UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

ECONOMICS

On nonparametric additive error models with discrete regressors

by

Katarzyna Maria Bech

A thesis submitted for the degree of Doctor of Philosophy

September, 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Economics

Doctor of Philosophy

ON NONPARAMETRIC ADDITIVE ERROR MODELS WITH DISCRETE REGRESSORS

by Katarzyna Maria Bech

This thesis contributes to the literature on nonparametric additive error models with discrete explanatory variables. Although nonparametric methods have become very popular in recent decades, research on the impact of the discreteness of regressors is sparse. Main interest is in an unknown nonparametric conditional mean function in the presence of endogenous explanatory variables. Under endogeneity, the identifying power of the model depends on the number of support points of the discrete instrument relative to that of the regressor. Under nonparametric identification failure, we show that some linear functionals of the conditional mean function are point-identified, while some are completely unconstrained. A test for point identification is suggested.

Observing that the simple nonparametric model can be interpreted as a linear regression, new approaches to testing for exogeneity of the regressor(s) are proposed. For the point-identified case, the test is an adapted version of the familiar Durbin-Wu-Hausman approach. This extends the work of Blundell and Horowitz (2007) to the case of discrete regressors and instruments. For the partially identified case, the Durbin-Wu-Hausman approach is not available, and the test statistic is derived from a constrained minimization problem. In this second case, the asymptotic null distribution is non-standard, and a simple device is suggested to compute the critical values in practical applications. Both tests are shown to be consistent, and a simulation study reveals that both have satisfactory finite-sample properties. The practicability of the suggested testing procedures is illustrated in applications to the modelling of returns to education.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Katarzyna Maria Bech, declare that this thesis titled, 'On nonparametric additive error models with discrete regressors' and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- Parts of this work have been published as:

  Bech, K., Hillier, G. (2015) "Nonparametric testing for exogeneity with discrete regressors and instruments" CeMMAP working paper CWP11/15

Signed: ________________________________

Date:__________________________________

Acknowledgements

The completion of this thesis would not have been possible without help, advice, guidance and support of many people.

First of all, I would like to thank my supervisor Professor Grant Hillier for his continuously great supervision and patience. His experience, enthusiasm and flexibility have inspired, encouraged and thought me in various ways. I will be forever grateful for his teaching, assistance and unlimited support throughout my studies in Southampton and will always carry a passion for econometrics, injected by him.

I am also grateful to my second supervisor Professor Anastasios Magdalinos for giving me useful comments and providing support on mathematical technicalities. Additionally, I would like to express my deepest appreciation and gratitude to Professor Peter Phillips for extensive suggestions and discussions. I also thank Raymond O'Brien for constructive ideas and his involvement in simulations.

I would like to thank the research and administrative staff at the University of Southampton for providing an inspiring atmosphere and the ESRC for their financial support.

I owe much to my friends: Dafni, Maria, Panos, Jason, Orphe, Andres, Nicholas, Carlos, Adriana, Jana and Elif, without whom days in Southampton would not have been so colorful.

Finally, to my family: my mum Wieslawa, for teaching me maths since I was 3 years old and my dad Bohdan, for teaching me to believe that I can achieve anything I dream about. Also I want to thank Piotr for his patience, love and self-sacrifice. Their support was invaluable. Thank you.

1 Introduction

Nonparametric and semiparametric methods have attracted a great deal of attention from econometricians and statisticians in the past few decades. Initially, researchers were interested in describing the relationships between two or more series, but with the recent expansions in economics, they are involved in measuring conditional probabilities of decisions, duration of events, etc. The main concern is the need to make precise assumptions about the nature of these relationships. Frequently, the assumptions imposed on the model are implausible, especially those that restrict the functional form of the relationship. It is argued that the nature of most econometric models is nonparametric and parametrization can be viewed as an approximation of the relationship, which is required for estimation. Economic theory rarely provides information on the shape of the relationship between a dependent variable and regressors. It is important to realize that functional form choices have an extensive impact on parameter estimates and inference. Hence, the need of developing new, more flexible nonparametric methods emerged.

The nonparametric methods seem attractive, since they have desirable efficiency properties that hold under relatively mild assumptions on the population of interest and data generating processes. The main advantage, over parametric specifications, is that nonparametric techniques do not rely on any functional form or underlying distributional restrictions. Hence, they can be widely applicable in cases when there is limited information about the studied sample and making a priori assumption might give inaccurate results, as restrictive parametric restrictions often fail to be valid. Additionally, the nonparametric techniques are typically easy to implement and, especially for non-statisticians, to understand and interpret.

Throughout this thesis, the main focus is on a simple additive error nonparametric model of the form

$$Y = h(X) + \varepsilon, \tag{1}$$

where Y is a dependent variable, X is a vector of regressors and ε is the disturbance term. In parametric econometrics the estimation of the unknown $h(\cdot)$ is carried out by assuming some functional form of h, typically a linear function. However, it is well known that any misspecification in the functional form leads to inconsistent estimates and affects the size and power of tests based on them. In view of this, it is desirable to consider estimation and testing without assuming any functional form. Several methods of estimating nonparametrically the unknown structure of interest $h(\cdot)$ have been proposed. The most popular technique is the nonparametric kernel estimation firstly proposed by Nadaraya (1964) and Watson (1964). Other methods involve smoothing splines techniques (Reinsch(1967)) and signal extraction (Rao 1986)). The basic nonparametric methods are summarized in Section 1.1.

Since nonparametric methods are very general, the estimation is frequently computationally intensive. In many applications, the problems do not have a closed form solution and need to be solved numerically. Therefore, from the empirical point of view, it is important to be able to simplify the complicated expressions, which define the nonparametric estimators, to make them convenient to apply. One of the contributions of this thesis is to provide a neat and compact way of presenting the nonparametric specification in terms of a linear model and to express the standard nonparametric estimators in the form of familiar linear regression estimators.

Along with estimation, the separate path taken by the literature on nonparametric models is concerned with obtaining statistical inference. Hypothesis testing and construction of confidence intervals have been discussed by many authors. Particularly interesting are tests regarding the correct model specification. In the standard linear regression analysis, to ensure consistent ordinary least squares estimation, one of the assumptions imposed on the model is that of exogeneity of explanatory variables, i.e. a lack of correlation between the disturbance term and regressors. If the exogeneity assumption is violated, other estimation techniques must be employed. In the nonparametric framework, the notion of exogeneity, although defined in a slightly different way, still plays a crucial role in ensuring the consistency of the standard estimators. Therefore, testing whether the regressors are exogenous or endogenous is necessary for choosing an appropriate nonparametric estimation method. This thesis provides such nonparametric exogeneity testing procedures.

Throughout this thesis, the main interest lies in nonparametric models defined by equation (1), with X being a set of discrete regressors. Even though the literature on nonparametric models has grown rapidly in recent decades and different estimation techniques and methods for obtaining statistical inference are developed in frameworks with continuous regressors, not enough attention is paid to the models with discrete explanatory variables.

In applied works, the model with discrete regressors can be applied in many economic problems. Variables such as gender, race, levels of education or household size typically take a discrete number of values. When X is binary it may indicate the occurrence of the event. In empirical applications, such regressors are called 'dummy variables' taking values 0 or 1, for instance, an individual is either male or female, working or unemployed. The discrete regressor might also be integer valued (for instance, indicating the number of children in a household or years of education an individual has completed) or ordered (for example, giving the position on an attitudinal scale). The nonparametric model with discrete regressors has been applied by Hu and Lewbel (2008) to identify and estimate the difference in average wages between individuals who falsely claim college experience and those who tell the truth about not completing college education; and by Delgado (2011) to examine the impact of voluntary pollution prevention programs on the level of pollution emissions. More recently, Iori, Kapar and Olmo (2014) use nonparametric methods to explain variation in the continuous variable (bank funding spreads) given the set of discrete regressors (bank characteristics, nationality, size and operating currency) in the European interbank money market.

1.1 Models with continuous regressors

This section briefly summarizes the existing literature on nonparametric additive error models with continuous regressors and provides a contrast to a discrete case, which is the main focus of this thesis.

Consider the nonparametric model (1), where Y, X and ε are continuously distributed random variables. Under the standard mean independence condition $E[\varepsilon|X] = 0$, the unknown function $h(\cdot)$ is determined by the conditional distribution of Y given X.

The first approach to a nonparametric estimation of the structural function $h(\cdot)$ was independently proposed by Nadaraya (1964) and Watson (1964), who suggest estimating the conditional mean of Y given X, using the sample observations y_i and x_i^s, by the kernel estimator:

$$\widehat{h}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i^s - x}{b}\right) y_i}{\sum_{i=1}^{n} K(\frac{x_i^s - x}{b})},$$

where the kernel function K is usually a univariate density function, assumed to be symmetric, and b (the bandwidth) is a function of the sample size n and goes to zero as $n \rightarrow \infty$. The proposed estimator is then a weighted average of the observations y_i with weights depending on the distance between x_i^s and x. An implicit assumption in nonparametric estimation is that y_i contains information about $h(x)$, if x_i^s is in the neighborhood of x. Alternatively, one might allow the window width b to vary across data points, and construct the recursive kernel estimator, as in Devroye and Wagner (1980) and Greblicki and Pawlak (1987).

The main difficulty of any kernel-based approach is the efficient choice of the smoothing parameter b. Although the shape of the nonparametric estimator is not very sensitive to the actual choice of the kernel function, it crucially depends on the bandwidth b. The popular bandwidth selection strategies are based on minimizing the integrated squared error or its expected value. However, these methods are computationally complex as they involve the estimation of unknown density derivatives.

Another problem in nonparametric approach to estimation is the "curse of dimensionality", that is the need for a large number of observations in the sample in order to obtain accurate estimators in high-dimensional spaces. As a result, the rates of convergence of standard nonparametric estimators are slower. For instance, the bias of the Nadaraya-Watson estimator is of order $O(b^2)$ and $Var(\widehat{h}(x)) = O(n^{-1}b^{-1})$.

Both problems highlighted above do not occur in models with discrete regressors. Firstly, the nonparametric estimation of the density function (probability mass function in the discrete case) does not require kernel smoothing as the probability masses are easily estimated from the data as sample proportions. Secondly, the problem is not infinite dimensional as there is a finite number of conditional means $E\left[Y|X = x_k\right]$, which need to be estimated. Therefore, the estimation of nonparametric additive error models with discrete regressors seems to be more straightforward.

1.2 Endogeneity in nonparametric models

The main focus of this thesis is the impact of the presence of endogenous regressors in the nonparametric model. The problem of endogeneity arises frequently in economics and occurs when the independent variable is correlated with the model error term. Typically, it is a result of omitting a relevant explanatory variable in a regression specification, simultaneity in the model or a measurement error in the regressor. When the regressors used in the model are exogenous, the standard nonparametric estimator of the conditional mean of the outcome given the explanatory variables is consistent. However, in many important economic applications, the regressors used in the analysis are endogenous and the consistent estimation of $h(\cdot)$ in (1) requires the implementation of different techniques.

The most popular method of dealing with endogeneity in econometric models is the instrumental variable (IV) estimation. Although IV methods are traditionally parametric in nature, the extension of the approach to more flexible nonparametric framework originated with Newey and Powell (2003). Further studies were also conducted by Hall and Horowitz (2005) and Darolles, Florens and Renault (2011), who propose a kernel estimator and derive optimal convergence rates. The main idea is that researchers should find a set of variables satisfying instrument relevance and exogeneity conditions and use them to consistently estimate the causal relationship between the dependent variable and endogenous regressors. However, the IV method involves some identification issues.

The nonparametric instrumental variable approach of Newey and Powell (2003) shows that the nonparametric point identification of h requires that the conditional distribution of X given Z is complete. The completeness condition can be seen as an intuitive generalization (or nonparametric analogue) of the rank condition for identification in linear specifications. However, not many researchers concentrate on providing evidence for or against these assumptions in datasets. In linear models under endogeneity, it is possible to test whether the rank condition for identification is satisfied or not. Yet, as shown by Canay, Santos and Shaikh (2013) the equivalent completeness condition in the nonparametric IV framework is untestable. Specifically, they show that when testing

$$\begin{aligned} H_0 &: \text{ completeness condition does not hold (lack of point identification) } vs. \\ H_1 &: \text{ completeness condition holds (model is point identified)} \end{aligned}$$

it is not possible to provide empirical evidence in favour of H_1. Such a test is equivalent to existing tests of rank conditions in linear specifications with the null hypothesis of rank condition failure. Hence, the arising conclusion is that even though it is possible to test for identification in linear models, the analogous nonparametric completeness condition is nontestable. As pointed out by the authors, since the data cannot provide supporting evidence for point identification, empirical researchers should present alternative arguments to justify their analysis being performed under point identification assumption. Nonetheless, the more rational way to proceed is to abandon completeness condition and employ statistical methods that allow for partial identification in this framework. These methods are briefly discussed in Section 1.3.

The problem with identification is particularly noticeable in nonparametric models with additive errors when the regressors are discrete. A typical example of an endogenous discrete regressor appears in treatment effect models, where the endogeneity of a treatment variable comes from self-selection. Other examples include a nonparametric version of classical supply and demand model with prices and discrete quantities, which are jointly determined by the model and endogeneity arising from measurement error, if the discrete regressor is misclassified. The identification and estimation of the nonparametric models with discrete endogenous regressors is discussed in Das (2005) and Florens and Malavolti (2003). They show that the necessary and sufficient condition for point identification is that the number of points of support of the discrete instrument is at least as large as the number of points of support of endogenous regressor. It implies that the problem of not being able to test whether the identification condition is satisfied is avoided, as given the instrument, the identification is straightforward to judge. If this nonparametric identification condition fails the unknown conditional mean function (CMF) is only partially identified. In the absence of additional information this means that certain functionals of the CMF can take arbitrary values, whilst others are point-identified. One contribution of this thesis is to provide a test for point-identifiability in this context.

Returning to the nonparametric estimation with endogenous regressors, Newey and Powell's (2003) nonparametric two-stage least squares estimator is based on solving the integral equation

$$\widehat{E}\left[Y|Z\right] = \int h(X)\widehat{F}(dX|Z). \tag{2}$$

As (2) is a Fredholm integral equation of the first kind, it creates a so called ill-posed inverse problem, as the mapping from the structural ($F_{X|Z}$) to the reduced form ($E[Y|Z]$) is not continuous. This means that $h(\cdot)$ cannot be estimated consistently by replacing the unknown population quantities with consistent estimators. In order to obtain a consistent estimator, it is necessary to regularize the mapping that identifies the unknown function of interest. Newey and Powell (2003) and Santos (2010) propose to assume compactness through smoothness restrictions on h. They restrict the true function to be an element of a compact set of functions, which makes the mapping from reduced form to structure continuous. One of the advantages of nonparametric models with discrete endogenous regressors is that they do not suffer from the ill-posed inverse problem. Restricting the endogenous regressors to be discrete eliminates the ill-posed inverse problem. The discrete specification is well-posed, and no regularization of the problem is required.

Even in simple linear models, the presence of endogenous regressors typically leads to the OLS coefficients being biased and, in many cases, inconsistent. Because of the severe consequences of endogeneity, applied researchers need to check whether the explanatory variables used are exogenous, before providing an inference of the parameters of interest. Following the work of Hausman (1978) the research on testing for exogeneity of the regressors has been growing rapidly.

Recently, with the expansion of nonparametric models, new testing procedures had to be developed. The problem of testing the correct specification of a nonparametric model given by equation (1) has been discussed by many authors including the work of Fan and Li (1996), Zheng (1996), Lavergne and Vuong (2000) and Lavergne and Patilea (2008). The tests fit in a conditional moment restriction testing framework and are based on an earlier work of Newey (1985) and Bierens (1990) among others. All the nonparametric tests of this type assume that the regressors are continuously distributed. The aim of this thesis is to provide a test for exogeneity in a nonparametric model with discrete explanatory variables.

Our benchmark study is that of Blundell and Horowitz (2007). They propose a consistent nonparametric test for exogeneity in model (1) with continuous regressors. The test is based on the comparison of the unknown function $h(\cdot)$ with the conditional mean function of Y given X. We will follow similar methodology in constructing the exogeneity test for discrete regressors.

1.3 Review of partial identification literature

Throughout the thesis, we broadly talk about partial identification. For the reader not familiar with that concept, we provide a brief summary. The following literature review comes from Bech (2011).

The literature on estimation and inference in partially identified models has been growing rapidly in the last two decades. It is clear that in applied econometrics data alone is not sufficient to deduct meaningful conclusions about the population of interest. Inference always requires making assumptions on the population behaviour via a hypothesis about the data generating mechanism. Until the late 1980's, parameters were only considered to be either point identified or not identified at all. Point identification was typically achieved by using assumptions, which were strong enough to identify the exact value of the parameters. However, by imposing weaker and more credible restrictions, researchers are able to partially identify some features of the model.

An interest parameter is said to be partially identified by the model if it is not uniquely determined by the distribution of the observed data. Hence, the partial identification approach states that even if the model cannot point identify parameters, it frequently contains some relevant message, which enables researchers to bound parameters in informative ways. The class of partially identified models also contains models that are identified in some parts, but the parameter is unidentified in others. For instance, in the simultaneous equations models, some equations might be identified and others are not. The typical example is the very simple supply-demand model, where

$$\begin{aligned} Q &= \beta_1 P + \varepsilon_S && \text{(supply equation)} \\ Q &= \alpha_1 P + \alpha_2 X + \varepsilon_D && \text{(demand equation)} \\ \text{with } Cov(\varepsilon_S, \varepsilon_D) &= 0 \text{ and } X \text{ exogenous.} \end{aligned}$$

Clearly, the parameters of the demand equation are not identified, but the slope parameter in supply equation is[1]. The nonparametric additive error model with discrete endogenous regressors under identification failure also falls in this class of partially identified models, as some linear functionals of the parameters of interest are point identified while other are undetermined.

The term "identification" was firstly introduced by Tjalling Koopmans in 1949. He developed the concept of the identifiability as the property of "a parameter that can be determined from a sufficient number of observations." (Koompans (1949)). Surprisingly, until the late 1980s, the impact of the failure of point identification and the main findings of the articles on partial identification were neglected. Although estimation and inference under unjustifiably assumed point identification might lead to distortions in the asymptotic theory of estimators, the first rare works on partial identification had almost no impact on empirical studies before the 1990s. The first main research on partial identification was conducted by Frisch (1934), who studies the problems of estimation when all variables are measured with error. In this framework, Frisch (1934) derives the bounds on the slope parameter of a linear regression i.e. the identified set, which can be estimated from the data. The second important studies were performed by Marschak and Andrews (1944). They show that the parameters of the production function can be bounded to sections within the parameter space.

In the late 1980s the new literature on partial identification was developed to confront the traditional approaches to inference with missing data models. Phillips (1989) explores the changes in the properties of common statistical procedures under point identification failure. Phillips (1989) also provides a list of models in which partial identification analysis is required e.g. the classical simultaneous equations model under rank condition failure, time series spurious regressions and microeconometric models with endogenous regressors.

The main pioneer of recent studies of partial identification is Charles F. Manski, who started his work in 1989 with analyzing the problem of self-selection into treatment. Manski's (1989) problem is to find the correct prior restrictions which

[1] That is true in the iid, cross-sectional framework. However, if the data are time series and nonstationary, then there is scope for structural identification and consistent estimation. In such cases, it is possible that the nonstationary variable (here price) can serve as its own instrument. This type of facility also occurs in the cross section framework, where there are location shifts as these can drive nonstationary behaviour. The same might occur in treatment effec models when there is a threshold shift.

point identify treatment effects. Horowitz and Manski (2000) study an inference with the missing outcome data using nonparametric prediction. They show that even without any restrictions on the distribution of the missing outcomes, data alone yields informative nonparametric bounds on treatment response. Manski and Tamer (2002) examine inference in regressions with interval data and develop two new estimators: Modified Maximum Score Estimator and Modified Minimum Distance Estimator. Manski and Pepper (2000) use partial identification in the estimation of wage as a function of schooling without assuming statistical independence between outcomes and instruments. Horowitz and Manski (1995) show that in errors in variables models under the assumption of robust estimation, the population parameters are not identified, but can be frequently bounded. They apply the findings to the income distribution model and conclude that estimating bounds consistently is often accessible. Other influential partial identification studies include Tamer (2003) who detects that with the least possible set of assumptions, incomplete econometric structural models contain useful information about the parameters of interest. He also shows how to obtain reasonable conditions for identification in the presence of multiple equilibria. Blundell et al. (2007) study the impact of non-random selection into work. They show that in the presence of censoring, the wage distribution is not point identified without strong assumptions. However, even the worst-case bounds can be informative and there is a way to tighten bounds using restrictions dictated by economic theory.

Within the broad class of partially identified nonparametric instrumental variables models, Severini and Tripathi (2006) derive semi-parametric bounds for the estimation of linear functionals of h. Santos (2010) develops methods for hypothesis testing and construction of the confidence sets under partial identification assumption, based on the techniques applied by Newey and Powell (2003). Chesher (2005) provides conditions for nonparametric set identification in nonseparable models with discrete endogenous regressors, and discusses an interval estimation in this framework. Chesher (2010) derives set identifying results for nonseparable IV models with discrete outcomes.

The recent literature on obtaining inference in partially identified models using different techniques includes results by Imbens and Manski (2004), who derive confidence intervals for the mean of a bounded random variable that asymptotically contain the true value of parameter with fixed probability. Chernozhukov et al. (2007) were the first to examine inference in more general setups by using the population objective function and sub-sampling methods. Similar sub-sampling procedures are employed by Romano and Shaikh (2008) who additionally present conditions for a uniform coverage of the confidence regions. Santos (2010) introduces methods for hypothesis testing in a nonparametric Instrumental Variables model under point identification failure and constructs the asymptotic distribution of a test statistic of a hypothesis that some elements of the identified set satisfy a given condition. Rosen (2008) introduces confidence sets for a parameter of interest in models composed of moment inequalities. Beresteanu and Molinari (2008) suggest an alternative way of obtaining inference in partially identified models by applying instruments from the random set theory (Bech (2011)).

1.4 Notation

Throughout this thesis, the notational conventions are as follows: the upper case letters Y, X, Z, ε represent scalar random variables, and $Y_n, X_n, Z_n, \varepsilon_n$ are the $n \times 1$ vectors of sample equivalents, with n being the sample size. Realized sample observations are denoted by x_i^s, z_i^s, $i = 1, ..., n$, and the observed value of Y_n by y. Symbols x_k for $k = 1, ..., K$ and z_j for $j = 1, ..., J$ denote the points of support of discrete random variables X and Z. $I(B)$ stands for an indicator function, which takes value 1 if the event B occurs, and is 0 otherwise. The probability density function of a continuous random variable Y is denoted by $f_Y(y)$ and the probability mass function of a discrete random variable X is $p_X(x)$. The cumulative distribution function is denoted by $F_X(x)$. For a matrix A of full column rank we define

$$\begin{aligned} P_A &= A\left(A'A\right)^{-1}A' \\ M_A &= I - P_A, \end{aligned}$$

both of which depend only on the space spanned by the columns of A (i.e. are invariant under $A \to AB$, with B a non-singular matrix). For any r, l_r denotes an r-vector of ones and C_r denotes an $r \times (r-1)$ matrix with the properties $C_r' l_r = 0$ and $C_r' C_r = I_{r-1}$.

1.5 Structure of the thesis

The plan of this dissertation is as follows. Chapter 2 introduces the nonparametric model of interest and presents the notation that enables us to interpret the nonparametric specification as a linear model. Additionally, it explains the identification problems in the presence of endogenous regressors and gives some basic estimation results. An important part of this chapter, Section 2.2.1, shows how the point-identifiability of linear functionals of the unknown function of interest might be tested under nonparametric point identification failure.

Chapter 3 deals with nonparametric testing for exogeneity. Section 3.1 presents the test for models that point identify the entire unknown function and establishes its asymptotic properties under the null and alternative hypothesis. Section 3.2 proposes a test for exogeneity in models that are partially identified, and again gives the asymptotic properties of the test statistic under the null and alternative hypothesis. In this second case, we also discuss the computation of critical values, because the asymptotic null distribution is non-standard. In Section 3.3, we present the results of the Monte Carlo investigation of the finite sample properties of the proposed tests.

Chapter 4 extends the results to models with additional exogenous regressors and multiple instruments. An important point that arises from this discussion is that nonparametric identification does not depend on the number of instrumental variables, but only on the number of support points.

In Chapter 5, we present empirical applications that illustrate the practical use of the proposed tests. We confirm that education is endogenous when estimating the returns to schooling in a standard wage equation and check whether any linear function of the conditional average wage is point identified.

Chapter 6 concludes and discusses further work. All proofs are in the Appendix A. Appendix B contains additional results on empirical power properties from Monte Carlo simulations and Appendix C presents supplementary results from empirical applications.

2 Nonparametric additive error model with discrete regressors: identification and estimation

2.1 Model and assumptions

This section introduces the model of interest and provides a neat way of representing the nonparametric model in terms of a familiar linear structure. We also discuss the identification issues arising in the models with discrete regressors and provide necessary and sufficient conditions for point identification of the unknown function $h(\cdot)$.

2.1.1 Model

We consider to begin with a simple additive error model, in which an observable scalar continuous random variable Y is determined by equation (1), with X, a single discrete regressor, and ε denotes a continuously distributed error term. The interest of econometricians typically lies in estimating the unknown structural function $h(\cdot)$. Consistent nonparametric estimation of $h(\cdot)$ is feasible under the assumption that the regressors are exogenous. Numerous definitions of exogeneity have been provided in the literature, see Deaton (2010). The standard exogeneity condition is that of an absence of correlation between regressor and the model error term. Here we employ the definition proposed by Blundell and Horowitz (2007) for nonparametric regressions:

Definition 2.1 *The explanatory variable X is exogenous if the conditional moment restriction*

$$E[\varepsilon|X = x_k] = 0$$

holds for all $k = 1, ..., K$.

Given exogeneity, $E[Y|X = x_k] = h(x_k)$, i.e. the conditional mean of the dependent variable given $X = x_k$ coincides with the structural function $h(x_k)$. This definition has the advantage that standard nonparametric regression of Y on X is then appropriate for consistent estimation of the unknown function of interest $h(\cdot)$.

In the presence of endogeneity of regressors, $h(\cdot)$ is unidentified in the absence of additional information. The common strategy to deal with the endogeneity problem is to apply the instrumental variable estimation. However, the IV solution to endogeneity is only possible if the discrete instrument has more support points than the endogenous discrete regressor. In the sparse support case, the instrument fails to fully identify $h(\cdot)$ (see Section 2.1.4).

The complete model of interest is characterized by the following set of assumptions:

Assumption 1 X is a discrete (scalar) random variable with support $\{x_1, ..., x_K\}$ with associated probabilities $p_k > 0$.

Assumption 2 There exists a discrete instrumental variable Z with support $\{z_1, ..., z_J\}$ and associated probabilities $q_j > 0$, with the property that

$$E[\varepsilon|Z = z_j] = 0, \quad j = 1, ..., J \tag{3}$$

which defines the instrument exogeneity condition[2].

Assumption 3 The matrix of joint probabilities P with elements

$$p_{jk} = \Pr\left[Z = z_j, X = x_k\right]; \; j = 1, ..., J; \; k = 1, ..., K$$

is of full rank K when $J \geq K$, and of full rank J when $J < K$.

Assumption 4 $E[X|Z = z_j]$ and $E[h(X)|Z = z_j]$ vary with z_j.

Assumption 5 The data consists of n *iid* observations on (Y, X, Z). Under exogeneity, for all j and k,

$$E[\varepsilon|X = x_k, Z = z_j] = 0 \text{ and } Var[\varepsilon|X = x_k, Z = z_j] = \sigma^2.$$

Assumption 2 and 4 are analogous to the standard assumptions for the validity of instruments in the single equation IV estimation (see, for example, Greene (1993), Section 20.4.3). The first condition in Assumption 4 (the instrument relevance conditional) together with (3) ensures that Z is a valid instrument.

[2]Notice that we include in the support of X and Z only points for which p_k and q_j are strictly positive.

The rank condition in Assumption 3 is effectively a completeness condition in Newey and Powell (2003). Assumption 5 implies that the following unconditional moments of the error term:

$$\begin{aligned} E[\varepsilon] &= E_{X,Z}\left[E\left[\varepsilon|X=x_k, Z=z_j\right]\right]=0 \\ Var\left[\varepsilon\right] &= E_{X,Z}\left[Var[\varepsilon|X=x_k, Z=z_j]\right]+Var_{X,Z}\left[E[\varepsilon|X=x_k, Z=z_j]\right] \\ &= E_{X,Z}\left[\sigma^2\right]+Var_{X,Z}\left[0\right]=\sigma^2. \end{aligned}$$

The complete model consists of equations (1) and (3). The assumption of exogeneity of the regressor, i.e. $E[\varepsilon|X=x_k]=0$ for all k is equivalent to

$$E[Y|X=x_k]=h(x_k), \quad k=1,...,K.$$

If this condition is satisfied the unknown function $h(\cdot)$ can be consistently estimated nonparametrically (see Section 2.3.1) and equation (3) is not needed for consistent estimation. If $E[\varepsilon|X=x_k]\neq 0$ i.e. the regressors are endogenous, the use of instruments is necessary and equation (3) plays a crucial role.

