Testing for Essential Heterogeneity*

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

Daniel Schmierer
University of Chicago

Sergio Urzua
University of Chicago

June 29, 2007

PRELIMINARY DRAFT

DO NOT CITE WITHOUT PERMISSION

*This project was supported by ..... We have received additional helpful comments from ..... Supplementary material for this paper is available at the website .....

Abstract

This paper examines the properties of instrumental variables (IV) applied to models with essential heterogeneity, that is, models where responses to interventions are heterogeneous and agents adopt treatments (participate in programs) with at least partial knowledge of their idiosyncratic response. We present several empirical examples demonstrating the importance of unobserved heterogeneity in economic applications. For each empirical example we implement a simple test for the presence of essential heterogeneity, we compare the IV estimates with other treatment parameters, and we analyze whether or not IV can be used to provide meaningful answers to well-posed economic questions.

JEL: C31

James Heckman
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637

jjh@uchicago.edu
773-702-0634

Daniel Schmierer
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637

dschmier@uchicago.edu
773-702-1787

Sergio Urzua
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637

surzua@uchicago.edu
773-702-1787

1 Introduction

Recent research has highlighted the difficulty in using the traditional instrumental variables methods to estimate treatment parameters in choice models. In particular, Heckman, Urzua, and Vytlacil (2006) show that when there is selection on the gain to treatment, or *essential heterogeneity*, then IV does not in general estimate any of the standard treatment parameters. In this paper, we seek to test for the presence of essential heterogeneity in a variety of choice settings. If we fail to find evidence for such heterogeneity in a given dataset, that suggests it may not be worthwhile for researchers to deal with the extra complications arising from taking this heterogeneity into account. However, when we find evidence for selection on the unobservables, it means that researchers should use caution in interpreting the results from IV methods. Indeed, we do find evidence from a variety of datasets for the presence of heterogeneous gains to treatment. Also, we discuss the properties of various tests for essential heterogeneity and present some results on the power of these tests using simulated data.

2 Prototypical Model of Potential Outcomes

We consider a setting where there are two possible outcomes for an individual, Y_1 or Y_0. For example, the outcome may be hourly wages and Y_1 may correspond to the case where the individual is a college graduate and Y_0 to the case where the individual has only a high school diploma. Then we say the treatment effect for this individual is given by $Y_1 - Y_0$. If we write

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0 \end{aligned} \tag{1}$$

then we can write the treatment effect as

$$Y_1 - Y_0 = \mu_1(X) - \mu_0(X) + U_1 - U_0$$

where X is a vector of observable characteristics for the individual. It may happen that controlling for the X, $Y_1 - Y_0$ is the same for all individuals. This is the case of homogenous treatment effects given X. More likely, individuals vary in their response to the treatment even after controlling for X. Ideally, we would

estimate the entire distribution of $Y_1 - Y_0$ given X.[1] Economists usually focus on various means of the distribution, however.

3 OLS and IV under Essential Heterogeneity

Using the Quandt (1958) switching regression framework, we could think about trying to identify the treatment effect using OLS or IV methods. Let $D = 1$ if the individual chooses Y_1 and $D = 0$ if the individual chooses Y_0. Then the outcome we observe for an individual is

$$\begin{aligned} Y &= DY_1 + (1-D)Y_0 \\ &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X) + U_1 - U_0)D + U_0 \end{aligned} \tag{2}$$

Rewriting this in regression notation, where the subscript i corresponds to the observation for individual i,

$$Y_i = \alpha + \beta D_i + \varepsilon_i \tag{3}$$

where $\alpha = \mu_0(X)$, $\beta = \mu_1(X) - \mu_0(X) + U_1 - U_0$ and $\varepsilon = U_0$. The case where β (given X) is the same for every individual is the familiar one and we develop it first. A least squares regression of Y on D (equivalently a mean difference in outcomes between individuals with $D = 1$ and individuals with $D = 0$), is possibly subject to a **selection bias**. Individuals that choose the treatment may be atypical in terms of their Y_0 $(= \alpha + \varepsilon)$. Thus if individuals that would have done well in terms of unobservable ε $(= U_0)$ even in the absence of the treatment are the ones that adopt the treatment, β estimated from OLS (or its nonparametric version—matching) is upward biased because $Cov(D, \varepsilon) > 0$.

Two main approaches have been adopted to solve this problem: (a) selection models (Gronau (1974); Heckman (1974, 1976a,b, 1979); Heckman and Robb (1985, 1986); Powell (1994)) and (b) instrumental variable models (Heckman and Robb (1985, 1986); Imbens and Angrist (1994); Angrist and Imbens (1995); Manski and Pepper (2000); Heckman and Vytlacil (1999, 2000, 2005)).

[1] See Carneiro, Hansen, and Heckman (2001, 2003), and the survey in Heckman, Lochner, and Todd (2006).

For the case with homogeneous responses, if there is an instrument Z with the properties that

$$Cov(Z, D) \neq 0 \tag{4}$$

and

$$Cov(Z, \varepsilon) = 0 \tag{5}$$

then standard IV identifies β, at least in large samples:

$$\text{plim } \hat{\beta}_{\text{IV}} = \frac{Cov(Z, Y)}{Cov(Z, D)} = \beta.^{2}$$

If other instruments exist, each identifies β. Z produces a controlled variation in D relative to ε. Randomization of assignment with full compliance with experimental protocols is an example of an instrument. From the instrumental variable estimator, we can identify the effect of choosing the treatment for any individual since all individuals respond to the treatment in the same way, controlling for their X.

However, it may be that even after conditioning on X there is still variation in β across individuals. In this case, there may be another problem in addition to the selection bias described above. We say that there is **essential heterogeneity** in response to the treatment if, after conditioning on the observables X, we have $Cov(\beta, D) \neq 0$. This will be the case if there is sorting on the gain to treatment. That is, if individuals make their decisions with at least partial knowledge of their idiosyncratic gain from the treatment, then the model contains essential heterogeneity. Heckman, Urzua, and Vytlacil (2006) show that in models with essential heterogeneity, standard instrument variables does not identify any meaningful treatment parameters. Therefore, a test for the presence of essential heterogeneity is necessary in order to determine whether IV can be used to recover a parameter that is meaningful to economists.

[2] The proof is straightforward. Under general conditions White (1984, see, e.g.),

$$\text{plim } \hat{\beta}_{\text{IV}} = \beta + \frac{Cov(Z, \varepsilon)}{Cov(Z, D)}$$

and the second term on the right hand side vanishes.

4 The Choice Model and the IV Approach

Let Z be a vector of instruments for D. We assume that choices are generated by a latent variable D^*, where

$$D^* = \mu_D(Z) - V \text{ and } D = \mathbf{1}(D^* \geq 0)$$

Then the propensity score, or choice probability is

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

where F_V is the distribution of V, which is assumed to be continuous. $P(z)$ is simply the probability that an individual chooses Y_1 when the instruments are fixed at value z. Note that because F_V is a strictly increasing function we can transform the latent variable which generates the choice as follows

$$\begin{aligned} D &= \mathbf{1}(D^* \geq 0) = \mathbf{1}(\mu_D(z) \geq V) \\ &= \mathbf{1}(F_V(\mu_D(z)) \geq F_V(V)) \\ &= \mathbf{1}(P(z) \geq U_D) \end{aligned}$$

where $U_D = F_V(V)$ is a Uniform$[0, 1]$ random variable by construction.

Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006) establish that the most fundamental treatment parameter is the marginal treatment effect (MTE), because all of the other treatment effects can be determined as weighted averages of it. The MTE is defined, for a given value of $X = x$, as

$$MTE(x, v) = E(Y_1 - Y_0|X = x, V = v)$$

That is, it is simply the mean treatment effect when the observables X are fixed at a value x and the unobservable V is fixed at a value v (the treatment effect still varies between individuals because of the unobservables U_1 and U_0). Note that we can also express the MTE in terms of the uniform random variable U_D as

$$MTE(x, u_D) = E(Y_1 - Y_0|X = x, U_D = u_D)$$

The other means of the distribution of the treatment effect that are commonly used in the literature and that we will consider are the average treatment effect (ATE) and the average effect of the treatment on the treated (TT), which are defined as

$$ATE = E(Y_1 - Y_0) \tag{6}$$

$$TT = E(Y_1 - Y_0|D = 1) \tag{7}$$

Both of these treatment effects can also be defined conditional on a given value of the observables X. We can now determine the implications of essential heterogeneity on the fundamental parameter, the marginal treatment effect. Again, assume that the instruments Z satisfy the following standard assumptions:

(IV-1) $Z \perp\!\!\!\perp \left(Y_0, Y_1, \{D(z)\}_{z\in\mathcal{Z}}\right)$ where $\mathcal{Z}$ is the set of possible values of Z. (**Independence**)

(IV-2) $\Pr(D = 1 \mid Z)$ depends on Z (**Rank**).

Note that these are the same assumptions about the instruments that we used to show that IV identifies the treatment effect in the homogeneous response case. Under these assumptions, we can write

$$\begin{aligned}
E(Y|X = x, Z = z) &= E(Y|X = x, P(Z) = p) \\
E(Y|X = x, P(Z) = p) &= E(DY_1 + (1 - D)Y_0|P(Z) = p, X = x) \\
&= E(Y_0|X = x) + E(D(Y_1 - Y_0)|X = x, P(Z) = p) \\
&= E(Y_0|X = x) + E(Y_1 - Y_0|X = x, D = 1)p \\
&= E(Y_0|X = x) + \int_0^p E(Y_1 - Y_0|X = x, U_D = u_D)du_D
\end{aligned}$$

This integrand is precisely the marginal treatment effect, $MTE(x, u_D) = E(Y_1 - Y_0|X = x, U_D = u_D)$. In the absence of essential heterogeneity, we know that $\beta \perp\!\!\!\perp D$, that is $(Y_1 - Y_0) \perp\!\!\!\perp D$ and therefore

$$\begin{aligned}
E(Y|P(Z) = p, X = x) &= E(Y_0|X = x) + \int_0^p E(Y_1 - Y_0|X = x, U_D = u_D)du_D \\
&= E(Y_0|X = x) + E(Y_1 - Y_0|X = x)p
\end{aligned}$$

This says that $E(Y|P(Z) = p, X = x)$ is linear in p in the absence of essential heterogeneity. However, under essential heterogeneity, $E(Y|P(Z) = p, X = x)$ will be a general nonlinear function of p. This

implication will form the basis for our test.

Note also that, as shown in Heckman, Urzua, and Vytlacil (2006), the MTE can be found using their so-called local instrumental variables (LIV), defined as

$$LIV = \frac{\partial E(Y|X = x, P(Z) = p)}{\partial p} = E(Y_1 - Y_0|X = x, U_D = p)$$

Therefore, in the absence of essential heterogeneity, the LIV estimator will give the constant $E(Y_1 - Y_0|X = x)$. So because all of the other treatment parameters we consider (ATE,TT and TUT) are all weighted averages of the MTE, they will all be identical in that case.

5 Testing for Linearity

As we showed above, $E(Y|P(Z) = p, X = x)$ will in general be some nonlinear function of the propensity score p. We keep the conditioning on X implicit in what follows. We can write

$$E(Y|P(Z) = p) = h(p) \tag{8}$$

for some general nonlinear function $h(\cdot)$. What we would like to do is allow for any possible functional form for $h(\cdot)$ and test the parametric null hypothesis

$$H_0 : \ h(p) = a + bp \text{ for some } a, b \in \mathbb{R}$$

against the composite alternative

$$H_1 : \text{ not } H_0$$

In implementing this test, we need to pick a specific alternative against which to test the null hypothesis of linearity. We find that the most powerful alternatives are simple polynomials in p.[3] That is, our alternative specification (conditional on X) is

$$h(p) = \sum_{j=0}^{d} \phi_j p^j$$

[3] In simulations with essential heterogeneity, we have tried testing against more general forms of nonlinearity, such as local linear regression and splines, but find that such tests have very low power.

Then the test for linearity is simply a test of

$$H_0 : \phi_j = 0 \text{ for } j = 2, \ldots, d$$

$$H_1 : \text{not } H_0$$

5.1 Implementing the Test of Linearity

The issues we face in implementing our test are how to condition on the X variables, and how to estimate the propensity scores, $P(Z)$. Ideally we would fully stratify the data by X values and estimate $E(Y|P(Z) = p)$ for each stratum separately. However, none of our data is rich enough to do this, given the number of controls X, which we would like to include. Therefore, we choose to incorporate the controls linearly and estimate the alternative specification as

$$Y_i = X_i\beta_0 + X_i(\beta_1 - \beta_0)P(Z_i) + \sum_{j=1}^{d} \phi_j P(Z_i)^j + \varepsilon_i \tag{9}$$

To see the rationale for the interaction between X and P note that we can write

$$\begin{aligned} E(Y|P(Z) = p, X = x) &= E(Y_0|P = p, X = x) + E(D(Y_1 - Y_0)|P = p, X = x) \\ &= \mu_0(x) + (\mu_1(x) - \mu_0(x))p + E(U_0|P = p, X = x) + E(U_1 - U_0|P = p, X = x)p \\ &= x\beta_0 + x(\beta_1 - \beta_0)p + \kappa(p) \end{aligned}$$

where the last equality comes from the IV independence assumption and the assumption that the μ's are linear in X. Our test then becomes a test of whether $\kappa(p)$ is linear in p.

We choose to estimate the propensity scores using a probit, because we find that the results are relatively robust to the estimation method.

In order to proceed systematically with the test for linearity, we propose a simple sequential method. That is, we suggest starting with just a linear term in P, then adding a quadratic term, then a cubic term, etc. If, after adding a quadratic term, one is already able to reject linearity, then one can stop and take that as evidence of essential heterogeneity. If not, then one can add a cubic term in P and test for both the significance of that term individually, as well as the joint significance of the quadratic and cubic terms. If either or both are significant, this provides some evidence of unobserved heterogeneity. Then,

continuing in this manner, we could potentially add ever higher degrees in P, however, the variance in the estimates quickly becomes very large. In practice, if using ordinary polynomials, the regression will fail due to collinearity of the higher order terms in P if the degree of the polynomial is high. In order to avoid this collinearity, we could use orthogonal polynomials, but we have found that it is unnecessary to use polynomials of degree high enough that collinearity becomes an issue. The variance in the estimates becomes large before the collinearity occurs.

Inherent in our test of linearity is the standard bias-variance tradeoff. That is, as we increase the number of polynomial terms we expect to get a more accurate approximation to the true MTE, however the standard errors will eventually begin increasing. In order to help choose, then, the optimal number of polynomial terms to include, we suggest constructing a nonparametric (or semiparametric) estimate of the MTE to use as a reference. While it is unlikely that a test of linearity on this nonparametric estimate directly would be able to reject linearity, it is useful to see how close the polynomial approximations are to this more flexible estimate. In our estimation below, we use a local polynomial approximation as our flexible functional form to which we compare our polynomial estimates.

