

Identification and Estimation of GMM Models by Combining Two Data Sets¹

Hidehiko Ichimura
UCL and CEMMAP²

Elena Martinez-Sanchis
UCL and CEMMAP

April 2005

¹This work has benefitted from discussions with Manolo Arellano, Walter Beckert, Richard Blundell, Andrew Chesher, Costas Meghir, Lars Nesheim, Geert Ridder and Marcos Vera-Hernandez for which we are very grateful. We would like to thank seminar participants at the Bank of Spain, CEMFI, Simon Fraser University, Universitat Autònoma de Barcelona, University of Alicante, UCL, University of Mannheim, University of Navarra, University of Pittsburgh, University of Warwick and University of Western Ontario. We also thank the Leverhulme Foundation for the support through their support for the CEMMAP. Ichimura thanks the research support of the ESRC. Martinez-Sanchis thanks the Fundacion Ramon Areces and IFS for their financial support. All errors are our own. Address for correspondence: Elena Martinez-Sanchis, CEMMAP, Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK. E-mail: elena_m@ifs.org.uk

²CEMMAP is a joint venture of IFS and UCL supported by the Leverhulme Foundation.

Abstract

In this paper we study an incomplete data problem in which the data set at hand contains only a strict subset of the list of variables relevant for an empirical analysis. The proposed method presumes the existence of another data set which contains a subset of variables in the original data set and the missing variables in the original data set. This assumption combined with a parametric structural assumption and some joint variation assumption on the variables in the auxiliary data set are shown to be sufficient to identify the effects of missing variables as well as those of the non-missing variables in the parametric structural relationship for a wide variety of problems. The main advantage of the framework we use here is that it extends the existing method for linear-in-parameters models in the incomplete data literature to more general models. General conditions under which the proposed estimators for the identified parameters exhibit consistency and asymptotic normality are developed.

JEL codes: C25, C51, C14

Keywords: data combination, identification, nonparametric estimation, missing data

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 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 works³. For non-linear in parameter models various

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

²Imbens 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).

³Early references also include Rubin (1974), which establishes maximum likelihood factorization methods

identification issues and estimation procedures have been insightfully discussed by Ridder and Moffitt (2003). It is hard for to give sufficient conditions for global identification a very general non-linear model (similarly to the identification in GMM non-linear model with complete data).

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. 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.⁴

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 identification of the effect of non-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 (and discrete) variable and we provide sufficient conditions under which coefficients of the missing regressors are identified. For the semiparametric binary choice model, similar conditions are given to identify the coefficients of the common regressors up to scale.

After explaining the framework of the incomplete data problem that we consider in section 2 we discuss identification issues and present a general estimation method in sections 3 and 4, respectively. The asymptotic theory for this framework is established in section 5.

Monte Carlo dealing with missing data problems. These methods however do not allow one to identify the effect of the missing regressor.

⁴Our 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)).

simulation results are presented and discussed in section 6. Section 7 concludes.

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 $\theta_0 \in \Theta \subset R^K$ 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)

$$E \{ \psi(\rho_a(Z_1, Z_c, Z_2, \theta), \rho_b(Z_1, Z_c, Z_2, \theta); \theta) | X_1, X_2, X_c \} = 0 \text{ almost surely in } X_1, X_2, X_c \text{ iff } \theta = \theta_0^C \quad (1)$$

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 (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 (1) identifies.

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

$$E \{h(Z_1, Z_c; \theta) | Z_c\} = 0 \text{ almost surely in } Z_c \text{ iff } \theta = \theta_0^I \quad (2)$$

where

$$h(z_1, z_c; \theta) = \psi(q_a(z_1, z_c, \theta), q_b(z_1, z_c, \theta); \theta) \quad (3)$$

$$q_j(z_1, z_c, \theta) = \int \rho_j(z_1, z_c, z_2, \theta) g(z_2 | z_c) d\mu \text{ for } j = a, b \quad (4)$$

The function $g(z_2 | z_c)$ is typically defined via $\gamma(z_2 | z_c)$. 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) 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 (1) is the same as the parameter that solves the moment condition with incomplete data in (2) (i.e. $\theta_0^C = \theta_0^I$)

The framework in (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 .

2.1 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, 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 $f(y_1, y_c | x_c; \theta)$.⁵ The moment condition that identifies the true value of the parameters θ_0 is

⁵The analogous likelihood can be formulated replacing the role of two data sets if in fact the data sets are symmetric as formulated above.

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 data⁶ 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:

$$\begin{aligned}
\rho_a(y_1, y_c, x_c, x_2; \theta) &= \int \nabla_{\theta} f(y_1, y_c, y_2 | x_c, x_2; \theta) dy_2 \\
\rho_b(y_1, y_c, x_c, x_2; \theta) &= \int f(y_1, y_c, y_2 | x_c, x_2; \theta) dy_2 \\
\psi(\rho_a(\cdot; \theta), \rho_b(\cdot; \theta)) &= \rho_a(\cdot; \theta) / \rho_b(\cdot; \theta) \\
q_a(y_1, y_c, x_c; \theta) &= \int \rho_a(y_1, y_c, x_c, x_2; \theta) g(x_2 | x_c) d\mu = \nabla_{\theta} f(y_1, y_c | x_c; \theta) \\
q_b(y_1, y_c, x_c; \theta) &= \int \rho_b(y_1, y_c, x_c, x_2; \theta) g(x_2 | x_c) d\mu = f(y_1, y_c | x_c; \theta) \\
h(y_1, y_c, x_c; \theta) &= \psi(q_a(\cdot; \theta), q_b(\cdot; \theta)) = \nabla_{\theta} f(y_1, y_c | x_c; \theta) / f(y_1, y_c | x_c; \theta)
\end{aligned}$$

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 3.

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

$$E[\psi(Y, X; \theta) | X] = 0 \text{ almost surely in } X \text{ iff } \theta = \theta_0 \quad (5)$$

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:⁷

$$\psi(z_1, z_c, z_2; \theta) = \psi_a(z_1, z_c, \theta) - \psi_b(z_c, z_2, \theta). \quad (6)$$

Under this separability assumption, the moment condition (5) can be integrated out to become

$$E[\psi(Z_1, Z_c, Z_2; \theta) | Z_c] = E[\rho_a(Z_1, Z_c; \theta) | Z_c] - E[\rho_b(Z_c, Z_2; \theta) | Z_c] \quad (7)$$

⁶We 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.

⁷For notational convenience the following expression changes the location of arguments in function T .

and each term on the right-hand side can be examined using the two data sets at hand (See Ridder and Moffitt (2003)).

In this formulation⁸

$$\begin{aligned} h(z_1, z_c; \theta) &= q_a(z_1, z_c; \theta) - q_b(z_c; \theta) \\ q_a(z_1, z_c; \theta) &= \int \rho_a(z_1, z_c; \theta) g(z_2 | z_c) dz_2 = \rho_a(z_1, z_c; \theta) \\ q_b(z_c; \theta) &= E[\rho_b(Z_c, Z_2; \theta) | z_c] \end{aligned}$$

2.2.1 Non-linear regression

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

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 $\psi_a(Z_1, Z_c, \theta) = \psi_a(Z_1) = Y_1$ and $\psi_b(Z_c, Z_2, \theta) = m(X_c, X_2; \theta_0)$. Since

$$E(Y_1 | X_c) = E[m(X_c, X_2; \theta_0) | X_c] \tag{8}$$

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. 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

$$h(Y_1, X_c; \theta) = Y_1 - E[m(X_c, X_2; \theta_0) | X_c]$$

The moment condition in (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

⁸We could also formulate

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

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

$$\begin{aligned}
\rho_a(y_1, x_c, x_2; \theta) &= (y_1 - m(x_c, x_2; \theta)) \\
\rho_b(x_c, x_2; \theta) &= \nabla_{\theta} m(x_c, x_2; \theta) \\
\psi(\rho_a(y_1, x_c, x_2; \theta), \rho_b(x_c, x_2; \theta); \theta) &= \rho_a(y_1, x_c, x_2; \theta) \cdot \rho_b(x_c, x_2; \theta) \\
q_a(y_1, x_c; \theta) &= \left(y_1 - \int m(x_c, x_2; \theta) g(x_2|x_c) d\mu \right) \\
q_b(x_c; \theta) &= E[\nabla_{\theta} m(X_c, X_2; \theta) | X_c = x_c] = \int \nabla_{\theta} m(x_c, x_2; \theta) g(x_2|x_c) d\mu \\
h(y_1, x_c; \theta) &= \psi(q_a(y_1, x_c; \theta), q_b(x_c; \theta); \theta) = q_a(y_1, x_c; \theta) \cdot q_b(x_c; \theta)
\end{aligned}$$

Thus, for a given model, there are alternative identifying moment conditions (which can be conditional or unconditional on the common exogenous regressors⁹) by defining in a different way the functions ψ , ρ_a and ρ_b .

3 Identification Conditions

The conditions under which the global identification of θ_0 holds in (2) are specific to each model.

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 $\theta \in \Theta$

$$\int \rho_b(z_1, z_c, z_2, \theta) g(z_2|z_c) d\mu = \int \rho_b(z_1, z_c, z_2, \theta^0) g(z_2|z_c) d\mu \text{ almost surely in } (Z_1, Z_c) \quad (9)$$

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

The linear regression model with incomplete data where $m(X_c, X_2, \theta) = X_c\theta_1 + X_2'\theta_2$ 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 .¹⁰ 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

⁹In the general definition for the estimators we use unconditional moments.

