TECHNICAL WORKING PAPER SERIES

IDENTIFICATION AND ESTIMATION OF TRIANGULAR
SIMULTANEOUS EQUATIONS MODELS
WITHOUT ADDITIVITY

Guido W. Imbens
Whitney K. Newey

Identification and Estimation of Triangular Simultaneous
Equations Models Without Additivity
Guido W. Imbens and Whitney K. Newey
NBER Technical Working Paper No. 285
November 2002

ABSTRACT

This paper investigates identification and inference in a nonparametric structural model with instrumental variables and non-additive errors. We allow for non-additive errors because the unobserved heterogeneity in marginal returns that often motivates concerns about endogeneity of choices requires objective functions that are non-additive in observed and unobserved components. We formulate several independence and monotonicity conditions that are sufficient for identification of a number of objects of interest, including the average conditional response, the average structural function, as well as the full structural response function. For inference we propose a two-step series estimator. The first step consists of estimating the conditional distribution of the endogenous regressor given the instrument. In the second step the estimated conditional distribution function is used as a regressor in a nonlinear control function approach. We establish rates of convergence, asymptotic normality, and give a consistent asymptotic variance estimator.

Guido Imbens
Department of Economics
University of California, Berkeley
549 Evans Hall, #3880
Berkeley, CA 94720-3880
and NBER
imbens@econ.berkeley.edu

Whitney K. Newey
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142-1347

1 Introduction

Structural models have long been of great interest to econometricians. Recently interest has focused on nonparametric identification under weak assumptions, in particular without functional form or distributional restrictions in a variety of settings (e.g., Roehrig 1988; Newey and Powell, 1988; Newey, Powell and Vella, 1999; Angrist, Graddy and Imbens, 2000; Darolles, Florens and Renault, 2000; Pinkse, 2000b; Blundell and Powell, 2000; Heckman, 1990; Imbens and Angrist, 1994; Altonji and Ichimura, 1997; Brown and Matzkin, 1996; Vytlacil, 2002; Das, 2000; Altonji and Matzkin, 2001; Athey and Haile, 2002; Bajari and Benkard, 2002; Chernozhukov and Hansen, 2002; Chesher, 2002; Lewbel, 2002). Even when relaxing functional form restrictions, much of the work on nonparametric identification of simultaneous equations models has maintained additive separability of the disturbances and the regression functions.[1] This is an restrictive condition because it rules out interesting economics such as the case where unobserved heterogeneity in marginal returns is the motivation for concerns about endogeneity of choices.

In this paper we focus on identification and estimation triangular simultaneous equations models with instrumental variables. We make two contributions. First, we present three new identification results that do not require additive separability of the disturbances in either the first stage regression or the main outcome equation. For our identification results we consider four assumptions: (i) the instrument and unobserved components are independent; (ii) the relation between the endogenous regressor and the instrument is monotone in the unobserved component; (iii) the instrument has sufficient power to move the endogenous regressor over its entire support; and (iv) the relation between the outcome of interest and the endogenous regressor is monotone in the unobserved component. The first identification result states that given the first and second of these assumptions the average conditional response is identified on the support of the endogenous regressor and the unobserved component. In our second identification result we show that if we also maintain the support condition, then the average structural function (introduced by Blundell and Powell (2001) as a generalization of the average treatment effect in the binary treatment case) is identified. The third identification results states that under the first, second, and fourth assumptions the entire structural relation between the outcome of interest and the endogenous regressor, as well as the joint distribution of the

---

[1] Exceptions include include Angrist, Graddy and Imbens (2000) who discuss conditions under which particular weighted average derivatives of the response functions can be estimated, Altonji and Matzkin (2001) who consider panel models with restrictions on the way the lagged explanatory variables enter the regression function, Das (2001) who uses a single index restriction combined with monotonicity, Chernozhukov and Hansen (2002) who use mainly restrictions on the outcome distributions, and Chesher (2001, 2002) who focuses on local identification (i.e., identification of average derivatives at specific values of the endogenous regressor).

disturbance and the endogenous regressor are identified on their joint support. Together these three identification results allow us to estimate the effect of many policies of interest.

Our second contribution is the development of a framework for estimation of these models. We employ a multi-step approach. The first step estimates the conditional distribution function of the endogenous regressor given the instrument. We evaluate this conditional distribution function at the observed values to obtain a residual that will be used as a generalized control function (e.g., Heckman and Robb, 1984; Newey, Powell and Vella, 1999). In the second step we regress the outcome of interest on the endogenous variable and the first-step residual to obtain what we label the average conditional response. Other estimands that can be written in terms of this average conditional response can then be obtained by by plugging in the estimated average conditional response function. For example, the average structural function is estimated by averaging the average conditional response over the marginal distribution of the first-step residual. We present specific results based on series estimators for the unknown functions, deriving convergence rates for each step of the estimation procedure. We also show asymptotic normality and give a consistent estimator of the asymptotic variance for some of the estimators.

2   The Model

We consider a two-equation triangular simultaneous equations model. The first equation, the "selection equation," relates an endogenous regressor or choice variable to an instrument and an unobserved disturbance:

$$X = h(Z, \eta). \tag{2.1}$$

The second equation, the "outcome equation," relates the primary outcome of interest to the endogenous regressor and an unobserved component:

$$Y = g(X, \varepsilon), \tag{2.2}$$

We are primarily interested in the relation between X and Y, as well as more generally in the effect of policies that change the distribution of X, on the distribution of Y. The unobserved component or disturbance in the first equation, η, is potentially correlated with ε, the unobserved component in the second equation. Thus ε and X are potentially correlated, implying that X is endogenous. The instrument Z is assumed to be independent of the pair of disturbances (η, ε). We assume X and Y are scalars, and allow Z to be a vector, although many of the results in the paper can be generalized to systems of equations. The unobserved component in the selection equation, η, is assumed to be a scalar. The unobserved component in the outcome equation, ε, can be a scalar or a vector. We will consider two special cases. In

the first ε is a scalar, potentially correlated with η. The second case, a generalization of the first has $\varepsilon = (\eta, \nu)$, with ν a scalar independent of η, so that we have

$$Y = g(X, \eta, \nu), \tag{2.3}$$

To see that this generalizes the case with scalar ε, define $\nu = F_{\varepsilon|\eta}(\varepsilon|\eta)$ and $g(X, \eta, \nu) = g(X, F_{\varepsilon|\eta}^{-1}(\nu, \eta))$.

The following two examples illustrates how such triangular systems may arise in economic models:

Example 1: (RETURNS TO EDUCATION)
This example is based on models for educational choices with heterogenous returns such as the one used by Card (2001) and Das (2001). Consider an educational production function, with life-time discounted earnings y a function of the level of education x and ability ε: $y = g(x, \varepsilon)$.

The level of education x is chosen optimally by the individual. Ability is not under the control of the individual, and not observed directly by either the individual or the econometrician. The individual chooses the level of education by maximizing expected life-time discounted earnings minus costs associated with acquiring education given her information set. The information set includes a noisy signal of ability, denoted by η, and a cost shifter z. This signal could be a predictor of ability such as test scores. The cost of obtaining a certain level of education depends on the level of education and on an observed cost shifter z.[2] Hence utility is

$$U(x, z, \varepsilon) = g(x, \varepsilon) - c(x, z),$$

and the utility maximizing level of education is

$$X = \text{argmax}_x \mathbb{E}\Big[U(x, Z, \varepsilon)|\eta, Z\Big] = \text{argmax}_x \Big[\mathbb{E}\Big[g(x, \varepsilon)|\eta, Z\Big] - c(x, Z)\Big],$$

leading to $X = h(Z, \eta)$.

Note the importance, in terms of the economic content of the model, of allowing the earnings function to be non-additive in ability. If the objective function $g(x, \varepsilon)$ were additive in ε, so that $g(x, \varepsilon) = g_0(x) + \varepsilon$, the marginal return to education, $\frac{\partial g}{\partial x}(x, \varepsilon)$, would be independent of ε. Hence the optimal level of education would be $\text{argmax}_x g_0(x) - c(x, Z)$, varying with the instrument but not with ε, so that the level of education would be exogenous. $\square$

Example 2: (PRODUCTION FUNCTION)
The second example is a non-additive extension of a classical problem in the estimation of production functions, e.g., Mundlak (1963). Consider a production function that depends on three

---

[2]Although we do not do so in the present example, we could allow the cost to depend on the signal η, if, for example financial aid was partly tied to test scores.

inputs: $y = g(x, \eta, \nu)$. The first input is observable to both the firm and the econometrician, and is variable in the short run (e.g., labor), denoted by x. The second input is observed only by the firm and is fixed in the short run, denoted by η. We will refer to this as the type of the firm.[3] The third input, ν, is not observed by the econometrician and unknown to the firm at the time the labor input is chosen. Weather conditions could be an example in an agricultural production function.

The level of the input x is chosen optimally by the firm to maximize expected profits. At the time the level of this input is chosen the firm knows the form of its production function, its type, and the value of a cost shifter for the labor input, e.g., an indicator of the cost of labor inputs, denoted by z. The third input ν is unknown at this point, and its distribution does not vary by the level of η. Profits are the difference between revenue (equal to production as the price is normalized to one) and costs, with the latter depending on the level of the input and the observed cost shifter z:[4]

$$\pi(x, z, \eta, \nu) = g(x, \eta, \nu) - c(x, z),$$

so that a profit maximizing firm solves the problem

$$X = \operatorname{argmax}_x \mathbb{E}\left[\pi(x, Z, \eta, \nu)|\eta, Z\right] = \operatorname{argmax}_x \left[\mathbb{E}\left[g(x, \eta, \nu)|\eta\right] - c(x, Z)\right], \qquad (2.4)$$

leading to $X = h(Z, \eta)$. Again, if $g(x, \eta, \nu)$ were additive in the unobserved type η, the optimal level of the input would be the solution to $\max_x \mathbb{E}[g(x, \nu) - c(x, Z)|\eta, Z]$. Because of independence of η and ν the optimal input level would in that case be uncorrelated with (η, ν) and X would be exogenous. $\square$

We are interested in two primitives of the model, the production function and the joint distribution of the input and disturbances, (X, ε, η) as well as in functions of these primitives. In simultaneous equations models researchers often focus solely on identification and estimation of the production function. Especially in the context of linear simultaneous equations models researchers traditionally limit their attention to the derivatives of the output with respect to the endogenous input. Many parameters of interest, however, depend on both the joint distribution of disturbances and endogenous regressors and the production function. To illustrate this point, consider the effect on average output of various interventions or policies that may be contemplated by policy makers. Similar to the binary endogenous regressor case[5] there is a

---

[3] This may in fact be an input that is variable in the long run such as capital or management, although in that case assessing whether the subsequent independence assumptions are satisfied may require modelling how its value was determined.

[4] More generally these costs may also depend on the type of the firm.

[5] See, for example, Heckman and Vytlacil, 2000; Manski, 1997; Angrist and Krueger, 2001; Blundell and Powell, 2001.

variety of such policies. Here we discuss five specific examples of parameters of interest that have either received attention before in the literature, or directly correspond to policies of interest, and demonstrate how these parameters depends on both the production function and the joint distribution of the endogenous regressors and disturbances.

A key role in the identification strategy will be played by the *average conditional response*, (ACR) function, denoted by $\beta(x, \eta)$:

$$\beta(x, \eta) \equiv \mathbb{E}\left[g(x, \varepsilon)|\eta\right] = \int g(x, \varepsilon) F_{\varepsilon|\eta}(d\varepsilon|\eta) \tag{2.5}$$

(Using model (2.1) and (2.3) the definition would be $\beta(x, \eta) \equiv \mathbb{E}\left[g(x, \eta, \nu)|\eta\right] = \int g(x, \eta, \nu) F_\nu(d\nu)$.) This function gives, for agents with type η, the average response to exogenous changes in the value of the endogenous regressor. As a function of x it is therefore causal or structural, but only for the subpopulation of agents with type η. Many of the policy parameters can be expressed conveniently in terms of ths function.

Policy I: FIXING INPUT LEVEL

Blundell and Powell (2000) focus on the identification and estimation of what they label the *average structural function* (ASF), the average of the structural function $g(x, \varepsilon)$ over the marginal distribution of ε.[6] A policy maker may consider fixing the input at a particular level x, say at $x = x_0$ or $x = x_1$. Evaluating the average outcome at these levels of the input requires knowledge of the function

$$\mu(x) = \mathbb{E}[g(x, \varepsilon)] = \int g(x, \varepsilon) F_\varepsilon(d\varepsilon), \tag{2.6}$$

at $x = x_0$ and $x = x_1$. The ASF can also be characterized in terms of the ACR:

$$\mu(x) = \int \int g(x, \varepsilon) F_{\varepsilon|\eta}(d\varepsilon|\eta) F_\eta(d\eta) = \int \beta(x, \eta) F_\eta(d\eta). \tag{2.7}$$

Note that the ASF $\mu(x)$ is *not* equal to the conditional expectation of Y given $X = x$,

$$\mathbb{E}[Y|X = x] = \int g(x, \varepsilon) F_{\varepsilon|X}(d\varepsilon|x),$$

because of the dependence between X and ε. If the production function is linear and additive, that is, $g(x, \varepsilon) = \beta_0 + \beta_1 \cdot x + \varepsilon$, then the average structural function is $\beta_0 + \beta_1 \cdot x$, and so the average effect of fixing the input at x_1 versus x_0 is $\beta_1 \cdot (x_1 - x_0)$. This slope coefficient β_1 is traditionally taken as the parameter of interest in linear simultaneous equations models. □

Policy II: AVERAGE MARGINAL PRODUCTIVITY

---

[6]This is a generalization of the widely studied average treatment effect in the binary treatment case.

A second parameter of interest corresponds to increasing for all units the value of the input by a small amount. The per-unit effect of such a change on average output is the average marginal productivity:

$$\mathbb{E}\left[\frac{\partial g}{\partial x}(X,\varepsilon)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\partial g}{\partial x}(X,\varepsilon)|X,\eta\right]\right] = \mathbb{E}\left[\int \frac{\partial g}{\partial x}(X,\varepsilon)F_{\varepsilon|\eta}(d\varepsilon|\eta)\right] = \mathbb{E}\left[\frac{\partial \beta}{\partial x}(X,\eta)\right], \quad (2.8)$$

where the last equality holds by interchange of differentiation and integration. This average derivative parameter is analogous to the average derivatives studied in Stoker (1986) and Powell, Stock and Stoker (1989) in the context of exogenous regressors. Although policies that would induce agents with heterogenous returns to all increase their input level by the same amount are rare,[7] the average of the marginal productivity (possibly in combination with its variance $\mathbb{V}(\frac{\partial g}{\partial x}(X,\varepsilon))$) can be an attractive way to summarize the distribution of marginal returns in a setting with heterogeneity. As in the case of the ASF, if the production function is linear and additive, that is, $g(x,\varepsilon) = \beta_0 + \beta_1 \cdot x + \varepsilon$, the average marginal return can be expressed directly in terms of the coefficients of the linear model. The marginal effect of a unit increase in x would be β_1, the coefficient on the input. Note that in general this average derivative cannot be inferred from the ASF $\mu(x)$. In particular, it is in general *not* equal to the expected value of the derivative of the ASF,

$$\mathbb{E}\left[\frac{\partial \mu}{\partial x}(X)\right] = \int \frac{\partial \mu}{\partial x}(x)F_X(dx) = \int \int \frac{\partial g}{\partial x}(x,\varepsilon)F_\varepsilon(d\varepsilon)F_X(dx),$$

unless either X and ε are independent (which is not a very interesting case because then X would be exogenous), or $g(x,\varepsilon)$ is additive in ε, which is one of the key assumptions we are attempting to relax. $\square$

Policy III: INPUT LIMIT
A third parameter of interest corresponds to imposing a limit, e.g., a ceiling or a floor, on the value of the input at $\overline{x}$. This changes the optimization problem of the firm in the production function example to

$$X = \text{argmax}_{x\leq\overline{x}}\mathbb{E}\left[\pi(x,Z,\eta,\nu)|\eta,Z\right] = \text{argmax}_{x\leq\bar{x}}\left[\mathbb{E}\left[g(x,\eta,\nu)|\eta\right] - c(x,Z)\right].$$

Those firms who in the absence of this restriction would choose a value for the input that is outside the limit now choose the limit $\bar{x}$ (under some conditions on the production and cost functions), and those firms whose optimal choice is within the limit are not affected by the

---

[7] An example of such a policy in the context of the relation between income and consumption or savings is a tax rebate that is fixed in nominal terms for all individuals.

policy, so that under these conditions $x = \min(h(z, \eta), \bar{x})$. Then the average production under such a policy would be, for $\ell(x) = \min(x, \bar{x})$,

$$\mathbb{E}\left[g(\ell(X), \eta, \nu)\right] = \mathbb{E}\left[\mathbb{E}\left[g(\ell(X), \eta, \nu)|X, \eta\right]\right] = \mathbb{E}\left[\int g(\ell(X), \eta, \nu) F_\nu(d\nu)\right] = \mathbb{E}\left[\beta(\ell(X), \eta)\right].$$
(2.9)

One example of such a policy would arise if the input is causing pollution, and the government is interested in restricting its use. Another example of such a policy is the compulsory schooling age, with the government interested in the effect raising the compulsory schooling age would have on average earnings. Note that even in the context of the standard additive and linear simultaneous equations model, knowledge of the regression coefficients would not be sufficient for the evaluation of such a policy; unless X is exogenous this would also require knowledge of the joint distribution of (X, η). □

