.

INSTRUMENTAL VARIABLES ESTIMATION OF NONPARAMETRIC MODELS WITH DISCRETE ENDOGENOUS REGRESSORS

Mitali Das

Dept of Economics, Columbia University, 420 W 118th street, New York, NY 10027
Telephone: (212) 854-4641, Fax: (212) 854-8059, Email: mitali.das@columbia.edu
This Version: January 2000

Abstract

This paper presents new instrumental variables estimators for nonparametric models with discrete endogenous regressors. The model specification is sufficiently general to include structural models, triangular simultaneous equations and certain models of measurement error. Restricting the analysis to discrete endogenous regressors is an integral component of the analysis since a similar model with continuously distributed endogenous regressors is ill-posed and cannot be identified. The central contribution of this paper is a consistent two-step nonparametric instrumental variables estimator of the model. Large sample results, including global convergence rates and asmptotic normality are also provided.

Discreteness of the regressors is shown to produce an additive representation of the model which leads to a simple verifiable condition for identification, and a restriction that is imposed in estimation. The proposed nonparametric two-step IV estimator is based on series estimation which is particularly amenable to additive models, and yields efficiency gains in imposing additivity. The first step constitutes nonparametric estimation of the instrument, while the second step constructs the IV estimator from a linear combination of an instrument matrix and a matrix of the regression covariates. Linear functionals of the estimator are shown to be asymptotically normal, including $\sqrt{n}$-consistent when certain regularity conditions hold.

Key words: Nonparametric Estimation, Instrumental variables estimation, Series Estimators

1 INTRODUCTION

This paper presents new instrumental variables (IV) estimators for nonparametric models with a univariate discrete endogenous regressor. The focus is on binary endogenous regressors, although an extension to non-binary endogenous regressors with finite support is also analyzed. In the specification, the error terms are not independent of the instruments, but satisfy a weaker conditional mean restriction. The model is sufficiently general to encompass structural models, triangular simultaneous equations systems, and certain nonparametric models with measurement error. A particular motivation for the specification is program evaluation models which frequently arise in economics applications.

Celebrated results for the identification and estimation of structural models such as Koopmans (1949) and Fisher (1966) depend strongly on linear parametrizations that are not part of prior knowledge of the structure. For the nonparametric model considered here, any parametrizations invoked to derive identification results will in general yield incorrect inferences about the identifiability of the model. Similarly, parametrizations for the purpose of estimation will result in inconsistent estimators of the unknown function. We illustrate both identifiability and estimability of the nonparametric model in the absence of any parametric restrictions.

A two-step instrumental variables estimator with an appropriately defined matrix of instruments is shown to be a consistent estimator of the model. The instruments are estimated nonparametrically as conditional means in the first step via series estimation. Nonparametric estimation of the instruments permits bypassing the specification of conditional distributions as in Newey (1990), but is heuristic and the subsequent large sample results for the IV estimator are not dependent on nonparametric estimation in the first step. The second step is a nonparametric generalization of the linear IV estimator, employing a matrix of the instruments and a matrix of the regression covariates to construct the nonparametric instrumental variables estimator.

Semiparametric and nonparametric instrumental variables estimators have been previously considered. Efficient IV estimation of linear models in which the instruments are estimated nonparametrically was proposed by Robinson (1976) and Newey (1990). A generalization of these models to semiparametric regression functions linear in the endogenous variables was considered in Newey (1985b) and Robinson (1988). Roehrig (1988) provided a general treatment of identification in nonparametric models where the errors are independent of the instruments. Estimation of nonparametric structural models was first studied by Newey and Powell (1989), who developed a nonparametric two-stage least squares estimator. However, there is no distribution theory for the proposed estimator, due to an inherent ill-posed nature of the problem. A slightly restrictive version of this model is studied in Newey, Powell and Vella (1999), who derive a two-step nonparametric IV estimator for triangular simultaneous equations models.

This paper extends the aforementioned body of work in a few ways. We consider a general class of models that nests linear and nonlinear specifications, including nonparametric triangular models and systems of equations. Some measurement error models are also shown to be a special case of the model. Motivated by dual considerations, however, this generality is limited by restricting the endogenous regressors to be discrete. First and more importantly, a problem which is evidently general in nonparametric systems of equations with continuously distributed endogenous regressors is a discontinuous mapping from the structural to the reduced form, raising an immediate problem for identification of the infinite-dimensional function (see Newey and Powell, 1989). Although some results may be obtained by restricting the domain of the data, such restrictions preclude estimation of functions with higher-order derivatives and appear to be problem-specific. Results may also be obtained by limiting the analysis to a sub-class of models, as is done in Newey *et al* (1999). Alternatively, this paper illustrates that restricting the endogenous regressors to be discrete suffices in addressing the problem, without sacrificing the generality of the model in other respects. Second, discrete data, which include a wide variety of qualitative and dummy variables, appear to form a large proportion of the available microeconometrics data, suggesting limited drawbacks from analyzing models in which endogeneity stems strictly from discrete regressors.

The key results of this paper rest on the existence of a transform which converts the infinite-dimensional problem to a finite-dimensional one, and is critical in deriving a one-to-one map between the structure and the reduced form, *i.e*, for identification. The transform parametrizes the endogenous component by extracting the discrete regressor from the unknown function, which is subsequently evaluated at each of the discrete variable's support points. This idea is illustrated by a simple binary regressor example below. Let d denote the binary regressor, and x represent a vector of continuously distributed variables. Then,

$$\begin{aligned} y &= m(x,d) + \varepsilon \\ &\equiv \bar{m}(x)d + \tilde{m}(x)(1-d) + \varepsilon \\ &= \tilde{m}(x) + [\bar{m}(x) - \tilde{m}(x)]d + \varepsilon \\ &= \alpha(x) + \beta(x)d + \varepsilon \end{aligned} \tag{1}$$

where ε denotes an error term and $m(\cdot)$ represents the nonparametric relation between y and (x,d). Equation (1) shows that an equivalent representation of the model is a specification in which $m(\cdot)$ is a linear combination of two additively separable functions $\alpha(\cdot)$ and $\beta(\cdot)d$. In Section 3, the above transform is utilized in deriving identification of the model. Although the specification is not *ex ante* additive, this representation of the model motivates the application of series estimators, which yield efficiency gains from imposing additivity.

The remainder of the paper is organized as follows. Section 2 introduces the model, motivating the specification with a number of examples. Section 3 discusses identification, deriving a

simple necessary condition that is verifiable from the data. Section 4 develops the two-step nonparametric IV estimator and illustrates that it is asymptotically well-defined. Mean square error (MSE) and uniform convergence rates are presented in Section 5. Section 6 derives asymptotic normality of linear functionals of the estimator, and discusses the construction of a consistent estimator of the asymptotic variance. Three extensions of the model are considered in Section 7. The first is to non-binary discrete endogenous variables with finite support, the second is to a semiparametric specification and the third extension is to a model with no exogenous covariates. Conclusions follow. All proofs are presented in an appendix to the paper.

2 THE CONDITIONAL MEAN MODEL

The model studied in this paper is summarized in the following way:

$$\begin{aligned} y &= m_o(x_1, d) + \varepsilon \\ E[\Lambda_o(v; m_o)|x] &= 0 \end{aligned} \tag{2}$$

where $v \equiv (y, x, d)$ denotes the data, d is a univariate dummy endogenous variable[1], x is a $s_x \times 1$ vector of instrumental variables which includes x_1 as a $s_1 \times 1$ subvector, m_o is the infinite-dimensional object of estimation, and $\Lambda_o(v; m_o)$ is a residual term which may coincide with ε in specific applications.

Using the transform described in equation (1), an equivalent representation of the regression model is

$$\begin{aligned} y &= m_o(x_1, d) + \varepsilon \\ &= \alpha_o(x_1) + \beta_o(x_1)d + \varepsilon. \end{aligned} \tag{3}$$

This model is a generalization of the limited information simultaneous equations model and focuses on the estimation of a single equation that may be part of a system of equations. The specification covers as special cases the standard linear and nonlinear simultaneous equation systems. For these cases, the conditional mean restriction in (2) is stronger than the typical zero covariance of the residual and instrument restriction necessary for consistent estimation. The generality of the above model is illustrated through several examples.

2.1 Example 1: Structural Estimation

Consider a nonparametric structural model, for example the classical model of supply and demand, where y and d respectively denote equilibrium prices and (discrete) quantities, and y and d are jointly determined in equilibrium as follows:

[1]For the remainder of the analysis, to simplify the exposition, d will denote a dummy variable. The extension to non-binary discrete variables will be considered in Section 7.

$$y = m_o(x_1, d) + \varepsilon, \qquad E[\varepsilon|x_1, x_2] = 0 \tag{4}$$
$$d = h_o(x_2, y) + \omega, \qquad E[\eta|x_1, x_2] = 0. \tag{5}$$

Here, x_1 and x_2 represent exogenous variables such as weather and income respectively and ε and ω are unobservable error terms.

This model may be thought of as the nonparametric generalization of dummy endogenous variable models studied previously (Heckman, 1978). Suppose the inverse demand function m_o is the object of estimation.[2] For instance, an estimate of m_o would be required as an intermediary step in the computation of the deadweight loss associated with the imposition of a tax (*e.g*, Hausman and Newey, 1995).

Equation (4) is a special case of (2) with $x = (x_1, x_2), v = (y, x, d)$ and $\Lambda_o(v; m_o) = y - m_o(x_1, d)$. Parallel methods to those described in the body of this paper for the estimation of m_o may be utilized to estimate h_o only if y is a discrete variable with finite support. Where y is continuously distributed, the mapping from the structural to the reduced form of h_o is discontinuous and prevents the construction of a consistent nonparametric estimator of h_o.[3]

2.2 Example 2: Triangular Simultaneous Equations Model

A second example of our generalized model is a nonparametric triangular simultaneous equations system, a model which commonly arises in program evaluation studies (*e.g*, Moffitt and Wolfe, 1992; Poterba, Venti and Wise, 1996; Heckman, 1997). For $x = (x_1, x_2)$, let

$$\begin{aligned} y &= m_o(x_1, d) + \varepsilon, & E[\varepsilon|\eta, x] = E[\varepsilon|\eta], \quad E[\varepsilon] = 0 \\ d &= \pi_o(x) + \eta, & \eta \text{ independent of } x \end{aligned} \tag{6}$$

where the object of estimation is m_o, and the second equation is is the reduced form for d. One well-known example of this model is Angrist (1990) where y denotes the post-war wage, d denotes the endogenously determined decision to draft in the Vietnam war, and x are conditioning variables. An implication of the stochastic restrictions is that $E[\varepsilon|x] = 0$.[4]

[2] For the remainder of the paper we will adopt the following convention: to represent the entire function, we will suppress the argument, *e.g*, m_o, while $m_o(x_1, d)$ will denote the value of the function at a point.