The statistical tests we use to test the coefficients in our regressions are t-tests and Wald tests. To get standard errors of the coefficients, we bootstrap 50 times and reestimate the propensity scores P, as well as the outcome equation $E(Y|P(Z) = p)$ in each bootstrap sample. We then use t-tests to test for the significance of individual coefficients and Wald tests to test for the joint significance of all of the nonlinear terms in P.

5.2 Testing for Heterogeneity Using LATE

Another way to test for the linearity of $E(Y|P(Z) = p)$ in p is to use the local average treatment effect (LATE) parameter of Imbens and Angrist (1994). This parameter is defined for an instrument Z as

$$LATE(z', z) = \frac{E(Y|Z = z') - E(Y|Z = z)}{Pr(D = 1|Z = z') - Pr(D = 1|Z = z)}$$

This measures the average treatment effect of those who are induced to switch into treatment by a shift in the Z variables from z to z' (under the uniformity assumption). Vytlacil (2002) shows that this can be written as

$$LATE(u'_D, u_D) = \frac{E(Y|P(Z) = u'_D) - E(Y|P(Z) = u_D)}{u'_D - u_D} \tag{10}$$

This means that $LATE$ is measuring the secant between two points on the $E(Y|P(Z)=p)$ curve. Clearly, if $E(Y|P(Z)=p)$ were linear then $LATE(v,w)$ would be the same for any points $v,w \in Supp(P(Z))$. Therefore, another testable implication of the absence of essential heterogeneity is the equality of $LATE$ at all evaluation points in the support of P.

In practice, however, estimating the $LATE$ over different intervals is difficult because it involves forming conditional expectations where we are conditioning on the value of a continuous variable (the propensity score). Doing so would require nonparametric methods that would likely have poor properties in the relatively small samples we are dealing with. In particular, they have poor power for the tests we are interested in because of large standard errors. We could think about modeling the expectations parametrically, and in particular using the polynomials that we used for the MTE. However, if we fit a global polynomial to model $E(Y|P)$ then when we restrict ourselves to comparing the estimates over two subintervals we get a biased estimate of the difference in $LATE$s.

Therefore, we take a different approach to test for heterogeneity using $LATE$s. We know, as shown in Heckman, Urzua, and Vytlacil (2006), that the linear IV estimator is just a weighted average of the MTE with weights integrating to 1. In particular, if we use the propensity score as an instrument then the weights are always positive. Therefore, we consider forming an IV estimate using just the data from a given interval of our propensity score. This estimate will be some weighted average of the MTE (in the population, in actuality it is a weighted average of $LATE$s). If we form another IV estimate over a different interval of P that will be a weighted average of a different portion of the MTE. However, the absence of essential heterogeneity implies that these IV estimates must be the same, because they are both weighted averages of the same quantity with weights summing to 1. This suggests a test of equality of the IV estimates across different intervals of P as a way to test for essential heterogeneity.

To implement this test we form, for two specified intervals $[\underline{p}_1, \overline{p}_1]$ and $[\underline{p}_2, \overline{p}_2]$,

$$\begin{aligned} IV\left(\underline{p}_1, \overline{p}_1\right) &= \frac{Cov\left(Y, P(Z) \mid P(Z) \in [\underline{p}_1, \overline{p}_1]\right)}{Var\left(P(Z) \mid P(Z) \in [\underline{p}_1, \overline{p}_1]\right)} \\ IV\left(\underline{p}_2, \overline{p}_2\right) &= \frac{Cov\left(Y, P(Z) \mid P(Z) \in [\underline{p}_2, \overline{p}_2]\right)}{Var\left(P(Z) \mid P(Z) \in [\underline{p}_2, \overline{p}_2]\right)} \end{aligned}$$

and then test

$$\begin{aligned} H_0 &: IV\left(\underline{p}_1, \overline{p}_1\right) = IV\left(\underline{p}_2, \overline{p}_2\right) \\ H_1 &: IV\left(\underline{p}_1, \overline{p}_1\right) \neq IV\left(\underline{p}_2, \overline{p}_2\right) \end{aligned}$$

Because there is estimation error from two stages (estimating $P(Z)$ and constructing this IV estimate), we bootstrap the difference between these estimates and check whether 0 lies in the tail of our bootstrapped distribution of the difference between the estimates. In practice, we use the intervals $[0, p_{med}]$ and $[p_{med}, 1]$ where p_{med} is the median of $P(Z)$ in the overall sample.

6 The Power of the Tests

The absence of essential heterogeneity has a number of testable implications, and therefore there are a wide variety of possible tests which we could carry out. Above, we developed two tests which we chose for their analytic simplicity and their ease of implementation. However, it remains to be seen whether these tests have any power – that is, whether they are able to reject false null hypotheses. Clearly, there are many variables on which the power of the tests will depend and in this section we hope to exposit some of that dependence. In our simulation results, which we present below, we find that sample size and the variance of the instrument are both very important in how well these tests perform. In particular, large sample sizes help the tests immensely, and a large variance of the instrument relative to the variance of the unobservable in the choice equation also improves the power of the test. The intuition behind both of these results is fairly straightforward. In this section we provide the results of carrying out our tests on simulated data which is generated from a fairly restrictive model – namely the Generalized Roy Model, where all errors are normal. However, we simply use this as our base case because it allows for the simple parameterization of essential heterogeneity. It lets us show how the power of the tests varies with a one-dimensional index, where that index is a simple function of variances and covariances of the unobservables in the model.

Our simulated data is generated from the same model given above, but with the unobservable terms

restricted to be normal. That is, our potential outcomes are

$$\begin{aligned} Y_0 &= \alpha_0 + \beta_{10} X_1 + \beta_{20} X_2 + U_0 \\ Y_1 &= \alpha_1 + \beta_{11} X_1 + \beta_{21} X_2 + U_1 \end{aligned}$$

and our choice equation is

$$D = \mathbf{1}(\alpha_d + \beta_d Z \geq V)$$

where

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

We also generate the regressors and the instrument as normal random variables with distribution

$$\begin{pmatrix} X_1 \\ X_2 \\ Z \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \sigma_{X_1 Z} \\ \sigma_{X_1 X_2} & \sigma_{X_2}^2 & \sigma_{X_2 Z} \\ \sigma_{X_1 Z} & \sigma_{X_2 Z} & \sigma_Z^2 \end{pmatrix} \right)$$

In this model, the marginal treatment effect is given by

$$MTE(X = x, P(Z) = p) = (\alpha_1 - \alpha_0) + (\beta_{11} - \beta_{10}) X_1 + (\beta_{21} - \beta_{20}) X_2 + \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \Phi^{-1}(p)$$

where $\Phi^{-1}(\cdot)$ is the inverse of a standard normal CDF. Therefore, the amount by which the MTE will vary in p depends solely on the term $\frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V}$. We can rewrite this term as $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$ where ρ_{1V} is the correlation between U_1 and V and ρ_{0V} is the correlation between U_0 and V. It is this index which lets us vary the degree of heterogeneity of treatment effects and trace out the power function in this dimension. This number also gives us a reference point for the degree of heterogeneity that exists in our real datasets. Notice that this model imposes some strong assumptions, including that the MTE is monotone in p, but we can estimate this selection term in our datasets and compare it to the power we have calculated from this base case model to get an idea of why we may or may not be able to reject the hypothesis of no essential heterogeneity.

To calculate the power of our tests we need to know the distribution of the test statistics under the

null hypothesis and under various alternative hypotheses. Because of the multiple steps in our estimation procedure, calculating asymptotic standard errors is difficult and prone to error. Therefore, we form the distributions of the test statistic using the nonparametric bootstrap. The test we are using in both the test of linearity and the testing using $LATE$ is a Wald test. Therefore, we know that, asymptotically, the test statistic has a χ^2_k distribution when we are testing k restrictions. However, in order to get the distribution of the test statistic under the alternative hypothesis (of essential heterogeneity), we need to somehow restrict the form of heterogeneity and completely specify the data generating process. In our model, we can parameterize the amount of heterogeneity using the term $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$ and hence we also know that under the null hypothesis $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = 0$. Therefore, for our model we can also simulate the exact distribution of the test statistic under the null using a bootstrap procedure. So our test becomes

$$\begin{aligned} H_0 &: \text{Generalized Roy model with } \rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = 0 \\ H_1 &: \text{Generalized Roy model with } \rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = k \end{aligned}$$

with σ_1 and σ_0 fixed. We use this bootstrap procedure and a grid of alternative values for k to trace out the power function for each of our tests in the one dimension of this index. We calculate the power functions using both the exact distribution of the test statistic under the null and the asymptotic (χ^2) distribution of the test statistic under the null.

6.1 Polynomial Test

First, we calculate the power of the simple test of the linearity of $E(Y|P(Z) = p)$ in p given above. 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 $P(Z)$. As described above, because with our model the data generating process is completely specified under the null, we can simulate the exact distribution of the test statistic under the null hypothesis. The procedure for doing this is:

1. Generate data under the parameterization such that $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = 0$.
2. Sample N observations with replacement from the empirical distribution of the data.
3. Estimate $\widehat{P}(Z_i^*)$ using a probit.
4. Run OLS on $Y_i^* = X_i^*\beta_0 + X_i^*(\beta_1 - \beta_0)\widehat{P}(Z_i^*) + \sum_{j=1}^{J}\phi_j\widehat{P}(Z_i^*)^j + \varepsilon_i$

5. Form V^*, a Huber-White robust estimator of the covariance matrix of the parameters.

6. Form the test statistic $W^* = \boldsymbol{\phi}' \left[RV^*R' \right]^{-1} \boldsymbol{\phi}$, where $\boldsymbol{\phi}$ is the vector of coefficients on the nonlinear term of $\widehat{P}(Z_i^*)$ and R is the $(J-1) \times k$ restriction matrix that picks out the coefficients on nonlinear terms of $\widehat{P}(Z_i^*)$.

7. Repeat steps two through six 1,000 times.

8. Find the 0.95 quantile of the distribution of W^* from the bootstrap samples, call this critical value $c^*_{0.95}$.

Then, for a given alternative hypothesis (k), the procedure to calculate the power of the test is:

1. Generate data under the parameterization such that $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = k$.

2. Repeat steps two through six from above 500 times – in each iteration calculating the test statistic W^*_{alt}.

3. For the power using the exact distribution of under the null, calculate the proportion of bootstrap samples in which $W^*_{alt} > c^*_{0.95}$.

4. For the power using the asymptotic χ^2 distribution under the null, calculate the proportion of the bootstrap samples in which $W^*_{alt} > Q_{\chi^2_{J-1}}(0.95)$ where $Q_{\chi^2_k}(\tau)$ is the τ-quantile of a χ^2 distribution with k degrees of freedom.

The results of these calculations for a quadratic polynomial in P are given in Figures 1 and 2, while the calculations for a cubic polynomial are given in figures 3 and 4. Figures 1 and 3 show how the power functions varying across different sample sizes, but holding the other parameters fixed, and figures 2 and 4 hold all of the parameters fixed, including the sample size and just vary the variance of the instrument. We can see from these figures that both a large sample size and a large explanatory power of the instrument (which in our case is a large variance of the instrument relative to the unobservable in the choice equation) are both necessary in order to have reasonable power.

6.2 LATE/IV Test

Calculating the power of the test using linear IV over separate intervals of the instrument is carried out in a similar way. The model and assumptions we are using for our power calculations imply that the correct

specification of an IV regression has as its dependent variable Y and has exogenous independent variables X and endogenous independent variables $X * D$. The optimal instrument for the endogenous variables $X * D$ is $X * P(Z)$. Therefore, when trying to test whether the IV estimate using just observations with $P(Z)$ below the median is different from the IV estimate using observations with $P(Z)$ above the median, the procedure we use to calculate the exact distribution of the test statistic under the null is:

1. Generate data under the parameterization such that $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = 0$.
2. Calculate the median of the propensity scores in the simulated data $p_{med} = Q_{\widehat{F}}(0.5)$.
3. Sample N observations with replacement from the empirical distribution of the data, $\widehat{F}$.
4. Estimate $\widehat{P}(Z_i^*)$ using a probit.
5. Regress $X * D$ on $X * P(Z)$ and $X * P(Z) * \mathbf{1}(P(Z) > p_{med})$ and calculate the fitted values, call them $\widehat{XD}$. Regress $X * D * \mathbf{1}(P(Z) > p_{med})$ on $X * P(Z)$ and $X * P(Z) * \mathbf{1}(P(Z) > p_{med})$ and calculate the fitted values, call them $\widehat{XD}^+$.
6. Regress Y on X, $X * \mathbf{1}(P(Z) > p_{med})$ and the fitted values $\widehat{XD}$ and $\widehat{XD}^+$.
7. Call $\underline{\gamma}$ the vector of coefficients on $\widehat{XD}$, which are the IV estimate using observations below the median, and $\overline{\gamma}$ the vector of coefficients on $\widehat{XD}$ plus the vector of coefficients on $\widehat{XD}^+$, this will be the IV estimate using observations above the median.
8. Form V^*, a Huber-White robust estimator of the covariance matrix of the parameters.
9. Form the test statistic $W^* = (\underline{\gamma} - \overline{\gamma})' \left[RV^*R'\right]^{-1} (\underline{\gamma} - \overline{\gamma})$ where R is restriction matrix that selects the relevant terms of the covariance matrix.
10. Repeat steps two through nine 1,000 times.
11. Find the 0.95 quantile of the distribution of W^* from the bootstrap samples, call this critical value $c^*_{0.95}$.

Then, for each alternative hypothesis, the procedure to calculate the power of the test is:

1. Generate data under the parameterization such that $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0 = k$.

2. Repeat steps two through nine above 500 times – in each iteration calculating the test statistic W^*_{alt}.

3. For the power using the exact distribution of under the null, calculate the proportion of bootstrap samples in which $W^*_{alt} > c^*_{0.95}$.

4. For the power using the asymptotic χ^2 distribution under the null, calculate the proportion of the bootstrap samples in which $W^*_{alt} > Q_{\chi^2_{\dim(X)}}(0.95)$ where $Q_{\chi^2_k}(\tau)$ is the τ-quantile of a χ^2 distribution with k degrees of freedom.

Finally, we also calculate the power of a test for whether the IV estimates above and below the median differ, but based on a specification which does not include all of the $X * D$ interactions. This specification just regresses Y on X and D and instruments D with $P(Z)$. The power functions for these two methods are shown in Figures 5 through 8. Figures 5 and 6 show the calculations for the correctly specified model with all of the $X * D$ interactions and figures 7 and 8 show the calculations for the misspecified model where we do not include the $X * D$ interactions and only instrument for D. As in the previous figures, the first set show the power functions vary across different sample sizes, while the second set show how the power functions varying with the explanatory power of the instrument. The results are fairly similar to the results for the polynomial test and show that both sample size and explanatory power of the instrument are crucial.