Since (1) can be represented as the linear function

$$Y=\sum_{k=1}^{K} I(X=x_k)h(x_k)+\varepsilon,$$

the unknown function $h(\cdot)$ is constrained by the set of J linear equations

$$E[Y|Z=z_j]=\sum_{k=1}^{K}\Pr[X=x_k|Z=z_j]h(x_k), \quad j=1,...,J. \tag{4}$$

Let β denote the K-vector with $\beta_k=h(x_k)$, $k=1,...,K$, π be the J-vector with the elements $E[Y|Z=z_j]$, $j=1,...,J$ and Π be the $J\times K$ matrix of conditional probabilities $\Pr[X=x_k|Z=z_j]$, $k=1,...,K$, $j=1,...,J$. Then, (4) can be written compactly as

$$\pi=\Pi\beta. \tag{5}$$

In the continuous case, equation (5) corresponds to the integral equation for the structural function, e.g. equation (2.2) in Blundell and Horowitz (2007)[3]. The nonparametric nature of the model is reflected in the fact that β, the vector of

[3] Similar conditions are given in Chesher (2004) and Freyberger and Horowitz (2014).

values of $h(\cdot)$ at the support points of X is completely unknown. It is worth noting that equation (5) always has a solution (for unknown β), since for each $j = 1, ..., J$ by definition

$$E[Y|Z = z_j] = \sum_{k=1}^{K} \Pr[X = x_k|Z = z_j]E[Y|X = x_k, Z = z_j],$$

which implies that π is certainly in the space spanned by the columns of Π. Throughout this thesis we assume that there exists a valid instrument Z for which (5) holds.

Let $n_k^X = \sum_{i=1}^n I(x_i^s = x_k)$ and $n_j^Z = \sum_{i=1}^n I(z_i^s = z_j)$, the multiplicity of x_k and z_j in the sample. Also $n_{jk} = \sum_{i=1}^n I(x_i^s = x_k)I(z_i^s = z_j)$. The important properties of sample multiplicities are

$$\sum_{k=1}^{K} n_{jk} = n_j^Z \text{ and } \sum_{j=1}^{J} n_{jk} = n_k^X.$$

The elements of the vector π can be consistently estimated from the data, by averaging those y_i that correspond to the observations with $z_i^s = z_j$ (standard nonparametric estimator of the conditional mean):

$$\widehat{\pi}_j = \frac{\frac{1}{n}\sum_{i=1}^n y_i I(z_i^s = z_j)}{\frac{1}{n}\sum_{i=1}^n I(z_i^s = z_j)} = \frac{1}{n_j^Z}\sum_{i=1}^{n} y_i I(z_i^s = z_j). \tag{6}$$

The elements of the matrix of conditional probabilities Π can be written as

$$\Pr[X = x_k|Z = z_j] = \frac{\Pr[X = x_k \cap Z = z_j]}{\Pr[Z = z_j]}$$

and can be consistently nonparametrically estimated by

$$\widehat{\Pi}_{jk} = \frac{\frac{1}{n}\sum_{i=1}^n I(x_i^s = x_k)I(z_i^s = z_j)}{\frac{1}{n}\sum_{i=1}^n I(z_i^s = z_j)} = \frac{n_{jk}}{n_j^Z}. \tag{7}$$

Therefore, π and Π can easily be learned from the data, and the problem is to use this information to make inference on $h(\cdot)$.

Remark 2.1 *In the discussion here, and also in what follows, it is implicitly assumed that all K support points of X, and all J of Z, occur in the sample. That is, that both n_k^X and n_j^Z are non-zero for all $k = 1, .., K$ and $j = 1, .., J$. This will ultimately (for large enough n) be the case with probability one. The alternative would be to define estimates for the π_j and Π_{jk} only for those points x_k and z_j that occur in the sample, say $K_s \leq K$ and $J_s \leq J$ points, and allow these to increase to K and J respectively, as n increases. This would make the arguments and derivations to follow considerably more cumbersome, without materially affecting the results, so instead we will tacitly assume throughout that n is large enough to ensure that $K_s = K$ and $J_s = J$.*

There is no difficulty in extending the results by allowing for additional discrete exogenous regressors and multiple instruments in the model. The results for these generalized models are presented in Chapter 4.

2.1.2 Linear Model Interpretation

The above setup can be represented compactly in terms of a linear model. We define the $n \times K$ matrix L_X with (i, k) element

$$(L_X)_{ik} = I(x_i^s = x_k),$$

so that $(L_X)_{ik} = 1$ if observation i corresponds to a value x_k for X, and is 0 otherwise. Likewise, define the $n \times J$ matrix L_Z with elements

$$(L_Z)_{ij} = I(z_i^s = z_j).$$

Note that the rows of both L_X and L_Z add up to 1, since there can be only one entry in each row that is equal to 1, and other entries in that row have to be 0. Both L_X and L_Z are random matrices, because both the number and position of the non-zero elements are determined randomly in the sample.

Let x denote the K-vector with elements x_k, $k = 1, ..., K$, the support points of the regressor. The vector $L_X x$ represents the n-vector of sample observations x_i^s. Finally, let y denote the n-vector of sample observations on Y.

Using the notation just introduced (6) can be written as

$$\widehat{\pi} = \left(L_Z' L_Z\right)^{-1} L_Z' y \tag{8}$$

and (7) becomes

$$\widehat{\Pi} = \left(L'_Z L_Z\right)^{-1} L'_Z L_X. \tag{9}$$

The inverse in (8) and (9) exists almost surely for large enough sample size, since $\Pr[Z = z_j] = q_j > 0$ by assumption. Clearly, for existence we require $n > K$ and $n > J$.

Observe that

$$n^{-1} L'_Z L_Z = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_1) & & 0 \\ & \ldots & \\ 0 & & \frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_J) \end{pmatrix} \to^p diag(q_j) := D_Z,$$

because $\frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_j) \to^p E[I(z_i^s = z_j)] = \Pr[Z = z_j]$ for all j by the $WLLN$. Hence, by the Slutsky Theorem

$$\left(n^{-1} L'_Z L_Z\right)^{-1} \to^p D_Z^{-1}.$$

Similarly, $n^{-1} L'_Z L_X$ is a consistent estimator for the joint probability matrix P. Therefore, $\widehat{\pi} \to^p \pi$ and $\widehat{\Pi} \to^p \Pi := D_Z^{-1} P$.

In terms of observables, the assumption of the exogeneity of X takes the form

$$E[Y_n | X = L_X x] = L_X \beta,$$

which is analogous to a linear model for the vector y with random regressors matrix L_X and unknown parameters $\beta_k = h(x_k)$, $k = 1, ..., K$. Our model can be therefore expressed as

$$y = L_X \beta + \varepsilon_n. \tag{10}$$

It means that although the model is purely nonparametric, it can be interpreted as a linear regression. Note that even though in the nonparametric specification there is only one discrete regressor X, the regression matrix L_X is $n \times K$ in the linear regression specification. Also, observe that the support points x_k only determine the points at which we can learn $h(\cdot)$, but do not appear elsewhere in the linear model. This familiar linear specification allow us to connect the nonparametric estimators with the well known regression estimators, particularly OLS and 2SLS.

2.1.3 A complication

There is a relationship between L_X and L_Z, which has an important implication on the further analysis. The problem is that every sample point must be associated with exactly one support point of both X and Z. It follows that, for any regressor X and any instrument Z, the row sums of both L_X and L_Z are all equal to one. Let us, for brevity, call this

Property C: $L_X l_K = L_Z l_J = l_n$.

Algebraically, this says that the column spaces of L_X and L_Z always have the vector l_n in common, and this needs to be taken into account in adapting existing procedures to the present problem.

Note that Property C implies, in particular,

$$M_{L_X} L_Z l_J = M_{L_X} l_n = 0.$$

As a consequence of Property C, some matrices involving both L_X and L_Z have reduced rank. Hence, special attention has to be paid when dealing with these matrices.

2.1.4 Identification

Newey and Powell (2003) and Das (2005) study identification of the unknown structural function $h(\cdot)$ in the presence of endogeneity of a discrete regressor X. Florens and Malavolti (2003) and Das (2005) consider estimation in this framework. They show that nonparametric point identification is achieved if the vector of instruments Z has at least as many points of support as the endogenous regressor X under a marginal covariation condition:

$$E[\varepsilon|Z = z] = c, \tag{11}$$

where c is a constant that is invariant with respect to Z. This follows from (5), since assuming that equations in (5) represent the only information about $h(\cdot)$ that the data contains, the point identification requires that the matrix Π has a rank K.

Using this marginal covariation restriction, one can normalize $c = 0$ producing the system of linear equations (5). Since the conditional expectations on the left hand side and probabilities on the right hand side are observables, (5) forms a set of linear equations in the unknown $h(x_k)$. Hence, the value of the vector β is identified if these linear equations have a unique solution.

Proposition 1 *(Newey and Powell (2003)) The necessary and sufficient condition for identification in the model* $Y = h(X) + \varepsilon$*, with discrete endogenous* X *and a discrete instrument* Z *satisfying* $E[\varepsilon|Z] = 0$*, both with finite support, is that* $rank(\Pi) = K$*, for which it is necessary that the number of points of support of the instrument* Z *is at least as large as the number of points of support of endogenous* X*.*[4]

Hence, if $J \geq K$, β is point-identified for known (π, Π) and $\beta = (\Pi'\Pi)^{-1} \Pi'\pi$.

2.2 Point-identifiability of linear functionals

When $J < K$, so that the identification condition fails, the model still has partial identifying power. The following proposition elaborates on partial identification, and slightly extends Proposition 1 in Freyberger and Horowitz (2014).

Proposition 2 *Let* $L(\beta) = c'\beta$ *be a linear functional of the elements of* β*. When* $rank(\Pi) = J < K$*, the following are true:*

1. *for any* c *orthogonal to the null space of* Π*,* $L(\beta)$ *is point-identified; the dimension of this set is* J*.*

2. *for* c *not orthogonal to the null space of* Π*,* $L(\beta)$ *is completely unconstrained; the dimension of this set is* $K - J$*.*

Thus, there is a space of linear functionals of dimension J that are point identified by (5), and a space of dimension $K - J$ about which we can hope to learn nothing, without the addition of further information. This space is larger the larger the difference $K - J$. Freyberger and Horowitz (2014) discuss the identifying power of additional restrictions on β in this unidentified case, in particular, shape restrictions on $h(\cdot)$.

[4] The result can also be found in Matzkin (2007), Chapter 73 in "Handbook of Econometrics".

Recall that β is a solution to equation (5) and let S_π denote a set of all possible solutions, i.e.

$$S_\pi = \{\beta : \Pi\beta = \pi\}.$$

Let V be a full rank $K \times (K - J)$ matrix satisfying $\Pi V = 0$ and let C be a full rank $K \times J$ matrix, such that $C'V = 0$, i.e. the columns of C are orthogonal to the columns of V. Let β_0 be any fixed vector satisfying $\Pi\beta_0 = \pi$. The set S_π can be represented as

$$S_\pi = \{\beta_0 + V\gamma, \gamma \in \mathbb{R}^{K-J}\}.$$

The set of identified functions $g(\cdot) : \mathbb{R}^K \longmapsto \mathbb{R}$ consists of just those functions that, when restricted to S_π, are constant. The following Proposition extends the results of Proposition 2 to general functions of the elements of β.

Proposition 3 *When rank*$(\Pi) = J < K$*, in the absence of additional restrictions, the set of point identified functions of* β *consists of those functions that depend on* β *only through* $C'\beta$*.*

The result in Proposition 3 arises because the set S_π is invariant under the transformations $\beta \to \beta + V\gamma, \gamma \in \mathbb{R}^{K-J}$, and it can be shown that $C'\beta$ is a maximal invariant under this group.

2.2.1 A test for point-identifiability

The condition required for point-identifiability in Proposition 2 can be tested. That allows us to construct a simple test for point-identifiability of some linear functionals of dimension J of the elements of β under identification failure, i.e. when the entire unknown vector remains undetermined.

Writing $\Pi = (\Pi_1, \Pi_2)$, with Π_1 $J \times J$ and non-singular, we may write[5]

$$L(\beta) = c_1'\beta_1 + c_2'\beta_2 = c_1'\Pi_1^{-1}\pi + (c_2' - c_1'\Pi_1^{-1}\Pi_2)\beta_2,$$

and the condition for point identification of $L(\beta)$ is:

$$c_2' - c_1'\Pi_1^{-1}\Pi_2 = 0', \tag{12}$$

[5]By considering the reduction of Π to upper echelon form it is easy to see that none of what follows depends on which columns of Π form the nonsingular component Π_1, if there is more than one choice. We omit details.

which, for given c, is a restriction on Π.

We can test whether this linear functional of interest is point identified by checking whether the sample equivalent of $c_2' - c_1'\Pi_1^{-1}\Pi_2$ is close to 0. Note that $\Pi_1 = D_Z^{-1}P_1$ and $\Pi_2 = D_Z^{-1}P_2$ where P_1 is a $J \times J$ non-singular matrix and P_2 is $J \times (K - J)$, so that the identifiability hypothesis for given c can be expressed in terms of the matrix of joint probabilities P :

$$H_0^c : \left(c_2' - c_1'P_1^{-1}P_2\right) = 0'.$$

The partition of c into (c_1, c_2) is induced by the choice of P_1, and, when $c_2 \neq 0$, $c_1'P_1^{-1}P_2$ cannot be zero if the corresponding $L(\beta)$ is to be point identified. In particular, if $c_2 \neq 0$, a necessary condition for identifiability of $c'\beta$ is that $c_1 \neq 0$.

Since P can be consistently estimated by $\widehat{P} = n^{-1}L_Z'L_X$, the sample equivalent of (12) is $c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2$, where $\widehat{P} = (\widehat{P}_1, \widehat{P}_2)$ is a suitable partition of $\widehat{P}$. A natural statistic to measure the departure of this vector from zero, and therefore to test H_0^c, is the quadratic form

$$G_n = n\left(c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2\right)V_{\widehat{P}}^{-1}\left(c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2\right)', \tag{13}$$

where V_P is the asymptotic covariance matrix of $\left(c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2\right)'$. The following result gives the asymptotic distribution of G_n under H_0^c, and the formula for the covariance matrix V_P :

Theorem 2.1 *Under H_0^c, and the assumptions above,*

(i)

$$\sqrt{n}\left(c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2\right) \to^d N(0, V_P),$$

with V_P given by

$$V_P = \begin{pmatrix} P_1^{-1}P_2 \\ -I_{K-J} \end{pmatrix}' \mathcal{D} \begin{pmatrix} P_1^{-1}P_2 \\ -I_{K-J} \end{pmatrix},$$

where

$$\mathcal{D} = diag\{c_1'P_1^{-1}D_kP_1'^{-1}c_1; k = 1, ..., K\},$$

in which D_k is a $J \times J$ diagonal matrix with the elements in column k of P on the diagonal.

(ii) Under H_0^c,

$$G_n \to^d \chi^2_{K-J}.$$

Large values of G_n provide evidence against H_0^c. The practical application of the above testing procedure is presented in Section 5.3, where the point-identifiability of the differences in returns to schooling for various educational levels is tested. The question is open whether an arbitrary function might be set identified under (5), in contrast to point identified functions dealt with in Proposition 3. Suppose that interest is in $g(\beta)$, a (given) function from $\mathbb{R}^K$ to $\mathbb{R}$. One can ask whether the restrictions (5) on β restrict $g(\beta)$ to a subset of $\mathbb{R}$. This could be addressed by using similar methods as in Freyberger and Horowitz (2014), by solving the extremum problems:

$$\max.(\min.)\ g(\beta)\ s.t.\ \pi = \Pi\beta.$$

If the maximum coincides with the minimum, $g(\beta)$ would be point identified, while if either is finite $g(\beta)$ would be set-identified. We do not develop this further here.

2.3 Estimation

This section presents some basic estimation results, and provides a link between standard nonparametric estimators and the estimators defined for our linear model. It also briefly summarizes the estimation techniques available in the literature for models that set identifies the function of interest.

2.3.1 Estimation under point identification

Firstly, we assume that $J \geq K$, i.e. that the model is point identified.

OLS under exogeneity Since, for any support point x_k of X, we have

$$E[Y|X = x_k] = h(x_k) + E[\varepsilon|X = x_k],$$

$\beta_k = h(x_k)$ can be nonparametrically estimated from the data by averaging the y_i corresponding to all x_i^s that equal x_k.

Under the exogeneity assumption, and given the linear interpretation of the model, we have the standard OLS estimator for β:

$$\widehat{\beta} = \left(L_X'L_X\right)^{-1} L_X'y = \begin{pmatrix} \frac{\sum_{i=1}^n y_i I(x_i^s = x_1)}{\sum_{i=1}^n I(x_i^s = x_1)} \\ \dots \\ \frac{\sum_{i=1}^n y_i I(x_i^s = x_K)}{\sum_{i=1}^n I(x_i^s = x_K)} \end{pmatrix}, \tag{14}$$

which coincides with the standard nonparametric estimator (see, for example Pagan and Ullah (1999), Section 3.2.2). The important observation is that the value of the conditional mean of Y given X, does not depend on the values x_k of X and the configuration of x_k in the sample (the position of non-zero elements in the matrix L_X) does not matter. The only thing that matters is the multiplicity of each x_k in the sample. Since $\frac{n_k^X}{n}$ is a sample proportion, it converges in probability to p_k i.e. the probability mass on the support point x_k.

Substituting the linear model $y = L_X\beta + \varepsilon_n$ in (14) gives

$$\widehat{\beta} = \beta + \left(L_X'L_X\right)^{-1} L_X'\varepsilon_n$$

and since $\left(n^{-1}L_X'L_X\right)^{-1} \to^p D_X^{-1}$ where D_X is $diag(p_k)$, the matrix of probability masses on each point of support of X on the main diagonal, and $n^{-1}L_X'\varepsilon_n \to^p E_X[L_X'E[\varepsilon_n|X]] = 0$ under the assumption of exogeneity, we obviously have

$$\widehat{\beta} \to^p \beta,$$

i.e. if X is exogenous, the OLS estimator $\widehat{\beta}$ is a consistent estimator of β. Using the linear interpretation of the model, we can easily establish the asymptotic distribution of the OLS estimator (or the standard nonparametric estimator).

Theorem 2.2 *Under assumptions above, if X is exogenous then the nonparametric (OLS) estimator $\widehat{\beta}$ is consistent and*

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \to^d N\left(0, \sigma^2 D_X^{-1}\right).$$

Remark 2.2 *The primitive components of the elements of $\widehat{\beta}$ are sums of random numbers of i.i.d.random variables, since the multiplicities and positions of the x_k in the sample are random. At first sight, therefore, one might expect to need a central limit theorem adapted to this situation, such as those of, for example,* *Robbins (1948), or Anscombe (1952), both of which deal with this case. However, the problem turns out to be more straightforward, and Theorem 2.2 can be proved by using a multivariate version of the Lindeberg-Feller central limit theorem (see Appendix A). Alternative ways to determine the asymptotic distribution using suggested above CLTs can also be found in the Appendix A.*

It can be shown that the covariance matrix $\sigma^2 D_X^{-1}$, under exogeneity achieves the asymptotic Cramer-Rao bound for the variance, and hence $\widehat{\beta}$ is asymptotically efficient. The unknown parameter σ^2 can be consistently estimated by the usual estimator used in a linear regression model: $n^{-1} y' M_{L_X} y \to^p \sigma^2$.

If $E[\varepsilon|X = x_k] \neq 0$ and X is endogenous, then

$$n^{-1} L_X' \varepsilon_n \to^p E_X[L_X' E[\varepsilon|X]] \neq 0$$

and $\widehat{\beta}$ is an inconsistent estimator for β.

IV under endogeneity If X is endogenous and $J \geq K$, the unknown function $h(\cdot)$ (or vector β) can be estimated using familiar IV methods. When the model point-identifies the structure of interest, the problem can be treated as a standard IV problem and the IV estimator for β is

$$\begin{aligned} \widehat{\beta}_{IV} &= \left(\widehat{\Pi}' L_Z' L_Z \widehat{\Pi}\right)^{-1} \widehat{\Pi}' L_Z' L_Z \widehat{\pi} \\ &= \left(L_X' L_Z \left(L_Z' L_Z\right)^{-1} L_Z' L_X\right)^{-1} L_X' L_Z \left(L_Z' L_Z\right)^{-1} L_Z' y \\ &= \left(L_X' P_{L_Z} L_X\right)^{-1} L_X' P_{L_Z} y. \end{aligned}$$

This is the IV estimator for β in the null model $y = L_X \beta + \varepsilon_n$, in the presence of the instrument matrix L_Z. Even though in the nonparametric specification there is only one discete instrument Z, we have J instrumental variables ($I(Z = z_j)$, $j = 1, ..., J$) in the linear model specification. The matrix of instruments corresponding to this interpretation of the model is L_Z, so the familiar requirements for the validity of the instruments are that $n^{-1} L_Z' L_X \to^p P$, a finite matrix of rank K, that $n^{-1} L_Z' \varepsilon_n \to^p 0$ and $n^{-1} L_Z' L_Z \to^p D_Z$, a positive definite matrix (e.g., Greene (1993), p.601). All these conditions are covered by Assumptions 2 and 3. However, crucially, when $J < K$ the IV estimator is no longer available.

The equivalence of $\widehat{\beta}_{IV}$ and the standard nonparametric IV estimator is shown by the following example.

Example 2.1 *Consider a simple nonparametric model with a scalar binary regressor X, taking values $x_1 = 0$ and $x_2 = 1$. In nonparametric literature, this model is typically reparametrized by defining $\alpha = h(x_1) = h(0)$ and $\beta = h(x_2) - h(x_1) = h(1) - h(0)$ as in Florens and Malavolti (2003). Given the instrument Z, the standard nonparametric IV estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are defined as*

$$\begin{aligned} \widehat{\alpha} &= \widehat{E}(\widehat{E}(Y|Z)|X=0) - \widehat{\beta}\widehat{E}(\widehat{E}(X|Z)|X=0) \\ \widehat{\beta} &= \frac{\widehat{E}(\widehat{E}(Y|Z)|X=1) - \widehat{E}(\widehat{E}(Y|Z)|X=0)}{\widehat{E}(\widehat{E}(X|Z)|X=1) - \widehat{E}(\widehat{E}(X|Z)|X=0)}, \end{aligned} \tag{15}$$

with

$$\begin{aligned} \widehat{E}(W|Z) &= \frac{\frac{1}{n}\sum_{i=1}^{n} w_i I(z_i^s = z)}{\frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z)} \\ \widehat{E}(\widehat{E}(W|Z)|X &= x_k) = \frac{1}{\sum_{m=1}^{n} I(x_m^s = x_k)} \sum_{m=1}^{n} \frac{\frac{1}{n}\sum_{i=1}^{n} w_i I(z_i^s = z_m^s)}{\frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_m^s)} I(x_m^s = x_k), \end{aligned}$$

where $W = Y$ or X.

Firstly, notice that $I(z_i^s = z_m^s) = 1$ if $z_i^s = z_m^s$, i.e. two observations are the same. This can only happen if both z_i^s and z_m^s take the same value z_j from the support of Z. Given the basic properties of the indicator function, we know that $I_{A\cap B} = I_A \cdot I_B$. Therefore

$$\begin{aligned} I(z_i^s &= z_m^s) = I(z_i^s = z_1)I(z_m^s = z_1) + \ldots + I(z_i^s = z_J)I(z_m^s = z_J) \\ &= \sum_{j=1}^{J} I(z_i^s = z_j)I(z_m^s = z_j). \end{aligned}$$

It follows that

$$\begin{aligned} \widehat{E}(\widehat{E}(W|Z)|X &= x_k) \\ &= \frac{1}{\sum_{m=1}^{n} I(x_m^s = x_k)} \sum_{m=1}^{n} \frac{\sum_{j=1}^{J} \frac{1}{n}\sum_{i=1}^{n} w_i I(z_i^s = z_j) I(z_m^s = z_j)}{\sum_{j=1}^{J} \frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_j) I(z_m^s = z_j)} I(x_m^s = x_k). \end{aligned}$$

Clearly, closed form solutions for $\widehat{h}(x_1)$ and $\widehat{h}(x_2)$ are very complicated, and even *if we wanted to state them here, it is impossible due to space limitations. Using the IV estimator obtained for a linear specification, the estimators are expressed in a short and elegant way as*

$$\begin{pmatrix} \widehat{h}(x_1) \\ \widehat{h}(x_2) \end{pmatrix} = \left(L'_X P_{L_Z} L_X\right)^{-1} L'_X P_{L_Z} y$$

Given that P_{L_Z} is a symmetric $n \times n$ matrix with the im^{th} element equal to

$$\sum_{j=1}^{J} \frac{I(z_i^s = z_j) I(z_m^s = z_j)}{n_j^Z},$$

the whole $L'_X P_{L_Z} L_X$ is a K- square symmetric matrix with the kl^{th} element equal to

$$\sum_{i=1}^{n} I(x_i^s = x_k) \sum_{m=1}^{n} I(x_m^s = x_l) \sum_{j=1}^{J} \frac{I(z_i^s = z_j) I(z_m^s = z_j)}{n_j^Z}.$$

The K- vector $L'_X P_{L_Z} y$ is simply

$$\begin{pmatrix} \sum_{i=1}^{n} y_i \sum_{m=1}^{n} I(x_m^s = x_1) \sum_{j=1}^{J} \frac{I(z_i^s = z_j) I(z_m^s = z_j)}{n_j^Z} \\ \ldots \\ \sum_{i=1}^{n} y_i \sum_{m=1}^{n} I(x_m^s = x_K) \sum_{j=1}^{J} \frac{I(z_i^s = z_j) I(z_m^s = z_j)}{n_j^Z} \end{pmatrix}.$$

The manipulation of terms in (15) shows that the IV estimator defined for the linear interpretation of the nonparametric model is equivalent to the standard nonparametric estimator. The advantage of our approach is that the estimator can be written in a compact matrix notation, which is easier to work with.

Provided $J \geq K$, the IV estimator is consistent in both scenarios: when X is exogenous and when it is endogenous, since $n^{-1} L'_Z L_X \rightarrow^p P$, $n^{-1} L'_Z L_Z \rightarrow^p D_Z$ and $n^{-1} L'_Z \varepsilon_n \rightarrow^p E_Z[L'_Z E[\varepsilon_n | Z]] = 0$. The last expression follows because of the instrument exogeneity condition (3).

The asymptotic normality of the IV estimator is established through:

Theorem 2.3 *Under assumptions above, the IV estimator $\widehat{\beta}_{IV}$ is consistent and*

$$\sqrt{n}\left(\widehat{\beta}_{IV} - \beta\right) \rightarrow^d N\left(0, \sigma^2 \left(P' D_Z^{-1} P\right)^{-1}\right).$$

Note that because K and J are fixed, we cannot estimate the entire unknown function $h(\cdot)$, but can only learn the values of $h(\cdot)$ at the support points of X. Additional information about $h(\cdot)$, could possibly be acquired if the support of the regressor (and instrument) were assumed to be increasing with the sample size, and this would also have implications for identifiability. Further study of this interesting possibility is beyond the scope of this thesis, but some basic ideas are outlined in the final conclusion of this thesis.

2.3.2 Estimation under partial identification

Now, we assume that $J < K$. Given the results in Section 2.2.1, if the null hypothesis of point identifiability of $L(\beta) = c'\beta$ is not rejected, the linear functional of interest can be estimated by

$$\widehat{L(\beta)} = c_1' \widehat{\Pi}_1^{-1} \widehat{\pi},$$

a consistent estimator if H_0^c is indeed true. On the other hand, rejection of the null hypothesis suggests that the linear functional of interest $L(\beta)$ cannot be consistently estimated without further information about $h(\cdot)$.

Several classes of additional restrictions on $h(\cdot)$, which produce bounds on certain linear functionals, so that the functionals are set identified, have been considered in the literature. Chesher (2004) gives conditions under which an informative bound on h can be consistently nonparametrically estimated from the data. He assumes that the structural model consists of equations (1) and (3), and additionally

$$X = g(Z, U),$$

where the continuously distributed error term U is normalized to $Unif(0,1)$ and is independent of the instrumental variable Z. The function $g(z,u)$ is the conditional quantile function of X given Z. Even though under standard marginal covariation condition (11), the value of $h(\cdot)$ is not point identified, it can be bounded in informative ways under the iterated covariation restriction of the following form:

$$E\left[\varepsilon | U = u, Z = z_j\right] = c(u), \tag{16}$$

where c is assumed to be a monotonic function. Since U and Z are assumed to be independent, condition (16) implies, but is not implied by (11). Therefore, the iterated covariation condition provides additional identifying information if $J < K$, but does not increase the identifying power of the model if $J \geq K$.

Suppose that there exist points z_{k-1}, z_k in the support of the instrument Z, such that for some $\overline{u} \in (0,1)$ and some $k = 1, ..., K$, we have

$$\Pr\left[X = x_k | Z = z_k\right] \leq \overline{u} \leq \Pr\left[X = x_{k-1} | Z = z_{k-1}\right]. \tag{17}$$

Under these additional assumptions given by equations (16) and (17), the following partial identification result is obtained:

$$\begin{aligned} & \min\left\{E\left[Y|X = x_k, Z = z_k\right], E\left[Y|X = x_k, Z = z_{k-1}\right]\right\} \\ \leq\ & h(x_k) + c(\overline{u}) \leq \\ & \max\left\{E\left[Y|X = x_k, Z = z_k\right], E\left[Y|X = x_k, Z = z_{k-1}\right]\right\}. \end{aligned}$$

This allow us to obtain the upper and lower bounds on differences $h(x_k) - h(x_j)$ by replacing the conditional expectations with sample averages. Therefore, even though the exact value of the vector β remains unknown, we are able to bound its value by the quantities that are easily estimated from the data. We present a practical application of Chesher's (2004) method in Section 5.3 to bound the differences in returns to schooling for various years of education using real data.

