M. Similarly, we obtain Taylor expansions for

$$
\frac{\hat{\phi}_1(X_i)\hat{\phi}_{Y^2(1-D)Z}(X_i) - \hat{\phi}_{Y^2(1-D)}(X_i)\hat{\phi}_Z(X_i)}{\hat{\phi}_1(X_i)\hat{\phi}_{DZ}(X_i) - \hat{\phi}_D(X_i)\hat{\phi}_Z(X_i)} - \frac{\phi_1(X_i)\phi_{Y^2(1-D)Z}(X_i) - \phi_{Y^2(1-D)}(X_i)\phi_Z(X_i)}{\phi_1(X_i)\phi_{DZ}(X_i) - \phi_D(X_i)\phi_Z(X_i)}
$$

and $\hat{\delta}_d(X_i) - \delta_d(X_i)$.

It follows that

$$
\begin{aligned}
&\frac{\hat{V}_1(X_i) - V_1(X_i)}{V_1(X_i)} \\
= \;& \frac{1}{A(X_i)} \times \frac{1}{n-1}\sum_{j\neq i}\big[\big(\Psi_{1ji} - \mathbb{E}(\Psi_{1i}|X_i)\big)(Z_j - \mathbb{E}(Z_i|X_i))K_h\big(X_j - X_i\big) - \text{Cov}(\Psi_{1i}, Z_i|X_i) f_X(X_i)\big] \\
- \;& \frac{1}{A(X_i)} \times \frac{1}{n-1}\sum_{j\neq i}\big[\big(D_j - \mathbb{E}(D_i|X_i)\big)(Z_j - \mathbb{E}(Z_i|X_i))K_h\big(X_j - X_i\big) - \text{Cov}(D_i, Z_i|X_i) f_X(X_i)\big] \\
+ \;& \frac{\text{Cov}(\Psi_{1i}, Z_i|X_i) - \text{Cov}(D_i, Z_i|X_i)}{A(X_i)} \times \frac{1}{n-1}\sum_{j\neq i}\big[K_h\big(X_j - X_i\big) - f_X(X_i)\big] + o_p(n^{-1/2}).
\end{aligned}
$$

Similarly, we obtain $\frac{\hat{V}_0(X_i)-V_0(X_i)}{V_0(X_i)}$. Because $\text{Cov}(\Psi_{1i}, Z_i|X_i) + \text{Cov}(\Psi_{0i}, Z_i|X_i) = \text{Cov}(\Psi_i, Z_i|X_i) = 0$, $\text{Cov}(D_i, Z_i|X_i) + \text{Cov}(1 - D_i, Z_i|X_i) = 0$, and the result obtains. □

Lemma 7. *Suppose the assumptions in Theorem 2 hold. Then,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} T_{ni} \left[\frac{\sqrt{|V_0(X_i)|}}{\sqrt{|\hat{V}_0(X_i)|}} - 1 \right] \frac{W_i \epsilon_i}{\sqrt{|C(X_i)|}} = o_p(1)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (T_{ni} - 1) \frac{W_i \epsilon_i}{\sqrt{|C(X_i)|}} = o_p(1).$$

Proof. Note that $\mathbb{E}[\frac{W\epsilon}{\sqrt{|C(X)|}}|X] = 0$. Then the result directly follows e.g. Andrews (1994) or Newey and McFadden (1994, Theorem 8.1). □

TABLE 4. Simulation Summary of IV Estimation (seed=7480)

Est.	Kernel	Sample size	Parameter	MB	MEDB	SD	RMSE
IV	NA	1000	β_0	0.1129	0.1095	0.0493	0.1232
			β_1	-0.0001	-0.0018	0.0234	0.0233
			β_2	-0.0720	-0.0706	0.0850	0.1113
		2000	β_0	0.1085	0.1089	0.0344	0.1138
			β_1	0.0003	0.0028	0.0167	0.0167
			β_2	-0.0670	-0.0678	0.0570	0.0879
		4000	β_0	0.1101	0.1090	0.0225	0.1124
			β_1	0.0003	0.0008	0.0122	0.0122
			β_2	-0.0673	-0.0673	0.0389	0.0777
EHIV	K_G	1000	β_0	0.0242	0.0140	0.0468	0.0526
			β_1	0.0017	0.0024	0.0606	0.0605
			β_2	-0.0271	-0.0199	0.0868	0.0909
		2000	β_0	0.0140	0.0096	0.0287	0.0319
			β_1	-0.0023	-0.0048	0.0397	0.0397
			β_2	-0.0157	-0.0133	0.0550	0.0572
		4000	β_0	0.0077	0.0060	0.0158	0.0176
			β_1	-0.0004	-0.0004	0.0245	0.0245
			β_2	-0.0099	-0.0091	0.0341	0.0354
	K_E	1000	β_0	0.0190	0.0149	0.0420	0.0461
			β_1	- 0.0005	-0.0020	0.0590	0.0589
			β_2	-0.0208	-0.0231	0.0851	0.0875
		2000	β_0	0.0165	0.0132	0.0292	0.0335
			β_1	-0.0021	-0.0017	0.0396	0.0396
			β_2	-0.0230	-0.0235	0.0592	0.0635
		4000	β_0	0.0120	0.0091	0.0201	0.0233
			β_1	0.0007	0.0033	0.0277	0.0277
			β_2	-0.0177	-0.0158	0.0405	0.0442

