NONLINEAR PANEL DATA MODELS WITH CONTINUOUS ENDOGENOUS REGRESSORS AND GENERAL INSTRUMENTS[1]

AMARESH K TIWARI[2],

Most papers studying nonlinear panel data models assume that conditional on the unobserved heterogeneity all covariates are exogenous. We relax this assumption and develop a control function method to handle heterogeneity and endogeneity of covariates. The control functions are based on "expected a posteriori" values of the correlated random effects. Average partial effects are identified and, unlike alternative control function approaches, our method allows for general instruments. The proposed method is applied to estimate the causal effects of household income and wealth on the incidence of child labor, and is contrasted with the standard binary response model for panel data.

KEYWORDS: Control Functions, Expected a Posteriori, Multidimensional Numerical Integration, Average Partial Effects, Child Labor, Midday Meal Scheme.

JEL Classification: C13, C18, C33, J4

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 (2010) discussing binary choice model shows that, 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. Weidner (2011) and Arellano and Bonhomme (2011) provide an overview, and categorize, of some of the methods developed to estimate the quantities of interest.

One of the leading methods in the literature is the fixed effect (FE) approach that treats individual effects as parameters to be estimated. But as the number of individual effects grow with the sample size, incidental parameter problem (see Lancaster, 2000, for a review) usually appears in fixed T estimation of nonlinear panel data models. It has been argued that the incidental parameter problem can be viewed as time-series finite-sample bias when T tends to infinity. Following this perspective, several approaches have been proposed to correct for the time-series bias. Some of the papers that follow the bias reduction technique for estimating the quantities of interest are Hahn and Newey

[1]This research was supported by IAP research network, grant nr. P7/06, of the Belgian Government's Belgian Science Policy. The author wishes to thank Stéphane Bonhomme, Cedric Heuchenne, Bernard Lejeune, Pierre Mohnen, Franz C. Palm, Sybrand Schim van der Loeff, seminar participants at the Center for Operations Research and Econometrics, Louvain-la-Neuve, and seminar participants at the Econometric Institute, Erasmus University Rotterdam, for their helpful comments. Finally, I would like to thank Soham Sahoo, without whose help this paper could not have been completed. All remaining errors are mine.

[2]University of Liege, A.Tiwari@ulg.ac.be

(2004), Arellano and Hahn (2007), Bester and Hansen (2009), Fernandez-Val (2009), and Hahn and Kuersteiner (2011).

Wooldridge (2009) 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 very 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 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. Some of the papers that deal with bound analysis are Honore and Tamer (2006) and Chernozhukov *et al.* (2013). However, with the exception of Honore and Tamer, the methods in 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)(HW), may not always hold true.

A partial list of papers that study nonparametric control function estimation of nonseparable models are Florens *et al.* (2008), Imbens and Newey (2009), and Torgovitsky (2012), where the focus is on estimating heterogeneous effect of endogenous treatment. However, in these papers all exogenous covariates are assumed to be independent of unobserved heterogeneity. Besides, these papers do not consider heterogeneity in the reduced form or the "treatment choice equation" as termed by Florens *et al.*. Also, with the exception of Altonji and Matzkin (2005), Papke and Wooldridge (2008)(PW), and HW the papers employing control function approach for panel data assume all covariates to be exogenous conditional on the unobserved heterogeneity.

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 account for endogeneity of such covariates. Heterogeneity is modeled as correlated random effects (CRE) and errors are assumed to be additively separable. Though we assume our model to be triangular, in many applications with additively separable errors and single index restriction the triangular representation can be achieved from a fully simultaneous system. The approach does entail restriction on the distribution of error components. But as Wooldridge argues, "estimation using CRE and FE involve trade-offs among assumptions and the type of quantities that can be estimated, and that no method provides consistent estimates of either parameters or APE's under a set of assumptions strictly weaker than the assumptions needed for the other procedures." Some papers that adopt the CRE approach to account for heterogeneity are Bester and Hansen (2007), PW, and Weidner (2011). While

Bester and Hansen and Weidner study semiparametric models, and do not specify the conditional distribution of the individual effects, PW assume a parametric form. In terms of imposed structures, our paper is closest to PW's.

Typically, in a simultaneous triangular system of equations with additive separability and without latent heterogeneity in the reduced form equations, the control functions are the unobserved time-varying errors in the reduced form equations. So that conditional on reduced form errors, which are proxied by the residuals, the structural parameters can be consistently estimated. In panel data setting, which allows us to account for unobserved individual effects, the residuals of the reduced form equations, defined as the observed value of the endogenous variables minus its expectation conditional on observed regressors and the unobserved individual effects, remain unidentified. This is plainly because the individual effects/heterogeneity in the reduced form equations, which are correlated with individual effects in the structural equation, are unobserved.

The novelty of our approach lies in integrating out the unobserved individual effects with respect to conditional distribution of the individual effects, which is obtained as the posterior distribution of the individual effects from the results from the first stage reduced form estimation. This leaves us with the "expected a Posteriori" (EAP) values of the individual effects, which can then be used to obtain control functions that are a function of the observed variables. The EAP values of individual effects are obtained through numerical integration with respect to distribution of the individual effects.

Our method, while being simple, makes a number of contributions to the literature. First, unlike most control function approaches that require the presence of continuous instruments, often with a large support, our method allows for general instruments. This is because the control functions, which are based on EAP value of individual effects, are functions of endogenous and exogenous variables from all time periods. Hence, conditional on contemporaneous endogenous variable the large, common support of the control functions needed to identify the average structural function (ASF) and APE is provided by the unrestricted continuous endogenous variables from other time periods. Thus, we find that once again panel data with multiple observations for each individual, which allows for accounting of unobserved heterogeneity, aids in identifying quantities of interest. In our case, as it happens, it allows for the possibility of control functions to employ general instruments for identification.

Secondly, we account for heterogeneity in the reduced form/ treatment choice equation. Finally, our model retains the attractive features of the PW, where no assumptions are made on the serial dependence among the outcome variable. Another interesting feature of our model is that only two time periods suffice to identify the structural measures of interest. Our model can be especially useful, see the application in this paper, when (i) only discrete instruments are available, and (ii) when one is faced with very short panels and FE and set identification approaches that require large T and fully non-separable models that require at least as many time periods as the number of regressors cannot be employed (see also HW).

Using data on India, the proposed estimator is employed to estimate causal effects of household income and wealth on child labor and their propensity to attend school. 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 and schooling. To demonstrate that our control function method can accommodate general instruments, most of the instruments employed in the estimation are discrete. We also look at how the Government of India's "Midday Meal Scheme", through which free cooked lunch is provided on working days for children in primary and upper primary classes (classes I to VIII) in Government schools, with the objectives of providing nutrition and encouraging children of poor and disadvantaged sections to attend school, affects children's propensity to participate in work and attend school.

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 section 3 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 4 we conclude. Technical proofs are relegated to appendix A. Due to space constraint, other technical details have been put in a supplementary appendix, which, among others, includes the derivation of the asymptotic covariance matrix and a note on multidimensional numerical integration.

2. MODEL SPECIFICATION

For the sake of exposition, we assume a binary choice model,

$$y_{it}^* = \boldsymbol{\varphi}(\boldsymbol{z}_{it}^y, \boldsymbol{x}_{it}) + \theta_i + \zeta_{it}, \tag{2.1}$$

where y_t^*[1] is the latent variable underlying the binary response outcome, $y_t = 1\{y_t^* > 0\}$; $1\{.\}$ is an indicator function that takes value 1 if the argument in the parenthesis holds true and 0 otherwise. We assume our model to be additively separable in the errors, θ and ζ_t, where θ is the unobserved time invariant individual effect and ζ_t is the idiosyncratic component. Conditional on θ, $\boldsymbol{z}_t^y$ is assumed to be independent of ζ_t for all t; that is, $\boldsymbol{z}_t^y$ is a vector of exogenous variables. Above, $\boldsymbol{x}_t$ is a vector of endogenous covariates; in other words, $\zeta_t \not\perp \boldsymbol{x}_t|\theta$. Also, $\boldsymbol{x}_t$ is continuous and is of dimension 'm'. To estimate the structural parameters in equation (2.1) we develop a two stage control function procedure. In the first stage, the parameters Θ_1 of the system of reduced form equations, equation (2.2), is estimated.

$$\boldsymbol{x}_{it} = \boldsymbol{\beta}(\boldsymbol{z}_{it}) + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{it}, \tag{2.2}$$

Equation (2.2) is the system of 'm' equations written in a reduced form for the endogenous variables $\boldsymbol{x}_t$. While we refer (2.2) as reduced form equations, it could be thought of as structural equations in a triangular system. However, with additive separability and single index restriction, the triangular representation in (2.1) and (2.2) can be derived from a fully simultaneous system involving $\boldsymbol{x}_t$ and y_t^*. In (2.2), $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is a vector of unobserved individual effects and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{mt})'$ is the vector of idiosyncratic error terms.

[1] In the rest of the paper, except when needed, we will drop the individual subscript i.

$\boldsymbol{\beta}(\boldsymbol{z}_t) = \{\boldsymbol{\beta}_1(\boldsymbol{z}_t), \ldots, \boldsymbol{\beta}_m(\boldsymbol{z}_t)\}'$, where $\boldsymbol{z}_t = (\boldsymbol{z}_t^{y\prime}, \tilde{\boldsymbol{z}}_t')'$ is the vector of exogenous variables. The crucial identification requirement is that the dimension of instruments, $\tilde{\boldsymbol{z}}_t$, which are excluded from the structural equation (2.1), be greater than or equal to the dimension of $\boldsymbol{x}_t$. Such exclusion restrictions have to be justified on economic grounds before $\tilde{\boldsymbol{z}}_t$ can be employed as instruments. Finally, define $\mathcal{Z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_T')'$, $\boldsymbol{X} = \{\boldsymbol{x}_1', \ldots, \boldsymbol{x}_T'\}'$, and $\boldsymbol{\epsilon} = \{\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_T'\}'$.

Before we proceed to discuss the identification and estimation of the quantities of interest, we first state and discuss some of the model's assumptions.

ASSUMPTION 1 *$\zeta_t \perp \mathcal{Z}|\theta$ and $\boldsymbol{\epsilon}_t \perp \boldsymbol{\alpha}, \mathcal{Z}$.*

The assumption that $\mathcal{Z}$ and $\boldsymbol{\alpha}$ are both independent of $\boldsymbol{\epsilon}_t$ is maintained because without it, in the correlated random effect framework that we employ, it is not possible to recover the distribution of $\boldsymbol{\alpha}_i$, which is required to obtain the control functions. When it cannot be argued that the triangular representation in (2.1) and (2.2) is in fact structural, and the triangular representation has to be obtained from a fully simultaneous system, then it would be required that both $\mathcal{Z}$ and θ be independent of ζ_t.

Now, one of the requirements of our control function method is to be able to recover, among others, the conditional distribution of $\boldsymbol{X}$ and $\boldsymbol{\alpha}$ given $\mathcal{Z}$. However, we do not know of any non or semiparametric estimator where the conditional distribution of endogenous variables and the random coefficients or effects are recovered when the problem is one that of estimating a system of regressions as in equation (2.2)[2]. Given this lack and the necessity of recovering certain conditional distributions, we assume a parametric specification to estimate the reduced form equations.

First, we will assume a single index form for $\boldsymbol{\beta}(\boldsymbol{z}_t)$ in the reduced form equations, so that $\boldsymbol{\beta}(\boldsymbol{z}_t) = \text{diag}(\boldsymbol{z}_t, \ldots, \boldsymbol{z}_t)'\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_m')'$. To account for the correlation of $\boldsymbol{\alpha}_i$ and $\mathcal{Z}_i$, we employ the correlated random effects (CRE) formulation in Chamberlain (1984). We assume that

ASSUMPTION 2

$$\text{E}(\boldsymbol{\alpha}_i|\mathcal{Z}_i) = \boldsymbol{\varrho}(\mathcal{Z}_i) = \textit{diag}(\bar{\boldsymbol{z}}, \ldots, \bar{\boldsymbol{z}})'\boldsymbol{\varrho}$$

where $\boldsymbol{\varrho} = (\boldsymbol{\varrho}_1', \ldots, \boldsymbol{\varrho}_m')'$, and $\bar{\boldsymbol{z}}$ could be either Chamberlain's or Mundlak's specification for CRE.