Manski and Pepper (2000) give conditions under which the upper and lower bounds on $h(x_k)$ and the upper bound on $h(x_k) - h(x_j)$ can be consistently estimated. Their "monotone treatment response" condition (analogous to monotonicity assumption (16)) ensures that for the outcomes $y^{(1)}$ and $y^{(2)}$ of the treatment values $x^{(1)}$ and $x^{(2)}$, $x^{(2)} \geq x^{(1)}$ implies $y^{(2)} \geq y^{(1)}$. The second assumption ("monotone treatment selection") replaces the standard assumption of availability of the relevant instruments and states that if $x^{(2)} \geq x^{(1)}$ then $E\left[Y|X_S = x^{(2)}\right] \geq E\left[Y|X_S = x^{(1)}\right]$, where X_S is the treatment selected by an individual. Under Manski and Pepper's (2000) additional restrictions, the value of $h(x_k)$ is bounded by

$$\begin{aligned} & \sum_{m:x_m<x_k} E\left[Y|X = x_m\right] p_m + E\left[Y|X = x_k\right] \Pr\left[X \geq x_k\right] \\ \leq\ & h(x_k) \leq \\ & \sum_{m:x_m>x_k} E\left[Y|X = x_m\right] p_m + E\left[Y|X = x_k\right] \Pr\left[X \leq x_k\right]. \end{aligned}$$

The third estimation method available in the literature under set identification is by Freyberger and Horowitz (2014), who study the identification and estimation of the linear functional $L(\beta) = c'\beta$. They use the shape restrictions on the unknown function of interest dictated by the economic theory, such as monotonicity of the demand function or convexity of cost functions, to obtain the bounds on $L(\beta)$ by solving linear programming problems. Their upper and lower bounds on $L(\beta)$ are equivalent to the bounds on the local average treatment effect obtained by Angrist and Imbens (1995) for systems of linear simultaneous equations.

2.4 Summary

This section has presented a new approach to estimation in the nonparametric additive errors model with discrete regressors. The fact that the explanatory variable only takes a finite number of distinct values enabled us to put the nonparametric structure into a well-known linear regression framework with $n \times K$ matrix of explanatory variables. Under standard assumptions, it has been shown that the simple nonparametric estimator of the conditional mean of the dependent variable given the set of regressors coincides with the ordinary least square estimator for the linear regression. When the explanatory variables are exogenous, this OLS estimator is consistent and follows the normal distribution asymptotically. In the presence of endogenous regressors, the instrumental variable estimation has been proposed. The identification study has revealed that the model point identifies the unknown structure of interest if the discrete instrument available to the researcher has at least as many points of support as the endogenous regressor. Under identification failure, nothing can be learned about the entire unknown function of interest without further restrictions. However, it has been shown that there exists some linear combinations of parameters of interest which might be point- identified. A test for point-identifiability of such functionals has been proposed. When the explanatory variables are endogenous, but the nonparametric identification condition is satisfied, the standard two stage least squares estimator for a linear regression coincides with the nonparametric IV estimator and has been proven to be consistent and to follow the normal distribution asymptotically.

3 Nonparametric testing for exogeneity with discrete regressors and instruments

The presence of endogenous regressors in a nonparametric model produces bias in the identified case, and in the partially identified case means that there are no consistent estimators for some parameters. We are particularly interested in testing exogeneity in models that are partially identified under the alternative hypothesis. There are many published applications in which the (assumed) endogenous regressor is instrumented by a variable with insufficient support. For instance, in Angrist and Krueger (1991) endogenous education is instrumented by the quarter of birth of an individual, and Bronars and Grogger (1994) use the twin birth indicator as an instrument for endogenous number of children. In these papers point identification is achieved by assuming a parametric (linear) specification. However, the parametric specification is an additional assumption, and the validity of such assumption should be tested. Parametric vs. nonparametric specification testing has been discussed by Donald, Imbens and Newey (2003) and Tripathi and Kitamura (2003). In nonparametric single equation IV models, the test proposed by Horowitz (2006) could be employed to check whether the parametric specification is appropriate for the available data. If the null hypothesis is rejected, then nonparametric estimation should be chosen. Typically, these specification tests are based on a comparison of parametric and nonparametric estimators. Since, under endogeneity, there exists no consistent estimator for the entire conditional mean function when the support of instrument is sparse relative to the support of endogenous regressor, the parametric specification hypothesis is not testable. This provides an incentive to use the nonparametric model. Alternatively, one could nonparametrically test for exogeneity of regressors, and given the outcome of the test, decide on functional form and estimation method.

Ideally, we would like to test whether $E[\varepsilon|X = x_k] = 0$ i.e. X is exogenous, which in terms of observables can be written as

$$H_0 : E[Y|X = x_k] = h(x_k), \quad k = 1, ..., K.$$

In equation (1) the function $h(\cdot)$ is unknown and if $h(x_k)$ were completely arbitrary, the null hypothesis would not impose any constraint on the conditional density function of Y given X, $f_{Y|X}(y|x)$ and would therefore be untestable. Thus, more information about $h(\cdot)$ than just equation (1) alone is required for H_0 to become a testable hypothesis. This additional information about $h(\cdot)$ is gained by using the fact that there exist a valid instrument Z satisfying (3) for any admissible z_j. The hypothesis H_0 imposes the constraint that the vector of conditional means $E[Y|X = x_k]$ is a solution to linear equations $\pi = \Pi\beta$, so in this case the null hypothesis imposes a restriction on the conditional density function $f_{Y|X}(y|x)$ and is therefore testable.

Remark 3.1 *There might be other restrictions that can be imposed on $h(\cdot)$ to make the null hypothesis testable. In order to make sure that $h(\cdot)$ is not entirely arbitrary, one could impose some shape restrictions dictated by economic theory. Such restrictions are already in use in the literature of nonparametric estimation, for example by Hall and Huang (2001) who estimate the conditional mean function subject to a monotonicity constraint. Monotone estimates are required in many empirical applications, when the theory suggests that the outcome should be monotonic in explanatory variables e.g. wage increasing in the years of schooling. Blundell, Horowitz and Parey (2012) use different shape restriction and provide a nonparametric estimator of the demand function assuming that the unknown function $h(\cdot)$ satisfies the Slutsky condition of consumer theory. The literature suggests that imposing shape restrictions improves the precision of nonparametric estimates, but in our case, it might also act as a tool to ensure that the hypothesis of exogeneity of regressors is testable.*

This chapter is organized as follows. Section 3.1 deals with nonparametric exogeneity testing under point identification. It presents a modified version of the Durbin-Wu-Hausman test-statistic, establishes the asymptotic distribution of the test-statistic under the null hypothesis and local alternatives. In Section 3.2, we provide the exogeneity testing procedure that can be applied in models that are partially identified. Additionally, we discuss various ways of computing critical values. In Section 3.3 the results of Monte Carlo simulations are presented.

3.1 Testing under point identification

Assume that $J \geq K$, so the model point identifies the entire unknown function of interest $h(\cdot)$.

3.1.1 Test statistic

We are interested in testing $E[\varepsilon|X = x_k] = 0$ through the following hypothesis of exogeneity of regressors:

$$H_0 : E[Y|X = x_k] = h(x_k), \quad k = 1, ..., K$$

or equivalently

$$H_0 : E[Y_n|X_n = L_X x] = L_X \beta.$$

In the previous chapter, we have shown that the OLS estimator $\widehat{\beta}$ is consistent and efficient if X is exogenous, but inconsistent otherwise. The IV estimator is consistent in both cases, but inefficient if X is exogenous. For this situation, then, the test is really just to decide which estimator to use (OLS or IV).

The OLS and IV estimators (with instrument Z) for β are:

$$\widehat{\beta} = (L'_X L_X)^{-1} L'_X y$$

$$\widehat{\beta}_{IV} = (L'_X P_{L_Z} L_X)^{-1} L'_X P_{L_Z} y$$

and the difference between them is therefore

$$\begin{aligned} \widehat{\beta}_{IV} - \widehat{\beta} &= (L'_X P_{L_Z} L_X)^{-1} L'_X P_{L_Z} y - (L'_X L_X)^{-1} L'_X y \\ &= (L'_X P_{L_Z} L_X)^{-1} L'_X P_{L_Z} M_{L_X} y. \end{aligned} \quad (18)$$

The covariance matrix of that difference is given by

$$Cov(\widehat{\beta}_{IV} - \widehat{\beta}) = (L'_X P_{L_Z} L_X)^{-1} L'_X P_{L_Z} M_{L_X} P_{L_Z} L_X (L'_X P_{L_Z} L_X)^{-1}. \quad (19)$$

An obvious test could be based on the Durbin-Wu-Hausman- type statistic. The standard Durbin-Wu-Hausman test is based on a quadratic form the difference (18), with the matrix of the quadratic form equal to the inverse of $Cov(\widehat{\beta}_{IV} - \widehat{\beta})$ (in order to produce a χ^2variable asymptotically). However, Property C implies that the relevant covariance matrix is in this case singular. To see this, observe that

$$\begin{aligned} l_K' \left(L_X' P_{L_Z} L_X \right) (\widehat{\beta}_{IV} - \widehat{\beta}) &= l_K' L_X' P_{L_Z} M_{L_X} y \\ &= l_n' P_{L_Z} M_{L_X} y \;\; \text{since } L_X l_K = 0 \\ &= l_n' M_{L_X} y \;\; \text{since } P_{L_Z} l_n = l_n \\ &= 0 \;\; \text{since } M_{L_X} l_n = 0. \end{aligned}$$

That is, for all L_X and L_Z there is an exact linear relation between the elements of $\widehat{\beta}_{IV} - \widehat{\beta}$, so its covariance matrix will always be singular.

We need to adapt the Durbin-Wu-Hausman test statistic to this situation. To do so we simply replace the inverse of the covariance matrix - the matrix that would normally be used in the quadratic form to produce an asymptotically χ^2 test statistic - by a generalized inverse of that matrix. The covariance matrix in (19) can be written as

$$\left(L_X' P_{L_Z} L_X \right)^{-1} C_K \left[C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K \right] C_K' \left(L_X' P_{L_Z} L_X \right)^{-1},$$

since $M_{L_X} P_{L_Z} L_X [l_K, C_K] = [0, M_{L_X} P_{L_Z} L_X C_K]$ and $[l_K, C_K]^{-1} = [K^{-1} l_K, C_K]'$[6]. The middle matrix $C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K$ is $(K-1)$ square matrix of full rank. Thus, the covariance matrix can be expressed as a matrix of the form $S = A^{-1} C B C' A^{-1}$, where C is $m \times p$, $C'C = I_p$, B is $p \times p$ nonsingular and symmetric and A is $m \times m$ nonsingular and symmetric. The generalized inverse of the matrix with this form is $S^{+} = A C B^{-1} C' A$. To verify that it is sufficient to check two conditions that define a generalized inverse, i.e. $SS^{+}S = S$ and $S^{+}SS^{+} = S^{+}$(both conditions hold).

Therefore, the generalized inverse of the covariance matrix is

$$S^{+} = \left(L_X' P_{L_Z} L_X \right) C_K \left[C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K \right]^{-1} C_K' \left(L_X' P_{L_Z} L_X \right).$$

[6]Essentially, what we are doing to construct the C_K matrix is the Gram-Schmidt ortogonalization procedure on the K-vector of ones.

Using this matrix to define the test statistic, we have

$$\begin{aligned} T_n^* &= (\widehat{\beta}_{IV} - \widehat{\beta})' \left[Cov(\widehat{\beta}_{IV} - \widehat{\beta}) \right]^+ (\widehat{\beta}_{IV} - \widehat{\beta}) \\ &= y' M_{L_X} P_{L_Z} L_X C_K \left[C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K \right]^{-1} C_K' L_X' P_{L_Z} M_{L_X} y \\ &= y' W_{XZ} \left(W_{XZ}' W_{XZ} \right)^{-1} W_{XZ}' y, \end{aligned}$$

where

$$\begin{aligned} W_{XZ} &= M_{L_X} P_{L_Z} L_X C_K \\ &= M_{L_X} L_Z \left(L_Z' L_Z \right)^{-1} L_Z' L_X C_K \end{aligned}$$

is $n \times (K-1)$.

Scaling to eliminate σ^2, we propose the test-statistic

$$T_n = \frac{y' W_{XZ} \left(W_{XZ}' W_{XZ} \right)^{-1} W_{XZ}' y}{n^{-1} y' M_{L_X} y}. \tag{20}$$

Observe that the values x_k of X and z_j of Z do not appear in the test statistic, nor does their configuration in the sample matter. The only things that appear are the multiplicities of each value in the sample, the n_k^X and n_j^Z, and the multiplicity of the joint event $(X = x_k, Z = z_j)$, n_{jk}. Note also that the numerator of the modified version of T_n is easily computed from a linear regression of y on W_{XZ}. Since W_{XZ} is easy to construct in practice, the value of the test-statistic can be efficiently calculated by any statistical software package.

Remark 3.2 *Using the generalized inverse is not the only way to deal with singularity of the covariance matrix. The naive approach would be to reduce the dimension of the test-statistic by eliminating for example, the first element in the difference (18) and picking up the lower-right corner of the covariance matrix in (19). Then the Durbin-Wu-Hausman test-statistic of the reduced dimension would follow the standard results. Alternative approach would be to use the Moore-Penrose inverse of the covariance matrix (built in all econometric software). All three approaches give similar values of the test-statistic, thus in applications, the researcher could choose the method that is the most convenient for them.*

3.1.2 Asymptotic distribution under the null hypothesis

In order to obtain the asymptotic distribution of the test statistic under H_0, observe that the primitive components the numerator of T_n are the two vectors $u_n = n^{-1}L_Z'\varepsilon_n$ and $v_n = n^{-1}L_X'\varepsilon_n$. In terms of these, the vector appearing in the numerator is

$$nC_K'L_X'L_Z(L_Z'L_Z)^{-1}\left(u_n - L_Z'L_X(L_X'L_X)^{-1}v_n\right).$$

Thus, we first consider the asymptotic behaviour of these two vectors, i.e. the asymptotic distribution of

$$w_n = \sqrt{n}\begin{pmatrix} u_n \\ v_n \end{pmatrix}.$$

This is given in:

Lemma 3.1 *Under H_0 and the given assumptions,*

$$w_n \to^d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\begin{bmatrix} D_Z & P \\ P' & D_X \end{bmatrix}\right).$$

This result will also be useful in the partially identified model later. Putting

$$z_{1n} = \sqrt{n}C_K'L_X'L_Z(L_Z'L_Z)^{-1}\left(u_n - L_Z'L_X(L_X'L_X)^{-1}v_n\right),$$

we have

Lemma 3.2 *Under H_0 and the given assumptions,*

$$z_{1n} \to^d N(0, \sigma^2\Sigma_{11})$$

where

$$\Sigma_{11} = C_K'P'D_Z^{-1}\left[D_Z - PD_X^{-1}P'\right]D_Z^{-1}PC_K$$

is positive definite.

Now

$$T_n = \frac{z_{1n}'\left[n^{-1}C_K'L_X'P_{L_Z}M_{L_X}P_{L_Z}L_XC_K\right]^{-1}z_{1n}}{n^{-1}y'M_{L_X}y},$$

and it is easily seen that

$$\left[n^{-1}C_K'L_X'P_{L_Z}M_{L_X}P_{L_Z}L_XC_K\right] \to_p \Sigma_{11}.$$

Thus, as expected, we have:

Theorem 3.1 *Under H_0, and the assumptions above,*

$$T_n \to^d \chi^2_{K-1}.$$

Evidently, critical values for the test can be easily obtained from statistical tables. The accuracy of this asymptotic result is examined in Section 3.3.

3.1.3 Test consistency

It is straightforward to see that, under suitable conditions on the class of alternative hypotheses, the test that rejects H_0 for large T_n is consistent, i.e. the power of the test approaches 1 as $n \to \infty$. Assume that the conditional expectation of the error term in (1) is given by:

$$E[\varepsilon_n | X_n = L_X x] = m, \tag{21}$$

where m is an $n-$vector that may be fixed or random, but *must not be in the column space of L_X*. That is, there must be variation in the elements of m that correspond to each of the support points x_k of X. If this is not the case the test will have no power. Under the alternative hypothesis we will have $E[\varepsilon | X = x_k] \neq 0$ for at least one value of k. Cases with no power occurs if, when the null hypothesis fails

$$E[Y|X = x_k] = h(x_k) + \gamma(x_k),$$

where $\gamma(x_k) = E[\varepsilon | X = x_k]$ depends only on x_k. In this case instead of (10), we will have the model

$$y = L_X(\beta + \gamma) + \widetilde{\varepsilon}_n,$$

where $\widetilde{\varepsilon}_n = \varepsilon_n - \gamma$, which is identical to the original model with the unknown h replaced by the also-unknown $h + \gamma$. Recall that the null hypothesis claims that the CMF of Y lies in the column space of L_X and the alternative (21) simply says that H_0 is false. Since we continue to assume the validity of the instrument, i.e., that $p\lim_{n\to\infty}(n^{-1}L_Z'\varepsilon_n) = 0$, we must have $p\lim_{n\to\infty}(n^{-1}L_Z'm) = 0$. We also assume that

$$p\lim_{n\to\infty}(n^{-1}L_X'm) = \mu,$$

a non-zero finite vector. Additionally, let σ^{*2} denote the probability limit of $n^{-1}y'M_{L_X}y$ under alternative hypothesis (21), and assume that $\sigma^{*2} < \infty$.

The following proposition establishes the consistency of the test against a fixed alternative hypothesis.

Proposition 4 *Under fixed alternatives (21) and the earlier assumptions, the proposed test is consistent, i.e., for any fixed constant* c_α,

$$\lim_{n\to\infty} \Pr\left(T_n > c_\alpha\right) = 1.$$

The consistency follows from the fact that

$$p\lim_{n\to\infty} n^{-1}T_n = \frac{\xi'\Sigma_{11}^{-1}\xi}{\sigma^{*2}} > 0, \tag{22}$$

where

$$\xi = C_K'P'D_Z^{-1}PD_X^{-1}\mu,$$

i.e. the value of the test statistic tends to $+\infty$ under fixed alternatives, implying that the power of the test tends to unity as $n \to \infty$. Note that the standard estimator $y'M_{L_X}y$ in the denominator of the test statistic (20) under-estimates σ^2. Since

$$n^{-1}y'M_{L_X}y = n^{-1}\varepsilon_n'M_{L_X}\varepsilon_n = n^{-1}\varepsilon_n'\varepsilon_n - n^{-1}\varepsilon_n'L_X(L_X'L_X)^{-1}L_X'\varepsilon_n,$$

and given the fact that

$$p\lim_{n\to\infty} n^{-1}\varepsilon_n'L_X(L_X'L_X)^{-1}L_X'\varepsilon_n = \mu'D_X^{-1}\mu,$$

it follows that

$$n^{-1}y'M_{L_X}y \to^p \sigma^2 - \mu'D_X^{-1}\mu \equiv \sigma^{*2} < \sigma^2.$$

This result has no impact on the consistency of the proposed test, as long as σ^{*2} is a finite positive constant.

3.1.4 Asymptotic distribution under local alternatives

Since the power of the test at fixed alternatives converges to 1, we examine the ability of the test to detect small deviations from the null hypothesis. We consider a sequence of local alternatives that converges to the null hypothesis at rate $O(n^{-\frac{1}{2}})$, i.e.

$$E[\varepsilon_n | X_n = L_X x] = n^{-\frac{1}{2}} m. \tag{23}$$

A simple generalization of Lemmas 3.1 and 3.2 gives

Lemma 3.3 *Under the sequence of local alternatives (23), and the assumptions on m just given,*

$$w_n \to^d N\left(\left(\begin{array}{c} 0 \\ \mu \end{array}\right), \sigma^2 \left[\begin{array}{cc} D_Z & P \\ P' & D_X \end{array}\right]\right)$$

and

$$z_{1n} \to^d N(-\xi, \sigma^2 \Sigma_{11}).$$

From familiar results for quadratic forms in normal vectors with non-zero mean, it immediately follows that

Theorem 3.2 *Under the sequence of local alternatives (23), and the assumptions above, the test statistic*

$$T_n \to^d Gamma\left(\beta, \lambda, \theta\right),$$

with the shape parameter $\alpha = \frac{K-1}{2}$, the scale parameter $\theta = 2\frac{\sigma^2}{\sigma^{*2}}$ and the non-centrality parameter $\lambda = 2\delta^2$, where

$$\delta^2 = \frac{\xi' \Sigma_{11}^{-1} \xi}{\sigma^2}.$$

The asymptotic behaviour of the test-statistic is captured by the non-central Gamma distribution. For a given size, the power of the test increases with noncentrality parameter δ^2. As in the standard Durbin-Wu-Hausman test, the value of this parameter depends on the the distance between the probability limits of the OLS and IV estimators. Hence, the test is more powerful if the probability limit of the OLS estimator is far from the true value of the parameter of interest.

3.2 Testing under partial identification

In this situation ($J < K$), there is no consistent estimator (in the conventional sense) for β if X is endogenous, so in this case the test is to decide whether point estimation of β is even possible. When $J < K$, the Durbin-Wu-Hausman approach to testing H_0 is not available. However, assuming the existence of an instrument Z with the properties given above, β is constrained to satisfy the linear equations $\pi = \Pi\beta$, but is not point identified by them. That is, there is a set of vectors β, a subset of $\mathbb{R}^K$, that satisfy these equations, of dimension $K - J$. The model maintains that β belongs to this set, and H_0 says that $E[Y_n|X_n = L_X x] = L_X\beta$.

3.2.1 Test statistic

Now, consider the empirical counterpart of the system $\pi = \Pi\beta$, namely $\widehat{\pi} = \widehat{\Pi}\beta$, and the vector β that, among all solutions to this system, minimizes $(y - L_X\beta)'(y - L_X\beta)$. That is, define

$$\widehat{\beta}_Z = \arg \min_{\beta:\widehat{\pi}=\widehat{\Pi}\beta} (y - L_X\beta)'(y - L_X\beta).$$

Straightforward familiar algebra gives

$$\begin{aligned}\widehat{\beta}_Z &= \widehat{\beta} + (L_X'L_X)^{-1}\widehat{\Pi}'\left(\widehat{\Pi}\,(L_X'L_X)^{-1}\widehat{\Pi}'\right)^{-1}\left(\widehat{\pi} - \widehat{\Pi}\widehat{\beta}\right) \\ &= \widehat{\beta} + (L_X'L_X)^{-1} L_X'L_Z\,(L_Z'P_{L_X}L_Z)^{-1} L_Z'M_{L_X}y,\end{aligned}$$

where $\widehat{\beta}$ is the OLS estimator defined earlier. The minimum achieved by this choice for β is therefore

$$\begin{aligned}Q_n &= (y - L_X\widehat{\beta}_Z)'(y - L_X\widehat{\beta}_Z) \\ &= y'M_{L_X}y + y'M_{L_X}L_Z\,(L_Z'P_{L_X}L_Z)^{-1} L_Z'M_{L_X}y.\end{aligned}$$

Intuitively, a large value for this minimum sum of squares is evidence against H_0, because it means that, among all solutions to $\widehat{\pi} = \widehat{\Pi}\beta$, none produces a small value of $(y - L_X\beta)'(y - L_X\beta)$. This suggests, not that $\pi \neq \Pi\beta$, because this is ruled out, but rather that $E[Y_n|X_n = L_X x] \neq L_X\beta$, i.e. that the null hypothesis is false. Normalizing Q_n by dividing by $n^{-1}y'M_{L_X}y$, this argument suggests rejecting H_0 when the statistic

$$R_n = \frac{y'M_{L_X}L_Z\left(L_Z'P_{L_X}L_Z\right)^{-1}L_Z'M_{L_X}y}{n^{-1}y'M_{L_X}y}$$

is large. Again, note that R_n is not regarded as a measure of whether $\pi = \Pi\beta$, but rather whether $\widehat{\beta}$, which embodies H_0, can satisfy these conditions.

Now, in view of Property C,

$$M_{L_X}L_Z[l_J, C_J] = [M_{L_X}l_n, M_{L_X}L_Z C_J] = [0, M_{L_X}L_Z C_J]$$

and, the $(2,2)$ block of

$$[[l_J, C_J]'\left(L_Z'P_{L_X}L_Z\right)[l_J, C_J]]^{-1} = \begin{bmatrix} n & l_n'L_Z C_J \\ C_J'L_Z'l_n & C_J'L_Z'P_{L_X}L_Z C_J \end{bmatrix}^{-1}$$

is given by

$$\left(C_J'L_Z'\left[P_{L_X} - P_{l_n}\right]L_Z C_J\right)^{-1}.$$

Thus, after taking account of Property C, R_n reduces to

$$R_n = \frac{y'M_{L_X}L_Z C_J\left(C_J'L_Z'\left[P_{L_X} - P_{l_n}\right]L_Z C_J\right)^{-1}C_J'L_Z'M_{L_X}y}{n^{-1}y'M_{L_X}y} \tag{24}$$

with the middle matrix being $(J-1)$ square. Thus, although at first sight a quadratic form involving J variables, the numerator of R_n in fact involves only $J-1$ terms.

3.2.2 Asymptotic distribution under null hypothesis

In this case the vector involved in the numerator of R_n is, in terms of the variates u_n and v_n dealt with in Lemma 3.1,

$$z_{2n} := \sqrt{n}C_J'(u_n - L_Z'L_X(L_X'L_X)^{-1}v_n),$$

and it follows at once that

$$z_{2n} \to^d N(0, \sigma^2 \Sigma),$$

where

$$\Sigma := C_J' \left(D_Z - P D_X^{-1} P' \right) C_J.$$

However, in this case the (normalized) matrix of the quadratic form converges to something other than the inverse of Σ, namely

$$n^{-1} C_J' L_Z' \left[P_{L_X} - P_{l_n} \right] L_Z C_J \to^p C_J' (P D_X^{-1} P' - p_Z p_Z') C_J := \Omega.$$

The following theorem gives the asymptotic distribution of the test statistic under the null hypothesis.

Theorem 3.3 *Under H_0 and the assumptions above,*

$$R_n \to^d z' \Omega^{-1} z \equiv \sum_{j=1}^{J-1} \omega_j \chi_j^2(1) \tag{25}$$

where $z \sim N(0, \Sigma)$, with Σ as defined above,

$$\Omega := C_J' (P D_X^{-1} P' - p_Z p_Z') C_J,$$

and the ω_j are positive eigenvalues satisfying

$$\det[\Sigma - \omega \Omega] = 0$$

with the $\chi_j^2(1)$ variables independent copies of a χ_1^2 random variable.

The asymptotic distribution of the proposed test with discrete regressors and instruments is similar to the distribution obtained by Blundell and Horowitz (2007) for the continuous case. Their test-statistic follows asymptotically the distribution of an infinite sum of weighted chi-square variables with 1 degree of freedom. When calculating critical values, they face the additional problem of approximating an infinite sum by a finite number of terms. In the discrete case, the asymptotic distribution is more straightforward, since it is based on a finite sum of terms due to the discrete nature of variables. Nonetheless, the distribution theory for such variables is complicated, and there is an incentive to use approximations, and several have been discussed extensively in the literature. In Section 3.2.5 we discuss the approximation proposed by Hall (1983) and further explored by Buckley and Eagleson (1988), which allows us to compute the critical values in practical applications.

3.2.3 Test consistency

Assume that H_0 is false and consider again the fixed alternative hypothesis (21). In the proof of Proposition 4 we have shown that

$$p \lim_{n \to \infty} n^{-\frac{1}{2}} w_n = \begin{pmatrix} 0 \\ \mu \end{pmatrix},$$

and it follows at once that

$$p \lim_{n \to \infty} n^{-\frac{1}{2}} z_{2n} = -C_J' P D_X^{-1} \mu = -\zeta,$$

say. Therefore,

$$p \lim_{n \to \infty} n^{-1} R_n = \frac{\zeta' \Omega^{-1} \zeta}{\sigma^{*2}} > 0.$$

Using the same argument as before, we obtain the consistency of the test:

Proposition 5 *Under fixed alternatives (21) and the earlier assumptions, the proposed test is consistent, i.e., for any fixed constant* c_α,

$$\lim_{n \to \infty} \Pr\left(R_n > c_\alpha\right) = 1.$$

3.2.4 Asymptotic distribution under local alternatives

Consider again the sequence of local alternatives (23). By Lemma 3.3, we obtain the asymptotic distribution of z_{2n}:

$$z_{2n} \to^d N\left(-\zeta, \sigma^2 \Sigma\right).$$

The following theorem establishes the asymptotic distribution of the test-statistic under local alternatives.

Theorem 3.4 *Under the sequence of local alternatives (23) and the assumptions above, the test statistic R_n converges to a distribution of a weighted sum of non-central chi-square random variables:*

$$R_n \to^d \frac{\sigma^2}{\sigma^{*2}} \sum_{j=1}^{J-1} \omega_j \chi_1^2(\delta_j^2) \tag{26}$$

with the noncentrality parameters

$$(\delta_1, ..., \delta_{J-1})' = \frac{S'\Sigma^{-\frac{1}{2}}\zeta}{\sigma},$$

where S denotes the orthogonal matrix of the eigenvectors of $\Sigma^{\frac{1}{2}}\Omega^{-1}\Sigma^{\frac{1}{2}}$.