[3] The structural and reduced forms are related by the equation $E(d|x) = \int h_o(x_2, y) f(x_2, y|x) dy$, where $f(\cdot)$ is a conditional density. Since both $E(d|x)$ and $f(\cdot)$ may be consistently estimated from the data, they can be treated as known components. Identification of the equation therefore rests on solving this linear integral equation to obtain a unique solution for h_o. A problem arises because this is an example of a linear integral equations of the first kind, for which it is difficult to establish a continuous map from h_o to $E(d|x)$ (see Newey and Powell (1989) for more details).

[4] This follows since $E(\varepsilon|x) = E(E(\varepsilon|\eta, x)|x) = E(E(\varepsilon|\eta)|x) = E(E(\varepsilon|\eta)) = E(\varepsilon) = 0$

This model is an example of (2) with $v = (y, x, d)$, and $\Lambda_o(v; m_o) = y - m_o(x_1, d)$. Conditions under which it is identified, and a consistent procedure to estimate it, will be discussed in the following sections. Estimation of a similar model (without the assumption that $E[\varepsilon] = 0$ and with $E[\eta|x] = 0$ in lieu of the independence assumption) is considered in Newey *et al* (1999); our results are complementary to those.

2.3 Example 3: Misclassification of a Dummy Regressor

A third example illustrates the significance of specifying a general residual term Λ_o. This is an example of nonparametric errors-in-variables, in which endogeneity is induced by misclassification of the dummy regressor. For example, miscoding $d = 1$ when the true value is zero. Let $x = (x_1, x_2)$, and

$$\begin{aligned} y &= m_o(x_1, d^*) + \varepsilon, & E[\varepsilon|x] = 0 \\ d &= d^* - \nu, & E[\nu|\varepsilon, x] = 0 \\ d^* &= \pi_o(x) + \eta, & \eta \text{ distributed independently of } x \end{aligned} \tag{7}$$

where d^* is the true unobserved response that is misclassified with some probability ψ, $(0 \leq \psi < 1)$, d denotes the misclassified response, ν is misclassification error and the third equation depicts a causal relation between the unobserved variable d^* and the instrumental variables, x.

Note that using equations (1) and (7),

$$\begin{aligned} y &= m_o(x_1, d^*) + \varepsilon \\ &\equiv \alpha_o(x_1) + \beta_o(x_1)d + \beta_o(x_1)\nu + \varepsilon. \end{aligned} \tag{8}$$

For $v = (y, x, d)$, this model fits equation (2) since $\Lambda_o(v; m_o) = y - \alpha_o(x_1) - \beta_o(x_1)d$ is mean zero conditional on x given the assumed restrictions on the error terms. A simple and consistent method of estimating this model is presented below. The result is complementary to those suggested for polynomial errors in variables model (*e.g*, Hausman *et. al*,1985, Hausman *et. al*,1989), but considers a slightly more general specification and is simpler to implement.

The above examples illustrate the generality of the model. Below, we present a two-step nonparametric IV estimator of this model. Analogous to the parametric IV estimator, the first step consists of constructing the instrument $\pi_o(x) \equiv E[d|x]$ via a nonparametric regression of d on x. The second step does not involve nonparametric regression, but instead employs a matrix of the functions of the instrumental variables along with a conformable (for matrix multiplication) matrix of (x_1, d) to generate an estimate of the structural function $m_o(x_1, d)$. In this form, the proposed estimator differs from nonparametric two-step estimators previously considered in the literature (*e.g*, Ahn 1995, Das, Newey and Vella 1999, Newey *et al* 1999).

3 IDENTIFICATION

This section develops theorems for the identification of the function of interest, m_o. The conditions and theorems presented are nonparametric in the sense that they pertain to the identification of a functional relationship, and not of parameters. The results presented here build on the early work of Koopmans (1949), Fisher (1966) and Brown (1983) for parametric models, and include them as special cases.

The first result is a nonparametric generalization of the rank condition that is sufficient for the identification of m_o. Partition x as $x = (x_1, x_2)$. Let $v = (y, x, d)$ as defined before, and define $\omega = d - \pi_o(x)$. Let d_y denote $\dim(y)$, $f = [f_1,\ f_2]' = [y - m_o(x_1, d) - \varepsilon,\ d - \pi_o(x) - \omega]'$ $f^* = [f_1^*,\ f_2^*] = y - m^*(x_1, d) - \varepsilon,\ d - \pi^*(x) - \omega]'$where m^* represents any other function that satisfies (2) and let $\Delta = [\partial f_1/\partial v,\ \partial f^*/\partial v]'$.

Theorem 3.1 *For the model in (2), if $m_o(x_1, d)$ is differentiable, rank $(\Delta) < d_y + 1$ and rank(x_2) =1 everywhere on $\mathcal{R}$, then $m_o(x_1, d)$ and $m^*(x_1, d)$ are observationally equivalent,* and *$m_o(x_1, d)$ is identified up to an additive constant.*

Proof: Appendix.

Theorem 3.1 is the standard rank condition that, in the linear simultaneous equation model, is both necessary and sufficient for identification (see Koopmans 1949, p.168-169). This result is closely related to Theorem 5.1 of Roehrig (1988), but does not require independence of the error term. While the stated rank condition is sufficient, it is not necessary for identification in a nonlinear model where the absence of any excluded exogenous variables, *i.e*, $\text{rank}(x_2) < 1$, can suffice for identification. For illustration, consider a simultaneous equation system, with no excluded exogenous variables:

$$\begin{aligned} y &= m_o(x_1, d) + \varepsilon_1, \qquad & E[\varepsilon_1 | x_1] = 0 \\ d &= h_o(x_1, y) + \varepsilon_2, \qquad & E[\varepsilon_2 | x_1] = 0. \end{aligned} \tag{9}$$

For m_o and h_o linear in the variables, this structural form is not identifiable. However, identification of this model can be obtained through nonlinearities in m_o and h_o. Consider identification of the first equation. By Theorem 3.1, the rank condition for identification of $m_o(x_1, d)$ is satisfied if

$$|\Delta| = 0 \quad where \Delta = \begin{bmatrix} \partial f_1/\partial y & \partial f_1/\partial d & \partial f_1/\partial x_1 \\ \partial f_1^*/\partial y & \partial f_1^*/\partial d & \partial f_1^*/\partial x_1 \\ \partial f_2^*/\partial y & \partial f_2^*/\partial d & \partial f_2^*/\partial x_1 \end{bmatrix}. \tag{10}$$

It follows that so long as $\partial f_2^*/\partial y$ or $\partial f_2^*/\partial x_1$ depend on y, the unique solution to (10) is $\partial m_o/\partial d = \partial m^*/\partial d$ and $\partial m_o/\partial x_1 = \partial m^*/\partial x_1$, which implies that $m_o(x_1, d)$ is identified, up to an additive constant.[5]

Identification of the model can also be linked to some of the results in the program evaluation literature. By (1) and (2)

$$\begin{aligned} E[y|x] &= E[m_o(x_1, d)|x] \\ &= \alpha_o(x_1) + \beta_o(x_1)\pi_o(x) \\ &= g_o(x). \end{aligned} \tag{11}$$

Since the function of interest is the sum of the two additive components α_o and $\beta_o\pi_o$, identification of m_o rests on the identification of each of these components. Note that conditional expectations are unique with probability one, implying that $g_o(x)$ and $\pi_o(x)$ are identified and may be thought of as known in analyzing the identification of m_o.

Without loss of generality, assume $x_2 \in (0, 1)$.

Lemma 3.1 *If* $0 < Var(\pi_o(x)|x_1) < \infty$, *then* $\beta_o(x_1)$ *is identified.*

Proof Since

$$\begin{aligned} E[y|x_1, x_2 &= 0] = \alpha_o(x_1) + \beta_o(x_1)E[d|x_1, x_2 = 0] \\ E[y|x_1, x_2 &= 1] = \alpha_o(x_1) + \beta_o(x_1)E[d|x_1, x_2 = 1], \end{aligned}$$

it follows that

$$\frac{E[y|x_1, x_2 = 0] - E[y|x_1, x_2 = 1]}{E[d|x_1, x_2 = 0] - E[d|x_1, x_2 = 1]} = \beta_o(x_1). \tag{12}$$

The Wald representation in equation (12) is an explicit expression for the component $\beta_o(x_1)$ in terms of the identified conditional expectations $E[y|x_1, x_2 = i]$ and $E[d|x_1, x_2 = i]$, $i = 0, 1$. Therefore, the key condition underlying the identification of $\beta_o(x_1)$ is that the denominator in (12) be bounded away from zero. This requires that there be sufficient variation in the conditional expectation $\pi_o(x)$ for different realizations of x_2, or $Var(\ \pi_o(x)|x_1) > 0$.

Although m_o is the stated object of estimation, frequently $\beta_o(x_1)$ is of interest as this component represents the incremental effect on the conditional mean of y for the subset of observations where $d = 1$ (the "treated" group). For example, in Angrist (1990) this component measures

[5] Where f and f^* are not continuously differentiable, assume that the discrete generalizations hold, *e.g*, $\Delta m_o = m_o(x_1, d + \Delta d)/\Delta d$

the earnings loss to Vietnam-era veterans relative to their civilian counterparts. In other program evaluation studies, $\beta_o(x_1)$ might measure the effects of Medicaid on labor supply (*e.g*, Moffitt and Wolfe, 1992) or wage effects of participation in the JTPA program (*e.g*, Heckman and Smith,1997). In an application where d is not endogenous, $\beta_o(x_1)$ represents the abnormal share price response to a firm that undertakes a leveraged buyout (*eg*, Chevalier 1994).

Contrary to a parametric specification, the above formulation allows the incremental effect on the "treated" population to differ by observation. A useful metric is an average of β_o over x_1, which is a functional we study in Section 6. Note that the condition in equation (12), which is a necessary condition for identification, is easily verified from the data. Based on Lemma 1, we have the following theorem for the identification of $m_o(x_1, d)$.

Theorem 3.2 *If Lemma 3.1 is satisfied, then $m_o(x_1, d)$ is identified up to an additive constant.*

Proof By Lemma 3.1, $\beta_o(x_1)$ is identified. Since $h_o(x)$ and $\pi_o(x)$ are identified by the uniqueness of conditional expectations, then $\alpha_o(x_1) = g_o(x) - \beta_o(x_1)\pi_o(x)$ is identified, implying that each of the additive components of $m_o(x_1, d)$ is identified. From this it follows that $m_o(x_1, d)$ is identified up to an additive constant.

