Tiwari, Amaresh

Book

Panel data binary response model in a triangular system

University of Tartu
Faculty of Social Sciences
School of Economics and Business Administration

Panel Data Binary Response Model in a Triangular System

Amaresh K Tiwari

Tartu 2018

ISSN-L 1406-5967
ISSN 1736-8995
ISBN 978-9985-4-1109-4 (pdf)
The University of Tartu FEBA
www.mtk.ut.ee/en/research/workingpapers

PANEL DATA BINARY RESPONSE MODEL IN A TRIANGULAR SYSTEM [1]

AMARESH K TIWARI[2]

We propose a new control function (CF) method for binary response outcomes in a triangular system with unobserved heterogeneity of multiple dimensions. The identified CFs are the expected values of the heterogeneity terms in the reduced form equations conditional on the endogenous, $X_i \equiv (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})$, and the exogenous, $Z_i \equiv (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iT})$, variables. The method requires weaker restrictions compared to traditional CF methods for triangular systems with imposed structures similar to ours, and point-identifies average partial effects with discrete instruments. We discuss semiparametric identification of structural measures using the proposed CFs. An application and Monte Carlo experiments compare several alternative methods with ours.

KEYWORDS: Control Functions, Unobserved Heterogeneity, Identification, Instrumental Variables, Average Partial Effects, Child Labor.

JEL Classification: C13, C18, C33

1. INTRODUCTION

Chamberlain (2010) and Arellano and Bonhomme (2011) point out that when panel data outcomes are discrete, serious identification issues arise when covariates are correlated with unobserved heterogeneity. Chamberlain shows that for binary choice model with fixed T, quantities of interest such as Average Partial Effect (APE) may not be point identified, or may not possess a $\sqrt{N}$ consistent estimator. Notwithstanding this underidentification result, various methods have been proposed to estimate the structural measures of interest.

Arellano and Bonhomme (2011) provide an overview, and categorize, of some of the methods developed to estimate the quantities of interest. These include the fixed effect (FE) approach that treat heterogeneity or individual effects as parameters to be estimated, where several approaches have been proposed to correct for bias due the incidental parameter problem. Wooldridge (2010), points out that the FE approach, though promising, suffers from a number of shortcomings. First, the number of time periods needed for the bias adjustments to work well is often greater than is available in many applications. Secondly, the recent bias adjustments methods require the assumptions of stationarity and weak dependence; in some cases, the strong assumption of serial independence (conditional on the heterogeneity) is maintained. However, in empirical work dealing with linear models, it has been found that idiosyncratic errors exhibit serial dependence. Also, "the requirement of stationarity is strong and has substantive restrictions as it rules out staples in empirical

[1] I would like to thank anonymous referees for helpful comments. Thanks are due to seminar participants at the The Bank of Estonia, the 9^{th} Nordic Econometric Meeting (Tartu), the Institute of Mathematics and Statistics (University of Tartu), and the Inaugural Baltic Economic Conference (Vilnius) for the same. I would especially like to thank Soham Sahoo for helping me with the data. All remaining errors are mine.

[2] University of Tartu, amaresh.kr.tiwari@gmail.com & amaresh.tiwari@ut.ee

work such as including separate year effects, which can be estimated very precisely given a large cross section."

There is another class of models that acknowledges the fact that many nonlinear panel data models are not point identified at fixed T and consequently discuss set identification (bound analysis) for certain quantiles of interest such as the marginal effects. These papers show that the bounds become tighter as the number of time periods, T, increases. However, the methods in most of these papers are still limited to discrete covariates. Moreover, these papers and papers utilizing FE approach assume that conditional on unobserved heterogeneity all covariates are exogenous or predetermined; this, as argued in Hoderlein and White (2012) (henceforth HW), may not always hold true.

In this paper, we relax the assumption of conditional exogeneity to allow for endogenous covariates that are continuous, and develop a control function method to identify and estimate structural measures of interest such as the Average Structural Function (ASF) and APE while accounting for endogeneity and heterogeneity in a triangular system.

Some of the papers that adopt the control function approach to study binary or fractional response outcomes are Blundell and Powell (2004) (BP), Papke and Wooldridge (2008) (PW) and Rothe (2009). A partial list of papers that study nonparametric control function estimation of nonseparable, including binary response, models are Altonji and Matzkin (2005) (AM), Florens *et al.* (2008), Imbens and Newey (2009), HW, and Torgovitsky (2015), where the focus is on estimating heterogeneous effect of endogenous treatment. Apart from PW, who specify the correlation between individual specific random effects and the exogenous variables, in these papers the exogenous covariates are assumed to be independent of unobserved heterogeneity.

Typically, in a simultaneous triangular system, unobserved heterogeneity in the reduced form equations (Florens *et al.* term the reduced form equations as "treatment choice equations") is assumed to be scalar, where the identifying assumption is that conditional on these scalar time-varying heterogeneity/errors or its CDF, which are identified, all covariates are independent of the heterogeneity in the structural equation. However, we know that economic models suggest heterogeneity in tastes, technologies, abilities, etc. that are unobserved. Also, some of these unobserved heterogeneity might as well be multidimensional. Kasy (2011) shows that for the existing control function methods, identification fails when the reduced form equations have multiple sources of the heterogeneity.

The exceptions are PW and HW, who consider panel data where heterogeneity, constituting of time invariant random effects and idiosyncratic errors, is multidimensional. While the imposed structures in PW's is similar to ours, they make the traditional control function assumption, and so their control function is scalar, whereas our control function is vector valued, whose dimension depend on the dimension of unobserved heterogeneity. And HW's specification of the triangular system does not nest ours.

We allow for multidimensional heterogeneity in form of separable error components comprising of time invariant correlated random effects (or random coefficients) and idiosyncratic terms, and assume that conditional on various error components of the reduced form equations the covariates are independent of the heterogeneities in the structural equation. But the various error components of the reduced form equations are not identified separately for them to be used as control functions.

However, the expected values of the error components of the reduced form equations conditional on the endogenous variables, $X_i \equiv (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})$, and the exogenous variables, $Z_i \equiv (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iT})$, are straightforwardly identified when the distributions of the error components are specified. Under our assumptions, this aids in identifying the measures of interest. We then propose that these conditional expected values of the error components of the reduced form equations be used as control functions, and argue that for triangular systems with setups similar to ours, these control functions imply a weaker restriction than the commonly made control function assumptions.

Our method, while being simple, makes a number of contributions to the literature. First, we allow for multiple sources of heterogeneity, albeit with restrictions, in the triangular system, where most papers, adopting the control function approach to handle endogeneity, do not.

Secondly, our method allows for instruments with small support, that is, instruments that are binary, discrete, or continuous, when point-identifying the ASF or the APEs. The two exceptions in the control function literature that we know of where point-identification in nonseparable triangular system is achieved when instruments have small support are D'Haultfœuille and Février (2015) and Torgovitsky. In our paper, we exploit separability of errors and panel data with repeated observations of the same unit for the purpose of identification when instruments have small support. Finally, our model retains the attractive features of PW's, where no assumptions are made on the serial dependence among the outcome variable.

Using data on India, the proposed estimator is employed to estimate causal effects of household income and wealth on the incidence of child labor. We find a strong effect of correcting for endogeneity, and show that the standard parametric models give a misleading picture of the causal effect of income and wealth on child labor.

The rest of the paper is organized as follows. In section 2 we introduce the model and discuss identification and estimation of structural measures of interests for a discrete response model in a triangular system with random effects. In section 3 we discuss the results of the Monte Carlo experiments, which have been conducted to compare our estimator with some of the existing methods for panel data binary response model, where the imposed structures are similar to ours. In section 4 we extend the model with random effects to allow for random coefficients. In section 5 we apply the proposed estimator to study income and wealth effects on work decision outcomes for children in the State of Andhra Pradesh of India, and finally in section 6 we conclude. Some technical details are to be found in the appendix. Due to space constraint, other technical details, which includes large sample properties of the estimator, have been put in a supplementary appendix.

2. MODEL SPECIFICATION AND IDENTIFICATION AND ESTIMATION OF STRUCTURAL MEASURES

Consider the following binary choice model in a triangular setup:

$$y_{it} = 1\{y^*_{it} = (\boldsymbol{w}'_{it}, \boldsymbol{x}'_{it})\boldsymbol{\varphi} + \theta_i + \zeta_{it} > 0\}, \tag{2.1}$$

where $1\{.\}$ is an indicator function that takes value 1 if the argument in the parenthesis holds true and 0 otherwise. In (2.1), θ_i is the unobserved time invariant individual effect and ζ_{it} is the idiosyncratic error component. The variables, $\boldsymbol{x}_{it}$, are endogenous in the sense that $\zeta_{it} \not\perp \boldsymbol{x}_{it}|\theta_i$; whereas most papers studying panel data binary choice model assume that $\zeta_{it} \perp \boldsymbol{x}_{it}|\theta_i$. We assume that each of the endogenous variables are continuous and have a large support. The dimension of $\boldsymbol{x}_{it}$ is d_x and the dimension of the exogenous variables, $\boldsymbol{w}_{it}$, is d_w.

The reduced form in the triangular system, which is estimated in the first stage, is a system of d_x linear equations,

$$\boldsymbol{x}_{it} = \pi \boldsymbol{z}_{it} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{it}. \tag{2.2}$$

In (2.2), π has a row dimension of d_x, $\boldsymbol{\alpha}_i \equiv (\alpha_{i1}, \ldots, \alpha_{id_x})'$ is the $(d_x \times 1)$ vector of unobserved random effects, $\boldsymbol{\epsilon}_{it} \equiv (\epsilon_{it1}, \ldots, \epsilon_{itd_x})'$ is the $(d_x \times 1)$ vector of idiosyncratic error terms, and $\boldsymbol{z}_{it} \equiv (\boldsymbol{w}'_{it}, \tilde{\boldsymbol{z}}'_{it})'$ is of dimension d_z. The dimension of the vector of instruments, $\tilde{\boldsymbol{z}}_{it}$, is greater than or equal to the dimension of $\boldsymbol{x}_{it}$. Such exclusion restriction, where $\tilde{\boldsymbol{z}}_{it}$ appears in the reduced form but not in the structural, are justified on economic grounds.

Since the presence or absence of the set of exogenous variables, $\boldsymbol{w}_{it}$, has no bearing on the identification results obtained in the paper, to ease notations we suppress it in the binary response model in the rest of the paper. All assumptions and results are to be understood as conditional on $\boldsymbol{w}_{it}$. Secondly, in the rest of the paper, except when needed, we will drop the individual subscript, i.

While we refer (2.2) as reduced form equation, it is possible that the triangular system in (2.1) and (2.2) is in fact fully simultaneous (see Blundell and Powell, 2004, for examples). However, even if a simultaneous system is not triangular, the triangular representation, such as the above, can be easily derived if the simultaneous equations involving y_t^* and $\boldsymbol{x}_t$ are linear and the errors are additively separable. Also, the triangular model can be generalized to allow for random coefficients instead of fixed coefficients. For the sake of exposition, we limit the analysis to fixed coefficients with random effects; a straightforward extension of the method to random coefficients is discussed in section 4.

We first define some notations: $X \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ is a $(d_x \times T)$ matrix, $Z \equiv (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ is of dimension $(d_z \times T)$, $\boldsymbol{\zeta} \equiv (\zeta_1, \ldots, \zeta_T)'$ a vector of idiosyncratic errors in the structural equation, and $\boldsymbol{\epsilon} \equiv (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)$ is a $(d_x \times T)$ matrix of idiosyncratic errors in the reduced form equations. Our first assumption toward identifying the structural measures of interest such as ASF and APE is:

AS 1 *(a)* $\boldsymbol{\zeta} \perp Z, \theta$ *&* $\boldsymbol{\epsilon} \perp Z, \boldsymbol{\alpha}$ *and (b)* $\theta \perp Z|\boldsymbol{\alpha}$.

In the above, while Z is independent of the idiosyncratic terms $\boldsymbol{\zeta}$ and $\boldsymbol{\epsilon}$, it is potentially correlated with the time invariant unobserved heterogeneities θ and $\boldsymbol{\alpha}$.

AS 2 *(a)*

$$\theta, \zeta_t|X, Z, \boldsymbol{\alpha} \sim \theta, \zeta_t|\boldsymbol{\epsilon}, Z, \boldsymbol{\alpha} \sim \theta, \zeta_t|\boldsymbol{\epsilon}, \boldsymbol{\alpha} \text{ where } \boldsymbol{\epsilon} = X - \mathrm{E}(X|Z, \boldsymbol{\alpha}).$$

(b) $\theta, \zeta_t \perp \boldsymbol{\epsilon}_{-t} | \boldsymbol{\alpha}, \boldsymbol{\epsilon}_t$,

In part (a) of AS 1, the assumption is that the dependence of the structural error terms θ and ζ_t on X, Z, and $\boldsymbol{\alpha}$ is completely characterized by the reduced form error components, $\boldsymbol{\epsilon}$ and $\boldsymbol{\alpha}$. The assumption in part (b), where only contemporaneous errors are correlated, has been made to ease exposition, and can be dropped.

Assumptions AS 1 and AS 2 are weaker than the identifying assumptions in traditional control function method, where both $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ are assumed independent of Z and it is assumed that $\zeta_t, \theta \perp \boldsymbol{x}_t | \boldsymbol{v}_t = \boldsymbol{\alpha} + \boldsymbol{\epsilon}_t$; such an assumption implies that heterogeneity in each of the d_x reduced form equations is scalar.

For identification, one of the requirements of our method is that we be able to recover the conditional distribution of $\boldsymbol{\alpha}$ given X and Z. However, we do not know of any semi or nonparametric estimator, and it is outside the scope of this paper to develop one, where the distributions or expectations of random effects/coefficients conditional on X and Z are estimated for a system of regressions. With parametric specification of the error components, however, this conditional distribution is obtained straightforwardly.

Biørn (2004) proposed a step-wise maximum likelihood method for estimating the systems of regression equations, where the distributions of error components are specified. We assume that the conditional distribution of $\boldsymbol{\alpha}$ given Z and the marginal distribution of $\boldsymbol{\epsilon}_t$ are as follows:

AS 3

$$\boldsymbol{\alpha}|Z \sim \mathrm{N}\left[\mathrm{E}(\boldsymbol{\alpha}|Z), \Lambda_{\alpha\alpha}\right] \ \ \textit{and}\ \boldsymbol{\epsilon}_t \sim \mathrm{N}\left[0, \Sigma_{\epsilon\epsilon}\right],$$

where $\mathrm{E}(\boldsymbol{\alpha}|Z) = \bar{\pi}\bar{\boldsymbol{z}}$ *could be either Chamberlain's or Mundlak's specification for correlated random effects.*

Thus the tail, $\boldsymbol{a} = \boldsymbol{\alpha} - \mathrm{E}(\boldsymbol{\alpha}|Z) = \boldsymbol{\alpha} - \bar{\pi}\bar{\boldsymbol{z}}$, is distributed normally with conditional mean zero and variance $\Lambda_{\alpha\alpha}$. Given assumption AS 3, we can write the reduced form in (2.2) as

$$\boldsymbol{x}_t = \pi \boldsymbol{z}_t + \bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{a} + \boldsymbol{\epsilon}_t. \tag{2.3}$$

When $d_x = 1$, the assumption that $\boldsymbol{a}$ and $\boldsymbol{\epsilon}_t$ are completely independent of Z can be weakened to allow for non-spherical error components. Baltagi *et al.* (2010) deal with heteroscedasticity in $\boldsymbol{a}$ and serial correlation in the idiosyncratic components and Baltagi *et al.* (2006) allow for heteroscedasticity in $\boldsymbol{a}$ and $\boldsymbol{\epsilon}_t$ but no serial correlation in the idiosyncratic component.

While we do derive control variables when the first-stage error components are non-spherical and $d_x = 1$, for the purpose of exposition, largely we will stick to assumptions AS 3, and estimate the first-stage parameters, $\Theta_1 = \{\pi, \bar{\pi}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}\}$, of the reduced form equations (2.3) by Biørn's step-wise maximum likelihood method. In the supplementary appendix, where we discuss large sample properties of the estimator, we briefly describe Biørn's methodology.

2.1. *Identification of Structural Measures*

Given AS 2, we have $\mathrm{E}(\theta + \zeta_t | X, Z, \boldsymbol{\alpha}) = \mathrm{E}(\theta + \zeta_t | \boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$ given by

$$\mathrm{E}(\theta|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t) + \mathrm{E}(\zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t) = (\boldsymbol{\rho}_{\theta\alpha}\boldsymbol{\alpha} + \boldsymbol{\rho}_{\theta\epsilon}\boldsymbol{\epsilon}_t) + (\boldsymbol{\rho}_{\zeta\alpha}\boldsymbol{\alpha} + \boldsymbol{\rho}_{\zeta\epsilon}\boldsymbol{\epsilon}_t) = \boldsymbol{\rho}_{\alpha}\boldsymbol{\alpha} + \boldsymbol{\rho}_{\epsilon}\boldsymbol{\epsilon}_t, \tag{2.4}$$

where, for example, $\boldsymbol{\rho}_{\theta\alpha}$ and $\boldsymbol{\rho}_{\theta\epsilon}$ respectively are vectors of population regression coefficients of θ on $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}_t$. The two $1 \times d_x$ matrices, $\boldsymbol{\rho}_{\alpha}$ and $\boldsymbol{\rho}_{\epsilon}$, when estimated give us a test of exogeneity of $\boldsymbol{x}_t$.

The above and AS 3 then imply that the conditional expectation of y_t^* given X, Z, and $\boldsymbol{\alpha}$ is given by

$$\begin{aligned}\mathrm{E}(y_t^*|X, Z, \boldsymbol{\alpha}) &= \boldsymbol{x}_t'\boldsymbol{\varphi} + \boldsymbol{\rho}_{\alpha}\boldsymbol{\alpha} + \boldsymbol{\rho}_{\epsilon}\boldsymbol{\epsilon}_t \\ &= \boldsymbol{x}_t'\boldsymbol{\varphi} + \boldsymbol{\rho}_{\alpha}(\bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{a}) + \boldsymbol{\rho}_{\epsilon}\boldsymbol{\epsilon}_t = \mathrm{E}(y_t^*|X, Z, \boldsymbol{a}).\end{aligned} \tag{2.5}$$

Let $\boldsymbol{v}_t = \boldsymbol{\alpha} + \boldsymbol{\epsilon}_t$ be the composite errors of the reduced form equations. In our model, the conditioning variables, $\boldsymbol{\alpha} = \bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{a}$ and $\boldsymbol{\epsilon}_t = \boldsymbol{v}_t - \boldsymbol{\alpha} = \boldsymbol{v}_t - (\bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{a})$, are, however, not identified because the $\boldsymbol{a}$'s are unobserved. It would be possible to estimate the structural parameters if we could integrate out $\boldsymbol{a}$ from $\mathrm{E}(y_t^*|X, Z, \boldsymbol{a})$ in (2.5) with respect to its conditional distribution, $f(\boldsymbol{a}|X, Z)$, and obtain

$$\begin{aligned}&\int \mathrm{E}(y_t^*|X, Z, \boldsymbol{a}) f(\boldsymbol{a}|X, Z) d\boldsymbol{a} \\ &= \boldsymbol{x}_t'\boldsymbol{\varphi} + \boldsymbol{\rho}_{\alpha}\bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{\rho}_{\epsilon}(\boldsymbol{v}_t - \bar{\pi}\bar{\boldsymbol{z}}) + \int (\boldsymbol{\rho}_{\alpha}\boldsymbol{a} - \boldsymbol{\rho}_{\epsilon}\boldsymbol{a}) f(\boldsymbol{a}|X, Z) d\boldsymbol{a} \\ &= \boldsymbol{x}_t'\boldsymbol{\varphi} + \boldsymbol{\rho}_{\alpha}\bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{\rho}_{\epsilon}(\boldsymbol{v}_t - \bar{\pi}\bar{\boldsymbol{z}}) + \boldsymbol{\rho}_{\alpha}\hat{\boldsymbol{a}} - \boldsymbol{\rho}_{\epsilon}\hat{\boldsymbol{a}} \\ &= \boldsymbol{x}_t'\boldsymbol{\varphi} + \boldsymbol{\rho}_{\alpha}\hat{\boldsymbol{\alpha}} + \boldsymbol{\rho}_{\epsilon}\hat{\boldsymbol{\epsilon}}_t = E(y_t^*|X, Z)\end{aligned} \tag{2.6}$$

where, in the second equality, $\hat{\boldsymbol{a}}_i = \mathrm{E}(\boldsymbol{a}_i|X_i, Z_i)$. In the third equality, $\hat{\boldsymbol{\alpha}}_i = \bar{\pi}\bar{\boldsymbol{z}} + \hat{\boldsymbol{a}}_i = \mathrm{E}(\boldsymbol{\alpha}_i|X_i, Z_i)$ and $\hat{\boldsymbol{\epsilon}}_{it} = \boldsymbol{v}_t - \hat{\boldsymbol{\alpha}}_i = \mathrm{E}(\boldsymbol{\epsilon}_{it}|X_i, Z_i)$.

In part (a) of Lemma 1 we show that[1]:

LEMMA 1 *If* $\boldsymbol{x}_t = \pi\boldsymbol{z}_t + \bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{a} + \boldsymbol{\epsilon}_t$, $t \in \{1, \ldots, T\}$, *where* $\boldsymbol{a}$ *and* $\boldsymbol{\epsilon}_t$ *are normally distributed with variances* $\Lambda_{\alpha\alpha}$ *and* $\Sigma_{\epsilon\epsilon}$ *respectively, then*

$$\mathrm{E}(\boldsymbol{a}|X, Z) = \hat{\boldsymbol{a}}(X, Z, \Theta_1) = \mathit{\Omega} \sum_{t=1}^{T} (\boldsymbol{v}_t - \bar{\pi}\bar{\boldsymbol{z}}),$$

where $\mathit{\Omega} = [T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}$ *and* $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi\boldsymbol{z}_t$.

PROOF OF LEMMA 1 Given in appendix A.

[1] In Lemma 1 we also derive the $\mathrm{E}(\boldsymbol{a}|X, Z)$ when a and ϵ_t are both scalar, a is heteroscedastic, and the distribution of ϵ_t is non-spherical.

Since $\hat{\boldsymbol{a}}(X,Z,\Theta_1)$ is a continuous function of Θ_1, and since the first stage consistent estimates, $\hat{\Theta}_1$, converge almost surly to Θ_1, $\hat{\hat{\boldsymbol{a}}}(X_i,Z,\hat{\Theta}_1)$, the estimated value of $\hat{\boldsymbol{a}}(X,Z,\Theta_1)$, converges almost surely to $\hat{\boldsymbol{a}}(X,Z,\Theta_1)$, hence

$$\boldsymbol{\rho}_\alpha\hat{\hat{\boldsymbol{\alpha}}}+\boldsymbol{\rho}_\epsilon\hat{\hat{\boldsymbol{\epsilon}}}_t \overset{a.s.}{\to} \boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t=\mathrm{E}(\theta+\zeta_t|X,Z) \text{ and } \boldsymbol{x}_t'\boldsymbol{\varphi}+\boldsymbol{\rho}_\alpha\hat{\hat{\boldsymbol{\alpha}}}_i+\boldsymbol{\rho}_\epsilon\hat{\hat{\boldsymbol{\epsilon}}}_t \overset{a.s.}{\to} \mathrm{E}(y_t^*|X,Z),$$

where $\hat{\hat{\boldsymbol{\alpha}}}$ and $\hat{\hat{\boldsymbol{\epsilon}}}_t$ are the estimated values of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ respectively.

If the population parameters, Θ_1, were known, we could write y_t^* in error form as

$$y_t^*=\mathrm{E}(y_t^*|X,Z)+\eta_t=\mathbb{X}_t'\Theta_2+\eta_t,$$

where $\mathbb{X}_t=(\boldsymbol{x}_t',\hat{\boldsymbol{\alpha}}',\hat{\boldsymbol{\epsilon}}_t')'$, $\Theta_2=(\boldsymbol{\varphi}',\boldsymbol{\rho}_\alpha',\boldsymbol{\rho}_\epsilon')'$, and $\eta_t=\theta+\zeta_t-\mathrm{E}(\theta+\zeta_t|X,Z)=\theta+\zeta_t-(\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t)$. Equation (2.1) then is written as

$$y_t=1\{\mathbb{X}_t'\Theta_2+\eta_t>0\}. \tag{2.7}$$

Since $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$, both, are of dimension d_x, the dimension of $\mathbb{X}_t$ is $3d_x$[2].

The identification conditions for Θ_2 in (2.7) to be identified when η_t is assumed to follow a known distribution (see Manski, 1988) are: (a) η_t be distributed independently of $\mathbb{X}_t$ and (b) there exists no $A\subseteq\mathbb{R}^{3d_x}$, $\mathbb{X}_t\in\mathbb{R}^{3d_x}$, such that A has probability 1 under $P_{\mathbb{X}}$, where $P_{\mathbb{X}}$ denotes the probability distribution of $\mathbb{X}$, and A is a proper linear subspace of $\mathbb{R}^{3d_x}$. In Lemma 2 we show that condition (b) is satisfied.

