Identification and Estimation of Triangular Models without Exclusion Restrictions

Monika Avila Márquez[1] and , Jaya Krishnakumar [2]

[1]School of Economics, University of Bristol
[2]Institute of Economics and Econometrics, University of Geneva

December 2023

Abstract

This paper studies different identification strategies for triangular simultaneous semiparametric equation models without exclusion restrictions. We start our study with an identification strategy based on a functional form assumption. In this setting, we derive an orthogonal score function and we provide a two-step estimator procedure using neural networks for the estimation of the nuisance parameter. Later, we relax the functional form assumption and impose restrictions on the unobservables to obtain additional moment conditions. The restrictions rely on asymmetrically distributed or heteroskedastic error terms. For estimation, we propose two-step semiparametric estimators. In the first step, we use neural networks for estimation of the nuisance parameters and in the second step we estimate the causal parameter of interest.

Keywords: Endogeneity, Artificial Neural Network, Identification, Control Function, Orthogonal score function

1 Introduction

Identification of models with endogenous variables usually relies on exclusion restrictions. When instruments are available, identification and estimation of the parameters of interest can be done using an instrumental variable approach or a control function approach. Conversely, in the absence of exclusion restrictions or available instrumental variables, one needs to adopt different identification strategies for the parameters of interest.

This paper proposes alternative identification strategies when no instrumental variables are available for triangular semiparametric models. We begin our study with a baseline triangular model composed of a linear structural equation and a nonlinear reduced form equation of the endogenous regressor. In this setting, we identify the parameters of interest of the main equation by exploiting a functional form assumption in combination with a control function approach. Later, we relax the linearity assumption in the main equation and we assume that it presents a partial linear form. For the identification of the parameters of interest in this more general model, we use two different approaches. The first one is an extension of the methodology proposed by Lewbel et al. (2020) for linear triangular models that requires that the distributions of the error terms are asymmetric. The second one extends the strategy proposed by Klein and Vella (2010) for linear triangular models and entails that the error terms are heteroskedastic. For estimation, we propose two-step semiparametric estimation methods using neural networks in the first stage to estimate the nuisance parameters. But other machine learning methods can be used for the estimation of the nuisance parameters in combination with cross-fitting as suggested by Chernozhukov et al. (2018).

More specifically, in the baseline model we use a control function approach that exploits the nonlinearity of the reduced form equation. This means we augment the main equation with estimated residuals that present independent variation from the included regressors. These residuals present independent variation because of the nonlinear relationship between the endogenous regressor and the exogenous variables. This situation could be plausible for linear structural equations with endogenous explanatory variables that are related to the included exogenous regressors. For instance, we are interested in studying the causal effect of school quality on the achievement of students and we do not have any instrumental variable. In this case, it is possible to assume a linear relationship between achievement, school quality, and socioeconomic characteristics such as household wealth and parents' educational attainment. However, we can assume that school quality has a nonlinear relationship with these socioeconomic characteristics. Then, we can add the predicted residuals of the reduced equation of school quality as a control function in the achievement equation. Let us note that the assumption that school quality has a nonlinear relationship with household wealth and parental education is plausible.

The described estimation procedure seems straightforward, but simple plugging in of the estimated residuals in the main equation only produces a good estimate of the marginal causal effect if the chosen neural network architecture is the correct one. In order to produce an estimator that retrieves consistent estimates irrespective of the neural network architecture, we follow Chernozhukov et al. (2018) by deriving an appropriate orthogonal score. This orthogonal score produces a $\sqrt{N}$-consistent estimate, with N representing the sample size, because it is not sensitive to the neural network architecture used for the estimation of the nuisance parameter. We call this estimator the Control Function-Orthogonal Score (CF-OS) estimator.

While identification and estimation of the parameters of the main equation exploiting an assumption of nonlinear reduced form equation in combination with an appropriate orthogonal score retrieve good estimates, a violation of the assumption of linearity in the structural equation invalidates the consistency of the proposed estimator. In order to deal with this problem, we relax the assumption of linearity in the main equation and assume that the structural equation has a partial linear form. In this new setting, we need to impose restrictions on the errors for identification. For this purpose, we extend Lewbel2020WP and Klein and Vella (2010) identification strategies to a semiparametric triangular model without exclusion restrictions.

In the first case, we follow Lewbel2020WP and assume that the error terms in the triangular model are asymmetrically distributed and exploit the moment conditions derived by Lewbel et al. (2020). The assumption that the errors are non-symmetrically distributed is plausible in several situations of interest. For instance, it might be more appropriate to assume that achievement is nonlinearly related to the socioeconomic variables and that its error term has an asymmetric probability density distribution. Similarly, it is possible that the distribution of the error term in the equation for the quality of school (the endogenous variable) is not symmetric. For estimation, we propose a two-step semiparametric estimator. In the first step, we estimate the conditional expectation of the dependent variable on the exogenous regressor, and the conditional expectation of the endogenous variable on the exogenous regressors. Then, we subtract these estimated conditional expectations from the independent variable and the endogenous variable respectively to construct residuals that are used in the second stage of the estimation. With these residuals, we estimate the parameter of interest using the moment conditions proposed by Lewbel et al. (2020). Therefore, the estimator proposed is a two-step semiparametric estimator with nuisance parameters estimated in the first step using neural networks.

In the second case, we assume that the error terms of both equations of the semiparametric triangular model are heteroskedastic instead of assuming that they are asymmetric. This is an extension of the identification strategy proposed by Klein and Vella (2010). The assumption of heteroskedastic error terms is plausible in settings where the dependent variable and endogenous regressor present dispersion as a function of the exogenous regressors. For instance, it is possible that the dispersion of achievement of students increases with wealth. In addition, the error in the equation for the quality of school could be heteroskedastic. In this setting, the proposed estimator is a two-step estimator. The first step is the same as the one for the estimator proposed using asymmetric error terms. In the second step, we need to estimate the coefficients by optimizing a loss function that considers the heteroskedasticity of the structural error term.

The simulation results show that the proposed methods have lower RMSE and bias than OLS. The estimator CF-OS outperforms all other methods when the data-generating process is a triangular model with a linear structural equation and a non-linear reduced equation. When the model is semiparametric with asymmetric error terms, the two-step semiparametric estimator using the moment conditions proposed by Lewbel et al. (2020) is the most appropriate. Finally, the simulation results for the estimator that exploits heteroskedasticity of the error terms show that the estimator produces better estimates than OLS when the model is semiparametric with heteroskedastic error terms.

This paper is related to different strands of literature: identification of triangular linear models without exclusion restrictions, control function approach, and orthogonal scores.

On one hand, the identification of triangular linear models without exclusion restrictions has been investigated by several authors. One can find identification strategies through heteroskedasticity that have been proposed by Klein and Vella (2010), Lewbel (2012), and Rigobon (2003). In addition, there is identification through restrictions on the distribution of the errors such as Lewbel et al. (2020). Identification through nonlinearity is not advised by Angrist and Pischke (2008) and Wooldridge (2010) if the predicted values of the endogenous variable are directly used in the main equation. In this paper, we contribute to the literature by providing identification through nonlinearity with an estimator that is robust to the estimation of the nonlinear reduced-form equation. On the other hand, the identification of triangular semiparametric models without exclusion restrictions has not been studied. This paper contributes to the literature by providing identification strategies that rely on restrictions on the unobservables.

According to Wooldridge (2015), the control function (CF) approach for a linear triangular model consists of two stages: in the first stage, the endogenous regressor is regressed on exogenous variables aiming to obtain the residuals and in the second stage, the estimated residuals are included in the structural equation to control for endogeneity. In order to identify all parameters of the augmented structural equation, it is necessary that the estimated residuals have independent variation from the included regressors in the main equation. In other words, the reduced equation must present excluded regressors. Wooldridge (2015) explains that when the reduced-form equation is nonlinear, the CF approach is more efficient than two stage least squares estimation method (2SLS) but it is nonrobust to misspecification of the nonlinear function. This paper contributes to the literature by making robust the CF approach to the misspecification of the nonlinear function. For this purpose, we derive an orthogonal score that allows us to estimate the parameter of interest without contamination from the estimation of the nonlinear reduced-form equation using neural networks. As mentioned, this robust CF approach is called CF-OS. In addition, the CF-OS method can be used when there are available instrumental variables, and when there is no endogeneity after a minor modification.

Chernozhukov et al. (2018) defines an orthogonal score as the one that presents a vanishing Gateaux derivative with respect to the nuisance parameters when evaluated at the true finite-dimensional parameter values. The orthogonal score is closely related to Robinson (1988). This Chapter contributes to the literature by providing an orthogonal score that makes the CF approach robust to the estimation of the reduced-form equation.

In Section 2, we present the baseline model. In Section 3 we describe the identification strategy. In Section 4 we describe the naive control function approach. In Section 5 we propose an orthogonal control function approach. In Section 6 we compare the proposed orthogonal control function estimator with 2SLS estimation. In Section 7 we relax the assumption of linearity of the main equation, we provide two identification strategies and their corresponding estimators. In Section 8 we present specification tests. In Section 9 we discuss the usefulness of the additional moment conditions. In Section 10 we present the simulation experiment and its results, and in Section 11 we present the conclusions of the paper.

Notation: We denote random variables with bold lowercase letters ($\mathbf{x}_1$), a realization of a random variable is denoted with light lowercase letters and indexed by i (x_{i1}). Vectors are denoted by light lowercase letters (x_1) and matrices by light uppercase letters (Z).

2 Theoretical setup

Assumption 1 *$\{y_i, x_{i1}, x_{i2}\}_{i=1}^N$ are identical and independent copies of $(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2) \in Y \times [-M, M] \times [-1, 1]$ with $\boldsymbol{x}_2$ a continuously distributed random variable with compact support $[-1, 1]$ and finite fourth moments, and $\boldsymbol{x}_1$ a continuously distributed random variable with compact support $X_1 \subset [-M, M]$ $M > 0$ and finite fourth moments.*

The assumption that $\mathbf{x}_1$ and $\mathbf{x}_2$ are scalar regressors is without loss of generality because the methods presented in this paper are easily generalized to settings with non-scalar regressors.

Assumption 2 *The structural equation has a linear form with and additive unobserved error term*

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_{1_o} + \boldsymbol{x}_2\beta_{2_o} + \varepsilon. \tag{1}$$

The unobserved random term ε has finite fourth moments. The unobserved parameter of interest β_{1_o} is a scalar structural parameter. Note that the parametric assumption of linearity is relaxed in Section 7.

Assumption 3 *The error term ε is mean independent of $\boldsymbol{x}_2$*

$$E[\varepsilon|\boldsymbol{x}_2] = E[\varepsilon] = 0. \tag{2}$$

Assumption 3 implies that the structural disturbance term ε is uncorrelated with $\mathbf{x}_2$ and any function of $\mathbf{x}_2$. Thus, the variable $\mathbf{x}_2$ is exogenous. This assumption is stronger than no correlation of ε with $\mathbf{x}_2$.

Assumption 4 *The conditional expectation of the regressor $\boldsymbol{x}_1$ on $\boldsymbol{x}_2$ has an unknown nonlinear functional form*

$$E[\boldsymbol{x}_1|\boldsymbol{x}_2] = g_o(\boldsymbol{x}_2) \neq E[\boldsymbol{x}_1]. \tag{3}$$

Assumption 4 states that $\mathbf{x}_1$ is mean dependent on $\mathbf{x}_2$, thus by definition $\mathbf{x}_1$ can be decomposed into two parts: the conditional expectation of $\mathbf{x}_1$ on $\mathbf{x}_2$ ($E[\mathbf{x}_1|\mathbf{x}_2]$) and an error $\mathbf{u}$ that has zero mean conditional on $\mathbf{x}_2$ ($E[\mathbf{u}|\mathbf{x}_2] = 0$).

$$\mathbf{x}_1 = g_o(\mathbf{x}_2) + \mathbf{u}. \tag{4}$$

In addition, Assumption 4 implies that $\mathbf{u}$ has independent variation from $\mathbf{x}_2$ and that $\mathbf{u}$ is uncorrelated with any function of $\mathbf{x}_2$. Therefore, $\mathbf{u}$ is a reduced-form disturbance.

Assumption 4 can be modified to include instrumental variables if they are available. Thus, the method CF-OS, proposed in section 5, can be also used with instrumental variables.

Decomposing $\mathbf{x}_1$ into a conditional expectation and an additive disturbance term makes sense under the assumption that $\mathbf{x}_1$ is continuous (Assumption 1). The first stage of the CF-OS method, presented in Section 5, holds under Assumptions 1 and 4. But Assumption 4 needs to be modified when $\mathbf{x}_1$ is a non-continuous endogenous explanatory variable. In this case, a more appropriate assumption is:

$$\mathbf{x}_1 = \tilde{g}_o(\mathbf{x}_2, \mathbf{u}).$$

In this situation, the first stage of the CF-OS method presented in Section 5 is not appropriate anymore because we are facing a nonseparable reduced-form equation. A possible solution could be retrieving $\mathbf{u}$ by following a procedure described by Matzkin (2016). She explains that if $\tilde{g}_o$ is strictly increasing on $\mathbf{u}$, $\mathbf{x}_2$ and $\mathbf{u}$ are independent, the cumulative density distribution of $\mathbf{u}$ is strictly increasing ($F_{\mathbf{u}}$), and normalizing $F_{\mathbf{u}} = \mathbf{u}$, we can show that $\tilde{g}_o(\mathbf{x}_2, \mathbf{u}) = F^{-1}_{\mathbf{x}_1|\mathbf{x}_2 = x_2}(\mathbf{u})$. Then, $\mathbf{u}$ can be obtained by $\mathbf{u} = F_{\mathbf{x}_1|\mathbf{x}_2 = x_2}(\mathbf{x}_1)$ with $F_{\mathbf{x}_1|\mathbf{x}_2 = x_2}$ equal to the conditional quantile function. This is left for further research.

Assumption 5 *The structural error term ε is mean dependent on the reduced-form error term $\boldsymbol{u}$*

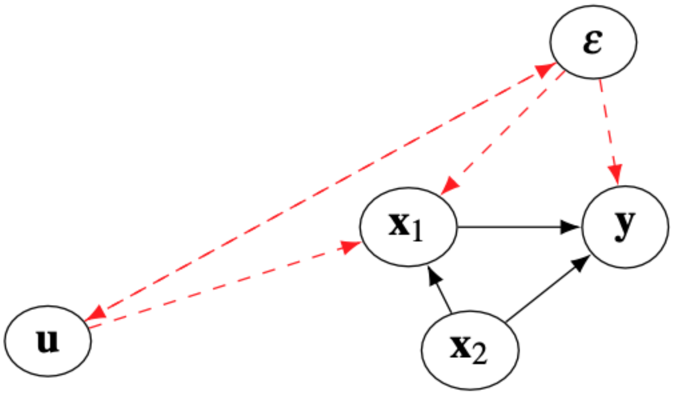
$$E[\varepsilon|\boldsymbol{u}] \neq E[\varepsilon] = 0. \tag{5}$$

Assumption 5 means that ε can be decomposed into two components: one that is mean dependent of $\mathbf{u}$ and an error term ω that has zero mean conditional on $\mathbf{u}$ as follows:

$$\varepsilon = E[\varepsilon|\mathbf{u}] + \omega.$$

Assumption 5 is not necessary for consistency of the estimation method CF-OS proposed in Section 5. Therefore, the CF-OS method presented in Section 5 can be used when the variable of interest is affected by observables (selection on observables) after a minor modification. The CF-OS has to be modified because there is no need to add the control function $\mathbf{u}\rho_o + \omega$. The reason is that ε^* is orthogonal to $\mathbf{u}$ when there is no endogeneity. More specifically, the orthogonal score is valid after replacing y^{**} by y^* (See Section 5 for definitions of ε^*, y^{**} and y^*).

Figure 1: DAG



Assumption 6

$$E[\varepsilon|\boldsymbol{u}] = \rho_o \boldsymbol{u}. \tag{6}$$

Assumption 6 states that the conditional expectation of ε on $\mathbf{u}$ has a linear functional form with ρ_o an unobserved parameter.
As consequence of Assumptions 4, 5, and 6, the structural error term ε is correlated with $\mathbf{x}_1$. The reason is that $E[\mathbf{x}_1\varepsilon] = E[(g_o(\mathbf{x}_2)+\mathbf{u})\varepsilon]$ by Assumption 4, and the last expression is equal to $E[(g_o(\mathbf{x}_2)+\mathbf{u})(\rho_o\mathbf{u}+\omega)] = \rho_o E[\mathbf{u}^2]$ by Assumptions 5 and 6.
The error term ω has zero mean conditional on $\mathbf{x}_1$ and $\mathbf{x}_2$ ($E[\omega|\mathbf{x}_1,\mathbf{x}_2] = 0$) because $E[\omega|\mathbf{x}_1,\mathbf{x}_2] = E[\omega|\mathbf{x}_2,\mathbf{u}] = E[\omega|\mathbf{u}] = 0$ where the first equality holds because $\mathbf{x}_1$ is a one-to-one function of $\mathbf{u}$ conditional on $\mathbf{x}_2$ (Wooldridge (2010)), the second equality holds by Assumption 3, and the last equality holds by Assumption 5.

The setup presents a structural equation that has a linear functional form with the outcome variable $\mathbf{y}$ explained by the endogenous variable $\mathbf{x}_1$ and the exogenous regressor $\mathbf{x}_2$. The endogenous regressor ($\mathbf{x}_1$) is mean dependent on the exogeneous regressor ($\mathbf{x}_2$.) The linear specification of the structural equation is commonly used in empirical studies and we relax it in Section 7. The model does not present external instrumental variables for $\mathbf{x}_1$, but this assumption is not needed for consistency of the estimation method CF-OS presented in 5. In fact, the estimation method CF-OS presented in 5 can be also used in settings where: 1) the regressor $\mathbf{x}_1$ is exogenous but affected by observables (selection on observables) or 2) there are available instrumental variables for the endogenous regressor $\mathbf{x}_1$. We illustrate the assumptions presented using a Directed Acyclic Graph (Fig. 1).

In addition, as mentioned the assumption that $\mathbf{x}_1$ and $\mathbf{x}_2$ are scalar regressors is without loss of generality because the methods presented in this paper are easily generalized to settings with non-scalar regressors. Finally, the assumption that the regressor $\mathbf{x}_1$ is a bounded variable is not more limiting than the assumption that $\mathbf{x}_1 = g_o(\mathbf{x}_2)+\mathbf{u}$ with g_o a bounded function (Farrell et al. (2021)).

3 Identification

The identification of the parameter of interest β_{1_o} relies on the linearity assumption of the structural equation (Assumption 2), the nonlinearity of the reduced form equation (Assumption 4), and the mean independence of the structural error term ε with the exogenous regressor $\mathbf{x}_2$ (Assumption 3). As a consequence of these assumptions, $\mathbf{u}$ presents independent variation from $\mathbf{x}_2$ and $\mathbf{x}_1$. Therefore, $\mathbf{u}$ can be used as a control function that renders $\mathbf{x}_1$ exogenous. Additionally, $\mathbf{u}$ is identifiable because has zero mean conditional on $\mathbf{x}_2$ (Assumption 3).

