Nonparametric Instrumental Variable Estimation using Complex Survey Data

Luc Clair

July 16, 2023

Abstract

The literature on nonparametric instrumental variable (IV) methods has been growing, and while the mathematics behind these methods are highly technical, researchers believe that these methods will soon emerge as viable alternatives to parametric approaches. The difficulty of these methods stems from the fact that the nonparametric instrumental variable estimator is the solution to a ill-posed inverse problem. The ill-posed inverse problem is solved by using regularization methods. Therefore there are two steps for solving the nonparametric (IV) problem: estimating conditional means and regularization. When analysis is performed using complex survey data, one must also consider the sampling design. When endogeneous sampling is present, traditional estimation methods will be inconsistent. I extend the theory of nonparametric IV models to account for sample design by estimating the conditional mean functions using a probability weighted local constant estimator.

1 Introduction

Violation of one or more of the assumptions made on the data generating process (DGP) of an econometric model leads to misspecification and the estimator may converge to a parameter that differs from the true population value. A common complication within health econometric models is the presence of endogenous predictor variables, which lead to inconsistent estimates using either traditional parametric or nonparametric regression methods. Simultaneity leads to complex relationships between health variables, blurring the direction of causality. A popular example in economics is the estimation of demand curves. Quantity demanded is a function of price; however, firms change their price based on quantity demanded. Contoyannis et al. (2005) encountered this problem when estimating the price elasticity for prescription drugs in Québec. For senior citizens over the age of 65, the provincial drug plan imposed a deductible of $14.30 per month, covered 31 percent of costs between $14.30 to $77.21, and covered 100 percent of costs above $77.21. As consumption increased, the price for additional units decreased after a certain amount paid, leading to a kinked budget constraint.

Mathematically, a model with at least one endogenous regressor is written as:

$$Y = g(X) + \epsilon;\ E(\epsilon|X) \neq 0. \tag{1.1}$$

In this case, the conditional mean $E(Y|X)$ no longer coincides with the object of interest, i.e. $E(Y|X) \neq g(X)$. If, however, there is a vector Z that is correlated with X and $E(\epsilon|Z) = 0$, one can apply instrumental variable (IV) techniques and obtain consistent estimates of $g(\cdot)$.

In a linear parametric model, we specify $g(X) = g(X, \beta) = X\beta$, greatly simplifying the analysis. In this linear specification, the parameter of interest β is either a scalar or a vector. Since the identifying mapping is continuous, we can apply two-staged least squares and estimate the model's parameters by replacing the unknown population distribution with a consistent sample analog (Horowitz 2013). This is known as a *well-posed* problem. The issue with using these methods is that economic theory may not support restricting the data generating process to $Y = X\beta + \epsilon$. Furthermore, these methods may mislead analysts by suggesting their results are more precise than

they actually are (Centorrino, Feve & Florens 2014).

This is the motivation behind a growing body of literature on nonparametric instrumental variable methods. Nonparametric models do not assume a specific functional form, instead the parameter of interest is the solution of an equation that is dependent on the DGP (Li & Racine 2007). The parameter in this case is a *function* ($g(X)$), not a vector or scalar. The challenge for most researchers who are considering using nonparametric IV methods is that they involve mathematical techniques that are not part of the 'standard toolkit' taught to econometricians, hence there are some fixed costs that must be incurred. However, these upfront fixed costs can reward the patient researcher with substantial dividends. Theoreticians acknowledge the hesitation of empirical researchers to use these methods but insist that nonparametric IV techniques are emerging as viable alternatives to parametric approaches (Darolles, Fan, Florens & Renault 2011). In contrast to parametric IV, the identifying mapping of the model's parameter (function) in nonparametric IV estimation is discontinuous and constitutes an *ill-posed inverse problem*. As will be shown in Section 4, the nonparametric IV model reduces to a Fredholm equation of the first kind with an infinite-dimensional operator function. In that way, we can think of solving for a parameter in a parametric model as a finite-dimensional problem and solving for a parameter in the nonparametric IV model as an infinite-dimensional problem.

In order to estimate the parameter (function) in an nonparametric IV model, we must first transform the discontinuous mapping into a continuous one, where the parameter can be uniquely identified. The process of converting an ill-posed problem into a well-posed problem is called 'regularization'. This additional step is the difference between nonparametric IV and conventional nonparametric regression: in addition to estimating conditional mean functions, we must now also regularize the mapping to solve for $g(\cdot)$. This requires the selection of two parameters, the smoothing parameter for estimating conditional mean functions and the regularization parameter.

In an applied setting, empirical economic studies often make use of large datasets that rely on complex sampling plans. When using data collected using unequal probability sampling, inference must take into account the sampling plan. If sampling is endogenous, such that the sampling criterion is related to the error term, estimation of conditional mean functions will be inconsistent

(Clair 2019, Magee, Robb & Burbidge 1998). Modified nonparametric kernel regression methods have been used to study the relationship between X and Y to incorporate survey weights. Sánchez-Borrego et al. (2014) introduced a local constant estimator with probability weighted kernel density functions for estimating conditional mean functions using mixed data types. Clair (2019) further investigated this probability weighted estimator and developed a least squares cross-validation method for selecting the smoothing parameters. Clair (2019)'s simulations showed that the probability weighted local constant had a lower mean squared error under endogenous stratification.

The purpose of this paper is to review the recent developments and advancements in nonparametric IV, present these methods in a language that is accessible to audiences outside of theoretical econometricians, and extend this theory to account for sampling design. To account for sample design, I propose using the probability-weighted local constant estimator from Sánchez-Borrego et al. (2014) and select the bandwidths using the method from Clair (2019). The rest of this paper is organized in six sections. Following this introduction, in Section 2, I provide an overview of the underlying theory of nonparametric instrumental variable regression. In Section 3, I review parametric instrumental variable methods, including two-staged least squares (2SLS) and generalized method of moments (GMM). Section 4 introduces the nonparametric IV framework and demonstrates how identification of $g(X)$ is the solution to an ill-posed inverse problem. This section reviews methods for estimating the conditional mean functions and reviews the probability weighted local constant estimator. In addition, this section also describes Landweber-Fridman regularization, an iterative method to correct a discontinuous mapping. In Section 5, I run simulations comparing probability weighted nonparametric IV methods to unweighted nonparametric IV methods and traditional nonparametric regression methods. Section 6 applies the probability-weighted nonparametric instrumental variable estimator to estimate the impact of supplemental insurance on the hours spent with a psychologist. Section 7 concludes.

2 Underlying Theory

The underlying theory of nonparametric instrumental variable estimation is complex. The theory contains concepts from differential and integral calculus, functional analysis, and linear integral equations. Functional analysis is a branch of mathematical analysis concerned with vector spaces with some limit-related structure and mappings between them. These mappings are called operators, denoted by $T : \mathcal{H} \to \mathcal{V}$, which project elements from the domain vector space $\mathcal{H}$ to the range vector space $\mathcal{V}$. The type of linear integral equation of interest in estimating nonparametric IV models is the Fredholm equation of the first kind:

$$\int_a^b f(x,z)g(x)dx = r(z),\ x \in [a,b] \tag{2.1}$$

where $g(x)$ is the unknown quantity we wish to solve for, $f(x,z)$ is a weight function, and $r(z)$ is a given function. In this case, both r and f are known. For a full investigation on linear integral equations, I refer you to Kress (1999).

In this section, I define relevant terms and concepts to understanding the problem of nonparametric IV and provide a narrative to facilitate the reader's understanding of the material. For a thorough and more technical review of the underlying concepts and their application to nonparametric instrumental variable estimation, I refer the reader to Carrasco et al. (2007). To start, I wish to review introductory concepts in functional analysis.

2.1 Functional Spaces

Fields and vector spaces are commonly worked with in applied economics as several concepts from linear algebra are relevant in econometric theory. Therefore, fields and vector spaces provide a helpful starting point to understand nonparametric IV methods. In mathematics, a field $\mathcal{F}$ refers to a set in which the addition, subtraction, multiplication, and division of the elements in any subset of $\mathcal{F}$ are also in $\mathcal{F}$ (assuming division is not by 0). That is, if $\alpha, \beta \in \mathcal{F}$ then $\alpha + \beta$, $\alpha - \beta$, $\alpha \times \beta$, and α/β ($\beta \neq 0$) are also in $\mathcal{F}$. The elements of a field are known as *scalars* and these scalars satisfy the *field axioms*.

Axiom 2.1 (**Field Axioms**). *If $\alpha, \beta, \gamma \in \mathcal{F}$, then they satisfy the following properties:*

(i) Associativity of Addition: $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$

(ii) Commutativity of Addition: $\alpha + \beta = \beta + \alpha$

(iii) Distributivity of Addition: $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$

(iv) Additive Identity: $\alpha + 0 = \alpha$.

(v) Additive Inverse: There exists a scalar $-\alpha$ *such that* $\alpha + (-\alpha) = 0$.

(vi) Associativity of Multiplication: $(\alpha\beta)\gamma = \alpha(\beta\gamma)$

(vii) Commutativity of Multiplication: $\alpha\beta = \beta\alpha$

(viii) Distributivity of Multiplication: $(\alpha + \beta)\gamma = \alpha\gamma + \beta\gamma$

(ix) Multiplicative Identity: $\alpha \cdot 1 = \alpha$.

(x) Multiplicative Inverse: There exists a scalar α^{-1} *such that* $\alpha\alpha^{-1} = 1$.