Policy IV: INPUT TAX

An alternative policy the government may consider to reduce the use of an input is to impose a tax on its use. Suppose the tax is τ per unit of the input. This changes the profit function from (2.4) to

$$\tilde{\pi}(x, z, \eta, \nu) = g(x, \eta, \nu) - c(x, z) - \tau \cdot x,$$

Note that the original cost function need not be linear in the input if there is nonlinear pricing, for example through quantity discounts. Maximizing the expected profit function, taking into account the tax, amounts to solving

$$X = \operatorname{argmax}_x \left[\beta(x, \eta) - c(x, Z) - \tau \cdot x\right].$$
(2.10)

Let $x = \tilde{h}(z, \eta, \tau)$ be the optimal level of the input given the new tax. We are interested in the average level of the output for a given level of the tax, or more generally in the distribution of output given the tax. The first order condition for the optimal input level in the absence of the tax was

$$\frac{\partial \beta}{\partial x}(x, \eta) = \frac{\partial c}{\partial x}(x, z).$$
(2.11)

Given the ACR $\beta(x, \eta)$, which is estimable on data without the tax under conditions discussed below, we can use equation (2.11) to derive the original cost function $c(x, z)$ up to a constant. Given the marginal cost function and the ACR we can derive the optimal level of the input given the tax, $\tilde{h}(z, \eta, \tau)$, by maximizing the profit function given the tax (2.10). Using the optimal input function we can then derive the new output distribution for a firm of type η and with input x, and, for example, the average output level, as $\mathbb{E}[\beta(\tilde{h}(Z, \eta, \tau), \eta)]$. □

Policy V: QUANTILE STRUCTURAL EFFECTS

Consider the case with ε scalar and $g(x, \varepsilon)$ strictly increasing in ε. A quantile analog of the ASF is the θth quantile of $g(x, \varepsilon)$ over the marginal distribution of ε holding x fixed. This quantile is equal to

$$\pi_Y(x, \theta) = g(x, \pi_\varepsilon(\theta)),$$

where $\pi_\varepsilon(\theta)$ is the θth quantile of the marginal distribution of ε. If we normalize the distribution of ε so that it is $U(0, 1)$, then $\pi_\varepsilon(\theta) = \theta$ and hence $\pi_Y(\theta, x) = g(x, \theta)$. Thus, we can interpret $g(x, \varepsilon)$ as describing how the εth quantile of the outcome varies with the exogenous changes in the endogenous regressor. This quantile effect is also considered by Chernozhukov and Hansen (2002). Under the uniform distribution normalization the ASF is equal to the integral of this quantile function over all quantiles. A similar interpretation is available for $g(x, \eta, \nu)$, as describing how the Y varies with x at the ηth and ν^{th} quantile for η and ν respectively, when both are normalized to have uniform distributions. This function was considered in Imbens and Newey (2001) and a local version of it by Chesher (2001, 2002). Our approach to identification and estimation of $g(x, \eta, \nu)$ differs from Chesher in that we use a control function approach where the first step variable η to control for endogeneity in the second step, whereas Chesher works with the quantile regression of the outcome on the endogenous regressor and the instrument. In a parametric model we would estimate the structural coefficient β from the quantile regression

$$Y = \beta \cdot X + \lambda \cdot \hat{\eta} + \nu,$$

where $\hat{\eta}$ is the first step residual from a quantile regression of X on Z. Chesher's approach would be to estimate $Y = \pi \cdot X + \gamma \cdot Z + \varepsilon$ and then solve for the structural coefficient β from this regression and the first stage regression of X on Z. We note here that the answer to which quantile effect to consider, $g(x, \varepsilon)$ or $g(x, \eta, \nu)$, depends critically on whether there are two structural disturbances or one. When $g(x, \varepsilon)$ is the correct model, $g(x, \eta, \nu)$ will be difficult to interpret, since ν is a function of the two structural errors. $\square$

3 Identification

In this section we present three new identification results. We are interested in restrictions on the outcome function $g(x, \varepsilon)$, the selection function $h(z, \eta)$, and the joint distribution of disturbances and instruments that in combination allow for identification of policy parameters or the outcome function over at least part of the support. Our results complement those in other recent studies of nonparametric identification in the combination of assumptions and estimands. In contrast to Roehrig (1988), Newey and Powell (1988), Newey, Powell and Vella

(1999), Darolles, Florens and Renault (2001) we allow for non-additive models. We make monotonicity assumptions that differ from (and neither imply, nor are implied by) those in Angrist, Graddy and Imbens (2000), allowing us to identify the average conditional response function. Altonji and Matzkin (2001) require panel data to achieve identification. Compared to Chernozhukov and Hansen (2002) we focus more on restrictions on the selection equation than on restrictions on the outcome equation, and exploit those to obtain identification results for the average conditional response as well as the joint distribution of the endogenous regressor and unobserved components. Compared to our assumptions Chesher (2002) imposes weaker independence conditions, but as a result he obtains only identification of the average derivative of the outcome equation at a point.

The first assumption we make is that the instrument is independent of the disturbances.

Assumption 3.1 (INDEPENDENCE) *The disturbances (ε, η) are jointly independent of Z.*

Note that as in, for example, Roehrig (1988) and Imbens and Angrist (1994), full independence is assumed, rather than the weaker mean-independence as in, for example, Newey and Powell (1988), Newey, Powell and Vella (1999) and Darolles, Florens and Renault (2001). Without an additive structure, such a mean-independence assumption is not meaningful. In the two examples in Section 2 this assumption could be plausible if the value of the instrument was chosen at a more aggregate level rather than at the level of the agents themselves. State or county level regulations could serve as such instruments, or natural variation in economic environment conditions, in combination with random location of firms. For the plausibility of the instrument variable assumption it is also important that the relation between the outcome of interest and the regressor is distinct from the objective function that is maximized by the economic agent, as pointed out in Athey and Stern (1998). To make the instrument correlated with the endogenous regressor it should enter the latter, but to make the independence assumption plausible it should not enter the former.

The second assumption requires the structural relation between the endogenous regressor and the instrument to be monotone in the unobserved disturbance.

Assumption 3.2 *(*MONOTONICITY OF ENDOGENOUS REGRESSOR IN THE UNOBSERVED COMPONENT*) The function $h(z, \eta)$ is strictly monotone in its second argument.*

This assumption is trivially satisfied if this relation is additive in instrument and disturbance, but clearly allows for general forms of non-additive relations. Matzkin (1999) considers nonparametric estimation of $h(z, \eta)$ under Assumptions 3.1 and 3.2 in a single equation exogenous regressor framework. Pinkse (2000b) refers to a multivariate version of this as "weak

separability". Das (2001) considers a stochastic version of this assumption to identify parameters in single index models with a single endogenous regressor.

It is interesting to compare this assumption to the monotonicity assumption used in Imbens and Angrist (1994) and Vytlacil (2002) in the binary regressor case. In terms of the current notation, Imbens-Angrist and Vytlacil focus on monotonicity of $h(z, \eta)$ in the observed component, the instrument z, rather than monotonicity in the unobserved component, the disturbance η. With a binary regressor and binary instrument weak monotonicity in z and weak monotonicity in η are in fact equivalent. However, in the multivalued regressor case, e.g., Angrist and Imbens (1995) and Angrist, Graddy and Imbens (2000), the two assumptions are distinct, with neither one implying the other. Assumption 3.2 has only weak testable implications. A slightly weaker form, requiring $h(z, \eta)$ to be monotone, rather than strictly monotone, in η has no testable implications at all. The testable implications for strict monotonicity version arise only when Z and/or X are discrete. With both Z and X continuous, there are no testable implications.

Das (2001) discusses a number of examples where monotonicity of the decision rule is implied by conditions on the economic primitives using monotone comparative statics results (e.g., Milgrom and Shannon, 1994; Athey, 2002). In the same vein, consider the education function example introduced in Section 2, and assume that $g(x, \varepsilon)$ is continuously differentiable. Suppose that (i), the educational production function is strictly increasing in ability ε, (ii) the return to formal education is strictly increasing in ability, so that $\partial g / \partial \varepsilon > 0$ and $\partial^2 g / \partial x \partial \varepsilon > 0$ (this would be implied by a Cobb-Douglas production function), and (iii) the signal η and ability ε are affiliated. Under those conditions the decision rule $h(z, \eta)$ is monotone in the signal η.[8]

Theorem 1: (IDENTIFICATION OF THE AVERAGE CONDITIONAL RESPONSE FUNCTION) *Suppose Assumptions 3.1 and 3.2 hold. Then the ACR $\beta(x, \eta)$ is identified on the joint support of X and η from the joint distribution of (Y, X, Z).*

All of our results are proved in the Appendices.

This result shows that $\beta(x, \eta)$ is identified by first calculating $\eta = F_{X|Z}(X, Z)$, then regressing Y on X and η. The key insight is that conditional on η the endogenous regressor X is independent of ε. This approach is essentially a nonparametric generalization of the control function approach (e.g., Heckman and Robb, 1984; Newey, Powell and Vella, 1999; Blundell and Powell, 2000), with the disturbance η playing the role of a generalized control function.

It is clear that we cannot identify $\beta(x, \eta)$ outside of the support of X and η, as we do not observe any outcomes at those values of x and η. For some of the parameters of interest

---

[8]Of course in this case one may wish to exploit these restrictions on the production function, as in, for example, Matzkin, 1993.

discussed in Section 2, however, it sufficient to know the average conditional response function on its support. For example, the average derivative parameter in (2.8) is equal to the expected value of the derivative of $\beta(x, \eta)$ with respect to x. Whether the parameter of interest in the input limit example can be identified from this result depends on the support of X and η. In the tax input example the impact of the tax can be identified for small changes in the tax parameter, although for larger changes the support of X and η may again prevent point identification. In general the ASF $\mu(x)$ can be identified only under a stronger assumption on the support. What makes the ASF, and the input limit parameter (and also the tax impact for larger values of the tax) more difficult to identify is that these policies require some firms to move away more than infinitesimal amounts from their optimal choices. In contrast, the average derivative parameter, and the tax impact for small values of the tax, require firms to move away from their currently optimal choices only by small amounts and hence it suffices to identify the average conditional response around optimal values.

The following assumption requires the conditional support of X given η to be the same for all values of η.

Assumption 3.3 (Support) *The support of X given η does not depend on the value of η.*

Assumption 3.3 is strong. Given the deterministic relation between Z and X given η, this implies that by changing the value of the instrument, one can induce any value of the endogenous regressor. In the binary endogenous variable case this implies that by changing the value of Z, one can induce both values for the endogenous regressor, similar to the "identification-at-infinity" results in Chamberlain (1986) and Heckman (1990). In the binary case that would immediately imply identification of the average outcome at both values of the endogenous regressor without the monotonicity assumption. In contrast, here the support condition in itself is not sufficient to identify the average structural function at all values of the regressor.

The next identification result is an extension of the results in Blundell and Powell (2000), allowing for a more flexible relation between the endogenous regressor and the instrument. Blundell and Powell (2000) allow for a general non-additively separable function $g(\cdot)$, but assume that $h(\cdot)$ is additive and linear.

Theorem 2: (Identification of the Average Structural Function)
Suppose Assumptions 3.1, 3.2 and 3.3 hold. Then the ASF $\mu(x)$ is identified from the joint distribution of (Y, X, Z).

Given identification of $\beta(x, \eta)$, implied by Theorem 1, identification of the ASF requires that one can integrate over the marginal distribution of η for all values of x. This is feasible

[11]

because of the support condition. Note that it is only in the last step where we average over the distribution of η, that we use the support condition. If the support condition does not hold, we cannot integrate over the marginal distribution of η, at least not at all values of X, because we can only estimate the ACR at values (X, η) with positive density. We may in that case be able to derive bounds on the average structural function if output Y is bounded itself, using the approach by Manski (1990, 1995).

The fourth assumption requires monotonicity of the production function in the second unobserved component.

Assumption 3.4 *(*Monotonicity of the Outcome in the Unobserved Component*)*
(i) The function $g(x, \varepsilon)$ is strictly monotone in its second argument.
(ii) The function $g(x, \eta, \nu)$ is strictly monotone in its third argument.

Again, this assumption is plausible in many economic models. For example, production functions are typically specified to be strictly monotone in all their inputs. Chernozhukov and Hansen (2002) use a similar assumption (without monotonicity of the selection equation) to obtain identification results for the outcome equation alone. The third identification result uses the additional monotonicity assumption to identify, for some values of X and ε, the unit-level structural function in combination with the joint distribution of endogenous regressor and unobserved components.

Theorem 3: (Identification of the structural response and joint distribution of endogenous regressor and unobserved components)
(i) *Suppose for model (2.1) and (2.2) Assumptions 3.1, 3.2, and 3.4(i) hold. Then the joint distribution of (X, η, ε) is identified, up to normalizations on the distributions of η and ε, and $g(x, \varepsilon)$ is identified on the joint support of (X, ε).*
(ii) *Suppose for model (2.1) and (2.3) Assumptions 3.1, 3.2, and 3.4(ii) hold. Then the joint distribution of (X, η, ν) is identified, up to normalizations on the distributions of η and ν, and $g(x, \eta, \nu)$ is identified on the joint support of (X, η, ν).*

As in Theorem 1, for this theorem we do not need a support condition. However, the identification of the production function is again limited to the joint support of the endogenous regressor and the disturbances.

4   Estimation

In this section we consider estimators of the ACR and functionals of it, such as the ASF. We will also discuss estimation of the structural functions $g(x, \varepsilon)$ and $g(x, \eta, \nu)$. In each case we

[12]

employ a multi-step estimator. The first step involves the construction of an estimator $\hat{\eta}_i$ of η_i. This estimator $\hat{\eta}_i$ is used as a control variable for nonparametric estimation in a second step, where Y is regressed on X and $\hat{\eta}$ exploiting the exogeneity of X conditional on η. Here $\hat{\eta}_i$ is the analog for a nonseparable model of the nonparametric regression residual control variate used in Heckman and Robb (1984), Newey, Powell, and Vella (1999) and Blundell and Powell (2000).

Throughout this discussion we will focus on the continuous η case and normalize η_i to be uniformly distributed on $(0,1)$. As shown in the proof of Theorem 1, with this normalization we can take $\eta = F_{X|Z}(X, Z)$. This variable can be estimated by $\hat{\eta}_i = \hat{F}_{X|Z}(X_i, Z_i)$ where $\hat{F}_{X|Z}(x, z)$ is a nonparametric estimator of the conditional CDF. Thus, the control variable we use in estimation is an estimate of the conditional CDF for the endogenous variable given the instrument. There are several ways of constructing $\hat{\eta}_i$. Below we will describe a series estimator. However, before doing so we will first give a general form for the second step of each estimator.

4.1 The ACR and ASF

To estimate the ACR we use the result that under Assumptions 3.1-3.2,

$$\mathbb{E}[Y|X, \eta] = \mathbb{E}[g(X, \varepsilon)|X, \eta] = \int g(X, \varepsilon) F_{\varepsilon|\eta}(d\varepsilon|\eta) = \beta(X, \eta),$$

where the second equality follows by independence of X and ε conditional on η. Thus, the ACR is equal to the conditional expectation of the outcome variable Y given X and the control variable η. It can be estimated by a nonparametric regression of Y on X and a nonparametric estimator $\hat{\eta}$,

$$\hat{\beta}(x, \eta) = \hat{\mathbb{E}}[Y|X, \eta].$$

The use of $\hat{\eta}$ rather than η in this nonparametric regression will not affect the consistency of the estimator, although it will affect the asymptotic distribution.

As we have discussed, a number of policy parameters are functionals of the ACR. Here we will give a brief description of corresponding estimators of these parameters. Under Assumptions 3.1 - 3.3 the ASF, average derivative, and input limit response satisfy equations (2.7), (2.8), and (2.9) respectively. We propose estimating them by

$$\hat{\mu}(x) = \int_0^1 \hat{\beta}(x, \eta) d\eta,$$

$$\mathbb{E}\left[\widehat{\frac{\partial g}{\partial x}(X, \varepsilon)}\right] = \frac{1}{n} \sum_{i=1}^n \frac{\partial \hat{\beta}}{\partial x}(x_i, \hat{\eta}_i),$$

$$\mathbb{E}[g(\widehat{\ell(X)}, \varepsilon)] = \frac{1}{n} \sum_{i=1}^n \hat{\beta}(\ell(x_i), \hat{\eta}_i).$$

[13]

Note that for the ASF we integrate the ACR over the (known) marginal distribution of η. For the other estimators we average over the estimated joint distribution of X and η.

For the series estimator we discuss below it is straightforward to calculate the integral in the ASF estimator as well as the sample averages for the other estimators. The ASF estimator has a partial mean form (Newey, 1994), as does the input limit response, so that they should have faster convergence rates than the ACR estimator $\hat{\beta}(x,\eta)$. This conjecture is shown below for a series estimator of the ASF. As in Powell, Stock, and Stoker (1989), we expect the average derivative estimator to be $\sqrt{n}$-consistent under appropriate conditions, which will include the density of x going to zero at the boundary of its support.

