ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Heckman, James Joseph; Schmierer, Daniel; Urzúa, Sergio

Working Paper

Testing the correlated random coefficient model

IZA Discussion Papers, No. 4525

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Heckman, James Joseph; Schmierer, Daniel; Urzúa, Sergio (2009) : Testing the correlated random coefficient model, IZA Discussion Papers, No. 4525, Institute for the Study of Labor (IZA), Bonn,
https://nbn-resolving.de/urn:nbn:de:101:1-20091105968

This Version is available at:
https://hdl.handle.net/10419/36304

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

DISCUSSION PAPER SERIES

IZA DP No. 4525

Testing the Correlated Random Coefficient Model

James J. Heckman
Daniel Schmierer
Sergio Urzua

October 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Testing the Correlated Random Coefficient Model

James J. Heckman
University of Chicago, University College Dublin, Yale University, American Bar Foundation and IZA

Daniel Schmierer
University of Chicago

Sergio Urzua
Northwestern University and IZA

Discussion Paper No. 4525
October 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 4525
October 2009

ABSTRACT

Testing the Correlated Random Coefficient Model

The recent literature on instrumental variables (IV) features models in which agents sort into treatment status on the basis of gains from treatment as well as on baseline-pretreatment levels. Components of the gains known to the agents and acted on by them may not be known by the observing economist. Such models are called correlated random coefficient models. Sorting on unobserved components of gains complicates the interpretation of what IV estimates. This paper examines testable implications of the hypothesis that agents do not sort into treatment based on gains. In it, we develop new tests to gauge the empirical relevance of the correlated random coefficient model to examine whether the additional complications associated with it are required. We examine the power of the proposed tests. We derive a new representation of the variance of the instrumental variable estimator for the correlated random coefficient model. We apply the methods in this paper to the prototypical empirical problem of estimating the return to schooling and find evidence of sorting into schooling based on unobserved components of gains.

JEL Classification: C31

Keywords: correlated random coefficient, testing, instrumental variables, power of tests based on IV

Corresponding author:

James J. Heckman
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago, IL 60637
USA
E-mail: jjh@uchicago.edu

1 Introduction

The correlated random coefficient model is the new centerpiece of a large literature in microeconometrics. For person i, it expresses outcome Y_i in terms of choice indicator D_i as

$$Y_i = \alpha_i + \beta_i D_i \tag{1}$$

where $D_i = 1$ if a choice is made; $D_i = 0$ if not and both the intercept, α_i, and the slope, β_i, vary among persons. In this expression both the α_i and β_i may depend on regressors X_i which we keep implicit.

β_i is the causal effect of D_i on Y_i holding α_i fixed. If agents make their choices to take treatment based on components of β_i that depend on variables not available to the observing economist, D_i is correlated with β_i even after conditioning on X_i. Most recent studies focus on estimating means or quantiles of the distribution of β_i.[1]

The model that motivated the research of a previous generation (see, e.g., Griliches, 1977) assumes no response heterogeneity ($\beta_i = \beta$). The correlated random coefficient model assumes that β_i varies in the population and in addition that

$$\text{Cov}\left(D_i, \beta_i\right) \neq 0. \tag{C-1}$$

The model also accounts for selection on intercepts, *i.e.* selection on pretreatment unobservables:

$$\text{Cov}\left(D_i, \alpha_i\right) \neq 0. \tag{C-2}$$

When (C-1) holds, marginal returns to an activity in general differ from average returns. When assumption (C-2) holds but D_i is independent of β_i, standard IV identifies the mean of β_i, which we denote by $\bar{\beta}$. This configuration of assumptions includes the case when β_i is

[1] Abbring and Heckman (2007) discuss methods for estimating the distribution of β_i.

random but independent of D_i and the case when β_i is the same for everyone.[2,3]

As first noted by Heckman and Robb (1985), instrumental variables (IV) applied to (1) when (C-1) holds produces an instrument-dependent parameter that, in general, is not $\bar{\beta}$.[4] In general, different instruments identify different parameters. Under conditions specified in Yitzhaki (1989),[5] Imbens and Angrist (1994), Heckman and Vytlacil (1999), and Heckman, Urzua, and Vytlacil (2006), IV estimates weighted averages of marginal effects. Heckman and Vytlacil (1999, 2001, 2005, 2007a) generalize the marginal treatment effect (MTE) introduced by Björklund and Moffitt (1987) and show that the MTE plays the role of a policy-invariant functional that is invariant to the choice of instrument. The MTE can be used to unify the literature on treatment effects.[6]

Heckman and Vytlacil (2001, 2005, 2007b) derive testable implications of the hypothesis that β_i is statistically independent of D_i given X_i:

$$H_0 : \beta_i \perp\!\!\!\perp D_i \mid X_i,$$

where $A \perp\!\!\!\perp B \mid C$ means A is independent of B given C. In this paper, we develop formal tests of this hypothesis and analyze their power. We apply our tests to a prototypical problem of estimating the return to schooling.

The paper proceeds as follows. Section 2 establishes the equivalence of the correlated random coefficient model with the Generalized Roy model. We state two testable implications of it. One test exploits the insight that, in general, in the case when H_0 is false, different instruments identify different parameters. We develop a test based on this principle in Section 3 after first presenting some new results on the sampling distribution of the instru-

[2]See Heckman and Vytlacil (1998), Heckman and Vytlacil (2007a,b). The standard "ability bias" problem (Griliches, 1977) assumes that $\beta_i = \beta$, a constant for all i, and that $\text{Cov}(D_i, \alpha_i) \neq 0$.

[3]Evidence from parametric models on the empirical relevance of (C-1) in a variety of areas of economics is presented in Heckman (2001, Table 3).

[4]See the discussion of the ensuing literature in Heckman, Urzua, and Vytlacil (2006) or Heckman and Vytlacil (2007a,b).

[5]Posted at website for Heckman, Urzua, and Vytlacil (2006), see `http://jenni.uchicago.edu/underiv/`.

[6]See Heckman and Vytlacil (2005).

mental variable estimator and considering the problem of constructing power functions for tests of hypotheses in the correlated random coefficient model. Section 4 develops tests for a second implication of the correlated random coefficient model. Section 5 analyzes the power of the proposed tests. Section 6 applies these tests to a prototypical problem: estimating the return to schooling using the model of Carneiro, Heckman, and Vytlacil (2006). Section 7 concludes.

2 Equivalence with the Generalized Roy Model and Two Testable Implications of H_0

An alternative way to represent equation (1) makes the link to economic choice theory more explicit. Individual i experiences outcome $Y_{1,i}$ if $D_i = 1$ and outcome $Y_{0,i}$ if $D_i = 0$, $i = 1, \ldots, I$. The observed outcome is $Y_i = D_i Y_{1,i} + (1 - D_i) Y_{0,i}$.[7] Let $\mu_j(X_i) = E(Y_{j,i} \mid X_i)$, $j \in \{0, 1\}$. One can write the model for potential outcomes conditional on X_i as $Y_{1,i} = \mu_1(X_i) + U_{1,i}$ and $Y_{0,i} = \mu_0(X_i) + U_{0,i}$ where $E(U_{j,i} \mid X_i) = 0$, $j \in \{0, 1\}$. In this notation, the observed outcome is

$$Y_i = \mu_0(X_i) + [\mu_1(X_i) - \mu_0(X_i) + U_{1,i} - U_{0,i}] D_i + U_{0,i}.$$

This is the correlated random coefficient model of equation (1) where the baseline outcome is $\alpha_i = \mu_0(X_i) + U_{0,i}$ and the gain is $\beta_i = \mu_1(X_i) - \mu_0(X_i) + U_{1,i} - U_{0,i}$ where, for notational simplicity, we suppress the dependence of α_i and β_i on X_i. To simplify the expressions, we drop the i subscripts throughout the rest of the paper unless their use clarifies the discussion. We define $\alpha = \overline{\alpha} + U_\alpha$ and $\beta = \bar{\beta} + U_\beta$ where $E(U_\alpha \mid X) = 0$ and $E(U_\beta \mid X) = 0$. Table 1 shows the equivalent parameters for the two models.

Whether the null hypothesis H_0 is true or not depends on the underlying choice model.

[7] This is in the form of a Quandt (1958) switching regression model.

Table 1: Equivalence of Notation Between the Correlated Random Coefficient Model and the Generalized Roy Model. All parameters are defined conditional on X_i, which is left implicit.

	Generalized Roy model	Correlated random coefficient model
Baseline outcome	$Y_{0,i} = \mu_0 + U_{0,i}$	α_i
Outcome in treated state	$Y_{1,i} = \mu_1 + U_{1,i}$	$\beta_i + \alpha_i$
Gain to treatment (Individual causal effect)	$Y_{1,i} - Y_{0,i} = \mu_1 - \mu_0 + U_{0,i} - U_{0,i}$	β_i
Outcome	$Y_i = Y_{0,i} + D_i (Y_{1,i} - Y_{0,i})$ $= \mu_{0,i} + (\mu_{1,i} - \mu_{0,i} + U_{1,i} - U_{0,i}) D_i + U_{0,i}$	$Y_i = \alpha_i + \beta_i D_i$

We postulate a threshold crossing model which assumes separability between observables Z that affect choice and an unobservable V: $D = \mathbf{1}(\mu_D(Z) - V \geq 0)$, where $\mathbf{1}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and is 0 otherwise, and μ_D is a deterministic function of Z.[8] Z can include components of X. Letting F_V be the distribution of V conditional on X, and assuming that $Z \perp\!\!\!\perp V \mid X$, the choice probability or "propensity score" is

$$P(z) = \Pr(D = 1 | Z = z) = F_V(\mu_D(z)),$$

where to simplify the notation, we keep the conditioning on X implicit. The choice equation can be written in several alternative and equivalent ways:

$$D = \mathbf{1}(\mu_D(Z) - V \geq 0) = \mathbf{1}(F_V(\mu_D(Z)) \geq F_V(V)) = \mathbf{1}(P(Z) \geq U_D)$$

[8] See, e.g., Thurstone (1927) and McFadden (1974, 1981). We do not strictly require separability, but we do require that the choice equation has one representation in separable form. See Heckman and Vytlacil (2007b).

where $U_D = F_V(V)$ so $U_D \sim \text{Uniform}[0,1]$.

We invoke the assumptions of Heckman and Vytlacil (2005, 2007b).[9] A fundamental treatment parameter introduced by Björklund and Moffitt (1987) is the marginal treatment effect (MTE). The MTE for a given value of $X = x$ is

$$MTE(x, u_D) = E(Y_1 - Y_0 \mid X = x, U_D = u_D) = E(\beta \mid X = x, U_D = u_D).$$

It is the mean effect of treatment when the observables X are fixed at a value x and the unobservable in the choice equation U_D is fixed at a value u_D. Heckman and Vytlacil (1999, 2001, 2005, 2007b) use the MTE to develop the following implication of H_0.

In the general case, the conditional expectation of Y given X and Z is

$$\begin{aligned} E(Y|X=x, Z=z) &= E(Y|X=x, P(Z)=p) \\ &= E(\alpha|X=x) + E(\beta D|X=x, P(Z)=p) \\ &= E(\alpha|X=x) + E(\beta|X=x, D=1)p \\ &= E(\alpha|X=x) + \int_0^p E(\beta|X=x, U_D=u_D)du_D, \end{aligned} \tag{2}$$

[9] Their conditions are:

(A-1) $(U_0, U_1, V) \perp\!\!\!\perp Z \mid X$. Alternatively, $(\alpha, \beta, V) \perp\!\!\!\perp Z \mid X$.

(A-2) The distribution of $\mu_D(Z)$ conditional on X is nondegenerate. Thus the distribution of $P(Z)$ is nondegenerate.

(A-3) The distribution of V is continuous (i.e., absolutely continuous with respect to Lebesgue measure). Thus $U_D = F_V(V)$ is uniform.

(A-4) $E|Y_1| < \infty$, and $E|Y_0| < \infty$, so defining $E(\beta) = \bar{\beta}, |\bar{\beta}| < \infty$.

(A-5) $1 > \Pr(D = 1 \mid X) > 0$.

Vytlacil (2002) shows that under mild regularity conditions, assumptions (A-1)-(A-5) are equivalent to the IV conditions of Imbens and Angrist (1994) used to define the local average treatment effect (LATE).

where the integrand in the final expression is the $MTE(x, u_D)$.[10] Under H_0,

$$E(\beta \mid X = x, U_D = u_D) = E(\beta \mid X = x),$$

so

$$E(Y \mid X = x, P(Z) = p) = E(\alpha \mid X = x) + E(\beta \mid X = x)p.^{11} \tag{3}$$

Thus the function $E(Y|X = x, P(Z) = p)$ is linear in p, conditional on $X = x$, which is a testable hypothesis.

A second implication of H_0 is that any standard instrument identifies $\overline{\beta} = E(\beta)$.[12] Thus under H_0 all valid instruments have the same estimand. Under conditions presented in this paper, comparing the estimates produced by different instruments tests the weaker hypothesis $H_0' : \text{Cov}(\beta, D \mid X) = 0$, which is an implication of the stronger hypothesis H_0. The analysis in this paper thus provides an alternative interpretation of standard tests of overidentification. A rejection of the null hypothesis that two instrumental variable estimands are different is not necessarily a rejection of the validity of one instrument. It could be interpreted as evidence in support of a correlated random coefficient model.

3 Tests Based on Comparing IV Estimates

Before presenting our test based on comparing the estimates from two IV estimators, we first discuss some general properties of the IV estimator in the correlated random coefficient model. We present a new representation of the sampling distribution of the IV estimator. We consider the problem of constructing the power of tests of several hypotheses using the sampling distribution of the IV estimator for the correlated random coefficient model before

[10] The first line follows from (A-1). The rest of the derivation comes from (1) and the law of iterated expectations.

[11] To see this, notice that $\beta \perp\!\!\!\perp D \mid X \iff \beta \perp\!\!\!\perp \mathbf{1}(P(Z) \geq U_D) \mid X \iff \beta \perp\!\!\!\perp U_D \mid X$.

[12] In the notation of equation (1), but dropping subscripts i, a standard instrument J has the two properties: (i) $\text{Cov}(J, D \mid X) \neq 0$ and (ii) $\text{Cov}((\alpha, \beta), J \mid X) = 0$. Note that J is shorthand for $J(Z)$. Note further that the condition $\text{Cov}(\beta, J \mid X) = 0$ only emerges as an interesting condition in a random coefficient model.

developing our test for H_0.

3.1 IV in the Correlated Random Coefficient Model

Consider an instrument $J(Z)$. Denote $J(Z)$ by J and define $\tilde{J} = J - \bar{J}$ where $\bar{J}$ is the sample mean of $J(Z)$. $E(J)$ is assumed to be finite. The IV estimator is

$$\widehat{\beta}_{IV,J} = \frac{\sum Y_i \tilde{J}_i}{\sum D_i \tilde{J}_i}.$$

Define $\text{Cov}(J, D) = \omega_J$ and let I denote the sample size. Under a weak law of large numbers, $\frac{1}{I}\sum D_i \tilde{J}_i \xrightarrow{p} \omega_J$ and $\bar{J} \xrightarrow{p} E(J)$. As shown in Heckman and Vytlacil (2005, 2007b), under the conditions (A-1)–(A-5) stated in Section 2,

$$\widehat{\beta}_{IV,J} \xrightarrow{p} \beta_{IV,J} = \int_0^1 E(\beta | U_D = u_D) h_J(u_D) du_D \tag{4}$$

where

$$h_J(u_D) = \frac{E[(J - E(J)) \mid P(Z) \geq u_D] \Pr(P(Z) \geq u_D)}{\omega_J}, \tag{5}$$

and we keep the conditioning on X implicit. Heckman and Vytlacil (2005) show that $\int_0^1 h_J(t)dt = 1$. Thus we can write

$$\beta_{IV,J} = \overline{\beta} + \int_0^1 E(U_\beta \mid U_D = u_D) h_J(u_D) du_D. \tag{6}$$

For later use we break out the component of $\beta_{IV,J}$ that depends on the instrument J:

$$\int_0^1 E(U_\beta \mid U_D = u_D) h_J(u_D) du_D = \Upsilon_J,$$

so $\beta_{IV,J} = \bar{\beta} + \Upsilon_J$. By definition, conditional on X, $\bar{\beta}$ does not depend on J.

Under independent sampling,

$$\sqrt{I}\left(\widehat{\beta}_{IV,J}-\beta_{IV,J}\right)\xrightarrow{d}N(0,\Omega_J)$$

where

$$\Omega_J=E\left[\alpha^2\right]\frac{\mathrm{Var}(J)}{\omega_J^2}+\int_0^1\left[2E\left(\alpha\beta\mid U_D=u_D\right)+E\left(\beta^2\mid U_D=u_D\right)\right]h_{\Omega_J}(u_D)du_D \tag{7}$$
$$-\left(\int_0^1 E(\beta\mid U_D=u_D)h_J(u_D)du_D\right)^2$$

and

$$h_{\Omega_J}(u_D)=\frac{1}{\omega_J^2}\int_{-\infty}^{\infty}(j-E(J))^2\int_{u_D}^1 f_{P,J}(P(z),j)dP(z)dj \tag{8}$$
$$=\frac{E[(J-E(J))^2\mid P(Z)\geq u_D]\Pr(P(Z)\geq u_D)}{\omega_J^2}.$$[13]

The weight $h_{\Omega_J}(u_D)$ does not necessarily integrate to 1:

$$\int_0^1 h_{\Omega_J}(t)dt=\frac{\mathrm{Cov}(\tilde{J}^2,D)}{[\mathrm{Cov}(\tilde{J},D)]^2}.$$

Appendix A presents the full derivation. The weight $h_{\Omega_j}(u_D)$ plays a role in determining the variance of the IV estimator that is analogous to the role of $h_J(u_D)$ in generating the probability limit of the IV estimator. $2E[\alpha\beta\mid U_D=u_D]+E[\beta^2\mid U_D=u_D]$ plays a role in generating the variance of the IV estimator analogous to the role of the MTE in generating the probability limit of the IV estimator. We use this representation to facilitate comparison of the power of the tests under alternative data generating processes and to consider the problem of the optimal choice of instruments.

[13] $f_{P,J}(P(z),j)$ is the density of $P(Z)$ and $J(Z)$ evaluated at $P(Z)=P(z)$ and $J(Z)=j$.

These formulae hold for general functions $J(\cdot)$ of instruments Z that satisfy assumptions (A-1)-(A-5) given in Section 2. For example, suppose that $J(Z)$ has discrete support on points $j_1, \ldots, j_K$ with corresponding values of the propensity score $p_1, \ldots, p_L$ with L possibly not equal to K. Let $p_0 = 0$. In this case, for $u_D \in [p_l, p_{l+1}]$ both h_J and h_{Ω_J} are constant so we can write

$$\Omega_J = E\left[\alpha^2\right] \frac{\mathrm{Var}(J)}{\omega_J^2} + \sum_{l=0}^{L-1} \lambda_{\Omega_l} \int_{p_l}^{p_{l+1}} \left[2E\left(\alpha\beta \mid U_D = u_D\right) + E\left(\beta^2 \mid U_D = u_D\right)\right] \frac{1}{p_{l+1} - p_l} du_D$$
$$- \left(\sum_{l=1}^{L-1} \lambda_l \int_{p_l}^{p_{l+1}} E(\beta \mid U_D = u_D) \frac{1}{p_{l+1} - p_l} du_D \right)^2.$$

The weights λ_{Ω_l} and λ_l are defined in the following way. Let j_i be the ith smallest value in the support of $J(Z)$, then

$$\lambda_{\Omega_l} = \frac{\sum_{i=1}^{K} [j_i - E(J)]^2 \sum_{t>l}^{L} f_{P,J}(p_t, j_i)}{\mathrm{Cov}(\tilde{J}(Z), D)^2} (p_{l+1} - p_l)$$
$$\lambda_l = \frac{\sum_{i=1}^{K} [j_i - E(J)] \sum_{t>l}^{L} f_{P,J}(p_t, j_i)}{\mathrm{Cov}(\tilde{J}(Z), D)} (p_{l+1} - p_l).$$

The special case of a binary instrument $J(Z)$ has two points of support, j_1 and j_2, corresponding to the points p_1 and p_2 in the propensity score distribution. Let $\Pr(J(Z) = j_1) = \Pr(P(Z) = p_1) = q$ and $\Pr(J(Z) = j_2) = \Pr(P(Z) = p_2) = 1 - q$. The λ_l are $\lambda_1 = 1$ and $\lambda_l = 0$, $l > 1$.[14] The weights for the variance simplify to

$$\lambda_{\Omega_0} = \frac{[j_1 - E(J)]^2 q + [j_2 - E(J)]^2 (1-q)}{\mathrm{Cov}(\tilde{J}(Z), D)^2} (p_1) \text{ and } \lambda_{\Omega_1} = \frac{[j_2 - E(J)]^2 (1-q)}{\mathrm{Cov}(\tilde{J}(Z), D)^2} (p_2 - p_1),$$

[14]
$$\lambda_1 = \frac{[j_2 - E(J)](1-q)}{\mathrm{Cov}(\tilde{J}(Z), D)} (p_2 - p_1) = \frac{(j_2 - j_1)(p_2 - p_1) q (1-q)}{\mathrm{Cov}(\tilde{J}(Z), D)} = \frac{\mathrm{Cov}(\tilde{J}(Z), P(Z))}{\mathrm{Cov}(\tilde{J}(Z), D)} = 1.$$

and

$$\lambda_{\Omega_0} + \lambda_{\Omega_1} = \frac{(j_1 - E(J))^2 qp_1 + (j_2 - E(J))^2 (1-q)p_2}{\text{Cov}(\tilde{J}, D)^2} = \frac{\text{Cov}(\tilde{J}^2, D)}{\text{Cov}(\tilde{J}, D)^2}.$$

Formula (4) extends the representation of IV as weighted averages of slopes of the underlying function, due to Yitzhaki (1989). It allows the instrument $J(Z)$ be different from the propensity score $P(Z)$ or a monotonic function of it. It reveals that, in general, different instruments identify different parameters. Thus, in general, $\beta_{IV,J} \neq \beta_{IV,J'}$ if J and J' apply different weights (5) to a common MTE.