Popular examples of fields include the set of real numbers, $\mathbf{R}$, the set of complex numbers $\mathbf{C}$, and the set of rational numbers[1] $\mathbf{Q}$.

A vector space is a set $\mathcal{V}$ whose elements, known as vectors, hold for two operations: vector addition, denoted by "+", and scalar multiplication, denoted by "$\cdot$". Vector addition refers to the property that any two elements $g, h \in \mathcal{V}$ can be added to give a third element in $\mathcal{V}$, i.e. $g + h \in \mathcal{V}$, and scalar multiplication is the property that for any $\alpha \in \mathcal{F}$ then $\alpha \cdot g \in \mathcal{V}$. In econometrics, the term vector is most closely associated with $n \times 1$ column matrices of values in $\mathbf{R}$, however, the elements of vector spaces can take multiple forms as long at they can be added together and scaled (multiplied by scalars), e.g. polynomials and other functions. Vector spaces follow axioms that allow more abstract vector spaces to behave like more geometric spaces, e.g. $\mathbf{R}^3$. These axioms are listed in Axiom 2.2.

[1] To reinforce this concept, it may help to see an example of a set that is not a field, e.g. the set of integers $\mathbf{Z}$. While $2, 3 \in \mathbf{Z}$ and $2 + 3 = 5, 2 - 3 = -1, 2 \times 3 = 6 \in \mathbf{Z}$, $3/2 = 1.5 \notin \mathbf{Z}$.

Axiom 2.2 (**Vector Space Axioms**). *If $f, g, h \in \mathcal{V}$ and $\alpha, \beta \in \mathbf{R}$ are scalars, then*

(i) Vector addition is associative and commutative

(ii) There exists an additive identity, θ, known as the zero vector such that (s.t.) $g + \theta = g$ for all $g \in \mathcal{V}$.

(iii) There exists an additive inverse $-g \in \mathcal{V}$ s.t. $g + (-g) = \theta$.

(iv) $\alpha(\beta g) = \alpha\beta(g)$ for all $\alpha, \beta \in \mathbf{R}$ and all $g \in \mathcal{V}$.

(v) $(\alpha + \beta)g = \alpha g + \beta g$ for all $\alpha, \beta \in \mathbf{R}$ and all $g \in \mathcal{V}$.

(vi) $\alpha(g + h) = \alpha g + \alpha h$ for all $\alpha \in \mathbf{R}$ and all $g, h \in \mathcal{V}$.

(vii) $0 \cdot g = \theta$, $1 \cdot g = g$.

Vectors in the form of column matrices of values in $\mathbf{R}$ are the easiest to visualize and are often thought of as "arrows". The values in a column matrix can be plotted as Cartesian coordinates and the length of the vector is given by the distance from the origin to point identified by the vector. When the elements of a vector space take the form of more complex functions, this visualization is no longer applies. To get a sense of the size and length of vectors in more abstract vector spaces, one can compute the norm. The norm, denoted by $||\cdot||$, is simply a function that transforms the vector to a scalar in $\mathbf{R}$ giving a positive value to the size of the elements in $\mathcal{V}$, where the functional form of the norm depends on the type of elements in the vector space. More formally, given $\alpha \in \mathcal{F}$ and $h, g \in \mathcal{V}$ the norm is a function $\mathcal{V} \to \mathbf{R}$ such that

(i) $||\alpha g|| = |\alpha|||g||$.

(ii) $||h + g|| \leq ||h|| + ||g||$.

(iii) $||g|| = 0$ if and only if $g \equiv \theta$ (Li & Racine 2007, Carrasco, Florens & Renault 2007).

Two relevant examples are the Euclidean norm, as this is the vector space most of us are familiar with, and the norm for the space of square integrable functions, as this is the canonical example for nonparametric IV analysis.

Example 2.1 (**Euclidean Norm**). *The norm of* $g = (g_1, g_2) \in \mathbf{R}^2$ *is given by:*

$$||g|| = (g_1^2 + g_2^2)^{1/2}, \tag{2.2}$$

which gives the ordinary distance from the origin to the point g.

Example 2.2 (**Space of Square Integrable Functions**). *Another example is the space of square integrable functions, denoted by* $\mathbf{L}^2$. *If* $h \in \mathbf{L}^2$, *then the norm of* h, $||h||$, *is given by:*

$$||h|| = \left\{ \int |h(x)|^2 dx \right\}^{1/2}. \tag{2.3}$$

When a vector space $\mathcal{V}$ has a defined norm, it is called a *normed vector space*. $\mathbf{R}^2$ and $\mathbf{L}^2$ are normed spaces with norms defined by Equations (2.2) and (2.3). A normed space is called a *Banach space* if for every sequence $\{g_n\} \in \mathcal{V}$

$$\lim_{n,m\to\infty} ||g_n - g_m|| = 0$$

(Li & Racine 2007)[2]. This is an important concept as a Banach space ensures there are enough limits in the space to allow the techniques of calculus to be used.

If one wishes to impose additional structure to define the length of a vector and/or the angle between two vectors, e.g. identify perpendicularity between two vectors, one needs to define the inner product. The inner product in a vector space $\mathcal{V}$, denoted by $\langle \cdot, \cdot \rangle$, is a function that returns a scalar value from the Cartesian product of two vector spaces, where the Cartesian product is a pairing of elements from the two sets[3]. The mapping is given by $\mathcal{V} \times \mathcal{V} \to \mathbf{R}$ with the property that, for $f, g, h \in \mathcal{V}$ and $\alpha, \beta \in \mathbf{R}$,

(i) $\langle \alpha g + \beta h, f \rangle = \alpha \langle g, f \rangle + \beta \langle h, f \rangle$.

(ii) $\langle g, h \rangle = \langle h, g \rangle$.

[2] A sequence with this property is a Cauchy sequence. When every Cauchy sequence in a vector space converges it is called complete. Banach spaces are complete normed vector spaces.

[3] An example is the space $\mathbf{R} \times \mathbf{R} = \mathbf{R}^2$ made of all coordinates in the Cartesian plane.

(iii) $\langle g, g \rangle = ||g||^2$.

If the inner product between two vectors is zero, they are said to be orthogonal. Orthogonality in $\mathbf{R}^2$ means that two vectors (arrows) are perpendicular. Therefore, orthogonality in this space allows one to apply well-known formulas, e.g. Pythagoras' Theorem, to understand the geometry of the space. Extending this concept to more complex vector spaces, orthogonality is a result that helps to simplify one's analysis when investigating the properties of a vector space. A Banach spaces where the inner product produces the the norm is called a *Hilbert space*.

Example 2.3 (Inner Product of $\mathbf{R}^3$)**.** *The inner product is defined by the dot product. Given two vectors $g, h \in \mathbf{R}^3$, the inner product is defined by:*

$$\langle g, h \rangle = (g_1, g_2, g_3) \cdot (h_1, h_2, h_3) = g_1 h_1 + g_2 h_2 + g_3 h_3.$$

The norm of $g \in \mathbf{R}^3$ is given by:

$$||g|| = (\langle g, g \rangle)^{1/2} = (g_1^2 + g_2^2 + g_3^2)^{1/2}.$$

Example 2.4 (Inner Product of $\mathbf{L}^2$)**.** $\mathbf{L}^2[0,1]$ *is a Hilbert space with inner product for functions $h_1, h_2 \in \mathbf{L}^2[0,1]$ defined by $\langle h_1, h_2 \rangle = \int_0^1 h_1(x) h_2(x) dx$ and norm $||h_1 - h_2|| = \{\langle h_1 - h_2, h_1 - h_2 \rangle\}^{1/2}$.*

Hilbert spaces encompass the concepts described above: they are complete vector spaces with defined norms induced by the inner product, which hold enough limits to perform calculus. Hilbert spaces include tangible vector spaces like $\mathbf{R}^2$ and more abstract vector spaces that may be finite dimensional or infinite dimensional. While elements in the finite dimensional $\mathbf{R}^2$ space can be specified using its Cartesian coordinates, elements in abstract Hilbert spaces are identified using an *orthonormal basis*.

An orthonormal basis is a sequence $\{g_n\}_{n=1}^{\infty}$ of nonzero vectors in Hilbert space $\mathcal{H}$ such that $\langle g_n, g_m \rangle = 0$ for $n \neq m$ and $\langle g_n, g_n \rangle = ||g_n|| = 1$. That is, all vectors in the sequence $\{g_n\}_{n=1}^{\infty}$ are unit vectors and orthogonal to each other (Li & Racine 2007). The sequence $\{g_n\}_{n=1}^{\infty} = 1$ is a

complete orthonormal basis of $\mathcal{H}$ if:

$$g = \sum_{j=1}^{\infty} \langle g, g_j \rangle g_j. \tag{2.4}$$

This result is known as the Fourier series for Hilbert spaces. The equality in (2.4) means that the right-hand-side of Equation (2.4) converges to g:

$$\lim_{n \to \infty} \left|\left| g - \sum_{j=1}^{n} \langle g, g_j \rangle g_j \right|\right| = 0.$$

Hilbert spaces are relevant as the nonparametric IV framework results in a mapping between two Hilbert spaces, where the mapping function is given by a Fredholm equation of the first kind. As will be shown below, the Fourier series for Hilbert spaces are used in developing the solution to a Fredholm equation of the first kind. Fredholm equations are often expressed using operator notation, which I discuss next.

