Testing Instrument Validity in Heterogeneous Treatment Effect Models with Covariates*

Anna Krumme
FernUniversität in Hagen
TU Dortmund

Matthias Westphal
FernUniversität in Hagen
RWI Essen

February 2025

Abstract

We propose a straightforward procedure to test the identifying assumptions of local treatment effect (LATE) estimation conditional on covariates. Using conditional distribution regressions, we identify group-specific distributions while controlling for the overall effect of covariates on the outcome. We derive bounds for unobserved mean potential outcomes from mixing outcome distributions to detect deviations from the mean-based testable implications of the LATE assumptions derived by Huber and Mellace (2015). We contribute to the literature by proposing an easy-to-implement procedure suitable for settings where conditioning on various covariates is essential. Furthermore, we validate the test performance in a brief simulation study and assess the method in two empirical labor market applications to illustrate its practical usefulness.

Keywords: Testing instrument validity, Heterogeneous causal effects, Bounds
JEL Classification: *C10, C12, C26*

Anna Krumme: FernUniversität in Hagen, Faculty of Business Administration and Economics, 58084 Hagen, Germany, E-mail: anna.krumme@fernuni-hagen.de.

Matthias Westphal: FernUniversität in Hagen, Faculty of Business Administration and Economics, 58084 Hagen, Germany, E-mail: matthias.westphal@fernuni-hagen.de.

*We thank Hendrik Schmitz, the participants of the internal RWI health research seminar and the CESA BrownBag Seminar for helpful comments and discussion.

1 Introduction

Instrumental variables (IV) research designs are essential for causal inference and can reveal important heterogeneities of economic agents (Mogstad and Torgovitsky, 2024). The prerequisites for valid IVs are well-known: They need to be exogenous, only affect the outcome through the treatment, and the treatment monotonically only in one direction. Yet, despite available tests for the validity of IVs, they are rarely conducted in practice. Besides the computational complexity of available non- or semiparametric estimation approaches and their often quite technical exposition, the main reason for this underuse is the difficulties of these approaches in dealing with covariates.

This paper aims to fill this gap by building on the testable implications of Huber and Mellace (2015) and combining them with conditional distribution regressions – a method that serves as the basis for IV quantile treatment effects (Chernozhukov and Hansen, 2005; Frandsen et al., 2012) or Lee (2009) bounds for IV (Dong, 2019; Westphal et al., 2022). Combining the testable implications and distributional regression allows for deriving bounds on effects for non-complying groups that cannot be affected by the IV assumptions – by conveniently employing covariates.

The inability to control for a larger number of covariates in most of the proposed tests, while remaining computationally feasible, further limits their applicability in practice, particularly in settings where the exogeneity assumption holds only conditionally on various covariates (e.g., models with fixed effects). Two different test bases exist. The mean-based testable implications by Huber and Mellace (2015) and the density-based conditions derived by Kitagawa (2015). For mean effects like the local average treatement effect (LATE), Huber and Mellace's (2015) test is optimal to refute IV validity, as shown by Laffers and Mellace (2017). However, all tests have in common that validity cannot be confirmed explicitly (only invalidity). Further literature mainly builds on the density-based approach. Mourifié and Wan (2017) reformulate the testable conditions and propose another testing procedure, Sun (2023) improves the Kitagawa (2015) test procedure and allows the treatment to be multi-valued, and Arai et al. (2022) extends the density-based approach to fuzzy regression discontinuity designs. An alternative approach to test the density-based conditions is provided by Farbmacher et al. (2022), who uses causal forests to detect local violations of the LATE assumptions. Carr and Kitagawa (2023) contribute to the literature by extending Kitagawas (2015) test to the marginal treatment effect framework and proposing the first IV validity test, which can accommodate a moderate number of covariates.[1]

[1] There are other papers in the literature that concentrate on violations of one or two of the validity assumptions, e.g. Angrist and Imbens (1995), Mogstad et al. (2021), Machado et al. (2019), De Chaisemartin (2017) and Kédagni and Mourifié (2020).

Our testing approach is most closely related to the one of Huber and Mellace (2015) but differs in two main ways. First, we reduce the number of conditions tested from 4 to 2 already when defining the parameters for testing to leave out any non-binding conditions. Second, even though the testing conditions are based on mean potential outcomes, we make use of group-specific conditional density functions for which the covariates are held fixed at the mean. This allows us to point identify mean potential outcomes for pure groups and partially identify bounds for the unobserved mean potential outcomes of mixed groups conditional on covariates. Conditioning on covariates with the approach of Huber and Mellace (2015) is limited as it requires running the procedure in covariate-specific subsamples.

We conduct a simulation study to evaluate the test performance in different settings that directly relate to the random assignment assumption and the exclusion restriction. We consider processes where the instrument assignment does not depend on covariates and processes where it does. Thus, in the latter case, the instrument is only randomly assigned when conditioning on covariates. For each case, we further distinguish processes where the exclusion restriction holds and where the instrument directly affects the outcome. The simulation results reveal a good test performance in terms of size and power in finite samples. We complement the simulation results with results for two empirical applications from the literature. First, we apply our procedure to test the validity of the draft eligibility instrument from Angrist (1991) to analyze the effect of military service on civil earnings. The second application is from Card (1993), which studies monetary returns to education using college proximity as an instrument. In line with previous tests, we cannot reject IV validity for the draft eligibility instrument. For the college proximity instrument, our results indicate that the instrument is not valid without the inclusion of covariates. In contrast, a model including the covariates used by Card (1993) does not allow for refuting IV validity.

This paper proceeds as follows. Section 2 introduces the general econometric setup, and presents the LATE assumptions and their testable implications. The testing procedure is explained in detail in Section 3. Section 4 presents the simulation results, before the results for the two empirical applications are shown in Section 5. Section 6 concludes.

2 Setting and Assumptions

With a binary treatment D and a binary instrument Z, the key estimator for causal inference on the outcome Y is the so-called Wald estimator

$$\mathrm{IV}_{Wald} = \frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)}. \tag{1}$$

Note that we suppress covariates X in this simple setup. Angrist and Imbens (1995) show that with an additional set of assumptions, this ratio of mean differences has a causal interpretation as the local average treatment effect (LATE):

$$\mathrm{IV}_{Wald} = E(Y^1 - Y^0 \mid D^1 > D^0) := LATE \tag{2}$$

Here, Y^d is the potential outcome for treatment state $d \in \{0,1\}$. Hence, for every individual, $Y^1 - Y^0$ is their specific treatment effect. The LATE averages this individual treatment effect for a specific group of individuals – individuals who take the treatment because of the instrument. To derive this, Angrist and Imbens introduce another potential outcome dimension for the treatment: D^z, indicating the potential treatment choice with a specific value of the instrument $z \in \{0,1\}$. And correspondingly for the outcome, Y^{dz}.

We will now introduce the assumptions necessary to go from Eq. (1) to (2), which sets the path for testable implications on these assumptions.