4.2   Estimating the Structural Functions

Here we will give a brief description of how the structural response functions $g(x,\varepsilon)$ and $g(x,\eta,v)$ can be estimated. Estimation of $g(x,\varepsilon)$ can be based on averaging over η as in the ASF. Let $F_{Y|X,\eta}(y,x,\eta) = \Pr(Y \leq y | X = x, \eta)$ denote the conditional distribution function of Y given X and η and $G(y,x) = \int_0^1 F_{Y|X,\eta}(y,x,\eta)d\eta$ be its integral over the (uniform) marginal distribution of η. Note that $Y \leq y$ if and only if $\varepsilon \leq g^{-1}(y,X)$. Then normalizing the marginal distribution of ε to be uniform on $(0,1)$ we have

$$
\begin{aligned}
g^{-1}(y,x) &= \Pr(\varepsilon \leq g^{-1}(y,x)) = \int_0^1 \Pr(\varepsilon \leq g^{-1}(y,x)|\eta)d\eta \\
&= \int_0^1 \Pr(\varepsilon \leq g^{-1}(y,x)|X=x,\eta)d\eta \\
&= \int_0^1 \Pr(g(x,\varepsilon) \leq y|X=x,\eta)d\eta = \int_0^1 Pr(Y \leq y|X=x,\eta)d\eta = G(y,x),
\end{aligned}
$$

where the third equality follows by conditional independence of X and ε given η. Inverting this relationship gives

$$
g(x,\varepsilon) = G^{-1}(\varepsilon,x).
$$

Thus we see that the structural function is the inverse of the integral over η of the conditional CDF of Y given X and η. An estimator can be obtained by plugging into this formula a nonparametric estimator $\hat{F}_{Y|X,\eta}(y,x,\eta)$ of the conditional CDF $F_{Y|X,\eta}(y,x,\eta)$ using Y_i, X_i, and $\hat{\eta}_i$, leading to

$$
\hat{g}(x,\varepsilon) = \hat{G}^{-1}(\varepsilon,x),
$$

where

$$
\hat{G}(y,x) = \int_0^1 \hat{F}_{Y|X,\eta}(y,x,\eta)d\eta.
$$

Like the ASF, this estimator is obtained by integrating over the control variate.

[14]

The function $g(x, \eta, \nu)$ can estimated using a conditional CDF approach similar to that for $g(x, \varepsilon)$, without integrating out η. To do this we normalize the distribution of ν to be uniform on $(0, 1)$. As before let $F_{Y|X,\eta}(y, x, \eta) = \Pr(Y \leq y | X = x, \eta)$ denote the conditional distribution function of Y given $X = x$ and η. Note that $Y \leq y$ if and only if $\nu \leq g^{-1}(y, X, \eta)$. Then the following equation is satisfied:

$$
\begin{aligned}
g^{-1}(y, x, \eta) &= \Pr(\nu \leq g^{-1}(y, x, \eta)) = \Pr(\nu \leq g^{-1}(y, x, \eta) | X = x, \eta) \\
&= \Pr(Y \leq y | X = x, \eta) = F_{Y|X,\eta}(y, x, \eta).
\end{aligned}
$$

where the third equality follows by independence of ν and (x, η). Inverting gives

$$
g(x, \eta, \nu) = F_{Y|X,\eta}^{-1}(\nu, x, \eta).
$$

Thus, $g(x, \eta, \nu)$ is the νth quantile of the conditional distribution of y given (x, η). This function can be estimated by plugging in a consistent estimator of F from nonparametric regression on x_i and $\hat{\eta}_i$ into this formula, giving

$$
\hat{g}(x, \eta, \nu) = \hat{F}_{Y|X,\eta}^{-1}(\nu | x, \eta).
$$

Of course, any other nonparametric estimator of the νth conditional quantile of Y given x and η, estimated from the observations Y_i, x_i, and $\hat{\eta}_i$, will also do.

4.3 Series Estimation

In order to operationalize the estimators we need to be specific about the form of nonparametric estimation carried out in each step. Here we will consider series estimators, although alternatives (such as kernel estimators) could be used. We focus on series estimators because of their computational convenience.

To describe the first step estimation of η_i let $q_{\ell L}(z)$, $(\ell = 1, ..., L; L = 1, 2, ...)$ denote approximating functions for the first step. Examples include power series or spline functions. Also, let $q^L(z) = (q_{1L}(z), ..., q_{LL}(z))'$ and $\hat{Q} = \sum_{i=1}^{n} q^L(z_i) q^L(z_i)'/n$. A series estimator of the conditional CDF at a particular x and z can be obtained as the predicted value from regressing an indicator function for $x_i \leq x$ on functions of z_i. It has the form

$$
\tilde{\eta} = \tilde{F}(x|z) = q^L(z)' \hat{Q}^- \sum_{j=1}^{n} q^L(z_j) 1(x_j \leq x)/n,
$$

where A^- denotes any generalized inverse of the matrix A. As is well known, the predicted values $\tilde{F}(x_i|z_i)$ will be invariant to the choice of generalized inverse, which is important here because we will allow for $\hat{Q}$ to be singular, even asymptotically.

[15]

One feature of this estimator $\tilde{\eta}$ is that it is not necessarily bounded between 0 and 1. We impose that restriction by fixed trimming. Let $\tau(\eta) = 1(\eta > 0)\min\{\eta, 1\}$ be the CDF of a uniform distribution. Then our estimate of the control function is given by

$$\hat{\eta}_i = \tau(\tilde{\eta}_i) = \tau(\tilde{F}(x_i|z_i)).$$

To describe the ACR estimator, let $w = (x, \eta)$ denote the entire vector of regressors in $\mathbb{E}[y|x, \eta]$. Let $p_{kK}(w)$, $(k = 1, ..., K; K = 1, 2, ...)$, be approximating functions of w, $p^K(w) = (p_{1K}(w), ..., p_{KK}(w))'$, $\hat{w}_i = (x_i, \hat{\eta}_i)$, and $\hat{P} = \sum_{i=1}^{n} p^K(\hat{w}_i)p^K(\hat{w}_i)'/n$. A nonparametric estimator of the ACR $\beta(w) = \mathbb{E}[y|w]$ is then

$$\hat{\beta}(w) = p^K(w)'\hat{\gamma},$$

where

$$\hat{\gamma} = \hat{P}^{-1} \sum_{j=1}^{n} p^K(\hat{w}_j)y_j/n.$$

This estimator can be used as described above to estimate the ASF, average derivative, or input limit response. It could also be used to estimate any other functional of the ACR.

An estimator of $F_{Y|X,\eta}(y, x, \eta)$ is needed for estimation of the response functions $g(x, \varepsilon)$ or $g(x, \eta, \nu)$. We could construct such an estimator by regressing the indicator function $1(Y \leq y)$ on $p^K(w)$. Although this estimator will be a step function as a function of y, as will the integral $\hat{G}(y, x)$ over ν, one can still work with a corresponding empirical quantile function, consisting of an appropriately defined inverse. It may be possible to use results similar to those of Doss and Gill (1992) to obtain theory for such estimators.

5 Large Sample Theory

We derive convergence rates and asymptotic normality results for the estimators. First we obtain convergence rates for the estimator of the first stage residual η. Second, we derive convergence rates for the average conditional response $\beta(x, \eta)$. Then we consider rates for functionals of the ACR. For brevity we focus on convergence rates for the ASF. Finally we prove asymptotic normality for the estimator of the ASF, and show that the variance can be estimated consistently for use in confidence intervals. Similar results, including asymptotic normality, could be obtained for other policy parameter estimators as well as for estimators of the structural functions.

5.1 Convergence Rates

To derive large sample properties of the estimator it is essential to impose some conditions. The first assumption imposes an approximation rate for the first step regression that is uniform

in both the arguments x and z of the conditional distribution function $F(x|z)$. Let $\mathcal{X}$ and $\mathcal{Z}$ denote the support of X_i and Z_i, respectively.

Assumption 5.1: *There exists $d_1, C > 0$ such that for every L there is a $L \times 1$ vector $\gamma^L(x)$ satisfying*

$$\sup_{x \in \mathcal{X}, z \in \mathcal{Z}} |F(x|z) - q^L(z)'\gamma^L(x)| \leq CL^{-d_1}.$$

This condition imposes an approximation rate for the CDF that is uniform in both its arguments. It is well known that such rates exist when higher order derivatives are bounded uniformly in x and the support of z is compact. In particular, it will be satisfied for both splines and power series with $d_1 = s_F/r_z$, if $F(x|z)$ has continuous derivatives up to order s_F, r_z is the dimension of z, and the spline order is at least s_F; see Schumaker (1981) or Lorentz (1986).

The following result gives a convergence rate for the first step:

Theorem 4: *If Assumption 5.1 is satisfied,*

$$\mathbb{E}\left[\sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2/n\right] = O(L/n + L^{1-2d_1}).$$

The two terms in rate result are variance (L/n) and bias (L^{1-2d_1}) terms respectively. In comparison with previous results for series estimators, this convergence result has L^{1-2d_1} in the rate rather than L^{-2d_1}. The "extra" L arises from the predicted values $\hat{\eta}_i$ being based on regressions with the dependent variables varying over the observations.

The following assumption is a normalization that is similar to that adopted by Newey (1997) and Newey, Powell, and Vella (1999). It is a joint restriction on the approximating functions and the distribution of x_i and η_i. Let $\mathcal{W}$ denote the support of $w_i = (X_i, \eta_i)$ and $\lambda_{\min}(A)$ denote the smallest eigenvalue of a symmetric matrix A.

Assumption 5.2: *There is a constant C and $\zeta(K)$, $\zeta_1(K)$ such that $\zeta(K) \leq C\zeta_1(K)$ and for each K there exists B such that $\tilde{p}^K(w) = Bp^K(w)$, $\lambda_{\min}(\mathbb{E}[\tilde{p}^K(w)\tilde{p}^K(w)']) \geq C$, $\sup_{w \in \mathcal{W}} \|\tilde{p}^K(w)\| \leq C\zeta(K)$, and $\sup_{w \in \mathcal{W}} \|\partial\tilde{p}^K(w)/\partial\eta\| \leq C\zeta_1(K)$.*

The size of the bounds $\zeta(K)$ and $\zeta_1(K)$ are known for some important cases. For example, if the joint density of w_i is bounded below and above on a rectangle then this condition will be satisfied for splines and power series with

$$\begin{aligned}
\zeta(K) &= \sqrt{K}, \zeta_1(K) = K^{3/2}; \text{ splines.} \\
\zeta(K) &= K, \zeta_1(K) = K^3; \text{ power series.}
\end{aligned}$$

To obtain a convergence rate, it is also important to specify a rate of approximation for $\beta(w)$. Such a rate is imposed in the following condition:

Assumption 5.3: $\beta(w)$ *is Lipschitz in η and there exists $d, C > 0$ such that for every K there is a α^K with*

$$\sup_{w \in \mathcal{W}} |\beta(w) - p^K(w)'\alpha^K| \leq CK^{-d}.$$

It is well known that this condition holds for polynomials and splines, where $\mathcal{W}$ is a compact rectangle and d is the ratio of number of continuous derivatives that exist to the dimension of w. In addition to these assumptions we also require the following variance condition, which is common in the series estimation literature;

Assumption 5.4: $Var(Y|X, Z)$ *is bounded.*

With these conditions in place we can obtain a convergence rate for the second-step estimator.

Theorem 5: *If Assumptions 5.1 - 5.4 are satisfied and $K\zeta_1(K)^2(L/n + L^{1-2d_1}) \to 0$ then*

$$\int \left[\hat{\beta}(w) - \beta(w)\right]^2 dF(w) = O_p(K/n + K^{-2d} + L/n + L^{1-2d_1})$$

$$\sup_{w \in W} |\hat{\beta}(w) - \beta(w)| = O_p(\zeta(K)[K/n + K^{-2d} + L/n + L^{1-2d_1}]^{1/2}).$$

This result gives both mean-square and uniform convergence rates. It is interesting to note that the mean-square rate is the sum of the first step convergence rate and the rate that would obtain for the second step if the first step was known. This result is similar to that of Newey, Powell, and Vella (1999), and results from inclusion of the first step dependent variable in the second step regression. Also, the first step and second step rates are each the sum of a variance term and a squared bias term.

To show an improved rate for the ASF estimator we assume a particular structure for $p^K(w)$, namely that for each K there is K_x, $p^{K_x}(x)$, K_η, and $p^{K_\eta}(\eta)$ such that

$$p^K(w) = p^{K_x}(x) \otimes p^{K_\eta}(\eta). \tag{5.1}$$

This structure implies restrictions on the values that K can take, namely it can only be equal to the product of integers. We ignore those restrictions in what follows. We also impose the following condition:

Assumption 5.5: *For all K there is c such that $c'p^{K_\eta}(\eta) \equiv 1$ and the constant matrix B in Assumption 5.2 can be chosen to have a Kronecker product form $B = B_x \otimes B_\eta$ such that for all K, $\lambda_{\min}(\int B_\eta p^K(\eta)p^K(\eta)'B_\eta' d\eta) \geq C$ and $\lambda_{\min}(\mathbb{E}[B_x p^{K_x}(x)p^{K_x}(x)'B_x']) \geq C$.*

Theorem 6: *If Assumptions 5.1 - 5.5 are satisfied, $K\zeta_1(K)^2(L/n + L^{1-2d_1}) \to 0$, and K_x/K_η is bounded and bounded away from zero then*

$$\int [\hat{\mu}(x) - \mu(x)]^2 F_X(dx) = O_p(K_x/n + K_x^{-4d} + L/n + L^{1-2d_1}).$$

In this result we see that the second step convergence rate is different, with the variance term being K_x/n rather than K/n, and the bias being K_x^{-4d}. These are exactly the terms that would be obtained in the rate of convergence for a series regression on only $p^{K_x}(x)$. Thus, the partial mean (i.e. integral) form of $\hat{\mu}(x)$ leads to the convergence rate for nonparametric regression just on x, as also occurs for kernel estimators (Newey, 1994).

5.2   Asymptotic Normality

We give conditions for asymptotic normality of linear functionals of the ACR, including the ASF. The general form of the estimand we consider is

$$\theta_0 = a(\beta_0),$$

where $a(\beta)$ is a linear mapping from functions of w to the real number line and the 0 subscript denotes true values. The ASF takes this form with $a(\beta) = \int_0^1 \beta(x, \eta)d\eta$. We restrict attention to linear functionals to keep the analysis relatively simple. We could extend the results to nonlinear functionals using an approach like that of Newey (1997).

An estimator $\hat{\theta}$ can be obtained by plugging in $\hat{\beta}$ in place of β_0, giving $\hat{\theta} = a(\hat{\beta})$. An asymptotic standard error, as needed for large sample confidence intervals, can be obtained by applying a formula for a second step least squares estimator, accounting for the presence of $\hat{\eta}_i$. Let $A = (a(p_{1K}), ..., a(p_{KK}))$. By linearity of $a(\beta)$, we have $\hat{\theta} = A\hat{\alpha}$. Thus, the functional estimator is a linear combination of second-step least squares coefficients, and standard errors can be computed accordingly. Let $\hat{p}_i = p^K(\hat{w}_i)$, $q_i = q^L(z_i)$, $\hat{u}_i = y_i - \hat{\beta}(\hat{w}_i)$, and

$$\hat{\Sigma} = \sum_{i=1}^{n} \frac{\hat{p}_i \hat{p}_i' \hat{u}_i^2}{n}, \qquad \hat{v}_{ji} = 1(x_i \leq x_j) - \tilde{F}(x_j|z_i),$$

$$\hat{\Sigma}_1 = \sum_{i=1}^{n} \hat{m}_i \hat{m}_i'/n, \qquad \hat{m}_i = \sum_{j=1}^{n} [\partial\hat{\beta}(\hat{w}_j)/\partial\eta]\hat{p}_j q_j' \hat{Q}^- q_i \hat{v}_{ji}/n.$$

An asymptotic variance estimator for $\sqrt{n}(\hat{\theta} - \theta_0)$ is then given by

$$\hat{V} = A\hat{P}^{-1}(\hat{\Sigma} + \hat{\Sigma}_1)\hat{P}^{-1}A'. \tag{5.2}$$

[19]

The $\hat{\Sigma}_1$ term corrects for the presence of the first step nonparametric estimators. It raises the estimated asymptotic variance because the first step is uncorrelated with the second step (see Newey and McFadden, 1994, Section 6). It takes a V-statistic projection form that is more complicated than the correction in Newey, Powell, and Vella(1999) because the left-hand side variable in the series regression, which is $1(x_j \leq x_i)$, varies across observations.

For asymptotic normality it is useful to use smooth trimming of the first step. Let ξ_n be a small positive number and $t_n(\eta) = (\eta + \xi_n)^2/4\xi_n$. In this section we assume that the control variable takes the form $\hat{\eta}_i = \tau_n(\tilde{\eta})$, where

$$
\tau_n(\eta) = \begin{cases}
1, & \eta > 1 + \xi_n, \\
1 - t_n(1 - \eta) & , \quad 1 - \xi_n < \eta \leq 1 + \xi_n, \\
\eta, & \xi_n \leq \eta \leq 1 - \xi_n, \\
t_n(\eta), & -\xi_n \leq \eta < \xi_n, \\
0, & \eta < -\xi_n.
\end{cases} \quad .
$$

This modification allows us to carry out expansions that lead to asymptotic normality.

Some additional conditions are important for the asymptotic normality results. The first condition restricts conditional moments of Y similarly to Newey (1997).

Assumption 5.6: $\mathbb{E}[|Y - \beta_0(w)|^4|X, Z]$ *is bounded and* $Var(Y|X, Z)$ *is bounded away from zero.*

It is also useful to impose a condition on the first stage approximating functions that is similar to Assumption 5.2.

Assumption 5.7: *There is a constant C and $\zeta(L)$, such that for each L there exists B such that $\tilde{q}^L(Z) = Bq^L(Z)$ satisfies $\lambda_{\min}(\mathbb{E}[\tilde{q}^L(Z)\tilde{q}^L(Z)']) \geq C$, $\sup_{w \in \mathcal{W}} \|\tilde{q}^L(Z)\| \leq C\zeta(L)$.*

The following condition is also useful.