2.2 Linear Operators

Define two Hilbert spaces $\mathcal{H}$ and $\mathcal{V}$ with a function $g \in \mathcal{H}$ and a function $r \in \mathcal{V}$. To simplify the notation when working with Fredholm equations, it is helpful to define an operator function $T : \mathcal{H} \to \mathcal{V}$ such that

$$T[g(x)](z) = \int_a^b f(x, z) g(x) dx,$$

for $g \in \mathcal{H}$. Then, (2.1) becomes:

$$\begin{aligned} r(z) &= \int_a^b f(x, z) g(x) dx \\ \Rightarrow r &= Tg \end{aligned} \tag{2.5}$$

The operator $T : \mathcal{H} \to V$ is called *linear* if, for every pair of elements g and h, and scalar $\alpha \in \mathbf{R}$, the operator is distributive across addition and $T(\alpha g) = \alpha(Tg)$. The domain of T is the subset of $\mathcal{H}$ on which T is defined and the range of T is the set $\mathcal{R}(T) = \{r \in \mathcal{V} : r = Tg \text{ for some } g \in \mathcal{H}\}$. If an element in $g \in \mathcal{H}$ does not map to $\mathcal{R}(T)$, then g is said to belong to the null space of T,

$\mathcal{N}(T)$.When $\mathcal{R}(T)$ is finite, T is said to be a finite dimensional operator and one may use linear algebra techniques to describe it. Furthermore, if T is a finite dimensional linear operator on $\mathcal{H}$ then

$$Tg = \sum_{i=1}^{N} \alpha_i \langle g, \phi_i \rangle \psi_i,$$

for vectors $\phi_1, \phi_2, ..., \phi_N \in \mathcal{H}$ and $\psi_1, \psi_2, ..., \psi_N \in \mathcal{V}$. $\{\phi_i\}_{i=1}^N \in \mathcal{H}$ and $\{\psi_i\}_{i=1}^N \in \mathcal{V}$ are orthonormal bases and $\alpha_i \in \mathbf{R}^+$. T is said to have a rank of $N < \infty$. An important linear operator in the nonparametric IV framework is the conditional expectations operator.

Example 2.5 (**Conditional Expectation Operator).** *The conditional expectation operator* $T : \mathbf{L}_X^2 \to \mathbf{L}_Z^2$*, which maps elements from the space of square integrable functions on* X *to the space of square integrable functions on* Z *is denoted by:*

$$T[g(x)](z) = E(g(x)|z) = \int g(x) f(x|z) dx, \tag{2.6}$$

where $g \in \mathbf{L}_X^2$ *and* $f(x|z)$ *is the conditional probability density function of* x *given* z*.*

The goal of Equation (2.5) is to solve for the unknown quantity $g(\cdot)$. A solution exists so long as $T : \mathcal{H} \to \mathcal{V}$ is *bijective*. That is, there is a mapping from $\mathcal{H}$ to every element in $\mathcal{V}$ and each $g \in \mathcal{H}$ maps to a unique element in $\mathcal{V}$. Another way of putting it is that the null space of a bijective operator is empty and every $Tg \in \mathcal{R}(T)$ is unique. Bijective operators are invertible, which means that for each $r \in T(\mathcal{H})$ there is only one element $g \in \mathcal{H}$ with $T^{-1}r = g$. If an operator T holds these properties, the problem is said to be *well-posed* (Hardamard 1923).

Definition 2.1 (**Well-posed problem).** *Let* $T : \mathcal{H} \to \mathcal{V}$ *be an operator from a subset* $\mathcal{H}$ *of a space* X *into a subset* $\mathcal{V}$ *of a space* Z*. The operator equation*

$$T[g(x)](z) = r(z)$$

is called well-posed if:

1. *T is bijective;*

2. The inverse operator $T^{-1} : \mathcal{V} \to \mathcal{H}$ is continuous.

When a problem is well-posed, the parameter of interest can be estimated consistently by replacing the unknown population distribution with a consistent sample analog. A continuous mapping preserves the consistency of the estimated parameter if the difference between the sample analog and true population distribution is small (Horowitz 2013). In order for an operator T to be continuous, it must be *bounded*. In other words, it must be that for all Tg in the range $\mathcal{V}$, the norm of $Tg \in \mathcal{R}(T)$ is bounded above by the scaled norm of $g \in \mathcal{H}$, $M||g||_{\mathcal{H}}$, where M is an arbitrary positive constant (Carrasco et al. 2007).

If for all $\{g_n\}$ in the domain $\mathcal{H}$, the sequence $\{Tg_n\}$ in the range $\mathcal{V}$ is such that $\lim_{n\to\infty} ||Tg_n - Tg|| = 0$, then T is said to be a *compact linear operator* (Gorodnik 2016)[4]. Finite dimensional operators are compact. In fact, a necessary condition for Equation (2.5) to be well-posed is that $\mathcal{H}$ and $\mathcal{V}$ must be finite dimensional (Kress 1999). Before finding a solution, one needs the concept of the adjoint operator.

2.2.1 Adjoint Operators

For a bounded operator $T : \mathcal{H} \to \mathcal{V}$, the *adjoint operator* $T^* : \mathcal{V} \to \mathcal{H}$ holds the property that: $\langle Tg, r\rangle_{\mathcal{V}} = \langle g, T^*r\rangle_{\mathcal{H}}$. Also, T^* is bounded with $||T|| = ||T^*||$.

Example 2.6 (Adjoint of the Conditional Expectation Operator). *Denote $T^* : \mathbf{L}_Z^2 \to \mathbf{L}_X^2$ as the adjoint operator for the conditional expectations operator in* (2.6)*. Then T^* is defined by:*

$$T^*[h(z)](x) = E(h(z)|x) = \int h(z)f(z|x)dz, \tag{2.7}$$

for some function $h \in \mathbf{L}_Z^2$.

If $T = T^*$, then T is called self-adjoint. Theorem 15.9 in Kress (1999) states that if T is a

[4]From Alexander Gorodnik's MATH 36202/M6202 course notes at the University of Bristol, Lecture 5. URL: https://people.maths.bris.ac.uk/ mazag/fa/lecture5.pdf

bounded self-adjoint operator, then the norm of $T : \mathcal{H} \to \mathcal{H}$ is given by:

$$||T|| = \sup_{|g|=1} |\langle Tg, g\rangle_{\mathcal{H}}|.$$

Note that the operator T^*T is self-adjoint. If T is self-adjoint and $||T|| > 0$, T is said to be a *positive operator*.

The solution for g in an operator equation $Tg = r$ where $T : \mathcal{H} \to \mathcal{V}$ is an operator projecting elements from Hilbert space $\mathcal{H}$ into Hilbert space $\mathcal{V}$ is known as *Picard's Theorem* (Kress 1999). The theorem makes use of singular value notation; the following is a brief overview of some of these concepts.

2.3 Singular Value Decomposition

Singular value decomposition refers to the ability to write the operator mapping of an element from a vector space as a scaled version of that element. When the mapping of a vector can be written as the vector multiplied by a scalar, the scalar is known as an *eigenvalue*. If the elements of the vector space are functions, the eigenvalues are known as eigenfunctions. More formally, an eigenvalue of a bounded operator $T : \mathcal{H} \to \mathcal{V}$ is a scalar λ such that $T\phi = \lambda\phi$, where ϕ is a non-zero vector in $\mathcal{H}$. The eigenvalues of a bounded self-adjoint positive operator T are nonnegative and bounded by the norm of T, where $\max(\lambda) = ||T||$. In contrast, the eigenvalues of an infinitely large self-adjoint operator shrink to zero.

Denoting $\{\lambda_j^2\}_{j=1}^n$ as the eigenvalues for the nonnegative self-adjoint compact operator $T^*T : \mathcal{H} \to \mathcal{H}$, the *singular values* are defined as $\lambda_j = \sqrt{\lambda_j^2}$. Important results on the singular value decomposition of linear operators have been summarized in the following theorem from Darolles et al. (2011).

Theorem 2.1 (From Darolles et al. (2011))**.** *Let $\{\lambda_j\}_{j=1}^n$ denote the sequence of the nonzero singular values of the compact linear operator T repeated according to their multiplicity. Then there exist orthonormal sequences $\{\phi_j\}_{j=1}^n$ of $\mathcal{H}$ and $\{\psi_j\}_{j=1}^n$ of $\mathcal{V}$ such that:*

(i) $T\phi_j = \lambda_j\psi_j, \;\; j \geq 0.$

(ii) $T^*\psi_j = \lambda_j\phi_j,\ j \geq 0.$

(iii) $\phi_0 = 1$ *and* $\psi_0 = 1.$

(iv) $\langle\phi_i, \psi_j\rangle = \lambda_i\delta_{ij},\ i, j \geq 0.$ $\delta_{ij} = 1$ *if* $i = j$ *and zero otherwise is known as the Kronecker symbol.*

(v) For all $g \in \mathcal{H}$, $g(x) = \sum_{j=0}^{\infty}\langle g, \phi_j\rangle\phi_j(x) + \bar{g}(x)$, *where* $\bar{g}(x) \in \mathcal{N}(T)$.

(vi) For all $h \in \mathcal{V}$, $h(x) = \sum_{j=0}^{\infty}\langle h, \psi_j\rangle\psi_j(x) + \bar{h}(x)$, *where* $\bar{h}(x) \in \mathcal{N}(T^*)$.