This implies that

$$\text{E}(\boldsymbol{x}_t|\mathcal{Z}) = \boldsymbol{Z}_t'\boldsymbol{\delta},$$

where $\boldsymbol{Z}_t = \text{diag}((\boldsymbol{z}_t', \bar{\boldsymbol{z}}')', \ldots, (\boldsymbol{z}_t', \bar{\boldsymbol{z}}')')$, and $\boldsymbol{\delta} = ((\boldsymbol{\beta}_1', \boldsymbol{\varrho}_1'), \ldots, (\boldsymbol{\beta}_m', \boldsymbol{\varrho}_m'))'$. The conditional distribution of $\boldsymbol{\alpha}_i$ given $\mathcal{Z}_i$ is assumed as

[2]For scalar $\boldsymbol{x}$, Arellano and Bonhomme (2012) have proposed a semiparametric random coefficient model, where such distributions can be estimated. Presumably, their results can be extended to identify the sort of control function proposed in this paper. But since their model pertains to scalar $\boldsymbol{x}$, we will not discuss or attempt to extend their results here.

ASSUMPTION 3

$$\boldsymbol{\alpha}_i|\mathcal{Z}_i \sim \mathrm{N}\left[\mathrm{E}(\boldsymbol{\alpha}_i|\mathcal{Z}_i), \Lambda_{\alpha\alpha}\right],$$

so that the tail, $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i - \mathrm{E}(\boldsymbol{\alpha}_i|\mathcal{Z}_i) = \boldsymbol{\alpha}_i - \boldsymbol{\varrho}(\mathcal{Z}_i)$, is distributed normally with conditional mean zero, variance $\Lambda_{\alpha\alpha}$, and is also assumed to be independent of $\mathcal{Z}_i$. The idiosyncratic error, $\boldsymbol{\epsilon}_t$, is assumed to be distributed as:

ASSUMPTION 4

$$\boldsymbol{\epsilon}_t \sim \mathrm{N}\left[0, \Sigma_{\epsilon\epsilon}\right].$$

Given the above, equation (2.2) can now be written as

$$\boldsymbol{x}_t = \boldsymbol{Z}_t'\boldsymbol{\delta} + \tilde{\boldsymbol{\alpha}} + \boldsymbol{\epsilon}_t. \tag{2.2a}$$

The parameters, $\Theta_1 = \{\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}\}$, of the modified equation (2.2a) can be estimated by a step-wise maximum likelihood method for system of regressions developed by Biørn (2004). Biørn's paper, however, does not account for any possible heteroscedasticity in $\tilde{\boldsymbol{\alpha}}$ or heteroscedasticity and serial correlation among $\boldsymbol{\epsilon}_t$. Also, we are not aware of any test for testing vector serial correlation in the idiosyncratic term of error component models for a system of regressions.

If $m = 1$, one can employ the methodology in Baltagi *et al.* (2010) to deal with heteroscedasticity in $\tilde{\boldsymbol{\alpha}}$ and serial correlation in the idiosyncratic components. For $m = 1$ Baltagi *et al.* (2006) allow for heteroscedasticity in $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\epsilon}_t$ but no serial correlation in the idiosyncratic component. In what follows, for the sake of exposition we will stick to the maintained assumptions in Biørn's for the reduced form equations (2.2a)[3]. In appendix A of the supplementary appendix we briefly describe the methodology in Biørn (2004).

2.1. *Identification and Estimation of Structural Parameters and Average Partial Effect*

2.1.1. *Identification*

The identification strategy that allows us to construct the control variables that correct for the bias, which arises due to endogeneity of $\boldsymbol{x}_t$ and the correlation of $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ with the unobserved heterogeneity, is based on the following conditional distribution restriction:

ASSUMPTION 5

$$\begin{aligned}\theta, \zeta_t|\boldsymbol{X}, \mathcal{Z}, \boldsymbol{\alpha} &\sim \theta, \zeta_t|\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}|\mathcal{Z}, \boldsymbol{\alpha}), \mathcal{Z}, \boldsymbol{\alpha} \\ &\sim \theta, \zeta_t|\boldsymbol{\epsilon}, \mathcal{Z}, \boldsymbol{\alpha} \\ &\sim \theta, \zeta_t|\boldsymbol{\epsilon}, \boldsymbol{\alpha}.\end{aligned}$$

[3]It is possible to modify Biørn's methodology to allow for limited order vector serial correlation and heteroscedasticity in error components. However, testing for nonspherical errors when one is dealing with a system of regressions with random effects might not be straightforward. One way, albeit somewhat inexact, might be to test for non-sphericity among error components by employing the tests developed in Baltagi *et al.* (2010), or in Wooldridge (2002), for each of the regressions separately in the system of regressions. If the tests confirm for non-sphericity, then the covariance matrices of the error components in Biørn's paper can be adjusted to allow for serial dependence and heteroscedasticity.

According to the above, the dependence of the structural error terms θ and ζ_t on $\boldsymbol{X}$, $\mathcal{Z}$, and $\boldsymbol{\alpha}$ is completely characterized by the reduced form error components $\boldsymbol{\epsilon}$ and $\boldsymbol{\alpha}$. This restriction is weaker than in most control function methods, where it is required that $\mathcal{Z}$ be independent of ζ_t conditional on $\boldsymbol{\epsilon}_t$.

The expectation of $\theta + \zeta_t$ given $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ is given by

$$\mathrm{E}(\zeta_t + \theta|\boldsymbol{\alpha}, \boldsymbol{\epsilon}) = \mathrm{E}(\zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t) + \mathrm{E}(\theta|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$$

The equality in the above follows from

ASSUMPTION 6 $\quad \theta, \zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon} \sim \theta, \zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t,$

which states that conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}_t$, θ and ζ_t are independent of $\boldsymbol{\epsilon}_{-t}$. A similar assumption that conditional on contemporaneous time varying reduced form errors and individual effects, the structural errors are independent of rest of the time varying reduced form errors has also been made in PW and Semykina and Wooldridge (2010)[4].

As we will see, without assuming a functional form for $\mathrm{E}(\zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$ and $\mathrm{E}(\theta|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$, it may be difficult to estimate the structural parameters of interest. Hence, as is common in parametric control function approach, we assume that

ASSUMPTION 7 $\quad \mathrm{E}(\theta|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$ *and* $\mathrm{E}(\zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$ *are linear in* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\epsilon}_t$.

That is,

$$\mathrm{E}(\theta|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t) + \mathrm{E}(\zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t) = (\Sigma_{\theta\alpha}\boldsymbol{\alpha} + \Sigma_{\theta\epsilon}\boldsymbol{\epsilon}_t) + (\Sigma_{\zeta\alpha}\boldsymbol{\alpha} + \Sigma_{\zeta\epsilon}\boldsymbol{\epsilon}_t) = \Sigma_\alpha\boldsymbol{\alpha} + \Sigma_\epsilon\boldsymbol{\epsilon}_t,$$

where the final expression is obtained by collecting the $\boldsymbol{\alpha}$ terms together and the $\boldsymbol{\epsilon}_t$ terms together. The two $(1 \times m)$ matrices Σ_α and Σ_ϵ respectively are

$$\Sigma_\alpha = \begin{pmatrix} \rho_{\alpha 1} & \dots & \rho_{\alpha m} \end{pmatrix} \text{ and } \Sigma_\epsilon = \begin{pmatrix} \rho_{\epsilon 1} & \dots & \rho_{\epsilon m} \end{pmatrix}.$$

The elements of Σ_ϵ and Σ_α, which are estimated in the second stage structural estimation, give us a test of the exogeneity of $\boldsymbol{x}_t$.

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

$$\begin{aligned} \mathrm{E}(y_t^*|\boldsymbol{X}, \mathcal{Z}, \boldsymbol{\alpha}) &= \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha\boldsymbol{\alpha} + \Sigma_\epsilon\boldsymbol{\epsilon}_t \\ &= \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha(\boldsymbol{\varrho}(\mathcal{Z}) + \tilde{\boldsymbol{\alpha}}) + \Sigma_\epsilon\boldsymbol{\epsilon}_t = \mathrm{E}(y_t^*|\boldsymbol{X}, \mathcal{Z}, \tilde{\boldsymbol{\alpha}}), \end{aligned} \tag{2.3}$$

where $\mathcal{X}_t = \{\boldsymbol{z}_t^{y\prime}, \boldsymbol{x}_t'\}'$. In a triangular setup with additive separability and without unobserved heterogeneity, residuals from the first stage reduced form regression form the estimates of $\boldsymbol{\epsilon}_t$. In our model, however, the residuals, $\boldsymbol{\epsilon}_t = \boldsymbol{x}_t - \mathrm{E}(\boldsymbol{x}_t|\mathcal{Z}, \tilde{\boldsymbol{\alpha}}) = \boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}}$,

[4]This assumption can be relaxed and one can specify the dependence of ζ_{it} and $\boldsymbol{\epsilon}_i$ without adding any conceptual difficulties.

and $\boldsymbol{\alpha} = \boldsymbol{\varrho}(\mathcal{Z}) + \tilde{\boldsymbol{\alpha}}$ are not identified because the $\tilde{\boldsymbol{\alpha}}_i$'s are unobserved[5]. But it may still be possible to estimate the structural parameters if we could integrate out $\tilde{\boldsymbol{\alpha}}_i$ with respect to its conditional distribution $\mathbf{f}(\tilde{\boldsymbol{\alpha}}_i|\boldsymbol{X}_i, \mathcal{Z}_i)$. To see this, consider $\mathrm{E}(y^*_{it}|\boldsymbol{X}_i, \mathcal{Z}_i, \tilde{\boldsymbol{\alpha}}_i)$ in (2.3):

$$\begin{aligned}
&\int \mathrm{E}(y^*_t|\boldsymbol{X}, \mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{f}(\tilde{\boldsymbol{\alpha}}_i|\boldsymbol{X}, \mathcal{Z})d\tilde{\boldsymbol{\alpha}}_i \\
&= \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha \boldsymbol{\varrho}(\mathcal{Z}) + \Sigma_\epsilon(\boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta}) + \int (\Sigma_\alpha \tilde{\boldsymbol{\alpha}} - \Sigma_\epsilon \tilde{\boldsymbol{\alpha}})\mathbf{f}(\tilde{\boldsymbol{\alpha}}|\boldsymbol{X}, \mathcal{Z})d\tilde{\boldsymbol{\alpha}} \\
&= \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha \boldsymbol{\varrho}(\mathcal{Z}) + \Sigma_\epsilon(\boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta}) + \Sigma_\alpha \hat{\tilde{\boldsymbol{\alpha}}} - \Sigma_\epsilon \hat{\tilde{\boldsymbol{\alpha}}} \\
&= \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha(\boldsymbol{\varrho}(\mathcal{Z}) + \hat{\tilde{\boldsymbol{\alpha}}}) + \Sigma_\epsilon(\boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta} - \hat{\tilde{\boldsymbol{\alpha}}}) \\
&= \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t = E(y^*_t|\boldsymbol{X}, \mathcal{Z}).
\end{aligned} \tag{2.4}$$

In the second equality $\hat{\tilde{\boldsymbol{\alpha}}}_i = \hat{\tilde{\boldsymbol{\alpha}}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_1)$ are the "expected a posteriori" (EAP) values of the time invariant individual effects $\tilde{\boldsymbol{\alpha}}_i$. In the fourth equality $\hat{\boldsymbol{\alpha}} = \boldsymbol{\varrho}(\mathcal{Z}) + \hat{\tilde{\boldsymbol{\alpha}}}$ and $\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta} - \hat{\tilde{\boldsymbol{\alpha}}}$.