¹⁰Note that this sufficient condition for identification is implicitly excluding nonlinear functions of X_c in the conditional mean model $E(Y_1|X_c, X_2; \theta)$.

the complete data model and also, when Z_2 is integrated out, in the incomplete data model.¹¹

We investigate sufficient conditions under which condition (9) holds for the parametric and semiparametric binary choice models.¹²

3.1.1 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. Consider the following model:

$$Y = 1\{\alpha_0 + X_c'\beta_0 + X_2'\gamma_0 + U > 0\} \quad (10)$$

where we denote $X = (X_c, X_2)'$. The number of elements in X is denoted by $d = d_1 + d_2$.

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 $F(\cdot|x_c, x_2)$ 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 $F(\cdot|x)$ 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 the distribution of X is tight and F is strictly monotonic.¹³ This is no longer the case when not all of the regressors are

¹¹Sufficient 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.

¹²In the rest of the paper, we consider that $\rho_b(z_1, z_c, z_2; \theta)$ is parametrically specified. Let $q(z_1, z_c) = \int \rho(z_1, z_c, z_2)g(z_2|z_c)dz_2$. 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 $E(Z_1|Z_c, Z_2)$.

¹³One could weaken this further by writing conditions explicitly in terms of the support of $X'\theta$ and that of U .

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 .

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

$$\int F(\alpha + X_c\beta + x_2'\gamma) g(x_2|X_c)d\mu = \int F(\alpha_0 + X_c\beta_0 + x_2'\gamma_0) g(x_2|X_c)d\mu \quad (11)$$

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. 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

$$\int F(\alpha + X_c\beta + x_2'\gamma|X_c, x_2) g(x_2|X_c)d\mu = \int F_0(\alpha_0 + X_c\beta_0 + x_2'\gamma_0|X_c, x_2) g(x_2|X_c)d\mu \quad (12)$$

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_1 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.

The following result complements the results of Manski and Tamer (2003). When we have a complementary data in the parametric case, multiple missing regressors are allowed and all regressors can be unbounded. Analogous result which complements their result for the semi-parametric case is also given below.

Theorem 1 *When there is complementary data to estimate the distribution of X_2 given X_c , β_0 of the parametric binary choice model defined by equation (10) is identified if Assumptions A.1 and A.3–A.5 hold.*

Proof. Suppose equality (11) holds. 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

$$\begin{aligned} 0 &= \int_{\Omega_2(X_c)} [F(\alpha + X_c\beta + x'_2\gamma) - F(\alpha_0 + X_c\beta_0 + x'_2\gamma_0)] g(x_2|X_c) d\mu \\ &\quad + \int_{\Omega_2^c(X_c)} [F(\alpha + X_c\beta + x'_2\gamma) - F(\alpha_0 + X_c\beta_0 + x'_2\gamma_0)] g(x_2|X_c) d\mu \end{aligned}$$

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 a regressor 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 γ by moving the regressor under consideration but holding other variables in X_c fixed, because x_2 and γ 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. ■

Next we turn to identification of α_0 and γ_0 . We assume that β_0 is identified so henceforth denote $s = x_c\beta_0$. The problem now is to show that for any α and γ

$$\int F(\alpha + s + x'_2\gamma) g(x_2|X_c) d\mu = \int F(\alpha_0 + s + x'_2\gamma_0) g(x_2|X_c) d\mu \quad (13)$$

a.s. in X_c if and only if $\alpha = \alpha_0$ and $\gamma = \gamma_0$.

When Z_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 α_1 and α_2 when almost surely in X_c

$$F(\alpha_1 + s) g(1|X_c) + F(\alpha_2 + s) g(2|X_c) = F(\alpha'_1 + s) g(1|X_c) + F(\alpha'_2 + s) g(2|X_c).$$

This implies that

$$\frac{g(2|X_c)}{g(1|X_c)} = \frac{F(\alpha_1 + s) - F(\alpha'_1 + s)}{F(\alpha'_2 + s) - F(\alpha_2 + s)}.$$

Therefore the two parameter combinations we need to consider are $\alpha_1 > \alpha'_1$ and $\alpha_2 < \alpha'_2$ and $\alpha_1 < \alpha'_1$ and $\alpha_2 > \alpha'_2$. Without any loss in generality, consider the first case. In this case, we have

$$g(1|X_c) = \frac{F(\alpha'_2 + s) - F(\alpha_2 + s)}{F(\alpha_1 + s) - F(\alpha'_1 + s) + F(\alpha'_2 + s) - F(\alpha_2 + s)}.$$

Note that the right-hand side defines a parametric model of the conditional probability $g(1|X_c)$ as a function of s using parameters α_1 , α'_1 , α_2 and α'_2 with parameter restrictions $\alpha_1 > \alpha'_1$ and $\alpha_2 < \alpha'_2$. Thus the identification condition is that $g(1|X_c)$ is not within this parametric model. This is certainly the case if there is a variable among X_c that does not appear in s .

More generally the following identification result holds:

Theorem 2 *If $g(1|X_c)$ is not an element of the parametric model expressed by*

$$g(1|X_c) = \frac{F(\alpha'_2 + s) - F(\alpha_2 + s)}{F(\alpha_1 + s) - F(\alpha'_1 + s) + F(\alpha'_2 + s) - F(\alpha_2 + s)}.$$

where $\alpha_1 > \alpha'_1$ and $\alpha_2 < \alpha'_2$ for the case of binary variable X_2 and by

$$= \frac{g(1|X_c)}{\frac{\begin{bmatrix} -\Delta F(\alpha_k + s) - [\Delta F(\alpha_2 + s) - \Delta F(\alpha_k + s)]g(2|X_c) - \dots \\ -[\Delta F(\alpha_{k-1} + s) - \Delta F(\alpha_k + s)]g(k-1|X_c) \end{bmatrix}}{[\Delta F(\alpha_1 + s) - \Delta F(\alpha_k + s)]}}.$$

where numbering of the regressors follow the order of $\alpha_j - \alpha'_j$ and that $\alpha_1 > \alpha'_1$ and $\alpha_k < \alpha'_k$ for the case of general discrete vector X_2 , then α_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

$$F(\alpha_1 + s)g(1|X_c) + \dots + F(\alpha_k + s)g(k|X_c) = F(\alpha'_1 + s)g(1|X_c) + \dots + F(\alpha'_k + s)g(k|X_c).$$

Without any loss in generality, assume that $\alpha_1 > \alpha'_1$ and $\alpha_k < \alpha'_k$ and that the index is ordered in decreasing order of $\alpha_j - \alpha'_j$. If this is not the case the equality will not hold almost surely in X_c . Let $\Delta F(\alpha_j + s)$ be $F(\alpha_j + s) - F(\alpha'_j + s)$. Since $g(j|X_c)$ over j sum to 1, we have

$$[\Delta F(\alpha_1 + s) - \Delta F(\alpha_k + s)]g(1|X_c) + \dots + [\Delta F(\alpha_{k-1} + s) - \Delta F(\alpha_k + s)]g(k-1|X_c) = -\Delta F(\alpha_k + s)$$

Note that $\Delta F(\alpha_1 + s) - \Delta F(\alpha_k + s) > 0$ so that

$$= \frac{g(1|X_c)}{\frac{\begin{bmatrix} -\Delta F(\alpha_k + s) - [\Delta F(\alpha_2 + s) - \Delta F(\alpha_k + s)]g(2|X_c) - \dots \\ -[\Delta F(\alpha_{k-1} + s) - \Delta F(\alpha_k + s)]g(k-1|X_c) \end{bmatrix}}{[\Delta F(\alpha_1 + s) - \Delta F(\alpha_k + s)]}}.$$

■

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.

For a more general regressor case, we have the following result:

Theorem 3 *Suppose for any δ not equal to 0 there is a subset in the support of X_c with positive probability such that holding s constant, the distribution of $X_2'\delta$ given $X_c = x_c$ stochastically dominates or dominated by that of $X_2'\delta$ given $X_c = x_c$. Then α_0 and γ_0 are identified.*

The condition does not hold if X_2 is collinear. Clearly the condition is more easily to be satisfied if there are excluded variables that are closely tied to X_2 .

Proof. For any γ and γ_0 take such a subset specified by the assumption for $\delta = \gamma - \gamma_0$. Without any loss of generality suppose there is a positive probability that $X_2'\delta$ given $X_c = x_c$ stochastically dominates that of $X_2'\delta$ given $X_c = x_c$ holding s . Then over this set, equation (13) cannot hold almost surely in X_c , thus leading to a contradiction. Thus $\gamma = \gamma_0$, which in turn implies that $\alpha = \alpha_0$. ■

If one is willing to make exclusion restrictions assumptions, then the identification does not need to rely on assumptions A.4 and A.5 on the support of the regressors. Thus, let consider the vector of exogenous common regressors as divided in two subvectors: $X_c = \{\tilde{X}_c, Z_c^e\}$ where \tilde{X}_c is the vector included in the conditional parametric model and Z_c^e is the exclusion restriction. The following theorem allows for the possibility of the common regressors to be discrete and the missing regressors to be continuous.