7 Applying the Tests to the Data

We implement our method of testing for essential heterogeneity in a wide variety of settings to show how ubiquitous the concept is. We consider some examples from labor economics, including the choices of college graduation, high school graduation, GED certification, and union membership, as well as an example from education, namely the effect of school vouchers on test scores. This section implements the method described above, uses this method to estimate the basic structural parameter – the marginal treatment effect (MTE), and discusses some of the difficulties in recovering other treatment effects, especially through the use of instrumental variables (IV). Table 1 presents summary statistics from the subsample of the PSID which we use for the union data, and Table 2 presents the summary statistics from the different subsamples of the NLSY79 which we use in the schooling examples.

THIS SECTION HAS NOT YET BEEN UPDATED TO REFLECT THE LATEST RESULTS.

7.1 School Vouchers

A topic that has been the subject of much recent research is the effect of school vouchers on school quality. In particular, Chile implemented reforms in 1981 and again in 1996 which changed the structure of schooling in that country. The 1981 reforms decentralized the administration of public schools, and established certain privately-run schools which received a fixed per-pupil payment from the government. There are many nuanced aspects to this system which need to be considered carefully, however, we are merely trying to show how an analyst might naively draw the wrong conclusions based on standard methods. See McEwan (2001) for a more detailed description of the Chilean educational system. We will consider the question of whether attending a voucher school rather than a public school increases a student's test scores. Our data for this application comes from the Sistema de Medición de la Calidad de la Educación (SIMCE), which is a national standardized test administered once a year to students entering the 4^{th} grade, the 8^{th} grade and the 10^{th} grade. We have data on 56,213 individual students.

The binary choice in this setting is whether a student attends a voucher school ($D = 1$) or a public school ($D = 0$). In our first stage we run a probit of voucher school attendance on the following independent variables (Z): number of family members in the household, the quality of the infrastructure of the school, indicators for various household income categories, mother's highest grade completed, father's highest grade completed, and region indicators. We use the fitted values from this probit as our estimates of the propensity score $P(Z)$.

The test we are using as our outcome measure has two components – a math score and a verbal score. We define as our outcome the average of the two scores. Once we have our estimated propensity scores, we regress the outcome on the following controls (X) in addition to polynomial terms in the propensity score P: gender, mother's age, father's age, indicators for various household income categories, mother's highest grade completed, father's highest grade completed, indicators for the number of books in the household, whether the child attended a preschool, how hard the child studies, whether the child has a job, whether the parents attend meetings with the child's teachers, whether the parents regularly communicate with the child's teachers, whether the parents participate in the schooling of the child, whether the child's school helps economically disadvantaged students, whether the child's math teacher adequately prepared

the child, whether the child's language teacher adequately prepared the child, an indicator for urban, and region indicators.

Our test for essential heterogeneity, as described in Section 5, is implemented by estimating (9) where Z is our vector of instruments and X is our vector of individual characteristics (plus a constant term). We estimate (9) using varying degrees of polynomials in P, from 2 to 5. We report the results for the tests for the significance of the individual higher order terms as well as the joint test of the significance of all higher order terms in Panel A of Table 4 for the two outcomes. We can see that all of the specifications provide fairly significant evidence against linearity, although the 5th degree polynomial would only reject at the 10% level. The fitted values from the 5th degree estimate and the 4th degree estimate are very similar, so we see the higher p-value for the 5th degree estimate as evidence that by the at that point we have gone too far for the optimal bias-variance tradeoff. These low p-values give strong evidence for essential heterogeneity, or selection on the gains, in voucher school choice in Chile. We have even more evidence, however, coming from the test on the equality of LATEs.

Panel B of Table 4 presents the p-values from the pairwise tests for the equality of LATEs across different intervals. These tests are from the LATEs calculated using the fourth degree in P estimate of $E(Y|P(Z) = p)$. The intervals that we use are the deciles of the distribution of our estimated propensity scores, $P(Z)$. Using bootstrap standard errors for the fitted values of $E(Y|P(Z) = p)$ we can test whether the LATEs over different intervals were drawn from the same distribution. In particular, we compare all pairs of intervals and then use a Bonferroni adjustment to account for the fact that we are testing multiple hypotheses. As we can see, nearly all of the pairwise tests reject equality of the LATEs, which provides further evidence that the gains to voucher schools vary with the propensity score. An F-test for the joint equality of all of the LATEs gives a p-value of 0.0000 for each of the outcomes.

In order to see the effect of this essential heterogeneity, we can estimate the common treatment effects, ATE and TT, and see how they compare to the results found using OLS or IV. We can calculate the treatment effects as weighted averages of our estimates of the MTE, by using the weights given in Heckman and Vytlacil (2005). Because our estimate of the MTE will differ based on the number of polynomial terms used to estimate $E(Y|P(Z) = p)$, our estimates of the treatment effects will also differ based on the degree of this polynomial. We give the estimates for the different degrees of polynomial in Panel C of Table 4. In that table we also show the estimate obtained using standard IV, namely the ratio of the covariances, described above with P as our instrument, as well as the IV estimate found using the implicit

weight that the IV estimator is placing on the MTE. These two estimates differ not only because our estimate of the MTE is not exact, but also because the weights themselves are merely an estimate.

Notice that depending on the procedure used, the estimates of the "effect" of voucher schools will vary widely. If we simply take the difference between the mean test score for the voucher schools and the mean test score for the public schools we get an effect of 17.5933 (the tests are standardized to get have a mean of 250 and a standard deviation of 50). Once we control for various characteristics of the students (X), however, the OLS estimate drops to 7.5377. If we instrument for voucher school attendance with our instruments Z, then running two-stage least squares gives an effect of 27.7239. If we first estimate the propensity scores, $P(Z)$, and then use those as an instrument, we get an effect of 25.9592, as seen in Panel C of Table 4. The rest of Panel C of Table 4 gives the estimates of the other treatment effects which we consider (ATE and TT) and we can see that the IV estimate is not generally capturing either.

The fitted values of $E(Y|P = p, X = x)$ for the mean X and it's derivative ($MTE(p, x)$) at the mean X, using the specification with a fourth degree polynomial in P, are plotted in Figure 1 for the average test scores. Also, Figure 1 plots the weights that the IV estimate (using $P(Z)$ as an instrument) is implicitly placing on the MTE. Finally, the last panel in Figure 1 shows the histogram of propensity scores, where we can see that in this case we have nearly full support.

THIS SECTION HAS NOT YET BEEN UPDATED TO REFLECT THE LATEST RESULTS.

7.2 Union Membership

There is a long literature focusing on the effect of unionism on wages. We focus on replicating, as closely as possible, the analysis of Lee (1978), which used data from the Survey of Economic Opportunity of 1967.[4] In our study we use instead the Panel Study on Income Dynamics (PSID) at a cross-section in 1988.[5] In this setting, the outcome variable is (log) weekly wages and the binary choice that agents face is whether to be a union member or not. In our sample, we include only men between the ages of 18 and 65 who are not enrolled in school and who worked at least one week in the previous year. This gives a sample size of $N = 4,081$ (out of a total PSID sample size that year of 7,114).

[4] See also Farber (1983), Duncan and Leigh (1985) and Robinson (1989) for other studies of unionism.

[5] We choose 1988 because the sample size is relatively large in that year, but the results seem to be fairly similar across years.

The first stage is a probit of union membership on a variety of worker characteristics. In our specification, $D = 1$ if the individual reports being a union member and $D = 0$ if he is not a union member, and the independent variables (Z) are indicators for residence in the northeast, midwest, south, and a metropolitan area of at least 250,000; indicators for years of education categories: 1 to 7 years, 9 to 11 years, 12 years, and 13 or more years; experience, experience squared, and indicator for white; indicators for having worked 1 to 26 weeks in the previous years and 48 to 52 weeks in the previous year; and indicators for the occupations: mining, construction, manufacturing durable goods, and manufacturing non-durable goods. The variables are chosen to match as closely as possible with those chosen by Lee (1978). We take the fitted values from this probit and use these as our estimates of the propensity score, $P(Z)$.

We then regress log weekly wages on the following controls (X), plus polynomial terms in P: indicators for residence in the northeast, midwest, south, and a metropolitan area of at least 250,000; indicators for years of education categories: 1 to 7 years, 9 to 11 years, 12 years, and 13 or more years; experience, experience squared, and indicator for white; and indicators for having worked 1 to 26 weeks in the previous years and 48 to 52 weeks in the previous year. Note that in this case the X variables are identical to the Z variables except the X does not include occupations.

Our test for essential heterogeneity, as described in Section 5, is implemented by estimating (9) where Z is our vector of instruments and X is our vector of individual characteristics (plus a constant term). We estimate this for varying degrees of the polynomial in P and Panel A of Table 5 gives the p-values of our test for different degrees of the polynomial. We can see that only once we have added a cubic term can we reject the linearity of $E(Y|P = p)$. However, once we add that cubic term, the results are highly significant and we can see that adding higher order polynomial terms only makes the estimates less precise.

The second test for linearity is that of the equality of the local average treatment effect (LATE) over the support of P. The intervals over which we calculate LATE are the deciles of distribution of the estimated propensity score. Using bootstrap standard errors for the fitted values of $E(Y|P(Z) = p)$ we can test whether the LATEs over different intervals were drawn from the same distribution. In particular, we compare all pairs of intevals and then use a Bonferroni adjustment to account for the fact that we are testing multiple hypotheses. Panel B of Table 5 provides the p-values from these pairwise tests. Also, an F-test for the joint equality of the means of the LATEs gives a p-value of 0.000.[6]

[6] We refrain from interpreting the results of these tests as overwhelming evidence of heterogeneity, because in some simulations without essential heterogeneity, these tests still rejected linearity. These tests do, however, provide some extra evidence of nonlinearity.

We discussed above how economists are usually interested in various means of the distribution of individual treatment effects. However, with a limited support of the propensity score, as is the case with our unionism data, we will be unable to calculate one of our desired mean treatment effects – ATE. In order to recover ATE, we need the support of $P(Z)$ to be the entire unit interval, which we can see from the histogram, it is not. Using the weights on the MTE from Heckman and Vytlacil (2005) makes it explicit that using our data with limited support we will be unable to recover either of these parameters. The problem manifests itself in general both in our inability to calculate the MTE outside of the support and our inability to calculate the weights that would be needed to find the treatment effects (in the case of ATE, however, we know the weights are 1). However, because we are simply using polynomials to estimate $E(Y|P)$ (and hence polynomials to estimate the MTE), we can extrapolate the MTE outside of the support of P, but these estimates of the MTE are not valid. These will allow use to estimate what we call the empirical ATE, which again will not be valid because of the extrapolation of the MTE. Panel C of Table 5 gives the estimated treatment effects found by weighting the MTE according to the weights in Heckman and Vytlacil (2005). In addition, Panel C gives the estimates found using traditional IV (the ratio of covariances described above), as well as the IV estimate found using the weights. These two IV estimates differ not only because of the inexact estimates of the MTE, but also because the weights themselves are estimated.

Looking at these treatment parameters we can see the danger in trying to calculate parameters that depend on the MTE outside of the support of P. For example, the ATE weights the MTE equally over the whole interval, but because in the upper part of the unit interval, that MTE is simply extrapolated from the support of P, it leads to ridiculous values.

The fitted values of $E(Y|P = p, X = x)$ for the mean X and it's derivative ($MTE(p, x)$) at the mean X, using the quadratic in P specification, are plotted in Figure 2. Also, Figure 2 plots the weights that the IV estimate (using $P(Z)$ as an instrument) is implicitly placing on the MTE. Finally, the histogram of propensity scores in Figure 2 shows the limited support of P.

THIS SECTION HAS NOT YET BEEN UPDATED TO REFLECT THE LATEST RESULTS.

7.3 GED vs. High School Dropout

Another choice setting we consider is whether or not an individual chooses to receive a GED. **Need a reference here to something.** The data we use for this application comes from the National Longitudinal Survey of Youth 1979 (NLSY79). We include only 30-year-old men from the "core" sample. We say that $D = 1$ if the individual is a GED recipient and $D = 0$ if the individual is a high school dropout. All other education categories are excluded. This leads to a sample size of 409. As our outcome variable, we use an average of log weekly wages at ages 29, 30 and 31.

In our first step, we estimate the propensity scores using a probit with D as the dependent variable and the following variables as independent variables (Z): standardized AFQT score, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, cost of GED, wages of local high school dropouts, unemployment of local high school graduates, indicators for black, hispanic, residence in the south at age 14, residence in an urban area at age 14 and year of birth indicators.

We form our propensity scores, P, as the fitted values from this probit. Then, we regress the outcome variable on polynomial terms in P, in addition to the following controls (X): job tenure, job tenure squared, experience, standardized AFQT score, standardized noncognitive test scores, highest grade completed, and indicators for black, hispanic and being married.

Again we implement our test for linearity by estimating (9) using the X controls and Z instruments described above. Panel A of Table 6 shows the p-values resulting from these tests for specifications of (9) with varying degrees of the polynomial. As we can see, in this case we need to add a cubic term in P before we see any significance of the nonlinear terms in P. Even though in the cubic specification neither the quadratic nor the cubic terms is individually significant, the joint test on both is fairly significant (at the 10% level), which we interpret as reasonable evidence for the existence of essential heterogeneity.

As above, we can also test for the equality of the LATE over different intervals in the support of P. The intervals over which we calculate LATE are the deciles of distribution of the estimated propensity score. We test for the equality of the LATEs calculated using the cubic specification over these intervals and we report the p-values from these tests in Panel B of Table 6. Also an F-test for the joint equality of the LATEs has a p-value of 0.0000.

Because in this data we have near full support of $P(Z)$ over the unit interval we can estimate all of the traditional treatment parameters using the weights from Heckman and Vytlacil (2005). We use the

approximate MTE calculated as the derivative of our polynomial estimate of $E(Y|P)$ to get the estimates of the treatment effects. Because we are only using approximations to the MTE, our estimates of the treatment effects differ depending on the degree of the polynomial used to estimate $E(Y|P)$. Also, we calculate the IV estimate using both the traditional ratio of covariances described above and using the weights from Heckman and Vytlacil (2005). These two IV estimates differ not only because of the inexact estimates of the MTE, but also because the weights themselves are estimated. Panel C of Table 6 reports these estimated treatment effects.

The fitted values of $E(Y|P = p, X = x)$ at the mean X and the $MTE(p, x)$ at the mean X, using the cubic specification, are plotted in Figure 3. In addition, Figure 3 plots the weights that IV is implicitly placing on the MTE, as well as histogram of the estimated propensity scores.

