REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2005 Name of Author MARTINEZ-SANCHIS, E.

Essays on Identification and Estimation of Structural Parametric and Semiparametric Models in Microeconometrics

Elena Martínez-Sanchís

A Dissertation submitted to the Department of Economics
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
University College London

August 2005, London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Contents

List of Tables

List of Figures

Abstract

This thesis focuses on identification and estimation of structural parametric and semiparametric models in microeconometrics. The analysis of the conditions under which -in the context of an econometric model- data can be informative about the parameters of interest of an economic process is essential and must be of high priority in any econometric work. When considering models with which to identify interesting features, emphasis should be placed on imposing the minimum set of restrictions in order to achieve identification, since inappropriate restrictions may lead to inconsistent estimates of the parameters of interest. For this reason in the literature one finds that some attention has been paid to relaxing parametric distributional assumptions on the unobservables or functional forms of the relationships between observables and unobservables.

To begin with, I examine how the parameters of interest of a general class of models can be identified and then estimated when not all of the relevant variables are jointly observed in the same dataset. To do so, the existence of an additional data set with information on both the missing variables and on some common variables in the original data set is necessary.

I then move on to an analysis of the identification of the preference parameters in a discrete choice demand model in which individuals only derive utility from the characteristics of the goods they consume. I discuss how this particular model makes the estimation of these parameters feasible without imposing distributional assumptions in the errors even if the number of goods in the choice set is very large.

Finally, I consider the comparison of nonparametric regression curves between different samples. I propose to estimate the parameters that explain these differences between the conditional mean functions by using an estimator developed in the semiparametric literature which avoids the computational problems faced by the previously proposed estimators.

Acknowledgements

I extend my most sincere thanks and appreciation to Hide Ichimura and Richard Blundell, my supervisors. Hide has been an incredibly important source of inspiration and learning during my years at UCL and in particular for this thesis. He was always accessible and I really learned a lot from my discussions with him during hours until he was satisfied with the solution we obtained. He was amazingly caring about all my learning process as a PhD student with a very clear idea in mind of what I should learn before undertaking certain tasks, which made me feel always very safe. Not only I learnt about econometrics from him but also about the right attitude as a researcher and academic. I will always be very grateful for all his support during this time.

I am really indebted to Richard whose optimism was always very helpful during my time at UCL and whose advises were crucial for this work. His deep understanding about economics and econometrics makes that any conversation with him always very fruitful and even his attendance to seminars was a source knowledge. I consider myself incredibly lucky to have spent so many years close to them and having learned so much from their unparalleled intellect.

I would like to thank everyone at IFS for their help and support. Not only they believed in my from the beginning giving me financial support, but also my experience here it has allowed me to see many interesting projects and how rigorous research can be used in practice in the policy analysis.

I am extremely grateful to the Fundacion Ramon Areces for the financial support it provided.

I am really indebted to cemmap which "adopted" me as a PhD scholar after two years at IFS. I am delighted of having shared some time , even though only as a student, with this excellent and very active group of econometricians. Their dedication to what should be a good practice in econometrics always impressed me. I would like to thank in particular Andrew Chesher, Ian Crawford and Costas Meghir for many interesting discussions, for being very helpful during my job search and for their incredible and permanent sense of humour.

I would like to thank in particular Simon Lee for being always available to receive questions and give advises to me at any time I needed it during these years.

During my time at UCL I spent many hours in study groups in math and statistics lead by Hide Ichimura, Simon Lee and Ian Preston. I thank all of them for the time they

dedicated to us and their patience when we were often stuck in theorems and proofs. The material I learnt from these groups definitely helped me in many stages of my research. My partners in these study groups (Carlos Henrique Courseil, Miguel Fogel and Pierre Hoonhout) were also a good incentive to keep working and fight against the solitude that this work involves. We discussed together during many hours until all of us understood everything and agreed with the same idea. I found in them also very good friends and they have been a huge support in many areas of my life. I really hope that the distance does not stop these discussions and many other happy moments together.

My time in London would not have been the same without Renata, Lars, Giovanni, Marco, Marcos and Arianna. Renata was my very first contact in London (even before I arrived) and I was always impressed by her hospitality and kindness. I must thank her for having enjoyed so many good meals from her cooking and many moments of laugh. I would never be able to thank Lars for all the very clever advises he has given me since he arrived and for many hours he spent with the problems of my papers. He has been always a source of good humour in all the moments we all shared together. Giovanni has been my officemate all this time, even though in an open space, he has been the one always close to me. It would not have been the same without his enthusiasm about economics and his eagerness in asking me to go out to the pub every time he was bored or demotivated which always coincide with my most productive times. Marco and Arianna have been my best flat mates of all (!) the ones I had in London. I would always bring in my memories the dinners and breakfast, respectively, that we enjoyed together. Marcos has been very helpful in many moments, workwise or lifewise, and I must thank him for his happiness because it is really contagious and I got many of his happy waves for sitting in front of him.

After almost 5 years in London I am amazed I can still be so close to all the friends I left behind in Spain. They have been really a very solid support in my low moments in London. Their first emails in the morning from all of them made me always smile from the first time everyday. Ana Vidal, Ana Bartolome, Claudia, Cristina, Dani, David, Diego, Edu (s), Eva (s), Hipolito, Javi, Jose, Juan (s) Lopez, Lia, Lola, Lorea, Manu, Manuel, Maria, Martin, Michael, Rosa, Vanessa Polo, Vicente ... I felt each of them always close to me and pushing me towards the end. I should specially mention Chitina, Ruth and Vanessa, whose friendship since we were children is still today a present for me. Fer was the essential support during almost all my years in London even though this experience was a huge sacrifice for both of us. It was almost impossible to continue without that

support and this thesis was at risk many times. However, I think it is fair to say that if I think of my happiest moments during these years in London, I see myself close to him.

I do not think I would have managed without Emla. She is the most incredible person I have met during my stay here. Her sweetness is unmatchable and she was always there for no matter what, to listen, to help me, to hug me and support me. I cannot even think of how sad I will be every lunch without her. I will bring her friendship always close to me.

One of the most important things I take with me of my stay in London is to have been able to meet Vero more in depth. She will never realize of how important she has been for me. She made bearable the hardest moments of this thesis being always there for me, being always so positive and believing in me more than myself and showing how much she cared about me. Her friendship is something to treasure. Her daily phone calls are a need for me now and I do not want to get used to live without them.

Sonia is the rock I depend on. She is the closest image I can think of the sister I never had. She gave me the strength that I needed to continue and she told me the words I needed to listen. I hope one day she realises of what incredible person she is.

Most of all, I thank my parents. It is incredible how close I have been feeling to them all these years away from home. At the hardest times, this separation was very painful for them too and they always tried to hide it in order not to make me feel worse. They have been always the source of my strength, the definite help to believe I could make it and the ones who can notice better -despite the distance- if there are some difficulties around. I would like to thank them from the bottom of my heart for all the caring love they have always given to me and for their deep understanding in the most stressful periods. I dedicate this work to them, with all my love.

Declaration

1. No part of this thesis has been presented to any University for any degree.

2. Chapter Two was undertaken as joint work with Hide Ichimura

Elena Martínez-Sanchís

Chapter 1

Introduction

In the forthcoming chapters, identification and estimation of both structural parametric and semiparametric models are considered. The minimum set of restrictions that should be imposed in defining an econometric model in order to achieve identification of the parameters of interest of an economic process, has received much attention in the econometric literature. In this literature, the main emphasis has been placed on relaxing the functional form restrictions on both the relationship between observables and unobservables, and on the distributional assumptions on the unobservables. However, a fully nonparametric approach to modelling the relationship between the observables in the model imposes enormous data requirements when the number of variables is high. Semiparametric modelling has become a very attractive tool to reduce this "curse of dimensionality" by imposing some parametric restrictions on the model while assuming that the functional form of many other parts of the model are unknown.

This thesis focuses on three very different aspects of this attempt in the literature to reduce the functional form restrictions needed in order to identify and estimate certain economic features of interest. The data one has access to determines which variables - amongst those that should be included in a model- must be considered to be unobserved to the econometrician. For a wide class of models, some stochastic restrictions on these unobservables -which often imply distributional assumptions- are usually imposed to identify and estimate the parameters of interest. However, there might exist information on the distribution of these unobserved variables in a different data set which would allow one to nonparametrically identify its distribution and consequently relax some of these restrictions (Chapter two). To study the economic interpretation of the assumptions placed on

the relationship between the observables and unobservables and the particular specification of the unobservables, is crucial. Chapter three analyzes this issue in the context of a discrete choice model used as a structural model of demand to obtain preference parameters. The last chapter studies a number of issues relating to the estimation of a specific semiparametric model in which linear and index restrictions are imposed in the unknown conditional mean function, in order to include a variable that determines to which group or sample each observation belongs. A brief description of each chapter follows.

In chapter two I study an incomplete data problem in which the data set available contains only a strict subset of the list of variables that are relevant in an empirical analysis. The proposed method presumes the existence of another data set which contains a subset of variables in the original data set as well as the variables that are missing from the original data set. The interest is on identifying the true effects on the dependent variable of the common variables between two data sets, once the omitted variable bias that arises from the missing variables is controlled for. Additionally, one would like to identify the effect of these missing variables on the dependent variable even if they are never jointly observed. In other words, one would like to find under which conditions one is able to identify the same parameters as in a complete data framework. I show that, for a wide variety of problems, having access to a complementary data set combined with a parametric structural restriction and some joint variation on the variables in the auxiliary data set, may be sufficient to identify the parameters of interest. The main advantage of the framework I use here is that it extends the existing method for linear-in-parameters models in the incomplete data literature to more general models. This generality is not assumed at any cost since when the model under consideration is nonlinear, the identification conditions are model-specific and it is difficult to provide conditions for global identification at this level of generality. For this reason, regarding the identification results, I focus on the parametric and semiparametric binary choice model which are leading examples of nonlinear models in econometrics. However, for the very general class of models I discuss the estimation of the parameters when no parametric restriction are imposed on the joint distribution of the variables observed in the complementary data set. General conditions under which the proposed estimators exhibit consistency and asymptotic normality are developed.

In chapter three I investigate the identification of preferences in a model which avoids the counterintuitive properties in policy analysis related to the introduction of new goods implied by standard discrete choice models that are used as structural models of de-

mand. The model analyzed in this chapter does not assume product specific unobserved tastes which are usually modeled in the standard approaches as an unbounded random term independent and identically distributed across individuals and alternatives. The implications is that in our model individuals derive utility from a finite set of product characteristics. Under these assumptions, the main contribution of this chapter is to relax the distributional restrictions on the random tastes over product characteristics without relying on the dimension of the product space. This contrasts with previous contributions in semiparametric multinomial discrete choice models in which the index dimension of the choice probabilities becomes intractable when the number of products in the market is high. When consumer-level data is available, I state the conditions under which the preference parameters are identified up to a scalar constant both when the distribution of the unobserved individual attributes is assumed to be known and unknown.

In chapter four I discuss the estimation of a semiparametric model in which parametric transformations exist that explain the differences between nonparametric conditional mean functions of different groups or samples. Instead of using a fully nonparametric approach, this model assumes a particular specification to explain how the variable defining the group or the subsample affects the endogenous variable. Thus, there exist two parameters explaining the differences between the nonparametric means for different groups: one implies a horizontal shift and a change of slope and the other parameter shifts the unknown functions vertically. This model is denoted in the literature as the shape invariant model and is equivalent to a single index model with a partial linear term in which one of the conditioning variables is the group or sample discrete variable. The previous estimators introduced in the literature for the shape invariant models face the computational difficulty that their objective functions only attain a local minimum at the true value of the parameters so that they must rely on intensive computational methods. We argue that the existing semiparametric least squares estimator constitutes a natural way of estimating the parametric transformations, along with solving this computational problem. To reduce this burden is important because this estimator makes the comparison of nonparametric curves with respect to more than one variable defining the group or the subsample, feasible. The asymptotic properties of this estimator in a semiparametric model with multiple equations are established. We also discuss the possibility of giving different weights to each combination of the equations and the optimal weight that makes the estimator efficient. Finally, a shape invariant model arises in the estimation of Engel curve relationships where the demographic composition is taken into account and I com-

pare the performance of the estimators discussed in this work with the estimates obtained from the British Family Expenditure Survey.

Chapter 2

Identification and Estimation of GMM Models by Combining Two Data Sets

2.1 Introduction

It is often the case that we do not have an ideal data set that contains all of the relevant variables that should be used in an empirical piece of work. In some cases a set of relevant variables is incompletely observed whilst in some other cases the variables are completely missing. As a consequence, empirical studies based on analogous data might yield incomparable results because the implicit models used are incomparable when the same variables in those data sets differ in their definition or a different set of conditioning variables is used. The object of the present research is to develop a general method that allows us to estimate a common model even when an available data set may be incomplete in itself.

We consider a special case of incomplete data problems in which a data set at hand contains only a strict subset of the list of variables relevant for empirical analysis. We tackle the problem by assuming that there is another data set which contains a subset of variables in the original data set as well as the missing variables in the original data set. We show that this assumption, combined with a parametric structural assumption and some joint variation assumption on the variables in the auxiliary data set, are often sufficient to identify the effects of missing variables as well as those of the non-missing variables in the

parametric structural relationship. We propose estimators for the identified parameters and establish the asymptotic properties of the estimators.

An empirical framework that allows one to consider a combination of two data sets may be important for many applications. A survey of individual finances might have detailed information on wealth but scarce information on consumption or labour market behavior. In fact, this is the case in the BHPS survey in the UK and the PSID in the US. On the other hand budget surveys, such as the CEX in the US and the FES in the UK, have rich information on individual decisions but have little or poor quality information on wealth. Both types of data sets could be complemented to estimate structural models in which both consumption and wealth are the relevant variables. Birth certificate data, health surveys, or consumer scanner data may be fruitfully combined with more general surveys as well to complement their general lack of information on household income.

Analysis under missing observations is a significant research area. An analogous problem to ours has been addressed and solved for the linear-in-parameter models by Glasser (1964), Gourieroux and Monfort (1981), Angrist and Krueger (1992) and Arellano and Meghir (1992).[1],[2] See a useful survey by Little (1992) for early works[3]. For non-linear in parameter models various identification issues and estimation procedures have been insightfully discussed by Ridder and Moffitt (2003). Our problem shares some properties with the literature that uses additional samples to correct for the measurement error in the regressors. The main difference with respect to our assumptions is that they do not assume joint observation of the missmeasured and variable measured without error. This makes that identification needs to rely on different conditions (See Hu and Ridder (2003), Chen, Hong and Tamer (2004), Schennach (2004)).

We develop a general framework that covers a wide class of non-linear models although the aim of the paper is not to provide identification conditions for each model belonging to this class. It is hard for to give sufficient conditions for global identification in a very

[1] See Carroll and Weil (1994), Lusardi (1996), Currie and Yelowitz (1997) and Dee and Evans (1997) for applications.

[2] Imbens and Lancaster (1994) study how to combine cross sectional data with information on (aggregate) population moments. We assume however the existence of two micro data sets (i.e. both of them with individual information).

[3] Early references also include Rubin (1974), which establishes maximum likelihood factorization methods dealing with missing data problems. These methods however do not allow one to identify the effect of the missing regressor.

general non-linear model (similarly to the identification in GMM non-linear model with complete data). We discuss though the general conditions that are required to compute the identifying moment condition with the incomplete data, and therefore to compute the estimators.

This level of generality allows one to establish the asymptotic distribution theory for a wide class of estimators when two data sets are needed in the estimation including estimators for linear and non-linear regression models, generalized method of moments (GMM) estimators and the maximum likelihood estimators (MLE). The results we obtain differ from the previous contributions in the literature because the sample analogue of the moment condition does not need to be separable in observations belonging to each of both data sets.

In order to provide specific conditions for global identification, we focus on a subclass of those non-linear models covered by the general framework by studying the identification of the parametric and semiparametric binary discrete choice model.

The identification results for the binary choice model complement the results of Manski and Tamer (2003). They consider the binary choice model with non-missing regressors with one incompletely observed regressor in the sense that the regressor value is known only to lie in an interval. Without assuming access to a complementary data set, but assuming that the variable affects the choice probability monotonically, they show point identification of the effect of non-missing and missing variables only when there is a positive probability of complete data, otherwise they only achieve partial identification. We show that for parametric models when there is a complementary data set, we can allow for more than one missing exogenous variable and we provide sufficient conditions under which coefficients of the missing regressors are identified. For the semiparametric binary choice model, some additional conditions need to be imposed on the distribution and the support of the common regressors in order to identify the parameters up to scale.

After explaining the framework of the incomplete data problem that we consider in section 2.2 we discuss identification issues and present a general estimation method in sections 2.3 and 2.4, respectively. The asymptotic theory for this framework is established in section 2.5.Monte Carlo simulation results are presented and discussed in section 2.6. Section 2.7 concludes.

2.2 General Framework for Combining Two Data Sets

All random (column) vectors and their realizations are denoted by upper and corresponding lower case letters respectively. Endogenous and exogenous random vectors are denoted by Y and X with subscripts respectively. We assume access to two data sets, data sets 1 and 2. Data set 1 contains observations on the random vector (Y_1, Y_c, X_1, X_c) and data set 2 contains observations on the random vector (Y_c, Y_2, X_c, X_2). Assume that $(Y_1, Y_c, Y_2, X_1, X_c, X_2)$ is needed to carry out a standard empirical analysis. Let $Z_1 = (Y_1, Y_c, X_1)'$, $Z_c = X_c$, and $Z_2 = (Y_2, X_2)'$ be random vectors of length m_1, m_c and m_2, respectively. Random vector Z_c includes only those exogenous variables that are common to both data sets and random vector Z_2 includes those variables that are exclusively observed in data set 2. The distribution of the missing variables in data set 1 conditional on the common variables, in particular conditional on the common exogenous variables, is assumed to be unknown but the second data set can be used to identify it. Thus, the distribution of interest to be identified from data set 2 is the conditional distribution of Z_2 given Z_c which we assume is dominated almost surely in Z_c by a fixed measure μ so that there is a conditional density $\gamma(z_2|z_c)$ for almost all z_c in the support of Z_c.

We are interested in estimating the structural parameter $\theta_0^C \in \Theta \subset R^K$ defined via the following moment conditions that can be computed with complete data (i.e. when Z_1, Z_c and Z_2 are jointly observed in the same dataset)

$$E\{\psi(\rho(Z_1, Z_c, Z_2, \theta); \theta)|X_1, X_c, X_2\} = 0 \quad \text{almost surely in } X_1, X_2, X_c \text{ iff } \theta = \theta_0^C \tag{2.1}$$

where function $\rho : R^{m_1} \times R^{m_2} \times R^{m_c} \times \Theta \to R^S$ and $\psi : R^S \times \Theta \to R^T$ where T is the number of moments and $T \geq K$.

However, if the data is incomplete and $Z_c = X_c$ are the only exogenous variables in common between both data sets, then we could only use conditional moments on Z_c. The conditional moment on the common exogenous regressors Z_c that is directly implied by (2.1) can only be computed with the data we have assumed we have access to if Z_2 and Z_1 are independent conditional on Z_c. Being able to write the moment condition is a necessary condition to identify θ_0 and without the mentioned conditional independence is not possible to do so, since the conditional distribution of (Z_1, Z_2) given Z_c cannot be identified from the incomplete data. This conditional independence assumption is however a strong assumption which would impose a strong restriction on the true value of

the parameters θ_0.

An alternative to the conditional independence assumption, which is used in this work, is to assume a conditional moment on Z_c which can be computed with the incomplete data. Then, we study the restrictions that need to be imposed on functions ψ, ρ_a and ρ_b in order for the parameter that it is identified through the conditional moment on Z_c with incomplete data to be the same as the true value of the parameter that moment (2.1) identifies.

The general framework that we consider defines the structural parameter $\theta_0^I \in \Theta \subset R^K$ via the following conditional moments given random vector Z_c, which are identified with incomplete data:

$$E\left\{h(Z_1, Z_c; \theta)|Z_c\right\} = 0 \text{ almost surely in } Z_c \text{ iff } \theta = \theta_0^I \tag{2.2}$$

where

$$h(z_1, z_c; \theta) = \psi(q(z_1, z_c, \theta); \theta) \tag{2.3}$$

$$q(z_1, z_c, \theta) = \int \rho(z_1, z_c, z_2, \theta) g(z_2|z_c) d\mu \tag{2.4}$$

The function $g\left(z_2|z_c\right)$ is typically defined via $\gamma\left(z_2|z_c\right)$. We motivate the formulation below but first note that in general both functions q cannot be interpreted as conditional mean functions of ρ given Z_c without further assumptions. This is because we do not use the conditional distribution of Z_2 given Z_c *and* Z_1 to integrate out Z_2. This alternative is impossible with the type of data we have assumed to have access to. Note that the moment conditions in (2.2) are not the only moments that can be identified given the model with complete data and the data sets in our hands.

Therefore, the identification problem we want to pursue in this paper is under which conditions we can ensure that the value of the parameter that uniquely solves moment condition (2.1) is the same as the parameter that solves the moment condition with incomplete data in (2.2) (i.e. $\theta_0^C = \theta_0^I$)

The framework in (2.2)–(2.3) covers general parametric conditional probability models, non-linear regression models and some generalized method of moment (GMM) models by defining for each particular case the form of functions ψ, ρ_a and ρ_b and the variables that should be included in Z_1, Z_c and Z_2.

2.2.1 Parametric models

Suppose a parametric conditional probability model is specified by

$$f(y_1, y_c, y_2|x_1, x_c, x_2; \theta)$$

Integrating out y_2, the model implies a parametric model $f(y_1, y_c|x_1, x_c, x_2; \theta)$. If there is no x_1, i.e. if all conditioning vector is observed jointly in the second data set, then integrating out x_2 using $g(x_2|x_c)$ would yield a parametric conditional probability model for the data that it is observed in the first data set $f(y_1, y_c|x_c; \theta)$.[4] The moment condition that identifies the true value of the parameters θ_0 is the score of the likelihood function using the conditional probability model with complete data $f(y_1, y_c|x_c, x_2; \theta)$. Therefore, the h function in this case corresponds to the first order condition of the maximum likelihood estimator (MLE) using the implied conditional probability model $f(y_1, y_c|x_c; \theta)$ for incomplete data[5] Let $g(x_2|x_c)$ denote the density of X_2 given X_c with respect to μ. The functions defined for the general framework take that following forms to identify the parameters imbedded in the conditional parametric model just outlined:

$$\begin{aligned}
\rho(y_1, y_c, x_c, x_2; \theta) &= \begin{bmatrix} \int \nabla_\theta f(y_1, y_c, y_2|x_c, x_2; \theta)\, dy_2 \\ \int f(y_1, y_c, y_2|x_c, x_2; \theta)\, dy_2 \end{bmatrix} \\
\psi(\rho(.; \theta)) &= \rho_1(.; \theta)/\rho_2(.; \theta) = \nabla_\theta f(y_1, y_c|x_c, x_2; \theta)/f(y_1, y_c|x_c, x_2; \theta) \\
q(y_1, y_c, x_c; \theta) &= \int \rho(y_1, y_c, x_c, x_2; \theta)\, g(x_2|x_c) d\mu = \\
&= \begin{bmatrix} \nabla_\theta f(y_1, y_c|x_c; \theta) \\ f(y_1, y_c|x_c; \theta) \end{bmatrix} \\
h(y_1, y_c, x_c; \theta) &= \psi(q(.; \theta)) = \\
&= q_1(.; \theta)/q_2(.; \theta) = \nabla_\theta f(y_1, y_c|x_c; \theta)/f(y_1, y_c|x_c; \theta)
\end{aligned}$$

where the subscripts of ρ and q denote the elements of these vectors. It is not clear if the original parameters are still identified after integrating out certain variables. We explicitly address this issue for some specific cases in section 2.3. Note that although the following moment condition

$$\int \psi(\rho(y_1, y_c, x_c, x_2; \theta))\, g(x_2|x_c) d\mu = 0$$

[4] The analogous likelihood can be formulated replacing the role of two data sets if in fact the data sets are symmetric as formulated above.

[5] We assume $f(y_1, y_c|x_c, \theta)$ to be dominated for each θ in its neighborhood by an integrable function with finite integral so that integration and differentiation can be interchanged.

could be computed with this setting, it does not arise from the maximisation of the log likelihood with incomplete data.

2.2.2 GMM

Define $Y = (Y_1, Y_c, Y_2)'$ and $X = (X_1, X_c, X_2)'$. In a GMM framework, the structural parameter θ_0 is defined by

$$E[\psi(Y, X; \theta)|X] = 0 \text{ almost surely in } X \text{ iff } \theta = \theta_0 \tag{2.5}$$

where X is a set of instrumental variables and ψ may have some exclusion restrictions so that not all elements of X need to appear directly as arguments in ψ.

We assume that ψ takes the following form and some elements of Z_c are excluded as arguments:[6]

$$\psi(z_1, z_c, z_2; \theta) = \rho_1(z_1, z_c, \theta) - \rho_2(z_c, z_2, \theta). \tag{2.6}$$

Under this separability assumption, the moment condition (2.5) can be integrated out to become

$$E[\psi(Z_1, Z_c, Z_2; \theta)|Z_c] = E[\rho_1(Z_1, Z_c; \theta)|Z_c] - E[\rho_2(Z_c, Z_2; \theta)|Z_c] \tag{2.7}$$

and each term on the right-hand side can be examined using the two data sets at hand (See Ridder and Moffitt (2003)). Although this separability assumption guarantees that the moments that arise from conditioning only on Z_c can be computed with the data, it is not a sufficient condition for the identification of θ_0.

In this formulation[7]

$$\begin{aligned} h(z_1, z_c; \theta) &= q_1(z_1, z_c; \theta) - q_2(z_c; \theta) \\ q_1(z_1, z_c; \theta) &= \int \rho_1(z_1, z_c; \theta) g(z_2|z_c) dz_2 = \rho_1(z_1, z_c; \theta) \\ q_2(z_c; \theta) &= E[\rho_2(Z_c, Z_2; \theta)|z_c] \end{aligned}$$

2.2.3 Non-linear regression

As an example of the GMM model with incomplete data, let consider the nonlinear regression model.

[6] For notational convenience the following expression changes the location of arguments in function ψ.

[7] We could also formulate

$$h(z_c, z_2; \theta) = E[\rho_1(Z_1, Z_c; \theta)|Z_c] - \rho_2(z_c, z_2; \theta).$$

In this model there is no Y_c or Y_2 and again assume that we always observe the regressor distribution so that there is no X_1. We consider here the asymmetric case where the only endogenous variable is exclusively observed in data set 1 so that $Z_1 = Y_1; Z_c = X_c$ and $Z_2 = X_2$. The parametric form of the conditional mean function is so that $E(Y_1|X_c, X_2) = m(X_c, X_2; \theta_0)$. The non-linear regression model obviously satisfies the separability condition mentioned above and using the previous notation $\rho_1(Z_1, Z_c, \theta) = \rho_1(Z_1) = Y_1$ and $\rho_2(Z_c, Z_2, \theta) = m(X_c, X_2; \theta_0)$. Since

$$E(Y_1|X_c) = E[m(X_c, X_2; \theta_0)|X_c] \tag{2.8}$$

the parametric conditional mean function is now $E[m(X_c, X_2; \theta)|X_c]$ and it is computable since we assume that the joint distribution of (X_c, X_2) can be estimated from data set 2. Note that even when a subset of variables in X_c does not appear in the function $m(X_c, X_2; \theta)$, it may appear in $E[m(X_c, X_2; \theta)|X_c]$ as it may be correlated with X_2. As pointed out by Angrist and Krueger (1992) and Arellano and Meghir (1992), this can help identification of θ_0 as discussed below.

The h function corresponding to the moment condition above is then

$$h(Y_1, X_c; \theta) = Y_1 - E[m(X_c, X_2; \theta)|X_c]$$

The moment condition in (2.8) identifies θ_0 through the mean independence of the error in the regression with the common regressors. However, there might be additional moment conditions that identify the parameters where function h should be defined in alternative ways. In particular, for the non-linear regression model, the true value of the parameter θ_0 uniquely solves the first order condition of the non-linear least squares objective function using the implied conditional mean function: using the same notation for $g(x_2|x_c)$, the function h is defined as follows in this case

$$\rho(y_1, x_c, x_2; \theta) = \begin{bmatrix} y_1 - m(x_c, x_2; \theta) \\ \nabla_\theta m(x_c, x_2; \theta) \end{bmatrix}$$

$$\psi(\rho(y_1, x_c, x_2; \theta); \theta) = \rho_1(y_1, x_c, x_2; \theta) \cdot \rho_2(x_c, x_2; \theta)$$

$$q(y_1, x_c; \theta) = \begin{bmatrix} y_1 - \int m(x_c, x_2; \theta) g(x_2|x_c) d\mu \\ \int \nabla_\theta m(x_c, x_2; \theta) g(x_2|x_c) d\mu \end{bmatrix}$$

$$h(y_1, x_c; \theta) = \psi(q(y_1, x_c; \theta); \theta) = q_1(y_1, x_c; \theta) \cdot q_2(x_c; \theta)$$

Thus, for a given model, there are alternative identifying moment conditions (which can be conditional or unconditional on the common exogenous regressors[8]) by defining in a different way the functions ψ and ρ.

2.3 Identification Conditions

The conditions under which the global identification of θ_0 holds in (2.2) are specific to each model.

2.3.1 Parametric models

Let $\Theta \subset R^p$ be the parameter space. A well known identification condition for this case is that within the parametric model the only probability distribution replicating the distribution of the data corresponds to the one with the true parameter: namely, for any $\theta \in \Theta$

$$\int \rho(z_1, z_c, z_2, \theta) g(z_2|z_c) d\mu = \int \rho(z_1, z_c, z_2, \theta^0) g(z_2|z_c) d\mu \text{ almost surely in } (Z_1, Z_c) \quad (2.9)$$

if and only if $\theta = \theta_0$ and where $\rho(z_1, z_c, z_2, \theta) = f(y_1, y_c|x_c, x_2; \theta)$.

The linear regression model with incomplete data where the m function in (2.8) is expressed as

$$m(X_c, X_2, \theta) = X_c\theta_1 + X_2'\theta_2$$

identifies θ^0 if and only if $E(X_2|X_c)$ is a nonlinear function of X_c and there is no proper linear subspace of R^{m_c} having probability one under the probability distribution of X_c. [9] Regarding identification of the nonlinear regression models and the nonlinear GMM models, sufficient conditions need to be given in each particular case to guarantee that global identification holds in the complete data model and also, when Z_2 is integrated out, in the incomplete data model. [10]

We investigate sufficient conditions under which condition (2.9) holds for the parametric and semiparametric binary choice models.[11]

[8] In the general definition for the estimators we use unconditional moments.

[9] Note that this sufficient condition for identification is implicitely excluding variables which are nonlinear functions of X_c in the conditional mean model $E(Y_1|X_c, X_2; \theta)$.

[10] Sufficient conditions for global identification in nonlinear-in-parameters models are difficult to obtain (Newey and McFadden (1994)). See Rothenberg (1971) for sufficient conditions for local identification in a neighborhood of θ^0 in nonlinear IV models.

[11] In the rest of the paper, we consider that $\rho_b(z_1, z_c, z_2; \theta)$ is parametrically specified. Let $q(z_1, z_c) =$

2.3.2 Identification conditions for the binary choice model

Let $\theta = (\alpha, \beta', \gamma')' \in \Theta$ be the parameters of the model and let the corresponding greek letters with the subscript 0 denote the true value. Let d_1 and d_2 denote the number of elements in β and γ, respectively. Let $Z_1 = Y$; $Z_2 = X_2$ and $Z_c = X_c$. Consider the following model:

$$Y = 1\{\alpha_0 + X_c'\beta_0 + X_2'\gamma_0 + U > 0\} \tag{2.10}$$

where we denote $X = (X_c', X_2')'$. The number of elements in X is denoted by $d = d_c + d_2$.

We consider the following two different sets of stochastic restrictions on the errors U, which define respectively a parametric and a semiparametric binary choice model. Let $F(.|x_c, x_2)$ denote the distribution function of U conditional on $X_c = x_c$ and $X_2 = x_2$.

Assumption A. 1 *U and X are statistically independent, the median of U is zero and $F(.|x)$ is known and strictly increasing.*

In this case we denote $F(\cdot|x)$ as $F(\cdot)$.

Assumption A. 2 *U conditional on X has zero median.*

Many empirical studies adopt Assumption A.1 with the logistic or normal cumulative distribution function F. With complete data the parameter θ_0 is identified as long as no proper linear subspace of R^d includes the support of X almost surely in X and F is strictly monotonic.[12] This is no longer the case when not all of the regressors are jointly observed with the dependent variable Y. Even for the parametric case one would need to impose stronger restrictions on the support of X.

We assume below that data set 1 includes variables (Y, X_c) and the second data set includes variables X.

$\int \rho(z_1, z_c, z_2) g(z_2|z_c) dz_2$. The discussion about nonparametric identification of unknown function $\rho(z_1, z_c, z_2)$ from the identified functions $q(z_1, z_c)$ and $g(z_2|z_c)$, is beyond the scope of this paper. However, there exist some results that are interesting to be considered in the incomplete data framework. If $\rho(Z_1, Z_c, Z_2) = E(Z_1|Z_c, Z_2)$ and Z_c has some exclusion restriction, the results from Newey and Powell (2003) can be applied and the conditional mean function is nonparametrically identified as long as $g(z_2|z_c)$ satisfies the completeness assumption. Without assuming exclusion restrictions in Z_c, Cross and Manski (2002) and Horowitz and Manski (1995) derive partial nonparametric identification results with the assumed data at our hand for the conditional cdf $\rho(Z_1, Z_c, Z_2) = F(Z_1|Z_c, Z_2)$ and consequently, partial nonparametric identification for $E(Z_1|Z_c, Z_2)$.

[12] One could weaken this further by writing conditions explicitly in terms of the support of $X'\theta$ and that of U.

The identification condition under the parametric model is that ,for any $\theta \in \Theta$, and for a given $F(\cdot|x_c, x_2)$ satisfying Assumption A.1

$$\int F\left(\alpha + X_c'\beta + x_2'\gamma\right) g(x_2|X_c)d\mu = \int F\left(\alpha_0 + X_c'\beta_0 + x_2'\gamma_0\right) g(x_2|X_c)d\mu \qquad (2.11)$$

a.s. in X_c if and only if $\theta = \theta_0$. Note that if there is no complementary data we would have to show identification without assuming that we have the same g function on both sides since g would be unknown in this case. This is the main source of identification that arises from the complementary data set.

For the semiparametric case, the identification condition becomes, for any $\theta \in \Theta$ and for a given $F_0(\cdot|x_c, x_2)$ and any $F(\cdot|x_c, x_2)$ satisfying Assumption A.2

$$\int F\left(\alpha + X_c'\beta + x_2'\gamma|X_c, x_2\right) g(x_2|X_c)d\mu = \int F_0\left(\alpha_0 + X_c'\beta_0 + x_2'\gamma_0|X_c, x_2\right) g(x_2|X_c)d\mu \qquad (2.12)$$

a.s. in X_c if and only if $\theta = \theta_0$.

Parametric Binary Choice Model The following assumptions are made for identification of β_0:

Assumption A. 3 *Θ is a bounded set in R^{d+1}.*

This assumption limits the potential effect of the missing regressors.

Assumption A. 4 *Random vector $X_2|X_c$ is tight uniformly over X_c*

The complement of a set A is denoted by A^c. Let S_c denote the support of X_c.

Assumption A. 5 *There is at least one element of X_c that has unbounded support given each of the other regressors.*

This condition allows us to find proper variation in X_c regardless of the missing variables. Let denote by X_{ck} the common regressor with unbounded support.

Theorem 1 *When there is complementary data to estimate the distribution of X_2 given X_c, θ_0 of the parametric binary choice model defined by equation (2.10) is identified with respect to any parameter $\theta \in \Theta$ such that $\beta_k \neq \beta_{0k}$ if Assumptions A.1 and A.3–A.5 hold.*

Proof. Suppose equality (2.11) holds. θ_0 is identified with respect to θ such that $\alpha \neq \alpha_0$, $\beta \neq \beta_0$ and $\gamma = \gamma_0$, since for all X_c

$$sign\left[\left(\alpha + X_c'\beta + x_2'\gamma\right) - \left(\alpha_0 + X_c'\beta_0 + x_2'\gamma_0\right)\right]$$

equals -1 or 1 uniformly over the support of X_2 given X_c. There is no need for an unbounded support variable if $\gamma = \gamma_0$ and only the identification conditions with complete data are required. Additional conditions are required to identify θ_0 with respect to θ such that $\gamma \neq \gamma_0$. Let consider this case. Since X_2 given X_c is uniformly tight on X_c, for any $\varepsilon > 0$, there is a uniformly bounded subsets $\Omega_2(x_c)$ of the support of X_2 given X_c for almost all x_c in the support of X_c with $\Pr\{X_2 \in \Omega_2(x_c) | X_c = x_c\} > 1 - \varepsilon$. Note that we have

$$\begin{aligned} 0 &= \int_{\Omega_2(X_c)} \left[F\left(\alpha + X_c\beta + x_2'\gamma\right) - F\left(\alpha_0 + X_c\beta_0 + x_2'\gamma_0\right)\right] g(x_2|X_c)d\mu \\ &\quad + \int_{\Omega_2^c(X_c)} \left[F\left(\alpha + X_c\beta + x_2'\gamma\right) - F\left(\alpha_0 + X_c\beta_0 + x_2'\gamma_0\right)\right] g(x_2|X_c)d\mu \end{aligned}$$

almost surely in X_c. Since F is a CDF, the absolute value of the second term on the right-hand side is bounded by 2ε almost surely in X_c. If the coefficients on regressor k in X_c are different, then since θ lies on a bounded set (Assumption A.3) and $\Omega_2(x_c)$ is uniformly bounded, the difference between $\alpha + x_c\beta + x_2'\gamma$ and $\alpha_0 + x_c\beta_0 + x_2'\gamma_0$ can be made positive or negative uniformly over x_2 and θ by moving the regressor under consideration but holding other variables in X_c fixed, because x_2 and θ are uniformly bounded on $\Omega_2(X_c)$ and Θ. This together with strict monotonicity of F, leads to a contradiction as $\varepsilon > 0$ can be chosen arbitrarily to be small. ■

The sufficient conditions for identification of θ_0 with respect to θ such that $\beta = \beta_0$ would require that the support of $\Omega_2(x_c)$ changes with x_c in a very restrictive way in order to be able to make the difference between $\alpha + x_2'\gamma$ and $\alpha_0 + x_2'\gamma_0$ positive or negative uniformly for $\Omega_2(x_c)$ for x_c belonging to a subset of S_c with positive probability and for each possible value of $\theta \in \Theta$ such that $\beta = \beta_0$.

When there is no common variable with unbounded support or θ_0 wants to be identified with respect to $\theta \in \Theta$ such that $\beta = \beta_0$, the next theorem provides identification of θ_0 when the missing regressors X_2 are discrete and the distribution of X_2 given X_c does not belong to a particular parametric family.

Suppose X_2 is a random variable which takes on two values, 1 and 2, we can reparametrize the model so that the problem is to identify $\alpha_1 = \alpha + \gamma$ and $\alpha_2 = \alpha + 2\gamma$ when

almost surely in X_c[13]

$$F\left(\alpha_1^0 + X_c'\beta^0\right) g(1|X_c) + F\left(\alpha_2^0 + X_c'\beta\right) g\left(2|X_c\right) = \tag{2.13}$$
$$F\left(\alpha_1 + X_c'\beta\right) g(1|X_c) + F\left(\alpha_2 + X_c'\beta\right) g\left(2|X_c\right)$$

This implies that

$$\frac{g\left(2|X_c\right)}{g(1|X_c)} = \frac{F\left(\alpha_1^0 + X_c'\beta^0\right) - F\left(\alpha_1 + X_c'\beta\right)}{F\left(\alpha_2 + X_c'\beta\right) - F\left(\alpha_2^0 + X_c'\beta^0\right)}$$

We should consider restrictions on the parameters for each value of X_c such that (i) $\alpha_1^0 + X_c'\beta^0 > \alpha_1 + X_c'\beta$ and $\alpha_2^0 + X_c'\beta^0 < \alpha_2 + X_c'\beta$ or (ii) $\alpha_1^0 + X_c'\beta^0 < \alpha_1 + X_c'\beta$ and $\alpha_2^0 + X_c'\beta^0 > \alpha_2 + X_c'\beta$. [14] Consider the case where the parameters satisfy one of the above sets of conditions for X_c, we have

$$g\left(1|X_c\right) = \frac{F\left(\alpha_2 + X_c'\beta\right) - F\left(\alpha_2^0 + X_c'\beta^0\right)}{F\left(\alpha_1^0 + X_c'\beta^0\right) - F\left(\alpha_1 + X_c'\beta\right) + F\left(\alpha_2 + X_c'\beta\right) - F\left(\alpha_2^0 + X_c'\beta^0\right)} \tag{2.14}$$

Note that the right-hand side defines a parametric model of the conditional probability $g\left(1|X_c\right)$ as a function of X_c using parameters $\{\alpha_1, \alpha_2, \beta\}$ and $\{\alpha_1^0, \alpha_2^0, \beta^0\}$ with parameter restrictions (i) or (ii) above. Thus the identification condition is that for each $\theta \in \Theta$, there is a value of X_c for which $g\left(1|X_c\right)$ is not within this parametric model. A sufficient condition for this is that there exist a value of X_c such that $g\left(1|X_c\right)$ does not belong to the parametric model in (2.14) for any value of $\theta \in \Theta$ satisfying one of the above conditions. Note that the same reasoning of the identification follows with $X_c'\beta = X_c'\beta^0 = s$ if one wants to identify θ^0 with respect to θ such that $\beta = \beta^0$.

This identification condition is satisfied if there is a variable among X_c that does not appear in the index $X_c'\beta$ or $X_c'\beta^0$. Therefore, it is clear in this case that the exclusion restrictions of common regressors are a sufficient restriction to guarantee the identification of the parameters. Denote by $\tilde{X}_c$ the random vector that excludes some variables of X_c. This exclusion restriction guarantees that one can identify the parameters $\{\alpha_1^0, \alpha_2^0, \beta^0\}$ without additional restrictions than in the complete data case, since

[13] Assume for simplicity that the support of X_2 given X_c is uniform in X_c. The same results would arise if this condition does not hold. In this case, the values of $\{\alpha_1, \alpha_2\}$ to be identified would be different for each X_c but under the conditions specified here the parameters α, β and γ would be identified.

[14] The rest of the parameters $\{\alpha_1, \alpha_2, \beta\}$ such that do not satisfy these restrictions are directly identified with respect to $\{\alpha_1^0, \alpha_2^0, \beta^0\}$. This is because they imply a negative value of the ratio of the two conditional probabilities, which implies that the equality (2.13) cannot hold.

$$\left[F\left(\alpha_1^0+\tilde{X}_c'\beta^0\right)-F\left(\alpha_1+\tilde{X}_c'\beta\right)\right]g(1|X_c)+$$
$$+\left[F\left(\alpha_2^0+\tilde{X}_c'\beta^0\right)-F\left(\alpha_2+\tilde{X}_c'\beta\right)\right]g\left(2|X_c\right)=0 \text{ a.s in } X_c$$

implies that $\alpha_1=\alpha_1^0,\alpha_2=\alpha_2^0$ and $\beta=\beta^0$ iff there does not exist a proper linear subspace of R^2 having probability 1 under the distribution function of $[g(1|X_c),g(2|X_c)]$. This latter condition is directly satisfied because the existence of a proper linear subspace implies that $g(1|X_c)=\kappa g(2|X_c)$ for scalar κ and for all X_c, or equivalently that $g(1|X_c)$ is constant over X_c.

More generally the following identification result holds:

Theorem 2 *If there exist X_c for each $\{\alpha_1,\alpha_2,\beta\}\in\Theta$ such that $g\left(1|X_c\right)$ is not an element of the parametric model expressed by*

$$g\left(1|X_c\right)=\frac{F\left(\alpha_2+X_c'\beta\right)-F\left(\alpha_2^0+X_c'\beta^0\right)}{F\left(\alpha_1^0+X_c'\beta^0\right)-F\left(\alpha_1+X_c'\beta\right)+F\left(\alpha_2+X_c'\beta\right)-F\left(\alpha_2^0+X_c'\beta^0\right)}$$

where $\alpha_1^0+X_c'\beta^0>\alpha_1+X_c'\beta$ and $\alpha_2^0+X_c'\beta^0<\alpha_2+X_c'\beta$ for the case of binary variable X_2 and by

$$\begin{aligned}&g(1|X_c)=\\ &=\\ &=\frac{\left[\begin{array}{c}-\Delta F\left(\alpha_k+X_c'\beta\right)-\left[\Delta F\left(\alpha_2+X_c'\beta\right)-\Delta F\left(\alpha_k+X_c'\beta\right)\right]g(2|X_c)-\cdots\\ -\left[\Delta F\left(\alpha_{k-1}+X_c'\beta\right)-\Delta F\left(\alpha_k+X_c'\beta\right)\right]g\left(k-1|X_c\right)\end{array}\right]}{\left[\Delta F\left(\alpha_1+X_c'\beta\right)-\Delta F\left(\alpha_k+X_c'\beta\right)\right]}\end{aligned}\tag{2.15}$$

where numbering of the regressors follow the order of $\alpha_j-\alpha_j^0$ and that $\alpha_1-\alpha_1^0>X_c'\left(\beta^0-\beta\right)$ and $\alpha_k-\alpha_k^0<X_c'\left(\beta^0-\beta\right)$ for the case of general discrete vector X_2, then α_0,β_0 and γ_0 are identified.

Proof. The binary case is shown above. Suppose for some integer $k\geq 3$ almost surely in X_c

$$F\left(\alpha_1+X_c'\beta\right)g(1|X_c)+\cdots+F\left(\alpha_k+X_c'\beta\right)g\left(k|X_c\right)=$$
$$F\left(\alpha_1^0+X_c'\beta^0\right)g(1|X_c)+\cdots+F\left(\alpha_k^0+X_c'\beta^0\right)g\left(k|X_c\right).$$

Without any loss in generality, assume that $\alpha_1-\alpha_1^0>X_c'\left(\beta^0-\beta\right)$ and $\alpha_k-\alpha_k^0<X_c'\left(\beta^0-\beta\right)$ and that the index is ordered in decreasing order of $\alpha_j-\alpha_j^0$. If this is

not the case the equality will not hold almost surely in X_c. Let $\Delta F(\alpha_j + X_c'\beta)$ be $F(\alpha_j + X_c'\beta) - F\left(\alpha_j^0 + X_c'\beta^0\right)$. Since $g(j|X_c)$ over j sum to 1, we have

$$\begin{aligned}
&-\Delta F(\alpha_k + X_c'\beta) = \\
&= \left[\Delta F(\alpha_1 + X_c'\beta) - \Delta F(\alpha_k + X_c'\beta)\right] g(1|X_c) + \cdots + \\
&+ \left[\Delta F(\alpha_{k-1} + X_c'\beta) - \Delta F(\alpha_k + X_c'\beta)\right] g(k-1|X_c)
\end{aligned}$$

Note that $\Delta F(\alpha_1 + X_c'\beta) - \Delta F(\alpha_k + X_c'\beta) > 0$ so that

$$\begin{aligned}
&g(1|X_c) \\
&= \frac{\left[\begin{array}{c} -\Delta F(\alpha_k + X_c'\beta) - \left[\Delta F(\alpha_2 + X_c'\beta) - \Delta F(\alpha_k + X_c'\beta)\right] g(2|X_c) - \cdots \\ - \left[\Delta F(\alpha_{k-1} + X_c'\beta) - \Delta F(\alpha_k + X_c'\beta)\right] g(k-1|X_c) \end{array}\right]}{\left[\Delta F(\alpha_1 + X_c'\beta) - \Delta F(\alpha_k + X_c'\beta)\right]}
\end{aligned}$$

■

In the analysis above, we have allowed a free parameter for each value of X_2. If there are restrictions across different values of X_2 the identification result certainly holds.

A comment on the need of joint variation of the regressors is in order once one compares the linear and non-linear in parameters case. One would need to assume the very restrictive assumption of independence between X_2 and X_c in order to be able to identify the parameters without further restrictions in the nonlinear model. There is no need of further restrictions because under independence of X_2 and X_c and under the assumption that the support of X_2 is uniform over X_c, the equality of probabilities in (2.11) cannot hold almost surely in X_c. It is interesting to point out that in the linear in parameter model, under independence of X_2 and X_c, the parameter associated to transformations of X_2 is not separately identified from the constant term. Thus, the nonlinearity in the parameters helps in the identification under the independence condition.

2.3.3 Identification conditions for the semiparametric binary choice model

For the semiparametric case, as we discussed, the identification condition is, for any $\theta \in \Theta$ and for a given $F_0(\cdot|x_c, x_2)$ and any $F(\cdot|x_c, x_2)$ satisfying Assumption A.2

$$\int F(\alpha + X_c'\beta + x_2'\gamma|X_c, x_2)\, g(x_2|X_c)d\mu = \int F_0(\alpha_0 + X_c'\beta_0 + x_2'\gamma_0|X_c, x_2)\, g(x_2|X_c)d\mu \tag{2.16}$$

a.s. in X_c if and only if $\theta = \theta_0$. The approach in the parametric model above fails because now we can choose F as well. However, an analogous result holds for up to scale

identification of the parameters.[15] Assumption A.5 should be replaced by the following one:

Assumption A. 6 *There exists at least one variable in X_c denoted by X_{cj} and $\beta_j^0 \neq 0$ such that given each of the other regressors has everywhere positive Lebesgue density.*

Theorem 3 *When there is complementary data to estimate the distribution of X_2 given X_c, θ_0 of the semiparametric binary choice model defined by equation (2.10) is identified up to scale with respect to any parameter $\theta \in \Theta$ such that $\beta_j \neq \beta_{0j}$ and $\beta_{0j} \neq 0$ if Assumptions A.2, A.3–A.4 and A.6 hold.*

Proof. Proceed exactly as in the parametric identification proof up to the point at which we obtain the inequality between $\alpha + X_c'\beta + x_2'\gamma$ and $\alpha + X_c'\beta_0 + x_2'\gamma_0$ uniformly over x_2 and θ. The true value of the parameter θ_0 is identifed if there is a set of values in the support of X_c with positive probability for which the following condition cannot hold

$$\int_{\Omega_2(x_c)} \left[F\left(\alpha + x_c'\beta + x_2'\gamma | x_c, x_2\right) - F_0\left(\alpha_0 + x_c'\beta_0 + x_2'\gamma_0 | x_c, x_2\right)\right] g(x_2|x_c) d\mu = 0$$

Thus, we need to find those values of x_c such that for any $\theta \in \Theta$ such that $\beta_j \neq \beta_{0j}$ and $\beta_{0j} \neq 0$ one of the following two inequalities

$$\begin{aligned} \alpha + x_c'\beta + x_2'\gamma &> 0 > \alpha_0 + x_c'\beta_0 + x_2'\gamma_0 \text{ or} \\ \alpha + x_c'\beta + x_2'\gamma &< 0 < \alpha_0 + x_c'\beta_0 + x_2'\gamma_0 \end{aligned}$$

holds uniformly over $x_2 \in \Omega_2(x_c)$. Since random variable X_{cj} has unbounded support, if $\beta_j \neq \beta_{0j}$ then following the proof for the parametric model, there are values of x_{cj} in the support of X_{cj} that can make the difference between $\alpha + x_c'\beta + x_2'\gamma$ and $\alpha_0 + x_c'\beta_0 + x_2'\gamma_0$ positive or negative uniformly in x_2 and θ. Let denote by H the set of these values of X_{cj}. The next step is to guarantee that 0 can lie between both indices for those values of $x_{cj} \in H$. Note that any of the above inequalities can hold if $\theta = a\theta_0$, for any scalar $a > 0$. Thus, the identification of θ_0 is up to scale. Denote $X_c = [X_j, X_{-j}]$ and $\beta = [\beta_j, \beta_{-j}]$. Without loss of generality let consider the case where $\beta_{0j} > 0$, then if X_{cj} has everywhere positive Lebesgue density by Assumption A.6, it follows that

$$\Pr\left\{ -\frac{1}{\beta_j}\left(\alpha + x_{-jc}'\beta_{-j} + x_2'\gamma\right) < x_{cj} < -\frac{1}{\beta_{0j}}\left(\alpha_0 + x_{-jc}'\beta_{-j0} + x_2'\gamma_0\right) \middle| x_{cj} \in H \right\} > 0 \tag{2.17}$$

[15] See Manski (1985) for the identification of this model in the complete data framework.

for $\beta_j > 0$,

$$\Pr\left\{ x_{cj} < -\frac{1}{\beta_{0j}}\left(\alpha_0 + x'_{-jc}\beta_{-j0} + x'_2\gamma_0\right); x_{cj} < -\frac{1}{\beta_j}\left(\alpha + x'_{-jc}\beta_{-j} + x'_2\gamma\right) \middle| x_{cj} \in H \right\} > 0 \tag{2.18}$$

for $\beta_j < 0$ and

$$\Pr\left\{ x_{cj} < -\frac{1}{\beta_{0j}}\left(\alpha_0 + x'_{-jc}\beta_{-j0} + x'_2\gamma_0\right); \left(\alpha + x'_{-jc}\beta_{-j} + x'_2\gamma\right) > 0 \middle| x_{cj} \in H \right\} > 0 \tag{2.19}$$

for $\beta_j = 0$.

This implies that there are values of $x_{cj} \in H$ with positive probability where the inequality $\alpha + x'_c\beta + x'_2\gamma > 0 > \alpha_0 + x'_c\beta_0 + x'_2\gamma_0$ holds uniformly in x_2 and θ. Then, the median independence assumptions (i.e. $F(0|x) = F_0(0|x) = 0.5$ a.s in x) leads to the contradiction since

$$\begin{aligned} F\left(\alpha + x'_c\beta + x'_2\gamma | x_c, x_2\right) &> 0.5 > F_0\left(\alpha_0 + x'_c\beta_0 + x'_2\gamma_0 | x_c, x_2\right) \text{ or} \\ F\left(\alpha + x'_c\beta + x'_2\gamma | x_c, x_2\right) &< 0.5 < F_0\left(\alpha_0 + x'_c\beta_0 + x'_2\gamma_0 | x_c, x_2\right) \end{aligned}$$

uniformly over θ and $z_2 \in \Omega_2(x_c)$ for those values of x_c where one of the above probabilities (2.17) - (2.19) is positive. ■

The previous result complement the results of Manski and Tamer (2003). They assume that X_2 is partially observed in the original data set so that there is only interval information about where the true value of the variable lies and there is not access to any complementary data set where X_2 and X_c are jointly observed. Under these conditions, the unbounded support assumption only allows them to identify the parameters associated to the unbounded variables X_{ck} while the rest of the parameters are only partially identified. They can only achieved point identification of the parameters when regressor X_2 is completely observed along with X_c and Y at least for some observations. When we have a complementary data in the parametric case, multiple (discrete) missing regressors are allowed, more importantly, it allows one to point identify the parameters associated to the missing regressors. Additional conditions on the support of the common and missing regressors need to be imposed with respect to the complete data case. With respect to the parametric binary choice case, a more strict assumption requiring at least one continuous common regressor needs to be imposed in order to semiparametrically identify the parameters up to scale.

It would be interesting to consider how the identification conditions could be relaxed if one considers a mixture of the setting in Manski and Tammer (2003) and the missing data problem we consider in this work. That it is, using both the interval information on X_2 in the original data set and the joint distribution of X_2 and X_c in the complementary data set. In this case, the interval information on X_2 could be considered as a natural exclusion restriction. We are interested in the effect of the complete variable X_2 on the dependent variable, but obviously this interval information is related with it and can be used as an instrument. As we pointed out before, having exclusion restrictions among the common regressors ensures that some of the above identification conditions are satisfied so that a weaker set of identification conditions could be studied when there exist those excluded variables.