Lemma 2 *If (i) $\nexists\ A_x\subseteq\mathbb{R}^{d_x}$ such that $\Pr_{P_x}(A_x)=1$ under P_x, where A_x is a proper linear subspace of $\mathbb{R}^m$; (ii) $\operatorname{rank}(\Pi)=d_x$, where $\Pi=\begin{pmatrix}\pi & \bar{\pi}\end{pmatrix}$; (iii) $\nexists\ A_z\subseteq\mathbb{R}^k$, where $k=\dim((\boldsymbol{z}_t',\bar{\boldsymbol{z}}')')$, such that $\Pr_{P_z}(A_z)=1$ under P_z, where A_z is a proper linear subspace of $\mathbb{R}^k$; and (iv) if the covariance matrices of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\alpha}$ are of full rank, then $\nexists$ an $A\subseteq\mathbb{R}^{3d_x}$, such that A has probability 1 under $P_{\mathbb{X}}$ and A is a proper linear subspace of $\mathbb{R}^{3d_x}$.*

Proof of Lemma 2 Given in appendix A.

Now, conditions (i) to (iii) in Lemma 2 are standard conditions for identification of $\boldsymbol{\varphi}$ in the traditional control function methods, where $\boldsymbol{\alpha}\perp Z$ and the control function is the composite error, $\boldsymbol{v}_t=\boldsymbol{\alpha}+\boldsymbol{\epsilon}_t=\boldsymbol{x}_t-\pi\boldsymbol{z}_t$. Our conditioning variables – which, as we argue in the next section, can be employed as control functions – however, are $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$, and are functions of $\Lambda_{\alpha\alpha}$ and $\Sigma_{\epsilon\epsilon}$. Positive definiteness of $\Lambda_{\alpha\alpha}$ and $\Sigma_{\epsilon\epsilon}$ in condition (iv) helps establish the statement of the Lemma to be true.

Since $\theta+\zeta_t=\mathrm{E}(\theta+\zeta_t|X,Z)+\eta_t=\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t+\eta_t$, for a given $\boldsymbol{x}_t$ we compute the average

[2]Note that we have suppressed $\boldsymbol{w}_t$ in $\mathbb{X}_t$, where $\boldsymbol{w}_t$ is of dimension d_w. So, in fact, the dimension of $\mathbb{X}_t$ is $3d_x+d_w$. Suppressing $\boldsymbol{w}_t$ in $\mathbb{X}_t$, however, results in no loss of generality.

structural function (ASF) as

$$
\begin{aligned}
\mathrm{E}_{\theta+\zeta}(1\{\boldsymbol{x}_t'\boldsymbol{\varphi}+\theta+\zeta_t>0\}) &= \mathrm{E}_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}},\eta}(1\{\boldsymbol{x}_t'\boldsymbol{\varphi}+\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t+\eta_t>0\})\\
&= \mathrm{E}_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}}(\mathrm{E}_{\eta|\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}}(1\{\boldsymbol{x}_t'\boldsymbol{\varphi}+\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t+\eta_t>0\}))\\
&= \mathrm{E}_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}}(\mathrm{E}_{\eta}(1\{\boldsymbol{x}_t'\boldsymbol{\varphi}+\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t+\eta_t>0\}))\\
&= \int \Pr(y_t=1|\boldsymbol{x}_t,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t)dF(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}})=G(\boldsymbol{x}_t), \qquad (2.8)
\end{aligned}
$$

where the second equality follows from the law of iterated expectations and the third equality follows from the assumption that

AS 4 $\eta_t=\theta+\zeta_t-\mathrm{E}(\theta+\zeta_t|X,Z)$ *is independent of* X *and* Z[3].

If η_t is distributed normally with variance σ^2, we can obtain the average partial effect (APE) of a variable, say w, as

$$
\frac{\partial G(\boldsymbol{x}_t)}{\partial w}=\int\frac{\varphi_w}{\sigma}\phi\left(\frac{\boldsymbol{x}_t'\boldsymbol{\varphi}+\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}+\boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t}{\sigma}\right)dF(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}), \qquad (2.9)
$$

where $\frac{\varphi_w}{\sigma}$ is the scaled coefficient of the variable, w, and $\phi(.)$ is the standard normal density function.

Following Lemma 2, since the components of $\boldsymbol{x}_t$ are continuous, with scale and location normalization, Θ_2 can be estimated by semiparametric methods without specifying the distribution of η_t (see Horowitz, 2009, for a review of identification results for semiparametric binary choice models).

2.2. $\mathrm{E}(\boldsymbol{\alpha}_i|X_i,Z_i)=\hat{\boldsymbol{\alpha}}_i$ *and* $\mathrm{E}(\boldsymbol{\epsilon}_{it}|X_i,Z_i)=\hat{\boldsymbol{\epsilon}}_{it}$ *as Control Functions*

Given that we were able to identify the ASF and APE when the structural model was augmented with $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$, we propose that $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$ be used as control functions. Since $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}_t$ are not identified separately, the traditional control function approach assumes that the composite error, $\boldsymbol{\upsilon}_t=\boldsymbol{\alpha}+\boldsymbol{\epsilon}_t$, is independent of Z and that conditional on $\boldsymbol{\upsilon}_t$, $\boldsymbol{x}_t$ is independent of $\theta+\zeta_t$[4]. That is, the heterogeneity, υ_{lt}, where $\upsilon_{l,t}\in(\upsilon_{1,t},\ldots,\upsilon_{d_x,t})\equiv\boldsymbol{\upsilon}_t$, is assumed to be scalar, whereas there could be multiple sources of heterogeneity.

[3]The assumption that the tail, $U-\mathrm{E}(U|W)$, is independent of the conditioning variable, W, has been made elsewhere such as in Chamberlain (1984) for the correlated random effects (CRE) probit model.

[4] In HW's model, 2 time periods are considered and $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2-\boldsymbol{\epsilon}_1$ are the control functions. To identify $\boldsymbol{\epsilon}_t$, HW consider the following triangular system:

$$
\begin{aligned}
&y_t=1\{y_t^*=\boldsymbol{x}_t'\boldsymbol{\varphi}+\boldsymbol{\alpha}+\zeta_t>0\}\\
&\boldsymbol{x}_t=f_0(\boldsymbol{z}_t)+f_1(\boldsymbol{z}_t,\boldsymbol{\alpha})\boldsymbol{\epsilon}_t,
\end{aligned}
$$

where the individual specific unobserved heterogeneity, $\boldsymbol{\alpha}$, is common to all equations and $\boldsymbol{\epsilon}_t\perp(\boldsymbol{z}_t,\boldsymbol{\alpha})$. They impose the normalizations: $\mathrm{E}(\boldsymbol{\epsilon}_t)=0$ and $\mathrm{Var}(\boldsymbol{\epsilon}_t)=1$. This permits them to solve for $\boldsymbol{\epsilon}_t$ as $\boldsymbol{\epsilon}_t=\mathrm{Var}(\boldsymbol{x}_t|\boldsymbol{z}_t)^{-1/2}[\boldsymbol{x}_t-\mathrm{E}(\boldsymbol{x}_t|\boldsymbol{z}_t)]$. The estimates of $\boldsymbol{\epsilon}_t$ are then obtained by estimating $\mathrm{E}(\boldsymbol{x}_t|\boldsymbol{z}_t)$ and $\mathrm{Var}(\boldsymbol{x}_t|\boldsymbol{z}_t)$. Though HW identify structural quantities semiparametrically, these assumptions imply that their model does not nest the model considered in this paper.

We propose $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$ to be employed as control function because the condition, $(\zeta_t, \theta) \perp \boldsymbol{x}_t | \hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}}$, for $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$ to qualify as control function is weaker than assuming $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi \boldsymbol{z}_t = \mathrm{E}(\boldsymbol{\alpha} + \boldsymbol{\epsilon}_t | X, Z)$ as control function. This is because, given that $\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{v}_t - \hat{\boldsymbol{\alpha}}$, there is one-to-one mapping between $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$ and $(\boldsymbol{v}_t, \hat{\boldsymbol{\alpha}})$, and therefore the conditioning σ-algebra, $\sigma(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$, is same as the σ-algebra, $\sigma(\boldsymbol{v}_t, \hat{\boldsymbol{\alpha}})$. Hence, conditioning on $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$ is equivalent to conditioning on $\boldsymbol{v}_t$ and additional individual specific information as summarized by $\hat{\boldsymbol{\alpha}}$. That is, in assuming $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$ as control function one is assuming that no information about (ζ_t, θ) is contained in $\boldsymbol{x}_t$ over and above that contained in $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$, while the same may not be true if one were to assume $\boldsymbol{v}_t = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\epsilon}}_t$ as the only control function.

Assuming $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$ as control functions would be equivalent to assuming the following:

ACF 1 *(a)*

$$\begin{aligned} \zeta_t, \theta | X, Z, \hat{\boldsymbol{\alpha}} &\sim \zeta_t, \theta | V, Z, \hat{\boldsymbol{\alpha}} \\ &\sim \zeta_t, \theta | V, \hat{\boldsymbol{\alpha}}, \end{aligned}$$

where $V \equiv (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T) = X - \pi Z$ *and* $\hat{\boldsymbol{\alpha}} = \mathrm{E}(\boldsymbol{\alpha} | X, Z)$.

(b) $\zeta_t, \theta \perp \boldsymbol{v}_{-t} | \boldsymbol{v}_t, \hat{\boldsymbol{\alpha}}$.

In part (a), the assumption is that the dependence of (θ, ζ_t) on X and Z is completely characterized by V and $\hat{\boldsymbol{\alpha}}$. The assumption in part (b), which has been made for expositional ease, can be dropped. Since $\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{v}_t - \hat{\boldsymbol{\alpha}}$, there is one-to-one mapping between $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$ and $(\boldsymbol{v}_t, \hat{\boldsymbol{\alpha}})$, hence $(\zeta_t, \theta) | (\boldsymbol{v}_t, \hat{\boldsymbol{\alpha}}) \sim (\zeta_t, \theta) | (\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$. Also, AS 1 and assumptions needed to identify $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$, too, would need to hold true.

Remark. When Z is independent of the heterogeneity terms, (ζ_t, θ) and $(\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$, as is commonly assumed, then, given $(\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$, $(\zeta_t, \theta) \perp \boldsymbol{z}_t | \boldsymbol{v}_t$, where $\boldsymbol{v}_t = \boldsymbol{\alpha} + \boldsymbol{\epsilon}_t$. Also, under such independence, $\hat{\boldsymbol{\alpha}} = \Omega \sum_{t=1}^{T} (\boldsymbol{\alpha} + \boldsymbol{\epsilon}_t)$, too, is independent of Z, and we have $(\zeta_t, \theta) \perp \boldsymbol{z}_t | \boldsymbol{v}_t, \hat{\boldsymbol{\alpha}}$. With $\boldsymbol{v}_t$ invertible in both $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$, there exists a one-to-one mapping between $(\boldsymbol{x}_t, \boldsymbol{z}_t, \hat{\boldsymbol{\alpha}})$, $(\boldsymbol{x}_t, \boldsymbol{v}_t, \hat{\boldsymbol{\alpha}})$, and $(\boldsymbol{z}_t, \boldsymbol{v}_t, \hat{\boldsymbol{\alpha}})$, which then also implies that $(\zeta_t, \theta) \perp \boldsymbol{x}_t | \boldsymbol{v}_t, \hat{\boldsymbol{\alpha}}$.

However, when Z is not independent of θ and $\boldsymbol{\alpha}$, as is the case in our model, then the condition, $(\zeta_t, \theta) \perp \boldsymbol{z}_t | \boldsymbol{v}_t$, does not hold in general for the traditional control function method to be applicable. In ACF 1 what we are proposing is that conditional on V and additionally on $\hat{\boldsymbol{\alpha}}$, Z is independent of (ζ_t, θ). Now,

$$\hat{\boldsymbol{\alpha}} = (I - T\Omega)\bar{\pi}\bar{\boldsymbol{z}} + \Omega \sum_{t=1}^{T} (\boldsymbol{\alpha} + \boldsymbol{\epsilon}_t) = \bar{\pi}\bar{\boldsymbol{z}} + \Omega \sum_{t=1}^{T} (\boldsymbol{x}_t - \pi \boldsymbol{z}_t - \bar{\pi}\bar{\boldsymbol{z}}),$$

a function of X and Z, summarizes certain individual specific information. The assumption, $(\zeta_t, \theta) \perp Z | V, \hat{\boldsymbol{\alpha}}$, or equivalently $(\zeta_t, \theta) \perp X | V, \hat{\boldsymbol{\alpha}}$, with $\hat{\boldsymbol{\alpha}}$ as an additional control function, is related to the dependence assumptions in AM, Bester and Hansen (2009) (BH) and HW, where the distribution of unobserved effects depends on the observed variables only through certain function of the observed variables. These functions, as BH argue, may be viewed as sufficient statistic. AM assume that (ζ_t, θ) is independent of X given certain summary statistics such as the mean, $T^{-1}\sum_{t=1}^{T}\boldsymbol{x}_t$, or index functions of summary statistics, while in BH these functions of observed variables are assumed to be unrestricted index functions. In our case, the control function, $(\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}})$, is motivated by the result that under the restrictions in AS 2 and (2.4), the mean of $\theta + \zeta_t$ given (X_i, Z_i) depends on (X_i, Z_i) through $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$.

When $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$ are assumed as control functions, then one need not specify the conditional distribution of $\theta + \zeta_t$ given $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$ and estimate the coefficients, $\boldsymbol{\varphi}$, and measures like the ASF by semiparametric method discussed in BP, which is an extension of the matching estimator of $\boldsymbol{\varphi}$ for the single-index model without endogeneity. Rothe develops a semiparametric maximum likelihood (SML) method for binary response model to handle endogeneity using control functions. These semiparametric methods, however, require that the instruments, $\boldsymbol{z} = \tilde{\boldsymbol{z}}$, be continuous with large support. If the instruments are discrete, the "rank condition" in BP and condition (ii) of Theorem 1 in Rothe[5], necessary for identification, are violated.

We do not pursue semiparametric estimation of binary choice models with the control functions developed in this paper any further. Semiparametric estimation and the large sample properties of the estimates are left for future research.

2.2.1. *Identification of Average Structural Function and Average Partial Effects*

If we assume $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ as control functions, then

$$\begin{aligned}\Pr(y_t = 1|\boldsymbol{x}_t, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t) &= \int 1\{-(\zeta_t+\theta) < \boldsymbol{x}_t'\boldsymbol{\varphi}\} dF(\zeta_t + \theta|\boldsymbol{x}_t, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)\\ &= \int 1\{-(\zeta_t+\theta) < \boldsymbol{x}_t'\boldsymbol{\varphi}\} dF(\zeta_t + \theta|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)\\ &= F(\boldsymbol{x}_t'\boldsymbol{\varphi}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t),\end{aligned}$$

where $F(\boldsymbol{x}_t'\boldsymbol{\varphi}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$ is the conditional CDF of $\zeta_t + \theta$ given $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$ evaluated at $\boldsymbol{x}_t'\boldsymbol{\varphi}$. The second equality follows because $\zeta_t + \theta \perp \boldsymbol{x}_t|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t$.

If conditional on $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$, $\zeta_t + \theta$ is assumed to be distributed normally with variance σ^2 and conditional mean, the regression function, $\boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}} + \boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t$, then

$$F(\boldsymbol{x}_t'\boldsymbol{\varphi}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t) = \Phi\left(\frac{\boldsymbol{x}_t'\boldsymbol{\varphi} + \boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}} + \boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_t}{\sigma}\right).$$

Given $\boldsymbol{x}_t$, averaging $F(\boldsymbol{x}_t'\boldsymbol{\varphi}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$ over $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$, we get the ASF:

$$\begin{aligned}G(\boldsymbol{x}_t) &= \int F(\boldsymbol{x}_t'\boldsymbol{\varphi}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t) dF(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}),\\ &= \int\left[\int 1\{\boldsymbol{x}_t'\boldsymbol{\varphi} + \theta + \zeta_t > 0\} dF(\theta + \zeta|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})\right] dF(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})\\ &= \mathrm{E}_{\theta+\zeta}(1\{\boldsymbol{x}_t'\boldsymbol{\varphi} + \theta + \zeta_t > 0\}).\end{aligned} \tag{2.10}$$

[5] Since conditioning on $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$ is equivalent to conditioning on $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi\boldsymbol{z}_t$ and $\hat{\boldsymbol{\alpha}}$, and if identification requires that conditional on the control variables, $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$, – hence, $\boldsymbol{v}_t$ and $\hat{\boldsymbol{\alpha}}$ – the vector $\boldsymbol{x}_t$ contains at least one, x_t^1, continuously distributed component with nonzero coefficient, then it would be necessary that $\boldsymbol{z}_t$ contains a continuously distributed regressor.

The APE of changing a variable, say w, from w_t to $w_t + \Delta_w$ can be obtained as

$$\frac{\Delta G(\boldsymbol{x}_t)}{\Delta_w} = \frac{G(\boldsymbol{x}_{t_{-w}}, (w_t + \Delta_w)) - G(\boldsymbol{x}_t)}{\Delta_w}. \tag{2.11}$$

To point-identify the ASF, $G(\bar{\boldsymbol{x}})$, it is required that $\mathrm{E}(y_t|\boldsymbol{x}_t = \bar{\boldsymbol{x}}, \hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t = \bar{\boldsymbol{\epsilon}})$ be evaluated at all values of $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\epsilon}})$ in the support of the unconditional distribution of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})$. This requires that the support of the conditional distribution of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})$ conditional on $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$ be equal to the support of the unconditional distribution (see Imbens and Newey; Florens *et al.*; Blundell and Powell for a discussion). For many triangular systems that employ the control function, $\boldsymbol{v}_{it}$, or $(F(x_{1,it}|\boldsymbol{z}_{it}) \ldots F(x_{d_x,it}|\boldsymbol{z}_{it}))$, where $F(x_{1,it}|\boldsymbol{z}_{it})$ is CDF of $x_{1,it}$ given $\boldsymbol{z}_{it}$, the requirement of common support necessitates that the set of instruments, $\boldsymbol{z}_{it}$, contains a continuous instrument with large support. In lemma 3 we show that:

Lemma 3 *The support of the conditional distribution of $\hat{\boldsymbol{\alpha}}(X, Z, \Theta_1)$ and $\hat{\boldsymbol{\epsilon}}_t(X, Z, \Theta_1)$, conditional on $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$, is same as the support of their marginal distribution.*

Proof of Lemma 3 Given in appendix A

In our approach, the control functions, $\hat{\boldsymbol{\epsilon}}_t$ and $\hat{\boldsymbol{\alpha}}$, are smooth, unbounded functions of $\boldsymbol{x}_t$'s, $t \in \{1, \ldots, T\}$. Therefore, because the $\boldsymbol{x}_s$'s, $s \neq t$, are unrestricted and are continuous with large supports, the ranges of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \pi \boldsymbol{z}_t - \hat{\boldsymbol{\alpha}}$ conditional on $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$ do not depend on $\boldsymbol{x}_t$. Since the result does not rely on any kind of restriction on $\boldsymbol{z}_t$'s support, our method circumvents the need to have a continuous instrument with large support to identify the ASF. In the absence of continuous instruments with large support, this result would also be useful for computing the ASF for the kind of triangular setups considered in Blundell and Powell (2003), where the errors in the structural equation are nonadditive, but additively separable in the reduced form.

2.3. *Estimation Of Probit Conditional Mean Function*

If η_t in (2.7) is assumed to follow a normal distribution, then

$$\mathrm{E}(y_t|X, Z) = \Phi((\mathbb{X}_t' \Theta_2)/\sigma),$$

where $\mathbb{X}_t = (\boldsymbol{x}_t', \hat{\boldsymbol{\alpha}}'(X, Z), \hat{\boldsymbol{\epsilon}}_t'(X, Z))$, $\Theta_2 = (\boldsymbol{\varphi}', \boldsymbol{\rho}_\alpha', \boldsymbol{\rho}_\epsilon')'$, and σ^2 is the variance of η_t. Since in probit models the coefficients can only be identified up to a scale, in this section with a slight abuse of notation we denote the scaled parameters, $\frac{1}{\sigma}\Theta_2$, by Θ_2. To estimate Θ_2, one can employ nonlinear least squares by pooling the data. However, as PW discuss, since $\mathrm{Var}(y_t|(X, Z))$ will most likely be heteroscedastic and since there will be serial correlation across time in the joint distribution, $F(y_0, \ldots, y_T|X, Z)$, the estimates, though consistent, will be estimated inefficiently resulting in biased standard errors. PW argue that modelling $F(y_0, \ldots, y_T|X, Z)$ and applying MLE methods, while possible, is not trivial. Moreover, if the model for $F(y_0, \ldots, y_T|X, Z)$ is misspecified but $\mathrm{E}(y_t|X, Z)$ is correctly specified, the MLE will be inconsistent for Θ_2 and the resulting APEs.

To account for heteroscedasticity and serial dependence when all covariates are exogenous, PW employ the method of multivariate weighted nonlinear least squares (MWNLS) to obtain efficient estimates of Θ_2. To get the correct estimates of the standard errors, the method requires is a parametric model of $\text{Var}(y_i|X_i, Z_i)$, where y_i is the $T \times 1$ vector of responses. Similar to PW's, we specify $\text{Var}(y_t|X, Z)$ as

$$\text{Var}(y_t|X, Z) = \tau \mathbf{m}(\mathbb{X}_t, \Theta_2)(1 - \mathbf{m}(\mathbb{X}_t, \Theta_2)), \tag{2.12}$$

where $\mathbf{m}(\mathbb{X}_t, \Theta_2) = \Phi(\mathbb{X}_t'\Theta_2)$ and $0 < \tau \leq 1$. For covariances, $\text{Cov}(y_t, y_r|X, Z)$, a "working" version, which can be misspecified for $\text{Var}(y|X, Z)$, is assumed. This, in the context of panel data, is what underlies the method of generalized estimating equation (GEE), as described in Liang and Zeger (1986). The main advantage of GEE lies in the consistent and unbiased estimation of parameters' standard errors even when the correlation structure is misspecified. Also, GEE and MWNLS are asymptotically equivalent whenever they use the same estimates of the $T \times T$ positive definite matrix, $\text{Var}(y|X, Z)$.

Generally, the conditional correlations, $\text{Cov}(y_t, y_s|X, Z)$, are a function of X and Z. In the GEE literature, the "working correlation matrix" is that which assumes the dependency structure to be invariant over all observations; that is, the correlations are not a function of X and Z. Here we will focus on a particular correlation matrix that is suited for panel data applications with small T . In the GEE literature it is called an "exchangeable" correlation pattern. Exchangeable correlation assumes constant time dependency, so that all the off-diagonal elements of the correlation matrix are equal. Though other correlation patterns such as "autoregressive", which assumes the correlations to be an exponential function of the time lag, or "stationary M", which assumes constant correlations within equal time intervals could also be assumed.

GEE method suggests that parameter, ρ, that characterize $\text{Var}(y|X, Z) = \mathbf{V}(X, Z, \Theta_2, \tau, \rho)$ can be estimated using simple functions of residuals, u_t,

$$u_t = y_t - \text{E}(y_t|X, Z) = y_t - \mathbf{m}(\mathbb{X}_t, \Theta_2),$$

where the mean function , $\text{E}(y_t|X, Z)$, is correctly specified. With the variance defined in (2.12), we can define standardized errors as

$$e_t = \frac{u_t}{\mathbf{m}(\mathbb{X}_t, \Theta_2)(1 - \mathbf{m}(\mathbb{X}_t, \Theta_2))}.$$

Then we have $\text{Var}(e_t|X, Z) = \tau$. The exchangeability assumption is that the pairwise correlations between pairs of standardized errors are constant, say ρ. This, to reiterate, is a "working" assumption that leads to an estimated variance matrix to be used in MWNLS. Neither consistency of the estimator of ρ, nor valid inference, will rest on exchangeability being true.

To estimate a common correlation parameter, let $\tilde{\Theta}_2$ be a preliminary, consistent estimator of Θ_2. $\tilde{\Theta}_2$ could be the pooled ML estimate of the heteroscedastic probit model. Define the residuals, $\tilde{u}_t$, as $\tilde{u}_t = y_t - \mathbf{m}(\mathbb{X}_t, \tilde{\Theta}_2)$ and the standardized residuals as

$$\tilde{e}_t = \frac{\tilde{u}_t}{\mathbf{m}(\mathbb{X}_t, \tilde{\Theta}_2)(1 - \mathbf{m}(\mathbb{X}_t, \tilde{\Theta}_2))}.$$

Then a natural estimator of a common correlation coefficient is

$$\tilde{\rho} = \frac{1}{NT(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s \neq t} \tilde{e}_{it} \tilde{e}_{is}. \tag{2.13}$$

Under standard regularity conditions, without any substantive restrictions on $\text{Corr}(e_t, e_s|X, Z)$, the plim of $\tilde{\rho}$ is

$$\text{plim}(\tilde{\rho}) = \frac{1}{[T(T-1)]} \sum_{t=1}^{T} \sum_{s \neq t} \text{E}(e_{it} e_{is}) \equiv \rho^*$$

If $\text{Corr}(e_t, e_s|X, Z)$ happens to be the same for all $t \neq s$, then $\tilde{\rho}$ consistently estimates this constant correlation. Generally, it consistently estimates the average of these correlations across all (t, s) pairs, which is defined as $\mathbf{C}(\tilde{\rho})$. Given the estimated $T \times T$ working correlation matrix, $\mathbf{C}(\tilde{\rho})$, which has unity down its diagonal and $\tilde{\rho}$ everywhere else, we can construct the estimated working variance matrix:

$$\mathbf{V}(X, Z, \tilde{\Theta}_2, \tilde{\rho}) = \mathbf{D}(X, Z, \tilde{\Theta}_2)^{1/2} \mathbf{C}(\tilde{\rho}) \mathbf{D}(X, Z, \tilde{\Theta}_2)^{1/2} = \mathbf{V}(X, Z, \tilde{\Upsilon})$$

where $\mathbf{D}(X, Z, \Theta_2)$ is the $T \times T$ diagonal matrix with $\mathbf{m}(\mathbb{X}_t, \Theta_2)(1 - \mathbf{m}(\mathbb{X}_t, \Theta_2))$ down its diagonal. (Note that dropping the variance scale factor, τ , has no effect on estimation or inference.)