The proof is based on standard results on the distribution of quadratic forms in normal vectors with non-zero mean, and hence omitted. Under local alternatives, the test statistic asymptotically follows the distribution of a weighted sum of non-central chi-square (1) variables. This result again corresponds to the distribution obtained by Blundell and Horowitz (2007) for the continuous case.

3.2.5 Critical values computation

The asymptotic distribution of the test-statistic is non-standard and depends on the weights ω_j, which, in practice, need to be estimated from the data. Since we cannot provide statistical tables with the appropriate tail probabilities and cut off points, it is essential to find a quick technique for calculating the critical values of the proposed test.

The distribution of a weighted sum of $\chi^2_{(1)}$ random variables has been studied in the literature since 1960's. Many authors derived explicit formulas for the probability density function and a cumulative distribution function of the process of interest. The results are typically obtained by examining the behaviour of a moment generating function, as in Mathai (1982) or by using mixtures approximations as in Solomon and Stephens (1977) or Oman and Zacks (1981). An explicit expression for the distribution function is given in Johnson and Kotz (1970). Let $q = \sum_{k=1}^{K} d_k Z_k^2 = z'Dz$, where $D = diag\{d_k\}$ and $z \sim N(0, I_k)$.

Then the probability density function of q is

$$f_q(q) = \frac{\exp\{-\frac{1}{2}\tau q\} q^{\frac{k}{2}-1}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})|D|^{\frac{1}{2}}} \sum_{j=0}^{\infty} \frac{q^j (\frac{1}{2})_j}{j! 2^j (\frac{k}{2})_j} C_j(\tau I_k - D^{-1}), \tag{27}$$

where τ is an arbitrary positive constant, and the cumulative distribution of q is given by

$$F_q(z) = \Pr[q \leq z] = \frac{1}{|\tau D|^{\frac{1}{2}}} \sum_{j=0}^{\infty} \frac{\left(\frac{1}{2}\right)_j}{j!} C_j(I_k - (\tau D)^{-1}) G_{k+2j}(\tau z), \tag{28}$$

where $G_m(.)$ denotes the cumulative distribution function of $\chi^2_{(m)}$ random variable. The function $C_j(M)$ is a top-order zonal polynomial and $(c)_j = c(c+1)...(c+j-1)$ is the Pochhammer symbol. $\left(\frac{1}{2}\right)_j C_j(M)$ are derived from the generating function

$$C(t) = |I - tM|^{-\frac{1}{2}}$$

as the coefficients on $\frac{t^j}{j!}$.

Obviously, (27) and (28) are rather complicated and difficult to handle in empirical applications. From the practical point of view, in order to calculate the critical values for the proposed test, it is crucial to be able to approximate the process of interest by a well known structure. Alternatively, one could use the inverse interpolation procedure of finding the critical values proposed by Sheil and Muircheartaigh (1977). However, this method is computationally intensive and requires specifying the upper and lower bounds on the weights, which we would like to avoid.

There are numerous ways of computing the critical values in this case. Letting $\widehat{\omega}_j$ be consistent estimators of the weights ω_j under H_0, the distribution of $\sum_{j=1}^{J-1} \widehat{\omega}_j \chi_j^2(1)$ can be simulated and appropriate $1-\alpha$ quantiles can be used as critical values in the standard rejection rule. However, our experiments show that this approach is computationally intensive and time consuming. The second method involves simulating the quadratic form

$$z'\widehat{\Omega}^{-1}z, \; z \sim N(0, \widehat{\Sigma})$$

and computing the quantiles. This method delivers satisfactory results and reduces the simulation time significantly. An alternative (and popular) procedure of obtaining the critical values, based on the numerical inversion of the characteristic function, was proposed by Imhof (1961). This procedure is much more computationally intensive, since it requires the knowledge of all eigenvalues of $\Sigma\Omega^{-1}$.

A final method is based on using an approximation to the distribution of a weighted sum of chi-square variables. Even though a linear combination of independent chi-squared variables is, under regularity conditions, known to be asymptotically normally distributed when the sample size tends to ∞ (Johnson, Kotz and Balakrishnan (1994), p.444), the simulations reveal the unsatisfactory performance of the normal approximation. Hence, we suggest applying the approximation proposed by Hall (1983) and further explored by Buckley and Eagleson (1988), where the distribution of a weighted sum of χ_1^2 random variables is approximated by the distribution of $\tilde{R} = a\chi_v^2 + b$ by choosing (a, b, v) so that the first three cumulants of R and $\tilde{R}$ agree.

The cumulants κ_l of a random variable are defined via the cumulant-generating function $K(t)$, which is the logarithm of the characteristic function $\phi(t)$ with the following expansion (Muirhead (1982), p.40)

$$K(t) = \log(\phi(t)) = \sum_{l=1}^{\infty} \kappa_l \frac{(it)^l}{l!}.$$

Since the characteristic function $\phi(t)$ of a chi-square random variable with r degrees of freedom is

$$\phi(t) = (1 - 2it)^{-\frac{r}{2}},$$

the cumulant generating function $K(t)$ of $\chi^2_{(r)}$ variable is

$$K(t) = -\frac{r}{2} \log(1 - 2it) = \frac{1}{2} r \sum_{l=1}^{\infty} \frac{(2it)^l}{l}$$

and the cumulants κ_l solve

$$\sum_{l=1}^{\infty} \kappa_l \frac{(it)^l}{l!} = \frac{1}{2} r \sum_{l=1}^{\infty} \frac{(2it)^l}{l}.$$

Let $R = \sum_{j=1}^{J-1} \omega_j \chi_j^2(1)$. The cumulants of this chi-squared-type mixture are given by[7]

$$\kappa_l(R) = 2^{l-1}(l-1)! \sum_{j=1}^{J-1} \omega_j^l.$$

Therefore, the first three cumulants of R are

$$\begin{aligned}
\kappa_1(R) &= E(R) = \sum_{j=1}^{J-1} \omega_j = trace(\Sigma\Omega^{-1}) \\
\kappa_2(R) &= Var(R) = 2\sum_{j=1}^{J-1} \omega_j^2 = 2trace\left(\left(\Sigma\Omega^{-1}\right)^2\right) \\
\kappa_3(R) &= E\left((R-E(R))^3\right) = 8\sum_{j=1}^{J-1} \omega_j^3 = 8trace\left(\left(\Sigma\Omega^{-1}\right)^3\right).
\end{aligned}$$

The cumulants of $\tilde{R} = a\chi_v^2 + b$ are:

$$\kappa_1(\tilde{R}) = av + b,\ \ \kappa_2(\tilde{R}) = 2a^2v,\ \ \kappa_3(\tilde{R}) = 8a^3v.$$

To determine the parameters a, b and v we set $\kappa_m(\tilde{R}) = \kappa_m(R)$ for $m = 1, 2, 3$ which leads to

$$\begin{aligned}
a &= \frac{\kappa_3(R)}{4\kappa_2(R)} \qquad (29) \\
b &= \kappa_1(R) - \frac{2\kappa_2^2(R)}{\kappa_3(R)} \\
v &= \frac{8\kappa_2^3(R)}{\kappa_3^2(R)}.
\end{aligned}$$

Hence the approximate cumulative distribution of R is

$$F_R(t) = \Pr(R \le t) \approx \Pr(\tilde{R} \le t) = \Pr\left(\chi_v^2 \le \frac{t-b}{a}\right).$$

The critical value c_α solves

$$1 - \Pr\left(\chi_v^2 \le \frac{c_\alpha - b}{a}\right) = \alpha$$

for $\alpha = 1\%$, 5% or 10%.

[7] See Severini (2005), Theorem 8.5, p. 245.

Note that parameter v is typically not an integer and the χ^2_v distribution here is in fact a gamma distribution with parameters $\frac{1}{2}$ and $\frac{v}{2}$. In practice, the matrix $\Sigma\Omega^{-1}$ is unknown and, in order to calculate the values of parameters in (29), it has to be replaced by its consistent estimate:

$$C_J'L_Z'M_{L_X}L_ZC_J\left[C_J'L_Z'(P_{L_X}-P_{l_n})L_ZC_J\right]^{-1}.$$

3.3 Monte Carlo simulations

In this section, we discuss the results of Monte Carlo simulations designed to examine the finite sample size and power properties of the proposed tests. We modify Blundell and Horowitz's (2007) setup by generating X and Z as discrete random variables.

3.3.1 Simulation design

In the experiments, realizations of (X, Z) are generated as $Z = Binomial(J-1, q)$ with $q = 0.5$ and X is a function of Z such that

$$X = x_k \text{ if } a_k < X^* \leq b_k, \tag{30}$$

where a_k and b_k are fixed for fixed K, and

$$X^* = \psi Z + (1-\psi^2)^{1/2}v$$

with $v \sim N(0,1)$ and $\psi \in \{0.35, 0.7\}$. That is, we partition the real line into segments and assign $X = x_k$ if X^* is in the interval $(a_k, b_k]$. The choice of x_k values is irrelevant. The selected values for a_k and b_k are presented in Table 1.

	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	a_5	b_5	a_6	b_6
K=2	$-\infty$	0	0	$+\infty$								
K=3	$-\infty$	-0.5	-0.5	0.5	0.5	$+\infty$						
K=5	$-\infty$	-0.25	-0.25	0	0	0.25	0.25	0.5	0.5	$+\infty$		
K=6	$-\infty$	-0.5	-0.5	-0.25	-0.25	0	0	0.5	0.5	1	1	$+\infty$

Table 1: The choice of cutoff points in the simulation design

Note that ψ measures the strength of the relationship between X and Z. Weak instruments are characterized by $\psi = 0.35$ and $\psi = 0.7$ characterizes strong instruments. Under this data generating process, probability masses on each support point x_k of X are given in Tables 2 and 3.

		$\psi = 0.35$					
		p_1	p_2	p_3	p_4	p_5	p_6
K=2	J=2	0.4272	0.5728				
	J=3	0.3590	0.6410				
	J=4	0.2971	0.7029				
K=3	J=3	0.1903	0.3712	0.4385			
	J=4	0.1490	0.3409	0.5101			
	J=5	0.1152	0.3056	0.5792			
K=5	J=2	0.3279	0.0993	0.1042	0.1021	0.3666	
	J=3	0.2680	0.0911	0.0999	0.1026	0.4385	
	J=4	0.2157	0.0813	0.0930	0.0998	0.5101	
K=6	J=3	0.1903	0.0777	0.0911	0.2025	0.1873	0.2513
	J=4	0.1490	0.0667	0.0813	0.1928	0.1953	0.3149
	J=5	0.1152	0.0560	0.0709	0.1787	0.1970	0.3822

Table 2: The probability masses of X with weak instruments

		$\psi = 0.7$					
		p_1	p_2	p_3	p_4	p_5	p_6
K=2	J=2	0.3317	0.6683				
	J=3	0.2130	0.7870				
	J=4	0.1334	0.8666				
K=3	J=3	0.0847	0.3256	0.5897			
	J=4	0.0491	0.2323	0.7186			
	J=5	0.0282	0.1587	0.8131			
K=5	J=2	0.2274	0.1043	0.1188	0.1233	0.4261	
	J=3	0.1393	0.0737	0.0918	0.1055	0.5897	
	J=4	0.0838	0.0496	0.0661	0.0819	0.7186	
K=6	J=3	0.0847	0.0546	0.0737	0.1973	0.2228	0.3669
	J=4	0.0491	0.0346	0.0496	0.1480	0.1977	0.5209
	J=5	0.0282	0.0214	0.0322	0.1050	0.1599	0.6532

Table 3: The probability masses of X with strong instruments

In order to show how these were obtained, we provide

Example 3.1 *Let $K = 3$, $J = 2$. Since Z is $Bin(J-1, q)$, it follows that $q_1 = \Pr[Z = 0] = 1 - q$, $q_2 = \Pr[Z = 1] = q$ and*

$$D_Z = \begin{bmatrix} 1-q & 0 \\ 0 & q \end{bmatrix}.$$

Knowing the data generating process for X, we obtain:

$$\begin{aligned}
p_1 &= \Pr[X=0] = (1-q)\Phi\left(\frac{-0.5}{(1-\psi^2)^{1/2}}\right) + q\Phi\left(\frac{-0.5-\psi}{(1-\psi^2)^{1/2}}\right) \\
p_2 &= \Pr[X=1] = (1-q)\left(\Phi\left(\frac{0.5}{(1-\psi^2)^{1/2}}\right) - \Phi\left(\frac{-0.5}{(1-\psi^2)^{1/2}}\right)\right) + \\
&\quad q\left(\Phi\left(\frac{0.5-\psi}{(1-\psi^2)^{1/2}}\right) - \Phi\left(\frac{-0.5-\psi}{(1-\psi^2)^{1/2}}\right)\right) \\
p_3 &= \Pr[X=2] = (1-q)\left(\Phi\left(\frac{-0.5}{(1-\psi^2)^{1/2}}\right)\right) + q\Phi\left(\frac{-0.5+\psi}{(1-\psi^2)^{1/2}}\right),
\end{aligned}$$

where Φ is the cumulative standard normal distribution function. The matrix of *joint probabilities P can be generated according to*

$$(P)_{jk} = \Pr[Z = z_j, X = x_k] = \Pr[X = x_k | Z = z_j] \Pr[Z = z_j],$$

for example

$$\begin{aligned}(P)_{11} &= \Pr[Z = z_1, X = x_1] = \Pr[Z = 0, X = 0] \\ &= \Pr[X = 0 | Z = 0] \Pr[Z = 0] = \Phi\left(\frac{-0.5}{(1-\psi^2)^{1/2}}\right)(1-q)\end{aligned}$$

and

$$\begin{aligned}(P)_{21} &= \Pr[Z = z_2, X = x_1] = \Pr[Z = 1, X = 0] \\ &= \Pr[X = 0 | Z = 1] \Pr[Z = 1] = \Phi\left(\frac{-0.5-\psi}{(1-\psi^2)^{1/2}}\right) q.\end{aligned}$$

The realizations of a continuous outcome Y are generated from

$$Y = \theta_0 + \theta_1 X + \varepsilon,$$

where $\varepsilon = \sigma_\varepsilon^2 \left(\eta v + (1-\eta^2)^{\frac{1}{2}} u\right)$ with $u \sim N(0,1)$ and $\theta_0 = 0$, $\theta_1 = 0.5$ and $\sigma_\varepsilon = 0.2$. The parameter η measures the strength of the relationship between X and ε, and its value varies across experiments. The null hypothesis is true if $\eta = 0$ and false otherwise. The experiments use sample sizes of $n = 100, 200, 400$ and 1000 (for the power analysis) observations and there are 2000 Monte Carlo replications in each experiment.

3.3.2 Size analysis $J \geq K$

Recall that under the null hypothesis $T_n \rightarrow^d \chi^2_{K-1}$, so the critical values are easily obtained from statistical tables. The empirical size of the proposed test for different combinations of J and K (satisfying $J \geq K$) is presented in Tables 4 and 5.

K=2		J=2			J=3			J=4		
ψ	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	100	0.90	4.95	10.05	1.00	4.95	9.35	0.95	5.35	10.50
	200	1.30	5.10	10.20	1.10	5.05	10.45	1.10	5.10	11.20
	400	0.85	5.20	10.20	1.10	5.05	10.45	1.10	5.10	10.25
0.7	100	0.80	4.50	9.55	0.85	4.95	10.10	1.25	4.95	9.65
	200	1.10	4.55	10.80	1.25	4.85	10.10	0.95	5.40	9.70
	400	1.25	5.20	10.10	0.95	4.95	10.50	1.10	5.15	10.40

Table 4: Proportion of rejections under the null hypothesis; K=2

K=3		J=3			J=4			J=5		
ψ	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	100	0.95	5.35	10.75	0.85	5.35	11.20	1.05	5.45	10.30
	200	0.80	4.75	10.10	1.25	5.25	10.95	1.25	5.50	10.45
	400	0.85	4.85	10.25	1.10	5.05	9.85	1.20	5.10	9.55
0.7	100	1.05	5.90	11.20	1.15	5.15	10.70	0.95	5.15	10.30
	200	1.00	5.40	10.65	1.25	5.05	10.35	1.10	5.75	10.70
	400	1.45	5.80	10.65	0.95	5.45	10.55	1.05	5.10	9.65

Table 5: Proportion of rejections under the null hypothesis; K=3

It is possible to construct simulation-based confidence intervals for the true size. For 2000 simulations, the 5% confidence interval for $\alpha = 1\%$ is $[0.56, 1.44]$, for $\alpha = 5\% : [4.04, 5.96]$ and for $\alpha = 10\% : [8.69, 11.31]$. As can be seen, all empirical sizes are within the calculated bounds, and hence reasonably close to the nominal values of 1%, 5% and 10%, even in small samples of 100 observations. The size seems not very sensitive to changes in the number of points of support of the endogenous regressor and instrument and do not vary with the strength of instrument.

3.3.3 Power analysis $J \geq K$

For the power analysis, the errors are generated as $\varepsilon = \sigma_\varepsilon^2 \left(\eta v + (1-\eta^2)^{\frac{1}{2}} u\right)$, $u \sim N(0,1)$. Recall that this specification excludes alternatives with $E[\varepsilon|X = x_k] = \gamma(x_k)$ in which the power is equal to the size of the test. The results of power analysis at 5% significance level for different sample sizes, but $K = J$, are summarized in Figures 1 and 2. Detailed empirical power results are included in the Appendix B.

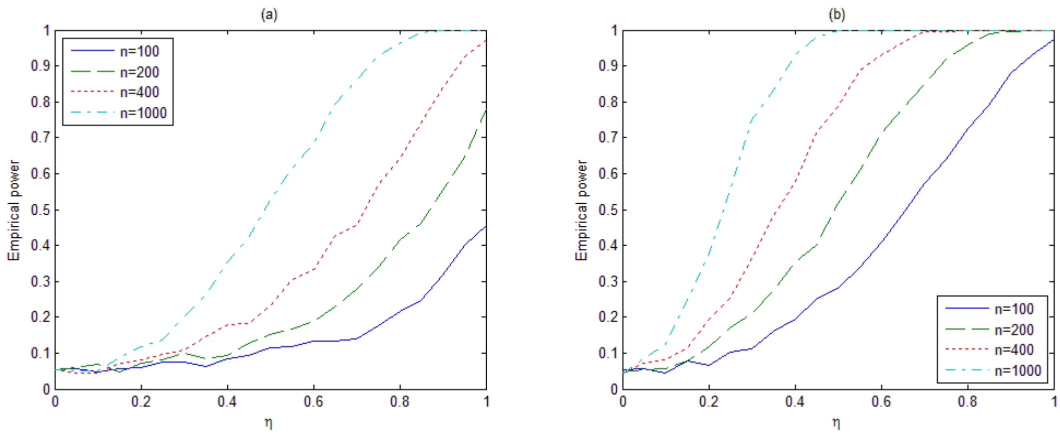

Figure 1: Empirical power for K=2 and J=2 with weak (a) and strong (b) instruments

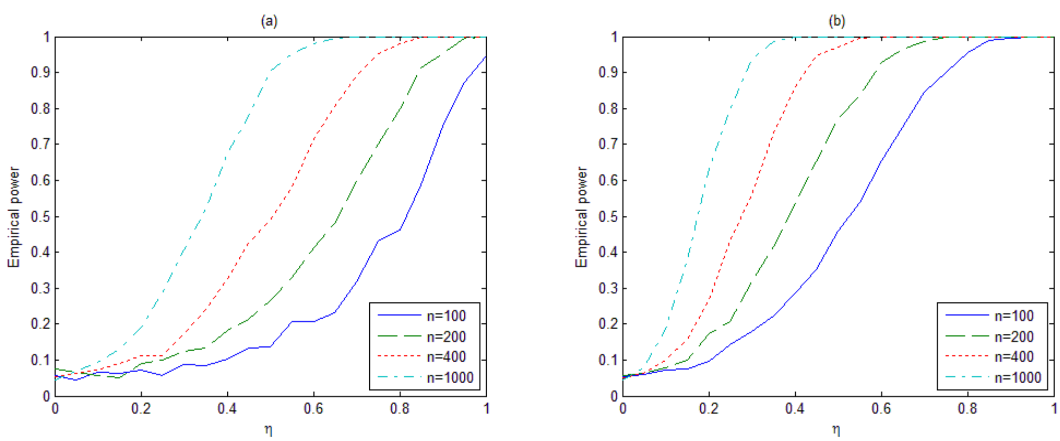

Figure 2: Empirical power for K=3 and J=3 with weak (a) and strong (b) instruments

Figures 3 and 4 show how the empirical power changes with the number of points of support of the instrument.

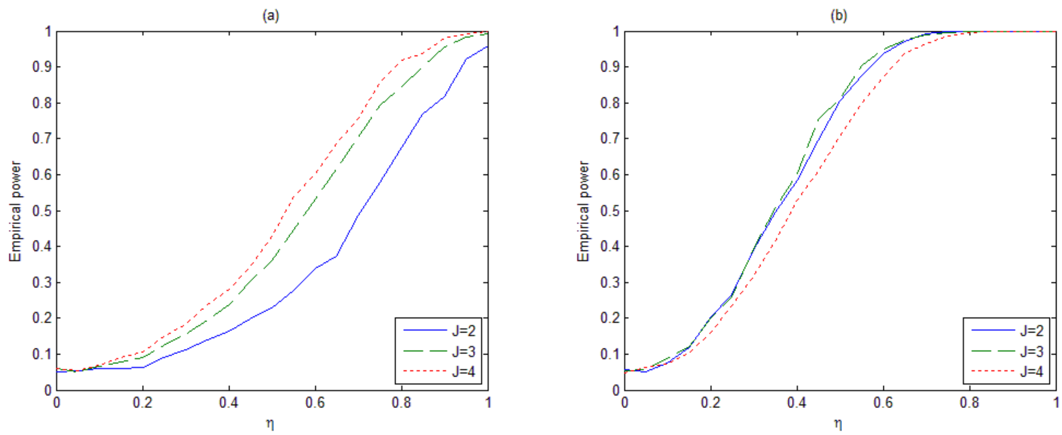

Figure 3: Empirical power for K=2 and n=400 with weak (a) and strong (b) instruments

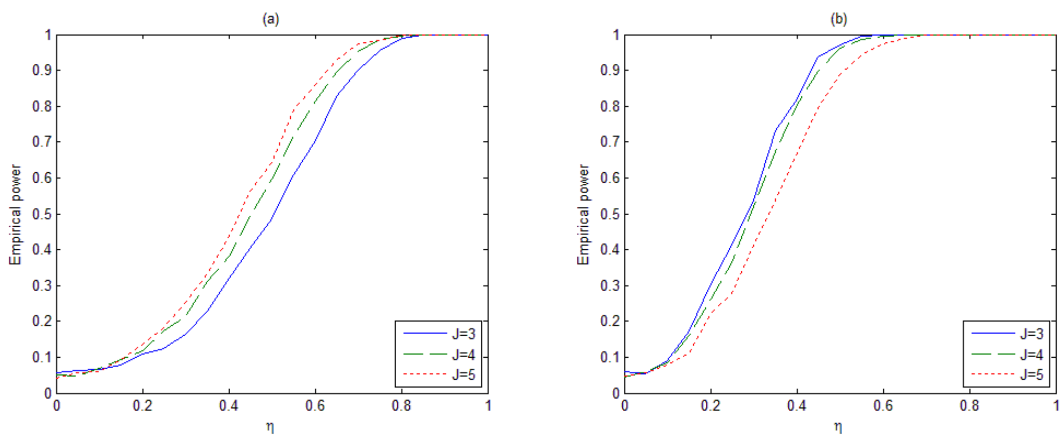

Figure 4: Empirical power for K=3 and n=400 with weak (a) and strong (b) instruments

The proposed test exhibits satisfactory power properties. The test is apparently unbiased, and its empirical power increases with the sample size and converges to 1 quickly. For a fixed number of support points of the endogenous regressor and instrument, the empirical power is higher if the instrument used in experiment is strong. The test has also higher power if the support of endogenous regressor is larger. If the instrument is weak, for fixed K, the empirical power of the test increases when additional points of support of the instrument are added. Therefore, for weak instruments, the larger the support of Z, the more powerful the test is. This suggests that in practice the researcher should look for an instrument with many support points to increase the probability of detecting the endogeneity of regressor. On the other hand, if the instrument is strong, the empirical power remains roughly the same if the difference between the numbers of support points of X and Z is small, but decreases slightly with the gap between J and K.

This unusual power function behaviour is due to our simulation design. Note that as J increases, the probability masses p_k become more and more unequal (see Tables 2 and 3). The problem is more apparent in the case of strong instruments. For example, in the case of 3 support points of X and 5 support points of Z, $\Pr[X = x_1] = 0.0282$. Given that the sample size considered here is $n = 400$, there are only a few sample points for which we observe the value x_1. As a result, the precision in computing the value of the test statistic is substantially reduced. In order to show that the test performs better if there is a similar number of observations for each support point x_k, in an additional small experiment we change the simulation design such that all $p_k = \frac{1}{K}$. The new cutoff points of (30) are given in Table 6.

		ψ= 0.35					
		a_1	b_1	a_2	b_2	a_3	b_3
K=2	J=2	$-\infty$	0.18	0.18	$+\infty$		
	J=3	$-\infty$	0.35	0.35	$+\infty$		
	J=4	$-\infty$	0.53	0.53	$+\infty$		
K=3	J=3	$-\infty$	-0.07	-0.07	0.77	0.77	$+\infty$
	J=4	$-\infty$	0.1	0.1	0.95	0.95	$+\infty$
	J=5	$-\infty$	0.27	0.27	1.13	1.13	$+\infty$
		ψ= 0.7					
K=2	J=2	$-\infty$	0.35	0.35	$+\infty$		
	J=3	$-\infty$	0.7	0.7	$+\infty$		
	J=4	$-\infty$	1.05	1.05	$+\infty$		
K=3	J=3	$-\infty$	0.32	0.32	1.08	1.08	$+\infty$
	J=4	$-\infty$	0.64	0.64	1.46	1.46	$+\infty$
	J=5	$-\infty$	0.96	0.96	1.84	1.84	$+\infty$

Table 6: The choice of cutoff points for equal point masses on the support points of X; K=2 and K=3

Figures 5 and 6 show how the empirical power changes with the number of points of support of the strong instrument under the new data generating process.

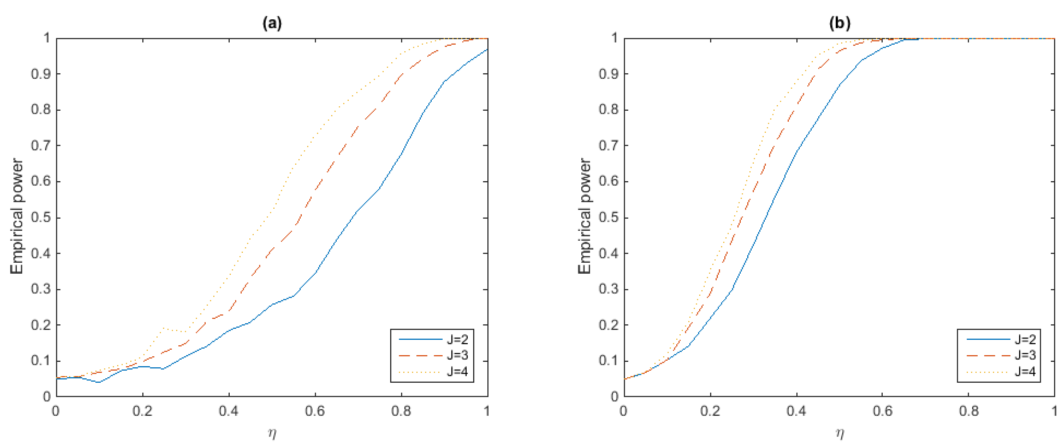

Figure 5: Empirical power for K=2 and n=400 with weak (a) and strong (b) instruments with equal point masses on the support points of X

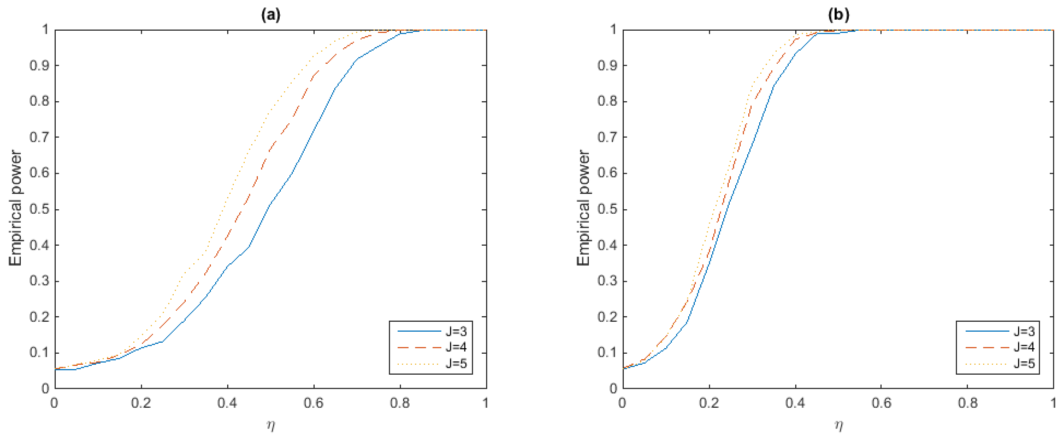

Figure 6: Empirical power for K=3 and n=400 with weak (a) and strong (b) instruments with equal point masses on the support points of X

The empirical power increases now with the number of the support points in the instrument. It is clear that the empirical power properties are improved, while the size is unaffected. The increase in power can also be noticed in the case of weak instruments, but the difference is smaller in magnitude, since with the original data generating process, p_k were closer to $\frac{1}{K}$ with weak instruments.