FIGURE 6. Estimation of $\sigma(d, x)$

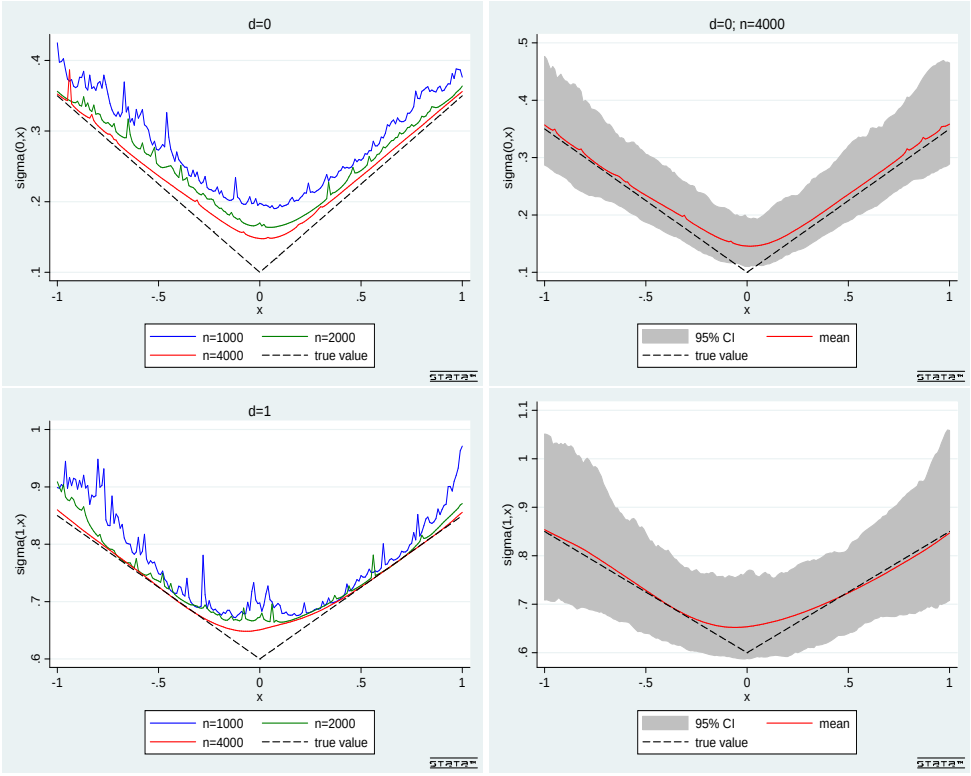


FIGURE 7. Estimation of ITE's density

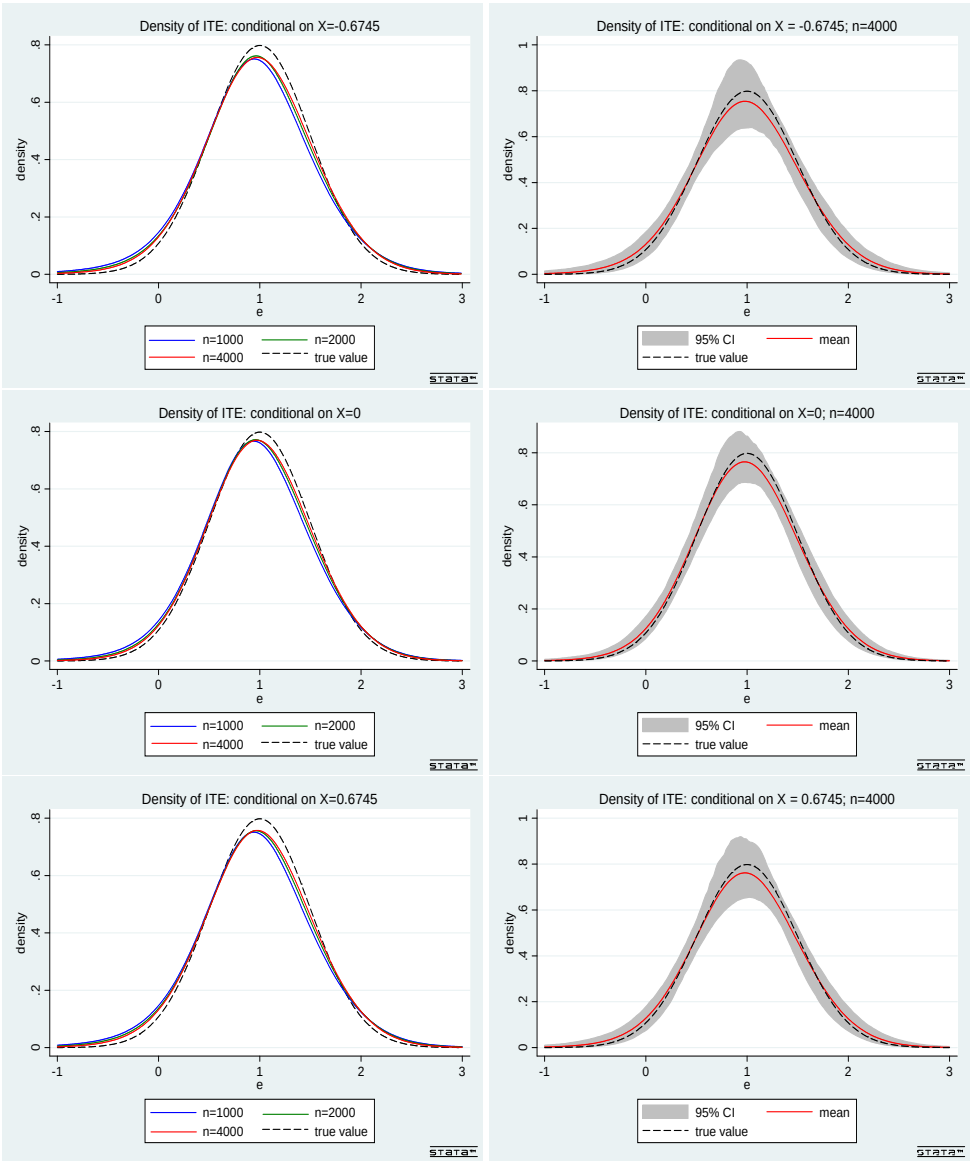


TABLE 5. Robust check: $\hat{\beta}_2$, $n = 4000$ (seed=7480)

r_0	ρ_0	λ_0	MB	MEDB	SD	RMSE
0.1	0.5	0.5	- 0.0001	- 0.0029	0.1433	0.1431
0.2			-0.0508	-0.0477	0.0918	0.1048
0.3			-0.0321	-0.0250	0.0646	0.0721
0.4			-0.0136	-0.0085	0.0445	0.0465
0.5			-0.0047	-0.0035	0.0343	0.0346
0.5	0.0	0.5	0.0032	0.0047	0.0317	0.0318
	0.1		0.0022	0.0042	0.0317	0.0318
	0.2		0.0010	0.0032	0.0315	0.0315
	0.3		-0.0006	0.0012	0.0321	0.0321
	0.4		-0.0022	-0.0021	0.0329	0.0329
	0.6		-0.0089	-0.0068	0.0383	0.0393
	0.7		-0.0145	-0.0110	0.0438	0.0461
	0.8		-0.0222	-0.0177	0.0520	0.0564
	0.9		-0.0309	-0.0259	0.0603	0.0677
0.5	0.5	0.00	0.0008	0.0010	0.0180	0.0180
		0.25	-0.0046	-0.0040	0.0264	0.0268
		0.75	-0.0045	-0.0027	0.0422	0.0424
		1.00	-0.0042	-0.0023	0.0503	0.0504
		1.25	-0.0040	-0.0020	0.0586	0.0587
		1.50	-0.0038	-0.0017	0.0673	0.0673