Theorem 4 *If there is a subset in the support of X_c denoted by Ω_{Z_c} with positive probability such that, holding \tilde{x}_c constant, the distribution of X_2 given $X_c = x_c$ stochastically dominates or dominated by that of X_2 given $X_c = \bar{x}_c$ for $\{x_c, \bar{x}_c\} \in \Omega_{Z_c}$. Then θ_0 is identified.*

Proof.

$$\int [F(\alpha + \tilde{x}'_c\beta + x'_2\gamma) - F(\alpha_0 + \tilde{x}'_c\beta_0 + x'_2\gamma_0)] g(x_2|x_c) d\mu = 0$$

and then consider another observation of the common regressors where only the value of the excluded variable changes (i.e. $\bar{x}_c = \{\tilde{x}_c, x_c^e\}$) so that

$$\int [F(\alpha + \tilde{x}'_c\beta + x'_2\gamma) - F(\alpha_0 + \tilde{x}'_c\beta_0 + x'_2\gamma_0)] g(x_2|\bar{x}_c) d\mu = 0$$

If there exist observations x_c and \bar{x}_c such that $g(z_2|x_c)$ dominates or is dominated by $g(z_2|\bar{x}_c)$, then the above equality cannot hold for both observations x_c and \bar{x}_c by the definition of first-order stochastic dominance. Note that either $[F(\alpha + \tilde{x}'_c\beta + x'_2\gamma) - F(\alpha_0 + \tilde{x}'_c\beta_0 + x'_2\gamma_0)]$ or $[F(\alpha_0 + \tilde{x}'_c\beta_0 + x'_2\gamma_0) - F(\alpha + \tilde{x}'_c\beta + x'_2\gamma)]$ must be an increasing function of x_2 , so that in one of the cases the first-order stochastic definition applies and identification is achieved. ■

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

$$\int F(\alpha + X_c\beta + x'_2\gamma|X_c, x_2)g(x_2|X_c)d\mu = \int F_0(\alpha_0 + X_c\beta_0 + x'_2\gamma_0|X_c, x_2)g(x_2|X_c)d\mu \quad (14)$$

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 for β_0 holds once one of the coefficients of X_c is normalized to 1.¹⁴

Theorem 5 *When there is complementary data to estimate the distribution of X_2 given X_c , after normalizing one of the coefficients to 1, β_0 of the semiparametric binary choice model defined by equation (10) is identified if Assumptions A.2 and A.3–A.5 hold.*

Proof. Proceed identically to the point at which we obtain the inequality between $\alpha + x_c\beta + x'_2\gamma$ and $\alpha_0 + x_c\beta_0 + x'_2\gamma_0$ uniformly over x_2 and γ . By adjusting the regressor whose coefficient is normalized to 1, we can make 0 lie between the two. This leads to a contradiction as $\varepsilon > 0$ can be chosen to be arbitrarily small. ■

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 $\Omega_{Z_1} \in R^{m_1}, \Omega_{Z_c} \in R^{m_c}, \Omega_{Z_2} \in R^{m_2}, \Theta \in R^p$. 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 $\Gamma_q \times \Gamma_q \times \Theta$. Formally, functions $q_a(Z_1, Z_c, \theta)$ and $q_b(Z_1, Z_c, \theta)$ are functions from $\Omega_{Z_1} \times \Omega_{Z_c} \times \Theta$ into R , and $\psi(q_a(Z_1, Z_c, \theta), q_b(Z_1, Z_c, \theta); \theta)$ is a mapping from $q_a(\cdot) \times q_b(\cdot) \times \Theta$ into R^S with $S \geq p$, where $\{q_a, q_b\} \in \Gamma_q$. S denotes the number of moment conditions. We consider the sup-norm for the space of functions Γ denoted by $\|\cdot\|_\Gamma$.

Define the sample analogue of the moment condition in (2) as

$$\hat{H}(\hat{q}_{N_2,a}(\cdot, \theta), \hat{q}_{N_2,b}(\cdot, \theta)) = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{I}_{N_1 i} \psi(\hat{q}_{N_2,a}(z_{1i}, z_{ci}, \theta), \hat{q}_{N_2,b}(z_{1i}, z_{ci}, \theta); \theta) = 0 \quad (15)$$

where

$$\hat{q}_{N_2,j}(z_{1i}, z_{ci}, \theta) = \int \rho_j(z_{1i}, z_{ci}, z_2; \theta) \hat{g}_{N_2}(z_2|z_{ci}) dz_2 \text{ for } j = \{a, b\} \quad (16)$$

and the trimming indicator¹⁵

¹⁴See Manski (1985) for the identification of this model in the complete data framework.

¹⁵We consider a fixed trimming term which does not change with the sample size, unlike in Robinson (1988).

$$\hat{I}_{N_1 i} = 1 \left\{ \hat{f}_{N_1}(z_{ci}) > b \right\} \quad (17)$$

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

Our estimator solves the following problem

$$\hat{\theta} = \inf_{\theta \in \Theta} \hat{H}(\theta, \hat{q}_a(\cdot, \theta), \hat{q}_b(\cdot, \theta))' \times \hat{W} \times \hat{H}(\theta, \hat{q}_a(\cdot, \theta), \hat{q}_b(\cdot, \theta)) \quad (18)$$

where \hat{W} is a $S \times S$ matrix that converges in probability to a positive definite matrix W .¹⁶

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 estimators of the q functions evaluated at the i – th observation

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

If, among other assumptions¹⁷, the s – th derivatives of ψ 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 $0 \leq j \leq s-1$ 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

$$\hat{q}_j(z_{1i}, z_{ci}, \theta) = \left(N h_{N_2}^{M_1} \right)^{-1} \sum_{r=1}^{N_2} \frac{\rho_j(z_{1i}, z_{ci}, z_{2r}; \theta) K_1 \left(\frac{z_{cr} - z_{ci}}{h_{N_2}} \right)}{\left(N h_{N_2}^{M_1} \right)^{-1} \sum_{r=1}^{N_2} K_1 \left(\frac{z_{cr} - z_{ci}}{h_{N_2}} \right)} + O(h_{N_2}^s) \text{ for } j = \{a, b\} \quad (19)$$

The estimator we propose here for the moment condition is a weighted average of the function ρ_j 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.

¹⁶Note that the estimator $\hat{\theta}$ and \hat{H} are function of both sample sizes n and N . The estimates of \hat{q} and \hat{g} are obtained from the data set 2, so that they are a function of N only. We ignore the different subindices for simplicity in the notation.

¹⁷We provide detailed conditions in Section (5).

If the distribution of Z_2 given Z_c is discrete where Z_2 takes R possible different values $\{v_1, \dots, v_R\}$, the sample analogue of the estimators for q_j , $j = \{a, b\}$ are then

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

where

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

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 (2.2.1) 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 model¹⁸ 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 regression¹⁹

$$y_{1i} = x_{ci}' \theta_1 + \hat{E}_{N_2}(X_2 | X_c = x_{ci})' \theta_2 + v_i \text{ for } i = 1, \dots, N_1 \quad (20)$$

$$v_i = u_i + (x_{2i} - E(X_2 | X_c = x_{ci}))' \theta_2 + \left(E(X_2 | X_c = x_{ci}) - \hat{E}_{N_2}(X_2 | X_c = x_{ci}) \right)' \theta_2 \quad (21)$$

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, $i = \{1, \dots, N_1\}$. In order for the OLS estimates of θ^0 from (20) 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

$$\begin{aligned} \hat{E}_{N_1}(Y_1 | X_c = x_{ci}) &= x_{ci}' \theta_1 + \hat{E}_{N_2}(X_2 | X_c = x_{ci})' \theta_2 + v_i \text{ for } i = 1, \dots, N_1 \quad (22) \\ v_i &= (E(X_2 | X_c = x_{ci}) - \hat{E}_{N_2}(X_2 | X_c = x_{ci}))' \theta_2 - (E(Y_1 | X_c = x_{ci}) - \hat{E}_{N_1}(Y_1 | X_c = x_{ci})) \end{aligned}$$

¹⁸

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

¹⁹The sub-indices in the expectations denote the sample size of the dataset in which each conditional mean is computed.

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 parameters from regression (20) 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

$$y_{1i} = \tilde{x}'_{ci}\theta_1 + \hat{E}_{N_2}^*(X_2|X_c = x_{ci})'\theta_2 + v_i \text{ for } i = 1, \dots, N_1 \quad (23)$$

$$v_i = u_i + \theta_2(x_{2i} - E^*(X_2|X_c = x_{ci})) + \theta_2\left(E^*(X_2|X_c = x_{ci}) - \hat{E}_{N_2}^*(X_2|X_c = x_{ci})\right) \quad (24)$$

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 (24) 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 $\hat{E}_{N_2}^*(X_2|X_c = x_c)$.²⁰ 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

²⁰Since

$$\begin{aligned} E(\varepsilon'X_c) &= E([X_2 - E^*(X_2|X_c)]'X_c) = \\ &= E([X_2 - X_cE(X_c'X_c)^{-1}E(X_c'X_2)]'X_c) = 0 \end{aligned}$$

and

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

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 $Var(X_2 - \hat{E}_{N_2}(X_2|X_c = x_c)|X_c = x_c)$. 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 (23) 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)

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \left(\frac{1}{N_1} (Y_1 - \tilde{X}'_{cN_1} \theta_1)' X_{cN_1} - \frac{1}{N_2} (X'_{2N_2} \theta_2)' X_{cN_2} \right)' \times \Omega_{N_1 N_2}^{-1} \times \quad (25)$$

$$\times \left(\frac{1}{N_1} (Y_1 - \tilde{X}'_{cN_1} \theta_1)' X_{cN_1} - \frac{1}{N_2} (X'_{2N_2} \theta_2)' X_{cN_2} \right) \quad (26)$$

where $\Omega_{N_1 N_2}$ is a matrix which converges to a non-singular positive definite matrix Ω .²¹ 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 $E(X'_2 X_c)$ 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 $E(X'_2 X_c)$, the estimator we have defined in (15) 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 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.