3.3.4 Size analysis $J < K$

We have experimented with different methods of computing the critical values for the proposed test. The three methods proposed in Section 3.2.5 produce very similar results for the empirical size and power of the test. In this section, we present the results based on the chi-square approximation, which minimizes the computational time. The empirical size of the proposed test is presented in Tables 7 and 8.

K=5		J=2			J=3			J=4		
ψ	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	100	1.10	5.00	10.45	1.20	5.65	10.85	0.85	5.25	10.80
	200	0.90	4.85	10.00	0.80	4.55	9.65	0.95	4.55	9.65
	400	0.85	5.50	10.50	1.15	6.15	10.55	1.10	5.50	9.75
0.7	100	1.20	6.10	11.50	1.35	5.85	11.35	1.55	5.80	11.20
	200	1.15	5.80	10.90	1.05	4.75	9.70	1.10	5.35	10.35
	400	1.20	5.65	9.80	1.05	5.30	10.10	0.95	4.95	10.50

Table 7: Proportion of rejections under the null hypothesis; K=5

K=6		J=3			J=4			J=5		
ψ	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	100	1.20	5.70	10.50	0.85	5.25	10.95	1.10	6.15	11.50
	200	1.05	4.80	10.15	1.00	5.60	11.10	1.50	5.55	10.40
	400	0.95	4.65	9.80	0.90	5.05	9.85	0.90	5.25	10.50
0.7	100	1.05	5.35	10.55	0.95	4.75	10.30	1.60	5.70	10.15
	200	1.15	5.10	9.85	0.85	5.05	9.90	1.65	5.90	11.20
	400	0.90	5.40	10.75	1.05	5.65	11.10	0.95	5.20	10.40

Table 8: Proportion of rejections under the null hypothesis; K=6

.

The test has adequate size in all cases, even in the small samples of 100 observations. The size is not sensitive to changes in the number of points of support and the strength of the relationship between endogenous regressor and the instrument.

3.3.5 Power analysis $J < K$

The results of a power analysis at 5% significance level are presented in Figures 7 and 8. Detailed empirical power results are included in the Appendix B.

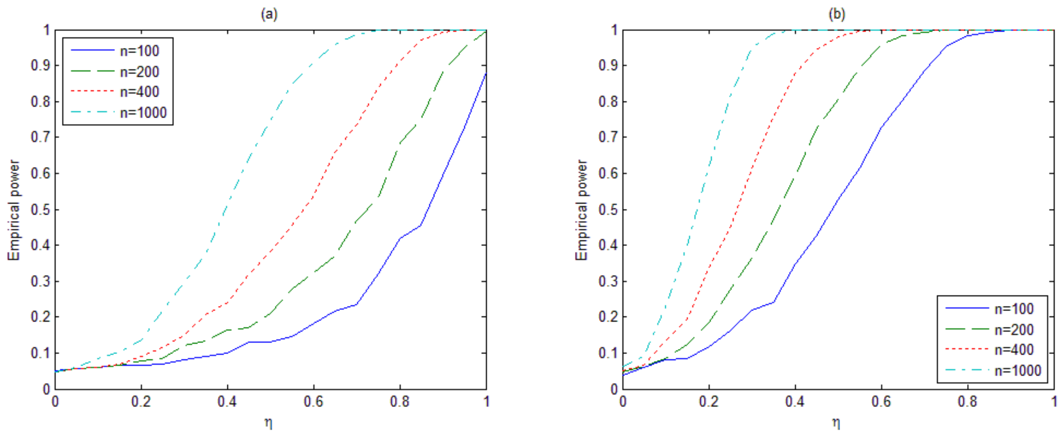

Figure 7: Empirical power for K=5 and J=2 with weak (a) and strong (b) instruments

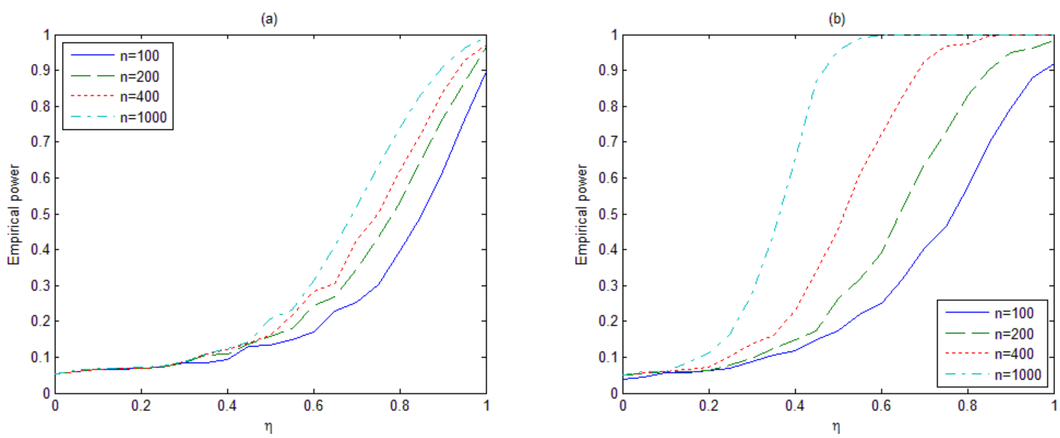

Figure 8: Empirical power for K=6 and J=3 with weak (a) and strong (b) instruments

Figures 9 and 10 show how the empirical power changes with the number of points of support of the instrument.

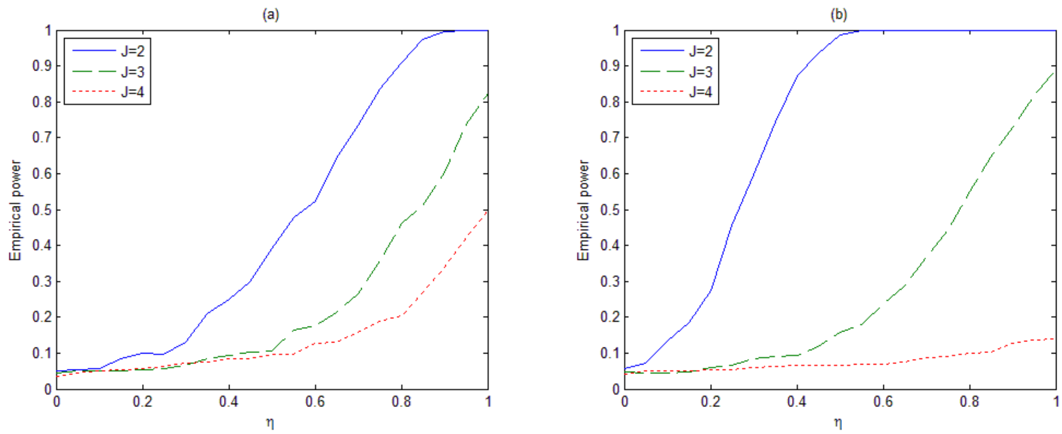

Figure 9: Empirical power for K=5 and n=400 with weak (a) and strong (b) instruments

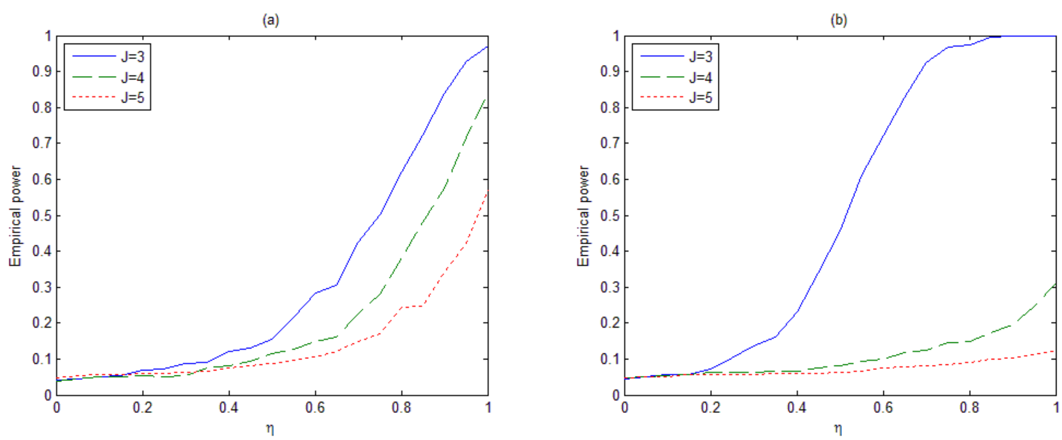

Figure 10: Empirical power for K=6 and n=400 with weak (a) and strong (b) instruments

The empirical power increases with the sample size and in some cases (strong instruments) converges quickly to 1. The test is again apparently unbiased, and performs particularly well if the instruments are strong. For a fixed number of points of support of the regressor, the proposed test detects endogeneity of the regressor better when the support of the instrument is smaller. Hence, for both weak and strong instruments, the power of the test is decreasing with the number of points of support of the instrument. This suggests that in applications, in order to obtain higher power in detecting endogeneity, among all the instruments available, that with the smallest number of support points should be chosen.

Note that if the gap between K and J is small, the test tends to be more powerful with weak instruments. This counter intuitive behaviour of the power function might be again due to simulation design. Allowing for equal point masses on the support points of X requires new cutoff points of (30), which are given by

		$\psi = 0.35$										
		b_1	a_2	b_2	a_3	b_3	a_4	b_4	a_5	b_5	a_6	b_6
K=5	J=2	−0.63	−0.63	−0.07	−0.07	0.42	0.42	0.98	0.98	$+\infty$		
	J=3	−0.47	−0.47	0.11	0.11	0.6	0.6	1.17	1.17	$+\infty$		
	J=4	−0.30	−0.30	0.28	0.28	0.78	0.78	1.36	1.36	$+\infty$		
K=6	J=3	−0.59	−0.59	−0.07	−0.07	0.35	0.35	0.77	0.77	1.29	1.29	$+\infty$
	J=4	−0.43	−0.43	0.10	0.10	0.53	0.53	0.95	0.95	1.48	1.48	$+\infty$
	J=5	−0.27	−0.27	0.27	0.27	0.7	0.7	1.13	1.13	1.67	1.67	$+\infty$
		$\psi = 0.7$										
K=5	J=2	−0.33	−0.33	0.15	0.15	0.56	0.56	1.03	1.03	$+\infty$		
	J=3	−0.04	−0.04	0.48	0.48	0.92	0.92	1.44	1.44	$+\infty$		
	J=4	0.25	0.25	0.81	0.81	1.29	1.29	1.85	1.85	$+\infty$		
K=6	J=3	−0.15	−0.15	0.32	0.32	0.7	0.7	1.08	1.08	1.55	1.55	$+\infty$
	J=4	0.13	0.13	0.64	0.64	1.05	1.05	1.46	1.46	1.97	1.97	$+\infty$
	J=5	0.42	0.42	0.96	0.96	1.4	1.4	1.84	1.84	2.38	2.38	$+\infty$

Table 9: The choice of cutoff points for equal point masses on the support points of X; K=5 and K=6

with $a_1 = -\infty$. Figures 11 and 12 show how the empirical power changes with the number of points of support of the strong instrument under new data generating process.

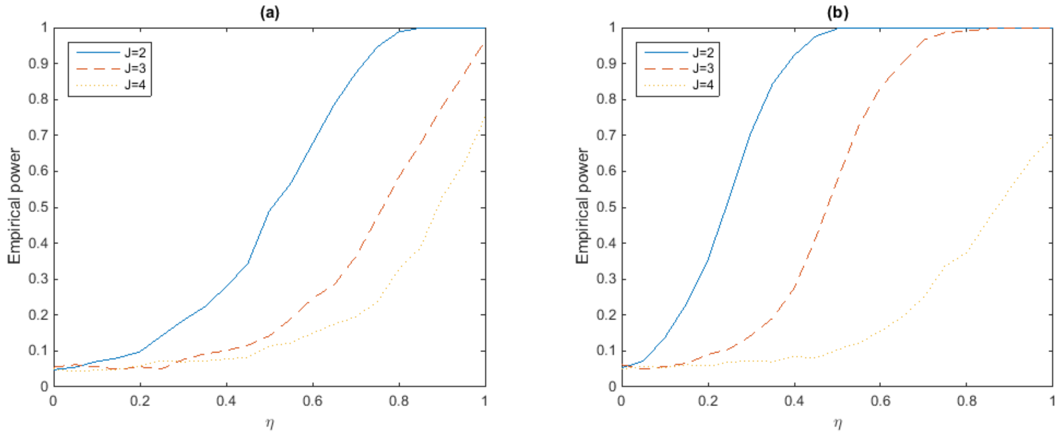

Figure 11: Empirical power for K=5 and n=400 with weak (a) and strong (b) instruments with equal point masses on the support points of X

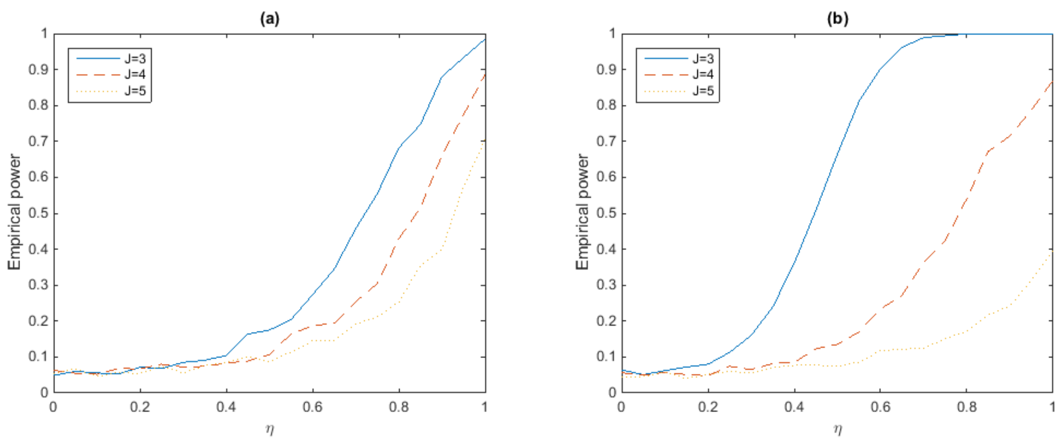

Figure 12: Empirical power for K=6 and n=400 with weak (a) and strong (b) instruments with equal point masses on the support points of X

The power properties are significantly improved and the size is unaffected. We can again conclude that the test performs better in samples with similar number of observations for each support point of the endogenous regressor. In general, these results are more than satisfactory, particularly so in view of the fact that the model is only partially identified under the alternative hypothesis, and its simplicity.[8]

[8]Some recently developed testing procedures for partially identified models are very complicated, and are useful only for a limited range of hypotheses.

3.4 Concluding remarks

The consistency of standard estimation procedures fails in the presence of endogeneity in the model. Therefore, in order to choose the consistent estimation technique, the applied researchers should test whether the explanatory variables used in the model are exogenous. In this chapter, we have proposed a test for exogeneity of regressors in nonparametric models with discrete explanatory variables and discrete instruments under the assumption that the model point identifies the unknown structure of interest. Using the linear interpretation of a nonparametric model, the test is built on a quadratic form of a difference between two estimators, one of which is consistent only under exogeneity and the other is consistent under both scenarios. This testing framework follows closely the Durbin-Wu-Hausman-type of test. It has been shown that under the null hypothesis of exogeneity, the test statistic follows chi-square distribution asymptotically. The consistency of the test has been established by showing that under the alternative hypothesis the test-statistic follows a noncentral chi-square distribution and that the asymptotic power of the test equals 1. The results of Monte Carlo simulations have shown satisfactory finite-sample properties of the proposed test. Based on our experiments, we can conclude that:

- both tests have correct size even in small samples, and are unbiased,
- empirical power increases with the sample size and converges to 1,
- using a strong instrument leads to better power properties,
- both tests are more powerful if the sample is balanced, i.e. there is an equal number of observations for each support point,
- empirical power changes with the number of support points of both endogenous regressor and instrument.

Particularly interesting is the fact that the power increases with the gap between the number of points of support in the variables. Therefore, assuming that there is a choice between valid instruments for the applied researcher, when $J \geq K$, one should choose that with the most points of support, and when $J < K$ choose the one with the smallest number of support points in order to increase the probability of detecting endogeneity of the regressor.

4 Generalizations

This section provides extensions of the testing procedures to models that are more realistic for practical applications. Firstly, we discuss exogeneity testing in the presence of an additional exogenous regressor, but still with a single instrument. Then we show that general results for testing the exogeneity of multiple regressors using multiple instruments in models with multiple exogenous explanatory variables have exactly the same structure, so that the tests are easily generalized.

4.1 Models with two discrete regressors

In most applications, it is reasonable to let the unknown $h(\cdot)$ be a function of more than one regressor. In this section we extend the model given by equations (1) and (3) to allow for a second explanatory variable.

4.1.1 Setup

We assume that the additional regressor is definitely exogenous, and write the extended model as

$$Y = h(W, X) + \varepsilon, \tag{31}$$

where Y is a scalar continuous dependent variable, X is a single discrete regressor that may be endogenous, and W denotes a discrete regressor whose exogeneity is not in question. We assume that W has D points of support.

In addition to equation (31), we assume that there exists a single discrete instrumental variable Z such that

$$E\left[\varepsilon|W = w_d, Z = z_j\right] = 0,\ \forall d, j \tag{32}$$

which generalizes (3) and represents the instrument exogeneity condition. Therefore, the extended model consists of equations (31) and (32). The data $(y_i^s, x_i^s, z_i^s, w_i^s)$ consists of n $i.i.d.$ observations on (Y, X, Z, W).

We are interested in testing the null hypothesis of exogeneity of the regressor X, i.e. $E[\varepsilon|W = w_d, X = x_k] = 0$ for all d and k. In terms of observables, the null hypothesis is

$$H_0 : E[Y|W = w_d, X = x_k] = h(w_d, x_k). \tag{33}$$

Under the null, the unknown function $h(\cdot)$ can be consistently estimated using standard nonparametric techniques. If condition (33) is violated, the choice of the consistent estimation method depends on the identification regime.

In order to generalize the results of the previous sections, we modify the notation by defining β to be the lexicographically ordered DK-vector with D blocks of K elements $h(w_d, x_k)$. In each block, w_d is fixed and x_k varies from x_1 to x_K, i.e.

$$\beta = \begin{pmatrix} h(w_1, x_1) \\ ... \\ h(w_1, x_K) \\ h(w_2, x_1) \\ ... \\ h(w_2, x_K) \\ ... \\ h(w_D, x_K) \end{pmatrix}.$$

Similarly, let π be the DJ-vector with D blocks of conditional means $E[Y|W = w_d, Z = z_j]$, and let $\Pi = \text{diag}\{\Pi_d, d = 1, .., D\}$ be the $DJ \times DK$ matrix in which

$$\Pi_d(j, k) = \Pr[X = x_k | Z = z_j, W = w_d].$$

As in the previous case, the restriction that the null hypothesis imposes on the conditional density of Y given (X, W) is that the vector of conditional means $E[Y|W = w_d, X = x_k]$ is a solution to the set of DJ linear equations

$$\pi = \Pi\beta \tag{34}$$

in the unknown β. (34) has a unique solution if and only if $rank(\Pi) = DK$, which requires $DJ \geq DK$, i.e., $J \geq K$. Therefore, the nonparametric identification condition remains the same as in the model without the additional regressor. Hence, allowing for additional exogenous explanatory variable does not affect the identifying power of the instrument.[9]

[9] This is intuitive: β can be partitioned into D sub-vectors with $K \times 1$ components $h_d(x_k)$. Therefore, we have D problems of the same type as in the model without W. For h to be point identified, each h_d must be, so the identification condition is unchanged.

4.1.2 Linear model representation

Let $n_{dk}^{WX} = \sum_{i=1}^{n} I(w_i^s = w_d) I(x_i^s = x_k)$ and $n_{dj}^{WZ} = \sum_{i=1}^{n} I(w_i^s = w_d) I(z_i^s = z_j)$ denote the multiplicities of pairs (w_d, x_k) and (w_d, z_j) in the sample. Note that

$$n_{d\cdot} = \sum_{k=1}^{K} n_{dk}^{WX} = \sum_{j=1}^{J} n_{dj}^{WZ}$$

represents the sample multiplicity of w_d.

The compact linear model representation of the above setup requires defining $n \times DK$ matrix L_{WX}, which is build from DK blocks, denoted by L_X^d, of dimensions $n_{d\cdot} \times K$. The observations are ordered according to the values of W, i.e. the upper left corner of L_{WX} consists of $n_{1\cdot}$ rows, which correspond to $n_{1\cdot}$ observations in which $W = w_1$. The elements of each block L_X^d are the indicators $I(w_i^s = w_d) I(x_i^s = x_k)$. Therefore, the matrix L_{WX} can be partitioned as

$$L_{WX} = \begin{bmatrix} L_X^1 & 0 & \dots & 0 \\ 0 & L_X^2 & & \dots \\ \dots & & .. & 0 \\ 0 & \dots & 0 & L_X^D \end{bmatrix}.$$

Likewise, define $n \times DJ$ matrix L_{WZ}, which is build from DJ blocks, denoted by L_Z^d, of dimensions $n_{d\cdot} \times J$. Both L_{WX} and L_{WZ} are random matrices with the row sums equal to 1. The linear model representation is therefore given by

$$y = L_{WX}\beta + \varepsilon_n,$$

where β can be consistently estimated by OLS under exogeneity. If X is endogenous and the point identification condition is satisfied, β can be consistently estimated by IV using the matrix L_{WZ} as instrument.

Note that there always exists a permutation of sample observations on Y, consistent with the construction of L_{WX} and L_{WZ}. Additionally, this permutation of the data is common to both L_{WX} and L_{WZ}, since the observations are ordered according to the values of W. It is assumed that all combinations of K support points of X, J support points of Z and D support points of W occur in the sample. That is, there is at least one observation (preferably more) in which $X = x_k, Z = z_j$ and $W = w_d$ for all k, j, d.

As before, there is a relationship between L_{WX} and L_{WZ}, that leads to rank deficiency of matrices appearing in the test-statistics. In models with an exogenous covariate, there are $D+1$ common vectors in the column spaces of L_{WX} and L_{WZ}. Since the row sums of both L_{WX} and L_{WZ} are equal to one,

$$L_{WX} l_{LK} = L_{WZ} l_{LJ} = l_n.$$

Additional to this relation, which has already been discussed in the simple model, note that for each block L_X^d of L_{WX} and L_Z^d of L_{WZ} we have

$$L_X^d l_K = L_Z^d l_J = l_{n_{d.}}.$$

Let m_d^K denote the DK-vector of D blocks with l_K placed as d^{th} block, 0 elsewhere. Similarly, let m_d^J be the DJ-vector with l_J as the d^{th} block and m_d^n be n-vector of $n_{d.}$ blocks with $l_{n_{d.}}$ as $n_{d.}^{th}$ block. The above relation implies that

$$L_{WX} m_d^K = L_{WZ} m_d^J = m_d^n,$$

for all d. Since there are D possible positions of l_K and l_J in m, there are D vectors common to column spaces of L_{WX} and L_{WZ}. Even though at first sight, there are $D+1$ relationships between L_{WX} and L_{WZ}, note that the vectors $m_1^n, ..., m_D^n$ add up to l_n. Therefore, there are only D linearly independent common vectors.

In order to deal with the rank deficiency, we introduce $DK \times (DK - D)$ matrix C_{DK}, which is orthogonal to $[m_1^K, ..., m_D^K]$ and $DJ \times (DJ - D)$ matrix C_{DJ} orthogonal to $[m_1^J, ..., m_D^J]$.

4.1.3 Test statistics in the extended model

Since allowing for an additional exogenous regressor in the model only increases the dimension of the objects used to construct the test-statistics presented in Sections 3.1 and 3.2. Thus, for testing the null hypothesis of exogeneity of X under point identification, i.e. in models with $J \geq K$, the test statistic becomes:

$$T_n = \frac{y' G_{XZ} (G'_{XZ} G_{XZ})^{-1} G_{XZ} y}{n^{-1} y' M_{L_{WX}} y}, \tag{35}$$

where

$$G_{XZ} = M_{L_{WX}} P_{L_{WZ}} L_{WX} C_{DK},$$

and under H_0 it is easy to show that $T_n \to^d \chi^2_{(DK-D)}$. For the case $J < K$, i.e., under partial identification, we have

$$R_n = \frac{y' M_{L_{WX}} L_{WZ} \left(L'_{WZ} P_{L_{WX}} L_{WZ}\right)^{-1} L'_{WZ} M_{L_{WX}} y}{n^{-1} y' M_{L_{WX}} y}, \tag{36}$$

where ω_j are the eigenvalues of $\Sigma\Omega^{-1}$ with, in this case,

$$\Sigma = p \lim_{n\to\infty} n^{-1} [C'_{DJ} L'_{WZ} M_{L_{WX}} L_{WZ} C_{DJ}],$$

and

$$\Omega = p \lim_{n\to\infty} n^{-1} (C'_{DJ} L'_{WZ} (P_{L_{WX}} - P_{\iota_W}) L_{WZ} C_{DJ}).$$

Again, it is easy to adapt the earlier proofs to show that, under H_0,

$$R_n \to^d \sum_{j=1}^{D(J-1)} \omega_j \chi^2_j(1).$$

Here, $C_{DJ} = \text{diag}\{C_{n_d}\}$ is $DJ \times D(J-1)$, and $\iota_W = \text{diag}\{l_{n_d}\}$ is $n \times D$. The asymptotic distributions are stated without proof, but easily follow from the proofs of Theorems 3.1 and 3.3 in the Appendix A.

4.2 Multiple endogenous regressors, multiple exogenous covariates and multiple instruments

In the most general version of the model one would have several X variables to be tested, several additional exogenous variables, Ws, and several instruments. The model would thus be:

$$Y = h(W_1, ..., W_M, X_1, ..., X_R) + \varepsilon,$$

we allow for R potentially endogenous discrete regressors and M exogenous discrete covariates, and we have, say, S discrete instruments $Z_1, ..., Z_S$. Additionally, we can allow each of these variables to have different numbers of support points.

Allowing for multidimensional regressors and instruments does not require any changes in the results presented in the previous section. To see this, let the number of support points of each X_r, W_m and Z_s be denoted by K_r, D_m and J_s, respectively, for all r, m and s. We can denote combinations of support points by sequences

$$\begin{aligned} x_\alpha &= (x_{\alpha_1}, ..., x_{\alpha_R}),\ 1 \le \alpha_r \le K_r, \\ w_\varphi &= \left(w_{\beta_1}, ..., w_{\beta_M}\right),\ 1 \le \varphi_m \le D_m, \\ z_\gamma &= \left(z_{\gamma_1}, ..., z_{\gamma_S}\right),\ 1 \le \gamma_s \le J_s. \end{aligned}$$

There are $K = \Pi_{r=1}^{R} K_r$ sequences α, $D = \Pi_{m=1}^{M} D_m$ sequences φ, and $J = \Pi_{s=1}^{S} J_s$ sequences γ, and the combinations x_α, w_β, and z_γ can be considered to be the support points of single "composite" random variables, and all results for this case therefore follow from the previous subsection, with J, K, and D replaced by appropriate products. A particularly interesting result following from this transformation is the nonparametric identification condition, summarized in:

Proposition 6 *The necessary and sufficient condition for identification in the model $Y = h(W_1, ..., W_M, X_1, ..., X_R) + \varepsilon$, with multiple discrete endogenous regressors $X_1, ..., X_R$, multiple discrete exogenous regressors $W_1, ..., W_M$ and multiple discrete instruments $Z_1, ..., Z_S$, satisfying $E[\varepsilon | W_1, ..., W_M, Z_1, ..., Z_S] = 0$, is that the product of the number of points of support of the instruments is at least as large as the product of the number of points of support of endogenous regressors, i.e., $J = \Pi_{s=1}^{S} J_s \ge K = \Pi_{r=1}^{R} K_r$.*

Remark 4.1 *This result differs from the standard identification condition, which requires that there are at least as many instrumental variables as endogenous regressors. When all variables are discrete, nonparametric identification does not depend on the actual number of instruments, but only on the number of (combined) support points. Hence, theoretically, many endogenous regressors can be instrumented by one variable, assuming that it is correlated with all endogenous regressors and the support of that instrument is large enough. On the other hand, with more instruments the point identification condition $J \ge K$ is more likely to be satisfied.*

5 Empirical applications

This section presents an empirical application of the testing procedures described in previous sections. We are interested in testing whether education is endogenous when estimating the returns to schooling in the standard wage equation using some classic applied work.

The impact of education on earnings is one of the most popular relationships studied in labour economics. Thousands of studies show that better- educated workers earn higher wages. However, this measured wage gap between more and less educated individuals cannot be interpreted as the estimate of the economic return to education. As individuals differ in their personal characteristics, their schooling choices vary and the causal effects of education on wage are difficult to uncover. To explain the earnings differences between workers with different levels of education, researchers typically use models, which builds on Becker (1967) where individuals face the problem of choosing the optimal level of education, associated with a given level of earnings. The optimal choice requires balancing the benefits of higher schooling, captured by the life-time earnings profile, with the cost of education (direct costs, such as tuition fees and indirect costs in the form of foregone earnings while still in education). From the practical point of view, the return to education is typically estimated from the cross-sectional regression of earnings on schooling. Since individuals have different tastes or attitudes towards education, their returns to schooling are different. As highlighted by Card (2001), the issue is that people with higher return to education have an incentive to get more schooling and the typical estimate of the average marginal return to education will be upward biased.