4 ESTIMATION

Estimation of (2) differs from regular nonparametric regression in two ways. First, estimation of $m_o(x_1, d)$ is not equivalent to the estimation of a conditional mean due to the correlation of d with the residual. Second, estimation of m_o does not correspond to a nonparametric generalization of two-stage least squares estimation, *i.e*, direct nonparametric estimation of m_o by replacing (x_1, d) with $(x_1, \pi(x))$. As is well known, the latter is invalid since replacing nonlinear functions of the endogenous regressors with nonlinear functions of their predicted values results in inconsistent estimation of m_o. Rather, in the estimation scheme outlined below, we present a generalization to the linear IV estimator, incorporating the restriction that $m_o(x_1, d)$ is additively separable as $\alpha_o(x_1) + \beta_o(x_1)d$.

A two-step IV procedure is proposed for the estimation of the model. As outlined above, the first step consists of a nonparametric regression of d on the instruments x to yield an estimate of the instrument $\hat{\pi}(x_i) = \hat{E}[d_i|x_i]$, $(i = 1, ..., n)$; the estimated $\hat{\pi}(x_i)$ are used in forming the instrument matrix. The second step is the construction of an IV estimator using the instrument matrix and a matrix of functions of (x_1, d), to generate an estimator of $m_o(x_1, d)$. This estimator is shown to be additive, consisting of a component that approximates $\alpha_o(x_1)$, and a component that approximates $\beta_o(x_1)d$.

All results are presented for series estimators which provide a convenient method of imposing the additive form present in (1). Alternative nonparametric estimators (*e.g*, kernel or nearest-neighbour estimation) could be applied as well. However, the additive structure is especially suitable to series estimation, where imposing additivity is computationally simple, and known to lead to higher efficiency and convergence rates of the estimator (Stone 1985, Andrews and Whang 1991).

In principle, first step estimation of the conditional expectation $\pi(x)$ could be carried out without nonparametric methods, for example, by least squares. A well-known result is the consistency of the IV estimator for arbitrary non-zero correlation asymptotically between the endogenous regressor and instrument (Newey, 1990). For example, if $d = \mathbf{1}(x'\gamma - \nu > 0)$ where $\mathbf{1}(\mathcal{A})$ is the indicator function for $\mathcal{A}$, and ν is known to be a normally distributed error term, maximum likelihood estimation of the model could generate the instrument $\Phi(x'\widehat{\gamma})$. For other models, however, estimation of $\pi(x)$ could require knowledge of the conditional distribution of d given x, and construction of the conditional expectation could require cumbersome computation. By employing nonparametric methods to estimate π, we avoid difficult computation or the reliance on distributional assumptions.

Consider the first step. For an integer $L > 0$, define an $L \times 1$ vector of approximating functions $r^L(x) = (r_{1L}(x), ..., r_{LL}(x))'$. The vector $r^L(x)$ will be used to approximate π, with the property that for large L, the approximation to π gets arbitrarily close in the Euclidean norm. As further discussed below, the approximating sequence in this paper will consist primarily of polynomials (*e.g*, power series), although some additional results will be provided for regression splines. Let n denote the number of observations. For $i = 1, ...n$, the first step series estimate $\widehat{\pi}_i = \widehat{\pi}(x_i)$ is obtained as

$$\widehat{\pi}_i = r^L(x_i)'\widehat{\gamma}, \qquad \widehat{\gamma} = (R'R)^- R'd, \qquad R = [r_1, ..., r_n]' \tag{13}$$

where $d = (d_1, ..., d_n)$, $\widehat{\gamma}$ is the least squares estimator from regressing d on $r^L(x)$, and $(R'R)^-$ represents a generalized inverse. This constitutes the first step of estimation.

For the second step, define a $K \times 1$ $(K > 0)$ vector of approximating functions $\bar{p}^K(x_1, d) = (p^K(x_1), p^K(x_1, d))'$, where $p^K(x_1)$ and $p^K(x_1, d)$ are each $(K/2 \times 1)$subvectors of the approximating sequence. Let the first element of $p^K(x_1)$ be the unit vector. To reflect the underlying additively separable form of the model, let the next $(K-1)/2$ functions in $p^K(x_1)$ depend only on x_1, and the $(K/2)$ functions in $p^K(x_1, d)$, comprise of interactions of each of the $(K-1)/2$ functions in $p^K(x_1)$ with d. Let k index the elements of the subvectors $p^K(x_1)$ and $p^K(x_1, d)$. It should be emphasized that in this formulation there are an equal number of elements in $p^K(x_1)$ and $p^K(x_1, d)$, and in particular, $p_k^K(x_1, d) = p_k^K(x_1)d$, $(k = 1, ..., K/2)$.

To specify the instrument matrix, we define an additional $K \times 1$ vector of approximating functions $\bar{q}^K(x_1, \pi(x)) = (q^K(x_1), q^K(x_1, \pi(x)))'$. The elements of the instrument matrix will be

isomorphic to $\bar{p}^K(x_1, d)$, with the instrument $\pi(x)$ replacing d, wherever it appears in $\bar{p}^K$. Set the subvector $q^K(x_1)$ to be equal to $p^K(x_1)$. Next, let the subvector $q^K(x_1, \pi(x))$ constitute the product of each of the elements of $q(x_1)$ with $\pi(x)$, implying that $q_k^K(x_1, \pi(x)) = q_k^K(x_1)\pi(x) = p_k^K(x_1)\pi(x)$, $(k = 1, ..., K/2)$.

As the predicted value from the first step is an estimated probability, it is useful to bound the values of $\hat{\pi}(x)$. Further, since polynomial series estimation yield poor approximations in the presence of outliers, restricting the domain of x_1 may also be important.[6] Let $w = (x_1, d, \pi(x))$. Let $\tau(w)$ be a trimming function, defined as

$\tau(w) = \prod_{j=1}^{s_x} 1(\mathrm{a}_j \leq x_{1j} \leq b_j)\mathbf{1}(a_{s_x+1} \leq \pi_j \leq b_{s_x+1})\mathbf{1}(a_{s_x+1} \leq d_j \leq b_{s_x+1})$ where a_j and b_j are either pre-specified constants or are themselves estimated, and $0 < a_{s_x+1} < b_{s_x+1} < 1$. For i indexing the observations, let $\hat{\tau}_i = \hat{\tau}(x_{1i}, d_i, \widehat{\pi}_i)$, $\hat{q}_i = \bar{q}^K(x_{1i}, \hat{\pi}_i(x))$, and $p_i = \bar{p}^K(x_{1i}, d_i)$. The instrumental variables estimator for $m_o(x_1, d)$ is defined by

$$\hat{m}(x_1, d) = \bar{p}^K(x_1, d)'\hat{\delta}, \qquad \hat{\delta} = (\hat{Q}'\hat{P})^- \hat{Q}'y \tag{14}$$

where $\mathrm{y}=[\mathrm{y}_1, ..., y_n]'$ $\hat{P} = [\hat{\tau}_1 p_1, ..., \hat{\tau}_n p_n]' = \begin{bmatrix} \hat{\tau}_1 p^K(x_{11}), & \hat{\tau}_1 p^K(x_{11}) d_1 \\ \vdots & \vdots \\ \hat{\tau}_n p^K(x_{1n}), & \hat{\tau}_n p^K(x_{1n}) d_n \end{bmatrix}$

$$\hat{Q} = [\hat{\tau}_1 \hat{q}_1, ..., \hat{\tau}_n \hat{q}_n]' = \begin{bmatrix} \hat{\tau}_1 p^K(x_{11}), & \hat{\tau}_1 p^K(x_{11}) \hat{\pi}_1 \\ \vdots & \vdots \\ \hat{\tau}_n p^K(x_{1n}), & \hat{\tau}_n p^K(x_{1n}) \hat{\pi}_n \end{bmatrix}.$$

The estimator $\hat{\delta}$ can be used to form estimators of either $\alpha_o(x_1)$ or $\beta_o(x_1)$ by extracting the subvector of $\bar{p}^K(x_1, d)$ that depends on only x_1 or the subvector that depends on both x_1 and d, respectively. Partition δ as $\hat{\delta} = (\hat{\delta}_\alpha', \hat{\delta}_\beta')'$ where $\hat{\delta}_\alpha$ corresponds to the first $K/2$ elements of $\hat{\delta}$, and $\hat{\delta}_\beta$ corresponds to the remaining terms. Then,

$$\hat{\alpha}(x_1) = p^K(x_1)'\hat{\delta}_\alpha, \qquad \hat{\beta}(x_1) = p^K(x_1) d' \hat{\delta}_\beta. \tag{15}$$

The constant term is identified, and its estimator is the first element of $\hat{\delta}_\alpha$, which corresponds to the unit vector in $p^K(x_1)$. Note that a constant is not estimated in $\hat{\beta}(x_1)$, since $\beta(x_1)$ represents the incremental effect on y for the subset of observations where $d = 1$.

For the proposed IV estimator to be asymptotically well-defined, a requisite condition is that the second moment matrix of approximating functions $\hat{Q}'\hat{P}$ be uniformly bounded away

[6] Although $d = \{0, 1\}$, we include it in the trimming function to be general. This generality is useful when the definition of d is extended to the case of discrete d with finite support.

from singularity. We show below that a necessary condition for this matrix to be non-singular is $Var(\pi(x)|x_1) > 0$. Note that $Var(\pi(x)|x_1) > 0$ is verifiable, and a necessary condition for identification as given in Lemma 3.1, so that once $m_o(x_1, d)$ is identified, the proposed estimator will be well-defined.

Let $G = E[Q'P]$. By $G = E[E[G|x]]$, G=$\begin{bmatrix} E[p^K(x_1)'p^K(x_1)] & E[p^K(x_1)'p^K(x_1)]E[d|x] \\ E[p^K(x_1)'p^K(x_1)\pi(x)] & E[p^K(x_1)'p^K(x_1)\pi(x)]E[d|x] \end{bmatrix}$

=E$\left[\begin{bmatrix} 1 & \pi(x) \\ \pi(x) & \pi^2(x) \end{bmatrix} \otimes p^K(x_1)'p^K(x_1)\right] = E\left[\pi \otimes p^K(x_1)'p^K(x_1)\right]$

=E[E$\left[\pi \otimes p^K(x_1)'p^K(x_1)\right] |x_1] = E[[\pi|x_1] \otimes p^K(x_1)'p^K(x_1)]$.