Estimation by MWNLS then involves solving for $\hat{\Theta}_2$ by minimizing the following with respect to Θ_2:

$$\min_{\Theta_2} \sum_{i=1}^{N} [\mathbf{y}_i - \mathbf{m}_i(X_i, Z_i, \Theta_2)]' [\mathbf{V}(X_i, Z_i, \tilde{\Upsilon})]^{-1} [\mathbf{y}_i - \mathbf{m}_i(X_i, Z_i, \Theta_2)], \tag{2.14}$$

where $\mathbf{m}_i(X_i, Z_i, \Theta_2)$ is the T vector with t^{th} element $\mathbf{m}(\mathbb{X}_{it}, \Theta_2)$.

The requirement of GEE is that the mean model, $\text{E}(y_t|X, Z)$, be correctly specified, else the GEE approach to estimation can give inconsistent results. We have, given our identifying assumptions, shown that $\text{E}(y_t|X, Z) = \Phi(\mathbb{X}_t' \Theta_2)$, and therefore we can employ GEE to account for serial correlation across time. Once the control functions have been estimated, one can then use the STATA command,"xtgee," which fits generalized linear models and allows one to specify the within-group correlation structure for the panels, to estimate Θ_2.

Once we have the consistent estimates, $\hat{\Theta}_2$, of Θ_2, the sample analog of the APE of a variable, say w, $\frac{\partial G(\boldsymbol{x}_t)}{\partial w}$ for any fixed $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$ can be computed as

$$\widehat{\frac{\partial G(\boldsymbol{x}_t)}{\partial w}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\varphi}_w \phi(\bar{\boldsymbol{x}}' \hat{\boldsymbol{\varphi}} + \hat{\rho}_\alpha \hat{\hat{\boldsymbol{\alpha}}}_i + \hat{\rho}_\epsilon \hat{\hat{\boldsymbol{\epsilon}}}_{it}). \tag{2.15}$$

Since

$$\hat{\varphi}_w \phi(\bar{\boldsymbol{x}}' \hat{\boldsymbol{\varphi}} + \hat{\rho}_\alpha \hat{\hat{\boldsymbol{\alpha}}}_i + \hat{\rho}_\epsilon \hat{\hat{\boldsymbol{\epsilon}}}_{it}) \overset{a.s.}{\to} \varphi_w \phi(\bar{\boldsymbol{x}}' \boldsymbol{\varphi} + \boldsymbol{\rho}_\alpha \hat{\boldsymbol{\alpha}}_i + \boldsymbol{\rho}_\epsilon \hat{\boldsymbol{\epsilon}}_{it}),$$

by the weak LLN $\widehat{\frac{\partial G(\boldsymbol{x}_t)}{\partial w}}$ converges in probability to $\frac{\partial G(\boldsymbol{x}_t)}{\partial w}$ as $NT \to \infty$. The APE of an exogenous dummy variable, w, can be computed by taking the difference in sample analogue of the ASF, $\widehat{G(\boldsymbol{x}_t)}$, computed at $(0, \bar{\boldsymbol{x}}'_{-w})'$ and $(1, \bar{\boldsymbol{x}}'_{-w})'$.

In appendix A of the supplementary appendix we derive the asymptotic covariance matrix of the second-stage coefficient estimates and the standard errors of the APEs when the first stage estimation involves estimating a system of regression using the method in Biørn (2004). However, firstly, because the expressions needed to compute the covariance matrices might be computationally involved, and secondly, because new expressions for the covariance matrix of the second-stage coefficient estimates will have to be derived when a different estimator for the first stage reduced form is employed, we suggest that bootstrapping procedure be employed to approximate the variance of the estimated coefficient[6].

3. MONTE CARLO EXPERIMENTS

In this section, we discuss results of Monte Carlo (MC) experiments, which we conduct to analyze the finite sample behaviour of our model with one endogenous variable, x. We also compare the performance of our estimator to the performances of alternative estimators with setups similar to ours. More specifically, we compare the estimates of APE of x from ours and alternative estimators to the true measure of the APE.

We consider the following data generating process (DGP):

$$y_{it} = 1\{\varphi x_{it} + \theta_i + \zeta_{it} > 0\} \text{ and 0 otherwise, where} \tag{3.1}$$

$$x_{it} = \pi z_{it} + \alpha_i + \epsilon_{it}, i = 1, \ldots, n, t = 1, \ldots, 5, \tag{3.2}$$

and where z_{it} is the instrument. We assume that $\varphi = -1$ and that $\pi = 1.5$. We allow the individual specific effects α_i and θ_i to be correlated with the vector of instruments, $Z_i = (z_{i1}, \ldots, z_{i5})'$. The variables, Z_i, α_i, and θ_i, are drawn from the following distribution: $(Z_i', \alpha_i, \theta_i)' \sim \mathrm{N}\left[0, \Sigma_{z\alpha\theta}\right]$, where

$$\Sigma_{z\alpha\theta} = \begin{bmatrix} \begin{bmatrix} \sigma_z^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_z^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_z^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_z^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_z^2 \end{bmatrix} & \begin{bmatrix} \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\theta}\sigma_z\sigma_\theta \\ \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\theta}\sigma_z\sigma_\theta \\ \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\theta}\sigma_z\sigma_\theta \\ \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\theta}\sigma_z\sigma_\theta \\ \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\theta}\sigma_z\sigma_\theta \end{bmatrix} \\ \begin{bmatrix} \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\alpha}\sigma_z\sigma_\alpha & \rho_{z\alpha}\sigma_z\sigma_\alpha \\ \rho_{z\theta}\sigma_z\sigma_\theta & \rho_{z\theta}\sigma_z\sigma_\theta & \rho_{z\theta}\sigma_z\sigma_\theta & \rho_{z\theta}\sigma_z\sigma_\theta & \rho_{z\theta}\sigma_z\sigma_\theta \end{bmatrix} & \begin{bmatrix} \sigma_\alpha^2 & \rho_{\alpha\theta}\sigma_\theta\sigma_\alpha \\ \rho_{\alpha\theta}\sigma_\theta\sigma_\alpha & \sigma_\theta^2 \end{bmatrix} \end{bmatrix},$$

where $\sigma_z = 5$, $\sigma_\alpha = 3$, $\sigma_\theta = 4$, $\rho_{z\alpha} = 0.4$, $\rho_{z\theta} = 0.2$, and $\rho_{\alpha\theta} = 0.5$. The above choice of correlation coefficients ensures that, conditional on α_i, the conditional correlation between z_{it} and θ_i, $\rho_{z\theta|\alpha} = \rho_{z\theta} - \rho_{z\alpha}\rho_{\alpha\theta} = 0$, which, in this case, also implies that conditional on α_i, $\theta_i \perp Z_i|\alpha_i$.

[6]STATA's command, xtgee, for generalized estimating equations has the option for bootstrap estimation of the standard errors, where one can define cluster variables for identifying panels.

Also, $(\zeta_{it}, \epsilon_{it})$ is assumed independent of $(Z_i', \alpha_i, \theta_i)'$, and is drawn from $(\zeta_{it}, \epsilon_{it}) \sim \mathrm{N}\left[0, \Sigma_{\zeta\epsilon}\right]$, where

$$\Sigma_{\zeta\epsilon} = \begin{bmatrix} \sigma_\zeta^2 & \rho_{\zeta\epsilon}\sigma_\zeta\sigma_\epsilon \\ \rho_{\zeta\epsilon}\sigma_\zeta\sigma_\epsilon & \sigma_\epsilon^2 \end{bmatrix} = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix}.$$

From this DGP we generate $(Z_i', \alpha_i, \theta_i)'$ and $(\zeta_{it}, \epsilon_{it})$ of varying size, n, with t fixed at $t = 5$. When generating $(Z_i', \alpha_i, \theta_i)'$, we first generate n number of α_i's, and then given α_i, we generate θ_i and $Z_i = (z_{i1}, \ldots, z_{i5})'$. We then discretized z_{it} to take value 1 if $z_{it} > 0$ and 0 otherwise. Having generated $(Z_i', \alpha_i, \theta_i)'$ and $(\zeta_{it}, \epsilon_{it})$, we generate x_{it} according to (3.2) and then y_{it} according to (3.1).

Our DGP assumptions, $\theta_i \perp Z_i|\alpha_i$ and $(Z_i', \alpha_i, \theta_i)' \perp (\zeta_{it}, \epsilon_{it})$, which are in line with AS 1, together satisfy AS 2[7], which in turn implies that $\theta_i + \zeta_{it} \perp x_{it}|\alpha_i, \epsilon_{it}$. This then implies that we can write (3.1) as

$$y_{it} = 1\{\varphi x_{it} + \rho_\alpha \alpha_i + \rho_\epsilon \epsilon_{it} + \tilde{\eta} > 0\}, \tag{3.3}$$

where, given the parameter values of the DGP, $\rho_\alpha = \rho_{\theta\alpha}\frac{\sigma_\theta}{\sigma_\alpha} = 0.6667$, $\rho_\epsilon = \rho_{\zeta\epsilon}\frac{\sigma_\zeta}{\sigma_\epsilon} = 0.75$, and $\tilde{\sigma}^2$, the variance of $\tilde{\eta}$, is $\tilde{\sigma}^2 = (1 - \rho_{\theta\alpha}^2)\sigma_\theta^2 + (1 - \rho_{\zeta\epsilon}^2)\sigma_\zeta^2 = 12.4375$.

In practice, since α_i and ϵ_{it} are unobserved and cannot be identified, we cannot estimate (3.3) as a probit model to obtain $\Pr(y_{it} = 1|x_{it}, \alpha_i, \epsilon_{it})$ and the scaled coefficient, $\varphi/\tilde{\sigma}$. Given this, we in section 2 developed a two-step method to consistently estimate the ASF and APE. We showed that by estimating

$$y_{it} = 1\{\varphi x_{it} + \rho_\alpha \hat{\alpha}_i + \rho_\epsilon \hat{\epsilon}_{it} + \eta_{it} > 0\} \tag{3.4}$$

as a probit model we can obtain estimates of the scaled parameter φ/σ, where σ^2 is the variance of η_{it}, which is independent of X_i and Z_i, and where $\rho_\alpha\hat{\alpha}_i + \rho_\epsilon\hat{\epsilon}_{it} = \mathrm{E}(\theta_i + \zeta_{it}|X_i, Z_i)$.

Having generated the data, the control variables, $\hat{\alpha}_i$ and $\hat{\epsilon}_{it}$, are constructed (see section 2.1) from the first-step estimates of the following modified reduced form equation:

$$x_{it} = \pi z_{it} + \bar{\pi}\bar{z}_i + a_i + \epsilon_{it}, \tag{3.5}$$

where $\bar{\pi}\bar{z}_i + a_i$ is equal to α_i of reduced form equation (3.2), where $\bar{\pi}\bar{z}_i = \bar{\pi}T^{-1}\sum_{i=1}^{T} z_{it}$ is the assumed specification for $\mathrm{E}(\alpha_i|Z_i)$ and a_i is the residual individual effect. Equation (3.5) is estimated as a panel data random effect model by the method of MLE.

But $\tilde{\sigma}^2$, the variance of $\tilde{\eta}_{it}$ in (3.3), is not the same as σ^2, the variance of η_{it} in (3.4). Therefore, $\varphi/\tilde{\sigma}$ is not equal to φ/σ, where the estimate of the latter is obtained by estimating (3.4) as a probit model. While, given the specified DGP, it could be possible to compute the value of σ and compare the estimates of φ/σ to its true value, computing the value of σ could be tedious.

[7]Given the DGP assumptions, it can be verified that $\zeta_{it} \perp Z_i|\alpha_i, \epsilon_{it}, \theta_i$ and $\theta_i \perp Z_i|\alpha_i, \epsilon_{it}$. The two together imply that $\theta_i, \zeta_{it} \perp Z_i|\alpha_i, \epsilon_{it}$.

However,

$$\int \Pr(y_{it} = 1|x_{it}, \alpha_i, \epsilon_{it})dF(\alpha, \epsilon) = \int \Pr(y_{it} = 1|x_{it}, \hat{\alpha}_i, \hat{\epsilon}_{it})dF(\hat{\alpha}, \hat{\epsilon}), \tag{3.6}$$

where the measure on the RHS, whether one follows the method outlined in section 2.1 or assume $(\hat{\alpha}_i, \hat{\epsilon}_{it})$ as control function, was shown equal to $\mathrm{E}_{\theta+\zeta}(1\{x_{it}\varphi + \theta_i + \zeta_{it} > 0\})$ (see equations (2.8) and (2.10)), and was referred to as the ASF evaluated at x_{it}. The equality in (3.6) holds because the LHS in (3.6) too is $\mathrm{E}_{\theta+\zeta}(1\{x_{it}\varphi + \theta_i + \zeta_{it} > 0\})$[8].

Since $\tilde{\eta}_{it}$ in (3.3) and η_{it} in (3.4) are normally distributed, differentiating throughout with respect to x in (3.6) we get

$$\int \frac{\varphi}{\tilde{\sigma}}\phi\left(\frac{\varphi x_{it} + \rho_\alpha \alpha_i + \rho_\epsilon \epsilon_{it}}{\tilde{\sigma}}\right)dF(\alpha, \epsilon) = \int \frac{\varphi}{\sigma}\phi\left(\frac{\varphi x_{it} + \rho_\alpha \hat{\alpha}_i + \rho_\epsilon \hat{\epsilon}_{it}}{\sigma}\right)dF(\hat{\alpha}, \hat{\epsilon}), \tag{3.7}$$

where the measure on the RHS in section 2.1 was referred to as APE of x at x_{it}.

For notational convenience, in this section, we will denote the true ASF, $\int \Pr(y_{it} = 1|x_{it}, \alpha_i, \epsilon_{it})dF(\alpha, \epsilon)$, by $G(x_{it})$, and the true APE, the APE on the LHS of (3.7), by $\frac{\partial G(x_{it})}{\partial x}$. Estimates of APE, for example the estimate of RHS in (3.7), from any of the model considered in this section will be denoted by $\widehat{\frac{\partial G(x_{it})}{\partial x}}$.

Now, while in practice the heterogeneity terms, (θ_i, ζ_{it}) and $(\alpha_i, \epsilon_{it})$, are unobserved, in MC experiments we do know what these values are. We could therefore compute the true measure of APE, $\frac{\partial G(x_{it})}{\partial x}$, by averaging $(\varphi/\tilde{\sigma})\phi((\varphi x_{it} + \rho_\alpha \alpha_i + \rho_\epsilon \epsilon_{it})/\tilde{\sigma})$ over α_i and ϵ_{it}. At $x_{it} = 1$, given the parameter values, it turns out that $\frac{\partial G(x_{it})}{\partial x} = -.09396$[9].

One of the alternative estimators, which has its setup similar to ours is the method proposed by PW. To address the issue of endogeneity, PW also propose a two-step control function method. In their model, they specify the conditional distribution of θ_i given X_i and the conditional distribution of α_i given Z_i. They assume that $\theta_i = \bar{\pi}_\theta \bar{x}_i + \tau_i$ and $\alpha_i = \bar{\pi}_\alpha \bar{z}_i + a_i$, where $\bar{\pi}_\theta \bar{x}_i$ is the specification for $\mathrm{E}(\theta_i|X_i)$ and $\bar{\pi}_\alpha \bar{z}_i$ is the specification for $\mathrm{E}(\alpha_i|Z_i)$. Given the assumptions, they write the triangular system in (3.1) and (3.2) as

$$y_{it} = 1\{\varphi x_{it} + \bar{\pi}_\theta \bar{x}_i + \tau_i + \zeta_{it} > 0\} \tag{3.8}$$

$$x_{it} = \pi z_{it} + \bar{\pi}_\alpha \bar{z}_i + \upsilon_{PWit}, \tag{3.9}$$

[8]Now

$$\int \Pr(y_{it} = 1|x_{it}, \alpha_i, \epsilon_{it})dF(\alpha, \epsilon) = \int \left[\int 1\{x_{it}\varphi + \theta_i + \zeta_{it} > 0\}dF(\theta + \zeta|\alpha, \epsilon)\right]dF(\alpha, \epsilon)$$
$$= \mathrm{E}_{\theta+\zeta}(1\{x_{it}\varphi + \theta_i + \zeta_{it} > 0\}).$$

[9]While the value of $\frac{\partial G(x_{it})}{\partial x}$ could be computed analytically, we compute its value numerically for every MC replication. Its value differs from $-.09396$ only at the 6^{th} decimal place as we vary the sample size.

where $\upsilon_{PWit} = a_i + \epsilon_{it}$. They then make the control function assumption that $\tau_i + \zeta_{it} \perp x_{it}|\upsilon_{PWit}$. This allows them to estimate the APE as

$$\int \frac{\varphi}{\sigma_{PW}}\phi\left(\frac{\varphi x_{it} + \bar{\pi}_\theta \bar{x}_i + \rho\upsilon_{PWit}}{\sigma_{PW}}\right)dF(\bar{\pi}_\theta\bar{x}, \upsilon_{PW}),$$

where ρ is population regression coefficient of $\tau_i + \zeta_{it}$ on υ_{PWit}, and where υ_{PWit} is obtained as residuals after estimating (3.9) in the first stage. The conditional distribution of $\tau_i + \zeta_{it}$ given υ_{PWit} is assumed to follow a normal distribution with variance σ^2_{PW}. If their method gives consistent estimates of APE, then it must be that the above measure is equal to $\frac{G(x_{it})}{\partial x}$.

In Chamberlain's correlated random effects (CRE) probit and in Chamberlain's conditional logit (CL), x_{it} is assumed to be independent of the idiosyncratic term, ζ_{it}. While in CRE probit model $\mathrm{E}(\theta_i|X_i)$ is specified, in the CL model the distribution of θ_i is left unspecified. Assuming that $\theta_i = \bar{\pi}_\theta\bar{x}_i + \tau_i$, where $\bar{\pi}_\theta\bar{x}_i$ is the specification for $\mathrm{E}(\theta_i|X_i)$, the structural equation for the CRE probit model is given by

$$y_{it} = 1\{\varphi x_{it} + \bar{\pi}_\theta\bar{x}_i + \tau_i + \zeta_{it} > 0\}, \text{ where } \tau_i = \theta_i - \mathrm{E}(\theta_i|X_i).$$

$\tau_i + \zeta_{it}$ is assumed independent of X_i and is distributed normally with variance σ^2_{CRE}. The CRE probit model is estimated as a probit model by pooling the data. If the CRE probit model too gives consistent measure of APE then it has to be that

$$\int \frac{\varphi}{\sigma_{CRE}}\phi\left(\frac{\varphi x_{it} + \bar{\pi}_\theta\bar{x}_i}{\sigma_{CRE}}\right)dF(\bar{\pi}_\theta\bar{x}) = \frac{\partial G(x_{it})}{\partial x},$$

where the LHS is the measure of APE of x pertaining to the CRE probit model.

The structural equation for the CL model is same as equation (3.1), where ζ_{it} follows a logistic distribution. The APE of x at x_{it} for the CL model is

$$\int \varphi\Lambda(x_{it}, \theta_i)(1 - \Lambda(x_{it}, \theta_i))dF(\theta),$$

where $\Lambda(x_{it}, \theta_i) = \Pr(y_{it} = 1|x_{it}, \theta_i) = \frac{\exp(\varphi x_{it} + \theta_i)}{1+\exp(\varphi x_{it} + \theta_i)}$. Once we have estimated φ by estimating the CL model, we can estimate the APE by averaging $\varphi\Lambda(x_{it}, \theta_i)(1 - \Lambda(x_{it}, \theta_i))$ over θ_i. Again, while in practice θ_i is unobservable, in MC experiments we know what these values are.

Table 1 provides the results for various sample size, n, with $m = 10000$ Monte Carlo replications. In the Table and in Figure 1 we compare the performance of our method, which we call EAP method[10], to the alternative estimators considered above.

In Figure 1 we plot the density of $m = 10000$ MC estimates of $\widehat{\partial G(x_{it})}/\partial x - \partial G(x_{it})/\partial x$ obtained for the four estimation methods for different sample sizes. It can be seen from

[10]We call it so because, as can be seen in Appendix 1, our conditioning variables to account for endogeneity are based on $\mathrm{E}(\boldsymbol{\alpha}|X, Z)$, which is computed as "expected a posteriori" (EAP) value of $\boldsymbol{\alpha}$, and where the required distributions to compute it are obtained from the first stage estimates.

the figure that as the sample size increases, the variance of $\widehat{\partial G(x_{it})}/\partial x - \partial G(x_{it})/\partial x$ for each of the method decreases. However, the figure also shows that for each of the alternative estimators, the APE of x is estimated with a bias, which persists as the sample size grows larger. Thus, even as the variance of $\widehat{\partial G(x_{it})}/\partial x - \partial G(x_{it})/\partial x$ for each of the alternative methods decreases, as can be seen from Table 1, the RMSE for alternative methods decreases quite slowly.

[**Table 1 about here**]

[**Figure 1 about here**]

Between the alternative estimators, the conditional logit model performs better than the CRE probit model and the control function method proposed by PW. HW too compare their estimator to the CL model by MC experiments. However, in their MC experiments, x_{it} is not correlated with transitory errors, ζ_{it}, whereas in our DGP the two are. Therefore in our experiments we find that the APE of x from the CL model is estimated with a bias. Even when the transitory error distribution is misspecified, the least bias of the estimates from the CL model among the alternative estimators could be because the correlation between x_{it} and θ_i is left unspecified. The misspecification of transitory error distribution, however, as HW argue, should not have a major effect as the logistic distribution is also a unimodal symmetric error distribution like the true error, which is normally distributed. But it has the largest variance, which could be because the CL model does not use all observations: those individuals for whom $\sum_{t=1}^{5} y_{it} = 0$ or $\sum_{t=1}^{5} y_{it} = 5$ are dropped.

Since the CL and CRE probit models do not account for the correlation between x_{it} and the transitory errors, ζ_{it}, the methods can give biased results. Unexpectedly, however, the method proposed by PW, which tries to accounts for the correlation of x_{it} with both θ_i and ζ_{it}, gives the least satisfactory results. This suggests that the assumptions of their model including the control function assumption, $\tau_i + \zeta_{it} \perp x_{it} | \upsilon_{PWit}$, may be too restrictive and are likely to be violated. Besides, in PW's model the APE's are not point identified when the instrument, z_{it}, is binary and υ_{PWit} the control function.

The results therefore imply that the additional assumption, AS 4, in section 2.1 or assuming $(\hat{\boldsymbol{\epsilon}}_{it}, \hat{\boldsymbol{\alpha}}_i)$ as control function, which allowed us to identity the ASF and APE, may not be restrictive, and that the developed method can yield consistent result.

To conclude, this finite sample study establishes the following:

(1) Our method performs well with sample sizes frequently encountered in practice.

(2) It performs better than the alternative estimators with setups similar to ours.

4. PANEL PROBIT MODEL WITH RANDOM COEFFICIENTS IN A TRIANGULAR SYSTEM

In this section we extend the model with random effect studied in section 2 to allow for random coefficients, and discuss identification of certain structural measures of interest. Consider the following binary choice random coefficient model

$$y_{it} = 1\{y_{it}^* = \mathcal{X}_{it}'\boldsymbol{\varphi}_i + \zeta_{it} > 0\}, \tag{4.1}$$

where $\mathcal{X}_{it} = (x_{it}, \boldsymbol{w}_{it}')'$ and ζ_{it} are the idiosyncratic errors. Here we consider a single continuous endogenous variable, x_{it}, with a large support. Let d_w be the dimension of the exogenous variables, $\boldsymbol{w}_{it}$. In (4.1), the random coefficients are $\boldsymbol{\varphi}_i = \boldsymbol{\varphi} + \boldsymbol{\theta}_i$, where $\boldsymbol{\varphi}$ is a $(1+d_w)\times 1$ vector of constant means and $\mathrm{E}(\boldsymbol{\theta}_i) = 0$. Thus, $\boldsymbol{\varphi}$ is the average slope that we might be interested in.

The reduced form in the triangular system is given by:

$$x_{it} = \boldsymbol{z}_{it}'\boldsymbol{\alpha}_i + \epsilon_{it}, \tag{4.2}$$

where $\boldsymbol{z}_{it} = (\boldsymbol{w}_{it}', \tilde{\boldsymbol{z}}_{it}')'$. The dimension of the vector of instruments, $\tilde{\boldsymbol{z}}_{it}$, d_z, is greater than or equal to 1. $\boldsymbol{\alpha}_i = \boldsymbol{\alpha} + \boldsymbol{a}_i$ is the $(d_w + d_z)\times 1$ vector of random coefficients, where $\boldsymbol{\alpha}$ is a vector of constants and $\boldsymbol{a}_i$ a vector of stationary random variables with zero means and constant variance-covariances. And finally, ϵ_{it} is a scalar idiosyncratic term.

The identifying distributional restrictions are summarized as follows:

RC 1 *(a)* $(\boldsymbol{\theta}_i, \boldsymbol{\zeta}_i), (\boldsymbol{a}_i, \boldsymbol{\epsilon}_i) \perp Z_i$ *and (b)* $\boldsymbol{\theta}_i, \boldsymbol{a}_i \perp \boldsymbol{\zeta}_i, \boldsymbol{\epsilon}_i$, *where* $Z_i \equiv (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iT})$ *is a* $T \times (d_w + d_z)$ *matrix,* $\boldsymbol{\zeta}_i \equiv (\zeta_{i1}, \ldots, \zeta_{iT})'$ *, and* $\boldsymbol{\epsilon}_i \equiv (\epsilon_{i1}, \ldots, \epsilon_{iT})'$.