2.4 Estimation

Let N_1 be the sample size of data set 1 and N_2 be the sample size of data set 2 and p the dimension of the vector of parameters. Let $\Omega_{Z_1} \in R^{m_1}, \Omega_{Z_c} \in R^{m_c}, \Omega_{Z_2} \in R^{m_2}, \Theta \in R^K$. Let Γ_q be a Banach space of functions on $R^{m_1} \times R^{m_c} \times \Theta$. Let Γ_ψ be a Banach space of functions on $\Gamma_q \times \Theta$. Formally, function $q(Z_1, Z_c, \theta)$ is a function from $\Omega_{Z_1} \times \Omega_{Z_c} \times \Theta$ into R^S, and $\psi(q(Z_1, Z_c, \theta); \theta)$ is a mapping from $q(.) \times \Theta$ into R^T with $T \geq K$, where $q \in \Gamma_q$. T denotes the number of moment conditions. We consider the sup-norm for the space of functions Γ denoted by $\|.\|_\Gamma$.

Define the sample analogue of the moment condition in (2.2) as

$$\hat{H}(\hat{q}_{N_2}(.,\theta)) = \frac{1}{N_1}\sum_{i=1}^{N_1} \hat{I}_{N_1 i}\psi(\hat{q}_{N_2}(z_{1i}, z_{ci}, \theta); \theta) = 0 \tag{2.20}$$

where

$$\hat{q}_{N_2}(z_{1i}, z_{ci}, \theta) = \int \rho(z_{1i}, z_{ci}, z_2; \theta)\hat{g}_{N_2}(z_2|z_{ci})dz_2 \tag{2.21}$$

and the trimming indicator[16]

$$\hat{I}_{N_1 i} = 1\left\{\hat{f}_{N_1}(z_{ci}) > b\right\} \tag{2.22}$$

In what follows, we omit the dependence of the estimators $\hat{f}$ and $\hat{g}$ of the sample sizes used for their estimation.

[16] We consider a fixed trimming term which does not change with the sample size, unlike in Robinson (1988).

Our estimator solves the following problem

$$\hat{\theta} = \inf_{\theta \in \Theta} \hat{H}(\theta, \hat{q}(.,\theta))' \times \hat{W} \times \hat{H}(\theta, \hat{q}(.,\theta)) \tag{2.23}$$

where $\hat{W}$ is a $T \times T$ matrix that converges in probability to a positive definite matrix W.[17]

Under the assumption that both regressors Z_c and Z_2 are continuous, and substituting $g(z_2|z_c)$ by its kernel nonparametric conditional density estimation, we obtain the following expression for the estimate of the q function evaluated at the $i-th$ observation

$$\hat{q}(z_{1i}, z_{ci}, \theta) = \int \rho(z_{1i}, z_{ci}, z_2; \theta) \frac{\left(N h_{N_2}^{m_c+m_2}\right)^{-1} \sum_{r=1}^{N_2} K_1\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right) K_2\left(\frac{z_{2r}-z_2}{h_{N_2}}\right)}{\left(N h_{N_2}^{m_c}\right)^{-1} \sum_{r=1}^{N_2} K_1\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)} dz_2$$

If, among other assumptions[18], the $s-th$ derivatives of ρ with respect to z_2 are continuous and the kernel function is of order s (such that $\int k(u)du = 1$, $\int k(u)u^j du = 0$ for $\leq j \leq s-1$ and $\int k(u)u^s du = 0$), then the usual change of variable of $t = (z_2 - z_{2r})/h_{N_2}$ in the above integral leads to

$$\hat{q}(z_{1i}, z_{ci}, \theta) = \left(N h_{N_2}^{M_1}\right)^{-1} \sum_{r=1}^{N_2} \frac{\rho(z_{1i}, z_{ci}, z_{2r}; \theta) K_1\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}{\left(N h_{N_2}^{M_1}\right)^{-1} \sum_{r=1}^{N_2} K_1\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)} + O(h_{N_2}^s) \tag{2.24}$$

The estimator we propose here for the moment condition is a weighted average of the function ρ where observations from both data sets are combined. For each possible combination of observations i from the first data set and r from the second data set, the kernel function gives more importance to those combinations in which the corresponding values of the common variable in both data sets are closer to each other.

If the distribution of Z_2 given Z_c is discrete where Z_2 takes R possible different values $\{v_1, ..., v_R\}$, the sample analogue of the estimate for q is then

$$\hat{q}(z_{1i}, z_{ci}, \theta) = \sum_{s=1}^{R} \rho(z_{1i}, z_{ci}, v_s; \theta) \hat{P}_{N_2}(Z_2 = v_s | z_{ci})$$

where

$$\hat{P}_{N_2}(Z_2 = v_s|z_{ci}) = \frac{\hat{f}_{N_2}(z_{ci}|Z_2 = v_s)\hat{P}_{N_2}(Z_2 = v_s)}{\hat{f}_{N_2}(z_{ci})} = \frac{\sum_{r=1}^{N_2} 1\{z_{2r} = v_s\} K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}{\sum_{r=1}^{N_2} K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}$$

[17] Note that the estimator $\hat{\theta}$ and $\hat{H}$ are function of both sample sizes N_1 and N_2. The estimates of $\hat{q}$ and $\hat{g}$ are obtained from the data set 2, so that they are a function of N_2 only. We ignore the different subindices for simplicity in the notation.

[18] We provide detailed conditions in Section (2.5).

In what follows, we present some examples of particular estimators.

Consider the linear regression model as a particular case of the nonlinear regression models explained in Section (2.2.3) with $m(X_c, X_2; \theta^0) = X_c'\theta_1^0 + X_2'\theta_2^0$. The moment conditions of the linear regression model[19] with incomplete data identify the true value of the parameters θ_0 as long as the conditional mean of $E(X_2|X_c)$ is nonlinear in X_c. These moment conditions suggest to estimate θ^0 from the following regression[20]

$$y_{1i} = x_{ci}'\theta_1 + \hat{E}_{N_2}(X_2|X_c = x_{ci})'\theta_2 + v_i \text{ for } i = 1, ..., N_1 \tag{2.25}$$

$$v_i = u_i + (x_{2i} - E(X_2|X_c = x_{ci}))'\theta_2 + \left(E(X_2|X_c = x_{ci}) - \hat{E}_{N_2}(X_2|X_c = x_{ci})\right)'\theta_2 \tag{2.26}$$

where $E(U|X_c) = 0$ and $\hat{E}_{N_2}(X_2|X_c = x_{ci})$ is a nonparametric estimation of the mean of X_2 using data set 2 conditional on each observation of the common regressors x_{ci} of data set 1,$i = \{1, ..., N_1\}$. In order for the OLS estimates of θ^0 from (2.25) to be consistent, we need to impose conditions that guarantee that for the generated regressor $\frac{1}{N_1}\sum_{i=1}^{N_1} \hat{E}_{N_2}(X_2|X_c = x_{ci})'v_i$ converges to zero in probability. The consistency of the nonparametric conditional mean and the nonlinearity of $E(X_2|X_c)$ in X_c ensure that these conditions are satisfied.

Alternatively, the same linear regression model suggests to estimate θ^0 from the following regression

$$\hat{E}_{N_1}(Y_1|X_c = x_{ci}) = x_{ci}'\theta_1 + \hat{E}_{N_2}(X_2|X_c = x_{ci})'\theta_2 + v_i \text{ for } i = 1, ..., N_1 \tag{2.27}$$

$$v_i = (E(X_2|X_c = x_{ci}) - \hat{E}_{N_2}(X_2|X_c = x_{ci}))'\theta_2 - (E(Y_1|X_c = x_{ci}) - \hat{E}_{N_1}(Y_1|X_c = x_{ci}))$$

If the conditional mean of X_2 given X_c is linear in X_c (as it is the case when both are jointly normal distributed), in order for the model to separately identify θ_1^0 and θ_2^0 the vector of regressors X_c needs to have some exclusion restrictions. For this additive model in the error term U, the separability conditions discussed in the GMM section are automatically satisfied. Denote $\tilde{X}_c$ as a strict subset of X_c. Again, the linear regression model with the conditional mean independence $E(U|X_c) = 0$ suggests to estimate the

[19]
$$E\left((Y_1 - \theta_1 X_c - \theta_2 E(X_2|X_c))\begin{pmatrix} X_c \\ E(X_2|X_c) \end{pmatrix}\right) = 0 \text{ iff } \theta = \theta^0$$

[20]The sub-indices in the expectations denote the sample size of the dataset in which each conditional mean is computed.

parameters from regression (2.25) where some variables in X_c are excluded in the linear part.

Therefore, when X_2 enters linearly in the model, the estimated parameters are obtained through the imputation of X_2 using its estimated conditional mean given the common variables X_c in both data sets.

The way X_2 is imputed using the observations of the common regressor X_c explains the differences between the estimator proposed by Arellano and Meghir (1992) and the one we propose here. They suggest to obtain an imputed value of the missing regressor by estimating the best linear prediction of X_2 given the common regressors X_c. Thus, they obtain their estimates from the following regression

$$y_{1i} = \tilde{x}'_{ci}\theta_1 + \hat{E}^*_{N_2}(X_2|X_c = x_{ci})'\theta_2 + v_i \text{ for } i = 1, ..., N_1 \tag{2.28}$$

$$v_i = u_i + \theta_2 (x_{2i} - E^*(X_2|X_c = x_{ci})) + \theta_2 \left(E^*(X_2|X_c = x_{ci}) - \hat{E}^*_{N_2}(X_2|X_c = x_{ci})\right) \tag{2.29}$$

where $E^*(X_2|X_c = x_c)$ is the best linear predictor of X_2 given a particular realization of X_c. It is important to point out that even if the structural equation that relates X_2 with X_c is nonlinear, the best linear prediction of X_2 given X_c allows one to obtain consistent estimates of the parameters of interest θ. This becomes clear when the correlation of X_c with each of the terms in v in (2.29) is analyzed.

The definition of the best linear predictor $E^*(X_2|X_c = x_c)$ defines an error $\varepsilon = x_2 - E^*(X_2|X_c = x_c)$, which by definition is uncorrelated with x_c and $\hat{E}^*_{N_2}(X_2|X_c = x_c)$.[21] Additionally, the consistent estimation of the best linear predictor ensures, by the law of large numbers, that the third term in v is not correlated with x_c. In terms of consistency then, there is no obvious advantage of using the nonparametric estimator of $E(X_2|X_c)$ instead of its linear projection, even if true conditional mean of X_2 is non-linear in X_c. However, it is not difficult to think of cases of nonlinear relationships of between X_2 and X_c where $Var(X_2 - \hat{E}^*_{N_2}(X_2|X_c = x_c)|X_c = x_c)$ is higher than $Var(X_2 - \hat{E}_{N_2}(X_2|X_c =$

[21] Since

$$E(\varepsilon' X_c) = E([X_2 - E^*(X_2|X_c)]' X_c) =$$
$$E([X_2 - X_c E(X'_c X_c)^{-1} E(X'_c X_2)]' X_c) = 0$$

and

$$E(\varepsilon' X_c (X'_c X_c)^{-1} X'_c X_2) = 0$$

$x_c)|X_c = x_c)$. For these cases, this would result in a higher efficiency of the estimator that approximates nonparametrically the conditional mean of X_2 given X_c.

The estimator obtained from the linear imputation method in (2.28) coincides with a two-stage least-squares estimator, where the first step uses observations from an auxiliary data set.

For the linear model with exclusion restrictions (and in general, for any given model), there is a number of different ways to write estimators for the parameters that this model identifies. For example, for the linear GMM model above defined from the moment condition $E(U|X_c = x_c) = 0$, Angrist and Krueger (1992) suggest the following alternative to the two-sample two-stage estimators discussed above. The sample analogue of the moment condition $E(U'X_c) = 0$ suggests the following estimator which is denoted in the literature of combining data sets as Two-Sample IV estimator (2SIV)

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta\in\Theta} \left(\frac{1}{N_1}(Y_1 - \tilde{X}'_{cN_1}\theta_1)'X_{cN_1} - \frac{1}{N_2}\left(X'_{2N_2}\theta_2\right)' X_{cN_2}\right)' \times \\ &\times\Omega^{-1}_{N_1N_2} \times \left(\frac{1}{N_1}(Y_1 - \tilde{X}'_{cN_1}\theta_1)'X_{cN_1} - \frac{1}{N_2}\left(X'_{2N_2}\theta_2\right)' X_{cN_2}\right) \end{aligned} \tag{2.30}$$

where $\Omega_{N_1N_2}$ is a matrix which converges to a non-singular positive definite matrix Ω.[22] The sub-indices N_1 and N_2 denote that the variable is taken from data set 1 or data set 2, respectively.

Since the moment condition is separable in Y_1 and X_2, the first part of this moment condition can be estimated using only the observations in data set 1 with sample size N_1 and the second part using data set 2 with sample size N_2. This estimator computes each of the sample analogue moments imbedded in criterion function with the observations of that data set that allows us to compute this moment. Hence, for example, the sample analogue of moment $E(X'_2X_c)$ is fully computed with observations in data set 2. However, there is an alternative estimation of this moment that combines both samples. Therefore, instead of computing moment $E(X'_2X_c)$, the estimator we have defined in (2.20) suggests to compute the sample analogue of the objective function by estimating $E(E(X'_2|X_c)X_c)$ using both data sets. Data set 2 is used to estimate nonparametrically the inner conditional mean and data set 1 is used to compute the outer expectation. In this way, we can link the estimation of this moment with the observations in data set 1 by conditioning on each observation there. This way of computing the sample analogue of the moment condition

[22] This weigthing matrix can be computed either using only dataset 1 or only using dataset 2 or both. That is the reason for the double sub-index $N_1$and N_2.

turns out to be more efficient than the estimator proposed by Angrist and Krueger (1992) in the Monte Carlo simulations we have performed in this paper.

Although the previous studies have focused on linear models which directly imputes the value of Z_2 and replace it by its estimated conditional mean given Z_c, the idea behind the Two-sample IV estimator can be extended to nonlinear models too as long as they are separable as in (2.6). First, consider the following moment conditions implied by (2.7):

$$E\left[Z_c'\left(\rho_1(Z_1, Z_c; \theta) - \rho_2(Z_c, Z_2; \theta)\right)\right] = 0$$

As it happened for the linear GMM model, there are different alternatives to construct the sample analogue of these unconditional moments with the data assumed at our hand. The first alternative computes the sample analogue of above expectation with that data set having full information on the variables inside each expectation. That it is, a valid estimator of θ^0 solves

$$\inf_{\theta\in\Theta} \hat{H}(\theta)'\hat{W}\hat{H}(\theta) \tag{2.31}$$
$$\text{with } \hat{H}(\theta) = \frac{1}{N_1}\sum_{i=1}^{N_1} z_{ci}'\rho_1(z_{1i}, z_{ci}, \theta) - \frac{1}{N_2}\sum_{r=1}^{N_2} z_{cr}'\rho_2(z_{cr}, z_{2r}, \theta)$$

The alternative estimator we propose is derived from expression (2.20). Thus, using the law of iterated expectations, we provide an alternative method of computing the sample analogues of moments associated with ρ_2 which uses also the information on Z_c in data set 1. In other words, it constructs a sample analogue of the conditional expectation $E_{Z_c}\left(E\left(Z_c'\rho_2(Z_c, Z_2, \theta)\middle| Z_c\right)\right)$ where the inner expectation is nonparametrically estimated using data set 2 and the outer expectation uses observations in data set 1. Thus, the sample analogue of the moment condition is as follows

$$\hat{H}(\theta) = \frac{1}{N_1}\sum_{i=1}^{N_1} z_{ci}'\rho_1(z_{1i}, z_{ci}, \theta) - \frac{1}{N_1}\sum_{i=1}^{N_1}\int z_{ci}'\rho_2(z_{ci}, s, \theta)\hat{g}(s|z_{ci})ds$$

or alternatively, once the bias associated to the estimation of $g(s|z_{ci})$ has been controlled for,

$$\hat{H}(\theta) \tag{2.32}$$
$$= \frac{1}{N_1}\sum_{i=1}^{N_1} z_{ci}'\rho_1(z_{1i}, z_{ci}, \theta) - \frac{1}{N_1}\frac{1}{N_2}\sum_{i=1}^{N_1}\sum_{r=1}^{N_2}\frac{\frac{1}{h_{N_2}^{m_c}} z_{ci}'\rho_2(z_{ci}, z_{2r}, \theta)K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}{\hat{f}(z_{ci})}$$

with $\hat{f}(z_{ci}) = (Nh_{N_2}^{m_c})^{-1} \sum_{r=1}^{N_2} K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)$

Alternatively, using unconditional moment condition, one can propose estimators of θ_0 by using the FOC of the sample analogue of the objective function

$$E\left((\rho_1(Z_1, Z_c; \theta) - E\left(\rho_2(Z_c, Z_2; \theta)|Z_c\right))^2\right)$$

that the true value of the parameter uniquely minimizes, where $E\left(\rho_2(Z_c, Z_2; \theta)|Z_c\right)$ is estimated using data set 2 for each conditioning observation of Z_c in data set 1. Thus, as mentioned before, given the moment condition $E(\rho_1(Z_1, Z_c) - \rho_2(Z_c, Z_2)|Z_c) = 0$, there is a wide variety of valid estimators of the parameters that can be constructed using different ways of building the sample analogue of this moment condition. It is difficult to determine a apriori which of these estimators is the most efficient. Our conjecture is that those estimators that use the law of iterated expectations to condition on observations of data set 1 are more efficient than those estimators that construct some sample analogues of moments using only data set 2. This is confirmed in the Monte Carlo simulation that we perform in this paper. Unfortunately, there is no result in this framework of incomplete data which can provide us with that estimator that attains the semiparametric efficiency bound. This constitutes an interesting topic for future research.

Regarding the Maximum Likelihood estimator, consider the parametric conditional probability model $f(y_1|x_c, x_2; \theta)$. The ML estimator can be defined by considering the score of the log likelihood of the model with incomplete data, i.e. $\log \int f(y_1|x_c, x_2; \theta) g(x_2|x_c) dx_2$. Thus, we define $\hat{\theta}$ as that value that solves

$$\frac{1}{N_1} \sum_{i=1}^{N_1} \left(Nh_{N_2}^{m_c}\right)^{-1} \sum_{r=1}^{N_2} \frac{\left[\nabla_\theta f(y_{1i}|x_{ci}, x_{2r}, \hat{\theta}) \Big/ \int f(y_{1i}|x_{ci}, x_2; \theta) \hat{g}(x_2|x_{ci}) dx_2\right] K_1\left(\frac{x_{cr}-x_{ci}}{h_{N_2}}\right)}{\left(Nh_{N_2}^{m_c}\right)^{-1} \sum_{r=1}^{N_2} K_1\left(\frac{x_{1r}-x_{ci}}{h_{N_2}}\right)} \tag{2.33}$$

$$+ O(h_{N_2}^s) = 0$$

And replacing $\hat{g}$ by its nonparametric estimation, finally we have that the ML estimator $\hat{\theta}$ solves

$$\frac{1}{nN} \sum_{i=1}^{N_1} \sum_{r=1}^{N_2} \frac{1}{h_{N_2}^{m_c}} \frac{\nabla_\theta f(y_{1i}|x_{1i}, x_{2r}, \hat{\theta}) K_1\left(\frac{x_{1r}-x_{ci}}{h_{N_2}}\right)}{\left(Nh_{N_2}^{m_c}\right)^{-1} \sum_{s=1}^{N_2} f(y_{1i}|x_{1i}, x_{2s}, \hat{\theta}) K_1\left(\frac{x_{1s}-x_{ci}}{h_{N_2}}\right)} +$$

$$+ O(h_{N_2}^2) = 0$$

2.5 Asymptotic Normality

In the theorem of this section, we state the sufficient conditions to show asymptotic normality of $\hat{\theta}$. Newey and McFadden (1994) discuss the asymptotic behavior for general two-step semiparametric estimators. We apply those general results for the case in which the first step is a kernel nonparametric estimator of $g(z_2|z_c)$ obtained from a different data set and the equation that defines the estimator does not depend linearly on the kernel estimator. We assume that both data sets are independent which makes the derivation of the asymptotics more straight forward.[23]

To motivate the asymptotic results, consider a Taylor's series expansion for $\hat{\theta}$ around θ_0 from the FOC of the objective function in (2.23)

$$\sqrt{N_1+N_2}\left(\hat{\theta}-\theta_0\right) = -\left[\nabla_\theta'\hat{H}(\hat{\theta},\hat{q}(.,\hat{\theta}))\times\hat{W}\times\nabla_\theta\hat{H}(\bar{\theta},\hat{q}(.,\bar{\theta}))\right]^{-1} \tag{2.34}$$
$$\times\left[\nabla_\theta'\hat{H}(\hat{\theta},\hat{q}(.,\hat{\theta}))\times\hat{W}\times\sqrt{N_1+N_2}\hat{H}(\theta_0,\hat{q}(.,\theta_0))\right]$$

where $\|\bar{\theta}-\theta_0\| \leq \left\|\hat{\theta}-\theta_0\right\|$

In what follows we denote by z_c and $\tilde{z}_c$ to the realized values of random variable Z_c in data sets 1 and 2, respectively. Equivalent notation is used for Z_2. Observations in the first data set are indexed by i and observations in the second data set are indexed by r, so that we have access to the following data: $\{z_{1i}, z_{ci}\}$ for $i = 1, ..., N_1$ and $\{\tilde{z}_{cr}, \tilde{z}_{2r}\}$ for $r = 1, ..., N_2$. This notation is useful to clarify how the projections of the U-statistic on the other sample that arise in the asymptotics are computed.

Consider the following assumptions:

Assumption B. 1 *The observations in data set 1* $\{z_{1i}, z_{ci}\}_{i=1}^{N_1}$ *are independent and identically distributed. The observations in data set 2* $\{\tilde{z}_{cr}, \tilde{z}_{2r}\}_{r=1}^{N_2}$ *are independent and identically distributed. Additionally both samples are independent*

Assumption B. 2 *The identification condition is satisfied so that* θ_0 *is the only value of the parameters that satisfies*

$$\int\int \psi(q(z_1, z_c, \theta_0); \theta_0) f(z_1, z_c) dz_1 dz_c = 0$$

[23] The case of independent samples is the typical situation that we face. It is very unlikely that there are common observations in both data sets. However, in this hypothetical case, one could identify the parameters using the observations that are in common and our conjecture is that there are some efficiency gains that would arise from these common observations. Also, the estimators would be different to the ones we present in this section.

Assumption B. 3 $E\left(\left|\psi(q(Z_1, Z_c, \theta_0); \theta_0)\right|^2\right) < \infty$

Assumption B. 4 *Let* $\lambda_1 = p\lim_{N_1,N_2\to\infty} \frac{N_1}{N_1+N_2}$ *and* $\lambda_2 = p\lim_{N_1,N_2\to\infty} \frac{N_2}{N_1+N_2}$ *so that* $\lambda_1 + \lambda_2 = 1$

Assumption B. 5 θ_0 *is an interior point of the compact set* $\Theta \in R^K$

Assumption B. 6 *The kernel* K *is a Borel measurable bounded real-valued function twice continuously differentiable and with second derivatives satisfying the Lipschitz continuity. Kernel* K *also satisfies:* $\int K(u)du = 1$; $\int u^j K(u)du = 0$ *for* $j = 1, ..., s-1$; $\int u^s K(u)du < \infty$; $\int |K(u)|\, du < \infty$; $|u|\,|K(u)| \to 0$ *as* $|u| \to \infty$; $\sup|K(u)| < \infty$; $\int K^2(u)du < \infty$

Assumption B. 7 l *is the maximum absolute moment (with* $l \geq 2$*) between* $\rho(Z_1, Z_c, Z_2; \theta_0)$ *and* $\frac{\partial\rho(Z_1,Z_c,Z_2;\theta_0)}{\partial\theta}$

Assumption B. 8 *Let* $r = \max\{2, m_c\}$ *and* $s > \frac{r}{4}$. *As* $N_1 \to \infty, N_2 \to \infty$, *the sequence of the bandwidths should satisfy* $h_{N_2} \to 0$; $(N_1 + N_2)h_{N_2}^{4s} \to 0$; $N_2 h_{N_2}^r \to \infty$; $\left(N_1 h_{N_2}^{m_c} b^2\right) / \log N_1 \to \infty$, $\frac{N_1 h_{N_2}^{2+\frac{2}{l-2}}}{(-\log h_{N_2})} \to \infty$

Assumption B. 9 *The* $s-th$ *order derivatives of* $\rho(z_1, z_c, z_2, \theta_0)$ *and* $\frac{\partial\rho(z_1,z_c,z_2,\theta_0)}{\partial\theta}$ *with respect to* z_c *and* z_2 *are Lipschitz continuous*

Assumption B. 10 $\psi(q;\theta)$ *is Frechet differentiable with respect to* θ *and* $q(.)$ *and the Frechet derivatives are Lipschitz continuous. with* $C_j(z_1, z_c) > 0$, $E\{C_j(z_1, z_c)\} < \infty$ *for* $j = \{1, 2, 3, 4\}$

$$\left|\frac{\partial\psi(q;\theta)}{\partial\theta} - \frac{\partial\psi(q';\theta')}{\partial\theta}\right| \leq C_1(z_1, z_c)\left|\theta - \theta'\right| + C_2(z_1, z_c)\left\|q - q'\right\|_{\Gamma_q}$$
$$\left|\frac{\partial\psi(q;\theta)}{\partial q} - \frac{\partial\psi(q';\theta')}{\partial q}\right| \leq C_3(z_1, z_c)\left|\theta - \theta'\right| + C_4(z_1, z_c)\left\|q - q'\right\|_{\Gamma_q}$$

Assumption B. 11 $\rho(z_1, z_c, z_2, \theta)$ *is continuously differentiable with respect to* θ *uniformly in a neighborhood of* θ_0

Assumption B. 12 *The* $s-th$ *order derivative of the density function of* Z_c *denoted by* $f(z_c)$ *is Lipschitz continuous. This density function also satisfies* $\sup_{z_c\in\Omega_{Z_c}}|f(z_c)| < \infty$ *and* $\inf_{z_c\in\Omega_{Z_c}}|f(z_c)| > 0$

Assumption B. 13 *The $s-th$ order derivatives with respect to z_c of the conditional densities $g(z_2|z_c)$ and $f(z_1|z_c)$ are continuous*

Assumption B. 14 *$plim_{N_1,N_2\to\infty}\hat{W} = W$ where W is symmetric and positive definite*

Henceforth we use the following shorthand notation. Let $q_{i\theta_0} = q(z_{1i}, z_{ci}, \theta_0)$ where subindex i denotes that z_1 and z_c are conditioned on the ith observation. Denote $\psi_{i\theta_0}(q_{i\theta_0}; \theta_0) = \psi(q(z_{1i}, z_{ci}, \theta_0); \theta_0)$ and $\rho_{i\theta_0}(z_2) = \rho(z_{1i}, z_{ci}, z_2; \theta_0)$ to indicate that the rest of the variables are all conditioned on the ith observation. Where necessary, we make explicit the argument of functions ψ, ρ and q.

Theorem 4 *Suppose that $\hat{\theta}$ is consistent to θ_0. Under Assumptions B. (1)-Assumptions B.14, if $V'WV$ is nonsingular with*

$$V = \int \nabla_\theta \psi(q(z_1, z_c, \theta_0); \theta_0) f(z_1, z_c) dz_1 dz_c \tag{2.35}$$

, then

$$\sqrt{N_1 + N_2}\left(\hat{\theta}_{N_1 N_2} - \theta_0\right) \xrightarrow{d} N\left(0, \left(V'WV\right)^{-1}\left(V'W\Sigma WV\right)\left(V'WV\right)^{-1}\right)$$

where $\Sigma = \frac{1}{\lambda_1}\Sigma_1 + \frac{1}{\lambda_2}\Sigma_2$ and

$$\Sigma_1 = Var\left(\psi(q(Z_1, Z_c; \theta_0); \theta_0)\right) \tag{2.36}$$

$$\Sigma_2 = Var\left(\int \left\{ \left[\rho(z_1, z_{cr}, z_{2r}, \theta_0) - \int \rho(z_1, z_{cr}, z_2, \theta_0) g(z_2|z_{cr}) dz_2\right]' \times \times \frac{\partial \psi(q(z_1, z_{cr}; \theta_0); \theta_0)}{\partial q} \right\} f(z_1|z_{cr}) dz_1 \right)$$

Proof. [Proof of Theorem (4)]

Consider the Taylor's series expansion in (2.34). The asymptotic distribution of $\hat{\theta}$ is shown in two parts. The first part shows the asymptotic distribution of the score term $\sqrt{N_1 + N_2}\hat{H}(\theta_0, \hat{q}(., \theta_0))$ and the second part shows that the conditions we state ensure the uniform convergence of the Jacobian to a positive definite matrix.

Part 1

We can focus on the distribution of a statistic which uses the trimming indicator based on the true density function since by Lemma A. 1 in the Appendix the above conditions on the sequence of bandwidths and the kernel function ensure that $\sup_i \left|\hat{I}_{Ni} - I_i\right| \xrightarrow{p} 0$ as

$N_1, N_2 \to \infty$. The expression below makes clear the sources of inefficiency that arise when Z_2 is not jointly observed with Z_1 and Z_c.

$$\hat{H}(\theta_0, \hat{q}_{N_2}(.,\theta_0)) =$$

$$= \frac{1}{N_1}\sum_{i=1}^{N_1} I_i \psi(\rho_{i\theta_0}(z_{2i});\theta_0) \tag{2.37}$$

$$+ \frac{1}{N_1}\sum_{i=1}^{N_1} I_i \left[\psi(q_{i\theta_0};\theta_0) - \psi(\rho_{i\theta_0}(z_{2i});\theta_0)\right] \tag{2.38}$$

$$+ \frac{1}{N_1}\sum_{i=1}^{N_1} I_i \left[\hat{q}_{i\theta_0} - q_{i\theta_0}\right]' \frac{\partial\psi(q_{i\theta_0};\theta_0)}{\partial q} \tag{2.39}$$

$$+ R_{N_1N_2} \tag{2.40}$$

Only term (2.37) would arise if Z_2 were jointly observed with Z_1 and Z_c. Term (2.38) reflects the efficiency loss due to not observing of $Z2$, since if Z_2 is observed there is no need to integrate out function ρ over the distribution of Z_2 given Z_c. The next term (2.39) represents the efficiency loss due to the estimation of the conditional distribution function of Z_2 given Z_c, $g(Z_2|Z_c)$ inside function q.

Lemma A. 2 in the Appendix shows under which some of the assumptions above $\sqrt{N_1+N_2}R_{N_1N_2} = o_p(1)$. From expression (2.39),we use the asymptotically linearity at rate $N_2^{-1/2}$ for kernel estimators of conditional expectations. Define $m\rho_{i\theta_0}(z_c) = \int \rho_{i\theta_0}(z_2)g(z_2|z_c)\,dz_2$ and its estimated counterparts by $\widehat{m\rho}_{i\theta_0}(Z_c)$. Thus, expression (2.39) can be written as

$$U_{N_1N_2} = \tag{2.41}$$

$$\frac{1}{N_1N_2h^{m_c}}\sum_{i=1}^{N_1}\sum_{r=1}^{N_2}\left\{\left[I_i\frac{\left\{\rho_{i\theta_0}(z_{2r}) - m\rho_{i\theta_0}(z_{cr})\right\}' K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}{f(z_{ci})}\right]\times\left[\frac{\partial\psi(q_{i\theta_0})}{\partial q}\right]\right\}$$

$$+ \frac{1}{N_1}\sum_{i=1}^{N_1} b_{qi} + o_p\left(\frac{1}{\sqrt{N_2}}\right) + O(h_{N_2}^s)$$

where

$$b_{qi} = \frac{1}{h_{N_2}^{m_c}}\frac{I_i}{f(z_{ci})}E_{Z_c}\left(\left[m\rho_{i\theta_0}(Z_c) - m\rho_{i\theta_0}(z_{ci})\right]' K\left(\frac{Z_c - z_{ci}}{h_{N_2}}\right)\right)\times\left[\frac{\partial\psi(q_{i\theta_0})}{\partial q}\right]$$

Since $m\rho_{i\theta_0}(Z_c)$ is differentiable with respect to Z_c by the $s-th$ order differentiability of $g(Z_2|Z_c)$ with respect to Z_c in Assumption B.13, one can show by the usual change of

variable and a Taylor's series expansion in kernel estimator that

$$p\lim\left(\frac{\sqrt{N_1+N_2}}{N_1}\sum_{i=1}^{N_1}b_{qi}\right)=p\lim\left(\sqrt{N_1+N_2}O(h_{N_2}^s)\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{I_i}{f(z_{ci})}\frac{\partial\psi(q_{i\theta_0})}{\partial q}\right)$$

which is equal to zero as long as $N_1h_{N_2}^{2s}\to 0$ and $N_2h_{N_2}^{2s}\to 0$ as $N_1\to\infty, N_2\to\infty$.

By Assumption B. 4, the last reminder term of $U_{N_1N_2}$ converges to zero in probability since $\sqrt{N_1+N_2}o_p\left(\frac{1}{\sqrt{N_2}}\right)=\frac{1}{\sqrt{\lambda_2}}o_p(1)$. We now compute the projection $\hat{V}_{N_1N_2}$ of terms (2.41) denoted henceforth as $V_{N_1N_2}$. These are two-sample U-statistics of order 1, since there is only one observation from each sample in each kernel[24]. Let define the kernels a in each of the U-statistic as

$$V_{N_1N_2}=\frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{r=1}^{N_2}a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})$$

where

$$a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})=\frac{I_i}{h_{N_2}^{m_c}}\left[\frac{\left\{\rho_{i\theta_0}(z_{2r})-m\rho_{i\theta_0}(z_{cr})\right\}'K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}{f(z_{ci})}\right]\times\left[\frac{\partial\psi(q_{i\theta_0})}{\partial q}\right]$$

Denote by $\Upsilon=E\left(a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})\right)$. The projection of statistic $(V_{N_1N_2}-\Upsilon)$ is defined as

$$\hat{V}_{N_1N_2}=\frac{1}{N_1}\sum_{i=1}^{N_1}E\left(a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})|\,z_{1i},z_{ci}\right)+\frac{1}{N_2}\sum_{r=1}^{N_2}E\left(a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})|\,z_{cr},z_{2r}\right)-2\Upsilon$$

It can be shown that the projection over both samples of the kernels are

$$E\left(a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})|\,z_{1i},z_{ci}\right)=0$$

$$E\left(a_{N_2}(z_{1i},z_{ci},z_{cr},z_{2r})|\,z_{cr},z_{2r}\right)=$$

$$I_rE_{Z_1|Z_c}\left[\left\{\rho_{\theta_0}(z_{2r})-m\rho_{\theta_0}(z_{cr})\right\}'\times\left[\frac{\partial\psi(q_{\theta_0})}{\partial q}\right]\middle|\,Z_c=z_{cr}\right]+O(h_{N_2}^s)$$

[25]The above conditions ensure that the later Taylor's series expansion can be done [26]. The Lemma A. 6 in the Appendix gives sufficient conditions for

$$\sqrt{N_1+N_2}\left[V_{N_1N_2}-\Upsilon-\hat{V}_{N_1N_2}\right]\xrightarrow{p}0$$

[24]For Central Limit Theorems for U-statistics, see Serfling (1980) and van der Vaart (1998)

[25]The projection of the statistic over the first sample becomes zero since when we condition on observation z_{cr} and integrate out using the distribution of $g(z_{2r}|z_{cr})$, the numerator of the projection becomes zero.

[26]Note that

$$\int\left[1\{f(z_{cr}+th_N)>b\}-1\{f(z_{cr})>b\}\right]K(t)\,dt\to 0$$

if $h_N\to 0$ as $N\to\infty$ because the indicator function has only finitely points of discontinuity in t and $K(t)$ is continuous in those points.

as $N_1 \to \infty$, $N_2 \to \infty$ and known as the U-statistics projection result. There we use the sufficient condition of $E\left(|a_{N_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r})|^2\right) = o(N_2)$ in Powell, Stock and Stoker (1989), which is satisfied as long as $N_2 h_{N_2}^{m_c} \to \infty$ as $N_2 \to \infty$.

The sufficient conditions in Assumption B. 8 guarantees that $N_2 h_{N_2}^{m_c} \to \infty$ and that also the conditions in Lemma A.2 in the Appendix are satisfied since $l \geq 2$.

Note also that because the projection on the first sample is zero, then $\Upsilon = 0$. Having used then the projection device to find the distribution of (2.39) we can conclude that the asymptotic distribution of $\hat{H}(\theta_0, \hat{q}_{N_2}(., \theta_0), \hat{g}_{N_2})$ is normally distributed as

$$\sqrt{N_1 + N_2}\hat{H}(\theta_0, \hat{q}_{N_2}(., \theta_0)) =$$
$$= \sqrt{N_1 + N_2}\left(\frac{1}{N_1}\sum_{i=1}^{N_1} I_i \psi(q_{i\theta_0}; \theta_0) + \hat{V}_{N_1 N_2}\right) + o_p(1) \to N(0, \Sigma)$$

with the expression in Σ as in expression (2.36).

Part 2

Under the differentiability conditions of function ψ and q with respect to θ in Assumptions B. 10 and B. 11 uniformly in a neighborhood of θ_0, the Taylor's series expansion in (2.34) is correctly done. With respect to the Jacobian term in (2.34), the uniform convergence arguments together with the consistency of $\hat{\theta}$ and $\hat{q}$ suggests that

$$\left|\nabla_\theta \hat{H}(\hat{\theta}, \hat{q}_{N_2}(., \hat{\theta})) - V\right| = o_p(1) \tag{2.42}$$

and consequently also, $\left|\nabla_\theta \hat{H}(\bar{\theta}, \hat{q}_{N_2}(., \bar{\theta})) - V\right| = o_p(1)$ where

$$V = \int \nabla_\theta \psi(q(z_1, z_c, \theta_0); \theta_0) f(z_1, z_c) dz_1 dz_c \tag{2.43}$$

The convergence in probability in (2.42) is shown in two steps. Lemma A.7 in the Appendix shows that

$$\left|\nabla_\theta \hat{H}(\theta, \tilde{q}(., \theta)) - \nabla_\theta \hat{H}(\theta_0, q(., \theta_0))\right| = o_p(1) \tag{2.44}$$

where θ and $\tilde{q}(., \theta)$ belongs to a neighborhood of the true value of the parameters θ_0 and the true function $q(., \theta_0)$. By the law of large numbers,

$$\left|\nabla_\theta \hat{H}(\theta_0, q(., \theta_0)) - V\right| = o_p(1) \tag{2.45}$$

By the continuity of the matrix inversion (given the nonsingularity of $V'WV$) and the Slutsky theorem, the result of the asymptotic variance arises.■

The main difference between the asymptotics that we have derived and those of previous approaches is that we allow for sample analogue moment conditions that are not necessarily separable in both data sets. Arellano and Meghir (1992) and Angrist and Krueger (1992) derive the asymptotic distribution for GMM problems when data sets are combined in which the criterion function is perfectly separable in variables observed in each of the available data sets.

2.6 Monte Carlo Evidence

We perform three different experiments to assess the performance of the estimator we propose in this work: a linear model without exclusion restrictions, a linear model with exclusion restrictions and a Probit model.

The first experiment consists of the linear model in (2.25) where the conditional mean model of X_2 given X_c is nonlinear in X_c. We consider the case of scalar X_c and X_2. The data generating process is $Y = \theta_0 + \theta_1 X_c + \theta_2 X_2 + U$ with $\theta = [0.5; 1.5; 2]$, $U \sim N(0,1)$ and $X_c \sim N(0,1)$; $X_2 = \beta_0 + \beta_1 X_c + \beta_2 X_c^2 + \varepsilon$ where $\beta = [1; 1; 1]$ and $\varepsilon \sim N(0, \sigma^2)$. We generate two different sets of variables $\{X_c, X_2\}$ from this data generating process with sample sizes $N_1 = 1000$ and $N_2 = 5000$, respectively. The conditional mean of X_2 given X_c is nonparametrically estimated from data set 2. The performance of the estimates of $\hat{\theta}$ depends on the goodness of fit of the regression of the missing regressor X_2 on X_c, which clearly depends on the value of σ^2. We perform different experiments for different values of σ^2. The results are presented in Tables (2.1)-(2.2). for values of $\sigma^2 = 1$ and $\sigma^2 = 3$, respectively. In each case, we report the mean, the quantiles and the MSE over the number of replications for each parameter and also the mean of the adjusted R^2 of the OLS regression of the quadratic equation of X_2. The data was trimmed from the boundary of the support of X_c so that 95% of the data were considered to evaluate the estimated conditional mean. This trimming defines an upper bound for the optimal bandwidth, which is obtained by Cross-Validation for each replication[27]. A third order

[27] The Cross Validation function was computed using the observations in dataset 2, since it is the only one in which Z_c and Z_2 are jointly observed. We want to evaluate the estimates of the conditional expectation for each observation in dataset 1. However, the CV function that we are able to construct minimises the estimated prediction error of the conditional mean function evaluated at the observations of Z_c in dataset 2. Since both datasets are generated from the same underlying population, the CV using the simulated dataset 1 and dataset 2 are very similar and also the optimal bandwidth that both provide.

kernel was used to reduce the order of the bias of the estimated conditional mean function. In particular, the kernel used is $K(u) = (4/3)k(u) - (1/6) * k(u/2)$ where $k(u)$ is a standard normal pdf. This helps in reducing the bias of the third component of v in (2.26). In each row of the last panel of Tables (2.1)-(2.2), the mean over replications of the components of v in (2.26) are reported and also the mean over replications of the correlation with the generated variables used in the regression.

These results illustrate that our estimator performs well in a model without exclusion restrictions as long as the true underlying conditional mean model is nonlinear in X_c. The performance of the estimator is worse when the model for X_2 is more noisy and X_c explains less of the variance of X_2, as can be seen when comparing the MSE of both simulations is Tables (2.1)-(2.2). The decomposition of the error components is useful to assess the source of asymptotic bias of the replications. The results below suggest that the main source arises from the difference between the true conditional mean and the estimated conditional mean. Both tables also report a decomposition of the variance of the error between its components. With respect to the full data case, the main source of inefficiency when X_2 is not jointly observed with Y is due to the fact that we replace X_2 by its conditional mean $E(X_2|X_c)$. The inefficiency that arises because this conditional mean is nonparametrically estimated is almost negligible in the results we report.[28]

The second experiment illustrates a model with excluded restrictions from structural equation. We consider both the just-identified and the overidentified case. X_c is an exogenous scalar variable, X_2 is the scalar missing regressor and W_1, W_2 are the excluded variables. The design of the experiment for the just identified case is the following. The common regressor and the excluded variable are independently normal : $X_c \sim N(0,1); W_1 \sim N(0,4)$ and the missing regressor relates to these two variables as follows: $X_2 = 1 + 2W_1 + X_c W_1 + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2); \sigma = 0.85$. The model for the dependent variable is $Y = \theta_0 + \theta_1 X_c + \theta_2 X_2 + U$ with $\theta = [0.5; 1.5; 2]$, $U \sim N(0,1)$. Our estimator in this case amounts to imputing the value of X_2 using its nonparametric conditional mean given X_c and W_1 as regression (2.25) suggests. We report these results in the upper panel of Table (2.3) and compare them with the results from the two-sample two-stage least squares where X_2 is linearly fitted using X_c and W_1 as in (2.28) in the second panel of results. We also present there results for two different versions of the two-sample IV estimator. The first version is reported in the third panel of Table (2.3) and uses only

[28] The simulations performed for the alternative linear model in (2.27) yield very similar results to the ones reported in Tables 1-2. For brevity, we omit these results here.

data set 2 to compute those moments that include X_2, as the estimator from moment condition (2.31) and (2.30) suggests. The second IV version is the estimator that solves (2.32) where instead of using the sample analogue of $E(X_2'W_1)$ from data set 2, uses the sample analogue of $E(E(X_2'|X_c, W_1)W_1)$ where the inner expectation is computed with data set 2 and the outer expectation is computed with data set 1. The results reported in Table (2.3) use only data set 1 to compute the weighting matrix of the IV estimator. Similar results where obtained when the weighting matrix used only observations from data set 2.

The design for the simulation of the overidentified model with exclusion restriction is similar except for the conditional mean model for the missing regressor $X_2 = 1 + 2W_1 + X_cW_1 + W_1W_2 + 2W_2 + \varepsilon$ with $W_1 \sim N(0,4), \varepsilon \sim N(0,\sigma^2)$. The corresponding results for the overidentified case can be found in Table (2.4).

The estimator we propose (i.e. those estimates in the first and fourth panel of Tables (2.3) and (2.4)) turns out to be more efficient than the two estimators we compare it with. Obviously, the design of the experiment helps in finding these results, because the conditional mean model is non linear in the conditioning variables. This induces a higher dispersion in the differences between X_2 and the estimated linear projection of X_2 given X_c, W_1, W_2 then in the differences between X_2 and the nonparametric estimate of the conditional mean $E(X_2|X_c, W_1, W_2)$. Table (2.5) compares the variance decomposition of v in (2.26) in each of its terms for both the estimator where the nonparametric conditional mean of X_2 given (X_c, W_1, W_2) and the estimator that uses the best predictor of X_2 given (X_c, W_1, W_2).[29] The mean over replications of these variance and covariances are reported. The analysis of this table reveals that the differences in efficiency between both estimators arise from the higher dispersion of the $(E(X_2|X_c, W_1, W_2) - \hat{E}^*_{N_2}(X_2|X_c, W_1, W_2))$ with respect to the dispersion of its nonparametric counterpart $(E(X_2|X_c, W_1, W_2) - \hat{E}_{N_2}(X_2|X_c, W_1, W_2))$. The IV estimator implied by our framework turns out to be also more efficient than the two-sample IV estimator proposed in the literature in both the just-identified and the over identified case.

In the third experiment we design the simulation of a probit model with a discrete and scalar missing regressor X_2. The data generating process of the regressors is $X_c \sim N(0,1)$

[29] We trim the observations when the value of X_2 is imputed using its nonparametric estimation of the conditional mean of $X_2|X_c$. For this reason, the comparison of this variance decomposition with the estimator in which $E(X_2|X_c)$ is linearly fitted is carried out using the same observations. As a consequence, the first two terms of v are equal and they only differ in the third term.

and $X_2 = 1\{1 + X_c + \varepsilon > 0\}$ where $\varepsilon \sim N(0,1)$ and $y = 1\{\theta_0^0 + \theta_1^0 X_c + \theta_2^0 X_2 + U > 0\}$ with $[\theta_0^0, \theta_1^0, \theta_2^0] = [1, 3, -3]$ and $U \sim N(0,1)$. The results of two estimators of this model are reported in Table (2.6). First, we estimate the parameters of a probit model where the dependent variable is generated using both X_c and X_2 as explained above but the estimations only use regressor X_c to estimate the model. These results are reported in the top panel of Table (2.6). The bottom panel reports the results of the ML estimator that combines two different data sets defined in (2.33). The high value of the coefficient of the parameters associated to X_2 induces a high omitted variable bias in the estimates of the probit model including only the available information in X_c. The use of an additional data set allows us to estimate more efficiently the model by reducing this omitted variable bias.[30]

For scalar X_2, we also provide identification results for a more general scalar and continuous X_2. Table (2.7) reports the simulation results of a binary choice model where X_2 is uniformly distributed $X_2 \sim U(0,1)$ and $X_c = 10(1 - X_2)M$ where $M \sim N(0,1)$ and $y = 1\{\theta_0^0 + \theta_1^0 X_c + \theta_2^0 X_2 + U\}$ with $[\theta_0^0, \theta_1^0, \theta_2^0] = [0.5, 1.5, -0.5]$ and $U \sim N(0,1)$. Again, these results suggest that even if the regressors are not jointly observed with the binary endogenous variable, our estimator helps in reducing the omitted variable bias that ignoring X_2 as a relevant variable of the model would induce.

2.7 Conclusions

In this paper, we have developed a framework that allows for identification and estimation of structural models in which not all of the relevant variables are jointly observed. This framework can be applied to those models that identify their parameters via zero moment restrictions. We exploit the joint variation of the variables in an additional data set together with a parametric restriction to identify the effects of the missing and non-missing variables in the parametric structural relationship under certain conditions. We present a general estimator for this class of models based on the nonparametric estimation of the conditional distribution function of the regressors which can obtained from the auxiliary

[30] The optimal bandwidth choice for this set up in which the estimator is defined as the maximiser of a least squares-type objective function where some nonparametric estimates are imbedded has been studied by Hardle, Hall and Ichimura (1993). How to select the bandwith when the objective function is a likelihood function involving some nonparametric estimation is an open question. We use a value of the bandwith for the sample sizes reported for the probit results of 0.75. Various sensitivity analysis exercises were carried out (having some contraints given the support of Z_c) and the results did not change substantially.

data set. This general setting encompasses a broad class of estimators such as linear and non-linear least squares, MLE and GMM. For linear regression and linear GMM models previous results are available in the literature. We compare the performance of our general estimator with the existing ones and we point out that the main differences arise in the way our estimator computes the conditional moments that need to be estimated from the auxiliary data set. There are no existing results in this framework of incomplete data which provides us with the semiparametric efficiency bound so that we cannot formally discuss in this work efficiency issues between all the possible estimators defined from a given set of moment conditions. This constitutes an interesting future application of this framework. Preliminary evidence based on Monte Carlo experiments indicates that in some familiar cases our estimator is more efficient than previous estimators.

The identification conditions are specific to each parametric model, therefore we provide detailed conditions for each case we discuss. In general, the identification results can be summarized as follows. For the linear model, the common regressors and the imputed value of the missing regressor given the common regressors must satisfy the usual rank condition. For the GMM, the moment condition must be separable in those variables that are not jointly observed in the same data set so that identification does not rely on strong conditional independence assumptions. This separability condition is automatically satisfied when the model is additively separable in the unobservables. For nonlinear regression models and nonlinear GMM models, sufficient identification conditions are harder to obtain because they are problem specific. Therefore, our main identification results for the general parametric model are limited to the parametric and semiparametric binary choice model.

For the binary choice model our results complement the work by Manski and Tamer (2003) by allowing for a vector-valued missing discrete regressor for both parametric and semiparametric models and in addition allowing for identification of the coefficients of those missing regressors. These results are obtained through the added information available from the auxiliary data. We present Monte Carlo results that illustrate how our data sets combination method reduces substantially the omitted variable bias that arises in the binary choice model when a relevant missing variable is excluded from estimation.

We also derive the asymptotic variance of this type of estimators for the general case and provide sufficient conditions that must checked to be satisfied for each particular case.