The main problem in the traditional estimation of returns to education is (potential) endogeneity of schooling variable. The most popular sources of correlation between education and the unobserved error term in the wage equation, highlighted in the literature are the "ability bias" and the measurement error. The problem of unobserved omitted ability comes from the fact that more able individuals can earn higher wages at any educational level and are more likely to acquire higher schooling, resulting in upward-biased estimates of return to education. On the other hand, the mismeasurement of schooling might lead to the downward bias in the estimates of the impact of education on earnings (see Griliches (1977)). As schooling is typically self-reported in available datasets and the reliability of self-reporting has been estimated to be 85-90% (Angrist and Krueger (1999)), the resulting bias might be severe.

Given potential endogeneity of education, the study of the causal effects of schooling on earnings requires the use of instrumental variables methods and finding the exogenous source of variation in schooling. The set of instruments typically used in the literature include the geographic proximity of schools (Card (1995) and Kane and Rouse(1993)) and the quarter of birth of an individual (Angrist and Krueger (1991) and Staiger and Stock (1997)). Many studies have also used the institutional features of the schooling system, such as tuition costs (Kane and Rouse (1993)) and changes in the minimum school leaving age (Harmon and Walker (1995)). The common finding appearing in these studies is that instrumental variable estimates of the return to education exceed the analogous OLS estimates, which suggests that the upward bias of the standard OLS estimates caused by omitted ability might be offset by the negative bias due to the measurement error. Another common feature of the above studies is that education is assumed to be endogenous, but it is not tested for endogeneity, even though methods of exogeneity testing in linear models are well known (e.g. Hausman (1978)). Therefore, our interest lies in investigating whether education is truly endogenous.

We discuss two classic models: Card (1995) and Angrist and Krueger (1991), which differ only in the choice of the instrumental variable and the set of exogenous explanatory variables. We consider the nonparametric version of their structural relationship between education and (log) earnings. In both examples, education is instrumented by a variable with the support that is insufficient for nonparametric point identification ($J < K$). The linear specifications estimated in the original studies are identified, because of the parametric restriction imposed. Before providing an inference on the parameters of interest, one should check whether the parametric regression is well-specified. Such testing procedures typically involve comparing the parametric and nonparametric estimates. If education is endogenous, the linear specifications used by Card (1995) and Angrist and Krueger (1991) are not testable, as the nonparametric point estimates do not exist, and hence cannot be compared by their parametric analogues. Therefore, we test for endogeneity of education nonparametrically and given the outcome of the test we propose an adequate estimation method.

5.1 Testing for exogeneity of education in Card (1995)

In Card (1995) education is assumed to be endogenous (due to omitted ability or measurement error) and the following linear model (the standard earnings function)

$$\ln(wage_i) = \beta_0 + \beta_1 Educ_i + \sum_{m=1}^{M} \gamma_m W_{mi} + \varepsilon_i$$

is estimated by Two Stage Least Squares using a binary instrument, which takes value 1 if there is an accredited 4-year public college in the neighborhood (in the "local labour market"), 0 otherwise. It is argued that the presence of a local college decreases the cost of further education (transportation and accommodation costs) and particularly affects the schooling decisions of individuals with poor family backgrounds. The set of exogenous explanatory regressors W includes variables like race, years of potential labour market experience, region of residence and some family background characteristics.

The dataset is available online[10], and consists of 3010 observations from the National Longitudinal Survey of Young Men. Education is measured by the years of completed schooling and varies we between 2 and 18 years. The (sample) support of education consists then of $K = 17$ possible values and the support of instruments is $J = 2$. In order to test for exogeneity of education we use the test-statistic given by (36) in different specifications (in terms of selected exogenous regressors). Note that the more exogenous covariates included, the less likely it is to have enough observations for all possible combinations of support points. For instance, there are no individuals in the sample with 17 years of schooling who are black and do not live in the capital of the state. To overcome this small sample problem, we group the years of education into four educational levels: less than high school, high school graduate, some college and post-college education (a modified version of Acemoglu and Autor (2010) education grouping), such that $K^* = 4 > J$. One of the important determinants of wages is the level of experience. Since the actual labour market experience is not available in the dataset, Card (1995) constructs a potential experience as $age - education - 6$. Since all individuals in the sample are of similar age $(24 - 34)$, people with the same years of schooling have similar levels of potential experience implying the lack of observations for all possible combinations of support points. Therefore,

[10] At http://davidcard.berkeley.edu/data_sets.html.

we group experience into two levels: low ($0-8$ years) and high ($9+$). In Card's (1995) linear specifications, the experience variable enters as a quadratic function to capture the nonlinear effect observed in many studies i.e. that wage increases with experience, but at the decreasing rate. One of the advantages of the nonparametric specification used here is that the shape of the relationship between wage and experience does not matter or that the nonparametric function captures all kinds of nonlinearities.

Table 10 summarizes the results of implementing our testing procedure in different specifications. In the first specification we use years of education without grouping. In the other specifications, we use levels of education and levels of experience (grouped) indicated by stars. The binary variable Race takes value 1 if an individual is black, 0 otherwise, and the binary SMSA equals 1 if an individual lived in a metropolitan area in 1976. The full set of family background variables is excluded from the analysis due to small sample limitations. The dependent variable is the log of hourly wages in 1976.

Covariates	R_n	$cv.1$	$cv.2$	$cv.3$	α
Educ	1.765	0.239	0.232	0.238	1%
		0.136	0.132	0.138	5%
		0.094	0.096	0.097	10%
Educ*, Exp*	4.147	1.221	1.259	1.217	1%
		0.715	0.696	0.719	5%
		0.511	0.500	0.515	10%
Educ*, Exp*, Race	3.572	1.771	1.692	1.688	1%
		1.107	1.131	1.108	5%
		0.849	0.871	0.860	10%
Educ*, Exp*, Race, SMSA	2.955	2.382	2.330	2.415	1%
		1.702	1.679	1.735	5%
		1.399	1.365	1.430	10%

Table 10: The value of the test-statistic and critical values for testing the exogeneity of education in different specifications

The critical values (columns $3, 4$ and 5) are computed by using three methods discussed in Section 3.2.5 : $cv.1$ are computed by simulating the weighted sum directly, $cv.2$ by simulating the quadratic form in normal vectors and $cv.3$ are obtained by chi-squared approximation. All these methods produce similar values and given relatively small sample size, their computation time is equally short. Hence, the choice of the most favorable approach is left to the practitioner. The value of the test statistic and critical values are large (in magnitude), because the estimated weights $\widehat{\omega}_j$ are large (e.g. $\widehat{\omega} = 35.84$ in the first specification). The results presented in the table are therefore scaled down (by dividing by 1000). Alternatively, we could scale down the test-statistic by dividing by the sum of the weights and adjust the asymptotic distribution accordingly.

In all four specifications, the calculated value of the test- statistic (column 2) exceeds the critical value at any significance level α and the null hypothesis of exogeneity is rejected. Therefore, we conclude that education is endogenous. Recall that under endogeneity, this model is nonparametrically not identified and the linear specification is critical for identifiability of parameters. However, as mentioned before, the parametric specification is not testable, and should not be used. Even though the nonparametric identification condition is violated, some linear functionals of the parameters of interested might be point identified and consistently estimated from the data. We will use the test proposed in Section 2.2.1 to check that in Section 5.3. Alternatively, some shape restrictions on the unknown function of interest might be imposed to partially identify and estimate informative bounds on the effects of education on earnings (see Section 5.3).

It can be argued that if education is endogenous, then so is experience, since the potential labour market experience variable is constructed as a function of education. Taking that into account, Card (1995) uses age as an instrument for (potentially) endogenous experience. Now, we are interested in testing jointly the exogeneity of two explanatory variables. Using the results discussed in Section 4.2, we treat this problem as the case with a single discrete regressor X with $K = 8$ points of support (4 values for educational levels and 2 values for levels of experience). Similarly, we have a single instrumental variable Z with $J = 22$ support points (11 values for age and 2 values for college proximity). Note that in this example we use one instrument for education and one for experience, but theoretically it is sufficient to use one instrument for both (as long as it satisfies the relevance condition). Since $J > K$ in this case, the nonparametric exogeneity can be tested using the statistic in equation (35). The results are summarized in Table 11. The reported critical values come from $\chi^2_{(DK-D)}$ distribution.

Covariates	T_n	null distribution	cv	α
Educ*, Exp*	173.47	$\chi^2_{(7)}$	18.48	1%
			14.07	5%
			12.02	10%
Educ*, Exp*, Race	185.11	$\chi^2_{(14)}$	29.14	1%
			23.68	5%
			21.06	10%
Educ*, Exp*, Race, SMSA	174.30	$\chi^2_{(28)}$	48.28	1%
			41.34	5%
			37.92	10%

Table 11: The value of the test-statistic and critical values for testing jointly the exogeneity of education and experience in different specifications

In all three specifications, the calculated value of the test-statistic exceeds the critical values at any significance level α. Hence, the null hypothesis of joint exogeneity of education and experience is rejected. Not surprisingly, education and experience turn out to be jointly endogenous. However, in this case ($J > K$) the endogeneity problem can be dealt with by using the nonparametric IV estimator discussed earlier.

5.2 Testing for exogeneity of education in Angrist and Krueger (1991)

The grouping of the years of education, which we were forced to impose because of the small sample size in Card's (1995) framework might be seen as a loss of information, since the marginal effect of an additional year of schooling cannot be identified. Therefore, we provide an additional example in which endogeneity of education is tested with the larger dataset used in Angrist and Krueger (1991). The data is available online[11] and consists of observations from 1980 Census documented in *Census of Population and Housing, 1980: Public Use Microdata Samples*. The sample consists of men born in the United States between 1930-1949 divided into two cohorts: those born in the 30's (329509 observations) and those born in the 40's (486926 observations). Angrist and Krueger (1991)

[11] At http://economics.mit.edu/faculty/angrist/data1/data/angkru1991.

estimate the conventional linear earnings function

$$\ln(wage_i) = \beta Educ_i + \sum_c \delta_c Y_{ci} + \sum_{s=1}^{S} \gamma_s W_{si} + \varepsilon_i \tag{37}$$

for each cohort separately, by $2SLS$ using the quarter of birth as an instrument for (assumed) endogenous education. They observe that individuals born earlier in the year (first two quarters) have less schooling than those born later in the year. It is a consequence of the compulsory schooling laws, as individuals born in the first quarters of the year reach the minimum school leaving age at the lower grade and might legally leave school with less education. The (sample) support of education consists of $K = 21$ values (years of education between 0 and 20) and the support of the instrument is $J = 4$. In the linear model (37), Y_{ci} is a dummy variable indicating whether an individual was born in year c, $c = 1, ..., 10$ and W denotes the standard set of exogenous covariates.

To begin an investigation of endogeneity of education, we consider a nonparametric simplified version of (37), in which (log) wage is a function of education and the year of birth (W is excluded). Hence, we have a model with one potentially endogenous regressor (education) and one exogenous covariate (year of birth). As mentioned in Section 4.1.1, we can divide the observations into subsamples corresponding to different values of exogenous regressors, i.e. different years of birth. Table 12 shows the results of testing for exogeneity of education in each subsample and for the full cohort. The critical values are calculated by using the chi-square approximation to the weighted sum of chi-square variables.

		critical values		
year of birth	R_n	1%	5%	10%
1930	0.6139	14.9057	9.6479	7.4103
1931	0.3670	17.7219	11.4240	8.7507
1932	8.4177	20.5510	13.6590	10.6751
1933	2.8771	18.9001	12.6327	9.9115
1934	5.8959	25.1629	16.4850	12.7607
1935	3.4247	19.9431	13.0420	10.0829
1936	9.2843	28.0781	17.8537	13.5379
1937	0.9321	12.9027	8.6420	6.7889
1938	0.7996	21.8865	14.4882	11.2930
1939	5.2867	21.5534	14.3601	11.2414
full cohort	38.044	85.933	72.138	65.465

Table 12: The value of the test-statistic and critical values for testing exogeneity of education for 1930's cohort

Surprisingly, the calculated value of the test statistic is lower than the obtained critical values at any significance level, which implies that the null hypothesis of exogeneity of education cannot be rejected. Since only education is used to explain the variation in wages, the omission of other relevant regressors should clearly result in endogeneity of schooling. However, the test suggests that education is exogenous, which we doubt to believe. Nonetheless, this questionable outcome does not suggest that our testing procedure fails in detecting endogeneity of education. The possible explanation is that the quarter of birth is not a valid instrument for education. It is obvious that the performance of the test is affected by the quality of instruments used, and provides reliable results only when instruments are appropriate. The main criticism of Angrist and Krueger (1991) analysis, pointed out by Bound, Jaeger and Baker (1995) is that the quarter of birth is a weak instrument. As discussed in Hahn, Ham and Moon (2011) even the standard Hausman (1978) test for exogeneity is invalid in the presence of weak instruments. Therefore, we might expect our testing procedure to provide misleading results if the instruments do not satisfy the instrumental relevance condition. A second criticism of Angrist and Krueger (1991) results, discussed by Bound and Jaeger (1996) is that quarter of birth might be correlated with unobserved ability and hence does not satisfy the instrumental exogeneity condition. This leads to the conclusion that the quarter of birth might not be a valid instrument and that the outcome of our exogeneity testing procedure remarkably depends on the validity of instruments.

This view might be supported by the analysis of the 1940's cohort summarized in Table 13.

		critical values		
year of birth	R_n	1%	5%	10%
1940	2.9915	22.6885	14.7948	11.4127
1941	12.3195	20.9773	13.7668	10.6693*
1942	8.1070	33.0513	20.9644	15.8623
1943	93.2736	54.4804*	34.6581*	26.2979*
1944	11.9499	32.8291	21.0770	16.0995
1945	33.8664	26.2149*	17.3642*	13.5449*
1946	19.1200	22.5649	14.6112*	11.2136*
1947	24.3138	32.4185	21.3369*	16.5725*
1948	38.2687	45.3827	28.8280*	21.8547*
1949	35.8060	34.5416*	22.5009*	17.3578*
full cohort	278.703	138.344*	114.551*	103.182*

Table 13: The value of the test-statistic and critical values for testing exogeneity of education for 1940's cohort

The calculated values of the test statistic are reported in column 1. The critical values lower than the calculated value of the test statistic are indicated with stars. The results for the cohort born in the 1940's differ slightly from what was presented for the 1930's cohort. In some subsamples, the calculated value of the test statistic exceeds the critical values and education is confirmed to be endogenous, as expected. The results of testing for exogeneity of education with other exogenous explanatory variables (race and region of residence) are presented in the Appendix C. Note that the dataset contains additional variables, which are relevant to explain the variation in wages, such as marital status and family background characteristics. These variables could not be included in the analysis, as there are not enough observations for all possible combinations of the values of these regressors. However, all of the results presented here confirm the endogeneity of education in some subsamples, but fail to detect it in others. An important lesson learnt from Angrist and Krueger (1991) example is that the validity of our test is highly related to the validity of instruments.

5.3 Testing for point-identifiability of linear functionals

In order to show how our test for point identifiability of some linear functionals might be use in practice, we go back to Card's (1995) model and consider a simplified model with endogenous education regressor, but without any exogenous explanatory variables. The interest of applied researchers typically lies in estimating the marginal effects i.e. the change in the dependent variable for a unit (or percentage) change in the regressor. Suppose that we want to recover these marginal effects in the nonparametric specification. Therefore, we want to estimate the difference in the conditional expectations of the dependent variable given some values of the regressor, i.e. the change in the average wage for one year increase in education. Additionally, given a linear regression specification the marginal effects are constant for all values of the explanatory variable. We might also be interested in checking whether these effects are constant in the nonparametric specification.

Recall that $\beta_k = h(x_k)$, where x_k denotes the years of schooling. In the sample, $x_1 = 2$ and $x_{17} = 18$ years of education. Let $L(\beta) = c'\beta = c_1'\beta_1 + c_2'\beta_2$ be a linear functional of the elements of β. From Proposition 2 it follows that the linear functional of dimension $J = 2$ might be point identified. Let

$$c_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

and c_2 be $(K - J) = 15$ vector of zeros. Suppose that we are interested in estimating the difference in earnings between individuals with 6 and 7 years of education. Partitioning β comfortably into $\beta_1 = [h(7)\ h(6)]'$ and β_2 with all remaining $h(x_k)$ provides

$$L(\beta) = [h(7) - h(6)]\,. \tag{38}$$

Using the test-statistic in (13), we test whether the linear functional (38) is point identified. The 5% critical value from χ^2_{15} distribution is 25.00. The calculated value of the test statistic equals 1.9332 which is lower than the critical value. Therefore, the null hypothesis that this linear functional is point identified cannot be rejected. This implies that the difference in earnings between individuals with 6 and 7 years of schooling is point identified and can be consistently estimated by

$$\widehat{L}(\beta) = c_1' \widehat{\Pi}_1^{-1} \widehat{\pi} = 0.1017, \tag{39}$$

but we can learn nothing from the data about any linear combination of all remaining $K - J$ average wages.

Table 14 summarizes the results of testing the point-identifiability of some differences in average wage of individuals with different educational levels. The estimates of differences which are point identified are given in column 3.

linear combination	G_n	$\widehat{L}(\beta)$
$h(3) - h(2)$	0.1356	0.0040
$h(4) - h(3)$	0.8188	0.0103
$h(5) - h(4)$	0.4673	0.0273
$h(6) - h(5)$	-	-
$h(7) - h(6)$	1.9332	0.1017
$h(8) - h(7)$	0.1494	0.2395
$h(9) - h(8)$	26.5527	-
$h(10) - h(9)$	75.2217	-
$h(11) - h(10)$	4.7003	0.1317
$h(12) - h(11)$	13.5271	0.2499
$h(13) - h(12)$	38.6814	-
$h(14) - h(13)$	61.5525	-
$h(15) - h(14)$	33.4315	-
$h(16) - h(15)$	80.0153	-
$h(17) - h(16)$	10.7344	-0.1900
$h(18) - h(17)$	74.1413	-

Table 14: The value of the test-statistic and esimated linear functional for differences in returns to schooling

Note that the point-identifiability of the difference $h(6) - h(5)$ could not be tested as the matrix of joint probabilities P_1 is singular in that case. There are some differences that are not point identified, and in order to learn something about them alternative methods (discussed in Section 2.3.2) must be employed. Here, we would like to compare our estimates in Table 14 with interval estimates obtained by using Chesher's (2004) method. Our exercise follows an empirical example 6.2 in Horowitz (2011). In order to apply Chesher's (2004) approach, the monotonicity condition (17) must be satisfied, which requires that

$$\Pr\left[X \leq x_k | Z = 1\right] \leq \Pr[X \leq x_{k-1} | Z = 0]. \tag{40}$$

Table 15 provides the relevant empirical probabilities that the years of education is less than or equal to certain values from 2 to 18 conditional on the instrument.

years of schooling	Z=1	Z=0
2	0.0005	0.0010
3	0.0015	0.002
4	0.002	0.0041
5	0.0044	0.0093
6	0.0083	0.0177
7	0.0141	0.0355
8	0.0282	0.0763
9	0.0482	0.1181
10	0.0862	0.1673
11	0.1359	0.2269
12	0.4559	0.5773
13	0.5504	0.6683
14	0.6459	0.7384
15	0.7068	0.7750
16	0.8656	0.9141
17	0.9192	0.9570
18	1	1

Table 15: Empirical probabilities for different years of education

There are some years of education for which condition (40) is satisfied, for example for $6, 7, 8, 10$ and 11 years of schooling. Therefore, we can estimate the bounds on the differences:

$$\begin{aligned} 0.0365 &\leq [h(7)-h(6)] \leq 0.2895 \\ -0.1732 &\leq [h(8)-h(7)] \leq 0.352 \\ -0.0.57 &\leq [h(11)-h(10)] \leq 0.3187. \end{aligned}$$

The analogous estimated values in Table 14 lie within the calculated intervals.

The presented exercise shows that both methods of estimating the differences in conditional expectations are complementary. Chesher's (2004) bounds can be used in cases when our method fails, for example, the difference $h(9) - h(8)$ is not point identified, but can be bounded between $[-0.2742, 0.1334]$. On the other hand, we provide consistent estimates for cases, which do not satisfy the monotonicity assumption required by Chesher (2004), e.g. when the years of schooling equals 3. For situations in which both methods could be applied, the results do not contradict each other. Clearly, there are some cases in which none of the proposed methods can be used and learning about the unknown function of interest requires different approaches.

6 Conclusions and further work

This thesis has been concerned with nonparametric additive error models with discrete regressors. Even though nonparametric techniques have become very popular recently, there is still a range of problems that remains unexplored. Our first contribution to the existing literature is the observation that the nonparametric model with discrete regressors can be reinterpreted as a linear model and the standard nonparametric estimators translate into standard regression estimators. This knowledge allows us to use the well-established regression techniques without making strong assumptions on the underlying population of interest and data generating process. We have discussed the notion of nonparametric exogeneity and presented the consequences of the presence of endogenous regressors in the nonparametric model. We also emphasized the necessity of testing whether the exogeneity assumption holds. Additionally, the identification analysis for models with endogenous discrete regressors was presented and the conditions for point and partial identification were summarized. The nonparametric point identification requires that the instrumental variable has at least as many support points as endogenous regressor. Under identification failure endogeneity implies the nonexistence of any consistent estimator for some interest-parameters without further assumptions. However, it has been shown that some linear functionals of the conditional mean function $h(\cdot)$ might be point identified, while the entire function remains unknown. One of the contributions of this dissertation was to provide a test for point-identifiability of these linear functionals. It has been demonstrated that the test can easily be applied in practical problems.

This dissertation has also provided two consistent tests for exogeneity. To the best of our knowledge, there exist no such tests for nonparametric models with discrete regressors. In the models that point identify the unknown function of interest, the test is built on a quadratic form of a difference between two estimators, one of which is consistent only under exogeneity and the other is consistent under both scenarios. This testing framework follows closely the Durbin-Wu–Hausman type of test. It has been shown that under the null hypothesis of exogeneity, the test statistic follows asymptotically a chi-square distribution with degrees of freedom equal to the number of points of support in the regressor X less 1. The test is consistent against fixed alternatives with well established asymptotic distribution under local alternatives.

In the models that partially identify the entire structure of interest, the Durbin-Wu-Hausman methodology is not available, as there exists no consistent estimator under the alternative hypothesis. Hence, the test-statisic is derived from a constrained minimization problem. We have shown that the proposed test is consistent, but with a non-standard asymptotic null distribution. Three possible ways of computing critical values for practical applications have been presented, and efficiency of these methods has been examined in Monte Carlo simulations. The experiments demonstrate that both tests have adequate size and satisfactory power properties in finite samples. An interesting finding is that the power of both tests increases with the difference between the number of points of support of the regressor and instrument. This observation might be useful for empirical researchers if they have a choice between different instrumental variables. Another useful, practical information revealed in simulations is that the probability of detecting endogeneity is higher if there is a similar number of observations for each support point of the regressor.

Generalizations of our testing procedures to models that are widely used in empirical applications have also been discussed. We have provided a testing framework for nonparametric models with multiple regressors, using multiple instruments, in the presence of multiple exogenous explanatory variables. An interesting result of our discussion is that in this case nonparametric identification does not depend on the actual number of instruments, but only on the product of the numbers of their support points. Therefore, theoretically, many endogenous regressors could be instrumented by a single variable with large enough support.

The empirical examples provided in the final section of this thesis, showed that the proposed tests are easily applicable in practice and confirmed endogeneity of education is some classic applied work concerned with estimating the returns to schooling in standard wage equations. Under endogeneity, models of Card (1995) and Angrist and Krueger (1991) are nonparametrically unidentified, but point-identification and the consistent estimation of differences in the average wage for distinct years of education were proved to be feasible.

The analysis presented above can be extended in many directions. Our first idea is to consider increasing dimensions of variables. Since the identification of the unknown function of interest $h(\cdot)$ depends on the support of the instrument Z relative to the support of the endogenous regressor X, it is interesting to examine the impact of changes in J and K, if we let them both to grow with the sample size.

Suppose that X is exogenous and we allow the support of X to increase with the sample size. The study of the asymptotic distribution of standard estimators as both K and n increases follows two routes. The first approach searches for the fastest growth rate of K that is consistent with standard asymptotic normality and consistency results. It has been shown that the condition $K = o(n)$ is necessary, but often insufficient.

The increasing K asymptotics were firstly discussed in the context of M-estimation. Huber (1973) shows that the standard OLS estimator is consistent and asymptotically normal, when K increases with n, but only if $\frac{K}{n} \to 0$ (this is a necessary condition). He proves normality of the M-estimator of the linear regression model under the stronger condition that $\frac{K^3}{n} \to 0$. This rate was improved by Yohai and Maronna (1979) to $\frac{K^2}{n} \to 0$ for consistency and $\frac{K^{5/2}}{n} \to 0$ for asymptotic normality, and by Portnoy (1984,1985) to $\frac{K \log K}{n} \to 0$ for consistency and $\frac{(K \log K)^{1.5}}{n} \to 0$ for asymptotic normality.

In our model, in the standard case with K fixed, if $E[\varepsilon_i] = 0$ and $E[\varepsilon_i^2] < \infty$ then for the OLS estimator $\widehat{\beta}$, we have $\left(\widehat{\beta} - \beta\right)' (L_X' L_X) \left(\widehat{\beta} - \beta\right) = O_p(1)$. The sufficient condition for the consistency of $\widehat{\beta}$ is that the smallest eigenvalue of $(L_X' L_X)$ tends to infinity. For the cases in which $K \to \infty$, the conditions for $K^{-1} \left(\widehat{\beta} - \beta\right)' (L_X' L_X) \left(\widehat{\beta} - \beta\right)$ to be bounded in probability can be derived from Huber's (1973) result.

In the model of form

$$y_i = \sum_{k=1}^{K} I(x_i^s = x_k)\beta_k + \varepsilon_i,$$

with ε_i *iid* $(0, \sigma^2)$, a necessary and sufficient condition for all least square estimates of form $\widehat{\theta} = \sum_{k=1}^{K} a_k \beta_k$ to be asymptotically normal is that

$$\max_i \gamma_{ii} \to 0 \text{ as } n \to \infty,$$

where γ_{ii} is the i^{th} diagonal element of P_{L_X}. Note that in our case,

$$\gamma_{ii} = \sum_{k=1}^{K} \frac{I(x_i^s = x_k)}{n_k} = \left\{ \begin{array}{c} \frac{1}{n_1} \text{ if } x_i^s = x_1 \\ \dots \\ \frac{1}{n_K} \text{ if } x_i^s = x_K \end{array} \right\}.$$

Hence $\max_{1 \leq i \leq n}(\gamma_{ii}) = \max_k\{\frac{1}{n_k}\} = \min_k\{n_k\}$, $k = 1, ..., K$. Note that $n_k, k = 1, ..., K$ are the eigenvalues of $(L_X' L_X)$. Therefore, the sufficient condition for consistency is $K \min_k\{n_k\} \rightarrow 0$.

The second approach of increasing K asymptotics looks for alternative asymptotic distributions of the estimators keeping $\frac{K}{n}$ positive. The assumption that K grows proportionally with n, i.e. $\frac{K}{n} \rightarrow \varpi$, with $0 < \varpi < 1$, is typically used in classical many instruments asymptotic theory of Bekker (1994) and in the theory of large random matrices (e.g. Bai (1999), Ledoit and Wolf (2004)). This framework rules out the cases of few regressors, as $K \rightarrow \infty$ and $\varpi > 0$ and the cases with $K = o(n)$ discussed above. In order to introduce the increasing dimension of X, the normalization technique employed by Cai (2007) in time varying-coefficient model and discussed by Feng et al. (2015) to deal with varying-coefficient panel data models, might be used. Suppose that $X \in \{0, 1, 2, ..., v(n) - 1\}$, where $v(n) \rightarrow \infty$ and $v(n)/n \rightarrow \varpi$ for $0 < \varpi < \infty$ as $n \rightarrow \infty$. In this case, model (1) can be written as

$$Y = h\left(\frac{X}{v(n)}\right) + \varepsilon,$$

and $h(\cdot)$ can be treated as a function with continuous covariates.

Under endogeneity of X, the identification of the unknown $h(.)$ depends on the support of the instrument Z. For fixed value of K, if we let J to grow with the sample size, it might be possible that the model that is only set identified in small samples ($J < K$) is going to point identify $h(.)$ in the large samples.

The cases of fixed K and increasing J were broadly discussed by Bekker (1994), Hansen et. al. (2008) and van Hasselt (2010) among others. They provide the multivariate approximations to the distributions of standard estimators (e.g. OLS and 2SLS), using a parameter sequence with the number of instruments increasing with the sample size.

The second possible extension of our work is to relax the *iid* assumption and to consider models with correlated errors. Although many researchers concentrated on exogeneity testing in nonparametric models, we are not aware of any attempt to incorporate the dependence in the error term. It is extremely important to extend our testing procedures to cases with correlated disturbances, since then our tests could be used with time-series data. Many modern nonparametric techniques are now applied in time series analysis. A natural extension of the model to a dynamic case with discrete time observations might be given by

$$Y_t = h(X_t) + \varepsilon_t,$$

where (Y_t, X_t) is a joint Markov process and

$$E[\varepsilon | Y_{t-1}, X_{t-1}] = 0.$$

If we use the lagged variables (Y_{t-1}, X_{t-1}) as instruments, all the theory presented above can be applied. The results can also be extended to stationary and ergodic processes and to nonstationary data.