Using the mentioned assumptions, it is tempting to rewrite the structural equation as follows:

$$\mathbf{y} = g_o(\mathbf{x}_2)\beta_{1_o} + \mathbf{x}_2\beta_{2_o} + \mathbf{u}\beta_{1_o} + \varepsilon. \tag{7}$$

In this equation, one can believe that estimation is straightforward (if we are convinced of the non-linearity assumption) because the new error term $\mathbf{u}\beta_{1_o} + \varepsilon$ in 7 is conditionally independent of $g_o(\mathbf{x}_2)$ and $\mathbf{x}_2$. Moreover, one can think that the best estimation procedure is to estimate $g_o(\mathbf{x}_2)$ in the first stage and in the second stage plug it in (7) and perform simple OLS. This is not possible because the regularization bias of the first stage estimator of $g_o(\mathbf{x}_2)$ contaminates the estimation of β_{1_o} and β_{2_o} (Chernozhukov et al. (2018), Robinson (1988)) and produces estimates that are not $\sqrt{N}$-consistent.

In order to deal with this problem, we can use a control function approach and augment the structural equation by controlling for the unobservable $\mathbf{u}$ as follows:

$$\mathbf{y} = \mathbf{x}_1\beta_{1_o} + \mathbf{x}_2\beta_{2_o} + \rho_o\mathbf{u} + \omega. \tag{8}$$

In model 8, we can guarantee linear independence of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{u}$ thanks to the nonlinearity assumption of the reduced form equation. In addition, we can set up the following moment conditions:

$$m_1(\mathbf{Z}; \theta_o, g_o(\mathbf{x}_2)) = E[\phi(\mathbf{Z}; \theta_o, g_o(\mathbf{x}_2))] = E[\omega\mathbf{Z}] = \mathbf{0}, \tag{9}$$

where $\phi(\mathbf{Z}; \theta_o, g_o(\mathbf{x}_2))$ is a vector of score functions equal to $\omega\mathbf{Z}$, $\theta_o = [\beta_{1_o} \quad \beta_{2_o} \quad \rho_o]'$, $\omega = \mathbf{y} - \mathbf{x}_1\beta_{1_o} - \mathbf{x}_2\beta_{2_o} - \rho_o\mathbf{u}$, $\mathbf{0} = [0 \quad 0 \quad 0]'$, and $\mathbf{Z} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{u}]'$.

These moment conditions can be seen as the first-order conditions of the minimization of the following loss function:

$$Q_o(\mathbf{Z}; \theta, g_o(\mathbf{x}_2)) = E[(\mathbf{y} - \mathbf{x}_1\beta_1 - \mathbf{x}_2\beta_2 - \mathbf{u}\rho)^2].$$

This interpretation is important for consistency of the estimator proposed in Section 4 because it permits to specify the necessary primitive condition that $Q_o(\mathbf{Z}; \theta, g_o(\mathbf{x}_2))$ has a unique global optimizer θ_o (Newey and McFadden (1994)).

Now, we might believe that it is possible to estimate the parameters of interest using the naive control function approach. But in Section 4, we show that estimation using the score $\phi(\mathbf{Z}; \theta_o, g_o(\mathbf{x}_2))$ is not optimal because it is sensible to the first stage estimation of $g_o(\mathbf{x}_2)$. In order to solve this problem, in Section 5 we derive an orthogonal score that retrieves $\sqrt{N}$-consistent estimation of the parameter of interest.

4 Naive Control Function Approach

As described in the previous section, we can write the structural equation (Assumption 2) for observation i as:

$$y_i = x_{i1}\beta_{1_o} + x_{i2}\beta_{2_o} + \rho_o u_i + \omega_i. \tag{10}$$

It is clear that if u_i is observed, we can identify and estimate the three parameters β_{1_o}, β_{2_o}, ρ_o because x_{i1}, x_{i2} and u_i are independent of the error term ω_i. Thus, we could retrieve the parameters using the moment conditions 9. But since u_i is not observed, we propose a two-step estimation procedure described in the following subsection.

4.1 Estimation

First Step:
In the first step, we estimate u_i using the residuals $\hat{u}_i = x_{i1} - \hat{g}_o(x_{i2})$ obtained from a first stage estimation of $g_o(x_{i2})$. For this purpose, we propose to learn $g_o(x_{i2})$ using a machine learning technique. In particular, a suitable method is a feed-forward multilayer perceptron RELU architecture. We use this specific machine learning method because its theoretical properties permit us to derive the asymptotic properties of our second step estimator of β_{1_o}.

Second Step:
In the second step, the estimator of θ_o is the solution of the following sample moment conditions:

$$\frac{1}{N}\sum_{i}^{N}\phi(\hat{Z}_i;\theta_o,\hat{g}_o(x_{i2}))) = 0, \tag{11}$$

where $\hat{Z}_i = [x_{i1} \quad x_{i2} \quad \hat{u}_i]'$, and $\hat{u}_i$ are the residuals obtained in the first stage.

This is equivalent to estimate θ_o by performing ordinary least squares regression of y_i on x_{i1}, x_{i2} and $\hat{u}_i$ or by minimization of the following quadratic loss function :

$$\hat{\theta} = {}^{argmin}_{\theta} Q_n(\hat{Z};\theta,\hat{g}_o(x_2)),$$

with $Q_n(\hat{Z};\theta,\hat{g}_o(x_2)) = \frac{1}{N}\sum_i^N (y_i - x_{i1}\beta_1 - x_{i2}\beta_2 - \hat{u}_i\rho)^2$, x_2 a $N \times 1$ vector collecting all observations of x_{i2}, and $\hat{Z}$ a $N \times 3$ matrix collecting all observations of x_{i1}, x_{i2}, and $\hat{u}_i$.

4.2 Statistical properties

Before evaluating the statistical properties of the estimator derived from the naive control function approach, we need to add the following assumption to guarantee the consistent estimation of $g_o(\mathbf{x}_2)$.

Assumption 7 *$g_o(\boldsymbol{x}_2)$ lies in a δ-Hölder ball $W^{\delta,\infty([-1,1])}$*

$$g_o(\boldsymbol{x}_2) \in W^{\delta,\infty([-1,1])} := \{g_o(\boldsymbol{x}_2) : {}_{\alpha,|\alpha|<\delta,}^{max} {}_{\boldsymbol{x}_2 \in [-1,1]}^{ess \quad sup} |D^{\alpha} g_o(\boldsymbol{x}_2)| \leq 1\}$$

Assumption 7 is equivalent to Assumption 2 of Farrell et al. (2021) and states that $g_o(\mathbf{x}_2)$ is a uniform continuous function with smoothness parameter $\delta \in \mathbb{N}_+$ and $D^{\alpha} g_o(\mathbf{x}_2)$ its weak derivative with $\delta > 1$. This assumption states that we consider a broad class of functions that are smooth.

4.2.1 1. Consistency of the first step estimator $\hat{g}_o(x_{i2})$

Lemma 1 ***Non-asymptotic high probability bound for $\hat{g}_o(x_{i2})$***
If i) assumptions 1, 4, and 7 hold, ii) $\hat{g}_o(x_{i2})$ is a deep Multilayer Perceptron RELU (MLP-RELU) network estimator restricted to the MLP architecture class (F_{MLP}) for a loss function $l(g(\boldsymbol{x}_2);\boldsymbol{x}_1,\boldsymbol{x}_2)$[1] that is Lipschitz in $g(\boldsymbol{x}_2)$, $l(g(\boldsymbol{x}_2);\boldsymbol{x}_1,\boldsymbol{x}_2)$ and obeys a curvature condition around $g(\boldsymbol{x}_2)$, such that for constants c_1,c_2,C_l bounded away from zero:

$$|l(g(\boldsymbol{x}_2);\boldsymbol{x}_1,\boldsymbol{x}_2) - l(g'(\boldsymbol{x}_2);\boldsymbol{x}_1,\boldsymbol{x}_2)| \leq C_l |g(\boldsymbol{x}_2) - g'(\boldsymbol{x}_2)|,$$

$$c_1 E(g(\boldsymbol{x}_2) - g_o(\boldsymbol{x}_2))^2 \leq E[l(g(\boldsymbol{x}_2);\boldsymbol{x}_1,\boldsymbol{x}_2] - E[l(g_o(\boldsymbol{x}_2);\boldsymbol{x}_1,\boldsymbol{x}_2] \leq c_2 E[g(\boldsymbol{x}_2) - g_o(\boldsymbol{x}_2)]^2,$$

iii) $\hat{g}_o(\boldsymbol{x}_2)$ has width $W \asymp n^{\frac{d}{2(\delta+d)}} log^2 n$ and depth $L \asymp logn$, then:

$$E[(\hat{g}_o(\boldsymbol{x}_2) - g_o(\boldsymbol{x}_2))^2] \leqslant C\{N^{-\frac{\delta}{\delta+d}} log^8 N + \frac{loglogN}{N}\},$$

where $C > 0$ is a constant, $\delta \in \mathbb{N}_+$ is a smoothness parameter of the Hölder ball, $d = 1$ when we only have one exogenous variable.

The proof of this lemma follows from Theorem 1 presented by Farrell et al. (2021).

[1] If x_{i1} is continuous, the loss function is the $l-2$ norm or quadratic loss.

4.2.2 2. Consistency of the second step estimator $\hat{\theta}_o$

Theorem 1 *If i) conditions of Lemma 1 are satisfied such that $\hat{g}_o(x_{i2}) - g_o(x_{i2}) = o_p(1)$, ii) (a) $E(|x_{i1}|^2) < \infty$, $E(|x_{i2}|^2) < \infty$, $E(|\hat{u}_i|^2) < \infty$, ii) (b) $plim N^{-1}\sum_i^N \hat{Z}_i\hat{Z}_i' = Q_{ZZ}$ is positive definite, then:*

$$\hat{\theta}_o - \theta_o = o_p(1).$$

Condition i) states that the first step estimator is consistent. Condition ii.a) states that the expected absolute value of the elements of $\hat{Z}_i\hat{Z}_i'$ are finite. Condition ii.b) states that u_i has independent variation from x_{i2} which is implied by assumptions 2, 3, and 4. This is equivalent to a full rank condition for Z_i. The proof of this Theorem is presented in the supplemental material.

4.2.3 3. Rate of convergence of $\hat{\theta}_o$ is slower than $1/\sqrt{N}$

Proposition 1 *If i) conditions of Lemma 1 are satisfied such that $\hat{g}_o(x_{i2}) - g_o(x_{i2}) = o_p(1)$, ii) $plim N^{-1}\sum_i^N \hat{Z}_i\hat{Z}_i' = Q_{ZZ}$ is positive definite and bounded, iii) $(\hat{g}_o(x_{i2}) - g_o(x_{i2}))u_i = o_p(N^{-1/2})$, then:*

$$|\sqrt{N}(\hat{\theta}_o - \theta_o)| \xrightarrow{p} \infty.$$

This proposition states that the estimator presents a rate of convergence that is slower than $1/\sqrt{N}$ even if the first step estimator is consistent, u_i has independent variation from x_{i2} as it is implied by assumptions 2, 3, and 4, and if the estimation error of $g_o(x_{i2})$ multiplied by the unobserved random variable u_i converges to 0 at the rate $N^{-1/2}$. The reason is that the estimation error of learning $g_o(x_{i2})$ is multiplied by x_{i1}, and x_{i2} which are non-centered at 0.

The proof of this Proposition is presented in the supplemental material.

5 Control Function Approach-Orthogonal Score (CF-OS)

The main problem of the estimation presented in the previous section is that the score functions $\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))$ are not Neyman-orthogonal to $g_o(\mathbf{x}_2)$ (For a detailed definition of Neyman-orthogonal scores see Chernozhukov et al. (2018)). Using an orthogonal score is necessary to obtain $\sqrt{N}$-consistent estimates of the finite-dimensional parameters in the presence of high-dimensional nuisance parameters (Robinson (1988)).

In order to determine whether the score functions $\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))$ have the Neyman orthogonal property, we need to check two conditions:

1. $E[\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))] = 0$.
2. The Gateaux derivative operator of the score $\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))$ in the direction $g(\mathbf{x}_2) - g_o(\mathbf{x}_2)$ vanishes.

Condition 1 is satisfied as explained in section 3. In order to check the second condition, we derive the Gateaux derivative of $\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))$ in the direction $g(\mathbf{x}_2) - g_o(\mathbf{x}_2)$:

$$\partial_{r=0}E[\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))] = E[(g(\mathbf{x}_2) - g_o(\mathbf{x}_2))\tilde{\mathbf{Z}}] \neq 0,$$

with $\tilde{\mathbf{Z}} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad 0]'$.

Then, the Gateaux derivative of $\phi(\mathbf{Z};\theta_o, g_o(\mathbf{x}_2))$ in the direction $g(\mathbf{x}_2) - g_o(\mathbf{x}_2)$ does not vanish. As a result, the method presented in subsection 4.1 is sensible to biases (overfitting or regularization) in the estimation of the nuisance parameter $g_o(\mathbf{x}_2)$. As a solution, we derive the Neyman orthogonal score using the following methodology.

First, we write the structural equation, after replacing the control function, in matrix form as follows:

$$y = \beta_{1_o}x_1 + \beta_{2_o}x_2 + \rho_o u + \omega. \tag{12}$$

Then, we concentrate out the parameter β_{2_o} by premultiplying equation 8 by the annihilator matrix $M_{x_2} = I_N - x_2(x_2'x_2)^{-1}x_2'$:

$$M_{x_2}y = M_{x_2}x_1\beta_{1_o} + \rho_o M_{x_2}u + M_{x_2}\omega, \tag{13}$$

Renaming the transformed variables of the model 13 yields the following triangular system:

$$y^* = x_1^* \beta_{1_o} + \rho_o u^* + \omega^*, \tag{14}$$

$$x_1 = g_o(x_2) + u, \tag{15}$$

with $E(\omega^{*\prime} u) = 0$, $g_o(x_2)$ and ρ_o considered as nuisance parameters [2].
Now, we follow Chernozhukov et al. (2018) and obtain the orthogonalised score function:

$$\psi(W_i; \beta_{1_o}, g_o(x_{i2})) = (y_i^{**} - x_{i1}^* \beta_{1_o})(x_{i1} - g_o(x_{i2})). \tag{16}$$

The orthogonal score 16 can be re-written as:

$$\psi(W_i; \beta_{1_o}, g_o(x_{i2})) = (y_i^{**} - x_{i1}^* \beta_{1_o}) u_i, \tag{17}$$

where $W_i = [y_i^{**} \quad x_{i1}^* \quad u_i]'$, y_i^{**} is a typical element of $y^* - \rho_o u^*$, and x_{i1}^* is a typical element of $x_1^* = M_{x_2} x_1$. We derive β_{1_o}, using the orthogonal score 17, as:

$$\beta_{1_o} = E[u_i x_{i1}^*]^{-1} E[u_i y_i^{**}]. \tag{18}$$

Now, we are left to check if the new score function $\psi(W_i; \beta_{1_o}, g_o(x_{i2}))$ has the Neyman orthogonal property. As before, we check two conditions:

1. $E[\psi(W_i; \beta_{1_o}, g_o(x_{i2}))] = 0$.
2. The expectation of the Gateaux derivative operator of the orthogonal score $\psi(W_i; \beta_{1_o}, g_o(x_{i2}))$ in the direction $g(\mathbf{x}_2) - g_o(\mathbf{x}_2)$ vanishes.

Condition 1 is clearly satisfied. In order to check the second condition, we derive the Gateaux derivative of $\psi(W_i; \beta_{1_o}, g_o(x_{i2}))$ in the direction $g(x_{i2}) - g_o(x_{i2})$ and express it in matrix form as follows:

$$\begin{aligned} \partial_{r=0} E[\psi(W; \beta_{1_o}, g_o(x_2))] = &- E[(\rho M_{x_2}(g(x_2) - g_o(x_2)))'(x_1 - g_o(x_2)) \\ &- (M_{x_2} y - M_{x_2} x_1 \beta_{1_o} - M_{x_2} u \rho_o)'(g(x_2) - g_o(x_2))] = 0. \end{aligned} \tag{19}$$

The latter result is obtained using the law of iterated expectations. The expectation of the first term is equal to $E[E[(\rho M_{x_2}(g(x_2) - g_o(x_2)))'(x_1 - g_o(x_2))|x_2]] = E[(\rho M_{x_2}(g(x_2) - g_o(x_2))'(E[x_1|x_2] - g_o(x_2)))] = 0$. A similar procedure applies to the second term.

The orthogonality of the score function 17 with respect to the high-dimensional nuisance parameter $g_o(x_2)$ allows to obtain a $\sqrt{N}$-consistent estimator of the parameter of interest (Chernozhukov et al. (2018), Robinson (1988)).

An estimator of 18 is unfeasible because u_i and ρ_o are unknown. Thus, we present a feasible estimation procedure in the following subsection.

5.1 Estimation

In order to make the estimation feasible, we need a two-step procedure.

First Step:
In the first step, we estimate u_i using the residuals $\hat{u}_i = x_{i1} - \hat{g}_o(x_{i2})$ obtained from a first stage estimation of equation $g_o(x_{i2})$. For this purpose, we propose to learn $g_o(x_{i2})$ using a machine learning technique. In particular, a suitable method is a feed-forward multilayer perceptron RELU architecture. We use this specific machine learning method because its theoretical properties permit us to derive the asymptotic properties of our second step estimator of β_{1_o}.

[2] The other nuisance parameters β_{2_o} has been removed by applying the Frisch-Waugh-Lovell theorem.

Second step
In the second step, we need to estimate parameters ρ_o and β_{1_o}. For this purpose, we use the residuals $\hat{u}_i$ obtained in the first stage and replace them in the following sample moment conditions:

$$\frac{1}{N} m(\hat{H};\hat{\rho}_o,\hat{g}_o(x_2))) = 0, \tag{20}$$

$$\frac{1}{N} \psi(\hat{W};\hat{\beta}_{1_o},\hat{g}_o(x_2))) = 0, \tag{21}$$

where $m(\hat{H};\hat{\rho}_o,\hat{g}_o(x_2))) = (M_{x_1^*}y^* - \hat{\rho}_o M_{x_1^*}\hat{u}^*)' M_{x_1^*}\hat{u}^*$ and $\psi(\hat{W};\hat{\beta}_{1_o},\hat{g}_o(x_2))) = \hat{u}'(\hat{y}^{**} - x_1^*\hat{\beta}_{1_o})$.
The estimators are the solutions to 20 and 21:

$$\hat{\rho}_o = (\hat{u}^{*\prime} M_{x_1^*}\hat{u}^*)^{-1}(\hat{u}^{*\prime} M_{x_1^*} y^*), \tag{22}$$

$$\hat{\beta}_{1_o} = (\hat{u}' x_1^*)^{-1}(\hat{u}'\hat{y}^{**}). \tag{23}$$

5.2 Statistical Properties

5.2.1 Consistency of the second step estimator $\hat{\rho}_o$

Theorem 2 *If i) conditions of Lemma 1 are satisfied such that $\hat{g}_o(x_2) - g_o(x_2) = o_p(1)$, ii) $plim N^{-1}\hat{u}^{*\prime} M_{x_1^*}\hat{u}^*$ is bounded away from 0 and finite, then:*

$$\hat{\rho}_o - \rho_o = o_p(1).$$

Condition i) states that the first step estimator $\hat{g}_o$ is consistent. Condition ii) states that $\hat{u}^*$ is not orthogonal to $x_1^* = M_{x_2}x_1$ which is satisfied by assumption 4.

The proof of this Theorem is presented in the supplemental material.

5.2.2 Consistency of the second step estimator $\hat{\beta}_{1_o}$

Theorem 3 *If i) conditions of Lemma 1 are satisfied such that $\hat{g}_o(x_2) - g_o(x_2) = o_p(1)$, ii) $plim N^{-1}\hat{u}' x_1^*$ is bounded away from 0 and finite, then:*

$$\hat{\beta}_{1_o} - \beta_{1_o} = o_p(1).$$

Condition i) states that the first step estimator is consistent. Condition ii) states that $\hat{u}^*$ is not orthogonal to $x_1^* = M_{x_2}x_1$ which is satisfied by assumption 4.