2.8 Tables

Table 2.1: Monte Carlo Experiment for a linear model without exclusion restriction $\sigma^2 = 1$

N_1=1000;N_2=5000; No. replications=100

$\sigma^2 = 1$, mean of adj-$R^2 = 0.7623$

θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.5197	0.3318	0.5575	0.6871	0.0827
1.5	1.4939	1.3875	1.4783	1.6045	0.0323
2	1.9980	1.8804	1.9879	2.0097	0.0219
sum of MSE	0.1369				
$\hat{E}(v)$	0.0167				
$\widehat{Var}(v)$	13.0251				
$\widehat{corr}(v, X_c)$	-0.0019				
$\widehat{corr}(v, \hat{E}_{N_2}(X_2\|X_c))$	-0.0017				
First component of v					
$\hat{E}(u)$	0.0143				
$\widehat{Var}(u)$	9.0270				
$\widehat{corr}(u, X_c)$	-0.0064				
$\widehat{corr}(u, \hat{E}_{N_2}(X_2\|X_c))$	-0.0037				
Second component of v					
$\theta_2\hat{E}(X_2 - E(X_2\|X_c))$	-0.0029				
$\widehat{Var}(X_2 - E(X_2\|X_c))$	3.9787				
$\widehat{corr}(\theta_2(X_2 - E(X_2\|X_c)), X_c)$	0.0049				
$\widehat{corr}(\theta_2(X_2 - E(X_2\|X_c)), \hat{E}_{N_2}(X_2\|X_c))$	-0.0024				
Third component of v					
$\theta_2\hat{E}((E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)))$	0.0159				
$\widehat{Var}(E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c))$	0.0056				
$\widehat{corr}((E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)), X_c)$	0.0192				
$\widehat{corr}((E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)), \hat{E}_{N_2}(X_2\|X_c))$	0.1257				
Covariances components of v					
$\widehat{cov}(u, (X_2 - E(X_2\|X_c)))$	0.0014				
$\widehat{cov}(u, (E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)))$	0.0002				
$\widehat{cov}((X_2 - E(X_2\|X_c)), (E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)))$	-0.0009				

Table 2.2: Monte Carlo Experiment for a linear model without exclusion restriction $\sigma^2 = 3$

N_1=1000;N_2=5000; No. replications=100

$\sigma^2 = 3$, mean of adj-$R^2 = 0.5164$

θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.5383	0.2777	0.5910	0.8029	0.1386
1.5	1.5123	1.3755	1.4984	1.6444	0.0530
2	1.9874	1.8657	1.9599	2.1223	0.3556
sum of MSE	0.5475				
$\hat{E}(v)$	0.0172				
$\widehat{Var}(v)$	21.0018				
$\widehat{corr}(v, X_c)$	0.0002				
$\widehat{corr}(v, \hat{E}_{N_2}(X_2\|X_c))$	-0.020				
First component of v					
$\hat{E}(u)$	0.0143				
$\widehat{Var}(u)$	9.0271				
$\widehat{corr}(u, X_c)$	-0.0064				
$\widehat{corr}(u, \hat{E}_{N_2}(X_2\|X_c))$	-0.0037				
Second component of v					
$\theta_2\hat{E}(X_2 - E(X_2\|X_c))$	-0.0050				
$\widehat{Var}(X_2 - E(X_2\|X_c))$	11.9363				
$\widehat{corr}(\theta_2(X_2 - E(X_2\|X_c)), X_c)$	0.0049				
$\widehat{corr}(\theta_2(X_2 - E(X_2\|X_c)), \hat{E}_{N_2}(X_2\|X_c))$	-0.0024				
Third component of v					
$\theta_2\hat{E}((E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)))$	0.0250				
$\widehat{Var}(E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c))$	0.0152				
$\widehat{corr}((E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)), X_c)$	0.0103				
$\widehat{corr}((E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)), \hat{E}_{N_2}(X_2\|X_c))$	0.0825				
Covariances components of v					
$\widehat{cov}(u, (X_2 - E(X_2\|X_c)))$	0.0013				
$\widehat{cov}(u, (E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)))$	0.0008				
$\widehat{cov}((X_2 - E(X_2\|X_c)), (E(X_2\|X_c) - \hat{E}_{N_2}(X_2\|X_c)))$	-0.0016				

Table 2.3: Monte Carlo Experiment for a linear model with exclusion restriction. Just identified case

N_1=1000;N_2=3000; No. replications=100					
	Two-Sample two stage (Nonparametric)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.7089	0.6019	0.7040	0.8511	0.0770
1.5	1.2384	0.9227	1.3382	1.5404	0.2545
2	2.1456	2.0450	2.1456	2.2134	0.0366
sum of MSE	0.3680				
	Two-Sample two stage (Linear Prediction)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.7933	0.6689	0.7857	0.9040	0.119
1.5	0.9231	0.6318	0.8787	1.1869	0.4747
2	2.1861	2.0477	2.1946	2.3045	0.0558
sum of MSE	0.6424				
	IV (using complete data set for moments)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.7388	0.6133	0.7311	0.8505	0.0836
1.5	0.8775	0.5573	0.8414	1.1523	0.5482
2	2.4009	2.2488	2.4114	2.5300	0.1864
sum of MSE	0.8181				
	IV (conditional moments from data set 2)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.6181	0.4997	0.6021	0.7362	0.0411
1.5	1.2757	1.0417	1.2143	1.5575	0.1522
2	2.4876	2.3877	2.4950	2.5657	0.2564
sum of MSE	0.4496				

Table 2.4: Monte Carlo Experiment for a linear model with exclusion restriction. Over identified case

N_1=1000;N_2=3000; No. replications=100					
	Two-Sample two stage (Nonparametric)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.8527	0.7159	0.8419	0.9910	0.1698
1.5	1.3111	1.0541	1.3872	1.5607	0.1553
2	2.1012	2.0236	2.1074	2.1685	0.0198
sum of MSE	0.3449				
	Two-Sample two stage (Linear Prediction)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.8264	0.7173	0.8081	0.9315	0.1321
1.5	0.9165	0.6419	0.8981	1.1725	0.4768
2	2.1847	2.0501	2.1971	2.3117	0.0545
sum of MSE	0.6634				
	IV (using complete data set for moments)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.7502	0.6431	0.7401	0.8469	0.0865
1.5	0.8708	0.5833	0.8465	1.1367	0.5491
2	2.3985	2.2522	2.4134	2.5371	0.1831
sum of MSE	0.8187				
	IV (conditional moments from data set 2)				
θ^0	Mean	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	MSE
0.5	0.2623	0.1585	0.2526	0.3724	0.1062
1.5	0.9046	0.6391	0.9145	1.1965	0.4998
2	2.2565	2.1563	2.2465	2.5131	0.0781
sum of MSE	0.6841				

Table 2.5: Variance decomposition of error term for Linear Model with exclusion restrictions

N_1=1000;N_2=3000; No. replications=100		
	Just identified model	Over identified model
$\widehat{Var}(u)$	0.9875	1.0003
$\widehat{Var}((X_2 - E(X_2\|Z))$	207.5141	293.5613
$\widehat{Var}((E(X_2\|Z) - \hat{E}^*_{N_2}(X_2\|Z))$	10.6006	95.7986
$\widehat{Var}((E(X_2\|Z) - \hat{E}_{N_2}(X_2\|Z))$	2.1836	1.0003
$\widehat{cov}(u, (X_2 - E(X_2\|Z))$	0.0037	-0.0082
$\widehat{cov}(u, (E(X_2\|Z) - \hat{E}^*_{N_2}(X_2\|Z))$	-0.0080	-0.5478
$\widehat{cov}(u, (E(X_2\|Z) - \hat{E}_{N_2}(X_2\|Z))$	-0.0013	-0.5424
$\widehat{cov}((X_2 - E(X_2\|Z), (E(X_2\|Z) - \hat{E}^*_{N_2}(X_2\|Z))$	0.0286	0.0017
$\widehat{cov}((X_2 - E(X_2\|Z), (E(X_2\|Z) - \hat{E}_{N_2}(X_2\|Z))$	-0.0060	0.0018

In this table, Z denotes vector $[Xc, W_1, W_2]$

Table 2.6: Monte Carlo Experiment for a probit model (Z2 dummy variable, Zc Normal)

N_1=1000;N_2=2000 No. replications=100							
Z_2 omitted							
θ^0	Mean	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	MSE
1	-0.8865	-0.9971	-0.9359	-0.8768	-0.8408	-0.7883	3.5632
3	1.6059	1.4674	1.5424	1.6020	1.6555	1.7282	1.9511
Likelihood combining Z_2	Mean	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	MSE
θ^0							
1	1.0092	0.4644	0.7997	1.0135	1.2105	1.5108	0.1081
3	3.0329	2.5410	2.8059	3.0255	3.2403	3.5864	0.1112
-3	-3.0474	-3.9902	-3.4259	-3.0584	-2.7467	-2.2948	0.3113

Table 2.7: Monte Carlo Experiment for a probit model (Z2 uniform, Zc Normal)

N_1=1000;N_2=2000							
No. replications=100							
Z_2 omitted							
θ^0	Mean	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	MSE
0.5	0.0080	-0.1081	-0.0466	0.0004	0.0638	0.1214	0.2473
1.5	1.4787	1.2498	1.3742	1.4926	1.5483	1.6966	0.0170
Likelihood combining Z_2	Mean	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	MSE
θ^0							
0.5	0.2758	0.2064	0.2458	0.2762	0.3024	0.3508	0.0520
1.5	1.4855	1.2568	1.2568	1.5100	1.5477	1.7035	0.0167
-0.5	-0.2664	-0.3117	-0.3117	-0.2713	-0.2461	-0.2142	0.0557

2.9 Appendix

Lemmas used in Part 1 of the Proof of Theorem (4)

Lemma 1 *Let* $\hat{I}_{Ni} = 1\{\hat{f}(z_{ci}) > b\}$ *and* $I_i = 1\{f(z_{ci}) > b\}$. *If* $\left(Nh_{N_2}^{m_c}b^2\right)/\log N_1 \to \infty$, $|K(0)| < \infty$ *and there is no positive probability that* $f(z_{ci}) = b$, *then*

$$\Pr\left\{\text{at least one } i \text{ such that } \hat{I}_{Ni} - I_i \neq 0\right\} \to 0 \text{ as } N_2, N_1 \to \infty$$

Proof. (see Ichimura (2003)) ■

Lemma 2 *Let Assumptions B.9, B.10, B.12, B.7, B.13 and B.6 be satisfied and consider the bandwidth sequence that satisfies*

$$(N_1 + N_2)h_{N_2}^{4s} \to 0 \tag{2.46}$$

$$\frac{N_2 h_{N_2}^{\frac{l}{l-2}}}{(-\log h_{N_2})} \to \infty \tag{2.47}$$

$$N_2 h_{N_2}^2 \to \infty \tag{2.48}$$

Then, $\sqrt{N_1 + N_2}R_{N_1N_2} = o_p(1)$

Proof.

The reminder term in (2.40) is expressed as

$$R_{N_1N_2} = \frac{1}{N_1}\sum_{i=1}^{N_1} I_i \left[\hat{q}_{i\theta_0} - q_{i\theta_0}\right]' \left[\frac{\partial\psi(\bar{q}_{i\theta_0};\theta_0)}{\partial q} - \frac{\partial\psi(q_{i\theta_0};\theta_0)}{\partial q}\right]$$

where $\|\bar{q}_{i\theta_0} - q_{i\theta_0}\|_{\Gamma_q} \leq \|\hat{q}_{i\theta_0} - q_{i\theta_0}\|_{\Gamma_q}$. From the Frechet differentiability of ψ with respect to q and the Lipschitz continuity conditions of its derivatives in assumption B.10, then

$$R_{N_1N_2} \leq \frac{1}{N_1}\sum_{i=1}^{N_1} I_i C_4(z_{1i}, z_{ci}) \left[\hat{q}_{i\theta_0} - q_{i\theta_0}\right]' \left[\hat{q}_{i\theta_0} - q_{i\theta_0}\right]$$

Thus, in this reminder term we have a nonparametric conditional mean function and in the expressions below we use notation for both the numerator and the denominator of

both estimators. Denote $\hat{q}_{i\theta_0} = \widehat{rq}_{i\theta_0}/\hat{f}_i$. Then,

$$R_{N_1N_2} \leq$$

$$\frac{1}{N_1}\sum_{i=1}^{N_1}\left[\begin{array}{c} I_i\frac{1}{f_i}\left[\widehat{rq}_{i\theta_0} - rq_{i\theta_0}\right] - \frac{rq_{i\theta_0}}{f_i^2}\left[\widehat{f}_i - f_i\right] + \\ +o_p\left(\widehat{rq}_{i\theta_0} - rq_{i\theta_0}\right) + o_p\left(\widehat{f}_i - f_i\right) \end{array}\right]' \times$$

$$\times \left[\begin{array}{c} I_i\frac{1}{f_i}\left[\widehat{rq}_{i\theta_0} - rq_{i\theta_0}\right] - \frac{rq_{i\theta_0}}{f_i^2}\left[\widehat{f}_i - f_i\right] + \\ +o_p\left(\widehat{rq}_{i\theta_0} - rq_{i\theta_0}\right) + o_p\left(\widehat{f}_i - f_i\right) \end{array}\right] \times C_4(z_{1i}, z_{ci}) \tag{2.49}$$

$$\tag{2.50}$$

To show that $\sqrt{N_1+N_2}R_{N_1N_2} = o_p(1)$ we follow the next steps. Lemma 3 shows that the order of the bias of the nonparametric estimators is $h_{N_2}^s$ so that $E\left(\widehat{rq}_{i\theta_0}\right) - rq_{i\theta_0} = O(h_{N_2}^s); E\left(\widehat{f}_i\right) - f_i = O(h_{N_2}^s)$. The differentiability conditions of Lemma 3 are stated in assumptions B.9, B.12 and B.13. From expression (2.50) and Lemmas 3 - 5 below, the reminder term converges in probability to zero if there exist positive sequences h_{N_2}, $\{\varepsilon_{rN_2}\}, \{\varepsilon_{fN_2}\}$ and $\{M_{N_2}\}$ such that

$$\begin{aligned} \sqrt{N_1+N_2}\varepsilon_{rN_2}\varepsilon_{fN_2} &\to 0 \\ \sqrt{N_1+N_2}\varepsilon_{rN_2}h_{N_2}^s &\to 0; \sqrt{N_1+N_2}\varepsilon_{fN_2}h_{N_2}^s \to 0 \\ \sqrt{N_1+N_2}h_{N_2}^{2s} &\to 0 \end{aligned} \tag{2.51}$$

as $N_1 \to \infty$, $N_2 \to \infty$ and such that these sequences satisfy the conditions of Lemmas 4 - 5 below. To see that these sequences exist, take $\varepsilon_{jN_2} = (-\log h_{N_2}/N_2h_{N_2})^{1/2} b_{jN_2}$ and $M_{jN_2} = (N_2h_{N_2}/(-\log h_{N_2}))^{1/2} b_{jN_2}^u$ for $0 < u < 1$ and for positive sequences b_{jN_2} that diverge to infinity for $j = \{r, f\}$. Then, the sequences satisfy the conditions in lemmas 4 - 5 as long as condition (2.47) holds.

The conditions in (2.51) hold if the sequences b_{jN_2} diverge at a slower rate than $o((-\log h_{N_2})^{-1/2})$ and if $(N_1+N_2)h_{N_2}^{4s} \to 0$ and $N_2h_{N_2}^2 \to \infty$. ■

Lemma 3 *Let* $E\left(\frac{1}{h_{N_2}^{m_c}}\varphi(z_{1i}, z_{ci}, Z_2; \theta_0)K\left(\frac{Z_c - z_{ci}}{h_{N_2}^{m_c}}\right)\right)$ *exists. The* $s-th$ *order derivatives of* $f(Z_c)$ *and* $\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)g(Z_2|Z_c)\,dZ_2$ *with respect to* Z_c *and the* $s-th$ *order derivatives of function* $\varphi(Z_1, Z_c, Z_2; \theta_0)$ *with respect to* Z_2 *are Lipschitz continuous. The*

kernel function satisfies Assumption B.6, then for $h_{N_2} > 0$ *and* $h_{N_2} \to 0$ *and* $N_2 \to \infty$

$$E\left(\left(\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)\hat{g}\left(Z_2|z_{ci}\right) dZ_2\right)\hat{f}(z_{ci})\right) -$$
$$-\left(\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0) g\left(Z_2|z_{ci}\right) dZ_2\right) f(z_{ci}) = O(h_{N_2}^s)$$

Let these conditions be satisfied for $\varphi(z_{1i}, z_{ci}, Z_2; \theta_0) = \rho(z_{1i}, z_{ci}, Z_2; \theta_0)$ *and* $\varphi(z_{1i}, z_{ci}, Z_2; \theta_0) = 1$.

Proof. Note that the expression for the estimator is

$$\left(\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)\hat{g}\left(Z_2|z_{ci}\right) dZ_2\right)\hat{f}(z_{ci}) =$$
$$\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)\frac{1}{N_2 h_{N_2}^{m_c+m_2}}\sum_{r=1}^{N_2} K\left(\frac{Z_2 - z_{2r}}{h_{N_2}}\right) K\left(\frac{z_{ci} - z_{cr}}{h_{N_2}}\right) dZ_2$$

After the change of variable $t_l = (Z_{2l} - z_{2rl})/h_{N_2}$ for $l = 1, ..., m_2$ and a Taylor's series expansion of order s of $\varphi(z_{1i}, z_{ci}, Z_2; \theta_0)$ around z_{2r}, then

$$\left(\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)\hat{g}\left(Z_2|z_{ci}\right) dZ_2\right)\hat{f}(z_{ci}) =$$
$$\frac{1}{N_2 h_{N_2}^{m_c}}\sum_{r=1}^{N_2} \varphi(z_{1i}, z_{ci}, z_{2r}; \theta_0) K\left(\frac{z_{ci} - z_{cr}}{h_{N_2}}\right) + O(h_{N_2}^s)$$

Taking now expectations from the above estimator with respect to variables Z_2 and Z_c and by the Law of Iterated Expectations,

$$E\left(\left(\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)\hat{g}\left(Z_2|z_{ci}\right) dZ_2\right)\hat{f}(z_{ci})\right) =$$
$$\int \frac{1}{h_{N_2}^{m_c}} E_{Z_2|Z_c}\left(\varphi(z_{1i}, z_{ci}, Z_2; \theta_0)|\, Z_c\right) K\left(\frac{z_{ci} - Z_c}{h_{N_2}}\right) f(Z_c) dZ_c + O(h_{N_2}^s)$$

which by the $s-th$ order continuously differentiability of $g\left(Z_2|Z_c\right)$ with respect to Z_c can be shown that

$$E\left(\left(\int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0)\hat{g}\left(Z_2|z_{ci}\right) dZ_2\right)\hat{f}(z_{ci})\right) = \int \varphi(z_{1i}, z_{ci}, Z_2; \theta_0) g\left(Z_2|z_{ci}\right) dZ_2 f(z_{ci}) + O(h_{N_2}^s)$$

■

Lemma 4 *Under assumptions B.7 and B.6, then*

$$\Pr\left\{\sup_{z_1, z_c, \theta \in \Omega_{Z_1} \times \Omega_{Z_c} \times \Theta} |\widehat{rq}_{i\theta} - E\left(\widehat{rq}_{i\theta}\right)| > \varepsilon_{rN_2}\right\} \to 0 \text{ as } N_2 \to \infty$$

if $\varepsilon_{rN_2} h_{N_2} M_{N_2}^{l-1} \to \infty$ *and* $(\log h_{N_2})(1 + M_{N_2}\varepsilon_{rN_2})/(N h_{N_2} \varepsilon_{rN_2}^2) \to 0$, *where* M_{N_2} *denotes a sequence for the support of the dependent variable* $\rho(z_1, z_c, z_2, \theta)$

Proof. See Ichimura (1993) Lemmas A.5 and A.8 in the Appendix ■

Lemma 5 *Under assumption B.6, then*

$$\Pr\left\{ \sup_{z_c \in \Omega_{Z_c}} \left| \widehat{f}_i - E\left(\widehat{f}_i\right) \right| > \varepsilon_{fN_2} \right\} \to 0 \text{ as } N_2 \to \infty$$

if $(\log h_{N_2})(1 + M\varepsilon_{fN_2})/(N_2 h_{N_2} \varepsilon_{fN_2}^2) \to 0$, *where* M *denotes an interval containing 1*

Proof. See Ichimura (1993) Lemmas A.5 and A.8 in the Appendix ■

Lemma 6 *If* $N_2 h_{N_2}^{m_c} \to \infty$ *as* $N_2 \to \infty$, *then* $E\left((a_{N_2}(Z_1, Z_c, Z_c, Z_2))^2\right) = o(N_2)$

Proof.

$$a_{N_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r}) = \frac{I_i}{h_{N_2}^{m_c}} \left[\frac{\left\{\rho_{i\theta_0}(z_{2r}) - m\rho_{i\theta_0}(z_{cr})\right\}' K\left(\frac{z_{cr} - z_{ci}}{h_{N_2}}\right)}{f(z_{ci})} \right] \times \left[\frac{\partial \psi(q_{i\theta_0})}{\partial q} \right]$$

Denote the conditional expectations of $a_{N_2}(Z_1, Z_c, Z_c, Z_2)^2$ on the realised values of Z_c in each data set as

$$v(z_{ci}, z_{cr}) = \int\int \left\{ \left[\begin{array}{c} \rho(Z_1, z_{ci}, Z_2; \theta_0) - \\ \int \rho(Z_1, z_{ci}, Z_2; \theta_0) g(Z_2 | z_{cr}) dZ_2 \end{array} \right]' \times \left[\frac{\partial \psi(q(Z_1, z_{ci}; \theta_0))}{\partial q} \right] \right\}^2 f(Z_1 | z_{ci}) g(Z_2 | z_{cr}) dZ_1 dZ_2$$

Then,

$$\begin{aligned} E\left((a_{N_2}(Z_1, Z_c, Z_c, Z_2))^2\right) &= \int \frac{1}{h_{N_2}^{2M_1}} \frac{v(z_{ci}, z_{cr}) K^2\left(\frac{z_{ci} - z_{cr}}{h_{N_2}}\right)}{f^2(z_{ci})} f(z_{ci}) f(z_{cr}) dz_{ci} dz_{cr} \\ &= \int \frac{1}{h_{N_2}^{m_c}} \frac{v(z_{ci}, z_{ci} + t h_{N_2}) K^2(t)}{f(z_{ci})} f(z_{ci} + t h_{N_2}) dz_c dt \\ &= O\left(N_2 \left(N_2 h_{N_2}^{m_c}\right)^{-1}\right) \end{aligned}$$

Consequently, we have that $E\left((a_{N_2}(Z_1, Z_c, Z_c, Z_2))^2\right) = o(N_2)$ if and only if $N_2 h_{N_2}^{m_c} \to \infty$ as $N_2 \to \infty$.■

Lemmas used in Part 2 of the Proof of Theorem (4)

Lemma 7 *Under Assumption B. 10 and B. 11, if* $\hat{\theta} \xrightarrow{p} \theta_0$ *and the bandwidth sequence satisfies* $h_{N_2} \to 0$ *and* $\frac{N_2 h_{N_2}^{2+\frac{2}{\Gamma-2}}}{(-\log h_{N_2})} \to \infty$ *as* $N_2 \to \infty$, *then*

$$\left|\nabla_\theta H(\hat{\theta}, \hat{q}(.,\hat{\theta})) - \nabla_\theta H(\theta_0, q(.,\theta_0))\right| = o_p(1)$$

Proof.

For θ and $\tilde{q}(.,\theta)$ belonging to a neighborhood of the true value of the parameters θ_0 and the true functions $q(.,\theta_0)$,

$$\begin{aligned}
&|\nabla_\theta H(\theta, \tilde{q}(.,\theta)) - \nabla_\theta H(\theta_0, q(.,\theta_0))| \leq \\
&|\nabla_\theta H(\theta, \tilde{q}(.,\theta)) - \nabla_\theta H(\theta_0, \tilde{q}(.,\theta))| + \qquad (2.52)\\
&+ |\nabla_\theta H(\theta_0, \tilde{q}(.,\theta)) - \nabla_\theta H(\theta_0, q(.,\theta_0))| \qquad (2.53)
\end{aligned}$$

It can be shown by the Lipschitz continuity conditions in Assumption B.10 that

$$\begin{aligned}
&|\nabla_\theta H(\theta, \tilde{q}(.,\theta)) - \nabla_\theta H(\theta_0, \tilde{q}(.,\theta))| \leq \\
\leq\ & \frac{1}{N_1}\sum_{i=1}^{N_1}\left[C_1(z_{1i}, z_{ci}) + C_3(z_{1i}, z_{ci})\left\|\frac{\partial \tilde{q}(z_{1i}, z_{ci}, \theta)}{\partial \theta}\right\|_{\Gamma_q}\right] \times \|\theta - \theta_0\|
\end{aligned}$$

For consistent estimators $\hat{\theta}$ and $\hat{q}$ of θ_0 and q, respectively, we can consider $\theta = \hat{\theta}$ and $\tilde{q} = \hat{q}$ in the above expression. Under the conditions on functions $C's$ on Assumption B 10 and the differentiability of function ρ in assumption B. 11, the first term in (2.52) converges to zero in probability

$$\left|\nabla_\theta H(\hat{\theta}, \hat{q}(.,\hat{\theta})) - \nabla_\theta H(\theta_0, \hat{q}(.,\hat{\theta}))\right| = o_p(1)$$

By assumption B.(10) and after some algebra, we obtain an upper bound for the second term in (2.52)

$$\begin{aligned}
&|\nabla_\theta H(\theta_0, \tilde{q}(.,\theta)) - \nabla_\theta H(\theta_0, q_0(.,\theta_0))| \leq \\
&\leq \frac{1}{N_1}\sum_{i=1}^{N_1} C_2(z_{1i}, z_{ci})\left\|\tilde{q}_{i\theta} - q_{i\theta_0}\right\|_{\Gamma_q} + \\
&+ \frac{1}{N_1}\sum_{i=1}^{N_1}\left[C_4(z_{1i}, z_{ci})\left\|\tilde{q}_{i\theta} - q_{i\theta_0}\right\|_{\Gamma_q}\right]\frac{\partial \tilde{q}_{i\theta}}{\partial \theta} + \\
&+ \frac{1}{N_1}\sum_{i=1}^{N_1}\frac{\partial \psi(q_{i\theta_0}; \theta_0)}{\partial q}\left\|\frac{\partial \tilde{q}_{i\theta}}{\partial \theta} - \frac{\partial q_{i\theta_0}}{\partial \theta}\right\|_{\Gamma_q} \qquad (2.54)
\end{aligned}$$

Denote as in Lemma 2 the numerator and denominator of the conditional mean expectation as $\hat{q}_{i\theta} = \widehat{rq}_{i\theta}/\hat{f}_i$ and for its first derivative as $\frac{\partial \hat{q}_{i\theta}}{\partial \theta} = \hat{q}_{i\theta}^{(1)} = \widehat{rq}_{i\theta}^{(1)}/\hat{f}_i$. To show that the upper bound in (2.54) converges to zero in probability, we should require consistency of $\hat{\theta}$. We also need the following results on uniform consistency

$$\Pr\left\{\sup_{i,\theta} \|\widehat{rq}_{i\theta} - rq_{i\theta}\| > \varepsilon_{rN_2}\right\} \to 0$$
$$\Pr\left\{\sup_{i} \left\|\hat{f}_i - f_i\right\| > \varepsilon_{fN_2}\right\} \to 0$$
$$\Pr\left\{\sup_{i,\theta} \left\|\widehat{rq}_{i\theta}^{(1)} - rq_{i\theta}^{(1)}\right\| > \varepsilon_{r_1N_2}\right\} \to 0$$

The bias order of these nonparametric conditional expectations is $O(h_{N_2}^s)$ as shown in Lemma 3 as long as the conditions there are satisfied for $\varphi(z_{1i}, z_{ci}, Z_2; \theta_0) = \frac{\partial q(z_{1i}, z_{ci}; \theta_0)}{\partial \theta}$. The above uniform convergence results hold, using Lemmas A.5, A.6, A.8 and A.9 in Ichimura (1993), if there exist sequences $\{\varepsilon_{rN_2}\}$ and $\{\varepsilon_{r_1N_2}\}$ such that

$$\varepsilon_{rN_2} h_{N_2} M_{rN_2}^{l-1} \to \infty; (\log h_{N_2})(1 + M_{rN_2}\varepsilon_{rN_2})/(N_2 h_{N_2} \varepsilon_{rN_2}^2) \to 0$$
$$\varepsilon_{r_1N_2} h_{N_2} M_{r_1N_2}^{l-1} \to \infty; (\log h_{N_2})(1 + M_{r_1N_2}\varepsilon_{r_1N_2})/(N_2 h_{N_2} \varepsilon_{r_1N_2}^2) \to 0$$

and where M_{rN_2} denotes a sequence of the support of the dependent variable $\rho(z_1, z_c, z_2, \theta)$, M_{1N} denotes a sequence containing 1, $M_{r_12N_2}$ denotes a sequence of the support of the dependent variable $\frac{\partial \rho(z_1, z_c, z_2, \theta)}{\partial \theta}$. These sequences exist as long as $\frac{N_2 h_{N_2}^{2+\frac{2}{l-2}}}{(-\log h_{N_2})} \to \infty$ and $\frac{N_2 h_{N_2}^{\frac{l}{l-2}}}{(-\log h_{N_2})} \to \infty$ as in (2.47). Note that if $l \geq 2$, the former condition on the sequence of bandwidths implies the latter.

Chapter 3

Identification of Preferences in the Pure Characteristics Demand Model with Microdata

3.1 Introduction

Differentiated product models have been widely applied to the adjustment of welfare indices to quality change, welfare analysis of the introduction of new goods, merger analysis and many other policy analysis where the estimation of price elasticities and substitution patterns play a central role.

In this paper we study the identification of preferences in pure hedonic discrete choice models of differentiated products with consumer-level data (i.e. when choices, product characteristics and individual attributes are jointly observed) where individuals derive utility from a finite set of product characteristics and rules out product specific unobserved preferences. These consumer preferences are used in the estimation of the price elasticities and substitution elasticities across products with different attributes.

Some recent papers (Berry and Pakes (2003), Bajari and Benkard (2003, 2005)) have pointed out that standard discrete choice econometric models used to estimate structural models of demand have some undesirable properties when the number of products becomes large, implying counterintuitive implications in policy analysis related with the introduction of new goods.

The main assumption that drives these properties is the existence of the i.i.d error

component -independent across individuals and products - with support in the entire real line that usually is interpreted as an unobserved individual taste for each specific product[1]. We argue that this error term can also be interpreted as unobserved tastes over a set of unobserved product characteristics whose dimension has to be equal to the number of alternatives in the market.

An alternative model proposed in the literature to overcome the counterintuive implications of the standard models brought to the data is a model which does not include the additive random error term with unbounded support. In this work, we study the identification conditions of the preference parameters in both parametric and semiparametric models. The semiparametric model allows one to relax the distributional assumptions on the taste coefficients for each product characteristic. For example, the normality assumption may not be a reasonable distribution for unobserved tastes and any other distribution may be more suitable depending on the characteristic in consideration (some tastes over characteristics may have a truncated distribution, skewed distribution, nonnegative distribution, etc). Estimation of the parameters in the utility function in the standard multinomial discrete choice models is not computational feasible without making assumptions about the distribution of this additive i.i.d term, since nonparametric estimations would suffer from the curse of dimensionality when the number of goods considered is very large. Therefore, in comparison with previous semiparametric models based on standard multinomial discrete choice models (see for example, Lee, L.F. (1995), Matzkin (1991)), this model has the advantage of reducing the dimension of the problem from the product space to the characteristics space[2]. This issue becomes important when markets with a high number of products are considered.

One of the counterintuitive implications of discrete choice models with iid random terms with unbounded support is that each individual utility increases to infinity as the number of products in the market becomes large and the predicted demand probabilities are always positive regardless of the characteristics of the products introduced and for every price vector. Although this last feature is attractive from the computational point of view, it has implications when computing welfare measures by computing the area

[1] Previous work using this specification include, for example, Berry (1994), Berry, Levinshon and Pakes (1995) and Nevo (2000) for aggregate data; Berry, Levinshon and Pakes (1998) and Goldberg (1995) for microdata; and Petrin (1998) using a combination of both.

[2] In the standard discrete choice models including an i.i.d random term, the choice probabilites are computed from a multiple integral whose dimension relies in the number of alternatives.

behind the demand curve, since demand is never zero even if the price tends to infinity. In fact, this property has been found to lead to an overestimation of welfare measures in empirical works by Petrin (2003) and Ackerberg and Rysman (2001). The unobserved differentiability of the products implied by the error structure also makes that no perfect substitutes of any product can be found even when the number of products becomes very large.

Thus, the assumptions about the additive i.i.d error term in the standard discrete choice econometric approaches seem to have been introduced in order to ease computation of the probabilities, although the existence of this additive error term is not directly implied by the pure characteristics framework where products are defined as a finite vector of product characteristics.

We also control for the omitted variables problem that arise due to unobserved (by the econometrician) product characteristics and which are likely to be correlated with product price. Thus, prices are higher for those products which display those unobserved characteristics that are desirable for consumers. Using consumer-level data we could potentially control for this endogeneity problem by estimating product-specific constants in the individual choice probabilities. The model considered in this work accounts only for a unidimensional unobserved product attribute. In some sense, this is a more restrictive model than the standard approach with an i.i.d term: the model here allows only one unobserved component but in the standard model infinite unobserved product factors have to be consider if the number of alternatives becomes infinitely large. We argue that a more general model with multiple unobserved product characteristics whose dimension is not linked to the number of alternatives would be a more desirable model. This constitutes an interesting topic for future research.

After discussing and justifying the model we use in this paper in Section (3.2), we derive the choice probabilities implied by the pure characteristics model in Section (3.3). Section (3.4) gives sufficient conditions for the identification of the preference parameters under both cases where the distribution of the unobserved individual taste drifters is assumed to be known or unknown.

3.2 Demand Model: Notation and Assumptions

Modelling demand using a discrete choice model has been the practice of many of the IO empirical works on demand of differentiated products. Products differentiate from each

other in their inherent characteristics or attributes. This discrete choice setting assumes that individuals are only able to choose one unit of their chosen product as opposed to the continuous demand model where individuals choose the amount of the good they want to consume.

This discrete choice framework differs also from the standard hedonic models[3] where consumers decide the amount of each characteristic to be present in the ideal product maximizing their utility (e.g. consumers can choose the different components or characteristics when purchasing a PC). In the discrete choice case, however, individuals take as given the possible product combination available in the market and choose that product that maximizes their utility. The discrete choice framework complicates the analysis but it is more realistic for those industries in which consumers cannot construct their ideal product or the available products are placed discretely in the characteristics space.

We follow a discrete approach in this paper and consider the choice of one product among a choice set $\Im$. It is assumed that the rest of the goods consumed by each individual or household constitute a composite good denoted by c. Consumers choose that product $j \in \Im$ and that level of consumption of the composite good c that yields the highest level of utility subject to their budget constraint. The utility attained by each individual depends on the level of c attained and on the product chosen by the individual. As in the hedonic literature, the individual utility derived from the consumption of a differentiated product is assumed to be defined over the product characteristics space. This allows one to reduce the number of parameters to be estimated in a demand system, since the substitution patterns across products is reduced to the characteristics dimension regardless of the number of products available.[4]

Moreover, we also allow for some of the product characteristics to be unobserved by the econometrician, but not by the consumers. Obviously, utility is also a function of some individual attributes that make the utility derived from the consumption of one product be different between individuals with different attributes (such as income, family size,

[3]See Rosen (1974), Brown and Rosen (1982), Epple (1987) Ekeland, Heckman and Nesheim (2004), Heckman, Matzkin and Nesheim (2005), for instance.

[4]The traditional consumer analysis prior to the work of Lancaster (1966) and Gorman (1980) worked with individual preferences or orderings defined over the product space. Thier work kead to a complete replacement of this old theory in this respect by assuming that consumer's preferences over products are a function of characteristics or properties intrinsec in marketed goods which allows to reduce significantly the number of elasticities to be computed in a demand system.

etc.). The heterogeneity in tastes over product characteristics induced by these individual characteristics makes individuals to choose different alternatives.

In what follows we assume we have access to consumer-level data or micro data. That it is, the structure of the data is such that information about individual attributes is matched with their choices and the characteristics of these choices. The random utility model underlying the demand choice that it is described in the next section could be defined exactly defined if aggregate data is available. In the aggregate case though, individual attributes need to be integrated out from the individual probabilities in order to obtain the market share for each alternative, since the observations consist of sales and product characteristics in each market.

3.2.1 Notation

We consider data from choices of N individuals in each market/period (indexed by $i \in I = \{1, ..., N\}$) on a choice set containing J different alternatives (indexed by $j \in \Im = \{1, ..., J\}$). The choice random variable by d_i (a $J-$dimensional vector where $d_{ij} = \{0, 1\}$ for all $j \in \Im$. The binary variable d_{ij} equals 1 if individual i chooses product j and $d_{ijt} = 0$ otherwise and where choices are mutually exclusive so that $\sum_{j=1}^{J} d_{ij} = 1$).

The vector of characteristics for product j denoted by X_j. This vector is divided in two components $X_j = \{x_j, \xi_j\}$ where x_j is the vector of observed product characteristics with dimension $(K \times 1)$ and ξ_j is the vector of unobserved product characteristics with dimension $(K' \times 1)$. This unobserved product specific characteristic represents product attributes that are difficult to measure, such as prestige, reliability,quality of any omitted product characteristics. [5]

We assume, as others do, that the observed product characteristics x (excluding price) are exogenous to the model. The matrix of characteristics for all the products is denoted by matrix $X = (X_1, ..., X_J)'$ of dimension $(J \times \kappa)$ where $\kappa = K + K'$.

Let Z be random vector of dimension M that represents observed consumer attributes. Its sample space is denoted by $\Omega_Z \in R^M$. An individual drawn at random from the population will have some attribute vector $z \in \Omega_Z$. Let ε be the vector of dimension E representing unobserved consumer attributes. Its sample space is denoted by $\Omega_\varepsilon \in R^E$. An individual drawn at random from the population will have a realization of unobserved

[5] This component is assumed to be prefectly observed by individuals so that there is no room for any learning preocess about the value of this variable.

random variable $e \in \Omega_E$. The vector of unknown parameters is denoted by θ and Θ denotes the parameter space.

Economic agents make their choices on which j to choose and the level of composite consumption c by maximizing their utility derive form the consumption of both subject to their budget constraint

$$\begin{aligned} \max_{(j \in J, c)} \; U_{ij} &= u(c, X_j, z_i, \varepsilon_i) \\ s.t. \; c + p_j &\leq y_i \end{aligned}$$

where y_i is the level of income of individual i and p_j is the price of product j.

If $u_c > 0$, for a selected product j the utility of individual i satisfies that

$$U_{ij} = u(y_i - p_j, X_j, z_i, \varepsilon_i) \tag{3.1}$$

Utility in (3.1) represents the indirect utility function conditional on the discrete choice j, which is the maximum level of utility that an individual with income y_i when he chooses alternative j. In order to analyze the determinants of the product choice and identify the preference parameters associated to product characteristics is convenient to focus only on the alternative choice. Considering indirect utility functions allows one to abstract from the choice of the other goods affecting individual utilities.

It should be noted that the indirect utility function in (3.1) is giving some information about the way income and price interact in the utility function, giving flexibility to the way in which the rest of individual attributes and product characteristics interact. For simplicity by now, we do not to distinguish between income and any other individual attributes that enter the utility function. In the same way, we treat product price as another product characteristic. We assume for the moment that the interaction between income and product price does not have any particular feature with respect to any other interaction between product characteristics and individual attributes.

The model analyzed in this work is

$$U_{ij} = \theta_i x_j + \xi_j \tag{3.2}$$

where

$$\theta_i = \left(\begin{bmatrix} \theta^1 & \theta^2 \end{bmatrix} \times \begin{bmatrix} z_i \\ \varepsilon_i \end{bmatrix} \right) \tag{3.3}$$

$$= \begin{bmatrix} \sum_{m=1}^{M} \theta^1_{m1} z_{im} + \sum_{e=1}^{E} \theta^2_{e1} \varepsilon_{ie} \\ \vdots \\ \sum_{m=1}^{M} \theta^1_{mK} z_{im} + \sum_{e=1}^{E} \theta^2_{eK} \varepsilon_{ie} \end{bmatrix}$$

and the matrices of parameters θ^1 and θ^2 are of dimension $(K \times M)$ and $(K \times E)$, respectively. The parameter θ^1_{mk} is the preference over product characteristic k driven by the observed individual product attribute m and parameter θ^2_{ek} is the preference over product characteristic k driven by the unobserved individual attribute e. Define the matrix of parameters of dimension $(K \times (M+E))$ as $\theta = \begin{bmatrix} \theta^1 & \theta^2 \end{bmatrix}$. Furthermore, we assume that the unobserved individual attributes are statistically independent of the observed individual attributes. However, the unobserved product characteristics (ξ_j) are not independent of the observed product characteristics (x_j). For example, price is unlikely to be independent of the unobserved product characteristics.

The most important difference of this model with the standard discrete choice models is the fact that here the utility function does not include an additive i.i.d (across individuals and alternatives) random error term with full support on the real line. [6] The next section justifies the assumption of restricting the randomness of the utility function to the random coefficient θ_i by explaining some counterintuitive implications of the model that includes the i.i.d random term with full support when the number of alternatives in the market is very high.

It should be noted that no particular distributional assumption has been imposed on the random error ε. In fact, as it will be discussed later in this work, this model allows us to identify the preference parameters without imposing any restriction on the distributional form of the random coefficients.

Some comments on the utility specification are in order. If vector of individual attributes z includes a constant term across individuals, then the utility function U_{ij} includes a mean utility term for product j (i.e. amount of utility that is common to all the individuals consuming product j).[7] Observed and unobserved heterogeneity in tastes are introduced

[6] This additive i.i.d term is usually assumed to be distributed as Extreme Value - Type I so that the closed form probabilities for the conditional logit model are obtained (McFadden (1974))

[7] It is assumed that random vector ε_i does not include any factor invariant accross individuals since the

in an additive way. If $\theta_2 \neq 0$, individuals with the same observed attributes may still differ in their preferences over product characteristics.

Also, the specification in (3.2) assumes that individual tastes on product characteristic are linear in observed and unobserved individual attributes. The non-linearity of tastes in observed attributes could be easily relaxed using our approach while the nonlinearity in unobserved attributes would imply more difficulties in the estimation here suggested. [8]

We also assume throughout this work that the unobserved product characteristics is unidimensional (i.e. $K' = 1$). This is a restrictive assumption since it implies that all omitted product characteristics can be summarized in a unique factor. There have been some attempts to extend the model to the case of multiple unobserved characteristics in the marketing literature (Elrod (1988), Chintagunta (1994) and Elrod and Keane (1995) using factor analysis). The mixed logit model used in the identification in Ben-Akiva, Bolduc and Walker (2003) and Walker (2002) allows for a multidimensional unobserved product-specific variables with heterogeneous tastes across individuals. In the latter papers, the identification however is based in a more restricted model with heteroskedasticity and where no random coefficients on observed product characteristics are considered. However, any of these papers considers the possibility that the omitted product characteristic may be correlated with the observed product characteristics.

When the endogeneity issue is considered, identification becomes harder and a unidimensional unobserved product characteristic is usually assumed in the literature. Under the assumption that ξ_j is correlated with x_j and $K' = 1$, Berry and Pakes (2003), BLP (1995) and Bajari and Benkard (2005)) study the identification of the parameters in the parametric case when only aggregate data is available.

It has been argued that these identification results under endogeneity of the product characteristics extend to the case where unobserved product characteristic is multivariate and the second term in (3.2) reduces to an index of unobserved factors (Bajari and Benkard (2005)). However, we think that the treatment that the unobserved product characteristic has received in the literature is imposing restrictions on the way random coefficients enter in the utility function and it is actually not straight forward to extend it to the case of multidimensional unobserved product characteristic as it has been argued.

Assuming a unidimensional unobserved factor ξ allows one to ignore the possibility of

mean utility parameter imbedded in θ_2 would not be separately identified from the mean utility parameter in θ_1.

[8] See Brown and Walker (1989) for non-linear assumption in function u.

individual specific taste on unobserved product characteristics. Suppose that we want to consider unobserved tastes on ξ in utility (3.2) so that

$$U_{ij} = \beta_i X_j + \alpha_i \xi_j \tag{3.4}$$

As long as the unobserved product characteristic is a vertical characteristic ($\alpha_i > 0, \forall i$, every individual positively values characteristic ξ), the individual utility can be rescaled by α_i without changing the ordering over the alternatives. Therefore, in the unidimensional case, the assumption in (3.2) of no heterogeneity in tastes on ξ is innocuous. However, when the dimension of the unobserved product characteristics is greater than one ($K' > 1$) this procedure is no longer valid and the distribution of the unobservables factors α_i need also to be taken into account when computing the choice probabilities.

However, rescaling the utility by α_i has some costs on the assumptions we need to impose on the random coefficient $\theta_i$in (3.2) and on the random coefficient before rescaling (denoted by β_i) if they are assumed to be linear in both z and ε.

The utility function in (3.4) is equivalent to (3.2) if

$$\beta_i = \left(\begin{bmatrix} \beta_{1i} & \beta_{2i} \end{bmatrix} \times \begin{bmatrix} z_i \\ \varepsilon_i \end{bmatrix} \right)$$

and either (i) β_{1i} and β_{2i} are proportional to α_i (i.e. $\frac{\beta_{1i}}{\alpha_i} = \theta_1; \frac{\beta_{2i}}{\alpha_i} = \theta_2$); or (ii) $\beta_{1i} = \beta_1$; $\beta_{2i} = \beta_2$; $\alpha_i = \alpha$. In other words, in order for utility expression (3.2) to be consistent with heterogeneous tastes on ξ, some linearity restrictions and proportionality assumptions need to be imposed on the random coefficient over observed product characteristic β_i.

In case (i), we should reinterpret the taste coefficients on observable product characteristics relatively to individual tastes on unobserved product characteristics ξ_j.

The additivity of ξ_j without individual random coefficient is needed for the identification results presented below.

In order to allow for a more general model, another factor (α_i) should be included in the model apart from the unobserved taste drifter on observed product characteristics ε_i.

3.2.2 Difference in assumptions on unobservables with the standard discrete choice models: Justification of our specification

The main difference of the model outlined above with the standard discrete choice models normally used in empirical applications is that the latter includes an additive random error in the utility function (3.2) specified above, say ϵ_{ij} with different realizations for each

individual and alternative, with full support in the entire real line which are i.i.d across individuals i and products j. The models including ϵ_{ij} have been denoted in the previous literature as *taste-for-products* models versus *taste-for-characteristics* models which eliminate this additive random term. The interpretation and implications of this additive random error have received some attention in the recent literature. The basic references are Berry and Pakes (2002) and Bajari and Benkard (2003, 2005).

It has been argued that the assumptions imposed on the additive random term ϵ_{ij} in the empirical applications are only justifiable in terms of reducing the complexity of solving the multidimensional integrals involved in the computation of the choice probabilities, since in some cases these assumptions give a closed form for the high-dimensional integrals. Although some restrictions on the distribution function of the unobservables are unavoidable in order to identify some of the structures of the model, some of the usual parametric assumptions on the unobservable ϵ_{ij} have some counterintuitive implications in welfare computations and effects of the introduction of new goods when the number of products in the market becomes large (Bajari and Benkard (2003), Berry and Pakes (2003), Caplin and Nabeluff (1991), Andersen, de Palma and Thisse (1992), Petrin (1998)). Bajari and Benkard (2003) list the properties of the models incorporating these restrictive assumptions and specify which particular assumptions yield each of the implications. We comment them in the remaining of this section.

One of the main assumptions driving these implications is that the additive error ϵ is a continuous random variable with unbounded support on the entire real line.

The first implication that should be pointed about this assumption is that the choice probabilities for each alternative are strictly positive regardless of the value of the product characteristics. This feature is particularly undesirable in a structural model of demand since implies that the choice probabilities are strictly positive for every vector of prices. Thus, even if there are two products with the same (observed and unobserved) product characteristics but with very different prices, there is some small positive probability that random error ϵ takes those values that are able to overcompensate for the price disutility so that higher utility is obtained from the consumption of the high priced product. The random term introduces some unobserved differentiation of each product with respect to the other marketed products. The fact that the choice probabilities are strictly positive for any value of the product characteristics, individual attributes and/or value of the parameters in the utility function is however highly convenient for computational

purposes[9].

If the random variable ϵ has full support on the real line, then the probability that ϵ_j is greater (in absolute value) that any $M < \infty$, conditioned on the errors of the rest of products smaller than ϵ_j, $|\epsilon_{-j}| \leq M$ is strictly positive (which is to say, $\Pr(|\epsilon_j| \leq M | |\epsilon_{-j}| \leq M) \leq 1 - \delta_M$ for $0 < \delta_M < 1$). When the number of alternatives becomes very large, there is a positive probability that the maximum of the values of ϵ across alternatives $j = 0, ..., J$ is greater than any finite number. Formally,

$$\begin{aligned} \Pr(\max_{0 \leq j \leq J} |\epsilon_j| \leq M) &= \Pr(|\epsilon_J| \leq M, |\epsilon_{J-1}| \leq M, ..., |\epsilon_1| \leq M) \qquad (3.5) \\ &= \Pr(|\epsilon_J| \leq M | |\epsilon_{-J}| \leq M) \cdot \Pr\left(|\epsilon_{J-1}| \leq M | |\epsilon_{-(J-1)}| \leq M\right) \cdot ... \\ &... \cdot \Pr(|\epsilon_2| \leq M | |\epsilon_1| \leq M) \cdot \Pr(|\epsilon_1| \leq M) \leq \\ &\leq (1 - \delta_M)^J \to 0 \text{ as } J \to \infty \text{ as } 0 < \delta_M < 1 \end{aligned}$$

Note that the assumption of i.i.d errors has not been imposed here. Therefore, even if one uses more advanced models that avoid the IIA assumption- such as the Nested Logit which relaxes the independence assumptions of alternatives within the same nest, the property in expression (3.5) holds when the number of alternatives is large (even if they are grouped in nests) as long as the unboundedness of the conditional distribution $\epsilon_j | \epsilon_{-j}$ is satisfied.

The property of ϵ in (3.5) makes that there exists a positive probability that the level of utility corresponding to the most preferred alternative goes to infinity. This has obvious implications for welfare evaluations of changes in the choice set since the consumers utility attached to those alternatives that have been added or removed from the choice set might be non-finite. Therefore, as the number of products increases in the market, the compensating variation for each individual of all products in the market tends to infinity. This is believed to be overestimating the welfare benefits associated to the variety of available products because of the high sensitivity of the utility to high realizations of ϵ_j.

If in addition to the full support assumption, the distribution function of ϵ has thick tails (formally, the hazard rate of ϵ does not tend to infinity when ϵ tends to the upper limit of its support), then the expected difference between the highest and the second

[9] As it will become clear in the explanation of the choice probabilities in the context of a pure characteristics demand model, the fact that for certain values of the parameter space the choice probabilites become zero increases the complexity of the computation since the likelihood function becomes discontinuous in the parameters.

highest tends to infinity as $J \to \infty$. Thus, consumers suffer from infinity welfare losses when their first choice is eliminated from the choice set, meaning that products do not become perfectly substitutes even if their number increases to infinity. When the number of alternatives tends to infinity, one would expect that products are allocated very close to each other in the characteristics space so that they become perfect substitutes. The reason is that the random error ϵ induces some unobserved differentiation of each product with respect to the rest and because of the full support assumption, the value attached to this specific differentiation of product j is likely to be above any finite number.

All these implications do not seem reasonable if the discrete choice framework wants to be used as a structural model of demand. However, some features of the structural model of demand would be more affected by this assumptions than others. Thus, large realizations of ϵ affect more welfare evaluations of changes in the number of products than the estimation of the unknown parameters in the utility function (Berry and Pakes (2003)).

These properties suggest that the model used in the empirical applications to estimate structural demand parameters in a demand framework impose certain assumptions on the error structure of the utility which have some undesirable properties. A model like the one described above in (3.2) and (3.3) does not imply any of these properties since no additive error term has been assumed. Instead of removing this random error, the model analysed in Ackerberg and Rysman (2001) where the variance of the error terms ϵ in a taste-for-products model depend on the number of available products is another alternative to the tastes-for-products model.

3.2.3 Interpretation of the i.i.d term ϵ_{ij} : Unobserved tastes over characteristics vs product specific unobserved tastes

It is important to clarify the different behavioral interpretations of the i.i.d error term ϵ_{ij} in the standard multinomial discrete choice model and the error term $\alpha_i \xi_j$ in the taste-for-characteristics model.

The term $\alpha_i \xi_j$ is the unobserved taste for individual i over unobserved product characteristic ξ_j. The unobserved characteristic ξ has different values across products but it is an inherent attribute of every marketed product j, although unobserved to the econometrician.

The standard interpretation given to i.i.d ϵ_{ij} is a product-specific unobserved taste. In

other words, each product offers a specific unobserved attribute that it is not possible to be obtained through the consumption of a different product and each individual has an unobserved taste over that attribute.

Take for example the design of one product as an unobserved characteristic by the econometrician. Suppose there exist a product such that its design is so specific that this attribute might be considered as a something very genuine of the product, over which each individual has different tastes. Thus, this standard interpretation of ϵ_{ij} means that each product has its own unobserved product characteristic that makes it specifically (and unobservably) different from the rest of the products.

The nature of the unobserved characteristics may be such that, for example, different products share the same type of design and with respect to this characteristic they are almost indistinguishable. In this case, the unobserved tastes over unobserved product characteristics would be preferably expressed as $\alpha_i \xi_j$. Which type of unobserved characteristic we have depends on the nature of the products we study.

One way of obtaining i.i.d. random terms ϵ_{ij} across i and j is by interacting product dummies with a $J-$dimensional vector of independent tastes for individual i[10]. In this way, it is easy to understand how the random errors are interpreted as product specific unobserved heterogeneity. However, this way of constructing ϵ_{ij} is only a sufficient condition in order for the errors to be i.i.d across i and j.

In order to find necessary conditions that need to be satisfied when random errors ϵ_{ij} are i.i.d across individuals and products, it is convenient to express the error component ϵ_{ij} as a linear combination of unobserved tastes for individual i (v_i) over a $Jth-$dimensional vector $\mu(.,j) = [\mu(1,j)...\mu(J,j)]$ of unobserved product characteristics for product j as follows

$$\epsilon_{ij} = \sum_{r=1}^{J} \mu(r,j) v_{ir} \tag{3.6}$$

where $\mu(r,j)$ corresponds to the $r-th$ unobservable characteristic of product j and v_{ir} is the $i-th$ individual preference for unobserved product attribute r.

Let $E_i = [\epsilon_{i1}\ \epsilon_{i2}...\epsilon_{iJ}]'$; $V_i = [v_{i1}\ v_{i2}...v_{iJ}]'$; $\Sigma_j = [\mu(1,j)\ ...\ \mu(J,j)]'$. Let denote the

[10] That it is, random term ϵ_{ij} can be expressed as

$$\epsilon_{ij} = \zeta_j * \eta_i$$

where ζ_j is a $J-$dimensional vector with zeros but a one in the jth element, and η_i is a $J-$dimensional taste vector.

matrix of unobserved product characteristics of dimension $J \times J$ as $\Sigma = [\Sigma_1 \ ... \ \Sigma_J]'$. Then, for individual i the matricial expression for the $J-th$ dimensional vector E_i is

$$E_i = \Sigma \times V_i$$

If the errors v_{ir} are independent across r and matrix Σ is diagonal, then obviously ϵ_{ij} is also i.i.d across i and j (See Appendix A.1) and it can be interpreted as above as a specific unobserved taste of individual i for product j. A diagonal Σ implies that each unobserved characteristic is specific of only one product. Thus, because $\mu(j,j) \neq 0$ and $\mu(r,j) = 0 \ \forall r \neq j$, the $jth-$unobserved product characteristic is specific of product j and it does not play any role in describing the preferences over any other product different from j.

However, the independence of V_i and diagonality of Σ are not necessary conditions to obtain i.i.d. errors ϵ_{ij} (See Appendix A.2). Take the normal case as an example, the random errors V_i do not need to be independent and they may have a covariance structure such that there exist some linear combinations (coefficients in Σ) which make the variance and covariance structure of the resulting errors equal to zero. The interpretation of ϵ_{ij} is different with this structure. The random error ϵ_{ij} can be interpreted then as an index of unobserved tastes (not necessarily independent) over a vector of unobserved product characteristics whose dimension coincides with the number of alternatives as it is clear from expression (3.6). The vector E_i would not be independent across j if the dimension of the vector of unobserved product characteristics is smaller than J (i.e. if the dimension of Σ is $J \times H$, where $H < J$).

This last interpretation of random terms ϵ_{ij} has the advantage of making easier the comparison between a *taste-for-products* model (including ϵ_{ij}) and a *taste-for-characteristics* model (with random coefficients for instance and unobserved product characteristics). While the latter can incorporate any number of unobserved product characteristics (although we study identification only for the unidimensional case), the i.i.d error term of the taste-for-products model can be considered also as an index of unobserved tastes on unobserved product characteristics but the number of product factors considered has to be equal to the number of alternatives. Therefore, while the dimension of ξ in the taste-for-characteristics model may remain constant as the number of products J increases, the dimension of matrix Σ in (3.6) increases when the number of products increases in the market.

The taste-for-product model and the taste-for-characteristic model with unidimen-

sional ξ_j are two extreme ways of describing preferences. Given that the dimension of the unobserved product characteristic has been restricted to the unidimensional case, it may be difficult to capture all the unobserved heterogeneity in tastes for unobserved products with only one attribute. A more realistic model and preferred to both models would be an extension of the taste-for-characteristics model which allows for a higher dimensional vector of unobserved product characteristics.

Even in this multidimensional factor framework, it would be interesting to develop tests that allow us to determine which model explains better the observed data. One would like to assess whether a model with specific taste over products has some additional explanatory power with respect to a model where there exists unobserved heterogeneity in tastes over a vector of unobserved product characteristics captured by ξ_j.

Which of the two models is considered as a more general model depends on which interpretation we take. If we interpret the error term as unobserved individual tastes on product dummies, then the taste-for-products can be regarded as a particular case of the taste-for-characteristics model where product dummmies are considered as product characteristics. If we interpret ϵ_{ij} as unobserved individual taste over a $J-$dimensional vector of unobserved product characteristics, then the taste-for-products characteristics can be considered as more general because it allows for a higher dimensional vector of unobserved components (although restrictive because the dimension of this vector needs to be the same as the number of alternatives).

3.2.4 Advantage of the approach in the Semiparametric Approach: Dimensionality Reduction

The main advantage of the model in (3.2) and (3.3) is that reduces the dimensionality of the problem and there is no need to introduce a different factor when the number of products in the market increases. Thus, if we believe that specific tastes over products does not explain more of the decision process when unobserved tastes for product characteristics have been included in the model, the model in (3.2) and (3.3) allows one to reduce the dimension of the problem and consequently relax some of the parametric assumptions on the distribution of the taste coefficients that have been usually imposed in the previous related literature.

In a taste for product model, the choice probabilities for product j depend on the unobserved taste components for the rest of the products in the choice set J (i.e. ϵ_{ir} for

$r \neq j$). In order to ease the computation of such probabilities, standard discrete choice models have imposed independence ϵ_{ij} across products and across individuals and have also assumed that each error term is distributed as Extreme Value- Type I, since in this case a reduced form of the choice probabilities can be obtained. Even with a small deviation from the Extreme Value distribution, choice probabilities need to rely on simulations and distributional assumptions that may or may not be appropriate to describe the tastes in our population.

There have been some previous works analyzing the semiparametric identification of parameters in a multinomial discrete choice model by relaxing the distributional assumption of the errors. Lee (1995), for example, introduces a multinomial version of Klein and Spady's semiparametric estimator. His model does not allow for random coefficients, he assumes an additive i.i.d random term and his estimator depends on $J-1$ random variables.[11] Although under certain conditions on the taste for products model, it may be possible to semiparametrically identify the preference parameters, its computation may be cumbersome due to the curse of dimensionality that would involve the computation of choice probabilities depending at least on $J-1$ random variables when the number of products in the market J is relatively high.