²¹This weighing 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 .

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 (6). First, consider the following moment conditions implied by (7): $E [Z'_c (\rho_a(Z_1, Z_c; \theta) - \rho_b(Z_c, Z_2; \theta))] = 0$. 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

$$\inf_{\theta \in \Theta} \hat{H}(\theta)' \hat{W} \hat{H}(\theta) \quad (27)$$

$$\text{with } \hat{H}(\theta) = \frac{1}{N_1} \sum_{i=1}^{N_1} z'_{ci} \rho_a(z_{1i}, z_{ci}, \theta) - \frac{1}{N_2} \sum_{r=1}^{N_2} z'_{cr} \rho_b(z_{cr}, z_{2r}, \theta) \quad (28)$$

The alternative estimator we propose is derived from expression (15). Thus, using the law of iterated expectations, we provide an alternative method of computing the sample analogues of moments associated with ρ_b which uses also the information on Z_c in data set 1. In other words, it constructs a sample analogue of the conditional expectation $E_{Z_c} (E (Z'_c \rho_b(Z_c, Z_2, \theta) | Z_c))$ 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

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

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

$$\hat{H}(\theta) \quad (29)$$

$$= \frac{1}{N_1} \sum_{i=1}^{N_1} z'_{ci} \rho_a(z_{1i}, z_{ci}, \theta) - \frac{1}{N_1} \frac{1}{N_2} \sum_{i=1}^{N_1} \sum_{r=1}^{N_2} \frac{\frac{1}{h_{N_2}^{m_c}} z'_{ci} \rho_b(z_{ci}, z_{2r}, \theta) K \left(\frac{z_{cr} - z_{ci}}{h_{N_2}} \right)}{\hat{f}(z_{ci})} \quad (30)$$

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

Alternatively, using unconditional moment condition, one can propose estimators of θ_0 by using the FOC of the sample analogue of the objective function $E \left((\rho_a(Z_1, Z_c; \theta) - E(\rho_b(Z_c, Z_2; \theta) | Z_c))^2 \right)$ that the true value of the parameter uniquely minimises, where $E(\rho_b(Z_c, Z_2; \theta) | Z_c)$ 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 $E(\rho_a(Z_1, Z_c) - \rho_b(Z_c, Z_2) | Z_c) = 0$, 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 priori 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

$$\frac{1}{N_1} \sum_{i=1}^{N_1} (Nh_{N_2}^{m_c})^{-1} \sum_{r=1}^{N_2} \frac{\left[\nabla_{\theta} f(y_{1i}|x_{ci}, x_{2r}, \hat{\theta}) / \int f(y_{1i}|x_{ci}, x_2; \theta) \hat{g}(x_2|x_{ci}) dx_2 \right] K_1 \left(\frac{x_{c_r} - x_c}{h_{N_2}} \right)}{(Nh_{N_2}^{m_c})^{-1} \sum_{r=1}^{N_2} K_1 \left(\frac{x_{1r} - x_c}{h_{N_2}} \right)} + O(h_{N_2}^s) = 0 \quad (31)$$

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

$$\frac{1}{nN} \sum_{i=1}^{N_1} \sum_{r=1}^{N_2} \frac{1}{h_{N_2}^{m_c}} \frac{\nabla_{\theta} f(y_{1i}|x_{1i}, x_{2r}, \hat{\theta}) K_1 \left(\frac{x_{1r} - x_c}{h_{N_2}} \right)}{(Nh_{N_2}^{m_c})^{-1} \sum_{s=1}^{N_2} f(y_{1i}|x_{1i}, x_{2s}, \hat{\theta}) K_1 \left(\frac{x_{1s} - x_c}{h_{N_2}} \right)} + O(h_{N_2}^2) = 0$$

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 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.²²

To motivate the asymptotic results, consider a Taylor's series expansion for $\hat{\theta}$ around θ_0 from

²²The 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.

the FOC of the objective function in (18)

$$\begin{aligned} \sqrt{N_1 + N_2} (\hat{\theta} - \theta_0) &= - \left[\nabla'_{\theta} \hat{H}(\hat{\theta}, \hat{q}_a(\cdot, \hat{\theta}), \hat{q}_b(\cdot, \hat{\theta})) \times \hat{W} \times \nabla_{\theta} \hat{H}(\bar{\theta}, \hat{q}_a(\cdot, \bar{\theta}), \hat{q}_b(\cdot, \bar{\theta})) \right]^{-1} \\ &\times \left[\nabla'_{\theta} \hat{H}(\hat{\theta}, \hat{q}_a(\cdot, \hat{\theta}), \hat{q}_b(\cdot, \hat{\theta})) \times \hat{W} \times \sqrt{N_1 + N_2} \hat{H}(\theta_0, \hat{q}_a(\cdot, \theta_0), \hat{q}_b(\cdot, \theta_0)) \right] \end{aligned} \quad (32)$$

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

In what follows we denote by z_c and \tilde{z}_c to the realised 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 $i = 1, \dots, N_1$ and $\{z_{cr}, \tilde{z}_{2r}\}$ for $r = 1, \dots, N_2$. 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 $\{z_{cr}, \tilde{z}_{2r}\}_{r=1}^{N_2}$ are independent and identically distributed. Additionally both samples are independent*

Assumption B. 2 *The identification condition is satisfied*

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

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

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

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

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 $j = 1, \dots, s - 1$; $\int u^s K(u) du < \infty$; $\int |K(u)| du < \infty$; $|u| |K(u)| \rightarrow 0$ as $|u| \rightarrow \infty$; $\sup |K(u)| < \infty$; $\int K^2(u) du < \infty$*

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

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

$$\frac{N_1 h_{N_2}^{2 + \frac{2}{l-2}}}{(-\log h_{N_2})} \rightarrow \infty$$

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

Assumption B. 10 $\psi(q_a, q_b; \theta)$ is Frechet differentiable with respect to θ , $q_a(\cdot)$ and $q_b(\cdot)$ and the Frechet derivatives are Lipschitz continuous. with $C_j(z_1, z_c) > 0$, $E\{C_j(z_1, z_c)\} < \infty$ for $j = 1, \dots, 9$

$$\begin{aligned} \left| \frac{\partial \psi(q_a, q_b; \theta)}{\partial \theta} - \frac{\partial \psi(q'_a, q'_b; \theta')}{\partial \theta} \right| &\leq C_1(z_1, z_c) |\theta - \theta'| + C_2(z_1, z_c) \|q_a - q'_a\|_{\Gamma_q} + C_3(z_1, z_c) \|q_b - q'_b\|_{\Gamma_q} \\ \left| \frac{\partial \psi(q_a, q_b; \theta)}{\partial q_a} - \frac{\partial \psi(q'_a, q'_b; \theta')}{\partial q_a} \right| &\leq C_4(z_1, z_c) |\theta - \theta'| + C_5(z_1, z_c) \|q_a - q'_a\|_{\Gamma_q} + C_6(z_1, z_c) \|q_b - q'_b\|_{\Gamma_q} \\ \left| \frac{\partial \psi(q_a, q_b; \theta)}{\partial q_b} - \frac{\partial \psi(q'_a, q'_b; \theta')}{\partial q_b} \right| &\leq C_7(z_1, z_c) |\theta - \theta'| + C_8(z_1, z_c) \|q_a - q'_a\|_{\Gamma_q} + C_9(z_1, z_c) \|q_b - q'_b\|_{\Gamma_q} \end{aligned}$$

Assumption B. 11 $\rho_j(z_1, z_c, z_2, \theta)$ for $j = \{a, b\}$ are 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 $\text{plim}_{N_1, N_2 \rightarrow \infty} \hat{W} = W$ where W is symmetric and positive definite

Henceforth we use the following shorthand notation. Let $q_{a, i\theta_0} = q_a(z_{1i}, z_{ci}, \theta_0)$ and $q_{b, i\theta_0} = q_b(z_{1i}, z_{ci}, \theta_0)$ where sub-index i denotes that z_1 and z_c are conditioned on the i th observation. Denote $\psi_{i\theta_0}(q_a, q_b) = \psi(q_a(z_{1i}, z_{ci}, \theta_0), q_b(z_{1i}, z_{ci}, \theta_0); \theta_0)$; $\rho_{a, i\theta_0}(z_2) = \rho_a(z_{1i}, z_{ci}, z_2; \theta_0)$ and $\rho_{b, i\theta_0}(z_2) = \rho_b(z_{1i}, z_{ci}, z_2; \theta_0)$ to indicate that the rest of the variables are all conditioned on the i th observation. Where necessary, we make explicit the argument of functions ψ , ρ_j and q_j for $j = \{a, b\}$.