Assumption 5.8: $\beta_0(w)$ *is twice continuously differentiable in w with bounded first and second derivatives, there is a constant C such $|a(\beta)| \leq C \sup_{w \in \mathcal{W}} |\beta(w)|$ and either i) there is $\delta(w)$ and $\tilde{\alpha}^K$ such that $\mathbb{E}[\delta(w)^2] < \infty$, $a(p_{kK}(\cdot)) = \mathbb{E}[\delta(w)p_{kK}(w)]$, $a(\beta_0(\cdot)) = \mathbb{E}[\delta(w)\beta_0(w)]$, and $\mathbb{E}[\{\delta(w) - p^K(w)'\tilde{\alpha}^K\}^2] \to 0$; or ii) for some α^K, $\mathbb{E}[\{p^K(w)'\tilde{\alpha}^K\}^2] \to 0$ and $a(p^K(\cdot)'\tilde{\alpha}^K)$ is bounded away from zero as $K \to \infty$.*

When condition i) of Assumption 5.8 is satisfied $\hat{\theta}$ will be $\sqrt{n}$-consistent and when condition ii) is satisfied it will not. The following growth rate conditions are also imposed.

Assumption 5.9: *There is a constant C such that $C^{-1}(L/n + L^{1-2d_1}) \leq \xi_n^3 \leq C(L/n + L^{1-2d_1})$. Also, each of the following converge to zero: nL^{1-2d_1}, nK^{-2d}, $K\zeta_1(K)^2L^2/n$, $\zeta(K)^6L^4/n$, $\zeta(K)^4\zeta(L)^4L/n$.*

[20]

For splines these conditions will require that K^4L^2/n and K^3L^4/n each converge to zero. This will hold if both K and L grow slower than $n^{1/7}$. A K and L satisfying this assumption will exist if $d_1 \geq 4$ and $d \geq 4$.

To state the asymptotic normality result we need to be specific about the form of the asymptotic variance. Let $p_i = p^K(w_i)$, $P = \mathbb{E}[p_i p_i']$, $q_i = q^L(z_i)$, $Q = \mathbb{E}[q_i q_i']$, $u_i = y_i - \beta_0(w_i)$, and

$$
\begin{aligned}
\Sigma &= \mathbb{E}[p_i p_i' u_i^2], v_{ji} = 1(x_i \leq x_j) - F(x_j|z_i), \\
\Sigma_1 &= \mathbb{E}[m_i m_i'], m_i = \mathbb{E}[\tau_n'(\eta_j)\{\partial\beta(w_j)/\partial\eta\}p_j q_j' Q^{-1} q_i v_{ji}|y_i, x_i, z_i], \\
V &= AP^{-1}(\Sigma + \Sigma_1)P^{-1}A
\end{aligned}
$$

Theorem 7: *If Assumptions 5.1 - 5.9 are satisfied then $\sqrt{n}(\hat{\theta} - \theta_0)/\sqrt{V} \xrightarrow{d} N(0,1)$.*

We can also obtain a result for the asymptotic variance estimator that allows us to do inference concerning θ_0, with the following condition holding.

Assumption 5.10: *There exists $\bar{d}$ and $\bar{\alpha}^K$ such that for each component w_j of w,*

$$
\sup_{w \in W} |\beta_0(w) - p^K(w)'\bar{\alpha}^K| = O(K^{-\bar{d}}), \sup_{w \in W} |\partial[\beta_0(w) - p^K(w)'\bar{\alpha}^K]/\partial w_j| = O(K^{-\bar{d}}).
$$

Also, $\zeta_1(K)^2 L K^{-2\bar{d}} \to 0$.

Theorem 8: *If Assumptions 5.1 - 5.10 are satisfied then $\hat{V}/V \xrightarrow{p} 1$.*

It follows from Theorems 7 and 8 and the Slutzky theorem that

$$
\sqrt{n}(\hat{\theta} - \theta_0)/\sqrt{\hat{V}} \xrightarrow{d} N(0,1).
$$

so that confidence intervals and test statistics can be formed from $\hat{\theta}$ and $\hat{V}$ in the usual way.

6   A Monte Carlo Example

To begin to investigate the small sample properties of these estimators we carried out a small Monte Carlo study. The model was

$$
Y = \exp(X + \varepsilon), X = \eta Z^{1-\eta}, \varepsilon = (\eta + \nu)/2,
$$

where $Z, \eta,$ and ν are mutually independent, each with a $U(0,1)$ distribution. We used power series estimates in both the first and second stages. We considered two different sample sizes,

$n = 100$ and $n = 400$. The number of replications was 250. We considered two different estimators of the ASF. The first was a linear instrumental variables (IV) estimator with right-hand side variables $(1, X)$ and instruments $(1, Z)$. The second was the series estimator we considered above with power series in both stages. The first stage used regressors z^j, with $j \leq 2$ for $n = 100$ and $j \leq 5$ for $n = 400$. The second stage used regressors $(1, x, \nu)$ for $n = 100$ and $(1, x, \nu, x^2, \nu^2, x\nu)$ for $n = 400$.

Figure 1 reports the results in graphs, one for each sample size and estimator. The figures plot the median of the $\hat{\mu}(x)$ as well as the upper and lower .05 quantiles for each x. We find that for $n = 100$, both estimators are quite biased. For $n = 400$ the bias of IV persists but the bias of the nonparametric estimator is largely eliminated, except for the upper range of x. The variance of our nonparametric estimator is substantially large than that of IV estimator, as a result of including nonlinear term in x and v. As a result of both bias and variance effects the true value of the ASF lies well inside the quantile range for the series estimator but outside the quantile range for the IV estimator for most values of x.

7 Conclusion

In this paper we presented several identification results for a triangular simultaneous equations model without additivity. Relaxing additivity assumption is important because such assumptions rarely follow from economic theory. Moreover, economic theory often implies that unless models are non-additive in unobserved components, regressors will be exogenous. Exploiting these identification results we develop estimators for the effects of policies of interest and for the underlying structural functions themselves. We derive convergence rates and show asymptotic normality and consistency of an asymptotic variance estiamtor.

A Proofs of Identification and Consistency

Proof of Theorem 1: We normalize the marginal distribution of η so that $Pr(\eta \leq c) = c$ for all c in the support of η. For continuous η this means normalization to a uniform distribution on the interval $[0, 1]$. Then, using the fact that $h(z, \eta)$ is one to one:

$$
\begin{aligned}
F_{X|Z}(x_0|z_0) &= Pr(X \leq x_0|Z = z_0) = Pr(h(Z, \eta) \leq x_0|Z = z_0) = Pr(\eta \leq h^{-1}(Z, x_0)|Z = z_0) \\
&= Pr(\eta \leq h^{-1}(z_0, x_0)|Z) = F_\eta(h^{-1}(z_0, x_0)) = h^{-1}(z_0, x_0).
\end{aligned}
$$

Since the conditional distribution function of X given Z is identified, so is $h^{-1}(z, x)$, and hence the function $h(x, \eta)$ itself. As a by-product we get the value of $\eta = h^{-1}(Z, X) = F_{X|Z}(X|Z)$

Since $(\eta, \varepsilon) \perp Z$, we have

$$\varepsilon \perp Z \mid \eta \implies \varepsilon \perp h(Z, \eta) \mid \eta \implies \varepsilon \perp X \mid \eta,$$

Hence

$$\beta(x, \eta) = \mathbb{E}[g(x, \varepsilon)|\eta] = \mathbb{E}[g(x, \varepsilon)|X = x, \eta] = \mathbb{E}[g(X, \varepsilon)|X = x, \eta] = \mathbb{E}[Y|X = x, \eta]$$

$$= \mathbb{E}[Y|X = x, F_{X|Z}(X|Z) = \eta],$$

which is identified from the joint distribution of (Y, X, Z). Q.E.D.

Proof of Theorem 2: Let $\mathcal{X}$ denote the support of X. By Theorem 1 $\beta(x, \eta)$ is identified on the support of X, which equals $\mathcal{X} \times [0, 1]$ by Assumptoin 3.3. Consequently, so is

$$\int_0^1 \beta(x, \eta)d\eta = \int_0^1 \int g(x, \varepsilon)F_{\varepsilon|\eta}(d\varepsilon|\eta)d\eta = \mu(x).$$

If η is discrete with support S_η, then $\beta(x, \eta)$ is identified on $\mathcal{X} \times S_\eta$, and so is the probability function of η, $f(\eta)$, and hence $\mu(x) = \sum_\eta \beta(x, \eta)f(\eta)$ is identified. Q.E.D.

Proof of Theorem 3(ii): We normalize the marginal distributions of η and ν to uniform distributions on the interval $[0, 1]$. Theorem 1 shows that $h(z, \eta)$ is identified. Next we follow the same procedure to estimate ν, since conditional on η, ν and X are independent:

$$
\begin{aligned}
F_{Y|X,\eta}(y_0, x_0, \eta_0) &= Pr(Y \le y_0|X = x_0, \eta = \eta_0) = Pr(g(X, \eta, \nu) \le y_0|X = x_0, \eta = \eta_0) \\
&= Pr(\nu \le g^{-1}(X, \eta, y_0)|X = x_0, \eta = \eta_0) = Pr(\nu \le g^{-1}(x_0, \eta_0, y_0)|X = x_0, \eta = \eta_0) \\
&= F_\nu(g^{-1}(x_0, \eta_0, y_0)) = g^{-1}(x_0, \eta_0, y_0).
\end{aligned}
$$

For all values (x_0, η_0) in the joint distribution of (X, η) this conditional distribution function is identified, and hence for all those values the inverse of the function $g(x, \eta, \nu)$ and thus the function itself is identified.

Given identification of $g(x, \eta, \nu)$, we can derive ε through the relation $\varepsilon = G^{-1}(y, x)$, where $G(y, x) = \int_0^1 F_{Y|X,\eta}(y, x, \eta)d\eta$ as in Section 4.2 Q.E.D.

Throughout the remainder of the Appendix, C will denote a generic positive constant that may be different in different uses. Also, with probability approaching one will be abbreviated as w.p.a.1, positive semi-definite as p.s.d., positive definite as p.d., $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, and $A^{1/2}$ will denote the minimum and maximum eigenvalues, and square root, of respectively of a symmetric matrix A. Let $\sum_i$ denote $\sum_{i=1}^n$. Also, let CS, M, and T refer to the Cauchy-Schwartz, Markov, and triangle inequalities, respectively. Also, let CM refer to the following result that we use without proof: If $\mathbb{E}[|Y_n||Z_n] = O_p(r_n)$ then $Y_n = O_p(r_n)$.

[23]

Before proving Theorem 4, we prove a preliminary result. Let $q_i = q^L(z_i)$, $v_{ij} = 1(x_j \leq x_i) - F(x_i|z_j)$.

Lemma A1: For $Z = (z_1, ..., z_n)$ and $L \times 1$ vectors of functions $b_i(Z)$, $(i = 1, ..., n)$, if $\sum_{i=1}^n b_i(Z)'\hat{Q}b_i(Z)/n = O_p(r_n)$ then

$$\sum_{i=1}^n \{b_i(Z)'\sum_{j=1}^n q_j v_{ij}/\sqrt{n}\}^2/n = O_p(r_n).$$

Proof: Note that $|v_{ij}| \leq 1$. Consider $j \neq k$ and suppose without loss of generality that $j \neq i$ (otherwise reverse the role of j and k because we cannot have $i = j$ and $i = k$). By independence of the observations,

$$\mathbb{E}[v_{ij}v_{ik}|Z] = \mathbb{E}[\mathbb{E}[v_{ij}v_{ik}|Z, x_i, x_k]|Z] = \mathbb{E}[v_{ik}\mathbb{E}[v_{ij}|Z, x_i, x_k]|Z] = \mathbb{E}[v_{ik}\mathbb{E}[v_{ij}|z_j, z_i, x_i]|Z]$$

$$= \mathbb{E}[v_{ik}\{\mathbb{E}[1(x_j \leq x_i)|z_j, z_i, x_i] - F(x_i|z_j)\}|Z] = 0.$$

Therefore, it follows that

$$\mathbb{E}[\sum_{i=1}^n \{b_i(Z)'\sum_{j=1}^n q_j v_{ij}/\sqrt{n}\}^2/n|Z] \leq \sum_{i=1}^n b_i(Z)'\{\sum_{j,k=1}^n q_j \mathbb{E}[v_{ij}v_{ik}|Z]q_k'/n\}b_i(Z)/n$$

$$= \sum_{i=1}^n b_i(Z)'\{\sum_{j=1}^n q_j \mathbb{E}[v_{ij}^2|Z]q_j'/n\}b_i(Z)/n$$

$$\leq \sum_{i=1}^n b_i(Z)'\hat{Q}b_i(Z)/n,$$

so the conclusion follows by CM. *Q.E.D.*

Proof of Theorem 4: Let $\delta_{ij} = F(x_i|z_j) - q_j'\gamma^L(x_i)$, with $|\delta_{ij}| \leq L^{-2d_1}$ by Assumption 5.1. Then for $\tilde{\eta}_i = \tilde{F}(x_i|z_i)$ and $\eta_i = F(x_i|z_i)$,

$$\tilde{\eta}_i - \eta_i = \Delta_i^I + \Delta_i^{II} + \Delta_i^{III},$$

where

$$\Delta_i^I = q_i'\hat{Q}^-\sum_{j=1}^n q_j v_{ij}/n,$$

$$\Delta_i^{II} = q_i'\hat{Q}^-\sum_{j=1}^n q_j \delta_{ij}/n,$$

$$\Delta_i^{III} = -\delta_{ii}.$$

Note that $|\Delta_i^{III}| \leq CL^{-d_1}$ by Assumption 5.1. Also, by $\hat{Q}$ p.s.d. and symmetric there exists a diagonal matrix of eigenvalues Λ and an orthonormal matrix B such that $\hat{Q} = B\Lambda B'$. Let

[24]

Λ^- denote the diagonal matrix of inverse of nonzero eigenvalues and zeros and $\hat{Q}^- = B\Lambda^- B'$. Then $\sum_i q_i'\hat{Q}^- q_i = tr(\hat{Q}^-\hat{Q}) \le CL$. By CS and Assumption 5.1,

$$
\begin{aligned}
\sum_{i=1}^n (\Delta_i^{II})^2/n &\le \sum_{i=1}^n (q_i'\hat{Q}^- q_i \sum_{j=1}^n \delta_{ij}^2/n)/n \le C \sum_{i=1}^n (q_i'\hat{Q}^- q_i)L^{-2d_1}/n \\
&= CL^{-2d_1} tr(\hat{Q}^-\hat{Q}) \le CL^{1-2d_1}.
\end{aligned}
$$

Note that for $b_i(Z) = q_i'\hat{Q}^-/\sqrt{n}$ we have

$$
\sum_{i=1}^n b_i(Z)'\hat{Q} b_i(Z)/n = tr(\hat{Q}\hat{Q}^-\hat{Q}\hat{Q}^-)/n = tr(\hat{Q}\hat{Q}^-)/n \le CL/n = O_p(L/n),
$$

so it follows by Lemma A1 that $\sum_{i=1}^n (\Delta_i^I)^2/n = O_p(L/n)$. The conclusion then follows by T and by $|\tau(\tilde{\eta}) - \tau(\eta)| \le |\tilde{\eta} - \eta|$, which gives $\sum_i (\hat{\eta}_i - \eta_i)^2/n \le \sum_i (\tilde{\eta}_i - \eta_i)^2/n$. Q.E.D.

Before proving other results we give some useful lemmas. For these results let $p_i = p^K(w_i)$, $\hat{p}_i = p^K(\hat{w}_i)$, $p = [p_1, ..., p_n]$, $\hat{p} = [\hat{p}_1, ..., \hat{p}_n]$, $\hat{P} = \hat{p}'\hat{p}/n$, $\tilde{P} = p'p/n$, $P = \mathbb{E}[p_i p_i']$. Note that in the statement of these results we allow $\hat{\eta}_i$ and η_i to be vectors. Also, as in Newey (1997) it can be shown that without loss of generality we can set $P = I_K$.

Lemma A2: *If Assumptions 3.1 - 3.2 are satisfied then* $\mathbb{E}[Y|X, Z] = \beta(X, \eta)$ *evaluated at* $\eta = F_{X|Z}(X|Z)$.

Proof: Recall $\eta = F_{X|Z}(X|Z)$ is a function of X and Z that is invertible in X with inverse $X = h(Z, \eta)$. By independence of Z and (ε, η), ε is independent of Z conditional on η, so that

$$
\begin{aligned}
\mathbb{E}[Y|X, Z] &= \mathbb{E}[Y|X, Z, \eta] = \mathbb{E}[g(X, \varepsilon)|X, Z, \eta] = \mathbb{E}[g(h(Z, \eta), \varepsilon)|\eta, Z] \\
&= \int g(h(Z, \eta), \varepsilon) F_{\varepsilon|\eta}(d\varepsilon|\eta) = \beta(X, \eta),
\end{aligned}
$$

at $\eta = F_{X|Z}(X|Z)$. Q.E.D.

Let $u_i = Y_i - \beta(X_i, \eta_i)$, and let $u = (u_1, \ldots, u_n)'$.