If a random error term iid with unbounded support ϵ_{ij} is added to the model in (3.2) and (3.3) the probability of choosing product j can be expressed as

$$P_{ij}^{P} = \Pr\left(\theta_i(x_j - x_r) + (\xi_j - \xi_r) \geq \epsilon_{ir} - \epsilon_{ij} \text{ for all } r \neq j \middle| x\right)$$

Under the i.i.d assumption of random errors ϵ_{ij},

$$P_{ij}^{P} = \int \left[\prod_{r \neq j} F_{\upsilon_r|x}\left(\theta_i(x_j - x_r) + (\xi_j - \xi_r)\right)\right] dF_{\theta|x}(t|x)$$

where $\upsilon_r = \epsilon_{ir} - \epsilon_{ij}$ is equally distributed for all $r \neq j$.

This probability depends on $(J-1)$ indices. These indices are the points at which we evaluate each of the $(J-1)$ distribution functions in the expression above. The existence of

[11] For the binary model, there exist other works studying the identification and estimation of the parameters without distributional assumptions imposed on the stochastic term (Cosslett (1983), Ichimura and Thompson (1998), Klein and Spady (1993), Manski(1975) and Matzkin (1992)). Matzkin (1992) and Kemp (2000) relax the restrictions on the structure of the systematic function of the exogenous observables in utility and Matzkin (1991) studies this issue in the multinomial case. In this work we do not deal with semiparametric estimation of the utility function. Instead we impose the linearity assumption for the utility function but we relax the assumption on the distribution of the unobserved factors.

a random term that is product specific forces one to rely on the product space dimension, without being able to reduce that dimensionality.

One plausible alternative that one would think is that assuming a logistic distribution for the product specific unobserved taste would allow us to reduce in some sense the number of indices of which the choice probability depends on. However, even in the case where $F_{\upsilon_r|x}$ is distributed as a Extreme Value-Type I and the distribution of θ is unknown, it is not possible to reduce the number of indices on which this probability depends. Under this parametric distributional assumption, the probabilty of choosing product j is expressed as

$$P_{ij}^{P} = \int \left[\frac{\exp(\theta_i x_j + \xi_j)}{\sum_{j=0}^{J} \exp(\theta_i x_j + \xi_j)} \right] dF(\theta|x)$$

The fact that our model does not include this alternative-specific additive random term allow us to find a semiparametric estimator whose dimension depends on the number of characteristics K used to define the products or equivalently on the dimension of the unobservable variables explaining the tastes over product characteristics. Thus, for those industries in which the number of marketed products is high and few observed characteristics can be used to describe them, the estimator we suggest is more tractable that the ones previously proposed in the literature. In this version though, we only study a simplified version of this more general with a unidimensional unobserved factor.

3.3 Choice Probabilities

The choice probabilities in the likelihood function are obtained from a model of utility-maximizing behavior of the decision-makers. Consider the random utility function for the taste-for-characteristics model in (3.2) and (3.3). In order to simplify the computation of the choice probabilities we assume that there exist only a unidimensional vector of unobserved tastes over product characteristics ε (i.e. $E = 1$). Let $F_\varepsilon : \Omega_\varepsilon \subset R \rightarrow [0,1]$ denote the distribution function of the unobserved individual attributes. The choice variable for each individual i and product j is

$$d_{ij} = 1\left\{(\theta_1 z_i)' x_j + (\theta_2 \varepsilon_i)' x_j + \xi_j \geq (\theta_1 z_i)' x_r + (\theta_2 \varepsilon_i)' x_r + \xi_r, \text{ for all } r \neq j\right\} \quad (3.7)$$

The observed choice for individual i with attributes z_i can be viewed as drawings from a multinomial distribution with selection probabilities $\Pr(d_j = 1|X, z_i; \theta, F_\varepsilon)$ for each $j \in \Im$ which are expressed as follows

$$\Pr\left(d_j = 1|X, z_i; \theta, F_\varepsilon\right) =$$

$$\Pr\left(\begin{array}{c}(\theta_1 z_i)' x_j + (\theta_2 \varepsilon_i)' x_j + \xi_j \geq \\ (\theta_1 z_i)' x_r + (\theta_2 \varepsilon_i)' x_r + \xi_r, \text{ for all } r \neq j|X, z_i; \theta, F_\varepsilon\end{array}\right) = \tag{3.8}$$

$$\Pr\left(\begin{array}{c}(\theta_1 z_i)' [x_j - x_r] + [\xi_j - \xi_r] \geq \\ (\theta_2 \varepsilon_i)' [x_r - x_j], \text{ for all } r \neq j|X, z_i; \theta, F_\varepsilon\end{array}\right) \tag{3.9}$$

The unidimensional assumption on ε is a strong restriction since it assumes that the random variable determining the random preference for product characteristics is the same across characteristics. The same realization of variable ε explains unobserved tastes across K product characteristics. Conditioning on observable attributes, this univariate random variable ε is the only element that generates unobserved heterogeneity in tastes and choices. Nonetheless, the particular utility specification we consider relaxes somehow this dimensionality restriction by assuming different coefficients for each product characteristic k. Different values of the coefficients in θ_2 allows one to obtain different variances for the random coefficients across product characteristics.

When ε is assumed to be unidimensional, the conditional choice probabilities for product j are

$$\Pr(d_j = 1|X, z_i; \theta, F_\varepsilon) =$$

$$\Pr\left(\begin{array}{c}(\theta_1 z_i)' [x_j - x_r] + [\xi_j - \xi_r] \geq \\ \theta_2' [x_r - x_j] \varepsilon_i, \text{ for all } r \neq j|X, z_i; \theta, F_\varepsilon\end{array}\right)$$

The exact expression of this probability choice depends on the sign of the inner product $\theta_2'(x_r - x_j)$. Conditioned on a particular value of θ_2, products can be ordered with respect to this inner product. Products with respect to this ordering are indexed by $(j)^{\theta_2}$ for $j \in \Im$. This notation reflects the fact that this ordering depends on the value of parameter θ_2

$$\theta_2' x_{(1)^{\theta_2}} < ... < \theta_2' x_{(j-1)^{\theta_2}} < \theta_2' x_{(j)^{\theta_2}} < \theta_2' x_{(j+1)^{\theta_2}} < ... < \theta_2' x_{(J)^{\theta_2}} \tag{3.10}$$

When computing the choice probabilities one should take into account the fact that there are products above and below product j with respect to this ordering, since when isolating ε the sign of the inner product is of relevance now.

If product j is an intermediate good with respect to ordering θ_2 i.e. $j \neq (1)^{\theta_2}$ and $j \neq (J)^{\theta_2}$,

$$\Pr\left(\begin{array}{l}\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\geq\varepsilon_i,\ \text{for all } (s)^{\theta_2}>j\\ \frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\leq\varepsilon_i,\ \text{for all } (s)^{\theta_2}<j\end{array}\right)$$

$$\Pr\left(\begin{array}{l}\max_{(s)^{\theta_2}<j}\left[\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\right]\leq\varepsilon_i\\ \leq\min_{(s)^{\theta_2}>j}\left[\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\right]\end{array}\right) \tag{3.11}$$

If product j is the first product with respect to the ordering θ_2, i.e. $j=(1)^{\theta_2}$,

$$\begin{aligned}&\Pr\left(\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\geq\varepsilon_i,\ \text{for all } s\neq j\right)\\ =\ &\Pr\left(\min_{s>(1)^{\theta_2}}\left[\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\right]\geq\varepsilon_i\right)\end{aligned} \tag{3.12}$$

If product j is the last product with respect to the ordering θ_2, i.e. $j=(J)^{\theta_2}$,

$$\begin{aligned}&\Pr\left(\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\leq\varepsilon_i,\ \text{for all } s\neq j\right)\\ =\ &\Pr\left(\max_{s<(J)^{\theta_2}}\left[\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\right]\leq\varepsilon_i\right)\end{aligned} \tag{3.13}$$

The notation to be used in the choice probabilities is introduced below. The two products that maximize and minimize the lower and upper bound for ε, respectively, in the choice probabilities for product j in (3.11) are denoted by

$$\begin{aligned}R_j(\theta,X_j,X_{-j,\theta_2},z_i)\ &\equiv\ \arg\max_{(s)^{\theta_2}<j}\left[\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\right]\\ r_j(\theta,X_j,X_{+j,\theta_2},z_i)\ &\equiv\ \arg\min_{(s)^{\theta_2}>j}\left[\frac{1}{\theta_2'(x_s-x_j)}\left[(\theta_1 z_i)'\left[x_j-x_s\right]+\left[\xi_j-\xi_s\right]\right]\right]\end{aligned}$$

where

$$\begin{aligned}X_{-j,\theta_2}\ &=\ \left\{\left(x_{(s)^{\theta_2}},\xi_{(s)^{\theta_2}}\right)\ \text{for } (s)^{\theta_2}<j\right\}\\ X_{+j,\theta_2}\ &=\ \left\{\left(x_{(s)^{\theta_2}},\xi_{(s)^{\theta_2}}\right)\ \text{for } (s)^{\theta_2}>j\right\}\end{aligned}$$

This notation indicates that the product characteristics that affect the probability of choosing product j are the characteristics corresponding to the products placed in the boundary of a specific order with respect to product j (i.e. products R and r). It should be also noted that the product denoted by r (R) depends on the value of the parameters

θ, on (unobserved and observed) characteristics of product j, (unobserved and observed) characteristics of products below (above) product j with respect to the ordering in (3.10) and also on the individual attributes. For simplicity of notation, we denote these two products only as a function of the value of the parameters (i.e. $R_{ij}(\theta)$ and $r_{ij}(\theta)$).

It is important to notice that the two sets of parameters affect in a different the determination of products $R_{ij}(\theta)$ and $r_{ij}(\theta)$. Thus, only the part of the parameters corresponding to θ_2 affect the ordering in (3.10) which selects those products lying below or above product j. However which products maximizes or minimizes the indices in (3.11) depend on the particular values of both θ_1 and θ_2.

Once we substitute these products in the indices in (3.11), then the choice probabilities can be expressed as functions of the cdf of the unobserved individual characteristic F_ε evaluated at those two indices.

Let denote by $\bar{\Delta}_j$ the upper bound of ε in the choice probability of product j

$$\bar{\Delta}_j(\theta, X_j, X_{+j,\theta_2}, z_i) = \frac{(\theta_1 z_i)' \left[x_j - x_{r_{ij}(\theta)}\right] + \left[\xi_j - \xi_{r_{ij}(\theta)}\right]}{\theta_2' \left(x_{r_{ij}(\theta)} - x_j\right)} \tag{3.14}$$

and denote by $\underline{\Delta}_j$ the lower bound

$$\underline{\Delta}_j(\theta, X_j, X_{-j,\theta_2}, z_i) = \frac{(\theta_1 z_i)' \left[x_j - x_{R_{ij}(\theta)}\right] + \left[\xi_j - \xi_{R_{ij}(\theta)}\right]}{\theta_2' \left(x_{R_{ij}(\theta)} - x_j\right)} \tag{3.15}$$

For $j \neq (1)^{\theta_2}$ and $j \neq (J)^{\theta_2}$,

$$\Pr(d_j = 1|X, z_i; \theta, F_\varepsilon) = 1\left\{\bar{\Delta} > \underline{\Delta}\right\} \times \left[F_\varepsilon\left(\bar{\Delta}_j(\theta, X_j, X_{+j,\theta_2}, z_i)\right) - F_\varepsilon\left(\underline{\Delta}_j(\theta, X_j, X_{-j,\theta_2}, z_i)\right)\right] \tag{3.16}$$

The indicator $1\left\{\bar{\Delta} > \underline{\Delta}\right\}$ avoids computing negative probabilities. This property however complicates the estimation of the model by maximum likelihood since the loglikelihood function is not defined for those values of the parameters for which there is at least one individual for which the probability of selecting his observed choice is equal to zero. Moreover, note that the choice probabilities are not continuous with respect to parameter θ_2. Note that a small change in θ_2 might alter the inner product ordering in (3.10). At the same time, this implies that the products that maximize and minimize the lower and upper bound indices in the choice probabilities should be obtained from different sets of products. Since both bounds depend on the product characteristics, which are assumed to be exogenous, there is not guarantee that the new interval for ε would imply a value of the

choice probability close enough to the probability evaluated at a slightly different value of θ_2. This feature of the model not only implies some difficulties in the identification of the model, but also sets some challenges in order to find those values of the parameter that maximizes that the likelihood function. [12]

For $j = (1)^{\theta_2}$, the choice probability is

$$\Pr(d_j = 1|X, z_i; \theta, F_\varepsilon) = F_\varepsilon\left(\bar{\Delta}_j(\theta, X_j, X_{-j,\theta_2}, z_i)\right)$$

and for $j = (J)^{\theta_2}$,

$$\Pr(d_{ij} = 1|X, z_i; \theta, F_\varepsilon) = 1 - F_\varepsilon\left(\underline{\Delta}_j(\theta, X_j, X_{+j,\theta_2}, z_i)\right)$$

3.4 Identification

Using notation as in Koopmans and Reiersol (1950), a structure belonging to the model described in (3.2) and (3.3) is defined as $S = (h, F_\varepsilon)$, where h is the structural relationship linking observable and unobservable variables that in this case is expressed as

$$d_{ij} = h(x, \xi, z_i, \varepsilon_i) \tag{3.17}$$

for all i and j and F_ε is the distribution function of the unobservables.

A model is defined by that set of structures that share a set of apriori knowledge or conditions on both h and F_ε. Let denote by $\Pr(d_j = 1|X, Z; S)$ the choice probabilities for alternative j generated by structure S.

Definition 1 *Two structures S and S' are observationally equivalent if*

$$\Pr(d_j = 1|X, Z; S) = \Pr(d_j = 1|X, Z; S')\forall j \in \Im \text{ a.e. in } Z$$

Definition 2 (Parametric Model) *A parametric model Γ_P is defined as that set of structures $S = (h, F_\varepsilon)$ such that (i) relationship h is*

$$d_{ij} = 1\left\{(\theta_1 z_i)' x_j + (\theta_2 \varepsilon_i)' x_j + \xi_j \geq (\theta_1 z_i)' x_r + (\theta_2 \varepsilon_i)' x_r + \xi_r, \text{ for all } r \neq j\right\}$$

[12] In some of the preliminar Montecarlo experiments of the maximum likelihood estimation of this model (not presented in this work) different alternatives have been tried in order to be able to obtain the globa maximum of the likelihood function. Some of them include the use of the simulated annealing algorithm as the optimisation method and also the MCMC method proposed by Chernozhukoz and Hong (2003)

Thus, the utility function is specified as in (3.2) and (3.3) with $K' = 1$ and $E = 1$ and its functional form is known up to a finite number of parameters θ; and (ii) F_ε is a known distribution function. The different structures belonging to this parametric model are defined giving specific values to the vector of parameters $\theta \in \Theta \subset \mathcal{R}^{\dim(\theta)}$

Let Φ denote the space of all probability distributions on the real line.

Definition 3 (Semiparametric Model) *A semiparametric model Γ_{SP} is defined as that set of structures $S = (h, F_\varepsilon)$ such that (i) relationship h is defined in the similarly to the parametric model with $K' = 1$ and $E = 1$; and (ii) $F_\varepsilon \in \Phi$ is a continuously differentiable unknown function, strictly increasing, $0 < F_\varepsilon(e) < 1$ for every $e \in \mathcal{R}$ and $F_\varepsilon(0) = 0.5$. The different structures belonging to this semiparametric model are defined giving specific values to the pair $\{\theta, F_\varepsilon\} \in \Theta \times \Phi$ satisfying the above conditions (i) and (ii)*

Let $\theta^0 \in int(\Theta)$ denote the true value of the parameters and $F_\varepsilon^0 \in \Phi$ denote the true distribution function of the unobserved variable ε

Definition 4 (Parametric Identification) *The true value of the parameter θ^0 is identified with respect to parameter value $\theta \neq \theta^0 \in \Theta$ if there exists at least one product $j \in \Im$ such that*

$$\Pr(z \in \pi_j(\theta)) > 0$$

where

$$\pi_j(\theta) \equiv \left\{z \in Z \text{ such that } \Pr(d_j = 1|X, z; \theta, F_\varepsilon^0) \neq \Pr(d_j = 1|X, z; \theta^0, F_\varepsilon^0)\right\} \tag{3.18}$$

Definition 5 (Semiparametric Identification) *The true value of the parameter θ^0 and the true distribution function F_ε^0 are identified with respect to another pair $(\theta, F_\varepsilon) \in \Gamma_{SP}$ such that $F_\varepsilon(\varepsilon) \neq F_\varepsilon^0(\varepsilon)$ a.e in $\varepsilon \in \Omega_\varepsilon$ and $\theta \neq \theta^0$ if there exists at least one product $j \in \Im$ such that*

$$\Pr(z \in \pi_j(\theta, F_\varepsilon)) > 0 \tag{3.19}$$

where

$$\pi_j(\theta, F_\varepsilon) \equiv \left\{z \in Z \text{ such that } \Pr(d_j = 1|X, z; \theta, F_\varepsilon) \neq \Pr(d_j = 1|X, z; \theta^0, F_\varepsilon^0)\right\}$$

13

[13] The identification of the parameters implies that the limiting likelihood function

$$L_\infty(\alpha) = E\left(\sum_{j=0}^{J} \Pr\left(d_j = 1|z, X, \theta^0; F_\varepsilon^0\right) \log \Pr\left(d_j = 1|z, X, \theta; F_\varepsilon\right)\right)$$

In the next sections, we discuss sufficient conditions for the identification of parameters θ^0 in the case that the distribution function F_ε is known and conditions in order to identify $(\theta^0, F_\varepsilon)$ in the semiparametric case when F_ε is unknown.

3.4.1 Identification conditions for the parametric model

The following assumptions are made for the identification of θ_0 when F_ε is assumed to be known:

Assumption 3. 1 *The unobserved individual attributes ε is a unidimensional factor ($E = 1$) with a known distribution function denoted by $F_\varepsilon : \Omega_\varepsilon \subset \mathcal{R} \rightarrow [0,1]$ such that (i) $F_{\varepsilon|z} = F_\varepsilon \ \forall z \in \Omega_Z$; (ii) differentiable; (iii) strictly monotonically increasing on its support and (iv) $0 < F_\varepsilon(\varepsilon) \leq 1$ for any $\varepsilon \in R$*

Assumption 3. 2 *$E(\xi_j|x) = \varphi(x_j) \neq 0$ for all $j \in \Im$*

Assumption 3. 3 *There exists at least one product characteristic $k \in K$, such that $x_{jk} \neq x_{gk}$ for each pair $j, g \in \Im$*

Assumption 3. 4 *$rank(X) = K$ with $K < J$ and there is no proper linear subspace of $\mathcal{R}^M$ having probability one under the probability distribution of Z, F_Z*

Assumption (3) rules out the possibility that there exist two products with exactly the same available observed characteristics. This implies that it is not possible to have two identical rows in the matrix of product characteristics X. This assumption is important in order for our choice probabilities to be well defined and for our identification strategy.

Assumption (4) ensures full column rank of the matrix of product characteristics and also a limiting rank condition for the individual characteristic random variable.

Theorem 5 ***(Parametric Identification of the Pure Characteristics Model without ξ)*** *Let consider the parametric model defined in definition (2). Under Assumptions 3.(1)-3.(4) and (i) $\xi_j = 0 \ \forall j \in \Im$; (ii) Z is a random vector of individual attributes which*

has a unique maximum at θ^0. A necessary conditions for this to be satisfied is that

$$E\left(\left|\sum_{j=0}^{J} \Pr\left(d_j = 1|z, X, \theta^0; F_\varepsilon^0\right) \log \Pr\left(d_j = 1|z, X, \theta; F_\varepsilon\right)\right|\right) < \infty$$

(Newey and McFadden (1994)). Weaker conditions are pointed out by Van der Vaart (1998).

dimension $M \geq 2$ where at least two attributes have unbounded support; (iii) there are at least two observed product characteristics $K \geq 2$; and (iv) there exist at least two individual attributes $s \neq m$ such that for at least one product characteristic k is satisfied that $\theta^0_{1,km} \neq \theta^0_{1,ks}$, then the true value of the parameters θ^0 (such that $\theta^0_2 \neq 0$) is identified up to a scalar scale with respect to any other $\theta \in \Theta$ (such that $\theta_2 \neq 0$)

Proof. See in Proof's Section ■

The intuition behind the identification proof is the following. The idea is to find those values of the individual attribute Z in (3.18) such that there exists at least one choice probability that differs for different values of the parameters, or equivalently, those values of z such that the equality of choice probabilities for all j imply also the equality of parameters. The difficulty that arises from the choice probabilities derived in section (3.3) is that they depend on two different indices (the upper bound and the lower bound for the unobserved taste variable ε). This is not the case in the standard discrete choice model where the probabilities depend on a unique index. For each product, we consider all its possible positions in the inner product ordering with respect to θ_2 and θ^0_2. We then define those sets of Z such that we can bring either the upper or the lower bound to infinity so that the choice probability of each product only depends on one of them. By inspecting the expression for the upper and lower bound indices in (3.14) and (3.15), it is easy to see that we can fix one of them and bring the value of the other one to infinity if the characteristics of the products on which the upper and lower bound depend are different and there exist at least two unbounded individual attributes. This guarantees that the set of z has a positive probability and the last step is to show that when the choice probabilities depend only on one index, the equality in probabilities implies equality of the parameters of interest as well. This last step is similar to the identification of standard parametric models with the particularity that now the indices consist of a ratio where both the numerator and the denominator depend on the parameters and consequently only identification up to a scalar is possible. Consequently, in order to identify the sign of the coefficients, we would need to impose some assumptions or restrictions on the sign of the parameter we decide to normalize.

It is interesting to note that if only one observed product characteristic is available ($K = 1$), the model only predicts positive probabilities for the first and the last products placed in the ordering in (3.10).

When the characteristics space is unidimensional, if $j \neq (1)^{\theta_2}, j \neq (J)^{\theta_2}$ then the upper and lower bound of the choice probabilities are

$$\bar{\Delta}\bar{\Delta}_j(z_i, x; \theta) = -\frac{\theta_1' z_i}{\theta_2} \text{ and } \underline{\Delta}\underline{\Delta}_j(z_i, x; \theta) = -\frac{\theta_1' z_i}{\theta_2}$$

and consequently, the choice probabilities $\Pr(d_j = 1|X, z; \theta, F_\varepsilon^0) = 0$ for all $j \neq (1)^{\theta_2}, j \neq (J)^{\theta_2}$. The only choice probabilities different from zero are

$$\begin{aligned} \Pr(d_j &= 1|X, z; \theta, F_\varepsilon^0) = F_\varepsilon^0\left(-\frac{\theta_1' z_i}{\theta_2}\right) \text{ if } j = (1)^{\theta_2} \\ \Pr(d_j &= 1|X, z; \theta, F_\varepsilon) = 1 - F_\varepsilon^0\left(-\frac{\theta_1' z_i}{\theta_2}\right) \text{ if } j = (J)^{\theta_2} \end{aligned}$$

These unattractive predictions of the model arise because the unobserved heterogeneity in tastes over a unique product characteristic is not enough to capture the diversity observed in demand in the data. We believe that there may exist a necessary relationship between the number of available choices in the data and the number of observable characteristics in order for the pure characteristic model to be able to predict reasonable choice probabilities.

In the next Theorem we consider the identification of the parameters θ when there are unobserved product characteristics (i.e. $\xi \neq 0$). If vector of individual attributes Z includes also a vector of ones, then there exists a mean utility term for each product that does not vary across individuals. Denote by $\bar{\theta}$ the parameter associated to this mean utility term, then the utility function assumed in (3.2) can be rewritten as in

$$U_{ij} = \bar{\theta} x_j + \theta_i x_j + \xi_j$$

By fitting product specific constants in the utility function one could identify all those elements in the utility that does not change across individuals. Let denote these product specific constants by δ_j which capture both the mean utility of the observed product characteristics plus the unobserved product characteristics

$$\delta_j = \bar{\theta} x_j + \xi_j \tag{3.20}$$

By Assumption 3.(2), the unobserved product characteristics ξ_j and the observed product characteristics x_j are not assumed to be independent. For example, when firms set prices, they consider all the product characteristics that may not the available to the econometrician and that are captured in ξ[14].The econometric issues that arise in this case in

[14] Although other product characteristics different from price are likely to be correlated with ξ, prices have been the variable typical used to illustrate the endogeneity problem in these models.

an aggregate model have been considered in Berry (1994). This simultaneity bias appears to be less important in a microdata model as the one considered here.

However, the fact that there are omitted variables is still a problem even when microdata is used. In this case, consistent estimates of the vector of product specific constants δ are obtained and IV can be used in (3.20) (using a consistent estimate of vector δ) to control for the endogeneity of prices in the estimation of parameter $\bar{\theta}$.[15] In the next theorem, we denote by θ the parameters imbedded in θ_i (i.e. θ_1 and θ_2) so that the product specific constants have separate notation.

Theorem 6 ***(Parametric Identification of the Pure Characteristics Model with*** ξ) *Let consider the parametric model defined in definition (2). Under Assumptions 3.(1)-3.(4) and (i)* $\xi_j \neq 0$ *for some* $j \in \Im$*; (ii)* Z *is a random vector of individual attributes which dimension* $M \geq 2$ *where at least two attributes have unbounded support; (iii) there are at least two observed product characteristics* $K \geq 2$*; and (iv) there exist at least two individual attributes* $s \neq m$ *such that for at least one product characteristic* k *is satisfied that* $\theta^0_{1,km} \neq \theta^0_{1,ks}$*, then the true value of the parameters* θ^0 *(such that* $\theta^0_2 \neq 0$*) is identified up to a scalar scale with respect to any other* $\theta \in \Theta$ *(such that* $\theta_2 \neq 0$*) and the differences of the product specific constants of each product with respect to a base product , say product 0,* $\tilde{\delta}_{j1} = \delta_j - \delta_1$ *are identified.*

Proof. See Proof's Section ■

3.4.2 Identification conditions for the semiparametric model

Theorem 7 (Semiparametric Identification without ξ **)** *Let consider the semiparametric model defined in definition (3). Under Assumptions 3.(1)-3.(4) and (i)* $\xi_j = 0 \ \forall$ $j \in \Im$*; (ii)* Z *is a random vector of individual attributes which dimension* $M \geq 2$ *where at least two attributes have unbounded support; (iii) there are at least two observed product characteristics* $K \geq 2$*; and (iv) there exist at least two individual attributes* $s \neq m$ *such that for at least one product characteristic* k *is satisfied that* $\theta^0_{1,km} \neq \theta^0_{1,ks}$*, and (iv) there exists at least one variable* z_m *with* $\theta^0_{1,km} \neq 0$ *for at least one* k *such that ,conditioned on the other elements of* Z*, the distribution of* z_m *has everywhere positive Lebesgue density.*

[15] It should be noted that the issue of endogenous prices is important to be considered only when an estimate of $\bar{\theta}$ is needed (for example to compute cross and own price and product characteristics elasticities). For those applications in which only an estimate of δ is needed, one can disregard the endogeneity problem.

Then, the true value of the parameters θ_{1m}^0 *for* $m = 2, ..., M$ *and* θ_2^0 *(such that* $\theta_2^0 \neq 0$*) is identified up to a scalar scale with respect to any other* $\theta \in \Theta$ *(such that* $\theta_2 \neq 0$*)*

Proof. See Proof's Section ■

In order to identify the finite dimensional preference parameters, only the median independence assumption of ε and Z is needed. However, if the stronger condition of statistical independence between the unobserved and observed individual attributes is assumed, then also the distribution of F_ε is identified up to scale in addition to the identification of the preference parameters in the utility function θ_{1m}^0, θ_2^0 for $m = 2, .., M$ (See Corollary 5. Proposition 2 in Manski (1988)). The identification up to scale implies that the taste preferences parameters for one of the attributes and for all the characteristics should be normalized to 1.

It should be pointed out that the additional assumption of continuity in at least one individual attribute is required when the distribution of the unobservables is assumed to be unknown. The way the proof evolves requires stronger conditions on the support of Z than in the parametric proof (i.e. a more strict subset of Ω_Z is required to have positive probability in order to be able to identify the parameters of interest). However, given that we are providing sufficient conditions, the same unboundedness condition on Z in the parametric proof is also sufficient in this case.

The access to microdata has allowed us to identify the distribution of the unobserved tastes over product characteristics. As it has been pointed out before in this work, this is important because it is usually difficult to know apriori the appropriateness of any parametric assumption imposed on the distribution of the heterogeneity in tastes. This constitutes the major advantage with respect to the setting where one only has access to aggregate data (as in Berry and Pakes (2003)). In that case, the semiparametric identification of the preferences parameters and the taste distribution is much harder to achieve, and the conditions under which this is possible have not been studied yet.

From the choice probabilities derived before, it can be checked that the dimensionality problem does not change with the number of alternative or products in the market but with the number of characteristics considered. [16]

[16] Tackling the estimation of the identified parameters in θ^0 and F_ε is an interesting and important issue that would be prioritary to incorporate in future versions of this work.

3.5 Conclusions

This work studies the conditions under which the preference parameters of a pure characteristics model with microdata are identified. The model we consider here does not include the iid random term with full support that is usually considered in the utility function in standard approaches. We justify this model in terms of the counterintuitive implications of standard assumption of the product specific unobserved heterogeneity, especially when the discrete choice model is used as a structural model of demand and policy issues related to the change of products in the market are under consideration. The differences in the interpretation of the unobserved tastes over products and characteristics between the standard model and the one considered here are discussed.

For the parametric case, we conclude that the parameters of the utility function are identified up to a scalar constant. The identification requires at least two of individual attributes capturing the observed heterogeneity in tastes with unbounded support. Other identification conditions include full support of the matrix of product characteristics, no identical products should exist on the basis of the observable characteristics and also multicollinearity between the different individual attributes should be ruled out. We also show results for the identification of the alternative specific constant, which include the unobserved product characteristics and the product specific mean utility. The possibility of estimating these product fixed effects allows us to control for the endogeneity problem of prices that arises in these models where unobserved product characteristics are considered.

One of the main advantages of the model considered in this work is that not only allows us to semiparametrically identify the preference parameters but also their estimation would be computationally feasible regardless of the number of products in the market (as apposed to the standard multinomial discrete choice models). For the semiparametric model, the sufficient conditions to identify the preference parameters (after normalizing the parameters for tastes over all observed product characteristics for one individual attribute) and the distribution of the unobserved consumer tastes up to scale need to be strengthen with respect to the parametric case to include at least one continuous individual attribute.

In terms of future work, the estimation of both the preference parameters and the distribution of the unobservables will be studied in both the parametric and the semiparametric model. A large amount of scanner data has become recently available to practitioners. This data includes repeated observations of the household purchases for an extensive list of products along with individual specific information about demographics,

income and consumption habits and detailed information about the characteristics of the products they consume. This constitutes an ideal setting to apply the model considered in this work. Additionally, it would allow us to analyze how repeated choice decisions for the same individual over time helps in the identification of the parameters of interest. Finally, we are also interested in deriving statistical tests able to assess which model of preferences describes better the observed choices in the data. We think that a more general model than the one studied here where the heterogeneity in tastes relies on more than one unobserved product characteristic would be a fruitful extension of the actual model and also less restrictive that the standard models where each product introduces a new dimension of unobserved differentiation.

3.6 Appendix

Appendix A1: Sufficient conditions to construct i.i.d ϵ_{ij} as a product of unobserved product characteristics and individual tastes

Consider the notation introduced in section (3.2.3) and let f_V be the joint density of the vector of unobserved tastes $(f_V(v_{i1}, ..., v_{iJ}))$.

Using the transformation technique we can obtain the joint density of vector E_i from f_V. Thus,

$$\begin{aligned} f_E(\epsilon_{i1}, ..., \epsilon_{iJ}) &= \left|\Sigma^{-1}\right| f_V\left(\Sigma^{-1} \times E_i\right) \\ &= \left|\Sigma^{-1}\right| f_V\left(\sum_{s=1}^{J} m(1,s)\epsilon_{si}, ..., \sum_{s=1}^{J} m(J,s)\epsilon_{si}\right) \end{aligned}$$

where $m(r,s)$ is $(r,s) - th$ element of matrix Σ^{-1}.

Assume (i) that v_{ir} are independent across r and i and identically distributed across i, so that the joint distribution of V_i is

$$f_V(v_{i1}, ..., v_{iJ}) = \prod_{s=1}^{J} \tilde{f}_s(v_{si})$$

and (ii) matrix Σ of unobserved product characteristics is diagonal ($\mu(r,s) \neq 0$ if $r = s$, 0 otherwise), which also implies matrix Σ^{-1} is diagonal ($m(r,s) \neq 0$ if $r = s$, 0 otherwise).

Under assumptions (i) and (ii), then

$$f_E(\epsilon_{i1}, ..., \epsilon_{iJ}) = \left|\Sigma^{-1}\right| \prod_{s=1}^{J} \tilde{f}_s(m(s,s)\epsilon_{si})$$

Therefore, conditions (i) and (ii) are sufficient conditions in order for vector E_i to be independent across products j. Since V_i is independent across i so is E_i.

Appendix A.2: Independence of V_i and diagonal Σ is not sufficient for i.i.d E_i

Consider the case of $J = 2$ where random errors $\{\epsilon_{i1}, \epsilon_{i2}\}$ are i.i.d with joint distribution

$$f_E(\epsilon_{i1}, \epsilon_{i2}) = g(\epsilon_{i1}) \cdot g(\epsilon_{i2})$$

If these errors are expressed as linear combinations of tastes derived from a $J-$dimensional vector of unobserved characteristics as in (3.6), we show that the independence of ϵ_{i1} and ϵ_{i2} does not directly imply independence of vector V_i and a diagonal matrix Σ. [17]Using the transformation technique we have the following equation for density functions of E_i and V_i

$$g(\epsilon_{i1}) \cdot g(\epsilon_{i2}) = \left|\Sigma^{-1}\right| f_V\left(m(1,1)\epsilon_{1i} + m(1,2)\epsilon_{2i}, m(2,1)\epsilon_{1i} + m(2,2)\epsilon_{2i}\right)$$

Using notation $\nu_{i1} = m(1,1)\epsilon_{1i} + m(1,2)\epsilon_{2i}$, $\nu_{i2} = m(2,1)\epsilon_{1i} + m(2,2)\epsilon_{2i}$ we obtain the following system of differential equations

$$\frac{g'(\epsilon_{i2})}{g(\epsilon_{i2})} f_V(\nu_{i1}, \nu_{i2}) - m(1,2)\frac{\partial f_V(\nu_{i1}, \nu_{i2})}{\partial \nu_{i1}} - m(2,2)\frac{\partial f_V(\nu_{i1}, \nu_{i2})}{\partial \nu_{i2}} = 0$$
$$\frac{g'(\epsilon_{i1})}{g(\epsilon_{i1})} f_V(\nu_{i1}, \nu_{i2}) - m(1,1)\frac{\partial f_V(\nu_{i1}, \nu_{i2})}{\partial \nu_{i1}} - m(2,1)\frac{\partial f_V(\nu_{i1}, \nu_{i2})}{\partial \nu_{i2}} = 0$$

Using both equations we obtain the following differential equation for ν_{i2}

$$\left[\frac{g'(\epsilon_{i2})}{g(\epsilon_{i2})} - \frac{m(1,2)}{m(1,1)}\frac{g'(\epsilon_{i1})}{g(\epsilon_{i1})}\right] f_V(\nu_{i1}, \nu_{i2}) -$$
$$\left[\frac{m(2,1)}{m(1,1)} + m(2,2)\right]\frac{\partial f_V(\nu_{i1}, \nu_{i2})}{\partial \nu_{i2}} = 0$$

Solving for the joint distribution of $f_V(\nu_{i1}, \nu_{i2})$, it can be checked that errors (ν_{i1}, ν_{i2}) do not need to be independent or $\mu(2,1) = 0$ or $\mu(1,2) = 0$ to generate independent $(\epsilon_{i1}, \epsilon_{i2})$

$$f_V(\nu_{i1}, \nu_{i2}) =$$
$$g(\mu(2,1)\nu_{i1} + \mu(2,2)\nu_{i2})^{\frac{1}{\mu(2,2)\left(\frac{m(2,1)}{m(1,1)} + m(2,2)\right)}} g(\mu(1,1)\nu_{i1} + \mu(1,2)\nu_{i2})^{\frac{-m(1,2)}{m(1,1)\mu(1,2)\left(\frac{m(2,1)}{m(1,1)} + m(2,2)\right)}}$$

[17]The notation used is $\Sigma = \begin{pmatrix} \mu(1,1) & \mu(1,2) \\ \mu(2,1) & \mu(2,2) \end{pmatrix}$

3.7 Proofs

Proof of Theorem (5) (Parametric Identification of the Pure Characteristics Model without ξ). For the case in which there does not exist unobserved product characteristics we need to redefine the indices at which the choice probability evaluate the distribution function F_ε. Let these indices without unobserved product characteristics be denoted by $\bar{\Delta}\bar{\Delta}_j(z_i, x; \theta)$ and $\underline{\Delta\Delta}_j(z_i, x; \theta)$ whose expression is exactly the same as for the homologous $\bar{\Delta}_j$ and $\underline{\Delta}_j$ but with $\xi_j = 0 \; \forall j \in \Im$.

Let $\otimes$ be the Kronecker product and let $W_{js_i(\theta)} = (x_j - x_{s_{ij}(\theta)}) \otimes z_i$ and $\Delta x_{js_i(\theta)} = (x_{s_{ij}(\theta)} - x_j)$ for $s = \{r, R\}$. Thus, vector $W_{jr_i(\theta)}$ contains all the interactions between the M individual attributes in Z for individual i and the differences between the K product characteristics of product j and the characteristics of that product that minimizes the upper index of the choice probability for j for the parameter value θ. The first M rows are interactions of the first product characteristic with all the individual attributes. Therefore, the indices evaluated at the true value of the parameters θ_0 are

$$\begin{aligned}
\bar{\Delta}\bar{\Delta}_j(z_i, x; \theta^0) &= \frac{\left(\theta_1^0 \times z_i\right)'(x_j - x_{r_{ij}(\theta^0)})}{\theta_2^{0\prime}(x_{jr_i(\theta^0)} - x_j)} = \frac{\tilde{\theta}_1^{0\prime} W_{jr_i(\theta^0)}}{\theta_2^{0\prime} \Delta x_{jr_i(\theta^0)}} \\
\underline{\Delta\Delta}_j(z_i, x; \theta^0) &= \frac{\left(\theta_1^0 \times z_i\right)'(x_j - x_{R_{ij}(\theta^0)})}{\theta_2^{0\prime}(x_{jR_i(\theta^0)} - x_j)} = \frac{\tilde{\theta}_1^{0\prime} W_{jR_i(\theta^0)}}{\theta_2^{0\prime} \Delta x_{jR_i(\theta^0)}}
\end{aligned}$$

where $\tilde{\theta}_1^0 = vec\left[\theta_1^0\right]$ and $\tilde{\theta}_2^0 = vec\left[\theta_2^0\right] = \theta_2^0$.[18]

Suppose the following equality of probabilities hold for $\theta \neq \theta^0$

$$\Pr(d_j = 1 | X, z; \theta^0) = \Pr(d_j = 1 | X, z; \theta) \text{ for all } j \in \Im \text{ a.e in } Z$$

Then, with probability 1 in Ω_Z and for every $j \in \Im$ except for $j = (1)^{\theta_2}, j = (J)^{\theta_2}, j = (1)^{\theta_2^0}$ and $j = (J)^{\theta_2^0}$ the above identification condition implies (using (3.16))

$$F_\varepsilon(t_{ji1}) - F_\varepsilon(t_{ji2}) = F_\varepsilon(t_{ji1} + s_{ji1}) - F_\varepsilon\left(t_{ji2} + s_{ji2}\right) \tag{3.21}$$

where,

$$\begin{aligned}
t_{ji1} &= \bar{\Delta}\bar{\Delta}_j(z_i, x; \theta^0) \qquad t_{ji2} = \underline{\Delta\Delta}_j(z_i, x; \theta^0) \\
s_{ji1} &= \bar{\Delta}\bar{\Delta}_j(z_i, x; \theta) - \bar{\Delta}\bar{\Delta}_j(z_i, x; \theta^0) \qquad s_{ji2} = \underline{\Delta\Delta}_j(z_i, x; \theta) - \underline{\Delta\Delta}_j(z_i, x; \theta^0)
\end{aligned}$$

[18] Define $vec(Y)$ as that operation which appends all the trasposed rows of a matrix Y of dimension $K_1 \times K_2$ in a column vector of dimension $K_1 K_2$. Note that since θ_2 is already a column vector $vec(\theta_2) = \theta_2$.

When j is one of extremes of the ordering with respect to θ_2 or θ_2^0, the following choice probabilities arise

$$F_\varepsilon(t_{ji1}) - F_\varepsilon(t_{ji2}) = F_\varepsilon(t_{ji1} + s_{ji1}) \text{ if } \left\{ j = (1)^{\theta_2}, j \neq (1)^{\theta_2^0}, j \neq (J)^{\theta_2^0} \right\} \tag{3.22}$$

$$F_\varepsilon(t_{ji1}) = F_\varepsilon(t_{ji1} + s_{ji1}) - F_\varepsilon(t_{ji2} + s_{ji2}) \text{ if } \left\{ j \neq (1)^{\theta_2}, j \neq (J)^{\theta_2}, j = (1)^{\theta_2^0} \right\} \tag{3.23}$$

$$F_\varepsilon(t_{ji1}) - F_\varepsilon(t_{ji2}) = 1 - F_\varepsilon(t_{ji2} + s_{ji2}) \text{ if } \left\{ j = (J)^{\theta_2}, j \neq (1)^{\theta_2^0}, j \neq (J)^{\theta_2^0} \right\} \tag{3.24}$$

$$1 - F_\varepsilon(t_{ji2}) = F_\varepsilon(t_{ji1} + s_{ji1}) - F_\varepsilon(t_{ji2} + s_{ji2}) \text{ if } \{ j \neq (J)^{\theta_2}, j \neq (J)^{\theta_2}, j = (J)^{\theta_2^0} \} \tag{3.25}$$

$$1 - F_\varepsilon(t_{ji2}) = 1 - F_\varepsilon(t_{ji2} + s_{ji2}) \text{ if } \{ j = (J)^{\theta_2}, j = (J)^{\theta_2^0} \} \tag{3.26}$$

$$F_\varepsilon(t_{ji1}) = F_\varepsilon(t_{ji1} + s_{ji1}) \text{ if } \{ j = (1)^{\theta_2}, j = (1)^{\theta_2^0} \} \tag{3.27}$$

$$1 - F_\varepsilon(t_{ji2}) = F_\varepsilon(t_{ji1} + s_{ji1}) \text{ if } \{ j = (1)^{\theta_2}, j = (J)^{\theta_2^0} \} \tag{3.28}$$

$$F_\varepsilon(t_{ji1}) = 1 - F_\varepsilon(t_{ji2} + s_{ji2}) \text{ if } \{ j = (J)^{\theta_2}, j = (1)^{\theta_2^0} \} \tag{3.29}$$

In cases (3.27) and (3.26) -where the product j is either the first or the last one with respect to the inner product ordering under both θ_2 and θ_2^0– the identification is easier. This is because expressions (3.27) and (3.26) imply $s_{ji1} = 0$ or $s_{ji2} = 0$, which as it is shown below implies identification of θ^0 with respect to θ. A special note deserve those values of the parameter that make reverse the order with respect to θ_2, as in (3.28) and (3.29).

For the identification of θ^0 with respect to parameters θ such that the ordering with respect to θ_2 is as in (3.21) -(3.25), let consider indices t_{j1} and t_{j2} as an inner product of the vector of individual attributes z_i as follows

$$t_{ji1} = \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{\tilde{\theta}_{1,km}^0 (x_{jk} - x_{r_{ij}(\theta_0)k})}{\tilde{\theta}_2^{0\prime} (x_{r_{ij}(\theta_0)} - x_j)} z_{im}$$

$$t_{ji2} = \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{\tilde{\theta}_{1,km}^0 (x_{jk} - x_{R_{ij}(\theta_0)k})}{\tilde{\theta}_2^{0\prime} (x_{R_{ij}(\theta_0)} - x_j)} z_{im}$$

By definition, $r_{ij(\theta^0)}$ and $R_{ij(\theta^0)}$ are two different products and by Assumption 3. (3) they differ at least in one characteristic. Therefore, both indices t_{ji1} and t_{ji2} vary in a different way with respect to vector z_i.

Therefore, as long as the dimension of the individual attributes vector is greater than 2 ($M \geq 2$) by condition (ii) of this theorem, the value of the index t_{ij1} can be held fixed

while the value of the index t_{ij2} changes. For example, for continuous Z_1 and Z_2 with $M = 2$, the slopes of isoquant curves for t_{ij1} and t_{ij2} in a two dimensional space are

$$\left.\frac{\partial z_{2i}}{\partial z_{1i}}\right|_{\overline{t_{ij1}}} = \frac{\sum_{k=1}^{K} \tilde{\theta}^0_{1,k2}\left(x_{jk} - x_{r_{ij}(\theta_0)k}\right)}{\sum_{k=1}^{K} \tilde{\theta}^0_{1,k1}\left(x_{jk} - x_{r_{ij}(\theta_0)k}\right)}$$

$$\left.\frac{\partial z_{2i}}{\partial z_{1i}}\right|_{\overline{t_{ij2}}} = \frac{\sum_{k=1}^{K} \tilde{\theta}^0_{1,k2}\left(x_{jk} - x_{R_{ij}(\theta_0)k}\right)}{\sum_{k=1}^{K} \tilde{\theta}^0_{1,k1}\left(x_{jk} - x_{R_{ij}(\theta_0)k}\right)}$$

If $\tilde{\theta}^0_{1,k2} \neq \tilde{\theta}^0_{1,k1}$ for some k as condition (iv) in this theorem requires (i.e. there is at least one product characteristic for which the individual tastes associated to Z_1 and Z_2 are different), then both slopes above differ from each other.

This example also illustrates the need of having more than one product characteristics, since otherwise both indices t_{ij1} and t_{ij2} would be equal.

Given that equations (3.21) to (3.25) hold a.e in Z, the key of the identification proof is to find a set of values $z \subset \Omega_Z$ with positive probability for which one can conclude the equality of the two vectors of parameters ($\theta = \theta^0$). This set is found by keeping either t_{ij1} (or t_{ij2}) fixed and driving t_{ij2} (or t_{ij1}) to infinity.

Let define

$$\begin{aligned} \tau_{1j}(\theta^0, \theta, X) &= \{z_i \in \Omega_Z \text{ such that } t_{ji1} = t_{1j}\} \\ \tau_{2j}(\theta^0, \theta, X) &= \{z_i \in \Omega_Z \text{ such that } t_{ji2} = t_{2j}\} \end{aligned}$$

Next, for each j, we define the subsets of τ_{2j} or τ_{1j} - denoted by ρ_j- which make the equality of choice probabilities at θ and θ^0 in (3.21)-(3.29) as a function of a single index (either t_{j1} or t_{j2}).

If $j \neq (1)^{\theta_2}, j \neq (J)^{\theta_2}, j \neq (1)^{\theta_2^0}$ and $j \neq (J)^{\theta_2^0}$, this set is defined as follows

$$\rho_j(\theta^0, \theta, X) = \{z_i \in \tau_{1j} \subset \Omega_Z \text{ such that } t_{ji2} \to \infty\}$$

The purpose of building this set is that for those values of Z, the choice probabilities in (3.21) can be simplified as

$$F_\varepsilon(t_{ji1}) = F_\varepsilon\left(t_{ji1} + s_{ji1}\right) \text{ if } z_i \in \rho_j(\theta^0, \theta, X)$$

Given that the cdf F_ε is assumed to be strictly increasing, the above equation implies $s_{ji1} = 0$.

For equations (3.21) to (3.23) index t_{ij2} needs to be driven to $+\infty$ or $-\infty$ in order to write the choice probabilities as a function of a unique index. For equations (3.24) to (3.25) we need to drive index t_{ij1} to $+\infty$. Equations (3.26) to (3.29) have immediate consequences over the parameters without restricting the support of Z.

Thus, when j is one of extremes of the ordering with respect to θ or θ^0, the set $\rho_j(\theta^0, \theta, X)$ is defined as follows [19]

$$\rho_j(\theta^0,\theta,X) = \{z_i \in \tau_{1j} \subset \Omega_Z \text{ such that } t_{ji2} \to -\infty\} \text{ if } \{j = (1)^{\theta_2}\} \text{ or } \{j = (1)^{\theta_2^0}\}$$
$$\rho_j(\theta^0,\theta,X) = \{z_i \in \Omega_Z\} \text{ if } \{j = (1)^{\theta_2}, j = (1)^{\theta_2^0}\}$$
$$\rho_j(\theta^0,\theta,X) = \{z_i \in \tau_{2j} \subset \Omega_Z \text{ such that } t_{ji1} \to \infty\} \text{ if } \{j = (J)^{\theta_2}, j \neq (J)^{\theta_2^0}\} \text{ or } \{j \neq (J)^{\theta_2}, j = (J)^{\theta_2^0}\}$$
$$\rho_j(\theta^0,\theta,X) = \{z_i \in \Omega_Z\} \text{ if } \{j = (J)^{\theta_2}, j = (J)^{\theta_2^0}\}$$

We separate the cases for which t_{ij1} has to go to $+\infty/-\infty$ (products belonging to the set $\Im_1$) from the cases in which t_{ij2} goes to $+\infty/-\infty$ (products belonging to the set $\Im_2$) as follows $\forall$

$$\Im_1(\theta,\theta^0) = \left\{ \begin{array}{c} j \text{ such that } \left\{j = (J)^{\theta_2}, j \neq (1)^{\theta_2^0}, j \neq (J)^{\theta_2^0}\right\} \\ \text{or } \{j \neq (J)^{\theta_2}, j \neq (J)^{\theta_2}, j = (J)^{\theta_2^0}\} \\ \text{or } \{j \neq (J)^{\theta_2}, j \neq (1)^{\theta_2}, j \neq (J)^{\theta_2^0}, j \neq (1)^{\theta_2^0}\} \end{array} \right\}$$

$$\Im_2(\theta,\theta^0) = \left\{ \begin{array}{c} j \text{ such that } \left\{j = (1)^{\theta_2}, j \neq (1)^{\theta_2^0}, j \neq (J)^{\theta_2^0}\right\} \\ \text{or } \left\{j = (1)^{\theta_2^0}, j \neq (1)^{\theta_2}, j \neq (J)^{\theta_2}\right\} \end{array} \right\}$$

Let define

$$P(\theta^0,\theta,X) = \cup_{j=1}^{J} \rho_j(\theta^0,\theta,X)$$

as the set of Ω_Z such that at least one of the product probabilities is written only as a function of t_{ij1} or t_{ij2}.

Then, for each value $z \in P(\theta^0, \theta, X)$, there exist at least one product j for which the probabilities in (3.21)-(3.27) become

$$F_\varepsilon(t_{2j}) = F_\varepsilon(t_{2j} + s_{ji2}) \text{ for } z \in \rho_j(\theta^0,\theta,X) \text{ where } j \in \Im_1(\theta,\theta^0) \quad (3.30)$$
$$F_\varepsilon(t_{1j}) = F_\varepsilon(t_{1j} + s_{ji1}) \text{ for } z \in \rho_j(\theta^0,\theta,X) \text{ where } j \in \Im_2(\theta,\theta^0) \quad (3.31)$$

[19] Note then that for the cases where $\{j = (1)^{\theta_2}, j = (1)^{\theta_2^0}\}$ and $\{j = (J)^{\theta_2}, j = (J)^{\theta_2^0}\}$, the set $\rho_j(\theta, \theta^0, X)$ does not restrict the sample space of random vector Z since the probabilities are already a function of a unique index.

Note that if the sets $\rho_j(\theta^0, \theta, X)$ are disjoint across j, then for each $z \in P(\theta^0, \theta, X)$ there exists only one possible product j for which its probability is written as a function of a unique index t_{ij}. However, in general each element z in $P(\theta^0, \theta, X)$ may belong to one or more sets $\rho_j(\theta^0, \theta, X)$.

If by condition (ii) in this theorem, there exists at least two variables in Z with large enough support on the entire real line, then we can ensure that

$$\Pr\left\{z \in \Omega_Z \cap P(\theta^0, \theta, X)\right\} > 0, \ \forall \theta \in \Theta, \theta \neq \theta^0 \text{ and } \theta_2 \neq 0$$

since either t_{ij1} or t_{ij2} can be brought to large values by increasing or decreasing the values of these particular attributes. This assumption though allows to have discrete individual attributes. The rest of the individual attributes are allowed to be discrete.

Therefore, for each z belonging to set $P(\theta^0, \theta, X)$, there exists at least one product j that depends only either on t_{j2} or t_{j1}.

Let define for each $z \in \rho_j$ that set of products for which the choice probabilities can be expressed as a function of a unique index t_{ij}

$$\begin{aligned} Q_{1i} &\equiv Q_1(z_i, \theta^0, \theta, X) = \left\{j \in \Im_1(\theta, \theta^0) \text{ such that } z_i \in \rho_j(\theta^0, \theta, X)\right\} \\ Q_{2i} &\equiv Q_2(z_i, \theta^0, \theta, X) = \left\{j \in \Im_2(\theta, \theta^0) \text{ such that } z_i \in \rho_j(\theta^0, \theta, X)\right\} \end{aligned}$$

Thus, if j belongs to Q_{1i} the choice probabilities of product j for individual i are only a function of t_{ij2} and s_{ij2} (and not of t_{ij1}); and viceversa for those products belonging to Q_{2i}.