The proof of this Theorem is presented in the supplemental material.

5.2.3 Asymptotic distribution of the estimator $\hat{\beta}_{1,o}$

Proposition 2 *If i) conditions of Lemma 1 are satisfied such that $\hat{g}_o(x_2) - g_o(x_2) = o_p(1)$, ii) $plim N^{-1}\hat{u}' x_1^*$ is finite and bounded away from 0 and finite, iii) $(\hat{g}_o(x_2) - g_o(x_2))' M_{x_2}\omega = o_p(N^{-1/2})$, iv) $(\hat{g}_o(x_2) - g_o(x_2))' M_{x_2}\xi = o_p(N^{-1/2})$ then:*

$$\sqrt{N}(\hat{\beta}_{1_o} - \beta_{1_o}) \xrightarrow{L} N(0, D^{-1}VD^{-1}),$$

with $D = plim N^{-1}\hat{u}' x_1^*$, and $V = plim N^{-1}E[\omega^2]u' M_{x_2}u$.

Condition i) states that the first step estimator is consistent. Condition ii) states that u^* is not orthogonal to $M_{x_2}x_1$ which is implied by assumption 4. Conditions iii) and iv) state that the inner products of the estimation error of $g_o(x_2)$ and the transformed unobserved random terms $M_{x_2}u$ and $M_{x_2}\omega$ converge in probability to 0 at the rate $N^{-1/2}$. The latter is possible because the complexity of the parameter space of $g_o(x_2)$ is bounded as shown by Farrell et al. (2021).

The proof of this Proposition is presented in the supplemental material.

6 Comparison of CF-OS with 2SLS

Estimation of the model presented in Section 2 using 2SLS, with $g_o(\mathbf{x}_2)$ and $\mathbf{x}_2$ as instrumental variables, is not possible. The reason is that 2SLS method relies on a score function that is not orthogonal to $g_o(\mathbf{x}_2)$. To see this clearer, we need to verify if the Gateaux derivative of the score function $E[\varepsilon\mathbf{d}] = 0$ with $\mathbf{d} = [g_o(\mathbf{x}_2) \quad \mathbf{x}_2]'$ vanishes. For this purpose, we derive the Gateaux derivative of the score function $E[\varepsilon\mathbf{d}] = 0$ in the direction of $g(\mathbf{x}_2) - g_o(\mathbf{x}_2)$ and obtain:

$$\partial_{r=0}E[\varepsilon g_o(\mathbf{x}_2)] = E[(g(\mathbf{x}_2) - g_o(\mathbf{x}_2))g_o(\mathbf{x}_2)] \neq 0,$$

$$\partial_{r=0}E[\varepsilon\mathbf{x}_2] = E[(g(\mathbf{x}_2) - g_o(\mathbf{x}_2)\mathbf{x}_2] \neq 0.$$

Thus, we can conclude that the 2SLS estimator (using as instrumental variables $g_o(\mathbf{x}_2)$ and $\mathbf{x}_2$) is nonrobust because it is sensitive to overfitting and regularization bias of learning $g_o(\mathbf{x}_2)$. Importantly, this conclusion does not change if there are available instrumental variables. Thus, the CF-OS is the most efficient estimation procedure and it is robust to overfitting and regularization bias in estimating the nonlinear reduced form (with or without available instrumental variables).

There are two options to avoid contamination of the 2SLS estimator due to biases in the learning of the unknown function $g_o(\mathbf{x}_2)$ (that can include or not available instrumental variables): 1) assume a parametric functional form of the reduced-form equation or 2) use polynomials of $\mathbf{x}_2$ (and available instrumental variables) as instrumental variables.

If we opt to assume a parametric functional form for the reduced-form equation, we can use nonlinear fitted values as instruments in the 2SLS procedure (Angrist and Pischke (2008)). 2SLS estimation is less efficient than CF estimation if the parametric functional form is correct (Wooldridge (2015)). Then, if the parametric form of the reduced form equation is correct, CF is preferred. But if the parametric functional form is incorrect, CF estimation is nonrobust (Wooldridge (2015)). On the contrary, the CF-OS does not need a parametric assumption of the reduced-form equation and it is the most efficient estimator.

Another option is using the 2SLS method with polynomials of $\mathbf{x}_2$ or/and possible instrumental variables. It is not clear if 2SLS, using polynomials of $\mathbf{x}_2$ as IVs, is better or worse than CF-OS. A study of the difference between 2SLS, using polynomials of $\mathbf{x}_2$ as IVs, and the CF-OS method is left for further research.

We are aware that Wooldridge (2010) does not necessarily advocate identification through nonlinearity of the reduced form. Hence, we would like to point out that our method proposed in Section 5 as well as the following ones, are still consistent and valid when there are available instrumental variables and the additional moment conditions can be used to test the validity of the instruments (See Section 9).

7 Relaxing the linear assumption

In this Section, we relax Assumption 2 by allowing a nonlinear functional form in the structural equation. We ease Assumption 2 in order to avoid spurious results when the true DGP is not the one presented in section 2.

Assumption 8 *The structural equation has a partial linear form with an additive unobserved error term.*

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_{1_o} + f_o(\boldsymbol{x}_2) + \varepsilon. \tag{24}$$

7.1 Identification and estimation through asymmetric error terms

7.1.1 A. Identification

Under the Assumptions 1 to 6, and with Assumption 8 replacing Assumption 2, β_{1_o} is not identifiable. The reason is that there are no available exclusion restrictions and the nonlinear functional form assumption does not allow to generate a reduced-form error with independent variation of $f_o(\mathbf{x}_2)$. Then, we need to impose further restrictions on the unobservables of the model (ε and $\mathbf{u}$) in order to achieve identification of β_{1_o}. For this purpose, in this Subsection, we add an assumption for ε and replace Assumption 6. More specifically, the new assumptions for the disturbance terms of the model are:

Assumption 9 *The disturbance term of the endogenous regressor* $\boldsymbol{u}$ *is partitioned into two terms.*

$$\boldsymbol{u} = \boldsymbol{u}_1 + \boldsymbol{u}_2.$$

Assumption 10 *The structural disturbance term presents a linear relationship with one of the components of the disturbance term of the endogenous regressor.*

$$\varepsilon = \gamma_o \boldsymbol{u}_1 + \varsigma.$$

Assumptions 9 and 10 imply that $\mathbf{u}_2$ is an unobserved exclusion restriction.

Assumption 11 *Sign restriction on* γ_o.

$$\gamma_o > 0$$

This assumption is needed to guarantee the identification of the parameter of interest. It means that the researcher needs to know the direction of the correlation between the endogenous regressor and the unobserved structural error.

Assumption 12 *The errors* $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, ς *have zero mean and are mutually independent.*

Assumption 13 $\boldsymbol{u}_2$ *has a finite variance and* $\boldsymbol{u}_1$,ς *have finite fourth moments.*

Finally, we add an assumption for the joint distribution of the observed regressors.

Assumption 14 *The joint distribution of the dependent variable and regressors is fully observed.*

Assumptions 9 to 14 are equivalent to the ones presented by Lewbel et al. (2020).
The assumptions presented imply the following moment conditions:

$$E[\mathbf{vu} - \gamma_o \sigma^2_{\mathbf{u}_1} - \beta_{1_o}(\sigma^2_{\mathbf{u}_1} + \sigma^2_{\mathbf{u}_2})] = 0, \tag{25}$$

$$E[\mathbf{u}^2 - \sigma^2_{\mathbf{u}_1} - \sigma^2_{\mathbf{u}_2}] = 0, \tag{26}$$

$$E[(\mathbf{v} - \beta_{1_o} u)^2 - \gamma_o^2 \sigma^2_{\mathbf{u}_1} - \sigma^2_{\omega}] = 0, \tag{27}$$

$$E[(\mathbf{v} - \beta_{1_o}\mathbf{u})(\mathbf{v} - (\beta_{1_o} + \gamma_o)\mathbf{u})\mathbf{u}] = 0, \tag{28}$$

$$E[(\mathbf{v} - \beta_{1_o}\mathbf{u})(\mathbf{v} - (\beta_{1_o} + \gamma_o)\mathbf{u})(\mathbf{u}^2 - \sigma^2_{\mathbf{u}_1} - \sigma^2_{\mathbf{u}_2}) - 2\gamma_o \sigma^2_{\mathbf{u}_1}(\mathbf{v} - \beta_{1_o}\mathbf{u})\mathbf{u}] = 0. \tag{29}$$

The moment conditions 25 to 29 are an extension of the moments provided by Lewbel2020WP to a setting suitable for a triangular semiparametric model with no instrumental variables. In our case, we use $\mathbf{v} = \mathbf{y} - E[\mathbf{y}|\mathbf{x}_2]$ and $\mathbf{u} = \mathbf{x}_1 - E[\mathbf{x}_1|\mathbf{x}_2]$ following Robinson (1988).

7.1.2 B. Estimation

The estimation procedure has two steps:

Step 1: Following Robinson (1988), we obtain the residuals $\hat{v}_i = y_i - \widehat{E[y_i|x_{i2}]}$ and $\hat{u}_i = x_{i1} - \widehat{E[x_{i1}|x_{i2}]}$ by learning the conditional expectations using a feed-forward neural network.

Step 2: We use the residuals obtained in Step 1 and the moment conditions provided by Lewbel et al. (2020) to estimate the parameters of interest.

7.1.3 C. Statistical properties

In this Subsection, we call the estimator of $E[\mathbf{y}|\mathbf{x}_2]$ as $\hat{h}_o(\mathbf{x}_2)$. Additionally, we collect moments 25 to 29 in a vector called $a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)$ where $\tilde{\mathbf{Z}} = [\mathbf{u} \quad \mathbf{v}]'$, $\zeta_o = [\beta_{1_o} \quad \gamma_o \quad \sigma^2_{u_1} \quad \sigma^2_{u_2} \quad \sigma^2_{\omega}]'$, $\tau_o = [g_o \quad f_o]'$. In addition, we need to add the following assumptions to guarantee the consistent estimation of $f_o(\mathbf{x}_2)$ and $E[\mathbf{y}|\mathbf{x}_2]$.

Assumption 15 *f_o lies in a δ_2-Hölder ball $W^{\delta_2,\infty([-1,1])}$.*

$$f_o(\boldsymbol{x}_2) \in W^{\delta_2,\infty([-1,1])} := \{f_o : \overset{max}{\alpha_2,|\alpha_2|<\delta_2}, \overset{ess \quad sup}{x_2\in[-1,1]} |D^{\alpha_2} f_o(\boldsymbol{x}_2)| \leq 1\}.$$

Assumption 16 *$E[\boldsymbol{y}|\boldsymbol{x}_2]$ lies in a δ_3-Hölder ball $W^{\delta_3,\infty([-1,1])}$.*

$$E[\boldsymbol{y}|\boldsymbol{x}_2] \in W^{\delta_3,\infty([-1,1])} := \{E[\boldsymbol{y}|\boldsymbol{x}_2] : \overset{max}{\alpha_3,|\alpha_3|<\delta_3}, \overset{ess \quad sup}{\boldsymbol{x}_2\in[-1,1]} |D^{\alpha_3} E[\boldsymbol{y}|\boldsymbol{x}_2]| \leq 1\}.$$

Assumptions 15 and 16 are in line with Farrell et al. (2021) and state that $f_o(\mathbf{x}_2)$ and $E[\mathbf{y}|\mathbf{x}_2]$ are uniform smooth continuous functions. Their smoothness parameters are $\delta_2 > 1$ and $\delta_3 > 1$ respectively. $D^{\alpha_2} g_o(\mathbf{x}_2)$ and $D^{\alpha_3} g_o(\mathbf{x}_2)$ are their weak derivatives.

1. Consistency of $\hat{h}_o(\mathbf{x}_2)$.

Lemma 2 ***Non-asymptotic high probability bound for $\hat{h}_o(\boldsymbol{x}_2)$***
If i) assumptions 1, 3, 8 to 16 hold, ii) $\hat{h}_o(\boldsymbol{x}_2)$ is a deep Multilayer Perceptron RELU (MLP-RELU) network estimator restricted to the Multilayer Perceptron architecture class (F_{MLP}) for a loss function $l(h(\boldsymbol{x}_2);\boldsymbol{y},\boldsymbol{x}_2)$ that is Lipschitz in $h(\boldsymbol{x}_2)$, $l(h(\boldsymbol{x}_2);\boldsymbol{y},\boldsymbol{x}_2)$ and obeys a curvature condition around $h(\boldsymbol{x}_2)$, such that for constants c_3,c_4,C_5 bounded away from zero:

$$|l(h(\boldsymbol{x}_2);y,\boldsymbol{x}_2) - l(h'(\boldsymbol{x}_2);\boldsymbol{y},\boldsymbol{x}_2)| \leq C_5|h(\boldsymbol{x}_2) - h'(\boldsymbol{x}_2)|,$$

$$c_3E(h(\boldsymbol{x}_2) - h_o(\boldsymbol{x}_2))^2 \leq E(l(h(\boldsymbol{x}_2);\boldsymbol{y},\boldsymbol{x}_2) - E(l(h_o(\boldsymbol{x}_2);\boldsymbol{y},\boldsymbol{x}_2) \leq c_4E(h(\boldsymbol{x}_2) - h_o(\boldsymbol{x}_2))^2,$$

iii) $\hat{h}_o(\boldsymbol{x}_2)$ has width $W \asymp n^{\frac{d}{2(\delta_3+d)}} log^2 n$ and depth $L \asymp logn$, then:

$$E[(\hat{h}_o(\boldsymbol{x}_2) - h_o(\boldsymbol{x}_2))^2] \leqslant C_6\{N^{-\frac{\delta_3}{\delta_3+d}} log^8 N + \frac{loglogN}{N}\},$$

where $C_6 > 0$ is a constant, $\delta_3 \in \mathbb{N}_+$ is a smoothness parameter of the Hölder ball, $d = 1$ when we only have one exogenous variable.

The proof of this lemma follows from Theorem 1 of Farrell et al. (2021).

2. Consistency of the second step estimator $\hat{\zeta}_o$.

Proposition 3 *if i) $E[a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)] = 0$, ii) $\zeta_o \in \Theta_\zeta$ a compact space, iii) $a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)$ stochastic equicontinuous, iv) $E[sup_\zeta ||a(\tilde{\mathbf{Z}}, \zeta, \tau)||] < \infty$, then*

$$\hat{\zeta}_o \xrightarrow{p} \zeta_o.$$

Condition i) is the identification condition and it is satisfied by Lewbel et al. (2020), condition ii) is the boundedness condition on the parameter set and it is substantive, condition iii) and iv) can be shown following Newey and McFadden (1994).

3. Asymptotic distribution of the second step estimator $\hat{\zeta}_o$.

Proposition 4 *If i) $\hat{\zeta}_o \xrightarrow{p} \zeta_o$, ii) $\hat{g}_o(\boldsymbol{x}_2) - g_o(\boldsymbol{x}_2) = o_p(1)$, $\hat{h}_o(\boldsymbol{x}_2) - h_o(\boldsymbol{x}_2) = o_p(1)$, iii) $\nabla_\zeta a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)$ satisfies conditions of lemma 4.3 of Newey and McFadden (1994), iv) $E[\nabla_\theta a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)] = A$ is non singular, then*

$$\sqrt{N}(\hat{\zeta}_o - \zeta_o) \to N(0, A^{-1}V_2A^{-1}),$$

where $V_2 = E(a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)a(\tilde{\mathbf{Z}}, \zeta_o, \tau_o)')$.
This proposition is based on Theorem 8.2 of Newey and McFadden (1994), and its proof is provided by the authors.

7.2 Identification and estimation through heteroskedasticity

In the previous Subsection, we presented an identification strategy that relies on the higher moments obtained from the moment-generating function of the unobserved errors. The main assumptions are that errors must not be symmetric nor normally distributed. These assumptions might be restrictive for certain empirical applications.

In order to relax these assumptions, we extend the approach proposed by Klein and Vella (2010) for linear triangular systems with no exclusion restrictions that exploits heteroskedasticity in the error terms. In our case, we assume that the DGP is a semiparametric triangular model with heteroskedastic disturbance terms.

In particular, we replace the assumptions 9 to 14 by the following ones:

Assumption 17

$$\varepsilon = \alpha_1(\boldsymbol{x}_2)\varepsilon^*,$$

where $\alpha_1(\boldsymbol{x}_2)$ is a smooth function of the regressor $\mathbf{x}_2$, $E[\varepsilon^*|\mathbf{x}_2] = 0$, and $E[\varepsilon^{*2}|\boldsymbol{x}_2] = 1$.

Assumption 17 states that the structural error term is heteroskedastic. The error term ε^* has zero mean conditional on the regressor $\mathbf{x}_2$, and it is homoskedastic.

Assumption 18

$$\boldsymbol{u} = \alpha_2(\boldsymbol{x}_2)\boldsymbol{u}^*,$$

where $\alpha_2(\boldsymbol{x}_2)$ is a smooth function of the regressor $\mathbf{x}_2$, $E[\boldsymbol{u}^*|\boldsymbol{x}_2] = 0$, and $E[\boldsymbol{u}^{*2}|\boldsymbol{x}_2] = 1$.

Assumption 18 states that the reduced-form equation has an error term that is heteroskedastic. The error term $\mathbf{u}^*$ is zero mean conditional on the regressor $\mathbf{x}_2$, and it is homoskedastic.

Assumption 19 $\varepsilon^* = \rho_o\boldsymbol{u}^* + \upsilon^*$

Assumption 19 states that the homoskedastic error term ε^* in the main equation can be decomposed in two parts: one that is correlated to the homoskedastic error term in the reduced-form equation and a second elements υ^* that is orthogonal to $\mathbf{u}^*$. As a consequence of this assumption, the regressor $\mathbf{x}_1$ is endogenous. The reason is that $E[\varepsilon\mathbf{x}_1] = E[(\alpha_1(\mathbf{x}_2)\varepsilon^*)(g_o(\mathbf{x}_2) + \alpha_2(\mathbf{x}_2)\mathbf{u}^*)]$ by Assumptions 4, 17, and 18, and $E[(\alpha_1(\mathbf{x}_2)\varepsilon^*)(g_o(\mathbf{x}_2) + \alpha_2(\mathbf{x}_2)\mathbf{u}^*)] = \alpha_1(\mathbf{x}_2)\alpha_2(\mathbf{x}_2)\rho_o$ by Asumption 17, 18, and 19.

7.2.1 A. Identification

In order to identify the parameter of interest β_{1_o}, we follow Robinson (1988) by substracting the conditional expectation on $\mathbf{x}_2$ from the dependent variable $(\mathbf{y} - E[\mathbf{y}|\mathbf{x}_2])$ and we obtain:

$$\eta = \beta_{1_o}\mathbf{u} + \varepsilon, \tag{30}$$

where $\eta = \mathbf{y} - E[\mathbf{y}|\mathbf{x}_2]$ and $\mathbf{u} = \mathbf{x}_1 - E[\mathbf{x}_1|\mathbf{x}_2]$.