Lemma A3: *If* $\sum_i \|\hat{\eta}_i - \eta_i\|^2/n = O_p(\Delta_n^2)$ *and Assumptions 5.1 - 5.4 are satisfied then*

$$
\begin{aligned}
(i), \quad \|\tilde{P} - P\| &= O_p(\zeta(K)\sqrt{K/n}), \qquad\qquad\qquad (A.1)\\
(ii), \quad \|p'u/n\| &= O_p(\sqrt{K/n}) \\
(iii), \quad \|\hat{p} - p\|^2/n &= O_p(\zeta_1(K)^2\Delta_n^2), \\
(iv), \quad \|\hat{P} - \tilde{P}\| &= O_p(\zeta_1(K)^2\Delta_n^2 + \sqrt{K}\zeta_1(K)\Delta_n), \\
(v), \quad \|(\hat{p} - p)'u/n\| &= O_p(\zeta_1(K)\Delta_n/\sqrt{n}).
\end{aligned}
$$

Proof: The first two results follow as the proof for Theorem 1 in Newey (1997). For *(iii)* a mean value expansion gives $\hat{p}_i = p_i + [\partial p^K(\bar{w}_i)/\partial \eta](\hat{\eta}_i - \eta_i)$, where $\bar{w}_i = (x_i, \bar{\eta}_i)$ and $\bar{\eta}_i$ lies in between $\hat{\eta}_i$ and η_i. Since $\hat{\eta}_i$ and η_i lie in $[0,1]$, it follows that $\bar{\eta}_i \in [0,1]$ so that by Assumption 5.2 $\|\partial p^K(\bar{w}_i)/\partial v\| \leq C\zeta_1(K)$. Then by CS, $\|\hat{p}_i - p_i\| \leq C\zeta_1(K)|\hat{\eta}_i - \eta_i|$. Summing up gives

$$\|\hat{p} - p\|^2/n = \sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n = O_p(\zeta_1(K)^2 \Delta_n^2). \tag{A.2}$$

For *(iv)*, by Assumption 5.2, $\sum_{i=1}^n \|p_i\|^2/n = O_p(\mathbb{E}[\|p_i\|^2]) = tr(I_K) = K$. Then by T, CS, and M,

$$\begin{aligned}
\|\hat{P} - \tilde{P}\| &\leq \sum_{i=1}^n \|\hat{p}_i\hat{p}_i' - p_ip_i'\|/n \leq \sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n + 2(\sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n)^{1/2}(\sum_{i=1}^n \|p_i\|^2/n)^{1/2}. \\
&= O_p(\zeta_1(K)^2 \Delta_n^2 + \sqrt{K}\zeta_1(K)\Delta_n).
\end{aligned}$$

Finally, for *(v)*, for $Z = (z_1, ..., z_n)$ and $X = (X_1, ..., X_n)$, it follows from Lemma A2 and Assumption 5.4 as in Newey 1997 that $\mathbb{E}[uu'|X, Z] \leq CI_n$, so that by p and $\hat{p}$ depending only on Z and X,

$$\begin{aligned}
\mathbb{E}[\|(\hat{p} - p)'u/n\|^2|X, Z] &= tr\{(\hat{p} - p)'\mathbb{E}[uu'|X, Z](\hat{p} - p)/n^2\} \\
&\leq C\|\hat{p} - p\|^2/n^2 = O_p(\zeta_1(K)^2 \Delta_n^2/n).
\end{aligned}$$

Q.E.D.

Lemma A4: *If Assumption 5.9 holds, then w.p.a.1, $\lambda_{\min}(\hat{P}) \geq C$, $\lambda_{\min}(\tilde{P}) \geq C$.*

Proof: By Lemma A3 and $\zeta(K)^2 K/n \leq CK\zeta_1(K)^2 L/n$, we have $\|\hat{P} - \tilde{P}\| \xrightarrow{p} 0$ and $\|\tilde{P} - P\| \xrightarrow{p} 0$, so the conclusion follows as in Newey (1997). Q.E.D.

Let $\beta = (\beta(w_1), \ldots, \beta(w_n))'$, and $\hat{\beta} = (\beta(\hat{w}_1), \ldots, \beta(\hat{w}_n))'$.

Lemma A5: *If $\sum_i \|\hat{\eta}_i - \eta_i\|^2/n = O_p(\Delta_n^2)$, Assumptions 5.1 - 5.4 are satisfied, $\sqrt{K}\zeta_1(K)\Delta_n \to 0$, and $K\zeta(K)^2/n \to 0$ then for $\tilde{\alpha} = \hat{P}^{-1}\hat{p}'\hat{\beta}/n, \bar{\alpha} = \hat{P}^{-1}\hat{p}'\beta/n$,*

$$(i) \; \|\hat{\alpha} - \bar{\alpha}\| = O_p(\sqrt{K/n}),$$

$$(ii) \; \|\tilde{\alpha} - \bar{\alpha}\| = O_p(\Delta_n),$$

$$(iii) \; \|\tilde{\alpha} - \alpha^K\| = O_p(K^{-d}).$$

Proof: For (i)

$$\begin{aligned}
\mathbb{E}[\|\hat{P}^{1/2}(\hat{\alpha} - \bar{\alpha})\|^2|X, Z] &= \mathbb{E}[u'\hat{p}\hat{P}^{-1}\hat{p}'u/n^2|X, Z] = tr\{\hat{P}^{-1/2}\hat{p}'\mathbb{E}[uu'|X, Z]\hat{p}\hat{P}^{-1/2}\}/n^2 \\
&\leq Ctr\{\hat{p}\hat{P}^{-1}\hat{p}'\}/n^2 \leq Ctr(I_K)/n = CK/n.
\end{aligned}$$

[26]

Since by Lemma A4, $\lambda_{\min}(\hat{P}) \geq C$ w.p.a.1, this implies that $\mathbb{E}[\|\hat{\alpha} - \bar{\alpha}\|^2 | X, Z] \leq CK/n$. Similarly, for *(ii)*,

$$\|\hat{P}^{1/2}(\tilde{\alpha} - \bar{\alpha})\|^2 \leq C(\hat{\beta} - \beta)'\hat{p}\hat{P}^{-1}\hat{p}'(\hat{\beta} - \beta)/n^2 \leq C\|\hat{\beta} - \beta\|^2/n = O_p(\Delta_n^2),$$

which follows from $\beta(w)$ being Lipschitz in η, so that also $\|\tilde{\alpha} - \bar{\alpha}\|^2 = O_p(\Delta_n^2)$. Finally for *(iii)*,

$$
\begin{aligned}
\|\hat{P}^{1/2}(\tilde{\alpha} - \alpha^K)\|^2 &= \|\tilde{\alpha} - \hat{P}^{-1}\hat{p}'\hat{p}\alpha^K/n\|^2 \leq C(\hat{\beta} - \hat{p}'\alpha^K)'\hat{p}\hat{P}^{-1}\hat{p}'(\hat{\beta} - \hat{p}'\alpha^K)/n^2 \\
&\leq \|\hat{\beta} - \hat{p}\alpha^K\|^2/n \leq C \sup_{w \in \mathcal{W}} |\beta_0(w) - p^K(w)'\alpha^K|^2 = O_p(K^{-2d}),
\end{aligned}
$$

so that $\|\hat{P}^{1/2}(\tilde{\alpha} - \alpha^K)\|^2 = O_p(K^{-2d})$. Q.E.D.

Proof of Theorem 5: Note that by Theorem 4, for $\Delta_n^2 = L/n + L^{1-2d_1}$, we have $\sum_i \|\hat{\eta}_i - \eta_i\|^2/n = O_p(\Delta_n^2)$, so by $K\zeta(K)^2/n \leq CK\zeta_1(K)^2L/n$ the hypotheses of Lemma A5 are satisfied. Also by Lemma A5 and T, $\|\hat{\alpha} - \alpha^K\|^2 = O_p(K/n + K^{-2d} + \Delta_n^2)$. Then

$$
\begin{aligned}
\int [\hat{\beta}(w) - \beta(w)]^2 F_w(dw) &= \int [p^K(w)'(\hat{\alpha} - \alpha^K) + p^K(w)'\alpha^K - \beta(w)]^2 F_w(dw) \\
&\leq C\|\hat{\alpha} - \alpha^K\|^2 + CK^{-2d} = O_p(K/n + K^{-2d} + \Delta_n^2).
\end{aligned}
$$

For the second part of Theorem 5,

$$
\begin{aligned}
\sup_{w \in \mathcal{W}} |\hat{\beta}(w) - \beta(w)| &= \sup_{w \in \mathcal{W}} |p^K(w)'(\hat{\alpha} - \alpha^K) + p^K(w)'\alpha^K - \beta(w)| \\
&= O_p(\zeta(K)(K/n + K^{-2d} + \Delta_n^2)^{1/2}) + O_p(K^{-d}) \\
&= O_p(\zeta(K)(K/n + K^{-2d} + L/n + L^{1-2d_1})^{1/2}).
\end{aligned}
$$

Q.E.D.

Proof of Theorem 6: First, we note that it can be assumed without loss of generality that $\mathbb{E}[B_x p^{K_x}(x_i) p^{K_x}(x_i)' B_x'] = I_{K_x}$ and $\mathbb{E}[B_\eta p^{K_\eta}(\eta_i) p^{K_\eta}(\eta_i)' B_\eta'] = I_{K_\eta}$ which can be shown as in Newey (1997). Also, since $c'p^{K_\eta}(\eta) \equiv 1$ for some c, for $\tilde{c} \equiv B_\eta^{-1'}c$ we have $\tilde{c}'B_\eta p^{K_\eta}(\eta) \equiv 1$. Note that $\tilde{c}'\tilde{c} = \tilde{c}'\mathbb{E}[B_\eta p_\eta^{K_\eta}(\eta_i) p_\eta^{K_\eta}(\eta_i)' B_\eta]\tilde{c} = 1$, so that there is an orthonormal matrix $\tilde{B}_\eta$ with $\tilde{c}'$ as its first row. Then $\tilde{p}^{K_\eta}(\eta) = \tilde{B}_\eta B_\eta p^{K_\eta}(\eta)$ is an orthonormal basis, $e_1'\tilde{p}^{K_\eta}(\eta) = \tilde{c}'B_\eta p^{K_\eta}(\eta) \equiv 1$, and $\int_0^1 \tilde{p}^{K_\eta}(\eta)d\eta = \mathbb{E}[\tilde{p}^{K_\eta}(\eta) \cdot 1] = e_1$. Then $\tilde{p}^K(w) \overset{def}{=} (I \otimes \tilde{B}_\eta)Bp^K(w) = \tilde{p}^{K_x}(x) \otimes \tilde{p}^{K_\eta}(\eta)$ satisfies Assumption 5.5 with $B = I$. For notational convenience let $p^K(w) = \tilde{p}^K(w)$. Note that

$$\bar{p}(x) \overset{def}{=} \int_0^1 p^K(w)d\eta = p^{K_x}(x) \otimes e_1, \quad \int \bar{p}(x)\bar{p}(x)'F_X(dx) = I_{K_x} \otimes e_1 e_1' \leq I_K. \tag{A.3}$$

[27]

As above, $\mathbb{E}[uu'|X, Z] \leq CI_n$, so that by Fubini's Theorem,

$$
\begin{aligned}
\mathbb{E}[\int \{\bar{p}(x)'(\hat{\alpha} - \bar{\alpha})\}^2 F_X(dx)|X, Z] &= \int \{\bar{p}(x)'\hat{P}^{-1}\hat{p}'\mathbb{E}[uu'|X, Z]\hat{p}\hat{P}^{-1}\bar{p}(x)\}F_X(dx)/n^2 \\
&\leq C \int \bar{p}(x)'\hat{P}^{-1}\bar{p}(x)F_X(dx)/n \\
&\leq C\mathbb{E}[\bar{p}(X)'\bar{p}(X)]/n \\
&= C\mathbb{E}[p^{K_x}(X)'p^{K_x}(X) \otimes e_1'e_1]/n = K_x/n.
\end{aligned}
$$

It then follows by CM that $\int \{\bar{p}(x)'(\hat{\alpha}-\bar{\alpha})\}^2 F_X(dx) = O_p(K_x/n)$. Note that $K^{-d} = (K_x^2[K_\eta/K_x])^{-d} \leq CK_x^{-2d}$. Then by Lemma A5, eq. (A.3), and T,

$$
\begin{aligned}
\int \{\bar{p}(x)'(\bar{\alpha} - \alpha^K)\}^2 F_X(dx) &\leq (\bar{\alpha} - \alpha^K)' \int \bar{p}(x)\bar{p}(x)'F_X(dx)(\bar{\alpha} - \alpha^K) \leq \|\bar{\alpha} - \alpha^K\|^2 \\
&= O_p(K_x^{-4d} + \Delta_n^2).
\end{aligned}
$$

Also, by CS,

$$
\int \{\bar{p}(x)'\alpha^K - \mu(x)\}^2 F_X(dx) \leq \int \int_0^1 \{p^K(w)'\alpha - \beta(w)\}^2 d\eta F_X(dx) = O(K^{-2d}) = O(K_x^{-4d}).
$$

Then the conclusion follows by T and

$$
\begin{aligned}
\int [\hat{\mu}(x) - \mu(x)]^2 F_0(dx) &= \int \{\bar{p}(x)'(\hat{\alpha} - \alpha^K) + \bar{p}(x)'\alpha^K - \mu(x)\}^2 F_X(dx) \\
&= O_p(K_x/n + K_x^{-4d} + \Delta_n^2) + O_p(K_x^{-4d}). \quad Q.E.D.
\end{aligned}
$$

B   Proofs of Asymptotic Normality and Consistent Standard Errors.

Throughout this Appendix we will take $P = I$ and $Q = I$, which is possible as discussed in Newey (1997), and $\Delta_n^2 = L/n + L^{1-2d_1}$, $\tilde{\Delta}_n^2 = \Delta_n^2 + \xi_n^3$, $\bar{\Delta}_n^2 = K/n + K^{-2\bar{d}} + \tilde{\Delta}_n^2$.

Lemma B0: *If Assumption 5.9 is satisfied then all of the following converge to zero: $\sqrt{n}\zeta_1(K)^2\tilde{\Delta}_n^2\Delta_n$, $\sqrt{nK}\zeta_1(K)\tilde{\Delta}_n\Delta_n$, $\sqrt{n}\zeta_1(K)\tilde{\Delta}_n\Delta_n$, $\sqrt{n}\zeta(K)\Delta_n^2/\xi_n$, $\sqrt{n}\zeta(K)\xi_n^2$, $\sqrt{n}\zeta(K)\tilde{\Delta}_n^2$, $\zeta(K)K^{1/2}L^{1/2}/\sqrt{n}$, $\zeta_1(K)\tilde{\Delta}_n$, $\zeta(K)^2L^{1-2d_1}$, $\zeta(K)^2\zeta(L)^2L^{1-2d_1}$, $\zeta(K)^2L\xi_n$, $\zeta(K)^2KL/n$, $\zeta(K)^2(K/n + K^{-2d} + \tilde{\Delta}_n)$, $\zeta(K)^4\tilde{\Delta}_n^4L$, $K\zeta_1(K)^2\tilde{\Delta}_n^2L$. If Assumption 5.10 is also satisfied, then also the following converge to zero: $\zeta_1(K)^2\bar{\Delta}_n^2L$, .*

Proof: Note first that by $nL^{1-2d_1} \to 0$ we have $\Delta_n^2 = L/n + (1/n)nL^{1-2d_1} \leq CL/n$. Also, by $C^{-1}\Delta_n^{2/3} \leq \xi_n \leq C\Delta_n^{2/3}$ we have $\Delta_n^2/\xi_n \leq C\Delta_n^{4/3} \leq C(L/n)^{2/3}$ and $\xi_n^2 \leq C(L/n)^{2/3}$. Then $\tilde{\Delta}_n^2 \leq CL/n$. Thus we have

$$\sqrt{n}\zeta_1(K)^2\tilde{\Delta}_n^2\Delta_n \leq C\zeta_1(K)^2L^{3/2}/n \to 0, \sqrt{nK}\zeta_1(K)\tilde{\Delta}_n\Delta_n \leq C[K\zeta_1(K)^2L^2/n]^{1/2} \to 0,$$

$$\sqrt{n}\zeta_1(K)\tilde{\Delta}_n\Delta_n \leq C\sqrt{nK}\zeta_1(K)\tilde{\Delta}_n\Delta_n \to 0, \sqrt{n}\zeta(K)\Delta_n^2/\xi_n \leq C[\zeta(K)^6L^4/n]^{1/6} \to 0,$$

$$\sqrt{n}\zeta(K)\xi_n^2 \leq C[\zeta(K)^6L^4/n]^{1/6} \to 0, \sqrt{n}\zeta(K)\tilde{\Delta}_n^2 \leq C(\zeta(K)^2L^2/n)^{1/2} \to 0,$$

$$\zeta(K)K^{1/2}L^{1/2}/\sqrt{n} \leq C[K\zeta_1(K)^2L^2/n]^{1/2} \to 0, \zeta_1(K)\tilde{\Delta}_n \leq C[\zeta_1(K)^2L/n]^{1/2}$$

$$\zeta(K)^2L^{1-2d_1} \leq [\zeta(K)^2/n]nL^{1-2d_1} \to 0, \zeta(K)^2\zeta(L)^2L^{1-2d_1} \leq [\zeta(K)^2\zeta(L)^2/n]nL^{1-2d_1} \to 0,$$

$$\zeta(K)^2L\xi_n \leq C(\zeta(K)^6L^4/n)^{1/3} \to 0, K\zeta_1(K)^2\tilde{\Delta}_n^2L \leq CK\zeta_1(K)^2L^2/n \to 0,$$

$$\zeta(K)^2KL/n \leq CK\zeta_1(K)^2L^2/n \to 0, \zeta_1(K)^4\tilde{\Delta}_n^4L \leq C(\zeta_1(K)^2L^{3/2}/n) \to 0$$

$$\zeta(K)^2(K/n + K^{-2d} + \tilde{\Delta}_n) \leq C\zeta_1(K)^2K/n + (\zeta(K)^2/n)(nK^{-2d}) + (\zeta(K)^4L/n)^{1/2} \to 0,$$

$$K\zeta_1(K)^2\tilde{\Delta}_n^2L \leq \zeta_1(K)^2KL^2/n \to 0.$$

If Assumption 5.10 is also satisfied then

$$\zeta_1(K)^2L\bar{\Delta}_n^2 \leq C\zeta_1(K)^2LK/n + C\zeta_1(K)^2LK^{-2\bar{d}} + C\zeta_1(K)^2L^2/n \to 0.$$

Lemma B1: $|\tau_n(\tilde{\eta}) - \tau_n(\eta)| \leq |\tilde{\eta} - \eta|$. *In addition,* $\tau_n(\eta)$ *is continuously differentiable with derivative* $\tau_n'(\eta)$ *satisfying* $|\tau_n'(\tilde{\eta}) - \tau_n'(\eta)| \leq |\tilde{\eta} - \eta|/2\xi_n$. *Also, for any integer* r, $\int_0^1 |\tau_n(\eta) - \eta|^r d\eta = O(\xi_n^{r+1})$ *and* $\int_0^1 |\tau_n'(\eta) - 1|^r d\eta = O(\xi_n)$.