The final generalization of nonparametric exogeneity testing framework is to relax the assumption of additivity of the error term and consider a nonseparable model of form

$$Y = h(X, \varepsilon). \tag{41}$$

These nonparametric models were broadly discussed by Roehrig (1988), Matzkin (2003), Imbens and Newey (2009), Chesher (2003,2005) among others. There is a convenient interpretation of h in (41). Under the assumption that ε is $Unif(0,1)$ and h is a function which is strictly increasing in ε, h represents a nonparametric conditional quantile function.

As in the additive error model, the consistent estimation of the unknown structural function $h(\cdot)$ is feasible if the regressors X are exogenous. The identification and estimation in this context is discussed, for example in Matzkin (2003). In the presence of endogeneity, instrumental variables estimation is recommended. An interesting feature of nonseparable models is that the identifying properties of the model depend on whether the endogenous regressor is continuos or discrete. If X is continuously distributed, under mild regularity conditions the model point identifies the structure of interest $h(\cdot)$ (see, Chesher (2003) and Imbens and Newey (2009)). However, when the endogenous regressor is discrete, the models are only set identifying (Chesher (2005)). Therefore, testing for exogeneity of regressors in (41) when X is discrete, is essentially testing point vs. partial identification. We believe that such tests would be a great contribution to existing knowledge, since the literature on specification testing in partially identified models is very sparse.

Appendix A: Proofs

Proof of Proposition 2

For some permutation of the elements of β the equations $\pi = \Pi\beta$ can be split into

$$\Pi_1\beta_1 + \Pi_2\beta_2 = \pi,$$

with Π_1 $J \times J$ and non-singular, and β_2 $(K-J)\times 1$. Thus, the solutions satisfy

$$\beta_1 = \Pi_1^{-1}(\pi - \Pi_2\beta_2).$$

Varying $\beta_2 \in \mathbb{R}^{K-J}$ generates all solutions to (5). Partitioning c conformably into $J \times 1$ vector c_1 and $(K-J)\times 1$ vector c_2,

$$L(\beta) = c_1'\beta_1 + c_2'\beta_2 = c_1'\Pi_1^{-1}\pi + (c_2' - c_1'\Pi_1^{-1}\Pi_2)\beta_2. \tag{42}$$

Vectors b in the null space of Π satisfy $b_1 = -\Pi_1^{-1}\Pi_2 b_2$, and c is orthogonal to $null(\Pi)$ iff $(c_2' - c_1'\Pi_1^{-1}\Pi_2)b_2 = 0$ for all b_2, i.e., $c_2' - c_1'\Pi_1^{-1}\Pi_2 = 0'$. Thus, for linear combinations c orthogonal to the null space, $L(\beta) = c'\beta = c_1'\Pi_1^{-1}\pi$ is point identified. On the other hand, for any vector c such that $(c_2' - c_1'\Pi_1^{-1}\Pi_2) \neq 0'$, $L(\beta)$ is completely unrestricted, because there is a b_2 such that$(c_2' - c_1'\Pi_1^{-1}\Pi_2)b_2 \neq 0$, and one can choose $\beta_2 = \gamma b_2$. Then, by varying γ, any value for $L(\beta)$ in equation (42) can be achieved. Note that c_1 corresponds to linear combinations that are point identified, and c_2 to those which are completely undetermined.

Proof of Proposition 3

It is clear that the condition is sufficient, since for $\beta \in S_\pi$

$$C'\beta = C'\beta_0 + C'V\gamma = C'\beta_0$$

is constant. To show that the condition is necessary, we need to show that $C'\beta$ is a maximal invariant under the group of transformations on S_π given by $\beta \to \beta + V\gamma, \gamma \in \mathbb{R}^{K-J}$. Invariance is obvious. Hence, we just need to show that, for $\beta_1, \beta_2 \in S_\pi$,

$$C'\beta_1 = C'\beta_2 \Rightarrow \beta_2 = \beta_1 + V\gamma^*, \text{ for some } \gamma^* \in \mathbb{R}^J.$$

This follows, because for $\beta_1, \beta_2 \in S_\pi$,

$$\begin{aligned} C'\beta_1 &= C'\beta_2 \iff C'(\beta_1 - \beta_2) = 0 \\ &\iff \beta_2 = \beta_1 + V\gamma, \text{ for some } \gamma \in \mathbb{R}^{K-J}, \end{aligned}$$

which is a transformation by a group element, as required. The results then follow from Theorem 6.1.4. in Muirhead (1982), which says that the functions that are invariant under a group action are functions only of a maximal invariant.

Proof of Theorem 2.1

For part (i), let $p = vec(P)$ and $\widehat{p} = vec(\widehat{P})$, both $JK \times 1$ vectors. Since $\widehat{P}$ is a matrix of sample proportions it follows from the multivariate CLT (Severini (2005), p.377-378) that

$$\sqrt{n}\left(\widehat{p} - p\right) \to^d N\left(0, D_p - pp'\right),$$

where $D_p = diag\{p\}$. Now,

$$\widehat{P}_1^{-1}\widehat{P}_2 = \left[P_1 + \left(\widehat{P}_1 - P_1\right)\right]^{-1}\left[P_2 + \left(\widehat{P}_2 - P_2\right)\right],$$

and in this expression both $\left(\widehat{P}_1 - P_1\right)$ and $\left(\widehat{P}_2 - P_2\right)$ are $O_p(n^{-\frac{1}{2}})$. Write the product as:

$$P_1^{-1}\left[I_J + \left(\widehat{P}_1 - P_1\right)P_1^{-1}\right]^{-1}\left[P_2 + \left(\widehat{P}_2 - P_2\right)\right].$$

Since $\left(\widehat{P}_1 - P_1\right)P_1^{-1}$ is a square matrix, for large n the inverse here can be expanded as[12]

$$\left[I_J + \left(\widehat{P}_1 - P_1\right)P_1^{-1}\right]^{-1} = I_J - \left(\widehat{P}_1 - P_1\right)P_1^{-1} + O_p(n^{-1}).$$

[12] Using the matrix geometric series (see e.g. Theorem 10.26 in Rosenblatt and Bell (1999)).

Therefore

$$\begin{aligned}\widehat{P}_1^{-1}\widehat{P}_2 &= \left[P_1^{-1} - P_1^{-1}\left(\widehat{P}_1 - P_1\right)P_1^{-1} + O_p(n^{-1})\right]\left[P_2 + \left(\widehat{P}_2 - P_2\right)\right] \\ &= P_1^{-1}P_2 + P_1^{-1}\left(\widehat{P}_2 - P_2\right) - P_1^{-1}\left(\widehat{P}_1 - P_1\right)P_1^{-1}P_2 + O_p(n^{-1}).\end{aligned}$$

Thus, for fixed (c_1, c_2),

$$c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2 = \left(c_2' - c_1'P_1^{-1}P_2\right) - c_1'P_1^{-1}\left(\widehat{P}_2 - P_2\right) + c_1'P_1^{-1}\left(\widehat{P}_1 - P_1\right)P_1^{-1}P_2 + O_p(n^{-1}),$$

and under $H_0^c, c_2' - c_1'P_1^{-1}P_2 = 0'$. By Theorem 2 in Magdalinos (1992), the distribution of $c_2' - c_1'\widehat{P}_1^{-1}\widehat{P}_2$ can be approximated by the asymptotic distribution of the (row) vector

$$c_1'P_1^{-1}\left[\left(\sqrt{n}\left(\widehat{P}_1 - P_1\right), \sqrt{n}\left(\widehat{P}_2 - P_2\right)\right)\begin{pmatrix} P_1^{-1}P_2 \\ -I_{K-J}\end{pmatrix}\right] = z'UA, \qquad (43)$$

say, where $z' = c_1'P_1^{-1}, U = \sqrt{n}(\widehat{P} - P)$, and $A = \binom{P_1^{-1}P_2}{-I_{K-J}}$.

Consider an arbitrary linear combination

$$g_n(t) = z'UAt = tr\left[z'Uv\right] = tr[vz'U],$$

where $v = At$. Given the relation between trace and vectorization operators, it follows that

$$g_n(t) = a'vec(U),$$

with $a' = (vec(zv'))' = (v_1z', ..., v_Kz')$. Since $vec(U) \to^d N(0, D_p - pp'))$, for every t,

$$g_n(t) \to^d N(0, a'(D_p - pp')a).$$

But the variance here is

$$a'(D_p - pp')a = a'D_pa - a'pp'a,$$

and

$$\begin{aligned}a'p &= (vec(zv'))'vec(P) = trace\left[vz'P\right] \\ &= trace[z'Pv] = z'Pv = 0,\end{aligned}$$

since $PA = (P_1 \ P_2)\binom{P_1^{-1}P_2}{-I_{K-J}} = 0$, so the variance reduces to

$$a' D_p a = t' A' \mathcal{D} A t,$$

where

$$\mathcal{D} = diag\{c_1' P_1^{-1} D_k P_1'^{-1} c_1; k = 1, ..., K\},$$

with D_k the $J \times J$ diagonal matrix with the elements in column k of P on the diagonal.

It follows from Cramer's characterization theorem that, under H_0^c,

$$\sqrt{n}\left(c_2' - c_1' \widehat{P}_1^{-1} \widehat{P}_2\right)' \to^d N(0, V_P),$$

with

$$V_P = \begin{pmatrix} P_1^{-1} P_2 \\ -I_{K-J} \end{pmatrix}' \mathcal{D} \begin{pmatrix} P_1^{-1} P_2 \\ -I_{K-J} \end{pmatrix},$$

as claimed. V_P can be consistently estimated by replacing unknown probability matrices with their sample equivalents. The null distribution of G_n given in part (ii) of Theorem 2.1 follows immediately.

Proof of Theorem 2.2

The asymptotic distribution of the OLS estimator $\widehat{\beta}$ follows immediately from Lemma 3.1 (the proof of Lemma 3.1 can be found below). As, under exogeneity

$$\frac{1}{\sqrt{n}} L_X' \varepsilon_n \to^d N(0, \sigma^2 D_X)$$

and $n^{-1} L_X' L_X \to^p D_X$, by Slutsky Theorem we have

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \left(\frac{L_X' L_X}{n}\right)^{-1} \frac{L_X' \varepsilon_n}{\sqrt{n}} \to^d N\left(0, \sigma^2 D_X^{-1}\right),$$

as required.

Alternative ways to determine the asymptotic distribution of the OLS estimator

Robbins' CLT

Suppose that we have a random variable $Y = X_1 + ... + X_N = \sum_{i=1}^{N} X_i$, but N is itself a random variable. Assume that $X_i's$ are iid and N is independent of the $X_i's$. The distribution function of N depends on a parameter λ and is determined by the values $w_l = \Pr[N = l]$ for $l = 0, 1, ...$, where w_l are functions of λ, such that $w_l \geq 0$ and $\sum_{l=0}^{\infty} w_l = 1$. Let

$$\begin{aligned}
\alpha &= E[N] = \sum_{l=0}^{\infty} w_l l, \\
\beta^2 &= E[N^2] = \sum_{l=0}^{\infty} w_l l^2 \text{ (finite)}, \\
\gamma^2 &= Var[N] = \beta^2 - \alpha^2
\end{aligned}$$

be a functions of parameter λ. The quantities independent of λ are

$$\begin{aligned}
a &= E[X_i], \\
b^2 &= E[X_i^2], \\
c^2 &= Var[X_i] = \left(b^2 - a^2\right) < \infty.
\end{aligned}$$

Then, we have the following moments of Y :

$$\begin{aligned}
E[Y] &= \alpha a, \\
Var[Y] &= \alpha c^2 + \gamma^2 a^2 \equiv \delta^2.
\end{aligned}$$

Let the normalized variable Z be

$$Z = \frac{Y - E[Y]}{[Var(Y)]^{\frac{1}{2}}} = \frac{Y - \alpha a}{\delta},$$

and assume that as $\lambda \to \infty$

$$\delta^2 \to \infty, \gamma = o(\delta^2). \tag{44}$$

Theorem 6.1 (Robbins (1948)) *If (44) holds and if as* $\lambda \to \infty$

$$a^2\gamma^2 = o(\alpha),$$

then $Z \to^d N(0,1)$ *and* Y *is asymptotically* $N(\alpha a, \delta^2)$.

Corollary 1 (Robbins (1948)) *If* N *is asymptotically normal* (α, γ^2) *then* Y *is asymptotically normal* $(\alpha a, \delta^2)$.

In our case, consider $\widehat{\beta}_k - \beta_k = \frac{1}{n_k^X}\sum_{i=1}^n \varepsilon_i I(x_i^s = x_1) \equiv \frac{1}{n_k^X}\sum_{n_k^X} \varepsilon_i$. And let $Y = \sum_{n_k^X} \varepsilon_i$, that is the sum of random number of random variables. Using the notation of Robbins (1948), we have

$$a = E[\varepsilon_i] = 0$$

and

$$c^2 = Var[\varepsilon_i] = \sigma^2.$$

Now n_k^X is a random variable that can take any value from 0 to n. In order to proceed, we need to know its distribution function. Since n_k^X is the multiplicity of x_k in the sample of n observations ("number of successes in n trials), n_k^X follows the Binomial distribution with the probability of success equal to $\Pr[X = x_k] = p_k$, i.e.

$$n_k^X \sim Bin(n, p_k).$$

Then

$$\begin{aligned} \alpha &= E[n_k^X] = np_k \\ \gamma^2 &= Var[n_k^X] = np_k(1-p_k). \end{aligned}$$

If n is large enough, the reasonable approximation of $Bin(n, p_k)$ is given by the normal distribution $N(np_k, np_k(1-p_k))$. Therefore

$$n_k^X \sim_{appr.} N(np_k, np_k(1-p_k)) \text{ as } n \to \infty.$$

Using the Corollary 1, we get

$$Y \to^d N(\alpha a, \delta^2),$$

where $\alpha a = 0$ and $\delta^2 = \alpha c^2 + \gamma^2 a^2 = np_k\sigma^2$. This implies that

$$\frac{Y}{\sqrt{n}} \to^d N(0, \sigma^2 p_k)$$

or

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i I(x_i^s = x_k) \to^d N(0, \sigma^2 p_k).$$

We have shown the marginal convergence of the each element of $\frac{1}{\sqrt{n}} L'_X \varepsilon_n$. Next, consider the components $\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i I(x_i^s = x_k), \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i I(x_i^s = x_l)\right)$ for $k \neq l$. By independence, we have

$$\begin{aligned} & E\left[\frac{1}{n}\left(\sum_{i=1}^{n} \varepsilon_i I(x_i^s = x_k)\right)\left(\sum_{i=1}^{n} \varepsilon_i I(x_i^s = x_l)\right)\right] \\ = \; & E\left[\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 I(x_i^s = x_k) I(x_i^s = x_l)\right] = 0, \end{aligned}$$

because $I(x_i^s = x_k)I(x_i^s = x_l) = 0$, the events cannot occur simultaneously. Therefore, the vector

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i I(x_i^s = x_1) \\ \ldots \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i I(x_i^s = x_K) \end{pmatrix} \to^d N(0, \sigma^2 D_X),$$

and the asymptotic distribution of $\sqrt{n}(\widehat{\beta} - \beta)$ follows instantly.

Anscombe's Theorem

An alternative way of obtaining the asymptotic distribution of sums of random numbers of *iid* random variables is given in

Theorem 6.2 (Anscombe (1952)) *Suppose that $\zeta_1, \zeta_2, \ldots, \zeta_n, \ldots$ are iid random variables with mean* 0 *and variance* 1*. Let $S_n = \zeta_1 + \zeta_2 + \ldots + \zeta_n$. Let further $v(t)$ denote a positive integer-valued random variable for any $t > 0$ such that $\frac{v(t)}{t}$ converges for $t \to \infty$ in probability to a constant $c > 0$. Then*

$$\lim_{t\to\infty} P\left(\frac{S_{v(t)}}{\sqrt{v(t)}} < x\right) = \Phi(x) = N(0, 1).$$

In our case, let $U_i \equiv \frac{\varepsilon_i}{\sqrt{Var(\varepsilon_i)}} = \frac{\varepsilon_i}{\sigma}$. Then $U_1, U_2, ..., U_n$ is a sequence of independent and identically distributed random variables with mean 0 and variance 1 (since $\varepsilon_i' s$ are iid with mean 0 and variance σ^2). Let $S_{n_k^X} = \sum_{i=1}^{n_k^X} U_i$, where n_k^X is a positive integer-valued random variable. We know that $\frac{n_k^X}{n} \rightarrow^p \Pr[X = x_k] = p_k > 0$. Then

$$\frac{S_{n_k^X}}{\sqrt{n_{X\,k}}} \rightarrow^d N(0,1).$$

This implies that

$$\frac{\sum_{n_k^X} \varepsilon_i}{\sqrt{n_k^X}} \rightarrow^d N(0,\sigma^2).$$

As $\sqrt{n_k^X}\left(\widehat{\beta}_k - \beta_k\right) = \frac{\sum_{n_k^X} \varepsilon_i}{\sqrt{n_k^X}}$, we have that

$$\sqrt{n_k^X}\left(\widehat{\beta}_k - \beta_k\right) \rightarrow^d N\left(0,\sigma^2\right),$$

and

$$\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right) \rightarrow^d N(0,\sigma^2 p_k^{-1}).$$

Proof of Theorem 2.3

The asymptotic distribution of the IV estimator $\widehat{\beta}_{IV}$ follows immediately from Lemma 3.1, which provides:

$$\frac{1}{\sqrt{n}} L_Z' \varepsilon_n \rightarrow^d N(0, \sigma^2 D_Z).$$

Additionally, we have already shown that

$$\frac{L_X' L_Z}{n} \rightarrow^p P' \text{ and } \left(\frac{L_Z' L_Z}{n}\right)^{-1} \rightarrow^p D_Z^{-1}.$$

Since $\widehat{\beta}_{IV}$ is given by

$$\widehat{\beta}_{IV} - \beta = \left(L_X' P_{L_Z} L_X\right)^{-1} L_X' P_{L_Z} \varepsilon_n = \left(L_X' L_Z (L_Z' L_Z)^{-1} L_Z' L_X\right)^{-1} L_X' L_Z (L_Z' L_Z)^{-1} L_Z' \varepsilon_n,$$

it follows that

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_{IV}-\beta) &= \left(\frac{L_X'L_Z}{n}\left(\frac{L_Z'L_Z}{n}\right)^{-1}\frac{L_Z'L_X}{n}\right)^{-1}\frac{L_X'L_Z}{n}\left(\frac{L_Z'L_Z}{n}\right)^{-1}\frac{L_Z'\varepsilon_n}{\sqrt{n}} \\ &\rightarrow {}^{d}N\left(0,\sigma^2\left(P'D_Z^{-1}P\right)^{-1}\right).\end{aligned}$$

Proof of Lemmas 3.1, 3.2 and Theorem 3.1

The crucial result underlying the distribution theory of the results presented in this thesis is the joint asymptotic distribution of $u_n = n^{-1}L_Z'\varepsilon_n$ and $v_n = n^{-1}L_X'\varepsilon_n$. Therefore, we first prove Lemma 3.1 and use it to obtain the other results.

Write

$$w_n = n^{-\frac{1}{2}}\sum_{i=1}^{n}\varepsilon_i A_i, \tag{45}$$

where A_i is the $i-th$ column of the $(J+K)\times n$ matrix $(L_Z, L_X)'$, i.e.

$$A_i = \begin{pmatrix} I(z_i^s = z_1) \\ \cdots \\ I(z_i^s = z_J) \\ I(x_i^s = x_1) \\ \cdots \\ I(x_i^s = x_K) \end{pmatrix},$$

and each A_i contains exactly two non-zero elements, both unity. The components $V_{ni} = n^{-\frac{1}{2}}\varepsilon_i A_i$ are independent, but not *iid*, so we need the Lindeberg-Feller CLT (see, e.g. van der Vaart (1998), Section 2.8) to establish asymptotic normality for w_n. Under H_0 it is clear that $E(w_n) = 0$, as $E[u_{nj}] = n^{-1}E_Z[\sum_i I(z_i^s = z_j)E[\varepsilon_i|z_i^s]] = 0$ for each $j = 1, ...J$, and for each $k = 1, ...K$,

$$E[v_{nk}] = n^{-1}\sum_i E_X[I(x_i^s = x_k)E[\varepsilon_i|x_i^s]] = 0.$$

The covariance matrix is, under Assumptions 1 - 5,

$$\begin{aligned} Cov(w_n) &= E_{Z,X} E_{\varepsilon|Z,X} \left[n^{-1} (L_Z, L_X)' \varepsilon_n \varepsilon_n' (L_Z, L_X) \right] \\ &= \sigma^2 E_{Z,X} \left[n^{-1} (L_Z, L_X)' (L_Z, L_X) \right] \\ &= \sigma^2 \begin{bmatrix} D_Z & P \\ P' & D_X \end{bmatrix}. \end{aligned}$$

which is clearly finite. To verify the asymptotic normality of w_n it remains to verify the Lindeberg condition. To do so, note that

$$\|V_{ni}\|^2 = \left\| n^{-\frac{1}{2}} \varepsilon_i A_i \right\|^2 = 2n^{-1} \varepsilon_i^2,$$

so the required condition is that, as $n \to \infty$, for all $\epsilon > 0$,

$$2n^{-1} \sum_{i=1}^{n} E \left[\varepsilon_i^2 I \left\{ |\varepsilon_i| > \sqrt{\frac{n}{2}} \epsilon \right\} \right] \to 0.$$

This sum contains n identical terms, and so is equal to

$$2E \left[\varepsilon_1^2 I \left\{ |\varepsilon_1| > \sqrt{\frac{n}{2}} \epsilon \right\} \right]$$

which evidently (since the variance of ε_1, the integral over the whole line, is finite) converges to zero as $n \to \infty$. The Lindeberg condition therefore does hold, proving the Lemma.

For Lemma 3.2 recall that

$$\begin{aligned} z_{1n} &= \sqrt{n} C_K' L_X' L_Z (L_Z' L_Z)^{-1} \left(u_n - L_Z' L_X (L_X' L_X)^{-1} v_n \right) \\ &= C_K' L_X' L_Z (L_Z' L_Z)^{-1} \left[I_J, \;\; - L_Z' L_X (L_X' L_X^{-1} \right] w_n \end{aligned}$$

and the asymptotic distribution follows at once by Slutsky's Theorem.

Finally, we need to prove that Σ_{11} is positive definite. To do so, first observe that neither the support of Z, nor that of X, can affect the properties of w_n. That is to say, such properties must be *invariant to the support of* Z (or X), and hence hold for arbitrary support vectors z (or x). Now, the key matrix in Σ_{11} is $D_Z - PD_X^{-1}P' = D_Z - PD_X^{-1}D_X D_X^{-1}P'$.

Let z denote an *arbitrary* J-vector of hypothetical support points of Z, and consider the quadratic form in the matrix $D_Z - PD_X^{-1}D_X D_X^{-1}P'$:

$$z'D_Z z - \left(z'PD_X^{-1}\right) D_X \left(D_X^{-1}P'z\right). \tag{46}$$

The first term is $E_Z[Z^2] = E_X[E_{Z|X}[Z^2|X]]$ - the second moment of Z when its support is z. The term $D_X^{-1}P'z$ is the vector of conditional means $E[Z|X = x_k]$, $k = 1, ..., K$, so the whole second term is $E_X[E_{Z|X}[Z|X]^2]$. Hence, the complete expression in (46) can be interpreted as

$$E_X\left[E_{Z|X}\left[Z^2 - E_{Z|X}[Z|X]^2\right]|X\right] = E_X[Var(Z|X)] > 0,$$

i.e. the expectation of the conditional variance of Z given X when the support of Z is z. Since this must hold for all z, it follows that the matrix $D_Z - PD_X^{-1}P'$ is positive definite as required. The only exception would be if the conditional variance of Z given X vanished for each value of X, which we rule out. The result in Theorem 3.1 follows immediately.

Proof of Proposition 4

Under fixed alternatives (21):

$$p\lim_{n\to\infty} n^{-\frac{1}{2}}w_n = p\lim_{n\to\infty}\begin{pmatrix} u_n \\ v_n \end{pmatrix} = \begin{pmatrix} 0 \\ \mu \end{pmatrix},$$

since $p\lim_{n\to\infty} v_n = p\lim_{n\to\infty}\left(n^{-1}L_X'\varepsilon_n\right) = E_X\left[L_X'E\left[\varepsilon_n|X_n = L_X x\right]\right] = E_X\left[L_X'\eta\right] = \mu$ by assumption. Additionally,

$$\begin{aligned} p\lim_{n\to\infty} n^{-\frac{1}{2}}z_{1n} &= p\lim_{n\to\infty} C_K'L_X'L_Z(L_Z'L_Z)^{-1}\left(u_n - L_Z'L_X(L_X'L_X)^{-1}v_n\right) \\ &= -C_K'P'D_Z^{-1}PD_X^{-1}\mu. \end{aligned}$$

It follows immediately that $p\lim_{n\to\infty} n^{-1}T_n$ is a positive constant.

Proof of Lemma 3.3 and Theorem 3.2

Under the sequence of local alternatives of order $n^{-\frac{1}{2}}$:

$$\begin{aligned} E\left[w_n\right] &= E\left[n^{-\frac{1}{2}}(L_Z, L_X)'\varepsilon_n\right] \\ &= n^{\frac{1}{2}}E\left[n^{-1}(L_Z, L_X)'\varepsilon_n\right] \\ &= \begin{pmatrix} 0 \\ \mu \end{pmatrix}, \end{aligned}$$

since $E\left[n^{-1}L_X'\varepsilon_n\right] = n^{-\frac{1}{2}}\mu$. The covariance matrix remains the same as under H_0, because the departures from the null distribution are only local (see, for example Cox and Hinkley (1974), p. 317-318). Although the mean of the asymptotic distribution has changed, the asymptotic variance remains the same. The asymptotic distribution follows again from the Lindeberg-Feller CLT (we omit the details). The distribution of z_{1n} is then straightforward to obtain by Slutsky Theorem. Then it follows immediately that as $n^{-1}y'M_{L_X}y \rightarrow^p \sigma^{*2}$:

$$T_n \rightarrow^d \frac{\sigma^2}{\sigma^{*2}}\chi^2_{(K-1)}(\delta^2).$$

Since the noncentral chi-squared distribution is a Poisson-weighted mixture of central chi-squared distribution and given the fact that $\chi^2_{(K-1)}$ is equivalent to $Gamma(\frac{K-1}{2}, 2)$, we use the scaling property for the gamma distribution to obtain the required asymptotic distribution.

Proof of Theorem 3.3

The argument is given in the text. We have already shown that Σ is positive definite, so the positivity of the ω_j will follow from the positive definiteness of Ω.

The argument is similar to that used to prove that Σ is positive definite, which we have already shown in the Proof of Lemma 3.1. Since the inverse of a positive definite matrix is positive definite itself, we only have to prove that Ω is positive definite. The first term in

$$a'\Omega a = a'PD_X^{-1}D_XD_X^{-1}P'a - a'q_Zq_Z'a$$

is familiar (it also appears in the proof of positive definiteness of Σ) and is

$$E_X \left[E_{Z|X}[Z|X]^2 \right].$$

The second term is simply $\left(E_Z\left[Z\right]\right)^2 = \left(E_X\left[E_{Z|X}[Z|X]\right]\right)^2$. Hence, the complete expression is

$$E_X \left(E_{Z|X}[Z|X]^2 \right) - \left(E_X \left[E_{Z|X}[Z|X] \right] \right)^2 = Var \left(E_{Z|X}[Z|X] \right) > 0,$$

the variance of the conditional expectation of Z given X. Since this must again be true for every support vector a. It follows that the matrix $\Omega = PD_X^{-1}P' - q_Z q_Z'$ is positive definite as required. The only case, in which this term would be zero is when $E_{Z|X}[Z|X]$ is a constant i.e. the expectation of Z does not vary with X.

Appendix B: More on Monte Carlo simulations

This Appendix contains additional results of Monte Carlo simulations. For various combinations of numbers of support points of the instrument and endogenous regressor, Tables 16 and 17 provide empirical power for $J \geq K$ case, and Tables 18, 19 and 20 for $J < K$ case. To produce results in Table 19, the critical values were obtained by the first method described in Section 3.2.5, i.e. by simulating the distribution of a weighted sum of chi-square (1) variables. For results in Tables 18 and 20, we computed critical values using the chi-square approximation. Notice that both methods produced very similar results, but the chi-square approximation approach was substantially faster.