THIS SECTION HAS NOT YET BEEN UPDATED TO REFLECT THE LATEST RESULTS.

7.4 High School Diploma vs. High School Dropout

Another choice setting of interest in labor economics is the return to a high school education. See Katz and Autor (1999) and Card (2001) for extensive treatments of the topic. For this application, we again use data from the NLSY79. We include only 30-year-old men from the "core" sample. We say that $D = 1$ if an individual's highest level of education is a high school diploma and $D = 0$ if the individual is a high school dropout (not a GED recipient). This gives a sample size of 1083. The outcome variable is the average of log hourly wages at ages 29, 30, and 31.

In order to estimate the propensity scores, we run a probit of D on the following independent variables (Z): standardized AFQT score, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, wages of local high school dropouts, wages of local high school graduates, unemployment of local high school dropouts, unemployment of local high school graduates, indicators for black, hispanic, residence in the south at age 14, residence in an urban area at age 14 and year of birth indicators.

Using the fitted values from this probit we form our estimates of the propensity score, $P(Z)$. We then regress the outcome variable on polynomials in P plus the following regressors (X): job tenure, job tenure squared, experience, standardized AFQT score, standardized noncognitive test scores, highest grade

completed, and indicators for black, hispanic and being married.

Again we implement our test by estimating (9) for different degrees of the polynomial in P. Table 7, panel A, contains the results from these tests on this data. As we can see, these tests do not reject the null hypothesis of linearity of $E(Y|P = p)$ for any of the polynomials of degree 2 to 5. This means that we cannot rule out the constant-MTE case and so it may not be necessary to deal with the additional complications of incorporating the essential heterogeneity.

The other test for linearity is the test of the equality of the LATEs over various intervals. The intervals over which we calculate LATE are the deciles of distribution of the estimated propensity score. The p-values from these tests are given in panel B of Table 7 and many of them do indeed reject equality. Also, an F-test for the the joint equality of means of the LATEs for any of the specifications gives a p-value of 0.000.[7]

Using the approximate MTE calculated above we can calculate the various treatment parameters by weighting the MTE by the weights given in Heckman and Vytlacil (2005) to get the treatment effects listed in panel C of Table 7.

The fitted values of $E(Y|P = p, X = x)$ for the mean X and it's derivative ($MTE(p, x)$) at the mean X, for the quadratic in P specification, are plotted in Figure 4. In addition, we give the weights that IV is implicitly placing on the MTE, and the histogram of estimated propensity scores.

THIS SECTION HAS NOT YET BEEN UPDATED TO REFLECT THE LATEST RESULTS.

7.5 College Degree vs. High School Diploma

Finally, we test for essential heterogeneity in the decision of whether or not to graduate from college. The data comes from the NLSY79. We include only 30-year-old men from the "core" sample. We consider $D = 1$ if the individual is a college graduate and $D = 0$ if the individual's highest educational attainment is a high school diploma. This leads to a sample with 1335 observations. As our outcome variable we use the average of log wages at ages 29, 30 and 31.

In the first stage we run a probit of D on the following independent variables (Z): standardized AFQT score, father's highest grade completed, mother's highest grade completed, number of siblings,

[7] Note the previously-mentioned caveat about interpreting these LATE rejections too strongly because some simulations indicate they may reject even in the absence of heterogeneity.

family income in 1979, wages of local high school graduates, wages of local some college, wages of local college graduates, unemployment of local high school graduates, unemployment of local some college, unemployment of local college graduates, indicators for black, hispanic, residence in the south at age 14, residence in an urban area at age 14 and year of birth indicators.

As our estimates of the propensity score, $P(Z)$, we use the fitted values from this probit. Using those fitted values, we regress the outcome variable on polynomials in the propensity score in addition to the following control variables (X): job tenure, job tenure squared, experience, standardized AFQT score, standardized noncognitive test scores, highest grade completed, and indicators for black, hispanic and being married.

We test for linearity by estimating (9) using this college data, for varying degrees of the polynomial in P. Panel A of Table 8 gives the p-values from the tests of nonlinearity described above. As we can see from the table, after adding only a quadratic term in P the test is able to strongly reject the linearity of $E(Y|P = p)$. We can interpret this as evidence of the fact that individuals are selecting into college based on their idiosyncratic gains from college. This means that the MTE is not constant and IV will not identify any meaningful treatment parameter.

We also test for essential heterogeneity by testing the equality of the LATEs across different intervals in the support of the propensity score, P. The intervals over which we calculate LATE are the deciles of distribution of the estimated propensity score. Comparing each pair of LATEs, using the quadratic specification, leads to the p-values reported in panel B of Table 8. Also, an F-test for the equality of means of the LATEs gives a p-value of 0.000. These tests provide more evidence of the essential heterogeneity in this choice setting.

Using the approximate MTE calculated above we can calculate the various treatment parameters by weighting the MTE by the weights given in Heckman and Vytlacil (2005). Panel C of Table 8 reports these treatment effects.

The fitted values of $E(Y|P = p, X = x)$ for the mean X and it's derivative ($MTE(p, x)$) at the mean X, for the quadratic in P specification are plotted in Figure 5. In addition, Figure 5 shows the weights that the IV estimator places on the MTE as well as the histogram of estimated propensity scores.

8 Summary and Conclusion

Although the recent literature has shown the theoretical difficulties in applying standard IV methods to data which contain essential heterogeneity, this paper seeks to determine whether such concerns are important in practice. Using data from four prototypical choice settings in labor economics, we have shown reasonable evidence that such heterogeneity is indeed present – i.e. that individuals are selecting into treatment on the basis of their idiosyncratic gain from the treatment. Our strongest results come from the data on union membership and college graduation, while the data on high school graduation and GED certification are less conclusive. Therefore, in these settings, and potentially many others, researchers need to use caution in interpreting estimates found using traditional IV methods. The underlying structural parameter of importance is the MTE, which, as we have shown, may not be constant, and therefore an IV estimate, which gives one number, will be unable to identify this parameter throughout its support.

Table 1: Specification of the Generalized Roy Model Used

To Calculate the Power of the Tests

Outcomes	Decision Rule:
$Y_0 = \alpha_0 + \beta_{10} X_1 + \beta_{20} X_2 + U_0$	$D = \mathbf{1}(\alpha_d + \gamma_d Z \geq V)$
$Y_1 = \alpha_1 + \beta_{11} X_1 + \beta_{21} X_2 + U_1$	

Observed $Y = DY_1 + (1 - D)Y_0$

with parameters:	with parameters:
$\alpha_0 = 0$, $\beta_{10} = 0.1$, $\beta_{20} = 0.3$	$\alpha_d = 0.2$
$\alpha_1 = 0.2$, $\beta_{11} = 0.2$, $\beta_{21} = 0.4$	$\gamma_d = 0.3$

Distribution of Unobservables:

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

The power function is traced out by varying ρ_{1V} from -0.7 to 0.7. Values outside this interval lead to a covariance matrix which is not positive definite.

Distribution of Observables:

$$\begin{pmatrix} X_1 \\ X_2 \\ Z \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 * \sqrt{10} & 0.5 * \sigma_Z \\ 0.5 * \sqrt{10} & 10 & 0.5 * \sigma_Z \\ 0.5 * \sigma_Z & 0.5 * \sigma_Z & \sigma_Z^2 \end{pmatrix} \right)$$

I calculate the power function for values of σ_Z^2 between 1 and 10.

References

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

Card, D. (2001, September). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Carneiro, P., K. Hansen, and J. J. Heckman (2001, Fall). Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Economic Policy Review 8*(2), 273–301.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422. 2001 Lawrence R. Klein Lecture.

Duncan, G. M. and D. E. Leigh (1985, July). The endogeneity of union status: An empirical test. *Journal of Labor Economics 3*(3), 385–402.

Farber, H. S. (1983). Worker preferences for union representation. In J. Reid (Ed.), *Research in Labor Economics*, Volume Supplement 2: New Approaches to Labor Unions. Greenwich, Connecticut: JAI Press.

Gronau, R. (1974, November-December). Wage comparisons – a selectivity bias. *Journal of Political Economy 82*(6), 1119–43.

Heckman, J. J. (1974, July). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. J. (1976a, December). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement 5*(4), 475–492.

Heckman, J. J. (1976b). Simultaneous equation models with both continuous and discrete endogenous variables with and without structural shift in the equations. In S. Goldfeld and R. Quandt (Eds.), *Studies in Nonlinear Estimation*, pp. 235–272. Cambridge, MA: Ballinger Publishing Company.

Heckman, J. J. (1979, January). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings equations and rates of return: The Mincer equation and beyond. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, pp. 307–458. Amsterdam: North-Holland.

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. and R. Robb (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples*, pp. 63–107. New York: Springer-Verlag. Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

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 (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 (2000, January). The relationship between treatment parameters within a latent variable framework. *Economics Letters 66*(1), 33–39.

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

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

Katz, L. F. and D. H. Autor (1999). Changes in the wage structure and earnings inequality. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, Chapter 25, pp. 1463–1555. New York: North-Holland.

Lee, L.-F. (1978, June). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review 19*(2), 415–433.

Manski, C. F. and J. V. Pepper (2000, July). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica 68*(4), 997–1010.

Powell, J. L. (1994). Estimation of semiparametric models. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Volume 4*, pp. 2443–2521. Amsterdam: Elsevier.

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.

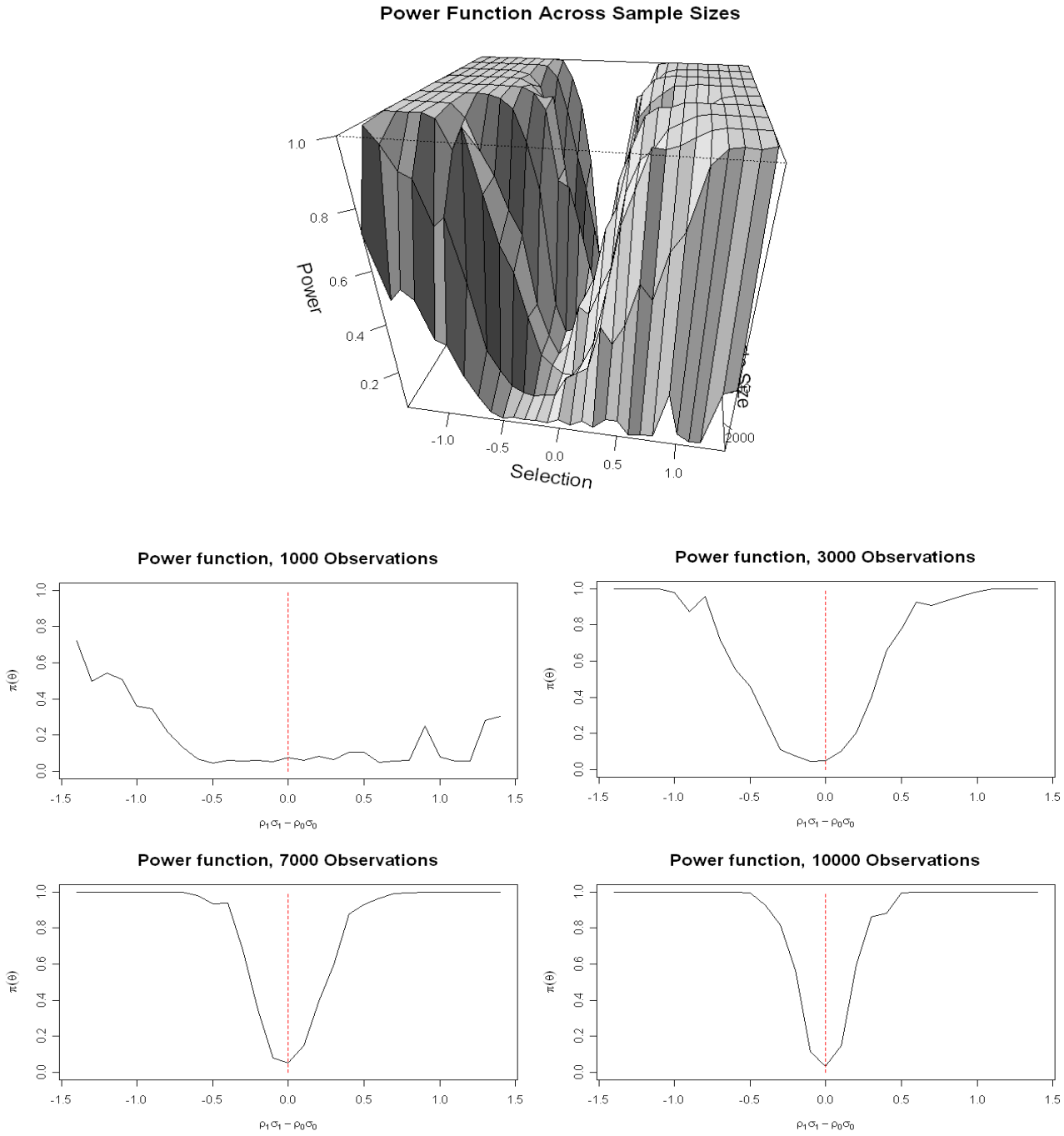
Robinson, C. (1989, June). The joint determination of union status and union wage effects: Some tests of alternative models. *Journal of Political Economy 97*(3), 639–667.

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

White, H. (1984). *Asymptotic Theory for Econometricians*. Orlando, FL: Academic Press.