Proof: The derivative of $\tau_n(\eta)$ is equal to 0, 1, $t_n'(1 - \eta)$, or $t_n'(\eta)$. For each of the pieces the derivative is bounded by 1. For the second conclusion, since $t_n'(\eta) = (\eta + \xi_n)/2\xi_n$, we have

$$\tau_n'(\eta) = \begin{cases} 0, & \eta > 1 + \xi_n, \\ t_n'(1 - \eta), & 1 - \xi_n < \eta \leq 1 + \xi_n, \\ 1, & \xi_n \leq \eta \leq 1 - \xi_n, \\ t_n'(\eta), & -\xi_n \leq \eta < \xi_n, \\ 0, & \eta < -\xi_n. \end{cases}$$

By inspection, $\tau_n'(\eta)$ is piecewise linear and continuous with maximum absolute slope $1/2\xi_n$, giving the first conclusion. For the third, note that by symmetry of the $t_n(\eta)$ around $\eta = -\xi_n$, we have

$$\int_0^1 |\tau_n(\eta) - \eta|^r d\eta = 2\int_0^{\xi_n} |t_n(\eta) - \eta|^r d\eta = \int_0^{\xi_n} |(\eta^2 + 2\eta\xi_n + \xi_n^2 - 4\eta\xi_n)/4\xi_n|^r d\eta$$

$$= (4\xi_n)^{-r}\int_0^{\xi_n} (\xi_n - \eta)^{2r} d\eta = -(2r+1)^{-1}(4\xi_n)^{-r}[(\xi_n - \eta)^{2r+1}]_0^{\xi_n}$$

$$= (2r+1)^{-1}4^{-r}\xi_n^{r+1}.$$

For the fourth conclusion, again by symmetry

$$\int_0^1 |\tau_n'(\eta) - 1|^r d\eta = 2\int_0^{\xi_n} |t_n'(\eta) - 1|^r d\eta = 2\int_0^{\xi_n} |(\eta - \xi_n)/2\xi_n|^r d\eta$$

$$= 2^{1-r}\xi_n^{-r}\int_0^{\xi} (\xi_n - \eta)^r d\eta = -(r+1)^{-1}2^{1-r}\xi_n^{-r}[(\xi_n - \eta)^{r+1}]_0^{\xi} = \xi_n 2^{1-r}(r+1)^{-1}.$$

[29]

Q.E.D.

Lemma B2: *For every i there is a $\bar{\eta}_i$ in between $\tilde{\eta}_i$ and η_i with*

$$\hat{\eta}_i - \eta_i = \tau_n(\eta_i) - \eta_i + \tau_n'(\eta_i)(\tilde{\eta}_i - \eta_i) + r_{in},$$
$$|r_{in}| = |\tau_n'(\bar{\eta}_i) - \tau_n'(\eta_i)||\tilde{\eta}_i - \eta_i| \leq C|\tilde{\eta}_i - \eta_i|^2/\xi_n.$$

Proof: Follows by the mean-value theorem and by Lemma B1. Q.E.D.

Lemma B3: *If Assumptions 5.1-5.8 are satisfied, $\sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2/n = O_p(\tilde{\Delta}_n^2)$.*
Proof: By $|\tau_n(\tilde{\eta}_i) - \tau_n(\eta_i)| \leq |\tilde{\eta}_i - \eta|$, Theorem 4, Lemma B1, and M,

$$\sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2/n \leq C \sum_{i=1}^n \{[\tau_n(\eta_i) - \eta_i]^2 + (\tilde{\eta}_i - \eta_i)^2\}/n = O_p(\xi_n^3) + O_p(\Delta_n^2).\text{Q.E.D.}$$

Note that by $P = I$ we have $V = A(\Sigma + \Sigma_1)A'$. Let $F = 1/\sqrt{V}$, $\hat{H} = FA\hat{P}^{-1}$, $\tilde{H} = FA\tilde{P}^{-1}$, $H = FA$, and $\beta_\eta(w) = \partial\beta(w)/\partial\eta$.

Lemma B4: (i) $|F| \leq C$, (ii) $\|H\| \leq C$, (iii) $\|\tilde{H}\| = O_p(1)$, (iv) $\|\hat{H}\| = O_p(1)$, (v) $\max_{i \leq n} |\hat{p}_i| \leq C\zeta(K)$,

$$(vi) \quad \{(\hat{H} - \tilde{H})\hat{P}(\hat{H} - \tilde{H})'\}^{1/2} = O_p(\zeta_1(K)^2\tilde{\Delta}_n^2 + \sqrt{K}\zeta_1(K)\tilde{\Delta}_n),$$
$$(vii) \quad \{(\tilde{H} - H)\tilde{P}(\tilde{H} - H)'\}^{1/2} = O_p(\zeta(K)\sqrt{K/n}), \quad (viii) \quad \tilde{H}\tilde{P}\tilde{H}' = O_p(1),$$
$$(ix) \quad \sum_{i=1}^n (\hat{H}\hat{p}_i - Hp_i)^2/n = O_p(\zeta_1(K)^4\tilde{\Delta}_n^4 + K\zeta_1(K)^2)\tilde{\Delta}_n^2 + \zeta(K)^2K/n).$$

Proof: By $Var(y|X, Z) \geq C$ we have $V \geq A\Sigma A' \geq CAA'$. It follows from Assumption 5.8 i) or ii) as in the proofs of Theorems 2 and 3 of Newey (1997) that AA' is bounded away from zero, showing that (i) holds. For (ii), $\|H\|^2 = AA'/V \leq C$. For (iii), by Lemmas A3 and A4,

$$\|\tilde{H}\|^2 = \|H + H(I - \tilde{P})\tilde{P}^{-1}\|^2 \leq \|H\|^2(1 + \|I - \tilde{P}\|) = O_p(1).$$

(iv) follows similarly. For (v), by $\hat{w}_i \in \mathcal{W}$ and Assumption 5.2, $\max_{i \leq n} |\hat{p}_i| \leq C\zeta(K)$. For (vi), note that by $P = I$

$$(\hat{H} - \tilde{H})\hat{P}(\hat{H} - \tilde{H})' \leq |(\hat{H} - \tilde{H})(\hat{P} - I)(\hat{H} - \tilde{H})'| + \|\hat{H} - \tilde{H}\|^2 \leq \|\hat{H} - \tilde{H}\|^2(\|\hat{P} - I\| + 1).$$

Furthermore, w.p.a.1 $\|\hat{H} - \tilde{H}\| = \|\hat{H}(\tilde{P} - \hat{P})\tilde{P}^{-1}\| \leq C\|\hat{H}\|\|\tilde{P} - \hat{P}\|$ by Lemma A3 and CS, so by Lemma A3, $(\hat{H} - \tilde{H})\hat{P}(\hat{H} - \tilde{H})' \leq \|\hat{P} - \tilde{P}\|^2 O_p(1)$. Applying Lemma A3 gives the conclusion. (vii) follows similarly. The next conclusion $(viii)$ holds by CS, Lemma A2, and w.p.a.1

$$\tilde{H}\tilde{P}\tilde{H}' \leq |\tilde{H}(\tilde{P} - I)\tilde{H}'| + \|\tilde{H}\|^2 \leq \|\tilde{H}\|^2(1 + \|\tilde{P} - I\|) \leq C\|\tilde{H}\|^2 = O_p(1).$$

[30]

The final conclusion follows by Lemmas A2 and

$$
\begin{aligned}
\sum_{i=1}^{n}(\hat{H}\hat{p}_i - Hp_i)^2/n &\leq C\|\hat{H}\|^2 \sum_{i=1}^{n}\|\hat{p}_i - p_i\|^2/n + (\hat{H} - H)\tilde{P}(\hat{H} - H)' \\
&\leq O_p(\zeta_1(K)^2\tilde{\Delta}_n^2) + \|\hat{H} - H\|^2(\|\tilde{P} - I\| + 1),
\end{aligned}
$$

and

$$
\begin{aligned}
\|\hat{H} - H\|^2 &\leq 2\|\hat{H} - \tilde{H}\|^2 + 2\|\tilde{H} - H\|^2 \leq C(\|\hat{P} - \tilde{P}\|^2 + \|\tilde{P} - P\|^2) \\
&= O_p(\zeta_1(K)^4\tilde{\Delta}_n^4 + K\zeta_1(K)^2)\tilde{\Delta}_n^2 + \zeta(K)^2K/n)
\end{aligned}
$$

Q.E.D.

Next, let $\mu_{ji} = -Hp_j\beta_\eta(w_j)\tau_n'(\eta_j)q_j'q_iv_{ji}$ and $\bar{\mu}_i = \mathbb{E}[\mu_{ji}|y_i, x_i, z_i], (j \neq i)$,

Lemma B5: *If Assumptions 5.1-5.9 are satisfied,*

$$
\mathbb{E}[|\mu_{ii}|] \leq C\zeta(L)L^{1/2}, \mathbb{E}[\mu_{ij}^2] \leq C\zeta(L)^2, \mathbb{E}[\bar{\mu}_i^4] \leq C\zeta(K)^4\zeta(L)^4L.
$$

Proof: By Lemma B4, boundedness of v_{ij}, $\beta_\eta(w_j)$, and $\tau_n'(\eta_j)$, and CS,

$$
\mathbb{E}[|\mu_{ii}|] \leq C\{\mathbb{E}[Hp_ip_i'H']\}^{1/2}\{\mathbb{E}[\{q_i'q_iv_{ii}\}^2]\}^{1/2}/n \leq C\zeta(L)L^{1/2},
$$

$$
\mathbb{E}[\mu_{ij}^2] \leq C\mathbb{E}[\{Hp_i\}^2q_i'v_{ij}^2q_jq_j'q_i] \leq C\mathbb{E}[\{Hp_i\}^2q_i'q_i] \leq C\zeta(L)^2\mathbb{E}[\{Hp_i\}^2] \leq C\zeta(L)^2,
$$

$$
\mathbb{E}[\bar{\mu}_i^4] \leq \mathbb{E}[\mu_{ij}^4] \leq C\mathbb{E}[\{Hp_iq_i'q_j\}^4] \leq C\zeta(K)^4\zeta(L)^4\mathbb{E}[q_i'q_jq_j'q_i] = C\zeta(K)^4\zeta(L)^4L.\text{Q.E.D.}
$$

Lemma B6: *If* $\sum_{i=1}^{n}s_i^2/n = O_p(1)$ *and* $\sum_{i=1}^{n}(\hat{s}_i - s_i)^2/n = O_p(r_n^2)$ *for* $r_n \to 0$ *then* $\sum_{i=1}^{n}|\hat{s}_i^2 - s_i^2|/n = O_p(r_n)$.
Proof: By T, CS, and $O_p(r_n^2) + O_p(1)O_p(r_n) = O_p(r_n)$,

$$
\begin{aligned}
\sum_{i=1}^{n}|\hat{s}_i^2 - s_i^2|/n &\leq \sum_{i=1}^{n}(|\hat{s}_i - s_i|^2 + 2|s_i||\hat{s}_i - s_i|)/n \\
&\leq \sum_{i=1}^{n}|\hat{s}_i - s_i|^2/n + 2\{\sum_{i=1}^{n}|s_i|^2/n\}^{1/2}\{\sum_{i=1}^{n}|\hat{s}_i - s_i|^2/n\}^{1/2} = O_p(r_n). \text{ Q.E.D.}
\end{aligned}
$$

Lemma B7: *If Assumptions 5.1-5.9 are satisfied,*

$$
\hat{H}\sum_{i=1}^{n}\hat{p}_i[\beta_0(w_i) - \beta_0(\hat{w}_i)]/\sqrt{n} = \sqrt{n}\sum_{i,j=1}^{n}\mu_{ji}/n^2 + o_p(1) = \sum_{i=1}^{n}\bar{\mu}_i/\sqrt{n} + o_p(1).
$$

Proof: By Lemma B0 it follows similarly to Lemma A2 that $\|\hat{Q} - I\| = O_p(L^{1/2}\zeta(L)/\sqrt{n}) \xrightarrow{p} 0$, and that w.p.a.1 $\hat{Q}$ is nonsingular and $\lambda_{\max}(\hat{Q}^{-1}) \leq C$. It follows by expanding $\hat{\beta}_i = \beta_0(\hat{w}_i)$ around $\beta_i = \beta_0(w_i)$ and straightforward algebra that w.p.a.1

$$\hat{H}\sum_{i=1}^n \hat{p}_i(\beta_i - \hat{\beta}_i)/\sqrt{n} = \sqrt{n}\sum_{i,j=1}^n \mu_{ij}/n^2 + \hat{R}, \tag{B.1}$$

where $\hat{R} = \sum_{j=1}^8 \hat{R}_j$ and for r_{in} as in Lemma B2,

$$\hat{R}_1 = (\hat{H} - \tilde{H})\sum_{i=1}^n \hat{p}_i\beta_\eta(w_i)\tau_n'(\eta_i)(\tilde{\eta}_i - \eta_i)/\sqrt{n},$$

$$\hat{R}_2 = \tilde{H}\sum_{i=1}^n (\hat{p}_i - p_i)\beta_\eta(w_i)\tau_n'(\eta_i)(\tilde{\eta}_i - \eta_i)/\sqrt{n}, \hat{R}_3 = \hat{H}\sum_{i=1}^n \hat{p}_i\beta_\eta(w_i)r_{in}/\sqrt{n},$$

$$\hat{R}_4 = \hat{H}\sum_{i=1}^n \hat{p}_i\beta_\eta(w_i)[\tau_n(\eta_i) - \eta_i]/\sqrt{n}, \hat{R}_5 = \hat{H}\sum_{i=1}^n \hat{p}_i\beta_{\eta\eta}(\bar{w}_i)(\hat{\eta}_i - \eta_i)^2/2\sqrt{n}.$$

$$\hat{R}_6 = \tilde{H}\sum_{i=1}^n p_i\beta_\eta(w_i)\tau_n'(\eta_i)(\Delta_i^{II} + \Delta_i^{III})/\sqrt{n},$$

$$\hat{R}_7 = \tilde{H}\sum_{i=1}^n p_i\beta_\eta(w_i)\tau_n'(\eta_i)q_i'(\hat{Q}^{-1} - I)\sum_{j=1}^n q_j v_{ij}/n\sqrt{n},$$

$$\hat{R}_8 = (\tilde{H} - H)\sum_{i=1}^n p_i\beta_\eta(w_i)\tau_n'(\eta_i)q_i'\sum_{j=1}^n q_j v_{ij}/n\sqrt{n},$$

where Δ_i^I and Δ_i^{II} are specified as in the proof of Theorem 4. Next, we consider each $\hat{R}_j$ in turn. By Lemmas A3, B0, B4, CS, and $\beta_\eta(w_i)\tau_n'(\eta_i)$ bounded,

$$|\hat{R}_1| \leq \sqrt{n}\{(\hat{H} - \tilde{H})\hat{P}(\hat{H} - \tilde{H})'\}^{1/2}\{\sum_{i=1}^n (\tilde{\eta}_i - \eta_i)^2/n\}^{1/2}$$

$$= O_p(\sqrt{n}[\zeta_1(K)^2\tilde{\Delta}_n^2 + \sqrt{K}\zeta_1(K)\tilde{\Delta}_n]\Delta_n) \xrightarrow{p} 0,$$

$$|\hat{R}_2| \leq C\sqrt{n}\|\hat{H}\|\sum_{i=1}^n \|\hat{p}_i - p_i\|\|\tilde{\eta}_i - \eta_i\|/n = O_p(\sqrt{n}\zeta_1(K)\tilde{\Delta}_n\Delta_n) \xrightarrow{p} 0.$$

Then by Lemmas B0, B2, B3, and B4,

$$|\hat{R}_3| \leq C\|\hat{H}\|\zeta(K)\sum_{i=1}^n |r_{in}|/\sqrt{n} = O_p(\sqrt{n}\zeta(K)\Delta_n^2/\xi_n) \xrightarrow{p} 0,$$