Then

$$T[g(X)](z) = \sum_{j=1}^{\infty}\lambda_j\langle g, \phi_j\rangle\psi_j(z) \tag{2.8}$$

and

$$T^*[h(Z)](x) = \sum_{j=1}^{\infty}\lambda_j\langle h, \psi_j\rangle\phi_j(x). \tag{2.9}$$

$\{\lambda_j, \phi_j, \psi_j\}$ *is called the singular system of* T. *Note that* λ_j^2 *are the eigenvalues of* TT^* *and* T^*T *associated with the eigenfunctions* ψ_j *and* ϕ_j, *respectively.*

Using Theorem 2.1, Equation (2.5) can be re-written as:

$$\sum_{j=1}^{\infty}\lambda_j\langle g, \phi_j\rangle\psi_j = \sum_{j=1}^{\infty}\langle r, \psi_j\rangle\psi_j \tag{2.10}$$

(Centorrino et al. 2014). The solution to the Fredholm equation of the first kind $Tg = r$ is then given by Picard's Theorem:

$$g = \sum_{j=1}^{\infty}\frac{1}{\lambda_j}|\langle r, \psi_j\rangle|\phi_j \tag{2.11}$$

(Kress 1999). If T is infinite dimensional, then $\lim_{j\to\infty}\lambda_j = 0$. It should now become clear that in order for (2.5) to be a well-posed problem, $\mathcal{H}$ and $\mathcal{V}$ must be finite dimensional. Horowitz (2011) provides the following rationale: Define two functions $r_1, r_2 \in \mathbf{L}^2$ as

$$r_1 = \sum_{j=1}^{\infty}\langle r_1, \psi_j\rangle\psi_j$$

and

$$r_2 = \sum_{j=1}^{\infty} \langle r_2, \psi_j \rangle \psi_j.$$

Next, define g_1 and g_2 as the approximate solutions for g given r_1 and r_2, respectively:

$$g_1 = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle r_1, \phi_j \rangle \psi_j$$

and

$$g_2 = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle r_2, \phi_j \rangle \psi_j.$$

If T is an infinite dimensional operator, the limit for the singular values for T^*T is zero, i.e. $\lambda_j \to 0$ as $j \to \infty$. For any arbitrarily small ε such that

$$||r_1 - r_2|| = \left[\sum_{j=1}^{\infty} (\langle r_1, \psi_j \rangle - \langle r_2, \psi_j \rangle)^2 \right]^{1/2} < \varepsilon$$

the difference

$$||g_1 - g_2|| = \left[\sum_{j=1}^{\infty} \left(\frac{\langle r_1, \phi_j \rangle - \langle r_2, \phi_j \rangle}{\lambda_j} \right)^2 \right]^{1/2}$$

can be made arbitrarily large. Therefore, small perturbations in the data can create vastly different estimates for g (Horowtiz 2011).

If T is finite dimensional, such that $r_1 = \sum_{j=1}^{J} \langle r_1, \psi_j \rangle \psi_j$ and $r_2 = \sum_{j=1}^{J} \langle r_1, \psi_j \rangle \psi_j$ for some finite J, $\lambda_j > 0$ for all $j = 1, ..., J$ and small changes in data cause only small changes in the estimated parameter. As mentioned above, finite-dimensional operators are compact and therefore continuous. The difference between the estimated and the true parameter values is small if the difference between the sample analog and true population distribution is small.

If at least one of the conditions in Definition 2.1 is not met, the problem is called *ill-posed*, and if the ill-posed nature of the problem is caused by inverting a continuous mapping, the problem is considered an *ill-posed inverse problem*. In this case, it is no longer true that the difference between the estimated and true parameter values is small because we replace the population distribution

with a consistent sample analog. When T is an infinite-dimensional operator, g in Equation (2.5) is the solution to an ill-posed problem.

3 Well-Posed Problems and Parametric IV

Parametric methods for solving the endogeneity problem in econometric models begin by specifying the functional form of $g(\cdot)$ in (1.1). If $g(\cdot)$ were known to a finite-dimensional parameter β, then we could re-write the equation as:

$$Y = g(X, \beta) + \epsilon; \ E(\epsilon|X) \neq 0, \tag{3.1}$$

so that some of the K variables in X may be correlated with ϵ. Y is scalar valued and the parameter of interest, β, is a Kx1 vector of coefficients. There are L variables, denoted by Z, that have two properties (Greene 2012):

1. uncorrelated with the disturbances, i.e. $E(\epsilon|Z) = 0$; and
2. they are correlated with the independent variables X.

If $g(\cdot)$ is specified to be linear, then:

$$Y = X\beta + \epsilon. \tag{3.2}$$

In this case, ordinary least squares estimates will be biased and inconsistent. Taking the expectation conditional on Z in Equation (3.2) we get:

$$E(Y|Z) = E(X|Z)\beta. \tag{3.3}$$

Define now, the linear operator $T : X \to E(X|Z)$. T is a conditional expectations operator, which projects X onto the space of the instrument Z. Setting $r = E(Y|Z)$, equation (3.3) could then be written as:

$$(TX)\beta = r, \tag{3.4}$$

because T is a linear operator.

Let $T^* : Z \to E(Z|X)$ denote the adjoint operator, which projects an object on the space of X. Multiplying both sides of (3.4) by the adjoint operator and solving for β:

$$\begin{aligned}(T^*T)\beta &= T^*r \\ \beta &= (T^*T)^{-1}T^*r. \end{aligned} \tag{3.5}$$

β is identified if T^*T is finite and nonsingular. While viewing this problem in operator notation is unconventional, it will help to have this foundation when approaching nonparametric IV methods. Assuming a linear relationship between X and Z, i.e. $X = Z\gamma + \eta$, β can be estimated by:

$$\begin{aligned}\hat{\beta} &= (\hat{T}^*\hat{T})^{-1}\hat{T}^*r \\ &= ([Z(Z'Z)^{-1}Z'X]'Z(Z'Z)^{-1}Z'X)^{-1}[Z(Z'Z)^{-1}Z'X]'Y \\ &= (X'P_ZX)^{-1}X'P_ZY, \end{aligned} \tag{3.6}$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix (Centorrino et al. 2014). The condition for identification is that $RANK(Z) = L \geq RANK(X) = K$. When $L = K$, β is said to be just identified. Assuming that the disturbance ϵ is homoskedastic, β can be estimated by:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y. \tag{3.7}$$

If, however, $L > K$ then the parameter is over-identified and two-staged least squares can be implemented:

$$\begin{aligned}\hat{\beta}_{2SLS} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y, \end{aligned} \tag{3.8}$$

where $\hat{X} = Z(Z'Z)^{-1}Z'X$ [5]. Equation (3.8) is a special case of the Generalized Method of Moments (GMM) estimator (Baum, Schaffer & Stillman 2003). The L instruments give us a set of L moment

[5] $P_Z = Z(Z'Z)^{-1}Z'$ is idempotent, i.e. $P_ZP_Z = P_Z$.

conditions:

$$g_i(\beta) = Z_i'(Y_i - X_i\beta). \tag{3.9}$$

The orthogonality conditions are satisfied at the true value of β, i.e. $E(g_i(\beta)) = 0$. We estimate the parameter using the L sample moments:

$$\bar{g}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}). \tag{3.10}$$

The GMM estimate of β is:

$$\hat{\beta}_{GMM} = (X'ZW_NZ'X)^{-1}X'ZW_NZ'Y, \tag{3.11}$$

where W_N is an $L \times L$ full rank symmetric weighting matrix, which must be estimated in order to derive $\hat{\beta}_{GMM}$ (see Cameron and Trivedi (2007), Section 6.4). When errors are homoskedastic, (3.11) reduces to (3.8) (Baum et al. 2003).

The GMM method can be extended in the case that $g(X, \beta)$ in Equation (3.1) is defined as nonlinear. The population moment conditions and sample moments are defined as:

$$E(g(X, \beta)) = 0$$

and

$$\frac{1}{n}Z'\epsilon = 0,$$

respectively. These GMM estimators are $n^{-1/2}$ consistent and asymptotically normal.

Equation (3.5) is an example of a well-posed problem. As Horowitz (2013) explains, Equation (3.5) uniquely determines β if the inverse matrices on the right-hand side are non-singular. Therefore, we can use (3.5) to identify β in (3.2). Furthermore, β is a continuous function of the population moments and the probability distributions of the random variables on the right-hand side of Equation (3.5). Following the above definition, we can consistently estimate β by replacing the unknown population moments with their sample moments, as is done in Equation (3.7) and

Equation (3.8) for the cases where $L = K$ and $L > K$, respectively.

4 Nonparametric Instrumental variables

The general framework in the nonparametric IV literature begins by defining $S = (Y, X, Z) \in \mathbf{R} \times \mathbf{R}^K \times \mathbf{R}^L$. S is characterized by a cumulative density function $F(S)$ with respect to the *Lebesgue measure*, and has probability distribution $f(S)$.