To obtain (2.4), we use Bayes rule to write $\mathbf{f}(\tilde{\boldsymbol{\alpha}}|\boldsymbol{X}, \mathcal{Z})$ as

$$\mathbf{f}(\tilde{\boldsymbol{\alpha}}|\boldsymbol{X}, \mathcal{Z}) = \frac{\mathbf{f}(\boldsymbol{X}, \mathcal{Z}|\tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})}{\mathbf{h}(\boldsymbol{X}, \mathcal{Z})} = \frac{\mathbf{f}(\boldsymbol{X}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{p}(\mathcal{Z}|\tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})}{\mathbf{h}(\boldsymbol{X}|\mathcal{Z})\mathbf{p}(\mathcal{Z})},$$

where $\mathbf{g}$ and $\mathbf{h}$ are density functions. By our assumption the residual time invariant individual effects, $\tilde{\boldsymbol{\alpha}}$, are independent of the exogenous variables $\mathcal{Z}$, hence $\mathbf{p}(\mathcal{Z}|\tilde{\boldsymbol{\alpha}}) = \mathbf{p}(\mathcal{Z})$. That is,

$$\mathbf{f}(\tilde{\boldsymbol{\alpha}}|\boldsymbol{X}, \mathcal{Z}) = \frac{\mathbf{f}(\boldsymbol{X}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})}{\mathbf{h}(\boldsymbol{X}|\mathcal{Z})} = \frac{\mathbf{f}(\boldsymbol{X}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})}{\int \mathbf{f}(\boldsymbol{X}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}.$$

Hence, we have

$$\begin{aligned}
\int \tilde{\boldsymbol{\alpha}}\mathbf{f}(\tilde{\boldsymbol{\alpha}}|\boldsymbol{X}, \mathcal{Z})d(\tilde{\boldsymbol{\alpha}}) &= \int \frac{\tilde{\boldsymbol{\alpha}}\mathbf{f}(\boldsymbol{X}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int \mathbf{f}(\boldsymbol{X}|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}} = \frac{\int \tilde{\boldsymbol{\alpha}}_i \prod_{t=1}^T \mathbf{f}(\boldsymbol{x}_t|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int \prod_{t=1}^T \mathbf{f}(\boldsymbol{x}_t|\mathcal{Z}, \tilde{\boldsymbol{\alpha}})\mathbf{g}(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}} \\
&= \hat{\tilde{\boldsymbol{\alpha}}}(\boldsymbol{X}, \mathcal{Z}, \Theta_1)
\end{aligned} \tag{2.5}$$

where the second equality follow from the fact that conditional on $\mathcal{Z}$ and $\tilde{\boldsymbol{\alpha}}$, each of the $\boldsymbol{x}_t$, $\boldsymbol{x}_t \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ are independent and normally distributed with mean $\boldsymbol{Z}_t'\boldsymbol{\delta} + \tilde{\boldsymbol{\alpha}}$ and

[5] To identify $\boldsymbol{\epsilon}_t$, HW assume the reduced form as

$$\boldsymbol{x}_t = f_0(\boldsymbol{z}_t) + f_1(\boldsymbol{z}_t, \boldsymbol{\alpha})\boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t \perp (\boldsymbol{z}_t, \boldsymbol{\alpha})$, and 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 necessarily nest the model considered here.

standard deviation $\Sigma_{\epsilon\epsilon}$. $\mathbf{g}(\tilde{\boldsymbol{\alpha}})$ by our assumption is normally distributed with mean zero and variance $\Lambda_{\alpha\alpha}$. Let $\check{\boldsymbol{\alpha}}(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$ denote the estimated EAP value of $\tilde{\boldsymbol{\alpha}}$. $\check{\boldsymbol{\alpha}}$ can be estimated by employing multidimensional numerical integration techniques with respect to $\tilde{\boldsymbol{\alpha}}$ at the estimated values, $\hat{\Theta}_1$. In appendix D of the supplementary appendix we provide a note on the numerical technique employed to estimate $\check{\boldsymbol{\alpha}}$.

Now, it can be shown that

LEMMA 1 *$\check{\boldsymbol{\alpha}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \hat{\Theta}_1)$ converges almost surly to $\hat{\tilde{\boldsymbol{\alpha}}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_1)$, where $\hat{\Theta}_1 = \{\hat{\boldsymbol{\delta}}, \hat{\Sigma}_{\epsilon\epsilon}, \hat{\Lambda}_{\alpha\alpha}\}$ is the consistent first stage estimate.*

PROOF OF LEMMA 1 Given in appendix A .

Lemma 1 implies that

$$\boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha \hat{\tilde{\boldsymbol{\alpha}}}_i + \Sigma_\epsilon \hat{\tilde{\boldsymbol{\epsilon}}}_t \xrightarrow{a.s} E(y_t^* | \boldsymbol{X}, \mathcal{Z}) = \int E(y_t^* | \boldsymbol{X}, \mathcal{Z}, \tilde{\boldsymbol{\alpha}}) \mathbf{f}(\tilde{\boldsymbol{\alpha}} | \boldsymbol{X}, \mathcal{Z}) d(\tilde{\boldsymbol{\alpha}}), \tag{2.6}$$

where $\hat{\tilde{\boldsymbol{\alpha}}} = \hat{\boldsymbol{\varrho}}(\mathcal{Z}) + \check{\boldsymbol{\alpha}}$ and $\hat{\tilde{\boldsymbol{\epsilon}}}_t = \boldsymbol{x}_t - \mathbf{Z}_t' \hat{\boldsymbol{\delta}} - \check{\boldsymbol{\alpha}}$. If population parameters, $\boldsymbol{\delta}$, $\Sigma_{\epsilon\epsilon}$, and $\Lambda_{\alpha\alpha}$, were known, the above implies that we could write the linear predictor of y_{it}^*, given $\boldsymbol{X}_i$ and $\mathcal{Z}_i$ in error form as

$$y_t^* = \boldsymbol{\varphi}(\mathcal{X}_t) + \Sigma_\alpha \tilde{\boldsymbol{\alpha}} + \Sigma_\epsilon \tilde{\boldsymbol{\epsilon}}_t + \tilde{\zeta}_t, \tag{2.7}$$

where $\tilde{\zeta}_t$ is distributed with conditional mean 0.

Had y_t^* been continuous and observed, with estimates $\hat{\tilde{\boldsymbol{\alpha}}}$ and $\hat{\tilde{\boldsymbol{\epsilon}}}_t$ in place, equation (2.7) could be estimated by employing GMM. However, when response outcomes are discrete and we have to deal with nonlinear models, additional assumptions than those made above may be required. For an individual i, we are interested in the Average Structural Function (ASF),

$$\mathrm{E}(y_t | \mathcal{X}_t) = G(\mathcal{X}_t) = \int H(\mathcal{X}_t, \theta, \zeta_t) dF_{\theta,\zeta}, \tag{2.8}$$

where $y_t = 1\{\varphi(\mathcal{X}_t) + \theta + \zeta_t > 0\} = H(\mathcal{X}_t, \theta, \zeta_t)$, and the Average Partial Effect (APE) of changing a variable, say w, in time period t from w_t to $w_t + \Delta_w$,

$$\frac{\Delta \mathrm{E}(y_t | \mathcal{X}_t)}{\Delta_w} = \frac{\Delta G(\mathcal{X}_t)}{\Delta_w} = \frac{\int \Big(H(\mathcal{X}_{t_{-w}}, (w_t + \Delta_w), \theta, \zeta_t) - H(\mathcal{X}_t, \theta, \zeta_t) \Big) dF_{\theta,\zeta}}{\Delta_w}, \tag{2.9}$$

where the average is taken over the marginal distribution of the error terms θ and ζ. However, the above could only be possible if the endogeneity of $\mathcal{X}_t$ were absent, that is, if the covariates $\mathcal{X}_t$ could be manipulated independently of the errors, θ and ζ_t.

To obtain the ASF, $G(\mathcal{X}_t)$, consider $\mathrm{E}(y_t|\mathcal{X}_t, \boldsymbol{X}, \mathcal{Z}) = \mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z})$. For an individual i, we have

$$\begin{aligned}\mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z}) &= \mathrm{E}(\mathrm{E}(H(\mathcal{X}_t, \theta, \zeta_t)|\boldsymbol{X}, \mathcal{Z}, \boldsymbol{\alpha})|\boldsymbol{X}, \mathcal{Z}) = \mathrm{E}(\mathrm{E}(H(\mathcal{X}_t, \theta, \zeta_t)|\boldsymbol{\epsilon}, \boldsymbol{\alpha})|\boldsymbol{X}, \mathcal{Z}) \\ &= \mathrm{E}(\mathrm{E}(H(\mathcal{X}_t, \theta, \zeta_t)|\boldsymbol{\epsilon}_t, \boldsymbol{\alpha})|\boldsymbol{X}, \mathcal{Z}) = \mathrm{E}(\tilde{H}(\mathcal{X}_t, \boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)|\boldsymbol{X}, \mathcal{Z}) \\ &= H^*(\mathcal{X}_t, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t) = \mathrm{E}(y_t|\mathcal{X}_t, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)\end{aligned}, \tag{2.10}$$

where the first equality is obtained by the Law of Iterated Expectation and the second follows from Assumption 5. The third equality follows from Assumption 6. In the fourth equality the intermediate regression function, $\tilde{H}(\mathcal{X}_t, \boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$, is the conditional CDF of $\theta + \zeta_t$ given $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\alpha}$, evaluated at $\boldsymbol{\varphi}(\mathcal{X}_t)$. That is

$$\tilde{H}(\mathcal{X}_t, \boldsymbol{\alpha}, \boldsymbol{\epsilon}_t) = F_{\theta+\zeta_t|\boldsymbol{\epsilon}_t, \boldsymbol{\alpha}}(\boldsymbol{\varphi}(\mathcal{X}_t)|\boldsymbol{\epsilon}_t, \boldsymbol{\alpha}).$$

Had we been able to identify $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}_t$ separately from one other and been able to provide estimates of them, we could have estimated $\tilde{H}(\mathcal{X}_t, \boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)$ as a multiple index model semi-parametrically as in Blundell and Powell (2004)(BP). Since we do not observe $\boldsymbol{\alpha}$ to identify $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}_t$, we resort to conditioning on the observables. Now, we have shown that

$$\mathrm{E}(\theta + \zeta_t|\boldsymbol{X}, \mathcal{Z}) = \mathrm{E}(\mathrm{E}(\theta + \zeta_t|\boldsymbol{\alpha}, \boldsymbol{X}, \mathcal{Z})|\boldsymbol{X}, \mathcal{Z}) = \mathrm{E}(\mathrm{E}(\theta + \zeta_t|\boldsymbol{\alpha}, \boldsymbol{\epsilon}_t)|\boldsymbol{X}, \mathcal{Z}) = \Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t.$$

To obtain the regression function, $H^*(\mathcal{X}_t, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$, the conditional CDF of $\theta + \zeta_t$ given $\boldsymbol{X}$ and $\mathcal{Z}$, we, like Chamberlain (1984), assume that

Assumption 8

$$\theta + \zeta_t|\boldsymbol{X}, \mathcal{Z} \sim \mathrm{N}\left[\mathrm{E}(\theta + \zeta_t|\boldsymbol{X}, \mathcal{Z}), \sigma_\zeta^2\right]$$

so that the tail, $\tilde{\zeta}_t = \theta + \zeta_t - \mathrm{E}(\theta + \zeta_t|\boldsymbol{X}, \mathcal{Z}) = \theta + \zeta_t - \Sigma_\alpha \hat{\boldsymbol{\alpha}} - \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t$ *is distributed normally with conditional mean 0.*

Given Assumption 8, we have

$$y_t = 1\{\mathcal{X}_t' \boldsymbol{\varphi} + \Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t + \tilde{\zeta}_t > 0\} \tag{2.11}$$

where $\mathcal{X}_t' \boldsymbol{\varphi}$ is the single index restriction on $\boldsymbol{\varphi}(\mathcal{X}_t)$.

If we do not assume a parametric distribution for the tail, $\tilde{\zeta}_t$, the structural quantities of interest, such as APE, could be obtained by nonparametric methods discussed in BP. If one is interested only in coefficients, $\boldsymbol{\varphi}$, the semiparametric maximum likelihood method developed in Rothe (2009) could be employed. Now, given that the control functions are generated regressors, employing these semiparametric methods would require that large sample properties of the estimates be worked out. Since the first stage parameters are obtained using parametric methods and the generated control functions are constructed

by using numerical integration technique, the large sample properties will be different from that developed in their papers.

The above cited two semiparametric control function methods are, however, not robust to heteroscedasticity and do not account for serial dependence among the response outcomes, which, when working with panel data, is likely to be an issue. Since these issues can be addressed in a straightforward manner in parametric models, we continue to resort to parametric specification.