In equation 30, we cannot identify the parameter of interest β_{1_o} because the error term is mean dependent on $\mathbf{u}$ and there are no available exclusion restrictions. While identification of β_{1_o} seems unattainable, one can exploit the heteroskedasticity assumptions. First, we can notice that $E[\varepsilon|\mathbf{u}] = E[\alpha_1(\mathbf{x}_2)\varepsilon^*|\alpha_2(\mathbf{x}_2)\mathbf{u}^*]$ by Assumptions 17 and 18. Now, we can re-write the last expression as $E[\varepsilon|\mathbf{u}] = E[\rho_o\alpha_1(\mathbf{x}_2)\mathbf{u}^* + \alpha_1(\mathbf{x}_2)v^*|\alpha_2(\mathbf{x}_2)\mathbf{u}^*]$ by Assumption 19. Thus, $E[\varepsilon|\mathbf{u}] = \rho_o\alpha_1(\mathbf{x}_2)\mathbf{u}^*$. Finally, knowing that $\mathbf{u} = \alpha_2(\mathbf{x}_2)\mathbf{u}^*$ leads to $E[\varepsilon|\mathbf{u}] = \rho_o\frac{\alpha_1(\mathbf{x}_2)}{\alpha_2(\mathbf{x}_2)}\mathbf{u}$. The last expression allows us to rewrite the transformed model as follows:

$$\eta = \beta_{1_o}\mathbf{u} + \rho_o\frac{S_\varepsilon}{S_\mathbf{u}}\mathbf{u} + v^{**}, \tag{31}$$

where $S_\varepsilon = \alpha_1(\mathbf{x}_2) = \sqrt{E[\varepsilon^2|\mathbf{x}_2]}$, $S_\mathbf{u} = \alpha_2(\mathbf{x}_2) = \sqrt{E[\mathbf{u}^2|\mathbf{x}_2]}$, and $v^{**} = \alpha_1(\mathbf{x}_2)v^*$.

Therefore, the identification of the parameter of interest is transformed into a problem that demands the identification of S_ε, $S_\mathbf{u}$ and ρ_o.

Now, identification of $S_\mathbf{u}$ is guaranteed by identification of $g_o(\mathbf{x}_2)$ since it is possible to obtain $\mathbf{u} = \mathbf{x}_1 - g_o(\mathbf{x}_2)$ and $E[\mathbf{u}^2|\mathbf{x}_2] = S^2_\mathbf{u}$.

The trickiest part is the identification of S_ε since it depends on the unknown parameter β_{1_o}. Following Klein and Vella (2010), we propose to solve the following minimization problem:

$$\overset{argmin}{\beta_{1_o},\rho_o}\sum_i(\eta_i - \beta_{1_o}u_i - \rho_o\sqrt{E[(\eta_i - \beta_{1_o}u_i)^2|x_{i2}]}\frac{u_i}{S_{ui}})^2. \tag{32}$$

7.2.2 B. Estimation

The minimization problem 32 is unfeasible because of the presence of the unknown terms $\sqrt{E[\varepsilon^2|\mathbf{x}_2]}$ and $\sqrt{E[\mathbf{u}^2|\mathbf{x}_2]}$. In order to make it feasible, we need to estimate $\sqrt{E[\varepsilon^2|\mathbf{x}_2]}$ and $\sqrt{E[\mathbf{u}^2|\mathbf{x}_2]}$. Since their nonparametric estimation is computing demanding, we follow Farré et al. (2013) and propose a simplification based on two additional assumptions:

Assumption 20

$$E[\boldsymbol{u}^2|\boldsymbol{x}_2] = exp(\boldsymbol{x}_2\kappa_1)^2.$$

Assumption 21

$$E[\varepsilon^2|\boldsymbol{x}_2] = exp(\boldsymbol{x}_2\kappa_2)^2.$$

Now, we are ready to set up the estimation procedure:
Step 1:
Following Robinson (1988), we obtain estimations of $u_i = x_{i1} - E[x_{i1}|x_{i2}]$ and $\eta_i = y_i - E[y_i|x_{i2}]$ by learning the conditional expectations using a feed-forward neural network.

Step 2:
Estimation of heteroskedastic variance ($S^2_\mathbf{u}$) of $\mathbf{u}$ using the following non-linear least squares procedure:

$$\hat{\kappa}_1 := \overset{argmin}{\kappa_1}\sum_i(\hat{u}_i^2 - exp(x_{i2}\kappa_1)^2)^2.$$

Then using $\hat{\kappa}_1$ generate $\hat{S}_{ui} = exp(x_{i2}\hat{\kappa}_1)$.

Step 3a:
Plug in the estimated residuals $\hat{u}_i$, obtained in step 1, and $\hat{S}_{ui}$, obtained in step 2, in the minimization problem 32. Thus, we obtain the feasible objective function:

$$\hat{\beta}_{1_o}, \hat{\rho}_o, \hat{\kappa}_2 := \overset{argmin}{\beta_{1_o},\rho_o,\kappa_2}\sum_i(\hat{\eta}_i - \beta_{1_o}\hat{u}_i - \rho_o exp(x_{i2}\kappa_2)\frac{\hat{u}_i}{\hat{S}_{ui}})^2. \tag{33}$$

Another possibility for estimation is replacing step 3a with the procedure called step 3b:

Step 3b:
1. Initialize the parameter $\beta_{1_o}^{(0)}$ and obtain first estimates of $\hat{\varepsilon}_i^{(0)} = \hat{\eta}_i - \hat{u}_i \beta_{1_o}^{(0)}$.
2. Use $\hat{\varepsilon}_i^{(0)^2}$ to estimate $S_{\varepsilon_i}^{(0)2}$.
3. Minimize 33 using $S_{\varepsilon_i}^{(0)}$ in place of $exp(x_{i2}\kappa_2)$.
4. Obtain an estimate $\beta_{1_o}^{(1)}$ and iterate until convergence.

8 Specification tests

The identification strategies presented in this paper rely on assumptions and restrictions on the structural equation and the reduced-form equation. Unfortunately, testing the assumptions on the structural equation is not possible in any of the models proposed. The reason is that we cannot identify the structural equation without the assumptions proposed when there are no instrumental variables. If there are instrumental variables, it is possible to test the assumptions on the structural equation. But in this case, it is a question of testing the validity of the available instruments.

More specifically, testing the assumption of linearity in the structural equation is not possible since its identification relies on this assumption. Because of this is the reason, Wooldridge (2010) does not advocate the use of functions of $\mathbf{x}_2$ for identification of the parameters of interest. But, the CF-OS method proposed in Section 5 is still consistent and useful for 1) models with endogenous explanatory variables that are nonlinear on available instrumental variables or 2) models without endogeneity and with regressors that present a nonlinear function with observables (Selection on observables). Similarly, testing the assumptions of the asymmetric error term in the structural equation or heteroskedastic error term in the structural equation is not possible when there are not available instrumental variables because its identification relies on one of these assumptions.

On the other hand, it is possible to test the different assumptions made for the reduced-form equation. We can find available tests for the hypothesis that the reduced-form equation is linear against a nonparametric specification (Horowitz (2006)). In addition, it is possible to test the assumption of heteroskedastic reduced-form error using the test proposed by Eubank and Thomas (1993) or Dette and Munk (1998). Eubank and Thomas (1993) presents a test for the heteroskedasticity of the error terms using a penalized spline estimator (There are no available heteroskedasticity tests using Neural Networks as estimators). Finally, a test for the hypothesis of symmetrically distributed error term against non-symmetrically distributed error term is proposed by Neumeyer and Dette (2007).

Combining strategies

Since it is challenging to test the different assumptions proposed, another option is combining the identification strategies and average estimates. If we believe that a semi-parametric structural model presents not only asymmetric error terms but also heteroskedastic error terms, it is possible to estimate the model using both strategies and then average the estimated parameters (The asymptotic distribution of this averaged estimator is left for further research). Further, it is not advisable to average the estimates obtained using a semiparametric structural model and a linear structural model. The reason is that both assumptions are incompatible when there are no available instrumental variables. In contrast, this could be done in the presence of instrumental variables. Another option could be combining the identification strategy of nonlinearity in the reduced form with heteroskedasticity and/or asymmetric error terms for a linear triangular system.

9 Test of valid instrumental variables

In the case of an available instrumental variable (we call it **i**), the methods proposed are still consistent. In this situation, we could use the additional moment conditions provided in this paper to test the validity of the available instrumental variables.

9.1 Triangular model with linear structural equation

In the case of the triangular model presented in section 2 (linear structural equation and a nonlinear reduced equation), the availability of an instrumental variable (**i**) provides the following moment conditions:

$$E[\varepsilon B] = 0, \tag{34}$$

with $B = [\mathbf{i} \quad \mathbf{x}_2]'$.

We could estimate the parameters of the structural equation 2 using the moment condition $E[\varepsilon B] = 0$ and call this estimator $\hat{\theta}_o^{(1)}$. Also, we can estimate the parameters of the structural equation 2 using only the moment conditions presented in Section 5 and call this estimator $\hat{\theta}_o^{(2)}$. Then, we can compare both estimators using a Hausman-type test procedure (Hausman (1983),Hahn and Ridder (2019)). The properties of this testing procedure are left for further research.

9.2 Triangular model with semiparametric equation

In the case of the triangular model presented in section 7 (semiparametric structural equation and a nonlinear reduced equation), the availability of an instrumental variable (**i**) provides the following moment conditions:

$$E[\omega \mathbf{u}] = 0. \tag{35}$$

We could estimate the parameter β_{1_o} using the moment condition $E[\omega \mathbf{u}] = 0$ and call it $\hat{\beta}_{1_o}^{(1)}$. We can also estimate β_{1_o} using the moments presented in Subsection 7.1 or using the method proposed in Subsection 7.2 and call it $\hat{\beta}_{1_o}^{(2)}$. Then, we can compare both estimators using a Hausman-type test procedure (Hausman (1983),Hahn and Ridder (2019)). The properties of this testing procedure are left for further research.

10 Monte Carlo simulation

10.1 The design

In order to test the different estimation methods proposed, we performed three Monte Carlo simulation experiments with different data-generating processes in each simulation experiment. In the following subsections, we present the design and results of the Monte Carlo simulations. More details are presented in the supplementary material.

10.1.1 A. Linear structural equation and non-linear secondary equation

In the first Monte Carlo simulation experiment, we generate data from 9 different data-generating processes.

DGP 1 - 4: The first 4 data generating processes are based on a triangular simultaneous model with a linear structural equation and a non-linear reduced-form equation (DGP 36). Both equations present additive disturbance terms. The structural equation presents one endogenous regressor x_1 and one exogenous covariate x_2. We generate 100 samples using this data-generating process (DGP) with different values of the parameters (Details in Table 1).

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \\ x_{i1} &= g(x_{i2}) + u_i, \end{aligned} \tag{36}$$

with:

$$\begin{pmatrix} x_{i2} \\ u_i \\ \varepsilon_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} var(x_2) & 0 & 0 \\ 0 & var(u) & cov_{u,\varepsilon} \\ 0 & cov_{u,\varepsilon} & var(\varepsilon) \end{pmatrix} \right). \tag{37}$$

The DGPs 1 to 3 are modifications of the DGPs used by Su and Ullah (2008) and Martins-Filho and Yao (2012). The main difference with respect to their DGPs is the linearity in the structural equation. We use different functional forms proposed by Su and Ullah (2008) and Martins-Filho and Yao (2012) for the unknown function for $g(x_2)$ in the reduced form or secondary equation (We cannot compare our results to theirs because they consider a semi-parametric triangular system with exclusion restrictions). Finally, we use a logarithmic function for the DGP 4.

DGP 5 - 6: The fifth and sixth data-generating processes are generated from a linear triangular model with limited explanatory variables. The DGP 5 uses a censored endogenous variable:

$$x_{i1} = (x_{i2} + u_i)\mathbb{1}(x_{i2} + u_i > 0).$$

The DGP 6 uses a binary endogenous variable (DGP 6):

$$x_{i1} = \mathbb{1}(x_{i2} + u_i > 0).$$

10.1.2 B. Semiparametric structural equation and non-linear secondary equation: nonsymmetric errors

In order to test the estimation method proposed in subsection 7.1, we generated data from a DGP similar to the one proposed by Lewbel et al. (2020) but modifying the structural equation using a semi-parametric model:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + f(x_{i2}) + \varepsilon_i, \\ x_{i1} &= g(x_{i2}) + u_i, \end{aligned} \tag{40}$$

with $u_i = u_{i1} + u_{i2}$, $u_{i1} \sim Gumbel(0, 1.64)$, $u_{i2} \sim logNormal(0, 1.72)$, $\varepsilon_i = u_{i2} + r_i$, $r_i \sim Gumbel(0, 1.64)$, $f(x_{i2}) = log(|x_{i2} - 1| + 1)sign(x_{i2} - 1)$, and $g(x_{i2}) = log(|x_{i2} - 1| + 1)sign(x_{i2} - 1)$.
This data-generating process is called DGP 7.

10.1.3 C. Semiparametric structural equation and non-linear secondary equation: heteroskedastic errors

In order to test the estimation methods proposed in subsection 7.2, we generated data from a DGP similar to the one proposed by Klein and Vella (2010) but modifying the triangular system as follows:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + f(x_{i2}) + 0.33exp(0.6x_{i2})v_i^* + z_i, \\ x_{i1} &= g(x_{i2}) + exp(0.2x_{i2})v_i^*, \end{aligned} \tag{41}$$

with $v_i^* \sim N(0,1)$, $z_i \sim N(0,1)$, $f(x_{i2}) = log(|x_{i2} - 1| + 1)sign(x_{i2} - 1)$, and $g(x_{i2}) = log(|x_{i2} - 1| + 1)sign(x_{i2} - 1)$.
This data-generating process is called DGP 8.

10.2 The results

In this section, we present the results of the simulation experiment. In figures 2, 3, and 4 we present the boxplots of the estimated parameter of interest β_1 of the DGPs 1, 7, and 8 when the true value of β_1 is equal to 1. The estimation methods used are OLS ignoring endogeneity (OLS), 2SLS using as IVs $\hat{g}(x_{i2})$ (2SLS), OLS p.: OLS using $\hat{g}(x_{i2})$ instead of x_{i1} (OLS Orth. Vble.), naive Control function approach (CF), naive Control function approach with sample splitting (CF SS), control function approach with orthogonal score (CF-OS), control function approach with orthogonal score with sample splitting (CF-OS SS), estimation with asymmetric errors (AE), and estimation with heteroskedastic errors (HE).
When the underlying data-generating process is a simultaneous triangular model with a linear main

Table 1: Monte Carlo Experiment: Identification through nonlinearity

Scenario	$y = f(x_1,x_2)$	$g(\theta x_2)$	β_1	$dim(x_2)$	β_2	$var(u)$	$var(x_2)$	$var(\varepsilon)$	$cov(u,eps)$
1.1	$y = x_1\beta_1 + x_2\beta_2$	$log(\lvert x_2-1\rvert+1)sign(x_2-1)$	1	1	1	1	1	1	0.9
1.2	$y = x_1\beta_1 + x_2\beta_2$	$log(\lvert x_2-1\rvert+1)sign(x_2-1)$	1	1	1	1	2	1	0.9
1.3	$y = x_1\beta_1 + x_2\beta_2$	$log(\lvert x_2-1\rvert+1)sign(x_2-1)$	1	1	1	2	1	1	0.9
1.4	$y = x_1\beta_1 + x_2\beta_2$	$log(\lvert x_2-1\rvert+1)sign(x_2-1)$	1	1	1	1	1	2	0.9
2.1	$y = x_1\beta_1 + x_2\beta_2$	$\frac{exp(x_2)}{1+3*exp(x_2)}$	1	1	1	1	1	1	0.9
2.2	$y = x_1\beta_1 + x_2\beta_2$	$\frac{exp(x_2)}{1+3*exp(x_2)}$	1	1	1	1	2	1	0.9
2.3	$y = x_1\beta_1 + x_2\beta_2$	$\frac{exp(x_2)}{1+3*exp(x_2)}$	1	1	1	2	1	1	0.9
2.4	$y = x_1\beta_1 + x_2\beta_2$	$\frac{exp(x_2)}{1+3*exp(x_2)}$	1	1	1	1	1	2	0.9
3.1	$y = x_1\beta_1 + x_2\beta_2$	$1+\frac{2*exp(x_2)}{1+exp(x_2)}$	1	1	1	1	1	1	0.9
3.2	$y = x_1\beta_1 + x_2\beta_2$	$1+\frac{2*exp(x_2)}{1+exp(x_2)}$	1	1	1	1	2	1	0.9
3.3	$y = x_1\beta_1 + x_2\beta_2$	$1+\frac{2*exp(x_2)}{1+exp(x_2)}$	1	1	1	2	1	1	0.9
3.4	$y = x_1\beta_1 + x_2\beta_2$	$1+\frac{2*exp(x_2)}{1+exp(x_2)}$	1	1	1	1	1	2	0.9
4.1	$y = x_1\beta_1 + x_2\beta_2$	$log(0.1+x_2^2)$	1	1	1	1	1	1	0.9
4.2	$y = x_1\beta_1 + x_2\beta_2$	$log(0.1+x_2^2)$	1	1	1	1	2	1	0.9
4.3	$y = x_1\beta_1 + x_2\beta_2$	$log(0.1+x_2^2)$	1	1	1	2	1	1	0.9
4.4	$y = x_1\beta_1 + x_2\beta_2$	$log(0.1+x_2^2)$	1	1	1	1	1	2	0.9
5.1	$y = x_1\beta_1 + x_2\beta_2$	censored endogenous regressor	1	1	1	1	1	1	0.9
5.2	$y = x_1\beta_1 + x_2\beta_2$	censored endogenous regressor	1	1	1	1	2	1	0.9
5.3	$y = x_1\beta_1 + x_2\beta_2$	censored endogenous regressor	1	1	1	2	1	1	0.9
5.4	$y = x_1\beta_1 + x_2\beta_2$	censored endogenous regressor	1	1	1	1	1	2	0.9
6.1	$y = x_1\beta_1 + x_2\beta_2$	binary regressor	1	1	1	1	1	1	0.9
6.2	$y = x_1\beta_1 + x_2\beta_2$	binary regressor	1	1	1	1	2	1	0.9
6.3	$y = x_1\beta_1 + x_2\beta_2$	binary regressor	1	1	1	2	1	1	0.9
6.4	$y = x_1\beta_1 + x_2\beta_2$	binary regressor	1	1	1	1	1	2	0.9

equation and nonlinear reduced-form equation (DGP 1 to DGP 6), the best method is the proposed CF-OS estimator. When the data generating process is a simultaneous semiparametric triangular model with asymmetric error terms, the best estimator is the two-step semiparametric estimator (AE) that exploits the moment conditions proposed by Lewbel et al. (2020) (DGP 7, Figure 3). When the data-generating process is a simultaneous semiparametric triangular model with heteroskedastic error terms, the best estimator is the two-step semiparametric estimator (HE) that requires the optimization of a nonlinear objective function (DGP 8, Figure 4). In tables 4 to 10 of Appendix B, we present the detailed results of bias and root mean squared error of the different estimation methods for the parameter of interest β_1 for all DGP.