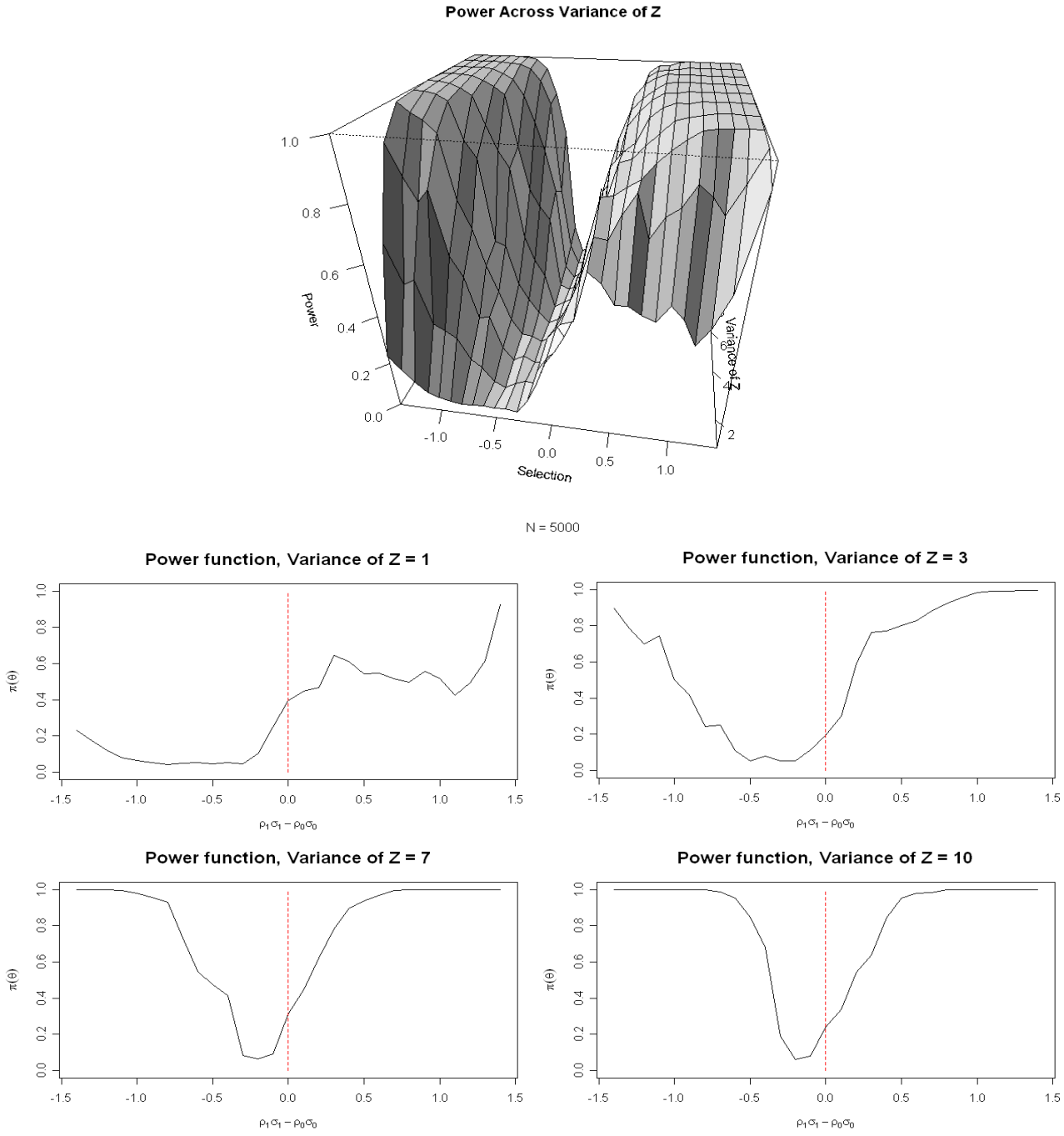
Figure 1: Power of the Test of Linearity of E(Y | P) Using Quadratic Polynomial in P, Varying Sample Size

(Using Chi-Square Distribution under the Null)



Note: The variance of the instrument is 10. For each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) we bootstrap the Wald statistic for the test that the coefficient on the P^2 term is zero 500 times and calculate what proportion of those test statistics lie outside the 95th percentile of a χ^2 distribution with 1 degree of freedom (we are testing one coefficient).

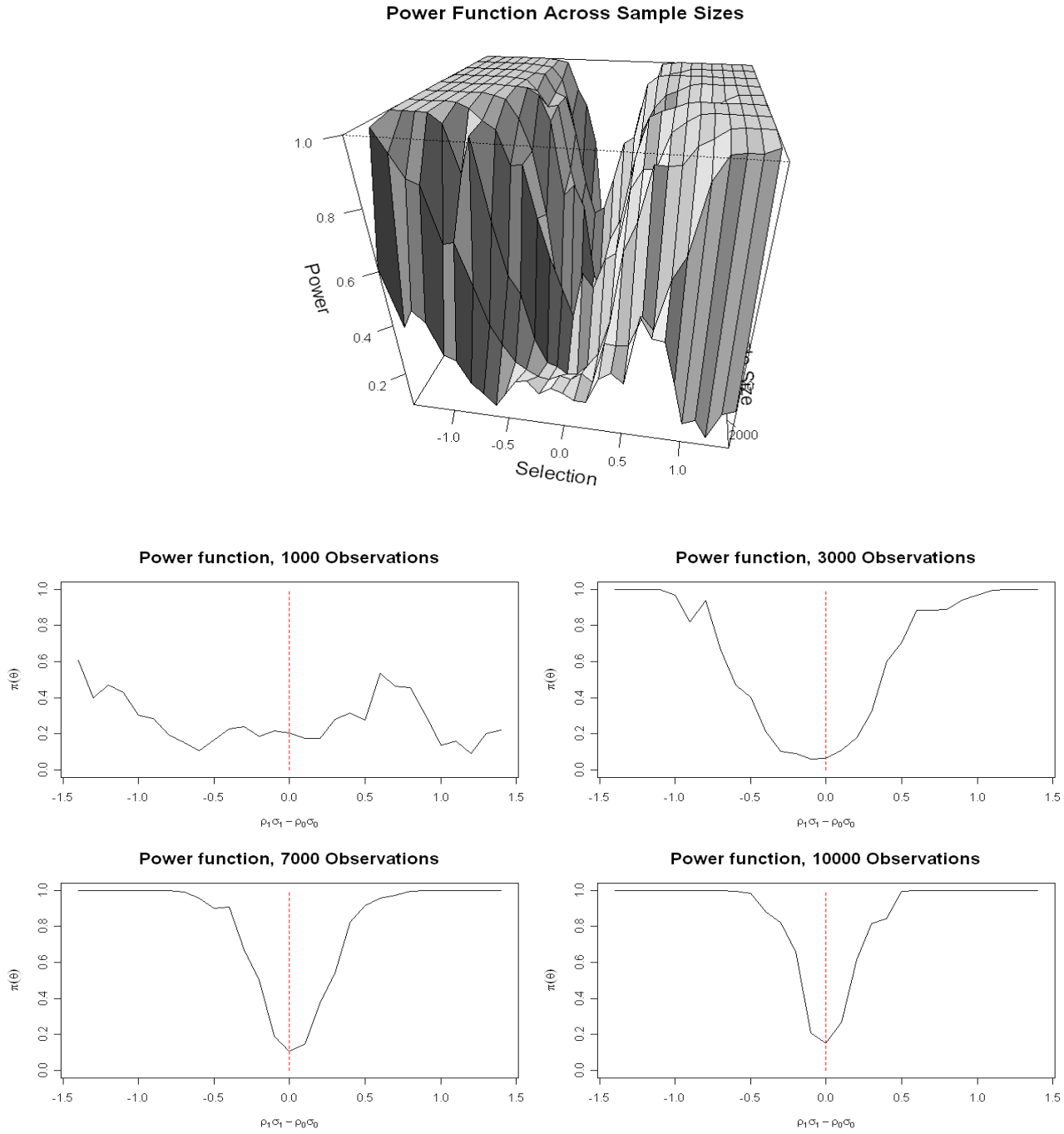
Figure 2: Power of the Test of Linearity of E(Y | P)
Using Quadratic Polynomial in P, Varying σ_Z
(Using Chi-Square Distribution under the Null)



Note: The sample size is 5,000. For each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) we bootstrap the Wald statistic for the test that the coefficient on the P^2 term is zero 500 times and calculate what proportion of those test statistics lie outside the 95th percentile of a χ^2 distribution with 1 degree of freedom (we are testing one coefficient).

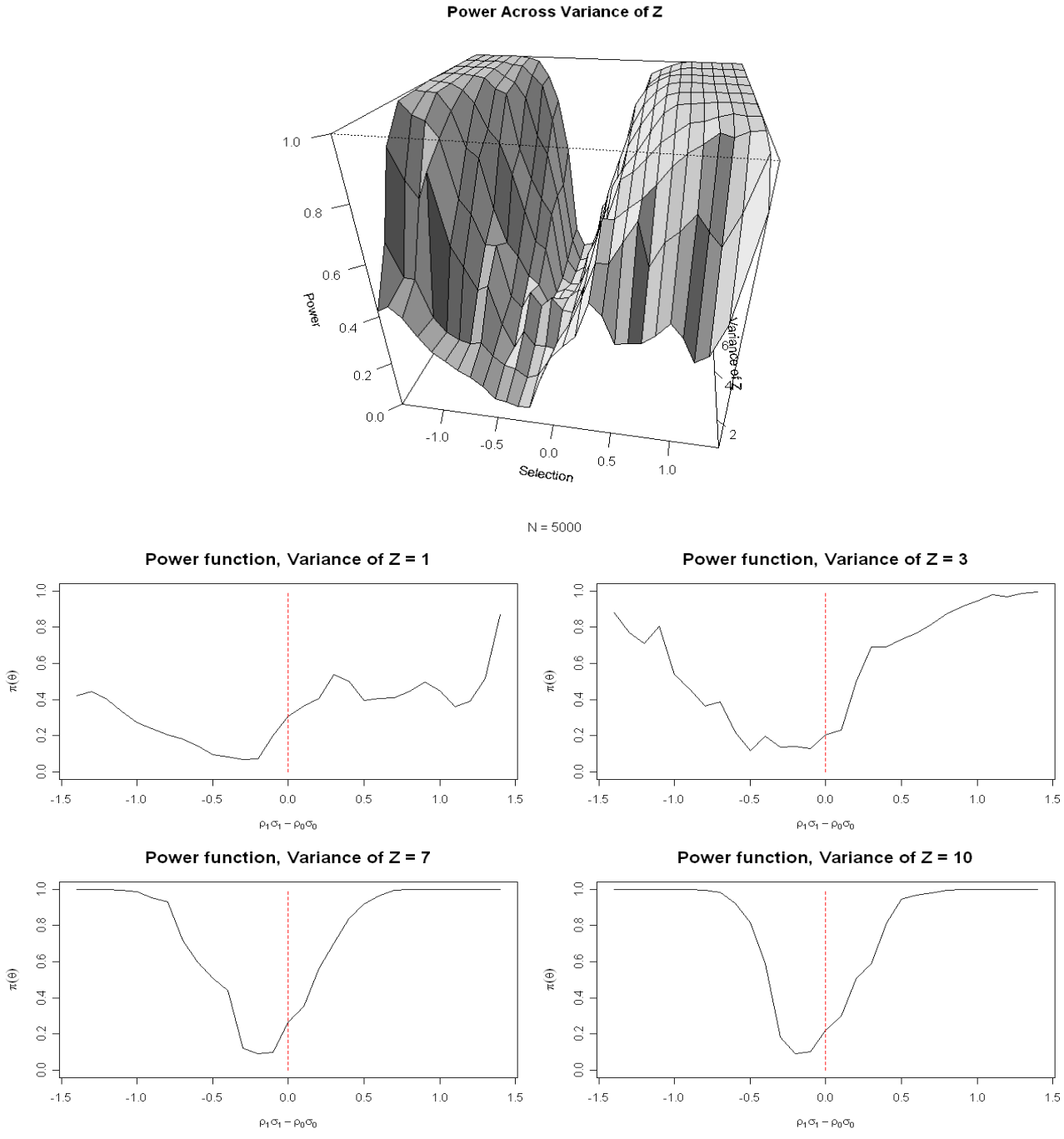
Figure 3: Power of the Test of Linearity of E(Y | P) Using Cubic Polynomial in P, Varying Sample Size

(Using Chi-Square Distribution under the Null)



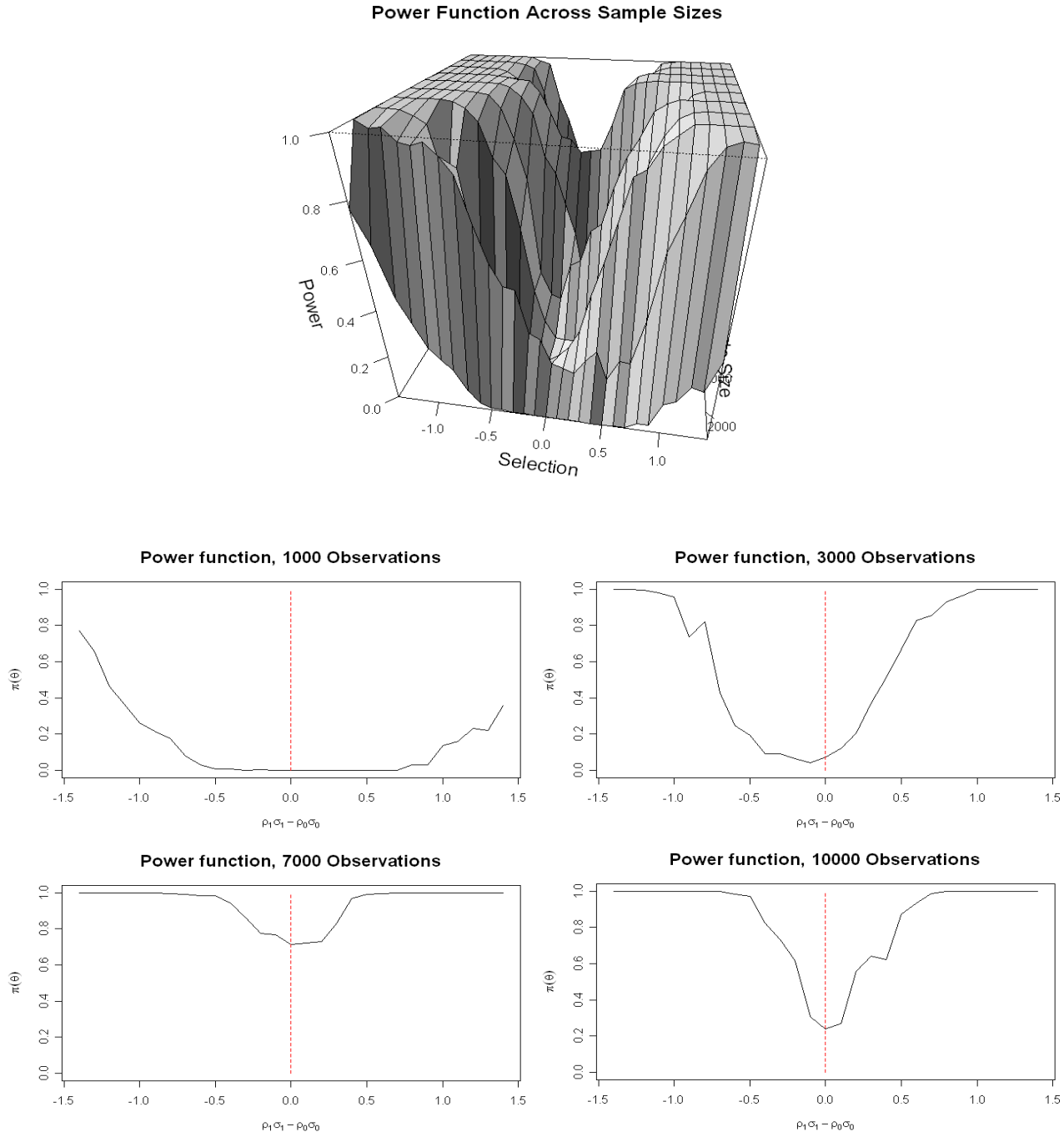
Note: The variance of the instrument is 10. For each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) we bootstrap the Wald statistic for the joint test that the coefficients on the P^2 and P^3 terms are zero 500 times and calculate what proportion of those test statistics lie outside the 95th percentile of a χ^2 distribution with 2 degrees of freedom (we are testing two coefficients).