Consequently, define those subsets of $P(\theta^0, \theta, X) \subset \Omega_Z$ implying either $t_{ij2} \to +\infty/-\infty$ or $t_{ij1} \to +\infty$

$$\begin{aligned} P_1(\theta^0, \theta, X) &= \cup_{j \in \Im_1(\theta, \theta^0)} \rho_j(\theta^0, \theta, X) \\ P_2(\theta^0, \theta, X) &= \cup_{j \in \Im_2(\theta, \theta^0)} \rho_j(\theta^0, \theta, X) \end{aligned}$$

[20]

By Assumption 3. (1) and from expression (3.30) we can conclude that

$$\begin{aligned} s_{ij2} &= 0 \ , \forall z_i \in P_1(\theta^0, \theta, X), \forall \ j \in Q_{1i} \\ s_{ij1} &= 0 \ , \forall z_i \in P_2(\theta^0, \theta, X), \forall \ j \in Q_{2i} \end{aligned}$$

[20] Note that $P(\theta^0, \theta, X) = P_1(\theta^0, \theta, X) \cup P_2(\theta^0, \theta, X)$

Using the notation introduced above for s_{ij1} and s_{ij2},

$$\begin{aligned} \underline{\Delta\Delta}_j(z_i, x; \theta) &= \underline{\Delta\Delta}_j(z_i, x; \theta^0) \ , \forall z_i \in P_1(\theta^0, \theta, X), \forall\ j \in Q_{1i} \\ \bar{\Delta}\bar{\Delta}_j(z_i, x; \theta) &= \bar{\Delta}\bar{\Delta}_j(z_i, x; \theta^0) \ , \forall z_i \in P_2(\theta^0, \theta, X), \forall\ j \in Q_{2i} \end{aligned} \tag{3.32}$$

This implies

$$\begin{aligned} \frac{\tilde{\theta}_1' W_{jR_i(\theta)}}{\theta_2' \Delta x_{jR_i(\theta)}} &= \frac{\tilde{\theta}_1^{0\prime} W_{jR_i(\theta^0)}}{\theta_2^{0\prime} \Delta x_{jR_i(\theta^0)}}, \forall z_i \in P_1(\theta^0, \theta, X), \forall\ j \in Q_{1i} \\ \frac{\tilde{\theta}_1' W_{jr_i(\theta)}}{\theta_2' \Delta x_{jr_i(\theta)}} &= \frac{\tilde{\theta}_1^{0\prime} W_{jr_i(\theta^0)}}{\theta_2^{0\prime} \Delta x_{jr_i(\theta^0)}} \ , \forall z_i \in P_2(\theta^0, \theta, X), \forall\ j \in Q_{2i} \end{aligned}$$

or equivalently,

$$\left(\tilde{\theta}_1' W_{jR_i(\theta)}\right)\left(\theta_2^{0\prime} \Delta x_{jR_i(\theta^0)}\right) - \left(\tilde{\theta}_1^{0\prime} W_{jR_i(\theta^0)}\right)\left(\theta_2' \Delta x_{jR_i(\theta)}\right) = 0 \tag{3.33}$$

$$, \forall z_i \in P_1(\theta^0, \theta, X), \forall\ j \in Q_{1i} \tag{3.34}$$

$$\left(\tilde{\theta}_1' W_{jr_i(\theta)}\right)\left(\theta_2^{0\prime} \Delta x_{jr_i(\theta^0)}\right) - \left(\tilde{\theta}_1^{0\prime} W_{jr_i(\theta^0)}\right)\left(\theta_2' \Delta x_{jr_i(\theta)}\right) = 0 \tag{3.35}$$

$$, \forall z_i \in P_2(\theta^0, \theta, X), \forall\ j \in Q_{2i} \tag{3.36}$$

Using the above notation for each of these vectors, expressions (3.33) and (3.35) above are equivalent respectively to

$$\begin{aligned} \sum_{m=1}^{M} \left[\begin{array}{c} \sum_{k=1}^{K} \left(\tilde{\theta}_{1km}^0 \theta_{2k} - \tilde{\theta}_{1km} \theta_{2k}^0\right)\left(x_{jk} - x_{R_i(\theta^0)k}\right) z_{im} \left(x_{R_i(\theta)k} - x_{jk}\right) + \\ + \sum_{k=1}^{K} \sum_{k<s} \left(\tilde{\theta}_{1km}^0 \theta_{2s} - \tilde{\theta}_{1sm} \theta_{2k}^0\right)\left(x_{jk} - x_{R_i(\theta^0)k}\right) z_{im} \left(x_{R_i(\theta)s} - x_{js}\right) \end{array} \right] &= 0 \\ \sum_{m=1}^{M} \left[\begin{array}{c} \sum_{k=1}^{K} \left(\tilde{\theta}_{1km}^0 \theta_{2k} - \tilde{\theta}_{1km} \theta_{2k}^0\right)\left(x_{jk} - x_{r_i(\theta^0)k}\right) z_{im} \left(x_{r_i(\theta)k} - x_{jk}\right) + \\ + \sum_{k=1}^{K} \sum_{k<s} \left(\tilde{\theta}_{1km}^0 \theta_{2s} - \tilde{\theta}_{1sm} \theta_{2k}^0\right)\left(x_{jk} - x_{r_i(\theta^0)k}\right) z_{im} \left(x_{r_i(\theta)s} - x_{js}\right) \end{array} \right] &= 0 \end{aligned}$$

Let express equations (3.33) and (3.35) in a matricial way. Let n_1 be the number of values of Z belonging to $P_1(\theta^0, \theta, X)$ and, analogously, n_2 be the number of values of Z belonging to $P_2(\theta^0, \theta, X)$.

Let consider the matricial expression for this system of equations

$$\begin{bmatrix} W_1 & V_1 \\ W_2 & V_2 \end{bmatrix} \begin{bmatrix} B_W \\ B_V \end{bmatrix} = 0 \tag{3.37}$$

where W_t is a matrix of dimension $(n_t(\sum_{i=1}^{n_t} Q_{it}) \times KM)$ for $t = \{1, 2\}$
and V_t is a matrix of dimension $\left(n_t(\sum_{i=1}^{n_t} Q_{it}) \times \frac{K(K-1)M}{2}\right)$ for $t = \{1, 2\}$.

A representative row of matrices W_1 and W_2 for an individual i is expressed as

$$w_{1i} = \left\{\left\{\left(x_{jk} - x_{R_i(\theta^0)k}\right) z_{im} \left(x_{R_i(\theta)k} - x_{jk}\right)\right\}_{k=1}^{K}\right\}_{m=1}^{M}, \; z_i \in P_1(\theta^0, \theta, X), \; j \in Q_{1i}$$

$$w_{2i} = \left\{\left\{\left(x_{jk} - x_{r_i(\theta^0)k}\right) z_{im} \left(x_{r_i(\theta)k} - x_{jk}\right)\right\}_{k=1}^{K}\right\}_{m=1}^{M}, \; z_i \in P_2(\theta^0, \theta, X), \; j \in Q_{2i}$$

A representative row of matrix V_1 and V_2 of an individual i is expressed as

$$v_{1i} = \left\{\left\{\left\{\left(x_{jk} - x_{R_i(\theta^0)k}\right) z_{im} \left(x_{R_i(\theta)s} - x_{js}\right)\right\}_{k=1}^{K}\right\}_{k<s}\right\}_{m=1}^{M}, \; z_i \in P_1(\theta^0, \theta, X), \; j \in Q_{1i}$$

$$v_{2i} = \left\{\left\{\left\{\left(x_{jk} - x_{r_i(\theta^0)k}\right) z_{im} \left(x_{r_i(\theta)s} - x_{js}\right)\right\}_{k=1}^{K}\right\}_{k<s}\right\}_{m=1}^{M}, \; z_i \in P_2(\theta^0, \theta, X), \; j \in Q_{2i}$$

B_W is a column vector of dimension MK and B_V is a column vector of dimension $\frac{(K-1)KM}{2}$ whose elements are expressed as

$$B_W = \left\{\left\{\tilde{\theta}^0_{1km}\theta_{2k} - \tilde{\theta}_{1km}\theta^0_{2k}\right\}_{k=1}^{K}\right\}_{m=1}^{M}$$

$$B_V = \left\{\left\{\left\{\tilde{\theta}^0_{1km}\theta_{2s} - \tilde{\theta}_{1sm}\theta^0_{2k}\right\}_{k=1}^{K}\right\}_{k<s}\right\}_{m=1}^{M}$$

The only way one can obtain multicollinearity across and between the columns of W and V is when matrix of product characteristics X does not have full rank and when there exists a proper linear subspace of the individual attributes Z. Since this is ruled out by Assumption 3. (4), then matrix $[WV]'[WV]$ is full rank in the limit, so that it can be concluded that $B_W = 0$ and $B_V = 0$.

Thus,

$$\tilde{\theta}^0_{1km}\theta_{2k} = \tilde{\theta}_{1km}\theta^0_{2k} \text{ for every } k = 1, .., K; m = 1, .., M$$

$$\tilde{\theta}^0_{1km}\theta_{2s} = \tilde{\theta}_{1km}\theta^0_{2s} \text{ for every } k = 1, .., K; k < s; m = 1, .., M$$

This system of equations does not have a unique solution. In fact, it has infinite ways of expressing θ as a function of θ^0 [21] Therefore, since not all the elements of the vector of parameters θ are identified, we need to impose some normalization and define

[21] The system of equations $[B_W \; B_V] = 0$ can be expressed as

$$C(\theta^0) \times \theta = 0$$

in order to solve for θ as a function of θ^0. Even for $K > 2$, matrix $C(\theta^0)$ has not full rank.

the parameters that can be identified. For simplicity, we normalized parameter $\theta_{21} = 1$ so that all the parameters are identified up to this scalar scale.

From the equations corresponding to $k = 1$

$$\tilde{\theta}_{11m} = \frac{\tilde{\theta}^0_{11m}}{\theta^0_{21}} \text{ for every } m$$

The parameters of $\tilde{\theta}_1$ corresponding to characteristic $k = 1$ (normalized characteristic) and for all the individual attributes are identified as the ratio of the true parameter and the normalized parameter.

From the equations corresponding to $s = 1$, we are able to identify the parameters $\tilde{\theta}_1$ associated to all the interactions between the K product characteristics and M individual attributes.

$$\tilde{\theta}_{1km} = \frac{\tilde{\theta}^0_{1km}}{\theta^0_{21}} \text{ for every } k, m \tag{3.38}$$

From equations corresponding to $s \neq 1$, we obtain

$$\tilde{\theta}_{1km} = \tilde{\theta}^0_{1km} \frac{\theta_{2s}}{\theta^0_{2s}} \text{ for every } k, m \tag{3.39}$$

From expressions (3.38) and (3.39), we obtain the identification of the unobserved taste parameters,

$$\theta_{2s} = \frac{\theta^0_{2s}}{\theta^0_{21}} \text{ for every } s \neq 1$$

If $\tilde{\theta}^0_{1km} = 0$ for some k and m, given that $\theta^0_2 \neq 0$, also the value of $\tilde{\theta}_{1km}$ is zero. QED.

Proof of Theorem (6) (Parametric Identification of the Pure Characteristics Model with ξ).

The upper and lower indices are now defined as

$$\begin{aligned}
\bar{\Delta}_j(z_i, x; \theta^0, \delta^0) &= \frac{(\delta^0_j - \delta^0_{r_{ij}(\theta^0)}) + (\theta^0_1 \times z_i)'(x_j - x_{r_{ij}(\theta^0)})}{\theta^{0\prime}_2 (x_{jr_i(\theta^0)} - x_j)} \\
\underline{\Delta}_j(z_i, x; \theta^0, \delta^0) &= \frac{(\delta^0_j - \delta^0_{R_{ij}(\theta^0)}) + (\theta^0_1 \times z_i)'(x_j - x_{R_{ij}(\theta^0)})}{\theta^{0\prime}_2 (x_{jR_i(\theta^0)} - x_j)}
\end{aligned}$$

The proof mimics the one without product fixed effects up to expressions (3.33) and (3.35) with some changes in the notation to make the upper and lower boundary products also

dependent on the fixed effects (i.e. $r_{ij(\theta^0,\delta^0)}$ and $R_{ij(\theta^0,\delta^0)}$). These two expressions become now

$$\begin{aligned}&\left(\tilde{\theta}_1' W_{jR_i(\theta,\delta)} + \left(\delta_j - \delta_{R_{ij}(\theta,\delta)}\right)\right)\left(\theta_2^{0\prime}\Delta x_{jR_i(\theta^0,\delta^0)}\right) - \\ &\left(\tilde{\theta}_1^{0\prime} W_{jR_i(\theta^0,\delta^0)} + \left(\delta_j^0 - \delta^0_{R_{ij}(\theta^0,\delta^0)}\right)\right)\left(\theta_2'\Delta x_{jR_i(\theta,\delta)}\right) = 0 \qquad (3.40)\\ &,\forall z_i \in P_1(\theta^0,\delta^0,\theta,\delta,X),\forall\ j \in Q_{1i}\end{aligned}$$

$$\begin{aligned}&\left(\tilde{\theta}_1' W_{jr_i(\theta,\delta)} + \left(\delta_j - \delta_{r_{ij}(\theta,\delta)}\right)\right)\left(\theta_2^{0\prime}\Delta x_{jr_i(\theta^0,\delta^0)}\right) - \qquad (3.41)\\ &\left(\tilde{\theta}_1^{0\prime} W_{jr_i(\theta^0,\delta^0)} + \left(\delta_j^0 - \delta^0_{r_{ij}(\theta^0,\delta^0)}\right)\right)\left(\theta_2'\Delta x_{jr_i(\theta,\delta)}\right) = 0\\ &,\forall z_i \in P_2(\theta^0,\delta^0,\theta,\delta,X),\forall\ j \in Q_{2i}\end{aligned}$$

Using the above notation for each of these vectors, expressions (3.40) and (3.41) above are equivalent respectively to

$$\left[\begin{array}{c}\sum_{m=1}^{M}\left[\begin{array}{c}\sum_{k=1}^{K}\left(\tilde{\theta}^0_{1km}\theta_{2k} - \tilde{\theta}_{1km}\theta^0_{2k}\right)\left(x_{jk} - x_{T_i(\theta^0,\delta^0)k}\right) z_{im}\left(x_{T_i(\theta,\delta)k} - x_{jk}\right) + \\ +\sum_{k=1}^{K}\sum_{k<s}\left(\tilde{\theta}^0_{1km}\theta_{2s} - \tilde{\theta}_{1sm}\theta^0_{2k}\right)\left(x_{jk} - x_{T_i(\theta^0,\delta^0)k}\right) z_{im}\left(x_{T_i(\theta,\delta)s} - x_{js}\right)\end{array}\right] \\ + \left(\delta_j - \delta_{T_{ij}(\theta,\delta)}\right)\sum_{k=1}^{K}\theta^0_{2k}\Delta x_{jT_i(\theta^0,\delta^0),k} - \left(\delta_j^0 - \delta^0_{T_{ij}(\theta^0,\delta^0)}\right)\sum_{k=1}^{K}\theta_{2k}\Delta x_{jT_i(\theta,\delta),k}\end{array}\right] = 0$$

for $T = \{R, r\}$

After adding and substracting $\left(\delta_j - \delta_{T_{ij}(\theta^0,\delta^0)}\right)\sum_{k=1}^{K}\theta^0_{2k}\Delta x_{jT_i(\theta,\delta),k}$ for $T = \{R, r\}$, we can express the condition above in a way that can help us in the identification,

$$\left[\begin{array}{c}\sum_{m=1}^{M}\left[\begin{array}{c}\sum_{k=1}^{K}\left(\tilde{\theta}^0_{1km}\theta_{2k} - \tilde{\theta}_{1km}\theta^0_{2k}\right)\left(x_{jk} - x_{T_i(\theta^0,\delta^0)k}\right) z_{im}\left(x_{T_i(\theta,\delta)k} - x_{jk}\right) + \\ +\sum_{k=1}^{K}\sum_{k<s}\left(\tilde{\theta}^0_{1km}\theta_{2s} - \tilde{\theta}_{1sm}\theta^0_{2k}\right)\left(x_{jk} - x_{T_i(\theta^0,\delta^0)k}\right) z_{im}\left(x_{T_i(\theta,\delta)s} - x_{js}\right)\end{array}\right] \\ + \left[\left(\delta_j - \delta_{T_{ij}(\theta,\delta)}\right)\sum_{k=1}^{K}\theta^0_{2k}\Delta x_{jT_i(\theta^0,\delta^0),k} - \left(\delta_j - \delta_{T_{ij}(\theta^0,\delta^0)}\right)\sum_{k=1}^{K}\theta^0_{2k}\Delta x_{jT_i(\theta,\delta),k}\right] \\ + \left[\left(\delta_j - \delta_{T_{ij}(\theta^0,\delta^0)}\right)\theta^0_{2k} - \left(\delta_j^0 - \delta^0_{T_{ij}(\theta^0,\delta^0)}\right)\theta_{2k}\right]\Delta x_{jT_i(\theta,\delta),k}\end{array}\right] = 0 \qquad (3.42)$$

for $T = \{R, r\}$

Let express equations (3.42) in a matricial way. Let define n_1 and n_2 as in Theorem (5).

Let consider the matricial expression for this system of equations

$$\left[\begin{array}{cccc} W_1 & V_1 & I_1 & C_1 \\ W_2 & V_2 & I_2 & C_2 \end{array}\right]\left[\begin{array}{c} B_W \\ B_V \\ B_I \\ B_C \end{array}\right] = 0 \qquad (3.43)$$

where matrices W_t and V_t are defined as in theorem (5), C_t is a matrix of dimension $(n_t(\sum_{i=1}^{n_t} Q_{it}) \times J(J-1)K)$, I_t is a matrix of dimension $\left(n_t(\sum_{i=1}^{n_t} Q_{it}) \times J(J-1)^2K\right)$ for $t = 1, 2$.

Generically, matrix C_1 contains the difference between the characteristics of the product in the lower boundary $R_{ij}(\theta, \delta)$ and product j, for each individual with $z_i \in P_1(\theta^0, \delta^0, \theta, \delta, X)$ and $j \in Q_{1i}$. The structure of C_1 consists in three blocks. There are J blocks, one for each product $j \in Q_{1i}$. Each of these blocks has K blocks for each product characteristics. Then, for each product j and characteristic k, there are $J-1$ columns where only one value is different from zero for each individual/product (row). Only column $R_{ij}(\theta, \delta)$ has value $(x_{R_i(\theta,\delta),k} - x_{jk})$ for each row i for the column block corresponding to $j \in Q_{1i}$ and k. Matrix C_2 is defined accordingly for $r_i(\theta, \delta)$.

The expression for B_C has to be defined accordingly to the definitions of C_1 and C_2. Thus, B_C is a column vector of dimension $J(J-1)K$ whose elements are expressed as

$$B_C = \left\{ \left\{ \left\{ (\delta_j - \delta_s)\, \theta^0_{2k} - \left(\delta^0_j - \delta^0_s\right) \theta_{2k} \right\}_{s=1}^{J-1} \right\}_{k=1}^{K} \right\}_{j=1}^{J} \tag{3.44}$$

Thus, B_I is a column vector of dimension $J(J-1)^2K$ whose elements are expressed as

$$B_I = \left\{ \left\{ \left\{ \left\{ (\delta_j - \delta_s)\, \theta^0_{2k} \Delta x_{jh,k} - (\delta_j - \delta_h)\, \theta^0_{2k} \Delta x_{js,k} \right\}_{s=1}^{J-1} \right\}_{h=1}^{J-1} \right\}_{k=1}^{K} \right\}_{j=1}^{J} \tag{3.45}$$

Matrices I_1 and I_2 are matrices of zeros except for certain cells that take value one. For each row/individual, there exist as many columns as combinations between δ_j, δ_s and δ_h. For example, for an individual with $z_i \in P_1(\theta^0, \delta^0, \theta, \delta, X)$ and $j \in Q_{1i}$, row of matrix I_1 has value one in the corresponding cell for all the characteristics of the combination of products $j, R_{ij}(\theta^0, \delta^0)$ and $R_{ij}(\theta, \delta)$.

Submatrices $\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$ have full column rank because different individuals have different products as upper and lower boundaries. And as before Assumptions 3.(3)-3.(4) ensure that the whole matrix $\begin{bmatrix} W & V & I & C \end{bmatrix}$ has full column rank. Then, $B_W = 0, B_V = 0, B_I = 0, B_C = 0$. The conditions $B_W = 0, B_V = 0$ allows one to identify θ^0 with respect to θ up to a scalar constant as before. We next show that $B_C = 0$ allows identification of the product specific constants.

$B_C = 0$ implies

$$(\delta_j - \delta_s)\, \theta^0_{2k} = \left(\delta^0_j - \delta^0_s\right) \theta_{2k} \text{ for } \forall k, \forall j \in \Im, \forall s \neq j$$

As in theorem (3.2.3), θ_{21} is normalized to one. Therefore, for $k = 1$

$$(\delta_j - \delta_s) = \frac{(\delta_j^0 - \delta_s^0)}{\theta_{21}^0} \; \forall k, \forall j \in \Im, \forall s \neq j$$

By normalizing, $\delta_1 = 0$, one can identify the specific constant for each $j \in \Im$, $j \neq 1$ up to scale and normalized with respect to δ_1.

Since we have shown that both θ^0 and δ^0 are identified, this implies that the boundary products are the same with the true value of the parameters and with the alternative value ($R_{ij}(\theta^0, \delta^0) = R_{ij}(\theta, \delta)$ and $r_{ij}(\theta^0, \delta^0) = r_{ij}(\theta, \delta)$). This is consistent with the conclusion that B_I equals zero. For each individual with $z_i \in P_1(\theta^0, \delta^0, \theta, \delta, X)$ and $j \in Q_{1i}$ for instance, the relevant element of B_I is

$$\left\{(\delta_j - \delta_s)\, \theta_{2k}^0 \Delta x_{jh,k} - (\delta_j - \delta_h)\, \theta_{2k}^0 \Delta x_{js,k}\right\}$$

with $s = R_{ij}(\theta, \delta), h = R_i(\theta^0, \delta^0)$. Thus, since $s = h$, one should expect then that elements of B_I that interact $s = h$ should be zero as we obtain. QED.

Proof of Theorem (7) (Semiparametric Identification of the Pure Characteristics Model without ξ)

Our definition of identification is equivalent to

$$\Pr\Big\{z \in Z \text{ such that } \theta = \theta_0 \text{ if } \Pr(\mathrm{d}_{ij} = 1|x, z, F_\varepsilon^0; \theta^0) = \Pr(\mathrm{d}_{ij} = 1|x, z, F_\varepsilon; \theta)$$
$$\text{for all j} \in J\Big\} > 0$$

The proof follows the same reasoning as in the parametric proof up to expressions (3.30) and (3.31), although throughout the proof different distribution functions F_ε and F_ε^0 should be used in the equalities. Thus, using the notation in the parametric identification theorem, these two expression become

$$\begin{aligned} F_\varepsilon^0(\underline{\Delta}_j(z_i, x; \theta^0)) &= F_\varepsilon(\underline{\Delta}_j(z_i, x; \theta)) \text{ for } z \in \rho_j(\theta^0, \theta, X) \text{ where } j \in \Im_1(\theta, \theta^0) \\ F_\varepsilon^0(\bar{\Delta}_j(z_i, x; \theta^0)) &= F_\varepsilon(\bar{\Delta}_j(z_i, x; \theta)) \text{ for } z \in \rho_j(\theta^0, \theta, X) \text{ where } j \in \Im_2(\theta, \theta^0) \end{aligned}$$

Recall the definition of the upper and lower bound indices

$$\begin{aligned} \bar{\Delta}_j(z_i, x; \theta^0) &= \Pi'_{0ij} z_i = \frac{\theta_1^{0\prime}(x_j - x_{r_{ij}(\theta_0)})}{\theta_2^{0\prime}(x_{r_{ij}(\theta_0)} - x_j)} z_i \\ \bar{\Delta}_j(z, x; \theta) &= \Pi'_{ij} z_i = \frac{\theta_1'(x_j - x_{r_{ij}(\theta)})}{\theta_2'(x_{r_{ij}(\theta)} - x_j)} z_i \end{aligned}$$

For each j, define the following set

$$Q_{\Pi j} = \left\{ z \subset \Omega_Z \ \ s.t. \ \left[\left(\Pi'_{0j} z < 0 \leq \Pi'_j z \right) \cup \left(\Pi'_j z < 0 \leq \Pi'_{0j} z \right) \right] \cap \rho_j(\theta^0, \theta, X) \right\}$$

Given the statistical independence of ε and z (and hence, median independence), conditions (ii) and (iv) in this theorem ensure that $Q_{\Pi j}$ has a positive probability for each j. Thus, the continuous variable z_m ensures that one can find values of this variable to reverse the sign of indices $\Pi'_j z$ and $\Pi'_{0j} z$ (see Manski (1985) in the identification proof of the semiparametric binary discrete model with median independence). Therefore, we can conclude from here that Π_{0ij} for $z_i \in Q_{\Pi j}$ is identified up to scale (Π_{0ij} is a $(M \times 1)$ vector and therefore we need to normalize of the these coefficients, say $\Pi_{0ij}, 1 = 1$). Next we show that the identification of Π_{0ij} implies identification of θ^0, since

$$\Pi_{0ij,m} = \frac{\theta^{0\prime}_{1m}(x_j - x_{r_{ij}(\theta_0)})}{\theta^{0\prime}_2(x_{r_{ij}(\theta_0)} - x_j)} = \frac{\theta'_{1m}(x_j - x_{r_{ij}(\theta)})}{\theta'_2(x_{r_{ij}(\theta)} - x_j)} = \Pi_{ij,m}$$

for all $m \neq 1$, for all $z_i \in \rho_j(\theta^0, \theta, X)$ and for all $j \in \Im_1(\theta, \theta^0)$ (for $j \in \Im_2(\theta, \theta^0)$, the same expression as above with $R_{ij}(\theta)$ and $R_{ij}(\theta_0)$ instead holds).

Solving for the parameter values the system of equations defined by

$$\left[\sum_{k=1}^{K} \theta^0_{1m,k} (x_j - x_{r_{ij}(\theta_0)})_k \right] \times \left[\sum_{k=1}^{K} \theta_{2,k} (x_j - x_{r_{ij}(\theta_0)})_k \right] -$$
$$- \left[\sum_{k=1}^{K} \theta_{1m,k} (x_j - x_{r_{ij}(\theta_0)})_k \right] \times \left[\sum_{k=1}^{K} \theta^0_{2,k} (x_j - x_{r_{ij}(\theta_0)})_k \right] = 0$$

for all $m \neq 1$, for all $z_i \in \rho_j(\theta^0, \theta, X)$ and for all $j \in \Im_1(\theta, \theta^0)$, and following the same procedure as in the parametric proof, Assumption 3.(4) on the full rank of matrix X guarantees that

$$\theta_{1km} = \frac{\theta^0_{1km}}{\theta^0_{21}}, \forall k = 1, ..., K, m = 2, ..., M$$
$$\theta_{2k} = \frac{\theta^0_{2k}}{\theta^0_{21}}, \forall k = 1, ..., K$$

QED.

Chapter 4

Semiparametric Least Squares Estimation of Shape Invariant Models with Multiple Equations: An Application to Engel Curves

4.1 Introduction

In the semiparametric literature much attention has been given to the estimation of shape invariant nonparametric regression curves (Lawton, Sylvestre and Maggio (1972), Hardle and Marron (1990) and Pinkse and Robinson (1995)). The main idea of this model is that although no parametric restrictions are desirable to be imposed on the regression curve, one might be interested in quantifying the differences between curves for different samples. The unknown conditional mean functions for different samples are related by some parametric transformations which are known up to a finite number of parameters. These parameters shift and scale the unknown function without altering its overall shape. Both the unrestricted conditional mean functions and the finite dimensional parameters that relate these functions for observations belonging to different samples are potential parameters of interest. In this work we focus on the identification and estimation of the finite dimensional parameters that imply a vertical and an horizontal shift of the nonparametric regression functions.

This paper proposes an alternative way of estimating the finite dimensional parameters

of the shape invariant model by the Semiparametric Least Squares estimator (henceforth, SLS) introduced by Ichimura (1993). We argue that this is a natural way of estimating the differences between unknown regression curves. Also, the estimators proposed in the early literature face some computational difficulties because the objective function attains only a local minimum at the true value of the parameters even for those models where the parameters are identified. This means that computational intensive methods should be used to find the local minimum close to the true value of the parameters. Although we find that SLS also faces some similar computational difficulties, the modification of the SLS estimator we propose here solves this problem and it is computationally less costly to obtain consistent estimates of the parameters[1]. Due to this property of the suggested estimator, it is feasible to deal with the comparison of the regression curves of more than two samples without adding much computational cost if this estimation method is used. It also extends the original framework of SLS to the estimation of a system of equations where there might exist correlation between the errors of the different equations.

We only consider in this work the case where both the transformation of the argument of the unknown function and the transformation of the function itself are linear. This represents a more restrictive model than the one considered by Hardle and Marron (1990) and Pinkse and Robinson (1995) which discuss the nonlinear transformation case. The advantage is that the linearity assumption of the parametric transformations allows us to be more specific about the identification conditions of the finite dimensional parameters.

Consumer demand and, in particular, the estimation of Engel curve relationships constitute an important area for the application of semi and non parametric methods. Shape invariant models arise in these applications for some specifications of the demographic composition. Early works in the analysis of Engel curves (Hardle and Jerison (1988) and Blundell and Duncan (1998)) use nonparametric techniques to estimate the unrestricted relationship between budget shares and total expenditure. When this relationship wants to be adjusted by observed heterogeneity (i.e. demographics) the way of modelling it in a semiparametric model becomes an important issue. Blundell, Duncan and Pendakur (1998) show that if the demographic composition enters in a partial linear way (as in Robinson (1988)) the conditions for the consistency of the consumer theory impose strong restrictions on the functional shape of the Engel curves. The shape invariant model arises because a generalization of this model needs to be considered to incorporate the hetero-

[1] Wilke (2003) proposes a modification of Pinkse-Robinson and Hardle-Marron estimator that also solves for the local minimum problem at the true value of the parameters.

geneity in tastes that arises from the demographics (Blackorby and Donaldson (1994)). This more general model rescales the total expenditure variable inside the unknown function by the demographic composition . To obtain a good estimate of the parameters that rescale the total expenditure for different demographic groups is important in order to semiparametrically estimate consumption based equivalence scales. Blundell, Duncan and Pendakur (1998) use the British Family Expenditure Survey (FES) to estimate Engel curve relationships with this extended partial linear model using the Pinkse and Robinson (1995) estimator. We use the same FES data and obtain estimates of the finite dimensional parameters of interest using SLS estimator for multiple equations. We compare our results with the previous estimates that can be found in the literature from the Pinkse-Robinson estimator and also from the modified estimator proposed by Wilke (2003). [2]

This paper is organized as follows. Section (4.2) introduces the notation and the model used throughout this work. Section (4.3) discusses the previous approaches to estimate the model and the computational problems they face. Section (4.4) proposes an alternative method to estimate the parameters of the model by using a modified version of SLS and Section (4.5) gives sufficient conditions for the identification of these parameters. Section (4.6) establish the large sample properties of the SLS estimator for a system of equations. The optimal weighting matrix for the different equations is discussed as well there. Section (4.7) shows some Montecarlo experiments that demonstrate the estimator we propose performs better than the early previous estimators proposed in the literature in finite samples when gradient methods are used to compute the minimum of the objective function. Section (4.8) applies the SLS estimator to the estimation of Engel curves using the British Family Expenditure Survey and Section (4.9) concludes.

4.2 Model and Notation

The shape invariant model is described as follows. We observe J different outcomes for the same individual. Random vector $W = [W_1, ..., W_J]$ denotes the J different outcomes and X and Z are exogenous variables. The supports of exogenous variables X and Z are denoted by Ω_X and Ω_Z, respectively. We also use $\Omega_X^{(0)}$ and $\Omega_X^{(1)}$ to denote the supports

[2] In this work, we abstract from the problem that might arise due to the endogeneity of the total expenditure, since it is likely to be simultaneously determined with budget shares. To adjust for this endogeneity problem has been found to be important as documented in Blundell, Chen and Kristensen (2003).

of the conditional random variables $X|Z=0$ and $X|Z=1$, respectively. Let denote by $Y=(W,X,Z)\in R^{d_y}$ the observable random variables with $d_y=J+d_x+d_z$ where d_x and d_z are the dimensions of X and Z, respectively. All the equations share the same exogenous variables and the system of equations is defined as follows

$$w_{ji}=\phi_j(x_i-c_0'z_i)+a_{0j}'z_i+\varepsilon_{ji} \tag{4.1}$$

for $j=1,..,J$ where $i=1,...,N$ refers to individuals and j refers to the outcomes. The vector of parameters in the linear part of the conditional mean is denoted by $a=[a_1',a_2',...,a_J']'$. Let $\mathcal{A}\subset\mathcal{R}^{Jd_z}$ and $\mathcal{C}\subset\mathcal{R}^{d_z}$ be the parameter space for a and c, respectively.

The function $\phi_j: R\rightarrow R$ for $j=1,...,J$ is not known. We assume conditional mean independence of the errors and the exogenous variables

$$E(\varepsilon_j|x,z)=0 \text{ for all } j.$$

The unknown conditional mean function differs across outcomes and also the coefficients a_{0j} in the linear part of the model are equation-specific. However, the parameter c_0 imbedded inside the unknown function ϕ is common for all the equations.

We assume that Z are binary discrete variables taking only values $\{0,1\}$. Let Z_r for $1\leq r\leq d_z$ be an element of random vector Z. This allows one to interpret the $r-th$ element of parameter a_{0j} as the vertical difference in the conditional mean $E(w_j|x)$ between observations with $Z_r=1$ and the sample with $Z_r=0$. The $r-th$ element of parameter c_0 implies an horizontal shift of the conditional mean function of observations with $Z_r=1$ with respect to observations with $Z_r=0$ and it also implies a change in the slope of both conditional mean functions. In this way, we can also compare the nonparametric regression curves of observations belonging to different groups defined, for example, by a combination of two different elements Z_r and Z_s for $r\neq s$ of random vector Z. Throughout the paper however, for simplicity, we restrict ourselves to cases where $d_x=1$ and $d_z=1$ so that parameters c and a_j are unidimensional parameters for $j=1,...,J$.

In the application of the estimation of Engel curves where demographic adjustments are taken into account, w_{ji} is the budget share for good j and household i, x_i denotes the logarithm of total expenditure and z_i is a discrete variable taking value {0,1} describing to which demographic group observation i belongs to.

4.3 Previous estimators for the shape-invariant model

We focus on the case where Z is unidimensional and can take only two possible values $\{0,1\}$. Let denote the conditional mean function of w_j given x for each of these two subsamples as

$$\begin{aligned} e_j^{(0)}(x_i) &= E(w_j|x_i, Z_i = 0) = \phi_j(x_i) \\ e_j^{(1)}(x_i) &= E(w_j|x_i, Z_i = 1) = a_{0j} + \phi_j(x_i - c_0) \end{aligned} \tag{4.2}$$

so that the following relationship holds

$$e_j^{(1)}(x_i) = a_{0j} + e_j^{(0)}(x_i - c_0) \text{ for } j = 1, ..., J \tag{4.3}$$

Hardle and Marron (1990) and Pinkse and Robinson (1995) discuss estimators of parameters a_0 and c_0 for a wider class of models than the model outlined above. Thus, expression (4.3) relates both conditional means by two linear transformations (known up to a finite number of parameters) of the regression function $e^{(0)}$: one for the argument and another one for the function itself. In this work, we restrict ourselves to this case. However, Hardle and Marron (1990) cover the more general case where both transformations are non-linear and Pinkse and Robinson (1995) consider the case where only the transformation to the function is linear, but the transformation of the argument of $e^{(0)}$ is not necessary linear.[3]

The relationship between both conditional means expressed in (4.3) allows one to understand the estimators proposed in the literature of shape invariant estimation. These estimators are obtained through the minimization of the following loss functions. Hardle and Marron (henceforth, HM) suggest to minimize with respect to a and c

[3] Although not discussed by the above mentioned authors, this framework allows for Z being discrete (not necessarily binary) variables and multidimensional Z. In the case of Z taking for example values $Z = \{0, 1, 2\}$, the following relationships hold

$$\begin{aligned} e_j^{(1)}(x_i) &= a_{0j} + e_j^{(0)}(x_i - c_0) \\ e_j^{(2)}(x_i) &= 2a_{0j} + e_j^{(0)}(x_i - 2c_0) \end{aligned}$$

for $j = 1, ..., J$. In the case of multidimensional Z, similar relationships can be found among the conditional mean functions defined as

$$e_j^{(st)}(x_i) = E(w_j|x_i, Z_{1i} = s, Z_{2i} = t)$$

for $s, t = \{0, 1\}$. The weight that each of the above equalities should be given in the estimation is rather important and not discussed in the original set up where these estimators were introduced.

$$L^{HM}(a,c) = \sum_{j=1}^{J} \int_{\underline{x}}^{\overline{x}} \left[\hat{e}_j^{(1)}(x) - a_j - \hat{e}_j^{(0)}(x-c)\right]^2 w(x)dx \tag{4.4}$$

where $w(x)$ is a weight function which is nonnegative and positive only on the interior of a compact interval $[\underline{x}, \bar{x}]$ and $\hat{e}_j^{(1)}(x)$ and $\hat{e}_j^{(0)}(x-c)$ are consistent estimators of the conditional mean functions defined in (4.2). Let define the conditional mean functions as the the ratio of the following two functions: $e_j^{(1)}(x) = r_j^{(1)}(x)/f^{(1)}(x)$ and $e_j^{(0)}(x-c) = r_j^{(0)}(x-c)/f^{(0)}(x-c)$. Pinkse and Robinson (henceforth, PR) suggest to modify this loss function by multiplying by $\hat{f}^{(0)}(x-c) * \hat{f}^{(1)}(x)$ so that

$$L^{PR}(a,c) = \sum_{j=1}^{J} \int_{\underline{x}}^{\overline{x}} \left[\hat{f}^{(0)}(x-c)\hat{r}_j^{(1)}(x) - a_j \hat{f}^{(1)}(x)\hat{f}^{(0)}(x-c) - \hat{f}^{(1)}(x)\hat{r}_j^{(0)}(x-c)\right]^2 w(x)dx \tag{4.5}$$

The reason Pinkse and Robinson argue in order to modify Hardle and Marron's loss function is for computational purposes in the derivation of the asymptotic properties, since the ratios of nonparametric estimators have been replaced by multiplications whose properties are easier to compute.

The weight function $w(x)$ not only selects the integration limits $[\underline{x}, \bar{x}]$ but also helps in the efficiency of the estimator. The appropriate choice of the integration limits is crucial to define both objective functions and for the performance of the estimator, even if $e^{(1)}$ and $e^{(0)}$ were known functions. This is because the integrand in (4.4) is not defined if $x - c$ does not lie in the support $\Omega_X^{(0)}$. Therefore, in order for the objective function to be well defined for each value c, the intersection of the supports $\Omega_X^{(0)} + c \cap \Omega_X^{(1)}$ should not be empty and $[\underline{x}, \bar{x}]$ should be chosen to lie in this intersection. This is because we need to integrate over values x belonging to the support $\Omega_X^{(1)}$ such that if we substract c still belongs to the support $\Omega_X^{(0)}$.When both random variables X_1 and X_0 have full support on the entire real line, this problem does not exist because $e^{(0)}$ can be defined at any value of x and c.

When functions $e^{(0)}$ and $e^{(1)}$ are unknown and need to be estimated, the loss functions should be integrated over a range of x such that both conditional means are consistently estimated and are well defined. Let consider the weighting function as the indicator function that takes only value 1 in the intersection of supports $\Omega_X^{(0)} + c$ and $\Omega_X^{(1)}$ (i.e. $w(x) = 1\{x \in \Omega_X^{(0)} + c \cap \Omega_X^{(1)}\}$). As Hardle and Marron point out, the fact that the weighting function and therefore the integration limits depends on the value of the parameter c makes that the loss function is minimized attaining value zero at those arbitrary high or small

values of c such that the intersection of the supports is empty and the indicator (weighting) function is always zero. Since they reckon that this feature imposes some difficulties in the computation of the estimator, they suggest to establish a priori a compact set $\mathcal{C}$ with feasible values for parameter c and then define the weighting function as

$$w(x) = \prod_{c \in \mathcal{C}} 1\{x \in \Omega_X^{(0)} + c \cap \Omega_X^{(1)}\}$$

There are two drawbacks of defining the weighting function in this way. First, to determine a reasonable set of values c for the transformation of the argument of the unknown function might not be easy for some applications, where depending on the shape of ϕ_j a graphical analysis beforehand may not be very informative. Second, if the variables X_0 and X_1 have compact support and the compact parameter space $\mathcal{C}$ is big enough, then the set of values of x where one can evaluate the loss function might become very small. If this is the case, some identification difficulties may arise then because as it is formally shown below, the parameters are identified under the nonlinearity assumption of ϕ_j for at least one $j = \{1, ..., J\}$. This nonlinearity assumption might be violated if the support where we evaluate ϕ is very small and the function is approximately linear over this range.

Alternatively, Pinkse and Robinson define the weighting function such that takes only nonnegative and positive values on the interior of a compact interval where all the points x satisfy $f^{(0)}(x - c) > 0$ and $f^{(1)}(x) > 0$ for $c \in \mathcal{C}$. The objective function takes value zero at those values of c for which there is no value of x such that densities $f^{(0)}(x - c)$ and $f^{(1)}(x)$ are both bounded away from zero. Again, if one knows that random variables X_0 and X_1 have full support on the real line, then the weighting function $w(x) = 1$ for all x belonging to the support of X.

In practical terms, the choice of the integration limits for x should guarantee that estimates $\hat{f}^{(0)}(x - c)$ and $\hat{f}^{(1)}(x)$ are consistently estimated away from zero for every $x \in [\underline{x}, \bar{x}]$ for each value of $c \in \mathcal{C}$. Even if X has full support on the real line, the observed supports in finite samples denoted by $\hat{\Omega}_X^{(0)}$ and $\hat{\Omega}_X^{(1)}$ are compact sets. Thus, although $f^{(0)}(x - c)$ may be bounded away from zero, it might be the case that $(x - c)$ is outside the observed (or estimated support) $\hat{\Omega}_X^{(0)}$, in which case $\hat{f}^{(0)}(x - c)$ is not going to be consistently estimated bounded away from zero. It is possible to estimate consistently both densities for those values of x belonging to the intersection of the observed (or estimated) supports $\hat{\Omega}_X^{(1)}$ and $\hat{\Omega}_X^{(0)} + c$.

Following the same reasoning as before, the minimum of the objective function is attained at zero for those values of c that make the intersection $\hat{\Omega}_X^{(1)} \cap \hat{\Omega}_X^{(0)} + c$ to be

empty. A look at the support of X for both demographic groups determine the set of values of c that could potentially be identified from the data by yielding a value of the loss function different from zero. The case in which we are interested (and also the one that implies some computational difficulties) is when

$$c_0 \in \tilde{C} \equiv \{c \in \mathcal{C} \text{ such that } \hat{\Omega}_X^{(1)} \cap \hat{\Omega}_X^{(0)} + c \neq \varnothing\}$$

such that the loss function attains only a local minimum at c_0.

To illustrate this, Figure (4.1) plots the nonparametric kernel densities of $f^{(0)}(x_i - c)$ and $f^{(1)}(x_i)$ using the FES data described below in the empirical section for each observation x_i of random variable X_1 (log total expenditure for demographic group Z=1) and for different values of $c = \{1, 2\}$. As parameter c increases, the points at which we should evaluate the nonparametric density of demographic group $Z = 0$ lie outside the observed support of X_0 in the data so that $f^{(0)}(x_i - c)$ is not consistently estimated away from zero for these points. Thus, in this particular case, if c is much higher than 2 there would not exist overlap between the observed supports $\hat{\Omega}_X^{(1)}$ and $\hat{\Omega}_X^{(0)} + c$ in the data.

The top graphs in Figure 4.2 show the loss function for PR and HM estimators with respect parameter c for the Monte Carlo simulations reported in Section 4.7 (see details there). Both random samples for X_1 and X_0 were drawn from different normal distributions. The loss functions are evaluated at a_0. It should be pointed out that the corresponding loss function with respect to parameter a behave nicely as expected being globally concave. Those values that do not belong to $\tilde{C}$ are easily identified from the graph since they give zero value to the loss function. Also the true loss function for HM where function ϕ is assumed to be known is plotted, but also choosing the integration limits for x as a function of c and of the supports $\hat{\Omega}_X^{(0)} + c$ and $\hat{\Omega}_X^{(1)}$. These graphs illustrate the difficulties that arise in the minimization of HM and PR loss function to find the local minima that it is close to the true value of the parameters (a_0, c_0).[4]

The minimization of both loss functions $L^{HM}(a, c)$ and $L^{PR}(a, c)$ with respect to a and c should be constrained so that $c \in \tilde{C}$. However, it is unlikely that this might solve the local minimum problem because of the particular behavior of the loss functions inside

[4]For each value of c, the intersection between the observed supports of the drawn data is computed. Once the integration limits are set, the integral is computed using the middpoint approximation (see Judd (1998)). Other alternatives include the use of the observations belonging to the computed intersections of the supports to compute their sample mean. The loss functions computed in this way perform slightly worse than the ones using the middle point approximation, since as c increases the number of observations in the intersection decreases quickly.

the parameter set $\tilde{C}$. The difficulties in the computation of the minimum of loss functions L^{PR} and L^{HM} is that they tend to zero as the intersection $\hat{\Omega}_X^{(1)} \cap \hat{\Omega}_X^{(0)} + c$ becomes very small as c increases in absolute value, until the intersection becomes empty and then the weighting function gives zero value to the loss function.

This local minimum problem at the true value of the parameter has been solved in different ways in the applications of these estimators. For example, sequential minimization in a and c has been proposed, using grid search for the optimization with respect to parameter c over a reasonable set of values (see Blundell, Duncan and Pendakur (1998)). However, other standard and less tedious optimization methods would perform very poorly given the behavior of the loss functions for PR and HM estimators.

However, grid search methods are computationally costly if the dimension of the parameters inside the unknown conditional mean function is high. That it is, if for example one is interested in estimating the parametric shifts in the Engel curves with respect to both the demographic composition of the household and the employment status of the head of the household, the dimension of parameter c in this case would increase the computational cost of doing grid search over a set of reasonable values for these parameters.

Wilke (2003) solves this problem by modifying the above objective function of HM (and also PR) by dividing for the density attained at the overlap of the corresponding supports, which makes that the loss function increases when the intersection becomes small and improves the performance of the estimator in finite samples. Thus, the loss function for this estimator is modified in the following way

$$L^W(a,c) = \frac{\sum_{j=1}^{J} \int_{\hat{\Omega}_X^{(1)} \cap \hat{\Omega}_X^{(0)}+c} \left[\hat{e}_j^{(1)}(x) - a_j - \hat{e}^{(0)}(x-c)\right]^2 w(x)dx}{\int_{\hat{\Omega}_X^{(1)} \cap \hat{\Omega}_X^{(0)}+c} \hat{f}^{(1)}(x) w(x)dx}$$

The third graph in Figure (4.2) shows this objective function and illustrates how his modification can help in the estimation of parameter c.

This M-shape of the PR and HM loss functions with respect to parameter c is due to the fact that the integration limits depend on c and also because the functions $f^{(0)}, e^{(0)}, f^{(1)}$ and $e^{(1)}$ need to be estimated using the observed supports $\Omega_X^{(1)}$ and $\Omega_X^{(0)}$. If we use the information on the parametric form of functions $e^{(1)}$ and $e^{(0)}$ and the full support of X_0 and X_1 on the whole real line ensures that $e^{(0)}$ is well defined uniformly for all c so that the integration range for x can be defined independently of c, the last graph in Figure (4.2) shows that the loss function would be globally concave also with respect to parameter c.

Figures (4.3) and (4.4) show the different shapes of the loss functions for PR and HM when the integration range for x is fixed for all c and lies on the support where both $f^{(0)}(x_i)$ and $f^{(1)}(x_i)$ are estimated consistently. These graphs illustrate that even if the integration limit for x does not depend on c, the loss function would decrease as the value of the parameter c increases in absolute value as the estimates of $\hat{f}^{(0)}(x-c)$ become arbitrary small. This shows that the choice of this integration limit is very important in order to obtain consistent estimates of the parameters.

4.4 Estimating the shape invariant model using SLS

We suggest to estimate the parameters of interest by using Semiparametric Least Squares (henceforth, SLS) proposed by Ichimura (1993). With respect to the estimators proposed by Pinkse-Robinson and Hardle-Marron, the modification of the estimator we introduce here solves the computational problem of finding the local minimum attained at the true value of the parameters. Additionally, with respect to all the estimators discussed in the previous section, SLS constitutes a natural way of estimating the parameters of interest and it helps to extend the idea of comparing nonparametric regression curves to more than two independent samples (since Z does not need to be binary) or when the comparison wants to be done in more than one dimension (when Z is multidimensional).

The identification conditions discussed in Section 4.5 ensure that the true value of the parameters uniquely solves the following loss function [5]

$$\{a_{01}, ..., a_{0J}, c_0\} = \arg\min_{(a,c)} L(a,c) \tag{4.6}$$

$$= \arg\min_{(a,c)} \sum_{j=1}^{J} E\left\{[w_j - za_j - E(w_j - za_j|x-cz)]^2\right\} \tag{4.7}$$

As discussed in Ichimura (1993), the variation in $w_j - za_j$ for each j comes from both variation of ε_j and $(x - c_0 z)$ and also from the variation in z if $a_j \neq a_{j0}$. Therefore, if index takes a constant value s, i.e. $(x - c_0 z) = s$ and $a_j = a_{0j}$, then the variation arises uniquely from ε_j. Thus, the variance in loss function (4.6) is minimized when $c = c_0$ and

[5] It is important to note the difference of this objective function with the one of NLLS where function m is known up to c. Both objective functions differ at parameter values different from the true values since

$$E(w_j - za_j|x - cz) \neq m(x - cz)$$

if $c \neq c_0$ and $a \neq a_0$

$a = a_{0j}$ for all j. The objective function above focuses only in the minimization of the variance of $w_j - a_j z$ for each j. However, the same argument could be used to argue that the covariance between $(w_j - za_j)$ and $(w_r - za_r)$ for $j \neq r$ is minimized at the true value of the parameters. Therefore, the argument above also suggest that the true value of the parameters also minimize

$$\{a_{01}, ..., a_{0J}, c_0\} = \tag{4.8}$$

$$\arg\min_{(a,c)} E\left\{ \begin{bmatrix} w_1 - za_1 - E(w_1 - za_1|x - cz) \\ \vdots \\ w_J - za_J - E(w_J - za_J|x - cz) \end{bmatrix}' V \begin{bmatrix} w_1 - za_1 - E(w_1 - za_1|x - cz) \\ \vdots \\ w_J - za_J - E(w_J - za_J|x - cz) \end{bmatrix} \right\}$$

where V is a semi-definite positive matrix of size $J \times J$ which can depend on the data. The minimum of function (4.8) is also attained at $\mathrm{E}\{\varepsilon' V \varepsilon\}$ where $\varepsilon = [\varepsilon_1; ..; \varepsilon_J]$ when $a = a_0$ and $c = c_0$.

If the conditional mean $E(w_j - za_j|x - cz)$ is known up to a and c, the identification conditions guarantee that the loss function is globally minimized at (a_0, c_0) as long as density function of the index $f_{X-cZ}(x - cz)$ is bounded away from zero uniformly in c, x and z, so that the conditional mean function is well defined. This last condition holds if the random variable $X|Z$ has full support on the real line uniformly on Z and $0 < \Pr(Z = 1) < 1$ as it can be checked from what follows. If the density function of the index is evaluated at observations such that $z = 1$

$$\begin{aligned} f_{X-cZ}(x - c) &= f_{X-cZ|Z}(x - c|Z = 1)\Pr(Z = 1) + f_{X-cZ|Z}(x - c|Z = 0)\Pr(Z = 0) = \\ &= f_X(x|Z = 1)\Pr(Z = 1) + f_X(x - c|Z = 0)\Pr(Z = 0) \end{aligned} \tag{4.9}$$

and if one evaluates at observations such that $z = 0$,

$$\begin{aligned} f_{X-cZ}(x) &= f_{X-cZ|Z}(x|Z = 1)\Pr(Z = 1) + f_{X-cZ|Z}(x|Z = 0)\Pr(Z = 0) = \\ &= f_X(x + c|Z = 1)\Pr(Z = 1) + f_X(x|Z = 0)\Pr(Z = 0) \end{aligned} \tag{4.10}$$

In the application of this model to the estimation of demand systems where X is the logarithm of total expenditure this condition holds since $X|Z$ has indeed full support for both $Z = 0$ and $Z = 1$. If this condition is not satisfied, then we would need to introduce an indicator function $I_Q = 1\{(x, z) \in Q\}$ inside the expectation in (4.6) where set Q is

defined as follows

$$\begin{aligned} Q &= \{(x,z) \in \Omega_X \times \Omega_Z \text{ s.t } f_{X-cZ}(x-cz) > 0 \text{ uniformly on } c \in \mathcal{C}\} \supset \\ &\supset \{(x,z) \in \Omega_X \times \Omega_Z \text{ s.t } f_X(x|Z=0) > 0 \text{ and } f_X(x|Z=1) > 0\} \end{aligned} \tag{4.11}$$

Thus, if X is not a random variable with full support we should guarantee that we evaluate the above expectations at values of $x \in \Omega_X$ belonging to the intersection of supports $\Omega_X^{(0)}$ and $\Omega_X^{(1)}$. Note that although both definitions of Q are equivalent, the second one does not depend on the parameter space $\mathcal{C}$.

Note that in the case of X with full support on the real line, the objective function of the previous estimators does not have a local minima problem at (a_0, c_0) when function ϕ is known (see graph 4 in Figure (4.2)). In this case, the integration limits could be defined independently of the value of c and the objective function is globally concave with respect to c. The problem with these estimators is that even in the case where ϕ is known, if random variable $X|Z$ does not have full support on the real line, the integration limits for x should be chosen so that $f^{(0)}(x-c)$ and $f^{(1)}(x)$ are well defined. Then, either one has a clear idea of the compact set where c_0 lies in and the weighting function and the integration limits are chosen uniformly for $c \in \mathcal{C}$, or the integration limits should be defined depending on c which leads to the unpleasant M-shape of the objective function with respect to parameter c. Also in finite samples, this problem is present even if $X|Z$ has full support because the estimator should be defined with respect to the observed supports $\hat{\Omega}_X^{(0)}$ and $\hat{\Omega}_X^{(1)}$.

In SLS, there exists still a computational problem when the conditional mean is unknown and needs to be estimated. Consider the following SLS estimator where nonparametric kernel estimators are used to obtain a sample analogue of (4.6)

$$\{\hat{a}_1, ..., \hat{a}_J, \hat{c}\} = \arg\min_{(a,c)} \hat{L}(a,c) \tag{4.12}$$

$$\text{where } \hat{L}(a,c) = \sum_{j=1}^{J} \sum_{i \in Q} \left[w_{ij} - z_i a_j - \hat{E}_h(w_j - z a_j | x_i - c z_i) \right]^2 \tag{4.13}$$

$$\text{and } \hat{E}_h(w_j - z a_j | x_i - c z_i) = \frac{\frac{1}{h_n}\frac{1}{n-1}\sum_{r \neq i}^{n} (w_{rj} - z_r a_j) K\left(\frac{(x_i - c z_i) - (x_r - c z_r)}{h_n}\right)}{\frac{1}{h_n}\frac{1}{n-1}\sum_{r \neq i}^{n} K\left(\frac{(x_i - c z_i) - (x_r - c z_r)}{h_n}\right)}$$

where K is a kernel function and h_n is a bandwidth sequence dependent on the sample size. Using only observations belonging to set Q helps to show the uniform convergence of $\hat{E}_h(w_j - za_j|x_i - cz_i)$.

Figure (4.5) shows the objective function $\hat{L}(c|a_0)$ with respect to parameter c conditioned at the true value for a. It should be first said that the function achieves the global minimum at the true value c_0 of the parameters when it is evaluated at the rest of the true value of the parameters, which for example constitutes a difference with respect to the PR function. It does not seem however that the optimization problem would encounter less problems here to find the global minimum than the minimization of the objective functions of PR and HM estimators in finding the local minimum close at (a_0, c_0).