By Lemmas B0, B1, B3, and M

$$|\hat{R}_4| \leq C\sqrt{n}\|\hat{H}\|\zeta(K)\sum_{i=1}^n |\tau_n(\eta_i) - \eta_i|/n = O_p(\sqrt{n}\zeta(K)\xi_n^2) \xrightarrow{p} 0,$$

$$|\hat{R}_5| \leq \sqrt{n}\|\hat{H}\|\zeta(K)\sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2/n = O_p(\sqrt{n}\zeta(K)\tilde{\Delta}_n^2) \xrightarrow{p} 0.$$

[32]

By Assumption 5.9, the proof of Theorem 4, CS, and Lemma B4,

$$|\hat{R}_6| \leq C\sqrt{n}\{\tilde{H}\tilde{P}\tilde{H}\}^{1/2}\{\sum_{i=1}^{n}[(\Delta_i^I)^2 + (\Delta_i^{II})^2]/n\}^{1/2} = O_p(\sqrt{n}L^{(1/2)-d_1}) \xrightarrow{p} 0.$$

Let $b_i(Z) = (\hat{Q}^{-1} - I)q_i$. Then

$$\sum_{i=1}^{n} b_i(Z)'\hat{Q}b_i(Z)/n \leq \sum_{i=1}^{n} q_i'(\hat{Q}^{-1} - I)\hat{Q}(\hat{Q}^{-1} - I)q_i/n = tr((I - \hat{Q})^2) = C\|I - \hat{Q}\|^2 \xrightarrow{p} 0.$$

It then follows by CS and Lemmas A1 and B4 that

$$|R_7| \leq C\{\tilde{H}\tilde{P}\tilde{H}'\}^{1/2}\{\sum_{i=1}^{n}[q_i'(\hat{Q}^{-1} - I)\sum_{j=1}^{n} q_j v_{ij}/\sqrt{n}]^2/n\}^{1/2} \xrightarrow{p} 0.$$

Next, for $b_i(Z) = q_i$,

$$\{\sum_{i=1}^{n} b_i(Z)'\hat{Q}b_i(Z)/n\}^{1/2} = tr(\hat{Q}^2)^{1/2} = \|\hat{Q}\| \leq \|\hat{Q} - I\| + \|I\| = O_p(L^{1/2}).$$

Therefore, we have by Lemmas A1, A3, B0, B4, CS, and CM,

$$|R_8| \leq C\{(\tilde{H} - H)\tilde{P}(\tilde{H} - H)\}^{1/2}\{\sum_{i=1}^{n}[q_i'\sum_{j=1}^{n} q_j v_{ij}/n\sqrt{n}]^2/n\}^{1/2} = O_p(\zeta(K)K^{1/2}L^{1/2}/\sqrt{n}) \xrightarrow{p} 0.$$

It then follows from T that $\hat{R} \xrightarrow{p} 0$ in equation (B.1), giving the first equality in the conclusion.

Next, $\mathbb{E}[\mu_{ij}|y_i, x_i, z_i] = 0$, and by Lemma B4,

$$\mathbb{E}[|\mu_{ii}|]/n \leq C\zeta(L)L^{1/2}/n \to 0, \mathbb{E}[\mu_{ij}^2]/n^2 \leq C\zeta(L)^2/n^2 \to 0.$$

The second equality of the Lemma then follows by the V-statistic result in Lemma 8.4 of Newey and McFadden (1994). Q.E.D.

Lemma B8: *If Assumptions 5.1-5.9 are satisfied, $\hat{H}\hat{p}'u/\sqrt{n} = Hp'u/\sqrt{n} + o_p(1)$.*
Proof: $\|\hat{H} - H\| \xrightarrow{p} 0$ follows from the proof of Lemma B7 (see R_1 and R_8). For $\tilde{W} = (z_1, x_1, ..., z_n, x_n)$, by B4 w.p.a.1

$$\mathbb{E}[\|(\hat{H} - H)p'u/\sqrt{n}\|^2|\tilde{W}] = (\hat{H} - H)p'\mathbb{E}[uu'|\tilde{W}]p(\hat{H} - H)'/n \leq C(\hat{H} - H)\tilde{P}(\hat{H} - H)' \xrightarrow{p} 0.$$

Then by Lemma A2 and B0, $\|(\hat{p} - p)'u/\sqrt{n}\| = O_p(\zeta_1(K)\tilde{\Delta}_n) \xrightarrow{p} 0$, so that by M and Lemma B4,

$$\|(\hat{H}\hat{p}'u - Hp'u)/\sqrt{n}\| \leq \|\hat{H}\|\|(\hat{p} - p)'u/\sqrt{n}\| + \|(\hat{H} - H)p'u/\sqrt{n}\| \xrightarrow{p} 0.\text{Q.E.D.}$$

Proof of Theorem 7: By Assumption 5.3, $(\hat{\beta}-\hat{p}'\alpha^K)'(\hat{\beta}-\hat{p}'\alpha^K)/n = \sum_{i=1}^{n}[\beta(\hat{w}_i)-p^K(\hat{w}_i)'\alpha^K]^2/n$
$= O_p(K^{-2d})$, so that by Lemma B4

$$|\hat{H}\hat{p}'(\hat{\beta}-\hat{p}'\alpha^K)/\sqrt{n}|^2 \le n\hat{H}\hat{P}\hat{H}'(\hat{\beta}-\hat{p}'\alpha^K)'(\hat{\beta}-\hat{p}'\alpha^K)/n = O_p(nK^{-2d}) \xrightarrow{p} 0.$$

Also, by Assumption 5.8, $|a(p^K(\cdot)'\alpha^K) - a(\beta_0)| = |a(p^K(\cdot)'\alpha^K - \beta_0(\cdot))| = O(K^{-d})$. Then by Lemmas B7 and B8,

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta)/\sqrt{V} &= \sqrt{n}[a(\hat{\beta})-a(\beta_0)]/\sqrt{V} = \hat{H}[\hat{p}'u + \hat{p}'(\beta-\hat{\beta}) + \hat{p}'(\hat{\beta}-\hat{p}'\alpha^K)]/\sqrt{n}\\ &\quad + \sqrt{n}[a(p^K(\cdot)'\alpha^K) - a(\beta_0)]/\sqrt{V} = \sum_{i=1}^{n}(Hp_iu_i + \bar{\mu}_i)/\sqrt{n} + o_p(1).\end{aligned}$$

Let $Z_{in} = (Hp_iu_i + \bar{\mu}_i)/\sqrt{n}$. Note that $\mathbb{E}[Z_{in}] = 0$ and $Var(Z_{in}) = 1/n$. Then by Lemma B5 and $\mathbb{E}[\|Hp_i\|^4|u_i|^4] \le C\zeta(K)^2K$, for any $\epsilon > 0$ we have

$$\begin{aligned}n\mathbb{E}[1(|Z_{in}| > \epsilon)Z_{in}^2] = n\epsilon^2\mathbb{E}[1(|Z_{in}| > \epsilon)(Z_{in}/\epsilon)^2] &\le n\epsilon^2\mathbb{E}[1(|Z_{in}| > \epsilon)(Z_{in}/\epsilon)^4]\\ &\le n\epsilon^2\mathbb{E}[(Z_{in}/\epsilon)^4] = n\epsilon^{-2}\mathbb{E}[|Z_{in}|^4]\\ &\le C(\mathbb{E}[\|Hp_i\|^4|u_i|^4] + \mathbb{E}[\bar{\mu}_i^4])/n \le [\zeta(K)^2K + \zeta(K)^4\zeta(L)^4L]/n \to 0.\end{aligned}$$

The conclusion then follows by the Lindberg-Feller central limit theorem. Q.E.D.

Lemma B9: *For $\hat{\mu}_i = \hat{H}\hat{m}_i$, if Assumptions 5.1-5.9 are satisfied then $\sum_{i=1}^{n}\hat{\mu}_i^2/n - \mathbb{E}[\bar{\mu}_i^2] \xrightarrow{p} 0$.*
Proof: Let $\hat{t}_i = \hat{H}\hat{p}_i\partial\hat{\beta}(\hat{w}_i)/\partial\eta$, $\delta_{ij} = F(x_i|z_j) - q_j'\alpha(x_i)$, $\hat{a}_{ij} = q_j'\hat{Q}^{-1}q_iv_{ji}$, $\hat{\beta}_\eta(w) = \partial\hat{\beta}(w)/\partial\eta$, and $\beta_{0\eta}(w) = \partial\beta_0(w)/\partial\eta$. Then by $\hat{Q}^{-1}$ existing w.p.a.1, $\hat{\mu}_i = \bar{\mu}_i + \sum_{t=1}^{9}\tilde{r}_{ti}$ for

$$\begin{aligned}\tilde{r}_{1i} &= -\sum_{j=1}^{n}\hat{t}_jq_j'\hat{Q}^{-1}q_i\delta_{ji}/n, \tilde{r}_{2i} = -\sum_{j=1}^{n}\hat{t}_jq_j'\hat{Q}^{-1}q_iq_i'\hat{Q}^{-1}\sum_{k=1}^{n}q_k\delta_{jk}/n^2,\\ \tilde{r}_{3i} &= -\sum_{j=1}^{n}\hat{t}_jq_j'\hat{Q}^{-1}q_iq_i'\hat{Q}^{-1}\sum_{k=1}^{n}q_kv_{jk}/n^2, \tilde{r}_{4i} = \sum_{j=1}^{n}\hat{t}_j[1-\tau_n'(\eta_j)]\hat{a}_{ij}/n,\\ \tilde{r}_{5i} &= \sum_{j=1}^{n}\hat{s}_j[\hat{\beta}_\eta(\hat{w}_j) - \beta_{0\eta}(\hat{w}_j)]\tau_n'(\eta_j)\hat{a}_{ij}/n, \tilde{r}_{6i} = \sum_{j=1}^{n}\hat{s}_j[\beta_{0\eta}(\hat{w}_j) - \beta_{0\eta}(w_j)]\tau_n'(\eta_j)\hat{a}_{ij}/n,\\ \tilde{r}_{7i} &= \sum_{j=1}^{n}(\hat{s}_j - s_j)\beta_{0\eta}(w_j)\tau_n'(\eta_j)\hat{a}_{ij}/n, \tilde{r}_{8i} = \sum_{j=1}^{n}s_j\beta_{0\eta}(w_j)\tau_n'(\eta_j)q_j'(\hat{Q}^{-1}-I)q_iv_{ji}/n,\\ \tilde{r}_{9i} &= \sum_{j=1}^{n}\mu_{ji}/n - \bar{\mu}_i.\end{aligned}$$

By Lemma B4, $|\hat{t}_i| \le C\zeta(K)$ and $\sum_{i=1}^{n}q_i'\hat{Q}^{-1}q_i/n = tr(\hat{Q}\hat{Q}^{-1}) = L$ w.p.a.1, so by Assumption

5.9 and CS,

$$\sum_{i=1}^{n} \tilde{r}_{1i}^2/n \;\leq\; \sum_{i,j=1}^{n} \hat{t}_j^2 \{q_j'\hat{Q}^{-1}q_i\}^2 \delta_{ji}^2/n^2 \leq C\zeta(K)^2 L^{-2d_1} \sum_{i,j=1}^{n} q_i'\hat{Q}^{-1}q_j q_j'\hat{Q}^{-1}q_i/n^2$$

$$= \; C\zeta(K)^2 L^{-2d_1} \sum_{i=1}^{n} q_i'\hat{Q}^{-1}q_i/n = \zeta(K)^2 L^{1-2d_1} \to 0.$$

Similarly, $q_i'\hat{Q}^{-1}q_i \leq C\zeta(L)^2$ w.p.a.1, so that by Assumption 5.9

$$\sum_{i=1}^{n} \tilde{r}_{2i}^2/n \;\leq\; \sum_{i,j,k=1}^{n} \hat{t}_j^2 \{q_j'\hat{Q}^{-1}q_i\}^2 \{q_i'\hat{Q}^{-1}q_k\}^2 \delta_{jk}^2/n^3$$

$$\leq \; C\zeta(K) \sum_{i,j} \{q_j'\hat{Q}^{-1}q_i\}^2 q_i'\hat{Q}^{-1}q_i/n^3 = O_p(\zeta(L)^2 L/n^2),$$

so that by CM and Assumption 5.9

$$\sum_{i=1}^{n} \tilde{r}_{3i}^2/n \leq C\zeta(K)^2 \sum_{i,j=1}^{n} \{q_j'\hat{Q}^{-1}q_i\}^2 \{q_i'\hat{Q}^{-1}\sum_{k=1}^{n} q_k v_{jk}/n\}^2/n^2 = O_p(\zeta(K)^2\zeta(L)^2 L/n^2) \xrightarrow{p} 0.$$

Next, by v_{ji} bounded, w.p.a.1,

$$\sum_{i,j=1}^{n} a_{ji}^2/n \leq C \sum_{i,j=1}^{n} q_i'\hat{Q}^{-1}q_j q_j'\hat{Q}^{-1}q_i/n = C \sum_{i,j=1}^{n} q_i'\hat{Q}^{-1}q_i/n = CL.$$

Also, by Lemma B1 $\mathbb{E}[|\tau_n'(\eta_j) - 1|^2] = O(\xi_n)$, so by CS and Assumption 5.9 we have

$$\sum_{i=1}^{n} \tilde{r}_{4i}^2/n \leq C(\sum_{j=1}^{n} \hat{t}_j^2 |\tau_n'(\eta_j) - 1|^2/n) \sum_{i,j=1}^{n} a_{ji}^2/n = O_p(\zeta(K)^2 L \xi_n) \xrightarrow{p} 0.$$

Also, it follows as in the proof of Lemma A5 that for $\bar{\alpha}^K$ from Assumption 5.10 and for $\bar{\Delta}_n^2 = K/n + K^{-2\bar{d}} + \tilde{\Delta}_n^2$, $\|\hat{\alpha} - \bar{\alpha}^K\| = O_p(\bar{\Delta}^2)$. Then

$$\sup_{w\in\mathcal{W}} |\hat{\beta}_\eta(w) - \beta_{0\eta}(w)| \;\leq\; \sup_{w\in\mathcal{W}} |[\partial p^K(w)/\partial\eta]'(\hat{\alpha} - \bar{\alpha}^K) + \partial\{p^K(w)'\bar{\alpha}^K\}/\partial\eta - \beta_{0\eta}(w)|$$

$$\leq \; \zeta_1(K)\|\hat{\alpha} - \bar{\alpha}^K\| + CK^{-\bar{d}} = O_p(\zeta_1(K)\bar{\Delta}_n).$$

By Lemma B0 and $\tau_n'(\eta)$ bounded,

$$\sum_{i=1}^{n} \tilde{r}_{5i}^2/n \leq \sup_{w\in\mathcal{W}} |\hat{\beta}_\eta(w) - \beta_{0\eta}(w)|^2 (\sum_{j=1}^{n} \hat{s}_j^2/n) \sum_{i,j=1}^{n} a_{ji}^2/n \leq O_p(\zeta_1(K)^2\bar{\Delta}_n^2 L) \xrightarrow{p} 0.$$

By Lemmas 4, B0, and B4,

$$\sum_{i=1}^{n} \tilde{r}_{6i}^2/n \;\leq\; C\{\max_{j\leq n} \hat{s}_i^2\}\sum_{j=1}^{n}(\hat{\eta}_j - \eta_j)^2/n \sum_{i,j=1}^{n} a_{ji}^2/n \leq O_p(\zeta(K)^2 \tilde{\Delta}_n^2 L) \xrightarrow{p} 0.$$

$$\sum_{i=1}^{n} \tilde{r}_{7i}^2/n \;\leq\; C\sum_{j=1}^{n}(\hat{s}_j - s_j)^2/n \sum_{i,j=1}^{n} a_{ji}^2/n$$

$$\leq\; O_p([\zeta(K)^2 K/n + \zeta_1(K)^4 \tilde{\Delta}_n^4 + K\zeta_1(K)^2 \tilde{\Delta}_n^2]L) \xrightarrow{p} 0.$$

By Lemma A1,

$$\sum_{i=1}^{n} \tilde{r}_{8i}^2/n \;\leq\; (\sum_{j=1}^{n} s_j^2/n)\sum_{i,j=1}^{n}\{q_j'(\hat{Q}^{-1} - I)q_i v_{ji}\}^2/n^2 \leq O_p(1)\sum_{i,j=1}^{n} q_j'(\hat{Q}^{-1} - I)q_i q_i'(\hat{Q}^{-1} - I)q_j/n^2$$

$$\leq\; C tr\{(\hat{Q} - I)^2\} = C\|\hat{Q} - I\|^2 \xrightarrow{p} 0.$$

Next, let $\rho_{ji} = \mu_{ji} - \bar{\mu}_i$ consider j and k with $j \neq k$. Assume without loss of generality that $k \neq i$. Then by independence of the observations $\mathbb{E}[\rho_{ki}|y_i, x_i, z_i, y_j, x_j, z_j] = \mathbb{E}[\rho_{ki}|y_i, x_i, z_i] = 0$, so by iterated expectations,

$$\mathbb{E}[\rho_{ji}\rho_{ki}] = \mathbb{E}[\rho_{ji}\mathbb{E}[\mu_{ki}|y_i, x_i, z_i, y_j, x_j, z_j]] = 0.$$

Then by the observations identically distributed,

$$\mathbb{E}[\sum_{i=1}^{n} \tilde{r}_{9i}^2]/n \;=\; \mathbb{E}[(\sum_{j=1}^{n} \rho_{ji}/n)^2] = \sum_{j,k=1}^{n} \mathbb{E}[\rho_{ji}\rho_{ki}]/n^2 \leq \mathbb{E}[\rho_{ji}^2]/n + \mathbb{E}[\rho_{ii}^2]/n^2$$

$$\leq\; \mathbb{E}[\mu_{ji}^2]/n + 2\mathbb{E}[\mu_{ii}^2]/n^2 \leq C\mathbb{E}[s_j^2 q_j' q_j]/n + C\mathbb{E}[s_j^2\{q_j' q_j\}^2]/n^2$$

$$\leq\; C(\zeta(L)^2/n + \zeta(L)^4/n^2)\mathbb{E}[s_j^2] \to 0.$$

so by M, $\sum_{i=1}^{n} \tilde{r}_{9i}^2/n \xrightarrow{p} 0$. Then by T,

$$\{\sum_{i=1}^{n}(\hat{\mu}_i - \bar{\mu}_i)^2/n\}^{1/2} \leq \sum_{t=1}^{9}\{\sum_{i=1}^{n} \tilde{r}_{ti}^2/n\}^{1/2} \xrightarrow{p} 0.$$

Since $\bar{\mu}_i = Hm_i$, we have $\mathbb{E}[\bar{\mu}_i^2] = H\Sigma_1 H' = A\Sigma_1 A'/V \leq 1$. Then by M and Lemma B6, $|\sum_{i=1}^{n} \hat{\mu}_i^2/n - \sum_{i=1}^{n} \bar{\mu}_i^2/n| \xrightarrow{p} 0$. Also, by Lemma B5 $\mathbb{E}[\bar{\mu}_i^4]/n \leq \mathbb{E}[\mu_{ji}^4]/n \to 0$, so that by Chebyshev's law of large numbers, $\sum_{i=1}^{n} \bar{\mu}_i^2/n - \mathbb{E}[\bar{\mu}_i^2] \xrightarrow{p} 0$. The conclusion holds by T. Q.E.D.