Theorem 6 Suppose that $\hat{\theta}$ is consistent to θ_0 . Under Assumptions B. (1)-Assumptions B.14, if $V'WV$ is nonsingular with

$$V = \int \nabla_{\theta} \psi(q_a(z_1, z_c, \theta_0), q_b(z_1, z_c, \theta_0); \theta_0) f(z_1, z_c) dz_1 dz_c \quad (34)$$

, then

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

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

$$\Sigma_1 = \text{Var}(\psi(q_a(Z_1, Z_c; \theta_0), q_b(Z_1, Z_c; \theta_0); \theta_0)) \quad (35)$$

$$\Sigma_2 = \text{Var}\left(\sum_{j \in \{a, b\}} \int \left\{ \left[\rho_j(z_1, z_{cr}, z_{2r}, \theta_0) - \int \rho_j(z_1, z_{cr}, z_2, \theta_0) g(z_2 | z_{cr}) dz_2 \right] \times \frac{\partial \psi(q_a(z_1, z_{cr}; \theta_0), q_b(z_1, z_{cr}; \theta_0); \theta_0)}{\partial q_j} \right\} f(z_1 | z_{cr}) dz_1 \right)$$

Proof of Theorem (6). Consider the Taylor's series expansion in (32). The asymptotic distribution of $\hat{\theta}$ is shown in two parts. The first part shows the asymptotic distribution of the score term $\sqrt{N_1 + N_2} \hat{H}(\theta_0, \hat{q}_a(\cdot, \theta_0), \hat{q}_b(\cdot, \theta_0))$ 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 $\sup_i |\hat{I}_{Ni} - I_i| \xrightarrow{p} 0$ as $N_1, N_2 \rightarrow \infty$. The expression below makes clear the sources of inefficiency that arise when Z_2 is not jointly observed with Z_1 and Z_c .

$$\begin{aligned} \hat{H}(\theta_0, \hat{q}_{a, N_2}(\cdot, \theta_0), \hat{q}_{b, N_2}(\cdot, \theta_0)) &= \\ &= \frac{1}{N_1} \sum_{i=1}^{N_1} I_i \psi(\rho_{a, i\theta_0}(z_{2i}), \rho_{b, i\theta_0}(z_{2i}); \theta_0) \end{aligned} \quad (36)$$

$$+ \frac{1}{N_1} \sum_{i=1}^{N_1} I_i [\psi(q_{a, i\theta_0}, q_{b, i\theta_0}; \theta_0) - \psi(\rho_{a, i\theta_0}(z_{2i}), \rho_{b, i\theta_0}(z_{2i}); \theta_0)] \quad (37)$$

$$+ \frac{1}{N_1} \sum_{i=1}^{N_1} I_i \frac{\partial \psi(q_{a, i\theta_0}, q_{b, i\theta_0}; \theta_0)}{\partial q_a} [\hat{q}_{a, i\theta_0} - q_{a, i\theta_0}] \quad (38)$$

$$+ \frac{1}{N_1} \sum_{i=1}^{N_1} I_i \frac{\partial \psi(q_{a, i\theta_0}, q_{b, i\theta_0}; \theta_0)}{\partial q_b} [\hat{q}_{b, i\theta_0} - q_{b, i\theta_0}] \quad (39)$$

$$+ R_{N_1 N_2} \quad (40)$$

Only term (36) would arise if Z_2 were jointly observed with Z_1 and Z_c . Term (37) reflects the efficiency loss due to not observing of Z_2 , since if Z_2 is observed there is no need to integrate out over the distribution of Z_2 given Z_c both functions ρ_a and ρ_b . The next terms (38) and (39) represent the efficiency loss due to the estimation of the conditional distribution function of Z_2 given Z_c , $g(Z_2 | Z_c)$ inside functions q_a and q_b .

Lemma A. 2 in the Appendix shows under which some of the assumptions above $\sqrt{N_1 + N_2} R_{N_1 N_2} = o_p(1)$. From expressions (38) to (39), we use the asymptotically linearity at rate $N_2^{-1/2}$ for kernel estimators of conditional expectations. Define $m\rho_{j, i\theta_0}(z_c) = \int \rho_{j, i\theta_0}(z_2) g(z_2 | z_c) dz_2$ and its

estimated counterparts by $\widehat{m}_{\rho_{j,i\theta_0}}(Z_c)$ for $j = \{a, b\}$. Thus, expressions (38) and (39) can be written as

$$\begin{aligned}
U_{N_1 N_2}^j &= \\
&\frac{1}{N_1 N_2 h^{m_c}} \sum_{i=1}^{N_1} \sum_{r=1}^{N_2} \left\{ \left[I_i \frac{\{\rho_{j,i\theta_0}(z_{2r}) - m_{\rho_{j,i\theta_0}}(z_{cr})\} K\left(\frac{z_{cr} - z_{ci}}{h_{N_2}}\right)}{f(z_{ci})} \right] \times \left[\frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0})}{\partial q_j} \right] \right\} \\
&+ \frac{1}{N_1} \sum_{i=1}^{N_1} b_{qi} + o_p\left(\frac{1}{\sqrt{N_2}}\right) + O(h_{N_2}^s)
\end{aligned} \tag{41}$$

where

$$b_{j,qi} = \frac{1}{h_{N_2}^{m_c}} \frac{I_i}{f(z_{ci})} E_{Z_c} \left([m_{\rho_{j,i\theta_0}}(Z_c) - m_{\rho_{j,i\theta_0}}(z_{ci})] K\left(\frac{Z_c - z_{ci}}{h_{N_2}}\right) \right) \times \left[\frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0})}{\partial q_j} \right]$$

for $j = \{a, b\}$

Since $m_{\rho_{j,i\theta_0}}(Z_c)$ 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

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

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

By Assumption B. 4, the last reminder terms of $U_{N_1 N_2}^1$ and $U_{N_1 N_2}^2$ 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 projections $\widehat{V}_{N_1 N_2}^j$, $j = \{a, b\}$ of terms (41) denoted henceforth as $V_{N_1 N_2}^j$. These are two-sample U-statistics of order 1, since there is only one observation from each sample in each kernel²³. Let define the kernels a in each of the U-statistic as

$$V_{N_1 N_2}^j = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{r=1}^{N_2} a_{jN_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r}) \text{ for } j = \{a, b\}$$

where

$$a_{jN_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r}) = \frac{I_i}{h_{N_2}^{m_c}} \left[\frac{\{\rho_{j,i\theta_0}(z_{2r}) - m_{\rho_{j,i\theta_0}}(z_{cr})\} K\left(\frac{z_{cr} - z_{ci}}{h_{N_2}}\right)}{f(z_{ci})} \right] \times \left[\frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0})}{\partial q_j} \right]$$

²³For Central Limit Theorems for U-statistics, see Serfling (1980) and van der Vaart (1998)

Denote by $\Upsilon_j = E(a_{jN}(z_{1i}, z_{ci}, z_{cr}, z_{2r}))$, $j = \{a, b\}$. The projections of both statistics ($V_{N_1 N_2}^j - \Upsilon_j$) for $j = \{a, b\}$ are defined as

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

It can be shown that the projections over both samples of the kernels for $j = \{a, b\}$ are

$$\begin{aligned} E(a_{jN_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r}) | z_{1i}, z_{ci}) &= 0 \\ E(a_{jN_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r}) | z_{cr}, z_{2r}) &= \\ I_r E_{Z_1 | Z_c} \left[\{\rho_{j\theta_0}(z_{2r}) - m\rho_{j\theta_0}(z_{cr})\} \times \left[\frac{\partial \psi(q_{a,\theta_0}, q_{b,\theta_0})}{\partial q_j} \right] \Big|_{Z_c = z_{cr}} \right] + O(h_{N_2}^2) \end{aligned}$$

²⁴The above conditions ensure that the later Taylor's series expansion can be done ²⁵. The Lemma A. 6 in the Appendix gives sufficient conditions for

$$\sqrt{N_1 + N_2} \left[(V_{N_1 N_2}^a - \Upsilon_a) + (V_{N_1 N_2}^b - \Upsilon_b) - (\hat{V}_{N_1 N_2}^a + \hat{V}_{N_1 N_2}^b) \right] \xrightarrow{p} 0$$

as $N_1 \rightarrow \infty$, $N_2 \rightarrow \infty$ and known as the U-statistics projection result. There we use the sufficient condition of $E(|a_{jN_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r})|^2) = o(N_2)$, $j = \{a, b\}$ in Powell, Stock and Stoker (1989), which is satisfied as long as $N_2 h_{N_2}^{m_c} \rightarrow \infty$ as $N_2 \rightarrow \infty$.

The sufficient conditions in Assumption B. (8) guarantees that $N_2 h_{N_2}^{m_c} \rightarrow \infty$ 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, both $\Upsilon_a = 0$ and $\Upsilon_b = 0$. Having used then the projection device to find the distribution of (38) and (39) we can conclude that the asymptotic distribution of $\hat{H}(\theta_0, \hat{q}_{N_2}(\cdot, \theta_0), \hat{g}_{N_2})$ is normally distributed as

$$\begin{aligned} \sqrt{N_1 + N_2} \hat{H}(\theta_0, \hat{q}_{a, N_2}(\cdot, \theta_0), \hat{q}_{b, N_2}(\cdot, \theta_0)) &= \\ = \sqrt{N_1 + N_2} \left(\frac{1}{N_1} \sum_{i=1}^{N_1} I_i \psi(q_{a, i\theta_0}, q_{b, i\theta_0}; \theta_0) + \hat{V}_{N_1 N_2}^a + \hat{V}_{N_1 N_2}^b \right) + o_p(1) &\rightarrow N(0, \Sigma) \end{aligned}$$

with the expression in Σ as in expression (35).

²⁴The 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.