As noted by Heckman and Vytlacil (2005, 2007b), while the weight in (5) integrates to 1, it is not necessarily non-negative for all values of u_D so the interpretation of the weighted average produced by IV is obscure. Even though the MTE is positive everywhere, the IV estimate may be negative.[15]

Some applied economists report tests based on IV sampling distributions as if they are testing the null hypothesis that $\bar{\beta} = 0$. Under H_0, i.e., the absence of a correlated random coefficient model, the sampling distribution of the standard IV estimator, $\hat{\beta}_{IV,J}$, can be used to consistently test the null hypothesis that $\bar{\beta} = 0$. However, when H_0 is false, a test of $\bar{\beta} = 0$ based on the sampling distribution of the IV estimator is, in general, inconsistent and biased because by (6), IV does not, in general, converge to $\bar{\beta}$.

Consider the following example based on the normal generalized Roy Model.

$$\begin{pmatrix} U_1 \\ U_0 \\ V \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{1V} \\ \sigma_{10} & \sigma_0^2 & \sigma_{0V} \\ \sigma_{1V} & \sigma_{0V} & \sigma_V^2 \end{pmatrix} \right), \tag{9}$$

and assume $X = 1$. Recalling that $u_D = F_V(v)$, when V is a normal random variable, the

[15] See the examples in Heckman, Urzua, and Vytlacil (2006).

marginal treatment effect is

$$MTE(U_D = u_D) = \bar{\beta} + \left(\frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \right) \Phi^{-1}(u_D) \tag{10}$$

where $\Phi^{-1}(\cdot)$ is the inverse of a standard normal CDF (hence $\Phi^{-1}(u_D) = v$). Alternatively, in terms of v,

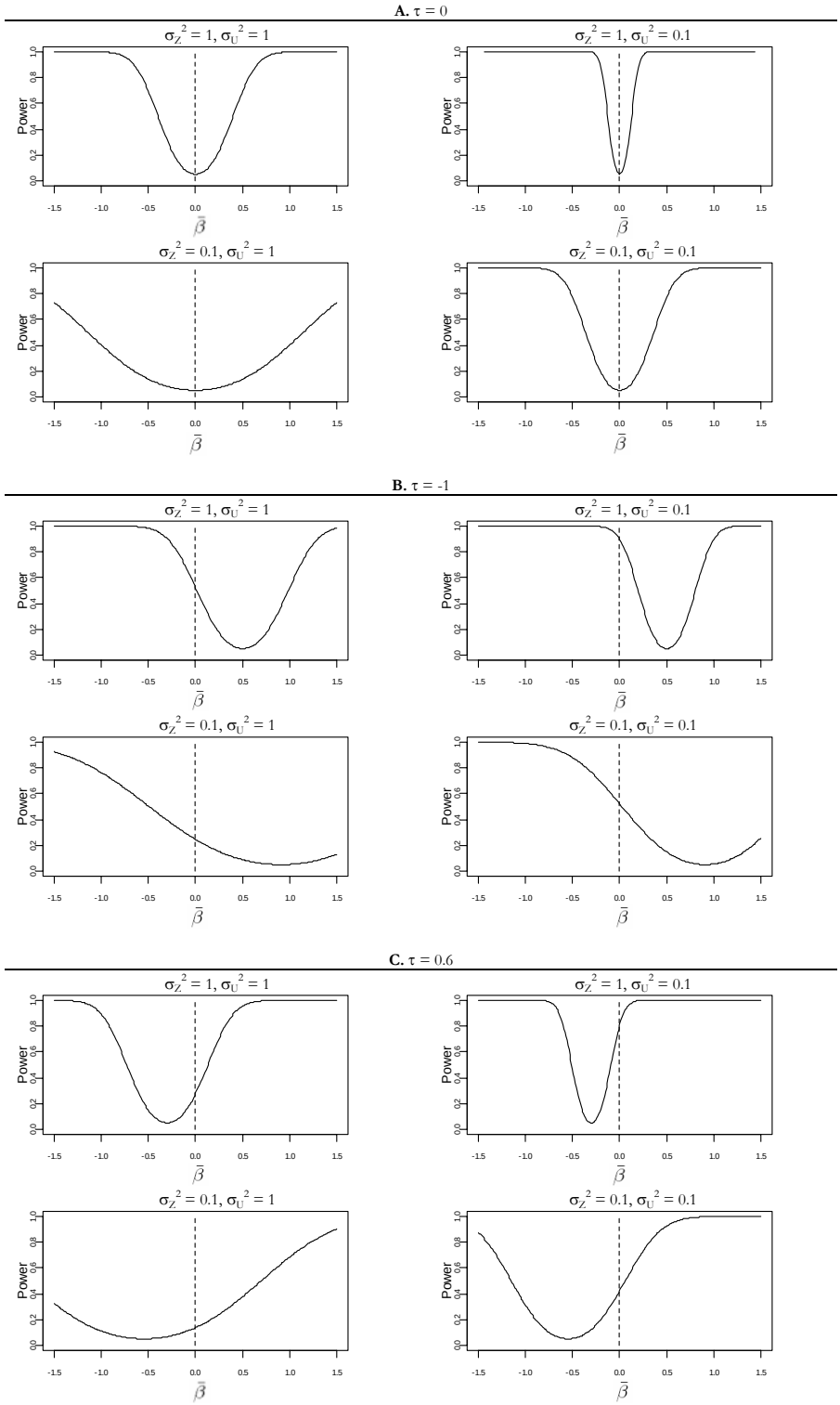
$$MTE(V = v) = \bar{\beta} + \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} v.$$

Let $\tau = \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V}$. A value of $\tau \neq 0$ produces a correlated random coefficient model. For such values $\operatorname{plim} \hat{\beta}_{IV,J} \neq \bar{\beta}$. The choice equation is assumed to be $D = \mathbf{1}(Z \geq V)$ where both Z (a single instrument) and V are normally distributed and $Z \perp\!\!\!\perp V$. Additionally, assume that $\sigma_1^2 = \sigma_0^2 = \sigma_U^2$, $\sigma_{10} = 0.5 \times \sigma_1 \times \sigma_0$ and $\sigma_V^2 = 1$.

Figure 1 plots the power of a Wald test of the hypothesis that $\bar{\beta} = 0$ based on $\hat{\beta}_{IV,J}$. We compute the power function for different values of $\bar{\beta}$. Recall from (6) that this is the component of $\beta_{IV,J}$ that does not depend on J. In Panel A, $\hat{\beta}_{IV,J}$ is a consistent estimator for $\bar{\beta}$. In the other two panels it is not. Thus in the top panel of the figures, when $\tau = 0$, and hence H_0 is true, the test of the hypothesis $\overline{\beta} = 0$ is unbiased and consistent and the size of the test is controlled.[16] As expected, smaller values of σ_U^2 produce higher power, and larger values of σ_Z^2 produce higher power. The bottom two panels plot the power of the test that $\overline{\beta} = 0$ when $\tau = -1$ and $\tau = 0.6$, respectively. In these two latter cases, $\operatorname{plim} \widehat{\beta}_{IV,J} = \beta_{IV,J} \neq \bar{\beta}$. Hence the tests are biased and inconsistent. The power and size of the test for the existence of an "effect" (i.e., whether $\overline{\beta} = 0$) can be badly distorted. Thus even if $\overline{\beta} = 0$, an "effect" can be detected, and if $\overline{\beta} \neq 0$, no "effect" can be detected.

[16] Although Figure 1 shows the power function only for one sample size, the consistency of the test is readily verified.

Figure 1: Power function for a Wald test of $\bar{\beta} = 0$ based on the sampling distribution of $\hat{\beta}_{IV,J}$.



Note: Each plot shows the power for a hypothetical sample size of 500. The size of the test is 0.05. The model is the normal generalized Roy model with the unobservables jointly normal with variance σ_U^2 and correlation 0.5. The choice equation is $D = \mathbf{1}(Z \geq V)$ where $V \sim N(0,1)$ and $Z \sim N(1, \sigma_Z^2)$. The power functions plot the power of the Wald test of $\beta_{IV,J} = 0$ for alternative values of $\bar{\beta}$. The vertical dashed lines denote the null hypothesis $\bar{\beta} = 0$. Each panel fixes $\tau = \mathrm{Cov}(\beta, V)/\mathrm{Var}(V)$ at a different level. When $\tau = 0$, $\mathrm{plim}\, \hat{\beta}_{IV,J} = \bar{\beta}_0$, which in these figures is zero, and hence the test is consistent. For all nonzero values of τ, the test is inconsistent.

3.2 Testing Hypotheses About Instrument-Dependent Parameters

More recently, many applied economists, following Imbens and Angrist (1994), interpret IV as a weighted average of "LATEs," or in our framework, a weighted average of MTEs, as in equation (3). It is understood that $\hat{\beta}_{IV,J}$ is not, in general, consistent for the true $\bar{\beta}$. Within this framework, economists often report tests of the hypothesis that $\beta_{IV,J} = 0$.

To calculate the power of such tests, we consider alternative values of $\beta_{IV,J} (= \bar{\beta} + \Upsilon_J$ from equation (6)) obtained by varying $\bar{\beta}$ holding Υ_J fixed. Notice that unlike the analysis in the preceding section, in this section we are not testing the hypothesis that $\bar{\beta} = 0$. Instead we are testing the hypothesis that $\beta_{IV,J} = 0$ (or some other specified value). We vary $\bar{\beta}$ to calculate the power of the test for alternative values of $\beta_{IV,J}$. This is a sensible way to proceed because $\bar{\beta}$ is instrument invariant. Investigating the power of the test in this fashion allows us to construct power functions for instrument-invariant alternatives.

Figure 2 plots the power function for the Wald test of the hypothesis $\beta_{IV,J} = 0$ as a function of $\beta_{IV,J}$ holding Υ_J fixed at -0.5. Consequently, the $\bar{\beta}$ compatible with the null hypothesis, $\bar{\beta}_0$, is 0.5. For the model of unobservables used in the previous subsection, keeping Υ_J fixed entails, among other things, holding $\tau = \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V}$ fixed along with the weighting function $h_J(u_D)$. For a given τ and a fixed IV weighting function $h_J(u_D)$, we vary the parameters of covariance matrix (9). These parameters affect the sampling distribution of $\hat{\beta}_{IV,J}$ and hence the power of the test.

Neither the IV estimand nor the variance of the IV estimator depends on σ_{10}. Therefore, the power of the test of the null hypothesis $\beta_{IV,J} = 0$ does not depend on σ_{10}. The only remaining parameters that can be changed without changing Υ_J are $\sigma_0^2, \sigma_1^2, \sigma_{1V}$ and σ_{0V}. To keep τ fixed, we can only vary σ_{1V} and σ_{0V} subject to a constraint that $\sigma_{1V} - \sigma_{0V}$ is constant.[17] For $\sigma_V = 1$, the four A panels of Figure 2 show the power of the test for different values of $\bar{\beta}$ when we vary σ_{1V} and σ_{0V} such that $\sigma_{1V} - \sigma_{0V} = -1$. The power of the test is

[17] Variations in σ_V^2 affect the denominator of the weights.

highest when σ_{1V} and σ_{0V} are both close to 0 (ie. straddling 0), and lowest when both are far from zero (either positive or negative). The panels in B vary $\beta_{IV,J}$ by varying $\bar{\beta}$ holding Υ_J fixed and hold fixed all of the elements of (9) except for σ_1^2, while the panels in C vary $\bar{\beta}$ hold fixed all of the parameters of (9) except σ_0^2. As expected, power decreases as both variances increase, in general at different rates.

There are other ways to calculate the power of the test that $\beta_{IV,J} = 0$ for alternative values that are obtained by varying $\bar{\beta}$ keeping Υ_J fixed. If the choice equation is

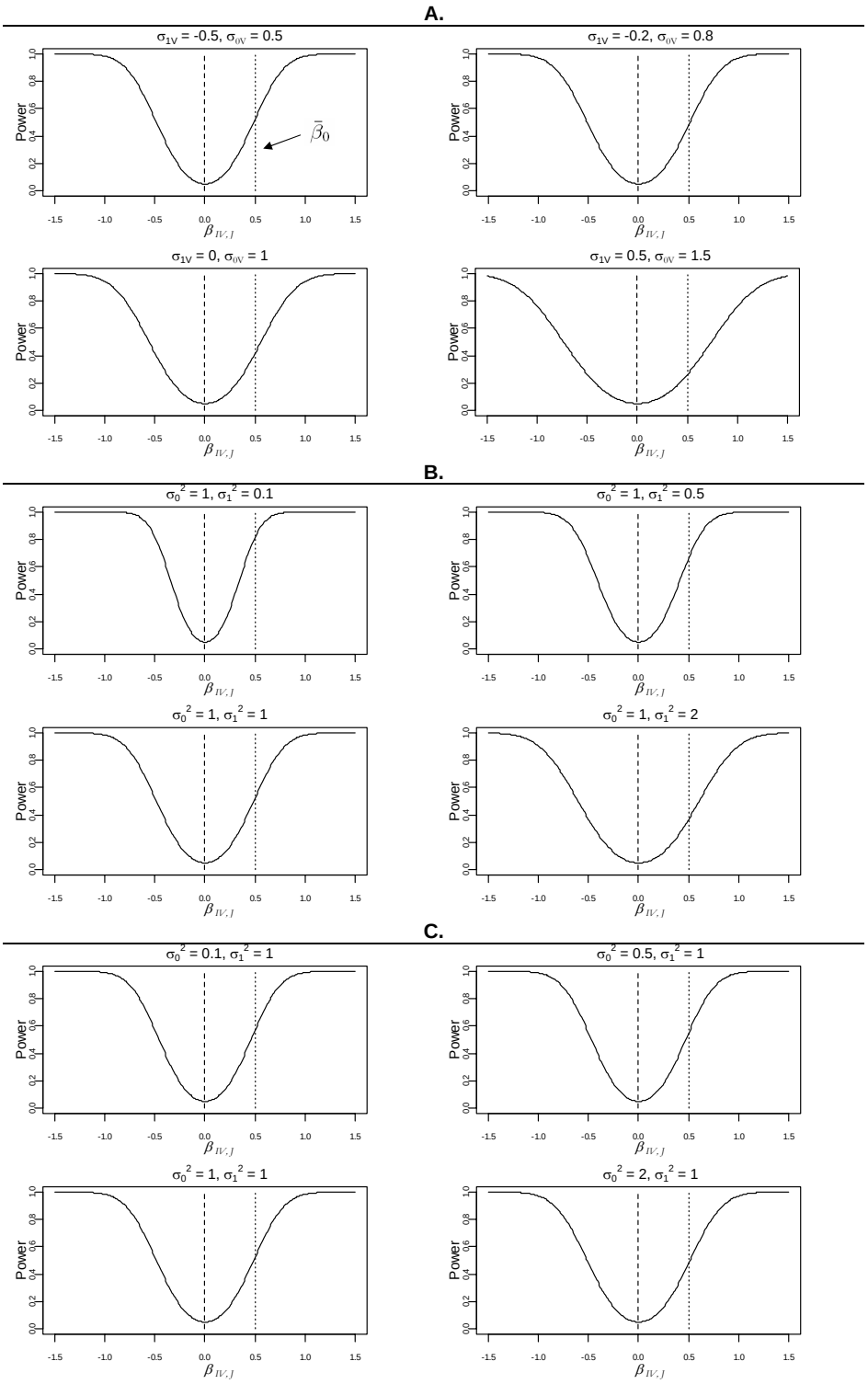
$$D = 1(Z\gamma \geq V)$$

and $Z \sim N(\overline{Z}, \Sigma_Z)$ and $V \sim N(0, \sigma_V^2)$, all instruments constructed from linear or affine transformations of Z have the same weight function (5) and hence have the same instrument-dependent value, $\beta_{IV,J}$. For proof of this claim, see Appendix B.[18]

This result implies that one can construct power functions for the hypothesis $\beta_{IV,J} = 0$ for different values of $\beta_{IV,J} = \bar{\beta} + \Upsilon_J$ for alternative choices of Σ_Z, holding $\gamma'\Sigma_Z\gamma$, the variance of the choice index, constant. The derivation in Appendix B shows that the IV estimand depends only on the distribution of the index $Z\gamma - V$. From assumption (A-1), $Z\gamma$ and V are statistically independent. σ_V^2 has to be held constant to keep Υ_J fixed. We keep this term fixed by varying components of Z while keeping $\gamma' Z\gamma$ fixed. An instrument with greater variance that obeys this constraint will produce greater power. Figure 3 plots power functions of the test of the hypothesis that $\beta_{IV,J} = 0$ using each component of a two-dimensional instrument $Z = (Z_1, Z_2)$. These plots show that for a given IV estimand $\beta_{IV,J}$, the power of the test is higher when using the instrument that accounts for more of the variance of the index $Z\gamma$. Going from top to bottom, the variance of Z_1 is increasing while the variance of Z_2 is decreasing. Accordingly, from top to bottom the power of the test $\beta_{IV,J} = 0$ using Z_1 as an instrument is increasing while the power of the test using Z_2

[18]This result is special to the case of $J(Z)$ linear or affine in Z with Z normally distributed, so $J(Z)$ is normally distributed and the further assumption (A-1) that $Z \perp\!\!\!\perp V$, where V is normally distributed. We have not analyzed more general conditions on Z and V under which the invariance holds.

Figure 2: Power function for the test of the hypothesis that $\operatorname{plim} \hat{\beta}_{IV,J} = 0$ when $\bar{\beta}_0 = 0.5$. Alternatives are different values of $\beta_{IV,J}$ obtained by fixing Υ_J and varying $\bar{\beta}$.



Note: Each plot shows the power for a hypothetical sample size of 500. The size of the test is 0.05. The instrument is normally distributed, $Z \sim N(1,1)$; $D = \mathbf{1}(Z \geq V)$. In panel A, the unobservables are generated with covariances given in the figure and $\sigma_V^2 = 1, \sigma_{10} = 0, \sigma_1^2 = 1, \sigma_0^2 = 1$. In panels B and C the unobservables are generated with variances given in the figure and $\sigma_V^2 = 1, \sigma_{10} = 0, \sigma_{1V} = -0.5, \sigma_{0V} = 0.5$. In all panels, under the null hypothesis $\bar{\beta}_0 = 0.5$, and alternative hypotheses are generated by changing $\bar{\beta}$. The vertical dashed line shows the value of $\operatorname{plim} \hat{\beta}_{IV,J} = 0$ under the null hypothesis and the vertical dotted line shows the value of $\bar{\beta}$ under the null hypothesis.

as an instrument is decreasing. Each panel shows the fraction of $\gamma' Z \gamma$ accounted for by the variance of the instrument used to construct the power function (either Z_1 or Z_2).[19] We now use the tools developed for IV in a correlated random coefficient model to test H_0.