The reason for the flat ends of the objective function in Figure (4.5) is that for arbitrary big or small values of c the estimated density for the index achieves its lower bound which is independent of c as explained in what follows. Consider the nonparametric estimation of the density of the index for a given value of the bandwidth h_n for observations i where $z_i = 0$. This density can be bounded below uniformly in c by

$$\begin{aligned}\inf_{c\in\mathcal{C}}\left|\hat{f}_{X-cZ}(x_i)\right| &= \inf_{c\in\mathcal{C}}\left|\frac{1}{(n-1)\,h_n}\left[\sum_{z_r=0}K\left(\frac{x_i-x_r}{h_n}\right)+\sum_{z_r=1}K\left(\frac{x_i-(x_r-c)}{h_n}\right)\right]\right| \geq \\ &\geq \frac{(n_0-1)}{(n-1)}\frac{1}{(n_0-1)\,h_n}\sum_{z_r=0}K\left(\frac{x_i-x_r}{h_n}\right) = \widehat{\Pr}(z=0)\hat{f}_X(x_i|z=0)\end{aligned}$$

where n_0 is the number of observations such that $z = 0$. And equivalently, it can be shown that the lower bound for the estimated density of the index evaluated at observations i such that $z_i = 1$ is given by

$$\begin{aligned}\inf_{c\in\mathcal{C}}\left|\hat{f}_{X-cZ}(x_i-c)\right| &= \inf_{c\in\mathcal{C}}\left|\frac{1}{(n-1)\,h_n}\left[\sum_{z_r=1}K\left(\frac{x_i-x_r}{h_n}\right)+\sum_{z_r=0}K\left(\frac{(x_i-c)-x_r}{h_n}\right)\right]\right| \geq \\ &\geq \widehat{\Pr}(z=1)\hat{f}_X(x_i|z=1)\end{aligned}$$

We denote this lower bound of the density of the index by $\widehat{lb}(x_i, z_i)$, so that

$$\begin{aligned}\widehat{lb}(x_i,0) &= \widehat{\Pr}(z=0)\hat{f}_X(x_i|z=0)\\ \widehat{lb}(x_i,1) &= \widehat{\Pr}(z=1)\hat{f}_X(x_i|z=1)\end{aligned}$$

If a finite kernel is used, then the above inequalities hold with equality since

$$\inf_{c\in\mathcal{C}}\left|\sum_{z_r=1}K\left(\frac{x_i-(x_r-c)}{h_n}\right)\right| = 0$$

when set $\mathcal{C}$ is large enough. In this case, the minimum estimated density for observation i is attained for arbitrary large or small values of c at $\widehat{lb}(x_i, z_i)$. Note that in the population $f_{X-cZ}(x - cz)$ is strictly bounded above from the lower bound $lb(x, z)$ if and only if $f_X(x - c|Z = 0)$ and $f_X(x + c|Z = 1)$ are bounded away from zero.

Let γ_i be the value of $c \in \mathcal{C}$ that minimizes the estimated density of the index evaluated at observation i (i.e.$\gamma_i = \arg\inf_{c\in\mathcal{C}} \left|\hat{f}_{X-cZ}(x_i - cz_i)\right|$). Then the loss function for SLS does not depend on parameter c for those values of c belonging to the intersection $\cap_{i=1}^n \gamma_i$,

$$\frac{\partial \hat{L}(a, c)}{\partial c} = 0 \text{ for } c \in \cap_{i=1}^n \gamma_i$$

since also the weights of the nonparametric kernel regression would not depend on c for those values $c \in \cap_{i=1}^n \gamma_i$.

If $c_0 \in \cap_{i=1}^n \gamma_i$, the true value of the parameters would not be easily identified in finite samples with respect to the other values of the parameters in $\cap_{i=1}^n \gamma_i$ since they would attain the same value of the objective function. Therefore, we work under the assumption that $c_0 \notin \cap_{i=1}^n \gamma_i$. In the estimation, we would ideally want to rule out values of c such that do not belong to $\cap_{i=1}^n \gamma_i$. One possibility is to restrict the estimation to those values of $c \notin \cap_{i=1}^n \gamma_i$ in the same way that the estimation was restricted for the computation of PR and HM estimators to those values of c such that the intersection $\Omega_X^{(1)} + c \cap \Omega_X^{(0)}$ is not empty.

However, even if the estimation procedure focuses on those values of c where the loss function is not flat, the minimization routine might have difficulties in achieving the local minimum located close to the true value of the parameters (a_0, c_0) for certain starting values due to the shape of the objective function even if one constrains the minimization routine to find values of the parameter such that $c \notin \cap_{i=1}^n \gamma_i$ In this sense, this is a problem shared also by the PR and HM objective functions. The objective function with respect to c decreases as c becomes arbitrary large or small because the density of the index attains its lower bound $\widehat{lb}(x_i, z_i)$ for more and more observations.

For computational purposes, we implement the SLS estimator modifying the objective function so that, for each c, we divide by the number of observations where the estimated density of the index does not attain its lowest bound $\widehat{lb}$

$$\hat{L}_2(a, c) = \frac{\frac{1}{n}\sum_{j=1}^J \sum_{i\in Q} \left[w_{ij} - z_i a_j - \hat{E}_h(w_j - z a_j | x_i - cz_i)\right]^2}{\frac{1}{n}\sum_{i\in H} 1\left\{\hat{f}_{X-cZ}(x_i - cz_i) - \widehat{lb}(x_i, z_i) > 0\right\}} \tag{4.14}$$

and $\hat{L}_2(a,c) = 0$ if $1\left\{\hat{f}_{X-cZ}(x_i - cz_i) - \widehat{lb}(x_i, z_i) > 0\right\} = 0$, $\forall i \in H$. The set H is defined as

$$H = \{(x,z) \in \Omega_X \times \Omega_Z \text{ s.t } f_X(x-c|z=0) > 0 \text{ and } f_X(x+c|z=1) > 0 \text{ uniformly on } c \in \mathcal{C}\} \tag{4.15}$$

This implies that for those observations $(x_i, z_i) \in H$, the indicator function evaluated at the true densities $1\{f_{X-cZ}(x_i - cz_i) - lb(x_i, z_i) > 0\}$ equals 1 uniformly on $c \in \mathcal{C}$. This helps in concluding that the asymptotic properties of the estimators that minimize $\hat{L}_2(a,c)$ and $\hat{L}(a,c)$ are similar. However, the indicator function evaluated at the nonparametrically estimated density of the index and its lower bound can be different from one for some observations belonging to set H.

The objective function is defined to be 0 for those values of $c \in \cap_{i\in H}\gamma_i$. Since for these values of the parameters, the objective function achieves the global minimum, we also constrained the optimization to those values of $c \notin \cap_{i\in H}\gamma_i$.[6]

It is important to note that the limiting objective function of objective function $\hat{L}_2(a,c)$ is that same as for $\hat{L}(a,c)$ as it will be formally shown in the Section on the asymptotic properties of the estimator. Thus, this modification of the objective function does not have any implication in terms of identification. Additionally, one could think that this modification has made that the objective function is non differentiable with respect to parameter c, which could potentially complicate the asymptotic properties of the estimator. However, since in the limit this indicator function would attain value 1 for all the observations $i \in H$ uniformly on c, also in the limit the denominator would not change over $c \in \mathcal{C}$. We discuss these arguments rigorously in Section 4.6.

[6] One could also think that a trimming indicator that selects those observations where the estimated density of the index does not attain its lower bound would help in the performance of the estimator with respect to parameter c. That it is an estimator that solves the following objective function

$$\hat{L}_2(a,c) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i\in Q} 1\left\{\hat{f}_{X-cZ}(x_i - cz_i) - \widehat{lb}(x_i, z_i) > 0\right\}\left[w_{ij} - z_i a_j - \hat{E}_h(w_j - za_j|x_i - cz_i)\right]^2$$

Some simulation exercises, similar to the ones presented later in this work, were performed for this estimator. Although in principle, the local minimum problem for arbitrary large values of c is solved, the estimator becomes unstable when parameter c increases in absolute value and less and less observations are used in the computation of the objective function. In finite samples, although the global minimum is attained at the true value of the parameters when this modification is done, the objective function might present some local minimum when very few observations are used. We decided not to include the analysis corresponding to this modification of the SLS objective function in this work since the estimators we proposed here perfomed better.

To assess the impact of the discontinuity of the indicator function with respect to parameter c introduced in the denominator in (4.14), we also examine the properties of an analogous estimator where the indicator function is substituted by a continuous although non-differentiable function that converges to the above indicator function as sample size increases. This alternative estimator minimizes the following objective function

$$\hat{L}_3(a,c) = \frac{\frac{1}{n}\sum_{j=1}^{J}\sum_{i\in Q}\left[w_{ij} - z_i a_j - \hat{E}_h(w_j - z a_j | x_i - c z_i)\right]^2}{\frac{1}{n}\sum_{i\in H} s_{h_n}\left\{\hat{f}_{X-cZ}(x_i - c z_i) - \widehat{lb}(x_i, z_i)\right\}} \tag{4.16}$$

where

$$s_{h_n}\{x\} = 1\{x \leq 0\}0 + 1\{x > 0\}\left[-\exp\left(-\frac{x}{n_n}\right) + 1\right] \tag{4.17}$$

and $\hat{L}_3(a,c) = 0$ if $s_{h_n}\left\{\hat{f}_{X-cZ}(x_i - c z_i) - \widehat{lb}(x_i, z_i)\right\} = 0, \forall i \in H$. Note that $\lim_{n\to\infty} s_{h_n}\{x\} = 1\{x > 0\}$ when $h_n \to 0$ as $n \to \infty$.[7]

Figure (4.6) shows the SLS loss function with respect to c for the true value a_0 and for the optimal bandwidth that minimizes the Cross Validation function evaluated at the true value (a_0, c_0) which is denoted by h_0. Two different objective functions are shown: (i) objective function with a constant trimming of the 2% of the smallest densities, (ii) modified objective function $\hat{L}_2(c|a_0)$ in expression (4.14). Figure (4.7) shows in addition the behavior of objective function $\hat{L}_3(c|a_0)$ with respect to parameter c.

Figure (4.8) shows the same objective functions and the graph below represents the number of observations for which the estimated density of the index does not attain its lower bound $\widehat{lb}(x_i, z_i)$ for each value of parameter c. As c increases in absolute value, for more and more observations the estimated density of the index attains its lower bound. As it was pointed out before, it is also important to note that the global minimum for

[7] An alternative to this function s_{h_n} is

$$s_{h_n}(x) = F\left(\frac{x}{h_n}\right)$$

where F is a normal cdf. This is a continuously differentiable function that also converges to the indicator function when $h_n \to 0$ as $n \to \infty$. Using the modified $\hat{L}_3(a,c)$ with this definition of s_{h_n} does not help in the computation of the global minimum in finite samples though. The reason is that for those observations where the estimated density of the index attains its lower bound, this definition of function s_{h_n} gives them value $F(0)$ not value 0. Because of this, it can be checked that the shape of the objective function in this case does not help in avoid the local minimum for arbitrary large or small values of c. Thus, the s function defined in (4.17) is a continuous although nondifferentiable function which also converges to the indicator function as $h_n \to 0$ but it gives value 0 for those observations where $\hat{f}_{X-cZ}(x_i - c z_i) - \widehat{lb}(x_i, z_i) = 0$.

these three objective functions is located near the true value of the parameters. However, the behavior of the modified objective functions here helps in obtaining a solution to the minimization problem near the true value of the parameters. It avoids that the minimization of the SLS loss function gives solutions for parameter c close to the constrained set and far away from c_0 as it would be the case if gradient methods were used in the minimization of $\hat{L}(a, c)$.

It is important to point out that in this work we do not consider the optimal choice of the bandwidth. The large sample properties discussed below are derived for a fixed sequence of the bandwidth that satisfies the conditions to be established below. Therefore, the bandwidth is not considered as an additional parameter with respect to which the objective function needs to be minimized. Hardle, Hall and Ichimura (1993) show that for single index models, solving for the optimal value of the bandwidth jointly with the parameters of interest is optimal for both the estimation of $\{a_0, c_0\}$ and function ϕ_j. Whether this result still holds for objective function $\hat{L}_2(a, c)$ and $\hat{L}_3(a, c)$ where the denominator depends both on h_n and c is left for future work together with the simulations and estimations based on this joint optimization.

4.5 Identification

Let consider the identification of (a_0, c_0) in the model given in (4.1). The true value of the parameters are identified if

$$za_j + E(w_j - a_j z | x - cz) = za_{0j} + \varphi_j(x - c_0 z) \text{ a.s. in } x, z \text{ for } j = 1, ..., J$$

implies $a_j = a_{0j}$ for $j = 1, ..., J$ and $c = c_0$

Theorem 8 *If there is a set $S \subset \Omega_X \times \Omega_Z$ such that*

$$0 < E_{Z|X - c_0 Z}(Z | x - c_0 z) < 1 \tag{4.18}$$

for $(x, z) \in S$ and there is at least one $j^* \in \{1, ..., J\}$ such that φ_{j^*} is differentiable and satisfies the following condition

$$\alpha + \varphi_{j^*}(t + \beta) = \varphi_{j^*}(t) \text{ a.s. in } t \text{ if and only if } \alpha = 0 \text{ and } \beta = 0 \tag{4.19}$$

, then both parameters a_0 and c_0 are identified.

Proof. Consider an alternative value of the parameters $\{a, c\}$ such that

$$za_j + E(w_j - a_j z|x - cz) = za_{0j} + \varphi_j(x - c_0 z) \text{ a.s. in } x, z \text{ for } j = \{1, ..., J\} \quad (4.20)$$

and denote by $t = x - cz$ and $E(w_{j^*} - a_{j^*} z|t) = \psi_{aj^*}(t)$, so that the above expression for $j = j^*$ becomes

$$\psi_{aj^*}(t) = z(a_{0j^*} - a_{j^*}) + \varphi_{j^*}(t + (c - c_0)z) \text{ a.s. in } t, z$$

Consider the case where Z takes only two possible values: $Z = \{0, 1\}$ so that

$$\begin{aligned} \psi_{aj^*}(t) &= (a_{0j^*} - a_{j^*}) + \varphi_{j^*}(t + (c_0 - c)) \text{ a.s. at } t \text{ for } z = 1 \\ \psi_{aj^*}(t) &= \varphi_{j^*}(t) \text{ a.s. at } t \text{ for } z = 0 \end{aligned}$$

Then,

$$(a_{0j^*} - a_{j^*}) + \varphi_{j*}(t + (c_0 - c)) = \varphi_{j^*}(t) \text{ a.s. } t \text{ if and only if } a_{j^*} = a_{0j^*} \text{ and } c = c_0$$

so that $\{a_{0j^*}, c_0\}$ are identified if function φ_{j^*} satisfies the identification condition in (4.19). The rest of the parameters a_{0j} for $j \neq j^*$ are identified as follows. Since c_0 is identified, then (4.20) implies

$$z(a_j - a_{0j}) + E(a_{0j} z - a_j z|x - c_0 z) = 0 \text{ a.s. in } x, z \text{ for } j^* \neq j$$

If the identification condition in (4.18) holds, the above expression implies that $a_j = a_{0j}$ for $j \neq j^*$ ■

Note that the identification condition (4.19) does not hold if φ_{j^*} is linear in its argument, since the above equality implies $(a_{0j^*} + c_0) - (a_{j^*} + c) = 0$. Thus, when φ_{j^*} is linear both parameters a_{j^*} and c cannot be separately identified. An implication is that the parameters of the model are not identified if function φ_j is linear for all the equations.

Note that an equivalent condition to (4.19) is that

$$\varphi'_{j^*}(t + \beta) = \varphi'_{j^*}(t) \text{ a.s. } t \text{ if and only if } \beta = 0$$

so that the parameters are not identified if φ_j is a cyclical function for all j. These identification conditions are similar to those obtained by Chen, Blundell and Kristensen (2003).

4.6 Large Sample Properties of the Estimator

We use the following shorthands for the conditional expectations

$$m_{j,i}(.,a,c) = E(w_j - az|x_i - cz_i)$$

and the corresponding nonparametric estimators

$$\hat{m}_{j,i}(.,a,c) = \hat{E}_h(w_j - az|x_i - cz_i)$$

Also, $m_i(.,a,c) = [m_{i1}(.,a,c),...,m_{iJ}(.,a,c)]$ and $\hat{m}_i(.,a,c) = [\hat{m}_{i1}(.,a,c),...,\hat{m}_{iJ}(.,a,c)]$. The density function of the index and its corresponding nonparametric estimation is denoted by $f_{c,i}(.,c) = f_{X-cZ}(x_i - cz_i|c)$ and $\hat{f}_{c,i}(.,c) = \hat{f}_{h,X-cZ}(x_i - cz_i|c)$, respectively. Let denote by $f_i(.) = f_X(x_i)$ and $\hat{f}_i(.) = \hat{f}_{h,X}(x_i)$ the true density function of X at x_i and its corresponding nonparametric estimation. Analogously, the estimated lower bound for the estimated density of the index is denoted by $\widehat{lb}_i = \widehat{lb}(x_i, z_i)$ and its population counterpart by $lb_i = lb(x_i, z_i)$. The subscript c tries to differ between the density function of the index (which depends on c) and the density function of random variable X (which does not depend on c). The m_j functions are infinite dimensional parameters and they are real-valued functions that depend on data (X, Z) and on the finite dimensional parameters (a, c). The $J-$real-valued-function m is assumed to belong to a Banach space $\mathcal{M}$ defining a class of some smooth functions defined over the domain of function m. Function $f_c(.,c)$ is a real-valued function that depends on data (X, Z) and on the finite dimensional parameter c while function $f(.)$ depends only on data X. The arguments of m, f_c and f are sometimes omitted for simplicity.

The estimators obtained from (4.12) are shown to be consistent regardless of the correlation of errors ε_j across j. However, if this correlation exists, a more efficient estimator is obtained by taking it into account and giving different weights to the correlation between different equations in an individual specific matrix $\hat{V}_{in}$ which is estimated from the data of order $(J \times J)$ and it is estimated from the data. With the aim to gain some efficiency, we define the following M-estimator of (a_0, c_0) that minimizes loss function

$$\hat{L}(a,c) = \frac{1}{n}\sum_{i=1}^{n} I_{iQ} l(y_i, a, \hat{m}_{i,h_n}(.,a,c)) \tag{4.21}$$

where

$$l(y_i, a, \hat{m}_i(., a, c)) = \tag{4.22}$$

$$= B(y_i, a, \hat{m}_{i,h_n}(., a, c))' \hat{V}_{in} B(y_i, a, \hat{m}_{i,h_n}(., a, c))$$

$$\text{where } B(y_i, a, \hat{m}_{i,h_n}(., a, c)) = \begin{pmatrix} w_{i1} - z_i a_1 - \hat{E}_{1,h_n}(w_1 - z a_1 | x_i - c z_i) \\ \vdots \\ w_{iJ} - z_i a_J - \hat{E}_{J,h_n}(w_J - z a_J | x_i - c z_i) \end{pmatrix}$$

and $I_{iQ} = 1\{(x_i, z_i) \in Q\}$ for set defined as in (4.11).

Therefore, function B is a known, $J-$vector-real-valued function of Y and unknown parameters $(a, c, m(.)) \in \mathcal{A} \times \mathcal{C} \times \mathcal{M}$. For simplicity in the notation, we also omit the dependence of the nonparametric estimators of the bandwidth and this dependence is assumed when the functions are estimated.

For the modification of the SLS estimator proposed above in expressions (4.14) and (4.16), the function l should be respectively defined as follows

$$l_2(y_i, a, c, \hat{m}_i(., a, c), \hat{f}_c, \widehat{lb}) = \frac{B(y_i, a, \hat{m}_i(., a, c))' \hat{V}_{in} B(y_i, a, \hat{m}_i(., a, c))}{\frac{1}{n} \sum_{i \in H} 1\left\{ \hat{f}_{X - cZ}(x_i - c z_i) - \widehat{lb}(x_i, z_i) > 0 \right\}} \tag{4.23}$$

$$l_3(y_i, a, c, \hat{m}_i(., a, c), \hat{f}_c, \widehat{lb}) = \frac{B(y_i, a, \hat{m}_i(., a, c))' \hat{V}_{in} B(y_i, a, \hat{m}_i(., a, c))}{\frac{1}{n} \sum_{i \in H} s_{h_n}\left\{ \hat{f}_{X - cZ}(x_i - c z_i) - \widehat{lb}(x_i, z_i) \right\}} \tag{4.24}$$

where $\hat{f}_c = [\hat{f}_{c,1}(., c), .., \hat{f}_{c,n}(., c)]'$ and $\widehat{lb} = [\widehat{lb}_1(.), .., \widehat{lb}_n(.)]'$. In the Consistency and Asymptotic Normality Sections we derive the asymptotic properties of the estimator defined in (4.12) and the sections below (4.6.1) and (4.6.2) point out the additional conditions that need to be added to show the asymptotic properties of this alternative estimator.

4.6.1 Consistency

We expect the probability limit of the objective function in (4.21) to be $L(a, c)$

$$L(a, c) = E\left[I_{iQ} B(y, a, m_i(., a, c))' V_i B(y_i, a, m_i(., a, c)) \right]$$

where $p\lim \hat{V}_{in} = V_i$. To show the convergence of $\hat{L}(a,c)$ to $L(a,c)$ uniformly on (a,c), define the following function

$$L^*(a,c) =$$
$$\frac{1}{n}\sum_{i=1}^{n} I_{iQ} B(y,a,m_i(.,a,c))' V_i B(y_i,a,m_i(.,a,c))$$

Since $\hat{L}(a,c) - L(a,c) = \hat{L}(a,c) - L^*(a,c) + L^*(a,c) - L(a,c)$, the uniform convergence is shown from the following two results

$$\left|\hat{L}(a,c) - L^*(a,c)\right| \xrightarrow{p} 0 \text{ and } |L^*(a,c) - L(a,c)| \xrightarrow{p} 0$$

uniformly over $(a,c) \in \mathcal{A} \times \mathcal{C}$

Assumption 4. 1 *The observed sample* $y_i = \{w_i, z_i, x_i\}_{i=1}^n$ *are i.i.d. and its first* $r-$*moments,* $r \geq 2$, *exists and there does not exist linear dependence among the explanatory variables*

Assumption 4. 2 *The unknown functions* $\{\varphi_1(t), ..., \varphi_J(t)\}$ *are continuous*

Assumption 4. 3 *The parameter space* $(\mathcal{A} \times \mathcal{C})$ *is compact*

Assumption 4. 4 *The set* Q *defined in (4.11) is a compact subset of* $\Omega_X \times \Omega_Z$

Assumption 4. 5 *The vector of expectations* $m_j(.,a,c) = E(w_j - a_j z | x - cz)$ *for* $j = 1, .., J$ *are continuous functions of* $(x - cz)$

Assumption 4. 6 *For each* $c \in \mathcal{C}$, *the index* $X - cZ$ *has an absolutely continuous distribution such that its density function* $f_c(u,c)$ *is continuous in* u

Assumption 4. 7 *The kernel function* $K(u)$ *is continuous and* $K(s) = 0$ *if* $s < -1$ *and* $s > 1$; $\int K(u)du = 1; |K(u)|$ *is bounded,* $K(u)$ *is continuously differentiable and* $\left|\frac{\partial K(u)}{\partial u}\right|$ *is uniformly bounded*

Assumption 4. 8 *For each observation* i, $plim_{n\to\infty}\hat{V}_{in} = V_i$, *where* V_i *is a positive semi-definite matrix*

The assumptions 4.(1) - 4.(4) ensure that the conditions for the Uniform Law of Large Number are satisfied (see Lemma 2.4 in Newey and McFadden (1994)) in order to show that $|L^*(a,c) - L(a,c)| \xrightarrow{p} 0$ uniformly in (a,c).

On the other hand, regarding the second uniform convergence result, note that

$$\left|\hat{L}(a,c) - L^*(a,c)\right| \leq$$

$$\leq \left|\frac{1}{n}\sum_{i=1}^{n} I_{iQ}B(y_i,a,\hat{m}_i(.,a,c))'\hat{V}_{in}B(y_i,a,\hat{m}_i(.,a,c)) - I_{iQ}B(y_i,a,\hat{m}_i(.,a,c))'V_iB(y_i,a,\hat{m}_i(.,a,c))\right| + \tag{4.25}$$

$$+ \left|\frac{1}{n}\sum_{i=1}^{n} I_{iQ}B(y_i,a,\hat{m}_i(.,a,c))'V_iB(y_i,a,\hat{m}_i(.,a,c)) - I_{iQ}B(y_i,a,m_i(.,a,c))'V_iB(y_i,a,m_i(.,a,c))\right| \tag{4.26}$$

The uniform convergence in probability to zero of these three terms above is studied formally in the proof of the consistency theorem. In particular, the uniform convergence of term (4.26) requires the uniform convergence of the nonparametric conditional means, which is satisfied under the assumptions above.

Theorem 9 *Under assumptions 4.(1)-4.(8) and the assumptions of Theorem 8 (Identification Theorem), if the bandwidth satisfies the following condition*

$$\lim_{n\to\infty} \frac{n}{\ln n} h_n^{(1+2/r)} = \infty \tag{4.27}$$

.then the estimator defined by

$$(\hat{a},\hat{c}) = \arg \min_{(a,c)\in\mathcal{A}\times\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} I_{iQ}B(y_i,a,\hat{m}_{i,h_n}(.,a,c))'\hat{V}_{in}B(y_i,a,\hat{m}_{i,h_n}(.,a,c))$$

is a consistent estimator of (a_0,c_0).

Proof. See Appendix. ■

Additional conditions for consistency of the modified SLS estimators

In order to show the uniform converge in probability of the objective function $\hat{L}_2(a,c)$ and $\hat{L}_3(a,c)$ to the limiting objective function, the convergence in probability of $\left|\hat{L}_2(a,c) - \hat{L}(a,c)\right|$ and $\left|\hat{L}_3(a,c) - \hat{L}(a,c)\right|$ uniformly in (a,c) needs to be added to the original consistency

proof. The theorem in this section shows this under the following additional conditions. Let denote by n_1, n_0 and n_H the number of observations with $z = 1$, $z = 0$ and belonging to set H, respectively.

Assumption 4. 9 $h_n \to 0$ *as* $n \to \infty$

Assumption 4. 10 *Let* $p_0 = \lim_{n\to\infty} \frac{n_0}{n}$ *and* $p_1 = \lim_{n\to\infty} \frac{n_1}{n}$

$f(x|z)$ is uniformly continuous in the real line and $\int |f(x|z)|\, dx < \infty$ uniformly in z

Assumption 4. 11 *The characteristic function of the kernel function* K *is absolutely integrable*

Assumption 4. 12 $lim_{n\to\infty} \frac{n_h}{n} = 1$

The following two lemmas are useful before showing the consistency of the alternative estimators.

Lemma 8 (Uniform Convergence in probability of the indicator function) *Under assumptions 4.(1)-4.(7) and 4.(9)-4.(12),*

$$\lim_{n\to\infty} \Pr\left(\sup_{i\in H, c\in\mathcal{C}} \left| 1\left\{ \hat{f}_{i,c}(.,c) - \widehat{lb}_i > 0 \right\} - 1\left\{ f_{i,c}(.,c) - lb_i > 0 \right\} \right| < \varepsilon \right) = 1$$

Proof. See Appendix ■

Lemma 9 (Uniform Convergence in probability of function s**)** *Under assumptions 4.(1)-4.(7) and 4.(9)-4,(12),*

$$\lim_{n\to\infty} \Pr\left(\sup_{i\in H, c\in\mathcal{C}} \left| s_{h_n}(\hat{f}_{i,c}(.,c) - \widehat{lb}_i) - 1 \right| < \varepsilon \right) = 1$$

Proof. See Appendix ■

The two lemmas above use the result of the uniform convergence in probability of the nonparametric estimators of the conditional mean and the uniform convergence of $\hat{f}_X(x|Z = 0)$ and $\hat{f}_X(x|Z = 1)$ to the true densities uniformly in x. The assumptions 4.(9)-4.(11) together with the following bandwidth conditions

$$n_0 h_n^2 \to \infty \text{ and } n_1 h_n^2 \to \infty \text{ as } n \to \infty$$

- which are indeed implied by condition in expression (4.27) and assumption 4.(10)- guarantee that this is the case (see Pagan and Ullah (1999)).

Theorem 10 (Consistency) *Under assumptions 4.(1)-4.(12) and the assumptions of the identification theorem (8), if the bandwidth satisfies the following condition*

$$\lim_{n\to\infty} \frac{n}{\ln n} h_n^{(1+2/r)} = \infty \tag{4.28}$$

,then the estimators defined by

$$\begin{aligned}(\hat{a}_2, \hat{c}_2) &= \arg \min_{(a,c)\in\mathcal{A}\times\mathcal{C}} \hat{L}_2(a,c) \\ (\hat{a}_3, \hat{c}_3) &= \arg \min_{(a,c)\in\mathcal{A}\times\mathcal{C}} \hat{L}_3(a,c)\end{aligned}$$

are consistent estimators of (a_0, c_0).

Proof. See Appendix ■

Note that the crucial uniform result in Lemma (8) does not hold if it is defined over a different set (not a subset) of H. This is because the uniform convergence of $|1\{f_{i,c}(.,c) - lb_i > 0\} - 1|$ on $c \in \mathcal{C}$ and on i might not hold if we do not restrict the set for the observations to set H, where the indicator equals one uniformly on c, by definition of this set. This is the reason why in the definition of the objective function $\hat{L}_2(a,c)$ and $\hat{L}_3(a,c)$ the sum of the corrected denominator is defined over set H.

4.6.2 Asymptotic Normality

The estimator that minimizes the objective function in $\hat{L}(a,c)$ solves the following system of equations

$$\frac{1}{n}\sum_{i=1}^{n} I_{iQ}\left[\nabla_{(a,c)} B(y_i, \hat{a}, \hat{m}_i(.,\hat{a},\hat{c}))\right]' \hat{V}_{in} B(y_i, \hat{a}, \hat{m}_i(.,\hat{a},\hat{c})) = 0$$

where

$$\nabla_{(a,c)} B(y_i, \hat{a}, \hat{m}_i(.,\hat{a},\hat{c})) = \left[D\left(-z_i - \frac{\partial \hat{m}_{ji}(\hat{a}_j,\hat{c})}{\partial a_j}\right), -\frac{\partial \hat{m}_i(\hat{a},\hat{c})}{\partial c}\right]$$

where $D(-z - \frac{\partial \hat{m}_j(\hat{a}_j,\hat{c})}{\partial a_j})$ is a matrix of order $(J \times J)$ where the only elements different from zero are the elements in the diagonal and the element $(j \times j)$ in the diagonal takes value $-z - \frac{\partial \hat{m}_j(\hat{a}_j,\hat{c})}{\partial a_j}$. Since by the assumptions above B is a differentiable function of the

parameters, doing the usual Taylor's series expansion of B around (a_0, c_0) we obtain

$$\sqrt{n}\begin{pmatrix} \hat{a} - a_0 \\ \hat{c} - c_0 \end{pmatrix} =$$

$$-\left[\frac{1}{n}\sum_{i=1}^{n} I_{iQ}\nabla_{(a,c)}B(y_i, \hat{a}, \hat{m}_i(.,\hat{a},\hat{c}))'\hat{V}_{in}\nabla_{(a,c)}B(y_i, \bar{a}, \hat{m}_i(.,\bar{a},\bar{c}))\right]^{-1} \times \quad (4.29)$$

$$\times\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\nabla_{(a,c)}B(y_i, \hat{a}, \hat{m}_i(.,\hat{a},\hat{c}))'\hat{V}_{in}B(y_i, a_0, \hat{m}_i(.,a_0,c_0))\right] \quad (4.30)$$

Two main results allow one to obtain the asymptotic distribution of $(\hat{a}, \hat{c})$: (i) the convergence in probability of the hessian term (4.29) to a positive definite matrix and (ii) the convergence in distribution of term (4.30) to a normal distribution.

Additionally to the assumptions needed for consistency, the following assumptions should be satisfied:

Assumption 4. 13 *The unknown functions* $\{\varphi_1(t), ..., \varphi_J(t)\}$ *are continuously differentiable of order* $q+1$ *where* $q > 2$

Assumption 4. 14 *The functions* $f_c(u, c_0)$ *and* $E(z|u)$ *are continuously differentiable in* u *of order* $q+1$

Assumption 4. 15 *The conditional expectation functions* $m_j(.,a,c) = E(w_j - a_j z|x - cz)$ *for* $j = 1, .., J$ *and* $E(z|x - cz)$ *are continuously differentiable in* $(x - cz)$

Assumption 4. 16 *The density function* $f_c(u, c)$ *is continuously differentiable in* u

Assumption 4. 17 $K(u)$ *is twice continuously differentiable and satisfies*

$$\int_{-\infty}^{+\infty} u^s K(u)du = 0$$

for $s = \{2, ..., q-1\}$

Theorem 11 (Asymptotic Normality) *Under assumptions 4.(1)-4.(17), if the bandwidth sequence satisfies the following conditions*

$$\begin{aligned} \lim_{n\to\infty} \frac{n}{\ln n} h_n^{2(1+2/r)+1} &= \infty \quad (4.31) \\ \lim_{n\to\infty} \sqrt{n} h_n^q &= 0 \\ \lim_{n\to\infty} \sqrt{n} h_n^2 &= \infty \end{aligned}$$

Then,the Hessian term in (4.29) converges to a positive definite matrix H

$$\frac{1}{n}\sum_{i=1}^{n} I_{iQ}\left[\nabla_{(a,c)}B(y_i,\hat{a},\hat{m}_i(.,\hat{a},\hat{c}))\right]'\hat{V}_{in}\left[\nabla_{(a,c)}B(y_i,\bar{a},\hat{m}_i(.,\bar{a},\bar{c}))\right] \xrightarrow{p} H$$

where

$$H = E\left[I_Q\Delta(y)'V\Delta(y)\right]$$

and

$$\Delta(y) = \left[D\left(-z - \frac{\partial m_j(a_{j0},c_0)}{\partial a_j}\right)\ ,\ -\frac{\partial m(a_0,c_0)}{\partial c}\right]$$

where $D(x)$ is a diagonal matrix where the elements in the diagonal are the elements in vector x, the score term in (4.30) converges in distribution to a normal random variable as follows

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\nabla_{(a,c)}B(y_i,\hat{a},\hat{m}_i(.,\hat{a},\hat{c}))'\hat{V}_{in}B(y_i,a_0,\hat{m}_i(.,a_0,c_0)) \xrightarrow{d} N(0,\Sigma)$$

where

$$\Sigma = E\left\{I_Q\left[\Delta(y)\right]'V\varepsilon\varepsilon'V\left[\Delta(y)\right]\right\}$$

so that the asymptotic distribution of the SLS estimator is

$$\sqrt{n}\begin{pmatrix}\hat{a}-a_0\\ \hat{c}-c_0\end{pmatrix} \xrightarrow{p} N(0, H^{-1}\Sigma H^{-1})$$

Proof. See Appendix. ■

Equivalent asymptotic distribution for the alternative estimators

In this section we show that the asymptotic distribution of the estimators $(\hat{a}_2,\hat{c}_2)$ and $(\hat{a}_3,\hat{c}_3)$ defined through the minimization of objective functions $\hat{L}_2(a,c)$ and $\hat{L}_3(a,c)$ is the same as the estimator $(\hat{a},\hat{c})$ defined through the original objective function $\hat{L}(a,c)$. Formally, we show that

$$\lim_{n\to\infty}\Pr\left(\left\|\sqrt{n}\left((\hat{a}_2,\hat{c}_2)-(\hat{a},\hat{c})\right)\right\|>\varepsilon_2\right) = 0 \tag{4.32}$$

$$\lim_{n\to\infty}\Pr\left(\left\|\sqrt{n}\left((\hat{a}_3,\hat{c}_3)-(\hat{a},\hat{c})\right)\right\|>\varepsilon_3\right) = 0 \tag{4.33}$$

for any $\varepsilon_2>0$ and $\varepsilon_3>0$. The probability in (4.32) can be rewritten as

$$\Pr\left(\begin{array}{c}\|\sqrt{n}\left((\hat{a}_2,\hat{c}_2)-(\hat{a},\hat{c})\right)\|>\varepsilon_2,\\ 1\left\{\hat{f}_{i,c}(.,c)-\widehat{lb}_i>0\right\}=1\left\{f_{i,c}(.,c)-lb_i>0\right\}\ \forall i\in H,\ \forall c\in\mathcal{C}\end{array}\right)+$$

$$+\Pr\left(\begin{array}{c}\|\sqrt{n}\left((\hat{a}_2,\hat{c}_2)-(\hat{a},\hat{c})\right)\|>\varepsilon_2,\\ 1\left\{\hat{f}_{i,c}(.,c)-\widehat{lb}_i>0\right\}\neq 1\left\{f_{i,c}(.,c)-lb_i>0\right\}\text{ for some } i\in H \text{ or some } c\in\mathcal{C}\end{array}\right)$$

The second probability in the expression above converges to zero by Lemma (8). Regarding the first probability, note that if the indicator function at the estimated density and lower bound equals one for all the observations in H and uniformly on $c \in \mathcal{C}$, then the objective function

$$\hat{L}_2(a,c) = \frac{n}{n_H}\hat{L}(a,c)$$

so that both estimators are equal in this case and also the first probability converges to zero.

A similar reasoning can be done to show the converges to zero of the probability in (4.33). In this case, Lemma (9) can be used in the same way to show that

$$\lim_{n\to\infty} \Pr\left(\left\|\sqrt{n}\left((\hat{a}_3,\hat{c}_3)-(\hat{a},\hat{c})\right)\right\| > \varepsilon_3\right) =$$

$$\lim_{n\to\infty} \Pr\left(\begin{array}{c} \left\|\sqrt{n}\left((\hat{a}_3,\hat{c}_3)-(\hat{a},\hat{c})\right)\right\| > \varepsilon_2, \\ s_{h_n}\left(\hat{f}_{i,c}(.,c)-\widehat{lb}_i\right) = 1\left\{f_{i,c}(.,c)-lb_i>0\right\} \ \forall i \in H, \ \forall c \in \mathcal{C} \end{array}\right) = 0$$

4.6.3 Optimal weighting matrix

The asymptotic variance of the SLS estimator defined in this paper depends on the limit V_i of the weighting matrix $\hat{V}_i$. In this section, we discuss the choice of this weighting matrix in order to find the efficient estimator in the class of estimators defined by the minimization of the sample analogue of objective function (4.8). When

$$V = E\left(\varepsilon\varepsilon'|y\right)^{-1}$$

then, the asymptotic variance of the estimator equals

$$E\left\{I_Q\left[\Delta(y)\right]' E\left(\varepsilon\varepsilon'|y\right)^{-1}\left[\Delta(y)\right]\right\}^{-1}$$

It can be then shown that if the $(J \times J)$ matrix $E\left(\varepsilon\varepsilon'|y\right)$ is not singular, then the estimator with $V_i = p\lim\left(\hat{V}_i\right) = E\left(\varepsilon\varepsilon'|y_i\right)^{-1}$ is asymptotically efficient for this class of estimators. Equivalently to the optimal minimum distance estimation, we need to show that

$$\begin{array}{l} \left(E\left\{I_Q\Delta(y)'V\Delta(y)\right\}\right)^{-1} E\left\{I_Q\Delta(y)'V\varepsilon\varepsilon'V\Delta(y)\right\}\left(E\left\{I_Q\Delta(y)'V\Delta(y)\right\}\right)^{-1} - \\ -\left(E\left\{I_Q\Delta(y)'E\left(\varepsilon\varepsilon'|y\right)^{-1}\Delta(y)\right\}\right)^{-1} \end{array}$$

is positive semi-definite for all positive semi-definite matrices V. Let define $s = I_Q\Delta(y)'E\left(\varepsilon\varepsilon'|y\right)^{-1}\varepsilon$ and $t = I_Q\Delta(y)'V\varepsilon$, then the above difference between matrices can be expressed as

$$
\begin{aligned}
& \left(E\left\{ts'\right\}\right)^{-1}\left(E\left\{tt'\right\}\right)\left(E\left\{st'\right\}\right)^{-1}-\left(E\left\{ss'\right\}\right)^{-1} \\
= \; & \left(E\left\{I_Q\Delta(y)'V\Delta(y)\right\}\right)^{-1}\times E\left[UU'\right]\times\left(E\left\{I_Q\Delta(y)'V\Delta(y)\right\}\right)^{-1}
\end{aligned}
$$

with $U = t - E\{ts'\}\left(E\{ss'\}\right)^{-1}s$. The above difference is positive semi-definite for all V since $E\left[UU'\right]$ is positive semi-definite.

4.6.4 Estimation of the Covariance Matrix

The asymptotic covariance matrix of the estimator can be consistently estimated by estimating both H and Σ as follows,

$$
\hat{H} = \frac{1}{n}\sum_{i=1}^{n} I_{iQ}\hat{\Delta}(y_i)'\hat{V}_i\hat{\Delta}(y_i) \tag{4.34}
$$

$$
\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} I'_{iQ}\hat{\Delta}(y_i)\hat{V}_i\hat{\varepsilon}_i\hat{\varepsilon}'_i\hat{V}_i\hat{\Delta}(y_i) \tag{4.35}
$$

with

$$
\begin{aligned}
\hat{\Delta}(y_i) &= \left[D\left(-z_i-\hat{E}\left(z|x_i-\hat{c}z_i\right)\right),\; -\frac{\partial\hat{E}\left(w-\hat{a}z|x_i-\hat{c}z_i\right)}{\partial c}\right] \\
\hat{\varepsilon}_i &= \begin{bmatrix} w_{i1}-\hat{a}_1 z_i-\hat{E}\left(w_1-\hat{a}_1 z|x_i-\hat{c}z_i\right) \\ \vdots \\ w_{iJ}-\hat{a}_J z_i-\hat{E}\left(w_J-\hat{a}_J z|x_i-\hat{c}z_i\right) \end{bmatrix}
\end{aligned}
$$

The asymptotically efficient estimator discussed in Section (4.6.3) can be constructed using a consistent estimator of $E\left(\varepsilon\varepsilon'|y\right)^{-1}$. The consistent estimation of this weighting matrix can be obtained from a first step consistent estimation $(\hat{a}^{(0)},\hat{c}^{(0)})$ -using for example $\hat{V}_i = I_J$ where I_J is the identity matrix of order $(J\times J)$– which is then used to construct a consistent estimator of the optimal weighting matrix $\hat{V}_i^*$

$$
\hat{V}_i^* = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^{(0)}\hat{\varepsilon}_i^{(0)\prime}\right)^{-1} \tag{4.36}
$$

where

$$
\hat{\varepsilon}_i^{(0)} = \begin{bmatrix} w_{i1}-\hat{a}_1^{(0)}z_i-\hat{E}\left(w_1-\hat{a}_1^{(0)}z|x_i-\hat{c}^{(0)}z_i\right) \\ \vdots \\ w_{iJ}-\hat{a}_J^{(0)}z_i-\hat{E}\left(w_J-\hat{a}_J^{(0)}z|x_i-\hat{c}^{(0)}z_i\right) \end{bmatrix}
$$

4.7 Monte Carlo Simulations

In this section we compare the performance of the estimators discussed in this work in finite samples. We simulate a model with only one good ($J = 1$) with 200 observations where the unknown function is

$$\varphi_1(x) = -0.7 + 1.4x - 0.14x^2$$

and the endogenous variable w_1 is generated as

$$w_1 = a_{01}z + \varphi_1(x + c_0 z) + e$$

with parameter values $a_0 = 0.3$ and $c_0 = -0.3$. The error e is normally distributed with mean 0 and standard deviation $\sigma_e = 0.2$. 60% of the sample is generated so that it belongs to the demographic group $Z = 1$. The distribution for random variable X is different for each demographic group. The random variable $X|Z = 0$ is randomly drawn from a normal distribution $N(4, \sigma_0^2); \sigma_0 = 1.5$ and $X|Z = 1$ is randomly drawn from a normal distribution $N(4.5, \sigma_1^2); \sigma_1 = 1$. The function $\varphi_1(.)$ satisfies the identification conditions.

The Montecarlo results with 300 trials for the different SLS introduced in this paper are presented here. The value of the bandwidth is not considered as an additional parameter and the results reported here are conditioned on the value of the bandwidth. The minimization of the objective function is done with respect to parameters (a, c) conditioned on that value of the bandwidth -denoted by $\hat{h}_0-$ that minimizes the objective function evaluated at the true value of the parameters $\hat{L}(h|a_0, c_0)$. The results for the estimators minimizing respectively $\hat{L}(a, c)$ (trimming 2% of those observations with the smallest values of $f_{X-cZ}(x-cz)$), $\hat{L}_2(a, c)$ and $\hat{L}_3(a, c)$ for three different starting values are presented in Tables (4.1) and (4.2). For each of these three exercises, the same starting value was used for each montecarlo replication. The first two objective functions are minimized subject to $\sum_{i=1}^{n} 1\left\{\hat{f}_{i,c}(.,c) - \widehat{lb}_i > 0\right\} > 0$ and the third objective function is minimized subject to $\sum_{i=1}^{n} s_{h_n}\left(\hat{f}_{i,c}(.,c) - \widehat{lb}_i\right) > 0$. Thus, for $\hat{L}(a, c)$ the solution is constrained to lie outside the set of the parameter space where the objective function is flat with respect to c. For objective functions $\hat{L}_2(a, c)$ and $\hat{L}_3(a, c)$, the solution is constrained to lie outside that set of parameter values where the objective function is exactly equal to zero

Different starting values deliver different results for each of the estimators. The first set of results corresponding to the estimator that minimizes the original objective function $\hat{L}(a, c)$ are notably sensitive to the starting value. Conditioned on c, the objective function

with respect to parameter a is globally concave, but the objective function with respect to c has local minima at arbitrary large or small values of c. As Figure (4.5) illustrates, the starting value in this case determines the local minima given as a solution if gradient optimization methods are used instead of grid search methods. For starting values $[2, 2]$ and $[-1, 1]$ -which are far away from (a_0, c_0)- this first estimator yields estimates with high bias, especially for parameter c. The bias of both parameters is substantially reduced as the starting value is closer to (a_0, c_0). The standard deviation over replications makes that the MSE is still high for parameter c even if the starting value is relatively close to the true value. A starting value that it is close enough to the true value ensures that the global minima is achieved by the optimization method. Thus, the starting value $[0.7, -0.7]$ yields relatively good results and the smallest MSE among the results with three different starting values.

Regarding the results for estimator defined through the minimization of $\hat{L}_2(a, c)$, if the starting value is further away from the true value the estimator still delivers a distribution of the estimated parameter c with high MSE. Although the corrected objective function $\hat{L}_2(c|a_0)$ in Figure (4.6) is globally concave with respect to c when it is evaluated at a_0, the behavior of this objective functions is not as nice when it is evaluated at a different value of a. Figure (4.9) shows the objective function with respect to parameters (a, c) conditioned on $\hat{h}_0$. As one moves a away from the true value, the objective function changes and it is not globally concave as when it is evaluated at a_0. This means that the starting value for a is also important in this case in order to obtain unbiased results. Comparing this results with the previous estimator in the same table, the results are better in the MSE sense even for the starting value that is further away. This second estimator $(\hat{a}_2, \hat{c}_2)$ performs well for the last two starting values and the MSE is very close to zero when the starting value is very close to (a_0, c_0).

The results for the SLS estimator that minimizes $\hat{L}_3(a, c)$ where the original objective function is divided by a continuous approximation of the indicator function are presented in Table (4.2). The distribution of the estimates over replications are quite robust to the different starting values in this case and the MSE of both parameters are very small. This is the estimator that works better in the simulations corresponding to one good.

Both the third estimator (where the original objective function is divided by s) and the second estimator (where the original objective function is divided by the indicator function) perform well even if the starting value is not placed close to the true values. The reason why the third estimator appears to be more robust to the different starting values

is that the objective function with respect to c is concave even if a is not close to the true value of a_0 as Figure (4.10) illustrates.

Tables (4.3) and (4.4) report the montecarlo results for the three previous estimators in the literature for shape invariant models: Pinkse-Robinson (1995) (PR), Hardle-Marron (1990) (HM) and the modification proposed by Wilke (2003). As expected from the shape of the objective functions with respect to c for Pinkse-Robinson and Hardle-Marron estimators, the solution is many times found at values of the parameters far away from the true values. For these two estimators, the best results based on the MSE over replications are found when the montecarlo experiment is designed so that the starting value is close to (a_0, c_0). The rest of the results show that if the starting point is not carefully chosen, this might result in high biased results. The estimator proposed by Wilke performs much better regardless of the starting value and additionally delivers smaller MSE values for the estimates than Pinkse-Robinson and Hardle and Marron estimators. The mean over replications for the estimator proposed by Wilke is paradoxically worse when the starting value is closer to the true value. This is due to isolated solutions for some replications that are away from the true value, which makes that the distribution of the parameter estimates is skewed. However, the median over replications is very close to (a_0, c_0).

We also consider the multiple equation case ($J = 2$). The specific parametric conditional mean functions φ are

$$\varphi_1(x) = -0.7 + 1.4x - 0.14x^2 \text{ and } \varphi_2(x) = 2 + \log(x)$$

and the dependent variable for each equation is generated with parameter values $[a_{01}, a_{02}, c_0] = [1, -2.5, 0.5]$. Here we present results for the case where the errors of each equation e_1 and e_2 are not independent. In this exercise, $e_1 \sim N(0, \sigma_1^2)$, $\sigma_1 = 0.2$ and $e_2 \sim N(0, \sigma_2^2)$, $\sigma_2 = 0.5$ with correlation coefficient $\rho_{12} = 0.6$. Given that for the single equation case we have concluded that the estimators defined through the minimization of objective functions $\hat{L}_2$ and $\hat{L}_3$ perform better than when the original objective function is used, we only present here results for these two estimators and for the estimators introduced previously in the literature. For the SLS estimators, the optimal weighting function $\hat{V}^*$ is used, which is estimated as described in (4.36) from a first step consistent estimators $[\hat{a}_1^{(0)}, \hat{a}_2^{(0)}, \hat{c}^{(0)}]$ where $W = I_J$ (identity matrix of order J) is used as a weighting matrix. The first step results for the modified SLS estimators are very similar to the second step results with slightly smaller standard deviation over the replications. We omit these first step results

here. The results for multiple equations ($J = 2$)[8] are presented in Table (4.5) and Table (4.6) for different starting points. The first set of results corresponds to a starting value which is not so close to the true parameters. For this reason, the results from the PR and HM estimators deliver quite high MSE since they must have found a solution in a global minima away from the true value. Their performance is significantly worse for parameter c and in the case of HM even for parameter a_2. Among the rest of parameters, the modification proposed by Wilke of the previous estimators produces very good results with MSE quite close to zero. Regarding the SLS estimators, they also perform better in this case than PR and HM estimators with the gradient computational methods used in the optimization. The MSE is quite low for parameters $a's$ and it is slightly larger for parameter c. The difference in the performance of the estimators obtained through the modification of the original objective function for SLS with respect to the estimator obtained through the modification of PR or HM objective function (Wilke) is due to the different curvature of these objective functions around the true value of the parameters. Thus, looking at the shape of both functions, the slope of the modified objective functions $\hat{L}_2$ and $\hat{L}_3$ with respect to parameter c is much more constant around (a_0, c_0) than the modified objective function proposed by Wilke. This explains the difference in accuracy in the estimates of $\hat{c}$ in the results presented for these three estimators. When the starting value is designed to be closer to the true value of the parameters (See Table (4.6)), the PR and HM results improve in the MSE sense though in the case of the PR estimator the bias corresponding to parameter c is still quite high. In this set of estimates, the SLS estimators provide a mean over replications that it is closer to the true value of the parameters while the standard errors are only slightly smaller and still higher than the dispersion over replications of the estimates provided by the estimator proposed by Wilke. Between the two alternative modifications of SLS proposed in this paper, it should be pointed out that better estimates are obtained from the minimization of $\hat{L}_3(a, c)$. The mean over replications is closer to the true value of the parameters and the variance of the estimates over replications is also smaller. Additionally, this estimator turns out to be more robust to the starting value used.

Therefore, in general, these simulation exercises suggest that when standard gradient methods are used in the optimization the right choice of the starting value is quite important for the results of the estimators previously suggested in the literature to estimate

[8] I should point out that qualititavely similar results were obtained in a simulation experiment with 3 equations.

shape invariant models, and also for the SLS estimator in its original formulation. If the starting value is not close to the true value, these computational methods deliver as valid solutions minima that are far away of the local or global minimum that is close to the true value of the parameters. Obviously, other types of optimization could be used if this property of the objective function is know a priori, such as grid search methods. However, these methods are computationally costly if parameter c is multidimensional, for example when the differences between nonparametric regression curves with respect to more than one variable want to be studied.

4.8 Empirical Application

This section applies the SLS estimation of shape invariant models to the estimation of Engel curves for different demographic groups using British Consumer data. To do so, we use the same data as in Blundell, Duncan and Pendakur (1998). Thus we use cross sections 1980-1982 of the British Family Expenditure Survey (FES). Only households with one child ($Z = 0$) or two children ($Z = 1$) and with married and cohabiting couples where the head of the household is employed are selected in this data. The selected sample is then homogeneous with respect to other demographic variables. It contains six categories of goods (food, domestic fuel, clothing, alcohol, transport and other goods). See Blundell, Duncan and Pendakur (1998) for more details on the selection of the sample. There is a total of 1519 observations where 594 of them belong to demographic group 0 (one child) and 925 of them belong to demographic group 1 (two children).

We use the alcohol budget shares to compare the estimators discussed in this paper in a single equation model. These results are shown in Table (4.7). Different starting values were used to guarantee the robustness of the results presented here and the solution that provided the minimum value function was selected. No grid search methods were used for parameter c. As the montecarlo experiments suggested, the estimates $(\hat{a}_3, \hat{c}_3)$ are more stable to different starting values than the estimates we obtained for $(\hat{a}_2, \hat{c}_2)$, but we find similar results for both of them after trying different starting values. We also present here for comparison the results of the estimator that minimizes the original SLS objective function $\hat{L}(a, c)$ with 2% of observations trimmed. The same set of results were obtained for different values of the bandwidth ($h = \{1, 0.5, 0.25, 0.1\}$) and the estimates attaining the smallest value function are presented here. For each estimator, the value of the bandwidth at which we find the minimum of the objective function is shown in the row

named by $\hat{h}$. The covariance matrix is estimated as described in (4.6.4). The covariance matrix estimation for HM and Wilke estimators is done using the asymptotic distribution computed in the corresponding papers. [9]

It can be checked that the estimates of c_0 are very different for each estimator considered. The values of the estimates for c are especially different for the estimates obtained through the minimization of the original SLS objective function (column (1)) and PR and HM estimators which , as we discuss in the montecarlo section, are likely to have achieved a local minima far away from the true value of the parameter. The sign of $\hat{c}$ in column (1) and in PR gives us also a hint that this might be the case. This is because c is interpreted as the equivalence scale between households with two children and households with one child. A negative value of this parameter does not have an economic interpretation given the two demographic groups under consideration. Given the relationship $e^{(1)}(x) = a + e^{(0)}(x - c)$, a positive value of c denotes the amount of additional total expenditure that should be

[9] Applying the asymptotic result of Hardle and Marron to the linear case we consider here, it can be shown that their estimator is asymptotically normally distributed with zero mean and variance given by V

$$V = 4\left[E\left(\left(w_0 - e^{(0)}(x)\right)^2\right) + E\left(\left(w_1 - e^{(1)}(x)\right)^2\right)\right]\int \begin{bmatrix} 1 & \partial e^{(0)}(x-c_0)/\partial c \\ \partial e^{(0)}(x-c_0)/\partial c & \left(\partial e^{(0)}(x-c_0)/\partial c\right)^2 \end{bmatrix} w(x)dx$$

which is estimated using the corresponding nonparametric estimates of $e^{(1)}, e^{(0)}$ and the estimate for c_0.

Wilke (2003) provides the following asymptotic distribution for the modified estimator he proposes in his work

$$\sqrt{n}\begin{pmatrix} \hat{a} - a_0 \\ \hat{c} - c_0 \end{pmatrix} \to N\left(0, H^{-1}VH^{-1}\right)$$

with

$$H = E\left(\begin{bmatrix} 1 & \partial e^{(0)}(x-c_0)/\partial c \\ \partial e^{(0)}(x-c_0)/\partial c & \left(\partial e^{(0)}(x-c_0)/\partial c\right)^2 \end{bmatrix} \middle| x \in \Omega_X^{(0)} + c \cap \Omega_X^{(1)} \right)$$

$$V = E\left(\left(\frac{(w_1 - e^{(1)}(x))^2}{f_1(x)^2} + \frac{(w_0 - e^{(0)}(x-c_0))^2}{f_0(x)^2}\right) \begin{bmatrix} 1 & \partial e^{(0)}(x-c_0)/\partial c \\ \partial e^{(0)}(x-c_0)/\partial c & \left(\partial e^{(0)}(x-c_0)/\partial c\right)^2 \end{bmatrix} \middle| x \in \Omega_X^{(0)} + c \cap \Omega_X^{(1} \right.$$

An estimate of this matrix of variance and covariance is obtained by replacing the conditional mean and density functions by its nonparametric estimation. Note that although the estimator is defined with the integral over the intersection of the supports $x \in \Omega_X^{(0)} + \hat{c} \cap \Omega_X^{(1)}$, the above matrices are computed with those observations in the data belonging to the intersection of $x \in \Omega_X^{(0)} + \hat{c} \cap \Omega_X^{(1)}$. This implies some inconsistency of the standard errors provided with respect to the estimation method used. Although we have computed Wilke's modified estimator by simulating the integral over x, he implements his estimator by minimimg the sum of square losses over the observations instead of computing the integral.

given to a family with two kids in order to have the same budget share on alcohol as the households in the reference demographic group (households with one kid).