In the above assumption, $\boldsymbol{z}_{it}$ is independent of the random coefficients, $(\boldsymbol{\varphi}_i, \boldsymbol{\alpha}_i)$, and the idiosyncratic errors, $(\zeta_{it}, \epsilon_{it})$. Also, as in the random effects model, we assume that the random coefficients and the idiosyncratic errors are independent of each other.

RC 2

$$\begin{aligned}\boldsymbol{\theta}_i, \zeta_{it} | X_i, Z_i, \boldsymbol{a}_i &\sim \boldsymbol{\theta}_i, \zeta_{it} | X_i - \mathrm{E}(X_i | Z_i, \boldsymbol{a}_i), Z_i, \boldsymbol{a}_i \\ &\sim \boldsymbol{\theta}_i, \zeta_{it} | \boldsymbol{\epsilon}_i, Z_i, \boldsymbol{a}_i \\ &\sim \boldsymbol{\theta}_i, \zeta_{it} | \boldsymbol{\epsilon}_i, \boldsymbol{a}_i,\end{aligned}$$

where $X_i \equiv (x_{i1}, \ldots, x_{iT})'$ *and* $\boldsymbol{\epsilon}_i = X_i - \mathrm{E}(X_i|Z_i, \boldsymbol{a}_i) = X_i - Z_i(\boldsymbol{\alpha} + \boldsymbol{a}_i)$.

In RC 2 the assumption is that the dependence of the structural error terms $\boldsymbol{\theta}_i$ and ζ_{it} on X_i, Z_i, and $\boldsymbol{a}_i$ is completely characterized by the reduced form error components, $\boldsymbol{\epsilon}_i$ and $\boldsymbol{a}_i$. If given $(\epsilon_{it}, \boldsymbol{a}_i)$ only contemporaneous correlations matter, then $\boldsymbol{\theta}_i, \zeta_{it} \perp \boldsymbol{\epsilon}_{i,-t} | (\epsilon_{it}, \boldsymbol{a}_i)$.

As in the model for random effects, we specify the marginal distributions of $\boldsymbol{a}_i$ and ϵ_{it}. We assume that

RC 3 $\boldsymbol{a}_i \sim N(0, \Sigma_a)$ *and that* $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$.

Let $\Theta_1 \equiv \{\boldsymbol{\alpha}, \Sigma_a, \sigma_\epsilon^2\}$ denote the set of parameters of random coefficient model in (4.2), the reduced form equation. The random coefficient model is a standard one, and most statistical packages have routines to estimate Θ_1.

Given assumption RC 2, we have

$$\begin{aligned}&\mathrm{E}(\boldsymbol{\theta}_i | X_i, Z_i, \boldsymbol{a}_i) = \mathrm{E}(\boldsymbol{\theta}_i | \boldsymbol{a}_i, \boldsymbol{\epsilon}_i) = \mathrm{E}(\boldsymbol{\theta}_i | \boldsymbol{a}_i) = \boldsymbol{\rho}_{\theta a}\boldsymbol{a}_i \text{ and} \\ &\mathrm{E}(\zeta_{it} | X_i, Z_i, \boldsymbol{a}_i) = \mathrm{E}(\zeta_{it} | \boldsymbol{a}_i, \boldsymbol{\epsilon}_i) = \mathrm{E}(\zeta_{it} | \boldsymbol{\epsilon}_i) = \boldsymbol{\rho}_{\zeta\epsilon}\boldsymbol{\epsilon}_i,\end{aligned} \tag{4.3}$$

where the second equality in each of the above follows from part (b) of assumption RC 1. In the above, $\boldsymbol{\rho}_{\theta a}$ is the $(d_w+1)\times(d_w+d_z)$ matrix of population regression coefficients of $\boldsymbol{\theta}_i$ on $\boldsymbol{a}_i$, and $\boldsymbol{\rho}_{\zeta\epsilon}$ is the population regression coefficient of ζ_{it} on $\boldsymbol{\epsilon}_i$.

Our assumptions and (4.3) then imply that the conditional expectation of y^*_{it} given X_i, Z_i, and $\boldsymbol{a}_i$ is given by

$$\mathrm{E}(y^*_{it}|X_i,Z_i,\boldsymbol{a}_i)=\mathcal{X}'_{it}\boldsymbol{\varphi}+\mathcal{X}'_{it}\boldsymbol{\rho}_{\theta a}\boldsymbol{a}_i+\boldsymbol{\rho}_{\zeta\epsilon}\boldsymbol{\epsilon}_i.$$

Because the stochastic part, $\boldsymbol{a}_i$, of the random coefficients in the reduced form equation are unobserved, the conditioning variable, $\boldsymbol{\epsilon}_i=X_i-Z_i(\boldsymbol{\alpha}+\boldsymbol{a}_i)$, too, is not identified. To estimate the structural parameters, as in the model with random effects, we first integrate out $\boldsymbol{a}_i$ from $\mathrm{E}(y^*_{it}|X_i,Z_i,\boldsymbol{a}_i)$ with respect to its conditional distribution, $f(\boldsymbol{a}_i|X_i,Z_i)$, to obtain

$$\begin{aligned}\mathrm{E}(y^*_t|X_i,Z_i)&=\int \mathrm{E}(y^*_{it}|X_i,Z_i,\boldsymbol{a}_i)f(\boldsymbol{a}_i|X_i,Z_i)d\boldsymbol{a}_i\\&=\mathcal{X}'_{it}\boldsymbol{\varphi}+\mathcal{X}'_{it}\boldsymbol{\rho}_{\theta a}\hat{\boldsymbol{a}}_i+\boldsymbol{\rho}_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_i,\end{aligned}\tag{4.4}$$

where $\hat{\boldsymbol{a}}_i=\mathrm{E}(\boldsymbol{a}_i|X_i,Z_i)$ and $\hat{\boldsymbol{\epsilon}}_i=X_i-Z_i(\boldsymbol{\alpha}+\hat{\boldsymbol{a}}_i)$. In part (c) of Lemma 1 in the appendix we show that

Lemma 1 (c) *If* $x_{it}=\boldsymbol{z}'_{it}\boldsymbol{\alpha}+\boldsymbol{z}'_{it}\boldsymbol{a}_i+\epsilon_{it}$, *where* $\boldsymbol{a}_i\sim N(0,\Sigma_a)$ *and* $\epsilon_{it}\sim N(0,\sigma^2_\epsilon)$, *then*

$$\mathrm{E}(\boldsymbol{a}_i|X_i,Z_i)=\hat{\boldsymbol{a}}_i(X_i,Z_i,\Theta_1)=[\sum_{t=1}^{T}\boldsymbol{z}_{it}\boldsymbol{z}'_{it}+\sigma^2_\epsilon\Sigma_a^{-1}]^{-1}\Bigg(\sum_{t=1}^{T}\boldsymbol{z}_{it}(x_{it}-\boldsymbol{z}'_{it}\boldsymbol{\alpha})\Bigg).$$

From (4.3) and (4.4), it therefore follows that

$$\mathrm{E}(\boldsymbol{\varphi}_i|X_i,Z_i)=\mathrm{E}(\boldsymbol{\varphi}+\boldsymbol{\theta}_i|X_i,Z_i)=\boldsymbol{\varphi}+\boldsymbol{\rho}_{\theta a}\hat{\boldsymbol{a}}_i\text{ and }\mathrm{E}(\zeta_{it}|X_i,Z_i)=\boldsymbol{\rho}_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_i.\tag{4.5}$$

Writing $\boldsymbol{\varphi}_i$ and ζ_{it} in error form as $\boldsymbol{\varphi}_i=\boldsymbol{\varphi}+\boldsymbol{\rho}_{\theta a}\hat{\boldsymbol{a}}_i+\tilde{\boldsymbol{\theta}}_i$ and $\zeta_{it}=\boldsymbol{\rho}_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_i+\tilde{\zeta}_{it}$ respectively, we can write the structural equation (4.1) as

$$y_{it}=1\{y^*_{it}=\mathcal{X}'_{it}\boldsymbol{\varphi}+\mathcal{X}'_{it}\boldsymbol{\rho}_{\theta a}\hat{\boldsymbol{a}}_i+\boldsymbol{\rho}_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_i+\mathcal{X}'_{it}\tilde{\boldsymbol{\theta}}_i+\tilde{\zeta}_{it}>0\}.\tag{4.6}$$

When $\tilde{\boldsymbol{\theta}}_i$ and $\tilde{\zeta}_{it}$ are independent of X_i and Z_i, and distributed normally with mean zero and variances Σ_θ and 1 respectively, then the parameters, $\Theta_2\equiv\{\boldsymbol{\varphi},\boldsymbol{\rho}_{\theta a},\boldsymbol{\rho}_{\zeta\epsilon},\Sigma_\theta\}$, of the above model can be estimated by integrated maximum likelihood method, where one can integrate out $\tilde{\boldsymbol{\theta}}_i$ using numerical multidimensional integration (see Heiss and Winschel, 2008). Alternatively, maximum simulated likelihood or Markov Chain Monte Carlo (MCMC) methods as discussed in Greene (2004), too, can be used to obtain Θ_2.

Once Θ_2 is estimated, the following measures of interest can be obtained. (A) The expected value, $\mathrm{E}(\boldsymbol{\varphi}_i|X_i,Z_i)=\boldsymbol{\varphi}+\boldsymbol{\rho}_{\theta a}\hat{\boldsymbol{a}}_i$. (B) The Average Partial Effect (APE) of changing a variable, say w, in time period t from w_{it} to $w_{it}+\Delta_w$ can be obtained as

$$\frac{\Delta G(\mathcal{X}_{it})}{\Delta_w}=\frac{G(\mathcal{X}_{it_{-w}},(w_t+\Delta_w))-G(\mathcal{X}_{it})}{\Delta_w},\text{ where}$$

$$G(\mathcal{X}_{it}) = \int \Phi\left(\frac{\mathcal{X}_{it}'\boldsymbol{\varphi} + \mathcal{X}_{it}'\boldsymbol{\rho}_{\theta a}\hat{\boldsymbol{a}}_i + \boldsymbol{\rho}_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_i}{1 + \mathcal{X}_{it}'\Sigma_\theta\mathcal{X}_{it}}\right) dF(\hat{\boldsymbol{a}}, \hat{\epsilon}).$$

Since we were able to identify the average slopes coefficients and the APEs when we augmented the structural equation with $\hat{\boldsymbol{a}}_i' \otimes \mathcal{X}_{it}'$ and $\hat{\boldsymbol{\epsilon}}_i$, as in the random effects case, we propose $(\hat{\boldsymbol{\epsilon}}_i, \hat{\boldsymbol{a}}_i)$ to be used as control function.

ACF 2

$$\begin{aligned}\zeta_{it}, \boldsymbol{\theta}_i | X_i, Z_i, \hat{\boldsymbol{a}}_i &\sim \zeta_{it}, \boldsymbol{\theta}_i | \hat{\boldsymbol{\epsilon}}_i, Z_i, \hat{\boldsymbol{a}}_i \\ &\sim \zeta_{it}, \boldsymbol{\theta}_i | \hat{\boldsymbol{\epsilon}}_i, \hat{\boldsymbol{a}}_i,\end{aligned}$$

where $\hat{\boldsymbol{\epsilon}}_i \equiv (\hat{\epsilon}_{i1}, \ldots, \hat{\epsilon}_{iT})' = X_i - Z_i(\boldsymbol{\alpha} + \hat{\boldsymbol{a}}_i)$ *and* $\hat{\boldsymbol{a}}_i = \mathrm{E}(\boldsymbol{a}_i | X_i, Z_i)$.

In the above, $\hat{\boldsymbol{\epsilon}}_i(X_i, Z_i)$ and $\hat{\boldsymbol{a}}_i(X_i, Z_i)$ are assumed to fully characterize the dependence of X_i and Z_i on the structural errors, ζ_{it} and $\boldsymbol{\theta}_i$, in (4.1). With $\hat{\boldsymbol{\epsilon}}_i$ and $\hat{\boldsymbol{a}}_i$ as control functions the semiparametric method in Hoderlein and Sherman (2015) can be employed to estimate the mean of $\boldsymbol{\varphi}_i$.

Kasy (2011) considers non-separable triangular systems for cross-sectional data to characterizes systems for which control functions – control functions such as $C(x, z) = x - \mathrm{E}(x|z)$ or $C(x, z) = F(x|z)$, where F is the conditional cumulative distribution function of x given z – exist. Kasy shows that when unobserved heterogeneity in first-stage reduced form equations is multi-dimensional, such as the reduced form equations with random coefficients, the errors in the structural equation are not independent of the endogenous covariates, x, or the instruments, z, given $C(x, z)$.

We consider panel data, where the random coefficient are time invariant, and our control functions, $\hat{\boldsymbol{\epsilon}}_i$ and $\hat{\boldsymbol{a}}_i$, are different from those considered in Kasy. Since $\hat{\boldsymbol{a}}_i$, a function of X_i and Z_i, summarizes certain individual specific information, as argued in section 2.2, the assumption in ACF 2 is that the dependence of $(\boldsymbol{\theta}_i, \zeta_{it})$ on (X_i, Z_i) can be reduced to dependence of $(\boldsymbol{\theta}_i, \zeta_{it})$ on $(\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{\epsilon}}_i)$, which is akin to dependence assumption in papers such as by AM and BH. The assumption is motivated by the result that under the restrictions in RC 1, RC 2, and (4.3), the expectations of ζ_{it} and $\boldsymbol{\theta}_i$ given (X_i, Z_i) depend on (X_i, Z_i) through $\hat{\boldsymbol{\epsilon}}_i$ and $\hat{\boldsymbol{a}}_i$ respectively.

5. IMPLICATIONS OF OWNERSHIP OF LAND AND FARM ASSETS ON CHILD LABOR

5.1. *Introduction*

Child labor is a pressing concern in all developing countries. According to International Labour Office's current (2016) estimates, 152 million children in the 5 to 17 years age group are working in economic activities throughout the world; 62 million of which are in the Asia-Pacific region. Conditions of child labor can vary. Many children work in hazardous industries, risking accident and injury, and there are others working in conditions that take a toll on their health. Moreover, when children work, they forego educating themselves[11], and, thus, human capital accumulation, with deleterious effect on their future earning potential. Furthermore, since there is positive externality to human capital accumulation, as argued by Baland and Robinson (2000) (henceforth BR), the social return to such accumulation, too, is not realized.

There is a huge literature, both empirical and theoretical, that has sought to understand the mechanism underlying child labor. What has emerged is that poverty (see Basu and Van, 1998; Baland and Robinson, 2000), along with imperfection in labor and land market (see Bhalotra and Heady, 2003; Dumas, 2007; Basu *et al.*, 2010) and capital market (see Baland and Robinson, 2000) to be the major causes of child labor. BR show that child labor increases when endowments of parents are low, and that when capital market imperfections exist and parents cannot borrow, child labor becomes inefficiently high.

Basu *et al.* (2010) (BDD) point out that papers like Bhalotra and Heady (2003)(BHy) and Dumas (2007) show that in some developing countries the amount of work the children of a household do increases with the amount of land possessed by the household. Since land is usually strongly correlated with a household's income, this finding seems to challenge the presumption that child labor involves the poorest households. They argue that these perverse findings are a facet of labor and land market imperfections, and that in developing countries, poor households in order to escape poverty want to send their children to work but are unable to do so because they have no access to labor markets close to their home. In such a situation, if the household comes to acquire some wealth, say land, its children, if only to escape penury, will start working. However, if the household's land ownership continues to rise, then beyond a point the household will be well-off enough and it will not want to make its children work.

BHy argue that on one hand there is the negative wealth effect of large landholding on child labor, whereby large landholding generate higher income and, thereby, makes it easier for the household to forgo the income that child labor would bring. On the other there is the substitution effect, where due to labor market imperfections, owners of land who are unable to productively hire labor on their farms have an incentive to employ their children. Since the marginal product of child labor is increasing in farm size, this incentive is stronger amongst larger landowners. The value of work experience will also tend to increase in farm size if the child stands to inherit the family farm. Furthermore, they argue that large landowners who cannot productively hire labor would want to sell their land rather than employ their children on it, but, because of land market failure, are unable to do so. Thus, land market failure reinforces labor market failure.

Cockburn and Dostie (2007) (CD) in their analysis of child labor in Ethiopia find that in presence of labor market imperfections, all assets need not be child labor enhancing. They find that certain productive assets that enable an increase in the total family income

[11] While school attendance may not be considered as the "inverse" of child labor, it can nevertheless be argued that whatever promotes school attendance is likely to deter child labor (see Baland and Robinson, 2000). Moreover, empirically there is a negative correlation between child labor and hours dedicated to schooling. This negative correlation between work and school attendance is also reflected in our data.

may not necessarily increase child labor. They show that assets such as oxen and ploughs that are operated by adults decrease child labor. To test this hypothesis, in our empirical specification we include an index of productive farm assets.

Now, while land and labor market imperfections may exist in developing countries, the extent of imperfection may not be uniform across all countries, or regions within a country. Hence, the relationship between child labor and different kinds of assets, such as landholding or agrarian assets, is an empirical question. The question is important because policy implications could be different under different relationships between various kinds of assets and child labor. For example, if one were to confirm the findings in BHy and BDD, then if monetary transfers are used to increase landholding or land redistribution is done in favor of the poor, child labor may in fact increase. On the other hand, when monetary transfers are used to increase agrarian assets, then is an inverse relationship between agrarian assets and child labor holds, such transfers could reduce the incidence of child labor.

In our data, we find the mean of non-agricultural income to be much higher than the mean agricultural income. This suggests that land is not the only source of income as in BHy and BDD. BDD assume that land is the only source of income and derive a regression equation where household income is left out. Since non-agricultural income constitutes a major portion of total household income, we also control for household income.

We also find that, overtime, land size distribution has become more unequal, which indicates that land market exists in the regions from where the data has been collected. Now, if land market exists, even if imperfect, then it is unlikely that land owned by households will be exogenous to a household's labor supply as in BHy and BDD, where land is mainly inherited, but endogenously determined along with household's, including children's, labor supply decisions. However, endogeneity could also arise due to omitted variables. These reasons would necessitate accounting for the endogeneity of landholding along with the endogeneity of productive assets and household income. To address the endogeneity problem we employ the method developed in the paper.

5.2. *Data and Empirical Model*

5.2.1. *Data*

We conduct our empirical analysis at the level of the child using two waves, 2006-07 and 2009-2010, of the data from Young Lives Study (YLS), a panel study from six districts of the state of Andhra Pradesh (henceforth AP) in India. We restrict our sample to children in the age group of 5 to 14 years in 2007 living in rural areas, and only a balanced panel is considered. Finally, excluding children for whom relevant information in either of the years was missing, we were left with 2458 children, which meant dropping about 23% children from the balanced panel. Table 2 and Table 3 have the relevant descriptive statistics.

[**Table 2 about here**]

[**Table 3 about here**]

The definition of work[12] includes (a) wage labor, (b) non-wage labor and (c) domestic

[12]Wage labor involves activities for pay, work done for money outside of household, or work done for

work. Children were asked how much time they spent in the reference period (a typical day in the last week) doing wage labor, non-wage labor, or domestic chores. If the answer was positive number of hours for any of the respective activities, then the binary variable $DWORK$ was assigned value 1, 0 otherwise. Similarly, if the child answered that the s/he spent positive number of hours at school, the binary variable, $DSCHOOL$, was assigned value 1 and 0 otherwise.

As can be seen from Table 2, the proportion of children working increased over the period of study. The major component of work (not reported here) is due to domestic chores. While both domestic and non-domestic work registered increase over the years, the increase in the proportion of children doing non-domestic work was higher. As far as schooling is concerned, the proportion of older children going to school dropped, but the proportion of younger children going to school saw increase over the years.

In Table 3, we can see that the mean annual household income (in 2009 rupees) increased during this period, and that non-agricultural income constitutes major proportion of the household income. We find that the average size of land owned increased over the years, and so did the index of farming related productive assets. The Asset Index is constructed by Principal Component Analysis of several variables, each of which indicate the number of farming related assets of each kind that the household owns. Farming assets constitute of agriculture tools, carts, pesticide pumps, ploughs, water pumps, threshers, tractors, and other farm equipments. Also, the size of landholding became more unequal. Among other variables, we see that the number of boys are slightly higher compared to the number of girls.

5.2.2. *Empirical Model*

We denote by $y_{it} = DWORK_{it}$, the binary outcome variable that takes value 1 if the child i decides[13] to work and 0 otherwise. We model the decision to work as

$$y_{it} = 1\{y^*_{it} = \mathcal{X}'_{it}\boldsymbol{\varphi} + \theta_i + \zeta_{it} > 0\}, \tag{5.1}$$

where y^*_{it} is amount of time devoted to work by child i in period t. In (5.1), $\mathcal{X}_{it} = (\boldsymbol{w}'_{it}, \boldsymbol{x}'_{it})'$, where $\boldsymbol{w}_{it}$ is a vector of strictly exogenous variables and $\boldsymbol{x}_{it}$ includes the endogenous variables: income of the household ($INCOME_{it}$), size of the land holdings ($LAND_{it}$), and the index of productive farm assets ($ASSET_{it}$).

To address the issue of endogeneity, we employ the two-step control function methodology developed in the paper, where we first estimate reduced form equations,

$$\boldsymbol{x}_{it} = \pi \boldsymbol{z}_{it} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{it}, \tag{5.2}$$

someone not a part the household. Non-wage labor includes tasks on family farm, cattle herding (household and/or community), other family business, shepherding, piecework or handicrafts done at home (not just farming), and domestic work includes tasks and chores such as fetching water, firewood, cleaning, cooking, washing, and shopping.

[13]There is a debate in the literature on whether working or attending school can be properly attributed to a child's own decision. See Edmonds (2007) to read more on the debate. Here we maintain that parents' decisions regarding their child is that of the child's.

where $\boldsymbol{z}_{it} = (\boldsymbol{w}'_{it}, \tilde{\boldsymbol{z}}'_{it})'$, $\tilde{\boldsymbol{z}}_{it}$ being the vector of instruments. The unobserved heterogeneity, $\boldsymbol{\alpha}_i = \bar{\pi}\bar{\boldsymbol{z}}_i + \boldsymbol{a}_i$, where $\bar{\pi}\bar{\boldsymbol{z}}_i = \mathrm{E}(\boldsymbol{\alpha}_i|Z_i)$ (see AS 3). The triangular representation (5.1) and (5.2) accounts for the fact that children's labor supply, y^*_{it}, household income, landholding, and productive farm assets are determined simultaneously.

We first estimate the parameters, Θ_1, of equation (5.2) using the stepwise ML method in Biørn (2004). Having estimated Θ_1, the control functions,

$$\hat{\boldsymbol{\alpha}}_i = \begin{bmatrix} \hat{\alpha}_{INCOME,i} \\ \hat{\alpha}_{LAND,i} \\ \hat{\alpha}_{ASSET,i} \end{bmatrix} \text{ and } \hat{\boldsymbol{\epsilon}}_{it} = \begin{bmatrix} \hat{\epsilon}_{INCOME,it} \\ \hat{\epsilon}_{LAND,it} \\ \hat{\epsilon}_{ASSET,it} \end{bmatrix},$$

based on Θ_1 are obtained.

The modified structural equation augmented with control functions to account for endogeneity and heterogeneity is given by

$$y_{it} = \{y^*_{it} = \mathcal{X}'_{it}\boldsymbol{\varphi} + \boldsymbol{\rho}_\alpha\hat{\boldsymbol{\alpha}}_i + \boldsymbol{\rho}_\epsilon\hat{\boldsymbol{\epsilon}}_{it} + \eta_{it} > 0\}, \tag{5.3}$$

where η_{it} is distributed normally with mean 0 and variance σ^2[14]. Inference about $\boldsymbol{\rho}_\alpha$ and $\boldsymbol{\rho}_\epsilon$ provides us with a test of exogeneity of the regressors, $\boldsymbol{x}$.

To identify the impact of the endogenous variables income, landholding, and asset holding on the decision to participate in work or go to school we employ the following instruments, $\tilde{\boldsymbol{z}}_{it}$: (1) $NREGS$, explained in the paragraph following, is the total $NREGS$ sanctioned amount at the mandal (region) level at the beginning of financial year (in 2008-09 prices), which Afridi *et al.* (2016) employ to instrument income in their paper, (2) $CASTE$, caste (social group) of the child, and (3) a set of four indicator variables that capture the level of infrastructural development in the household's locality/settlement.