²⁵Note that

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

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

Part 2

Under the differentiability conditions of function ψ , q_a and q_b with respect to θ in Assumptions B. 10 and B. 11 uniformly in a neighborhood of θ_0 , the Taylor's series expansion in (32) is correctly done. With respect to the Jacobian term in (32), the uniform convergence arguments together with the consistency of $\hat{\theta}$, \hat{q}_a and \hat{q}_b suggests that

$$\left| \nabla_{\theta} \hat{H}(\hat{\theta}, \hat{q}_{a,N_2}(\cdot, \hat{\theta}), \hat{q}_{b,N_2}(\cdot, \hat{\theta})) - V \right| = o_p(1) \quad (42)$$

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

$$V = \int \nabla_{\theta} \psi(q_a(z_1, z_c, \theta_0), q_b(z_1, z_c, \theta_0); \theta_0) f(z_1, z_c) dz_1 dz_c \quad (43)$$

The convergence in probability in (42) is shown in two steps. Lemma A.7 in the Appendix shows that

$$|\nabla_{\theta} H(\theta, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) - \nabla_{\theta} H(\theta_0, q_a(\cdot, \theta_0), q_b(\cdot, \theta_0))| = o_p(1) \quad (44)$$

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

$$|\nabla_{\theta} H(\theta_0, q_a(\cdot, \theta_0), q_b(\cdot, \theta_0)) - V| = o_p(1) \quad (45)$$

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.

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 (20) 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 $n=1000$ and $N=5000$, 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 1-2 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 replication²⁶. A third order 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 (21). In each row of the last panel of Tables 1-2, the mean over replications of the components of v in (21) 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 in Table 1 and 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 error variance 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.²⁷

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

²⁶The CV 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.

²⁷The simulations performed for the alternative linear model in (22) yield very similar results to the ones reported in Tables 1-2. For brevity, we omit these results here.

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 (20) suggests. We report these results in the upper panel of Table 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 (23) 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 3 and uses only data set 2 to compute those moments that include X_2 , as the estimator from moment condition (27) and (25) suggests. The second IV version is the estimator that solves (29) 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 3 use only data set 1 to compute the weighting matrix of the IV estimator. Similar results were 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 $X_2 = 1 + 2W_1 + X_c W_1 + W_1 W_2 + 2W_2 + \varepsilon$ 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 4.

The estimator we propose (i.e. those estimates in the first and fourth panel of Tables 3 and 4) turns out to be more efficient than the two estimators we compare it with. This is true for the estimators defined from the sample analogues of the two sets of moment conditions. 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 the true conditional mean and its conditional linear fitting (i.e. third additive term in (20)). Table 5 compares the variance decomposition of v in (21) in each of its terms for both the estimator where the nonparametric conditional mean of X_2 given X_c and the estimator that uses the best predictor of X_2 given X_c .²⁸ 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 $(E(X_2 | X_c) - \hat{E}_{N_2}^*(X_2 | X_c))$ with

²⁸We 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.

respect to the dispersion of its nonparametric counterpart ($E(X_2|X_c) - \hat{E}_{N_2}(X_2|X_c)$). 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 $X_c \sim N(0, 1)$ and $X_2 = 1\{1+X_c+\varepsilon\}$ where $\varepsilon \sim N(0, 1)$ and $y = 1\{\theta_0^0 + \theta_1^0 X_c + \theta_2^0 X_2 + U\}$ with $[\theta_0^0, \theta_1^0, \theta_2^0] = [1, 3, -3]$ and $U \sim N(0, 1)$. The results of two estimators of this model are reported in Table 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 6. The bottom panel reports the results of the ML estimator that combines two different data sets defined in (31). 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.²⁹

For scalar X_2 , we also provide identification results for a more general scalar and continuous X_2 . Table 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.

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

²⁹The 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 bandwidth when the objective function is a likelihood function involving some nonparametric estimation is an open question. We use a value of the bandwidth for the sample sizes reported for the probit results of 0.75. Various sensitivity analysis exercises were carried out (having some constraints given the support of Z_c) and the results did not change substantially.

of models based on the nonparametric estimation of the conditional distribution function of the regressors which can be obtained from the auxiliary 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 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 for the parametric models. 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 be checked to be satisfied for each particular case.

8 Appendix

8.1 Lemmas used in Part 1 of the Proof of Theorem (6)

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

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

Proof. (see Ichimura (2003)) ■

Lemma A. 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*

$$(N_1 + N_2)h_{N_2}^{4s} \rightarrow 0 \quad (46)$$

$$\frac{N_2 h_{N_2}^{\frac{1}{l-2}}}{(-\log h_{N_2})} \rightarrow \infty \quad (47)$$

$$N_2 h_{N_2}^2 \rightarrow \infty \quad (48)$$

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

Proof. The reminder term in (40) is expressed as

$$\begin{aligned} R_{N_1 N_2} &= \frac{1}{N_1} \sum_{i=1}^{N_1} I_i \left[\frac{\partial \psi(\bar{q}_{a,i\theta_0}, \bar{q}_{b,i\theta_0}; \theta_0)}{\partial q_a} - \frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0}; \theta_0)}{\partial q_a} \right] [\hat{q}_{a,i\theta_0} - q_{a,i\theta_0}] + \\ &\quad + \frac{1}{N_1} \sum_{i=1}^{N_1} I_i \left[\frac{\partial \psi(\bar{q}_{a,i\theta_0}, \bar{q}_{b,i\theta_0}; \theta_0)}{\partial q_b} - \frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0}; \theta_0)}{\partial q_b} \right] [\hat{q}_{b,i\theta_0} - q_{b,i\theta_0}] \end{aligned}$$

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

$$R_{N_1 N_2} \leq \frac{1}{N_1} \sum_{i=1}^{N_1} I_i \left\{ \begin{aligned} &[C_5(z_{1i}, z_{ci})] [\hat{q}_{a,i\theta_0} - q_{a,i\theta_0}]^2 + [C_9(z_{1i}, z_{ci})] [\hat{q}_{b,i\theta_0} - q_{b,i\theta_0}]^2 + \\ &+ [C_6(z_{1i}, z_{ci}) + C_8(z_{1i}, z_{ci})] [\hat{q}_{a,i\theta_0} - q_{a,i\theta_0}] [\hat{q}_{b,i\theta_0} - q_{b,i\theta_0}] \end{aligned} \right.$$

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

estimators. Denote $\hat{q}_{j,i\theta_0} = \hat{r}q_{j,i\theta_0}/\hat{f}_i$ for $j = \{a, b\}$. Then,

$$\begin{aligned}
R_{N_1 N_2} &\leq \\
&\frac{1}{N_1} \sum_{i=1}^{N_1} \left[I_i \frac{1}{\hat{f}_i} \left[\hat{r}q_{a,i\theta_0} - r q_{a,i\theta_0} \right] - \frac{r q_{a,i\theta_0}}{\hat{f}_i^2} \left[\hat{f}_i - f_i \right] + \right. \\
&\quad \left. + o_p \left(\hat{r}q_{a,i\theta_0} - r q_{a,i\theta_0} \right) + o_p \left(\hat{f}_i - f_i \right) \right]^2 C_5(z_{1i}, z_{ci}) + \\
&+ \frac{1}{N_1} \sum_{i=1}^{N_1} \left[I_i \frac{1}{\hat{f}_i} \left[\hat{r}q_{b,i\theta_0} - r q_{b,i\theta_0} \right] - \frac{r q_{b,i\theta_0}}{\hat{f}_i^2} \left[\hat{f}_i - f_i \right] \right. \\
&\quad \left. + o_p \left(\hat{r}q_{b,i\theta_0} - r q_{b,i\theta_0} \right) + o_p \left(\hat{f}_i - f_i \right) \right]^2 C_9(z_{1i}, z_{ci}) + \\
&+ \frac{1}{N_1} \sum_{i=1}^{N_1} \left[I_i \frac{1}{\hat{f}_i} \left[\hat{r}q_{a,i\theta_0} - r q_{a,i\theta_0} \right] - \frac{r q_{a,i\theta_0}}{\hat{f}_i^2} \left[\hat{f}_i - f_i \right] \right. \\
&\quad \left. + o_p \left(\hat{r}q_{a,i\theta_0} - r q_{a,i\theta_0} \right) + o_p \left(\hat{f}_i - f_i \right) \right] \times \\
&\times \left[I_i \frac{1}{\hat{f}_i} \left[\hat{r}q_{b,i\theta_0} - r q_{b,i\theta_0} \right] - \frac{r q_{b,i\theta_0}}{\hat{f}_i^2} \left[\hat{f}_i - f_i \right] \right. \\
&\quad \left. + o_p \left(\hat{r}q_{b,i\theta_0} - r q_{b,i\theta_0} \right) + o_p \left(\hat{f}_i - f_i \right) \right] [C_6(z_{1i}, z_{ci}) + C_8(z_{1i}, z_{ci})]
\end{aligned} \tag{49}$$