3.3 Testing H_0 Using Instrumental Variables

Armed with the preceding results, we now return to the main theme of this paper and study how to use different IVs to test H_0. Under H_0, the probability limits of any two IV estimators are identical, because for any choice of J,

$$\text{plim}\, \widehat{\beta}_{IV,J} = \beta_{IV,J} = \int_0^1 E(\beta | U_D = u_D) h_J(u_D) du_D = \bar{\beta} \int_0^1 h_J(u_D) du_D = \bar{\beta}.$$

If H_0 is false, in general any two IV estimators will differ. Excluding the case of equal IV weights for the two instruments, our IV test forms two estimators $\widehat{\beta}_{IV,1}$ and $\widehat{\beta}_{IV,2}$, based on $J_1(Z)$ and $J_2(Z)$ respectively, and tests the null hypothesis

$$H_0^{IV} : \beta_{IV,1} - \beta_{IV,2} = 0$$

against the alternative hypothesis

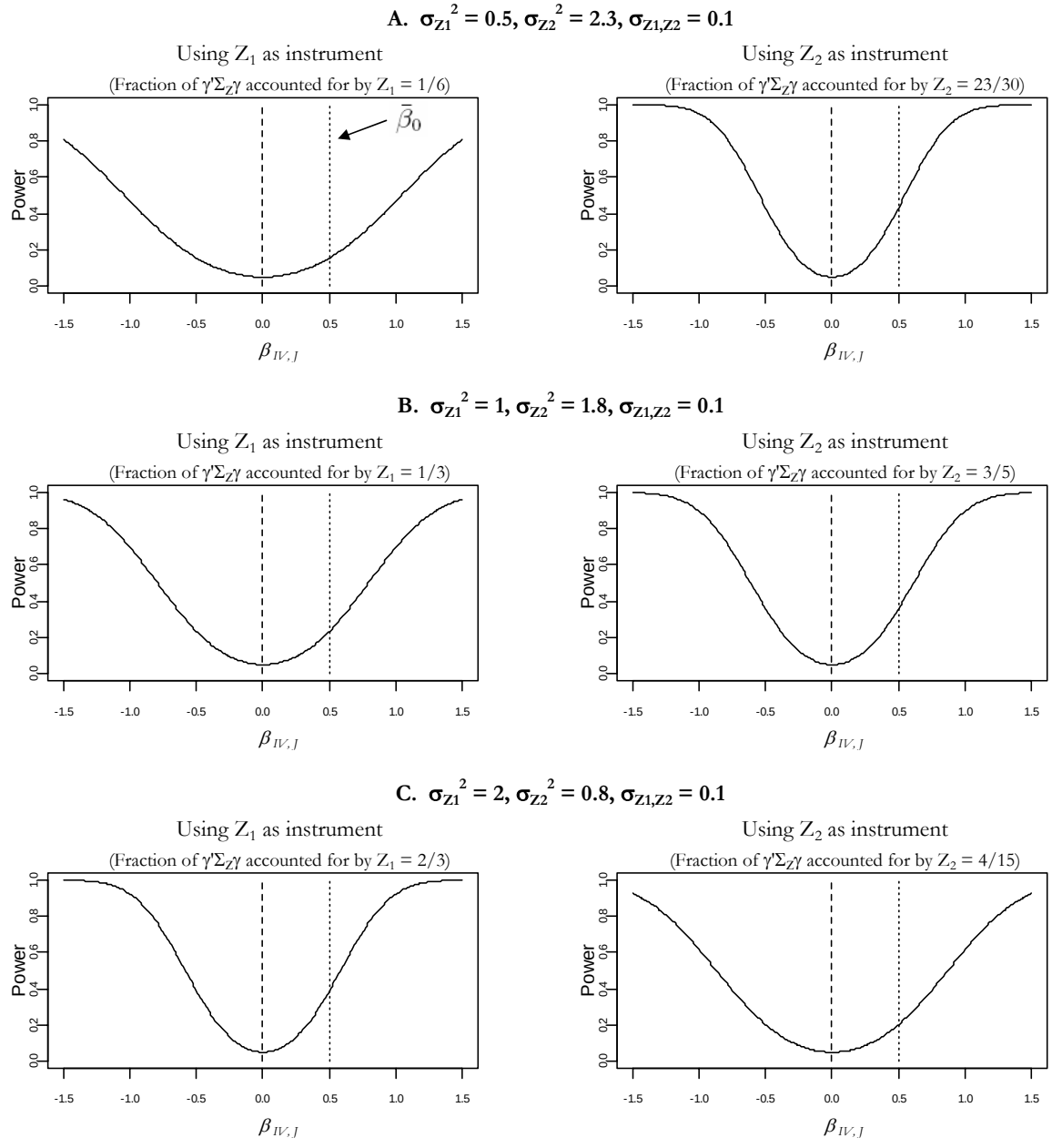
$$H_A^{IV} : \beta_{IV,1} - \beta_{IV,2} \neq 0.$$

This test is identical to a standard test for overidentification. However, within the context of a correlated random coefficient model, we do not interpret rejections of the null hypothesis as evidence of the violation of the assumptions required for the validity of an instrument. Rather, it is interpreted as evidence of selection on heterogeneous gains to treatment.

Under the null hypothesis, the Wald test statistic is asymptotically distributed as a

[19] Note that in a given row, the fractions do not sum to 1 because there is a covariance (of 0.1) between Z_1 and Z_2.

Figure 3: Power functions for the test of the hypothesis that $\beta_{IV,J} = \text{plim}\,\hat{\beta}_{IV,J} = 0$ for $\bar{\beta}_0 = 0.5$. Alternatives are different values of $\beta_{IV,J}$ obtained by fixing Υ_J and varying $\bar{\beta}$.



Note: Each plot shows the power for a hypothetical sample size of 500 varying $\bar{\beta}$ keeping Υ_J fixed. The size of the test is 0.05. The instruments are distributed normally, $Z_1 \sim N(1, \sigma^2_{Z1})$, $Z_2 \sim N(1, \sigma^2_{Z2})$ and $\text{Cov}(Z_1, Z_2) = \sigma_{Z1,Z2} = 0.1$; $D = \mathbf{1}(Z_1 + Z_2 \geq V)$ so $\gamma = (1, 1)$. The distribution of the index is held fixed and is distributed $N(2, 3)$. The unobservables are jointly normally distributed with and $\sigma^2_V = 1, \sigma_{10} = 0.5, \sigma^2_1 = 1, \sigma^1_0 = 1, \sigma_{1V} = -0.5, \sigma_{0V} = 0.5$. In all panels, under the null hypothesis $\bar{\beta} = 0.5$, and alternative hypotheses are generated by changing $\bar{\beta}$. The vertical dashed line shows the value of $\beta_{IV,J}$ under the null hypothesis and the vertical dotted line shows the value of $\bar{\beta} = \bar{\beta}_0$ under the null hypothesis being considered, ie. that $\beta_{IV,J} = \bar{\beta} + \Upsilon_J$.

χ^2_1. Under the alternative, in the general case, the Wald statistic converges to a noncentral chi-square distribution. Let $h_1(\cdot)$ and $h_2(\cdot)$ denote the weights (akin to $h_J(\cdot)$ above) corresponding to $J_1(Z)$ and $J_2(Z)$, respectively. To simplify the notation, we suppress the Z argument. Define $\tilde{J}_1 = J_1 - \overline{J}_1$ and $\tilde{J}_2 = J_2 - \overline{J}_2$ as the demeaned values of the instruments. Let $\tilde{\mathbf{J}}_\mathbf{1} = (\tilde{J}_{11}, \ldots, \tilde{J}_{1I})'$ and $\tilde{\mathbf{J}}_\mathbf{2} = (\tilde{J}_{21}, \ldots, \tilde{J}_{2I})'$ be the matrices of demeaned instruments stacked across individuals. Let $\mathbf{D} = (D_1, \ldots, D_I)'$ be the stacked values of the choice variable D_i. Under random sampling, and the assumptions of Section 2, $\frac{\tilde{\mathbf{J}}_\mathbf{1}'\mathbf{D}}{I} \xrightarrow{p} \omega_1$ and $\frac{\tilde{\mathbf{J}}_\mathbf{2}'\mathbf{D}}{I} \xrightarrow{p} \omega_2$ for some finite constants ω_1 and ω_2. Under $H_A^{IV} : \beta_{IV,1} - \beta_{IV,2} = \left[\int_0^1 MTE(u_D)(h_1(u_D) - h_2(u_D))du_D\right]/\sqrt{I}$, the noncentrality parameter of the chi-square distribution of the test statistic is

$$\lambda_{\text{IV},1,2} = \frac{1}{2}\left(\int_0^1 MTE(u_D)(h_1(u_D) - h_2(u_D))du_D\right)^2 \Psi_{1,2}^{-1} \tag{11}$$

where

$$\begin{aligned}\Psi_{1,2} = {} & E\left(\alpha^2\right)\left[\frac{\text{Var}\left(J_1\right)}{\omega_1^2} - \frac{2\,\text{Cov}(J_1, J_2)}{\omega_1\omega_2} + \frac{\text{Var}\left(J_2\right)}{\omega_2^2}\right] \\ & + \int_0^1 \left[2E(\alpha\beta \mid U_D = u_D) + E(\beta^2 \mid U_D = u_D)\right] h_{\Omega,J_1,J_2}(u_D)du_D \\ & - \left[\int_0^1 MTE(u_D)(h_1(u_D) - h_2(u_D))du_D\right]^2.\end{aligned} \tag{12}$$

Defining $J_1^* = J_1 - E(J_1)$ and $J_2^* = J_2 - E(J_2)$, the weight $h_{\Omega,J_1,J_2}(\cdot)$ is given by

$$\begin{aligned}h_{\Omega,J_1,J_2}(u_D) &= \int_{u_D}^1 \int_{-\infty}^{\infty} \left(\frac{J_1^*}{\omega_1} - \frac{J_2^*}{\omega_2}\right)^2 f_{(J_1-J_2),P}(j_1 - j_2, P(z))\, d(j_1 - j_2)\, dP(z) \\ &= E\left[\left(\frac{J_1^*}{\omega_1} - \frac{J_2^*}{\omega_2}\right)^2 \mid P(Z) \geq u_D\right] \Pr(P(Z) \geq u_D).^{20}\end{aligned}$$

The derivation follows a logic similar to that used to derive (7).[21] Notice that not only will the difference in the IV estimands depend on the alternative under consideration, but the variance of the difference between the IV estimators will also depend on the alternative under consideration.

We present this characterization of the variance in order to understand the properties of tests of H_0 based on IV estimators. This expression for the variance is not meant as a guide for how to implement such tests. In practice the analyst would form the test statistic using a standard estimator of the variance of the vector of IV estimates.

In general, the weights presented above do not have simple analytical expressions. They do in the case of a model with normal error terms with normally distributed instruments and a linear index structure for the choice equation. However, for this case, the proposed IV test has no power, because, and as previously discussed and as established in Appendix B, in this case $\beta_{\text{IV},J_1} \equiv \beta_{\text{IV},J_2}$ irrespective of the truth or falsity of H_0. For this case, the noncentrality parameter of the asymptotic chi-square distribution of the test statistic will be zero so the power of the test equals its size. To have a test with any power, we have to rule out instruments with equal weights. Since the weights can be constructed from the data on Z, it is possible to check this condition in any sample.[22]

We do not formally analyze conditions that guarantee that the two instruments J_1 and J_2, constructed from Z, optimize the power function of the test. From the expression for the noncentrality parameter, one can see the ingredients required to construct an asymptotically most powerful test. Let $Z \in \mathbb{R}^k$ be the vector of available instruments and let $\mathcal{J} = \left\{ J \mid J : \mathbb{R}^k \to \mathbb{R} \right\}$ be the space of functions which map the vector of instruments to the

[20] $f_{(J_1-J_2),P}(j_1 - j_2, P(z))$ is the joint density of $J_1 - J_2$, and $P(Z)$ evaluated at $J_1 - J_2 = j_1 - j_2$ and $P(Z) = P(z)$.

[21] The logic is not, however, identical. Using $(J_1 - J_2)$ as an instrument and testing if $\beta_{\text{IV},J_1-J_2} = 0$ is *not* equivalent to the test presented in the main text of the paper. The denominators of the IVs differ in the two approaches.

[22] It would be desirable to develop a formal test for equality of the two IV weights. The required ingredients are in the literature. We leave the formal derivation for another occasion.

real line. Then for a given MTE, the optimal choice of J_1 and J_2 solves the problem

$$\max_{J_1 \in \mathcal{J}, J_2 \in \mathcal{J}} \frac{1}{2} \left(\int_0^1 MTE(u_D)(h_1(u_D) - h_2(u_D)) du_D \right)^2 \Psi_{1,2}^{-1}.$$

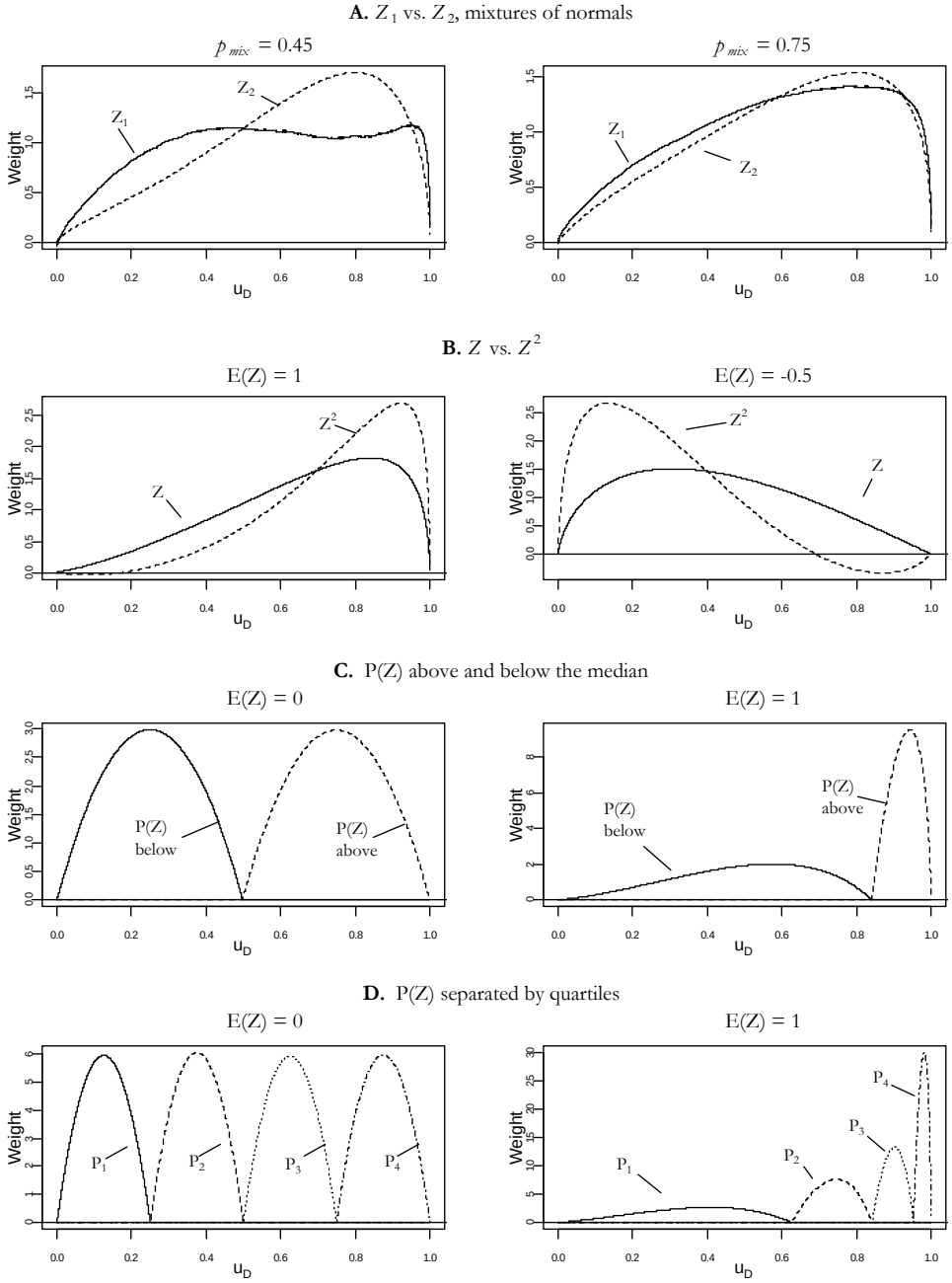
The optimal choice of instruments will generally depend on the shape of the MTE(u_D).[23]

We present an example with two nonnormal instruments in Figure 4. Specifically, let $D = \mathbf{1}(\gamma_1 Z_1 + \gamma_2 Z_2 \geq V)$ where the vector $Z = (Z_1, Z_2)$ is distributed as a multivariate *mixture* of normals with the distribution given at the base of the figure. The unobservables are assumed to be generated by a normal generalized Roy model. The test of equality of the IV estimators constructed using these two instruments has power to detect deviations from H_0. Figure 4A plots the weights $h_1(\cdot)$ and $h_2(\cdot)$ which the IV estimator places on the MTE, using Z_1 or Z_2 respectively. The weights must differ for the test based on the difference in IV estimators to have power to detect deviations from H_0. When the mixing proportion in the mixture of normals is 0.45, the instruments are highly nonnormal and the IV weights differ substantially. However, when the mixing proportion is 0.75, the instruments become closer to normal, the weights become very similar, and the test of H_0 loses power, as we demonstrate more systematically in Section 5 below.

Another example of a test that has power to detect deviations from H_0, even with normal instruments, constructs IV estimators using nonlinear functions of Z. We consider a normal generalized Roy model where there is one Z variable in the choice equation that is normally distributed, $D = \mathbf{1}(Z \geq V)$. We plot the weights of the IV estimators based on Z and Z^2. Figure 4B plots the weights for these two choices of instruments. The weights differ, and in addition the amount by which they differ generally depends on the distribution of Z. We plot the weights for two choices of the mean of Z presented in the figure. These choices clearly affect the weights and hence will generally affect the power of a test of H_0 based on these IV estimators.

[23]More generally, one could use multiple instruments and base a test on multiple contrasts of the set of instruments. We do not develop this test in this paper.

Figure 4: IV weights for alternative choices of the instrument.



Note: Panel A plots the weights of IV estimates constructed using either Z_1 or Z_2 as an instrument where (Z_1, Z_2) is distributed as a multivariate mixture of normals, with $D = \mathbf{1}(\gamma_1 Z_1 + \gamma_2 Z_2 \geq V)$. To construct these results, we assume

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim p_{mix} \times N\left(\begin{pmatrix} -0.8 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{pmatrix}\right) + (1 - p_{mix}) \times N\left(\begin{pmatrix} -0.8 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{pmatrix}\right)$$

and the coefficients in the choice equation are γ_1=0.2, γ_2=1. In the left plot of Panel A we let p_{mix} = 0.45 and in the right plot p_{mix} = 0.75. Panel B plots the weights of IV estimates constructed using either Z or Z^2 as an instrument where $Z \sim N(\mu_Z, 1)$, μ_Z = 1 or μ_Z = -0.5, and $D = \mathbf{1}(Z \geq V)$. Panel C plots the weights of IV estimates constructed using either P(Z) below the median or P(Z) above the median as instruments. Panel D plots the weights of IV estimates constructed using P(Z) in different quartiles of its distribution as instruments. In Panels C and D, $Z \sim N(\mu_Z, 1)$, μ_Z = 0 or μ_Z = 1, and $D = \mathbf{1}(Z > V)$. In all of the plots, we set $\sigma_V^2 = 1$.

Another choice of instruments uses $P(Z)$ on disjoint intervals of the support of $P(Z)$ as two instruments. Form two disjoint intervals $[\underline{p}_1, \overline{p}_1]$ and $[\underline{p}_2, \overline{p}_2]$, and construct IV estimators over these intervals as sample analogs to

$$\beta_{IV,(\underline{p}_1,\overline{p}_1)} = \frac{\text{Cov}\left(Y, P(Z) \mid P(Z) \in [\underline{p}_1, \overline{p}_1]\right)}{\text{Var}\left(P(Z) \mid P(Z) \in [\underline{p}_1, \overline{p}_1]\right)}$$

and

$$\beta_{IV,(\underline{p}_2,\overline{p}_2)} = \frac{\text{Cov}\left(Y, P(Z) \mid P(Z) \in [\underline{p}_2, \overline{p}_2]\right)}{\text{Var}\left(P(Z) \mid P(Z) \in [\underline{p}_2, \overline{p}_2]\right)}$$

and test

$$\begin{aligned} H_0^{IV} &: \beta_{IV,(\underline{p}_1,\overline{p}_1)} = \beta_{IV,(\underline{p}_2,\overline{p}_2)} \\ H_A^{IV} &: \beta_{IV,(\underline{p}_1,\overline{p}_1)} \neq \beta_{IV,(\underline{p}_2,\overline{p}_2)}. \end{aligned}$$

There is no *a priori* guidance on which intervals to use so we consider two ways to construct intervals over which to form IV estimates: (1) use the intervals $[0, p_{med}]$ and $[p_{med,}1]$ where p_{med} is the sample median of $P(Z)$, and (2) use the intervals $[0, p_{q1}]$, $[p_{q1}, p_{q2}]$, $[p_{q2}, p_{q3}]$ and $[p_{q3}, 1]$, where p_{qj} is the jth sample quartile of the distribution of $P(Z)$ and form all pairwise contrasts between these estimates. Note that even though we split the propensity score into four intervals, we are still conducting pairwise tests. However, because there is a multiplicity of pairwise tests, we must control the size of the test. We do this by using the stepdown procedure of Romano and Wolf (2005). Figures 4C and 4D plot the weights for the instruments constructed in this manner. These weights are nonoverlapping by construction and will also depend on the distribution of the instrument Z.

The power of the test of H_0 based on IV estimators also depends on the variance (12), which determines the denominator of the noncentrality parameter. The important terms

which are affected by the choice of instruments are the variance of the difference in the instruments $\left[\frac{\text{Var}(J_1)}{\omega_1^2} - \frac{2\,\text{Cov}(J_1,J_2)}{\omega_1\omega_2} + \frac{\text{Var}(J_2)}{\omega_2^2}\right]$ and the variance weight $h_{\Omega,J_1,J_2}(\cdot)$. The variance of the difference in the instruments is identified from the distribution of Z given X. The weights $h_{\Omega,J_1,J_2}(\cdot)$, can also be estimated from the data but are less transparent. For each of the examples presented in Figure 4, we plot the variance weights $h_{\Omega,J_1,J_2}(\cdot)$. In the case of the normal generalized Roy model, the weights are more intuitive and more easily calculated when conditioning directly on $V = v$ (rather than $U_D = u_D$), so we plot them as a function of v. Figure 5 plots the variance weights. *Ceteris paribus*, the larger the variance weights, the larger is the variance of the difference in the IV estimators and hence the lower the power of a test based on this difference. In Panel A of Figure 5 we see that when the mixing proportion is 0.45 the variance of the difference in the estimators is higher than when the mixing proportion is 0.75 due to the fact that the IV weights covary highly when the instruments are closer to normal so the variance of their difference is smaller. In Panel B, the variance weights are roughly similar for $E(Z) = 1$ and $E(Z) = -0.5$. Finally, in Panel C the variance weights are much larger when $E(Z) = 1$ than when $E(Z) = 0$. This demonstrates that even when the IV weights are nonoverlapping, as is the case in both examples in Panel C, the variance of the difference in the IV estimators will generally depend on the distribution of Z.