Lemma B10: *If Assumptions 5.1-5.10 are satisfied then for $\hat{s}_i = \hat{H}\hat{p}_i$ and $s_i = Hp_i$ we have*

$$\sum_{i=1}^{n} \hat{s}_i^2 \hat{u}_i^2/n - \mathbb{E}[s_i^2 u_i^2] \xrightarrow{p} 0.$$

[36]

Proof: Let $\check{\Delta}_n^2 = K/n + K^{-2d} + \tilde{\Delta}_n^2$. It follows similarly to the proof of Theorem 5 that $\sum_{i=1}^n [\hat{\beta}(\hat{w}_i) - \beta_0(\hat{w}_i)]^2/n = O_p(\check{\Delta}_n^2)$, so that by $\beta_0(w)$ Lipschitz,

$$
\begin{aligned}
\sum_{i=1}^n [\hat{u}_i - u_i]^2/n &\leq 2\sum_{i=1}^n [\hat{\beta}(\hat{w}_i) - \beta_0(\hat{w}_i)]^2/n + 2\sum_{i=1}^n [\beta_0(\hat{w}_i) - \beta_0(w_i)]^2/n \\
&\leq O_p(\check{\Delta}_n^2) + C\sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2/n = O_p(\check{\Delta}_n^2).
\end{aligned}
$$

Then by Lemmas B0, B4 and B6,

$$
\sum_{i=1}^n \hat{s}_i^2 |\hat{u}_i^2 - u_i^2|/n \leq C\zeta(K)^2 \sum_{i=1}^n |\hat{u}_i^2 - u_i^2|/n \leq O_p(\zeta(K)^2 \check{\Delta}_n) \overset{p}{\to} 0.
$$

Now, since $\hat{s}_i$ and s_i are functions only of X and Z and $\mathbb{E}[u_i^2|X,Z] = \mathbb{E}[u_i^2|X_i,Z_i] \leq C$ we have $\mathbb{E}[|\hat{s}_i^2 - s_i^2|u_i^2|X,Z] = |\hat{s}_i^2 - s_i^2|\mathbb{E}[u_i^2|X,Z] \leq C|\hat{s}_i^2 - s_i^2|$. Also, $\sum_{i=1}^n s_i^2/n = O_p(1)$ and, as shown in the proof of Lemma B9, $\sum_{i=1}^n (\hat{s}_i - s_i)^2/n \overset{p}{\to} 0$. Then by Lemma B6,

$$
\mathbb{E}[|\sum_{i=1}^n \hat{s}_i^2 u_i^2/n - \sum_{i=1}^n s_i^2 u_i^2/n||X,Z] \leq C\sum_{i=1}^n |\hat{s}_i^2 - s_i^2|/n \overset{p}{\to} 0.
$$

Hence $\sum_{i=1}^n \hat{s}_i^2 u_i^2/n - \sum_{i=1}^n s_i^2 u_i^2/n \overset{p}{\to} 0$ by CM. Next, note that $|s_i| \leq C\zeta(K)$, so by Lemma B4,

$$
\mathbb{E}[s_i^4 u_i^4]/n \leq \mathbb{E}[s_i^4 \mathbb{E}[u_i^4|X_i,Z_i]]/n \leq C\mathbb{E}[s_i^4]/n \leq C\zeta(K)^2 \mathbb{E}[s_i^2]/n \to 0.
$$

Therefore, by Chebyshev's law of large numbers, $\sum_{i=1}^n s_i^2 u_i^2/n - \mathbb{E}[s_i^2 u_i^2] \overset{p}{\to} 0$, so the conclusion follows by T. Q.E.D.

Proof of Theorem 8: Note that

$$
\sum_{i=1}^n \hat{s}_i^2 \hat{u}_i^2/n = \hat{H}\hat{\Sigma}\hat{H}' = A\hat{P}^{-1}\hat{\Sigma}\hat{P}^{-1}A'/V, \mathbb{E}[s_i^2 u_i^2] = A\Sigma A'/V,
$$

$$
\sum_{i=1}^n \hat{\mu}_i^2/n = \hat{H}\hat{\Sigma}_1\hat{H}' = A\hat{P}^{-1}\hat{\Sigma}_1\hat{P}^{-1}A'/V, \mathbb{E}[\bar{\mu}_i^2] = A\Sigma_1 A'/V.
$$

Then by T and Lemmas B9 and B10,

$$
\begin{aligned}
\frac{\hat{V}}{V} - 1 &= \frac{\hat{V} - V}{V} = \frac{A\hat{P}^{-1}\hat{\Sigma}\hat{P}^{-1}A' - A\Sigma A'}{V} + \frac{A\hat{P}^{-1}\hat{\Sigma}_1\hat{P}^{-1}A' - A\Sigma_1 A'}{V} \\
&= \sum_{i=1}^n \hat{s}_i^2 \hat{u}_i^2/n - \mathbb{E}[s_i^2 u_i^2] + \sum_{i=1}^n \hat{\mu}_i^2/n - \mathbb{E}[\bar{\mu}_i^2] \overset{p}{\to} 0. \text{ Q.E.D.}
\end{aligned}
$$

REFERENCES

[37]

Altonji, J., and R. Matzkin (2001), "Panel Data Estimators for Nonseparable Models with Endogenous Regressors", Department of Economics, Northwestern University.

Altonji, J., and H. Ichimura, (1997), "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables," mimeo, Northwestern University.

Angrist, J., G.W. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association 91, 444-472.*

Angrist, J., K. Graddy, and G.W. Imbens (2000): "The Interpretation of Instrumental Variabel Estimators in Simultaneous Equations Models with An Application to the Demand for Fish," *Review of Economic Studies 67*, 499-527.

Athey, S. (2002), "Monotone Comparative Statics Under Uncertainty" *Quarterly Journal of Economics*, 187-223.

Athey, S., and P. Haile (2002), "Identification of Standard Auction Models", *Econometrica* 70, 2107-2140.

Athey, S., and S. Stern, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", NBER working paper 6600.

Bajari, P., and L. Benkard (2001), "Demand Estimation with Heterogenous Consumers and Unobserved Product Characteristics: A Hedonic Approach," unpublished paper, Department of Economics, Stanford University.

Blundell, R., and J.L. Powell (2000): "Endogeneity in Nonparametric and Semiparametric Regression Models," invited lecture, 2000 World Congress of the Econometric Society.

Brown, D., and R. Matzkin, (1996): "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," mimeo, Northwestern University.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 34, 305-334.

Chesher, A. (2001), "Quantile Driven Identification of Structural Derivatives," Cemmap working paper CWP08/01.

Chesher, A. (2002), "Local Identification in Nonseparable Models," Cemmap working paper CWP05/02.

[38]

Darolles, S., J.-P., Florens, and E. Renault, (2001), "Nonparametric Instrumental Regression".

Das, M. (2000): "Nonparametric Instrumental Variable Estimation with Discrete Endogenous Regressors," Working Paper, Department of Economics, Columbia University.

Das, M. (2001): "Monotone Comparative Statics and the Estimation of Behavioral Parameters," Working Paper, Department of Economics, Columbia University.

Doss, H. and R.D. Gill (1992): "An Elementary Approach to Weak Convergence for Quantile Processes, With Applications to Censored Survival Data," *Journal of the American Statistical Association* 87, 869-877.

Hausman, J.A. and W.K. Newey (1995), "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," with J.A. Hausman, *Econometrica* 63, 1445-1476.

Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings* 80.

Heckman, J., and E. Vytlacil, (2000), "Local Instrumental Variables", Chapter 1, in Hsiao, Morimune, and Powell, (eds.) *Nonlinear Statistical Modelling*, Cambridge University Press, Cambridge.

Imbens, G.W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62, 467-476.

Lewbel, A., (2002); "Endogenous Selection or Treatment Model Estimation," unpublished working paper.

Lorentz, G., (1986), *Approximation of Functions*, New York: Chelsea Publishing Company.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80:2, 319-323.

Manski, C. (1995): *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Manski, C. (1997): "The Mixing Problem in Program Evaluation," *Review of Economic Studies* 64, 537-553.

Mark, S, and J. Robins, "Estimating the Causal Effect of Smoking Cessation in the Presence of Confounding Factors Using a Rank-Preserving Structural Failure Time Model," *Statistics in Medicine*, 12, 1605-1628.
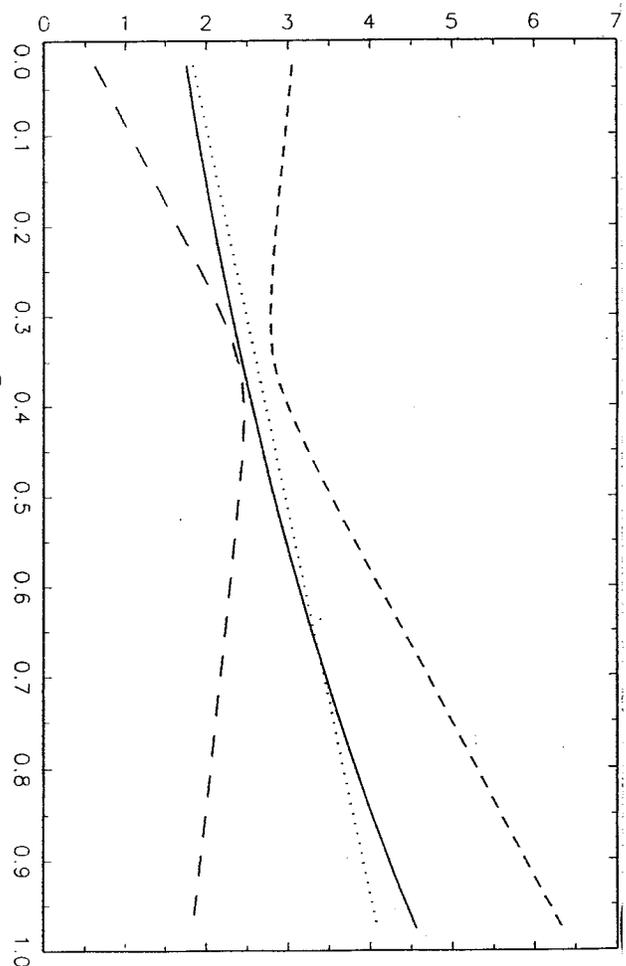
Matzkin, R. (1993), "Restrictions of Economic Theory in Nonparametric Models" *Handbook of Econometrics*, Vol IV, Engle and McFadden (eds.)

Matzkin, R. (1999), "Nonparametric Estimation of Nonadditive Random Functions", Department of Economics, Northwestern University.

Milgrom, P., and C. Shannon, (1994), "Monotone Comparative Statics," *Econometrica*, 58, 1255-1312.

Mundlak, Y., (1963), "Estimation of Production Functions from a Combination of Cross-Section and Time-Series Data," in *Measurement in Economics, Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, C. Christ (ed.), 138-166.

Newey, W.K. (1994), "Kernel Estimation of Partial Means and a Variance Estimator", *Econometric Theory* 10, 233-253.

Newey, W.K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

Newey, W.K. and J.L. Powell (2003): "Nonparametric Instrumental Variables Estimation," *Econometrica*, forthcoming.

Newey, W.K., J.L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica* 67, 565-603.

Pearl, J. (2000), *Causality*, Cambridge University Press, Cambridge, MA.

Pinkse, J., (2000a): "Nonparametric Two-step Regression Functions when Regressors and Error are Dependent," *Canadian Journal of Statistics* 28, 289-300.

Pinkse, J. (2000b): "Nonparametric Regression Estimation Using Weak Separability", University of British Columbia.

Powell, J., J. Stock, and T. Stoker, "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

Robins, J. (1995): "An Analytic Method for Randomized Trials with Informative Censoring: Part 1, *Lifetime Data Analysis* 1, 241-254.

Roehrig, C. (1988): "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica* 55, 875-891.

Schumaker (1981): *Spline Functions*, Wiley, New York.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1481.
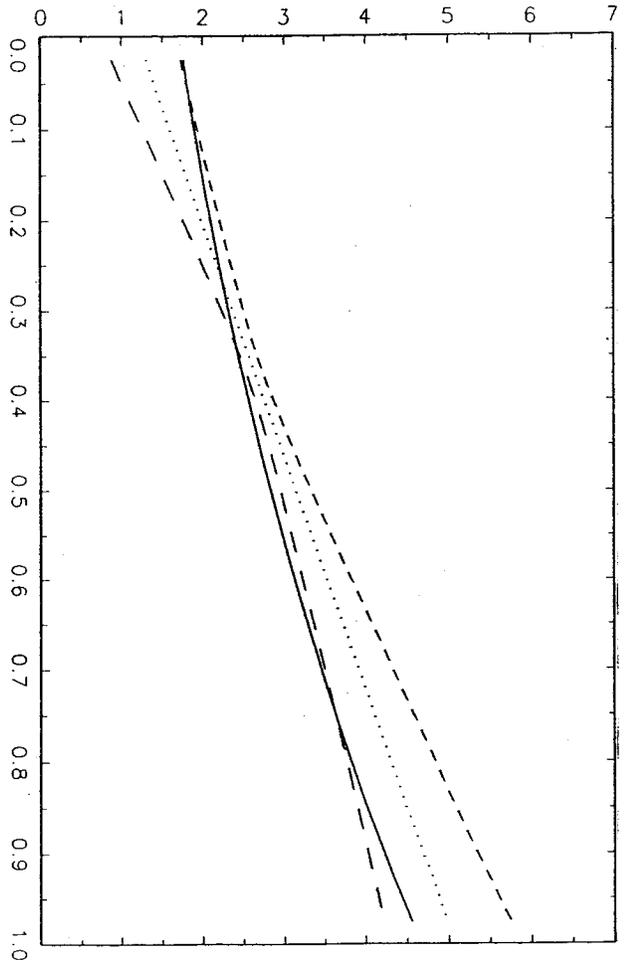
Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Variable Models: An Equivalence Result," *Econometrica* 70, 331-342.
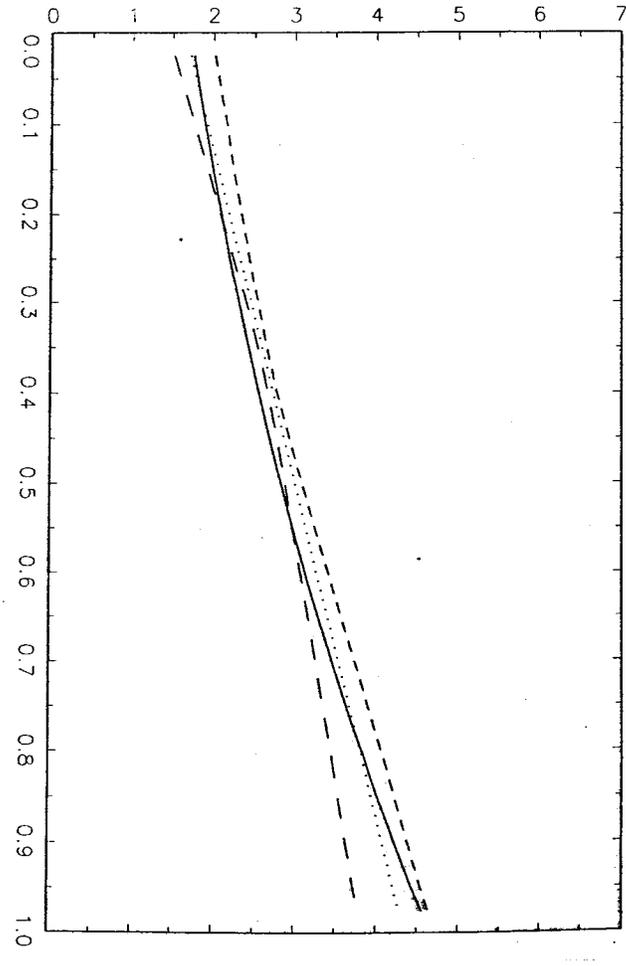
FIGURE ONE