To show that $\sqrt{N_1 + N_2} R_{N_1 N_2} = o_p(1)$ we follow the next steps. Lemma A.3 shows that the order of the bias of the nonparametric estimators is $h_{N_2}^s$ so that $E(\hat{r}q_{j,i\theta_0}) - r q_{j,i\theta_0} = O(h_{N_2}^s)$ for $j = \{a, b\}$; $E(\hat{f}_i) - f_i = O(h_{N_2}^s)$. The differentiability conditions of Lemma A.3 are stated in assumptions B.9, B.12 and B.13. From expression (49) and Lemmas A.3 - A.5 below, the reminder term converges in probability to zero if there exist positive sequences h_{N_2} , $\{\varepsilon_{aN_2}\}$, $\{\varepsilon_{bN_2}\}$, $\{\varepsilon_{1N_2}\}$, $\{M_{aN_2}\}$ and $\{M_{bN_2}\}$ such that

$$\begin{aligned}
\sqrt{N_1 + N_2} \varepsilon_{aN_2} \varepsilon_{1N_2} &\rightarrow 0; \sqrt{N_1 + N_2} \varepsilon_{aN_2} \varepsilon_{bN_2} \rightarrow 0; \sqrt{N_1 + N_2} \varepsilon_{1N_2} \varepsilon_{bN_2} \rightarrow 0 \\
\sqrt{N_1 + N_2} \varepsilon_{aN_2} h_{N_2}^s &\rightarrow 0; \sqrt{N_1 + N_2} \varepsilon_{bN_2} h_{N_2}^s \rightarrow 0; \sqrt{N_1 + N_2} \varepsilon_{1N_2} h_{N_2}^s \rightarrow 0 \\
\sqrt{N_1 + N_2} h_{N_2}^{2s} &\rightarrow 0
\end{aligned} \tag{50}$$

as $N_1 \rightarrow \infty$, $N_2 \rightarrow \infty$ and such that these sequences satisfy the conditions of Lemmas A.4 - A.5 below. To see that these sequences exist, take $\varepsilon_{jN_2} = (-\log h_{N_2}/N_2 h_{N_2})^{1/2} b_{jN_2}$ and $M_{jN_2} = (N_2 h_{N_2}/(-\log h_{N_2}))^{1/2} b_{jN_2}^u$ for $0 < u < 1$ and for positive sequences b_{jN_2} that diverge to infinity and for $j = \{a, b, 1\}$. Then, the sequences satisfy the conditions in lemmas A.4 - A.5 as long as condition (47) holds.

The conditions in (50) hold if the sequences b_{jN_2} diverge at a slower rate than $o((-\log h_{N_2})^{-1/2})$ and if $(N_1 + N_2)h_{N_2}^{4s} \rightarrow 0$ and $N_2 h_{N_2}^2 \rightarrow \infty$ ■

Lemma A.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 $h_{N_2} \rightarrow 0$ and $N_2 \rightarrow \infty$

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

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

Proof. Note that the expression for the estimator is

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

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

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

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

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

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

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

■

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

$$\Pr \left\{ \sup_{z_1, z_c, \theta \in \Omega_{Z_1} \times \Omega_{Z_c} \times \Theta} |\hat{r}q_{j,i\theta} - E(\hat{r}q_{j,i\theta})| > \varepsilon_{j,N_2} \right\} \rightarrow 0 \text{ as } N_2 \rightarrow \infty \text{ for } j = \{a, b\}$$

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

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

Lemma A. 5 Under assumption B.6, then

$$\Pr \left\{ \sup_{z_c \in \Omega_{z_c}} \left| \widehat{f}_i - E(\widehat{f}_i) \right| > \varepsilon_{1N_2} \right\} \rightarrow 0 \text{ as } N_2 \rightarrow \infty$$

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

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

Lemma A. 6 If $N_2 h_{N_2}^{m_c} \rightarrow \infty$ as $N_2 \rightarrow \infty$, then $E \left((a_{jN_2}(Z_1, Z_c, Z_c, Z_2))^2 \right) = o(N_2)$ for $j = \{a, b\}$

Proof.

$$a_{jN_2}(z_{1i}, z_{ci}, z_{cr}, z_{2r}) = \frac{I_i}{h_{N_2}^{m_c}} \left[\frac{\{\rho_{j,i\theta_0}(z_{2r}) - m\rho_{j,i\theta_0}(z_{cr})\} K\left(\frac{z_{cr}-z_{ci}}{h_{N_2}}\right)}{f(z_{ci})} \right] \times \left[\frac{\partial\psi(q_{a,i\theta_0}, q_{b,i\theta_0})}{\partial q_j} \right]$$

Denote the conditional expectations of $a_{jN_2}(Z_1, Z_c, Z_c, Z_2)^2$ on the realised values of Z_c in each data set as

$$v_j(z_{ci}, z_{cr}) = \int \int \left\{ \left[\frac{\rho_j(Z_1, z_{ci}, Z_2; \theta_0) - \int \rho_j(Z_1, z_{ci}, Z_2; \theta_0) g(Z_2|z_{cr}) dZ_2}{\int \rho_j(Z_1, z_{ci}, Z_2; \theta_0) g(Z_2|z_{cr}) dZ_2} \right]^2 \times \left[\frac{\partial\psi(q_a(Z_1, z_{ci}; \theta_0), q_b(Z_1, z_{ci}; \theta_0))}{\partial q_j} \right]^2 \right\} f(Z_1|z_{ci}) g(Z_2|z_{cr}) dZ_1 dZ_2$$

Then, for $j = \{a, b\}$

$$\begin{aligned} E \left((a_{jN_2}(Z_1, Z_c, Z_c, Z_2))^2 \right) &= \int \frac{1}{h_{N_2}^{2M_1}} \frac{v_j(z_{ci}, z_{cr}) K^2\left(\frac{z_{ci}-z_{cr}}{h_{N_2}}\right)}{f^2(z_{ci})} f(z_{ci}) f(z_{cr}) dz_{ci} dz_{cr} \\ &= \int \frac{1}{h_{N_2}^{m_c}} \frac{v_j(z_{ci}, z_{ci} + th_{N_2}) K^2(t)}{f(z_{ci})} f(z_{ci} + th_{N_2}) dz_{ci} dt \\ &= O \left(N_2 \left(N_2 h_{N_2}^{m_c} \right)^{-1} \right) \end{aligned}$$

Consequently, we have that $E \left((a_{jN_2}(Z_1, Z_c, Z_c, Z_2))^2 \right) = o(N_2)$ if and only if $N_2 h_{N_2}^{m_c} \rightarrow \infty$ as $N_2 \rightarrow \infty$. ■

8.2 Lemmas used in Part 2 of the Proof of Theorem (6)

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

$$\left| \nabla_{\theta} H(\hat{\theta}, \hat{q}_a(\cdot, \hat{\theta}), \hat{q}_b(\cdot, \hat{\theta})) - \nabla_{\theta} H(\theta_0, q_a(\cdot, \theta_0), q_b(\cdot, \theta_0)) \right| = o_p(1)$$

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

$$\begin{aligned} & \left| \nabla_{\theta} H(\theta, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) - \nabla_{\theta} H(\theta_0, q_a(\cdot, \theta_0), q_b(\cdot, \theta_0)) \right| \leq \\ & \left| \nabla_{\theta} H(\theta, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) - \nabla_{\theta} H(\theta_0, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) \right| + \end{aligned} \quad (51)$$

$$+ \left| \nabla_{\theta} H(\theta_0, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) - \nabla_{\theta} H(\theta_0, q_a(\cdot, \theta_0), q_b(\cdot, \theta_0)) \right| \quad (52)$$

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

$$\begin{aligned} & \left| \nabla_{\theta} H(\theta, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) - \nabla_{\theta} H(\theta_0, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) \right| \leq \\ & \leq \frac{1}{N_1} \sum_{i=1}^{N_1} \left[C_1(z_{1i}, z_{ci}) + C_4(z_{1i}, z_{ci}) \left\| \frac{\partial \tilde{q}_a(z_{1i}, z_{ci}, \theta)}{\partial \theta} \right\|_{\Gamma_q} + C_7(z_{1i}, z_{ci}) \left\| \frac{\partial \tilde{q}_b(z_{1i}, z_{ci}, \theta)}{\partial \theta} \right\|_{\Gamma_q} \right] \times \|\theta - \theta_0\| + \end{aligned}$$

For consistent estimators $\hat{\theta}$, \hat{q}_a and \hat{q}_b of θ_0 , q_a and q_b , respectively, we can consider $\theta = \hat{\theta}$, $\tilde{q}_a = \hat{q}_a$ and $\tilde{q}_b = \hat{q}_b$ 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 (51) converges to zero in probability

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

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

in (51)

$$\begin{aligned}
& |\nabla_{\theta} H(\theta_0, \tilde{q}_a(\cdot, \theta), \tilde{q}_b(\cdot, \theta)) - \nabla_{\theta} H(\theta_0, q_{0a}(\cdot, \theta_0), q_{0b}(\cdot, \theta_0))| \leq \\
& \leq \frac{1}{N_1} \sum_{i=1}^{N_1} C_2(z_{1i}, z_{ci}) \|\tilde{q}_{a,i\theta} - q_{a,i\theta_0}\|_{\Gamma_q} + C_3(z_{1i}, z_{ci}) \|\tilde{q}_{b,i\theta} - q_{b,i\theta_0}\|_{\Gamma_q} \\
& + \frac{1}{N_1} \sum_{i=1}^{N_1} \left[C_5(z_{1i}, z_{ci}) \|\tilde{q}_{a,i\theta} - q_{a,i\theta_0}\|_{\Gamma_q} + C_6(z_{1i}, z_{ci}) \|\tilde{q}_{b,i\theta} - q_{b,i\theta_0}\|_{\Gamma_q} \right] \frac{\partial \tilde{q}_{a,i\theta}}{\partial \theta} \\
& + \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0}; \theta_0)}{\partial q_a} \left\| \frac{\partial \tilde{q}_{a,i\theta}}{\partial \theta} - \frac{\partial q_{a,i\theta_0}}{\partial \theta} \right\|_{\Gamma_q} \tag{53}
\end{aligned}$$

$$+ \frac{1}{N_1} \sum_{i=1}^{N_1} \left[C_8(z_{1i}, z_{ci}) \|\tilde{q}_{a,i\theta} - q_{a,i\theta_0}\|_{\Gamma_q} + C_9(z_{1i}, z_{ci}) \|\tilde{q}_{b,i\theta} - q_{b,i\theta_0}\|_{\Gamma_q} \right] \frac{\partial \tilde{q}_{b,i\theta}}{\partial \theta} \tag{54}$$

$$+ \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{\partial \psi(q_{a,i\theta_0}, q_{b,i\theta_0}; \theta_0)}{\partial q_b} \left\| \frac{\partial \tilde{q}_{b,i\theta}}{\partial \theta} - \frac{\partial q_{b,i\theta_0}}{\partial \theta} \right\|_{\Gamma_q} \tag{55}$$