Definition 4.1 (**Lebesgue measure**). *A Lebesgue measure m is a measure defined on the real line $\mathbf{R}$ with the property that for any interval $[a, b]$ $(b \geq a)$, $m([a, b]) = b - a$. That is the Lebesgue measure of an interval equals the length of the interval. Any single point has a Lebesgue measure of zero (Li & Racine 2007).*

As in the literature, the space considered is the Hilbert space of square integrable functions $\mathbf{L}^2$ of S, given by:

$$\mathbf{L}^2 = \left\{ h : \int h(s)^2 dx < \infty \right\}$$

Let $||\cdot||$ and $\langle \cdot, \cdot \rangle$ denote the norm and inner product in $\mathbf{L}^2$, respectively:

$$||h|| = \left[\int h(s)^2 ds \right]^{1/2}$$

$$\langle g, h \rangle = \int g(s)h(s)ds$$

Consider again the model in Equation (1.1), only now g satisfies regularity conditions but is otherwise unknown. The first step in estimating $g(\cdot)$ is to derive a mapping from S to g that identifies g. Taking the expectation of Equation (1.1) conditional on Z we get:

$$E(Y|Z) = E(g(X)|Z) \tag{4.1}$$

$$E(Y|Z) = \int g(x) f_{X|Z}(x, z) dx$$

$$E(Y|Z) = \int g(x) \frac{f_{XZ}(x, z)}{f_Z(z)} dx, \tag{4.2}$$

where $f_{X|Z}(x,z)$, $f_{XZ}(x,z)$, and $f_Z(z)$ denote the density of X conditional on Z, the joint density of (X,Z), and density of Z, respectively. Note the similarity between equations (4.2) and (2.1); the solution to a nonparametric instrumental variable problem is the solution to a Fredholm equation of the first kind.

Denote $\mathbf{L}_i^2$ as the set of square integrable function equations on variable i only, $i = Y, X, Z$. Define T as an operator as the conditional expectations operator on $\mathbf{L}^2$ such that:

$$T : \mathbf{L}_X^2 \to \mathbf{L}_Z^2,$$

$$T[g(x)](z) = E(g(X)|Z) = \int g(x) f_{X|Z}(x,z) dx.$$

The adjoint operator of T, denoted by T^* is a mapping such that:

$$T^*[h(z)](x) = E(h(Z)|X) = \int h(z) f_{Z|X}(z,x) dz$$

$h \in L_Z^2$. Defining $E(Y|Z) = r(z)$ we can write Equation (4.2) as:

$$r = Tg. \tag{4.3}$$

If T was finite dimensional and T^*T was non-singular, then we could get a solution for g as:

$$g = (T^*T)^{-1} T^* r. \tag{4.4}$$

However, we are not restricting g other than following basic regularity conditions; g may be infinite-dimensional. The conditional expectation r is finite dimensional and can be estimated from the data. Therefore, T^* and T^*T are potentially infinite dimensional. Let $\{\lambda_j : j = 1, 2, ...\}$ denote the eigenvalues for T^*T. As $j \to \infty$, $\lambda_j \to 0$, rendering T^*T singular (Horowitz 2013). This means the inverse mapping from r to g is discontinuous. This is not an issue if T and T^* were finite dimensional, because the eigenvalues of a non-singular finite dimensional matrix are bounded away from zero (See Section 2.3).

Therefore, the problem of estimating g in Equation (4.3) is an ill-posed inverse problem. One can construct a stable approximate solution by using regularization methods. Regularizaton is the method by which we find an approximation of the unbounded inverse operator $T^{-1} : Tg \to g$ by a bounded linear operator (Kress 1999). The identification process of g in a nonparametric instrumental variable framework can be generalized to two steps, estimation of conditional mean functions and regularization.

4.1 Estimating conditional means

There are two central methods for estimating conditional mean functions in nonparametric regression analysis: kernel based methods and series methods. Both methods require the selection of a smoothing parameter which can be either plug-in values or derived from data-driven techniques (Li & Racine 2007). Here, I focus only on kernel based methods as (to the best of my knowledge) there is no fully developed probability weighted series method for estimating conditional means to account for sample design.

In order to use kernels to estimate r and T, we must first estimate the probability density functions f_{XZ}, f_X, and f_Z. Partition the predictor variables X into two types, continuous and discrete, i.e. $X = (X^c, X^d)$, where the superscripts c and d denote continuous and discrete variables, respectively. I use X_{is}^c to denote the s^{th} component of X_i^c and X_{is}^d for the sth component of X_i^d and assume that X_j^d takes $c_t \geq 2$ different values in $\mathcal{D}_t = \{0, 1, ..., c_t - 1\}$, $t = 1, ..., r$. To define a univariate kernel function for a discrete variable, one may use a variation of the Aitchison and Aiken (1976) kernel described in Li and Racine (2007, p.131):

$$l(x_{is}^d, x_{js}^d, \lambda_s^d) = (\lambda^d)_s^{1-\mathbf{1}(x_{is}^d = x_{js}^d)}, \tag{4.5}$$

which takes a value of 1 if $x_{is}^d = x_{js}^d$ and λ_s^d otherwise, where λ_s^d is the smoothing parameter for x_s^d. The smoothing parameter λ_s^d satisfies $0 \leq \lambda_s^d \leq 1$, with $\lambda_s^d = 0$ corresponding to an indicator function and $\lambda_s^d = 1$ giving equal weights to all values of its arguments. The product kernel for

discrete variables X^d is

$$L(x_i^d, x_j^d, \lambda^d) = \prod_{s=1}^{r} l(x_{is}^d, x_{js}^d, \lambda_s^d). \tag{4.6}$$

For continuous variables, let $k(\cdot)$ denote a symmetric, univariate kernel weight function. The product kernel for continuous variables X^c is defined as:

$$W(x_{is}^c, x_{js}^c) = \prod_{s=1}^{K} \frac{1}{h_{x,s}} k\left(\frac{x_{is}^c - x_{js}^c}{h_{x,s}}\right), \tag{4.7}$$

where $h_{x,s}$ is the bandwidth for variable X_s, $s = 1, ..., K$. Similarly, the product kernel for the instrumental variables Z is given by: $W(z_{is}^c, z_{js}^c) = \prod_{s=1}^{L} h_{z,s}^{-1} k((z_{is}^c - z_{js}^c)/h_{z,s})$, where $h_{z,s}$ is the bandwidth for variable Z_s, $s = 1, ..., L$. A mixed-data product kernel for predictor variables X is given by: $K_{\gamma,ix} = W(x_{is}^c, x_{js}^c) L(x_i^d, x_j^d, \lambda^d)$.

Denote $\tilde{f}_{XZ}$, $\tilde{f}_X$, and $\tilde{f}_Z$ as the kernel density estimators for f_{XZ}, f_X, and f_Z, repectively, then:

$$\tilde{f}_{XZ} = \frac{1}{n} \sum_{i=1}^{n} K_{\gamma,ix} W(z_{is}^c, z_{js}^c), \tag{4.8}$$

$$\tilde{f}_X = \frac{1}{n} \sum_{i=1}^{n} K_{\gamma,ix}, \tag{4.9}$$

$$\tilde{f}_Z = \frac{1}{n} \sum_{i=1}^{n} W(z_{is}^c, z_{js}^c). \tag{4.10}$$

Then, the estimators of T, T^*[6], and r are given:

$$\tilde{T}[g(X)](z) = \int g(x) \frac{\tilde{f}_{XZ}(x, z)}{\tilde{f}_Z(z)} dx, \tag{4.11}$$

$$\tilde{T^*}[h(Z)](x) = \int h(z) \frac{\tilde{f}_{XZ}(x, z)}{\tilde{f}_X(x)} dz, \tag{4.12}$$

$$\tilde{r}(z) = \frac{\sum_{i=1}^{n} y_i W(z_{is}^c, z_{js}^c)}{\sum_{i=1}^{n} W(z_{is}^c, z_{js}^c)}. \tag{4.13}$$

[6]Carrassco et al. (2007) showed that an operator between these two Hilbert spaces has the property that the adjoint of the estimated operator is equivalent to the estimated adjoint operator.

4.2 Complex Surveys

Consider now a finite population $U = \{1, ..., N\}$ of N units. For each $j \in U$ the outcome variable $Y_j \in \mathbf{R}$ and auxiliary variables $X_j = (X_j^d, X_j^c) = (X_{j1}^d, ..., X_{jr}^d, X_{j1}^c, ..., X_{jq}^c) \in \mathbf{R}^K$ are observed. X_j is a $(q + r) \times 1$ vector where the superscripts d and c denote that the variable is discrete and continuous, respectively. To make inferences on the population, a probability sample $\mathcal{S}$ of size $n_{\mathcal{S}}$ is drawn based on a complex sampling plan with each $j \in U$ having probability π_j of being selected in the sample. In order to account for the sampling design in the nonparametric IV estimation, I propose replacing the population totals from equations (4.8), (4.9), and (4.10), with their probability-weighted estimators:

$$\hat{f}_{XZ} = \frac{1}{n} \sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,ix} W(z_{is}^c, z_{js}^c), \tag{4.14}$$

$$\hat{f}_X = \frac{1}{n} \sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,ix}, \tag{4.15}$$

$$\hat{f}_Z = \frac{1}{n} \sum_{i=1}^{n} \pi_i^{-1} W(z_{is}^c, z_{js}^c). \tag{4.16}$$

giving model-assisted estimators for r, T, and T^*:

$$\hat{r}(z) = \frac{\sum_{i=1}^{n} \pi_i^{-1} y_i W(z_{is}^c, z_{js}^c)}{\sum_{i=1}^{n} \pi_i^{-1} W(z_{is}^c, z_{js}^c)}, \tag{4.17}$$

$$\hat{T}[g(X)](z) = \frac{\sum_{i=1}^{n} \pi_i^{-1} \hat{r}(z_i) W(x_{is}^c, x_{js}^c)}{\sum_{i=1}^{n} \pi_i^{-1} W(x_{is}^c, x_{js}^c)}, \tag{4.18}$$

$$\hat{T^*}[h(Z)](x) = \frac{\sum_{i=1}^{n} \pi_i^{-1} \hat{E}(\hat{r}(z)|x)_i W(z_{is}^c, z_{js}^c)}{\sum_{i=1}^{n} \pi_i^{-1} W(z_{is}^c, z_{js}^c)}. \tag{4.19}$$

If the sample was taken using simple random sampling (SRS), it is fair to assume the data is independently and identically distributed (i.i.d.) and f_{XZ}, f_X, and f_Z can be estimated using (4.8), (4.9), and (4.10), respectively. If the sample was selected such that the probability of inclusion was not equal across all units, then the sampling design should be taken into account.