Under standard regularity conditions given below, $E[p^K(x_1)'p^K(x_1)]$ will have full column rank, so that the non-singularity of G depends strictly on bounding the smallest eigenvalue of $\pi|x_1$ (denoted $\lambda_{\min}(\pi|x_1)$), below by zero. Evaluating the determinant of $\pi|x_1$, we have

$$|E[\pi|x_1]| = E[\pi^2(x)|x_1] - E[\pi(x)|x_1]^2 = Var(\pi(x)|x_1) > 0 \tag{16}$$

since by Lemma 3.1 (necessary for identification), $\exists\ \xi, (0 < \xi < \infty)$ such that $Var(\pi(x)|x_1) \geq \xi$. However, equation (16) is insufficient to show that $\lambda_{\min}(\pi|x_1)$ is strictly positive (as required by positive definiteness of G). Represent the vectors of eigenvalues of $\pi|x_1$ as $\lambda_1(\pi)$ and $\lambda_2(\pi)$. Since the trace of $(\pi|x_1)$ is the sum of its eigenvalues,

$$\lambda_1(\pi) + \lambda_2(\pi) = 1 + [\pi^2(x)|x_1] > 0. \tag{17}$$

Further, by implication of equation (16),

$$\lambda_1(\pi)'\lambda_2(\pi) = |E[\pi|x_1]| > 0. \tag{18}$$

Equations (17) and (18) jointly imply that $\lambda_1(\pi) > 0, \lambda_2(\pi) > 0$, from which it follows that $\lambda_{\min}(\pi|x_1) > 0$. Hence, the proposed IV estimator is well-defined asymptotically.

The two series estimators we consider are power series and smooth piecewise polynomials, or splines, with evenly spaced break-points. Such estimators have been considered previously in the literature (Porter 1996, Newey *et al* 1999). Power series are known to provide good approximations to smooth functions and are simple to compute, although they may be adversely affected by outliers in the data. Consider the first stage approximating sequence, $r^L(x)$. A power series approximation is typically modeled as increasing powers of a single function, for example,

$$r_{lL} = f(x)^{l-1}. \qquad l = 1, \ldots, L; \qquad L = 1, 2, \ldots \tag{19}$$

for some function f that is chosen according to the context of the model. A convenient choice for our first stage is f with bounded range (as similarly suggested in the nonparametric selection model by Das, Newey and Vella, 1999),

$$r_{lL} = \mathcal{F}(x)^{l-1} \tag{20}$$

where $\mathcal{F}(x) = \left[\frac{\phi(\Phi^{-1}(x)}{\pi}\right]$, $\phi(\cdot)$ is the standard normal density and $\Phi(\cdot)$ represents the standard normal distribution function. Comprehensive discussions of these and alternative(*e.g*, orthonormal polynomials) power series estimators are presented in Andrews (1991) and Newey (1997).

We also provide results for spline estimation. Splines are smooth piecewise polynomials with fixed joining points (or *knots*) for the polynomial. An advantage of splines relative to power series are the better approximations when either the underlying function is assumed to be discontinuous, or outliers are present. The knots may be placed at points where the underlying function is changing rapidly, or they may simply be placed equidistant. A typical m^{th} degree spline for univariate x, with J known knot points, $j_1, \ldots, j_J$, may be expressed as

$$\rho_{jJ} = x^{j-1}, 1 \leq j \leq m+1 1(x \geq j_i)(x - j_i)^m, 1 \leq i \leq J \tag{21}$$

For multivariate x, the approximating functions are products of the functions in the single variable case. We restrict ourselves to evenly spaced knots in the support of x whose range is therefore treated as known. For exposition consider the case where the support of x is on the interval $[-1, 1]$. Denote for a scalar c, $(c)_+ = 1(c > 0)c$. An m^{th} degree spline with $J-1$ knots is defined to be a linear combination of

$$\rho_{jJ} = x^{j-1},\ 1 \leq j \leq m+1([x+1-2(j-m-1)/J])_+^m,\ m+2 \leq j \leq m+J \tag{22}$$

For $s_X \equiv \dim(x)$, let $\mu = (\mu_{\ 1} ... \mu_{1X})$ denote a (1 x s_X) vector of non-negative integers, and let $\mu(l)_{l=1}^{\infty}$ denote a sequence of such vectors. Then, for a set of vectors $\mu(l)$ with the restriction that $\mu_j(l) \leq (m+J)$ for each j and each l, the approximating series can be expressed as

$$r_{lL} = \prod_{j=1}^{s_X} \rho_{\mu_j(L),J(j)}(x_j),\ k = 1, \ldots K \tag{23}$$

where J_j represents the number of knots for the j^{th} component of x. Similarly, in the second step, spline approximations can be applied by imposing additivity by considering terms that depend only on x_1 or (x_1, π), but no combinations of the two.

5 CONVERGENCE RATES

The integrated mean-square error and uniform convergence rates of the estimator $\hat{m}(x_1, d)$ are derived in this section. These results are complementary to, and share several features of, rates previously derived in Stone (1982), Andrews (1991) and Newey (1995, 1997). However, important differences arise here as well. In particular, we illustrate that the rate at which the first step is estimated is irrelevant for the second step MSE convergence rate, yielding the same result for the second step irrespective of the rate at which $\pi(x)$ is esimated. However, the estimator does not attain Stone's (1982) optimal rates, in contrast to alternative nonparametric two-step series estimators (*e.g*, Das *et. al.*, 1999, Newey *et. al*,1999).These results are discussed below.

To derive the results, some regularity conditions are imposed on the model. For a random matrix A, let $\|A\|_v = E[\|A\|^v]^{1/v}$ $\forall v < \infty$, and let $\|A\|_\infty$ denote the infimum of constants C, such that $\Pr(\|A\|_v < C) = 1$. For a matrix D let $\|D\| = [tracE[D\prime D)]^{1/2}$. Let $X = (x_1, d)$, $\mathcal{X}$ denote the support of X and $\mathcal{W}$=$\{w : \tau(w) = 1\}$. Our first assumption is a standard bounded conditional variance assumption.

Assumption 1*: The triple* $\{(y_1, x_1, d_1\}, ...(y_n, x_n, d_n)\}$ $(i = 1, ..., n)$ *is i.i.d, and* $Var(y|w)$ *and* $Var(d|x)$ *are bounded.*

The next assumption bounds each of the approximating sequences $\bar{p}^K$ and $\bar{q}^K$ away from singularity, also limiting the rate at which the supremum norms of the approximating functions grow. This condition is required, in conjunction with equations (17)-(18), to ensure nonsingularity of the second moment matrix of approximating functions.

Assumption 2: *For every* K *there are nonsingular matrices* B_P *and* B_Q *such that for* $\bar{P}^K(x_1, d) = \bar{p}^K(x_1, d)B_P$ *and* $\bar{Q}^K(x_1, \pi(x)) = \bar{q}^K(x_1, \pi(x))B_Q$: (i) *the smallest eigenvalues of* $E[\bar{P}^K(x_1, d)\bar{P}^K(x_1, d)']$ and $E[\bar{Q}^K(x_1, d)\bar{Q}^K(x_1, d)']$ *are bounded away from zero uniformly in* K*, and* (ii) *for every* $\mu \geq 0$*, there are the sequence of constants* $\zeta_\mu(K)$ *and* $\Xi_\mu(K)$ *satisfying* $\max_{|g|\leq\mu} \sup_{X\in\mathcal{X}} ||\partial^g \bar{P}^K(x, d)|| \leq \zeta_\mu(K)$*, and* $\max_{|g|\leq\mu} \sup_{X\in\mathcal{X}} ||\partial^g \bar{Q}^K(x, d)|| \leq \Xi_\mu(K)$.

Recall that $\dim(x_1) = s_1$. Our next assumption regulates the bias of the estimator and its derivatives by specifying the rate at which the uniform approximation error to m_o and its derivatives falls as K grows. This rate is controlled by both the smoothness of the function and its dimensionality, $\bar{s} = s_1 + 1$.

Assumption 3: $m_o(x_1, d)$ is continuously differentiable of order ξ on $\mathcal{X}$, and there exists μ, ψ, (δ_K) such that $\max_{|g|\leq\mu} \sup_{X\in\mathcal{X}} |\partial^g(m_o - \bar{p}^K(x, d)'\delta_K)|_\mu = O(K^{-\psi})$ as $K \to \infty$.

For the case where $\mu = 0$, it follows from Newey (1997; p. 157) that $\psi = \xi/\bar{s}$. Let $F_o(w)$ denote the cumulative distribution function of w, and $\mathcal{W}_1$ denote the coordinate projection of $\mathcal{W}$ on (x_1, d).

Theorem 5.1
If Assumptions 1-3 are satisfied, $\mu = 0$ and $\zeta_o(K)^2 K/n \to 0$, *then*

$$\int \tau(w)[\hat{m}(x_1, d) - m_o(x_1, d)]^2 dF_o(w) = O_p(K/n + K^{1-2\psi})$$

and for $q = 0.5$ for splines, and $q = 1$ for power series

$$\sup_{X \in \mathcal{W}_1} |\hat{m}(x_1, d) - m_o(x_1, d)| = O_p(K^q[\sqrt{K/n} + K^{1-\psi}])$$

Proof: Appendix

The first result of Theorem 5.1 illustrates that the MSE rate has a structure similar to the optimal rate derived in Stone (1982), although it is the presence of the larger bias term here which prevents the estimator from reaching the optimal rate. In particular, if we consider choosing $K = n^{1/2\psi}$ (at which the variance and bias components approach zero at the same rate), then the convergence rate of the estimator is $n^{(1-2\psi)/\psi}$. For $\mu = 0$, and $\psi = \xi/\bar{s}$ from Assumption 3, this rate is $n^{(\bar{s}-2\xi)/2\xi}$, which is disparate from Stone's (1982) optimal rate, $n^{-2\xi/(\bar{s}+2\xi)}$.

Although the proposed estimator is not asymptotically optimal, we can analyze its rate relative to an optimal one. For $0 < c < \bar{s}/\xi$, the estimator is strictly slower than the optimal rate, and approaches the optimal one from below only as $\bar{s}/\xi \to 0$. This occurs because, as is well-known, convergence rates approach zero faster as the function gets arbitrarily smooth (relative to its dimensionality). In the current context, it is apparent that as $\bar{s}/\xi \to 0$,the bias term of this estimator, $K^{1-2\psi}$, approaches the bias term of an estimator that has the best rate. To illustrate the relative improvement from using an estimator that achieves the optimal rate, consider an illustrative example where $\bar{s} = 6, \xi = 4$, and $n = 1000$. Then, $n^{-2\xi/(\bar{s}+2\xi)}/n^{(\bar{s}-2\xi)/2\xi} = 0.109$, suggesting a 10.9% improvement in the convergence rate.

There is a simple intuition for why the derived convergence rate is slower (and the bias term larger) than the optimal one. We know that if d were not endogenous, ordinary least squares estimation with the sequence $\bar{p}^K(x_1, d)$ would suffice for series estimation of this model and the optimal rate of convergence would be attainable (see Newey 1997). Since the rate of the estimator where d has to be predicted must be slower than when d does not have to be predicted,

the rate of the nonparametric IV estimator must be slower than the corresponding estimator when d is exogenous. Further, since the best convergence rate of either estimator is attained when both the variance and bias terms go to zero at the same rate, and since the variance terms are identical under both d endogenous and d exogenous, the bias term in the IV estimator must be larger, in order for this estimator to have a lower convergence rate.