The National Rural Employment Guarantee Scheme (NREGS) was initiated in 2006 by the Government of India with the objective to alleviate rural poverty. NREGS legally entitles rural households to 100 days of employment in unskilled manual labour (on public work projects) at a prefixed wage. Afridi *et al.* argue that more funds sanctioned would mean more work opportunity in NREGS, which will have a positive effect on household income. Now, it can be seen in Table 3 that over the period of our study, the proportion of children with either parent working in NREGS almost doubled. This increase in participation was accompanied by a rise in the number of days of work on NREGS projects as well. Afridi *et al.* in claiming $NREGS$ to be a valid instrument for income, argue that since fund sanctioned at the beginning of the financial year is not be affected by current demand for work, the funds sanctioned is exogenous and more funds imply more work opportunity

[14] For many children, as we know, the optimal choice of y^*_{it} is the corner solution, $y^*_{it} = 0$. For corner solution outcomes, we are interested in features of the distribution such as $\int \Pr(y^*_{it} > 0|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it})dF(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})$ and $\int \mathrm{E}(y^*_{it}|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it})dF(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})$, where

$$\begin{aligned} \mathrm{E}(y^*_{it}|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it}) &= \Pr(y^*_{it} = 0|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it}).0 + \Pr(y^*_{it} > 0|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it}).\mathrm{E}(y^*_{it}|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it}, y^*_{it} > 0) \\ &= \Pr(y^*_{it} > 0|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it})E(y^*_{it}|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it}, y^*_{it} > 0). \end{aligned}$$

Due to lack of space, in this application we study only $\int \Pr(y^*_{it} > 0|\mathcal{X}_{it}, \hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it})dF(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}})$.

in NREGS, which can have a positive effect on household income. Also, the total fund allocation to NREGS increased during the period 2007-2010. However, this increase was not uniform across the 15 *mandal*s[15].

Our second instrument is the caste, a system of social stratification, to which the child belongs. India is beleaguered with a caste system. Within this caste system, historically, the Scheduled Castes and Scheduled Tribes (SC/ST's) have been economically backward and concentrated in low-skill (mostly agricultural) occupations in rural areas. Moreover, they were also subject to centuries of systematic caste based discrimination, both economically and socially. The historical tradition of social division through the caste system created a social stratification along education, occupation, income, and wealth lines that has continued into modern India[16]. Fairing better than SC/ST's are those belonging to the "Other Backward Classes " (OBC)[17]. Hence, given the fact that income and wealth, both land and productive assets, vary with caste, we choose $CASTE$ as our second instrument, which is a discrete variable that takes three values: 1 if the child belongs to SC/ST household, 2 if the child belongs to OBC, and 3 if the child does not belong to SC/ST or OBC group, which we label as "Others" (OT). The variable $CASTE$, thus defined, is likely to be a good predictor of household income and wealth, where the average SC/ST household is likely to be poor, followed by the OBC's, and those in the OT group being the wealthiest.

[**Table 4 about here**]

We claim that $CASTE$ is a valid instrument for landholding because, though average wealth and income are evidently distributed along caste lines, we do not find a significant variation in child labor or school enrollment across caste or social group to which the child belongs (see Table 4). In other words, no social group is inherently disposed to make their children work or send them to school. This could be because rising awareness, overtime, about returns from education persuades families of all castes to send their children to school. We find support for the assertion in the literature too. Hnatkovska *et al.* (2012) find significant convergence in the education attainment levels and occupation choice of SC/ST's and non-SC/ST's between 1983 and 2004-2005. Moreover, the convergence in education level has been highest for the youngest cohort.

Our assertion that awareness about higher returns to education has been rising among all section of the society is also supported by the data. In the first wave of the data, the following question was asked: "Imagine that a family in the village has a 12 year old son/daughter who is attending school full-time. The family badly needs to increase the

[15] Data on the sanctioned funds at the *mandal* level is obtained from the Andhra Pradesh Government's website on NREGS (http://nrega.ap.gov.in/).

[16] In fact, this stratification was so endemic that the constitution of India aggregated these castes into a schedule of the constitution and provided them with affirmative action cover in both education and public sector employment. This constitutional initiative was viewed as a key component of attaining the goal of raising the social and economic status of the SC/STs to the levels of the non-SC/ST's.

[17] The Government of India classifies, a classification based on social and economic conditions, some of its citizen as Other Backward Classes (OBC). The OBC list is dynamic (castes and communities can be added or removed) and is supposed to change from time to time depending on social, educational and economic factors. In the constitution, OBC's are described as "socially and educationally backward classes", and government is enjoined to ensure their social and educational development.

household income. One option is to send the son/daughter to work but the son/daughter wants to stay in school. What should the family do?" An overwhelming percentage of the respondents answered that they should let children be at school; moreover, there was little difference in the response across caste groups – 90% of SC/ST's, 87% of OBC's, and 93% of OT's wanted that sons of such distressed families be kept at school. For daughters, the corresponding figures are: 87% of SC/ST's, 87% of OBC's, and 91% of OT's. Also, 96% of SC/ST households expected their children to complete a minimum of high school. The corresponding figure for OBC's and OT's are 95% and 98% respectively.

Our third set of instruments is the set of four dummy variables, which indicate (1) if drinkable water is provided in the locality/settlement, (2) if the services of a national bank are provided in the locality, (3) if private hospitals exist in the locality, and (4) if access to the locality is via an engineered road. As in BHy, these variables, which indicate the level of infrastructure development, are employed to instrument the index of productive farm assets.

5.3. *Discussion of Results*

We begin by discussing the results of the first stage reduced form equations in (5.2). The results in Table 5 suggest that our instruments are good predictors of the endogenous variables, income and wealth. First, corroborating the results in Afridi *et al.*, we too find that an increase in the amount sanctioned for NREGS projects in a mandal increases the household income. Secondly, as expected, $CASTE$ does, on an average, correctly predict the economic status of household in the regression of income, land holding, and assets on $CASTE$. Finally, the dummy variables indicating the level of infrastructure development are positively correlated with the index of productive farm assets.

[**Table 5 about here**]

The second-stage estimates of the structural equation (5.3) are illustrated in Table 6. Here, we would like to state that (a) our specification is parsimonious, where the only exogenous explanatory variables are the age and the sex of the children. (b) To estimate the slope coefficients in (5.3), we employ STATA's routine for generalized estimating equation. (c) All the specifications include district dummies, a time dummy, and the interaction of the two to account for the fact that the districts to which children belong may have different economic growth trajectories as well as trends related to work and education. The time dummy allows us to control for changes in demand and supply of work over time. (d) The average partial effects (APEs) of variables were computed at the mean of variables from the second round (2010) of data. (e) The standard errors were estimated using the analytical expression of the covariance matrix derived in appendix A of the supplementary appendix.

[**Table 6 about here**]

We begin by comparing the results from standard correlated random effect (CRE) probit model with the results developed in this paper. It can be evinced from Table 6 that most of the control functions – $\hat{\alpha}_{INCOME}$, $\hat{\alpha}_{LAND}$, $\hat{\alpha}_{ASSET}$, $\hat{\epsilon}_{INCOME}$, $\hat{\epsilon}_{LAND}$, $\hat{\epsilon}_{ASSET}$ – are significant. This suggests that income and ownership of wealth, be it land or productive assets, are endogenously determined along with household's labor supply, including that of the child's, decisions. When income and wealth are not instrumented, as in the CRE probit, the coefficient estimate for household income, in the light of the discussion in the paper, has an incorrect sign. Moreover, the result of CRE probit suggests that ownership of land and farm assets do not affect child labor, which, given the many recent evidences, is unlikely in a developing country. The results, thus, make clear the importance of accounting for endogeneity of income, landholding, and farm asset.

The estimates from the control function method suggest that children of households that have a higher landholding are more likely to engage in work[18]. This is in conformity with the findings in BDD, BHy and CD, where, due to presence of land, labor, and credit market imperfections, ownership of large amount land provides incentives for children to work. As far as income is concerned, we find that higher household income reduces the chances of child labor, which again confirms poverty to be a cause of child labor.

We find that ownership of productive farm assets leads to a significantly high reduction in children's participation in work for the average family. Dumas, BHy and CD argue that an increase in asset holding that increases the marginal productivity of labor induces two opposite effects on labor. While the income effect of increased wealth tends to reduce the labor time, the substitution effect, due to the absence of labor market, provides incentives for work, and tends to increase children's labor time. Our results suggest that the wealth effect of farm assets, which are not likely to be operated by children, dominate to reduce children's labor time. Secondly, since the prevalence of farm assets is high in those regions where there has been infrastructure development, it seems that lack of infrastructure development that impedes access to, or does not provide incentives to acquire, productive farm assets may be an important factor determining child labor[19]. Finally, we find that older children and boys are more likely to work.

6. CONCLUDING REMARKS

The primary objective of the paper has been to develop a method to point estimate structural measures of interest for panel data binary response model in a triangular system while accounting for multi-dimension unobserved heterogeneity. The unobserved heterogeneity terms constitute of time invariant random effects/coefficients and idiosyncratic errors. We first identify the expected values – conditional on the endogenous variables, $X_i \equiv (\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{iT})'$, and the exogenous variables, $Z_i \equiv (\boldsymbol{z}'_{i1}, \ldots, \boldsymbol{z}'_{iT})'$ – of the heterogeneity terms of the reduced form equations, and show that given these expected values, the

[18]Though we do not report here, we did not find that nonlinear terms of income, land, and productive assets to be significant

[19]In a separate set of regressions that included only the exogenous variables, we tried to assess if the infrastructure variables had independent impacts on work and schooling decisions of children. These variables turned out to be insignificant, suggesting that the demand for child labor or opportunities for schooling were not affected by infrastructure development or its lack in rural AP. In other words, infrastructure had its effect on work and schooling outcomes only through its impact on the economic conditions of certain households. This also validates using infrastructure variables as instruments for farm assets.

measures of interest are identified. We then propose that these conditional expected values of the heterogeneity terms in the reduced form equations be used as control functions.

The proposed method makes a number of interesting contribution to the literature. Apart from achieving identification of measures such as the average partial effects in a triangular system with multi-dimensional heterogeneity, among the class of triangular system with imposed structures similar to ours, the proposed control function method requires weaker restrictions than the traditional control function methods. Finally, the method allows for instruments with small support, which was possible due to panel data and time invariance of certain heterogeneity terms. Also, Monte Carlo experiments show that compared to alternative panel data binary choice models similar to ours, our method performs better.

The estimator was applied to estimate the causal effects of income and wealth – land and farm assets – on the incidence of child labor. We found that income and household ownership of productive farm assets significantly lower the incidence of child labor, suggesting a strong income effect of productive farm assets. Secondly, large landholding increases the likelihood of child labor, suggesting a substitution effect of land ownership. Thirdly, a test of exogeneity revealed that ownership of land is determined endogenously along with household labor supply decisions, contrary to what most empirical studies on child labor in developing countries assume.

Finally, we would like to note that for a more general semiparametric method, an important generalization would be to identify and estimate the proposed control functions without making distributional assumptions about the heterogeneity terms of the reduced form equations.

REFERENCES

Afridi, F., Mukhopadhyay, A. and Sahoo, S. (2016). Female Labor Force Participation and Child Education in India: Evidence from the National Rural Employment Guarantee Scheme. *IZA Journal of Labor & Development*, **5:7**, doi:10.1186/s40175–016–0053–y.

Altonji, J. G. and Matzkin, R. L. (2005). Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors. *Econometrica*, **73**, 1053–1102.

Arellano, M. and Bonhomme, S. (2011). Nonlinear Panel Data Analysis. *Annual Review of Economics*, **3**, 395–424.

Baland, J. M. and Robinson, J. A. (2000). Is Child Labor Inefficient? *Journal of Political Economy*, **108**, 663–679.

Baltagi, B. H., Bresson, G. and Pirotte, A. (2006). Joint LM Test for Heteroskedasticity in a One-way Error Component Model. *Journal of Econometrics*, **134**, 401–417.

—, Song, S. H. and Jung, B. C. (2010). Testing for Heteroskedasticity and Serial Correlation in a Random Effects Panel Data Model. *Journal of Econometrics*, **154**, 122–124.

Basu, K., Das, S. and Dutta, B. (2010). Child Labor and Household Wealth: Theory and Empirical Evidence of an Inverted-U. *Journal of Development Economics*, **91**, 8–14.

— and Van, P. H. (1998). The Economics of Child Labor. *American Economic Review*, **88**, 412–427.

Bester, A. and Hansen, C. (2009). Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model. *Journal of Business and Economic Statistics*.

Bhalotra, S. and Heady, C. (2003). Child Farm Labor: The Wealth Paradox. *World Bank Economic Review*, **17**, 197–227.

Biørn, E. (2004). Regression Systems for Unbalanced Panel Data: A Stepwise Maximum Likelihood Procedure . *Journal of Econometrics*, **122**, 281–291.

Blundell, R. and Powell, J. (2003). Endogeneity in Nonparametric and Semiparametric Regression Models. In M. Dewatripont, L. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress*, vol. 2, Cambridge: Cambridge University Press.

— and — (2004). Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies*, **71**, 655–679.

Chamberlain, G. (1984). Panel Data. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, Elsevier.

— (2010). Binary Response Models for Panel Data: Identification and Information. *Econometrica*, **78**, 159–168.

Cockburn, J. and Dostie, B. (2007). Child Work and Schooling: The Role of Household Asset Profiles and Poverty in Rural Ethiopia. *Journal of African Economies*, **16**, 519–563.

D'Haultfœuille, X. and Février, P. (2015). Identification of Nonseparable Triangular Models with Discrete Instruments. *Econometrica*, **83** (3), 1199–1210.

Dumas, C. (2007). Why do Parents make their Children Work? A Test of the Poverty Hypothesis in Rural Areas of Burkina Faso. *Oxford Economic Papers*, **59**, 301–329.

Edmonds, E. V. (2007). Child Labor. In P. Schultz and J. A. Strauss (eds.), *Handbook of Development Economics*, vol. 4, North Holland: Elsevier, pp. 3607–3709.

Florens, J., Heckman, J. J., Meghir, C. and Vytlacil, E. (2008). Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects. *Econometrica*, **76**, 11911206.

Greene, W. (2004). Convenient Estimators for the Panel Probit Model: Further Results. *Empirical Economics*, **29**, 21–47.

Heiss, F. and Winschel, V. (2008). Likelihood Approximation by Numerical Integration on Sparse Grids. *Journal of Econometrics*, **144**, 6280.

Hnatkovska, V., Lahiri, A. and Paul, S. (2012). Castes and Labor Mobility. *American Economic Journal: Applied Economics*, **4**, 274307.

Hoderlein, S. and Sherman, R. (2015). Identification and Estimation in a Correlated Random Coefficients Binary Response Model. *Journal of Econometrics*, **188**, 135–149.

— and White, H. (2012). Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects. *Journal of Econometrics*, **168**, 300–314.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, 2nd edn.

Imbens, G. W. and Newey, W. K. (2009). Identification and Estimation of Triangular Simultaneous Equations Models without Additivity. *Econometrica*, **77**, 1481–1512.

Kasy, M. (2011). Identification in Triangular Systems Using Control Functions. *Econometric Theory*, **27** (3).

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*, **73**, 1322.

Manski, C. F. (1988). Identification of Binary Response Models. *Journal of the American Statistical Association*, **83**, 729–738.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, **46**, 69–85.

Papke, L. E. and Wooldridge, J. M. (2008). Panel Data Methods for Fractional Response Variables with an application to Test Pass Rates. *Journal of Econometrics*, **145**, 121–133.

Rothe, C. (2009). Semiparametric Estimation of Binary Response Models with Endogenous Regressors. *Journal of Econometrics*, **153**, 51–64.

Torgovitsky, A. (2015). Identification of Nonseparable Models Using Instruments With Small Support. *Econometrica*, **83**, 1185–1197.

Wooldridge, J. M. (2010). Correlated Random Effects Models with Unbalanced Panels, Michigan State University, Department of Economics, Working Paper.

APPENDIX A PROOFS

LEMMA 1 *(a) Suppose $\boldsymbol{x}_t$ is specified as*

$$\boldsymbol{x}_t = \pi \boldsymbol{z}_t + \bar{\pi}\bar{\boldsymbol{z}}_t + \boldsymbol{a} + \boldsymbol{\epsilon}_t, t \in \{1, \dots, T\},$$

where $\boldsymbol{a}$ and $\boldsymbol{\epsilon}_t$ are normally distributed with variances $\Lambda_{\alpha\alpha}$ and $\Sigma_{\epsilon\epsilon}$ respectively, then

$$\mathrm{E}(\boldsymbol{a}|X,Z) = \hat{\boldsymbol{a}}(X,Z,\Theta_1) = [T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}\sum_{t=1}^{T}(\boldsymbol{x}_t - \pi\boldsymbol{z}_t - \bar{\pi}\bar{\boldsymbol{z}}_t).$$

(b) Suppose we have a single endogenous variable, x_t, given by

$$x_t = \pi \boldsymbol{z}_t + a + \epsilon_t, t = 1, \dots, T,$$

where the errors, $\boldsymbol{\epsilon} \equiv (\epsilon_1, \dots, \epsilon_T)'$, are non-spherical such that $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \Omega$, a $T \times T$ matrix, and a is normally distributed with variance σ_α^2, then

$$\mathrm{E}(\boldsymbol{a}|X,Z) = \hat{a}(X,Z,\Theta_1) = (x_1 - \pi\boldsymbol{z}_1)\omega_1 + \ldots + (x_T - \pi\boldsymbol{z}_T)\omega_T,$$

where $(\omega_1, \dots, \omega_T)' = \frac{\Omega^{-1}e}{(e'\Omega^{-1}e + \sigma_\alpha^{-2})}$ and e is a vector of ones of dimension T.

(c) If $x_{it} = \boldsymbol{z}_{it}'\boldsymbol{\alpha} + \boldsymbol{z}_{it}'\boldsymbol{a}_i + \epsilon_{it}$, where $\boldsymbol{a}_i \sim N(0, \Sigma_a)$ and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$, then

$$\hat{\boldsymbol{a}}_i(X_i, Z_i, \Theta_1) = [\sum_{t=1}^{T}\boldsymbol{z}_{it}\boldsymbol{z}_{it}' + \sigma_\epsilon^2\Sigma_a^{-1}]^{-1}\left(\sum_{t=1}^{T}\boldsymbol{z}_{it}(x_{it} - \boldsymbol{z}_{it}'\boldsymbol{\alpha})\right).$$

PROOF 1

(a) To obtain $\hat{\boldsymbol{a}}$, using Bayes rule we first write $f(\boldsymbol{a}|X,Z)$ as

$$f(\boldsymbol{a}|X,Z) = \frac{f(X,Z|\boldsymbol{a})g(\boldsymbol{a})}{h(X,Z)},$$

where g and h are density functions. The above can be written as

$$\frac{f(X,Z|\boldsymbol{a})g(\boldsymbol{a})}{h(X,Z)} = \frac{f(X|Z,\boldsymbol{a})p(Z|\boldsymbol{a})g(\boldsymbol{a})}{h(X|Z)p(Z)}.$$

Since Z is independent of the residual individual effects, $\boldsymbol{a}$, $p(Z|\boldsymbol{a}) = p(Z)$; that is,

$$f(\boldsymbol{a}|X,Z) = \frac{f(X|Z,\boldsymbol{a})g(\boldsymbol{a})}{h(X|Z)} = \frac{f(X|Z,\boldsymbol{a})g(\boldsymbol{a})}{\int f(X|Z,\boldsymbol{a})g(\boldsymbol{a})d\boldsymbol{a}}.$$

Given the above, we can obtain $\hat{\boldsymbol{a}}_i(X,Z,\Theta_1) = \mathrm{E}(\boldsymbol{a}_i|X_i,Z_i)$ as

$$\hat{\boldsymbol{a}}(X,Z,\Theta_1) = \int \boldsymbol{a}f(\boldsymbol{a}|X,Z)d(\boldsymbol{a}) = \int \frac{\boldsymbol{a}f(X|Z,\boldsymbol{a})g(\boldsymbol{a})d\boldsymbol{a}}{\int f(X|Z,\boldsymbol{a})g(\boldsymbol{a})d\boldsymbol{a}}. \tag{A-1}$$

Since $\boldsymbol{a}$ is normally distributed with mean zero and variance $\Lambda_{\alpha\alpha}$, $g(\boldsymbol{a}) = \phi(\boldsymbol{a})$, where $\phi(\boldsymbol{a})$ is the multivariate normal density function of $\boldsymbol{a}$. Given that $\boldsymbol{x}_t = \pi\boldsymbol{z}_t + \bar{\pi}\bar{\boldsymbol{z}}_t + \boldsymbol{a} + \boldsymbol{\epsilon}_t$, where $\boldsymbol{a}$ and $\boldsymbol{\epsilon}_t$ are normally distributed with variances $\Lambda_{\alpha\alpha}$ and $\Sigma_{\epsilon\epsilon}$ respectively, conditional on Z and $\boldsymbol{a}$ each of the $\boldsymbol{x}_t$'s of $X \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ are independently and normally distributed with mean $\pi\boldsymbol{z}_t + \bar{\pi}\bar{\boldsymbol{z}}_t + \boldsymbol{a}$ and standard deviation $\Sigma_{\epsilon\epsilon}$. Thus for an individual i the expected a posteriori value of $\boldsymbol{a}_i$ is given by

$$\hat{\mathfrak{a}}(X, Z, \Theta_1) = \frac{\int \boldsymbol{a} \prod_{t=1}^{T} f(\boldsymbol{x}_t | Z, \boldsymbol{a}) \phi(\boldsymbol{a}) d\boldsymbol{a}}{\int \prod_{t=1}^{T} f(\boldsymbol{x}_t | Z, \boldsymbol{a}) \phi(\boldsymbol{a}) d\boldsymbol{a}} = \frac{\int \boldsymbol{a} \exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{v}_t - \boldsymbol{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{v}_t - \boldsymbol{a}))\phi(\boldsymbol{a}) d\boldsymbol{a}}{\int \exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{v}_t - \boldsymbol{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{v}_t - \boldsymbol{a}))\phi(\boldsymbol{a}) d\boldsymbol{a}}. \tag{A-2}$$

In (A-2), for convenience we have, with a slight abuse of notation, defined $\boldsymbol{v}_t$ as $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi\boldsymbol{z}_t - \bar{\pi}\bar{\boldsymbol{z}}_t$, whereas in the main text $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi\boldsymbol{z}_t$. Also, the dimension of $\boldsymbol{x}_t$, $d_x = m$.

Since $\phi(\boldsymbol{a}) = \frac{1}{\sqrt{(2\pi)^m |\Lambda_{\alpha\alpha}|}} \exp(\boldsymbol{a}'\Lambda_{\alpha\alpha}^{-1}\boldsymbol{a})$, the expression, $\exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{v}_t - \boldsymbol{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{v}_t - \boldsymbol{a}))\phi(\boldsymbol{a})$, in the numerator and denominator of $\hat{\mathfrak{a}}(X, Z, \Theta_1)$ can be written as

$$\exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{v}_t - \boldsymbol{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{v}_t - \boldsymbol{a}))\phi(\boldsymbol{a}) = \frac{1}{\sqrt{(2\pi)^m|\Lambda_{\alpha\alpha}|}} \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right)\right]$$
$$\exp\left[-\frac{1}{2}\left(-\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{a} - \boldsymbol{a}'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right) + \boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}\right)\right],$$

where $\Sigma^{-1} = T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}$.