Assumption 1 (Mean independence):
$E(Y^{dz} \mid Z = 1) = E(Y^{dz}|Z = 0)$ and $E(D^z \mid Z = 1) = E(D^z|Z = 0) \quad \forall d, z\{0,1\}$

> **Remark:** As we identify mean effects, not quantiles or probability densities, we only need this mean independence. In contrast, density-based testing approaches following Kitagawa (2015), base their tests on the stronger assumption of full independence: $Y^{d1}, Y^{d0}, D^1, D^0 \perp\!\!\!\perp Z$.

By the independence assumption, we can write for the numerator of Eq. (1):

$$E(Y \mid Z = 1) - E(Y \mid Z = 0) = E(Y^{d1} - Y^{d0}),$$

which simply is the causal effect of Z on Y (also called intent-to-treat or reduced-form effect). The expression $E(Y^{d1} - Y^{d0})$ means that the treatment state d is unrestricted and may vary from individual to individual in this difference, whereas the instrument state z is fixed. Analogously, we can rearrange the denominator of Eq. (1) through the independence assumption as follows:

$$E(D \mid Z = 1) - E(D \mid Z = 0) = E(D^1 - D^0)$$

We can then decompose the average causal effect of Z on D based on counterfactual treatment behavior.

$$\begin{aligned} E(D^1 - D^0) &= \Pr(D^1 = 1) - \Pr(D^0 = 0) \\ &= \Pr(D^1 = 1, D^0 = 1) + \Pr(D^1 = 1, D^0 = 0) \\ &\quad - \Big[\Pr(D^0 = 0, D^1 = 1) + \Pr(D^0 = 0, D^1 = 0)\Big] \end{aligned}$$

In principle, we can define and label the four possible types as always-takers (AT, defined by $D^1 = D^0 = 1$), compliers (C, $D^1 > D^0$), defiers (DF, $D^1 < D^0 = 1$) and never-takers (NT, $D^1 = D^0 = 0$). With this compact notation, we can simplify the equation above as:

$$\pi_{AT} + \pi_C - [\pi_{AT} + \pi_{DF}] \quad = \quad \pi_C - \pi_{DF}$$

Using these types, we can also decompose the numerator of Eq. (1) as

$$\begin{aligned} E(Y^{d1} - Y^{d0}) \quad = \quad & \pi_{NT} E(Y^{01} - Y^{00} | D^1 = D^0 = 0) + \pi_{AT} E(Y^{01} - Y^{00} | D^1 = D^0 = 1) \\ & + \pi_C E(Y^{01} - Y^{00} | D^1 = 1, D^0 = 0) + \pi_{DF} E(Y^{01} - Y^{10} | D^1 = 0, D^0 = 1) \end{aligned}$$

Conditional on the type, we only need the value of the instrument to infer treatment takeup. Thus, we use δ^z_{type} to denote the corresponding expected outcome. Then we write the above equation as:

$$E(Y^{d1} - Y^{d0}) = \pi_{NT}[\delta^1_{NT} - \delta^0_{NT}] + \pi_{AT}[\delta^1_{AT} - \delta^0_{AT}] + \pi_C[\delta^1_C - \delta^0_C] + \pi_{DF}[\delta^1_{DF} - \delta^0_{DF}]$$

Now, we use this notion, to rewrite Eq. (2) as:

$$\text{IV}_{Wald} \quad = \quad \frac{\pi_{NT}[\delta^1_{NT} - \delta^0_{NT}] + \pi_{AT}[\delta^1_{AT} - \delta^0_{AT}] + \pi_C[\delta^1_C - \delta^0_C] + \pi_{DF}[\delta^1_{DF} - \delta^0_{DF}]}{\pi_C - \pi_{DF}} \tag{3}$$

This expression is more complicated than Eq. (2). To give it the desired interpretation, we need to make additional assumptions.

Assumption 2 (Mean exclusion restriction):
$E(Y^{d,1}) = E(Y^{d,0})$ for $d \in \{0, 1\}$.

> **Remark:** Again we only need the exclusion restriction to hold in expectation for the identification of mean effects. IV validity conditions of Kitagawa (2015) require $Y^{d,1} = Y^{d,0}$ for $d \in \{0, 1\}$.

By the exclusion restriction, the instrument only affects Y through D, such that effects for always-takers and never-takers are nonexistent:

$$\text{IV}_{Wald} \quad = \quad \frac{\pi_C E(Y^{01} - Y^{00} | D^1 = 1, D^0 = 0) - \pi_{DF} E(Y^{10} - Y^{01} | D^1 = 0, D^0 = 1)}{\pi_C - \pi_{DF}} \tag{4}$$

The last step uses

Assumption 3 (Monotonicity): $Pr(D^1 \geq D^0) = 1$

By monotonicity, $\pi_{DF} = 0$, and the above expression simplifies to Eq. (2). Although the testable implications discussed in this paper may detect joint violations of assumptions 1–3, we will assume that monotonicity holds for the sake of notational clarity and because a violation of the monotonicity assumptions must be substantial to be detected. We refer the reader to De Chaisemartin (2017) for more details of such a violation. If we condition the expectations above additionally on D, not (the remaining) three, only one or two types contribute to these expectations. This insight, first used by Imbens and Rubin (1997), is the first step to seeing the consequences when an assumption is violated. For instance, if we condition on $D = 1$ and $Z = 1$, always-takers and compliers enter the expectation:

$$E(Y \mid D = 1, Z = 1) = \frac{\pi_C}{\pi_C + \pi_{AT}} \delta_C^1 + \frac{\pi_{AT}}{\pi_C + \pi_{AT}} \delta_{AT}^1 \tag{5}$$

If $Z = 0$, always-takers exclusively enter the expectation:

$$E(Y \mid D = 1, Z = 0) = \delta_{AT}^0$$

For the untreated case with $Z = 1$, only never-takers must contribute to the expectation:

$$E(Y \mid D = 0, Z = 1) = \delta_{NT}^1$$

If $Z = 0$, the expectation is mixed with never-takers and compliers:

$$E(Y \mid D = 0, Z = 0) = \frac{\pi_C}{\pi_C + \pi_{NT}} \delta_C^0 + \frac{\pi_C}{\pi_C + \pi_{NT}} \delta_{NT}^0 \tag{6}$$

3 Testable Implications, Testing Procedure, and Estimation

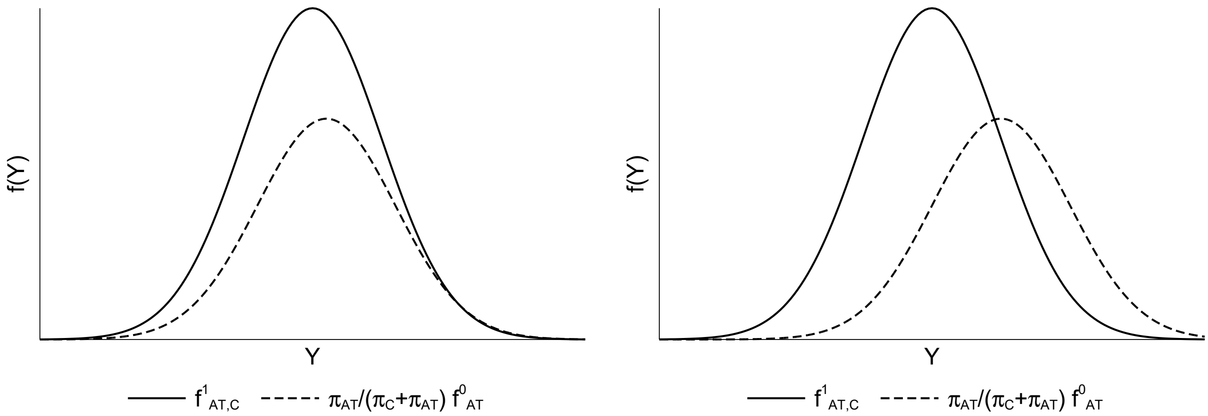
By assumptions 1–3, we take the always-takers' mean when $Z = 0$, δ_{AT}^0, and use Eq. (5) to infer the mean for the treated compliers, δ_C^1. This works, because the assumptions imply $\delta_{AT}^0 = \delta_{AT}^1$. Analogously, we can use the never-takers' mean δ_{NT}^1, equate it to δ_{NT}^0, and infer the mean of the untreated compliers according to Eq. (6).