A special feature of the derived rate is the absence of the number of first-step approximating functions, L, in the rate formulas. This lies in contrast to alternate two step estimators (*e.g*, Das *et al*, 1999; Newey *et al* 1999), where the rates depend on both K and L.[7] This result is isomorphic to the parametric IV case, where it is well- known that if estimates of the instruments converge "reasonably fast" to the unknown conditional mean, estimation of the instrument does not affect the limiting distribution of the IV estimator. Correspondingly, Theorem 5.1 reflects that the rate at which the instruments are estimated are irrelevant for the second step rate. In particular, this result illustrates that the second-step rate would be no different for an ordinarily least squares estimator of the instrument which converged at $n^{1/2}$, or a nonparametrically estimated instrument which converged at n^r, $0 < r < 1/2$.

The second result of Theorem 5.1 relates to the uniform convergence rate of the estimator. Results are given for both power series and regression splines. The derived uniform rate has a form similar to other series estimators and, like those, does not attain Stone's (1982) bound. In addition, however, these uniform rates are slower than alternate two-step series estimators (*e.g*, Das *et. al*. 1999) due to the presence of the larger bias term $K^{1-\psi}$.

6 ASYMPTOTIC NORMALITY OF LINEAR FUNCTIONALS

In economics applications, the structural function $m_o(x_1, d)$ is not often the object of inference and the purpose of estimation lies in some numerical characteristic of m_o. For example, if $m_o(x_1, d)$ represents log of annual hours worked, x_1 denotes non-wage income and d is the endogenously determined decision to participate in the AFDC program, then a public policy issue of interest is $\omega_o = \partial m_o(x_1)/\partial x_1$, or the estimated labor supply elasticity (*e.g*, Moffitt, 1983). In this section we consider such functionals of m_o, which include point estimates of $m_o(x_1, d)$, the components $\alpha_o(x_1)$ and $\beta_o(x_1)$, and the function m_o itself.

Let $\theta_o = a(m_o)$ represent a function of the function m_o, i.e, a functional. We restrict our focus to linear functionals of m_o. Below, we present asymptotic normality results for an estimator $\hat{\theta} = a(\hat{m})$ of θ_o, and derive the asymptotic standard errors of $\hat{\theta}$ so that we may develop large

[7] Although the rates derived in both Das *et al* (1999) and Newey *et al* (1999) depend on both K and L, by choosing K and L in particular ways, it is possible to make the rates derived therein to depend only on K if certain smoothness and dimensionality considerations of the second step vis-a-vis the first step hold.

sample (pointwise) confidence intervals for $\hat{\theta}$. These results are important and necessary for doing inference.

Several interesting functionals may be studied in the context of the dummy endogenous regressor model. Both $\alpha_o(x_1)$ and $\beta_o(x_1, d)$ are functionals with economic content. In equation (12) we considered the functional $\beta_o(x_1)$ derived from m_o by evaluating the model at each of the support points of x_2. This functional is important as it identifies the wage gap (for example, veterans versus non-veterans in Angrist (1990)). A useful measure would be the weighted average of this function, for example, with $f(x_1)$ denoting the density function of x_1,

$$\bar{\beta} = \int \hat{\beta}(x_1) f(x_1) dx_1 = \int \left(\frac{\Delta \hat{m}(X)}{\Delta d} \right) f(x_1) dx_1. \tag{24}$$

Although (24) is not an example of Newey's (1994) partial means estimator of m, it may be considered a partial means estimator of $\Delta m / \Delta d$. A related functional that is of interest is the weighted average derivative estimator (see Stoker 1986). This functional represents a summary of the average change in y over some range of X, *e.g*,

$$a(\hat{m}) = \int v(X) \left(\frac{\partial \hat{m}(X)}{\partial X} \right) dX \tag{25}$$

where $v(X)$ is a weighting function. Under the conditions given below, this functional will be $\sqrt{n}$-consistent, while the average incremental effect in (24) will not.

Below we state the conditions needed to derive the asymptotic behavior of the linear functionals $\hat{\theta}$ and show that the functionals are asymptotically normal. The utility of studying the asymptotics is that with a consistent estimate of the asymptotic variance we may do inference on the estimated functionals. We show that $\sqrt{n}\hat{\Omega}_K^{-1/2}(\hat{\theta} - \theta_o) d \longrightarrow N(0, I)$, where $\hat{\Omega}_K$ is the variance estimate of $\hat{\theta}$ so that in large samples inference on $\hat{\theta}$ using the variance estimate $\hat{\Omega}_K / n$ is valid.

Because the analysis is restricted to linear functionals, $A = [a(p^K(x_1)), a(p^K(x_1, d))]$ and $\hat{\theta} = A\hat{\delta}$. Our notation is as follows:

$$\hat{J} = \hat{Q}'\hat{Q}/n, \qquad \hat{G} = \hat{Q}'\hat{P}/n$$

$$\hat{\Sigma}_K = \sum_{i=1}^{n} \hat{\tau}_i(\hat{w}_i) \hat{q}_i \hat{q}_i' [y_i - \hat{m}(x_1, d)]^2 / n \tag{26}$$

The variance of $\hat{\theta}$ is $A \{Var(\hat{m})\} A'$,or

$$\begin{aligned}\hat{\Omega}_K &= A\left[\sum_{i=1}^{n}\tau(\hat{w}_i)\hat{q}_i p_i'/n\right]^{-1}\hat{\Sigma}_K\left[\sum_{i=1}^{n}\tau(\hat{w}_i)\hat{q}_i p_i'/n\right]^{-1}A' \\ &= A\hat{G}^{-1}\hat{\Sigma}_K\hat{G}^{-1\prime}A'.\end{aligned}$$

The regularity conditions below are required to state the asymptotic normality theorem. The first condition strengthens Assumption 1 from a restriction on the second conditional moment of the error term to a restriction on the fourth conditional moment.

Assumption 4 : $E[||y_i - m(x_{1i}, d_i)||^4|w]$ *is bounded and the smallest eigenvalue of* $Var(y_i|w_i)$ *is bounded away from zero.*

The next assumption requires the functional to be continuous in the uniform norm $|\cdot|$, but not in the mean square norm $||\cdot||$. This distinction will be important for illustrating $\sqrt{n}$-consistency of the functionals, since MSE convergence implies the convergence of second moments (required for $\sqrt{n}$-consistency), which is not implied by the uniform norm.

Assumption 6: $a(m_o)$ *is a scalar linear functional such that* $|a(m_o)| < |m_o|_\delta$ *and there exists* δ_K *such that for* $K \to \infty$, *(i)* $m_K = p^{K\prime}\delta_K$ *is bounded away from zero but (ii)* $E[\tau(x_1, d)\{m_K\}^2] \to 0$.

Theorem 6.1 (*Asymptotic Normality*)

If Assumptions 1-6 are satisfied, $\sqrt{n}K^{-\psi} \to 0$ *and* $\zeta_o(K)^2K/n \to 0$,

$$\sqrt{n}\Omega_K^{-1/2}(\theta - \theta)d{\to}N(0,1), \qquad \sqrt{n}\hat{\Omega}_K^{-1/2}(\theta - \theta)d{\to}N(0,1).$$

Proof: Appendix

The result in Theorem 6.1 gives asymptotic normality of linear functionals of m_o, which permits the construction of large sample confidence intervals for inference. The theorem is sufficiently broad to cover asymptotic normality of functionals such as point estimates and β_o, but excludes results for the weighted average derivative or partial means examples, for which an additional condition must hold. The additional condition pertains to mean-square continuity of the functionals (which is not implied by Assumption 6), and is shown to be sufficient for $\sqrt{n}$-consistency if it is satisfied.

Assumption 7: *There exists* $b(x)$ *and* $\tilde{\delta}_K$ *such that* $E[\tau(x)b(x)b(x)'] < \infty$, $a(m_o) = E[\tau(x)b(x)m_o(x_1, d)]$, *and* $E[\tau(w)||b(x) - \tilde{\delta}_K\ \bar{q}^K(x_1, \pi)||^2] \to 0$ as $K \to \infty$.

Assumption 7 requires the existence of a function $b(x) \in \wp$, where $\wp$ denotes the set of functions approximable in mean square by $\bar{q}^K(x_1, \pi)$ as $K \to \infty$. Importantly, Assumption 7 requires the function $b(x)$ to satisfy the condition that

$$a(m_o) = E[\tau(x)b(x)m_o(x_1, d)] \tag{27}$$

$\forall\ m_o(x_1, d) \in \wp$. The representation of the functional as the product of the functions $b(x)$ and $m_o(x_1, d)$ in (27) is equivalent to the requirement that $a(m_o)$ be mean-square continuous. When Assumption 7 holds (either in lieu of, or in addition to, Assumption 6), the asymptotic distribution of$\sqrt{n}(\hat{\theta} - \theta)$ will be normal, with variance $\bar{\Omega}$, where

$$\bar{\Omega} = E[\tau(x)b_K(x, d)b_K(x, d)'var(y|w)] \tag{28}$$

with $b_K(x, d) = A\bar{q}^K(x_1, \pi)$ and $A = E[b(x)\bar{p}^K(x_1, d)']G^{-1}$.

Theorem 6.2 ($\sqrt{n}$-consistency)

If Assumptions 1-6 are satisfied, $\sqrt{n}K^{-\psi} \to 0$ and $\zeta_o(K)^2K/n \to 0$,

$$\sqrt{n}(\theta - \theta)d{\to}N(0, \bar{\Omega}), \qquad \hat{\Omega}_K \to \bar{\Omega}.$$

Proof: Appendix

7 Extensions

7.1 Discrete Endogenous Regressors

The dichotomous nature of the endogenous regressor is an integral component of the analysis above. While binary regressors are sufficient to convert the ill-posed problem into a tractable one, it is shown next that the results extend easily to non-binary discrete endogenous regressors with finite support. Essentially, the problem consists of utilizing a transform that extracts the discrete variable and evaluates the modified function at each point in the support of the discrete variable.

Let d denote a discrete variable with finite support $\in \{s_1, s_j\}$, $|s_1|, |s_j| < \infty$, then

$$\begin{aligned} y &= m_o(x_1, d) + \varepsilon \\ &= m_1(x_1)\mathbf{1}_{\{d=s_1\}} + ... + m_j(x_1)\mathbf{1}_{\{d=s_j\}} + \varepsilon \end{aligned} \tag{29}$$

where $1(\cdot)$ is the indicator function, and $supp\{d\}$ consists of J observations. Equation (29) yields,

$$E[y|x] = m_1(x_1)\Pr(d = s_1|x) + ... + m_j(x_1)\Pr(d = s_J|x) + \varepsilon \tag{30}$$

where each $\pi_j(x) = E[\mathbf{1}_{\{d=s_j\}}|x] = \Pr(d = s_j|x)$, $(j = 1, ..., J)$, can be estimated as a series of binary dependent variables models. It is obvious that the dummy variable model is a special case of this more general specification, with $j = 2$, $s_1 = 0$, $s_2 = 1$.