Denote as in Lemma A.2 the numerator and denominator of the conditional mean expectation as $\hat{q}_{j,i\theta} = \hat{r}\hat{q}_{j,i\theta}/\hat{f}_i$ and for its first derivative as $\frac{\partial \hat{q}_{j,i\theta}}{\partial \theta} = \hat{q}_{j,i\theta}^{(1)} = \hat{r}\hat{q}_{j,i\theta}^{(1)}/\hat{f}_i$ for $j = \{a, b\}$. To show that the upper bound in (55) converges to zero in probability, we should require consistency of $\hat{\theta}$. We also need the following results on uniform consistency

$$\begin{aligned}
& \Pr \left\{ \sup_{i,\theta} \|\hat{r}\hat{q}_{j,i\theta} - rq_{j,i\theta}\| > \varepsilon_{j,0N} \right\} \rightarrow 0 \\
& \Pr \left\{ \sup_i \|\hat{f}_i - f_i\| > \varepsilon_{1N} \right\} \rightarrow 0 \\
& \Pr \left\{ \sup_{i,\theta} \|\hat{r}\hat{q}_{j,i\theta}^{(1)} - rq_{j,i\theta}^{(1)}\| > \varepsilon_{j,2N} \right\} \rightarrow 0
\end{aligned}$$

The bias order of these nonparametric conditional expectations is $O(h_{N_2}^2)$ as shown in Lemma A.3 as long as the conditions there are satisfied for $\varphi(z_{1i}, z_{ci}, Z_2; \theta_0) = \frac{\partial q_j(z_{1i}, z_{ci}; \theta_0)}{\partial \theta}$ for $j = \{a, b\}$. The above uniform convergence results hold, using Lemmas A.5, A.6, A.8 and A.9 in Ichimura (1993), if there exist sequences $\{\varepsilon_{j,0N_2}\}$ and $\{\varepsilon_{j,2N_2}\}$ such that

$$\begin{aligned}
& \varepsilon_{j,0N_2} h_{N_2} M_{j,0N_2}^{l-1} \rightarrow \infty; (\log h_{N_2})(1 + M_{j,0N_2} \varepsilon_{j,0N_2}) / (N_2 h_{N_2} \varepsilon_{j,0N_2}^2) \rightarrow 0 \text{ for } j = \{a, b\} \\
& \varepsilon_{j,2N_2} h_{N_2} M_{j,2N_2}^{l-1} \rightarrow \infty; (\log h_{N_2})(1 + M_{j,2N_2} \varepsilon_{j,2N_2}) / (N_2 h_{N_2} \varepsilon_{j,2N_2}^2) \rightarrow 0 \text{ for } j = \{a, b\}
\end{aligned}$$

and where $M_{j,0N_2}$ denotes a sequence of the support of the dependent variable $\rho_j(z_1, z_c, z_2, \theta)$, M_{1N} denotes a sequence containing 1, $M_{j,2N_2}$ denotes a sequence of the support of the dependent variable $\frac{\partial \rho_j(z_1, z_c, z_2, \theta)}{\partial \theta}$. These sequences exist as long as $\frac{N_2 h_{N_2}^{2+\frac{2}{l-2}}}{(-\log h_{N_2})} \rightarrow \infty$ and $\frac{N_2 h_{N_2}^{\frac{1}{l-2}}}{(-\log h_{N_2})} \rightarrow \infty$ as in

(47). Note that if $l \geq 2$, the former condition on the sequence of bandwidths implies the latter.

■

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

n=1000;N=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				

Note: The data generating process is $Y=0.5+1.5X_1+2X_2+U$ with $U \sim N(0,1)$ and $Z_c \sim N(0,1); X_2=1+X_c+X_c+\varepsilon$ and $\varepsilon \sim N(0,\sigma^2)$

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

n=1000;N=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				

Note: The data generating process is $Y=0.5+1.5X_1+2X_2+U$ with $U \sim N(0,1)$ and $X_c \sim N(0,1); X_2=1+X_c+X_c+\varepsilon$ and $\varepsilon \sim N(0,\sigma^2)$

Table 3. Monte Carlo Experiment for a linear model with exclusion restriction
Just-identified case

n=1000;N=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 4. Monte Carlo Experiment for a linear model with exclusion restriction
Over-identified case

n=1000;N=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 5. Variance decomposition of error term for Linear Model with exclusion restrictions

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

Table 6. Monte Carlo Experiment for a probit model (Z_2 dummy variable, Z_c Normal)

n=1000;N=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 7. Monte Carlo Experiment for a probit model (Z_2 uniform, Z_c Normal)

n=1000;N=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

References

- [1] Amemiya, T. (1985), "Advanced Econometrics", Harvard University Press
- [2] 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
- [3] 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
- [4] Carroll, C.D. and D. N. Neil (1993), "Saving and Growth: A Reinterpretation", NBER Working papers, No. 4470
- [5] Chen, X. , Hong, H. and E. Tamer (2004), "Measurement Error with Auxiliary Data", forthcoming *Review of Economic Studies*
- [6] Cross, P.J. and Manski, C. (2002), "Regressions, Short and Long", *Econometrica*, vol. 70, pp. 357-368
- [7] Currie, J. and A. Yelowitz, (2002) "Are Public Housing Projects Good for Kids?", *Journal of Public Economics* 75, pp.99-124
- [8] Dee, T.S. and W.N. Evans (1997), "Teen Drinking and Educational attainment: Evidence from Two-Sample Instrumental Variables", NBER No. 6082
- [9] Glasser, M. (1964) "Linear Regression Analysis with Missing Observations Among Independent Variables," *Journal of the American Statistical Association*, Vol. 51, pp.834-844.
- [10] Gourieroux, C. and A. Monfort (1981), "On the Problem of Missing Data in Linear Models", *Review of Economic Studies*, 48, pp. 579-586
- [11] Hardle, W. , Hall, P. and H. Ichimura (1993), "Optimal Smoothing in Single-Index Models", *Annals of Statistics*, 21, March 21, pp.157-178.
- [12] Horowitz, J. and Manski, C. (1995), "Identification and Robustness with Contaminated Data and Corrupted Data", *Econometrica*, vol. 63 pp.281-302
- [13] Hu, Y and Ridder, G. (2003), "Estimation of Nonlinear Models with Measurement Errors Using Marginal Information", Working paper, CLEO, University of Southern California

- [14] Imbens, G. W. and T. Lancaster (1994), "Combining Micro and Macro Data in Microeconomic Models", *Review of Economic Studies* (1994), 61- pp. 655-680
- [15] Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and weighted SLS Estimation of Single-Index Models", *Journal of Econometrics* 58, pp. 71-120
- [16] Ichimura, H. (2004), "Computation of Asymptotic Distribution for Semiparametric GMM Estimators", Department of Economics, University College London, mimeo.
- [17] Little, R.A. (1992), "Regression with Missing X's: A Review", *Journal of the American Statistical Association*, vol. 87 No. 420, pp.1227-1237
- [18] 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
- [19] Manski, C. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, pp.313-333.
- [20] 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
- [21] Newey, W. and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing", *Handbook of Econometrics*, vol4, Chapter 36
- [22] Newey, W. and Powell, J. (2003), "Instrumental Variable Estimation of Nonparametric Models".*Econometrica*, vol 71, No 5, pp. 1565 - 1578
- [23] Powell, J., Stock , J. H. and Stoker, T. (1989), "Semiparametric Estimation of Index Coefficients", *Econometrica*, vol. 57, No. 6, pp. 1403-1430
- [24] Ridder, G. and Moffit, R. (2003), "The Econometrics of Data Combination", Chapter for the *Handbook of Econometrics* (forthcoming)
- [25] Robinson, P.M. (1988), "Root-N-Consistent Semiparametric Regression", *Econometrica* 56, pp. 931-954
- [26] Rothenberg, T.J. (1971), "Identification in Parametric Models" *Econometrica* 39, pp. 377-592
- [27] 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.
- [28] Schennach, S. (2004), "Estimation of Nonlinear Models with Measurement Error", *Econometrica*, 72 pp.33-75

- [29] Serfling, R. J. (1980), "Approximation Theorems of Mathematical Statistics", Wiley Series in Probability and Statistics
- [30] van der Vaart, A. W. (1998), "Asymptotic Statistics", Cambridge Series in Statistical and Probabilistic Mathematics