Now, since in probit models the parameters are identified only up to a scale, we have

$$\mathrm{E}(y_t|\boldsymbol{X},\mathcal{Z}) = H^*(\mathcal{X}_t,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t) = \Pr(y_t=1|\mathcal{X}_t,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t) = \Phi\Big(\{\mathcal{X}_t'\boldsymbol{\varphi} + \Sigma_\alpha\hat{\boldsymbol{\alpha}} + \Sigma_\epsilon\hat{\boldsymbol{\epsilon}}_t\}\sigma_\zeta^{-1}\Big),$$

where the σ_ζ^2 is the variance of $\tilde{\zeta}_{it}$, which could be heteroscedastic. Having obtain $H^*(\mathcal{X}_t,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t)$, the measure ASF, $G(\mathcal{X}_t)$, can be obtained by averaging over $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$.

$$G(\mathcal{X}_t) = \Pr(y_t=1|\mathcal{X}_t) = \int H^*(\mathcal{X}_t,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t)dF_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t} = \int \Phi\Big(\frac{\mathcal{X}_t'\boldsymbol{\varphi} + \Sigma_\alpha\hat{\boldsymbol{\alpha}} + \Sigma_\epsilon\hat{\boldsymbol{\epsilon}}_t}{\sigma_\zeta}\Big)dF_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\epsilon}}_t} \tag{2.12}$$

Given (2.12), the APE, $\frac{\Delta G(\mathcal{X}_t)}{\Delta_w}$ in (2.9), of changing a variable, say w_t, from w_t to $w_t+\Delta_w$ can be obtained by dividing the difference in the ASF's at w_t and $w_t+\Delta_w$ by Δ_w. In our case, since the integrand is a smooth function of its arguments, in the limit when Δ_w tends to zero we can change the order of differentiation and integration in (2.12) to get

$$\frac{\partial G(\mathcal{X}_t)}{\partial w} = \frac{\partial \Pr(y_t=1|\mathcal{X}_t)}{\partial w} = \int \frac{\varphi_w}{\sigma_\zeta}\phi\Big(\frac{\bar{\mathcal{X}}'\boldsymbol{\varphi} + \Sigma_\alpha\hat{\boldsymbol{\alpha}} + \Sigma_\epsilon\hat{\boldsymbol{\epsilon}}_t}{\sigma_\zeta}\Big)dF_{\hat{\boldsymbol{\epsilon}}_t,\hat{\boldsymbol{\alpha}}}, \tag{2.13}$$

where ϕ is the density function of a standard normal and φ_w is the coefficient of w.

Before we proceed, we state Lemma 2, where we show that $\hat{\boldsymbol{\alpha}}_i(\boldsymbol{X}_i,\mathcal{Z}_i,\Theta_1)$ and $\hat{\boldsymbol{\epsilon}}_{it}(\boldsymbol{X}_i,\mathcal{Z}_i,\Theta_1)$ satisfy the the properties of a control function.

LEMMA 2 *Conditional on $\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X},\mathcal{Z},\hat{\Theta}_1)$ and $\hat{\boldsymbol{\alpha}}(\boldsymbol{X},\mathcal{Z},\hat{\Theta}_1)$, θ and ζ_t are independent of $\mathcal{X}_t$.*

PROOF OF LEMMA 2 Given in appendix A

Now, in order for the ASF $G(\mathcal{X}_t)$ and APE $\frac{\partial G(\mathcal{X}_t)}{\partial w}$ to be identified from the partial mean formulation in (2.12) and (2.13) for a particular value $\bar{\mathcal{X}}$ of $\mathcal{X}_t$, the support of the conditional distribution of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ given $\bar{\mathcal{X}}$ must be the same as the support of the marginal distribution of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ (conditionally on the exogenous $\mathcal{Z}$), (see Florens *et al.*, 2008; Imbens and Newey, 2009). In our approach, because $\hat{\boldsymbol{\alpha}}$ is a continuous and monotonic functions of $\boldsymbol{x}_t$, $\forall t$ (see Lemma 3) and because $\boldsymbol{x}_s$, $s\neq t$, which is unrestricted and has an unbounded support, the support of the conditional distribution – conditional on $\bar{\mathcal{X}}$ – of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta} - \hat{\boldsymbol{\alpha}}$ are unbounded too.

LEMMA 3 *The support of the conditional distribution of* $\hat{\boldsymbol{\alpha}}(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$ *and* $\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$, *given* $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$[6], *is the same as the marginal distribution of* $\hat{\boldsymbol{\alpha}}(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$ *and* $\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$ *(conditionally on* $\mathcal{Z}$*).*

PROOF OF LEMMA 3 Given in appendix A

The consequence of Lemma 3 is that

$$\mathrm{E}(y|\bar{\boldsymbol{x}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t) = \mathrm{E}(y|\bar{\boldsymbol{x}}, \hat{\boldsymbol{\alpha}}(\bar{\boldsymbol{x}}), \hat{\boldsymbol{\epsilon}}_t(\bar{\boldsymbol{x}})) \tag{2.14}$$

for all $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ in the unconditional support of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ and where $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$. That is, we can replace $\mathrm{E}(y|\bar{\boldsymbol{x}}, \hat{\boldsymbol{\alpha}}(\bar{\boldsymbol{x}}), \hat{\boldsymbol{\epsilon}}_t(\bar{\boldsymbol{x}}))$ by $\mathrm{E}(y|\bar{\boldsymbol{x}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$, which allows us to obtain the ASF $G(\bar{\boldsymbol{x}}) = \int \mathrm{E}(y = 1|\bar{\boldsymbol{x}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t) dF_{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t}$, or for that matter the APE. We can do this because according to Lemma 3, for any given $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}(\boldsymbol{x}_t = \bar{\boldsymbol{x}}_t, \boldsymbol{x}_s = \bar{\boldsymbol{x}}_s, \boldsymbol{X}_{-t,s})$, we can find an $\boldsymbol{x}_s = \boldsymbol{x}_s^*$, such that $\hat{\boldsymbol{\alpha}}(\boldsymbol{x}_t^*, \boldsymbol{x}_s^*, \boldsymbol{X}_{-t,s}) = \hat{\boldsymbol{\alpha}}^*$ for any $\boldsymbol{x}_t = \boldsymbol{x}_t^*$. The same, by Lemma 3, holds true for any $\hat{\boldsymbol{\epsilon}}_t = \hat{\boldsymbol{\epsilon}}^*$.

Moreover, since the result in (2.14) was obtained conditionally on the exogenous $\mathcal{Z}$, our method circumvents the need to have continuous instruments, often with large support, as is required in most semi and nonparametric control function methods in the literature.

Finally, we would like to note that the identification results derived here for binary choice model are easily extended to identify structural measures of interest for other nonlinear models such as, to name a few, selection, bivariate probit, and tobit models. Also, the structural equation (2.1) can allow for limited number of lagged values of exogenous variables. This would require that the reduced form equations be estimated with lagged values of $\boldsymbol{z}_t^y$. Provided the number of time periods is sufficient, the reduced form equations can be estimated with lagged values of exogenous variables as described above.

2.1.2. *Estimation*

We know that in probit model, heteroscedasticity in the latent variable when unaccounted leads to inconsistent maximum likelihood estimates of the coefficients and of the covariance matrix. To obtain consistent estimates of the structural parameters, we address this source of inconsistency by modeling heteroscedasticity as a variation of Harvey's "multiplicative heteroscedasticity" approach. We assume that the conditional variance of $\tilde{\zeta}_t$ as $\sigma_\zeta^2(\boldsymbol{X}_i, \mathcal{Z}_i) = \exp(h(W_i))^2$, where h is assumed to be linear. To include elements of $\boldsymbol{X}_i$ and $\mathcal{Z}_i$, $h(W_i)$ can be specified with a Mundlak or a Chamberlain type specification, and/or W_i can include $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ and their squares[7]. Assuming that W_i includes $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\epsilon}}_{it}$, in what follows, we will denote $h(W(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t))$ with $h(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$.

[6]Since $\boldsymbol{z}_t^y$ of $\mathcal{X}_t$ are also elements of $\mathcal{Z}$, upon which $\hat{\boldsymbol{\alpha}}$ is already conditioned, this implies that we consider the marginal and the conditional only with respect to $\boldsymbol{x}_t$.

[7]Since

$$\begin{aligned}\mathrm{Var}(\zeta_t + \theta|\boldsymbol{X}, \mathcal{Z}) &= \mathrm{E}((\zeta_t + \theta)^2|\boldsymbol{X}, \mathcal{Z}) - (\mathrm{E}(\zeta_t|\boldsymbol{X}, \mathcal{Z}) + \mathrm{E}(\theta|\boldsymbol{X}, \mathcal{Z}))^2 \\ &= \mathrm{E}((\zeta_t + \theta)^2|\boldsymbol{X}, \mathcal{Z}) - (\Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t)^2,\end{aligned}$$

$\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$ and their squares can be included in the specification for conditional variance.

If w belongs to both $\mathcal{X}_{it}$ and W_i then the APE of w on the probability of $y_t = 1$ at $\mathcal{X}_t = \bar{\mathcal{X}}$ is given by

$$\frac{\partial G(\mathcal{X}_t)}{\partial w} = \int \frac{\varphi_w - h'_w(\bar{\mathcal{X}}'\boldsymbol{\varphi} + \Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t)}{\exp(h(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t))} \phi\left(\frac{\bar{\mathcal{X}}'\boldsymbol{\varphi} + \Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t}{\exp(h(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t))}\right) dF_{\hat{\boldsymbol{\epsilon}}_t, \hat{\boldsymbol{\alpha}}}, \tag{2.15}$$

where h'_w is the derivative of $h(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$ with respect to w.

To obtain the parameters of interest, $\Theta_2 = \{\boldsymbol{\varphi}', \Sigma'_\alpha, \Sigma'_\epsilon, \Theta_{2h}\}'$, where Θ_{2h} is the set of parameters of $h(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t)$, one can employ nonlinear least squares by pooling the data. However, as PW discuss, since $\mathrm{Var}(y_t|\boldsymbol{X}, \mathcal{Z})$ will most likely be heteroscedastic and since there will be serial correlation across time in the joint distribution, $F(y_0, \ldots, y_T|\boldsymbol{X}, \mathcal{Z})$, the estimates, though consistent, will be estimated inefficiently resulting in biased standard errors. PW argue that modeling $F(y_0, \ldots, y_T|\boldsymbol{X}, \mathcal{Z})$ and applying MLE methods, while possible, is not trivial. Moreover, if the model for $F(y_0, \ldots, y_T|\boldsymbol{X}, \mathcal{Z})$ is misspecified but $\mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z})$ is correctly specified, the MLE will be inconsistent for Θ_2 and the resulting APEs.

To account for heteroscedasticity and serial dependence for the case where all covariates are exogenous, PW employ multivariate weighted nonlinear least squares (MWNLS) to obtain efficient estimates of Θ_2. To get the correct estimates of the standard errors of the estimates, what is required is a parametric model of $\mathrm{Var}(y_i|\boldsymbol{X}_i, \mathcal{Z}_i)$, where y_i is the $T \times 1$ vector of responses. For the conditional variances, $\mathrm{Var}(y_t|\boldsymbol{X}, \mathcal{Z})$, PW specify

$$\mathrm{Var}(y_t|\boldsymbol{X}, \mathcal{Z}) = \tau \mathbf{m}(\mathcal{W}_t, \Theta_2)(1 - \mathbf{m}(\mathcal{W}_t, \Theta_2)), \tag{2.16}$$

where $\mathcal{W}_t = \{\mathcal{X}'_t, \hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\epsilon}}'_t, W'\}'$, $\mathbf{m}(\mathcal{W}_t, \Theta_2) = \Phi\left(\frac{\mathcal{X}'_t\boldsymbol{\varphi} + \Sigma_\alpha \hat{\boldsymbol{\alpha}} + \Sigma_\epsilon \hat{\boldsymbol{\epsilon}}_t}{\exp(h(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\epsilon}}_t))}\right)$ and τ is such that $0 < \tau \leq 1$. For the covariance terms, $\mathrm{Cov}(y_t, y_s|\boldsymbol{X}, \mathcal{Z})$, in $\mathrm{Var}(y|\boldsymbol{X}, \mathcal{Z})$ a "working" version, which can be misspecified, is assumed. This is the approach underlying the generalized estimating equation (GEE) literature as described in Liang and Zeger (1986). The main advantage of GEEs 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 $\mathrm{Var}(y|\boldsymbol{X}, \mathcal{Z})$.