Estimation of this model is carried out by specifying a $K \times 1$ approximating sequence which consists of functions of x_1 interacted with $d_j = \mathbf{1}_{\{d=s_i\}}, (j = 1, ..., J)$, each of the J dummy variables created from d, and an additional $K \times 1$ approximating sequence of the instrumental variables consisting of x_1 interacted with $\pi_j(x)$, $\forall j = 1, ..., J$. For example,

$$\bar{p}^K(x_1, d) = [1, \bar{p}_{1K}(x_1), ..., \bar{p}_{KK}(x_1)]', \quad \bar{q}^K(x_1, d) = [1, \bar{q}_{1K}(x_1), ..., \bar{q}_{KK}(x_1)]' \tag{31}$$

and, for $k = 1, ..., K$,

$$\bar{p}_{kK}(x_1) = [p_{kK}(x_1)d_1, ..., p_{kK}(x_1)d_J], \quad \bar{q}_{kK}(x_1) = [p_{kK}(x_1)\pi_1, ..., p_{kK}(x_1)\pi_J] \tag{32}$$

The regularity conditions presented in Sections 5 and 6 are sufficiently general to cover the discrete endogenous regressors case. Therefore, asymptotic normality of the estimator will follow from the conditions and theorems stated in Section 6.

7.2 Semiparametric Estimation

A second extension is to a semiparametric formulation. Partition $x = (x_{10}, x_{11})$; an example of a semiparametric model is $y = m_o(x_1, d) + \varepsilon = m_{1o}(x_{10}, d) + x_{11}\bar{\delta} + \varepsilon$. These models are an attractive alternative to fully nonparametric models as they reduce the deleterious effects of the curse of dimensionality when the analyst has *a priori* information about the linear effects of some subset of the covariates on y.

Consider estimation of the example above. To reflect its partially linear form, the approximating series $\bar{p}^K(x, d)$ will consist of the (linear) components of x_{11}, with the remaining terms resembling the series for the fully nonparametric model, replacing x with x_{10} where it appeared previously. Similarly, the elements of $\bar{q}^K(x, \pi)$ will be equated to the elements of $\bar{p}^K(x, d)$, with $\pi(x)$ replacing d. Let $K = K_0 + K_1 + 1$, K_0 denote $\dim(x_{10})$ and K_1 denote $\dim(x_{11})$. Then, the approximating series can be defined as

$$\bar{p}^K(x_1, d) = [1, x_{11}, p_{1Ko}(x_{10}), ..., p_{K_0K_0}(x_{10})]', \quad \bar{q}^K(x_1, d) = [1, x_{11}, q_{1K_0}(x_{11}), ..., q_{K_0K_0}(x_{11})]'.$$

The assumptions of the previous section will yield asymptotic normality of the nonparametric IV estimator of m_o, and also for estimators of the functional $\bar{\delta}$. In addition, this functionals can be shown to satisfy the condition in (46), so that its estimator will be $\sqrt{n}$-consistent under the conditions of Theorem 6.1.

7.3 No Exogenous Covariates

Consider a model with no exogenous covariates, and a single endogenous regressor, d. Analysis of this model is a trivial exercise, and reduces to a parametric specification. Since information for each i, $i(= 1, ..., n)$, is derived purely from whether the indicator takes on the value 1 or 0, the incremental effect is constant across observations, and cannot vary for different values of covariates. Thus, this model is

$$y = c + d\gamma + \varepsilon \tag{33}$$

where γ represents the incremental effect for those observations in which $d = 1$. It follows that γ corresponds to the function $\beta_o(x_1)$, and identification of γ follows from (12). [8] An estimator of γ here is the efficient I.V estimator with a nonparametric first step such as Newey (1990), and $\sqrt{n}$-consistency follows from conditions given therein.

8 Conclusion

This paper develops instrumental variables estimators for nonparametric models with discrete endogenous regressors. Restricting the analysis to discrete endogenous regressors is an integral component of this work and the primary motivation for studying this class of models comes from the intractability of similar models with continuously distributed endogenous regressors. An additional motivation is the enormous empirical literature on program evaluation, which is a special case of the specification. The analysis focuses on the binary endogenous regressor case; the main results are shown to extend to discrete regressors more generally in an extension.

Discreteness of the regressors is shown to produce an additive representation of the model which leads to a simple verifiable condition for identification, and a restriction that is imposed in estimation. The proposed nonparametric two-step IV estimator is based on series estimation which is particularly amenable to additive models, and yields efficiency gains in imposing additivity. The first step constitutes nonparametric estimation of the instrument, while the second step constructs the IV estimator from a linear combination of an instrument matrix and a matrix of the regression covariates. Linear functionals of the estimator are shown to be asymptotically normal, including $\sqrt{n}$-consistent when certain regularity conditions hold. A logical extension of the current work is to efficient instrumental variables estimation of the model. This, and related research, is currently underway by the author.

Columbia University, Department of Economics, 420 West 118th Street, New York, NY 10027.

[8] Consider, without loss of generality, a binary instrument x. Then, since conditional expectations are uniquely identified with probability one, $\frac{E[y|x=1]-E[y|x=0]}{E[d|x=1]-E[d|x=0]} = \gamma$.

References

[1] Ahn, H., (1995): "Nonparametric Estimation of Conditional Choice Probabilities in a Binary Choice Model Under Uncertainty", *Journal of Econometrics*, 67, 337-378.

[2] Amemiya, T. (1978): "The nonlinear two-stage least squares estimator", *Journal of Econometrics*, 2, 105-110.

[3] Andrews, D.W.K (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Models", *Econometrica*, 59, 307-345.

[4] Andrews, D.W.K and Y. J. Whang (1990): " Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality", *Econometric Theory*, 6, 466-480.

[5] Angrist, J. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80, 313-336.

[6] Brown, Bryan (1983): "The Identification Problem in Systems Nonlinear in the Variables", *Econometrica*, 51, 175-196.

[7] Chevalier, J. (1995) : "Capital Structure and Product-Market Competition: Empirical Evidence from the Supermarket Industry", *American Economic Review*, 85, 415-440.

[8] Das, Mitali, W.K. Newey and F. Vella (1999), "Nonparametric Estimation of the Sample Selection Model", MIT Department of Economics working paper.

[9] Fisher, F. (1966): *The Identification Problem in Econometrics*. New York: McGraw Hill.

[10] Hausman, J.A and W.K. Newey (1995): "Nonparametric Estimation of exact Consumer Surplus and Deadweight Loss", *Econometrica*, 63, 1445-1476.

[11] Hausman, J.A, Newey, W.K and J.L. Powell (1995): "Nonlinear errors in variables", *Journal Of Econometrics,* 65, 1994, 205-233.

[12] Hausman, J.A, Newey, W.K, Powell, J.L and H. Ichimura (1991) "Identification and estimation of polynomial errors-in-variables models", *Journal Of Econometrics*, 50, 273-295.

[13] Heckman, J.J (1978): "Dummy Endogenous Variables in a Simultaneous Equation System", *Econometrica*, 46, 931-959.

[14] Heckman, J.J and J. Smith (1997): "The Sensitivity of Experimental Impact Estimates: Evidence|$from the National JTPA Study''$, $Working Paper$, $University of Chicago$.

[15] Koopmans, T.J (1950): "Identification Problems in Economic Model Construction" *Econometrica*, 17, 159-177.

[16] Moffitt, R (1983): "An economic model of Welfare Stigma", American Economic Review, 73, 1023-1035.

[17] Moffitt, R. and B. Wolfe (1991): "The Effect of the Medicaid Program on Welfare Participation and Labor Supply", *The Review of Economics and Statistics*, 74, 615-634.

[18] Newey, W.K (1985b): "Semiparametric Estimation of limited dependent variable models with endogenous explanatory variables", *Annales de l'INSEE*, 59/60, 219-237.

[19] ——— (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models", *Econometrica*, 58, 809-837.

[20] ——— (1994): "Kernel Estimation of Partial Means and a General Variance Estimator", *Econometric Theory*, 10, 233-253.

[21] ——— (1997): "Convergence Rates and Asymptotic Normality of Series Estimators", *Journal of Econometrics*, 79, 147-168.

[22] Newey, W.K. and J.L. Powell (1989): "Nonparametric Instrumental Variables Estimation", MIT Department of Economics Working Paper.

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

[24] Poterba, J.M, S. Venti and D.Wise (1996): "How Retirement Saving Programs Increase Saving", *Journal of Economics Perspectives*, 10, 91-116.

[25] Robinson, P.M (1976):"Instrumental Variables Estimation of Differential Equations", *Econometrica*, 4, 756-776.

[26] ——— (1988): "Root-n-consistent Semiparametric Regression", *Econometrica*, 56, 931-954.

[27] Roehrig, C.S (1988):"Conditions for Identification in Nonparametric and parfametric Models", *Econometrica*, 56, 433-447.

[28] Rothenberg, T.J (1971): "Identification in Parametric Models", *Econometrica*, 39, 577-592.

[29] Stoker, T.M (1986):"Consistent Estimation of Scaled Coefficients", *Econometrica*, 1461-1481.

[30] Stone, C.J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression", *Annals of Statistics*, 10, 1040-1053.

[31] ——— (1985): "Additive Regression and Other Nonparametric Models", *Annals of Statistics*, 13, 689-705.

9 Appendix

Proof of Theorem 3.1

Let $f = [f_1\ f_2]' = [y - m_o(x_1,d) - \varepsilon,\ d - \pi_o(x) - \nu]'$, and $f^* = [f_1^*\ f_2^*]' = [y - m^*(x_1,d) - \varepsilon,\ d - \pi^*(x) - \nu]'$ for any other function f^* that satisfies (2). Assume that f, $f^*$are suitably differentiable (where they are not continuously differentiable, assume that the discrete generalizations of the partial derivatives exist), and let $v = (x_1, d, y, x_2)$ and Δ be defined as before. Partition $v = (v_1, v_2)$, with $v_1 = (x_1, d)$, $v_2 = (y, x_2)$. Let

$$\Delta^* = \begin{bmatrix} \partial(f_o - f^*)/\partial v \\ \partial f^*/\partial v \end{bmatrix} = \begin{bmatrix} \partial(f_o - f^*)/\partial v_1 & 0 \\ \partial f^*/\partial v_1 & \partial f^*/\partial v_2 \end{bmatrix}$$

where the rank of $\partial f^*/\partial v_2$ is d_y+1 since $\text{rank}(x_2) = 1$. Since the rank of Δ^* is the same as the rank of Δ, note that $\text{rank}(\Delta^*) < d_y+1$ implies that $\partial f_o/\partial v_1 = \partial f^*/\partial v_1$, which proves that m_o is identified, up to an additive constant.