If either one of the assumptions does not hold, $\delta_{AT}^1 \neq \delta_{AT}^0$ and/or $\delta_{NT}^1 \neq \delta_{NT}^0$. We can test whether this is likely to be fulfilled by using the type and Z-specific probability distribution functions, $f_{type}^z(Y)$, together with the fact that the equations do not only need to hold in expectation but also in distribution. Hence, the two treated expectations become:

$$\begin{aligned} f(Y \mid D = 1, Z = 1) &= \frac{\pi_C}{\pi_C + \pi_{AT}} f_C^1(Y) + \frac{\pi_{AT}}{\pi_C + \pi_{AT}} f_{AT}^1(Y) \\ &:= f_{AT,C}^1(Y) \end{aligned}$$

$$f(Y \mid D=1, Z=0) = f^0_{AT}(Y)$$

The implication of IV validity is $f^0_{AT}(Y) = f^1_{AT}(Y)$ for the treated case with $D = 1$ and $f^1_{NT}(Y) = f^0_{NT}(Y)$ for the untreated case with $D = 0$, This equivalence has testable implications for the observed distributions: the joint distributions of $f^1_{AT,C}(Y)$ and $f^0_{NT,C}(Y)$ must nest the normalized single-type distributions $\frac{\pi_{AT}}{\pi_{AT}+\pi_C} f^0_{AT}(Y)$ and $\frac{\pi_{NT}}{\pi_{NT}+\pi_C} f^1_{NT}(Y)$, respectively.



(a) No violation

(b) Violation

Figure 1: Graphical test of IV validity

Notes: Own illustration.

Figure 1 visualizes the two possible scenarios in Panels (a) and (b). Panel (a) displays a type-specific density (dashed line) that is compatible with the joint density (solid line). The two densities do not cross. Panel (b) displays the case where the two densities do cross. Testing whether the densities cross is the idea of the Kitagawa (2015) test with the underlying assumption of full conditional independence.

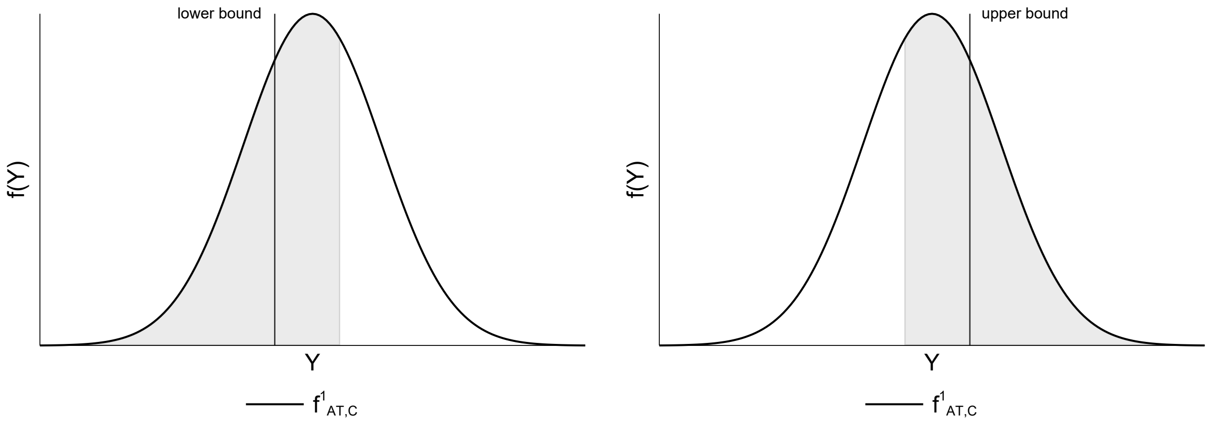
A similar (but not equivalent) implication of an incompatible distribution is that the mean of f^0_{AT} (δ^0_{AT}) must lie within the lower and upper bound of extreme-case scenarios. To keep notation and testing simple, we assume the outcome to be continuous. The lowest possible mean of the unobserved δ^1_{AT} results from $f(Y|D=1, Z=1)$ if the unobserved always-takers are placed in the lowest possible ranks of the distribution. As we know the share of always-takers, the lowest possible ranks are the first $q = \frac{\pi_{AT}}{\pi_{AT}+\pi_C}$ quantiles. This extreme-case scenario assumes the always-takers place from quantile 0 to q in the joint distribution, resulting in the lowest possible mean $\delta^{1,LB}_{AT}$. Formally, this reads

$$\delta^{1,LB}_{AT} = \int_0^{\frac{\pi_{AT}}{\pi_{AT}+\pi_C}} y \, dF(Y = y \mid D=1, Z=1). \tag{7}$$

The converse extreme case scenario is when the always-takers are placed in the highest q quantiles for the upper bound. This yields

$$\delta_{AT}^{1,UB} = \int_{\frac{\pi_{AT}}{\pi_{AT}+\pi_C}}^{1} y \, dF(Y = y \mid D = 1, Z = 1) \tag{8}$$

Figure 2 visualizes the lower and upper bounds for the joint treated distribution, which is the mean produced by the gray part of the distribution.



(a) Lower bound

(b) Upper bound

Figure 2: Graph of upper and lower bound

Notes: Own illustration. Shaded areas equals the q's proportion of the integral located in the lower (a) or upper (b) tail of the distribution. The vertical solid lines indicate the lower and upper bound of $E(Y^{1,1}|T^Z = AT^1, X)$.

For the untreated distributions, the extreme-case scenarios form if the never-takers place in the lowest or highest $r = \frac{\pi_{NT}}{\pi_{NT}+\pi_C}$ ranks of the joint $f_{NT,C}^0$ distribution.

$$\delta_{NT}^{0,LB} = \int_{0}^{\frac{\pi_{NT}}{\pi_{NT}+\pi_C}} y \, dF(Y = y \mid D = 0, Z = 0). \tag{9}$$

$$\delta_{NT}^{0,UB} = \int_{\frac{\pi_{NT}}{\pi_{NT}+\pi_C}}^{1} y \, dF(Y = y \mid D = 0, Z = 0) \tag{10}$$

We now have the two admissible intervals, which we use to compare the pure always-takers and never-taker means δ_{AT}^0 and δ_{NT}^0. The means are either

- compatible if $\delta_{AT}^0 \in \left[\delta_{AT}^{1,LB}, \delta_{AT}^{1,UB}\right]$ and $\delta_{NT}^1 \in \left[\delta_{NT}^{0,LB}, \delta_{NT}^{0,UB}\right]$. Then, we cannot reject IV validity. Or
- incompatible if either $\delta_{AT}^0 \notin \left[\delta_{AT}^{1,LB}, \delta_{AT}^{1,UB}\right]$ or $\delta_{NT}^1 \notin \left[\delta_{NT}^{0,LB}, \delta_{NT}^{0,UB}\right]$. Then, we can reject IV validity as one of assumptions 1–3 must be violated.