Figure 4: Power of the Test of Linearity of E(Y | P)
Using Cubic Polynomial in P, Varying σ_Z
(Using Chi-Square Distribution under the Null)



Note: The sample size is 5,000. For each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) we bootstrap the Wald statistic for the joint test that the coefficients on the P^2 and P^3 terms are zero 500 times and calculate what proportion of those test statistics lie outside the 95th percentile of a χ^2 distribution with 2 degrees of freedom (we are testing two coefficients).

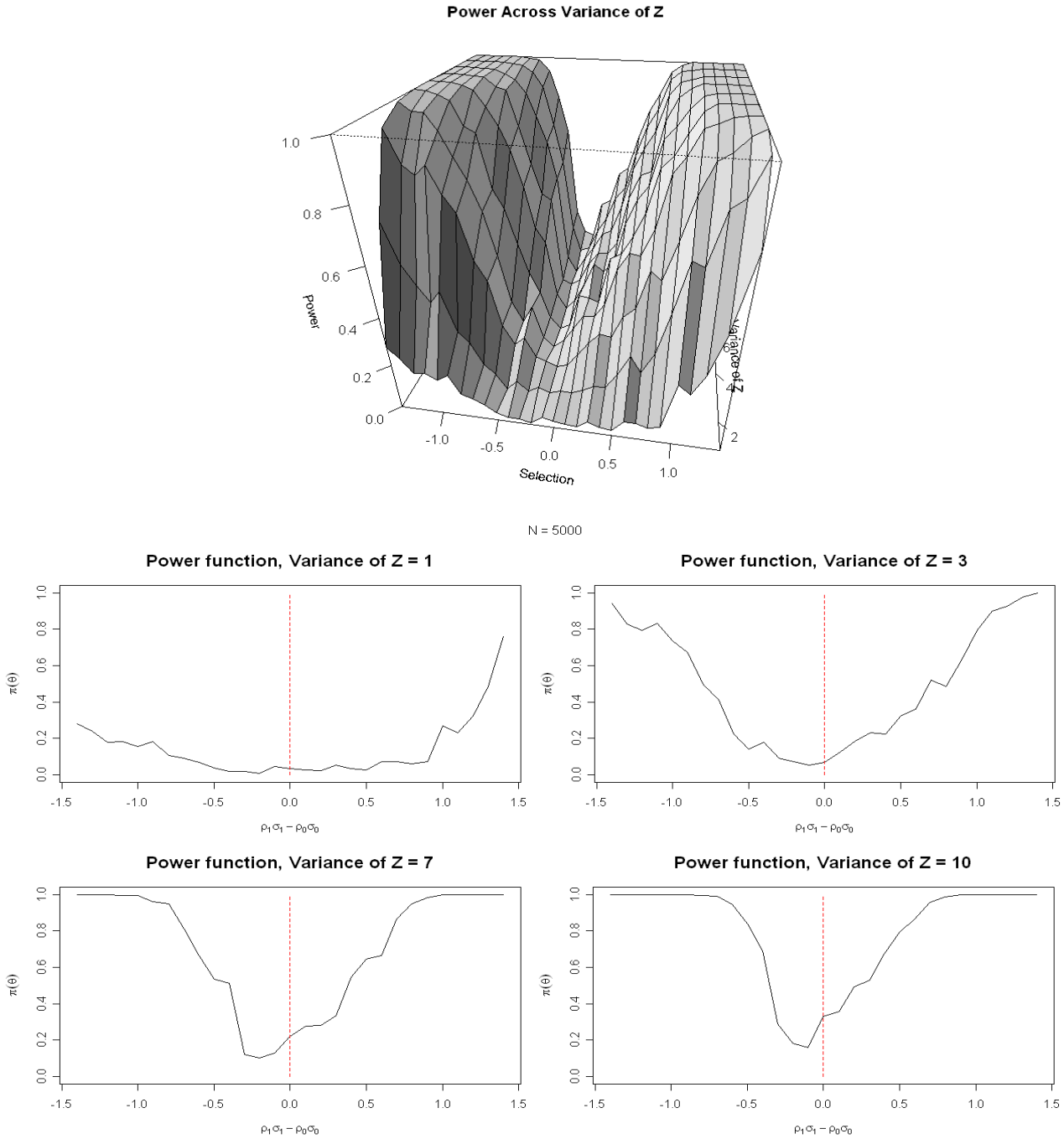
Figure 5: Power of the Test of Equality of IV Estimates Using Propensity Scores Above and Below the Median, Varying Sample Size
(Using Chi-Square Distribution under the Null)



Note: The variance of the instrument is 10. The power is calculated for each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) by bootstrapping the Wald statistic 500 times calculating what proportion of the test statistics lie outside the 95th percentile of a χ^2 distribution with 3 degrees of freedom (we are testing 3 coefficients).

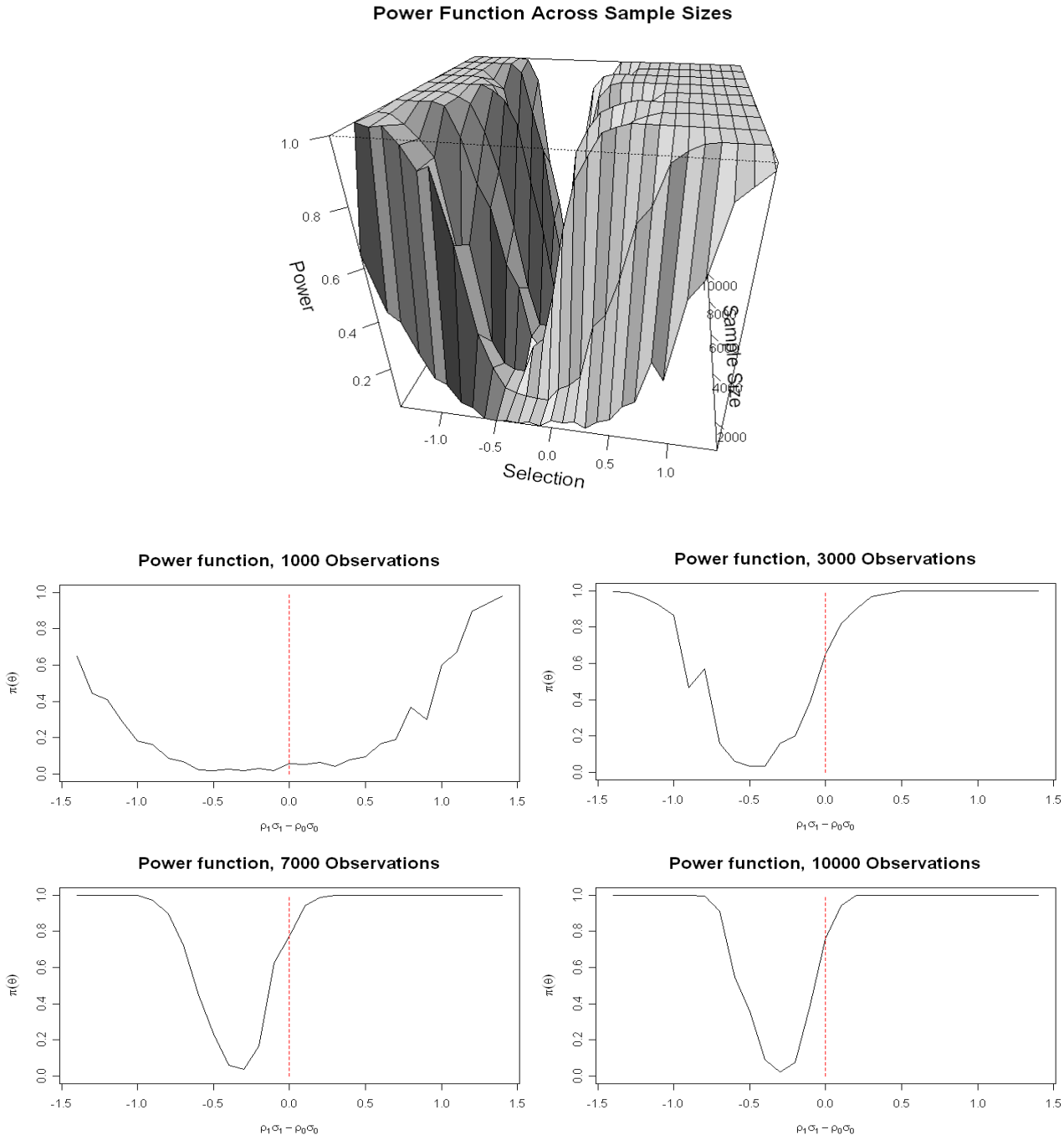
Figure 6: Power of the Test of Equality of IV Estimates Using Propensity Scores Above and Below the Median, Varying σ_Z

(Using Chi-Square Distribution under the Null)



Note: The sample size is 5,000. The power is calculated for each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) by bootstrapping the Wald statistic 500 times calculating what proportion of the test statistics lie outside the 95th percentile of a χ^2 distribution with 3 degrees of freedom (we are testing 3 coefficients).

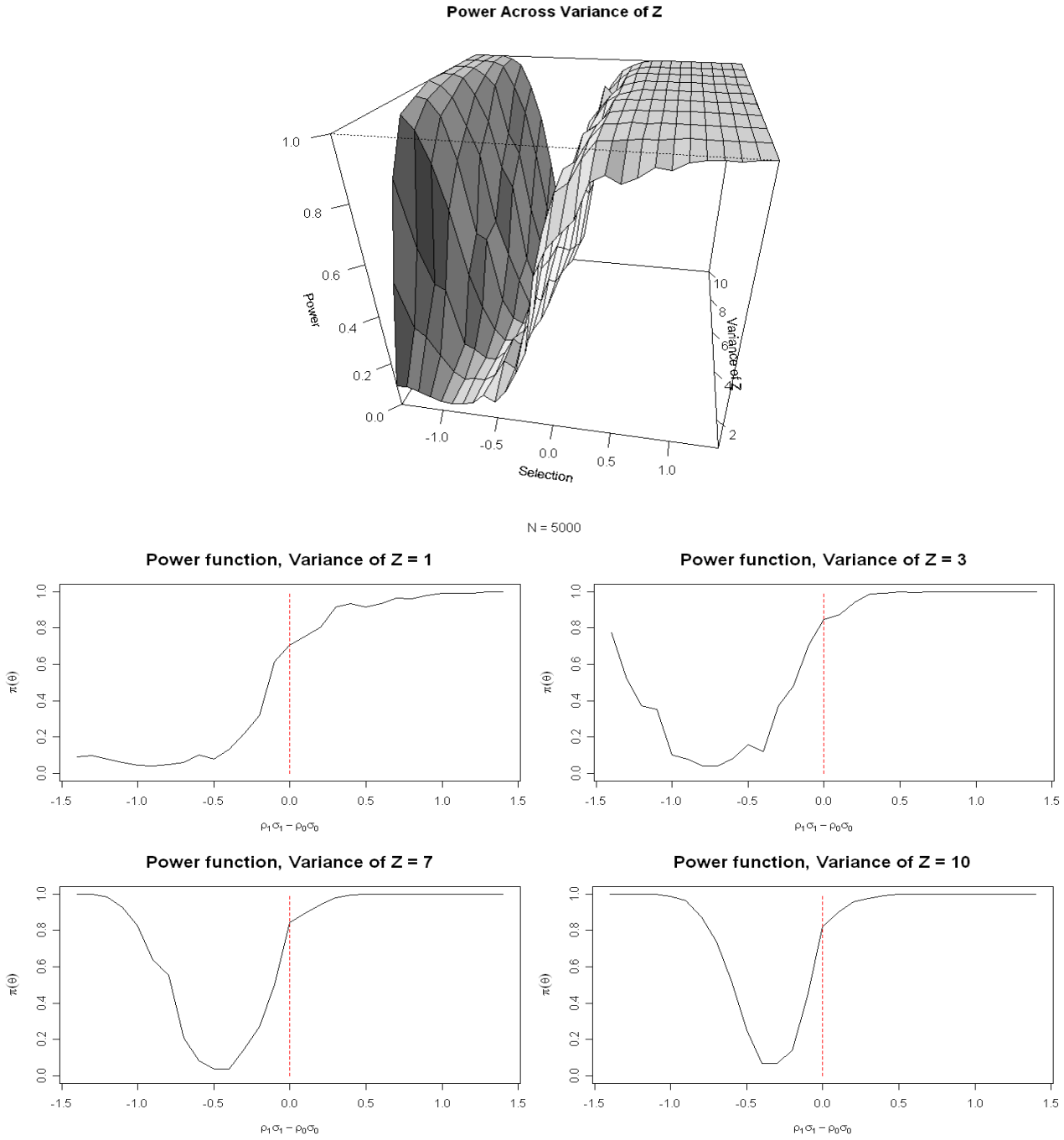
Figure 7: Power of the Test of Equality of Simple IV Estimates Using Propensity Scores Above and Below the Median, Varying Sample Size
(Using Chi-Square Distribution under the Null)



Note: The variance of the instrument is 10. The power is calculated for each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) by bootstrapping the Wald statistic 500 times calculating what proportion of the test statistics lie outside the 95th percentile of a χ^2 distribution with 1 degree of freedom (we are testing 1 coefficient). These IV estimates are the coefficient on D and contain no interactions with X (so they are misspecified).