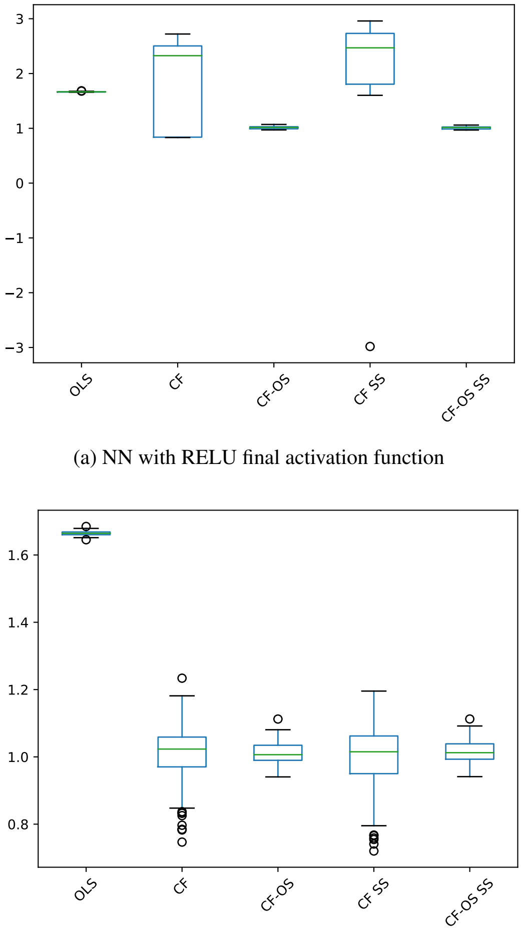
For the estimation of the nuisance parameters, we used deep neural networks with 2 different architectures for the prediction of the endogenous regressor. The neural network architectures are presented in Table 2. In both architectures, we use a dropout rate [3] equal to 0. Additionally, the optimization algorithm used is RMSprop with an adaptive learning rate that starts at 0.01 using a minibatch equal to 128 and the number of epochs is equal to 10 (Hinton et al. (2012)). In both architectures, we used two hidden layers with rectified linear units (RELU). The activation function RELU is a stepwise function given by $max(0,z)$ that was first proposed by Nair and Hinton (2010). Finally, we use linear activation function for the output layer of the second architecture and RELU for the first architecture.

It is important to notice that using RELU in the output layer misspecifies the endogenous variable with support in an interval of values that includes 0. In spite of this misspecification, we observe that the estimation results using this neural network architecture still leads to good estimates using the orthogonal scores. On the contrary, using a RELU activation function in the output layer performs better for censored or truncated endogenous variables (DGP). When the DGP includes a binary endogenous variable, the activation function of the output layer is the sigmoid function. Importantly, estimation of the reduced form of a binary or truncated variable using neural networks with only x_2 as a regressor means that we ignore the non-separability of the function. In spite of this, CF-OS is the best estimator for triangular models with truncated or binary regressors.

[3] Dropout is a regularization technique used in neural network training with the purpose of avoiding overfitting. According to Srivastava et al. (2014), a neural network that overfits the data predicts perfectly the training sample but not the evaluation set. This means the training error is low while the validation error is high.

Table 2: Neural Network Architectures

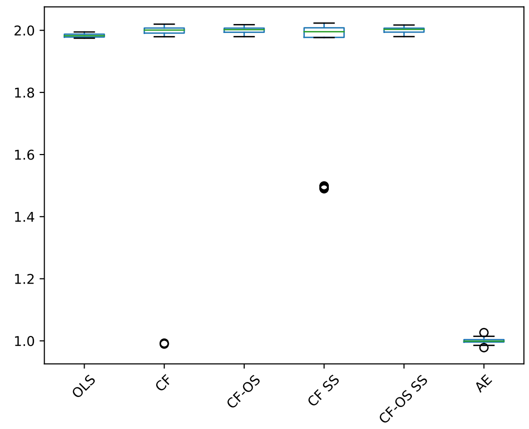
NNa	Hidden Layers	Nodes	Activation function for hidden layers	Activation function for output layer	Dropout Rate	Optimization Algorithm	Learning Rate
1	2	10	max(0,z)	max(0,z)	0	RMsprop	0.01
2	2	10	max(0,z)	$z'w$	0	RMsprop	0.01



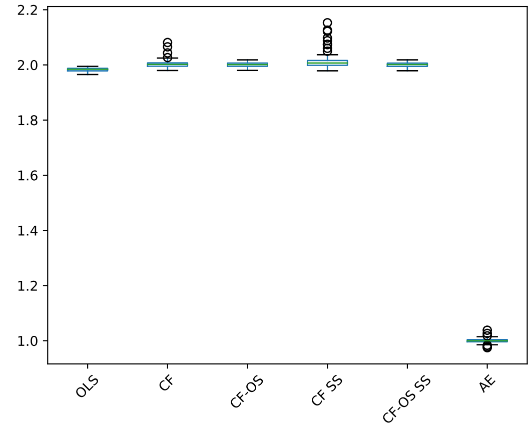
(a) NN with RELU final activation function

(b) NN with linear final activation function

Figure 2: Linear structural equation and non-linear reduced-form equation: DGP 1-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).

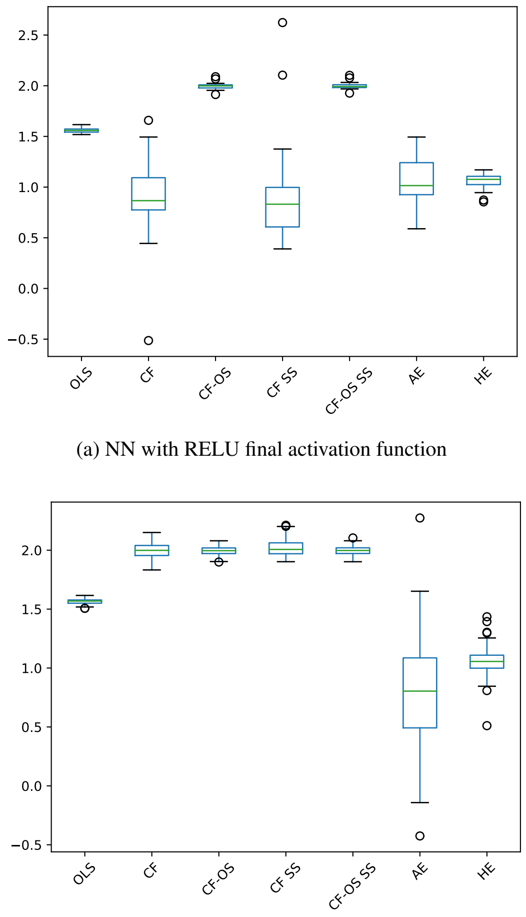


(a) NN with RELU final activation function



(b) NN with linear final activation function

Figure 3: Linear structural equation and non-linear reduced-form equation: DGP 7-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).



(a) NN with RELU final activation function

(b) NN with linear final activation function

Figure 4: Linear structural equation and non-linear reduced-form equation: DGP 8-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).

11 Conclusions

This paper proposes alternative identification strategies when no instrumental variables are available for triangular semiparametric models. We begin our study with a baseline triangular model composed of a linear structural equation and a reduced-form equation of the endogenous regressor that is nonlinear. In this setting, we identify the parameters of interest of the main equation by exploiting a functional assumption in combination with a control function approach. Later, we relax the linearity assumption in the main equation and we assume that it presents a partial linear form. For the identification of the parameters of interest in this more general model, we use two different approaches. The first one is an extension of the methodology proposed by Lewbel et al. (2020) for linear triangular models that requires that the distributions of the error terms are asymmetric. The second one extends the strategy proposed by Klein and Vella (2010) for linear triangular models and entails that the error terms are heteroskedastic. For estimation, we propose two-step semiparametric estimation methods using neural networks in the first stage to estimate the nuisance parameters. But other machine learning methods can be used for the estimation of the nuisance parameters in combination with cross-fitting as suggested by Chernozhukov et al. (2018). The simulation results show that the proposed methods have lower RMSE and bias than OLS. The estimator CF-OS outperforms all other methods when the data-generating process is a triangular model with a linear structural equation and a non-linear reduced equation. When the model is semiparametric with asymmetric error terms, the two-step semiparametric estimator (AE) using the moment conditions proposed by Lewbel et al. (2020) is the most appropriate (Section 7.1). Finally, the simulation results for the two-step estimator (HE) that exploits heteroskedasticity of the error terms show that the estimator produces better estimates than OLS (Section 7.2).

12 Annex

12.1 Proofs

12.1.1 Proof of Theorem 1

The estimator $\hat{\theta}_o$ is given by:

$$\hat{\theta}_o = (\sum_{i=1}^{N} \hat{Z}_i \hat{Z}_i')^{-1} (\sum_{i=1}^{N} \hat{Z}_i y_i). \tag{42}$$

We know that y_i is equal to $Z_i \theta_o + \omega_i$. Adding and subtracting $\hat{Z}_i' \theta_o$ gives $\hat{Z}_i' \theta_o + (Z_i - \hat{Z}_i)' \theta_o + \omega_i$. Thus, $y_i = \hat{Z}_i' \theta_o + \hat{\omega}_i$ with $\hat{\omega}_i = (Z_i - \hat{Z}_i)' \theta_o + \omega_i$. Replacing the last expression in 42 and re-arranging gives:

$$\hat{\theta}_o - \theta_o = (\sum_{i=1}^{N} \hat{Z}_i \hat{Z}_i')^{-1} (\sum_{i=1}^{N} \hat{Z}_i (Z_i - \hat{Z}_i)' \theta_o + \sum_{i=1}^{N} \hat{Z}_i \omega_i). \tag{43}$$

First, we expand the term within the inverse in expression 43:

$$\sum_{i=1}^{N} \hat{Z}_i \hat{Z}_i' = \sum_{i=1}^{N} Z_i Z_i' + \sum_{i=1}^{N} Z_i (\hat{Z}_i - Z_i)' + \sum_{i=1}^{N} (\hat{Z}_i - Z_i)' Z_i + \sum_{i=1}^{N} (\hat{Z}_i - Z_i)(\hat{Z}_i - Z_i)'.$$

And calling $\hat{Z}_i - Z_i$ as $\tilde{\xi}_i$ gives:

$$\sum_{i=1}^{N} \hat{Z}_i \hat{Z}_i' = \sum_{i=1}^{N} Z_i Z_i' + \sum_{i=1}^{N} Z_i \tilde{\xi}_i' + \sum_{i=1}^{N} \tilde{\xi}_i' Z_i + \sum_{i=1}^{N} \tilde{\xi}_i \tilde{\xi}_i'. \tag{44}$$

In expression 44, we can name the four terms in the RHS as 44.i), 44.ii), 44.iii), 44.iv). Now, we can analyze each one of these terms as follows:
44.i) $\sum_{i=1}^{N} Z_i Z_i' = O_p(N)$ because $\{Z_i Z_i'\}_{i=1}^{N}$ is an i.i.d. sequence by Assumption 1. The elements of $Z_i Z_i'$ have finite expected absolute values by condition ii.a of Theorem 1 and by Theorem 3.1 of White (1984) $\sum_{i=1}^{N} N^{-1} Z_i Z_i' \overset{a.s.}{\rightarrow} Q_{ZZ}$. By Theorem 1 Ferguson (1996), $\sum_{i=1}^{N} N^{-1} Z_i Z_i' \overset{a.s.}{\rightarrow} Q_{ZZ} \Rightarrow \sum_i N^{-1} Z_i Z_i' \overset{p}{\rightarrow} Q$
44.ii) $\sum_{i=1}^{N} Z_i \tilde{\xi}_i' = O_P(N^{1-\tilde{\delta}/2})$ because $\sum_{i=1}^{N} Z_i \tilde{\xi}_i' = \sum_{i=1}^{N} \begin{bmatrix} 0 & 0 & x_{i1}\xi_i \\ 0 & 0 & x_{i2}\xi_i \\ 0 & 0 & u_i\xi_i \end{bmatrix}$ with $\xi_i = \widehat{g_o(x_{i2})} - g_o(x_{i2})$. By triangle inequality, $E|\sum_{i=1}^{N} x_{i1}\xi| \leq \sum_{i=1}^{N} E|x_{i1}\xi|$ and using Holder's inequality:

$$E|\sum_{i=1}^{N} x_{i1}\xi| \leq \sum_{i=1}^{N} (E|x_{i1}|^2)^{1/2} (E|\xi_i|^2)^{1/2} = O_p(N^{1-\tilde{\delta}/2}).$$

The last equality holds because $E|\xi_i|^2 = O_p(N^{-\tilde{\delta}})$ by Lemma 1, $E|x_{i1}|^2 = O_p(1)$ by Assumption 1 and because $\{x_{i1}, x_{i2}, \varepsilon_i\}_{i=1}^{N}$ is an i.i.d. sequence we have that $\sum_{i=1}^{N} x_{i1}\xi_i = O_p(N^{1-\tilde{\delta}/2})$. A similar reasoning applies to the other elements.
44.iii) $\sum_{i=1}^{N} \tilde{\xi}_i' Z_i = O_p(N^{1-\tilde{\delta}/2})$ using a similar reasoning as in point 44. ii).
44.iv)$\sum_{i=1}^{N} \tilde{\xi}_i \tilde{\xi}_i' = O_p(N^{1-\tilde{\delta}})$ because $\sum_{i=1}^{N} \tilde{\xi}_i \tilde{\xi}_i' = \sum_{i=1}^{N} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \xi_i^2 \end{bmatrix}$. Then, $\sum_{i=1}^{N} \xi_i^2 = O_p(N^{1-\tilde{\delta}})$ by Lemma 1 and because ξ_i are i.i.d.

Secondly, we analyze each element of the second term of the expression 43. Then,

$$\sum_{i=1}^{N}\hat{Z}_i(Z_i-\hat{Z}_i)'\theta_o=\sum_{i=1}^{N}\hat{Z}_i\tilde{\xi}_i'\theta_o,=\sum_{i=1}^{N}Z_i\tilde{\xi}_i'\theta_o+\sum_{i=1}^{N}\tilde{\xi}_i\tilde{\xi}_i'\theta_o, \tag{45}$$

45.i) $\sum_{i=1}^{N}Z_i\tilde{\xi}_i'\theta_o=O_p(N^{1-\tilde{\delta}/2})$ using a similar reasoning as in point 44.ii).
45.ii)$\sum_{i=1}^{N}\tilde{\xi}_i\tilde{\xi}_i'\theta_o=O_p(N^{1-\tilde{\delta}})$ using a similar reasoning as in point 44.iv).

Also,

$$\sum_{i=1}^{N}\hat{Z}_i\omega_i=\sum_{i=1}^{N}Z_i\omega_i+\sum_{i=1}^{N}\tilde{\xi}_i\omega_i, \tag{46}$$

46.i) $\sum_{i=1}^{N}Z_i\omega_i=O_p(N^{1/2})$ because each term of $\sum_{i=1}^{N}Z_i\omega_i$ is $O_p(N^{1/2})$ by Markov's inequality.
46.ii) $\sum_{i=1}^{N}\xi_i\omega_i=O_p(N^{1-\tilde{\delta}/2})$ because $E[\sum_{i=1}^{N}\xi_i\omega_i]\leq\sum_{i=1}^{N}E[\xi_i\omega_i]$ by triangle inequality. $E[\sum_{i=1}^{N}\xi_i\omega_i]\leq NE[\xi_i\omega_i]$ because $\{x_{i1},x_{i2},\varepsilon_i\}_{i=1}^{N}$ is an i.i.d. sequence, $E[\sum_{i=1}^{N}\xi_i\omega_i]\leq NE[|\xi_i^2|)^{1/2}E[|\omega_i|^2]^{1/2}$ by Hölder's inequality with $E[|\xi_i^2|]=O_p(N^{-\tilde{\delta}})$ by Lemma 1.

Then, $(\sum_{i=1}^{N}\hat{Z}_i\hat{Z}_i')^{-1}$ of expression 43 is $O_p(N^{-1})$, and the term $(\sum_{i=1}^{N}\hat{Z}_i(Z_i-\hat{Z}_i)'\theta_o+\sum_{i=1}^{N}\hat{Z}_i\omega_i$ of expression 43 is $O_p(N^{1-\tilde{\delta}/2})$. Thus, we conclude that $\hat{\theta}_o$ converges to the true value with order $N^{-\tilde{\delta}/2}$. This order is lower than the parametric one of $-1/2$ because $\tilde{\delta}/2<1/2$ even if $d=1$ and $\delta=1$ since from Lemma 1, $\tilde{\delta}=\frac{\delta}{\delta+d}$.

12.1.2 Proof of Proposition 1

Using the stabilizing rate $N^{1/2}$, we have that:

$$\sqrt{N}(\hat{\theta}_o-\theta_o)=(\underbrace{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\hat{Z}_i\hat{Z}_i'}_{(A)})^{-1}(\underbrace{\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{N}\hat{Z}_i(Z_i-\hat{Z}_i)'\theta_o}_{(B)}+\underbrace{\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^{N}\hat{Z}_i\omega_i}_{(C)}).$$

Let us analyze term by term:

(A) $\frac{1}{N}\sum_{i=1}^{N}\hat{Z}_i\hat{Z}_i'=\frac{1}{N}\sum_{i=1}^{N}(Z_i+(\hat{Z}_i'-Z_i))(Z_i+(\hat{Z}_i'-Z_i))'$.

As in the proof of Theorem 1, we call $\tilde{\xi}_i=\hat{Z}_i'-Z_i$. Using this change of notation and expanding term (A) we obtain:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{Z}_i\hat{Z}_i'=\frac{1}{N}\sum_{i=1}^{N}(Z_iZ_i'+2\tilde{\xi}_iZ_i'+\tilde{\xi}_i\tilde{\xi}_i'),$$

$$\frac{1}{N}\sum_{i=1}^{N}\hat{Z}_i\hat{Z}_i'=Q_{ZZ}+o_p(1)<\infty.$$

where Q_{ZZ} is a positive definite non-stochastic and bounded matrix with full rank (This result is demonstrated in the proof of Theorem 1).