4.2.1 Cross-validation

An important aspect of kernel estimation methods is the selection of the smoothing parameters $\gamma = (h, \lambda^d)$. It is generally preferred that one uses a data-driven method for selecting the bandwidths for sound analysis. Furthermore, there are no rule-of-thumb selection methods for the bandwidths of discrete variables, λ^d. When estimating conditional mean functions using survey data, Clair (2019) and Breidt and Opsomer (2005) suggest using least squares cross-validation. For example, in estimating the conditional mean function $r(z) = E(Y|Z)$ using (4.17), one would choose h_z that minimizes the following cross-validation objective function:

$$CV(h_z) = \sum_{i \in \mathcal{S}} (y_i - \hat{r}_{-i}(z_i))^2 M(z_i), \tag{4.20}$$

where $r_{-i}(z_i) = \sum_{j \in \mathcal{S}, j \neq i} \pi_j^{-1} y_j W_{z,ij} / (\sum_{j \in \mathcal{S}, j \neq i} \pi_j^{-1} W_{z,ij})$ is the leave-one-out kernel estimator of $r(z_i)$ and $W_{z,ij} = \Pi_{j \neq i}^{q} h_{z,s}^{-1} k((z_{is}^c - z_{js}^c)/h_{z,s})$. $0 < M(z_i) < 1$ is a weight function which serves to avoid difficulties caused by dividing by zero.

4.3 Landweber-Fridman Regularization

Regularization is the process of converting a discontinuous mapping into a continuous one. Landweber-Fridman (LF) regularization seeks to restore a continuous inverse mapping by using an iterative scheme for finding a fixed-point solution for $g(X)$. Therefore, LF regularization solves the ill-posed problem without the need to invert the $\hat{T}^* \hat{T}$ matrix.

To implement LF regularization, first define a constant c where $0 < c < 1/||T||^2$. Because T is a compact linear operator, there exists an operator

$$A_\alpha = c \sum_{k=0}^{\alpha} (I - c\hat{T}^* \hat{T})^k \hat{T}^*$$

where $\hat{g}_\alpha = A_\alpha \hat{r}$ (See Kress (1999) Theorem 15.27). The approximation $\hat{g}_\alpha = A_\alpha \hat{r}$ corresponds to

α steps of the iteration scheme:

$$\begin{aligned}\hat{g}_m &= \hat{g}_{m-1} + c\hat{T}^*(\hat{r} - \hat{T}\hat{g}_{m-1}) \\ &= \hat{g}_{m-1} + c\hat{E}[\hat{E}[Y - \hat{g}_{m-1}|Z]|X],\end{aligned} \tag{4.21}$$

$m = 1, ..., \alpha$. α is the regularization parameter; it determines the number of iterations of the LF regularization procedure. The role of c is to ensure the convergence of the iterative solution. In nonparametric IV estimation, T is defined as the conditional expectations operator and therefore, $||T|| = 1$. As long as $c < 1$, the solution will converge (Centorrino et al. 2014). In order to solve for g, the conditional expectations in Equation (4.21) are estimated nonparametrically. The algorithm is as follows:

1. The process requires a starting value for $g(X)$; Centorinno et al. (2014) suggest $\hat{g}_0(X) = c\hat{T}\hat{r} = c\hat{E}(\hat{E}(Y|Z)|X)$.

2. Next, given $\hat{g}_0$, we can compute $\hat{E}(Y - \hat{g}_0(X)|Z)$.

3. Next, compute $\hat{E}(\hat{E}(Y - \hat{g}_0(X)|Z)|X)$.

4. From 4.21, compute $\hat{g}_1(X)$ as:

$$\hat{g}_1 = \hat{g}_0 + c\hat{E}[\hat{E}[Y - \hat{g}_0|Z]|X].$$

5. For $m = 2, 3, ...$ repeat steps 2 to 4, until a stopping rule stabilizes from iteration to iteration. As a stopping rule, Florens and Racine (2012) propose iterating until the following objective function is minimized:

$$SSR(m) = m\sum_{m=1}^{n}[\hat{E}(Y|Z) - \hat{E}(\hat{g}_m(X)|Z)]^2,\ m = 1, 2, ...$$

5 Monte Carlo Simulations

Simulations that compare nonparametric IV methods to parametric IV methods for different functional forms of the DGP have been conducted in Centorrino et al. (2015). In this section, I compare the finite-sample properties of the weighted nonparametric IV estimator to the model based nonparametric IV estimator and other nonparametric estimators that do not correct for an endogenous regressor. I base my simulations on those by Centorrino et al. (2015) and fix the joint distribution of the endogenous variable and the instrument and I perform sensitivity analysis on the level of noise in the model. The population is set to $N = 10,000$ and the main data generating process is given by:

$$Y = 0.5X_1^2 + X_2 + \epsilon,$$

where $E(\epsilon|X_1) \neq 0$. Endogeneity is modelled by first independently simulating the instrument Z, the exogenous predictor X_2, and two disturbances η and u. Then, I define the error ϵ as a function of η and u:

$$Z \sim N(0,1) \tag{5.1}$$

$$X_2 \sim U(0,1) \tag{5.2}$$

$$\eta \sim N(0, 0.27) \tag{5.3}$$

$$u \sim N(0, \sigma) \tag{5.4}$$

$$\epsilon = -0.5\eta + u \tag{5.5}$$

X_1 is then given by a function of the instrument Z and η:

$$X_1 = 0.2Z + \eta.$$

A sample of size n is drawn based on one of three sample plans: SRS, stratification of the Y variable, and stratification of the X_2 variable. Stratifying based on Y induces endogenous stratification into the model and stratifying based on X_2 induces exogenous stratification. In both stratification

schemes, the sample criterion is divided into three separate strata. From each strata, a sample of size $n/3$ is taken. Because the strata sizes differ, there are unequal inclusion probabilities across units in the population (See Table 1).

I estimate $g(X) = 0.5X_1^2 + X_2$ using four estimators: a probability weighted nonparametric IV (WNPIV) estimator, a nonparametric IV (NPIV) estimator, the modified local constant estimator (WLC) proposed by Sánchez-Borrego et al. (2014)

$$WLC = \frac{\sum_{i=1}^{n} \pi_i^{-1} y_i K_{\gamma,xi}}{\sum_{i=1}^{n} \pi_i^{-1} K_{\gamma,xi}}, \tag{5.6}$$

and the local constant (LC) estimator proposed by Nadaraya (1964) and Watson (1964)

$$LC = \frac{\sum_{i=1}^{n} y_i K_{\gamma,xi}}{\sum_{i=1}^{n} K_{\gamma,xi}}. \tag{5.7}$$

For WNPIV estimation, I use the weighted local estimator to estimate the conditional mean functions as in equations (4.17) to (4.19) and I use the least squared cross-validation method from Clair (2019) described in Section 4.2.1 to select the smoothing parameters for each regression. For the NPIV estimation, I use the traditional local constant estimator to estimate the conditional mean functions as in (4.11) to (4.13). The bandwidths are selected using the least squares cross-validation method for mixed-data types described in Li and Racine (2004) and (2007). For both instrumental variable estimators, I use the Landweber-Fridman approach to derive the regularized solution and set the maximum number of iterations to 1000. I then compute the mean squared error (MSE) for each estimator. I run 1000 Monte Carlo replications varying the sample size $n = 300, 600, 900$ and varying the level of noise, σ=0.05,0.10,0.25, in Equation (5.4).

Table 2 presents the results from SRS. The values in each cell are the median MSE values and the values in the round brackets are the median absolute deviations (MAD) of the MSE where

$$\text{MAD}[MSE(\hat{g}(X))] = \text{Median}\big[\big|MSE(\hat{g}(X)) - \text{Median}[MSE(\hat{g}(X))]\big|\big],$$

where $\hat{g}(X)$ is one of WNPIV, NPIV, WLC, or, LC. As is expected, the median values for the WN-

PIV estimator are equal to those of the NPIV estimator. Under SRS, the weighted nonparametric local constant estimator from Sánchez-Borrego et al. (2014) reduces to the traditional local constant proposed by Nadaraya (1964) and Watson (1964). Also, note that the median MSEs for the for the instrumental variable estimators are lower than the MSEs for the local constant estimators which do not correct for the endogeneity caused by X_1. Keeping σ constant, as n increases, the median MSEs for both IV estimators decrease, providing evidence that WNPIV and NPIV are consistent estimators. The values for WLC and LC, however, are increasing or remaining unchanged. Looking at the 95 percent confidence intervals of the MSEs in Table 3, as n increases (keeping σ constant) the lower and upper bounds of the MSEs for the IV estimators decrease and the confidence intervals do not intersect. The lower and upper bounds for the WLC and LC estimators are equal for all values of n, keeping σ constant, providing further evidence these estimators are inconsistent with an endogenous predictor variable.