These testing equations are equivalent to but expressed in a different way than the testable implications derived by Huber and Mellace (2015). They are optimal to refute IV validity defined by assumptions 1-3 as long as the outcome is continuous (see Laffers and Mellace, 2017). Yet, as well as the Kitagawa (2015) testing conditions, they cannot verify IV validity. The probability of detecting a violation increases the narrower the bounds. Greater shares of always or never-takers compared to complier shares correspond to tighter bounds. Additionally, conditioning on covariates, especially those explaining most variation in the treatment selection or outcome, can tighten the bounds (Lee, 2009; Semenova, 2020). Huber and Mellace (2015) show how mean dominance assumptions can tighten the bounds or even result in equality constraints. This holds likewise for our approach, as we test the same identifying assumptions (conditional on covariates), which can help increase testing power. However, this might not be relevant in many applied settings, where mean dominance assumptions are less plausible than the IV validity conditions. Only one of the two conditions can be tested in settings with one-sided non-compliance that rule out the existence of always or never-takers.

With this notation, we can define the parameters that we test as

$$\theta_1 = \begin{cases} \delta^0_{AT} - \delta^{1,UB}_{AT} & \text{if } \delta^{1,LB}_{AT} < \delta^0_{AT} \\ \delta^{1,LB}_{AT} - \delta^0_{AT} & \text{else.} \end{cases}$$

for the treated case and

$$\theta_0 = \begin{cases} \delta^1_{NT} - \delta^{0,UB}_{NT} & \text{if } \delta^{0,LB}_{NT} < \delta^1_{NT} \\ \delta^{0,LB}_{NT} - \delta^1_{NT} & \text{else.} \end{cases}$$

for the untreated case. If IV validity is violated, θ_1 and/or θ_0 are strucurally larger than zero, meaning that the δ^0_{AT} and/or δ^1_{NT} lie outside their corresponding bounds. This defines our hypothesis as

$$H_0 : \begin{pmatrix} \theta_1 \\ \theta_0 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{11}$$

A positive θ indicates that δ^0_{AT} or δ^1_{AT} lie outside the admissible bounds, i.e., the means are incompatible with IV validity.

Estimation

Now, for the estimation approach covariates are explicitly expressed, as their implementation into an easy testing procedure is the key contribution of this paper. To determine θ_1

and θ_0, we need to estimate the type shares π_{AT}, π_{NT} and π_C. We do so by estimating the following first-stage equation

$$D_i = \pi_{AT} + \pi_C Z_i + \widetilde{X}_i'\delta + U_{Di} \tag{12}$$

where $\widetilde{X}$ indicates the demeaned covariate vector X. Since the covariates are held constant at their means and both D and Z are binary, the constant can be interpreted as the share of always-takers (always $D = 1$), and π_C as the share of compliers (D varies with Z).[2] Consequently, as the shares sum up to one, the share of never-takers is given by $\pi_{NT} = 1 - \pi_{AT} - \pi_C$.

Furthermore, for the conditional expected values entering θ_0 and θ_1, we estimate the conditional densities $f^1_{AT,C}(Y)$, $f^0_{AT}(Y)$, $f^1_{NT}(Y)$, and $f^0_{NT,C}(Y)$. To derive the conditional pdfs, we start by estimating the conditional cdfs for each observable group (determined by possible combinations of D and Z) given covariates with a distribution regression approach. $F(y) = Pr(Y \leq y|D = d, Z = z, \widetilde{X})$ is a binary choice model with the dependent variable $\mathbb{1}[Y \leq y]$ for an arbitrary threshold y.[3] Therefore, we run repeated binary choice models of the form

$$\begin{aligned}\mathbb{1}[Y \leq y] = F^0_{NT,C}(y)\mathbb{1}[D = 0]\mathbb{1}[Z = 0] + F^0_{AT}(y)\mathbb{1}[D = 1]\mathbb{1}[Z = 0]\\ + F^1_{NT}(y)\mathbb{1}[D = 1]\mathbb{1}[Z = 0] + F^1_{AT,C}(y)\mathbb{1}[D = 1]\mathbb{1}[Z = 1] + \widetilde{X}'\lambda + v\end{aligned} \tag{13}$$

with different thresholds y in the support of Y. Note that $F^0_{NT,C}$, F^0_{AT}, F^1_{NT}, and $F^1_{AT,C}$ are parameters estimated by this regression. They measure the share of observations conditional on $D = d$ and $Z = z$ below the threshold y, while all $\widetilde{X}$ are set to zero (and are, hence, fixed).[4] Repeating this regression for many y on the support of Y approximates the group-specific conditional cdf. By choosing a finer grid of values for y, one can approve the chance to describe $F(y)$ accurately. As the pdf is the derivative of the cdf, we estimate the slope of the conditional cdfs at each value for y. The slopes at every evaluation point can be estimated with kernel-weighted local polynomial regressions. This requires the choice of a kernel function and bandwidth. We follow Mourifié and Wan (2017) and use the rule-of-thumb choice by Fan and Gijbels (1996).[5]

Calculating the θs based on the estimated density function yields the estimated parameters $\widehat{\theta}_1$ and $\widehat{\theta}_0$. Still, bootstrap-based inference is needed to test the H_0 at given significance levels. Therefore, we generate B bootstrap samples of size N (number of observations)

[2] This interpretation is valid as long as Assumptions 1 and 3 hold. Without covariates, the (sum of) shares can easily be calculated with $\pi_{AT} = Pr(D = 1|Z = 0)$, $\pi_{AT} + \pi_C = Pr(D = 1|Z = 1)$, $\pi_{NT} = Pr(D = 0|Z = 1$, and $\pi_{NT} + \pi_C = \Pr(D = 0|Z = 0)$.

[3] Without further indication, it is implicit that all cdfs are given conditional on covariates.

[4] One could, instead of linear models, run, e.g., repeated logit models and use predictive margins for each group.