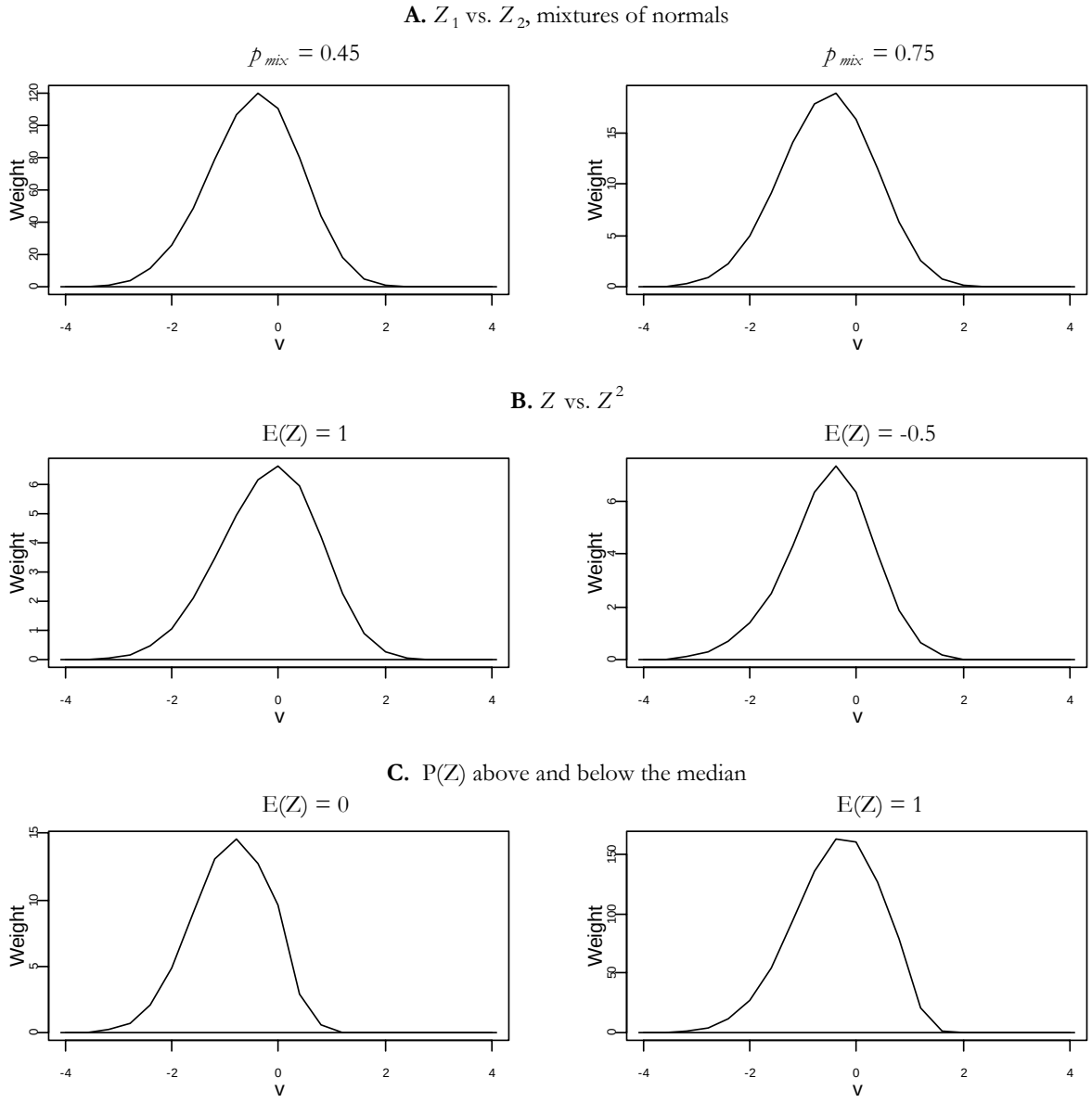
We emphasize that the specific comparisons of IV estimators presented in this section are illustrative examples. Our formal analysis is completely general and allows for any choice of valid instruments which satisfy (A-1)–(A-5).

4 Testing H_0 by Testing for Linearity

We next consider tests of H_0 based on linearity in p. Keeping the conditioning on X implicit, we can write (3) as

$$E(Y \mid P(Z) = p) = \mu + g(p) \tag{13}$$

Figure 5: IV variance weights ($h_{\Omega,J_1,J_2}(\cdot)$) as a function of $V = v$ for alternative choices of instruments.



Note: Panel A plots the variance weights of the difference in the IV estimates constructed using either Z_1 or Z_2 as an instrument where (Z_1, Z_2) is distributed as a multivariate mixture of normals, with D = **1**($\gamma_1 Z_1 + \gamma_2 Z_2 \geq V$). To construct these results, we assume

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim p_{mix} \times N\left(\begin{pmatrix}-0.8\\1\end{pmatrix}, \begin{pmatrix}1.4 & 0.5\\0.5 & 1.4\end{pmatrix}\right) + (1 - p_{mix}) \times N\left(\begin{pmatrix}-0.8\\1\end{pmatrix}, \begin{pmatrix}0.6 & -0.3\\-0.3 & 0.6\end{pmatrix}\right)$$

and the coefficients in the choice equation are γ_1=0.2, γ_2=1. In the left plot of Panel A we let p_{mix} = 0.45 and in the right plot p_{mix} = 0.75. Panel B plots the variance weights of the difference in the IV estimates constructed using either Z or Z^2 as an instrument where Z ~ N(μ_Z,1), μ_Z = 1 or μ_Z = -0.5, and D = **1**(Z ≥ V). Panel C plots the variance weights of the difference in the IV estimates constructed using either P(Z) below the median or P(Z) above the median as instruments. In Panel C, Z ~ N(μ_Z,1), μ_Z = 0 or μ_Z = 1, and D = **1**(Z ≥ V). In all of the plots, we set σ_V^2 = 1.

for some general nonlinear function $g(\cdot)$ where μ and g may depend on X. Our test for the absence of selection on the gain to treatment is a test of whether the function $g(\cdot)$ belongs to the linear parametric family $\mathcal{F} = \{a + bp, (a, b) \in \mathbb{R}^2\}$. Let $\mathcal{P}$ be the support of $P(Z)$, with typical element $p \in \mathcal{P}$. The null hypothesis of linearity can be written as

$$H_0^L : \text{There exists some } (a, b) \in \mathbb{R}^2 \text{ such that } g(p) = a + bp \text{ for almost all } p \in \mathcal{P},$$

while the alternative is

$$H_A^L : \text{There exists no } (a, b) \in \mathbb{R}^2 \text{ such that } g(p) = a + bp \text{ for almost all } p \in \mathcal{P}.$$

There is a large and still unsettled literature in econometrics and statistics dealing with specification tests of this type.[24] These tests proceed in one of two ways: (i) testing orthogonality restrictions implied by the parametric model, or (ii) comparing a nonparametric estimate of $g(p)$ with a parametric estimate, $\hat{a} + \hat{b}p$. We implement and explore the properties of both types of tests and briefly discuss a third test due to Li and Nie (2007).

Linearity Test 1: Wald Test Based on Series

The first test of linearity of $E(Y|P(Z) = p)$ in p determines whether terms in addition to p are required to fit the data. It is instructive to consider the case of the normal selection model as a baseline. When the data are generated from the normal generalized Roy model, we can characterize $E(Y|P(Z) = p)$ by

$$E(Y|P(Z) = p) = \bar{\alpha} + \bar{\beta}p + \tau \int_0^p \Phi^{-1}(u_D) du_D.$$

Figure 6 plots $E(Y|P(Z) = p)$ for alternative values of τ. This figure provides, in addition to the plots of $E(Y|P(Z) = p)$, the R^2 of a regression of $E(Y|P(Z) = p)$ on a linear term in p as

[24] See, e.g., Horowitz and Spokoiny (2001) and the references therein. The properties of particular tests depend on the specification of alternatives.

well as the R^2 after adding a quadratic term in p. The R^2 will depend on the distribution of the propensity score (the regressor). We present the R^2 for two different distributions of the propensity score. In column A of Figure 6 the distribution of the propensity score is uniform, while in column B the propensity score is concentrated in one half of the unit interval. When the propensity score is concentrated in one part of the unit interval, a linear function of p is a very good approximation to $E(Y|P(Z) = p)$. Note, however, that no matter what the distribution of the $P(Z)$, a quadratic function is able to closely approximate $E(Y|P(Z) = p)$. This suggests that for a normal alternative one can use a quadratic function in p to estimate $E(Y|P(Z) = p)$.

In the general non-normal case, we use polynomials to approximate classes of smooth alternatives for the function $g(\cdot)$. We estimate $E(Y|P(Z) = p)$ using polynomials of degree 2 or higher. Polynomials approximate well a broad class of functions. Exploring power in this class gives us an indication of the power of our procedures against such alternatives.[25] Our specification of a more general, but still parametric, alternative model is

$$g(P(Z)) = \sum_{l=0}^{L} \phi_l (P(Z))^l,$$

where L is assumed to be known.[26] Our test for linearity is

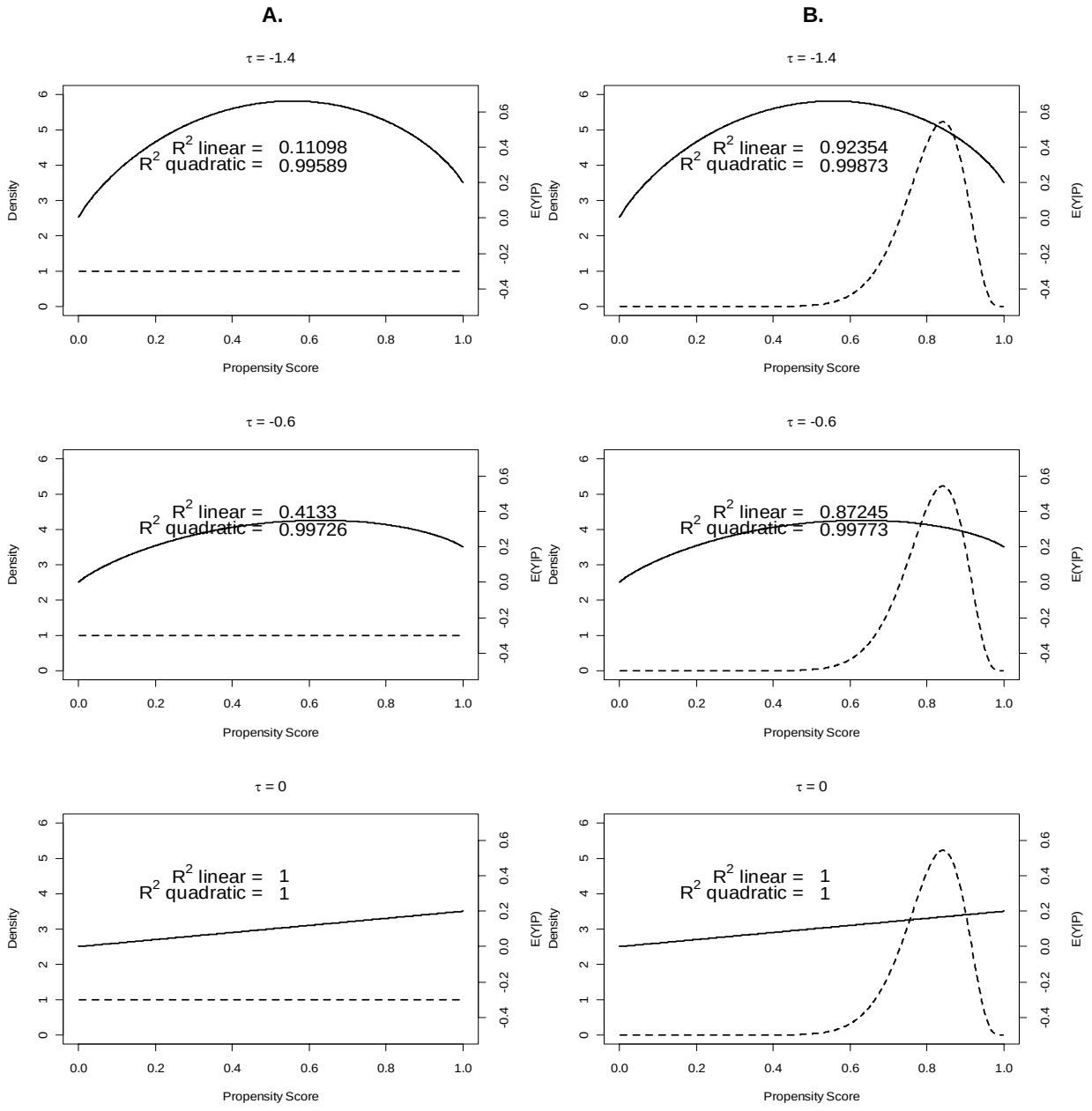
$$H_0 : \ \phi_l = 0 \text{ for } l = 2, \ldots, L$$
$$H_A : \ \phi_l \neq 0 \text{ for some (or all) } l = 2, \ldots, L.$$

We fit models with many different choices for the degree of the polynomial. We interpret a rejection of linearity in any of these models as a rejection of the null hypothesis of linearity. However, we need to use caution in constructing the critical values for our test statistics. If we conduct tests for linearity in each model separately, as we add more tests (i.e., more

[25] Ichimura and Todd (2007) discuss the properties of series estimators. Newey (1997) establishes convergence rates and proves asymptotic normality of such estimators.

[26] Below, we discuss a procedure when L is unknown.

Figure 6: $E(Y|P)$ (the solid line) and the distribution of propensity scores (the dashed line) in the normal generalized Roy model.



Note: The R^2 linear is calculated as the R^2 of a regression of the true E(Y|P) on a linear term in P(Z). The R^2 quadratic is the R^2 of a regression of the true E(Y|P) on a quadratic polynomial in P(Z). In column A, the single instrument Z is distributed N(0, 1), the unobservable in the choice equation, V, is distributed N(0, 1) and the choice equation is given by D = **1**(Z ≥ V), which results in uniformly distributed propensity scores. In column B, the single instrument Z is distributed N(1, 0.1), V is distributed N(0, 1) and the choice equation is D = **1**(Z ≥ V), which results in the shifted distributed of propensity scores shown.

polynomials for series estimators of ever higher degree), we would, in general, increase the probability of type I error. However, analogous to the results in the literature on multiple hypothesis testing, we construct critical values that allow us to control the probability of type I error.

Specifically, suppose that we estimate the function $g(\cdot)$ using $M(= L-1)$ different models corresponding to adding additional polynomial terms. We seek to use the information from all of these estimators to test for the linearity of $g(\cdot)$. We have statistics for the tests of linearity for each of those models separately. Call them $T_1, \ldots, T_M$. For each model, we know the asymptotic distribution under the null hypothesis. Because in our case these test statistics will not in general have the same asymptotic distribution under the null, to make them comparable, we convert them into p-values, denoted $q_1, \ldots, q_M$, which are distributed unit uniform under the null. Since we are looking for a deviation from linearity in any of the models, the only p-value that will be relevant for this decision will be the smallest one, namely

$$q^* = \min_{m \in \{1,\ldots,M\}} \{q_m\}.$$

This will be the p-value corresponding to the most significant test statistic. To make the size of the test 0.05, one cannot simply compare q^* to 0.05. To obtain the distribution of this statistic we use a bootstrap procedure developed in Romano and Wolf (2005) and Romano and Shaikh (2006). Their procedure works in this application because the test used in this paper can be viewed as the first step in their "stepdown" procedure.[27,28] However, in the results presented below we do not report stepdown p values for the remaining hypotheses because we are only interested in testing the first round null hypothesis that no higher order polynomial in p enters (14).

The Web Appendix, Section 7, presents a detailed description of the testing procedure

[27] Notice that we are not testing individual coefficients for the M different polynomials, but an entire class of polynomials of order 2 or higher, up to order $k \leq M + 1$.

[28] Alternatively we could use a χ^2 test (or F-test in small samples) of the hypothesis that the coefficients associated with the polynomials of order two and higher are all zero.

used in this paper and shows how we construct the critical value against which to compare q^*. In our applications of series estimators, we use four polynomials in $P(Z)$ of increasing degree from degree 2 to degree 5 ($M = 4$). In simulations available on our website, we confirm that this procedure is able to control the size of the test.[29,30]

In order to examine the power of this test in detecting deviations from H_0, for simplicity we will consider the test which fits a quadratic polynomial to $E(Y|P(Z) = p)$ and tests whether the coefficient on the quadratic term is zero. We test the significance of this coefficient using a Wald test. Again, we use the normal generalized Roy model as a simple base case. As shown in Figure 6, a quadratic polynomial very closely approximates the function $\int_0^{P(Z)} \Phi^{-1}(t)dt$ and therefore we expect this test to have power to detect deviations from H_0

[29]In the Web Appendix, we present simulations summarized in Figures A11 and A12, which are discussed below. They show that for a variety of configurations of the parameters, under the null hypothesis our procedure never rejects more than 5% of the time at the 0.05 level. See `http://jenni.uchicago.edu/testing_random/`.

[30]To implement the test, the econometrician would also need to account for conditioning variables X, which we have thus far kept implicit, and for the estimation of propensity scores, $P(Z)$. Many of the X variables that we use are categorical or binary. For X variables that are categorical we suggest stratifying the data on X and perform tests within X cells. However, when the cells are too thin to allow the stratification on X and still expect to have reasonable power in small samples, we instead suggest incorporating the conditioning variables X as linear regressors and estimate an alternative specification:

$$Y_i = X_i\delta_0 + X_i(\delta_1 - \delta_0)P(Z_i) + \sum_{l=1}^{L} \phi_l P(Z_i)^l + \varepsilon_i, \tag{14}$$

where it is assumed that $E(\varepsilon_i \mid X_i, Z_i) = 0$. The rationale for the interaction between X and $P(Z)$ arises from noting that

$$\begin{aligned}
&E(Y|P(Z) = p, X = x) \\
&\quad = E(\alpha|P(Z) = p, X = x) + E(\beta D|P(Z) = p, X = x) \\
&\quad = \left[\bar{\alpha}(x) + \bar{\beta}(x)p\right] + E(U_\alpha|P(Z) = p, X = x) + E(U_\beta|D = 1, P(Z) = p, X = x)p \\
&\quad = x\delta_0 + x(\delta_1 - \delta_0)p + \kappa(p)
\end{aligned}$$

where $\kappa(p) = E(U_\alpha \mid P(Z) = p) + E(U_\beta \mid D = 1, P(Z) = p)p$. The last equality comes from independence assumption (A-1) presented in Section 2 and the assumption that α and β are linear functions of X.

When estimating $\kappa(p)$ using polynomials in p one can simply regress Y on X, $X \times P(Z)$ and polynomials in $P(Z)$. We do this in our example presented in Section 6. We estimate $P(Z)$ using a probit model. Our results are robust to alternative specifications of the choice equation.

in this model.[31] We estimate the following specification:

$$Y_i = \alpha + \beta P(Z_i) + \eta [P(Z_i)]^2 + \varepsilon_i.$$

Under the null hypothesis $\eta = 0$, the Wald test statistic formed using the least squares estimator $\hat{\eta}$ is asymptotically distributed as a χ^2_1. Under the alternative, the Wald statistic will converge to a noncentral chi-square distribution. Let $\mathbf{g}_i = (1, P(Z_i), [P(Z_i)]^2)$ be the row vector of regressors for individual i and $G = (\mathbf{g}_1, \ldots, \mathbf{g}_I)'$ denote the matrix of regressors stacked across individuals. Let $c = [0 \;\; 0 \;\; 1]$. Under our conditions, $\frac{G'G}{n} \xrightarrow{p} \Gamma$. In the normal generalized Roy model, under alternative $\tau = \frac{\tau_A}{\sqrt{I}}$, the noncentrality parameter of the chi-square distribution of the test statistic is

$$\lambda_{\text{quad}} = \frac{1}{2} \tilde{\eta}^2 \left[\sigma^2_\varepsilon c \Gamma^{-1} c' \right]^{-1} \tag{15}$$

where

$$\tilde{\eta} = \frac{\tau_A \operatorname{Cov}\left(\left[\int_0^{P(Z)} \Phi^{-1}(t) dt \right]_\perp, [P(Z)]^2_\perp \right)}{\operatorname{Var}([P(Z)]^2_\perp)}$$

and the subscript $\perp$ denotes the residuals of a variable after projecting it onto a linear function of $P(Z)$.[32] In Section 5 we examine the power of this test and its determinants, taking into account that in applications $P(Z)$ is estimated in a first stage.

[31] In the Web Appendix we derive the power of a direct test of $\tau = 0$ using a correctly specified $E(Y|P(Z) = p)$ function.