Table 4 presents the results from stratification on the Y variable. When the sampling scheme is such that there are unequal probabilities of being selected in the sample and sampling is endogenous, improvements in efficiency can be made by using a probability weighted estimator. For each combination of n and σ, the median mean squared errors are lowest for the probability weighted IV estimator. The 95 percent confidence intervals for WNPIV in Table 5 have upper bounds smaller than the lower bounds of the MSEs for NPIV for all combinations of n and σ. Results in Table 4 also suggest that the endogeneity imposed by an endogenous regressor is more severe than endogeneity imposed by endogenous sampling. The median values for the MSEs of NPIV, which corrects for the endogenous regressor X_1, are smaller than the median values of MSEs for WLC, which corrects for endogeneity caused by a correlation between the sampling criterion and the error term. Again, keeping σ constant, as n increases the median MSEs for WLC and LC remain unchanged in Table 4 with intersecting confidence intervals in Table 5.

Finally, Table 6 displays the results from stratification on the exogenous variable X_2. Because X_2 is included on the right hand side, sample design is controlled for and the results from the WNPIV estimator are equal to those of the unweighted NPIV estimator. Similarly, the median values for the WLC and LC estimators are equal for all values of n and σ. These results coincide with

Clair (2019). Overall, these simulations show that ordinary nonparametric regression estimators are inconsistent in the presence of an endogenous regressor and that weighting each observation by the inverse of the unit's probability of inclusion does not reduce efficiency and may produce a great benefit if sampling is endogenous. As shown in Centorrino et al. (2015), if one chooses a linear parametric specification to estimate a quadratic or more complex function, the estimated coefficients will be biased and inconsistent. Therefore, nonparametric instrumental variable methods provide an alternative to parametric methods with a functional form specification advantage.

6 Application

The application considered here is based on the work of Mulvale and Hurley (2008) who looked at the effect of supplemental insurance on the use of psychologists using data from the 2002 Canadian Community Health Survey. The authors used a two part model: first, looking at the effect of supplemental insurance on the probability of using a psychologist and second, conditional on having seen a psychologist, what was the effect of supplemental insurance on the number of visits to a psychologist. In Canada, visits to a psychologist are not covered by the national health insurance plan and are paid out-of-pocket. In addition, visits to a psychologist can be expensive; in Ontario, the cost to see a psychologist is $225 per hour. This creates an incentive to hold supplemental health insurance to help cover some or all of the costs of accessing non-physician services. This raises the concern that insurance status may be endogenous as those who spend more time with psychologist are more likely to have supplemental insurance to help cover the costs. In order to control for the endogeneity of insurance status in predicting the use of psychologists, the authors estimated the model by instrumental variable regression using the marginal tax rate as the instrument. Using the marginal tax rate as an instrument for supplemental health insurance was first introduced by Stabile (2001) who looked at the demand for supplemental health insurance and the use of physician and hospital services.

The data used for this application is from the 2012 Canadian Community Health Survey - Mental Health component (CCHS-MH). The purpose of this survey was to investigate mental health in

Canada including the prevalence of mental illness, utilization of mental health care services, life experiences with mental illness. The population of interest was Canadians over the age of 15 years excluding those on native reserve, military personnel, and those in institutions. To collect the sample, the CCHS-MH used a multi-stages sampling design by first dividing the country into strata based on provinces and then selecting a random sample of clusters within each strata. From the chosen clusters a random sample of households were selected where only one member of each household was asked to respond. The weights in this survey not only represent the sampling scheme but were also calibrated for nonresponse.

The survey contains both information on supplemental insurance status and use of psychologist for mental health reasons. While this survey asked whether or not a person visited a psychologist, the question on the intensity of use was "how many hours did you spend with a psychologist?" This differs from the 2002 version of the questionnaire used in Hurley and Mulvale (2008), which asked about the number of visits with a psychologist. The CCHS-MH also contained information on income, therefore, I can compute the marginal tax rates. In this application, I will estimate the effect of supplemental insurance on the number of hours spent with a psychologist, conditional on having visited a psychologist. As I am only considering those who have visited a psychologist, the sample size is 541. I will also include two exogenous predictor variables, age and social provision score. The social provision score is a measure of one's social support system ranging from 10 to 40, where a value of 10 means they have little to no social support and a value of 40 means the person has a strong social support system. I will estimate the model using two estimators: the probability-weighted IV estimator (WNPIV) and the nonparametric IV estimator (NPIV). The marginal tax rates are computed from *Tax Facts and Figures 2011* from Pricewaterhouse Coopers Canada.

To start the analysis, I conduct a Durbin-Wu-Hausman test to see if insurance status is endogenous. This test has two stages. In the first stage, I run a probit with insurance status as the outcome variable and the marginal tax rate, age, and social provision score as the predictor variables. Next I regress hours spent with a psychologist against insurance status, age, social provision score, and the residual from the first stage. The null hypothesis is that the coefficient on the

residual from the fist stage is zero and insurance status is exogenous. If I reject the null, it means that insurance status is endogenous. With a p-value = 0.0075 on the error term, I reject the null at the 0.05 significance level and conclude that insurance status is endogenous.

To assess the strength of the marginal tax rate as an instrument for insurance status, I run a nonparametric conditional density estimation with insurance status as the outcome variable and the marginal tax rate, age, and social provision score as the predictor variables. Letting X and Y denote a vector of predictor variables and an outcome variable, respectively, the conditional density of Y given X is:

$$f(Y|X) = \frac{f(Y,X)}{f(X)},$$

where $f(Y,X)$ is the joint density of X and Y and $f(X)$ is the marginal density of X. The nonparametric conditional density estimator replaces $f(Y,X)$ and $f(X)$ by their respective kernel density estimators. Figure 1 presents the probability that someone has supplemental health insurance versus the marginal tax rate, keeping age and social provision score constant at their medians. There is a clear positive relationship between the marginal tax rate and the probability of holding insurance.

Table 7 presents the results from estimating the effect of insurance status on hours spent with a psychologist. The important measure here is the relative difference in time spent with a psychologist between those with insurance and those without. Both models report that people with insurance spent more time with a psychologist than those without. However, the weighted estimator report a higher percent difference that the unweighted estimators. The WNPIV estimator reports that people with insurance spend 65.8 percent more time with psychologists that those without, while NPIV says they spend 44.3 percent more time.

7 Conclusion

This paper provides a overview of the underlying theory of nonparametric instrumental variables and presents a guide to implementing these methods when using data collected via complex sampling plans. Nonparametric instrumental variable estimation reduces the likelihood of specification errors and provides an alternative to restrictive parametric methods. Simulations showed that using a

probability weighted estimator to estimate conditional mean functions did not decrease efficiency under SRS or exogenous sampling. However, efficiency gains can be made by using a weighted estimator when sampling is endogenous.

References

Baum, C., Schaffer, M. & Stillman, S. (2003), 'Instrumental variables and gmm: Estimation and testing', *Stata Journal* **3**(1), 1–31(31).
URL: *http://www.stata-journal.com/article.html?article=st0030*

Carrasco, M., Florens, J.-P. & Renault, E. (2007), *Linear Inverse Problems in Structural Econometrics Estimation based on spectral decomposition and regularization*, Vol. 6, Elsevier Science B.V., Amsterdam.

Centorrino, S., Feve, F. & Florens, J.-P. (2014), Implementations, simulations, and bootstrap in nonparametric instrumental variable estimation of additive regression models, Working paper, Stony Brook University.

Clair, L. (2019), 'Local constant regression with mixed data types in complex survey data', *Advances in Econometrics* **39**.

Darolles, S., Fan, Y., Florens, J. & Renault, E. (2011), 'Nonparametric instrumental regression', *Econometrica* **79**(5), 1541–1565.

Gorodnik, A. (2016), Lecture 5. Accessed July 1, 2016.
URL: *https://people.maths.bris.ac.uk/mazag/fa/lecture5.pdf*

Greene, W. H. (2012), *Econometric Analysis: Seventh Edition*, Prentice Hall, Saddle River, NJ.

Hardamard, J. (1923), *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven.

Horowitz, J. (2013), Ill-posed inverse problems in economics, Working Paper CWP37/13, The Institute for Fiscal Studies Department of Economics, UCL.

Horowtiz, J. (2011), 'Applied nonparametric instrumental variable estimation', *Econometrica* **79**(2), 347–394.

Kress, R. (1999), *Linear Intergral Equations Third Edition*, Springer, New York, NY.

Li, Q. & Racine, J. S. (2007), *Nonparametric Econometrics*, Princeton University Press, Princeton, NJ.

Magee, L., Robb, A. & Burbidge, J. (1998), 'On the use of sampling weights when estimating regression models with survey data', *Journal of Econometrics* **84**, 251–271.

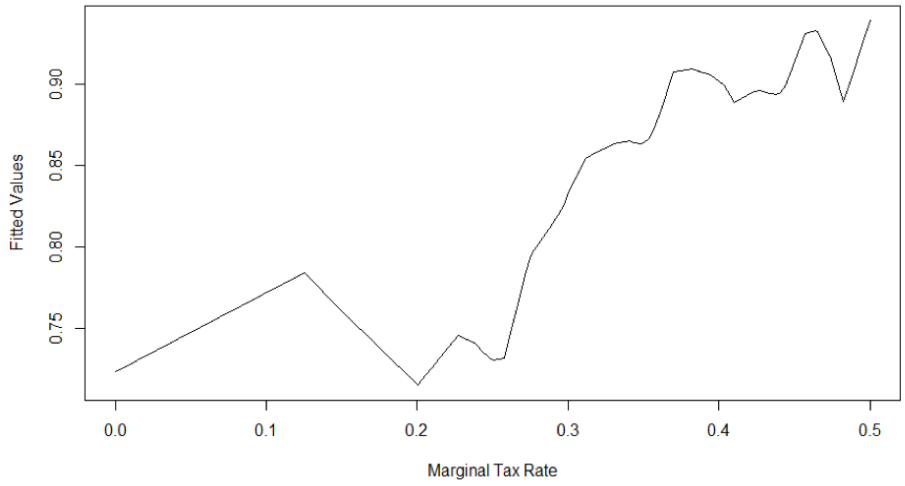
Mulvale, G. & Hurley, J. (2008), 'Insurance coverage and the treatment of mental illness: Effect on medication and provider use', *Journal of Mental Health Policy and Economics* **11**(4), 177–199.