Generally, the conditional correlations, $\mathrm{Cov}(y_t, y_s|\boldsymbol{X}, \mathcal{Z})$, are a function of $\boldsymbol{X}$ and $\mathcal{Z}$. 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 $\boldsymbol{X}$ and $\mathcal{Z}$. 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 literature suggests that parameter ρ that characterize $\mathrm{Var}(y|\boldsymbol{X}, \mathcal{Z}) = \mathbf{V}(\boldsymbol{X}, \mathcal{Z}, \Theta_2, \tau, \rho)$

can be estimated using simple functions of residuals u_t

$$u_t = y_t - \mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z}) = y_t - \mathbf{m}(\mathcal{W}_t, \Theta_2),$$

where the mean function $\mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z})$ is correctly specified. With variance having been defined in (2.16), we can define standardized errors as

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

Then we have $\mathrm{Var}(e_t|\boldsymbol{X}, \mathcal{Z}) = \tau$. 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 the 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 as $\tilde{u}_t = y_t - \mathbf{m}(\mathcal{W}_t, \tilde{\Theta}_2)$ and the standardized residuals as

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

Then, a natural estimator of a common correlation coefficient is

$$\tilde{\rho} = \frac{1}{NT(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s \neq t} \tilde{e}_{it} \tilde{e}_{is}. \tag{2.17}$$

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

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

If $\mathrm{Corr}(e_t, e_s|\boldsymbol{X}, \mathcal{Z})$ 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:

$$\mathbf{V}(\boldsymbol{X}, \mathcal{Z}, \tilde{\Theta}_2, \tilde{\rho}) = \mathbf{D}(\boldsymbol{X}, \mathcal{Z}, \tilde{\Theta}_2)^{1/2} \mathbf{C}(\tilde{\rho}) \mathbf{D}(\boldsymbol{X}, \mathcal{Z}, \tilde{\Theta}_2)^{1/2} = \mathbf{V}(\boldsymbol{X}, \mathcal{Z}, \tilde{\Upsilon}) \tag{2.18}$$

where $\mathbf{D}(\boldsymbol{X}, \mathcal{Z}, \Theta_2)$ is the $T \times T$ diagonal matrix with $\mathbf{m}(\mathcal{W}_t, \Theta_2)(1 - \mathbf{m}(\mathcal{W}_t, \Theta_2))$ down its diagonal. (Note that dropping the variance scale factor, τ , has no effect on estimation or inference.) We can now proceed to the estimation of by MWNLS, that solves for $\hat{\Theta}_2$ by minimizing the following with respect to Θ_2.

$$\min_{\Theta_2} \sum_{i=1}^{N} [\mathbf{y}_i - \mathbf{m}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_2)]' [\mathbf{V}(\boldsymbol{X}_i, \mathcal{Z}_i, \tilde{\Upsilon})]^{-1} [\mathbf{y}_i - \mathbf{m}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_2)] \tag{2.19}$$

where $\mathbf{m}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_2)$ is the T vector with t^{th} element $\mathbf{m}(\mathcal{W}_{it}, \Theta_2)$.

In their model, PW, however, do not make any assumption about how the expectation $\mathrm{E}(y_t|\boldsymbol{x}_t, \mathcal{Z})$ would change if they condition the expectation on $\boldsymbol{x}_s$, $s \neq t$, also. The requirement of GEE is that the mean model, $\mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z})$, be correctly specified, else the GEE approach to estimation can give inconsistent results. Hence, when the covariates are endogenous, the model in PW is not suited for GEE estimation. Herein lies the *advantage* of our model compared to PW's. We have, given our identifying assumptions, been able to show that $\mathrm{E}(y_t|\boldsymbol{X}, \mathcal{Z}) = H^*(\mathcal{X}_t, \hat{\boldsymbol{\alpha}}(\boldsymbol{X}, \mathcal{Z}), \hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X}, \mathcal{Z}))$, and therefore we can employ GEE to account for serial correlation across time in presence of endogeneity.

Once the consistent estimates of Θ_2 are estimated, the sample analog of ASF $G(\mathcal{X}_t)$, for any fixed $\mathcal{X}_{it} = \bar{\mathcal{X}}$ can be computed as

$$\hat{G}(\mathcal{X}_t) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Phi\left(\frac{\bar{\mathcal{X}}'\hat{\boldsymbol{\varphi}} + \hat{\Sigma}_\alpha\hat{\hat{\boldsymbol{\alpha}}}_i + \hat{\Sigma}_\epsilon\hat{\hat{\boldsymbol{\epsilon}}}_{it}}{\exp(\hat{h}(\hat{\hat{\boldsymbol{\alpha}}}_i, \hat{\hat{\boldsymbol{\epsilon}}}_{it}))}\right). \tag{2.20}$$

By applying Lemma 1 it can be also shown that

$$\Phi\left(\frac{\bar{\mathcal{X}}'\hat{\boldsymbol{\varphi}} + \hat{\Sigma}_\alpha\hat{\hat{\boldsymbol{\alpha}}}_i + \hat{\Sigma}_\epsilon\hat{\hat{\boldsymbol{\epsilon}}}_{it}}{\exp(\hat{h}(\hat{\hat{\boldsymbol{\alpha}}}_i, \hat{\hat{\boldsymbol{\epsilon}}}_{it}))}\right) \xrightarrow{a.s.} \Phi\left(\frac{\bar{\mathcal{X}}'\boldsymbol{\varphi} + \Sigma_\alpha\hat{\boldsymbol{\alpha}}_i + \Sigma_\epsilon\hat{\boldsymbol{\epsilon}}_{it}}{\exp(h(\hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\epsilon}}_{it}))}\right).$$

This implies that by the weak LLN, $\hat{G}(\mathcal{X}_t)$ will converge in probability to $G(\mathcal{X}_t)$ as $NT \to \infty$. Similarly, for any fixed $\mathcal{X}_{it} = \bar{\mathcal{X}}$, an estimate of the APE of w can be computed by taking the sample analog of (2.15). The APE of a dummy variable, w, can be computed by taking the difference in ASF computed at $\{0, \bar{\mathcal{X}}'_{-w}\}'$ and $\{1, \bar{\mathcal{X}}'_{-w}\}'$.

While consistent second stage structural parameters are obtained when first stage estimates of Θ_1 are consistent, to obtain correct inference about the structural parameters, one has to account for the fact that instead of true value of Θ_1, we use its estimated value. In appendix B of the supplementary appendix we derive the asymptotic covariance matrix of the estimated second stage coefficients; the standard errors of the APE's are derived in appendix C.1.

3. AN APPLICATION: IMPLICATIONS OF POVERTY AND MIDDAY MEAL SCHEME ON CHILD LABOR & SCHOOLING

3.1. *Introduction*

Child labor is a pressing concern in all developing countries. According to International Labour Office's (ILO) current estimates, 168 million children in the 5 to 14 year age group are working in economic activities throughout the world; 78 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[8], and,

[8]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.

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) are 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 some papers like Bhalotra and Heady (2003) (BH) 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.

BH 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 forego the income that child labor would bring. On the other, 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 and this is especially relevant 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 may not necessarily increase child labor. They point out that in presence of land and labor market imperfections, ownership of land and livestock generate incentives for child labor, but assets such as oxen and ploughs that are operated by adults decrease child labor. To test this hypothesis, in our empirical specification we include the size of landholding, as well as 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 BH 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 in situations where an inverse relationship between agrarian assets and child labor hold, such transfers could reduce the incidence of child labor.

In our data set we find mean non-agricultural income to be much higher than agricultural income. This suggests that land is not the only source of income as in BDD and BH. BDD based on the assumption that land is the only source of income, derive a regression equation where landholding is the only measure of well being. Given that 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. Now, if land market exists, no matter how imperfect, in the regions from where our data has been collected, then it is unlikely that land owned by household will be exogenous to household labor supply as in BH and BDD, where land is mainly inherited, but endogenously determined along with household's, including children's, labor supply decisions. This would necessitate accounting for the endogeneity of landholding along with the endogeneity of productive assets and household income. To solve the endogeneity problem we employ the method developed in the paper.

BR show that child labor, which hampers human capital accumulation, can be socially inefficient, and that the family cannot be expected to solve this source of inefficiency on its own. They show that ban on child labor, or, more generally, government policies that seek to alleviate child labor could be welfare enhancing. One such program is the Government of India's "Midday Meal Scheme". It involves provision for free lunch on working days for children in primary classes (classes I to V), which in 2008-9 was extended to include upper primary classes (classes VI to VIII), in Government and Government aided schools. The scheme aims to provide cooked meal to children, with the objectives of providing nutrition to children, encouraging poor children and those belonging to disadvantaged sections to attend school more regularly, so that enrollment, retention and attendance rates increase[9]. Hence, keeping in view the suggested importance of policy interventions, we investigate if the midday meal scheme affects the incidence of child labor and the propensity of children to attend school.

[9] According to the Government of India, the Midday Meal Scheme is the world's largest school feeding programme. To find more about the scheme, visit http://mdm.nic.in.

3.2. *Data and Empirical Model*

3.2.1. *Data*

We conduct our empirical analysis at the level of the child using the 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 information on relevant covariates in either of the years was missing, we were left with 2458 children, which meant dropping about 23% children from the balanced panel of rural children.

Table 1 and 2 describes the relevant descriptive statistics for 2007 and 2010.

[Table 1, 2, and 3 about here]

The definition of work[10] includes (a) wage labor, (b) non-wage labor and (c) domestic 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[11]. 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 1, the proportion of children working has increased over the period of study. The major component of work (not reported here) is due to domestic chores. But, 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, we find proportion of older children going to school has dropped, but the proportion of younger children going to school has increased over the years.

In Table 2 we can see that the mean annual household income (in 2009 rupees) increased during this period, and that the non-agricultural income constitutes the major portion of the the household income. However, the increase in the mean agriculture income has been higher than non-agricultural income. We also find that the size of the mean landholding has increased over the years, and so has the index of farming related productive assets[12]. Also,

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

[11] For a discussion on whether or not to include domestic work in child labor, see Basu *et al.* and Edmonds. It has been argued that domestic work is often light and can entail learning essential skills. On the other hand, some domestic work such as cooking, cleaning, or taking care of younger siblings can be exhausting. Further, not including domestic work in child labor creates the false impression that girls do less work than boys. In fact, if we define work to exclude domestic chores, a major part of child labor is ignored. Besides, as pointed out by Edmonds, household chores may not be interpreted as non-economic work since the associated activities are not inelastic with respect to economic factors.

[12] 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. The assets constitute of agriculture tools, carts, pesticide pumps, ploughs, water pumps, threshers, tractors, and other farm

it can be evinced from Table 2 that the size of landholding has become more unequal. Among other variables, we see that the number of boys in the data are slightly higher compared to number of girls, and that the average number of years of education of fathers is higher than that of mothers. We also see that the household size has remained more or less unchanged during this period.

The information on whether midday meal was provided at school or not is available only for the "Index Children", who are the ones that have been followed in all YL surveys and will be followed in the subsequent ones. In Table 3 we present some statistics for the Index Children. As stated earlier, the midday meal scheme was initially meant only for the students at the primary level (standard I-V), and it was only in 2008-9, when the second round of survey was being conducted, that the scheme was extended for students of the upper primary level (standard VI-VIII) in all regions in India. Consequently, we do not find older children, aged 14 and 15 years in 2010, being served midday meal at schools. Nevertheless, in 2010 a higher proportion of young children report that their schools served midday meal.

3.2.2. *Empirical Model*

We denote by $y_t = DWORK_t$, the binary outcome variable that takes value 1 if the child decides[13] to work and 0 otherwise. We model the decision to work as

$$y_t = 1\{y_t^* = \mathcal{X}_t'\boldsymbol{\varphi} + \theta + \zeta_t > 0\}, \tag{3.1}$$

where y_t^* is amount of time devoted to work. When we study the decision to spend time at school, $y_t = DSCHOOL_t$. Now, a household sends its child to school if present value of its lifetime utility due to enhanced human capital accumulation from sending the child to school is higher than present value of its lifetime utility when the child does not attend school but works (see Gunnarsson *et al.*, 2006). It can be argued that the difference between the present value of household's lifetime utility when the child goes to school in time period t and the same when the child works translates proportionately into the number of hours spent schooling. So, when $y_t = DSCHOOL_t$, the latent y_t^* can be construed as the number of hours spent at school.

In (3.1) $\mathcal{X}_t = \{\boldsymbol{z}_t^{y\prime}, \boldsymbol{x}_t'\}'$, where $\boldsymbol{z}_t^y$ is a set of strictly exogenous variables and $\boldsymbol{x}_t$ are endogenous. Here, $\boldsymbol{x}_t = \{IN_t, LN_t, W_t\}'$, where IN_t is income of the household to which the child i belongs, LN_t is the size of the landholding, and W_t is the index of productive farm assets.