[32] To justify the expression for $\tilde{\eta}$, note that we can write

$$\begin{aligned} Y_i &= \alpha + \beta P(Z_i) + \tau_A \int_0^{P(Z_i)} \Phi^{-1}(t) dt + \varepsilon_i \\ &= \alpha + \beta P(Z_i) + \eta [P(Z_i)]^2 + \left[\tau_A \int_0^{P(Z_i)} \Phi^{-1}(t) dt - \eta [P(Z_i)]^2 + \varepsilon_i \right] \end{aligned}$$

where the term in brackets is the error term. The true η is equal to 0 but $\operatorname{plim} \hat{\eta} = \tilde{\eta} \neq 0$ due to an omitted variable bias. This bias is given by the formula in the text and is derived from first residualizing on $P(Z_i)$ and then solving for the probability limit of $\hat{\eta}$.

Linearity Test 2: Bierens Conditional Moment Test

We also consider a test of the validity of representation (3) which relies on orthogonality restrictions implied by the parametric model. We use the conditional moment (CM) test of Bierens (1990).[33] This test uses the fact that under the null hypothesis the following moment condition must be satisfied

$$E[Y - a_0 - b_0 P(Z) \mid P(Z)] = 0$$

for the true parameter vector $(a_0, b_0) \in \mathbb{R}^2$. This conditional moment restriction implies the set of unconditional moment restrictions

$$E[(Y - a_0 - b_0 P(Z)) \exp(t' \Lambda(P(Z)))] = 0 \tag{16}$$

for all $t \in \mathbb{R}$, for some bounded one-to-one, mapping Λ from $\mathbb{R}$ into $\mathbb{R}$. A test can be constructed using the sample analog of the left-hand side of (16). Bierens (1990) shows how one can use sample analogs to construct a test statistic which, under the null hypothesis, converges in distribution to a χ^2_1 and under the alternative diverges to infinity.

In analyzing the power of this test, we first consider the power of a test using the sample analog of (16) for a single t in the context of the normal generalized Roy model and then discuss how to generalize to the test we actually use. For a single t, let $\hat{Q}(t)$ denote the sample analog of the moment condition (16). Bierens (1990) shows that $\sqrt{I}\hat{Q}(t)$ is asymptotically normal and under the null has mean zero. Denote its asymptotic variance under the null by $\sigma^2(t)$, with estimator $\hat{\sigma}^2(t)$. Under H_0, the test statistic, $I(\hat{Q}(t))^2/\hat{\sigma}^2(t)$ is asymptotically distributed as a central χ^2_1. Let $g(P(Z), \theta_0) = a_0 + b_0 P(Z)$ denote the parametric function we are testing. Note that the gradient of this function with respect to the parameter vector $\frac{\partial g(P(Z),\theta)}{\partial \theta} = [1, P(Z)]'$ does not depend on the point of evaluation of θ. Consider a local

[33] See also Bierens (1982) and Bierens and Ploberger (1997) for related tests. Newey (1985) discusses conditional moment tests more generally.

alternative $\tau = \tau_A/\sqrt{I}$ in a normal generalized Roy model. Under this alternative, the test statistic is distributed asymptotically as a noncentral chi-square with one degree of freedom and noncentrality parameter

$$\lambda_{CM}(t) = \frac{1}{2}(Q_A(t))^2[\sigma_A^2(t)]^{-1}.^{34} \tag{17}$$

This expression gives the asymptotic power of a test for a single choice of t. The Bierens test maximizes the test statistic over t. Under the alternative, the test statistic is not a simple noncentral chi-square. We explore the power of this test using simulation.[35]

All of the Preceding Tests are Conditional Moment Tests[36]

This paper seeks to test if the outcome equation is generated by a model of the form

$$E(Y \mid P(Z)) = a + bP(Z)$$

which is equivalent to

$$E\left[h(P(Z))\left[Y - a - bP(Z)\right]\right] = 0$$

[34]

$$Q_A(t) = E\left[\left(\bar{\alpha} + \bar{\beta}P(Z) + \tau_A \int_0^{P(Z)} \Phi^{-1}(u_D)du_D + \varepsilon - g(P(Z), \hat{\theta})\right)\exp(t'\Lambda(P(Z)))\right]$$

$$\sigma_A^2(t) = E\left[\left(\bar{\alpha} + \bar{\beta}P(Z) + \tau_A \int_0^{P(Z)} \Phi^{-1}(u_D)du_D + \varepsilon - g(P(Z), \hat{\theta})\right)^2 \times \left[\exp(t'\Lambda(P(Z))) - \xi_A(t)\Omega^{-1}\begin{bmatrix} 1 \\ P(Z) \end{bmatrix}\right]\right]$$

$$\xi_A(t) = E\left[\frac{\partial}{\partial\theta'}g(P(Z), \theta_A)\exp(t'\Lambda(P(Z)))\right] = E\left[\begin{bmatrix} 1 \\ P(Z) \end{bmatrix}'\exp(t'\Lambda(P(Z)))\right]$$

$$\Omega = E\left[\frac{\partial}{\partial\theta'}g(P(Z), \theta_A)\frac{\partial}{\partial\theta}g(P(Z), \theta_A)\right] = E\begin{bmatrix} 1 & P(Z) \\ P(Z) & [P(Z)]^2 \end{bmatrix}$$

where $\hat{\theta}$ is the least squares estimate of θ in a regression of Y on a constant and $P(Z)$.

[35]We modify his test statistic for our applications because we need to account for the fact that the propensity scores $P(Z)$ are estimated in a first stage. Therefore, when we form the test statistic, we use an estimate of the variance of the sample analog of (16) using a bootstrap procedure in which we reestimate $P(Z)$ in each sample, rather than an estimate based on the approximate asymptotic variance of (16) derived in Bierens (1990).

[36]We thank Edward Vytlacil for suggesting this unifying approach.

for any function h. Conditional moment tests would typically use a vector of functions $h(P(z))$ to construct tests.

All of the tests previously discussed are based on different choices of $h(P)$. For the Bierens test, we use $h(P(Z)) = \exp(t'\Lambda(P(Z)))$. The test of linearity based on polynomials takes $h(P(Z)) = \left[1, P(Z), (P(Z))^2, \ldots, (P(Z))^L\right]$. The IV test can also be cast in this framework.

The *plim* of the IV estimator obtained using $J_k(Z)$, $k = 1, \ldots, K$, as an instrument are the values of (a_k, b_k) that solve

$$E\left[J_k(Z)\left[Y - a_k - b_k D\right]\right] = 0$$

and

$$E\left[Y - a_k - b_k D\right] = 0, \qquad k = 1, \ldots, K.$$

By the law of iterated expectations, this is equivalent to solving

$$E\left[J_k(Z)\left[E(Y \mid Z) - a_k - b_k P(Z)\right]\right] = 0$$
$$E\left[E(Y \mid Z) - a_k - b_k P(Z)\right] = 0,$$

which is equivalent to solving

$$E\left[J_k(Z)\left[Y - a_k - b_k P(Z)\right]\right] = 0$$
$$E\left[Y - a_k - b_k P(Z)\right] = 0.$$

For one instrument there is no test, but for two or more ($K \geq 2$), one can test if a common pair of (a, b) satisfies all of the moment conditions produced from using different instrumental variables. This is the classical test of overidentification. Thus, all of the tests previously discussed can be viewed as conditional moment tests.

Linearity Test 3: A Semiparametric Test Based on Local Linear Regression[37]

A potential problem with the test based on series estimators (Linearity Test 1) is that it assumes that the degree of the highest order polynomial in $P(Z)$ is finite and known. A semiparametric approach that did not rely on strong functional form assumptions about the generator model would be more desirable.

Recently, Li and Nie (2007) have used local linear regression methods to develop a test for linearity of an unknown parametric function in a semiparametric model. They develop a test of linearity of the unknown nonparametric component (linearity in $P(Z)$ in our setup) that can be applied to the problem analyzed in this paper if it is adapted to the case of an estimated $P(Z)$. If $P(Z)$ is parametric and its coefficients are $\sqrt{N}$ estimable, their analysis can be applied directly. The case where $P(Z)$ is estimated nonparametrically is left for another occasion.

Li and Nie (2007) conduct a Monte Carlo study of their approach. They show good size and power properties for their test statistic. Their test can be interpreted as a local conditional moment test.

Conditioning on X

Throughout, we have conditioned on X. An important practical problem not addressed in this paper but common to all empirical models is picking the appropriate conditioning set, and determining how to explicitly model the dependence of Y on X.

5 The Power of the Tests

We use the generalized Roy model as our base case because it allows for a simple parameterization of the correlation between β and D. We consider a more general polynomial alternative in Section 5.4. Our results quantify the sample size needed to produce tests with

[37] We thank Xiaohong Chen for directing us to this paper and clarifying our thinking about semiparametric approaches to testing for linearity.

substantial power. We establish that the distribution of the propensity scores, $P(Z)$, is an important determinant of power.[38]

5.1 Normal Generalized Roy Model as the Data Generating Process

We simulate data from the generalized Roy model for a range of parameter values given in Table 2. As a base case, we consider a scalar normal instrument Z, ie. $Z \sim N(0, \sigma_Z^2)$. However, we also analyze models with multiple instruments and a variety of distributions for the instruments.

Table 2: Initial Specification used to calculate the power of the tests.

Outcomes	Decision Rule:
$Y_0 = \mu_0 + U_0$ $Y_1 = \mu_1 + U_1$	$D = \mathbf{1}(\alpha_D + \gamma Z \geq V)$
	Observed $Y = DY_1 + (1-D)Y_0$
with parameters: $\mu_0 = 0$ $\mu_1 = 0.2$	with parameters: $\alpha_D = 0$ $\gamma = 1$

Distribution of Unobservables:

$$\begin{pmatrix} U_1 \\ U_0 \\ V \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_U^2 & 0 & \rho_{1V} \\ 0 & \sigma_U^2 & -\rho_{1V} \\ \rho_{1V} & -\rho_{1V} & 1 \end{pmatrix} \right)$$

where $\rho_{1V} = \mathrm{Cov}(U_1, V)/\sqrt{\mathrm{Var}(U_1)\,\mathrm{Var}(V)}$ and we assume $\rho_{0V} = -\rho_{1V}$. We vary σ_U^2 between 0.1 and 2. We consider values of ρ_{1V} from -0.7 to 0.7. Values outside of this interval result in a covariance matrix that is not positive definite.

Distribution of Observables:

Normal case: $Z \sim N(\mu_Z, \sigma_Z^2)$

We calculate the power function for values of σ_Z^2 between 0.1 and 2 and $\mu_Z \in \{0, 1\}$.

Mixture of Normals case:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim p_{mix} \times N\left(\begin{pmatrix} -0.8 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{pmatrix} \right) + (1 - p_{mix}) \times N\left(\begin{pmatrix} -0.8 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{pmatrix} \right)$$

We calculate the power function for values of $p_{mix} \in \{0.45, 0.75, 0.95\}$.

[38] We do not investigate the power of the Li and Nie (2007) test because they already conduct a Monte Carlo study of their test for linearity.

The parameterization in Table 2 makes what seem to be two fairly restrictive assumptions about the covariance structure of the unobservables in the model. It assumes that $\rho_{10} = 0$, that is, that the covariance between the two potential outcomes is zero, and $\rho_{0V} = -\rho_{1V}$. These assumptions are not as restrictive as they may appear to be for calculating the power of our tests. In the normal generalized Roy model, the nonconstancy of the MTE in u_D (or v) depends only on the term $\tau = \rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$ and so for a given value of that index, the value of ρ_{10} does not affect the power of the test. The derivation of the power of the tests shows that it depends on τ, the distribution of $P(Z)$ and the precision with which we can form our estimators (which will depend on the variances of the unobservables). We allow all of these quantities to vary in our simulations.

5.2 The Power of the IV Tests

We investigate the following model without regressors:

$$Y = \mu_0 + (\mu_1 - \mu_0)D + \xi$$

where $\xi = (U_1 - U_0)D + U_0$, and $\xi \not\perp\!\!\!\perp D$. As explained in Section 3.1, we construct a test based on the difference in IV estimators. We could potentially use any two IV estimators to conduct this test. We do not derive the optimal pair or set of instruments that maximize the power against a given alternative. We consider the three examples from Section 3.1: (i) Z_1 and Z_2, with at least one being non-normal; (ii) Z is scalar and we use Z and Z^2, a nonlinear function of the instrument; and (iii) $P(Z)$ over distinct intervals of its support.

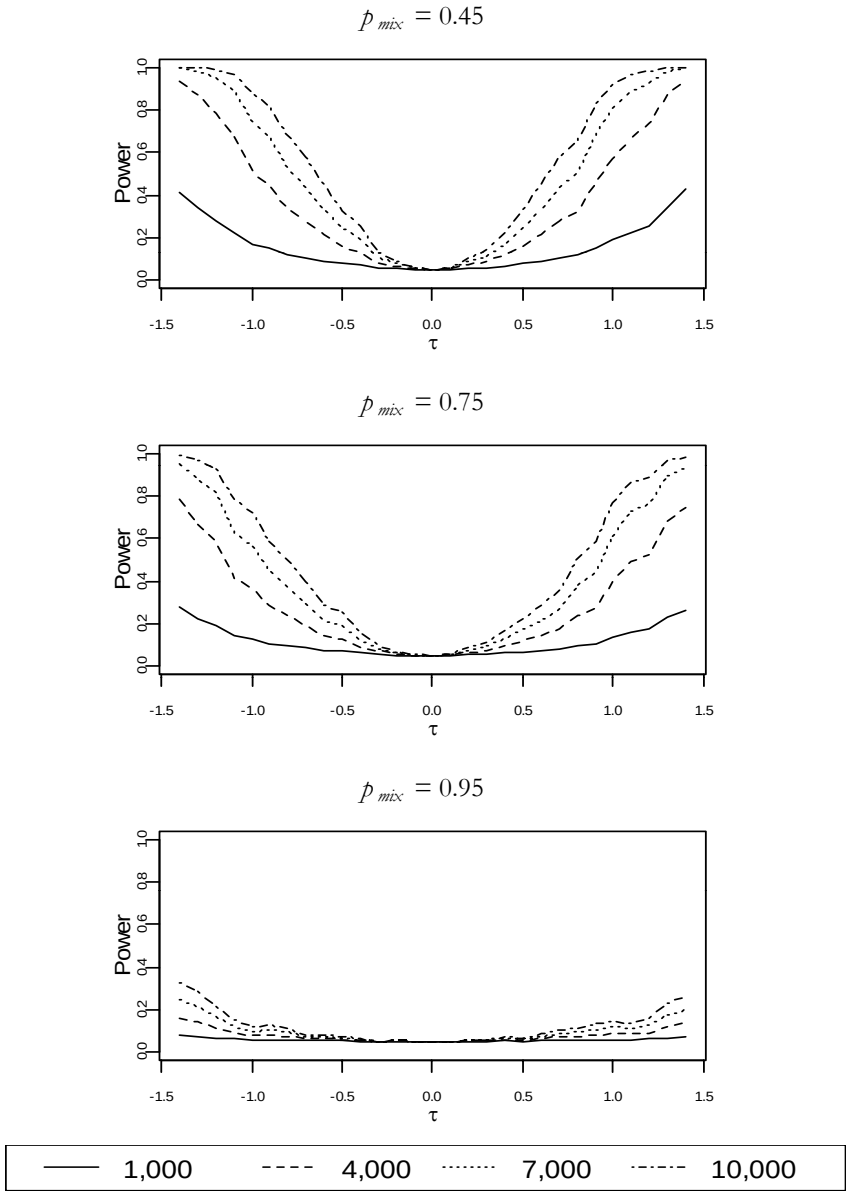
Consider first the power of a test based on the equality of IV estimators using Z_1 and Z_2 as instruments, where (Z_1, Z_2) is distributed as a mixture of bivariate normal distributions. This test was discussed in Section 3.1, where we show the weights that each of the IV estimators places on the MTE, as well as the variance weights of the difference between the two estimators. Combining this information, we can, for alternative parameterizations of the

normal generalized Roy model, calculate the power of a Wald test based on the difference between these two IV estimators. Figure 7 plots the power functions for this test. Each of the plots shows the power as a function of τ, which is a measure of deviation from the null of a flat MTE in the generalized Roy model. The three plots each fix p_{mix} at alternative values, from 0.45 to 0.95. Moving from 0.45 to 0.75 the power of the test declines. When $p_{mix} = 0.75$, the IV weights are more similar than when $p_{mix} = 0.45$ but the variance of the difference in the instruments becomes smaller. As the mixing proportion approaches 1, and the instruments approach normality, the power of the test greatly diminishes. This is predicted by our analysis that establishes the lack of power of normal instruments in a normal generalized Roy model.

We next examine the power of a test based on a nonlinear function of the single instrument Z – that is using Z vs. Z^2. In this case $D = \mathbf{1}(Z \geq V)$ and $Z \sim N(\mu_Z, \sigma_Z^2)$. This test was discussed in Section 3.1, where we plotted the weights of each of the IV estimators as well as the variance weight for the difference between the two estimators. Figure 8 plots the power of this test for different distributions of the instrument Z. We consider values of $\mu_Z \in \{-0.5, 1\}$ and values of $\sigma_Z^2 \in \{0.5, 1, 2\}$. The variance of the unobservable in the choice equation, σ_V^2, is fixed at 1. The left column of Figure 8 plots the power for $\mu_Z = 1$. For a fixed μ_Z, the power of the test is increasing in σ_Z^2. The figures in the right column plot the power for $\mu_Z = -0.5$. The power is uniformly higher when $\mu_Z = -0.5$ than when $\mu_Z = 1$, which is in accordance with the plots of the weights provided in Section 3.1.

Finally, we consider a test of equality of IV estimates formed using observations with $P(Z)$ in different intervals of the support of $P(Z)$. We separate the observations according to whether they lie above or below the sample median of $P(Z)$. As in the example considered in Section 3.1, we let $D = \mathbf{1}(Z \geq V)$ where $Z \sim N(\mu_Z, \sigma_Z^2)$. Figure 9 plots the power of this test for alternative values of $\mu_Z \in \{0, 1\}$ and $\sigma_Z^2 \in \{0.5, 1, 2\}$. As with the test based on IV using Z and Z^2, the power of this test is increasing in σ_Z^2. The power of the test is uniformly lower when $\mu_Z = 1$ than when $\mu_Z = 0$ because when $\mu_Z = 1$, one of the IV estimators places

Figure 7: Power of a Wald test of the equality of IV estimators formed using Z_1 and Z_2, mixtures of normals instruments, as a function of τ.

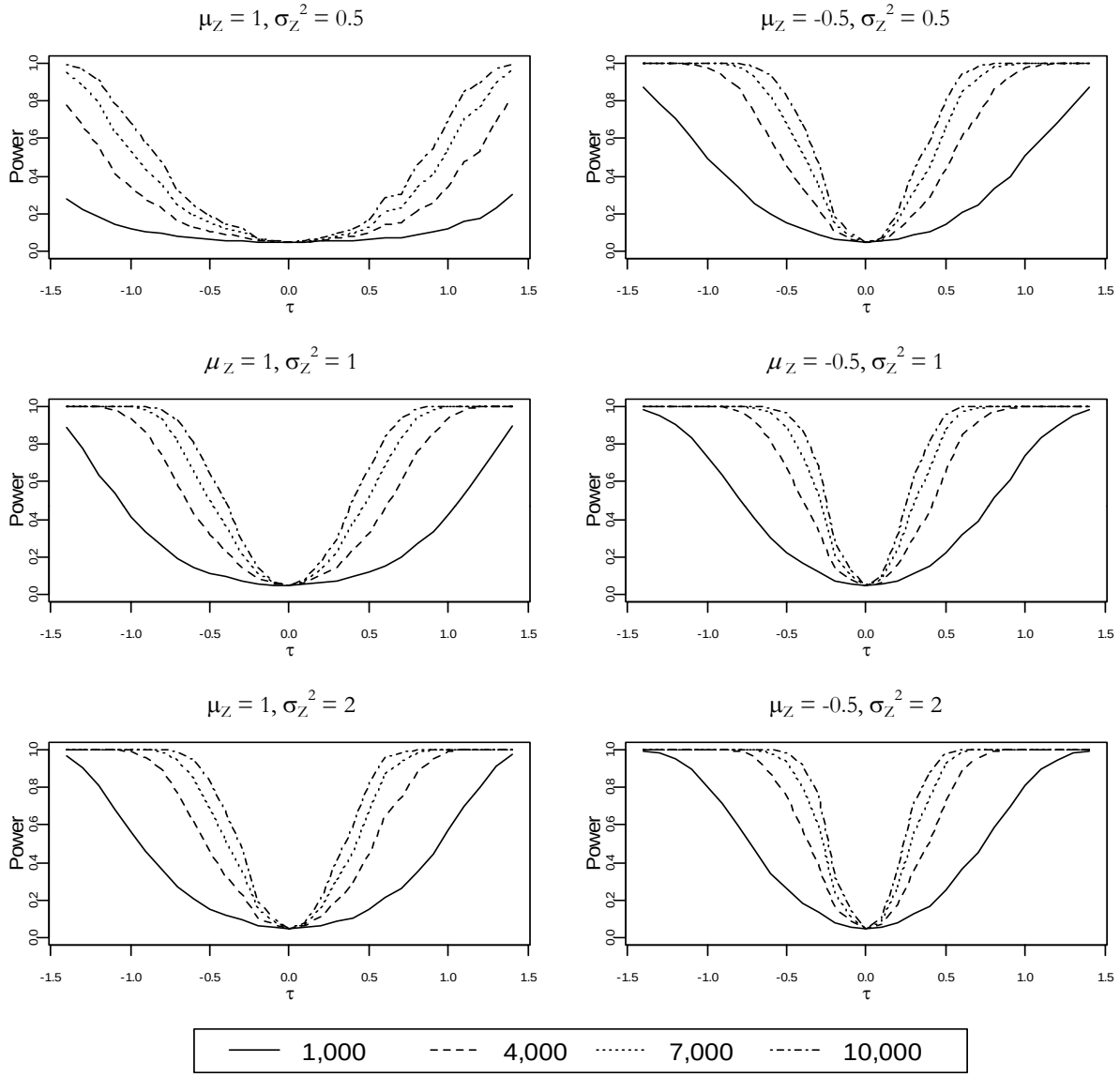


Note: Standard errors are computed from the formulae developed in the text. Each line plots the power for a different sample size, as indicated by the legend. The data generating process is the normal generalized Roy model. The variance of the unobservables in the outcome equations is fixed at 1, as is the variance of the unobservable in the choice equation. The choice equation is $D = \mathbf{1}(\gamma_1 Z_1 + \gamma_2 Z_2 \geq$

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim p_{mix} \times N\left(\begin{pmatrix} -0.8 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{pmatrix}\right) + (1 - p_{mix}) \times N\left(\begin{pmatrix} -0.8 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{pmatrix}\right)$$

and the coefficients in the choice equation are γ_1=0.2, γ_2=1. The power functions plotted are the power of a Wald test of the equality of IV estimates formed using Z_1 and Z_2 as instruments.