[5] This rule-of-thumb bandwidth choice is implemented in several STATA packages; for example, it is the default of the *lpoly* package.

randomly drawn from the original sample with replacement and indicated with $b \in \{1, 2, ..., B\}$. $\widehat{\theta}_{1,b}$ and $\widehat{\theta}_{0,b}$ denote the estimates calculated within every sample. Our p-value-based test is very similar to the simple bootstrap test with Bonferroni adjustment applied by Huber and Mellace (2015) except that we reduce the number of constraints already when defining the test parameters. To obtain p-values we recenter the parameter from each bootstrap sample, such that $\widetilde{\theta}_{1,b} = \widehat{\theta}_{1,b} - \widehat{\theta}_1$ and $\widetilde{\theta}_{0,b} = \widehat{\theta}_{0,b} - \widehat{\theta}_0$. This step, suggested by Hall and Wilson (1991), increases testing power if bootstrap samples are drawn from populations that do not satisfy H_0. To test the constraints of the H_0 against an upper-tailed alternative hypothesis separately, the bootstrap p-values for the treated and untreated cases are then given by

$$
\begin{aligned}
p_{\widehat{\theta}_1} &= \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \mathbb{1}[\widetilde{\theta}_{1,b} > \widehat{\theta}_1] \\
p_{\widehat{\theta}_0} &= \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \mathbb{1}[\widetilde{\theta}_{0,b} > \widehat{\theta}_0].^{6}
\end{aligned}
\tag{14}
$$

However, we want to perform a joint test on θ_1 and θ_0. The more conditions are tested, the higher the probability of obtaining an unusually high test statistic at random. Therefore, we apply the Šidák or Dunn-Šidák correction where the significance level for each test is set to $\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}}$ with m being the number of tests and α the overall significance level (Šidák, 1967). For the p-value of the joint test follows that $\widehat{p} = 1 - (1 - \min(p_{\widehat{\theta}_1}, p_{\widehat{\theta}_0}))^m$. Even though slightly less conservative than the Bonferroni correction, note that the Šidák correction can still be too conservative when the m is large and the test statistics are positively correlated (MacKinnon, 2009). With $m = 2$ in our case, we have the least conditions tested simultaneously. If the test statistics are not independent, the resulting p-value $\widehat{p}$ is still an upper bound and $\min(p_{\widehat{\theta}_1}, p_{\widehat{\theta}_0})$ the lower bound in the extreme case of perfectly correlated statistics (MacKinnon, 2009). Hence, consulting $p_{\widehat{\theta}_1}$ and $p_{\widehat{\theta}_0}$ as well as the Šidák corrected p-value for the joint test $\widehat{p}$ should be enough to judge on the H_0 or not in most settings.

To summarize, we conduct the following step-by-step implementation:

1. Demean covariates to get $\widetilde{X}$.
2. Estimate shares of types with first stage regression (Eq. (12)) and calculate q and r.
3. Set a grid for evaluation points within the support of Y (e.g., quantiles of the observed distribution of Y).[7]

[6]This follows from the fact that we want to reject our H_0 when the observed value of our test statistic $\widehat{T}$ is in the upper tail of $F(T)$, the cdf of T under the H_0. The distribution of the bootstrap test statistics $\widehat{T}_b$ gives the empirical distribution function $\widehat{F}$, i.e., the asymptotic approximation of F. Then, the bootstrap p-value is $p_{\widehat{\theta}} = 1 - \widehat{F}(\widehat{T}) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}[\widehat{T}_b > \widehat{T}]$ (see MacKinnon, 2009). Plugging in $\widehat{T}_b = \sqrt{N}\,(\widehat{\theta}_b - \widehat{\theta})/\sigma_{\widehat{\theta}}$ and $\widehat{T} = \sqrt{N}\,\widehat{\theta}/\sigma_{\widehat{\theta}}$ yields the simplified version in Eq. (14).

[7]As quantiles use to bunch in the middle of a unimodal distribution, one might want to use more dense evaluation points in the tails of the distribution.

4. Estimate conditional cdfs with repeated regressions of binary choice models (Eq. (13)).
5. Determine the slopes at the evaluation points to get conditional pdfs (e.g., with local linear regression).
6. Calculate conditional means δ^0_{AT} and δ^1_{NT} as well as lower and upper bounds $\delta^{1,LB}_{AT}$, $\delta^{1,UB}_{AT}$, $\delta^{0,LB}_{NT}$ and $\delta^{0,UB}_{NT}$ according to equations (7–10).
7. Determine θ_1 and θ_0 by plugging in results from step 6.
8. Conduct inference on both parameters, i.e., derive bootstrapped inference by repeating steps 1 to 7 with B bootstrap samples of size N of the original sample (B =number of bootstrap repetitions, N =number of observations), derive the corresponding p-values according to Eq. (14) and apply the Šidák method to obtain one p-value for the joint test

4 Simulation

We perform Monte Carlo exercises to evaluate our testing procedure's size and power (the probabilities of falsely and correctly rejecting H_0, respectively). We consider the general data-generating process (DGP) similar to the simulation study in Carr and Kitagawa (2023). We simulate this DPG $S = 1000$ times, with potentially different random parameters for each simulation. We bootstrap-replicate each simulation $B = 99$ times with the same parameter value to generate a p-value. The DGP reads

$$Y = X'\beta_X + \beta_D D + \beta_Z Z + U$$

$$D = \mathbb{1}[\pi_0 + \pi_1 Z + X'\pi_X + U_D \geq 0]$$

$$\text{with } \pi_0 = \Phi^{-1}(0.45) \text{ and } \pi_1 = \Phi^{-1}(0.55) - \Phi^{-1}(0.45)$$
$$\text{(implying } \pi_{AT} = 0.45, \pi_C = 0.1, \text{ and } \pi_{NT} = 0.45)$$

$$Z = \mathbb{1}[X'\gamma + U_Z \geq 0]$$

$$X = (X_1, X_2, X_3);\ X_j \sim N(0, I)\forall j \in \{1, 2, 3\}$$

$$U_Z \sim N(0, 1)$$

$$U, U_D \sim N(0, \Sigma) \qquad \text{with } \Sigma = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix},$$

where $\Phi^{-1}()$ is the inverse of the standard normal distribution to better control the (non)complier shares. For each simulation draw, each element of π_X $(\pi_{X,1}, \pi_{X,2}, \pi_{X,3})$ and β_X $(\beta_{X,1}, \beta_{X,2}, \beta_{X,3})$ is drawn from a uniform distribution on the $[-1, 1]$ interval.

Monotonicity (Assumption 3) is fulfilled by construction, as α_0 and α_1 are constant, i.e., the same for every i.

Our simulation focuses on potential violations of Assumptions 1 and 2. We distinguish cases where the independence assumption only holds conditional on X and cases where the exclusion restriction is violated, meaning that the instrument directly affects Y. Specifically, we assess the implications of varying γ and β_Z:

- **Violation of the independence assumption (assumption 1):**
  - Independence holds unconditionally: $\gamma_1 = \gamma_2 = \gamma_3 = 0$
  - Independence assumption potentially violated when not conditioning on X: $\gamma_j \sim U(-1,1)\ \forall j \in \{1,2,3\}$
- **Violation of the exclusion restriction (assumption 2)**
  - Exclusion restriction holds: $\beta_Z = 0$
  - Exclusion restriction violated: $\beta_Z = 1$

We apply our test procedure to each simulated or replicated sample with and without conditioning on covariates X. We summarize results by the rejection rates, which we compute as

$$\text{Rejection rate} = \frac{1}{S}\sum_{s\in S} \mathbb{1}\left[p_{\widehat{\theta}} \leq \text{Nominal size}\right].$$

We consider different nominal sizes $\in \{0.1, 0.05, 0.01\}$ for $S = 30$ simulations (which we currently extend to 1000). Šidák adjusted p-values $p_{\widehat{\theta}}$ are calculated based on $B = 99$ bootstrap repetitions. We present results for sample sizes of 250 and 1000 in Table 1.