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 for IN_t, LN_t, and

equipments.

[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 to read more on the debate. Here we maintain that parents' decisions regarding their child is that of the child's.

W_t given by

$$\boldsymbol{x}_{it} = \begin{bmatrix} IN_{it} \\ LN_{it} \\ W_{it} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_{it} & 0 & 0 \\ 0 & \boldsymbol{z}_{it} & 0 \\ 0 & 0 & \boldsymbol{z}_{it} \end{bmatrix}' \begin{bmatrix} \beta_{IN} \\ \beta_{LN} \\ \beta_{W} \end{bmatrix} + \begin{bmatrix} \alpha_{INi} \\ \alpha_{LNi} \\ \alpha_{Wi} \end{bmatrix} + \begin{bmatrix} \epsilon_{INit} \\ \epsilon_{LNit} \\ \epsilon_{Wit} \end{bmatrix}$$
$$= \text{diag}(\boldsymbol{z}_{it}, \ldots, \boldsymbol{z}_{it})'\boldsymbol{\beta} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{it}, \tag{3.2}$$

where $\boldsymbol{z}_t = \{\boldsymbol{z}_t^{y\prime}, \tilde{\boldsymbol{z}}_t'\}'$ are strictly exogenous. $\tilde{\boldsymbol{z}}_t$ is the vector of instruments, whose dimension is greater than or equal to that of $\boldsymbol{x}_t$; and $\boldsymbol{z}_t$ is correlated with unobserved heterogeneity, $\boldsymbol{\alpha}_i$. With additive separability of errors, the triangular representation in (3.1) and (3.2) can derived from a fully simultaneous system, where landholding, productive assets and child's labor supply, y_t^*, determine household income.

In this application we have assumed that $\boldsymbol{\epsilon}_t$'s are distributed independently across time periods[14]. Given Assumption 2 and Assumption 3 we can write (3.2) as

$$\boldsymbol{x}_t = \boldsymbol{Z}_t'\boldsymbol{\delta} + \tilde{\boldsymbol{\alpha}} + \boldsymbol{\epsilon}_t, \tag{3.2a}$$

where the distributions of $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\epsilon}_t$ have been specified in section 2. We estimate the parameters, Θ_1, of equation (3.2a) using Biørn stepwise ML method. Having estimated Θ_1, the modified structural equation augmented with control functions, $\hat{\boldsymbol{\alpha}}_i(\Theta_1)$ and $\hat{\boldsymbol{\epsilon}}_{it}(\Theta_1)$, that eliminates the bias due to presence of endogenous regressors, is given by

$$y_t = 1\{\mathcal{X}_t'\boldsymbol{\varphi} + \Sigma_\alpha\hat{\boldsymbol{\alpha}} + \Sigma_\epsilon\hat{\boldsymbol{\epsilon}}_t + \tilde{\zeta}_t > 0\}, \tag{3.3}$$

where $\tilde{\zeta}_t$ by our assumption is distributed normally with mean 0 and is allowed to be heteroscedastic. To estimate the slope coefficients in (3.3), we simply pool the data and employ ML method. Inference about Σ_ϵ and Σ_α provides us with a test of exogeneity of the regressors, $\boldsymbol{x}$.

To identify the impact of the three endogenous variables income, IN, landholding, LN, and asset holding, W, on the decision to participate in work or go to school we employ the following instruments ($\tilde{\boldsymbol{z}}_t$ in (3.2)): (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.* (2013) 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 of the households's locality/settlement.

[14] As stated earlier, we have not found any test that can test for vector autocorrelation of the idiosyncratic component when one is dealing with a system of regressions with individual effects. We also alluded to an approximate method, where one might test for serial dependence for each of the regressions separately in the system of regressions (see footnote 3). However, these existing tests require the presence of at least three waves of data, and we have only two. Hence, without being able to test for serial dependence, assuming limited order serial dependence among the idiosyncratic component also risks misspecification. Besides, the two waves in our data are separated by three time periods. Hence, it seems unlikely that the serial dependence in idiosyncratic term will be strong over the span of three years.

The National Rural Employment Guarantee Scheme (NREGS) was initiated in 2006 by the Government of India, whose objective is to alleviate rural poverty. NREGS legally entitles rural households to 100 days of employment in unskilled manual labor (on public work projects) at a prefixed wage. Afridi *et al.* argue that more funds sanctioned at the *mandal* level would mean more work opportunity in NREGS, which will have a positive effect on household income. Now, it can be seen in Table 2 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 in any region at the beginning of the financial year is not affected by current demand for work, the funds sanctioned is exogenous and more funds imply more work opportunity in NREGS, which can have a positive effect on household income. The last row in Table 2 suggests that the total fund allocation to NREGS increased during the period 2007-2010. However, this increase was not uniform across the 15 *mandals*[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 Class" (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

[15]Data on the sanctioned funds at the *mandal* level has been 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/ST group 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 Class (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 conditions of the communities. In the constitution, OBC's are described as "socially and educationally backward classes", and government is enjoined to ensure their social and educational development.

wealth and income are evidently distributed along caste lines[18], 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[19]. 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, occupation choice, wages and consumption 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, and that the overall consumption and wage convergence between the groups has been driven significantly by convergence in the educational choices of the two groups.

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 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 BH, these variables, which indicate the level of infrastructure development, are employed to instrument the index of productive farm assets.

3.3. *Discussion of Results*

We begin by discussing the results of the first stage reduced form equations (4.2), which was estimated using Biørn's stepwise maximum likelihood method for system of equations. 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

[18]For more on why SC/ST's and OBC'c continue to lag behind economically, see Iyer *et al.* (2013) and Munshi (2011).

[19]The figures in Tables 4 are based on a slightly larger data set than what was used to obtain the main results. This is because information on productive assets was not available for every household.

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 results of the household income and wealth implication for child labor are illustrated in Table 6. Here, we would like to state that 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 for and supply of schooling or work over time. Secondly, in Table 6 we also compare Chamberlain's method of panel probit with correlated random effects with the one developed in this paper, termed "Control Function" method. Thirdly, the basis for choosing the specification for the heteroscedastic variance for the Control Function method was simply the significance of the variables in the variance specification. Fourthly, all the average partial effects (APE's) of variables on the decision to work or go to school were computed at the mean of variables from the second round (2010) of data.

[Table 6 about here]

To begin with, we find that all the control functions – α_{IN}, α_{LN}, α_W, ϵ_{IN}, ϵ_{LN}, ϵ_W – 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. The coefficient estimate suggest that children of households that have a higher landholding are more likely to engage in work[20]. This is in conformity with the findings in BDD, BH and CD, where, due to presence of land and labor market imperfections, ownership of large amount land provides incentives for children to work. However, since the APE of landholding is not significant, it is unlikely that an increase in the landholding of an "average family" will have any effect the participation decision of the child.

While it is beyond the scope of this paper to empirically test for land and labor market imperfections, it can nonetheless be argued that in AP, including the rural areas, there has been a weakening of imperfections in the two markets. First, the insignificance of APE of landholding could be because the average family in rural AP, with higher non-agricultural income compared to agriculture, does not rely solely on land for its income. Besides, there is evidence that higher levels of non-agricultural development and access to non-agricultural income, which is brought about by improvement in other factor markets, make land sales market more perfect. Deininger *et al.* (2007) report that better access to technology tends to improve farmers' ability to acquire land through sales markets. Moreover, Deininger *et al.*, confirming other studies, find that the land sales market is much more active in southern India, where our data is from.

Besley and Burgess (2004) (from 1952 to 1992) and Aghion *et al.* (2008) (from 1980 to 1997) code state level amendments to the Industrial Disputes Act of 1947, a central legislation governing labor market regulation, as pro-worker, neutral or pro-employer to graph

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

the history of regulatory change across states in India. They find that AP has consistently amended the Act to institute labor market flexibility, and that trade liberalization benefited those states where reforms to the labor market institutions were carried out to make labor market more flexible[21]. Iarossi (2009), who developed a Investment Climate Index aimed at summarizing the aspects of the business environment that entrepreneurs consider when deciding whether to invest, finds AP to be among the more favorable state (ranked 4^{th}) as far as private investment is concerned.

Greater reliance of the average family on non-agricultural income and the above facts concerning weakening of land and labor market imperfection could explain why increase in the size of landholding does not significantly increase the propensity of child labor for the average family.

As far as income is concerned, the coefficient estimates suggests that as household income rises, the probability of the household's children working decreases. While this does confirm poverty to be a cause of child labor, the APE of income for an average family is not significant. We find that ownership of productive farm assets, distinct from land, leads to a significant reduction in children's participation in work for the average family. Dumas, BH 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[22].

In other results, we find that older children and boys are more likely to work. Judging by the APE of household size, it seems that household size does not play any significant role in determining child labor. This could be because, in our sample, while decisions of children to work and attend school have changed overtime, household size and the number

[21]In India, manufacturing is comprised of two sub-sectors: an unregistered (informal) sector of small firms and a registered (formal) sector of larger firms. Firms in the registered sector are covered by the Industrial Disputes Act, while firms in the informal sector are not covered by labor regulations. The organized sector, however, provides employment to only 6% of the total. Virtually all employment in agriculture is within the unorganized sector. But even if agriculture is excluded, unorganized sector employment is as much as 83% of all non-farm employment. This is true even when the value generated by manufacturing in 2008 by the organised sector constituted about 55% of the total. The findings in Besley and Burgess and Aghion *et al.* pertain mostly to the organised sector; Kotwal *et al.* (2011) discuss how the above labor reform measures could have impacted the informal sector.

[22]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.

of children within households have remained more or less the same.

Our result suggests that, while father's education level does have a significant negative effect on child labor, children of mothers who are more educated are more likely to work. Now, what we find in our sample is that 16.3% of mothers who have had no education were employed in the non-agricultural sector. The corresponding percentages for those mothers who have had education up to high school and those with education level above high school are 18.3% and 17.9% respectively[23]. The others either work at home or in the agricultural sector. Hence, it seems that in rural AP higher level of education received does not translate into better work opportunity, and therefore higher income, for women. While an increase in mothers' income could improve their children's outcomes purely due to an income effect, it is also possible, as pointed out by Afridi *et al.*, that mother's say in household resource allocation decisions increases due to her higher earned income. This allows a greater weight being attached to her preferences, which includes investing more in their children's health and education relative to what fathers prefer, in resource allocation decisions of the household. Since higher education does not help mothers avail better work opportunity, which could improve children's well being, the positive effect of mother's education on her child's decision to work seems entirely spurious.

In Table 6 we compare the results obtained using Chamberlain's method with that developed in this paper. The results make clear the importance of accounting for endogeneity of income, landholding, and asset possession. When income and wealth are not instrumented, land or farm assets, does not significantly affect the incidence of child labor. Moreover, in the light of the discussion in the paper, the coefficient estimate for household income has an incorrect sign.

[Table 7 about here]

Finally, in Table 7 we show how availability of midday meal at schools affects the incidence of child labor and school attendance for the "Index Children", who are those that are being followed in every YL survey. Baland and Robinson show how a ban on child labor or subsidizing human capital creation, can be Pareto improving when child labor is inefficiently high and socially suboptimal. While both coefficient estimates and the APE of midday meal for school attendance are significant, the APE of midday meal for work is not. Our analysis, thus, shows that the provision of free midday meal at schools provides a strong incentive for children to attend school, but the evidence that this provision reduces the incidence of child labor is weak.

[Table 8 about here]

Also, as can be seen from Table 8, possession of productive farm assets and provision of midday meal have heterogeneous impact on school attendance. We find that the children belonging to SC/ST group are more likely than any other social group to attend school if there is availability of midday meal at schools. Similarly, at their mean level of wealth, an increase in the wealth level of an SC/ST family increases the likelihood of their children attending school more than that of children belonging to any other social group. This

[23]These figures may mask the income differential among educated and uneducated mothers, but we do not have information on income earned from their primary activity.

informs us that policy interventions such as the provision of midday meal at schools or monetary transfers, targeted for acquisition of productive farm assets, can be very effective in promoting education among children of socially and economically disadvantaged groups, who face greater obstacles than others in pulling themselves out of poverty.