		η= 0.2			η= 0.5			η= 0.9		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
ψ	sample size	J=2								
0.35	100	1.60	6.05	11.90	2.85	9.55	17.55	12.75	30.50	42.35
	200	1.65	6.15	12.05	5.05	14.60	22.75	31.50	55.65	67.40
	400	1.95	7.25	13.80	9.05	23.15	33.80	64.65	85.35	91.45
	1000	3.80	11.15	18.30	26.55	50.70	63.70	98.35	99.70	99.80
0.7	100	1.90	8.20	14.65	11.95	27.55	39.40	69.50	86.90	93.05
	200	3.60	12.50	20.85	26.60	51.20	62.95	96.65	99.20	99.65
	400	6.30	17.50	27.15	57.50	80.10	87.75	100	100	100
	1000	19.30	39.85	51.35	97.45	99.40	99.65	100	100	100
		J=3								
0.35	100	1.45	6.55	12.60	2.90	11.25	18.70	17.70	38.35	50.05
	200	1.55	7.10	13.35	8.05	20.20	30.25	47.10	71.45	80.90
	400	2.70	10.45	17.20	16.65	36.50	49.30	85.55	94.95	97.50
	1000	6.60	17.90	26.30	51.30	76.15	84.55	100	100	100
0.7	100	2.15	7.95	15.05	11.70	29.40	42.20	61.10	83.90	90.75
	200	2.75	11.75	19.85	28.15	51.05	64.60	95.30	99.05	99.55
	400	6.15	18.90	29.80	61.45	82.35	90.35	100	100	100
	1000	21.05	42.10	55.15	98.00	99.60	99.85	100	100	100
		J=4								
0.35	100	1.20	5.95	11.35	3.30	11.40	19.20	21.15	41.80	54.25
	200	1.60	7.60	13.10	9.90	24.45	35.10	52.70	76.30	84.70
	400	3.35	10.70	18.15	20.55	43.25	55.65	91.65	97.40	98.80
	1000	7.10	19.15	27.45	61.60	82.60	90.35	100	100	100
0.7	100	1.65	6.90	13.75	7.35	20.75	31.20	37.85	67.15	80.50
	200	2.65	10.40	18.45	20.60	42.10	54.40	82.95	94.85	97.60
	400	4.25	14.80	24.40	48.65	70.95	82.30	99.60	100	100
	1000	15.20	33.05	46.90	92.85	98.40	99.25	100	100	100

Table 16: Proportion of rejections under the alternative hypothesis in the point identified model; K=2

		η= 0.2			η= 0.5			η= 0.9		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
ψ	sample size	J=3								
0.35	100	1.85	7.70	13.05	4.75	16.40	26.15	46.45	71.85	82.40
	200	1.90	8.25	14.55	9.45	24.95	36.95	87.50	96.60	98.20
	400	2.80	11.10	18.65	26.30	50.80	63.15	100	100	100
	1000	6.45	18.45	29.10	71.60	88.45	93.45	100	100	100
0.7	100	2.35	10.40	17.50	21.40	44.85	57.10	97.50	99.65	99.80
	200	5.75	16.40	25.75	56.10	76.35	84.10	100	100	100
	400	10.75	27.15	39.55	90.85	97.60	98.75	100	100	100
	1000	36.35	60.25	72.55	100	100	100	100	100	100
		J=4								
0.35	100	1.40	5.20	10.85	5.30	18.25	28.40	56.60	76.75	84.55
	200	2.15	8.05	15.80	13.55	32.20	44.25	91.30	97.90	98.85
	400	3.60	12.25	19.05	35.60	60.85	73.30	99.90	100	100
	1000	9.60	25.25	35.30	86.95	95.75	97.95	100	100	100
0.7	100	2.65	9.70	18.05	16.75	38.55	52.15	88.70	97.45	99.15
	200	4.80	14.05	23.50	47.10	69.10	79.85	100	100	100
	400	9.25	24.10	36.00	85.90	95.60	97.55	100	100	100
	1000	30.30	56.30	68.60	100	100	100	100	100	100
		J=5								
0.35	100	1.15	6.60	12.70	5.45	18.10	28.25	52.50	75.65	84.75
	200	1.80	8.90	16.30	15.10	35.00	48.10	92.50	98.00	99.10
	400	3.45	13.05	21.35	43.70	66.45	77.25	100	100	100
	1000	12.30	29.25	41.10	91.15	97.05	98.40	100	100	100
0.7	100	1.95	8.40	15.65	11.85	28.60	41.45	62.65	85.35	92.15
	200	2.75	10.90	19.55	30.55	56.55	68.45	97.95	99.45	99.75
	400	7.10	19.85	29.80	69.90	88.55	94.00	100	100	100
	1000	23.30	46.50	59.60	99.70	100	100	100	100	100

Table 17: Proportion of rejections under the alternative hypothesis in the point identified model; K=3

		η= 0.2			η= 0.5			η= 0.9		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
ψ	sample size	J=2								
0.35	100	1.75	7.25	12.85	3.75	11.95	20.00	33.40	58.95	70.25
	200	2.05	7.70	13.10	7.90	20.90	30.50	69.30	86.25	92.45
	400	2.80	8.60	14.70	17.65	36.65	48.65	95.95	99.10	99.90
	1000	5.05	15.80	24.05	50.90	74.30	83.65	100	100	100
0.7	100	3.45	12.25	19.65	28.15	53.10	64.80	99.35	99.95	100
	200	6.30	18.50	28.75	60.65	81.90	88.25	100	100	100
	400	14.00	30.50	42.75	92.70	97.65	99.05	100	100	100
	1000	39.85	65.10	76.45	100	100	100	100	100	100
		J=3								
0.35	100	1.05	5.50	11.00	2.60	9.70	18.45	20.25	40.10	52.10
	200	1.20	5.80	11.30	3.25	11.35	21.00	29.10	51.85	65.65
	400	1.25	5.85	12.30	3.30	11.90	22.10	35.60	61.60	75.90
	1000	1.40	6.80	12.45	3.75	13.70	24.05	40.85	67.90	79.10
0.7	100	0.95	5.65	11.35	2.60	11.35	18.90	15.85	40.85	57.15
	200	1.10	5.70	11.60	2.70	11.50	20.45	22.75	50.20	66.65
	400	1.35	5.80	11.75	3.35	14.70	27.90	41.10	74.65	88.10
	1000	1.45	6.15	13.25	6.75	29.25	53.90	86.35	97.60	99.35
		J=4								
0.35	100	1.05	5.30	10.65	2.10	8.20	13.85	11.95	26.55	38.00
	200	1.15	5.35	11.65	2.15	8.25	14.60	14.75	30.55	42.60
	400	1.20	6.05	11.90	2.20	8.35	14.75	15.20	32.70	46.05
	1000	1.35	6.50	12.35	2.25	8.40	15.85	15.45	34.60	49.25
0.7	100	0.95	4.65	9.65	1.15	6.00	11.50	3.10	11.15	18.45
	200	1.10	5.65	10.80	1.50	6.15	12.25	3.95	11.30	18.60
	400	1.35	5.70	11.10	1.65	7.35	13.30	4.05	11.80	20.10
	1000	2.00	6.40	11.55	2.30	7.90	13.65	4.15	12.60	20.85

Table 18: Proportion of rejections under the alternative hypothesis in the partially identified model; K=5; with approximated critical values

		η= 0.2			η= 0.5			η= 0.9		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
ψ	sample size	J=2								
0.35	100	1.95	7.25	13.45	4.20	13.30	21.90	34.40	58.30	70.05
	200	2.00	7.65	13.90	8.10	20.80	31.00	69.90	87.65	92.75
	400	2.15	9.35	16.90	17.95	39.65	52.45	96.65	99.15	99.70
	1000	5.60	16.75	25.70	50.85	73.80	82.95	100	100	100
0.7	100	3.65	12.85	21.15	27.45	51.90	65.40	99.25	99.95	100
	200	6.10	16.95	27.95	59.10	82.15	88.95	100	100	100
	400	14.00	31.95	43.75	93.10	98.15	99.50	100	100	100
	1000	41.95	66.50	77.20	100	100	100	100	100	100
		J=3								
0.35	100	1.25	5.35	10.85	2.40	11.35	19.45	21.05	40.95	53.45
	200	1.30	5.60	11.75	2.90	12.05	21.20	32.05	53.75	66.10
	400	1.35	6.00	11.85	3.40	12.40	21.25	37.25	62.05	74.60
	1000	1.50	6.55	12.25	4.00	13.10	21.65	44.00	68.50	79.90
0.7	100	1.15	5.45	11.15	2.10	9.95	18.70	17.05	40.90	55.65
	200	1.35	6.00	11.35	2.70	11.10	21.70	21.90	51.50	68.50
	400	1.45	6.20	11.75	3.25	14.50	26.75	43.10	74.50	87.25
	1000	1.65	7.25	13.75	6.95	31.20	54.35	87.30	97.80	99.65
		J=4								
0.35	100	0.90	5.10	10.20	1.75	7.15	13.05	13.25	29.00	40.30
	200	1.05	5.30	10.35	2.10	8.60	16.25	14.95	31.70	43.30
	400	1.15	5.30	10.95	2.25	9.10	16.30	15.65	34.65	47.45
	1000	1.40	6.70	12.30	2.30	9.75	17.10	16.55	35.60	49.40
0.7	100	0.85	4.95	9.35	1.35	6.35	12.40	2.35	7.45	17.55
	200	1.15	5.00	9.80	1.65	6.80	13.45	3.35	11.05	19.55
	400	1.30	5.25	9.85	1.90	7.25	13.60	3.65	11.80	19.95
	1000	1.55	6.85	11.60	1.95	7.65	14.20	3.75	11.95	20.10

Table 19: Proportion of rejections under the alternative hypothesis in the partially identified model; K=5

		η= 0.2			η= 0.5			η= 0.9		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
ψ	sample size	J=3								
0.35	100	0.85	5.95	11.90	3.55	12.90	21.70	37.40	61.55	74.25
	200	1.20	6.55	12.10	4.60	14.50	23.90	52.40	74.00	84.65
	400	1.30	6.75	12.45	4.95	17.20	29.40	63.50	83.60	91.40
	1000	1.45	6.80	12.55	5.90	20.15	32.60	74.20	90.20	95.05
0.7	100	1.35	6.60	12.60	4.20	18.55	31.30	49.05	78.55	89.75
	200	1.40	6.75	13.30	5.90	23.55	42.90	77.60	94.60	98.30
	400	1.75	7.00	14.50	13.75	45.30	67.90	97.20	99.65	99.85
	1000	2.10	11.10	21.65	62.10	95.05	95.05	100	100	100
		J=4								
0.35	100	0.85	5.35	10.40	2.60	9.70	17.55	23.80	43.70	58.40
	200	1.15	5.55	11.65	2.90	9.95	19.00	28.70	49.60	63.75
	400	1.35	6.20	11.85	2.90	10.60	19.45	32.35	57.35	70.80
	1000	1.50	6.50	12.85	3.20	11.35	20.55	37.90	65.10	78.45
0.7	100	1.25	5.45	10.85	1.95	7.70	14.40	6.10	17.50	30.10
	200	1.35	5.75	11.35	2.15	8.40	14.65	6.25	18.40	30.45
	400	1.40	6.05	11.75	2.25	8.45	15.30	6.70	20.35	34.65
	1000	1.45	6.35	12.55	2.85	9.30	17.05	8.80	27.95	44.75
		J=5								
0.35	100	0.90	4.80	10.20	1.70	9.05	15.70	13.40	29.60	41.80
	200	1.20	5.05	11.15	2.15	9.25	15.70	16.75	33.15	45.20
	400	1.40	5.15	11.90	2.40	9.65	16.25	17.45	34.85	46.65
	1000	1.55	6.30	12.60	2.75	9.90	17.90	18.75	38.25	50.85
0.7	100	0.95	4.70	9.90	1.20	5.85	10.25	2.50	8.70	14.85
	200	1.20	5.05	10.40	1.30	6.35	11.85	3.35	10.75	16.75
	400	1.35	5.70	11.35	2.35	6.90	12.75	4.45	11.50	17.35
	1000	1.70	6.40	11.55	2.45	8.65	13.15	6.05	11.85	18.55

Table 20: Proportion of rejections under the alternative hypothesis in the partially identified model; K=6; with approximated critical values

Appendix C: More on empirical applications

This Appendix contains additional results on testing for exogeneity of education in Angrist and Krueger (1991) dataset.

		critical values		
	R_n	1%	5%	10%
1930	0.4090	11.9661	8.0781	6.3941
1931	1.2137	12.8069	8.5118	6.6681
1932	10.5727	25.2758	16.9509	13.3383
1933	5.9939	20.7990	13.8662	10.8841
1934	7.9563	22.6713	15.2252	11.9953
1935	2.3214	21.9704	14.4460	11.2338
1936	8.3587	24.4301	15.8717	12.2359
1937	1.1957	18.1416	12.0474	9.4237
1938	3.9824	18.0718	12.2732	9.7452
1939	5.2154	21.7390	14.7925	11.7624
full cohort	47.3552	87.3456	74.1745	67.7834

Table 21: The value of the test-statistic and critical values for testing exogeneity of education for the 1930's cohort in Angrist and Krueger (1991) with race as an exogenous covariate

In Table 21, in all cases, the calculated value of the test statistic is lower than the critical values at any significance level. The null hypothesis of exogeneity of education is not rejected.

		critical values		
	R_n	1%	5%	10%
1940	2.9650	22.7767	14.9524	11.6072
1941	15.3934	28.2247	18.0054	13.7108*
1942	4.4649	30.9700	20.0039	15.3551
1943	99.6390	50.2469*	32.0104*	24.3425*
1944	-			
1945	42.6505	27.6174*	18.3706*	14.3926*
1946	17.1414	20.1883	13.4045*	10.4866*
1947	24.9107	32.2705	21.6397*	17.0442*
1948	31.3270	39.3458	26.1098*	20.4329*
1949	25.9228	28.5664	19.1873*	15.1345*
full cohort	-			

Table 22: The value of the test-statistic and critical values for testing exogeneity of education for the 1940's cohort in Angrist and Krueger (1991) with race as an exogenous covariate

In Table 22, note that the results for individuals born in 1944 are not available, as there are no observations with 1 year of education and race=1 in this subsample. Therefore, the result for the full cohort could not be obtained. The critical values lower than the value of the test statistic are indicated with stars. In some subsamples, the null hypothesis of exogeneity of education is rejected, as expected.

		critical values		
	R_n	1%	5%	10%
1930	1.0472	13.6545	9.4202	7.5827
1931	-			
1932	7.1465	14.2702	10.1066	8.2666
1933	6.9480	18.1071	12.4645	10.0219
1934	7.7902	21.1294	14.2104	11.2440
1935	3.2778	19.5331	13.4976	10.8701
1936	5.8818	24.2121	15.9695	12.4676
1937	1.2403	15.4051	10.7305	8.6858
1938	5.7139	17.4845	12.4219	10.1790
1939	-			
full cohort	-			

Table 23: The value of the test-statistic and critical values for testing exogeneity of education for the 1930's cohort in Angrist and Krueger (1991) with race and region of residence as exogenous covariates

In Table 23, note that there are no individuals in the sample born in 1931 with 20 years of education, race=1 and smsa=1, and no individuals in the sample born in 1939 with race=1, smsa=1 and years of education equal to 3 or 19. The results for the full cohort could not be calculated. The null hypothesis of exogeneity is not rejected in any subsample.

References

[1] Acemoglu, D., Autor, D. (2010) "Skills, tasks and technologies: implications for employment and earnings" NBER Working paper No. 16082

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

[3] Angrist, J.D., Krueger, A.B. (1991) "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, Vol. 106, No. 4, 979-1014

[4] Angrist, J.D., Krueger, A.B. (1999) "Empirical strategies in labor economics" in *Handbook of Labor Economics*, Volume 3A, Amsterdam and New York: North Holland

[5] Anscombe, F.J. (1952) "Large sample-theory of sequential estimation" *Proceeding of Cambridge Philosophical Society*, 48, 600-607

[6] Bai, Z.D. (1999) "Methodologies in spectral analysis of large dimensional random matrices, a review" *Statistica Sinica,* 9, 611-677

[7] Bech, K.M. (2011) "Partial identification: set identified parameters" MSc Dissertation, University of Southampton

[8] Becker, G.S. (1967) "Human capital and the personal distribution of income" Ann Arbor Michigan: University of Michigan Press

[9] Bekker, P.A. (1994) "Alternative approximations to the distributions of instrumental variable estimators" *Econometrica,* 62, 657-681

[10] Beresteanu, A., Molinari, F. (2008) ”Asymptotic properties for a class of partially identified models” *Econometrica*, 76, 763-814

[11] Bierens, H.J. (1990) "A consistent conditional moment test of functional form" *Econometrica*, 58, 1443-1458

[12] Blundell, R.W., Gosling, A., Ichimura, H., Meghir, C. (2007) "Changes in distribution of male and female wages accounting for employment composition using bounds" *Econometrica*, 75, 323-363

[13] Blundell, R.W., Horowitz, J.L. (2007) "A nonparametric test of exogeneity" *Review of Economic Studies*, 74, 1035-1058

[14] Blundell, R.W., Horowitz, J.L., Parey, M. (2012) "Measuring the price responsiveness of gasoline demand: economic shape restrictions and nonparametric demand estimation" *Quantitative Economics*, 3, 29-51

[15] Bound, J., Jaeger, D.A. (1996) "On the validity of season of birth as an instrument in wage equations: a comment on Angrist and Krueger's 'Does compulsory school attendance affect schooling and earnings?'" NBER Working Paper 5835

[16] Bound, J., Jaeger, D.A., Baker, R.M. (1995) "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak" *Journal of the American Statistical Association,* 90, 443-450

[17] Bronars, S.G., Grogger, J. (1994) "The economic consequences of unwed motherhood: using twin births as a natural experiment" *American Economic Review*, Vol. 84, No.5, 1141-1156

[18] Buckley, M.J., Eagleson, G.K. (1988) "An approximation to the distribution of quadratic forms in normal random variables" *Australian Journal of Statistics*, 30A, 150-159

[19] Cai, Z. (2007) "Trending time-varying coefficient time series models with serially correlated errors" *Journal of Econometrica,* 136, 163-188

[20] Canay, I.A., Santos, A., Shaikh, A.M. (2013) "On the testability of identification in some nonparametric models with endogeneity" *Econometrica,* 81, 2535-2559

[21] Card, D. (1995) "Using geographic variation in college proximity to estimate the return to schooling" in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp,* Toronto: University of Toronto Press, 201-222

[22] Card, D. (2001) "Estimating the return to schooling: progress on some persistent econometric problems" *Econometrica,* 69, 1127-1160

[23] Chernozhukov, V., Hong, H., Tamer, E. (2007) "Estimation and confidence regions for parameter sets in econometric models" *Econometrica*, 75, 1243-1284

[24] Chesher, A. (2003) "Identification in nonseparable models" *Econometrica,* 71, 1405-1441

[25] Chesher, A. (2004) "Identification in additive error models with discrete endogenous variables" CeMMap working paper, CWP11/04

[26] Chesher, A. (2005) "Nonparametric identification under discrete variation" *Econometrica,* 73, 1525-1550

[27] Chesher, A. (2010) "Instrumental variable models for discrete outcomes" *Econometrica,* 78, 575-601

[28] Cox, D.R., Hinkley, D.V. (1974) "Theoretical statistics" Chapman and Hall, London

[29] Darolles, S., Fan Y., Florens, J., Renault, E. (2011) "Nonparametric instrumental regression" *Econometrica*, 79, 1541-1565

[30] Das, M. (2005) "Instrumental variables estimators of nonparametric models with discrete endogenous regressors" *Journal of Econometrics*, 124, 335-361

[31] Deaton, A. (2010) "Instruments, randomization and learning about development" *Journal of Economic Literature*, 48, 424-455

[32] Delgado, M.S. (2011) "Nonparametric estimation with binary endogenous regressors: an application to voluntary pollution abatement" under review at the *Journal of Applied Econometrics*

[33] Devroye, L.P., Wagner, T.J. (1980) "Distribution-free consistency results in nonparametric discrimination and regression function estimation" *Annals of Statistics*, 8, 231-239

[34] Donald, S.G., Imbens, G., Newey, W.K. (2003) "Empirical likelihood estimation and consistent tests with conditional moment restrictions" *Journal of Econometrics,* 117, 55-93

[35] Fan, Y., Li, Q. (1996) "Consistent model specification tests: omitted variables, parametric and semiparametric functional forms" *Econometrica*, 64, 865-890

[36] Feng, G., Gao, J., Peng, B., Zhang, X. (2015) "A varying-coefficient panel data model with fixed effects: theory and an application to U.S. Commercial Banks" Working Paper 09/15, Department of Econometrics and Business Statistics, Monash University

[37] Florens, J.P., Malavolti L. (2003) "Instrumental regression with discrete endogenous variables" Working paper, GREMAQ, Universite des Sciences Sociales, Toulouse

[38] Frisch, R. (1934) ”Statistical Confluence Analysis” Publication No.5, University Institute of Economics, Oslo

[39] Freyberger, J., Horowitz, J.L. (2014) "Identification and shape restrictions in nonparametric instrumental variables estimation", Manuscript, October 2014 (revised version of CeMMAP working paper CWP15/12)

[40] Greblicki, W., Pawlak, M. (1987) "Necessary and sufficient consistency conditions for a recursive kernel regression estimate" *Journal of Multivariate Analysis,* 23, 67-76

[41] Greene, W.H. (1993) "Econometric Analysis", Second Edition, Macmillan Publishing Company, New York

[42] Griliches, Z. (1977) "Estimating the returns to schooling: some econometric problems" *Econometrica*, 45, 1-22

[43] Hahn, J., Ham, J., Moon, H.R. (2011) "The Hausman test and weak instruments" *Journal of Econometrics,* 160, 289-299

[44] Hall, P. (1983) "Chi squared approximations to the distribution of a sum of independent random variables" *Annals of Probability*, 11, 1028-1036

[45] Hall, P., Huang, L.S. (2001) "Nonparametric kernel regression subject to monotonicity constraints" *Annals of Statistics*, 29, 624-647

[46] Hall, P., Horowitz, J. (2005) ”Nonparametric methods for inference in the presence of instrumental variables” *Annals of Statistics*, 33, 2904-2929

[47] Hansen, C., Hausman, J., Newey, W.K. (2008) "Estimation with many instrumental variables" *Journal of Business and Economic Statistics,* 26, 398-422

[48] Harmon, C., Walker, I. (1995) "Estimates of the economic return to schooling for the United Kingdom" *American Economic Review*, 85, 1278-1286

[49] Hausman, J.A. (1978) "Specification tests in econometrics" *Econometrica*, 46, 1251-1272

[50] Horowitz, J.L. (2006) "Testing a parametric model against a nonparametric alternative with identification through instrumental variables" *Econometrica*, 74, 521-538

[51] Horowitz, J.L. (2011) "Applied nonparametric instrumental variables estimation" *Econometrica,* 79, 347-394

[52] Horowitz, J.L., Manski, C.F. (1995) ”Identification and robustness with contaminated and corrupted data” *Econometrica*, 63, 281-302

[53] Horowitz, J.L., Manski, C.F. (2000) ”Nonparametric analysis of randomized experiments with missing covariate and outcome data” *Journal of the American Statistical Association,* 95, 77-84

[54] Hu, Y., Lewbel, A. (2008) "Idenifying the returns to lying when the truth is unobserved" CeMMap working paper, CWP06/08

[55] Huber, P.J. (1973) "Robust regression: asymptotics, conjectures and Monte Carlo" *Annals of Statistics,* 1, 799-821

[56] Imbens, G.W., Manski, C.F. (2004) ”Confidence intervals for partially identified parameters” *Econometrica*, 72, 1845-1857

[57] Imbens, G.W., Newey, W.K. (2009) "Identification and estimation of triangular simultaneous equations models without additivity" *Econometrica,* 77, 1481-1512

[58] Iori, G., Kapar, B., Olmo, J. (2014) "Bank Characteristics and the Interbank Money Market: A Distributional Approach" working paper, Discussion Paper Series of the Economics Department, University of Southampton

[59] Imhof, J.P. (1961) "Computing the distribution of quadratic forms in normal variables" *Biometrika*, 48, 419-426

[60] Johnson, N.L., Kotz, S. (1970) "Distributions in statistics: continuous univariate distributions" Wiley, New York

[61] Johnson, N.L., Kotz, S., Balakrishnan, N. (1994) "Continuous univariate distributions. Volume 1", Second Edition, John Wiley & Sons Inc., New Jersey

[62] Kane, T.J., Rouse, C.E. (1993) "Labor market returns to two- and four-year colleges: is a credit a credit and do degrees matter?" NBER Working Paper 4268

[63] Koopmans, T.C. (1949) ”Identification problems in economic model construction” *Econometrica*, 17, 125-144

[64] Lavergne, P., Patilea V. (2008) "Breaking the curse of dimensionality in nonparametric testing" *Journal of Econometrics*, 143, 103-122

[65] Lavergne, P., Vuong, Q. (2000) "Nonparametric significance testing" *Econometric Theory*, 16, 576-601

[66] Ledoit, O., Wolf, M. (2004) "A well-conditioned estimator for large-dimensional covariance matrices" *Journal of Multivariate Analysis,* 88, 365-411

[67] Magdalinos, M.A. (1992) "Stochastic expansions and asymptotic approximations" *Econometric Theory*, 8, 343-367

[68] Manski, C.F. (1989) "Anatomy of the selection problem" *The Journal of Human Resources*, 24, 343-360

[69] Manski, C.F., Pepper, J. (2000) "Monotone instrumental variables: with an application to the returns to schooling" *Econometrica*, 68, 997-1010

[70] Manski, C.F., Tamer, E. (2002) "Inference on regressions with interval data on a regressor or outcome" *Econometrica*, 70, 519-547

[71] Marschak, J., Andrews, W.H. (1944) "Random simultaneous equations and the theory of production" *Econometrica*, 12, 143-203

[72] Mathai, A.M. (1982) "Storage capacity of a dam with gamma type inputs" Annals of the Institute of Statistical Mathematics, 34, 591-597

[73] Matzkin, R.L. (2003) "Nonparametric estimation of nonadditive random functions" *Econometrica,* 71, 1339-1376

[74] Matzkin, R.L. (2007) "Nonparametric identification" in *Handbook of Econometrics*, Volume 6, Part B, 5307-5368

[75] Muirhead, R.J. (1982) "Aspects of multivariate statistical theory" John Wiley & Sons Inc., Hoboken, New Jersey

[76] Nadaraya, E. (1964) "On regression estimators" *Theory of Probability and Applications*, 9, 157-159

[77] Newey, W.K. (1985) "Maximum likelihood specification testing and conditional moment tests" *Econometrica*, 53, 1047-1070

[78] Newey, W.K., Powell, J.L. (2003) "Instrumental variable estimation of nonparametric models" *Econometrica,* 71, 1565-1578

[79] Oman, S.D., Zacks, S. (1981) "A mixture approximation to the distribution of a weighted sum of chi-squared variables" *Journal of Statistical Computation and Simulation*, 13, 215-224

[80] Pagan, A., Ullah, A. (1999) "Nonparametric econometrics" Cambridge University Press

[81] Phillips, P.C. (1989) "Partially identified econometric models" *Econometric Theory*, 5, 181-240

[82] Portnoy, S. (1984) "Asymptotic behaviour of M-estimators of p regression parameters when p^2/n is large. I Consistency" *Annals of Statistics,* 12, 1298-1309

[83] Portnoy, S. (1984) "Asymptotic behaviour of M-estimators of p regression parameters when p^2/n is large. II Normal approximation" *Annals of Statistics,* 13, 1403-1417

[84] Rao, C.R. (1986) "Some recent results in signal detection" Research Report 23, University of Pittsburgh

[85] Reinsch, H. (1967) "Smoothing by spline functions" *Numerische Mathematik*, 10, 177-183

[86] Robbins, H. (1948) "The asymptotic distribution of the sum of a random number of random variables" *Bulletin of the American Mathematical Society,* 54, 1151-1161

[87] Roehrig, C.S. (1988) "Conditions for identification in nonparametric and parametric models" *Econometrica*, 56, 433-447

[88] Romano, J.P., Shaikh, A. (2008) "Inference for identifiable parameters in partially identified econometric models" *Journal of Statistical Planning and Inference*, 138, 2786-2807

[89] Rosen, A.M. (2008) "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities" *Journal of Econometrics*, 146, 107-117

[90] Rosenblatt, J., Bell, S. (1999) "Mathematical analysis for modeling" CRC Press, Boca Raton

[91] Santos, A. (2010) "Inference in nonparametric instrumental variables with partial identification" working paper, University of California, San Diego

[92] Severini, T.A. (2005) "Elements of distribution theory" Cambridge University Press

[93] Severini, T.A., Tripathi, G. (2006) ”Some identification issues in nonparametric linear models with endogenous regressors” *Econometric Theory*, 22, 258-278

[94] Sheil, J., Muircheartaigh, I. (1977) "Algorithm AS106: The distribution of non-negative quadratic forms in normal variables" *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26, 92-98

[95] Solomon, H., Stephens, M.A. (1977) "Distribution of a sum of weighted ch-square variables" *Journal of the Americal Statistical Association*, 72, 881-885

[96] Staiger, D., Stock, J.H. (1997) "Instrumental variables regression with weak instruments" *Econometrica,* 65, 557-586

[97] Tamer, E. (2003) ”Incomplete bivariate discrete response model with multiple equilibria” *The Review of Economic Studies*, 70, 147-167

[98] Tripathi, G., Kitamura, Y. (2003) "Testing conditional moment restrictions" *Annals of Statistics*, 31, 2059-2095

[99] van der Vaart, A.W. (1998) "Asymptotic statistics" Cambridge University Press

[100] van Hasselt, M. (2010) "Many instruments asymptotic approximations under nonnormal error distributions" *Econometric Theory,* 26, 571-583

[101] Watson, G.S. (1964) "Smooth regression analysis" *Sanikhya: The Indian Journal of Statistics*, Series A 26 (4): 359–372

[102] Yohai, V.J., Maronna, R.A. (1979) "Asymptotic behaviour of M-estimators for the linear model" *Annals of Statistics,* 7, 258-268

[103] Zheng, J.X. (1996) "A consistent test of functional form via nonparametric estimation techniques" *Journal of Econometrics*, 75, 263-289