Under IV validity ($\beta_Z = 0$ and either with covariates or $\gamma_j = 0$), our testing procedure with covariates yields rejection rates below nominal size independent of the γs already for the smaller sample size of 250. The rejection rates are lower than nominal sizes because the DGP defines a setup that is not at the boundary of the testing condition to hold. Without conditioning on covariates, the test still performs well as long as there is no effect of X on Z (columns 1-3). For elements of γ being non-zero (columns 4-6), X and Z are not independent, which (depending on the employed covariates) may violate mean independence (Assumption 1). These potential violations are reflected in the higher rejection rates for both sample sizes. They can be interpreted as evidence for the importance of including covariates in the model and test whenever the assignment of Z is not unconditionally random, and the potential confounder (here X) is observed. When the exclusion restriction is violated ($\beta_Z = 1$), the instrument is invalid, i.e., H_0 should be rejected. Regardless of whether Z is independent of X, rejection rates are well above the nominal size for both sample sizes. Overall, the results show that the test performs

	Z and X are independent ($\gamma_1 = \gamma_2 = \gamma_3 = 0$)			**Z depends on X** ($\gamma_1, \gamma_2, \gamma_3 \sim U(-1,1)$)		
Nominal size:	0.1	0.05	0.01	0.1	0.05	0.01
Exclusion restriction holds: $\beta_Z = 0$						
w/ covariates						
N=250	0.03	0.03	0.00	0.03	0.00	0.00
N=1000	0.00	0.00	0.00	0.00	0.00	0.00
w/o covariates						
N=250	0.07	0.03	0.00	0.30	0.07	0.03
N=1000	0.07	0.03	0.03	0.40	0.33	0.27
Exclusion restriction violated: $\beta_Z = 1$						
w/ covariates						
N=250	0.90	0.83	0.70	0.80	0.70	0.50
N=1000	1.00	1.00	1.00	1.00	0.97	0.93
w/o covariates						
N=250	0.73	0.67	0.60	0.67	0.60	0.40
N=1000	0.97	0.97	0.93	0.87	0.83	0.80

Notes: The rejection rates are based on the Šidák adjusted p-values. When $\beta_Z = 0$, the instrument is valid (white cells) except in columns 4-6 without including conditioning covariates (light gray cells), where X and Z are not independent. When $\beta \neq 0$, the exclusion restriction does not hold; hence, the instrument is invalid (dark gray cells). Additionally, X and Z are not independent in columns 4-6 without conditioning on covariates (row 4).

Table 1: Simulations

well in size and power and is superior to the test without covariates whenever there are (observed) confounders correlated with Z and Y.

5 Applications

We apply our testing approach to two well-known settings that have also been considered by Mourifié and Wan (2017), Kitagawa (2015), Sun (2023) and Huber and Mellace (2015) to show the performance of their testing procedures. The first relies on the Vietnam-era draft lottery instrument by Angrist (1991), and the second one from Card (1993) exploits the college proximity as an instrument.

5.1 Earning effects of military service – Draft Lottery Instrument

In the first empirical application, we use the draft eligibility instrument from Angrist (1991) to study the effect of veteran status on earnings. An IV approach is applied here because of the self-selection mechanism into military service that is potentially related to later earnings. The instrument is a binary variable for draft eligibility. As the instrument's assignment procedure is a lottery based on the individual's birth month, the instrument should be randomly assigned, meaning independence holds. The monotonicity condition is also very credible in this setting, as the existence of defiers is hard to imagine. However, the exclusion restriction could be violated. Young men eligible for the draft might have intended to escape or at least defer military service, e.g., by staying in college longer than they would have otherwise. More years of education, in turn, might increase wages, which is why a positive effect of the instrument for the never-takers seems plausible here.

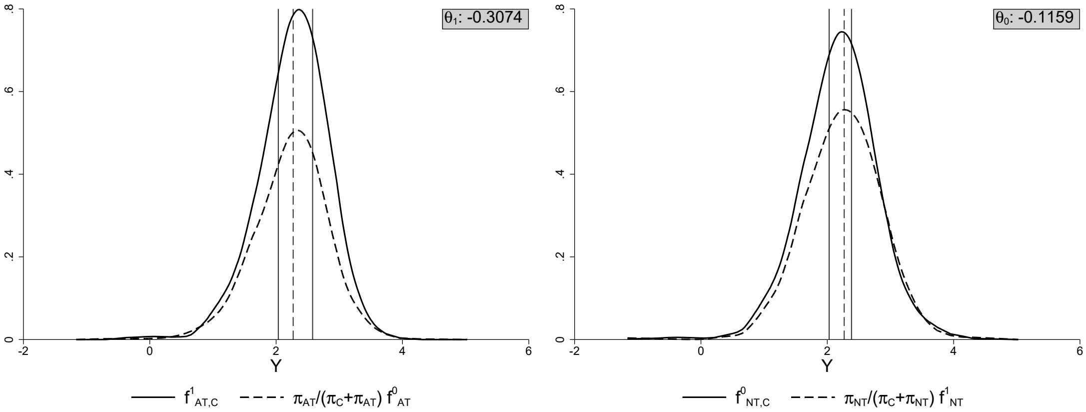
The data we use is from the 1984 Survey of Income and Program Participation (SIPP).[8] The final sample without missings consists of 3,071 individuals. The treatment is $D = 1$ if the individual has a veteran status, and the instrument is $Z = 1$ if the individual was eligible for the draft. The outcome Y is given as the logarithm of weekly wages. Following Angrist (1990), who studies the effect of the lottery on lifetime earnings, we add dummies for the birth cohort and a race indicator as covariates. Graphical results are presented in figure 3 where panel (a) shows the densities and (bounds of) mean potential outcomes without and panel (b) with covariates. The left graph for each panel belongs to the treated state, i.e., relevant for θ_1, and the right graph to the untreated state, i.e., relevant for testing θ_0. The estimates for the corresponding θs are given in the upper right corner. The graphical evidence shows that the validity conditions hold, as the dashed vertical lines lie within the solid vertical lines, indicating the bounds. This holds with and without conditioning on covariates. Hence, we cannot refute IV validity here. However, comparing panels (a) and (b), we see that conditioning on covariates narrows the bounds.

The same result can be seen in the left side of table 2, where θ_1 and θ_0 are negative without and with conditioning on covariates. The p-values for both θs in both models, as well as the Šidák corrected p-values, equal 1. Therefore, these results do not allow for a rejection of IV validity. This aligns with the findings of Kitagawa (2015) and Mourifié and Wan (2017), who apply their tests to the same setting without conditioning on covariates. The results for individual subgroups distinguished by race and educational attainment from Mourifié and Wan (2017) are also not interpreted as evidence against a valid instrument.