4. CONCLUDING REMARKS

The primary objective of the paper has been to develop a control function estimation procedure to account for unobserved heterogeneity and endogeneity in nonlinear panel data models. Most papers studying nonlinear panel data models assume all covariates to be exogenous conditional on the unobserved heterogeneity. In this paper we relax this assumption to allow for endogenous covariates . The control functions are based on expected a Posteriori (EAP) values of correlated random effects. To compute the EAP values, numerical integration with respect to the estimated conditional distribution of unobserved heterogeneity is performed. The conditional distribution is obtained as the posterior distribution of the unobserved heterogeneity from the estimates of the first stage reduced form equations.

The proposed method makes a number of interesting contribution to the literature. First, the method allows for general instruments, a feature not available with most semi or nonparametric control function estimators in the literature. Second, the method accounts for heterogeneity in the reduced form/ treatment choice equation. Finally, the methodology makes no assumption about the serial dependence in the response outcomes and provides an estimation strategy to account for it.

The estimator was applied to estimate the causal effects of income and wealth – land and farm assets – on the incidence of child labor and school attendance, where most of the instruments used to obtain the estimates were discrete. We found that household ownership of productive farm assets significantly lowers the probability of child labor and significantly increases their chances of attending school, suggesting a strong income effect of farm assets. Secondly, we found little evidence that large landholding increases the incidence of child labor, a phenomenon attributed to land and labor market imperfections. Thirdly, a test of exogeneity reveled that ownership of land and farm assets are determined simultaneously with household labor supply decisions, contrary to what most empirical studies on child labor in developing countries assume. Finally, our results strongly suggest that policy interventions, such as the provision of free midday meal at school, increases the probability of school attendance, but found weak evidence that it reduces the incidence of child labor.

There are a number of extensions and generalizations that are desirable. First, allowing and testing for nonspherical errors component for the reduced form system of regressions could be a valuable extension. Alternatively, devising estimation strategy that is robust to misspecification of distributional assumptions in the reduced form equations could also be an important contribution. Secondly, it would be worthwhile to investigate if the proposed control functions could be estimated without making distributional assumptions, which could then lead to semiparametric estimation of the structural quantities of interest.

REFERENCES

Afridi, F., Mukhopadhyay, A. and Sahoo, S. (2013). Female Labor Force Participation and Child Education in India: The Effect of the National Rural Employment Guarantee Scheme, Economics and Planning Unit, Indian Statistical Institute, Manuscript.

Aghion, P., Burgess, R., Redding, S. J. and Zilibotti, F. (2008). The Unequal Effects of Liberalization: Evidence from Dismantling the License Raj in India. *American Economic Review*, **98**, 1397–1412.

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.

— and — (2012). Identifying Distributional Characteristics in Random Coefficients Panel Data Models. *Review of Economic Studies*, **79**, 987–1020.

— and Hahn, J. (2007). Understanding Bias in Nonlinear Panel Models: Some Recent Developments. In R. Blundell, W. Newey and T. Persson (eds.), *Advances in Economics and Econonometrics: Theory and Applications, Ninth World Congress*, Cambridge: Cambridge University Press.

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.

Besley, T. and Burgess, R. (2004). Can Labor Regulation Hinder Economic Performance? Evidence from India. *The Quarterly Journal of Economics*, **119**, 91–134.

Bester, C. A. and Hansen, C. (2007). Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity, Technical Report.

— and — (2009). A Penalty Function Approach to Bias Reduction in Non-linear Panel Models with Fixed Effects. *Journal of Business and Economic Statistics*, **27**, 131–148.

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. (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.

Chernozhukov, V., Fernandez-val, I., Hahn, J. and Newey, W. (2013). Quantile and Average Effects in Nonseparable Panel Models. *Econometrica*, **81**, 535–580.

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.

Deininger, K., Jin, S. and Nagarajan, H. K. (2007). *Determinants and Consequences of Land Sales Market Participation: Panel Evidence from India*. Tech. rep., World Bank Policy Research, Working Paper 4323.

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.

Fernandez-Val, I. (2009). Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models. *Journal of Econometrics*, **150**, 71–85.

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.

Gunnarsson, V., Orazem, P. F. and Sánchez, M. A. (2006). Child Labor and School Achievement in Latin America. *World Bank Economic Review*, **20**, 31–54.

Hahn, J. and Kuersteiner, G. (2011). Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects. *Econometric Theory*.

— and Newey, W. (2004). Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica*, **72**, 1295 – 1319.

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

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

Honore, B. and Tamer, E. (2006). Bounds on Parameters in Panel Dynamic Discrete Choice Models. *Econometrica*, **74**, 611–629.

Iarossi, G. (2009). *The Investment Climate in 16 Indian States*. Tech. rep., World Bank Policy Research, Working Paper 4817.

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

Iyer, L., Khanna, T. and Varshney, A. (2013). Caste and Entrepreneurship in India. *Economic and Political Weekly*, **48**, 52–60.

Kotwal, A., Ramaswami, B. and Wadhwa, W. (2011). Economic Liberalization and Indian Economic Growth: What's the Evidence? *Journal of Economic Literature*, **49**, 1152–99.

Lancaster, E. (2000). The Incidental Parameter Problem since 1948. *Journal of Econometrics*, **95**, 391–413.

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

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

Munshi, K. (2011). Strength in Numbers: Networks as a Solution to Occupational Traps. *Review of Economic Studies*, **78**, 1069 – 1101.

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.

Semykina, A. and Wooldridge, J. (2010). Estimating Panel Data Models in the presence of Endogeneity and Selection. *Journal of Econometrics*, **157**, 375–380.

Torgovitsky, A. (2012). Identification of Nonseparable Models with General Instruments, Northwestern University, Working paper.

Weidner, M. (2011). Semiparametric Estimation of Nonlinear Panel Data Models with Generalized Random Effects, University College London, Working Paper.

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

— (2009). Correlated Random Effects Models with Unbalanced Panels, Michigan State University, Department of Economics, Working Paper.

APPENDIX A: PROOFS

LEMMA 1 *$\check{\boldsymbol{\alpha}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \hat{\Theta}_1)$ converges a.s. to $\hat{\check{\boldsymbol{\alpha}}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_1^*)$, where $\hat{\Theta}_1 = \{\hat{\boldsymbol{\delta}}', vech(\hat{\Sigma}_{\epsilon\epsilon})', vech(\hat{\Lambda}_{\alpha\alpha})'\}'$ is consistent first stage estimates and Θ_1^* is the true population parameter.*

PROOF 1

Now for an individual i

$$\begin{aligned}\hat{\check{\boldsymbol{\alpha}}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_1) &= \frac{\int \tilde{\boldsymbol{\alpha}} \exp(-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{r}_t - \tilde{\boldsymbol{\alpha}})'\Sigma_{\epsilon\epsilon}^{-1}(\mathbf{r}_t - \tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int \exp(-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{r}_t - \tilde{\boldsymbol{\alpha}})'\Sigma_{\epsilon\epsilon}^{-1}(\mathbf{r}_t - \tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}} \\ &= \frac{\int C\mathfrak{a} \exp(-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{r}_t - C\mathfrak{a})'\Sigma_{\epsilon\epsilon}^{-1}(\mathbf{r}_t - C\mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}}{\int \exp(-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{r}_t - C\mathfrak{a})'\Sigma_{\epsilon\epsilon}^{-1}(\mathbf{r}_t - C\mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}} \\ &= \frac{\int \Sigma(\Theta_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}}{\int \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}}, \end{aligned} \tag{A-1}$$

where $\tilde{\boldsymbol{\alpha}} = C\mathfrak{a}$, CC' being the Cholesky decomposition of the covariance matrix $\Lambda_{\alpha\alpha}$. Hence, $d\tilde{\boldsymbol{\alpha}} = |C|d\mathfrak{a} = |\Lambda_{\alpha\alpha}|^{1/2}d\mathfrak{a}$, $\Omega(\Theta_1, \mathfrak{a}) = C\mathfrak{a}$, and finally $r(\Theta_1, \mathfrak{a}) = \sum_{t=1}^{T}(\boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta} - C\mathfrak{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{x}_t - \boldsymbol{Z}_t'\boldsymbol{\delta} - C\mathfrak{a})$.

First consider the expression in the numerator $\int \Omega(\Theta_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}$. Now, $\Omega(\Theta_1, \mathfrak{a}) = C\mathfrak{a}$ is an $m \times 1$ matrix and continuous in Θ_1 and $\mathfrak{a}$. Let $\Omega_l(\Theta_1, \mathfrak{a})$ be the l^{th} element of $\Omega(\Theta_1, \mathfrak{a})$. Now, by the assumptions of MLE we know that $\boldsymbol{\Theta_1}$ is a compact set, where $\Theta_1 \in \boldsymbol{\Theta_1}$, and also for a given $\mathfrak{a}$, $|\Omega_l(\Theta_1, \mathfrak{a})|$, $|.|$ being the absolute value of its argument, is continuous in Θ_1. Therefore $|\Omega_l(\Theta_1, \mathfrak{a})|$ attains its supremum on $\boldsymbol{\Theta_1}$. Let

$$\Theta_{l1}^{\mathfrak{a}} = \underset{\Theta_1 \in \boldsymbol{\Theta_1}}{\operatorname{argmax}} |\Omega_l(\Theta_1, \mathfrak{a})|,$$

then by an application of the Maximum Theorem we can conclude that $|\Omega_l(\Theta_{l1}^{\mathfrak{a}}, \mathfrak{a})|$ is continuous in $\mathfrak{a}$. The above then implies that $|\Omega_l(\Theta_{l1}^{\mathfrak{a}}, \mathfrak{a})| \geq \Omega_l(\Theta_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a})) \ \forall \Theta_1 \in \boldsymbol{\Theta_1}$. We also know that $\hat{\Theta}_1 \xrightarrow{a.s.} \Theta_1^*$, and since each of the $\Omega_l(\Theta_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))$, $l \in \{1, \ldots, m\}$, is continuous in Θ_1 and $\mathfrak{a}$, $\Omega_l(\hat{\Theta}_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\hat{\Theta}_1, \mathfrak{a})) \xrightarrow{a.s.} \Sigma_l(\Theta_1^*, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1^*, \mathfrak{a}))$ for any given $\mathfrak{a}$. Thus by an application of Lebesque Dominated Convergence Theorem we can conclude that $\int \Omega_l(\hat{\Theta}_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\hat{\Theta}_1, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a} \xrightarrow{a.s.} \int \Omega_l(\Theta_1^*, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1^*, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}$.

Define $\Omega(\Theta_1^{\mathfrak{a}}, \mathfrak{a}) = \{|\Omega_1(\Theta_{11}^{\mathfrak{a}}, \mathfrak{a})|, \ldots, |\Omega_m(\Theta_{m1}^{\mathfrak{a}}, \mathfrak{a})|\}'$, then $\Omega(\Theta_1^{\mathfrak{a}}, \mathfrak{a}) \geq \Omega(\Theta_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a})) \ \forall \Theta_1 \in \boldsymbol{\Theta_1}$, and Lebesque Dominated Convergence Theorem implies that

$$\int \Omega(\hat{\Theta}_1, \mathfrak{a}) \exp(-\frac{1}{2}r(\hat{\Theta}_1, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a} \xrightarrow{a.s.} \int \Omega(\Theta_1^*, \mathfrak{a}) \exp(-\frac{1}{2}r(\Theta_1^*, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}.$$

Also, since $1 \geq \exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))$, we can conclude that

$$\int \exp(-\frac{1}{2}r(\hat{\Theta}_1, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a} \xrightarrow{a.s.} \int \exp(-\frac{1}{2}r(\Theta_1^*, \mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}.$$

Given that both the numerator and the denominator in (A-1) defined at $\hat{\Theta}_1$ converge almost surly to the same defined at Θ_1^*, it can now be easily shown that

$$\check{\boldsymbol{\alpha}}(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1) \xrightarrow{a.s.} \hat{\check{\boldsymbol{\alpha}}}(\boldsymbol{X}, \mathcal{Z}, \Theta_1^*).$$

LEMMA 2 *Conditional on $\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$ and $\hat{\boldsymbol{\alpha}}(\boldsymbol{X}, \mathcal{Z}, \hat{\Theta}_1)$, ζ_t and θ are independent of $\mathcal{X}_t$.*

PROOF 2

Now,

$$\begin{aligned}\mathrm{E}[\zeta_t|\mathcal{X}_t,\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X},\mathcal{Z},\Theta_1),\hat{\boldsymbol{\alpha}}(\boldsymbol{X},\mathcal{Z},\hat{\Theta}_1)] &= \mathrm{E}[\zeta_t|\mathcal{X}_t,\boldsymbol{X},\mathcal{Z}] = \mathrm{E}[\zeta_t|\boldsymbol{X},\mathcal{Z}]\\ &= \mathrm{E}[\mathrm{E}[\zeta_t|\boldsymbol{\alpha},\boldsymbol{X},\mathcal{Z}]|\boldsymbol{X},\mathcal{Z}] = \mathrm{E}[\mathrm{E}[\zeta_t|\boldsymbol{\epsilon},\boldsymbol{\alpha}]|\boldsymbol{X},\mathcal{Z}]\\ &= \mathrm{E}[\Sigma_{\zeta\alpha}\boldsymbol{\alpha}+\Sigma_{\zeta\epsilon}\boldsymbol{\epsilon}_t|\boldsymbol{X},\mathcal{Z}]\\ &= \int(\Sigma_{\zeta\alpha}\boldsymbol{\alpha}+\Sigma_{\zeta\epsilon}\boldsymbol{\epsilon}_t)dF_{\hat{\boldsymbol{\alpha}}|\boldsymbol{X},\mathcal{Z}}\\ &= \Sigma_{\zeta\alpha}\hat{\boldsymbol{\alpha}}(\boldsymbol{X},\mathcal{Z},\Theta_1)+\Sigma_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X},\mathcal{Z},\Theta_1),\end{aligned}$$

where the second equality follows from the fact that $\mathcal{X}_t$ belongs to $\boldsymbol{X},\mathcal{Z}$ and the third from the law of iterated expectation. The fourth equality follows from Assumption 6, the fifth from Assumption 7, and the last two from the relationships in equations (2.4) and (2.5). Thus we have shown that conditional on $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$, ζ_t is mean independent of $\mathcal{X}_t$.