In the proofs that follow, C will be the generic notation for a constant that may take different values in different steps. Since our estimator $\hat{m}$ is invariant to nonsingular linear transforms of $\bar{p}^K$ and $\bar{q}^K$, we make simplifying choices for the matrices B_P and B_Q from Assumption $A2$, and for $J = E(\bar{q}^K(x)\bar{q}^K(x)')$. We assume $B_P = I$, and $B_Q = I$, so that $\bar{P}^K(x_1,d) = \bar{p}^K(x_1,d)$, $\bar{Q}^K\ (x_1,\pi) = \bar{q}^K(x_1,\pi)$ and we choose $J = I$, since for the symmetric square root $J^{-1/2}$ of J^{-1}, $J^{-1/2}\bar{q}^K(x)$ is a nonsingular linear transform of $\bar{q}^K(x_1,\pi)$ that satisfies Assumption 2 with

$$\Xi_\mu(K) = \max_{|g|\leq\mu} \sup_{X\in\mathcal{X}} ||\partial^g J^{-1/2}\bar{q}^K(x,d)|| \leq C\Xi_\mu(K) \tag{34}$$

Therefore, for the remainder of the analysis it will be assumed without loss of generality that $J = I$.

As defined in the body of the paper, let $\hat{Q} = [\hat{\tau}_1 q_1, ..., \hat{\tau}_n q_n]'$, $\hat{P} = [\hat{\tau}_1 p_1, ..., \hat{\tau}_n p_n]'$, $(\hat{Q}'Q/n) = \hat{J}$, $\hat{G} = (\hat{Q}'\hat{P}/n)$, $G = E(Q_i P_i')$, $J = E(Q_i Q_i')$, and $\hat{m}(x,d) = \bar{p}^K(x,d)'\hat{\beta}^{IV}$, $\hat{\beta}^{IV} = (\hat{Q}'\hat{P})^{-1}\hat{Q}'y$. Note that

$$E[\|\hat{G} - G\|^2] \leq \sum_{k=1}^{K}\sum_{j=1}^{J} E[p_{kK}^2 q_{jK}^2]/n = E[\sum_{k=1}^{K} p_{kK}^2 \sum_{j=1}^{J} q_{jK}^2]/n$$

$$\leq \quad n^{-1}\zeta_o(K)^2 E[\sum_{j=1}^{J} q_{kK}^2] = n^{-1}\zeta_o(K)^2 tr(J) = n^{-1}\zeta_o(K)^2 K \to 0,$$

so that

$$\|\hat{G} - G\| = O_p(\zeta_o(K)K^{1/2}/\sqrt{n}) = o_p(1). \tag{35}$$

Similarly,

$$
\begin{aligned}
E[\|\hat{J} - I\|^2] &= \sum_{k=1}^{K}\sum_{j=1}^{J} E[q_{kK}^2 q_{jK}^2/n - Ijk] \leq \sum_{k=1}^{K}\sum_{j=1}^{J} E[q_{kK}^2 q_{jK}^2/n] \\
&= E[\sum_{k=1}^{K} q_{kK}^2 \sum_{j=1}^{J} q_{jK}^2]/n \leq n^{-1}\zeta_o(K)^2 E[\sum_{k=1}^{K} q_{kK}^2] \\
&= n^{-1}\Xi_o(K)^2 tr(I) = n^{-1}\Xi_o(K)^2 K \to 0
\end{aligned}
$$

giving

$$
\|\hat{J} - I\| = O_p(\zeta_o(K)K^{1/2}/\sqrt{n}) = o_p(1). \tag{36}
$$

Proof of Theorem 5.1

For $M = (m_o(x_1, d_1), \ldots, m_o(x_n, d_n))$ and $\epsilon = Y - M$, the assumption of bounded $Var(y|w)$ and independence of the observations together imply $E(\varepsilon\varepsilon') \leq CI$. Therefore,

$$
\begin{aligned}
E\|\hat{Q}'\epsilon/n\|^2 &= E[E[\|\hat{Q}\epsilon/n\|^2|X]] \\
&= E[\sum_{i=1}^{n} \hat{q}_i'\hat{q}_i E[\epsilon\epsilon'|X]/n^2] \\
&\leq CE[\hat{q}_i'\hat{q}_i/n] = C \;\; tr(J)/n = CK/n.
\end{aligned}
$$

Then, by the Markov inequality, $\|\hat{Q}'\epsilon/n\|^2 = O_p(K/n)$. Notice that conditioning the matrix G on X yields, in expectation, the matrix of instruments J. Therefore,

$$
\begin{aligned}
E[\|\hat{G}^{-1}\hat{Q}'/\sqrt{n}\|^2] &= E[\hat{G}^{-1}\hat{J}\hat{G}^{-1}] \\
&= E[E[\hat{G}^{-1}\hat{J}\hat{G}^{-1}|X]] \\
&= E[E[\hat{G}^{-1}|X]E[\hat{J}]E[\hat{G}^{-1}|X]] \\
&= E[J^{-1}JJ^{-1}] \\
&= E[q_i'q_i] = tr(J) = K
\end{aligned}
$$

yielding $\|\hat{G}^{-1}\hat{Q}'/\sqrt{n}\|^2 = O_p(K)$. Notice that since $E[\hat{G}^{-1}|X] = \hat{J}$, $\|\hat{G}^{-1}Q'/\sqrt{n}\|^2$ is also $O_p(K)$. It follows that

$$
\begin{aligned}
E\|\hat{G}^{-1}\hat{Q}'\epsilon/n\|^2 &= E[E[\hat{G}^{-1}\hat{Q}'\epsilon\epsilon'\hat{Q}\hat{G}^{-1}/n^2|X]] \\
&= E[E[\hat{G}^{-1}|X]\sum_{i=1}^{n} \hat{q}_i'\hat{q}_i Var(y|X)E[\hat{G}^{-1}|X]/n^2]
\end{aligned}
$$

$$
\begin{aligned}
&\leq CE[J^{-1}\sum_{i=1}^{n}\hat{q}_i'\hat{q}_iJ^{-1}/n^2] \\
&= Ctr(J)/n = CK/n. \qquad (37)
\end{aligned}
$$

Therefore, $\|\hat{G}^{-1}\hat{Q}'\epsilon/n\|^2 = O_p(K/n)$, and

$$
\begin{aligned}
\|\hat{G}^{-1}Q'\epsilon\|^2 &\leq \|G^{-1}\hat{Q}'\epsilon/n\|^2 + \|[\hat{G}^{-1}-G^{-1}]\hat{Q}'\epsilon/n\|^2 \\
&= O_p(K/n) + o_p(1)O_p(K/n) = O_p(K/n) + o_p(1) = O_p(K/n).
\end{aligned}
$$

Let β satisfy $sup_{\mathcal{X}}|m_o(x,d) - \bar{p}^K\beta| = O(K^{-\psi})$ as given in $A3$. Then, by the comment following (),

$$
\begin{aligned}
\|\hat{G}^{-1}\hat{Q}'(M-P\beta)/n\|^2 &\leq \|\hat{G}^{-1}\hat{Q}'/\sqrt{n}\|^2\|(M-P\beta)/\sqrt{n}\|^2 \\
&= [\hat{G}^{-1}\hat{J}\hat{G}^{-1}][(M-P\beta)(M-P\beta)/n] \\
&= O_p(K)O(K^{-2\psi}) \\
&= O_p(K^{1-2\psi}).
\end{aligned}
$$

Therefore, by $(\hat{\beta}^{IV}-\beta) = \hat{G}^{-1}\hat{Q}'\epsilon + \hat{G}^{-1}\hat{Q}'(M-P\beta)$, the MSE of $\hat{\beta}^{IV}$ is given by

$$
\begin{aligned}
\|\hat{\beta}^{IV}-\beta\|^2 &\leq \|\hat{G}^{-1}\hat{Q}'\epsilon\|^2 + \| + \hat{G}^{-1}\hat{Q}'(M-P\beta)\|^2 \\
&= O_p(K/n + K^{1-2\psi}). \qquad (38)
\end{aligned}
$$

Our first conclusion follows from the triangle inequality,

$$
\begin{aligned}
\int(\hat{m}(x,d)-m_o(x,d))^2 &= \int \bar{p}^K(x,d)'(\hat{\beta}^{IV}-\beta) + \bar{p}^K(x,d)'\beta - m_o(x,d))^2dF(w) \\
&\leq \|\hat{\beta}^{IV}-\beta\|^2 + \int(\bar{p}^K(x,d)-m_o(x,d))^2 \\
&= O_p(K/n+K^{1-2\psi}) + O(K^{-2\psi}) \\
&= O_p(K/n+K^{1-2\psi}). \qquad (39)
\end{aligned}
$$

For the second conclusion, note that

$$
\begin{aligned}
|\hat{m}(x,d)-m_o(x,d)|_o &\leq |\bar{p}^K(\hat{\delta}-\delta)|_o + \bar{p}^K\delta - m_o| \\
&\leq \zeta_o(K)||\hat{\delta}-\delta|| + O(K^{1-\psi}) \\
&= O_p(\zeta_o(K)[\sqrt{K/n}+K^{1-\psi}]) \qquad (40)
\end{aligned}
$$

Explicit formulas have been derived for $\zeta_o(K)$ by Andrews (1991) and Newey (1997), based on the principle that the estimators are invariant to location and scale shifts in the approximating functions. Tranform X to have support [-1,1]^s and replacing the components of the series in $\bar{p}^K$

to be orthonormal w.r.t the uniform distribution (for the case of polynomials) and to B-splines in the case of splines, which are described in the body of the paper. Then it follows from Andrews (1991) and Newey (1997) that $\zeta_d(K) = CK^{1+2d}$ for polynomials, $\zeta_d(K) = CK^{0.5+d}$ for splines, completing the proof of Theorem 5.1. *Q.E.D.*

Proof of Theorem 6.1

Let $\Omega_K^{-1/2} = (A'G^{-1}\Sigma G^{-1}A)^{-1/2}$ denote a symmetric square root of Ω_K^{-1}. Deduce that $\Sigma \geq CI$ by $\text{Var}(y|w)$ bounded from Assumption1. Then, $\Omega_K = A'G^{-1}\Sigma G^{-1}A \geq C\|A'G^{-1}\|^2$. Further, $E[\|G\|] = E[E\|G\||X] = E[\|J\|] = tr(J)^{1/2} = K^{1/2}$, yielding $\|G\| = O_p(K^{1/2})$. It follows that

$$\begin{aligned}
\|\Omega_K^{-1/2}A\|^2 &= \|\Omega_K^{-1/2}AG^{-1}G\|^2 \\
&\leq \|\Omega_K^{-1/2}AG^{-1}\|^2\|G\|^2 \\
&= tr(\Omega_K^{-1/2}AG^{-1}G^{-1'}A'\Omega_K^{-1/2})\|G\|^2 \\
&\leq tr(C\Omega_K^{-1/2}\Omega_K^{-1}\Omega_K^{-1/2})O_p(K) \\
&= CO_p(K) = O_p(K) \qquad (41)
\end{aligned}$$