B) $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{Z}_i(Z_i-\hat{Z}_i)'\theta_o=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(Z_i+\tilde{\xi}_i)\tilde{\xi}_i'\theta_o,$

C) $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{Z}_i\omega_i=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}Z_i\omega_i+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\xi}_i\omega_i.$

According to Chernozhukov et al. (2018), one can bound terms $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\xi}_i\omega_i$ using empirical process methods because the complexity of the parameter space of $g_o(x_2)$ is controlled. This is shown by Farrell et al. (2021) who finds upper bounds on the complexity of the parameter space using localization analysis and shows that the integral of its entropy is upper bounded. In addition, the term $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\xi}_i\tilde{\xi}_i'\theta_o$

converges to 0 because $\tilde{\delta} > 1/2$ even if $\tilde{\delta}/2 < 1/2$ (Chernozhukov et al. (2018), pg. 5). This means that $1/4 < \tilde{\delta}/2 < 1/2$.
$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(Z_i\omega_i)\xrightarrow{L}N(0,\sigma^2_{\omega}Q_{ZZ})$ by Lindeberg-Levy CLT. Thus, $\sqrt{N}(\hat{\theta}_o-\theta_o)\to N(0,\sigma^2_{\omega}Q_{ZZ}^{-1})$.
Finally, $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}Z_i\tilde{\xi}_i'\theta_o$ diverges because ξ_i is multiplied by x_{i1}, and x_{2i1} which are not centered at 0. As a result of the presence of this term, we conclude that:

$$|\sqrt{N}(\hat{\theta}_o-\theta_o)|\xrightarrow{p}\infty.$$

12.1.3 Proof of Theorem 2

The estimator of $\hat{\rho}_o$ is equal to:

$$\hat{\rho}_o = (\hat{u}^{*\prime}M_{x_1^*}\hat{u}^*)^{-1}(\hat{u}^{*\prime}M_{x_1^*}y^*). \tag{47}$$

We know that $y^* = x_1^*\beta_{1o}+\rho_o u^*+\omega^*$. Adding and substracting $\rho_o\hat{u}^*-\rho_o\hat{u}^*$ gives:

$$y^* = x_1^*\beta_{1o}+\rho_o\hat{u}^*+\tilde{\omega}^*, \tag{48}$$

with $\tilde{\omega}^* = \omega^*-\rho_o\hat{u}^*$.
Replacing 48 into 47 and re-arrenging gives:

$$\hat{\rho}_o = \rho_o+(\hat{u}^{*\prime}M_{x_1^*}\hat{u}^*)^{-1}(\hat{u}^{*\prime}M_{x_1^*}(\rho_o(u^*-\hat{u}^*)+\omega^*)). \tag{49}$$

Knowing that $\hat{u} = u+(g_o(x_2)-\hat{g}_o(x_2)) = u+\xi$, we can re-express $\hat{u}^*$ as $u^*+\xi^*$ because $\hat{u}^* = M_{x_2}\hat{u}$. Then, replacing $\hat{u}^* = u^*+\xi^*$ in 49 and re-arrenging gives:

$$\hat{\rho}_o = \rho_o+\underset{(1)}{(\hat{u}^{*\prime}M_{x_1^*}\hat{u}^*)^{-1}}(\underset{(2)}{-\rho_o u^{*\prime}M_{x_1^*}\xi^*}\underset{(3)}{-\rho_o\xi^{*\prime}M_{x_1^*}\xi^*}+\underset{(4)}{u^{*\prime}M_{x_1^*}\omega^*}+\underset{(5)}{\xi^{*\prime}M_{x_1^*}\omega^*}). \tag{50}$$

Now, we can analyze each term of 50:

1. The denominator of 50 is $O_p(N)$. To see why, we expand term (1) and obtain:

$$\hat{u}^{*\prime}M_{x_1^*}\hat{u}^* = \underset{(1.1)}{u^{*\prime}M_{x_1^*}u^*}+\underset{(1.2)}{u^{*\prime}M_{x_1^*}\xi^*}+\underset{(1.3)}{\xi^{*\prime}M_{x_1^*}u^*}+\underset{(1.4)}{\xi^{*\prime}M_{x_1^*}\xi^*}. \tag{51}$$

Now, analyzing element by element of expression 51:

i. Term (1.1) $u^{*\prime}M_{x_1^*}u^* = O_p(N)$ because:

$$u^{*\prime}M_{x_1^*}u^* = \underset{(1.1.1)}{u'M_{x_1^*}u}-\underset{(1.1.2)}{u'M_{x_1^*}P_{x_2}u}-\underset{(1.1.3)}{u'P_{x_2}M_{x_1^*}u}+\underset{(1.1.4)}{u'P_{x_2}M_{x_1^*}P_{x_2}u}.$$

A. Term (1.1.1) $u'M_{x_1^*}u = u'u-u'P_{x_1^*}u = O_p(N)$ since:

- $u'u = O_p(N)$ because $E[u_i^2] < \infty$ by Assumption 1 and by Markov's inequality.
- $u'P_{x_1^*}u \leq ||P_{x_1^*}u||||P_{x_1^*}u|| \leq ||u||||u||$ where the first inequality holds by Cauchy-Schwarz inequality and the second one by Pythagoras' theorem (Chudik and Pesaran (2015)). Finally, $||u||||u|| = O_p(N^{1/2})O_p(N^{1/2})$ because $||u||^2 = O_p(N)$ by Markov's inequality.

B. Term (1.1.2) $u'M_{x_1^*}P_{x_2}u = O_p(N)$ since:

$$u'M_{x_1^*}P_{x_2}u = u'P_{x_2}u-u'P_{x_1^*P_{x_2}}u = O_p(N),$$

by similar reasoning as in point 1.i.A).

C. Term (1.1.3) $u'P_{x_2}M_{x_1^*}u = O_p(N)$ since:

$$u'P_{x_2}M_{x_1^*}u = u'P_{x_2}u-u'P_{x_2}P_{x_1^*}u = O_p(N),$$

by similar reasoning as in point 1.i.A).

D. Term (1.1.4) $u'P_{x_2}M_{x_1^*}P_{x_2}u = O_p(N)$ since:

$$u'P_{x_2}M_{x_1^*}P_{x_2}u = u'P_{x_2}P_{x_2}u - u'P_{x_2}P_{x_1^*}P_{x_2}u = O_p(N),$$

by similar reasoning as in point 1.i.A).

ii. Term (1.2) $\xi^{*\prime}M_{x_1^*}u^* = O_p(N^{1-\tilde{\delta}/2})$, because if we expand it and analyze each part:

$$\xi^{*\prime}M_{x_1^*}u^* = \underset{(1.2.1)}{u'M_{x_1^*}\xi} - \underset{(1.2.2)}{u'M_{x_1^*}P_{x_2}\xi} - \underset{(1.2.3)}{u'P_{x_2}M_{x_1^*}M_{x_2}\xi} + \underset{(1.2.4)}{u'P_{x_2}M_{x_1^*}P_{x_2}\xi}$$

A. Term (1.2.1) $u'M_{x_1^*}\xi = u'\xi - u'P_{x_1^*}\xi = O_p(N^{1-\tilde{\delta}/2})$ because:

- $u'\xi = O_p(N^{1-\delta/2})$ since by triangle inequality, $E|\sum_{i=1}^N u_i\xi_i| \leq \sum_{i=1}^N E|u_i\xi|$ and using Holder's inequality:

$$E|\sum_{i=1}^N u_i\xi_i| \leq \sum_{i=1}^N (E|u_i|^2)^{1/2}(E|\xi_i|^2)^{1/2} = O_p(N^{1-\tilde{\delta}/2}).$$

  The last equality holds because $E|\xi_i|^2 = O_p(N^{-\tilde{\delta}})$ by Lemma 1, $E|u_i|^2 = O_p(1)$ by Assumption 1 and because $\{x_{i1}, x_{i2}, \varepsilon_i\}_{i=1}^N$ is an i.i.d. sequence we have that $\sum_{i=1}^N u_i\xi_i = O_p(N^{1-\tilde{\delta}/2})$.

- $u'P_{x_1^*}\xi = O_p(N^{1-\tilde{\delta}/2})$ because $u'P_{x_1^*}\xi \leq ||(P_{x_1^*}u)||||(P_{x_1^*}\xi)||$ by Cauchy-Schwarz inequality and $||(P_{x_1^*}u)||||(P_{x_1^*}\xi)|| \leq ||u||||\xi||$ by Pythagoras' theorem (Chudik and Pesaran (2015)).
  Finally, $||u||||\xi|| = O_p(N^{1/2})O_p(N^{1/2-\tilde{\delta}/2}) = O_p(N^{1-\tilde{\delta}/2})$ because $E[\sum u_i^2] = \sum_{i=1}^N E[u_i^2] = O_p(N)$ since $\{x_{i1}, x_{i2}, \varepsilon_i\}_{i=1}^N$ is an i.i.d. sequence and $E[u_i^2] = O_p(1)$ by Assumption 1. Similarly, we have that $||\xi|| = \sqrt{\sum_{i=1}^N \xi_i^2} = O_p(N^{1/2-\tilde{\delta}/2})$.

B. Term (1.2.2) $u'M_{x_1^*}P_{x_2}\xi = u'P_{x_2}\xi - u'P_{x_1^*}P_{x_2}\xi = O_p(N^{1-\tilde{\delta}/2})$ by similar reasoning as in point 1.a.ii.A.

C. Term (1.2.3) $u'P_{x_2}M_{x_1^*}M_{x_2}\xi = u'P_{x_2}M_{x_1^*}\xi - u'P_{x_2}M_{x_1^*}P_{x_2}\xi = O_p(N^{1-\tilde{\delta}/2})$ by similar reasoning as in point 1.a.ii.A.

D. Term (1.2.4) $u'P_{x_2}M_{x_1^*}P_{x_2}\xi = u'P_{x_2}P_{x_2}\xi - u'P_{x_2}P_{x_1^*}P_{x_2}\xi = O_p(N^{1-\tilde{\delta}/2})$ by similar reasoning as in point 1.a.ii.A.

iii. Term (1.3) $u^{*\prime}M_{x_1^*}\xi^* = O_p(N^{1-\tilde{\delta}/2})$ because it is a scalar which means that $(u^{*\prime}M_{x_1^*}\xi^*)' = \xi^{*\prime}M_{x_1^*}u^*$.

iv. Term (1.4) $\xi^{*\prime}M_{x_1^*}\xi^* = O_p(N^{1-\tilde{\delta}})$ because if we expand it and analyze each part:

$$\xi^{*\prime}M_{x_1^*}\xi^* = \underset{(1.4.1)}{\xi'M_{x_1^*}\xi} - \underset{(1.4.2)}{\xi'M_{x_1^*}P_{x_2}\xi} - \underset{(1.4.3)}{\xi'P_{x_2}M_{x_1^*}\xi} + \underset{(1.4.4)}{\xi'P_{x_2}M_{x_1^*}P_{x_2}\xi}$$

A. Term (1.4.1) $\xi'M_{x_1^*}\xi = \xi'\xi - \xi'P_{x_1^*}\xi = O_p(N^{1-\tilde{\delta}})$ because:

- $\xi'\xi = O_p(N^{1-\tilde{\delta}})$ because $E[\xi'\xi] = \sum_{i=1}^N E[\xi_i^2] = NE[\xi_i^2]$ because ξ_i is i.i.d and $E[\xi_i^2] = O_p(N^{\tilde{\delta}})$ by Lemma 1.
- $\xi'P_{x_1^*}\xi = O_p(N^{1-\tilde{\delta}})$ because $\xi'P_{x_1^*}\xi \leq ||\xi||||\xi||$ by Cauchy-Schwarz inequality and Pythagoras' theorem. Then, $||\xi|| = O_p(N^{1/2-\tilde{\delta}/2})$.

B. (1.4.2)$\xi'M_{x_1^*}P_{x_2}\xi = \xi'P_{x_2}\xi - \xi'P_{x_1^*}P_{x_2}\xi = O_p(N^{1-\tilde{\delta}})$ similar as point 1.a.iv.A.

C. (1.4.3)$\xi'P_{x_2}M_{x_1^*}\xi = \xi'P_{x_2}\xi - \xi'P_{x_2}P_{x_1^*}\xi = O_p(N^{1-\tilde{\delta}})$ similar as point 1.a.iv.A.

D. (1.4.4)$\xi'P_{x_2}M_{x_1^*}P_{x_2}\xi = \xi'P_{x_2}P_{x_2}\xi - \xi'P_{x_2}P_{x_1^*}P_{x_2}\xi = O_p(N^{1-\tilde{\delta}})$ similar as point 1.a.iv.A.

2. The numerator of 50 is $O_p(N^{1-\tilde{\delta}/2})$. To see why, we expand each of its terms as follows:

(a) Term (2) of 50 is $O_p(N^{1-\tilde{\delta}/2})$. We expand term (2) and we obtain:

$$-\rho_o u^{*\prime} M_{x_1^*} \xi^* = \underset{(2.1)}{-\rho_o u' M_{x_1^*} \xi} + \underset{(2.2)}{\rho_o u' M_{x_1^*} P_{x_2} \xi} + \underset{(2.3)}{\rho_o (P_{x_2} u)' M_{x_1^*} \xi} - \underset{(2.4)}{\rho_o (P_{x_2} u)' M_{x_1^*} P_{x_2} \xi}. \tag{52}$$

If we evaluate each term of 52, we have that:

i. Term (2.1) $\rho_o u' M_{x_1^*} \xi = \rho_o u' \xi + \rho_o u' P_{x_1^*} \xi = O_p(N^{1-\tilde{\delta}/2})$ because:

A. $\rho_o u' \xi = O_p(N^{1-\tilde{\delta}/2})$ because by triangle inequality, $E|\sum_{i=1}^N u_i \xi_i| \leq \sum_{i=1}^N E|u_i \xi|$ and using Holder's inequality:

$$E|\sum_{i=1}^N u_i \xi_i| \leq \sum_{i=1}^N (E|u_i|^2)^{1/2} (E|\xi_i|^2)^{1/2} = O_p(N^{1-\tilde{\delta}/2}).$$

The last equality holds because $E|\xi_i|^2 = O_p(N^{-\tilde{\delta}})$ by Lemma 1, $E|u_i|^2 = O_p(1)$ by Assumption 1 and because $\{x_{i1}, x_{i2}, \varepsilon_i\}_{i=1}^N$ is an i.i.d. sequence we have that $\sum_{i=1}^N u_i \xi_i = O_p(N^{1-\tilde{\delta}/2})$.

B. $\rho_o u' P_{x_1^*} \xi = O_p(N^{1-\tilde{\delta}/2})$ because $\rho_o u' P_{x_1^*} \xi \leq ||(P_{x_1^*} u)|| ||(P_{x_1^*} \xi)||$ by Cauchy-Schwarz inequality and $||(P_{x_1^*} u)|| ||(P_{x_1^*} \xi)|| \leq ||u|| ||\xi||$ by Pythagoras' theorem (Chudik and Pesaran (2015)). Finally, $||u|| ||\xi|| = O_p(N^{1/2}) O_p(N^{1/2-\tilde{\delta}/2}) = O_p(N^{1-\tilde{\delta}/2})$ because $E[\sum u_i^2] \leq \sum_{i=1}^N E[u_i^2] = O_p(N)$ where the inequality holds because $\{x_{i1}, x_{i2}, \varepsilon_i\}_{i=1}^N$ is an i.i.d. sequence and $E[u_i^2] = O_p(1)$ by Assumption 1. Similarly, we have that $||\xi|| = \sqrt{\sum_{i=1}^N \xi_i^2} = O_p(N^{1/2-\tilde{\delta}/2})$ by Lemma 1 and because ξ_i is i.i.d.

ii. Term (2.2) $\rho_o u' M_{x_1^*} P_{x_2} \xi = \rho_o u' P_{x_2} \xi - \rho_o u' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}/2})$ because:

A. $\rho_o u' P_{x_2} \xi = O_p(N^{1-\tilde{\delta}/2})$ because $u' P_{x_2} \xi \leq ||P_{x_2} u|| ||P_{x_2} \xi||$ by Cauchy-Schwarz inequality. And $||P_{x_2} u|| ||P_{x_2} \xi|| \leq ||u|| ||\xi||$ by Pythagoras' theorem. Then, $||u|| ||\xi|| = O_p(N^{1-\tilde{\delta}/2})$ as shown in 2.a.i.A).

B. $\rho_o u' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}/2})$ by following a similar reasoning as in point 2.a.ii.A).

iii. Term (2.3) $\rho_o (P_{x_2} u)' M_{x_1^*} \xi = \rho_o (P_{x_2} u)' \xi + \rho_o (P_{x_2} u)' P_{x_1^*} \xi = O_p(N^{1-\tilde{\delta}/2})$, because:

A. $\rho_o u' P_{x_2}' \xi = O_p(N^{1-\tilde{\delta}/2})$ by following a similar reasoning as in point 2.a.ii.A).

B. $\rho_o (P_{x_2} u)' P_{x_1^*} \xi = O_p(N^{1-\tilde{\delta}/2})$ by following a similar reasoning as in point 2.a.ii.A).

iv. Term (2.4) $-\rho_o u' P_{x_2}' M_{x_1^*} P_{x_2} \xi = \rho_o u' P_{x_2} \xi - \rho_o u' P_{x_2}' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}/2})$, because:

A. $\rho_o u' P_{x_2} \xi = O_p(N^{1-\tilde{\delta}/2})$ by following a similar reasoning as in point 2.a.ii.A).

B. $\rho_o u' P_{x_2} P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}/2})$ by following a similar reasoning as in point 2.a.ii.A).

(b) Term (3) of 50 is $O_p(N^{1-\tilde{\delta}})$. To see why, we expand term (3) and we obtain:

$$-\rho_o \xi^{*\prime} M_{x_1^*} \xi^* = \underset{(3.1)}{-\rho_o \xi' M_{x_1^*} \xi} + \underset{(3.2)}{\rho_o \xi' M_{x_1^*} P_{x_2} \xi} + \underset{(3.3)}{\rho_o (P_{x_2} \xi)' M_{x_1^*} \xi} - \underset{(3.4)}{\rho_o (P_{x_2} \xi)' M_{x_1^*} P_{x_2} \xi}, \tag{53}$$

If we evaluate each term of 52, we have that:

i. Term (3.1) $-\rho_o \xi' M_{x_1^*} \xi = -\rho_o \xi' \xi + \rho_o \xi' P_{x_1^*} \xi = O_p(N^{1-\tilde{\delta}})$ because:

A. $\rho_o \xi' \xi = O_p(N^{1-\tilde{\delta}})$ since $E[\xi' \xi] = \sum_{i=1}^N E[\xi_i^2] = NE[\xi_i^2]$ because ξ_i is i.i.d. and $E[\xi_i^2] = O_p(N^{\tilde{\delta}})$ by Lemma 1.

ii. Term (3.2) $\rho_o \xi' M_{x_1^*} P_{x_2} \xi = \rho_o \xi' P_{x_2} \xi - \rho_o \xi' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}})$ because:

A. $\rho_o \xi' P_{x_2} \xi = O_p(N^{1-\tilde{\delta}})$ since $\xi' P_{x_2} \xi \leq ||\xi|| ||\xi||$ by Cauchy-Schwarz inequality and Triangle Inequality. Then, $||\xi|| = O_p(N^{1/2-\tilde{\delta}/2})$ as shown in point 2.a.i.B.

B. $\rho_o \xi' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}})$ because $\xi' P_{x_1^*} P_{x_2} \xi \leq ||\xi|| ||\xi||$ by Cauchy-Schwarz inequality and Triangle Inequality. Then, $||\xi|| = O_p(N^{1/2-\tilde{\delta}/2})$ as shown in point 2.a.i.B.

iii. Term (3.3) $\rho_o (P_{x_2}\xi)' M_{x_1^*} P_{x_2} \xi = \rho_o \xi' P_{x_2} \xi - \rho_o (P_{x_2}\xi)' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}})$ because:

A. $\xi' P_{x_2} \xi = O_p(N^{1-\tilde{\delta}})$ as point 2.b.ii.A.

B. $\rho_o (P_{x_2}\xi)' P_{x_1^*} P_{x_2} \xi = O_p(N^{1-\tilde{\delta}})$ similar as point 2.b.ii.A.

(c) Term (4) of 50 is $O_p(N^{1/2})$. To see why, we expand term (4) and we obtain:

$$u^{*\prime} M_{x_1^*} \omega^* = \underset{(4.1)}{u'\omega} - \underset{(4.2)}{u' P_{x_2} \omega} - \underset{(4.3)}{u' P_{x_1^*} \omega} + \underset{(4.4)}{u' P_{x_1^*} P_{x_2} \omega} + \underset{(4.5)}{u' P_{x_2} P_{x_1^*} \omega} - \underset{(4.6)}{u' P_{x_2} P_{x_1^*} P_{x_2} \omega},$$

because:

i. Term (4.1) $u'\omega = O_p(N^{1/2})$ by Markov's inequality and independence of u and ω.

ii. Term (4.2) $u' P_{x_2} \omega = \sum_{i=1}^N u_i \omega_i p_{ii} + \sum_{i=1}^N \sum_{i \neq j} u_i \omega_j p_{ij} = O_p(N^{1/2})$ by Markov's Inequality and independence of u_i and ω_i (with p_{ij} a typical element of P_{x_2}).

iii. Term (4.3) $u' P_{x_1^*} P_{x_2} \omega = O_p(N^{1/2})$, by applying a similar reasoning as point in 2.c.ii.

iv. Term (4.4) $u' P_{x_1^*} P_{x_2} \omega = O_p(N^{1/2})$ by applying a similar reasoning as point in 2.c.ii.

v. Term (4.5) $u' P_{x_2} P_{x_1^*} \omega = O_p(N^{1/2})$ by applying a similar reasoning as point in 2.c.ii.

vi. Term (4.6) $u' P_{x_2} P_{x_1^*} P_{x_2} \omega = O_p(N^{1/2})$ by applying a similar reasoning as point in 2.c.ii.