Figure 8: Power of a Wald test of the equality of IV estimators formed using Z and Z^2, as a function of τ.



Note: Standard errors are computed from the formulae developed in the text. Each line plots the power for a different sample size, as indicated by the legend. The data generating process is the normal generalized Roy model. The variance of the unobservables in the outcome equations is fixed at 1, as is the variance of the unobservable in the choice equation. The choice equation is $D = \mathbf{1}(Z \geq V)$. The single instrument Z is distributed $N(\mu_Z,\sigma_Z^2)$. The power functions plotted are the power of a test of equality of the IV estimate formed using Z as an instrument and the IV estimate formed using Z^2 as an instrument.

all of its weight on a relatively small segment of the MTE. In further examples presented below, we show that, as expected, σ_Z^2 is an important determinant of the power of all of the tests we consider.

In the next section, we analyze the power of an IV test which separates $P(Z)$ across distinct intervals of its support into quartiles of $P(Z)$. This test considers any rejection of pairwise equality in any of the comparisons as a rejection of the null. In order to control for the size of the test, we use the stepdown procedure of Romano and Wolf (2005). We analyze the performance of this test along with other test procedures.

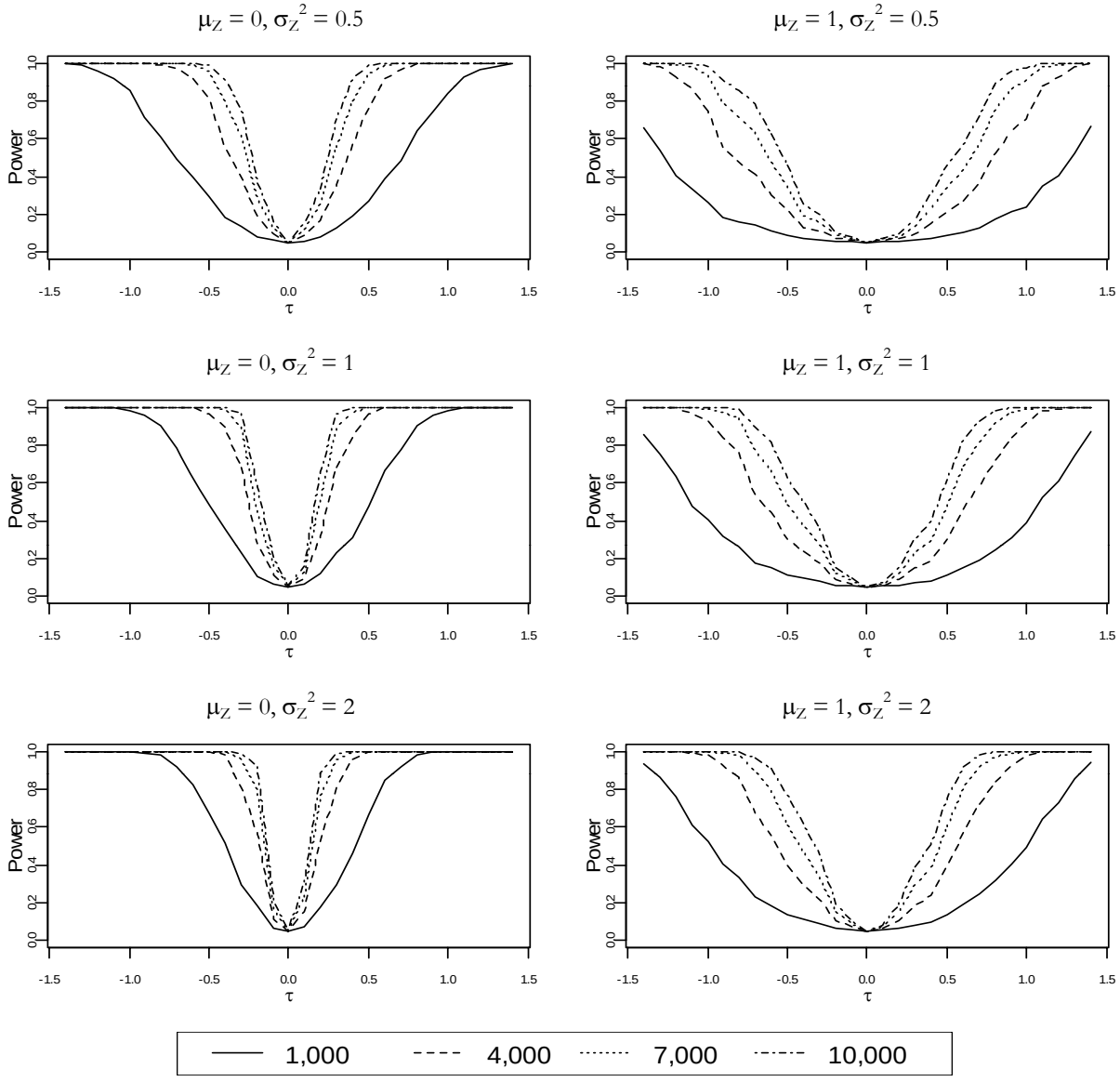
5.3 Comparative Power

We now compare the power of IV-based tests to power for the first two tests of linearity described in Section 4. The power is calculated for each test using a Monte Carlo procedure described in the Web Appendix, Section 8. Figure 10 plots the power for all of the tests analyzed in this paper.

We start the discussion with the IV test. We use the IV estimators which separate the observations across distinct intervals of the propensity score. We discuss two IV tests constructed on this principle. They vary dramatically in their power. One has relatively high power and the other relatively low power.

The power of the test of the equality of the estimates using the propensity score above and below the median as the instruments is given by the dotted line. The power of the test separating the data by quartiles of the propensity score is given by the dot-dashed line. Panel A of that figure fixes all of the parameters of the model and varies the sample size. As expected, the test is consistent. The size of the departures from the null can be gauged by noticing that with our parameterization, the restriction that the covariance matrix of the unobservables is positive definite implies $|\tau| < 1.4$. Therefore we can interpret a value of $\tau = 0.7$ as a large deviation from the null. The IV test based on separating the data into smaller intervals and conducting more pairwise comparisons tends to have lower power than

Figure 9: Power of a Wald test of the equality of IV estimators using $P(Z)$ as an instrument when the sample is separated by whether $P(Z)$ lies above or below the median.



Note: Standard errors are computed from the formulae developed in the text. Each line plots the power for a different sample size, as indicated by the legend. The data generating process is the normal generalized Roy model. The variance of the unobservables in the outcome equations is fixed at 1, as is the variance of the unobservable in the choice equation. The choice equation is D = **1**(Z ≥ V). The single instrument Z is distributed N(μ_Z,σ_Z^2). The power functions plotted are the power of a test of equality of the IV estimate formed using P(Z) as an instrument where the sample is separated according to whether P(Z) is above or below the median of P(Z).

the test based on the estimates above and below the median. Although one can test more distinct intervals, the estimates in each interval are less precisely estimated. In addition, the IV test based on the estimates above and below the median has markedly higher power than the series test, especially for small sample sizes.

Panel B of Figure 10 plots the power of the test examining another dimension. These plots hold the sample size fixed but change the "signal-to-noise ratio" which we define as

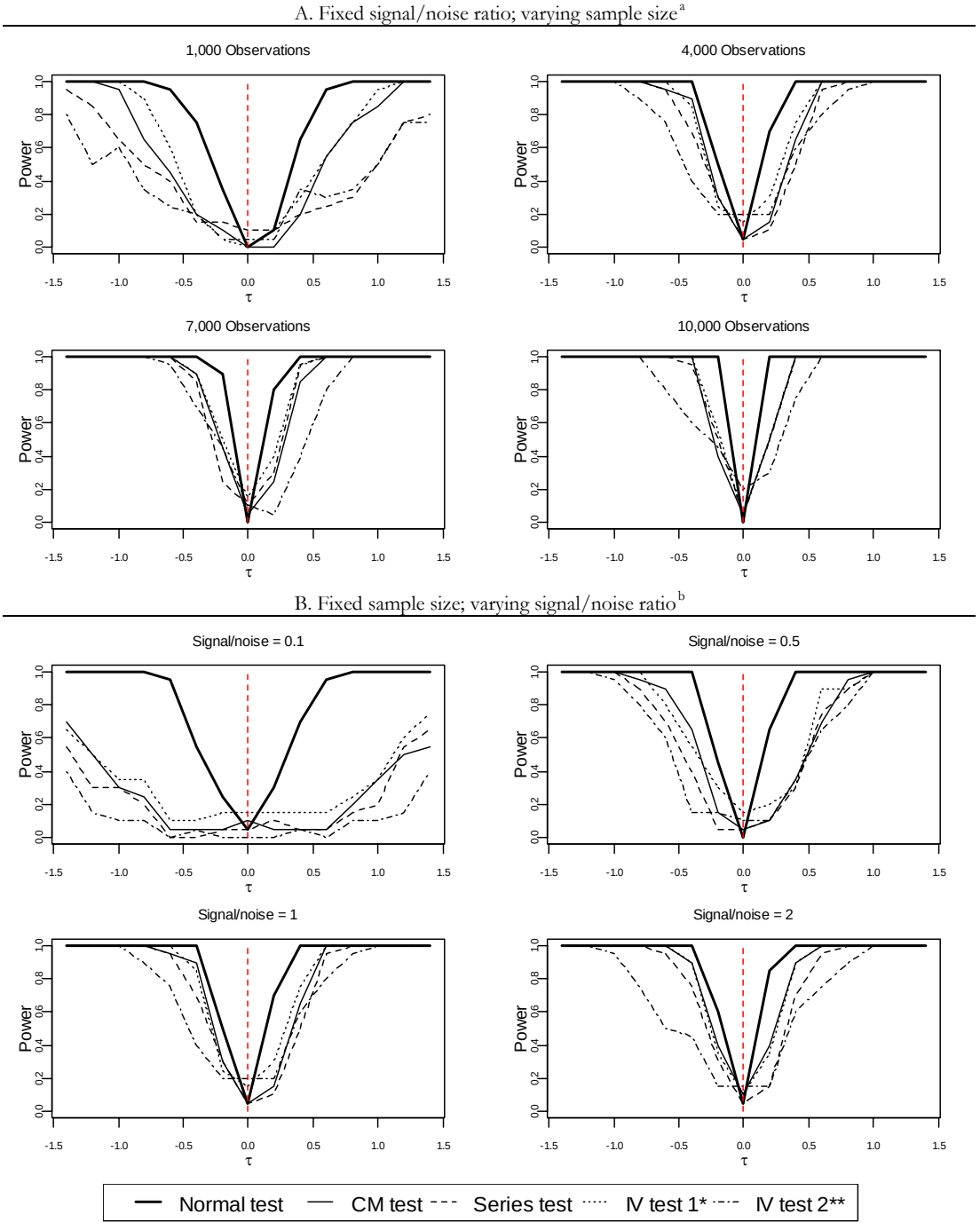
$$\frac{\text{Signal}}{\text{Noise}} = \frac{\text{Var}(Z\gamma)}{\text{Var}(V)} = \frac{\gamma^2 \sigma_Z^2}{\sigma_V^2}.$$

This parameter measures the predictive power of the instrument in determining the treatment choice. A relatively weak instrument corresponds to a small value of the signal-to-noise ratio. We can see from these plots that this ratio substantially affects the power of the test. In particular, when the signal-to-noise ratio is low, the test is unable to detect very large deviations from the null, even at a sample size of 4,000.

The signal-to-noise ratio affects the power of the tests through the distribution of propensity scores. A low signal-to-noise ratio results in very little dispersion in the propensity scores. This implies that estimates of $E(Y|P(Z) = p)$ will be based on a narrow range of values of $P(Z)$. This makes nonlinearity more difficult to detect. A linear function may be a good approximation to the true function in a small neighborhood of $g(p)$. Over a large stretch of values of p, linearity is a good approximation to the normal MTE; see figure 6.

As a reference, we plot the power of a test of H_0 based on whether or not $\tau = 0$ using maximum likelihood in a normal generalized Roy model. The power of this test is given by the heavy line. This test outperforms the other tests, as expected, but especially so when the signal-to-noise ratio is low. This is because with a low signal-to-noise ratio, the range of values of $P(Z)$ is small. This small range degrades the power of the less parametric tests but does not significantly hurt the parametric test which completely specifies the form of the MTE.

Figure 10: Power of the tests for selection on the gain to treatment as a function of τ.



Note: The data are generated according to the normal generalized Roy model given in the text. The choice equation is $D = \mathbf{1}(Z \geq V)$ where $Z \sim N(0,\sigma_Z^2)$. Standard errors are obtained from the bootstrap.

[a] Panel A shows the power functions when fixing the signal to noise ratio at 1 (this means $Var(Z\gamma)=1$, compared to the variance of the unobservable in the choice equation of 1) and changing the sample size.

[b] Panel B shows the power functions for the fixed sample size of 4,000, but changing the signal to noise ratio from 0.1 to 2. The signal to noise ratio is defined as $Var(Z\gamma)/Var(V)$.

* IV test 1 separates the observations into those with propensity scores in the interval $[0,P_{median}]$ and those with propensity scores in the interval $[P_{median},1]$ and testing the equality of the IV estimates based on those subsamples.

** IV test 2 separates the observations into quartiles of the propensity score. It tests for pairwise equality of the IV estimates calculated on each of those four subsamples and rejects the null if equality is rejected in any of the pairwise comparisons. It controls the size of the test using the method of Romano and Wolf (2005), as described in the text.

To investigate the relationship between the signal-to-noise ratio, the deviation from the null hypothesis (in the form of τ), the departure from linearity and the power of the tests, we define the following R^2 measure:[39]

$$1 - R^2 = \frac{E\left[(E(Y|P(Z) = p) - E^*(Y|P(Z) = p))^2\right]}{\text{Var}(E(Y|P(Z) = p))}$$

where $E^*(Y|P(Z) = p)$ is the linear projection of Y on $P(Z)$. This quantity is one minus the R^2 from a regression of the true expectation $E(Y|P(Z) = p)$ on a linear function in $P(Z)$. Under H_0, $1 - R^2 = 0$. Under the alternative, the magnitude of $1 - R^2$ is a measure of the degree to which the alternative differs from the null. We use $1 - R^2$ as a summary statistic of deviation from the null and plot the power of all of our tests as a function of it. It is a useful summary statistic because it does not rely on the normality of the base model. We use it for other specifications of the model which we discuss below.

Figure 11 follows the same general format as Figure 10, but graphs power as a function of $1 - R^2$, rather than τ.[40] Each plot presents only the right half of the corresponding plot in figure 10 (the $\tau > 0$ halves) because $1 - R^2$ is symmetric in τ. Panel A fixes the signal-to-noise ratio and plots the power for different sample sizes. At small sample sizes, the series test and the IV test using quartiles of the propensity score have the lowest power, but at larger sample sizes, the IV test using quartiles has lower power than all of the other tests.[41]

Panel B fixes the sample size and shows the power functions for different signal-to-noise ratios. In these plots, the greater the departure of the null from linearity, the greater the power of the tests. In addition, changing the signal-to-noise ratio changes $1 - R^2$. Therefore, to the extent that the power is different for different signal-to-noise ratios in Figure 10, those differences arise due to differences in the distribution of propensity scores. That is, at a

[39] Measuring deviations from the null in this fashion was suggested to us by Edward Vytlacil.

[40] Although we define $1 - R^2$ as a population quantity, we use the sample analog, calculated from a sample of 10,000. We simply want to choose a large number of observations in order to get a good estimate of the distribution of the propensity score $f_{P(Z)}$ with which to calculate $1 - R^2$.

[41] At sample size 10,000, the power functions for the conditional moment test, the series test and the test of the equality of the IV estimates above and below the median lie on top of each other.

given level of $1 - R^2$, the signal-to-noise ratio has no effect on the power of the tests (up to simulation error).[42]

5.3.1 Tests for Linearity

Test 1: Wald Test Based on Series Estimators

We next calculate the power of the test of the linearity of $E(Y|P(Z) = p)$ in p which compares the estimate based on a parametric (linear) estimator to a more general polynomial series estimator.[43] First, we calculate analytically the power of the test using only a quadratic polynomial in p based on expression (15) in Section 4. This allows us to precisely determine the power of the test across the dimensions we find to be important. Figure 12 plots the power of this test at different sample sizes for fixed alternatives. Inspection of these power functions shows that, as in the IV tests, the power is increasing in σ_Z^2. As expected, it is decreasing in $\sigma_0^2 = \sigma_1^2 = \sigma_U^2$.

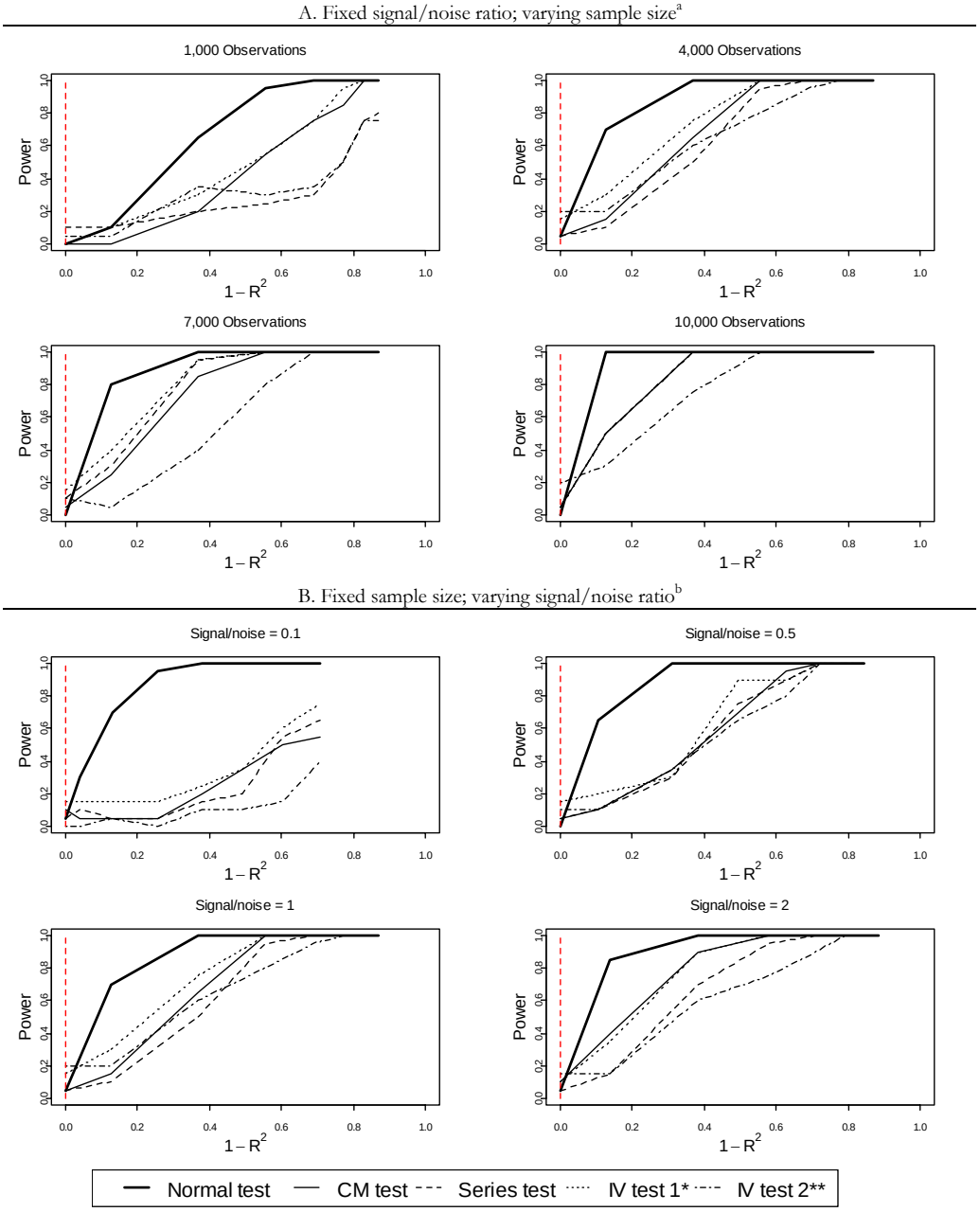
We have shown that in the case of a model with normal errors, a quadratic polynomial in $P(Z)$ is able to approximate $E(Y|P(Z) = p)$ well, and therefore will have high power in detecting deviations from H_0. However, in the non-normal case, $E(Y|P(Z) = p)$ is a general function which may not be well-approximated by a quadratic polynomial. Therefore, we construct a test which relies on fitting series estimators of $E(Y|P(Z) = p)$ of varying degrees.