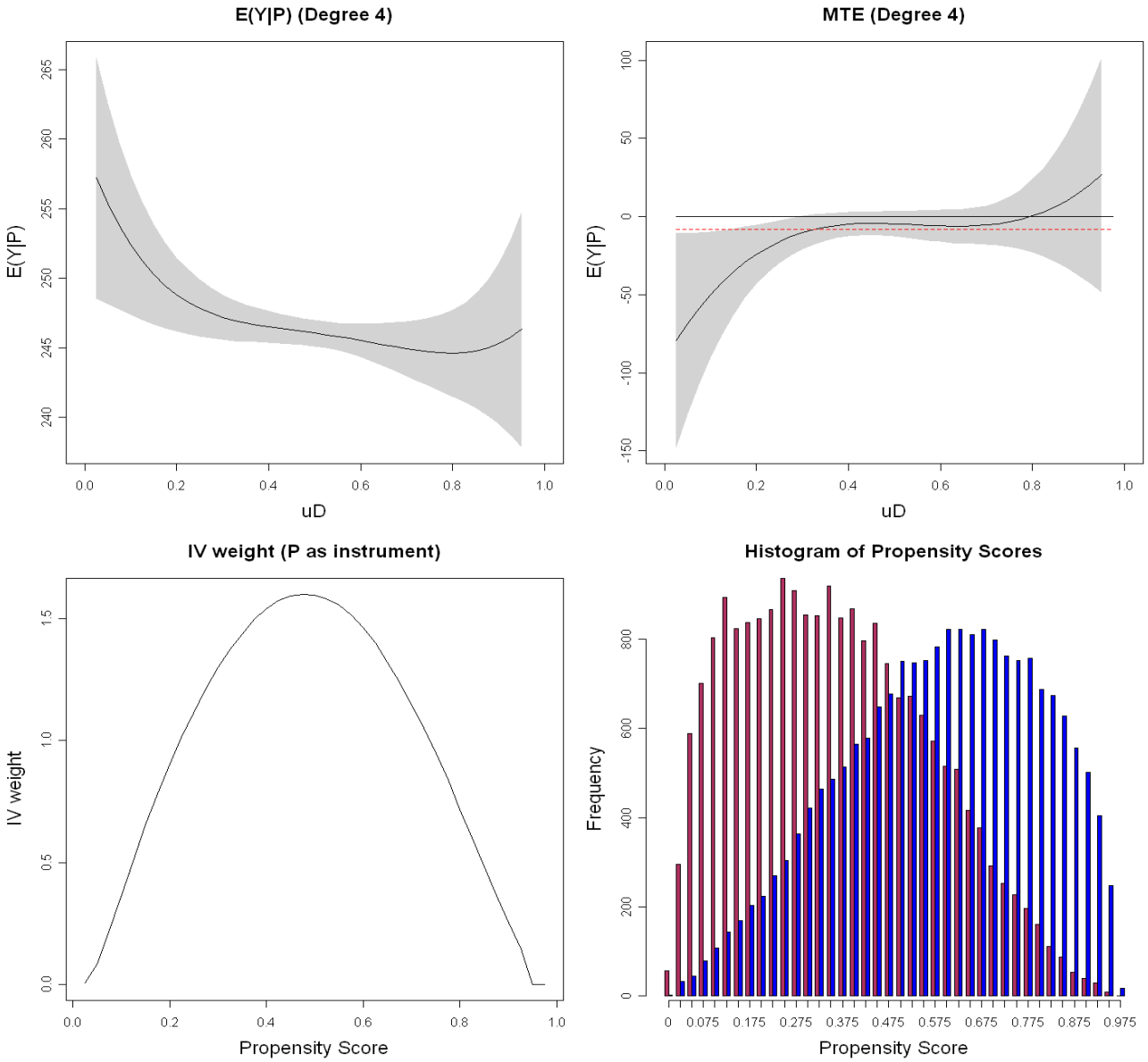
Figure 8: Power of the Test of Equality of Simple IV Estimates Using Propensity Scores Above and Below the Median, Varying σ_Z

(Using Chi-Square Distribution under the Null)



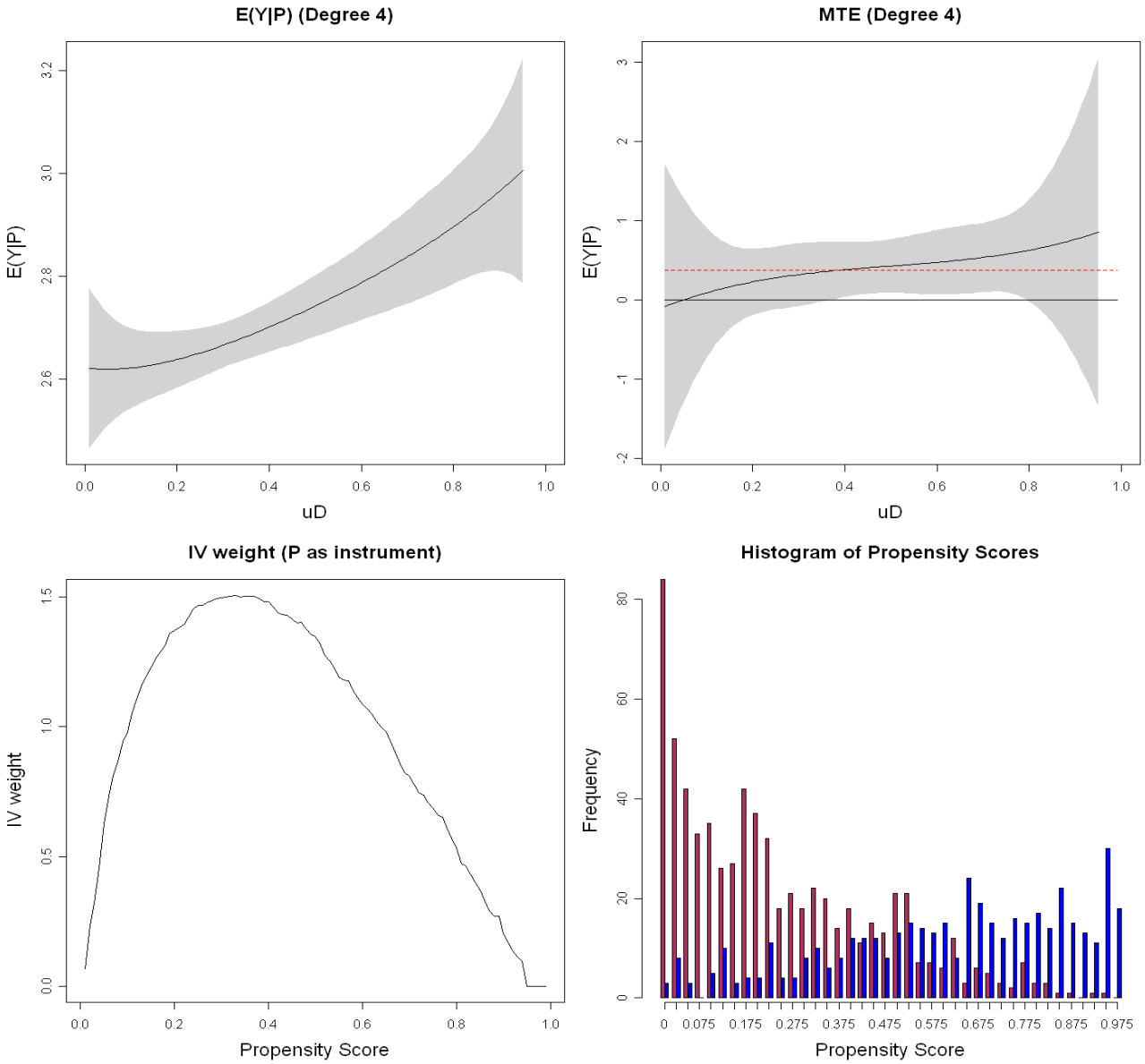
Note: The sample size is 5,000. The power is calculated for each alternative hypothesis (each value of $\rho_{1V}\sigma_1 - \rho_{0V}\sigma_0$) by bootstrapping the Wald statistic 500 times calculating what proportion of the test statistics lie outside the 95th percentile of a χ^2 distribution with 1 degree of freedom (we are testing 1 coefficient). These IV estimates are the coefficient on D and contain no interactions with X (so they are misspecified).

Figure 9: Chile School Vouchers -- Math Score



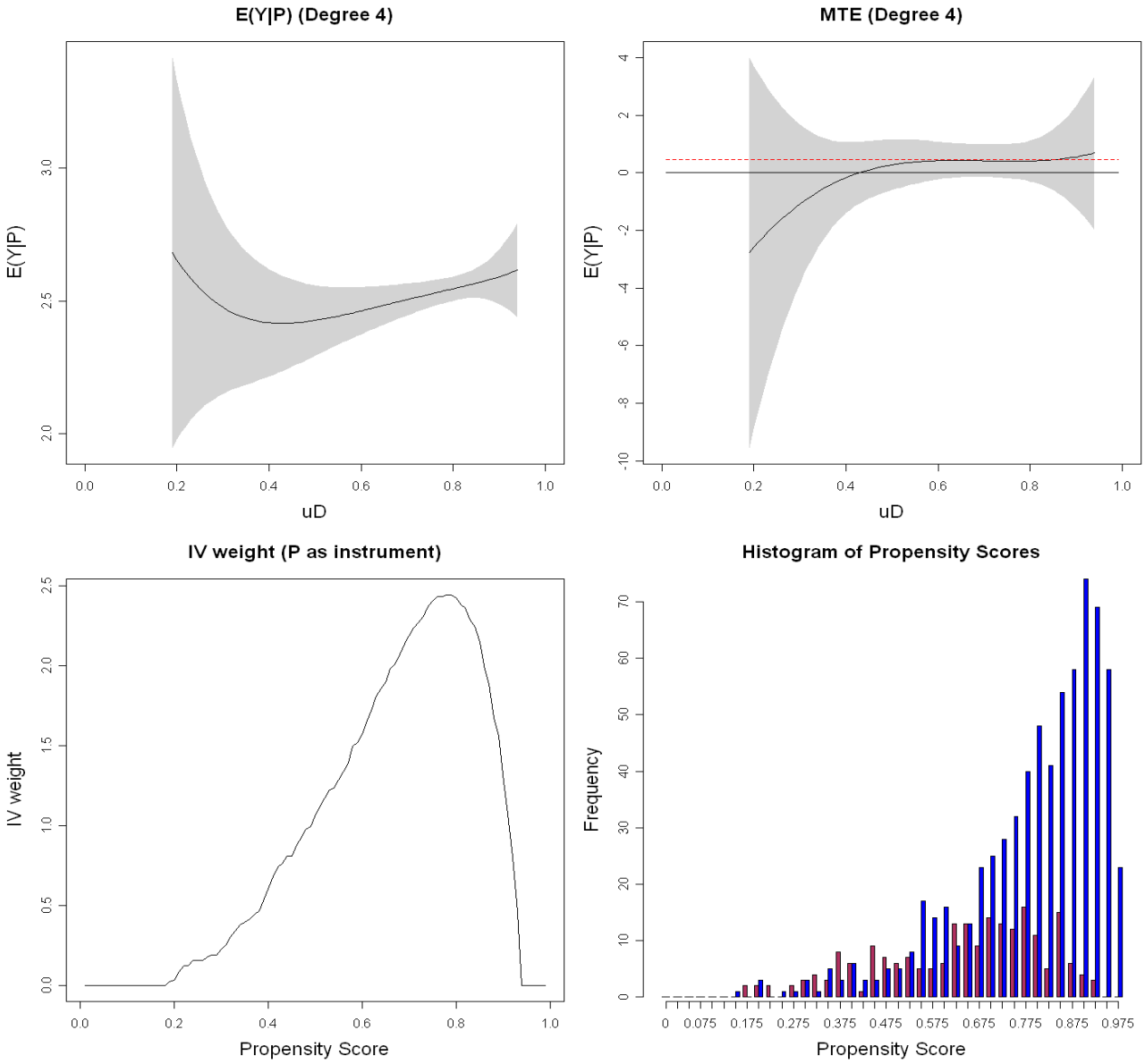
Note: The covariates in the outcome equations are: gender, mother's highest grade completed, father's highest grade completed, number of family members, an indicator for urban residence, household income categories and region indicators. The instruments are: the proportion of schools in one's municipality that were voucher schools in 2002, the difference in average test scores between the voucher schools and the public schools in one's municipality in 2002, in addition to all of the X variables. The dependent variable in the probit is 1 if the individual is enrolled in a voucher school, and 0 if the individual is enrolled in a public school. The $E(Y|P,X)$ curve is found by regressing log hourly wages on the X's, P, P^2, P^3, and P^4. The confidence intervals are found using 100 bootstraps. In the MTE graph, the horizontal red line indicates the IV estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 40,501.

Figure 10: 4-Year College Graduate vs. High School Wages



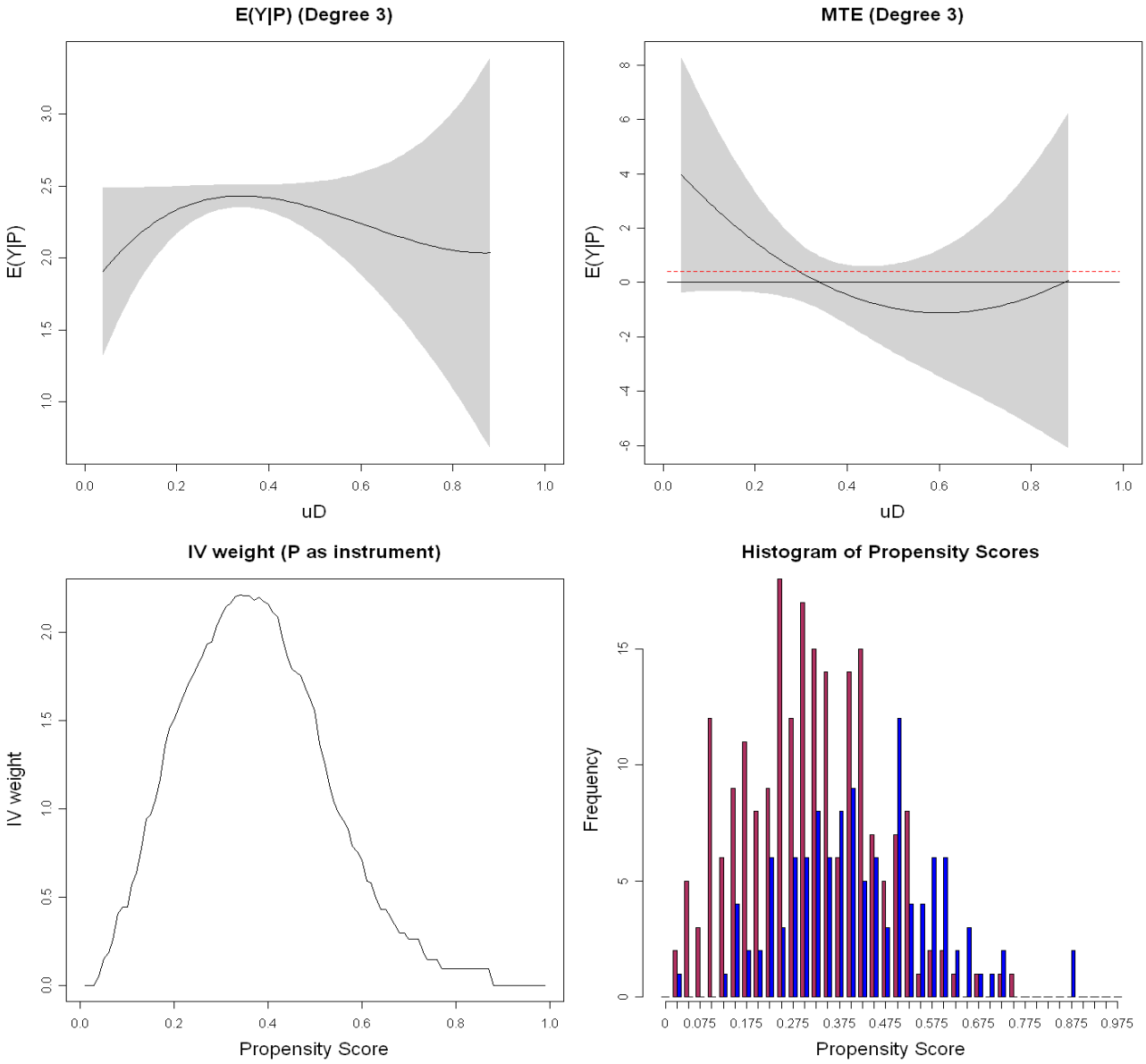
Note: The covariates in the outcome equations are: job tenure, job tenure squared, experience, experience squared, AFQT score, noncognitive score, marital status, indicators for black and hispanic, and year of birth indicators. The instruments are: AFQT score, noncognitive score, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, wages and unemployment rates of local high school graduates, wages and unemployment rates of local some college, wages and unemployment rates of local college graduates, indicators for black and hispanic, indicators for south residence and urban residence at age 14, and year of birth indicators. The dependent variable in the probit is 1 if the individual graduated from a 4-year college, and 0 if the individual's highest education is a high school diploma (GEDs are excluded). The E(Y|P,X) curve is found by regressing log hourly wages on the X's, P, P^2, P^3, and P^4. The confidence intervals are found using 100 bootstraps. In the MTE graph, the horizontal red line indicates the IV estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 1,144.