(d) Term (5) of 50 is $O_p(N^{1-\tilde{\delta}/2})$. To see why, we expand term (5) and we obtain:

$$\xi^{*\prime} M_{x_1^*} \omega^* = \underset{(5.1)}{\xi' M_{x_1^*} \omega} + \underset{(5.1)}{\xi' M_{x_1^*} \omega} - \underset{(5.2)}{\xi' P_{x_2} \omega} + \underset{(5.3)}{\xi' P_{x_1^*} P_{x_2} \omega} - \underset{(5.4)}{\xi' P_{x_2} \omega} + \underset{(5.5)}{\xi' P_{x_2} P_{x_1^*} \omega} + \underset{(5.6)}{\xi' P_{x_2} P_{x_2} \omega} - \underset{(5.7)}{\xi' P_{x_2} P_{x_1^*} P_{x_2} \omega}$$

because:

i. Term (5.1) $\xi' M_{x_1^*} \omega = O_p(N^{1-\tilde{\delta}/2})$ because:

A. $\xi'\omega = O_p(N^{1-\tilde{\delta}/2})$ because by Triangle inequality $E|\sum \omega_i \xi_i| \leq \sum_{i=1}^N E|\omega_i \xi_i|$ and by Hölder inequality $\sum_{i=1}^N E|\omega_i^2|^{1/2}|\xi_i|^{21/2} = NE|\omega_i^2|^{1/2} E|\xi_i^2|^{1/2}$ where the equality holds because ω_i and ξ_i are i.i.d. Finally, $E|\omega_i^2| = O_p(1)$ because of Assumption 1 and$E|\xi_i^2| = O_P(N^{-\tilde{\delta}})$ by Lemma 1.

B. $\xi' P_{x_1^*} \omega = O_p(N^{1-\tilde{\delta}/2})$ because $\xi' P_{x_1^*} \omega \leq ||P_{x_1^*}\omega|| ||P_{x_1^*}\xi'||$ by Cauchy Schwarz inequality and $|P_{x_1^*}\omega|| ||P_{x_1^*}\xi'|| \leq ||\omega|| ||\xi||$ by Pythagoras' theorem.

ii. Term (5.3) $\xi' P_{x_2} \omega = O_p(N^{1-\tilde{\delta}/2})$ by similar reasoning as in point 2.d.i.B).

iii. Term (5.4) $\xi' P_{x_1^*} P_{x_2} \omega = O_p(N^{1-\tilde{\delta}/2})$, by similar reasoning as in point 2.d.i.B).

iv. Term (5.5) $\xi' P_{x_2} \omega = O_p(N^{1-\tilde{\delta}/2})$, by similar reasoning as in point 2.d.i.B).

v. Term (5.6) $\xi' P_{x_2} P_{x_1^*} \omega = O_p(N^{1-\tilde{\delta}/2})$, by similar reasoning as in point 2.d.i.B).

vi. Term (5.7) $\xi' P_{x_2} P_{x_2} \omega = O_p(N^{1-\tilde{\delta}/2})$, by similar reasoning as in point 2.d.i.B).

vii. Term (5.8) (5.8)$\xi' P_{x_2} P_{x_1^*} P_{x_2} \omega = O_p(N^{1-\tilde{\delta}/2})$, by similar reasoning as in point 2.d.i.B).

We can conclude that:

$$\hat{\rho}_o - \rho_o = o_p(1)$$

because the denominator is $O_p(N)$ and the numerator is $O_p(N^{1-\tilde{\delta}/2})$.

12.1.4 Proof of Theorem 3

The estimator of β_{1_o} is equal to:

$$\hat{\beta}_{1_o} = (\hat{u}'x_1^*)^{-1}(\hat{u}'\hat{y}^{**}), \tag{54}$$

with $\hat{u} = x - \hat{g}_o(x_2)$, $\hat{u}^* = M_{x_2}\hat{u}$, $x_1^* = M_{x_2}x_1$ and

$$y^* = x_1^*\beta_{1_o} + \rho_o u^* + \omega^*, \tag{55a}$$

$$\hat{y}^{**} = y^* - \hat{\rho}_o\hat{u}^*. \tag{55b}$$

Replacing 55 into $\hat{\beta}_{1_o}$ gives:

$$\hat{\beta}_{1_o} = \beta_{1_o} + \underset{(1)}{(\hat{u}'x_1^*)^{-1}}\underset{(2)}{(\rho_o\hat{u}'u^* + \hat{u}'\omega^* - \hat{\rho}_o\hat{u}'\hat{u}^*)}. \tag{56}$$

Now, if we analyze each term of 56:

1. The denominator of 56 is $O_p(N)$. To see why, we expand term (1) and obtain:

$$\hat{u}'x_1^* = \underset{(1.1)}{u'M_{x_2}x_1} + \underset{(1.2)}{\xi' M_{x_2}x_1}.$$

   i. Term (1.1) $u'M_{x_2}x_1 = u'M_{x_2}u + u'M_{x_2}g_o(x_2) = O_p(N)$ because:
      - $u'M_{x_2}u = u'u + u'P_{x_2}u = O_p(N)$ with: $u'u = O_p(N)$ by Markov's Inequality and Assumption 1, and $u'P_{x_2}u = O_p(N)$ by Pythagoras' theorem, Cauchy-Schwarz inequality and Assumption 1.

   ii. $u'M_{x_2}g_o(x_2) = u'g_o(x_2) + u'P_{x_2}g_o(x_2)$ with $u'g_o(x_2) = O_p(N^{1/2})$ by Markov' Inequality and Assumption 1, and $u'P_{x_2}g_o(x_2) = O_p(N^{1/2})$ because $u'P_{x_2}g_o(x_2) = \sum_{i=1}^N u_i g_o(x_{i2})p_{ii} + \sum_i\sum_j u_i g_o(x_{j2})p_{ji} = O_p(N^{1/2})$ by Markovs' inequality (with p_{ij} a typical element of P_{x_2}).

   iii. Term (1.2) $\xi' M_{x_2}x_1 = \xi' M_{x_2}u + \xi' M_{x_2}g(x_2) = O_p(N^{1-\tilde{\delta}})$ because:

      A. $\xi' M_{x_2}u = \xi' u + \xi' P_{x_2}u$ with $\xi' u = O_p(N^{1-\tilde{\delta}/2})$ by Triangle Inequality, Hölder's inequality and Lemma 1, and $\xi' P_{x_2}u = O_p(N^{1-\tilde{\delta}/2})$ by Cauchy-Schwarz inequality, Pythagoras' theorem and Lemma 1.

2. The numerator of 56 is $O_p(N^{1-\tilde{\delta}})$. To get this result we replace $\hat{\rho}_o = \rho_o + o_p(1)$ in Term (2) and obtain:

$$(2) = \underset{(2.1)}{u'M_{x_2}\omega} + \underset{(2.2.)}{\xi' M_{x_2}\omega} - \underset{(2.3)}{\rho_o u'M_{x_2}\xi} - \underset{(2.4)}{\rho_o\xi' M_{x_2}\xi} - o_p(1)$$

   Analyzing term by term:

   (a) Term (2.1) $u'M_{x_2}\omega = u'\omega - u'P_{x_2}\omega = O_p(N^{1/2})$ because $u'\omega = O_p(N^{1/2})$ by Markov's Inequality and independence of u_i and ω_i. $u'P_{x_2}\omega = \sum_{i=1}^N u_i\omega_i p_{ii} + \sum_{i=1}^N\sum_{i\neq j} u_i\omega_j p_{ij} = O_p(N^{1/2})$ by Markov's Inequality and independence of u_i and ω_i.

   (b) Term (2.2) $\xi' M_{x_2}\omega = \xi'\omega - \xi' P_{x_2}\omega = O_p(N^{1-\tilde{\delta}/2})$ because $\xi'\omega = O_p(N^{1-\tilde{\delta}})$ by Hölder's Inequality, Assumption 1 and Lemma 1, $\xi' P_{x_2}\omega$ by Pythagoras' theorem, Cauchy-Schwarz Inequality, Assumption 1 and Lemma 1.

   (c) Term (2.3) $\rho_o u'M_{x_2}\xi = \rho_o u'\xi - \rho_o u'P_{x_2}\xi = O_p(N^{1-\tilde{\delta}/2})$ because $\rho_o u'\xi = O_p(N^{1-\tilde{\delta}/2})$ by Hölder's Inequality, Assumption 1 and Lemma 1 and $\rho_o u'P_{x_2}\xi = O_p(N^{1-\tilde{\delta}/2})$ by Pythagoras' theorem, Cauchy-Schwarz Inequality Assumption 1 and Lemma 1.

   (d) Term (2.4) $\rho_o\xi' M_{x_2}\xi = \rho_o\xi'\xi - \rho_o\xi' P_{x_2}\xi = O_p(N^{1-\tilde{\delta}})$ because $\rho_o\xi'\xi = O_p(N^{1-\tilde{\delta}})$ by Lemma 1, and $\rho_o\xi' P_{x_2}\xi = O_p(N^{1-\tilde{\delta}})$ by Pythagoras' theorem, Cauchy-Schwarz Inequality, and Lemma 1.

   Then, we can conclude that $\hat{\beta}_{1_o} - \beta_{1_o} = o_p(1)$.

12.1.5 Proof of Proposition 2

In order to obtain the asymptotic distribution of our estimator, we pre-multiply it by the stabilizing rate $\sqrt{N}$:

$$\sqrt{N}(\hat{\beta}_{1_o} - \beta_{1_o}) = \underset{(1)}{(\tfrac{1}{N}\hat{u}'x_1^*)^{-1}}\underset{(2)}{(\tfrac{1}{\sqrt{N}}\rho_o\hat{u}'u^* + \tfrac{1}{\sqrt{N}}\hat{u}'\omega^* - \tfrac{1}{\sqrt{N}}\hat{\rho}_o\hat{u}'\hat{u}^*)}.$$

In the proof of Theorem 3, we show that $(\frac{1}{N}\hat{u}'x_1^*) = O_p(N)$. Thus, if we divide it by N and apply the plim operator we obtain:

$$plim\frac{1}{N}\hat{u}'x_1^* = plim\frac{1}{N}u'M_{x_2}u + plim\frac{1}{N}u'M_{x_2}g_o(x_2) = D,$$

where $D < \infty$ is a constant.
While the term (2):

$$(\frac{1}{\sqrt{N}}\rho_o\hat{u}'u^* + \frac{1}{\sqrt{N}}\hat{u}'\omega^* - \frac{1}{\sqrt{N}}\hat{\rho}_o\hat{u}'\hat{u}^*),$$

is equal to:

$$(2) = \frac{1}{\sqrt{N}}(u'M_{x_2}\omega + \xi'M_{x_2}\omega - \rho_o u'M_{x_2}\xi - \rho_o\xi'M_{x_2}\xi - o_p(1)).$$

By Lindeberg-Feller CLT:

$$\frac{1}{\sqrt{N}}u'M_{x_2}\omega \xrightarrow{L} N(0,V).$$

According to Chernozhukov et al. (2018), one can bound the terms $\frac{1}{\sqrt{N}}\xi'M_{x_2}\omega$ and $\frac{1}{\sqrt{N}}\rho_o u'M_{x_2}\xi$ using empirical process methods if the complexity of the parameter space of $g_o(x_2)$ is controlled. Moreover, Farrell et al. (2021) finds upper bounds on the complexity of the parameter space using localization analysis and proving that the integral of its entropy is upper bounded.
Finally, the term $\frac{1}{\sqrt{N}}\rho_o\xi'M_{x_2}\xi$ converges to 0 as shown in the proof of Proposition 1.
Thus,

$$\sqrt{N}(\hat{\beta}_{1_o} - \beta_{1_o}) \xrightarrow{L} N(0, D^{-1}VD^{-1})$$

with $D = plimN^{-1}\hat{u}'x_1^*$, and $V = E[\omega^2]u'M_{x_2}u$.

12.2 DGP used in simulations

In this annex, we present detailed results for the bias and RMSE of the estimation methods used are: OLS ignoring endogeneity (OLS), 2SLS using as IVs $\hat{g}(x_{i2})$ (2SLS), OLS p.: OLS using $\hat{g}(x_{i2})$ instead of x_{i1} (OLS Orth. Vble.), naive Control function approach (CF), naive Control function approach with sample splitting (CF SS), control function approach with orthogonal score (CF-OS), control function approach with orthogonal score with sample splitting (CF-OS SS), estimation with asymmetric errors (AE), and estimation with heteroskedastic errors (HE).

12.2.1 The results: details

Table 3: Results: Identification through non-linearity

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
1.1	1	Bias	0.6649	$2x10^3$	21.1024	$1x10^3$	-176.6393	0.0047	0.0029	0.2339
		RMSE	0.6649	$4x10^4$	347.9398	$5x10^3$	$7x10^4$	0.0289	0.0273	0.9058
	2	Bias	0.6649	0.00134	-0.0017	0.0043	-0.0026	0.0097	0.0170	0.5817
		RMSE	0.6649	0.0280	0.1056	0.0937	0.1011	0.0312	0.0368	0.5996
1.2	1	Bias	0.6897	$8x10^3$	14.2120	$1x10^3$	353.7859	0.0171	0.0151	0.6967
		RMSE	0.6897	$4x10^4$	194.0042	$6x10^3$	$6x10^3$	0.0340	0.0287	0.7224
	2	Bias	0.6897	0.0034	-0.0224	0.0258	-0.0034	0.0242	0.0338	0.5448
		RMSE	0.6897	0.0250	0.1085	0.0968	0.0947	0.0490	0.0537	0.6118
1.3	1	Bias	0.3828	$-1x10^3$	18.5853	-558.2235	$1x10^3$	0.0044	0.0030	0.1735
		RMSE	0.3830	$6x10^4$	137.3490	$6x10^3$	$8x10^3$	0.0272	0.0270	0.3045
	2	Bias	0.3828	0.0032	0.0023	0.0037	0.01197	0.01296	0.01197	0.0572
		RMSE	0.3830	0.02762	0.1033	0.0532	0.0513	0.0295	0.0305	0.1920
1.4	1	Bias	0.6670	$3x10^3$	-50.0306	$1x10^3$	$2x10^3$	0.0059	0.0099	0.3293
		RMSE	0.6672	1.4996	42.6451	$9x10^3$	$1x10^4$	0.0396	0.0490	0.4500
	2	Bias	0.6670	0.0027	-0.0062	0.0118	-0.0178	0.0148	-0.0178	0.5087
		RMSE	0.6672	0.0384	0.1232	0.0999	0.1080	0.0453	0.0464	0.5999

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IV $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 4: Results: Identification through non-linearity

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
2.1	1	Bias	0.8514	$1x10^3$	-0.2942	-198.0749	975.9478	0.0061	0.01197	0.4562
		RMSE	0.8514	$1x10^4$	1.6223	$5x10^3$	$6x10^3$	0.0716	0.0703	0.5032
	2	Bias	0.8514	-0.0031	0.0049	-0.0111	-0.0418	0.0094	0.0215	0.5811
		RMSE	0.8514	0.0747	0.1904	0.1756	0.1712	0.0738	0.0732	0.5995
2.2	1	Bias	0.8532	$2X10^3$	-0.6772	936.9051	702.8788	0.0059	0.0237	0.3061
		RMSE	0.853	$3X10^4$	8.9066	$7x10^3$	$7x10^3$	0.0665	0.0775	0.4491
	2	Bias	0.8532	-0.0005	0.0106	-0.0142	-0.0069	0.0200	0.0268	0.6011
		RMSE	0.8532	0.0622	0.2105	0.2234	0.1411	0.0735	0.0635	0.6255
2.3	1	Bias	0.4375	855.6320	-6.1070	-95.9539	-381.1568	-0.0048	-0.0017	0.0223
		RMSE	0.4376	$2X10^4$	55.3283	$8x10^3$	$7x10^3$	0.0732	0.0988	0.1996
	2	Bias	0.4375	-0.0077	0.0234	-0.0215	-0.0405	0.0034	0.0085	0.079
		RMSE	0.4376	0.0711	0.2569	0.1307	0.1294	0.0674	0.0670	0.2297
2.4	1	Bias	0.8529	-0.3089	-0.3257	$1X10^3$	$1x10^3$	0.0097	0.0190	0.2046
		RMSE	0.8530	0.5398	0.6096	$1x10^4$	$1x10^4$	0.0978	0.1108	0.3826
	2	Bias	0.8529	-0.0027	0.0145	-0.0154	-0.0363	0.0137	0.0184	0.5225
		RMSE	0.8529	0.1041	0.2389	0.19465	0.1813	0.1056	0.1028	0.6146

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IV $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 5: Results: Identification through non-linearity

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
3.1	1	Bias	0.1797	-928.2880	-0.3099	-445.6430	271.4341	0.0004	0.0027	0.1920
		RMSE	0.1797	$4x10^4$	0.5580	$3x10^3$	$6x10^3$	0.0084	0.0098	0.6014
	2	Bias	0.1797	-0.0003	0.0354	-0.0326	-0.0748	0.0009	0.0046 0.5658	
		RMSE	0.1797	0.0083	0.1939	0.1745	0.2572	0.0082	0.0105	0.5860
3.2	1	Bias	0.1789	$4x10^3$	-0.2818	-203.8318	-858.9137	0.0004	0.004	-0.0129
		RMSE	0.1790	$4x10^4$	0.5398	$7x10^3$	$7x10^3$	0.0068	0.0113	0.6554
	2	Bias	0.1790	-0.0005	0.01797	-0.0167	-0.0840	0.0009	0.0020	0.0045
		RMSE	0.1790	0.0067	0.14392	0.1259	0.2667	0.0069	0.0093	0.5657
3.3	1	Bias	0.1494	$2x10^3$	-0.4034	258.1177	$1x10^3$	-0.0004	0.0011	-0.1582
		RMSE	0.1495	$1x10^5$	0.6334	$7x10^3$	$6x10^3$	0.0081	0.0095	0.4556
	2	Bias	0.1494	-0.0009	0.04645	-0.0222	-0.0026	-0.00016	0.00180	0.0817
		RMSE	0.1495	0.0081	0.2177	0.0980	0.1700	0.0081	0.0088	0.1963
3.4	1	Bias	0.1796	$-5x10^3$	-0.3335	$1x10^3$	-907.0396	0.0001	0.0018	0.1341
		RMSE	0.1798	$7x10^4$	0.5755	$5x10^3$	$1x10^4$	0.0119	0.0354	0.0164
	2	Bias	0.1796	-0.0003	0.07633	-0.0712	-0.0766	0.0006	0.0048	0.4388
		RMSE	0.1798	0.0118	0.2523	0.2349	0.2679	0.0121	0.01493	0.5707

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IV $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 6: Results: Identification through non-linearity