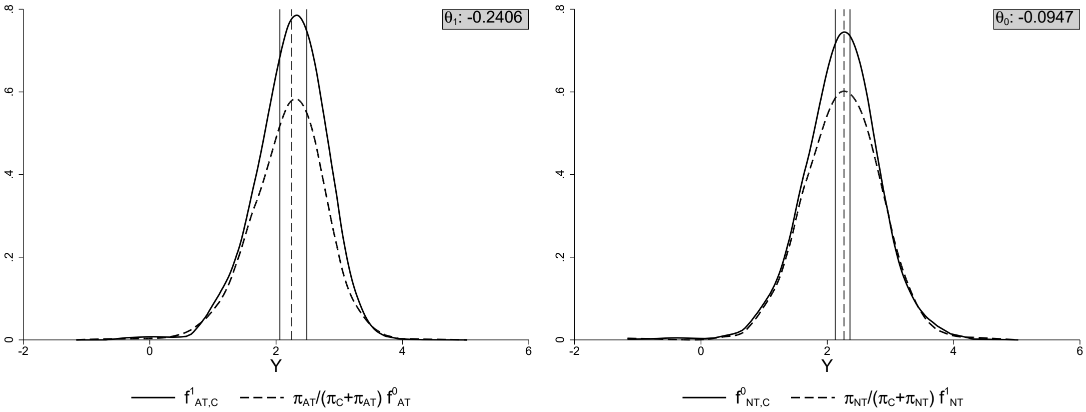
5.2 Returns to Education – College Proximity Instrument

The second empirical example we apply our procedure to is from Card (1993), who analyzed the effect of college education on earnings by exploiting college proximity as a source of external variation. Unobserved individual characteristics like innate ability

[8]The data is available in the Review of Economics and Statistics Dataverse as replication data for Mourifié and Wan (2017).



(a) Without covariates



(b) With covariates[a]

Figure 3: Graphs for the draft lottery instrument

Notes: Own illustration based on SIPP data. Pdfs for the mixed groups are given by the solid curves, and for the single groups, they are given by the dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta^{1,LB/UB}_{AT}$ (left) and $\delta^{0,LB/UB}_{NT}$ (right), and the vertical dashed lines display the conditional mean δ^{0}_{AT} (left) and δ^{1}_{NT} (right). f^{0}_{AT} and f^{1}_{NT} are down-weighted by their relative shares. This does not affect the mean potential outcome given by the dashed vertical line.
[a]Dummies for the birth cohort and a dummy for being non-white are used as covariates.

are likely to correlate with educational choice and later wages, yielding an endogeneity problem. Proximity to a college is employed as an instrumental variable based on the premise that a nearby college lowers the cost of pursuing college education by enabling students to live at home. In this setting, compliers are individuals from lower-income families who would not have attended college without the option to live with their parents. Unobserved individual abilities are assumed to be independent of their residential location during teenage age. However, the instrument may be correlated with factors like local labor market conditions or family background, which could also affect the outcome. By including several covariates in his model, this has been regarded by Card (1993).

The data is derived from the National Longitudinal Survey of Young Men (NLSYM), which followed a cohort of men aged 14–24 in 1966 with follow-up surveys through 1981[9]. Based

[9]The prepared dataset is available in the Review of Economics and Statistics Dataverse as replication data for Mourifié and Wan (2017)

	Draft lottery		College proximity	
	w/o covariates	w/ covariates[a]	w/o covariates	w/ covariates[b]
θ_1	-0.307	-0.241	-0.211	-0.082
$p_{\hat{\theta}_1}$	1.000	1.000	1.000	0.964
θ_0	-0.116	-0.095	0.086	0.013
$p_{\hat{\theta}_0}$	1.000	1.000	0.000	0.349
Šidák corrected $\hat{p}$	1.000	1.000	0.000	0.577
Shares				
π_C	0.139	0.088	0.069	0.035
π_{AT}	0.265	0.288	0.225	0.248
π_{NT}	0.596	0.623	0.707	0.718
No. evaluation points	256		344	
Observations	3027		3010	

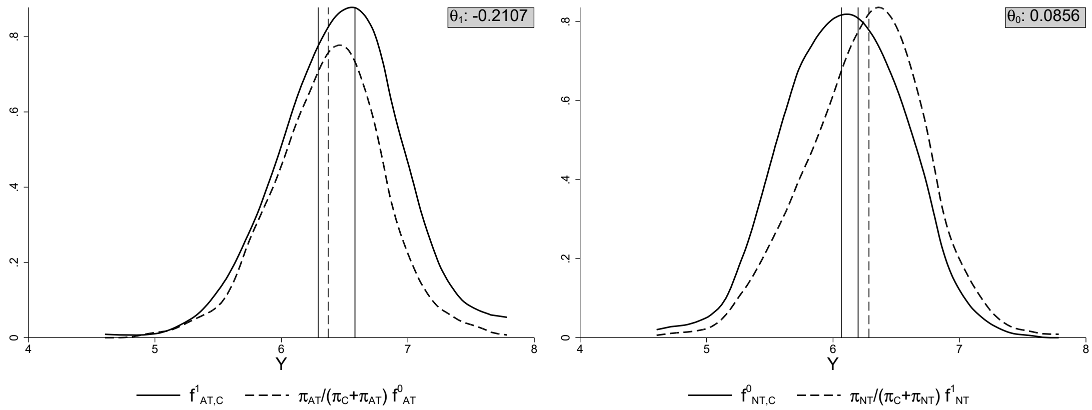
Notes: Tests are based on 999 bootstrap samples. [a]Dummies for birth cohorts and a dummy for non-white. [b]Dummy variables indicating race being black, residence in a standard metropolitan area (SMSA) in 1966 and 1976, region of residence in 1966, living in the south in 1976, living with both parents at age 14, and living with the mother only at age 14. Variables representing parents' years of education take on the value of the overall mean if they are missing. Dummies for missing fathers' and mothers' education have also been added.

Table 2: Results of the empirical applications

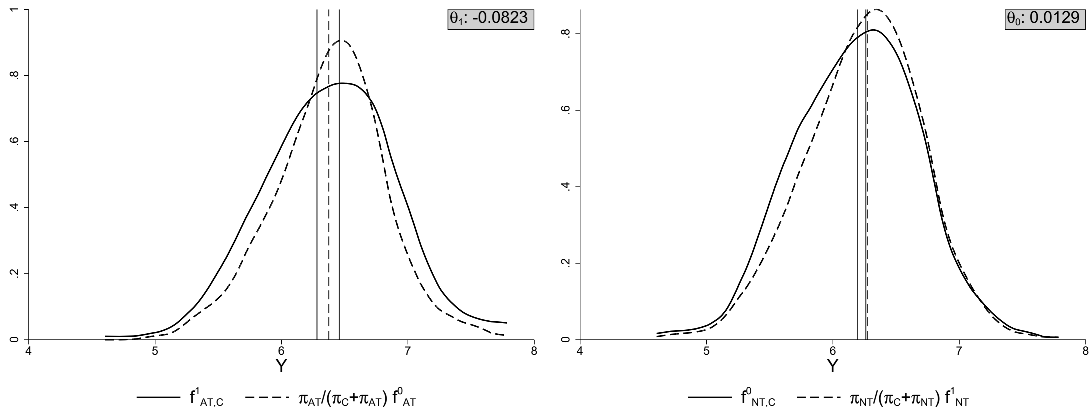
on the respondent's county of residence in 1966, the dataset includes a binary instrumental variable on the availability of a four-year college in the local labor market. Information on educational attainment and wages is used from the 1976 follow-up survey. We deviate from the original study and follow Kitagawa (2015) by defining a binary treatment D for having 16 or more years of education in 1976, approximating a four-year college degree measure. The binary instrument Z indicates if the individual grew up near a four-year college. The logarithm of weekly earnings in 1976 is used as the outcome variable Y. We apply the test without covariates first, then conditional on race, region, residence in a metropolitan area, family structure at age 14, and parents' education to increase the credibility of the random assignment assumption.[10] Thereby, we include all covariates determined prior to treatment assignment used by Card (1993), himself, besides interactions of parents' education. The final sample size after dropping observations with missing wages is 3,010.