Almost all the estimators (with the exception of HM) yield negative estimates of $a_{alcohol}$ which implies that for a given value of total expenditure, households with two kids devote a smaller budget share to alcohol than families with one kid.

Given the evidence obtained from the Montecarlo experiments, the estimates to be viewed as more reliable are those in columns (2), (3) and Wilke. The estimated standard errors are relatively small in these three cases, especially for the SLS estimates. Given that these three estimators give estimates that are almost statistically different, some goodness of fit criteria allowing one to choose among these three different estimates is left for future versions of this work.

Table (4.8) reports the estimates from the SLS estimator with multiple equations that minimizes the objective function $\hat{L}_2(a, c)$ which divides the original objective function by the number of observations for which the estimated density is above its lower bound. Four results are presented for different values of the bandwidth $h = \{0.1, 0.25, 0.5, 1\}$ and different starting values were used. For each value of the bandwidth, the solution yielding the smallest value of the objective function is reported. It is important to note that the estimates are not robust to changes in the value of the bandwidth. In many of the cases, especially for the estimates corresponding to the linear part (parameters a), there are even changes in the sign of the estimates we obtain. This implies that the optimal choice of the smoothing parameters turns out to be very important for this particular estimator. The results corresponding to $h = 0.5$ attain the smallest value of the objective function. Comparing these results with the ones reported in Blundell, Duncan and Pendakur (1998) (for reference they can be found in Table 4.10)) corresponding to the Pinkse-Robinson estimator[10], it can be checked that the estimates for parameter a are very similar for all the goods although the estimates for parameter c are significantly different in both cases. The SLS estimator yields an estimate of $\hat{c} = 0.0558$ while the estimate given by PR estimator is $\hat{c} = 0.2590$. The estimated standard errors for the SLS estimates lead us to conclude that for $h = 0.5$, the estimates of parameter c are not statistically significantly different from zero. The estimate of c presented in Blundell, Duncan and Pendakur (1998) is more precisely estimated [11]. This difference is even bigger if one compares the results

[10] In order to obtain these results the authors describe that the optimisation method involves gridsearch over parameter c over a reasonable set of vlaues.

[11] Since grid search is used for the computation of the estimate of c, the bootstrap standard errors for c

with the ones reported in Wilke (2003). More similar results with respect to parameter c ($\hat{c} = 0.2010$) can be found when the smoothing parameter is set to $h = 0.1$ (for which the value function at the estimates is slightly higher than for the global minimum) and the estimated standard error for parameter c is also smaller in this case. Thus, the conclusions from the results of this estimator is that the estimates (and their corresponding standard errors) are not robust across different values of the bandwidth and the optimal choice here seems important. The estimated c for this version of SLS is smaller than the previous estimates that can be found in the literature for both the global minimum found when $h = 0.5$ and for estimated values close to the global minimum when $h = 0.1$.

Table (4.9) presents the results from the SLS estimator that minimizes the objective function $\hat{L}_3(a, c)$, which divides the original objective function by the sum of continuos functions s converging to the indicator function as the sample size increases. In contrast to the previous SLS estimator, the results in this case are much more stable across different values of the smoothing parameter h. The objective function attains its minimum at the estimated parameters when $h = 0.1$. The estimates were also very robust across the different starting values. When comparing these results with PR and the modified estimator proposed by Wilke, one can check that the estimate of c is smaller for the SLS than for the previous two estimators. The standard error for c in our case is smaller than the bootstrapped standard errors given by Blundell, Duncan and Pendakur (1998), however the parameter is estimated more precisely by Wilke. Note though that there exits a trade off in the standard errors he reports for parameters a and c. While the parameters in part a are much more precisely estimated by the proposed estimator in this work, his estimation of c is very precise.

The similarities of the estimates for a of this version of SLS with respect to PR and Wilke differ across the goods. The coefficients belonging to the linear part of the specification have the expected sign given the definition of the two demographic groups under consideration. In the PR and Wilke estimates, those coefficients whose sign is opposite to the expected one turn out to be not significantly different.

The results we observe from the data are in line with the Montecarlo experiments. The estimator proposed by Wilke and the SLS estimator that minimizes objective function $\hat{L}_3(a, c)$ are robust with respect to different starting values and in the case of the SLS, more stable for different values of the bandwidth. The estimates of the equivalence of

are generated through repetition of the gridsearch process for 500 bootstrap samples.

scale are statistically different though for both estimators. The modified SLS we propose gives an estimate $\hat{c} = 0.1936$ (with standard error 0.0245) and the modified PR estimator proposed by Wilke gives an estimate of $\hat{c} = 0.3926$ (with standard error 0.0086).The reported estimate of c by Pinske and Robinson lies between both values. It should be pointed out that this estimate has been obtained at a higher computational cost since the computational procedure involves grid search methods and much different results would have been obtained if gradient methods were used. The other two estimators however obtain their estimates at a lower computational cost.

4.9 Conclusion

In this work we have provided an alternative way of estimating the parametric transformations of a shape invariant model that relate nonparametric regression curves for different samples by using the Semiparametric Least Squares (SLS) estimator (Ichimura (1993)). The previous estimators that can be found in the literature (Hardle and Marron (1990) and Pinkse and Robinson (1995) imply extensive computational methods because their corresponding objective functions only attain a local minimum at the true value of the parameters and, in addition, their shape do not help in obtaining consistent estimates. In order to avoid these disadvantages, some modifications of these estimators have been already proposed by Wilke (2003). The shape invariant model can be interpreted as a single index model and therefore SLS constitutes a natural and efficient way of estimating the parameters of interest.

We also find that the objective function in the original framework of SLS involves some computational difficulties as well. Thus, the partial derivative of the objective function with respect to the parameters imbedded in the unknown function is zero for arbitrary large or small values of this parameter, where the objective function attains a local minimum. The intuition is that when this parameter is arbitrary large in absolute value, the nonparametric estimation of the conditional mean function evaluated at a particular observation gives zero weight to those observations belonging to a different sample. In other words, only the observations belonging to the same sample are used and therefore the objective function does not change with respect to the parameter that captures the horizontal shift among regression functions for different samples . For this reason, we propose two possible modifications of the objective function that the SLS estimator minimizes which help in computing estimates for this parameter. The idea of both modifications is

that the objective function is divided by the amount of observations for which the estimated conditional mean function still depends on the value of the vertical shift parameter. The greater this parameter in absolute terms, the fewer this observations for which this property holds and this makes the objective function increase for these values of the parameter and ease the computation of the minimum. One of the modifications implies that the objective function is divided by an indicator function that depends on the value of the parameters and the another proposed estimator corrects the objective function by a continuos function that converges to the indicator function as the sample size increases.

In the Montecarlo experiments we perform, this second modified estimator performs better than the original formulation and than the corrected objective function by the indicator function. We find that when gradient methods are used in the computation, the performance of the previous suggested estimators highly depends on the starting values. If these are not carefully chosen, the shape of the objective functions makes that its minimization yields local minimum as valid solutions which are far away of the true value of the parameters.

The limit of both corrected objective functions coincides with the limiting function in original framework and also the corrections do not depend on the parameter value in the limit. This implies that the identification conditions and the asymptotic properties are equivalent for the three cases.

Additionally, it can deal automatically with the comparison of regression curves for more than two samples or with respect to more than one variable and because grid search methods are not needed in the computation, this increase in the dimension is computationally feasible.

Given that we consider the case where we observe multiple endogenous variables for each individual, we adjust the SLS estimator to deal with the estimation of a single index model with multiple equations. This allows one to account for the existence of correlation of the errors among equations for each individual. We also give sufficient conditions that the unknown functions in the system of equations should satisfy in order to identify the finite dimensional parameters of the model. We establish the asymptotic properties of the SLS estimator with multiple equations and discuss the optimal weighting matrix across equations in order to obtain an efficient estimator for this subclass of estimators.

A single index model with the form of a shape invariant model arises in the estimation of Engel curve relationships (Blundell, Duncan and Pendakur (1998)). We use the British Family Expenditure Survey to apply the different estimators discussed in this work to

estimate consumption based equivalence scale between households with two children and households with one child. In this empirical application, we find that the suggested estimator where the objective function is divided by the indicator function is less stable than when the denominator depends on a continuous function. And it is also less robust to the choice of the smoothing parameter. The estimator that divides the objective function by the continuous function converging to the indicator function together with the estimator proposed by Wilke (2003) were the estimators that performed better in the simulation. In the data though, we find different estimates of the equivalence scale parameter.

There are some issues that are not addressed in this paper and that would be worthwhile to investigate. First, given the alternative estimators for the parameters of the shape invariant model, it would be interesting to compare their asymptotic efficiency and study whether under some weighting schemes the SLS with multiple equations achieves the efficiency bound for single-index models (Newey (1990)). This would allow us to conclude that the SLS estimator with multiple equations is asymptotically more efficient that the early proposed estimators and that even the modified versions of them. A second point would be to study the large sample properties of the estimation of the infinite dimensional parameter given by the unknown function for each equation. And finally, as the Montecarlo simulations and the empirical application pointed out, the choice of the bandwidth might be important for the robustness of the estimates. For this reason, analysis of the optimal choice of the bandwidth for the modified estimators and the asymptotic properties of the estimator that considers the bandwidth as an additional parameter are in order.

4.10 Tables

Table 4.1: Simulation results for SLS: Part 1 . 300 trials. n=200 and J=1

Statistics over replications	mean	st. dev.	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}	MSE
(1) $(\hat{\mathbf{a}}, \hat{\mathbf{c}}) = \arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}(\mathbf{a}, \mathbf{c})$								
starting value : $[2, 2]$								
$a_0 = 0.3$	-0.403	0.489	-1.345	-0.625	-0.259	-0.104	0.020	0.732
$c_0 = -0.3$	4.613	1.719	2.605	2.706	5.290	6.129	6.490	27.090
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.045							
starting value : $[-1, 1]$								
$a_0 = 0.3$	0.371	0.353	0.072	0.250	0.294	0.330	1.170	0.130
$c_0 = -0.3$	-0.920	2.483	-7.104	-0.412	-0.218	-0.025	1.325	6.549
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.041							
starting value : $[0.7, -0.7]$								
$a_0 = 0.3$	0.272	0.195	0.192	0.277	0.304	0.336	0.368	0.039
$c_0 = -0.3$	-0.133	1.098	-0.743	-0.434	-0.296	-0.149	0.205	1.234
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.040							
(2) $(\hat{\mathbf{a}}_2, \hat{\mathbf{c}}_2) = \arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}_2(\mathbf{a}, \mathbf{c})$								
starting value : $[2, 2]$								
$a_0 = 0.3$	-0.114	0.110	-0.199	-0.156	-0.110	-0.071	-0.011	0.183
$c_0 = -0.3$	2.665	0.421	2.556	2.654	2.698	2.727	2.792	8.967
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.071							
starting value : $[-1, 1]$								
$a_0 = 0.3$	0.275	0.085	0.081	0.249	0.297	0.323	0.370	0.008
$c_0 = -0.3$	-0.104	0.549	-0.608	-0.429	-0.277	0.001	1.280	0.340
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.043							
starting value : $[0.7, -0.7]$								
$a_0 = 0.3$	0.309	0.050	0.232	0.285	0.311	0.339	0.384	0.003
$c_0 = -0.3$	-0.347	0.257	-0.747	-0.496	-0.348	-0.193	0.090	0.068
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.042							

Notes: These objective functions are conditioned on the bandwidth that minimizes the Cross-Validation function evaluated at the true value of the parameters, or equivalently $\hat{L}(h|a_0, c_0)$, denoted by $\hat{h}_0$. The value function $L(\hat{a}, \hat{c}, \hat{h}_0)$ reports the mean over replications of the value of the function at the solution

Table 4.2: Simulation results for SLS: Part 2 . 300 trials. n=200 and J=1

Statistics over replications	mean	st. dev.	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}	MSE
(2) $(\hat{\mathbf{a}}_3, \hat{\mathbf{c}}_3) = \arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}_3(\mathbf{a}, \mathbf{c})$								
starting value : $[2, 2]$								
$a_0 = 0.3$	0.320	0.082	0.000	0.313	0.0336	0.359	0.385	0.007
$c_0 = -0.3$	-0.504	0.184	-0.750	-0.601	-0.524	-0.444	-0.000	0.075
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.309							
starting value : $[-1, 1]$								
$a_0 = 0.3$	0.337	0.0401	0.272	0.316	0.337	0.361	0.389	0.003
$c_0 = -0.3$	-0.534	0.158	-0.760	-0.621	-0.546	-0.468	-0.331	0.080
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.328							
starting value : $[0.7, -0.7]$								
$a_0 = 0.3$	0.309	0.040	0.280	0.319	0.341	0.363	0.394	0.003
$c_0 = -0.3$	-0.557	0.122	-0.766	-0.639	-0.556	-0.479	-0.380	0.081
$L(\hat{a}, \hat{c}, \hat{h}_0)$	0.321							

Notes: These objective functions are conditioned on the bandwidth that minimizes the Cross-Validation function evaluated at the true value of the parameters, or equivalently $\hat{L}(h|a_0, c_0)$, denoted by $\hat{h}_0$. The value function $L(\hat{a}, \hat{c}, \hat{h}_0)$ reports the mean over replications of the value of the function at the solution.

Table 4.3: Simulation results for previous estimators: Part 1. 300 trials. n=200 and J=1

	mean	st. dev.	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}	MSE
starting value : $[2, 2]$								
Hardle-Marron estimator								
$a_0 = 0.3$	-0.523	0.099	-0.655	-0.588	-0.528	-0.482	-0.346	0.687
$c_0 = -0.3$	5.815	0.395	5.618	5.722	5.723	5.877	6.492	37.547
$L(\hat{a}, \hat{c})$	0.001							
Pinkse-Robinson estimator								
$a_0 = 0.3$	0.300	0.053	0.223	0.270	0.298	0.328	0.398	0.003
$c_0 = -0.3$	4.485	0.271	4.282	4.475	4.511	4.546	4.587	22.972
$L(\hat{a}, \hat{c})$	0.002							
Wilke estimator								
$a_0 = 0.3$	0.309	0.070	0.228	0.289	0.314	0.339	0.380	0.005
$c_0 = -0.3$	-0.310	0.471	0.727	-0.501	-0.336	-0.167	0.116	0.222
$L(\hat{a}, \hat{c})$	0.019							
starting value : $[-1, -1]$								
Hardle-Marron estimator								
$a_0 = 0.3$	1.486	1.805	0.251	0.859	1.04	1.150	5.905	4.665
$c_0 = -0.3$	-2.526	4.467	-5.720	-5.720	-5.720	-0.142	6.493	24.914
$L(\hat{a}, \hat{c})$	0.018							
Pinkse-Robinson estimator								
$a_0 = 0.3$	0.104	0.043	0.038	0.073	0.102	0.137	0.173	0.040
$c_0 = -0.3$	1.743	0.111	1.672	1.718	1.745	1.781	1.827	4.184
$L(\hat{a}, \hat{c})$	0.0001							
Wilke estimator								
$a_0 = 0.3$	0.308	0.066	0.228	0.288	0.314	0.339	0.378	0.004
$c_0 = -0.3$	-0.285	0.611	-0.721	-0.496	-0.335	-0.166	0.127	0.374
$L(\hat{a}, \hat{c})$	0.019							

Note: For these three estimators the integral over a range of x needs to be computed. This is done using the midpoint formula to compute the integral of the objective function with 100 points of support and they only use those values of X belonging to the intersection between $\hat{\Omega}_X^{(0)} + c \cap \hat{\Omega}_X^{(1)}$, taking value 0 for any value of x outside this support. Cross-validated bandwidth was used for the estimation of the nonparametric mean functions and density functions used in the loss functions of these estimators.

Table 4.4: Simulation results for previous estimators: Part 2 . 300 trials. n=200 and J=1

	mean	st. dev.	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}	MSE
starting value : $[0.7, -0.7]$								
Hardle-Marron estimator								
$a_0 = 0.3$	0.2451	0.2144	-0.3064	0.2788	0.3136	0.3412	0.3785	0.0490
$c_0 = -0.3$	0.4497	2.2773	-0.7370	-0.5646	-0.3570	-0.1230	6.4926	5.7480
$L(\hat{a}, \hat{c})$	0.0164							
Pinkse-Robinson estimator								
$a_0 = 0.3$	0.3267	0.0384	0.2731	0.3044	0.3266	0.3493	0.3835	0.0022
$c_0 = -0.3$	-0.7333	0.0430	-0.7481	-0.7401	-0.7356	-0.7309	-0.7248	0.1896
$L(\hat{a}, \hat{c})$	0.41e-6							
Wilke estimator								
$a_0 = 0.3$	0.2431	0.2270	-0.3209	0.2796	0.3144	0.3413	0.3796	0.0548
$c_0 = -0.3$	0.3838	2.1913	-0.7342	-0.5633	-0.3583	-0.1250	6.4926	5.2692
$L(\hat{a}, \hat{c})$	0.0169							

Note: For these three estimators the integral over a range of x needs to be computed. This is done using the midpoint formula to compute the integral of the objective function with 100 points of support and they only use those values of X belonging to the intersection between $\hat{\Omega}_X^{(0)} + c \cap \hat{\Omega}_X^{(1)}$, taking value 0 for any value of x outside this support. Cross-validated bandwidth was used for the estimation of the nonparametric mean functions and density functions used in the loss functions of these estimators.

Table 4.5: Simulation results for SLS for multiple equations J=2 . 300 trials. n=200

	mean	st. dev.	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}	MSE
starting value $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{c}] = [\mathbf{1}, \mathbf{1}, \mathbf{1}]$								
$(\hat{\mathbf{a}}_2, \hat{\mathbf{c}}_2) = \arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}_2(\mathbf{a}, \mathbf{c})$								
$a_{01} = 1$	0.782	0.347	0.553	0.629	0.709	0.894	1.052	0.168
$a_{02} = -2.5$	-2.849	0.272	-3.232	-3.058	-2.896	-2.612	-2.398	0.196
$c_0 = 0.5$	0.981	0.731	-0.377	0.374	0.954	1.403	2.255	0.766
$L(\hat{a}, \hat{c})$	0.988							
$(\hat{\mathbf{a}}_3, \hat{\mathbf{c}}_3) = \arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}_3(\mathbf{a}, \mathbf{c})$								
$a_{01} = 1$	1.194	0.051	1.095	1.167	1.193	1.223	1.261	0.040
$a_{02} = -2.5$	-2.227	0.102	-2.401	-2.281	-2.223	-2.168	-2.055	0.085
$c_0 = 0.5$	0.102	0.196	-0.262	-0.061	0.095	0.215	0.444	0.197
$L(\hat{a}, \hat{c})$	0.651							
Hardle-Marron estimator								
$a_{01} = 1$	0.750	0.495	0.0325	0.386	0.715	1.025	1.427	0.308
$a_{02} = -2.5$	-22.018	14.820	-34.160	-33.089	-32.140	-2.582	-2.363	600.58
$c_0 = 0.5$	4.384	3.235	0.387	0.793	6.217	6.493	6.493	25.547
$L(\hat{a}, \hat{c})$	0.040							
Pinkse-Robinson estimator								
$a_{01} = 1$	0.958	0.026	0.913	0.940	0.955	0.975	1.002	0.002
$a_{02} = -2.5$	-1.326	0.096	-1.492	-1.373	-1.314	-1.256	-1.192	1.388
$c_0 = 0.5$	5.303	0.075	5.191	5.258	5.297	5.345	5.387	23.077
$L(\hat{a}, \hat{c})$	4.2e-5							
Wilke estimator								
$a_{01} = 1$	1.020	0.405	0.946	0.999	1.020	1.048	1.079	0.002
$a_{02} = -2.5$	-2.492	0.094	-2.665	-2.555	-2.480	-2.429	-2.352	0.009
$c_0 = 0.5$	0.646	0.156	0.382	0.545	0.639	0.751	0.901	0.046
$L(\hat{a}, \hat{c})$	0.123							

Table 4.6: Simulation results for SLS for multiple equations J=2 . 300 trials. n=200

	mean	st. dev.	Q_5	Q_{25}	Q_{50}	Q_{75}	Q_{95}	MSE
starting value $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{c}] = [\mathbf{0.5}, \mathbf{-1}, \mathbf{0.7}]$								
$(\hat{a}, \hat{c}) = \arg\min_{(a,c)} \hat{L}_2(a, c)$								
$a_{01} = 1$	0.989	0.044	0.908	0.958	0.992	1.021	1.053	0.002
$a_{02} = -2.5$	-2.515	0.103	-2.708	-2.575	-2.513	-2.439	-2.364	0.011
$c_0 = 0.5$	0.656	0.372	0.0285	0.3916	0.6172	0.8727	1.2540	0.1627
$L(\hat{a}, \hat{c})$	0.375							
$(\hat{a}, \hat{c}) = \arg\min_{(a,c)} \hat{L}_3(a, c)$								
$a_{01} = 1$	1.165	0.041	1.091	1.139	1.163	1.193	1.226	0.029
$a_{02} = -2.5$	-2.270	0.089	-2.434	-2.322	-2.269	-2.211	-2.127	0.061
$c_0 = 0.5$	0.401	0.181	0.079	0.263	0.386	0.515	0.689	0.043
$L(\hat{a}, \hat{c})$	0.549							
Hardle-Marron estimator								
$a_{01} = 1$	1.020	0.041	0.9444	0.999	1.020	1.048	1.073	0.002
$a_{02} = -2.5$	-2.492	0.095	-2.684	-2.560	-2.480	-2.427	-2.360	0.009
$c_0 = 0.5$	0.654	0.162	0.357	0.556	0.641	0.769	0.909	0.050
$L(\hat{a}, \hat{c})$	0.1132							
Pinkse-Robinson estimator								
$a_{01} = 1$	0.998	0.039	0.934	0.971	1.003	1.021	1.065	0.002
$a_{02} = -2.5$	-2.633	0.120	-2.856	-2.720	-2.634	-2.546	-2.442	0.032
$c_0 = 0.5$	1.679	0.121	1.455	1.606	1.677	1.748	1.878	1.405
$L(\hat{a}, \hat{c})$	0.0001							
Wilke estimator								
$a_{01} = 1$	1.022	0.041	0.946	1.002	1.020	1.051	1.074	0.002
$a_{02} = -2.5$	-2.490	0.095	-2.680	-2.555	-2.478	-2.425	-2.352	0.009
$c_0 = 0.5$	0.644	0.158	0.352	0.546	0.628	0.764	0.884	0.046
$L(\hat{a}, \hat{c})$	0.122							

Table 4.7: Estimation using FES data for one equation. Alcohol Engel Curves. Results for all the estimators functions

	(1)	(2)	(3)	HM	PR	Wilke
$\hat{a}$	-0.0357	-0.0139	-0.0112	0.0149	-0.0263	-0.0060
	(0.0245)	(0.0038)	(0.0036)	(0.0047)	-	(0.0042)
$\hat{c}$	-1.9919	0.1020	0.1818	2.1772	-0.9417	0.2989
	(0.0360)	(0.0519)	(0.0479)	(0.4834)	-	(0.1157)
$\hat{h}$	0.1	0.25	0.1			
$\hat{L}(\hat{a}, \hat{c})$	0.0025	0.0040	0.0044	1.5e-05	1.8e-05	0.0001

Note: Column (1) : $(\hat{a}, \hat{c}) = \arg\min_{(a,c)} \hat{L}(a, c)$; Column (2) : $(\hat{a}_2, \hat{c}_2) = \arg\min_{(a,c)} \hat{L}_2(a, c)$; Column (3): $(\hat{a}_3, \hat{c}_3) = \arg\min_{(a,c)} \hat{L}_3(a, c)$. Standard errors in parenthesis.

Table 4.8: Estimation using FES data for multiple equations: Engel Curves. SLS Estimation when objective function is divided by the sum of indicator functions

Parameter	$\arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}_2(\mathbf{a},\mathbf{c})$			
Bandwidth value	$h = 0.1$	$h = 0.25$	$h = 0.5$	$h = 1$
a_{food}	−0.0054	0.0333	0.0247	−0.0850
	(0.0056)	(0.0064)	(0.0065)	(0.0085)
a_{fuel}	−0.0085	0.0017	−0.0015	−0.0396
	(0.0028)	(0.0029)	(0.0029)	(0.0043)
a_{cloth}	0.0118	−0.0052	0.0002	0.0640
	(0.0051)	(0.0054)	(0.0056)	(0.0074)
$a_{alcohol}$	−0.0104	−0.0145	−0.0123	0.0067
	(0.0034)	(0.0035)	(0.0035)	(0.0052)
$a_{transport}$	−0.0045	−0.0125	−0.0100	0.0199
	(0.0055)	(0.0057)	(0.0056)	(0.0090)
c	0.2010	0.0053	0.0558	1.0248
	(0.0245)	(0.362)	(0.0624)	(0.0608)
$\hat{L}_2(\hat{a},\hat{c})$	0.0451	0.0457	0.0448	0.0459

Note: Standard Errors in Parenthesis. Estimated Covariance matrix was obtained as described in expressions (4.34) and (4.35). Consistent estimators with weighting matrix $V_i = I_J$ are presented here.

Table 4.9: Estimation using FES data for Multiple equations: Engel Curves. SLS Estimation when objective function is divided by the sum of s functions

Parameter	$\arg\min_{(\mathbf{a},\mathbf{c})} \hat{\mathbf{L}}_3(\mathbf{a},\mathbf{c})$			
Bandwidth value	$h = 0.1$	$h = 0.25$	$h = 0.5$	$h = 1$
a_{food}	0.0069	0.0095	0.0184	0.0201
	(0.0055)	(0.0061)	(0.0067)	(0.0054)
a_{fuel}	−0.0078	−0.0067	−0.0038	−0.0034
	(0.0028)	(0.0029)	(0.0096)	(0.0027)
a_{cloth}	0.0112	0.0096	0.0041	0.0035
	(0.0050)	(0.0052)	(0.0056)	(0.0049)
$a_{alcohol}$	−0.0104	−0.0098	−0.0111	−0.0111
	(0.0035)	(0.0035)	(0.0035)	(0.0034)
$a_{transport}$	−0.0042	−0.0051	−0.0081	−0.0088
	(0.0055)	(0.0056)	(0.0056)	(0.0056)
c	0.1936	0.1739	0.1130	0.1033
	(0.0245)	(0.0331)	(0.0656)	(0.1198)
$\hat{L}_3(\hat{a},\hat{c})$	0.0487	0.0632	0.0960	0.1989

Note: Standard Errors in Parenthesis. Estimated Covariance matrix was obtained as described in expressions (4.34) and (4.35). Consistent estimators with weighting matrix $V_i = I_J$ are presented here.

Table 4.10: Estimates reported by Blundell, Duncan and Pendakur (1998) and Wilke (2003) using FES data for multiple equations: Engel Curves.

Parameter		
Bandwidth value	Blundell, Duncan and Pendakur (1998)	Wilke (2003)
a_{food}	0.0281	−0.0292
	(0.0048)	(0.2423)
a_{fuel}	−0.0013	−0.0176
	(0.0025)	(0.0336)
a_{cloth}	−0.0018	0.0209
	(0.0045)	(0.1238)
$a_{alcohol}$	−0.0121	−0.0009
	(0.0032)	(0.0520)
$a_{transport}$	−0.0100	0.0149
	(0.0053)	(0.1502)
c	0.2590	0.3926
	(0.0809)	(0.0086)

Note: Blundell, Duncan and Pendakur (1998) report results for FES data using Pinkse and Robinson (1995) estimator. Standard errors in parenthesis.

4.11 Appendix

We next present some lemmas that are used in the proof of Theorem (9) for consistency.

Lemma 10 *Under assumptions (1)-(7), if*

$$\lim_{n\to\infty} \frac{n}{\ln n} h_n^{(1+2/r)} = \infty$$

,then

$$\sup_{(x,z,a,c)\in Q\times\mathcal{A}\times\mathcal{C}} |\hat{m}_{j,h_n}(x,z;a,c) - m_j(x,z;a,c)| \xrightarrow{p} 0$$

for $j = 1, ..., J$

Proof.

Assumptions 4.(1)-4.(7) and the bandwidth condition in (4.27) are sufficient to apply the Uniform Law of Large Numbers for U-statistics indexed by bandwidths (See in Appendix in Ichimura and Lee (1991)) so that

$$\frac{1}{n-1}\sum_{r\neq i}^{n} \frac{1}{h_n} K\left(\frac{(x_i - cz_i) - (x_r - cz_r)}{h_n}\right) \xrightarrow{p} f_c(x_i - cz_i, c) \tag{4.37}$$

$$\frac{1}{n-1}\sum_{r\neq i}^{n} \frac{1}{h_n} (w_{rj} - z_r a_j) K\left(\frac{(x_i - cz_i) - (x_r - cz_r)}{h_n}\right) \xrightarrow{p} m_j(x_i, z_i, a, c) f_c(x_i - cz_i, c) \tag{4.38}$$

uniformly in $(x_i, z_i, a, c) \in Q \times \mathcal{A} \times \mathcal{C}$. To show the uniform convergence of the conditional expectation, consider

$$\sup_{(x,z,a,c)\in Q\times\mathcal{A}\times\mathcal{C}} |\hat{m}_{j,h_n}(x,z;a,c) - m_j(x,z;a,c)| \leq$$

$$\leq \frac{1}{\inf_{(x,z,c)\in Q\times\mathcal{C}} \left(\hat{f}_c(x - cz, c)\right)} \left\{ \sup_{(x,z,a,c)\in Q\times\mathcal{A}\times\mathcal{C}} \left| \begin{array}{c} \hat{m}_{j,h_n}(x,z;a,c)\hat{f}_c(x - cz, c) - \\ -m_j(x,z;a,c) f_c(x - cz, c) \end{array} \right| \right.$$

$$\left. + \sup_{(x,z,a,c)\in Q\times\mathcal{A}\times\mathcal{C}} |m_j(x,z;a,c)| \sup_{(x,z,a,c)\in Q\times\mathcal{A}\times\mathcal{C}} \left|\hat{f}_c(x - cz, c) - f_c(x - cz, c)\right| \right\}$$

The conditional expectation is continuous by assumption 4.(5) in the index and $Q \times \mathcal{A} \times \mathcal{C}$ is compact, then $m_j(x, z; a, c)$ is uniformly bounded.

The density $f_c(x_i - cz_i, c)$ is bounded away from zero uniformly in $(x, z, c) \in Q \times \mathcal{C}$ by the definition of compact set in expression (4.11). This implies that for $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left(\left|\inf_{(x,z,c)\in Q\times\mathcal{C}} \left(\hat{f}_c(x - cz, c)\right)\right| > \varepsilon\right) = 1$$

Therefore, applying the two uniform convergence results above in (4.38) and (4.37), the uniform convergence of the conditional expectation follows. ■

Proof. [Proof of Theorem (**9**)]

Given the assumptions in Theorem (**9**), the limiting objective function $L(a, c)$ is a continuous function of the parameters (a, c) and the identification conditions are also satisfied so that (a_0, c_0) are the unique minimizers of the limiting objective function $L(a, c)$. Therefore, in order to show consistency, the only condition that it is left to be satisfied is the uniform convergence in probability of the objective function to the limiting objective function. Let consider the application the Uniform Law of Large Numbers to

$$g(y; a, c) = I_Q B(y, a, m(., a, c))' V B(y, a, m(., a, c))$$

The only condition from Lemma 2.4 in Newey and McFadden (1994) that it is left to be satisfied is $E\left(\sup_{(a,c)\in(\mathcal{A}\times\mathcal{C})} |g(y; a, c)|\right) < \infty$. By the existence of the moments in assumption 4.(1), the compactness of $(\mathcal{A} \times \mathcal{C})$ and the continuity of g with respect to y for each value of (a, c), this dominance condition is satisfied. This shows that the ULLN can be applied to $g(y; a, c)$ and consequently,

$$|L^*(a, c) - L(a, c)| \xrightarrow{p} 0 \text{ uniformly in } (a, c)$$

In order to show the uniform convergence of $\left|\hat{L}(a, c) - L^*(a, c)\right|$ in probability to zero, the uniform convergence of terms (4.25)-(4.26) is studied.

Regarding the uniform convergence of term (4.25), assumption 4.(8) guarantees that this term converges in probability to zero uniformly.

Regarding the uniform convergence of term (4.26), consider instead the uniform convergence in probability of

$$\left| \begin{array}{l} \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n I_{iQ} B(y_i, a, \hat{m}_i(., a, c))_i' V_i B(y_i, a, \hat{m}_i(., a, c))\right]^{1/2} - \\ \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n I_{iQ} B(y_i, a, m_i(., a, c))' V_i B(y_i, a, m_i(., a, c))\right]^{1/2} \end{array} \right| =$$

$$\frac{1}{\sqrt{n}} \left[\|I_{iQ} \times S_i \times B(y_i, a, \hat{m}_i(., a, c))\| - \|I_{iQ} \times S_i \times B(y_i, a, m_i(., a, c))\|\right] \leq$$

$$\leq \frac{1}{\sqrt{n}} \|I_{iQ} \times S_i \times [B(y_i, a, \hat{m}_i(., a, c)) - B(y_i, a, m_i(., a, c))]\| = \frac{1}{\sqrt{n}} \|I_{iQ} \times S_i \times [m_i(., a, c) - \hat{m}_i(., a, c)]\|$$

where the positive definite matrix V_i has been expressed as $V_i = S_i' S_i$. Therefore the uniform convergence in probability of (4.26) is satisfied when the uniform convergence of $\hat{m}_i(.,a,c)$ converges in probability to $m_i(.,a,c)$ uniformly in $i \in Q$ and (a,c), which is satisfied under the above assumptions as stated in Lemma (10).

Proof. [Proof of Theorem (**9**)]

Given the assumptions in Lemma (**9**), the limiting objective function $L(a,c)$ is a continuous function of the parameters (a,c) and the identification conditions are also satisfied so that (a_0, c_0) are the unique minimizers of the limiting objective function $L(a,c)$. Therefore, in order to show consistency, the only condition that it is left to be satisfied is the uniform convergence in probability of the objective function to the limiting objective function. Let consider the application the Uniform Law of Large Numbers to

$$g(y;a,c) = I_Q B(y,a,m(.,a,c))' \Omega B(y,a,m(.,a,c))$$

The only condition from Lemma 2.4 in Newey and McFadden (1994) that it is left to be satisfied is $E\left(\sup_{(a,c)\in(\mathcal{A}\times\mathcal{C})} |g(y;a,c)|\right) < \infty$. By the existence of the moments in assumption 4.(1), the compactness of $(\mathcal{A}\times\mathcal{C})$ and the continuity of g with respect to y for each value of (a,c), this dominance condition is satisfied. This shows that the ULLN can be applied to $g(y;a,c)$ and consequently,

$$|L^*(a,c) - L(a,c)| \xrightarrow{p} 0 \text{ uniformly in } (a,c)$$

In order to show the uniform convergence of $\left|\hat{L}(a,c) - L^*(a,c)\right|$ in probability to zero, the uniform convergence of terms (4.25)-(4.26) is studied.

Regarding the uniform convergence of term (4.25), assumption 4.(8) guarantees that this term converges in probability to zero uniformly.

Regarding the uniform convergence of term (4.26), consider instead the uniform convergence in probability of

$$\left| \begin{array}{l} \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n I_{iQ} B(y_i,a,\hat{m}_i(.,a,c))_i' \Omega_i B(y_i,a,\hat{m}_i(.,a,c))\right]^{1/2} - \\ \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n I_{iQ} B(y_i,a,m_i(.,a,c))' \Omega_i B(y_i,a,m_i(.,a,c))\right]^{1/2} \end{array} \right| =$$

$$\frac{1}{\sqrt{n}} \left[\|I_{iQ} \times S_i \times B(y_i,a,\hat{m}_i(.,a,c))\| - \|I_{iQ} \times S_i \times B(y_i,a,m_i(.,a,c))\|\right] \leq$$

$$\leq \frac{1}{\sqrt{n}} \|I_{iQ} \times S_i \times [B(y_i,a,\hat{m}_i(.,a,c)) - B(y_i,a,m_i(.,a,c))]\| = \frac{1}{\sqrt{n}} \|I_{iQ} \times S_i \times [m_i(.,a,c) - \hat{m}_i(.,a,c)]\|$$

where the positive definite matrix Ω_i has been expressed as $\Omega_i = S_i' S_i$. Therefore the uniform convergence in probability of (4.26) is satisfied when the uniform convergence of

$\hat{m}_i(.,a,c)$ converges in probability to $m_i(.,a,c)$ uniformly in $i \in Q$ and (a,c), which is satisfied under the above assumptions as stated in Lemma (10). ■

Proof. [Proof of Lemma (8): Uniform Convergence in probability of indicator function] Lemma (10) states the bandwidth conditions that are required for the uniform convergence in probability of the kernel density estimator of the index. The particular assumption required for the bandwidth is $\frac{n}{\log n} h_n \to \infty$, which is implied by the bandwidth condition in the consistency theorem. The following result for the uniform convergence in probability of the nonparametric estimated densities

$$\Pr\left(\sup_x \left|\hat{f}(x|z=j) - f(x|z=j)\right| > \varepsilon\right) \to 0 \text{ as } n_j \to \infty \text{ for } j = \{0.1\}$$

for any $\varepsilon > 0$ can be shown under the additional conditions in Assumptions 4.(9)-4.(11) and if $n_0 h_n^2 \to \infty$ and $n_1 h_n^2 \to \infty$ as $n \to \infty$ (which is indeed implied by the assumptions above and by the bandwidth condition in the consistency theorem in expression (4.27)). The probabilities $\Pr(Z=1)$ and $\Pr(Z=0)$ are consistently estimated when both $n_0 \to \infty$ and $n_1 \to \infty$. These last two results imply the convergence in probability of $\widehat{lb}(x,z)$ to $lb(x,z)$ uniformly on (x,z).

Define $\hat{t}(x,z,c) = \hat{f}_c(x - cz, c) - \widehat{lb}(x,z)$ and $t(x,z,c) = f_c(x - cz, c) - lb(x,z)$, then the two uniform convergence results above guarantee the uniform converge in probability to zero of $(\hat{t}(x,z,c) - t(x,z,c))$ uniformly in (x,z,c)

$$\Pr\left(\sup_{(x,z,c)\in H\times\mathcal{C}} \left|\hat{f}_c(x-cz,c) - \hat{f}(x-cz,c)\right| + \sup_{(x,z)\in H} \left|\widehat{lb}(x,z) - lb(x,z)\right| > \varepsilon\right) \geq$$

$$\geq \Pr\left(\sup_{(x,z,c)\in H\times\mathcal{C}} \left|\hat{t}(x,z,c) - t(x,z,c)\right| > \varepsilon\right)$$

The same uniform convergence result applies to the following transformation of t

$$1\{\hat{t}_h(x,z,c) > 0\} \xrightarrow{p} 1\{t(x,z,c) > 0\}$$

uniformly on $(x,z,c) \in H \times \mathcal{C}$, since the indicator function $\vartheta(u) = 1\{u > 0\}$ is continuous for all $u > 0$ and by the definition of set H in (4.15), $t(x,z,c) > 0$ for all $(x,z) \in H$ and $c \in \mathcal{C}$ ■

Proof. [Proof of Lemma (9): Uniform Convergence in probability of function s]

$$\sup_{i\in H,c\in\mathcal{C}}\left|s_{h_n}(\hat{f}_{i,c}(.,c)-\widehat{lb}_i)-1\right|\leq$$

$$\leq \sup_{i\in H,c\in\mathcal{C}}\left|s_{h_n}(\hat{f}_{i,c}(.,c)-\widehat{lb}_i)-1\left\{\hat{f}_{i,c}(.,c)-\widehat{lb}_i>0\right\}\right|+ \tag{4.39}$$

$$+\sup_{i\in H,c\in\mathcal{C}}\left|1\left\{\hat{f}_{i,c}(.,c)-\widehat{lb}_i>0\right\}-1\left\{f_{i,c}(.,c)-lb_i>0\right\}\right|+ \tag{4.40}$$

$$+\sup_{i\in H,c\in\mathcal{C}}\left|1\left\{f_c(x_i-cz_i,c)-f(x_i)>0\right\}-1\right| \tag{4.41}$$

The convergence in probability of term (4.39) is satisfied by the fact that $\lim_{n\to\infty}s_{h_n}(x)=1\{x>0\}$ uniformly in x if $h_n\to 0$ as $n\to\infty$. To show the convergence in probability of term (4.40) see Lemma (8). The last term equals zero by definition of set H. ■

Proof. [Proof of Theorem (10)] Let first show the uniform convergence in probability of function $\hat{L}_2(a,c)$ to $\hat{L}(a,c)$. Function $\hat{L}_2$ is set to zero for those values of c such that $\sum_{i=1}^n 1\left\{\hat{f}_{i,c}(.,c)-\widehat{lb}_i>0\right\}=0$. Define (a^*,c^*) as

$$(a^*,c^*)=\arg\sup_{(a,c)}\left|\hat{L}_2(a,c)-\hat{L}(a,c)\right|$$

Then,

$$\Pr\left(\sup_{a,c}\left|\hat{L}_2(a,c)-\hat{L}(a,c)\right|>\varepsilon\right)=$$

$$\Pr\left(\sup_{a,c}\left|\hat{L}_2(a,c)-\hat{L}(a,c)\right|>\varepsilon\middle|\sum_{i\in H}1\left\{\hat{f}_{i,c^*}(.,c^*)-\widehat{lb}_i>0\right\}>0\right)\Pr\left(\sum_{i\in H}1\left\{\hat{f}_{i,c^*}(.,c^*)-\widehat{lb}_i>0\right\}>0\right)+ \tag{4.42}$$

$$+\Pr\left(\sup_{a,c}\left|0-\hat{L}(a,c)\right|>\varepsilon\middle|\sum_{i\in H}1\left\{\hat{f}_{i,c^*}(.,c^*)-\widehat{lb}_i>0\right\}=0\right)\Pr\left(\sum_{i\in H}1\left\{\hat{f}_{i,c^*}(.,c^*)-\widehat{lb}_i>0\right\}=0\right) \tag{4.43}$$

We next argue that term (4.43) converges to zero as $n\to\infty$. Note that by Lemma (8) we can conclude that

$$\Pr\left(\sum_{i\in H}1\left\{\hat{f}_{i,c^*}(.,c^*)-\widehat{lb}_i>0\right\}=0\right)\to 0 \text{ as } n\to\infty$$

Denote by $S_n(c) = \frac{1}{n}\sum_{i\in H} 1\left\{\hat{f}_{i,c}(.,c) - \widehat{lb}_i > 0\right\}$. Therefore, we focus on the behavior of the difference of $\left|\hat{L}_2(a,c) - \hat{L}(a,c)\right|$ in that part such that $S(c^*) > 0$. Then,

$$\Pr\left(\sup_{a,c}\left|\hat{L}_2(a,c) - \hat{L}(a,c)\right| > \varepsilon \middle| S_n(c^*) > 0\right) \leq$$
$$\Pr\left(\sup_{c}|1 - S_n(c)|\frac{1}{\inf_c|S_n(c)|}\sup_{a,c}\left|\hat{L}(a,c)\right| > \varepsilon \middle| S_n(c^*) > 0\right) \tag{4.44}$$

This probability converges to zero if $\sup_c|1 - S_n(c)|$ converges to zero in probability and if $\inf_c|S_n(c)|$ is bounded away from zero. Note that

$$\sup_c|1 - S_n(c)| \leq \sup_c\left|1 - \frac{n_H}{n}\right| +$$
$$+ \sup_{c,i\in H}\left|1\{f_{i,c}(.,c) - lb_i > 0\} - 1\left\{\widehat{f}_{i,c}(.,c) - \widehat{lb}_i > 0\right\}\right|$$

where $n_N = \sum_{i=1}^{n} 1\{i \in H\}$. The second term from the above inequality converges to zero in probability by Lemma (8). The first term in the above inequality converges to zero under assumption 4.(12). Thus, as the number of observations increases we require that the number of observations where the true density function of the index is strictly bounded above from lb increases at the same rate as n.

Since $1\left\{\widehat{f}_{i,c}(.,c) - \widehat{lb}_i > 0\right\}$ converges to $1\{f_{i,c}(.,c) - lb_i > 0\}$ uniformly in c and

$$\frac{1}{n}\sum_{i\in H} 1\{f_{i,c}(.,c) - lb_i > 0\} = \frac{n_H}{n}$$

for all $c \in \mathcal{C}$, then we can conclude that $\inf_c|S_n(c)|$ is bounded away from zero.

An equivalent reasoning can be used to show the uniform convergence in probability of $\left|\hat{L}_3(a,c) - \hat{L}(a,c)\right|$ to zero. By the uniform convergence of the nonparametric estimator of the density of the index and of the estimator of the lower bound, by the definition of set H and by the limit of function s_{h_n} to the indicator function if $h_n \to 0$ as $n \to \infty$, Lemma (9) in the Appendix shows

$$\Pr\left(\sup_{i\in H, c\in\mathcal{C}}\left|s_{h_n}(\hat{f}_{i,c}(.,c) - \widehat{lb}_i) - 1\right| > \varepsilon\right) \to 0$$

which implies that term $\Pr\left(\sum_{i\in H} s_{h_n}\left\{\hat{f}_{i,c^*}(.,c^*) - \widehat{lb}_i\right\} = 0\right)$ converges to zero as $n \to \infty$. Therefore, in the limit, one should focus on the behavior of $\left|\hat{L}_3(a,c) - \hat{L}(a,c)\right|$ when $\sum_{i\in H} s_{h_n}\left\{\hat{f}_{i,c^*}(.,c^*) - \widehat{lb}_i\right\} > 0$. In order to show that in this case, the difference $\left|\hat{L}_3(a,c) - \hat{L}(a,c)\right|$

converges to zero in probability uniformly in (a, c) and following the same reasoning as in expression (4.44), it is sufficient to show that the following term converges in probability to zero

$$\sup_c \left| 1 - \frac{1}{n} \sum_{i \in H} s_{h_n} \left\{ \hat{f}_{i,c^*}(., c^*) - \widehat{lb}_i \right\} \right| \leq \sup_c \left| 1 - \frac{n_H}{n} \right| +$$
$$+ \sup_{c, i \in H} \left| 1 \left\{ f_{i,c}(., c) - lb_i > 0 \right\} - s_{h_n} \left\{ \widehat{f}_{i,c}(., c) - \widehat{lb}_i \right\} \right|$$

The uniform convergence to zero of this upper bound is again implied by Lemma (9) and under under assumption 4. (12). ■

Proof. [Proof of Theorem (11)] First we show the convergence in probability of the Hessian to a positive definite matrix. Note that the hessian term in expression (4.29) can be expressed as

$$\frac{1}{n} \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, \hat{a}, \hat{m}_i(., \hat{a}, \hat{c})) \right]' \hat{V}_{in} \left[\nabla_{(a,c)} B(y_i, \bar{a}, \hat{m}_i(., \bar{a}, \bar{c})) \right] =$$

$$\sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right]' V_i \left[\nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right] + \tag{4.45}$$

$$+ \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right]' \left(\hat{V}_{in} - V_i \right) \left[\nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right] + \tag{4.46}$$

$$+ \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right]' \hat{V}_{in} \left[\nabla_{(a,c)} B(y_i, \bar{a}, m_i(., \bar{a}, \bar{c})) - \nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right] \tag{4.47}$$

$$+ \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, \hat{a}, \hat{m}_i(., \hat{a}, \hat{c})) - \nabla_{(a,c)} B(y_i, \hat{a}, m_i(., \hat{a}, \hat{c})) \right]' \hat{V}_{in} \left[\nabla_{(a,c)} B(y_i, \bar{a}, m_i(., \bar{a}, \bar{c})) \right] \tag{4.48}$$

$$+ \sum_{i=1}^{n} I_{iQ} \nabla_{(a,c)} B(y_i, \hat{a}, \hat{m}_i(., \hat{a}, \hat{c}))' \hat{V}_{in} \left[\nabla_{(a,c)} B(y_i, \bar{a}, \hat{m}_i(., \bar{a}, \bar{c})) - \nabla_{(a,c)} B(y_i, \bar{a}, m_i(., \bar{a}, \bar{c})) \right] \tag{4.49}$$

The first term (4.45) converges in probability to

$$H = E \left[I_Q \left[D \left(-z - \frac{\partial m_j(a_{j0}, c_0)}{\partial a_j} \right) , - \frac{\partial m(a_0, c_0)}{\partial c} \right]' V \left[D \left(-z - \frac{\partial m_j(a_{j0}, c_0)}{\partial a_j} \right) , - \frac{\partial m(a_0, c_0)}{\partial c} \right] \right]$$

by the Uniform Law of Large Numbers (Newey and McFadden (1994), Theorem 2.4) and by the consistency of $(\hat{a}, \hat{c})$.

The second term converges to zero in probability by assumption 4.(8).The third term converges in probability to zero by the consistency of $(\hat{a}, \hat{c})$ and the continuity of function B with respect to the parameters. In order to show that the terms (4.48) and (4.49) converge to zero in probability note that

$$\nabla_{(a,c)} B(y_i, a, \hat{m}_i(., a, c)) - \nabla_{(a,c)} B(y_i, a, m_i(., a, c)) =$$
$$\left[D \left(\frac{\partial m_j(a_j, c)}{\partial a_j} - \frac{\partial \hat{m}_j(a_j, c)}{\partial a_j} \right), \frac{\partial m(a, c)}{\partial c} - \frac{\partial \hat{m}(a, c)}{\partial c} \right]$$

and moreover,

$$\frac{\partial m_j(a_j, c)}{\partial a_j} - \frac{\partial \hat{m}_j(a_j, c)}{\partial a_j} = \hat{E}(z|x - cz) - E(z|x - cz) \text{ for all } j$$

Therefore the uniform convergence of the derivatives of the nonparametric estimators of the conditional mean, i.e.

$$\sup_{(x,z,a,c) \in Q \times \mathcal{A} \times \mathcal{C}} \left| \hat{E}(z|x - cz) - E(z|x - cz) \right| \xrightarrow{p} 0 \text{ and } \sup_{(x,z,a,c) \in Q \times \mathcal{A} \times \mathcal{C}} \left| \frac{\partial \hat{m}_j(a_j, c)}{\partial c} - \frac{\partial m_j(a_j, c)}{\partial c} \right| \xrightarrow{p} 0$$

for all j, guarantees the convergence in probability to zero of the last two terms. Under the assumptions of theorem (11), if $\lim_{n \to \infty} \frac{n}{\log(n)} h_n^{2(1+2/r)+1} = \infty$, this uniform convergence result holds (see Lemma 4 in Ichimura and Lee (1991)).

We next show the convergence in distribution of the score term in expression (4.30). First note that by the continuous differentiability of function B and the kernel function K, by the consistency of estimators $(\hat{a}, \hat{c})$ and by the convergence in probability of $\hat{V}_{in}$ to V_i uniformly on i, one can focus on the following term

$$\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, a_0, \hat{m}_i(., a_0, c_0)) \right]' V_i B(y_i, a_0, \hat{m}_i(., a_0, c_0)) \right] =$$
$$\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, a_0, m_i(., a_0, c_0)) \right]' V_i \varepsilon_i \right] + \tag{4.50}$$
$$+ \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, a_0, \hat{m}_i(., a_0, c_0)) - \nabla_{(a,c)} B(y_i, a_0, m_i(., a_0, c_0)) \right]' V_i \varepsilon_i \right] + \tag{4.51}$$
$$+ \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, a_0, m_i(., a_0, c_0)) \right]' V_i \left[m_i(., a_0, c_0) - \hat{m}_i(., a_0, c_0) \right] \right] + \tag{4.52}$$
$$+ \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{iQ} \left[\nabla_{(a,c)} B(y_i, a_0, \hat{m}_i(., a_0, c_0)) - \nabla_{(a,c)} B(y_i, a_0, m_i(., a_0, c_0)) \right]' V_i \left[m_i(., a_0, c_0) - \hat{m}_i(., a_0, c_0) \right] \right] \tag{4.53}$$

Central limit theorem applied to (4.50) shows that this term converges to normal distribution with asymptotic variance given by Σ. The convergence to zero in probability of term (4.52) is similar to term (4.51) and hence we omit it.