Let $\mathfrak{a} = C'\boldsymbol{a}$, CC' being the Cholesky decomposition of the matrix, $\Sigma^{-1} = T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}$. Hence, $\boldsymbol{a} = C'^{-1}\mathfrak{a}$, $\boldsymbol{a}'\Sigma^{-1}\boldsymbol{a} = \mathfrak{a}'\mathfrak{a}$ and $d\boldsymbol{a} = |C'^{-1}|d\mathfrak{a}$. Thus we can write the second expression in the above equation as

$$\exp\left[-\frac{1}{2}\left(-\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{a} - \boldsymbol{a}'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right) + \boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}\right)\right] =$$
$$\exp\left[-\frac{1}{2}\left(-\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}\mathfrak{a} - \mathfrak{a}'(C'^{-1})'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right) + \mathfrak{a}'\mathfrak{a}\right)\right].$$

Subtracting and adding the term, $\frac{1}{2}\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}(C'^{-1})'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right)$, inside the square parenthesis, we get

$$\exp\left[-\frac{1}{2}\left(-\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{a} - \boldsymbol{a}'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right) + \boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}\right)\right] = \exp\left[\frac{1}{2}\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}(C'^{-1})'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right)\right]$$
$$\exp\left[-\frac{1}{2}\left(\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}(C'^{-1})'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right) - \left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}\mathfrak{a} - \mathfrak{a}'(C'^{-1})'\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right) + \mathfrak{a}'\mathfrak{a}\right)\right],$$

where the RHS simplifies to

$$\exp\Big[\frac{1}{2}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t'\Big)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big]\exp\Big[-\frac{1}{2}\Big(\mathfrak{a}-C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big)'\Big(\mathfrak{a}-C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big)\Big].$$

Therefore the expression, $\exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{v}_t-\boldsymbol{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{v}_t-\boldsymbol{a}))\phi(\boldsymbol{a})$, in the numerator and denominator of $\hat{\boldsymbol{a}}(X,Z,\Theta_1)$ becomes

$$\begin{aligned}&\exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{v}_t-\boldsymbol{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{v}_t-\boldsymbol{a}))\phi(\boldsymbol{a})=\\ &\frac{1}{\sqrt{|\Lambda_{\alpha\alpha}|}}\exp\Big[-\frac{1}{2}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)'\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big]\exp\Big[\frac{1}{2}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t'\Big)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big]\\ &\frac{1}{\sqrt{(2\pi)^m}}\exp\Big[-\frac{1}{2}\Big(\mathfrak{a}-C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big)'\Big(\mathfrak{a}-C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big)\Big].\end{aligned}\tag{A-3}$$

First, note that the term,

$$\frac{1}{\sqrt{|\Lambda_{\alpha\alpha}|}}\exp\Big[-\frac{1}{2}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)'\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big]\exp\Big[\frac{1}{2}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t'\Big)\Sigma_{\epsilon\epsilon}^{-1}C'^{-1}C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big],$$

on the RHS of equation (A-3) is independent of $\mathfrak{a}$. Therefore it can be taken out of the integration in expression for $\hat{\boldsymbol{a}}_i(X_i,Z_i,\Theta_1)$. And since it is in both the numerator and the denominator, the term cancels out. The second thing to note is that the expression,

$$\frac{1}{\sqrt{(2\pi)^m}}\exp\Big[-\frac{1}{2}\Big(\mathfrak{a}-C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big)'\Big(\mathfrak{a}-C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\Big)\Big],$$

on the RHS of the equation (A-3) is the normal density function of $\mathfrak{a}$, $\phi_m(\mathfrak{a})$, with mean $C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)$ and variance I_m. Thus we get

$$\begin{aligned}\hat{\boldsymbol{a}}_i(X_i,Z_i,\Theta_1)=\frac{C'^{-1}\int\mathfrak{a}\phi_m(\mathfrak{a})d\mathfrak{a}}{\int\phi_m(\mathfrak{a})d\mathfrak{a}}&=C'^{-1}C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\\ &=\Sigma\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)\\ &=[T\Sigma_{\epsilon\epsilon}^{-1}+\Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big),\end{aligned}\tag{A-4}$$

where the expression after the second equality is obtained because $\int\phi_m(\mathfrak{a})d\mathfrak{a}=1$ and $\int\mathfrak{a}\phi_m(\mathfrak{a})d\mathfrak{a}=C^{-1}\Sigma_{\epsilon\epsilon}^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{v}_t\Big)$. Finally,

$$C'^{-1}C^{-1}=(CC')^{-1}=(\Sigma^{-1})^{-1}=\Sigma=[T\Sigma_{\epsilon\epsilon}^{-1}+\Lambda_{\alpha\alpha}^{-1}]^{-1}.$$

(b) While discussing the restrictions imposed on the reduced form equation, we had stated that when $d_x = 1$, the assumption that a and ϵ_t are completely independent of Z can be weakened to allow for non-spherical error components. Suppose that $\epsilon_t, t = 1, \dots, T$ are serially dependent such that $\boldsymbol{\epsilon} \equiv (\epsilon_1, \dots, \epsilon_T)'$ normally distributed with $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \Omega$ and a is normally distributed and heteroscedastic as in Baltagi *et al.* (2010). Given the distribution of the error terms, the expected a posteriori value of a_i is given by

$$\begin{aligned}\hat{a}(X, Z, \Theta_1) &= \frac{\int af(X|Z,a)\phi(a)da}{\int f(X|Z,a)\phi(a)da}\\ &= \frac{\int a\exp(-\frac{1}{2}(\boldsymbol{\upsilon} - ae)'\Omega^{-1}(\boldsymbol{\upsilon} - ae))\phi(a)da}{\int \exp(-\frac{1}{2}(\boldsymbol{\upsilon} - a)'\Omega^{-1}(\boldsymbol{\upsilon} - ae))\phi(a)da},\end{aligned}$$

where e is a vector of ones with dimension T and $\boldsymbol{\upsilon} \equiv ((x_1 - \pi\boldsymbol{z}_1), \dots, (x_T - \pi\boldsymbol{z}_T))'$.

Now, the expression, $\exp(-\frac{1}{2}(\boldsymbol{\upsilon}-ae)'\Omega^{-1}(\boldsymbol{\upsilon}-ae))\phi(a)$, in the numerator and denominator of $\hat{a}(X, Z, \Theta_1)$ can be written as

$$\begin{aligned}\exp(-\frac{1}{2}(\boldsymbol{\upsilon} - ae)'\Omega^{-1}(\boldsymbol{\upsilon} - ae))\phi(a) = \frac{1}{\sqrt{2\pi}\sigma_\alpha}\exp\Big[-\frac{1}{2}\boldsymbol{\upsilon}'\Omega^{-1}\boldsymbol{\upsilon}\Big]\\ \exp\Big[-\frac{1}{2}\Big(-\boldsymbol{\upsilon}'\Omega^{-1}ae - ae'\Omega^{-1}\boldsymbol{\upsilon} + a^2\sigma^2\Big)\Big],\end{aligned}$$

where $\sigma^2 = e'\Omega^{-1}e + \frac{1}{\sigma_\alpha^2}$. As in part (a), if we let $\mathfrak{a} = a\sigma$, then we have $a = \frac{\mathfrak{a}}{\sigma}$, $a^2\sigma^2 = \mathfrak{a}^2$ and $da = \frac{d\mathfrak{a}}{\sigma}$. Following the same procedure as in part (a) it can be shown that

$$\hat{a}(X, Z, \Theta_1) = (x_1 - \pi\boldsymbol{z}_1)\omega_1 + \ldots + (x_T - \pi\boldsymbol{z}_T)\omega_T,$$

where $(\omega_1, \dots, \omega_T)' = \frac{\Omega^{-1}e}{(e'\Omega^{-1}e + \sigma_\alpha^{-2})}$.

(c) As in the model with random effects, using Bayes rule we can obtain $\hat{\boldsymbol{a}}_i(X_i, Z_i, \Theta_1) = \mathrm{E}(\boldsymbol{a}_i|X_i, Z_i)$ as

$$\hat{\boldsymbol{a}}_i(X_i, Z_i, \Theta_1) = \int \boldsymbol{a}_i f(\boldsymbol{a}_i|X_i, Z_i)d(\boldsymbol{a}_i) = \int \frac{\boldsymbol{a}_i f(X_i|Z_i, \boldsymbol{a}_i)\phi(\boldsymbol{a}_i)d\boldsymbol{a}_i}{\int f(X_i|Z_i, \boldsymbol{a}_i)\phi(\boldsymbol{a}_i)d\boldsymbol{a}_i}, \tag{A-5}$$

where $\phi(\boldsymbol{a}_i)$ denotes the normal distribution of $\boldsymbol{a}_i$, which is distributed with mean zero and variance Σ_a, and conditional on Z_i and $\boldsymbol{a}_i$, each of the x_{it}'s of $X_i \equiv (x_{i1}, \dots, x_{iT})'$ are independently normally distributed with mean $\boldsymbol{z}_{it}'\boldsymbol{\alpha} + \boldsymbol{z}_{it}'\boldsymbol{a}_i$ and standard deviation σ_ϵ. In equation (A-5) we have written the expected value of $\boldsymbol{a}_i$ given X_i, Z_i, which, dropping the subscript i, is given by

$$\begin{aligned}\hat{\boldsymbol{a}}(X, Z, \Theta_1) &= \frac{\int \boldsymbol{a}\prod_{t=1}^T f(x_t|Z, \boldsymbol{a})\phi(\boldsymbol{a})d\boldsymbol{a}}{\int \prod_{t=1}^T f(x_t|Z, \boldsymbol{a})\phi(\boldsymbol{a})d\boldsymbol{a}}\\ &= \frac{\int \boldsymbol{a}\exp(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^T(\upsilon_t - \boldsymbol{z}_t'\boldsymbol{a})^2)\phi(\boldsymbol{a})d\boldsymbol{a}}{\int \exp(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^T(\upsilon_t - \boldsymbol{z}_t'\boldsymbol{a})^2)\phi(\boldsymbol{a})d\boldsymbol{a}},\end{aligned} \tag{A-6}$$

where $\upsilon_t = x_t - \boldsymbol{z}_t'\boldsymbol{\alpha}$.

Since $\phi(\boldsymbol{a}) = \frac{1}{\sqrt{(2\pi)^m|\Sigma_a|}}\exp(\boldsymbol{a}'\Sigma_a^{-1}\boldsymbol{a})$, the expression, $\exp(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}(\upsilon_t - \boldsymbol{z}_t'\boldsymbol{a})^2)\phi(\boldsymbol{a})$, in the numerator and denominator of $\hat{\boldsymbol{a}}(X_i, Z, \Theta_1)$ can be written as

$$\exp(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}(\upsilon_t - \boldsymbol{z}_t'\boldsymbol{a})^2)\phi(\boldsymbol{a}) = \frac{1}{\sqrt{(2\pi)^m|\Sigma_a|}}\exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}\upsilon_t^2\Big)$$
$$\exp\Big[-\frac{1}{2}\Big(-\frac{2}{\sigma_\epsilon^2}\sum_{t=1}^{T}\upsilon_t\boldsymbol{z}_t'\boldsymbol{a} + \boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}\Big)\Big],$$

where $\Sigma^{-1} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}\boldsymbol{z}_t\boldsymbol{z}_t' + \Sigma_a^{-1}$.

Since Σ_a^{-1} and $\sum_{t=1}^{T}\boldsymbol{z}_t\boldsymbol{z}_t'$ are both positive definite, Σ^{-1} too is positive definite. Let $\boldsymbol{\mathfrak{a}} = C'\boldsymbol{a}$, CC' being the Cholesky decomposition of the matrix, Σ^{-1}. Hence, $\boldsymbol{a} = C'^{-1}\boldsymbol{\mathfrak{a}}$, $\boldsymbol{a}'\Sigma^{-1}\boldsymbol{a} = \boldsymbol{\mathfrak{a}}'\boldsymbol{\mathfrak{a}}$ and $d\boldsymbol{a} = |C'^{-1}|d\boldsymbol{\mathfrak{a}}$. Thus we can write the second expression in the above equation as

$$\exp\Big[-\frac{1}{2}\Big(-\frac{2}{\sigma_\epsilon^2}\sum_{t=1}^{T}\upsilon_t\boldsymbol{z}_t'\boldsymbol{a} + \boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}\Big)\Big] = \exp\Big[-\frac{1}{2}\Big(-\Big(\frac{2}{\sigma_\epsilon^2}\sum_{t=1}^{T}\upsilon_t\boldsymbol{z}_t'\Big)C'^{-1}\boldsymbol{\mathfrak{a}} + \boldsymbol{\mathfrak{a}}'\boldsymbol{\mathfrak{a}}\Big)\Big].$$

Subtracting and adding the term, $\frac{1}{2}\Big(\frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}\upsilon_t\boldsymbol{z}_t'\Big)C'^{-1}(C'^{-1})'\Big(\frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}\boldsymbol{z}_t\upsilon_t\Big)$, inside the square parenthesis, and then proceeding as in part (a), we can show that

$$\hat{\boldsymbol{a}}_i(X_i, Z_i, \Theta_1) = [\sum_{t=1}^{T}\boldsymbol{z}_{it}\boldsymbol{z}_{it}' + \sigma_\epsilon^2\Sigma_a^{-1}]^{-1}\Big(\sum_{t=1}^{T}\boldsymbol{z}_{it}(x_{it} - \boldsymbol{z}_{it}'\boldsymbol{\alpha})\Big).$$

Lemma 2 *If (i) $\nexists\ A_x \subseteq \mathbb{R}^{d_x}$ such that $\Pr_{P_x}(A_x) = 1$ under P_x, where A_x is a proper linear subspace of $\mathbb{R}^{d_x}$; (ii) $\operatorname{rank}(\Pi) = d_x$, where $\Pi = \begin{pmatrix}\pi & \bar{\pi}\end{pmatrix}$; (iii) $\nexists\ A_z \subseteq \mathbb{R}^k$, where $k = \dim((\boldsymbol{z}_t', \bar{\boldsymbol{z}}')')$, such that $\Pr_{P_z}(A_z) = 1$ under P_z, where A_z is a proper linear subspace of $\mathbb{R}^k$; and (iv) if the covariance matrices of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\alpha}$ are of full rank, then there exists no $A \subseteq \mathbb{R}^{3d_x}$, $\mathbb{X}_t \in \mathbb{R}^{3d_x}$, such that A has probability 1 under $P_{\mathbb{X}}$ and A is a proper linear subspace of $\mathbb{R}^{3d_x}$.*

Proof 2

Now, condition (i) of the lemma is the "rank condition" for the standard probit model when $\boldsymbol{x}_{it}$ in (2.1) is exogenous and the object of interest is $\boldsymbol{\varphi}$ or marginal effects. The condition is assumed to hold true. Similarly, condition (iii) is the rank condition for the identification of the reduced form coefficients, $\Pi = \begin{pmatrix}\pi & \bar{\pi}\end{pmatrix}$, which is also assumed to hold true.

To show that the statement of the theorem is true, we have to show that:

$$\nexists\ A = \left\{ \begin{pmatrix} \boldsymbol{x}_t \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{pmatrix} \in \mathbb{R}^{3d_x} \middle| \boldsymbol{x}_t'\boldsymbol{c}_x + \hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\alpha + \hat{\boldsymbol{\epsilon}}'\boldsymbol{c}_\epsilon = 0, \begin{pmatrix} \boldsymbol{c}_x \\ \boldsymbol{c}_\alpha \\ \boldsymbol{c}_\epsilon \end{pmatrix} \neq 0 \right\} \text{ such that } \Pr_{P_{\mathbb{X}}}(A) = 1. \tag{A-7}$$

In other words, we have to show that $\boldsymbol{x}_t'\boldsymbol{c}_x + \hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\alpha + \hat{\boldsymbol{\epsilon}}'\boldsymbol{c}_\epsilon \neq 0$ almost surely ($a.s.$) whenever $(\boldsymbol{c}_x', \boldsymbol{c}_\alpha', \boldsymbol{c}_\epsilon')' \neq 0$.

To begin with, without loss of generality assume that $\boldsymbol{z}_t$ is uncorrelated with the individual effects $\boldsymbol{\alpha}$ so that $\bar{\pi}\bar{\boldsymbol{z}} = 0$. This implies that we can ignore $\bar{\boldsymbol{z}}$ in the reduced form equation (2.3) and consider only the dimension of $\boldsymbol{z}_t$, which is d_z, in condition (iii) of the lemma, and that

$$\Pi = \pi,\ \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{a}},\ \hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \pi\boldsymbol{z}_t - \hat{\boldsymbol{a}},\ k = d_x, \text{ and } \pi \text{ a } d_x \times d_z \text{ matrix.}$$

First, condition (i) of the lemma implies that

$$\text{for any } \boldsymbol{c} \in \mathbb{R}^{d_x} \text{ and } \boldsymbol{c} \neq 0,\ \boldsymbol{x}_t'\boldsymbol{c} \neq 0\ a.s.. \tag{A-8}$$

Thus in (A-7), $\boldsymbol{x}_t'\boldsymbol{c}_x \neq 0\ a.s.$.

By condition (ii), because rank of π' is d_x, for any $\boldsymbol{c} \in \mathbb{R}^{d_x}$ and $\boldsymbol{c} \neq 0$,

$$\pi'\boldsymbol{c} = \bar{\boldsymbol{c}} \neq 0. \tag{A-9}$$

By condition (iii) of the lemma, we have

$$\text{for any } \boldsymbol{c} \in \mathbb{R}^{d_z} \text{ and } \boldsymbol{c} \neq 0,\ \boldsymbol{z}_t'\boldsymbol{c} \neq 0\ a.s.. \tag{A-10}$$

Now, $\hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\alpha$ in (A-7) is

$$\begin{aligned} \hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\alpha &= \left([T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}\left(\sum_{t=1}^{T}\boldsymbol{v}_t\right)\right)'\boldsymbol{c}_\alpha = \left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\Sigma_{\epsilon\epsilon}^{-1}[T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\boldsymbol{c}_\alpha \\ &= \left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\bar{\boldsymbol{c}}_\alpha, \end{aligned}$$

where $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi\boldsymbol{z}_t$. Because $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$, the covariance matrices of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\alpha}$ respectively, are symmetric positive definite matrices, $\Sigma_{\epsilon\epsilon}^{-1}[T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}$ is nonsingular. This implies that $\Sigma_{\epsilon\epsilon}^{-1}[T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\boldsymbol{c}_\alpha = \bar{\boldsymbol{c}}_\alpha \neq 0$. By (A-8) then, $\boldsymbol{x}_t'\bar{\boldsymbol{c}}_\alpha \neq 0$, and by (A-9) and (A-10), $\boldsymbol{z}_t'\pi'\bar{\boldsymbol{c}}_\alpha \neq 0\ a.s.$. Thus $(\boldsymbol{x}_t' - \boldsymbol{z}_t'\pi')\bar{\boldsymbol{c}}_\alpha = \boldsymbol{v}_t'\bar{\boldsymbol{c}}_\alpha \neq 0$, when $\boldsymbol{x}_t \neq \pi\boldsymbol{z}_t\ a.s.$. Therefore,

$$\hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\alpha = \left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\bar{\boldsymbol{c}}_\alpha \neq 0\ a.s.. \tag{A-11}$$

We now show that $\hat{\boldsymbol{\epsilon}}_t'\boldsymbol{c}_\epsilon \neq 0$ *a.s.* when $\boldsymbol{c}_\epsilon \neq 0$. Now

$$\hat{\boldsymbol{\epsilon}}_t'\boldsymbol{c}_\epsilon = \boldsymbol{x}_t'\boldsymbol{c}_\epsilon - \boldsymbol{z}_t'\pi'\boldsymbol{c}_\epsilon - \hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\epsilon.$$

By condition (i) of the lemma, we know that $\boldsymbol{x}_t'\boldsymbol{c}_\epsilon \neq 0$ *a.s.*, by (A-8) we have that $\hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\epsilon \neq 0$ *a.s.*, and by (A-9) and (A-10), $\boldsymbol{z}_t'\pi'\boldsymbol{c}_\epsilon \neq 0$ *a.s.*. These imply that

$$\hat{\boldsymbol{\epsilon}}'\boldsymbol{c}_\epsilon \neq 0 \ a.s.. \tag{A-12}$$

Thus (A-8), (A-11), and (A-12) together imply that $\boldsymbol{x}_t'\boldsymbol{c}_x + \hat{\boldsymbol{\alpha}}'\boldsymbol{c}_\alpha + \hat{\boldsymbol{\epsilon}}'\boldsymbol{c}_\epsilon \neq 0$ *a.s.* whenever $(\boldsymbol{c}_x', \boldsymbol{c}_\alpha', \boldsymbol{c}_\epsilon')' \neq 0$.

Lemma 3 *The support of the conditional distribution of $\hat{\boldsymbol{\alpha}}(X, Z, \hat{\Theta}_1)$ and $\hat{\boldsymbol{\epsilon}}_t(X, Z, \hat{\Theta}_1)$, conditional on $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$, is same as the support of their marginal distribution.*

Proof 3

(a) We have shown that the expected value of $\boldsymbol{\alpha} = \bar{\pi}\bar{\boldsymbol{z}} + \boldsymbol{a}$ and $\boldsymbol{\epsilon}_t$ given Z and X, where $\boldsymbol{a}$ and $\boldsymbol{\epsilon}_t$ are normally distributed with variances $\Lambda_{\alpha\alpha}$ and $\Sigma_{\epsilon\epsilon}$ respectively, are given

$$\mathrm{E}(\boldsymbol{\alpha}|X, Z) = \hat{\boldsymbol{\alpha}} = \bar{\pi}\bar{\boldsymbol{z}} + \hat{\boldsymbol{a}} = \bar{\pi}\bar{\boldsymbol{z}} + \sum_{t=1}^{T} \mathit{\Omega}(\boldsymbol{x}_t - \Pi Z_t) \text{ and}$$
$$\mathrm{E}(\boldsymbol{\epsilon}_t|X, Z) = \hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \Pi Z_t - \sum_{t=1}^{T} \mathit{\Omega}(\boldsymbol{x}_t - \Pi Z_t) \text{ respectively,}$$

where $\Pi = (\bar{\pi}, \bar{\pi})$, $Z_t = (\boldsymbol{z}_t', \bar{\boldsymbol{z}}')'$ (see equation (2.3)), and $\mathit{\Omega} = [T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}$.

When $\boldsymbol{x}_t - \Pi Z_t \neq$ *constant* almost surely or when $\boldsymbol{z}_t$ has a small support, because $\boldsymbol{x}_t$'s have unbounded support in $\mathbb{R}^{d_x}$ and $\mathit{\Omega}$ is a $d_x \times d_x$ nonsingular matrix,

$$\mathrm{Supp}(\hat{\boldsymbol{\alpha}}) = \mathrm{Supp}(\hat{\boldsymbol{\epsilon}}_t) = \mathbb{R}^{d_x}.$$

Now fix $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$. Then, because the $\boldsymbol{x}_s$'s, $s \neq t$, are not restricted, we have

$$\mathrm{Supp}(\hat{\boldsymbol{\alpha}}|\boldsymbol{x}_t = \bar{\boldsymbol{x}}) = \mathrm{Supp}\left(\bar{\pi}\bar{\boldsymbol{z}} + \mathit{\Omega}(\bar{\boldsymbol{x}} - \Pi Z_t) + \sum_{s\neq t} \mathit{\Omega}(\boldsymbol{x}_s - \Pi Z_s)\right) = \mathbb{R}^{d_x} \text{ and}$$
$$\mathrm{Supp}(\hat{\boldsymbol{\epsilon}}_t|\boldsymbol{x}_t = \bar{\boldsymbol{x}}) = \mathrm{Supp}\left([I_m - \mathit{\Omega}](\bar{\boldsymbol{x}} - \Pi Z_t) - \sum_{s\neq t} \mathit{\Omega}(\boldsymbol{x}_s - \Pi Z_s)\right) = \mathbb{R}^{d_x}.$$

(b) When we have a single endogenous variable, x_t, given by

$$x_t = \Pi Z_t + a + \epsilon_t, t = 1, \ldots, T,$$

where the errors, $\boldsymbol{\epsilon} \equiv (\epsilon_1, \ldots, \epsilon_T)'$, are non-spherical such that $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \Omega$, a invertible $T \times T$ matrix, and a is normally distributed with variance σ_α^2, then we showed that

$$\hat{a}(X, Z, \Theta_1) = (x_1 - \Pi Z_1)\omega_1 + \ldots + (x_T - \Pi Z_T)\omega_T,$$

where $(\omega_1, \ldots, \omega_T)' = \frac{\Omega^{-1}e}{(e'\Omega^{-1}e + \sigma_\alpha^{-2})}$ and e is a vector of ones of dimension T.

Given that x_t's have large supports, using a similar argument as in part (a), we get

$$\mathrm{Supp}(\hat{a}|x_t = \bar{x}) = \mathbb{R} \text{ and } \mathrm{Supp}(\hat{\epsilon}_t|x_t = \bar{x}) = \mathbb{R}.$$

7. TABLES AND FIGURES

TABLE 1

Performance of the APE, $\frac{\partial G(x_{it})}{\partial x}\big|_{x_{it}=1}$, for alternative estimators.

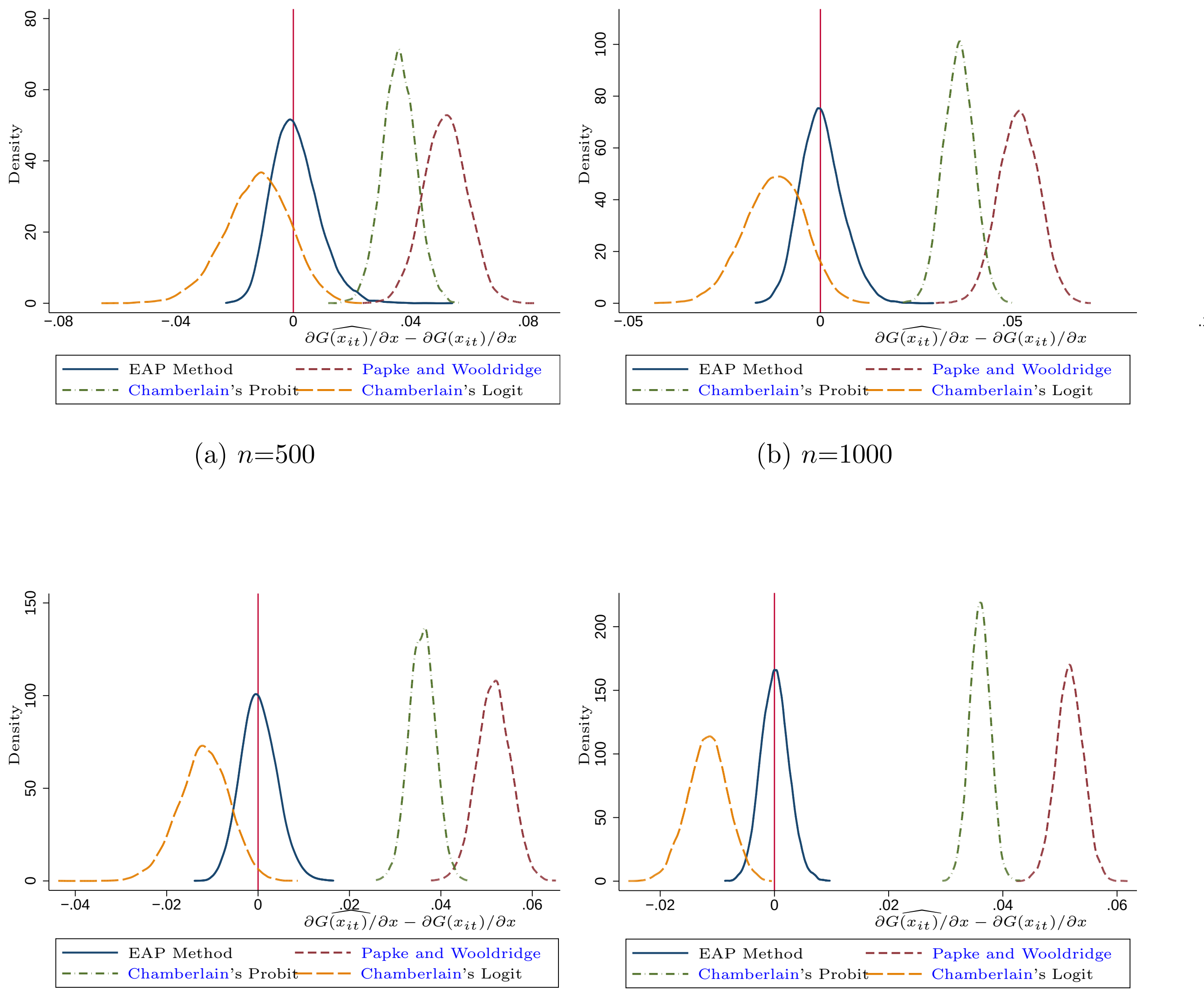
	EAP Method		Papke and Wooldridge		Chamberlain's CRE Probit		Chamberlain's Logit	
	RMSE	Median	RMSE	Median	RMSE	Median	RMSE	Median
n= 200	.01321	-.09358	.05312	-.04204	.03719	-.05771	.02276	-.10633
n= 500	.00808	-.09383	.05232	-.04216	.03645	-.05798	.01667	-.10582
n= 1000	.00555	-.09389	.05196	-.04225	.03620	-.05792	.01435	-.10571
n= 2000	.00392	-.09394	.05171	-.04238	.03599	-.05806	.01316	-.10576
n= 5000	.00243	-.09392	.05175	-.04226	.03603	-.05796	.01222	-.10561

TABLE 2

Work Status by Age Group

Year 2007				Year: 2010			
Age Group	Not Working	Working	Total	Age Group	Not Working	Working	Total
5 to 7 years	45.25	5.02	50.27	8 to 10 years	31.25	19.03	50.27
8 to 14 years	22.88	26.85	49.73	11 to 17 years	14.98	34.75	49.73
Total	68.13	31.87	100.00	Total	46.23	53.77	100.00

The figures are in percentage. Total number of children in each period: 2458



(a) n=500

(b) n=1000

(c) n=2000

(d) n=5000

Figure 1: Comparison with Alternative Estimators: Density of $\partial \widehat{G(x_{it})}/\partial x - \partial G(x_{it})/\partial x$ for different Sample Size.