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
4.1	1	Bias	0.3355	-0.3021	0.0094	115.9586	164.1640	-0.1218	-0.4279	-0.6845
		RMSE	0.3355	0.54714	0.7159	$5x10^3$	$6x10^3$	0.1307	0.1261	1.0997
	2	Bias	0.3355	-0.0006	0.0524	-0.0430	-0.1643	-0.0488	-0.1643	0.4787
		RMS	0.3355	0.0118	0.1144	0.0906	0.0300	0.2136	0.1788	0.5720
4.2	1	Bias	0.3176	-0.3703	-0.0955	846.4034	-41.8410	-0.4188	-0.4676	-0.5346
		RMSE	0.3177	0.7167	0.8895	$6x10^3$	$6x10^3$	0.59238	0.9269	1.9621
	2	Bias	0.3176	-0.0006	0.1409	-0.1023	-0.2812	-0.1218	-0.4279	0.4202
		RMSE	0.3177	0.0118	0.3126	0.2073	0.4321	0.2208	0.7546	0.6294
4.3	1	Bias	0.2446	-0.4117	-0.1621	-239.0584	-336.7148	-0.0629	-0.0753	-0.3791
		RMSE	0.2447	0.6411	0.7771	$5x10^3$	$5x10^3$	0.0912	0.0926	0.8051
	2	Bias	0.2446	-0.0002	0.0861	-0.0337	-0.0979	-0.0363	-0.0839	0.0466
		RMSE	0.2447	0.0120	0.1825	0.0673	0.1276	0.0584	0.0970	0.2912
4.4	1	Bias	0.3360	-0.3814	-0.1427	369.0024	$3x10^3$	-0.0954	-0.1120	-0.1239
		RMSE	0.3361	0.6169	0.7525	$7x10^3$	$1x10^4$	0.1287	0.1383	0.3682
	2	Bias	0.3360	0.0005	0.0438	-0.0328	-0.1533	-0.0404	-0.1402	0.3306
		RMSE	0.3361	0.0200	0.1085	0.0792	0.1893	0.0774	0.1580	0.5347

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IV $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 7: Results: Identification through non-linearity

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
5.1	3	Bias	0.6057	$1x10^3$	1.7134	$1x10^3$	135.1169	-0.0005	0.0018	-0.6044
		RMSE	0.6060	$4x10^4$	12.6218	$1x10^4$	$1x10^4$	0.0305	0.0339	1.3416
	2	Bias	0.6057	-0.0016	0.0095	-0.0143	-0.0269	0.0037	0.0066	0.6060
		RMSE	0.6060	0.0280	0.0949	0.1127	0.0925	0.0331	0.0361	0.8420
5.2	3	Bias	0.4485	$3x10^3$	8.2462	-125.4860	-637.0997	0.0025	-0.0003	-0.4141
		RMSE	0.5248	$2x10^4$	30.9276	$8x10^3$	$1x10^4$	0.0225	0.0228	1.3551
	2	Bias	0.5245	-0.0018	0.0340	-0.0336	-0.0274	0.0015	-0.0010	-0.1332
		RMSE	0.5248	0.0179	0.1102	0.1162	0.1324	0.0269	0.0236	1.8098
5.3	3	Bias	0.4634	-290.4078	0.6734	-544.4481	-663.3685	-0.0011	0.0012	-0.7514
		RMSE	0.4637	$3x10^4$	9.1896	$6x10^3$	$1x10^4$	0.0235	0.0243	1.2807
	2	Bias	0.4634	-0.0030	-0.0001	-0.0051	-0.0178	-0.0006	0.0019	0.0986
		RMSE	0.3593	0.0228	0.0903	0.0619	0.0589	0.0237	0.0237	0.7200
5.4	3	Bias	0.6051	15.4014	5.4780	$1x10^3$	-300.6532	-0.0003	0.0010	0.1042
		RMSE	0.6059	$4x10^4$	28.2858	$1x10^4$	$1x10^4$	0.0411	0.0443	1.1745
	2	Bias	0.6051	-0.0027	0.0210	-0.0303	-0.0215	0.0010	0.0053	0.7554
		RMSE	0.6059	0.0401	0.0939	0.1118	0.1048	0.0443	0.0417	0.9908

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IV $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 8: Results: Identification through non-linearity

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
6.1	1	Bias	0.6005	-890.3404	0.4323	558.5078	-616.3224	0.0012	0.0051	0.5243
		RMSE	0.6006	$3x10^4$	6.2552	$7X10^3$	$1X10^4$	0.0338	0.0386	1.1550
	2	Bias	0.6005	-0.0018	0.0073	-0.0165	-0.0116	0.0017	0.0048	0.6060
		RMSE	0.6006	0.0337	0.0716	0.1174	0.1120	0.0346	0.0348	0.8420
6.2	1	Bias	0.4485	-0.2974	4.4691	-525.6529	-586.2450	-0.0003	0.0033	0.6075
		RMSE	0.4486	0.5484	22.7238	$7X10^3$	$5X10^3$	0.0226	0.0403	1.3727
	2	Bias	0.4485	-0.0022	-0.0005	-0.0057	-0.0551	0.0002	0.0004	0.4059
		RMSE	0.4486	0.0269	0.0814	0.1320	0.1275	0.0300	0.0312	1.2859
6.3	1	Bias	0.3591	$3x10^3$	-0.3090	$1x10^3$	$5x10^3$	-0.0016	-0.0007	0.2016
		RMSE	0.3592	$3x10^4$	0.5612	$1X10^4$	$2X10^4$	0.0324	0.0366	0.7777
	2	Bias	0.3591	-0.0036	0.0027	-0.0091	-0.0184	-0.0003	0.0015	0.2365
		RMSE	0.3592	0.0327	0.0924	0.0913	0.0834	0.0328	0.0345	0.9094
6.4	1	Bias	0.6017	$1x10^3$	0.7729	775.5588	$-2X10^3$	-0.0006	0.0010	0.1425
		RMSE	0.6020	$4x10^4$	7.2574	$1X10^4$	$2X10^4$	0.0433	0.0560	1.3331
	2	Bias	0.6017	-0.0028	0.0035	-0.0129	-0.0252	0.0016	0.0029	1.2737
		RMSE	0.6020	0.0470	0.0917	0.1324	0.1090	0.0477	0.0486	1.3936

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IVs $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 9: Results: Identification through non symmetric errors

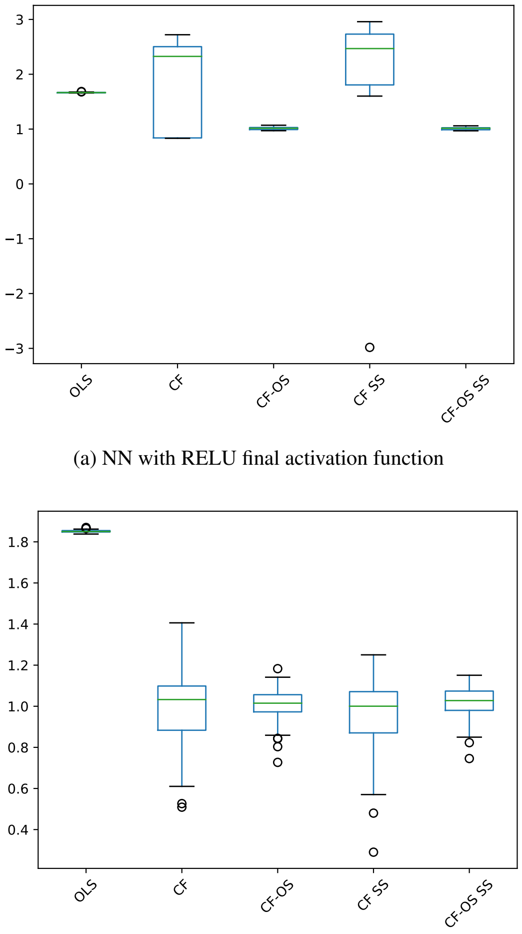
Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE
7.1	1	Bias	0.9762	$1x10^3$	0.4953	146.8609	-304.7458	1.000	1.001	-0.0000
		RMSE	0.9832	1.0079	1.0310	1,799.0315	1,134.1874	1.0004	1.0017	0.0087
	2	Bias	0.9762	1.010	0.9251	1.0055	1.0092	1.0070	1.0247	-0.0017
		RMSE	0.9832	1.010	1.0542	796.8959	3,683.3361	1.0005	1.0016	0.0193
7.2	1	Bias	0.9832	0.4679	0.2865	-276.0129	128.8530	1.0006	1.0018	-0.0000
		RMSE	0.9832	1.0080	1.0568	1,823.3477	1,995.9748	0.7887	1.0019	0.0194
	2	Bias	0.9832	1.011	1.0429	1.0009	1.0014	1.0003	1.0015	-0.0002
		RMSE	0.9832	1.011	1.0555	1.0009	1.0014	1.0004	1.0017	0.0081
7.3	1	Bias	0.9832	0.5268	0.3998	42.2218	-129.9619	1.0005	1.0019	-0.0025
		RMSE	0.9832	1.0069	1.0502	1,474.3769	2,431.0240	0.8249ß	1.0020	0.0182
	2	Bias	0.9832	1.0098	1.1822	1.0021	1.0010	1.0003	1.0016	-0.0007
		RMSE	0.9832	1.0099	1.5399	1.0022	1.0011	1.0003	1.0017	0.0082
7.4	1	Bias	0.9832	0.3676	0.3335	-294.0027	-116.4456	1.0004	1.0017	-0.0034
		RMSE	0.9832	1.0077	1.0446	1,706.1568	1,353.6921	1.0005	1.0018	0.0172
	2	Bias	0.9832	1.0113	1.0532	1.0009	1.0007	1.0003	1.0016	-0.0019
		RMSE	0.9832	1.0115	1.0679	1.0009	1.0007	1.0004	1.0017	0.0098

NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IVs $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors.

Table 10: Results: Identification through non symmetric errors

Scenario	NNa		OLS	2SLS	OLS p.	CF	CF-SS	CF-OS	CF-OS SS	AE	HE
8.1	1	Bias	0.8631	-1.7457	1.0873	0.5553	1.5123	1.0005	0.9981	-0.3601	0.0743
		RMSE	0.8632	2.1172	3.1759	0.9457	31.4379	1.0007	0.9988	1.2729	0.1019
	2	Bias	0.5645	0.9839	1.0049	1.0021	1.0069	0.9996	1.0052	-0.3048	0.0518
		RMSE	0.5649	0.9931	1.0261	1.0043	1.0086	1.0002	1.0078	0.7385	0.1106

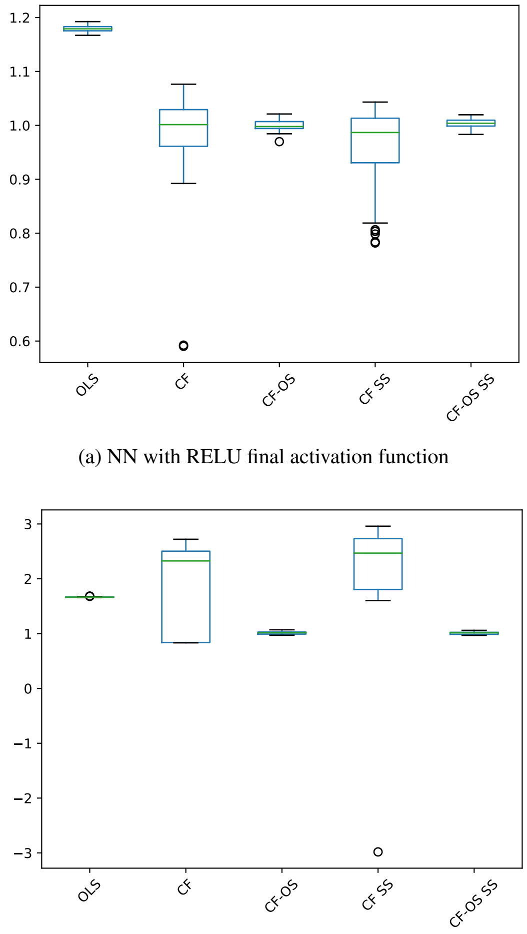
NNa: Neural Network Architecture, OLS: OLS ignoring endogeneity, 2SLS: 2SLS using as IVs $\hat{g}(x_{i2})$, OLS p.: OLS with $\hat{g}(x_{i2})$ in lieu of x_{i1}, CF: Naive Control function approach, CF SS: Naive Control function approach with sample splitting, CF-OS: Control function approach with orthogonal score, CF-OS SS: Control function approach with orthogonal score with sample splitting, AE: estimation with asymmetric errors, HE: estimation with heteroskedastic errors.



(a) NN with RELU final activation function

(b) NN with linear final activation function

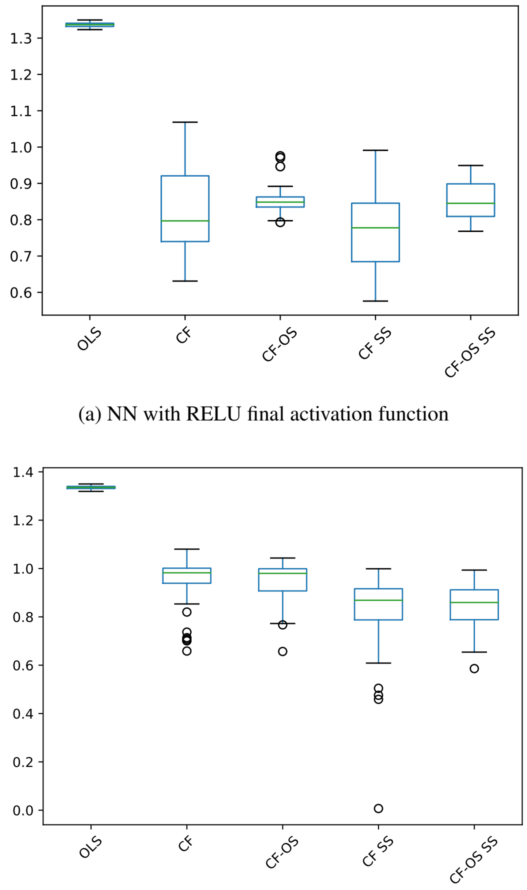
Figure 5: Linear structural equation and non-linear reduced-form equation: DGP 2-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).



(a) NN with RELU final activation function

(b) NN with linear final activation function

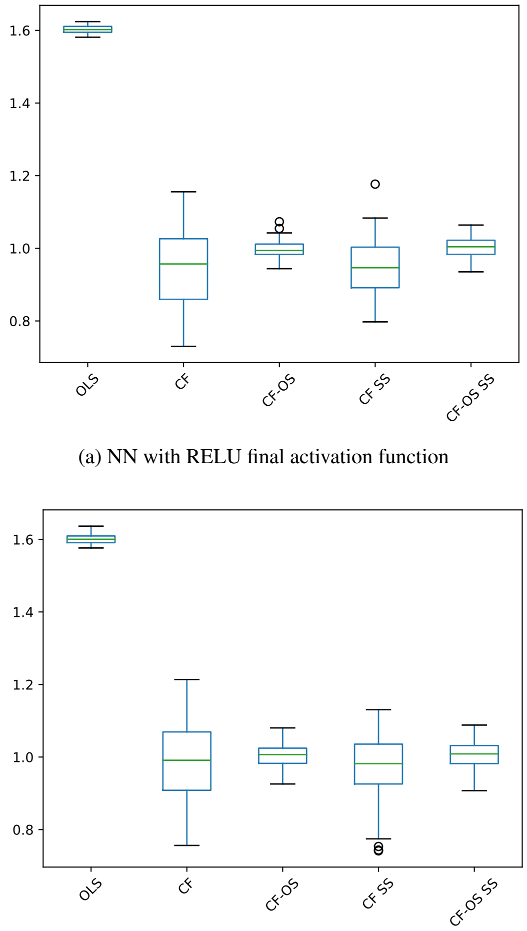
Figure 6: Linear structural equation and non-linear reduced-form equation: DGP 3-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).



(a) NN with RELU final activation function

(b) NN with linear final activation function

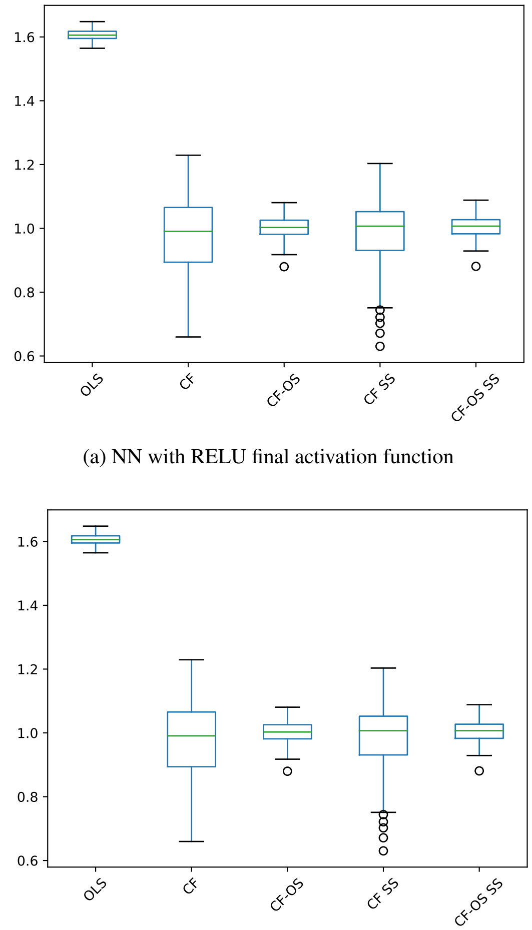
Figure 7: Linear structural equation and non-linear reduced-form equation: DGP 4-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).



(a) NN with RELU final activation function

(b) NN with linear final activation function

Figure 8: Linear structural equation and non-linear reduced-form equation: DGP 5-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).



(a) NN with RELU final activation function

(b) NN with linear final activation function

Figure 9: Linear structural equation and non-linear reduced-form equation: DGP 6-Scenario 1
Note: The true parameter value is 1 ($\beta_1 = 1$).

References

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics*. Princeton university press.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chudik, A. and M. H. Pesaran (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics 188*(2), 393–420.

Dette, H. and A. Munk (1998). Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(4), 693–708.

Eubank, R. L. and W. Thomas (1993). Detecting heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(1), 145–155.

Farré, L., R. Klein, and F. Vella (2013). A parametric control function approach to estimating the returns to schooling in the absence of exclusion restrictions: an application to the nlsy. *Empirical Economics 44*(1), 111–133.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Ferguson, T. S. (1996). *A course in large sample theory*. Routledge.

Hahn, J. and G. Ridder (2019). Three-stage semi-parametric inference: Control variables and differentiability. *Journal of econometrics 211*(1), 262–293.

Hausman, J. A. (1983). Chapter 7 specification and estimation of simultaneous equation models. Volume 1 of *Handbook of Econometrics*, pp. 391–448. Elsevier.

Hinton, G., N. Srivastava, and K. Swersky (2012). Neural networks for machine learning: Lecture 6a.

Horowitz, J. L. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica 74*(2), 521–538.

Klein, R. and F. Vella (2010). Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics 154*(2), 154–164.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics 30*(1), 67–80.

Lewbel, A., S. M. Schennach, and L. Zhang (2020). Identification of a Triangular Two Equation System Without Instruments.

Martins-Filho, C. and F. Yao (2012). Kernel-based estimation of semiparametric regression in triangular systems. *Economics Letters 115*(1), 24–27.

Matzkin, R. L. (2016). On independence conditions in nonseparable models: Observable and unobservable instruments. *Journal of Econometrics 191*(2), 302–311. Innovations in Measurement in Economics and Econometrics.

Nair, V. and G. E. Hinton (2010). Rectified linear units improve restricted boltzmann machines. In *Icml*.

Neumeyer, N. and H. Dette (2007). Testing for symmetric error distribution in nonparametric regression models. *Statistica Sinica*, 775–795.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier.

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics 85*(4), 777–792.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research 15*(1), 1929–1958.

Su, L. and A. Ullah (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics 144*(1), 193–218.

White, H. (1984). *Asymptotic theory for econometricians*. Academic press.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources 50*(2), 420–445.