Figure 11: High School vs. Dropout Wages



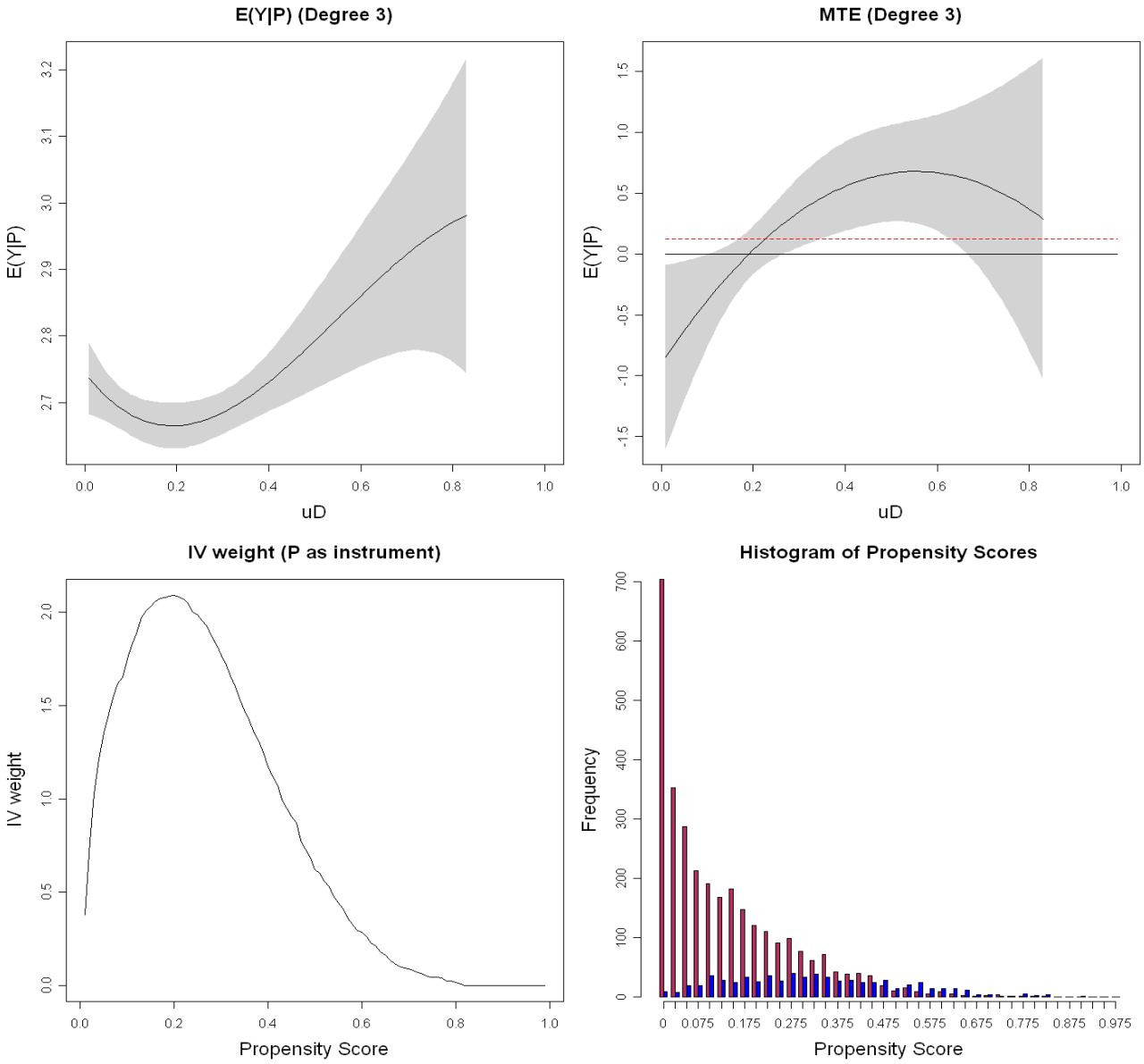
Note: The covariates in the outcome equations are: job tenure, job tenure squared, experience, experience squared, AFQT score, noncognitive score, marital status, indicators for black and hispanic, and year of birth indicators. The instruments are: AFQT score, noncognitive score, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, wages and unemployment rates of local dropouts, wages and unemployment rates of local high school graduates, indicators for black and hispanic, indicators for south residence and urban residence at age 14, and year of birth indicators. The dependent variable in the probit is 1 if the individual's highest education is a high school diploma, and 0 if the individual is a high school dropout (GEDs are excluded). The $E(Y|P,X)$ curve is found by regressing log hourly wages on the X's, P, P^2, P^3, and P^4. The confidence intervals are found using 100 bootstraps. In the MTE graph, the horizontal red line indicates the IV estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 1,144.

Figure 12: GED vs. Dropout Wages



Note: The covariates in the outcome equations are: job tenure, job tenure squared, AFQT score, noncognitive score, marital status, indicators for black and hispanic, and year of birth indicators. The instruments are: AFQT score, noncognitive score, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, local cost of the GED, wages of local dropouts, unemployment rates of local high school graduates, indicators for black and hispanic, indicators for south residence and urban residence at age 14, and year of birth indicators. The dependent variable in the probit is 1 if the individual's highest education is a GED, and 0 if the individual is a high school dropout. The $E(Y|P,X)$ curve is found by regressing log hourly wages on the X's, P, P^2 and P^3. The confidence intervals are found using 100 bootstraps. In the MTE graph, the horizontal red line indicates the IV estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 331.

Figure 13: Union Wages



Note: The covariates in the outcome equations are: experience, experience squared, various education categories, indicators for region of the country, indicator for urban, indicator for white, indicator for weeks worked between 1 and 26, indicator for weeks worked between 48 and 52 weeks. The instruments are: all of the X variables in addition to indicators for two-digit occupation codes. The dependent variable in the probit is 1 if the individual is a union member, and 0 if the individual is not a union member. The $E(Y|P,X)$ curve is found by regressing log hourly wages on the X's, P, P^2 and P^3. The confidence intervals are found using 100 bootstraps. In the MTE graph, the horizontal red line indicates the IV estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 3815.

Table 1: Chile Vouchers -- Math Score

A. P-values from sequentially adding polynomial terms

Degree of Polynomial	2	3	4	5
P	0.0000	0.0000	0.0014	0.0696
P^2	0.3550	0.0890	0.0395	0.3885
P^3		0.1593	0.0767	0.5795
P^4			0.1208	0.7376
P^5				0.8667
Joint test of nonlinear terms	0.3550	0.0875	0.0205	0.8796

B. Treatment Effects

Degree of Polynomial	2	3	4	5	Normal	Semipar.
ATE	-7.2307	-12.6136	-10.9667	-11.7628	-4.7516	-7.7067
TT	-12.9179	-18.6105	-28.4809	-29.4113	-8.5286	-12.924
TUT	-0.7015	-9.5799	5.5403	4.1173	-0.6851	-5.9742
IV	-8.0559	-8.0559	-8.0559	-8.0559	-8.0559	-8.0559
IV (using weights)	-8.3914	-9.0786	-10.9789	-11.2349	-5.4911	-4.2007

Note: The p-values in panel A are from t-tests in the case of the individual coefficients and Wald tests for the joint tests. The standard errors are calculated using 50 bootstrap samples. The treatment effects in panel B are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate is using P(Z), the propensity score, as the instrument. The IV estimate is calculated both using the weights from Heckman and Vytlacil (2005) and using the traditional ratio of covariances. The estimates differ not only because the estimate of the MTE is inexact, but also because the weights are estimated. In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

Table 2: 4-Year College Graduate vs. High School Wages

A. P-values from sequentially adding polynomial terms

Degree of Polynomial	2	3	4	5
P	0.7101	0.7264	0.6813	0.5703
P^2	0.3324	0.7575	0.7448	0.3386
P^3		0.9467	0.8095	0.2804
P^4			0.8172	0.2633
P^5				0.2570
Joint test of nonlinear terms	0.3324	0.6047	0.7989	0.6972

B. Treatment Effects

Degree of Polynomial	2	3	4	5	Normal	Semipar.
ATE	0.4275	0.4182	0.4296	0.5500	0.3369	0.3609
TT	0.2351	0.2210	0.1940	0.3092	0.2617	0.2265
TUT	0.5763	0.5642	0.6028	0.8095	0.4002	0.4633
IV	0.3764	0.3764	0.3764	0.3764	0.3764	0.3764
IV (using weights)	0.3411	0.3421	0.3404	0.3468	0.3008	0.3128

Note: The p-values in panel A are from t-tests in the case of the individual coefficients and Wald tests for the joint tests. The standard errors are calculated using 50 bootstrap samples. The treatment effects in panel B are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate is using P(Z), the propensity score, as the instrument. The IV estimate is calculated both using the weights from Heckman and Vytlacil (2005) and using the traditional ratio of covariances. The estimates differ not only because the estimate of the MTE is inexact, but also because the weights are estimated. In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

Table 3: High School Graduate vs. Dropout Wages

A. P-values from sequentially adding polynomial terms

Degree of Polynomial	2	3	4	5
P	0.9766	0.5533	0.6785	0.8442
P^2	0.3789	0.4515	0.6739	0.8629
P^3		0.5520	0.7251	0.8877
P^4			0.7626	0.9020
P^5				0.9108
Joint test of nonlinear terms	0.3789	0.5188	0.8640	0.9384

B. Treatment Effects

Degree of Polynomial	2	3	4	5	Normal	Semipar.
ATE	-0.0017	-0.4598	-0.9649	-1.5510	0.2420	0.2398
TT	-0.2118	-0.7159	-1.4474	-2.1723	0.0915	0.2411
TUT	0.8954	0.3059	0.7462	0.4588	1.0078	0.1933
IV	0.4641	0.4641	0.4641	0.4641	0.4641	0.4641
IV (using weights)	0.2371	0.2605	0.2237	0.2236	0.3547	0.2636

Note: The p-values in panel A are from t-tests in the case of the individual coefficients and Wald tests for the joint tests. The standard errors are calculated using 50 bootstrap samples. The treatment effects in panel B are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate is using P(Z), the propensity score, as the instrument. The IV estimate is calculated both using the weights from Heckman and Vytlacil (2005) and using the traditional ratio of covariances. The estimates differ not only because the estimate of the MTE is inexact, but also because the weights are estimated. In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

Table 4: GED vs. Dropout Wages

A. P-values from sequentially adding polynomial terms

Degree of Polynomial	2	3	4	5
P	0.4640	0.4123	0.6893	0.9723
P^2	0.6947	0.5529	0.9544	0.8231
P^3		0.6469	0.9052	0.7851
P^4			0.8311	0.7887
P^5				0.8061
Joint test of nonlinear terms	0.6947	0.8205	0.9096	0.9631

B. Treatment Effects

Degree of Polynomial	2	3	4	5	Normal	Semipar.
ATE	-0.5278	0.3563	0.3088	-2.1581	0.0637	-0.0939
TT	0.7162	1.7196	1.7843	0.2074	-0.0239	0.0960
TUT	-1.4192	-0.4096	-0.5354	-4.1226	0.1216	-0.2121
IV	0.3934	0.3934	0.3934	0.3934	0.3934	0.3934
IV (using weights)	0.5152	1.5113	1.5689	0.7140	-0.0089	0.0968

Note: The p-values in panel A are from t-tests in the case of the individual coefficients and Wald tests for the joint tests. The standard errors are calculated using 50 bootstrap samples. The treatment effects in panel B are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate is using P(Z), the propensity score, as the instrument. The IV estimate is calculated both using the weights from Heckman and Vytlacil (2005) and using the traditional ratio of covariances. The estimates differ not only because the estimate of the MTE is inexact, but also because the weights are estimated. In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

Table 5: Union Wages

A. P-values from sequentially adding polynomial terms

Degree of Polynomial	2	3	4	5
P	0.0096	0.1606	0.2803	0.4140
P^2	0.0041	0.0302	0.5305	0.9094
P^3		0.1065	0.8824	0.9146
P^4			0.9412	0.8728
P^5				0.8724
Joint test of nonlinear terms	0.0041	0.0144	0.0294	0.0311

B. Treatment Effects

Degree of Polynomial	2	3	4	5	Normal	Semipar.
ATE	0.6215	0.2437	0.3192	1.8325	0.2149	0.1510
TT	-0.1187	-0.2225	-0.2323	-0.1458	-0.0959	-0.0413
TUT	0.8512	0.3645	0.4683	2.4845	0.3083	0.2153
IV	0.1249	0.1249	0.1249	0.1249	0.1249	0.1249
IV (using weights)	0.0593	0.0498	0.0487	0.0566	0.0031	0.0064

Note: The p-values in panel A are from t-tests in the case of the individual coefficients and Wald tests for the joint tests. The standard errors are calculated using 50 bootstrap samples. The treatment effects in panel B are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate is using P(Z), the propensity score, as the instrument. The IV estimate is calculated both using the weights from Heckman and Vytlacil (2005) and using the traditional ratio of covariances. The estimates differ not only because the estimate of the MTE is inexact, but also because the weights are estimated. In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).