TABLE 3

Descriptive Statistics

	2007		2010	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
Child characteristics				
Sex (Male=1, Female=0)	0.52	0.50	0.52	0.50
Age (yrs.)	8.07	2.97	11.07	2.97
Household characteristics				
Parents participated in NREGS (Yes=1 & No=0)	0.33	0.47	0.66	0.47
Total number of days parents worked in NREGS	9.21	21.44	36.00	48.10
Land Owned (acre)	2.32	3.42	3.86	43.53
Asset Index	-0.13	0.98	0.22	1.46
Gini Coefficient for Land Owned	0.62		0.74	
Total Income of Household (in Thousand ₹)	30.91	34.35	48.88	60.24
Annual non-agricultural income (₹)	20787	35813	29013	62225
Annual agricultural income (₹)	5060	23319	9936	42746
Does a household own farm assets (Yes=1 & No=0)	0.69	0.46	0.91	0.29
Number of farm assets	4.70	11.06	6.29	9.01
Community (Mandal) characteristics				
Total NREGS amount sanctioned (in Million ₹)	7.25	8.30	20.19	19.17
Infrastructure Variables				
Engineered Road to the Locality (Yes=1 & No=0)	0.32	0.47	0.58	0.49
Drinkable Water in the Locality (Yes=1 & No=0)	0.87	0.34	0.86	0.34
National Bank in the Locality (Yes=1 & No=0)	0.23	0.41	0.08	0.27
Hospital in the Locality (Yes=1 & No=0)	0.37	0.89	0.38	0.48

Total number of children/observations in each period: 2458

TABLE 4

Descriptive Statistics of some Variables by Caste

		Scheduled Castes/Tribes	Other Backward Classes	Others
Year: 2007	Household Income (in Thousand ₹)	31.22 (33.94)	31.64 (34.29)	43.21 (48.59)
	Land Owned in acre	1.58 (2.12)	2.32 (3.51)	3.08 (4.53)
	Index of Productive Farm Asset	-0.22 (0.71)	-0.14 (1.02)	0.04 (1.17)
	School Dummy $DSCHOOL = 1$	0.90 (0.29)	0.89 (0.32)	0.96 (0.19)
	Work Dummy $DWORK = 1$	0.33 (0.47)	0.33 (0.47)	0.29 (0.45)
Year: 2010	Household Income (in Thousand ₹)	45.99 (45.51)	50.22 (66.35)	64.76 (70.26)
	Land Owned in acre	2.10 (1.95)	2.79 (15.82)	10.90 (108.71)
	Index of productive Farm Asset	0.12 (1.16)	0.29 (1.56)	0.54 (1.89)
	School Dummy $DCHOOL = 1$	0.89 (0.31)	0.87 (0.33)	0.94 (0.23)
	Work Dummy $DWORK = 1$	0.52 (0.50)	0.57 (0.49)	0.48 (0.50)
Number of Children/observations in each period:		906	1269	283

Standard errors in parentheses.

TABLE 5

First Stage Reduced Form Estimates: Joint Estimation of Income, Land, and Wealth Equation

	Income	Landholding	Farm Asset
Total NREGS amount sanctioned (in Million ₹)	0.047***	-0.008	-0.0003
	(0.009)	(0.007)	(0.0002)
Caste (SC/ST = 1, OBC = 2, OT = 3)	9.220***	2.278***	0.171***
	(1.217)	(0.726)	(0.0300)
Drinkable Water in the Locality (Yes=1 & No=0)	5.417	-1.879	0.341**
	(5.703)	(4.260)	(0.150)
National Bank in the Locality (Yes=1 & No=0)	-2.785	4.684**	0.046
	(3.099)	(2.315)	(0.082)
Engineered Road to the Locality (Yes=1 & No=0)	0.159	2.413	0.182***
	(2.130)	(1.591)	(0.0561)
Hospital in the Locality (Yes=1 & No=0)	-0.689	-4.143***	0.056*
	(1.248)	(0.932)	(0.033)
Other Exogenous Variables of the Structural Equations: Age and Sex of the Child	Yes	Yes	Yes

Total number of observations : 5140

Biørn's Stepwise MLE was employed to obtain these estimates. All the specifications include time dummy, district dummies, and the interaction of time and district dummies.

Standard errors in parentheses

Significance levels : $*$: 10% $**$: 5% $***$: 1%

TABLE 6

Household Income and Wealth Effect on Child's Decision to Work

	CRE Probit	**Control Function Method**	
	Coefficients	**Coefficients**	**APEs**
Income	0.003*** (0.0008)	-0.0234*** (0.0028)	-0.0054*** (0.0008)
Landholding	0.002 (0.002)	0.031*** (0.007)	0.0071*** (0.0017)
Farm Asset Index	-0.011 (0.0279)	-0.976*** (0.169)	-0.226*** (0.0302)
Age	2.019*** (0.072)	0.402*** (0.057)	0.093*** (0.009)
Sex	0.644*** (0.042)	0.394*** (0.0473)	0.0908*** (0.0129)
		Control Functions	
		$\hat{\alpha}_{INCOME}$	0.005 (0.00302)
		$\hat{\alpha}_{LAND}$	-0.015** (0.0065)
		$\hat{\alpha}_{ASSET}$	1.512*** (0.129)
		$\hat{\epsilon}_{INCOME}$	0.0275*** (0.003)
		$\hat{\epsilon}_{LAND}$	-0.031*** (0.0075)
		$\hat{\epsilon}_{ASSET}$	0.882*** (0.185)

Total number of children: 2458
Total number of observations: 4916. Total number of observations with positive outcome: 2128
CRE Probit is Chamberlain's Correlated Random Effect model for panel data binary choice outcomes.
All the specifications include time dummy, district dummies, and the interaction of the two.
Standard errors (SE) in parentheses
Significance levels : $*$: 10% $**$: 5% $***$: 1%

SUPPLIMENTRY APPENDIX FOR PANEL DATA BINARY RESPONSE MODEL IN A TRIANGULAR SYSTEM

The supplementary appendix is not meant to be included with the main text of the paper.

APPENDIX A: ASYMPTOTIC COVARIANCE MATRIX FOR STRUCTURAL PARAMETERS

Though obtaining the parameters of the second stage, given the first stage consistent estimates $\hat{\Theta}_1$, is asymptotically equivalent to estimating the subsequent stage parameters had the true value of Θ_1^* been known, to obtain correct inference about the structural parameters, one has to account for the fact that instead of true values of first stage reduced form parameters, we use their estimated value. Here we are assuming that the first stage estimation involves the estimation of system of regression using Biørn's method and that in the second stage a probit model, using the method of multivariate weighted nonlinear least squares (MWNLS), is estimated.

Newey (1984) has shown that sequential estimators can be interpreted as members of a class of Method of Moments (MM) estimators and that this interpretation facilitates derivation of asymptotic covariance matrices for multi-step estimators. Let $\Theta = (\Theta_1', \Theta_2')'$, where Θ_1 and Θ_2 are respectively the parameters to be estimated in the first and second step estimation of the sequential estimator. Following Newey (1984) we write the first and second step estimation as an MM estimation based on the following population moment conditions:

$$\mathrm{E}(\mathcal{L}_{i\Theta_1}) = E\frac{\partial \ln L_i(\Theta_1)}{\partial \Theta_1} = 0$$

$$\mathrm{E}(H_{i\Theta_2}(\Theta_1, \Theta_2)) = 0$$

and where $L_i(\Theta_1)$ is the likelihood function for individual i for the first step system of reduced form equations and $\mathrm{E}(H_{i\Theta_2}(\Theta_1, \Theta_2))$ is the population moment condition for estimating Θ_2 given Θ_1.

The estimates for Θ_1 and Θ_2 are obtained by solving the sample analog of the above population moment conditions. The sample analog of moment conditions for the first step estimation is given by

$$\frac{1}{N}\mathcal{L}_{\Theta_1}(\hat{\Theta}_1) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathcal{L}_i(\hat{\Theta}_1)}{\partial \Theta_1} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln L_i(\hat{\Theta}_1)}{\partial \Theta_1}$$

where $\mathcal{L}_i(\Theta_1)$ and the first order conditions with respect to $\Theta_1 = (\boldsymbol{\delta}', \mathrm{vec}(\Lambda_{\alpha\alpha})', \mathrm{vec}(\Sigma_{\epsilon\epsilon})')'$[1] are given in appendix B of this supplementary appendix, and N is the total number of individuals.

[1] While we have written our reduced form equation as

$$\boldsymbol{x}_{it} = \pi \boldsymbol{z}_{it} + \bar{\pi}\bar{\boldsymbol{z}}_i + \boldsymbol{a}_i + \boldsymbol{\epsilon}_{it}, \text{ Biørn writes it as } \boldsymbol{x}_{it} = Z_{it}'\boldsymbol{\delta} + \boldsymbol{a}_i + \boldsymbol{\epsilon}_{it},$$

where $Z_{it} = \mathrm{diag}((\boldsymbol{z}_{it}', \bar{\boldsymbol{z}}_i')', \ldots, (\boldsymbol{z}_{it}', \bar{\boldsymbol{z}}_i')')$ and $\boldsymbol{\delta} = (\mathrm{vec}(\pi)', \mathrm{vec}(\bar{\pi})')'$.

The sample analog of population moment condition for the second step estimation is given by

$$\frac{1}{N}H_{\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2) = \frac{1}{N}\sum_{i=1}^{N} H_{i\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2).$$

We have shown that the structural equations augmented with the control functions $\hat{\boldsymbol{\alpha}}_i(X_i, Z_i, \Theta_1)$ and $\hat{\boldsymbol{\epsilon}}_{it}(X_i, Z_i, \Theta_1)$ leads to the identification of Θ_2. Let Θ_2^* be the true values of Θ_2. Under the assumptions we make, solving $\frac{1}{N}\sum_{i=1}^{N} H_{it\Theta_2}(\hat{\Theta}_1, \Theta_2) = 0$ is asymptotically equivalent to solving $\frac{1}{N}\sum_{i=1}^{N} H_{it\Theta_2}(\Theta_1^*, \Theta_2) = 0$, where $\hat{\Theta}_1$ is a consistent first step estimate of Θ_1. Hence $\hat{\Theta}_2$ obtained by solving $\frac{1}{N}H_{\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2) = 0$ is a consistent estimate of Θ_2.

To derive the asymptotic distribution of the second step estimates $\hat{\Theta}_2$, consider the stacked up sample moment conditions:

$$\frac{1}{N}\begin{bmatrix} \mathcal{L}_{\Theta_1}(\hat{\Theta}_1) \\ H_{\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2) \end{bmatrix} = 0. \tag{A-1}$$

A series of Taylor's expansion of $\mathcal{L}_{\Theta_1}(\hat{\Theta}_1)$, $H_{\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2)$ and around Θ^* gives

$$\frac{1}{N}\begin{bmatrix} \mathcal{L}_{\Theta_1\Theta_1} & 0 \\ H_{\Theta_2\Theta_1} & H_{\Theta_2\Theta_2} \end{bmatrix}\begin{bmatrix} \sqrt{N}(\hat{\Theta}_1 - \Theta_1^*) \\ \sqrt{N}(\hat{\Theta}_2 - \Theta_2^*) \end{bmatrix} = -\frac{1}{\sqrt{N}}\begin{bmatrix} \mathcal{L}_{\Theta_1} \\ H_{\Theta_2}. \end{bmatrix} \tag{A-2}$$

In matrix notation the above can be written as

$$B_{\Theta\Theta_N}\sqrt{N}(\hat{\Theta} - \Theta) = -\frac{1}{\sqrt{N}}\Lambda_{\Theta_N},$$

where Λ_{Θ_N} is evaluated at Θ^* and $B_{\Theta\Theta_N}$ is evaluated at points somewhere between $\hat{\Theta}$ and Θ^*. Under the standard regularity conditions for Generalized Method of Moments (GMM) (see Newey, 1984) $B_{\Theta\Theta_N}$ converges in probability to the lower block triangular matrix $B_* = \lim \mathrm{E}(B_{\Theta\Theta_N})$. B_* is given by

$$B_* = \begin{bmatrix} \mathbb{L}_{\Theta_1\Theta_1} & 0 \\ \mathbb{H}_{\Theta_2\Theta_1} & \mathbb{H}_{\Theta_2\Theta_2} \end{bmatrix}$$

where $\mathbb{L}_{\Theta_1\Theta_1} = \mathrm{E}(\mathcal{L}_{i\Theta_1\Theta_1})$, $\mathbb{H}_{\Theta_2\Theta_1} = \mathrm{E}(H_{i\Theta_2\Theta_1})$. $\frac{1}{\sqrt{N}}\Lambda_N$ converges asymptotically in distribution to a normal random variable with mean zero and a covariance matrix $A_* = \lim \mathrm{E}\frac{1}{N}\Lambda_N\Lambda_N'$, where A_* is given by

$$A_* = \begin{bmatrix} V_{LL} & V_{LH} \\ V_{HL} & V_{HH} \end{bmatrix},$$

and a typical element of A_*, say V_{LH}, is given by $V_{LH} = \mathrm{E}[\mathcal{L}_{i\Theta_1}(\Theta_1)H_{i\Theta_2}(\Theta_1, \Theta_2)']$. Under the regularity conditions $\sqrt{N}(\hat{\Theta} - \Theta^*)$ is asymptotically normal with zero mean and covariance matrix given by $B_*^{-1}A_*B_*^{-1\prime}$, that is

$$\sqrt{N}(\hat{\Theta} - \Theta^*) \overset{a}{\sim} \mathrm{N}[(0), (B_*^{-1}A_*B_*^{-1\prime})]. \tag{A-3}$$

By an application of partitioned inverse formula and some matrix manipulation we get the asymptotic covariance matrix of $\sqrt{N}(\hat{\Theta}_2 - \Theta_2)$, V_2^* , where

$$
\begin{aligned}
V_2^* =& \mathbb{H}^{-1}_{\Theta_2\Theta_2} V_{HH} \mathbb{H}^{-1}_{\Theta_2\Theta_2} + \mathbb{H}^{-1}_{\Theta_2\Theta_2} \mathbb{H}^{-1}_{\Theta_2\Theta_1} \{\mathbb{L}^{-1}_{\Theta_1\Theta_1} V_{LL} \mathbb{L}^{1\prime}_{\Theta_1\Theta_1}\} \mathbb{H}^{1\prime}_{\Theta_2\Theta_1} \mathbb{H}^{-1\prime}_{\Theta_2\Theta_2} \\
&- \mathbb{H}^{-1}_{\Theta_2\Theta_2} \{\mathbb{H}_{\Theta_2\Theta_1} \mathbb{L}^{-1}_{\Theta_1\Theta_1} V_{LH} + V_{HL} \mathbb{L}^{-1\prime}_{\Theta_1\Theta_1} \mathbb{H}'_{\Theta_2\Theta_1}\} \mathbb{H}^{-1\prime}_{\Theta_2\Theta_2}
\end{aligned} \tag{A-4}
$$

To estimate V_2^*, sample analog of the B^*, B_N given in (A-2), and sample analog of A^*, $A_N = \frac{1}{N}\Lambda_N\Lambda_N$, have to be computed. A typical element of A_N, say V_{LH_N} , is given by $V_{LH_N} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{i\Theta_1}(\Theta_1) H_{i\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2)'$. The first and the second order conditions of the first-stage likelihood function for estimating Θ_1, which are used to compute the sample analog of $\mathbb{L}_{\Theta_1\Theta_1}$ and to compute A_N, are provided in appendix B of this supplementary appendix.

For binary response model, the score function pertaining to the minimand in equation (2.14) of the main text is given by

$$
\begin{aligned}
H_{i\Theta_2}(\Theta_1, \Theta_2) &= -\nabla_{\Theta_2} \mathbf{m}_i(X_i, Z_i, \Theta_2)' [\mathbf{V}(X_i, Z_i, \tilde{\Upsilon})]^{-1} [\mathbf{y}_i - \mathbf{m}_i(X_i, Z_i, \Theta_2)] \\
&= -\nabla_{\Theta_2} \mathbf{m}_i(\Theta_1, \Theta_2)' \tilde{\mathbf{V}}^{-1} \mathbf{u}_i,
\end{aligned}
$$

where $\mathbf{m}_i(\Theta_1, \Theta_2) \equiv \mathbf{m}_i(X_i, Z_i, \Theta_2)$ is a T vector with t^{th} element being $\mathbf{m}(\mathbb{X}_{it}, \Theta_2) = \Phi(\boldsymbol{x}'_{it}\boldsymbol{\varphi} + \boldsymbol{\rho}_\alpha \hat{\boldsymbol{\alpha}}_i + \boldsymbol{\rho}_\epsilon \hat{\boldsymbol{\epsilon}}_{it}) \equiv \mathbf{m}_{it}(\Theta_1, \Theta_2)$, $\mathbf{u}_i$ is a T vector with t^{th} element being $y_{it} - \mathbf{m}_{it}(\Theta_1, \Theta_2)$, and $\tilde{\mathbf{V}} \equiv \mathbf{V}(X_i, Z_i, \tilde{\Upsilon})$. Now

$$
\nabla_{\Theta_2} \mathbf{m}_{it}(\Theta_1, \Theta_2) = \phi(\mathbb{X}'_{it}\Theta_2)\mathbb{X}'_{it}
$$

where $\mathbb{X}_{it} = (\boldsymbol{x}'_{it}, \hat{\boldsymbol{\alpha}}'_i(\Theta_1), \hat{\boldsymbol{\epsilon}}'_{it}(\Theta_1))'$ and $\Theta_2 = (\boldsymbol{\varphi}', \boldsymbol{\rho}'_\alpha, \boldsymbol{\rho}'_\epsilon)'$.

Wooldridge (2002) and Wooldridge (2003) show (see Problem 12.11) that $\mathbb{H}_{\Theta_2\Theta_2}$ of B^* is given by

$$
\mathbb{H}_{\Theta_2\Theta_2} = \mathrm{E}[H_{i\Theta_2\Theta_2}(\Theta_1, \Theta_2)] = \mathrm{E}[\nabla_{\Theta_2} \mathbf{m}_i(\Theta_1, \Theta_2)' \tilde{\mathbf{V}}^{-1} \nabla_{\Theta_2} \mathbf{m}_i(\Theta_1, \Theta_2)],
$$

which can be approximated as

$$
\frac{1}{N}\sum_{i=1}^{N} \nabla_{\Theta_2} \mathbf{m}_i(\hat{\Theta}_1, \hat{\Theta}_2)' \hat{\mathbf{V}}^{-1} \nabla_{\Theta_2} \mathbf{m}_i(\hat{\Theta}_1, \hat{\Theta}_2),
$$

where $\hat{\mathbf{V}} = \mathbf{V}(X_i, Z_i, \hat{\Upsilon}) = \mathbf{V}(X_i, Z_i, \hat{\Theta}_2, \hat{\rho})$.

We now compute $H_{\Theta_2\Theta_1} = \sum_{i=1}^{N} H_{i\Theta_2\Theta_1} = \sum_{i=1}^{N} \frac{\partial H_{i\Theta_2}(\Theta_1, \Theta_2)}{\partial \Theta'_1}$ in order to compute the sample analog of $\mathbb{H}_{\Theta_2\Theta_1}$. Now,

$$
\begin{aligned}
\frac{\partial H_{i\Theta_2}(\Theta_1, \Theta_2)}{\partial \Theta'_1} = -\Bigg[& [\mathbf{u}'_i \tilde{\mathbf{V}}^{-1} \otimes I] \frac{\partial \mathrm{vec}(\nabla_{\Theta_2} \mathbf{m}_i(\Theta_1, \Theta_2)')}{\partial \Theta'_1} \\
&+ [\mathbf{u}_i \otimes \nabla_{\Theta_2} \mathbf{m}_i(\Theta_1, \Theta_2)'] \frac{\partial \mathrm{vec}(\tilde{\mathbf{V}}^{-1})}{\partial \Theta'_1} \\
&- \nabla_{\Theta_2} \mathbf{m}_i(\Theta_1, \Theta_2)' \tilde{\mathbf{V}}^{-1} \nabla_{\Theta_1} \mathbf{m}_i(\Theta_1, \Theta_2) \Bigg].
\end{aligned}
$$

Taking expectation of the above, we find that the first two terms are zero. Hence we have

$$\mathbb{H}_{\Theta_2\Theta_1} = \mathrm{E}[H_{i\Theta_2\Theta_1}(\Theta_1,\Theta_2)] = \mathrm{E}[\nabla_{\Theta_2}\mathbf{m}_i(\Theta_1,\Theta_2)'\tilde{\mathbf{V}}^{-1}\nabla_{\Theta_1}\mathbf{m}_i(\Theta_1,\Theta_2)],$$

which can be approximated by

$$\frac{1}{N}\sum_{i=1}^{N}\nabla_{\Theta_2}\mathbf{m}_i(\hat{\Theta}_1,\hat{\Theta}_2)'\hat{\mathbf{V}}^{-1}\nabla_{\Theta_1}\mathbf{m}_i(\hat{\Theta}_1,\hat{\Theta}_2).$$

The constituents, $\nabla_{\Theta_1}\mathbf{m}_{it}(\Theta_1,\Theta_2)$, of $\nabla_{\Theta_1}\mathbf{m}_i(\Theta_1,\Theta_2)$ are given by

$$\nabla_{\Theta_1}\mathbf{m}_{it}(\Theta_1,\Theta_2) = \phi(\mathbb{X}_{it}'\Theta_2)\Theta_2'\frac{\partial\mathbb{X}_{it}}{\partial\Theta_1'},$$

which is row matrix with dimension that of Θ_1, and where

$$\frac{\partial\mathbb{X}_{it}}{\partial\Theta_1'} = \begin{bmatrix} \frac{\partial\boldsymbol{x}_{it}}{\partial\boldsymbol{\delta}'} & \frac{\partial\boldsymbol{x}_{it}}{\partial\mathrm{vec}(\Lambda_{\alpha\alpha})'} & \frac{\partial\boldsymbol{x}_{it}}{\partial\mathrm{vec}(\Sigma_{\epsilon\epsilon})'} \\ \frac{\partial\hat{\boldsymbol{\alpha}}_i}{\partial\boldsymbol{\delta}'} & \frac{\partial\hat{\boldsymbol{\alpha}}_i}{\partial\mathrm{vec}(\Lambda_{\alpha\alpha})'} & \frac{\partial\hat{\boldsymbol{\alpha}}_i}{\partial\mathrm{vec}(\Sigma_{\epsilon\epsilon})'} \\ \frac{\partial\hat{\boldsymbol{\epsilon}}_{it}}{\partial\boldsymbol{\delta}'} & \frac{\partial\hat{\boldsymbol{\epsilon}}_{it}}{\partial\mathrm{vec}(\Lambda_{\alpha\alpha})'} & \frac{\partial\hat{\boldsymbol{\epsilon}}_{it}}{\partial\mathrm{vec}(\Sigma_{\epsilon\epsilon})'} \end{bmatrix}.$$

Since $\boldsymbol{x}_{it}$ is not a function of Θ_1, $\frac{\partial\boldsymbol{x}_{it}}{\partial\Theta_1'} = \mathbf{0}_{\boldsymbol{x}}$, where $\mathbf{0}_{\boldsymbol{x}}$ is a null matrix with row dimension that of column vector $\boldsymbol{x}_{it}$ and column dimension that of column vector Θ_1. Using the following matrix results:

$$\partial\mathrm{vec}(\Omega\boldsymbol{b}) = (\boldsymbol{b}'\otimes I_m)\partial\mathrm{vec}(\Omega),\ \ \partial\mathrm{vec}(\Omega^{-1}) = -(\Omega'^{-1}\otimes\Omega^{-1})\partial\mathrm{vec}(\Omega) \text{ and}$$
$$\frac{\partial\mathrm{vec}(\Omega)}{\partial\mathrm{vec}(\Omega)} = I_{mm},$$

where $\boldsymbol{b}$ is a vector of dimension m, Ω is a symmetric $m\times m$ matrix and I_{mm} is the $mm\times mm$ identity matrix, it can be shown that

$$\frac{\partial\hat{\boldsymbol{\alpha}}_i}{\partial\boldsymbol{\delta}'} = \frac{\partial(\mathrm{diag}(\bar{\boldsymbol{z}}_i',\ldots,\bar{\boldsymbol{z}}_i')'\mathrm{vec}(\bar{\pi}) + \hat{\boldsymbol{a}}_i)}{\partial\boldsymbol{\delta}'} = \mathbb{O}_{Zi}' - [T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}Z_{it}',$$
$$\frac{\partial\hat{\boldsymbol{\epsilon}}_{it}}{\partial\boldsymbol{\delta}'} = \frac{\partial(\boldsymbol{x}_{it} - Z_{it}'\boldsymbol{\delta} - \hat{\boldsymbol{a}}_i)}{\partial\boldsymbol{\delta}'} = -Z_{it}' + [T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}\Sigma_{\epsilon\epsilon}^{-1}Z_{it}',$$

$$\frac{\partial\hat{\boldsymbol{\alpha}}_i}{\partial\mathrm{vec}(\Lambda_{\alpha\alpha})'} = -\left(\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\otimes I_m\right)\left[\left(\Sigma_{\epsilon\epsilon}^{-1}\otimes I_m\right)\left(\Sigma'\otimes\Sigma\right)\right]I_{mm},$$

$$\frac{\partial\hat{\boldsymbol{\alpha}}_i}{\partial\mathrm{vec}(\Sigma_{\epsilon\epsilon})'} = -\left(\left(\sum_{t=1}^{T}\boldsymbol{v}_t'\right)\otimes I_m\right)\left[\left(I_m\otimes\Sigma\right)\left(\Sigma_{\epsilon\epsilon}^{-1}\otimes\Sigma_{\epsilon\epsilon}^{-1}\right)+\left(\Sigma_{\epsilon\epsilon}^{-1}\otimes I_m\right)\left(\Sigma'\otimes\Sigma\right)\right]TI_{mm},$$

$$\frac{\partial \hat{\boldsymbol{\epsilon}}_{it}}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} = \frac{-\partial \hat{\boldsymbol{\alpha}}_i}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}, \text{ and } \frac{\partial \hat{\boldsymbol{\epsilon}}_{it}}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{-\partial \hat{\boldsymbol{\alpha}}_i}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'},$$

where $\mathbb{O}_{Zi} = \text{diag}((0_z', \bar{\boldsymbol{z}}_i')', \ldots, (0_z', \bar{\boldsymbol{z}}_i')')$, 0_z denoting a vector of zeros with dimension that of $\boldsymbol{z}_{it}$, $\boldsymbol{v}_t = \boldsymbol{x}_t - \pi \boldsymbol{z}_t$, and $\Sigma = [T\Sigma_{\epsilon\epsilon}^{-1} + \Lambda_{\alpha\alpha}^{-1}]^{-1}$.