The test statistic is distributed as the maximum over a set of noncentral chi-square

[42] In the Web Appendix, we plot additional figures which trace out the power of the specific IV tests more thoroughly across the three dimensions which we find are important in this model – the deviation from the null (as measured by τ), sample size, and the signal to noise ratio. Figures A2 through A4 are plots for the tests based on the IV estimates above and below the median and Figures A5 through A7 for the tests based on IV estimates for separate quartiles. We also present the power of a test of whether the IV estimate using only observations with propensity scores below the 40th percentile and the IV estimate using only observations with propensity scores above the 60th percentile are equal in Figures A8 through A10. Using more intervals and fewer observations in general reduces power.

[43] Here we describe a procedure for a model that does not include conditioning variables X since in our simulations we do not have such covariates. However, in the empirical examples discussed below, we use such regressors. The test consists of regressing Y on X, X interacted with $P(Z)$ and a polynomial in $P(Z)$ and testing for the joint significance of the coefficients on the nonlinear terms in $P(Z)$. When there is no X, the model simplifies appropriately.

Figure 11: Power of the tests for selection on the gain to treatment as a function of $1 - R^2$, normal instrument.



Note: The data are generated according to the normal generalized Roy model given in the text. The choice equation is $D = \mathbf{1}(Z \geq V)$ where $Z \sim N(0,\sigma_Z^2)$. Standard errors are obtained from the bootstrap.

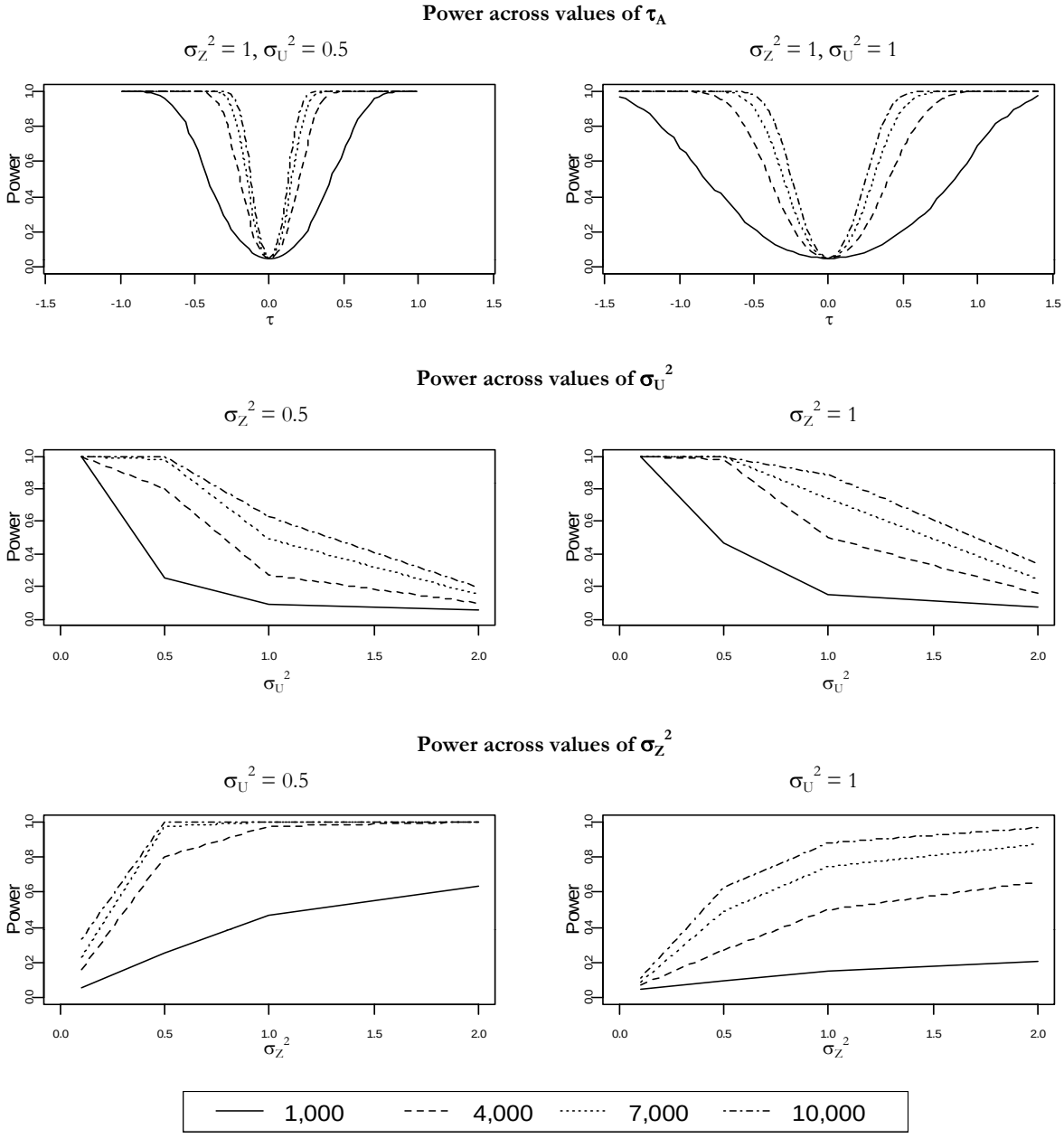
[a] Panel A shows the power functions when fixing the signal to noise ratio at 1 (this means $Var(Z\gamma)=1$, compared to the variance of the unobservable in the choice equation of 1) and changing the sample size. For a given configuration of parameters, R^2 is the R^2 from a regression of the true $E(Y|P)$ on a linear function of P.

[b] Panel B shows the power functions for the fixed sample size of 4,000, but changing the signal to noise ratio from 0.1 to 2. The signal to noise ratio is defined as $Var(Z\gamma)/Var(V)$.

* IV test 1 separates the observations into those with propensity scores in the interval $[0,P_{median}]$ and those with propensity scores in the interval $[P_{median},1]$ and testing the equality of the IV estimates based on those subsamples.

** IV test 2 separates the observations into quartiles of the propensity score. It tests for pairwise equality of the IV estimates calculated on each of those four subsamples and rejects the null if equality is rejected in any of the pairwise comparisons. It controls the size of the test using the method of Romano and Wolf (2005), as described in the text.

Figure 12: Power of a Wald test for the null hypothesis that the coefficient on $[P(Z)]^2$ is zero in the generalized Roy model.



Note: Standard errors are obtained analytically. The data are generated according to the normal generalized Roy model given in the text. The choice equation is $D = \mathbf{1}(Z \geq V)$ where $Z \sim N(0,\sigma_Z^2)$. These figures plot the power of a test of the significance of $[P(Z)]^2$ in a regression of Y on P(Z) and $[P(Z)]^2$. The size of the test is fixed at 0.05, and each line plots the power at a different sample size.

distributions, each with noncentrality parameter analogous to (15). Below we use simulation methods to explore the distribution of this test statistic under various alternatives. Note that although we do not have an exact expression for the distribution of the test statistic for this general test, we control the size of the test at $\alpha = 0.05$ using the Romano-Wolf stepdown procedure. Specifically, we simulate the distribution of the smallest (most significant) p-value under the null, as discussed above. This method applies the first stage of Romano and Wolf (2005) and is described in the Web Appendix, Section 7. The Monte Carlo algorithm we use is described in the Web Appendix, Section 8.

We limit the degree of the polynomial to 5. We add polynomials in increasing order from 2 to 5, keeping all lower order terms when the order of the polynomial is increased. This procedure produces a test of size at most 0.05 at $\tau = 0$. Results from these simulations are also given in Figures 10 and 11. The dashed lines in the figures plot the power of this test. Figure 10 shows that under the null ($\tau = 0$), we reject the null (up to simulation error) 5% of the time. Panel A of the figure shows the power function across different sample sizes, and panel B fixes the sample size and shows the power function across different values of the signal to noise ratio. The results indicate that sample sizes above 1,000 are necessary in order for the series test to have power and that with a relatively weak instrument, this test will fail to have power against most alternatives.[44,45,46]

Test 2: Bierens Conditional Moment Test

Finally, we examine the power of the conditional moment test of Bierens (1990) in detecting selection on the gain to treatment. This test uses the fact that under the null hypothesis of linearity of $g(p)$, moment condition (16) must hold for all $t \in \mathbb{R}$ and any bounded, one-to-one

[44] In our web appendix, we report analyses for the series test comparable to those reported in Figures A2 through A4. The results are qualitatively similar to what is obtained for the IV test. (See Figures A11-A13 in the Web Appendix.)

[45] Apparent inconsistencies of the test under the null in Figures 10 and 11 are due to sampling error.

[46] An alternative approach to testing the order of the polynomial is to fit a model with L polynomial terms and to conduct a joint test that the coefficients of the polynomials above order 1 are statistically significantly different from zero. We have not compared the power and size properties of these two approaches.

mapping $\Lambda(\cdot)$. Following Bierens (1990), we use $\Lambda(P(Z)) = \tan^{-1}((P(Z) - \bar{P}(Z))/s_{P(Z)})$ where $\bar{P}(Z)$ and $s_{P(Z)}$ are the sample mean and standard deviation of $P(Z)$, recognizing that this is just one of many possible Λ. To implement the test, we must also choose a region of values of t over which to calculate the sample analog of (16). We search over a grid of values of t between -10 and 10 spaced at intervals of 0.1. The method for constructing the test statistic presented in Bierens (1990) requires the choice of two arbitrary real numbers, which determine a cutoff value used in a decision rule for how to form the test statistic. These must be chosen independently of the data generating process. We choose parameter values that Bierens uses in his Monte Carlo simulations.[47] The procedure used to calculate the power of the test for different values of τ is described in section 8 of the Web Appendix. The power of the moment test is plotted as the solid line in Figures 10 and 11.[48]

5.4 Power under Alternative Data Generating Processes

Thus far, our Monte Carlo simulations have only considered the properties of tests under the specific alternative generated by different parameterizations of a normal generalized Roy Model. We choose this model because it allows for a simple parameterization of alternatives and is a useful baseline in empirical work. As a check to the robustness of our analysis, we have explored alternative data generating processes for the instruments as well as alternative specifications for the MTE. The results of these simulations are placed in the Web Appendix to this paper. They suggest that the power of the tests does not depend fundamentally on the normality assumptions made in the previous section. In particular, when the effective deviation from the null is measured by $1 - R^2$, as defined above, our results indicate that the power of the tests is relatively invariant to the process generating choices, given $1 - R^2$.

[47] See Bierens (1990), p. 1453. We choose, in Bierens' notation, $\gamma = 1$ and $\rho = 0.5$. These are not to be confused with other uses of γ and ρ in this paper. Alternative parameter values produce qualitatively the same results as we report in this paper. Results are available on request from the authors.

[48] In the Web Appendix, we plot additional figures which trace out the power of the moment test more thoroughly across the three dimensions which we find are important in this model – the deviation from the null (as measured by τ), the sample size, and the signal to noise ratio. Figures A14 through A16 contain these plots.

The parametric model always has the highest power, followed by IV in $P(Z)$ partitioned below or above the median. Generally, the power functions of the series and CM tests are somewhat below those for IV in $P(Z)$ above and below the median, with no clear ranking between them. Partitioning IV into quartiles generally produces the lowest power functions. We next consider a prototypical application of the null the hypothesis of selection on the gain to treatment.

6 An Analysis of a Prototypical Problem in Microeconometrics

This section draws on the empirical analysis of Carneiro, Heckman, and Vytlacil (2006) to test for the presence of a correlated random coefficient model in estimating the returns to college and to explore the power of the main tests considered in this paper in a prototypical economic problem. The data come from the National Longitudinal Survey of Youth, 1979 (NLSY79), and it contains 1,747 observations for white men from the random sample. The outcome of interest is the average of deflated (to 1983) hourly wages reported in 1989, 1990, 1991, and 1992.[49] The choice variable is $D = 1$ if the individual has attended some college or has completed any schooling beyond high school and $D = 0$ if the individual has a high school diploma or has completed 12 years of schooling but has never attended college. Our sample of 1,747 individuals contains 882 observations with $D = 0$ and 865 observations with $D = 1$. See the Web Appendix, Section 9 for further description of the sample.

Following Carneiro, Heckman, and Vytlacil (2006), in the first stage we run a probit for D using the following predictors of choice (Z): individual cognitive ability,[50] mother's education, number of siblings, an indicator for urban residence at age 14, average unemployment

[49] Following Carneiro, Heckman, and Vytlacil (2006) we delete all wage observations below one or above 100 dollars.

[50] We proxy cognitive ability using the Armed Forces Qualification Test (AFQT). The AFQT has been corrected for the effect of schooling at the time of the test using the method of Hansen, Heckman, and Mullen (2001).

in the state of residence, average log earnings in the Standard Metropolitan Statistical Area (SMSA) of residence, local wages and unemployment rates at age 17, an indicator for the presence of a college in the county of residence at age 14, and cohort dummies.[51]

For our estimates of the propensity score, $P(Z)$, we use the fitted values from this probit and in the second stage we regress the outcome variable on polynomials in the propensity score in addition to the following control variables (X): years of experience, cognitive ability, mother's education, number of siblings, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, local wages in 1991, and local unemployment rate in 1991.

Panels A and B of Table 3 give the results from the tests of the equality of IV estimators described above. We consider two formulations of the test: one based on IV estimates using $P(Z)$ as an instrument above and below the median, and one based on IV estimates that separate the data by quartile of $P(Z)$. Panel A shows that the test of H_0 based on IV estimators separated by the median of $P(Z)$ has a p-value of 0.0527. Panel B separates $P(Z)$ into quartiles. The smallest p-value of the pairwise tests of equality is 0.2407. This result is consistent with our investigation of the power of these two tests. Recall that the test based on separating by the median of $P(Z)$ is more powerful than the test based on separating by quartiles of $P(Z)$.

In addition, we have carried out tests of linearity described above. When we fit a quadratic polynomial to $E(Y|P(Z) = p)$, a Wald test for the signficance of the quadratic term in $P(Z)$ has a p-value of 0.046 – evidence against the null of linearity. However, when we add higher degrees of the polynomial and use a stepdown method to control for the size of the test, the critical value to control the size of the test at 0.05 is 0.024 against which we com-

[51]The complete list of instruments is: mother's education, number of siblings, urban residence at age 14, AFQT, mother's education squared, number of siblings squared, AFQT squared, local average wage, local average squared, local average unemployment rate, local average unemployment rate squared, year of birth dummies, presence of a local college at age 14, presence of a local college at age 14 interacted with AFQT, mother's education and number of siblings, local wage at age 17, local wage at 17 interacted with AFQT, mother's education and number of siblings, local unemployment rate at age 17, and local unemployment rate at age 17 interacted with AFQT, mother's education and number of siblings.

pare the p-value of 0.046. Allowing for the possibility of different degrees of the polynomial beyond the second, we lose the ability to reject linearity. In addition, the more traditional forward selection procedure for choosing the degree of the polynomial which starts with the quadratic polynomial and adds higher degree terms as long as they are statistically significant leads to choice of the quadratic polynomial as well. That is, if we test the significance the additional $[P(Z)]^3$ term, we obtain a p-value of 0.374. The conditional moment test produces a p-value of 0.5525 and fails to reject the null hypothesis. This is consistent with our evidence on its low power.

Under the assumptions of the normal selection model, we can test for selection on the gains to treatment by testing the coefficient on the selection term in the second stage regression. The test strongly rejects H_0 (see Table 3, Panel C). Our evidence for rejection of H_0 is consistent with the analysis of Carneiro, Heckman, and Vytlacil (2006), who show that for the same data a MTE based on a normal model is consistent with a nonparametrically estimated MTE.

Figure 13 plots the MTE estimated using different degrees of polynomial approximation, the MTE estimated assuming normality, and a nonparametric estimate of the MTE based on local polynomial regression. As shown by Carneiro, Heckman, and Vytlacil (2006), the normal model does a remarkably good job in describing the marginal treatment effect for this sample. Figure 13 also shows the weights that the IV estimator places on the MTE as well as the histogram of estimated propensity scores.

6.1 Monte Carlo Analysis for this Example

The Monte Carlo analyses presented in Section 5 were for general parameter values. To examine the power issue in the context of the college vs. high school example, it is useful to start with the benchmark model just presented and conduct a Monte Carlo study for this prototypical model. It confirms in this setting the general Monte Carlo analysis discussed in Section 5. All of the tests have relatively low power. This means that tests of the null of H_0

Table 3: College participation vs. stopping at high school: IV tests for selection on the gain to treatment. (Standard errors are obtained from the bootstrap.)

A. IV estimates above and below the median[a]

		Whole sample	Below	Above
Estimate:		0.1266	0.9112	-0.2445
Standard error:		(0.1500)	(0.5148)	(0.3553)
p-value of test:	0.0527			

B. IV estimates by quartiles of the propensity score[b]

	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	1.8842	0.5583	0.3076	-0.9497
Standard error:	(4.2662)	(3.4309)	(0.5950)	(5.2114)
Smallest p-value from pairwise tests:	0.2407			

C. Test of heterogeneity in normal selection model[c]

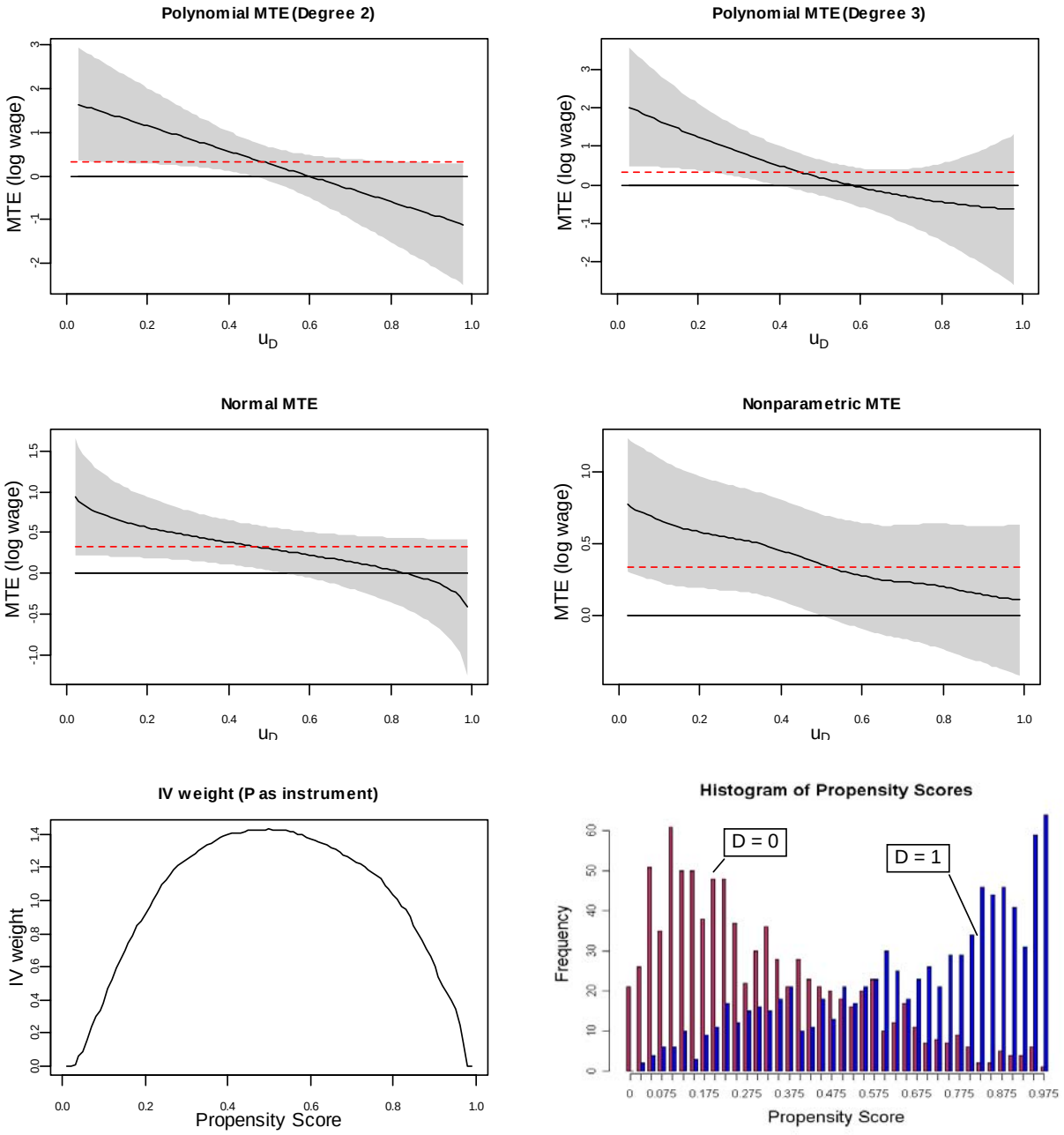
Probability value of test:	0.011

[a] These IV estimates do not include interactions between the treatment and X. The test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[b] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[c] The p-value in this panel is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

Figure 13: College participation vs. stopping at high school: estimates of marginal treatment effect for different models, IV weights and support of the estimated propensity score. (Standard errors are obtained from the bootstrap.)