Pricewaterhouse Coopers (2011), 'Tax facts and figures'.
URL: *https://www.pwc.com/ca/taxfacts*

Sánchez-Borrego, I., Opsomer, J., Rueda, M. & Arcos, A. (2014), 'Nonparametric estimation with mixed data types in survey sampling', *Rev Mat Complut* **27**, 685–700.

Stabile, M. (2001), 'Private insurance subsidies and public health care markets: evidence from canada', *Canadian Journal of Economics* **34**(4), 921–942.

Figure 1: Estimated Probability that Someone holds Supplemental Insurance Conditional on their Marginal Tax Rate



A Figures

B Tables

Table 1: Strata Borders for Endogenous and Exogenous Sampling Schemes

Sample Criterion	Strata Border	Inclusion Probability	Sample Weight
Y	$Y \leq$ %50 quantile	0.0002	5000
	%50 quantile $< Y \leq$ %85 quantile	0.00029	3448.3
	$Y >$ %85 quantile	0.0004	2500
X	$X \leq$ %40 quantile	0.00025	4000
	%40 quantile $< X \leq$ %75 quantile	0.00029	3448.3
	$X >$ %75 quantile	0.0007	1428.6

Table 2: Median, Mean, MAD, and 95 Percent Confidence Interval of the Mean Squared Error of Weighted and Unweighted Nonparametric IV Estimators under Simple Random Sampling

n	σ	MSE[$WNPIV$]	MSE[$NPIV$]	MSE[WLC]	MSE[LC]
300	0.05	0.0029	0.0029	0.0111	0.0112
		(0.0014)	(0.0012)	(0.0010)	(0.0011)
300	0.10	0.0034	0.0034	0.0111	0.0109
		(0.0014)	(0.0015)	(0.0013)	(0.0013)
300	0.25	0.0053	0.0053	0.0108	0.0109
		(0.0015)	(0.0014)	(0.0022)	(0.0023)
600	0.05	0.0017	0.0017	0.0113	0.0113
		(0.0008)	(0.0008)	(0.0007)	(0.0007)
600	0.10	0.0020	0.0021	0.0111	0.0111
		(0.0010)	(0.0009)	(0.0009)	(0.0009)
600	0.25	0.0039	0.0039	0.0109	0.0108
		(0.0016)	(0.0017)	(0.0016)	(0.0015)
900	0.05	0.0013	0.0013	0.0113	0.0113
		(0.0006)	(0.0006)	(0.0006)	(0.0006)
900	0.10	0.0014	0.0015	0.0111	0.0111
		(0.0006)	(0.0007)	(0.0007)	(0.0007)
900	0.25	0.0029	0.0029	0.0111	0.0109
		(0.0013)	(0.0015)	(0.0013)	(0.0013)

Values in round brackets are the median absolute deviations.

Table 3: 95 Percent Confidence Interval of the Mean Squared Error of WNPIV, NPIV, WNP, and NP Estimators under Simple Random Sampling

		MSE[$WNPIV$]		MSE[$NPIV$]		MSE[WLC]		MSE[LC]	
n	σ	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
300	0.05	0.0031	0.0033	0.0030	0.0033	0.0111	0.0113	0.0112	0.0114
300	0.10	0.0034	0.0037	0.0034	0.0037	0.0111	0.0114	0.0109	0.0112
300	0.25	0.0049	0.0051	0.0050	0.0052	0.0110	0.0115	0.0110	0.0115
600	0.05	0.0019	0.0021	0.0020	0.0022	0.0112	0.0114	0.0113	0.0114
600	0.10	0.0023	0.0025	0.0023	0.0025	0.0111	0.0113	0.0111	0.0113
600	0.25	0.0038	0.0041	0.0038	0.0040	0.0109	0.0114	0.0109	0.0112
900	0.05	0.0014	0.0015	0.0014	0.0016	0.0113	0.0114	0.0112	0.0114
900	0.10	0.0016	0.0018	0.0017	0.0018	0.0111	0.0112	0.0111	0.0113
900	0.25	0.0031	0.0033	0.0031	0.0033	0.0110	0.0112	0.0110	0.0113

Table 4: Median, Mean, MAD, and 95 Percent Confidence Interval of the Mean Squared Error of Weighted and Unweighted Nonparametric IV Estimators under Endogenous Sampling

n	σ	MSE[$WNPIV$]	MSE[$NPIV$]	MSE[WLC]	MSE[LC]
300	0.05	0.0028	0.0066	0.0147	0.0168
		(0.0014)	(0.0013)	(0.0011)	(0.0010)
300	0.10	0.0032	0.0078	0.0146	0.0175
		(0.0014)	(0.0016)	(0.0013)	(0.0014)
300	0.25	0.0052	0.0183	0.0144	0.0251
		(0.0023)	(0.0023)	(0.0026)	(0.0029)
600	0.05	0.0018	0.0054	0.0150	0.0168
		(0.0008)	(0.0007)	(0.0008)	(0.0007)
600	0.10	0.0018	0.0065	0.0146	0.0178
		(0.0008)	(0.0010)	(0.0010)	(0.0010)
600	0.25	0.0034	0.0164	0.0140	0.0253
		(0.0016)	(0.0021)	(0.0020)	(0.0020)
900	0.05	0.0013	0.0049	0.0150	0.0169
		(0.0006)	(0.0006)	(0.0007)	(0.0006)
900	0.10	0.0014	0.0060	0.0147	0.0178
		(0.0006)	(0.0007)	(0.0008)	(0.0008)
900	0.25	0.0024	0.0154	0.0141	0.0252
		(0.0011)	(0.0016)	(0.0016)	(0.0016)

Values in round brackets are the median absolute deviations.

Table 5: 95 Percent Confidence Interval of the Mean Squared Error of WNPIV, NPIV, WNP, and NP Estimators under Endogenous Sampling

		MSE[$WNPIV$]		$MSE[NPIV]$		MSE[WLC]		MSE[LC]	
n	σ	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
300	0.05	0.0033	0.0036	0.0067	0.0070	0.0147	0.0150	0.0167	0.0169
300	0.10	0.0037	0.0040	0.0080	0.0083	0.0145	0.0148	0.0175	0.0178
300	0.25	0.0053	0.0057	0.0179	0.0184	0.0144	0.0150	0.0249	0.0255
600	0.05	0.0021	0.0023	0.0055	0.0057	0.0149	0.0150	0.0168	0.0169
600	0.10	0.0022	0.0024	0.0066	0.0068	0.0145	0.0148	0.0177	0.0179
600	0.25	0.0038	0.0042	0.0165	0.0169	0.0141	0.0145	0.0252	0.0257
900	0.05	0.0015	0.0016	0.0050	0.0051	0.0150	0.0151	0.0168	0.0170
900	0.10	0.0017	0.0018	0.0061	0.0062	0.0147	0.0148	0.0178	0.0179
900	0.25	0.0028	0.0030	0.0156	0.0159	0.0141	0.0144	0.0253	0.0256

Table 6: Median, Mean, MAD, and 95 Percent Confidence Interval of the Mean Squared Error of Weighted and Unweighted Nonparametric IV Estimators under Exogenous Sampling

n	σ	MSE[$WNPIV$]	MSE[$NPIV$]	MSE[WLC]	MSE[LC]
300	0.05	0.0888	0.0889	0.0962	0.0929
		(0.0044)	(0.0039)	(0.0026)	(0.0026)
300	0.10	0.0889	0.0891	0.0959	0.0932
		(0.0040)	(0.0041)	(0.0029)	(0.0025)
300	0.25	0.0898	0.0898	0.0969	0.0933
		(0.0037)	(0.0037)	(0.0032)	(0.0029)
600	0.05	0.0887	0.0883	0.0975	0.0945
		(0.0030)	(0.0032)	(0.0018)	(0.0017)
600	0.10	0.0885	0.0887	0.0972	0.0946
		(0.0034)	(0.0031)	(0.0019)	(0.0017)
600	0.25	0.0898	0.0896	0.0980	0.0946
		(0.0029)	(0.0028)	(0.0023)	(0.0021)
900	0.05	0.0885	0.0884	0.0979	0.0950
		(0.0027)	(0.0028)	(0.0016)	(0.0013)
900	0.10	0.0881	0.0886	0.0979	0.0951
		(0.0029)	(0.0025)	(0.0016)	(0.0014)
900	0.25	0.0892	0.0884	0.0985	0.0952
		(0.0027)	(0.0026)	(0.0019)	(0.0017)

Values in round brackets are the median absolute deviations.

Table 7: Estimated Hours Spent with a Psychologist Versus Insurance Status from Weighted and Unweighted Nonparametric IV Estimation

Model	Insurance	No Insurance	Percent Change
NPIV	11.90	8.25	44.30
WNPIV	15.60	9.41	65.81