Since $H_{i\Theta_2\Theta_1}(\hat{\Theta}_1, \hat{\Theta}_2)$ and $H_{i\Theta_2\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2)$ converge almost surly to $H_{i\Theta_2\Theta_1}(\Theta_1^*, \Theta_2^*)$ and $H_{i\Theta_2\Theta_2}(\Theta_1^*, \Theta_2^*)$ respectively, by the weak LLN $\frac{1}{N}\sum_{i=1}^N H_{i\Theta_2\Theta_1}(\hat{\Theta}_1, \hat{\Theta}_2)$ will converge in probability to $\text{E}(H_{i\Theta_2\Theta_1}(\Theta_1^*, \Theta_2^*)) = \mathbb{H}_{\Theta_2\Theta_1}$ and $\frac{1}{N}\sum_{i=1}^N H_{i\Theta_2\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2)$ will converge in probability to $\text{E}(H_{i\Theta_2\Theta_2}(\Theta_1^*, \Theta_2^*)) = \mathbb{H}_{\Theta_2\Theta_2}$.

A.1. *Hypothesis Testing of Average Partial Effects*

In section 2 we discussed the estimation of average partial effect (APE) of a variable w. To test various hypothesis in order to draw inferences about the APE's we need to compute the standard errors of their estimates. From (2.15) of the main text we know that the estimated APE of w on the probability of $y_{it} = 1$ given $\boldsymbol{x}_{it} = \bar{\boldsymbol{x}}$ is given by

$$\frac{\partial \widehat{\Pr}(y_{it} = 1|\bar{\boldsymbol{x}})}{\partial w} = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \hat{\varphi}_w \phi\left(\bar{\mathbb{X}}_{it}' \hat{\Theta}_2\right) \equiv \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T g_{wit}(\hat{\Theta}_2),$$

where $\bar{\mathbb{X}}_{it} = (\bar{\boldsymbol{x}}', \hat{\boldsymbol{\alpha}}_i(\hat{\Theta}_1)', \hat{\hat{\boldsymbol{\epsilon}}}_{it}(\hat{\Theta}_1)')'$ and $\hat{\Theta}_2 = (\hat{\boldsymbol{\varphi}}', \hat{\boldsymbol{\rho}}_\alpha, \hat{\boldsymbol{\rho}}_\epsilon)'$. Now, we know that by the linear approximation approach (delta method), the asymptotic variance of $\frac{\partial \widehat{\Pr}(y_{it}=1|\bar{\boldsymbol{x}})}{\partial w}$ can be estimated by computing

$$\left[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}_2'}\right] \hat{V}_2^* \left[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}_2'}\right]', \tag{A-5}$$

where $\hat{V}_2^*$, the second stage error adjusted covariance matrix of Θ_2 estimated at $\hat{\Theta}_2$, is given in (A-4). $\frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}_2'}$ in (A-5) turns out to be

$$\frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}_2'} = \phi(\bar{\mathbb{X}}_{it}'\hat{\Theta}_2)[e_w - \hat{\varphi}_w(\bar{\mathbb{X}}_{it}'\hat{\Theta}_2)\bar{\mathbb{X}}_{it}]$$

where e_w is a column vector having the dimension of Θ_2' and with 1 at the position of φ_w in Θ_2 and zeros elsewhere.

If w is a dummy variable then the estimated APE of w is given by

$$\begin{aligned}
\Delta_w \Pr(y_{it} = 1) &= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \Phi(\bar{\boldsymbol{x}}_{-w}, w = 1, \hat{\hat{\boldsymbol{\alpha}}}_i, \hat{\hat{\boldsymbol{\epsilon}}}_{it}) - \Phi(\bar{\boldsymbol{x}}_{-w}, w = 0, \hat{\hat{\boldsymbol{\alpha}}}_i, \hat{\hat{\boldsymbol{\epsilon}}}_{it}) \\
&= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \Phi_{it}(w = 1) - \Phi_{it}(w = 0) \\
&= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \Delta_w \Phi_{it}().
\end{aligned}$$

To obtain the variance of the above, again by the delta method we have the estimate of the asymptotic variance of $\Delta_w \Pr(y_{it} = 1)$ given by

$$\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial \Delta\Phi_{it}(.)}{\partial \Theta_2'}\right]\hat{V}_2^*\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial \Delta\Phi_{it}(.)}{\partial \Theta_2'}\right]', \tag{A-6}$$

where

$$\begin{aligned}\frac{\partial \Delta\Phi_{it}(.)}{\partial \Theta_2'} &= \frac{\partial \Phi_{it}(w=1)}{\partial \Theta_2'} - \frac{\partial \Phi_{it}(w=0)}{\partial \Theta_2'}\\ &= \phi_{it}(w=1)\begin{bmatrix}\bar{\mathbb{X}}_{it-w}\\ 1\end{bmatrix}' - \phi_{it}(w=0)\begin{bmatrix}\bar{\mathbb{X}}_{it-w}\\ 0\end{bmatrix}'.\end{aligned}$$

APPENDIX B: MAXIMUM LIKELIHOOD ESTIMATION OF THE REDUCED FORM EQUATIONS

In this section we briefly describe Biørn (2004) step wise maximum likelihood procedure to estimate the reduced form system of equation

$$\boldsymbol{x}_{it} = Z_{it}'\boldsymbol{\delta} + \boldsymbol{a}_i + \boldsymbol{\epsilon}_{it}, \tag{B-1}$$

where $Z_{it} = \mathrm{diag}((\boldsymbol{z}_{it}', \bar{\boldsymbol{z}}_i')', \ldots, (\boldsymbol{z}_{it}', \bar{\boldsymbol{z}}_i')')$ and $\boldsymbol{\delta} = (\mathrm{vec}(\pi)', \mathrm{vec}(\bar{\pi})')'$. While Biørn (2004) deals with unbalanced panel, here we assume that our panel is balanced. Let N be the total number of individuals. Let $\mathcal{N}$ be the total number of observations, i.e., $\mathcal{N} = NT$. Let $\boldsymbol{x}_{i(T)} = (\boldsymbol{x}_{i1}', \ldots \boldsymbol{x}_{ip}')'$, $Z_{i(T)} = (Z_{i1}', \ldots Z_{iT}')'$ and $\boldsymbol{\epsilon}_{i(T)} = (\boldsymbol{\epsilon}_{i1}', \ldots \boldsymbol{\epsilon}_{iT}')'$ and write the model as

$$\boldsymbol{x}_{i(T)} = Z_{i(T)}'\boldsymbol{\delta} + (e_p \otimes \boldsymbol{a}_i) + \boldsymbol{\epsilon}_{i(T)} = Z_{i(T)}'\boldsymbol{\delta} + \boldsymbol{u}_{i(T)}, \tag{B-2}$$

Now,

$$\mathrm{E}(\boldsymbol{u}_{i(T)}\boldsymbol{u}_{i(T)}') = I_T \otimes \Sigma_{\epsilon\epsilon} + E_T \otimes \Lambda_{\alpha\alpha} = K_T \otimes \Sigma_{\epsilon\epsilon} + J_T \otimes \Sigma_{(T)} = \Omega_{u(T)}$$

where

$$\Sigma_{(T)} = \Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha},$$

where I_T is the T dimensional identity matrix, e_T is the $(T \times 1)$ vector of ones, $E_T = e_T e_T'$, $J_T = (1/T)E_T$, and $K_T = I_T - J_T$. The latter two matrices are symmetric and idempotent and have orthogonal columns, which facilitates inversion of $\Omega_{u(T)}$.

B.1. *GLS estimation*

Before addressing the maximum likelihood problem, we consider the GLS problem for $\boldsymbol{\delta}$ when Λ_α and $\Sigma_{\epsilon\epsilon}$ are known. Define $Q_{i(T)} = \boldsymbol{u}_{i(T)}'\Omega_{u(T)}^{-1}\boldsymbol{u}_{i(T)}$, then GLS estimation is the problem of minimizing $Q = \sum_{i=1}^{N} Q_{i(T)}$ with respect to $\boldsymbol{\delta}$. Since $\Omega_{u(T)}^{-1} = K_T \otimes \Sigma_{\epsilon\epsilon}^{-1} + J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}$, we can rewrite Q as

$$Q = \sum_{i=1}^{N} \boldsymbol{u}'_{i(T)}[K_T \otimes \Sigma_{\epsilon\epsilon}^{-1}]\boldsymbol{u}_{i(T)} + \sum_{i=1}^{N} \boldsymbol{u}'_{i(T)}[J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}]\boldsymbol{u}_{i(T)}.$$

GLS estimator of $\boldsymbol{\delta}$ when $\Lambda_{\alpha\alpha}$ and $\Sigma_{\epsilon\epsilon}$ are known is obtained from $\partial Q/\partial\boldsymbol{\delta} = 0$, and is given by

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{GLS} = & \Bigg[\sum_{i=1}^{N} Z'_{i(T)}[K_T \otimes \Sigma_{\epsilon\epsilon}^{-1}]Z_{i(T)} + \sum_{i=1}^{N} Z'_{i(T)}[J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}]Z_{i(T)}\Bigg]^{-1} \times \\ & \Bigg[\sum_{i=1}^{N} Z'_{i(T)}[K_T \otimes \Sigma_{\epsilon\epsilon}^{-1}]\boldsymbol{x}_{i(T)} + \sum_{i=1}^{N} Z'_{i(T)}[J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}]\boldsymbol{x}_{i(T)}\Bigg].\end{aligned} \tag{B-3}$$

B.2. *Maximum Likelihood Estimation*

Now consider ML estimation of $\boldsymbol{\delta}$, $\Sigma_{\epsilon\epsilon}$, and $\Lambda_{\alpha\alpha}$. Assuming normality of the individual effects and the disturbances, i.e., $\boldsymbol{a}_i \sim \mathrm{IIN}(0, \Lambda_{\alpha\alpha})$ and $\boldsymbol{\epsilon}_{it} \sim \mathrm{IIN}(0, \Sigma_{\epsilon\epsilon})$, then $\boldsymbol{u}_{i(T)} = (e_T \otimes \boldsymbol{a}_i) + \boldsymbol{\epsilon}_{i(T)} \sim \mathrm{IIN}(0_{mT,1}, \Omega_{u(T)})$. The log-likelihood functions of all $\boldsymbol{x}$'s conditional on all $\boldsymbol{Z}$'s for an individual and for all individuals in the data set then become, respectively,

$$\mathcal{L}_i = \frac{-mT}{2}\ln(2\pi) - \frac{1}{2}\ln|\Omega_{u(T)}| - \frac{1}{2}Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}), \tag{B-4}$$

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_i = \frac{-mNT}{2}\ln(2\pi) - \frac{1}{2}N\ln|\Omega_{u(T)}| - \frac{1}{2}\sum_{i=1}^{N} Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}), \tag{B-5}$$

where

$$Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}) = [\boldsymbol{x}_{i(T)} - Z'_{i(T)}\boldsymbol{\delta}]'[K_T \otimes \Sigma_{\epsilon\epsilon}^{-1} + J_T \otimes (\Sigma_{\epsilon\epsilon} + p\Lambda_{\alpha\alpha})^{-1}][\boldsymbol{x}_{i(T)} - Z'_{i(T)}\boldsymbol{\delta}],$$

and $|\Omega_{u(T)}| = |\Sigma_{(T)}||\Sigma_{\epsilon\epsilon}|^{T-1}$.

Biørn splits the problem of estimation into: (A) *Maximization of $\mathcal{L}$ with respect to $\boldsymbol{\delta}$ for given $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$* and (B) *Maximization of $\mathcal{L}$ with respect to $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$ for given $\boldsymbol{\delta}$*. *Subproblem* (A) is identical with the GLS problem, since maximization of $\mathcal{L}$ with respect to $\boldsymbol{\delta}$ for given $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$ is equivalent to minimization of $\sum_{i}^{N} Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha})$, which gives (B-3). To solve *subproblem*(B) Biørn derives expressions for the derivatives of both $\mathcal{L}_i$ and $\mathcal{L}$ with respect to $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$. The complete stepwise algorithm for solving jointly subproblems (A) and (B) then consists in switching between (B-3) and minimizing (B-5) with respect to $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$ to obtain $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$ and iterating until convergence.

The first order conditions for the log-likelihood function for an individual i with respect to $\boldsymbol{\delta}$, vech$(\Sigma_{\epsilon\epsilon})$ and vech$(\Lambda_{\alpha\alpha})$ are:

$$\frac{\partial \mathcal{L}_i}{\partial \boldsymbol{\delta}} = [\boldsymbol{x}_{i(T)} Z'_{i(T)}\boldsymbol{\delta}]'[K_T \otimes \Sigma^{1}_{\epsilon\epsilon} + J_T \otimes (\Sigma_{\epsilon\epsilon} + p\Sigma_{\alpha\alpha})^{1}]Z'_{i(T)},$$

$$\frac{\partial \mathcal{L}_i}{\partial \mathrm{vech}(\Sigma_{\epsilon\epsilon})} = -\frac{1}{2}L_m \mathrm{vec}\left[\Sigma^{-1}_{(T)} + (T-1)\Sigma^{-1}_{\epsilon\epsilon} - \Sigma^{-1}_{(T)}B_{ui(T)}\Sigma^{-1}_{(T)} - \Sigma^{-1}_{\epsilon\epsilon}W_{ui(T)}\Sigma^{-1}_{\epsilon\epsilon}\right],$$

and

$$\frac{\partial \mathcal{L}_i}{\partial \mathrm{vech}(\Lambda_{\alpha\alpha})} = -\frac{1}{2}L_m \mathrm{vec}\left[T\Sigma^{-1}_{(T)} - T\Sigma^{-1}_{(T)}B_{ui(T)}\Sigma^{-1}_{(T)}\right],$$

where vech$(\Sigma_{\epsilon\epsilon})$ and vech$(\Lambda_{\alpha\alpha})$ are column-wise vectorization of the lower triangle of the symmetric matrix $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$, and L_m is an elimination matrix. $W_{ui(T)}$ and $B_{ui(T)}$ respectively are defined as follows

$$W_{ui(T)} = \tilde{E}_{i(T)}K_T\tilde{E}'_{i(T)} \text{ and } B_{ui(T)} = \tilde{E}_{i(T)}J_T\tilde{E}'_{i(T)},$$

where $\tilde{E}_{i(T)} = [\boldsymbol{u}_{i1}, \ldots, \boldsymbol{u}_{iT}]$ is a $(m \times T)$ matrix and $\boldsymbol{u}_{i(T)} = \mathrm{vec}(E_{i(T)})$, 'vec' being the vectorization operator. That is, the disturbances defined in (B-2) for an individual i has been arranged in $(m \times T)$ matrix, $\tilde{E}_{i(T)}$.

The second order conditions are:

$$\frac{\partial^2 \mathcal{L}_i}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}'} = -Z_{i(T)}[K_T \otimes \Sigma^{-1}_{\epsilon\epsilon} + J_T \otimes (\Sigma_{\epsilon\epsilon} + p\Sigma_{\alpha\alpha})^{-1}]Z'_{i(T)}$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \boldsymbol{\delta}\partial \mathrm{vec}(\Lambda_{\alpha\alpha})'} = -T(\boldsymbol{u}_{i(T)} \otimes Z_{i(T)})(I_T K_{m,T} \otimes I_m)(\mathrm{vec}(J_T) \otimes \Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)})$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \boldsymbol{\delta}\partial \mathrm{vec}(\Sigma_{\epsilon\epsilon})'} = -(\boldsymbol{u}_{i(T)} \otimes Z_{i(T)})(I_T \otimes K_{m,T} \otimes I_m)(\mathrm{vec}(K_T) \otimes \Sigma^{-1}_{\epsilon\epsilon} \otimes \Sigma^{-1}_{\epsilon\epsilon} + \mathrm{vec}(J_T) \otimes \Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)})$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \mathrm{vec}(\Lambda_{\alpha\alpha})\partial \boldsymbol{\delta}'} = -\frac{T}{2}(\Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)})[(\tilde{E}_{i(T)}J_T \otimes I_m) + (I_m \otimes \tilde{E}_{i(T)}J_T)K_{m,T}]Z'_{i(T)}$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \mathrm{vec}(\Lambda_{\alpha\alpha})\partial \mathrm{vec}(\Lambda_{\alpha\alpha})'} = \frac{T^2}{2}[(\Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)}) - \Sigma^{-1}_{(T)}B_{ui(T)}\Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)} - \Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)}B_{ui(T)}\Sigma^{-1}_{(T)}]$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \mathrm{vec}(\Lambda_{\alpha\alpha})\partial \mathrm{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{T}{2}[(\Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)}) - \Sigma^{-1}_{(T)}B_{ui(T)}\Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)} - \Sigma^{-1}_{(T)} \otimes \Sigma^{-1}_{(T)}B_{ui(T)}\Sigma^{-1}_{(T)}]$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \mathrm{vec}(\Sigma_{\epsilon\epsilon})\partial \boldsymbol{\delta}'} = -\frac{1}{2}(\Sigma_{(T)}^{-1} \otimes \Sigma_{(T)}^{-1})[(\tilde{E}_{i(T)} J_T \otimes I_m) + (I_m \otimes \tilde{E}_{i(T)} J_T) K_{m,T}] Z'_{i(T)}$$
$$- \frac{1}{2}(\Sigma_{\epsilon\epsilon}^{-1} \otimes \Sigma_{\epsilon\epsilon}^{-1})[(\tilde{E}_{i(T)} K_T \otimes I_m) + (I_m \otimes \tilde{E}_{i(T)} K_T) K_{m,T}] Z'_{i(T)}$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \mathrm{vec}(\Sigma_{\epsilon\epsilon})\partial \mathrm{vec}(\Lambda_{\alpha\alpha})'} = \frac{T}{2}[(\Sigma_{(T)}^{-1} \otimes \Sigma_{(T)}^{-1}) - \Sigma_{(T)}^{-1} B_{ui(T)} \Sigma_{(T)}^{-1} \otimes \Sigma_{(T)}^{-1} - \Sigma_{(T)}^{-1} \otimes \Sigma_{(T)}^{-1} B_{ui(T)} \Sigma_{(T)}^{-1}]$$

$$\frac{\partial^2 \mathcal{L}_i}{\partial \mathrm{vec}(\Sigma_{\epsilon\epsilon})\partial \mathrm{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{1}{2}[\Sigma_{(T)}^{-1} \otimes \Sigma_{(T)}^{-1} + (T-1)\Sigma_{\epsilon\epsilon} \otimes \Sigma_{\epsilon\epsilon} - \Sigma_{\epsilon\epsilon} B_{ui(T)} \Sigma_{(T)} \otimes \Sigma_{(T)}^{-1}$$
$$- \Sigma_{(T)}^{-1} \otimes \Sigma_{(T)}^{-1} B_{ui(T)} - \Sigma_{\epsilon\epsilon}^{-1} W_{ui(T)} \Sigma_{\epsilon\epsilon}^{-1} \otimes \Sigma_{\epsilon\epsilon}^{-1} - \Sigma_{\epsilon\epsilon}^{-1} \otimes \Sigma_{\epsilon\epsilon}^{-1} W_{ui(T)} \Sigma_{(T)}^{-1}].$$

REFERENCES

Biørn, E. (2004). Regression Systems for Unbalanced Panel Data: A Stepwise Maximum Likelihood Procedure . *Journal of Econometrics*, **122**, 281–291.

Newey, W. K. (1984). A Method of Moment Interpretation of Sequential Estimators. *Economics Letters*, **14**, 201–206.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

— (2003). *Solutions Manual and Supplementary Materials for Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

KOKKUVÕTE

Paneelandmete Binaarne Mudel Triangulaarses Süsteemis

Käesoleva artikli peamine eesmärk oli arendada välja meetod punkthinnangute saamiseks struktuursetele indikaatortele nagu töötlusefektid (treatment effects) paneelandmete binaarsetes mudelites triangulaarses süsteemis, võttes seejuures arvesse mitmedimensioonilist mittevaadeldavat heterogeensust. Mittevaadeldava heterogeensuse liikmete hulka kuuluvad ajas mittevarieeruvad juhuslikud efektid ja idiosünkraatilised vealiikmed. Me esmalt identiftseerime taandatud kujul võrrandite heterogeensuse liimete oodatavad väärtused tingimuslikuna endogeensete ja eksogeensete muutujate suhtes kõikidel ajaperioodidel, ja seejärel näitame, et antud oodatavate väärtuste juures on meid huvitavad näitajad identifitseeritavad. Seejärel pakume välja lähenemise, et need heterogeensuse liikmete tingimuslikud oodatavad väärtused on kasutatavad taandatud kujul võrrandites kontrollfunktsioonina.

Antud väljapakutud meetod panustab mitmel huvitaval viisil teaduskirjandusse. Esiteks, saavutab antud meetod keskmiste osaliste efektide identifitseerimise triangulaarses mitmedimensioonilise heterogeensusega süsteemis. Teiseks, triangulaarsetes süsteemides meie lähenemisega sarnaste eelduste korral nõuab meie välja pakutud kontrollfunktsiooni meetod nõrgemaid kitsendusi kui traditsioonilised kontrollfunktsiooni meetodid. Kolmandaks võimaldab antud meetod kasutada väikese mittenulliliste väärtuste piirkonnaga (support) instrumente, mis sai võimalikuks tänu paneelandmete kasutamisele ja teatud heterogeensuse liikmete ajas mitte varieerumisele. Samuti, Monte-Carlo eksperimendid näitavad, et meie väljapakutud meetod töötab paremini võrreldes alternatiivsete paneelandmete binaarsete regressioonimeetoditega.

Antud hinnangufunktsiooni rakendati artiklis uurimaks taludes sissetulekute ja rikkuse (maa ja muude talus tootmises kasutatavate varade) põhjuslikku mõju lapstööjõu kasutamisele Indias. Artiklis leiti, et kõrgem sissetuleku ja kõrgem tootlike põllumajanduslike varade omamine oluliselt alandasid lapstööjõu kasutamise esinemissagedust, mis viitab tootlike põllumajanduslike varade tugevale sissetulekuefektile. Teiseks, suurem maaomandus tõstab lapstööjõu esinemissagedust, mis viitab lapstööjõu asendusefektile. Kolmandaks, eksogeensuse test näitas, et maaomandus on määratud endogeenselt koos majapidamise tööpakkumise otsustega, mis on vastupidine tulemus eeldusele, mida on kasutanud enamik arengumaades lapstööjõu kasutamist uurinud empiirilistest töödest.