Note that $\sqrt{n}\Omega_K^{-1/2}(\hat{\theta} - \theta)$ can be decomposed into the sum of terms, of which all but one converge in probability to zero, and one converges in distribution to $N(0,1)$,

$$\begin{aligned}
\sqrt{n}\Omega_K^{-1/2}(\hat{\theta} - \theta_o) &= \sqrt{n}\Omega_K^{-1/2}(a(\bar{p}^{K\prime}\hat{\beta}^{IV}) - a(m_o)) \\
&= \sqrt{n}\Omega_K^{-1/2}(\bar{p}^{K\prime}[\hat{G}^{-1}Q'(M+\epsilon)/n] - m_o \\
&= \Omega_K^{-1/2}A'[\hat{G}^{-1}Q'\epsilon/n + \hat{G}^{-1}Q'M/\sqrt{n} - \\
&\quad \sqrt{n}\hat{G}^{-1}\hat{G}\beta_K + \sqrt{n}\beta_K] - \sqrt{n}\Omega_K^{-1/2}M_o \\
&= \Omega_K^{-1/2}A'\hat{G}^{-1}Q'\epsilon/\sqrt{n}) + \sqrt{n}\Omega_K^{-1/2}A'\hat{G}^{-1}Q'(M - P\beta_K/n \\
&\quad + \sqrt{n}\Omega_K^{-1/2}(M_K - M_o). \qquad (42)
\end{aligned}$$

By equation (42)and the Cauchy-Schwartz inequality,

$$\begin{aligned}
\|\sqrt{n}\Omega_K^{-1/2}A'\hat{G}^{-1}Q'(M - P\beta_K)/n\| &\leq \sqrt{n}\|\Omega_K^{-1/2}A'\|\|\hat{G}^{-1}Q'/\sqrt{n}\|\|(M - P\beta_K)/\sqrt{n}\| \\
&= \sqrt{n}O_p(K^{1/2})\ \ tr(\hat{G}^{-1}Q'Q\hat{G}^{-1}/n)^{1/2}[(M - P\beta_K)(M - P\beta_K)]^{1/2} \\
&= \sqrt{n}O_p(K^{1/2})O_p(K^{1/2})O(K^{\psi}) \\
&= O_p(\sqrt{n}K^{1-\psi}) \to 0.
\end{aligned}$$

By the assumption in Theorem 6.1, $|\sqrt{n}\Omega_K^{-1/2}(M_o - P\beta_K)| = O_p(K^{1/2})\sqrt{n}|(M_o - P\beta_K)| = O_p(K^{1/2})\ \sqrt{n}K^{-\psi} \to 0$. Next, let $Z_{in} = \Omega_K^{-1/2}A'\hat{G}^{-1}q_i\varepsilon_i/\sqrt{n}$, yielding $E(Z_{in}) = 0$. Note that each Z_{in} $(i = 1, ..., n)$ is i.i.d, $\sum_i Z_{in} = \Omega_K^{-1/2}A'\hat{G}^{-1}Q'\varepsilon/\sqrt{n}$, $\sum_i E(Z_{in}^2) = 1$, and for some constant γ

$$\begin{aligned} & nE[1(|Z_{in}| > \gamma)Z_{in}^2] = n\gamma^2 E[1(|Z_{in}/\gamma| > 1)(Z_{in}/\gamma)^2] \leq n\gamma^2 E[(Z_{in}/\gamma)^4] \\ = \ & n\gamma^2 E[\|Z_{in}^4\|]/\gamma^4 = nE[\|\Omega_K^{-1/2}A'Q'\epsilon/\sqrt{n}\|^4/\gamma^2 \leq n\|\Omega_K^{-1/2}A'\|^4 E[Q'QE(\epsilon\epsilon'|X)]^2/n^2\gamma^2 \\ = \ & n\gamma^{-2}) \leq C\|\Omega_K^{-1/2}A'\|^4\zeta_o(K)^2E(\|Q\|^2)/n^2 = CO_p(K^2)\zeta_o(K)^2K/n \to 0 \end{aligned}$$

Therefore, due to the Lindberg-Feller central limit theorem, $\sum_i Z_{in} d\to N(0, I)$, which implies that $\sqrt{n}\Omega_K^{-1/2}(\hat{\theta} - \theta_o)d\to N(0, 1)$. This gives the first conclusion of Theorem 6.1.

Next, consider the last conclusion of Theorem 6.1. Define $\hat{h} = \hat{G}^{-1}\hat{A}\Omega_K^{-1/2}$, and $h = A\Omega_K^{-1/2}$. Then, by $||\hat{h}|| = [\hat{G}^{-1}\hat{A}\Omega_K^{-1/2}\Omega_K^{-1/2}\hat{A}'\hat{G}^{-1}]^{1/2} = tr(\Omega_K^{-1/2}AG^{-1}G^{-1'}A'\Omega_K^{-1/2}) \leq tr(C\Omega_K^{-1/2}\Omega_K^{-1}\Omega_K^{-1/2}) = O_p(1)$, and

$$\begin{aligned} ||\hat{h} - h|| &= ||\Omega_K^{-1/2}(\hat{G}^{-1}\hat{A} - GA)|| \leq ||\Omega_K^{-1/2}||\ \ ||\hat{G}^{-1}(\hat{A} - A)|| + ||A(\hat{G} - G)|| \\ &= O_p(1)(o_p(1) + o_p(1))p\to 0. \end{aligned} \tag{43}$$

By $\Sigma \leq CI$, the largest eigenvalue of Σ is bounded from above. Let $\tilde{\Sigma} = \sum_i \hat{q}_i\hat{q}_i'(y_i - m_o(x_{1i}, d_i))$. Then, by the fourth conditional moment bounded in Assumption 4, it follows that $||\tilde{\Sigma} - \Sigma||p\to 0$, and $||\hat{h}'\tilde{\Sigma}\hat{h} - \hat{h}'\Sigma\hat{h}||p\to 0$.

Define $\Delta_i = m_o(x_{1i}, d_i) - \hat{m}(x_{1i}, d_i)$. Since $\zeta_o(K)[\sqrt{K/n} + K^{1-\psi}] = [\zeta_o(K)^2K/n]^{1/2}\ (1 + \sqrt{n}K^{-\psi}) \to 0$, by Theorem 5.1, $\max_{i\leq n}|\Delta_i| \leq |\hat{m} - m_o|_o p\to 0$. Let $\hat{S} = n^{-1}\sum_i \hat{q}_i\hat{q}_i'|y_i - m(x_{1i}, d_i)|$ and $S = E[\hat{q}_i\hat{q}_i'|y_i - m(x_{1i}, d_i)|] = E[\hat{q}_i\hat{q}_i'E[|y_i - m(x_{1i}, d_i)|w] \leq CI$. Similar to the argument that $||\hat{J} - J||p\to 0$ in (36), and by $Var(y|w)$ bounded in Assumption 1, $||\hat{S} - S||p\to 0$. Then, it follows from Newey (1997; p. 166 A.10) that $|\Omega_K^{-1/2}\hat{\Omega}_K\Omega_K^{-1/2} - \hat{h}'\tilde{\Sigma}\hat{h}|p\to 0$. Further, by the triangle inequality, $||\hat{h}'\tilde{\Sigma}\hat{h} - \hat{h}'\Sigma\hat{h}||p\to 0$ and (43), $|\Omega_K^{-1/2}\hat{\Omega}_K\Omega_K^{-1/2} - 1|p\to 0$, implying

$$\begin{aligned} & (\Omega_K^{-1/2})^2\hat{\Omega}_K p\to 1, \\ \sqrt{n}\hat{\Omega}_K^{-1/2}(\theta - \theta) &= \sqrt{n}\Omega_K^{-1/2}(\theta - \theta)/(\Omega_K^{-1}\hat{\Omega}_K)^{1/2}d\to N(0, 1) \end{aligned} \tag{44}$$

giving the second conclusion. *Q.E.D.*

Proof of Theorem 6.2

Without loss of generality, the theorem is proved for scalar $a(m_o)$, since for any vector c with $||c|| = 1$, $c'\sqrt{n}(\theta - \theta)d \to N(0, c'\bar{V}c)$ by the Cramer-Wold device, and Assumption 7 holds with $c'b(x)$ replacing $b(x)$. Write $\bar{q}^K(x)$ for $\bar{q}^K(x_1, \pi)$. Let $b_K(x,d) = A\bar{q}^K(x) = E(\tau(x)b(x)\bar{p}^K(x_1,d)')G^{-1}\bar{q}^K(x)$. Note that $E[b_K(x)\, var(y|w) b_K(x)'] = E[\tau(x)b(x)\bar{p}^K(x_1,d)'\, G^{-1}\bar{q}^K(x) var(y|w)\bar{q}^K(x)'G^{-1\prime}\bar{p}^K(x_1,d)b(x)']$ $= E[b(x)m_o(x_1,d)]^2 = \Omega_K$ by Assumption 7 and (27). Suppress the arguments so that $b = b(x)$, $b_K = b_K(x,d)$.

Since b is approximable by $\bar{q}^K(x)$ in mean square, $E[(b - b_K)^2] = E[((b - b_K)^2|x] \leq E[\{b - \tilde{\delta}_K \bar{q}^K(x)\}] p \to 0$. Next, by Cauchy-Schwartz,

$$\begin{aligned} |\Omega_K - \bar{\Omega}| &\leq E[|b_K^2 - b^2|] \leq E[(b_K - b)^2] + 2E[|b||b_K - b|] \\ &\leq o(1) + 2(E[b^2])^{1/2}(E[(b - b_K)^2])^{1/2} \to 0, \end{aligned} \tag{45}$$

implying that $\Omega_K p \to \bar{\Omega}$. Also, by the proof of Theorem 6.1, $\Omega_K^{-1/2}\ \hat{\Omega}_K^{1/2} p \to 1$,which gives $\hat{\Omega}_K$ $p \to \bar{\Omega}$ by squaring. Since $var(y|w)$ and b_K are bounded away from zero, $\bar{\Omega}$ is bounded away from zero. The critical difference from Theorem 6.1 now follows as the convergence rate of θ is has a sharp bound, *i.e*, $\Omega_K^{-1/2} p \to 1/\sqrt{\bar{\Omega}}$. Then, by the conclusion of Theorem 6.1,

$$\sqrt{n}(\hat{\theta} - \theta_o) = \Omega_K^{1/2}\sqrt{n}\Omega_K^{-1/2}(\hat{\theta} - \theta_o) = N(0, \bar{\Omega}) \tag{46}$$

Q.E.D.