Convergence in probability to zero of term (4.53)

This vector of dimension $(J+1)\times 1$ can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\begin{bmatrix} \left(\frac{\partial m_{1i}(.,a_{10},c_0)}{\partial a_1}-\frac{\partial \hat{m}_{1i}(.,a_{10},c_0)}{\partial a_1}\right)\sum_{j=1}^{J} V_{i,1j}\left[m_{ij}(.,a_0,c_0)-\hat{m}_{ij}(.,a_0,c_0)\right] \\ \vdots \\ \left(\frac{\partial m_{Ji}(.,a_{J0},c_0)}{\partial a_J}-\frac{\partial \hat{m}_{Ji}(.,a_{J0},c_0)}{\partial a_J}\right)\sum_{j=1}^{J} V_{i,Jj}\left[m_{ij}(.,a_0,c_0)-\hat{m}_{ij}(.,a_0,c_0)\right] \\ \sum_{j=1}^{J}\sum_{s=1}^{J}\left(\frac{\partial m_{ji}(.,a_{j0},c_0)}{\partial c}-\frac{\partial \hat{m}_{ji}(.,a_{j0},c_0)}{\partial c}\right) V_{i,js}\left[m_{is}(.,a_0,c_0)-\hat{m}_{is}(.,a_0,c_0)\right] \end{bmatrix} \tag{4.54}$$

where $V_{i,js}$ represents the element (j,s) of matrix V_i. Let introduce the following notation for the estimated and true conditional mean functions. Let $m_{ij}(.,a_j,c)=r_{ij}(.,a_j,c)/f_{i,c}(.,c)$ and $\hat{m}_{ij}(.,a_j,c)=\hat{r}_{ij}(.,a_j,c)/\hat{f}_{i,c}(.,c)$ and define

$$C_{ij}(.,a_0,c_0)=\frac{1}{(n-1)\,h_n}\sum_{m\neq i}\left(\varphi_j(x_i-c_0z_i)-(w_{mj}-a_jz_m)\right)K\left(\frac{(x_i-c_0z_i)-(x_m-c_0z_m)}{h_n}\right)$$

Let consider the element corresponding to the derivative with respect to parameter a_s, which can be rewritten as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\left(\frac{\partial m_{si}(.,a_{s0},c_0)}{\partial a_s}-\frac{\partial \hat{m}_{si}(.,a_{s0},c_0)}{\partial a_s}\right)\sum_{j=1}^{J} V_{i,sj}\frac{1}{\hat{f}_{i,c}(.,c_0)}C_{ij}(.,a_0,c_0)\leq$$

$$\leq \sup_i\left|\frac{1}{\hat{f}_{i,c}(.,c_0)f_{i,c}(.,c_0)}\right|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\left(\frac{\partial r_{si}(.,a_{s0},c_0)}{\partial a_s}-\frac{\partial \hat{r}_{si}(.,a_{s0},c_0)}{\partial a_s}\right)\sum_{j=1}^{J} V_{i,sj}C_{ij}(.,a_0,c_0)+ \tag{4.55}$$

$$+\sup_i\left|\frac{1}{\hat{f}_{i,c}(.,c_0)^2 f_{i,c}(.,c_0)}\right|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\left(\frac{\partial r_{si}(.,a_{s0},c_0)}{\partial a_s}-\frac{\partial \hat{r}_{si}(.,a_{s0},c_0)}{\partial a_s}\right)\left(f_{i,c}(.,c_0)-\hat{f}_{i,c}(.,c_0)\right)\times \tag{4.56}$$

$$\times\sum_{j=1}^{J} V_{i,sj}C_{ij}(.,a_0,c_0)+$$

$$+\sup_i\left|\frac{\partial r_{si}(.,a_{s0},c_0)/\partial a_s}{\hat{f}_{i,c}(.,c_0)f_{i,c}(.,c_0)^2}\right|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\left(f_{i,c}(.,c_0)-\hat{f}_{i,c}(.,c_0)\right)\sum_{j=1}^{J} V_{i,sj}C_{ij}(.,a_0,c_0)+ \tag{4.57}$$

$$+\sup_i\left|\frac{\partial r_{si}(.,a_{s0},c_0)/\partial a_s}{\hat{f}_{i,c}(.,c_0)\bar{f}_{i,c}(.,c_0)^3}\right|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\left(f_{i,c}(.,c_0)-\hat{f}_{i,c}(.,c_0)\right)^2\sum_{j=1}^{J} V_{i,sj}C_{ij}(.,a_0,c_0) \tag{4.58}$$

where $\bar{f}_{i,c}$ lies between $\hat{f}_{i,c}$ and $f_{i,c}$. The factors in the expressions above are bounded in probability since $\hat{f}_{i,c}(.,c_0)$ converges in probability to $f_{i,c}(.,c_0)$ uniformly in i under the above assumptions, which is uniformly bounded away from zero. Also,

$$\sup_i|\partial r_{si}(.,a_{s0},c_0)/\partial a_s| = \sup_i|-E\left(z|x_i-c_0z_i\right)f_{i,c}\left(x_i-c_0z_i,c_0\right)|$$

is uniformly bounded in probability by the continuity of the conditional mean functions and the density of the index and by the compactness of $Q\times\mathcal{A}\times\mathcal{C}$. In order to show that term (4.55) converges to zero in probability, apply the Markov and Cauchy inequality to

obtain

$$\Pr\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J} I_{iQ}\left|\left(\frac{\partial r_{si}(.,a_{s0},c_0)}{\partial a_s}-\frac{\partial \hat{r}_{si}(.,a_{s0},c_0)}{\partial a_s}\right)V_{i,sj}C_{ij}(.,a_0,c_0)\right|>\varepsilon\right)\leq$$

$$\leq\sum_{j=1}^{J}\frac{\sqrt{n}}{\varepsilon}\left\{I_Q E\left(C_{ij}(.,a_0,c_0)^2\right)E\left(V_{i,sj}^2\right)E\left(\left|\frac{\partial r_{si}(.,a_{s0},c_0)}{\partial a_s}-\frac{\partial \hat{r}_{si}(.,a_{s0},c_0)}{\partial a_s}\right|^2\right)\right\}^{1/2} \tag{4.59}$$

It can be shown (see Ichimura and Lee (1991)) that if function φ and f are q continuously differentiable and a kernel function of order q is used, then

$$\begin{aligned} E\left(C_{ij}(.,a_0,c_0)^2\right) &= E\left(Var\left(C_{ij}(.,a_0,c_0)|i\right)+E^2\left(C_{ij}(.,a_0,c_0)|i\right)\right) \\ &= O\left(\frac{1}{(n-1)h_n}\right)+O\left(h_n^{2q}\right) \end{aligned}$$

Analogously, it can be shown that

$$E\left(\left(\frac{\partial r_{si}(.,a_{s0},c_0)}{\partial a_s}-\frac{\partial \hat{r}_{si}(.,a_{s0},c_0)}{\partial a_s}\right)^2\right)=O\left(\frac{1}{(n-1)h_n}\right)+O\left(h_n^{2q}\right)$$

Therefore, in order for term (4.59) to converge to zero in probability the following conditions for the sequence of the bandwidth need to be satisfied: $h_n\to 0, \sqrt{n}h_n\to\infty$ and $\sqrt{n}h_n^q\to 0$ as $n\to\infty$. It can be shown analogously that under the same conditions for the bandwidth, term (4.57) converges to zero in probability. Regarding terms (4.56) and (4.58), they converge to zero in probability if $n^{2/3}h_n\to\infty$ and $n^{1/3}h_n^q\to 0$ as $n\to\infty$, which are indeed implied by the above bandwidth conditions. When one considers the elements of vector (4.54) that involve the derivative with respect to parameter c, the above derivations should change slightly since in this case the derivative of the nonparametric conditional mean involve the derivative of the kernel function. Therefore, in this case it can be shown that

$$E\left(\left(\frac{\partial r_{si}(.,a_{s0},c_0)}{\partial c}-\frac{\partial \hat{r}_{si}(.,a_{s0},c_0)}{\partial c}\right)^2\right)=$$

$$E\left(\left(\frac{\partial\left[E\left(z|x_i-c_0z_i\right)f_{i,c}(.,c_0)\right]}{\partial c}-\frac{\partial\left[\hat{E}\left(z|x_i-c_0z_i\right)\hat{f}_{i,c}(.,c_0)\right]}{\partial c}\right)^2\right)=$$

$$O\left(\frac{1}{(n-1)h_n^3}\right)+O\left(h_n^{2q}\right)$$

for which we require an additional order of differentiability of the conditional mean functions φ and $f_c(.,c)$ (order of differentiability $q+1$) so that the bias can still have this order. Applying again the Markov and Cauchy inequality in this case, the corresponding expression is bounded above by terms which are

$$n\left[O\left(\frac{1}{(n-1)\,h_n^3}\right)+O\left(h_n^{2q}\right)\right]\left[O\left(\frac{1}{(n-1)\,h_n}\right)+O\left(h_n^{2q}\right)\right] =$$
$$n\left[O\left(\frac{1}{(n-1)\,h_n^2}\right)+O\left(h_n^{2q-1}\right)\right]\left[O\left(\frac{1}{(n-1)\,h_n^2}\right)+O\left(h_n^{2q+1}\right)\right] =$$

which converges to zero in probability if he following conditions for the sequence of the bandwidth need to be satisfied: $h_n \to 0$, $\sqrt{n}h_n^2 \to \infty$ and $\sqrt{n}h_n^q \to 0$ as $n \to \infty$.

Convergence to zero in probability of term (4.51)

This $(J+1)\times 1$ term can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{iQ}\begin{bmatrix} \left(\frac{\partial m_{1i}(.,a_{10},c_0)}{\partial a_1}-\frac{\partial \hat{m}_{1i}(.,a_{10},c_0)}{\partial a_1}\right)\sum_{j=1}^{J} V_{i,1j}\varepsilon_{ij} \\ \vdots \\ \left(\frac{\partial m_{Ji}(.,a_{J0},c_0)}{\partial a_J}-\frac{\partial \hat{m}_{Ji}(.,a_{J0},c_0)}{\partial a_J}\right)\sum_{j=1}^{J} V_{i,Jj}\varepsilon_{ij} \\ \sum_{j=1}^{J}\sum_{s=1}^{J}\left(\frac{\partial m_{ji}(.,a_{j0},c_0)}{\partial c}-\frac{\partial \hat{m}_{ji}(.,a_{j0},c_0)}{\partial c}\right)V_{i,js}\varepsilon_{is} \end{bmatrix}$$

Then, the elements corresponding to the derivatives corresponding to parameter c in the vector above can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{s=1}^{J} I_{iQ}\left(\frac{\partial r_{ij}(.,a_{j0},c_0)/\partial c}{f_{i,c}(.,c_0)}-\frac{\partial \hat{r}_{ij}(.,a_{j0},c_0)/\partial c}{\hat{f}_{i,c}(.,c_0)}\right)V_{i,js}\varepsilon_{is} - \tag{4.60}$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{s=1}^{J} I_{iQ}\left(\frac{r_{ij}(.,a_{j0},c_0)}{f_{i,c}(.,c_0)}\frac{\partial f_{i,c}(.,c_0)}{\partial c}-\frac{\hat{r}_{ij}(.,a_{j0},c_0)}{\hat{f}_{i,c}(.,c_0)}\frac{\partial \hat{f}_{i,c}(.,c_0)}{\partial c}\right)V_{i,js}\varepsilon_{is} \tag{4.61}$$

The convergence to zero in probability of term (4.61) can be analogously shown as term (4.60). By doing a Taylor's series expansion in (4.60) for some value of $\bar{f}_{i,c}$ between

$f_{i,c}(.,c_0)$ and $\hat{f}_{i,c}(.,c_0)$, it can be shown that in order to show the following four conditions

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{s=1}^{J} I_{iQ}\left(\frac{\partial r_{ij}(.,a_{j0},c_0)}{\partial c}-\frac{\partial \hat{r}_{ij}(.,a_{j0},c_0)}{\partial c}\right)\frac{1}{f_{i,c}(.,c_0)}V_{i,js}\varepsilon_{is}=o_p(1) \tag{4.62}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{s=1}^{J} I_{iQ}\left(\frac{\partial r_{ij}(.,a_{j0},c_0)}{\partial c}-\frac{\partial \hat{r}_{ij}(.,a_{j0},c_0)}{\partial c}\right)\left(\frac{1}{f_{i,c}(.,c_0)}-\frac{1}{\hat{f}_{i,c}(.,c_0)}\right)V_{i,js}\varepsilon_{is}=o_p(1) \tag{4.63}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{s=1}^{J} I_{iQ}\left(\frac{f_{i,c}(.,c_0)-\hat{f}_{i,c}(.,c_0)}{f_{i.c}(.,c_0)^2}\right)\frac{\partial r_{ij}(.,a_{j0},c_0)}{\partial c}V_{i,js}\varepsilon_{is}=o_p(1) \tag{4.64}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{s=1}^{J} I_{iQ}\left(f_{i,c}(.,c_0)-\hat{f}_{i,c}(.,c_0)\right)^2\frac{1}{\bar{f}_{i,c}(.,c_0)^3}\frac{\partial r_{ij}(.,a_{j0},c_0)}{\partial a_j}V_{i,js}\varepsilon_{is}=o_p(1) \tag{4.65}$$

The convergence to zero in probability of term (4.64) is similar to term (4.62) and hence omitted. To show the convergence in probability of term (4.62) note that

$$\frac{1}{f_{i,c}(.,c_0)}\left(\frac{\partial r_{ij}(.,a_{j0},c_0)}{\partial c}-\frac{\partial \hat{r}_{ij}(.,a_{j0},c_0)}{\partial c}\right)=\frac{1}{(n-1)h_n^2}\sum_{m\neq i}\psi_{im}^j$$

where $\psi_{im} = \frac{1}{f_{i,c}(.,c_0)}\left[(z_i-z_m)\,K'\left(\frac{(x_m-c_0z_m)-(x_i-c_0z_i)}{h_n}\right)-h_n^2\frac{\partial E(z|x_i-c_0z_i)f_{i,c}(.,c_0)}{\partial c}\right]$

Then the variance of term (4.62) can be written as

$$\frac{1}{n(n-1)^2h_n^4}E\left(I_{iQ}\left(\sum_{j=1}^{J}\sum_{s=1}^{J}\sum_{i=1}^{n}\sum_{i\neq m}\psi_{im}^{j}V_{i,js}\varepsilon_{is}\right)^2\right) =$$

$$= \frac{1}{n(n-1)^2h_n^4}E\left(I_{iQ}J^3\left(\sum_{i=1}^{n}\sum_{i\neq m}\psi_{im}^{j}V_{i,js}\varepsilon_{is}\right)\left(\sum_{i=1}^{n}\sum_{i\neq m}\psi_{im}^{j}V_{i,jl}\varepsilon_{il}\right)\right) +$$

$$+\frac{1}{n(n-1)^2h_n^4}E\left(I_{iQ}J^2\left(\sum_{i=1}^{n}\sum_{i\neq m}\psi_{im}^{j}V_{i,js}\varepsilon_{is}\right)^2\right)$$

$$+\frac{1}{n(n-1)^2h_n^4}E\left(I_{iQ}J^2\left(\sum_{i=1}^{n}\sum_{i\neq m}\psi_{im}^{j}V_{i,js}\varepsilon_{is}\right)\left(\sum_{i=1}^{n}\sum_{i\neq m}\psi_{im}^{s}V_{i,sj}\varepsilon_{ij}\right)\right)$$

$$= \frac{(n-2)}{(n-1)\,h_n^4}E\left(I_{iQ}\left[J^3\psi_{im}^{j}\psi_{ik}^{j}V_{i,js}V_{i,jl}\varepsilon_{is}\varepsilon_{il} + J^2\psi_{im}^{j}\psi_{ik}^{j}V_{i,js}^2\varepsilon_{is}^2 + J^2\psi_{im}^{j}\psi_{ik}^{s}\varepsilon_{is}\varepsilon_{ij}V_{i,js}V_{i,sj}\right]\right) +$$

$$+\frac{1}{(n-1)\,h_n^4}E\left(I_{iQ}\left[J^3\left(\psi_{im}^{j}\right)^2V_{i,js}V_{i,jl}\varepsilon_{is}\varepsilon_{il} + J^2\left(\psi_{im}^{j}\right)^2V_{i,js}^2\varepsilon_{is}^2 + J^2\left(\psi_{im}^{j}\right)^2\varepsilon_{is}\varepsilon_{ij}V_{i,js}V_{i,sj}\right]\right)$$

$$+\frac{1}{(n-1)\,h_n^4}E\left(I_{iQ}\left[J^3\psi_{im}^{j}\psi_{mi}^{j}V_{i,js}V_{m,jl}\varepsilon_{is}\varepsilon_{ml} + J^2\psi_{im}^{j}\psi_{mi}^{j}V_{i,js}V_{m,js}\varepsilon_{is}\varepsilon_{ms} + J^2\psi_{im}^{j}\psi_{mi}^{s}V_{i,js}V_{m,sj}\varepsilon_{is}\varepsilon_{rr}\right.\right.$$

where $i, m, k \in \{1, ..., n\}$ are the subindices for different observations and are different and $j, s, l \in \{1, ..., J\}$ are the subindices for the different equations.

It can be shown that under the above conditions $E\left(\frac{1}{h_n^2}\psi_{im}^{j}|i\right) = O(h_n^q)$ for all $j \in \{1, ..., J\}$, therefore the first term in the above expression converges to zero in probability if $h_n \to 0$ while the second and the third terms converge to zero in probability if $nh_n^4 \to \infty$ as $n \to \infty$

Regarding the elements corresponding to the derivatives with respect to parameter a_s for $s = \{1, .., J\}$, the convergence in probability to zero can be shown in a very similar way. The only difference is that analogous of (4.62) with respect to a_s requires the following definition of ψ_{im}

$$\frac{1}{f_{i,c}(.,c_0)}\left(\frac{\partial r_{is}(.,a_{s0},c_0)}{\partial a_s} - \frac{\partial \hat{r}_{is}(.,a_{s0},c_0)}{\partial a_s}\right) = \frac{1}{(n-1)h_n}\sum_{m\neq i}\psi_{im}$$

$$\text{where } \psi_{im} = \left[z_mK\left(\frac{(x_m - c_0z_m) - (x_i - c_0z_i)}{h_n}\right) - h_nE(z|x_i - c_0z_i)f_{i,c}(.,c_0)\right]$$

In this case the convergence to zero of the analogous term of (4.62) requires a less strict condition on the bandwidth, i.e. $nh_n^2 \to \infty$ as $n \to \infty$.

The convergence in probability to zero of terms (4.63) and (4.65) can be done in a similar way as it was shown that the elements of the $(J+1) \times 1$ vector in (4.53) converge to zero in probability (see above in this proof). ■

4.12 Figures

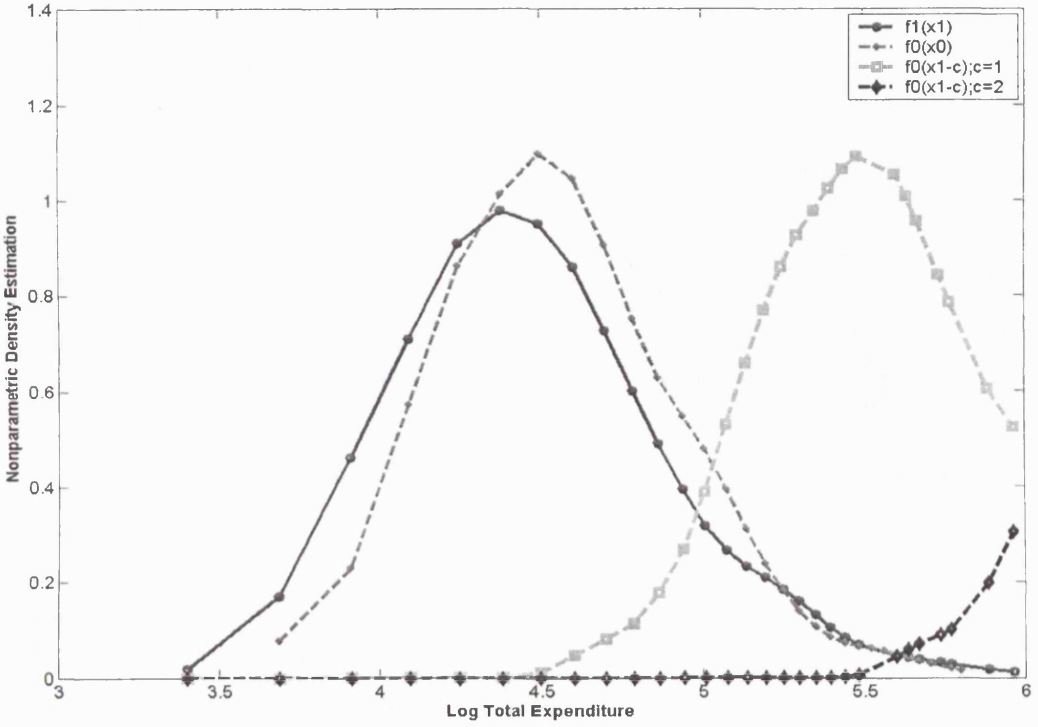

Figure 4.1: Nonparametric Kernel Densities for Log Total Expenditure for different demographic groups and different values of parameter c

Note: Data Drawn from 1980-1982 Family Expenditure Surveys for couples with one kid ($z = 0$) and two kids ($z = 1$). Line $f1(x1)$ plots the nonparametric density of log total expenditure for demographic group with $z = 1$ over the points of the support of $\Omega_X^{(1)}$; line $f0(x0)$ plots the nonparametric density of log total expenditure for demographic group with $z = 0$ over the points of the support of $\Omega_X^{(1)}$; line $f0(x1 + c)$ for $c = \{-1, -2\}$ plots the nonparametric density of log total expenditure for demographic group $z = 0$ over the points of the support of $X_1 - c$. Gaussian kernel used and Silverman's optimal bandwidth.

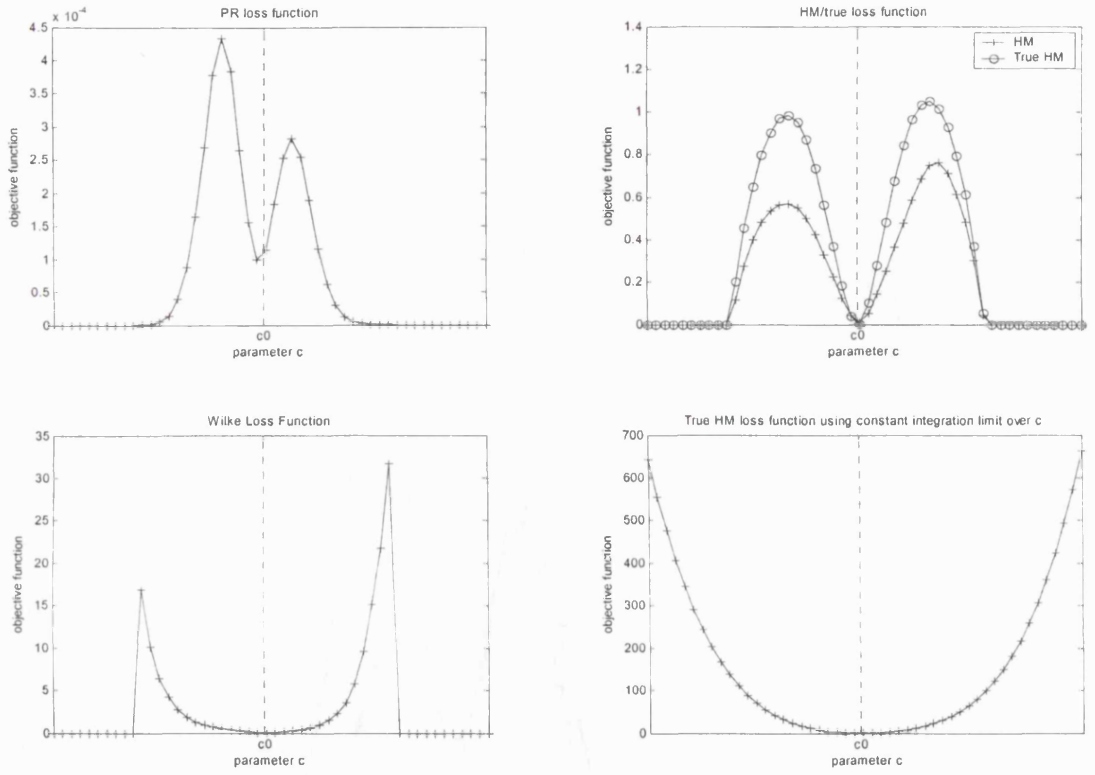

Figure 4.2: Loss functions for parameter c ($L(c|a_0)$) proposed by Pinkse and Robinson, Hardle and Marron, Wilke and the Loss function using knowledge of m function ($c_0 = -0.3$)

Note: For each value of the parameter c, the above functions compute the overlap of the supports $\Omega_X^{(1)} \cap \Omega_X^{(0)} + c$ from the observed data. The integration limits of the objective functions above are set to cover this intersection. The integration to compute the loss function is done using middlepoint approximation for integrals. The true loss function of Hardle and Marron uses the true known function m instead of its nonparametric estimation as in the HM loss function. The weight function $w(x) = 1$ if $x \in \{\Omega_X^{(1)} \cap \Omega_X^{(0)} + c\}$

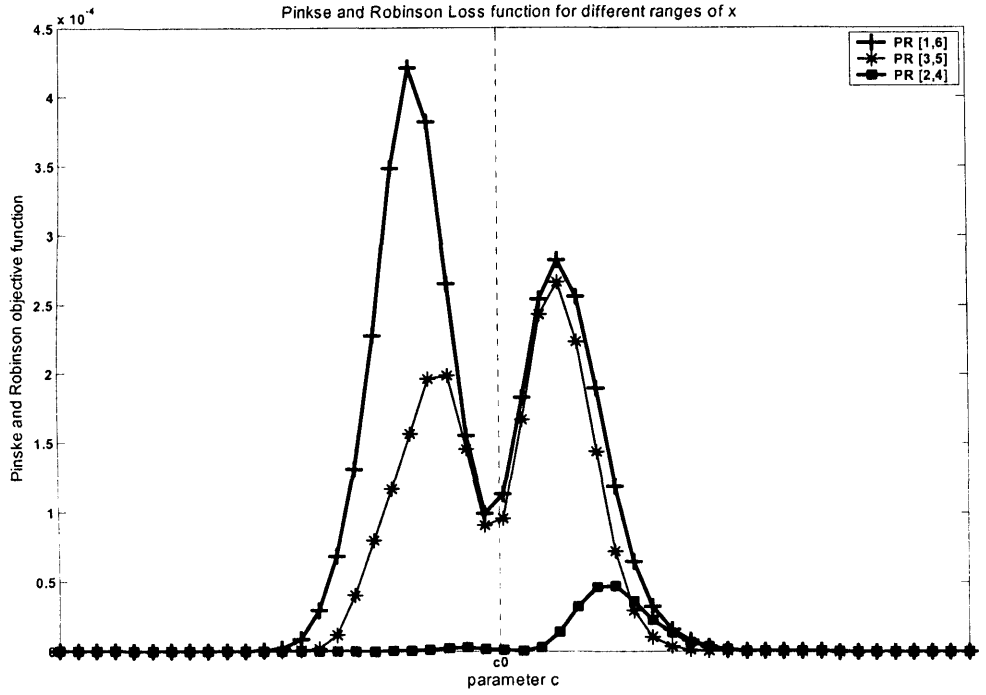

Figure 4.3: PR Loss Function $L^{PR}(c|a_0)$ conditioned on the true value of parameter a, for different values of the integration limits $[\underline{x}, \bar{x}]$

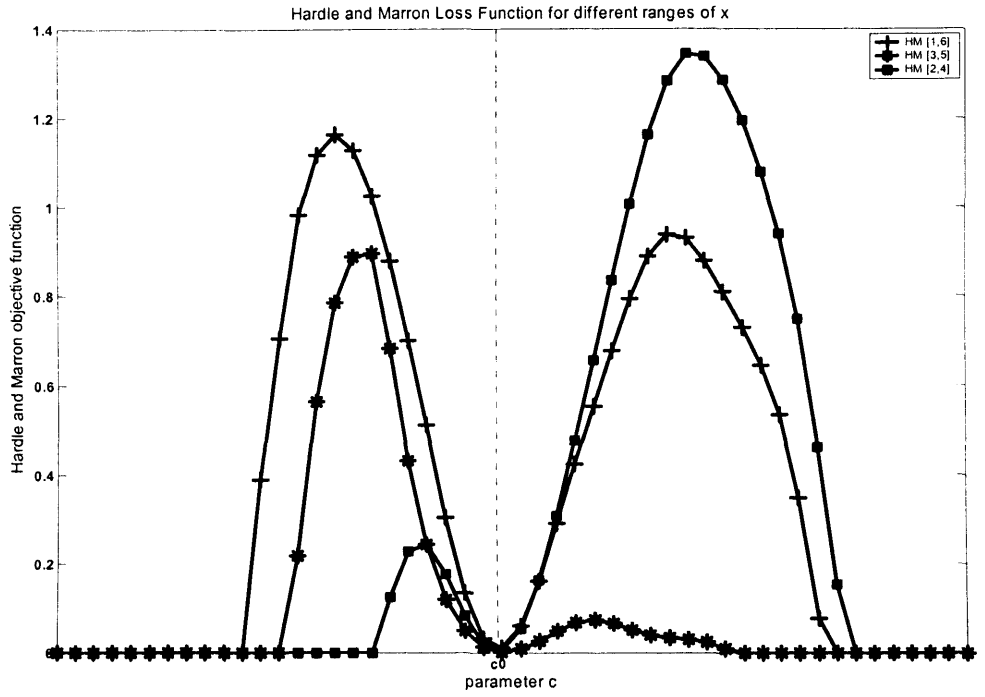

Figure 4.4: HM Loss Function $L^{HM}(c|a_0)$ conditioned on the true value of parameter a, for different values of the integration limits $[\underline{x}, \bar{x}]$

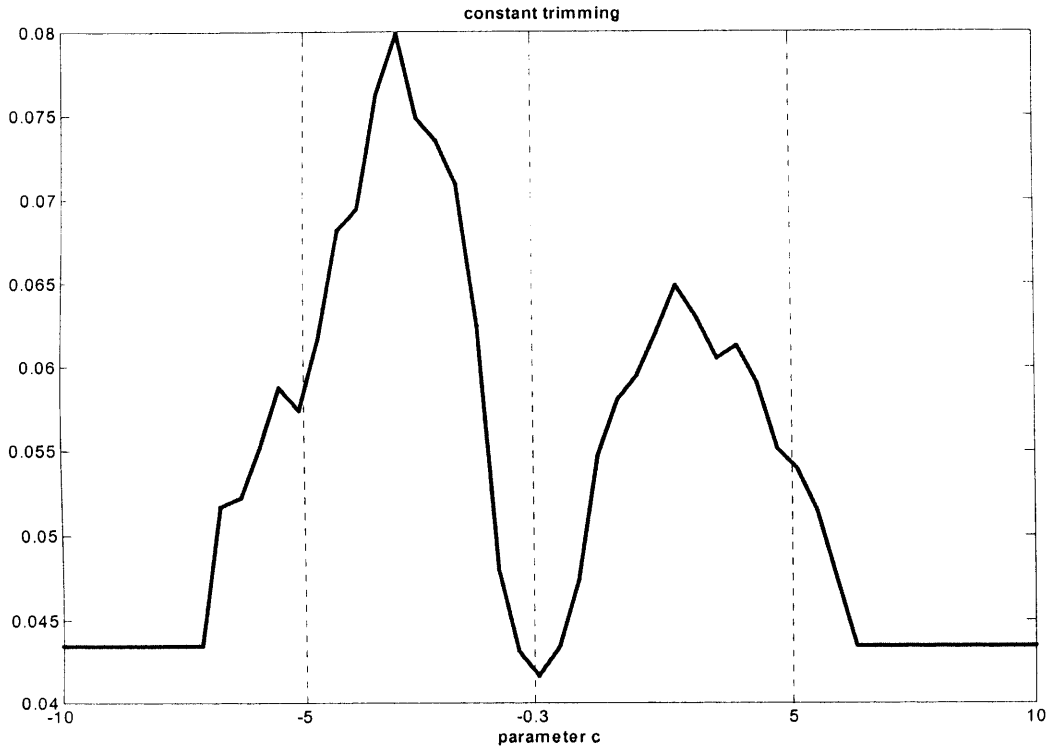

Figure 4.5: SLS Objective function as a function of c -$L^{SLS}(c|a_0, h_0)$ where h_0 is the optimal CV-bandwidth for (a_0, c_0) - for simulated data for one good

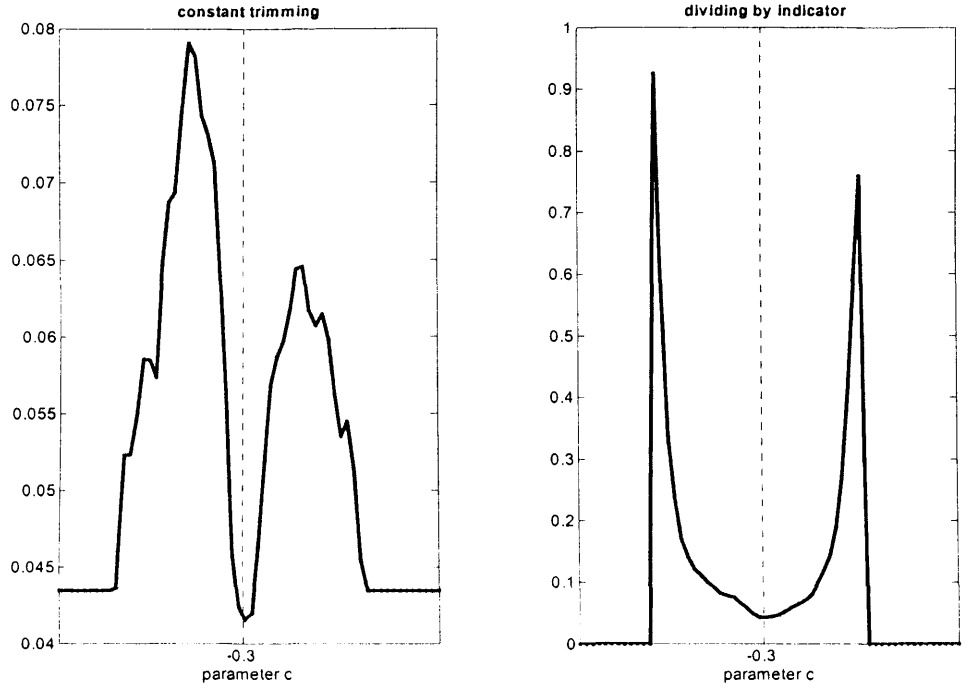

Figure 4.6: SLS Objective function as a function of c -$L^{SLS}(c|a_0, h_0)$ where h_0 is the optimal CV-bandwidth for (a_0, c_0) - for simulated data for one good.

Left Graph: objective function including a constant trimming of the 2% of the smallest densities (expression (4.12)); Right Graph: objective function dividing by the number of observation where the estimated density of the index does not attain its lower bound (expression (4.14)).

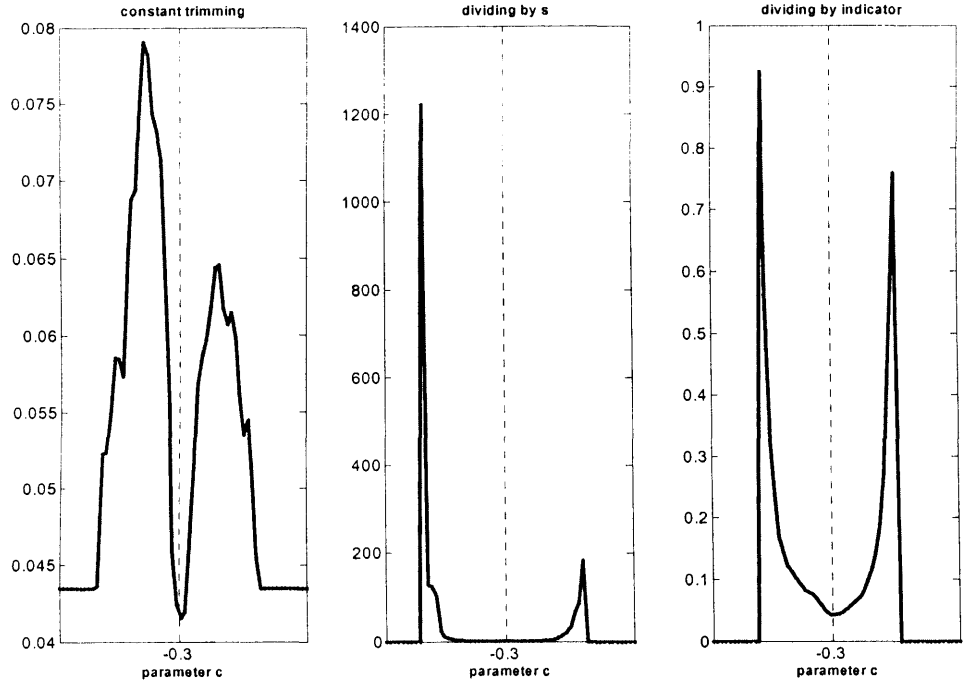

Figure 4.7: SLS Objective function as a function of c -$L^{SLS}(c|a_0, h_0)$ where h_0 is the optimal CV-bandwidth for (a_0, c_0) - for simulated data for one good.

Left Graph: objective function including a constant trimming of the 2% of the smallest densities (expression (4.12)); Right Graph: objective function dividing by the number of observation where the estimated density of the index does not attain its lower bound (expression (4.14)). Middle Graph: objective function dividing by continuous function s as in (expression (4.16)).

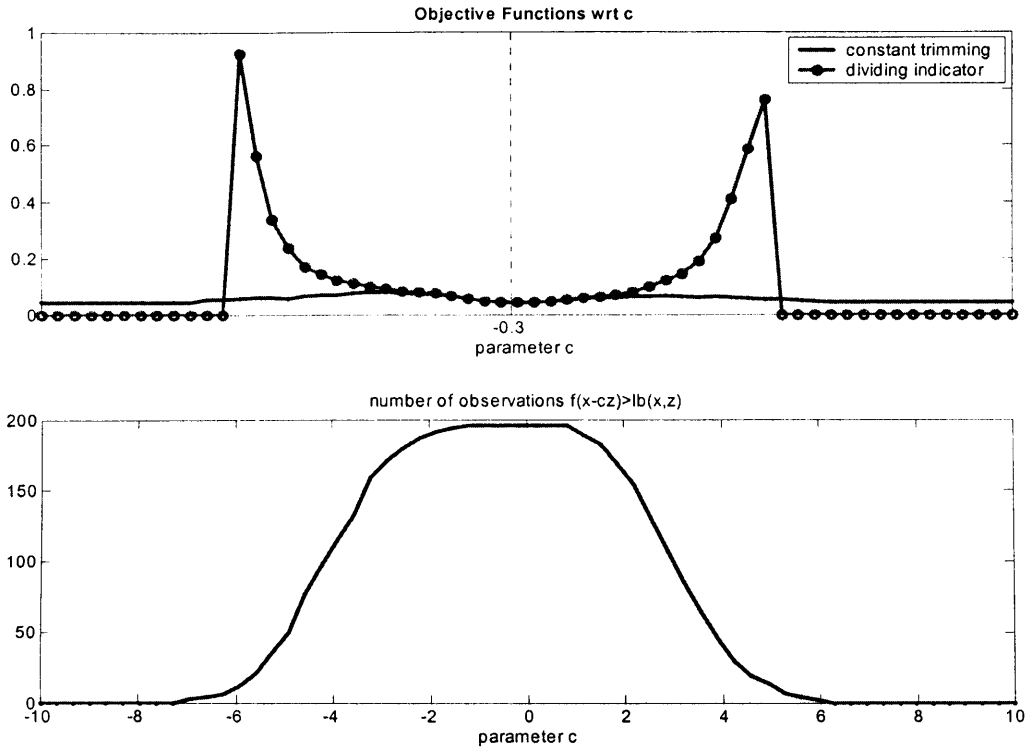

Figure 4.8: $L(c|a_0, h_0)$ SLS Objective function as a function of c and number of observations where estimated density is above its lower bound. Simulated Data One good.

(Top Graph) SLS Objective function as a function of c -$L^{SLS}(c|a_0, h_0)$ where h_0 is the optimal CV-bandwidth for (a_0, c_0) - for simulated data for one good for (i) objective function including a constant trimming of the 2% of the smallest densities (expression (4.12)); (ii) objective function dividing by the number of observation where the estimated density of the index does not attain its lower bound (expression (4.14)). (Bottom Graph) Number of observations such that $\hat{f}_{X-cZ}(x - cz)$ is strictly greater than $\widehat{lb}(x_i, z_i)$

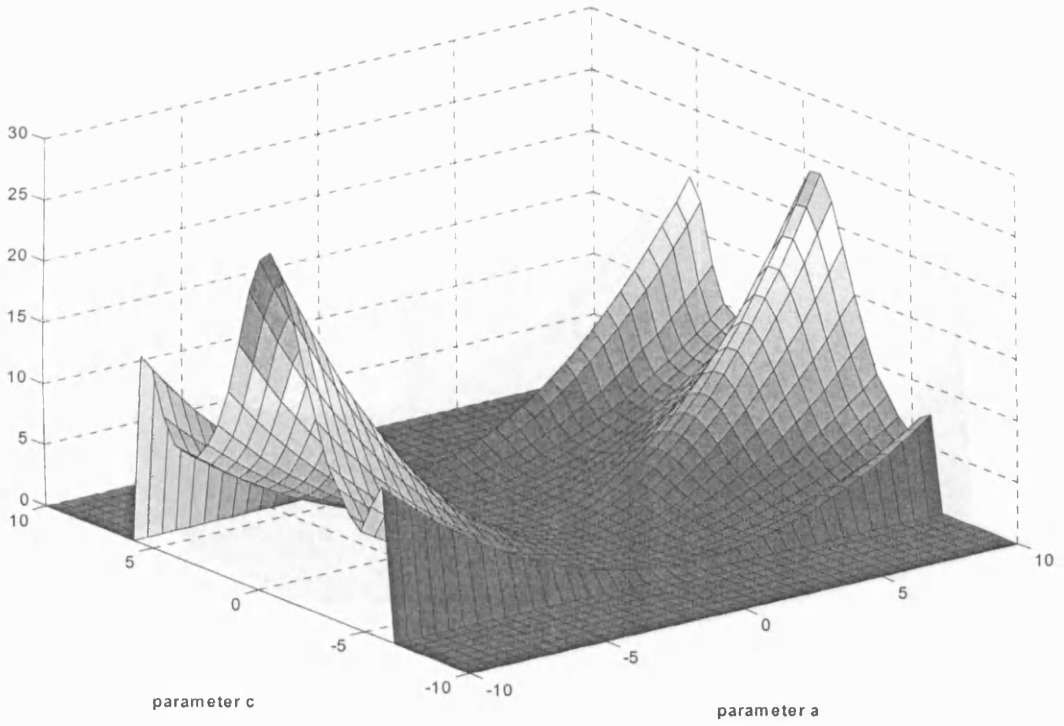

Figure 4.9: Objective Function $\hat{L}_2(a, c|\vec{h}_0)$ as a function of both parameters using simulated data for one good

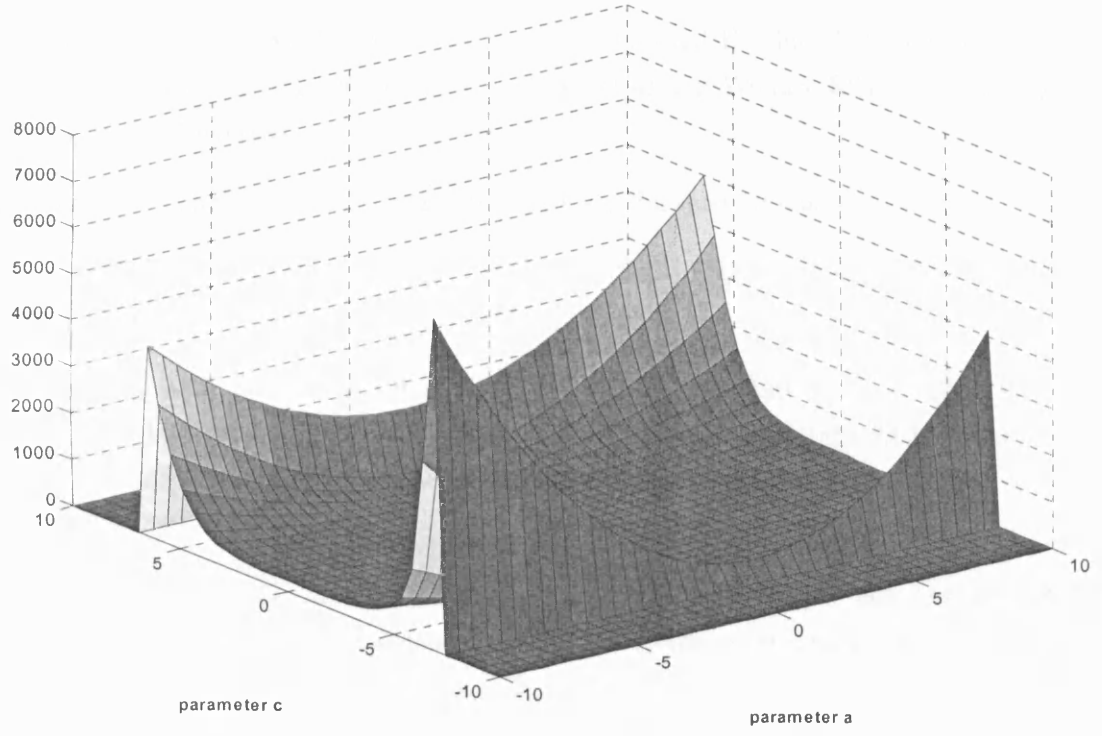

Figure 4.10: Objective Function $\hat{L}_3(a, c|\hat{h}_0)$ as a function of both parameters using simulated data for one good

Bibliography

[1] ACKERBERG, D. AND M. RYSMAN (2001), "Unobserved Product Differentiation in Discrete Choice Models: Estimating Price Elasticities and Welfare Effects", Working Paper, Boston University.

[2] AMEMIYA, T. (1985), "Advanced Econometrics", Harvard University Press

[3] ANDERSON, S., DE PALMA, A. AND J.THISSE (1992), "Discrete Choice Theory of Product Differentiation", Cambridge: MIT Press.

[4] ANGRIST, G. AND A. KRUEGER (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples", Journal of the American Statistical Association, Vol. 87, No. 418, pp. 328-336

[5] ARELLANO, M. AND C. MEGHIR (1992), "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets", The Review of Economics Studies, Vol. 59, Issue 3, pp.537-559

[6] BAJARI, P. AND L. BENKARD (2003), "Discrete Choice Models as Structural Models of Demand: Some Economic Implications of Common Approaches", mimeo.

[7] BAJARI, P. AND L. BENKARD (2005), "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach", mimeo.

[8] BEN-AKIVA, M., D. BOLDUC AND J. WALKER (2003), "Specification, Identification and Estimation of the Logit Kernel (or Continuous Mixed Logit) Model", Working Paper, MIT.

[9] BERRY, S. (1994), "Estimating Discrete-Choice Models of Product Differentiation", RAND Journal of Economics, vol. 25 (2)

[10] BERRY, S., LEVINSHON, J. AND A. PAKES (1995), "Automobile Prices in Market Equilibrium" Econometrica, 63 (4), pp. 841-90

[11] BERRY, S., LEVINSHON, J. AND A. PAKES (1998), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market", Journal of Political Economy, 112(1), p. 68-105

[12] BERRY, S. AND A. PAKES (2003), "The Pure Characteristics Discrete Choice Model of Differentiated Products Demand", mimeo.

[13] BLACKORBY, C. AND D. DONALDSON (1994), "Measuring the Cost of Children: a Theoretical Framework" in The Measurement of Household Welfare, ed. by R. Blundell, I. Preston and I. Walker, Chapter 2, pp. 51-69. Cambridge University Press.

[14] BLUNDELL, R. X. CHEN AND D. KRISTENSEN (2003), "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves", cemmap Working paper CWP 15/03

[15] BLUNDELL, R. AND A. DUNCAN (1998), "Kernel Regression in Empirical Microeconomics", Journal of Human Resources, 33, pp. 62-87

[16] BLUNDELL, R., A. DUNCAN AND K. PENDAKUR (1998), "Semiparametric Estimation and Consumer Demand", Journal of Applied Econometrics, 13, pp.435-461

[17] BROWN, J. AND H. ROSEN (1982), "On the Estimation of Structural Hedonic Price Models", Econometrica 50, pp.765-769

[18] BROWN, B.W. AND M.B. WALKER (1989), "The Random Utility Hypothesis and Inference in Demand Systems", Econometrica (57), pp. 815-829.

[19] CAPLIN, A. AND B. NALEBUFF (1991), "Aggregation and Imperfect Competition: On the Existence of Equilibrium", Econometrica, 59, pp.1-23.

[20] CARROLL, C.D. AND D. N. NEIL (1993), "Saving and Growth: A Reinterpretation", NBER Working papers, No. 4470

[21] CHEN, X. , HONG, H. AND E. TAMER (2004), "Measurement Error with Auxiliary Data", forthcoming Review of Economic Studies

[22] CHERNOZHUKOV, V. AND H. HONG (2003), "An MCMC Approach to Classical Estimation" Journal of Econometrics, 115(2), pp. 293-346

[23] CHINTAGUNTA, P. (1994), "Investigating Logit Model Implication for Brand Positioning", Journal of Marketing Research, 31, pp. 304-11.

[24] COSSLETT, S.R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model", Econometrica 51(3), pp.765-782

[25] CROSS, P.J. AND MANSKI, C. (2002), "Regressions, Short and Long", Econometrica, vol. 70, pp. 357-368

[26] CURRIE, J. AND A. YELOWITZ, (2002) "Are Public Housing Projects Good for Kids?", Journal of Public Economics 75, pp.99-124

[27] DEE, T.S. AND W.N. EVANS (1997), "Teen Drinking and Educational attainment: Evidence from Two-Sample Instrumental Variables", NBER No. 6082

[28] EKELAND, I., J. HECKMAN AND L. NESHEIM (2004), "Identification and Estimation of Hedonic Models", Journal of Political Economy, 112 (S1):S60-S109

[29] ELROD, T. (1988), "Choice Map: inferring a product Market Map for Panel Data", Marketing Science, 7, pp. 21-40.

[30] ELROD, T. AND M.P. KEANE (1995), "A factor-analytic Probit Model for Representing the Market Structure in Panel Data", Journal of Marketing Research, 32, pp. 1-16.

[31] EPPLE, D. (1987), "Hedonic Prices and Implicit Markets: Estimating Demand and Supply functions for Differentiated Products", Journal of Political Economy, 95, pp. 59-80

[32] GLASSER, M. (1964) "Linear Regression Analysis with Missing Observations Among Independent Variables," Journal of the American Statistical Association, Vol. 51, pp.834-844.

[33] GOLDBERG, P. (1995) "Product Differentitation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry", Econometrica, 63(4), pp.

[34] GORMAN, W (1956), " A Possible Procedure for Analyzing Quality Differentials in the Egg-Market", Review of Economic Studies, 47, pp. 843-856

[35] GOURIEROUX, C. AND A. MONFORT (1981), "On the Problem of Missing Data in Linear Models", Review of Economic Studies, 48, pp. 579-586

[36] HARDLE, W. , HALL, P. AND H. ICHIMURA (1993), "Optimal Smoothing in Single-Index Models", Annals of Statistics, vol. 21, pp.157-178.

[37] HARDLE, W. AND M. JERISON (1988), "Cross Section Engel Curves over time", Discussion paper no. A-160. SFB 303, University of Bonn

[38] HARDLE, W. AND J.S. MARRON (1990), "Semiparametric Comparison of Regression Curves", Annals of Statistics, vol. 18, pp.63-89.

[39] HOROWITZ, J. AND MANSKI, C. (1995), "Identification and Robustness with Contaminated Data and Corrupted Data", Econometrica, vol. 63 pp.281-302

[40] HU, Y AND RIDDER, G. (2003), "Estimation of Nonlinear Models with Measurement Errors Using Marginal Information", Working paper, CLEO, University of Southern California

[41] IMBENS, G. W. AND T. LANCASTER (1994), "Combining Micro and Macro Data in Microeconometric Models", Review of Economic Studies (1994), 61- pp. 655-680

[42] ICHIMURA, H. (1993), "Semiparametric Least Squares (SLS) and weighted SLS Estimation of Single-Index Models", Journal of Econometrics 58, pp. 71-120

[43] ICHIMURA, H. (2004), "Computation of Asymptotic Distribution for Semiparametric GMM Estimators", Department of Economics, University College London, mimeo.

[44] ICHIMURA, H. AND L.F. LEE (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation", Nonparametric and Semiparametric methods in Econometrics and Statistics, Ed. W.A. Barnett, J. Powell and G. Tauchen. Cambridge University Press.

[45] ICHIMURA, H. AND THOMPSON, T.S. (1998), "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution" Journal of Econometrics, vol. 86, 269-295

[46] JUDD, K. (1998), Numerical methods in economics. MIT Press: Cambridge, MA.

[47] KEMP, G. (2000), "Semiparametric Estimation of a Logit Model", University of Essex, mimeo.

[48] KLEIN, R.W. AND SPADY, R. H. (1993), "An Efficient Semiparametric Estimator for Binary Response Models", 61(3), pp.387-421

[49] KOOPMANS, T.C. AND REIERSOL, O. (1950), "The identification of Structural Characteristics", Annals of Mathematical Statistics , vol. 21, pp. 165-181.

[50] LANCASTER, K. (1966), "A New Approach to Consumer Theory", Journal of Political Economy, 74, pp. 132-157

[51] LAWTON, W.H., E.A. SYLVESTRE AND M.S. MAGGIO (1972), "Self Modelling Nonlinear Regression", Technometrics, 14 pp. 513-532

[52] LEE, L.F (1995), "Semiparametric Maximum Likelihood Estimation of Polychotomus and Sequential Choice Models", Journal of Econometrics 65, pp. 381-428

[53] LITTLE, R.A. (1992), "Regression with Missing X's: A Review", Journal of the American Statistical Association, vol. 87 No. 420, pp.1227-1237

[54] LUSARDI, A. (1996),"Permanent Income, Current Income and Consumption: Evidence from Two Panel Data Sets", Journal of Business and Economic Statistics, 14, pp. 81-90

[55] MCFADDEN, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior" in P. Zarembka (ed.), Frontiers in Econometrics, pp. 105-142, Academic Press: New York, 1974.

[56] MC FADDEN, D. AND K. TRAIN (2000), "Mixed MNL Models for Discrete Response" Journal of Applied Econometrics, Vol. 15, Is. 5, 447-470

[57] MANSKI, C. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," Journal of Econometrics, 27, pp.313-333.

[58] MANSKI, C. (1988), "Identification of Binary Response Models", Journal of the American Statistical Association, Vol. 83, No. 403 , pp. 729-738.

[59] MANSKI, C. AND TAMER, E. (2003), "Inference on Regressions with Interval Data on a Regressor or Outcome", Econometrica, vol. 70, No. 2, pp. 519-546

[60] MATZKIN, R. (1991), "Semiparametric Estimation of Monotone and Concave Utility Function for Polychotomuous Choice Models", vol 59 (5), pp.1375-1327

[61] MATZKIN, R. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Choice and the Threshold Crossing Models" Econometrica, vol 60(2), pp.239-270

[62] NEVO, A. (2000), "Mergers with Differentiated Products: The Case of the Reasy-to-Eat Industry", The RAND Journal of Economics, 31(3)

[63] NEWEY, W. (1990), "Semiparametric Efficiency Bounds", Journal of Applied Econometrics, vol 5(2), pp.99-135

[64] NEWEY, W. AND MCFADDEN, D. (1994), "Large Sample Estimation and Hypothesis Testing", Handbook of Econometrics, vol4, Chapter 36

[65] NEWEY, W. AND POWELL, J. (2003), "Instrumental Variable Estimation of Nonparametric Models".Econometrica, vol 71, No 5, pp. 1565 - 1578

[66] PAGAN, A. AND A. ULLAH (1999), "Nonparametric Econometrics", Themes in Modern Econometrics, Cambridge University Press

[67] PETRIN, A. (2003), "Quantifying the Benefits of New Products: the Case of the Minivan", Journal of Political Economy, vol.110 no. 4

[68] PINKSE, C. AND P. ROBINSON (1995), "Pooling Nonparametric Estimates of Regression Functions with a Similar Shape" in G. Maddala, P. Phillips and T.N. Srinivasan (eds), Advances in Econometrics and Quantitative Economics, pp.172-195

[69] POWELL, J., STOCK , J. H. AND STOKER, T. (1989), "Semiparametric Estimation of Index Coefficients", Econometrica, vol. 57, No. 6, pp. 1403-1430

[70] RIDDER, G. AND MOFFIT, R. (2003), "The Econometrics of Data Combination", Chapter for the Handbook of Econometrics (forthcoming)

[71] ROBINSON, P.M. (1988), "Root-N-Consistent Semiparametric Regression", Econometrica 56, pp. 931-954

[72] ROSEN, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", Journal of Political Economy, 82 pp. 34-55

[73] ROTHENBERG, T.J. (1971), "Identification in Parametric Models" Econometrica 39, pp. 377-592

[74] RUBIN, D. R., (1974), "Characterizing the Estimation of the Parameters in Incomplete-Data Problems", Journal of the American Statistical Association. Vol. 69, No. 346, pp. 467-474.

[75] SCHENNACH, S. (2004), "Estimation of Nonlinear Models with Measurement Error", Econometrica, 72 pp.33-75

[76] SERFLING, R. J. (1980), "Approximation Theorems of Mathematical Statistics", Wiley Series in Probability and Statistics

[77] VAN DER VAART, A. W. (1998), "Asymptotic Statistics", Cambridge Series in Statistical and Probabilistic Mathematics

[78] WALKER, J. (2002), "The Mixed Logit (or Logit Kernel) Model: Dispelling Misconceptions of Identification", Transportaton Research Recond 1805, pp. 86-98.

[79] WILKE, R. (2003), "Semiparametric Estimation of Regression Functions under Shape Invariance Restrictions", ZEW Discussion Paper No. 63-54