[a] In the MTE graphs, the dashed line indicates the IV estimate. In the histogram, the dark bars correspond to the D=1 group and the light bars to the D=0 group. The sample size is 1,747. The confidence intervals are found using 100 bootstrap samples. The covariates in the outcome equations are: years of experience, corrected AFQT, mother's education, number of siblings, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, local wages in 1991, and local unemployment rate in 1991. The instruments are: corrected AFQT, mother's education, number of siblings, an indicator for urban residence at age 14, average unemployment in the state of residence, average log earnings in the SMSA of residence, local wages and unemployment rates at age 17, an indicator for the presence of a college in the county of residence at age 14, and cohort dummies. The dependent variable in the probit is 1 if the individual reports having attended college or completed any schooling past 12 years, and 0 if the individual has a high school diploma or has completed 12 years of school but not attended college (GEDs are excluded).

tend to be conservative in the sense that they are likely not to reject when the null is false.

In these simulations, we take the regressors (X) and instruments (Z) from the NLSY79 data, using the 1,747 observations of the independent variables in the Carneiro et al. model. Assuming the normal generalized Roy model, we generate outcomes Y and choices D based on alternative parameterizations of the unobservables of the model. That is, we take X and Z from the data and we generate Y and D according to

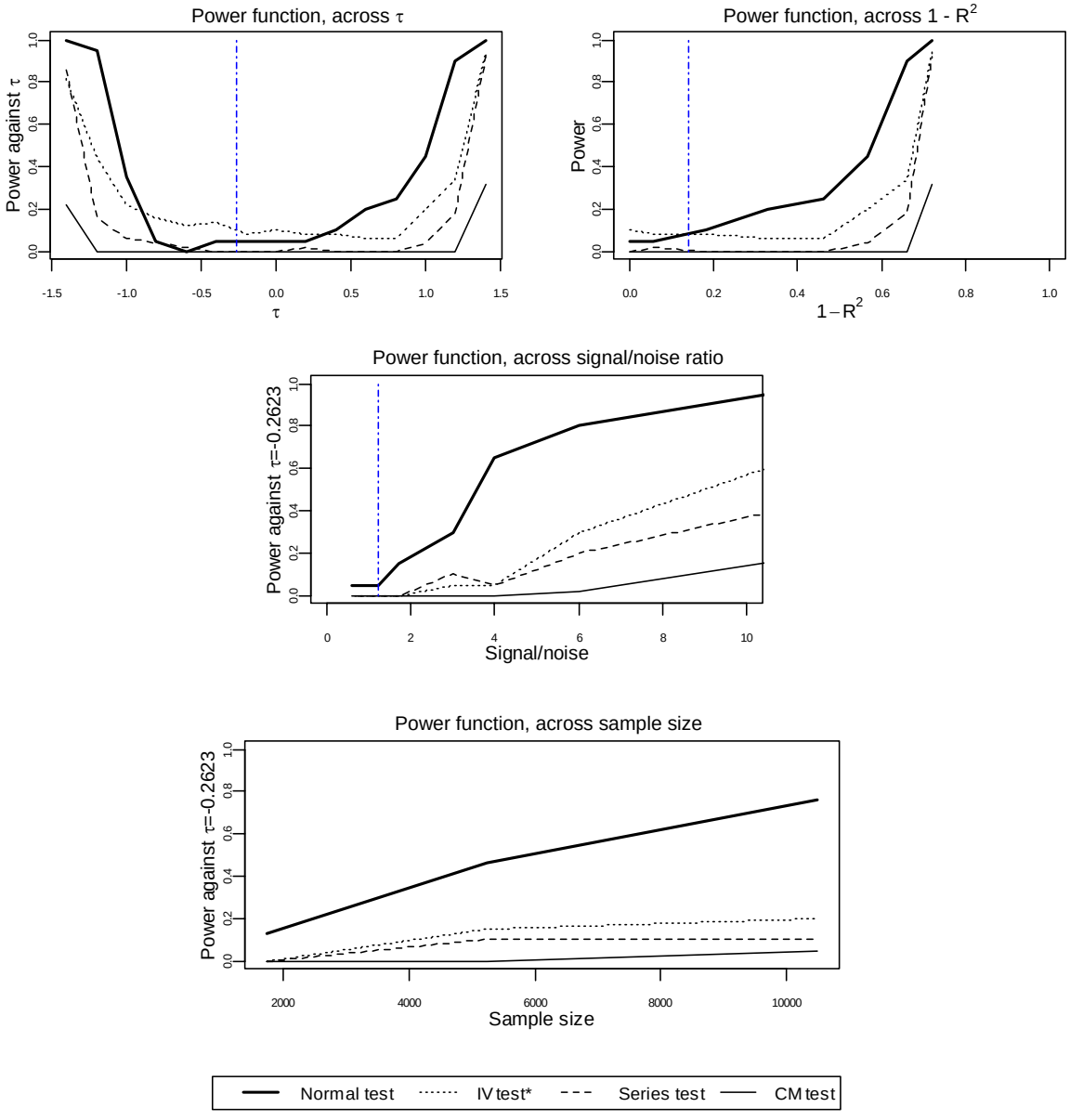
$$\begin{aligned} Y_1 &= \alpha_1 + X\beta_0 + U_1 \\ Y_0 &= \alpha_0 + X\beta_1 + U_0 \\ D &= \mathbf{1}(\alpha_D + Z\gamma \geq V) \\ Y &= DY_1 + (1-D)Y_0 \end{aligned}$$

where α_1, α_0, β_0, β_1, α_D and γ are calculated from the NLSY79 data.

These simulations use the same regressors and instruments as in the Carneiro et al. study. In order to generate the data under various alternative hypotheses, we specify alternative parameter values for the distributions of the unobservables U_1, U_0 and V. We assume that the unobservables are jointly normally distributed as in Table 2, but we vary the variances and covariances of these variables. Therefore, the only way in which these simulated datasets differ from the actual dataset is the values of Y and D and in the assumption of normal errors. Carrying out simulations in this fashion lets us examine how the power of the tests depends, in this specific example, on the departure from the null hypothesis (in the sense of τ) and on the sample size. We use the values of σ_1 and σ_0 calculated from the NLSY79 data. To gauge the explanatory power of our instruments in the NLSY79 data, we use the pseudo-R^2 of the first stage probit. The pseudo-R^2 in this application is 0.2986. For comparison to the power calculations, we note that this corresponds to a signal to noise ratio (as described in section 3.1) of about 1.2.

Figure 14 shows the power of three of our tests for different values of τ and different

Figure 14: Power of the tests from Monte Carlo simulations on NLSY79 data. (Standard errors are obtained from the bootstrap.)



Parameterization	
N= 1747	σ_{10} = 0
σ_1= 0.4857	σ_V= 1
σ_0= 0.3927	pseudo R^2 = 0.2986

Signal/noise ratio = 1.2

τ = -0.2623

Note: In these simulations we use the real data from the NLSY79 on the effect of a college education on wages for regressors (X) and instruments (Z). We parameterize the unobservables to match the estimated values from the data and we generate outcomes Y and choices D for different combinations of parameters. We have 1,747 observations and we fix the std. dev. of U_1 at 0.4857, the std. dev. of U_0 at 0.3927. In each of the top three panels, the dot-dashed vertical line indicates the estimates of the parameters calculated in the NLSY79 data. The series test estimates polynomials of degrees 2 through 5 and in each model tests for the joint significance of the nonlinear terms. The size of the test is controlled using the bootstrap procedure described in the text.

* The IV test used in these simulations tests for the equality of the IV estimates constructed using the propensity score as the instrument separately for observations below the median of P(Z) and above the median of P(Z).

sample sizes. Note that with our actual sample size of 1,747 and our estimated $\tau = -0.2623$ we are in a range where all of the tests have very low power, although for a fixed τ, the IV test generally has more power than the series test, which in turn has more power than the conditional moment (CM) test. In addition, the IV test shows greater increase in power with the signal/noise ratio than do the other tests. These results are in accordance with the Monte Carlo simulations conducted in Section 5.[52]

For comparison with the power calculations made in the previous sections, we also present these results in terms of the R^2 of a regression of the fitted $E(Y|P(Z) = p)$ on a linear term in $P(Z)$. This allows us to compare the value of the R^2 we calculate in the data with values of the R^2 for other parameterizations as shown in Figure 14. We find that for samples consistent with those found in practice the estimated $1 - R^2 = 0.1375$. It is represented by the vertical dot-dashed lines in Figure 14.

Overall, the simulations based on the data from this example indicate that for this sample it would take a large deviation from the null in order for us to reject H_0. Our evidence on power gives us greater confidence in rejecting the null hypothesis for the Carneiro, Heckman, and Vytlacil (2006) data. We reject H_0 in their sample, even though the power of the tests is low.

7 Summary and Conclusion

This paper develops and applies tests for the presence of a correlated random coefficient model. All of the tests we consider can be interpreted as conditional moment tests. We have developed the sampling distribution of the IV estimator using the marginal treatment effect and its extensions to higher moments of the distribution of the heterogeneity on which agents select. We investigate the power of general tests based on the correlated random coefficient model, where the parameter of interest is determined by an instrument.

We develop and evaluate instrumental variable tests for the null hypothesis of the absence

[52] An exception is that the CM test falls appreciably in power in this example.

of a correlated random coefficient model. We examine the power of these tests and the power of additional tests of the null hypothesis based on linearity in p for $E(Y \mid P(Z) = p, X = x)$.

In sample sizes common in empirical economics, the power of the proposed tests is low. The degradation of power of less parametric tests from the parametric tests is substantial. Such tests are conservative in the sense that they often do not reject the null of no correlated random coefficient model (H_0) when, in fact, a correlated random coefficient model describes the data. Among the tests we consider, an IV test based on partitioning the propensity score above and below the median has the best performance in terms of power in samples commonly encountered in practice. An analysis of the optimal choice of instruments to maximize the power functions for the tests is left for future work.

We test if the additional complexity of a correlated random coefficient model is required to describe data on the returns to college. We find support for the correlated random coefficient model using our testing procedure. This evidence is strengthened by our study of the power of these tests. We reject H_0 in a situation where the power of the tests we use is low.

This paper analyzes the case of a binary treatment. Heckman, Urzua, and Vytlacil (2006) and Heckman and Vytlacil (2007b) analyze the cases of a multiple treatment model generated by an ordered choice model with stochastic thresholds and a multiple treatment model generated by an unordered choice model. In all of these cases, IV produces an instrument-dependent parameter so the IV test for selection on unobserved gains based on comparing the estimands of two different IVs developed in this paper carries over in general to these settings. A test of linearity of the conditional expectation of Y given P in (a vector of) P is developed for the outcome model for multiple treatments generated by the ordered choice model in Heckman, Urzua, and Vytlacil (2006). It also applies to the unordered multiple choice model that identifies the treatment effect of a gain option compared to the next best option which Heckman, Urzua and Vytlacil show is a direct extension of the binary model.

A The Variance of Linear IV in the Correlated Random Coefficient Model

The IV estimator, using instrument $J(Z)$, is

$$\widehat{\beta}_{IV,J} = \frac{\sum Y_i \tilde{J}_i}{\sum D_i \tilde{J}_i}$$

and hence

$$\sqrt{I}\widehat{\beta}_{IV,J} = \frac{\frac{1}{\sqrt{I}}\sum (J_i - \bar{J})(\alpha_i + \beta_i D_i)}{\frac{1}{I}\sum D_i \tilde{J}_i}.$$

Invoking standard central limit theorems,

$$\sqrt{I}\left(\widehat{\beta}_{IV,J} - \beta_{IV,J}\right) \xrightarrow{d} N(0, \Omega_J).$$

Defining $J^* = J - E(J)$, where Ω_J is given by

$$\begin{aligned}
\Omega_J &= E\left\{\left[\frac{YJ^*}{\omega_J} - E\left(\frac{YJ^*}{\omega_J}\right)\right]^2\right\} \\
&= E\left[\frac{Y^2(J^*)^2}{\omega_J^2}\right] - \left(E\left[\frac{YJ^*}{\omega_J}\right]\right)^2 \\
&= E\left[\frac{Y^2(J^*)^2}{\omega_J^2}\right] - \left(\int_0^1 E(\beta \mid U_D = u_D) h_J(u_D) du_D\right)^2 \\
&= \frac{1}{\omega_J^2} E\left[(\alpha + \beta D)^2 (J^*)^2\right] - \left(\int_0^1 E(\beta \mid U_D = u_D) h_J(u_D) du_D\right)^2 \\
&= \frac{1}{\omega_J^2} E\left[\alpha^2 (J^*)^2\right] + \frac{2}{\omega_J^2} E\left[\alpha\beta D (J^*)^2\right] + \frac{1}{\omega_J^2} E\left[\beta^2 D (J^*)^2\right] \\
&\quad - \left(\int_0^1 E(\beta \mid U_D = u_D) h_J(u_D) du_D\right)^2.
\end{aligned}$$

Using the law of iterated expectations as well as the assumption that α is independent of Z, this expression can be written as

$$\Omega_J = E\left[\alpha^2\right]\frac{\text{Var}(J)}{\omega_J^2} + \frac{1}{\omega_J^2}2E\left[\alpha\beta(J^*)^2 \mid D=1\right]\Pr(D=1) + \frac{1}{\omega_J^2}E\left[\beta^2(J^*)^2 \mid D=1\right]\Pr(D=1) - \left(\int_0^1 E(\beta \mid U_D = u_D)h_J(u_D)du_D\right)^2$$

where

$$h_J(u_D) = \frac{E[J^* \mid P(Z) \geq u_D]\Pr(P(Z) \geq u_D)}{\omega_J}.$$

Under the conditions of Fubini's Theorem, we can exchange the order of integration and write

$$\begin{aligned}\Omega_J &= E\left[\alpha^2\right]\frac{\text{Var}(J)}{\omega_J^2} \\ &\quad + \int_0^1 \left[2E\left(\alpha\beta \mid U_D = u_D\right) + E\left(\beta^2 \mid U_D = u_D\right)\right]\frac{E((J^*)^2 \mid P(Z) \geq u_D)\Pr(P(Z) \geq u_D)}{\omega_J^2}du_D \\ &\quad - \left(\int_0^1 E(\beta \mid U_D = u_D)h_J(u_D)du_D\right)^2 \\ &= E\left[\alpha^2\right]\frac{\text{Var}(J)}{\omega_J^2} + \int_0^1 \left[2E\left(\alpha\beta \mid U_D = u_D\right) + E\left(\beta^2 \mid U_D = u_D\right)\right]h_{\Omega_J}(u_D)du_D \\ &\quad - \left(\int_0^1 E(\beta \mid U_D = u_D)h_J(u_D)du_D\right)^2\end{aligned}$$

where

$$\begin{aligned}h_{\Omega_J}(u_D) &= \frac{1}{\omega_J^2}\int_{-\infty}^{\infty}(j - E(J))^2\int_{u_D}^1 f_{P,J}(P(z), j)dP(z)dj \\ &= \frac{E[(J - E(J))^2 \mid P(Z) \geq u_D)]\Pr(P(Z) \geq u_D)}{\omega_J^2}\end{aligned}$$

which is the expression in the text.

B Proof of Invariance of the IV Estimand to the Choice of a Linear Instrument under Normality with a Linear Index Choice Equation

Suppose that the choice equation has a linear index structure, so that

$$D = \mathbf{1}(Z\gamma \geq V)$$

where $Z \sim N(\bar{Z}, \Sigma_Z)$, an L-dimensional multivariate normal random variable, γ an $L \times 1$ vector and $V \sim N(0, \sigma_V^2)$. Consider the instrument $J(Z)$, which is a linear function of Z, say $Z'\eta$. In this case, the IV estimand (written as a weighted average over the support of V) is

$$\beta_{IV,J} = \int_{-\infty}^{\infty} MTE(v) h_J(v) \phi\left(\frac{v}{\sigma_V}\right) dv$$

where $\phi(\cdot)$ is a standard normal pdf and the IV weight is

$$h_J(v) = \frac{E[J(Z) - E(J(Z))|Z\gamma > v] \Pr(Z\gamma > v)}{\mathrm{Cov}(J(Z), D)}.$$

Under the assumption of multivariate normality for the instruments,

$$h_J(v) = \frac{\frac{\mathrm{Cov}(J(Z), Z\gamma)}{\sqrt{\mathrm{Var}(Z\gamma)}} \phi\left(\frac{v - \bar{Z}\gamma}{\sqrt{\mathrm{Var}(Z\gamma)}}\right)}{\frac{\mathrm{Cov}(J(Z), Z\gamma - V)}{\sqrt{\mathrm{Var}(Z\gamma - V)}} \phi\left(\frac{-\bar{Z}\gamma}{\sqrt{\mathrm{Var}(Z\gamma - V)}}\right)}.$$

Under assumption (A-1) in Section 2, $\text{Cov}(J(Z), Z\gamma) = \text{Cov}(J(Z), Z\gamma - V)$, and we obtain

$$h_J(v) = \frac{\frac{1}{\sqrt{\text{Var}(Z\gamma)}}\phi\left(\frac{v-\bar{Z}\gamma}{\sqrt{\text{Var}(Z\gamma)}}\right)}{\frac{1}{\sqrt{\text{Var}(Z\gamma-V)}}\phi\left(\frac{-\bar{Z}\gamma}{\sqrt{\text{Var}(Z\gamma-V)}}\right)}.$$

That is, the IV weights, and hence the IV estimand, are the *same* for all $J(Z) = Z'\eta$ for any η.

Acknowledgments

This research was supported by NIH R01-HD043411, NSF SES-024158, the American Bar Foundation and the Geary Institute, University College Dublin, Ireland. The views expressed in this paper are those of the authors and not necessarily those of the funders listed here. We have received helpful comments from Pedro Carneiro, Jeremy Fox, Joel Horowitz, Benjamin Moll, Azeem Shaikh, Christopher Taber, Edward Vytlacil, the editor, Steve Durlauf, and an anonymous referee and participants in workshops at the University of Wisconsin and Northwestern University. In the final round of revisions, we received additional very helpful suggestions from Stéphane Bonhomme, Xiaohong Chen, Azeem Shaikh and Edward Vytlacil. Supplementary material for this paper is available at the Website `http://jenni.uchicago.edu/testing_random/`.

References

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 5145–5303. Amsterdam: Elsevier.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics 20*(1), 105–134.

Bierens, H. J. (1990, November). A consistent conditional moment test of functional form. *Econometrica 58*(6), 1443–1458.

Bierens, H. J. and W. Ploberger (1997, September). Asymptotic theory of integrated conditional moment tests. *Econometrica 65*(5), 1129–1151.

Björklund, A. and R. Moffitt (1987, February). The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics 69*(1), 42–49.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2006). Estimating marginal and average returns to education. Under revision.

Griliches, Z. (1977, January). Estimating the returns to schooling: Some econometric problems. *Econometrica 45*(1), 1–22.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2001). Ordered discrete choice models with stochastic shocks. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. (2001, August). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy 109*(4), 673–748.

Heckman, J. J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions. In J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, Volume 10, pp. 156–245. New York: Cambridge University Press.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. J. Vytlacil (1998, Fall). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources 33*(4), 974–987.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*, 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001). Local instrumental variables. In C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pp. 1–46. New York: Cambridge University Press.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4779–4874. Amsterdam: Elsevier.

Heckman, J. J. and E. J. Vytlacil (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman

and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam: Elsevier.

Horowitz, J. L. and V. G. Spokoiny (2001, May). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica 69*(3), 599–631.

Ichimura, H. and P. E. Todd (2007). Implementing nonparametric and semiparametric estimators. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B. Amsterdam: Elsevier.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Li, R. and L. Nie (2007, November). Efficient statistical inference procedures for partially nonlinear models and their applications. *Biometrics 64*(3), 904–911.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Academic Press.

McFadden, D. (1981). Econometric models of probabilistic choice. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

Newey, W. K. (1985, September). Maximum likelihood specification testing and conditional moment tests. *Econometrica 53*(5), 1047–1070.

Newey, W. K. (1997, July). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Quandt, R. E. (1958, December). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association 53*(284), 873–880.

Romano, J. P. and A. M. Shaikh (2006, August). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics 34*(4), 1850–1873.

Romano, J. P. and M. Wolf (2005, March). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review 34*, 273–286.

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

Yitzhaki, S. (1989). On using linear regression in welfare economics. Working Paper 217, Department of Economics, Hebrew University.