Figure 4 provides the graphical results for the college proximity instrument. As in figure 3, panel (a) shows the densities and (bounds of) mean potential outcomes without and panel (b) with covariates. Again, the estimates for the θs are shown in the right upper corner. The graphical evidence from the left side shows that the validity condition for the treated state holds with and without conditioning on covariates. For the untreated case on the right side, the mean for the never takers with $Z = 1$, δ^1_{NT} (dashed vertical line), lies outside

[10]The detailed list of covariates is shown under table 2.



(a) Without covariates



(b) With covariates[a]

Figure 4: Graphs for the college proximity instrument

Notes: Own illustration based on NLSYM data. Pdfs for the mixed groups are given by the solid curves, and for the single groups, they are given by the dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta_{AT}^{1,LB/UB}$ (left) and $\delta_{NT}^{0,LB/UB}$ (right), and the vertical dashed lines display the conditional mean δ_{AT}^{0} (left) and δ_{NT}^{1} (right). f_{AT}^{0} and f_{NT}^{1} are down-weighted by their relative shares. This does not affect the mean potential outcome given by the dashed vertical line. [a]Dummy variables indicating race being black, residence in a standard metropolitan area (SMSA) in 1966 and 1976, region of residence in 1966, living in the south in 1976, living with both parents at age 14, and living with the mother only at age 14. Variables representing parents' years of education take on the value of the overall mean if they are missing. Additionally, dummies for missing father's and missing mother's education are added.

the bounds for the mean for the never takers with $Z = 0$, $\delta_{NT}^{0,LB}$ and $\delta_{NT}^{0,UB}$ (solid vertical lines). Hence, we reject IV validity based on graphical evidence. The positive θ_0 estimate here indicates the distance from the dashed vertical line to the closer solid vertical line, i.e., the deviation of the testable condition.

Including covariates narrows bounds and also lowers the deviation in the untreated case from 0.0856 to 0.0129. As the inclusion of covariates should decrease any concerns about the random assignment of the instrument, a lower deviation is expected. Finally, inference on the deviation for the untreated state is necessary to conclude whether the H_0 can be rejected. Results from the right side of table 2 show a p-value of 0 for θ_0 and the Šidák correction without conditioning on covariates and, thus, are interpreted as evidence against the H_0 of a valid instrument. Including covariates yields a p-value of 0.349 for θ_0 and a Šidák corrected p-value of 0.577 for multiple testing. Both values do not allow for a rejection of the H_0. Hence, we conclude that once we control for covariates, we cannot reject the validity of the college proximity instrument. As the bounds are quite narrow and H_0 is not rejected, it seems very plausible that the instrument is truly valid. Kitagawa (2015) and Huber and Mellace (2015) draw the same conclusion on their results with and without controlling for covariates. Whereas Mourifié and Wan (2017) still rejects IV validity by testing in different subsamples, thereby controlling for three covariates. This result can be attributed to their limited number of controls, especially not controlling for parents' education.

6 Conclusion

This paper proposes an easily implementable testing procedure based on distribution regressions that allows testing the LATE assumptions conditional on covariates without drastically increasing computation times. We use group-specific conditional distribution estimates to derive bounds on unobserved mean potential outcomes that we compare to observed mean potential outcomes for testing the mean-based testable implications derived by Huber and Mellace (2015). Performing Monte Carlo exercises, we showed that the testing procedure performs well in finite sample sizes. We applied the test to the draft eligibility and college proximity instruments from the literature. We could not reject IV validity for the draft eligibility, even when including covariates. For the college proximity instrument, instead, we find that the rejection of the instrument's validity depends on the inclusion of conditioning covariates.

References

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, 80(3):313–336.

Angrist, J. D. (1991). The draft lottery and voluntary enlistment in the vietnam era. *Journal of the American Statistical Association*, 86(415):584–595.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Arai, Y., Hsu, Y.-C., Kitagawa, T., Mourifié, I., and Wan, Y. (2022). Testing identifying assumptions in fuzzy regression discontinuity designs. *Quantitative Economics*, 13(1):1–28.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research.

Carr, T. and Kitagawa, T. (2023). Testing instrument validity with covariates. Papers, arXiv.org.

Chernozhukov, V. and Hansen, C. (2005). An IV Model of Quantile Treatment Effects. *Econometrica*, 73(1):245–261.

De Chaisemartin, C. (2017). Tolerating defiance? Local average treatment effects without monotonicity. *Quantitative Economics*, 8(2):367–396. Publisher: Wiley Online Library.

Dong, Y. (2019). Regression discontinuity designs with sample selection. *Journal of Business & Economic Statistics*, 37(1):171–186.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press.

Farbmacher, H., Guber, R., and Klaassen, S. (2022). Instrument validity tests with causal forests. *Journal of Business & Economic Statistics*, 40(2):605–614.

Frandsen, B. R., Frölich, M., and Melly, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, 168(2):382–395.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2):757–762.

Huber, M. and Mellace, G. (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics*, 97(2):398–411.

Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.

Kédagni, D. and Mourifié, I. (2020). Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675.

Laffers, L. and Mellace, G. (2017). A note on testing instrument validity for the identification of late. *Empirical Economics*, 53:1281–1286.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Machado, C., Shaikh, A. M., and Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*, 212(2):522–555.

MacKinnon, J. G. (2009). *Bootstrap Hypothesis Testing*, chapter 6, pages 183–213. John Wiley Sons, Ltd.

Mogstad, M. and Torgovitsky, A. (2024). Chapter 1 - Instrumental variables with unobserved heterogeneity in treatment effects. In Dustmann, C. and Lemieux, T., editors, *Handbook of Labor Economics*, volume 5, pages 1–114. Elsevier.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–98.

Mourifié, I. and Wan, Y. (2017). Testing local average treatment effect assumptions. *The Review of Economics and Statistics*, 99(2):305–313.

Semenova, V. (2020). Better lee bounds. *arXiv preprint arXiv:2008.12720*.

Sun, Z. (2023). Instrument validity for heterogeneous causal effects. *Journal of Econometrics*, 237(2, Part A):105523.

Westphal, M., Kamhöfer, D. A., and Schmitz, H. (2022). Marginal college wage premiums under selection into employment. *The Economic Journal*, 132(646):2231–2272.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.