By Assumption 8 it we know that

$$\zeta_t|\mathcal{X}_t,\hat{\boldsymbol{\alpha}}(\boldsymbol{X},\mathcal{Z},\Theta_1),\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{X},\mathcal{Z},\Theta_1)\sim\zeta_t|\mathcal{X}_t,\boldsymbol{X},\mathcal{Z}\sim\zeta_t|\boldsymbol{X},\mathcal{Z}\sim\mathrm{N}[\Sigma_{\zeta\alpha}\hat{\boldsymbol{\alpha}}+\Sigma_{\zeta\epsilon}\hat{\boldsymbol{\epsilon}}_t,\sigma^2_\zeta].$$

Similarly, it can be shown that conditional on $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\epsilon}}_t$, θ is independent of $\mathcal{X}_t$.

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

Proof 3

Differentiating $\hat{\boldsymbol{\alpha}}$ with respect to $\boldsymbol{x}_t$ we get

$$\frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\boldsymbol{x}_t'}=\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}}{\partial\boldsymbol{x}_t'}=\left[\frac{\int\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}-\frac{\int\tilde{\boldsymbol{\alpha}}\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}\frac{\int\tilde{\boldsymbol{\alpha}}'\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}\right]\Sigma^{-1}_{\epsilon\epsilon}=\Sigma_{\alpha\alpha}\Sigma^{-1}_{\epsilon\epsilon},$$

where the expression in the square brackets, $\Sigma_{\alpha\alpha}$, is the second posterior moment of $\tilde{\boldsymbol{\alpha}}$, which is a positive definite matrix. Therefore $\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}(\boldsymbol{x}_t)}{\partial\boldsymbol{x}_t'}=\Sigma_{\alpha\alpha}\Sigma^{-1}_{\epsilon\epsilon}$ is invertible for all $\boldsymbol{x}_t\in\mathbb{R}^m$, for all t.

Without loss of generality assume that $T=2$ and let $\boldsymbol{x}_t$ be given, $\boldsymbol{x}_t=\bar{\boldsymbol{x}}$. Now, given that $\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}}{\partial\boldsymbol{x}_t'}$ is invertible for all t, by Inverse Function theorem $\hat{\tilde{\boldsymbol{\alpha}}}(\bar{\boldsymbol{x}},\boldsymbol{x}_{-t})$ is one-to-one on the open set U in $\mathbb{R}^m$, where $\boldsymbol{x}_{-t}\in U$, and there exists the inverse function $\mathcal{A}$ such that $\mathcal{A}(\hat{\tilde{\boldsymbol{\alpha}}}(\bar{\boldsymbol{x}},\boldsymbol{x}_{-t}))=\boldsymbol{x}_{-t}$.

For the sake of exposition assume that there is one endogenous variable x, so that, given $x_t=x_1=\bar{x}$, we have

$$\frac{\partial\hat{\alpha}(\bar{x},x_{-t})}{\partial x_{-t}}=\frac{1}{\sigma^2_\epsilon}\left[\frac{\int\alpha^2\exp(.)\phi(\tilde{\alpha})d\tilde{\alpha}}{\int\exp(.)\phi(\tilde{\alpha})d\tilde{\alpha}}-\left(\frac{\int\tilde{\alpha}\exp(.)\phi(\tilde{\alpha})d\tilde{\alpha}}{\int\exp(.)\phi(\tilde{\alpha})d\tilde{\alpha}}\right)^2\right]>0$$

because the expression in the square brackets, which is the second posterior moment of $\tilde{\alpha}$, is always positive. Therefore $\hat{\alpha}(\bar{x},x_{-t})$ is a one-to-one function of x_{-t}, and since x_{-t} has unbounded support, so does $\hat{\alpha}(\bar{x},x_{-t})$ for all $\bar{x}\in\mathbb{R}$.

Since $\hat{\epsilon}=x_t-\boldsymbol{Z}_t'\boldsymbol{\delta}-\hat{\tilde{\alpha}}$, for a given $x_t=\bar{x}$, $\hat{\epsilon}$ is also monotonic in x_{-t}, and hence has an unbounded support for all $\bar{x}\in\mathbb{R}$.

SUPPLEMENTARY MATERIALS

The supplementary material consists of the following files:

1 A supplementary appendix, attached at the end of the paper, has some additional technical details. Due to lack of space, these details could not be included in the main text of the paper.

2 A "rar" file, "children.rar", that contains the folder "children". The folder has the STATA data set and the STATA codes that can be used to replicate Table 5 and Table 6 of the paper. The file children.rar can be found at http://orbi.ulg.ac.be/handle/2268/169169.

TABLE 1

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

TABLE 2

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
Household Size	5.82	2.13	5.83	2.2
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 (Thousand Rupees)	30.91	34.35	48.88	60.24
Annual non-agricultural income (Rupees)	20787	35813	29013	62225
Annual agricultural income (Rupees)	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
Parents' characteristics				
Mother's Education (years spent)	2.53	4.35	2.53	4.35
Father's Education (years spent)	4.35	4.87	4.35	4.87
Community (Mandal) characteristics				
Total NREGS amount sanctioned (Rupees 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 3

Availability of Midday Meal and Schooling and Work Status for Index Children

		Young Children*	Older Children**
Year: 2007	Working (Yes=1, No=0)	0.061	0.645
	Attending School (Yes=1, No=0)	0.927	0.872
	Midday Meal available at School (Yes=1, No=0)	0.495	0.290
Year: 2010	Working (Yes=1, No=0)	0.326	0.781
	Attending School (Yes=1, No=0)	0.988	0.801
	Midday Meal available at School (Yes=1, No=0)	0.647	0.0

*Young Children: Aged 5 and 6 years in 2007
**Older Children: Aged 11 and 12 years in 2007
Number of Young Children in each Year: 882
Number of Older Children in each Year: 383

TABLE 4

Descriptive Statistics of some Variables by Caste

		Scheduled Caste/Scheduled Tribe	Other Backward Class	Others
Year: 2007	Household Income in Thousand Rs.	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 Rs.	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 (Rs. in Million)	0.0453***	-0.00843	-0.000308
	(0.00907)	(0.00683)	(0.000241)
Caste (SC/ST = 1, OBC = 2, OT = 3)	6.890***	2.177***	0.161***
	(1.157)	(0.743)	(0.0301)
Drinkable Water in the Locality (Yes=1 & No=0)	5.091	-1.907	0.339**
	(5.658)	(4.259)	(0.150)
National Bank in the Locality (Yes=1 & No=0)	-2.722	4.689**	0.0466
	(3.074)	(2.314)	(0.0816)
Engineered Road to the Locality (Yes=1 & No=0)	-0.0475	2.395	0.180***
	(2.114)	(1.591)	(0.0561)
Hospital in the Locality (Yes=1 & No=0)	-0.585	-4.134***	0.0563*
	(1.239)	(0.932)	(0.0329)
Other Exogenous Variables of the Structural Equations: Age, Sex of the Child, Mother's Education, Father's Education, Household Size	Yes	Yes	Yes

Standard errors in parentheses
Total number of observations : 4916
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.

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

TABLE 6.— **Household Income and Wealth Effect on Child's Decision to Work**

	Panel Probit†	**Control Function Approach**			
	Coefficients	**Coefficients**	**APE's**	**Control Functions**	
Income	0.00336***	-0.0231***	-0.001	$\hat{\alpha}_{INCOME}$	0.0261***
	(0.000902)	(0.00685)	(0.0053)		(0.00682)
Landholding	0.00785	0.0375**	0.0017	$\hat{\alpha}_{LAND}$	-0.0350***
	(0.00922)	(0.0163)	(0.0051)		(0.0133)
Asset Index	-0.00970	-0.942***	-0.0422***	$\hat{\alpha}_{ASSET}$	2.300***
	(0.0296)	(0.346)	(0.0166)		(0.361)
Age	2.105***	2.191***	0.0982***	$\hat{\epsilon}_{INCOME}$	0.0242***
	(0.114)	(0.310)	(0.0068)		(0.00720)
Sex	0.747***	1.006***	-0.0452**	$\hat{\epsilon}_{LAND}$	-0.0402***
	(0.0537)	(0.140)	(0.0203)		(0.0155)
Mother's Education	0.0195***	0.0399***	0.0017***	$\hat{\epsilon}_{ASSET}$	0.708**
	(0.00680)	(0.0130)	(0.0006)		(0.355)
Father's Education	-0.0234***	-0.0340**	-0.0014**		
	(0.00587)	(0.0140)	(0.0006)	**Specification for Heteroscedasticity**	
Household Size	-0.0729**	-0.0977**	-0.0044	Mother's Education	-0.0111
	(0.0337)	(0.0415)	(0.006)		(0.00692)
$\ln(\sigma_\theta^2)$‡	-1.352***			Father's Education	0.0116*
	(0.248)				(0.00669)
				$\hat{\alpha}_{INCOME}$	-0.00579***
					(0.00168)
				$\hat{\alpha}_{LAND}$	0.00910*
					(0.00500)
Total number of children: 2456					
Total number of observations: 4912. Total number of observations with positive outcome: 2126					

†Panel Probit is the Chamberlain's method with Correlated Random Effects.
‡σ_θ^2 is the panel-level standard deviation (see STATA command 'xtprobit').
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 7

Income, Wealth, and Midday Meal Effect on Decision to Work and Attend School for Index Children

	Work ($DWORK$)		**School** ($DSCHOOL$)	
	Coefficients	**APE's**	**Coefficients**	**APE's**
Income	-0.0589***	-0.0017	-0.00155	-0.0004
	(0.0240)	(0.0071)	(0.00720)	(0.0631)
Landholding	0.179***	0.0052	0.0531	0.0124
	(0.0692)	(0.0055)	(0.0364)	(0.0277)
Farm Asset Index	-3.362**	-.0873***	1.124***	0.2932***
	(1.327)	(0.0336)	(0.391)	(0.0994)
Age	5.739***	0.1494***	-1.294***	-0.3275
	(1.432)	(0.0296)	(0.285)	(0.6855)
Sex	1.851***	0.0533	-0.0940	-0.0246
	(0.586)	(0.071)	(0.103)	(0.049)
Mother's Education	0.0722	-0.0003	0.0530***	0.0138
	(0.0440)	(0.0019)	(0.0192)	(0.061)
Father's Education	-0.0511	-0.0015	0.0527***	0.0137
	(0.0414)	(0.007)1	(0.0167)	(0.0616)
Household Size	-0.283*	-0.0149**	-0.0562	-0.0147
	(0.172)	(0.0065)	(0.0432)	(0.0696)
Midday Meal	-0.816*	0.075	2.546***	0.6642***
	(0.464)	(0.0565)	(0.639)	(0.1922)
$\hat{\alpha}_{INCOME}$	0.0565**		-0.00529	
	(0.0240)		(0.00735)	
$\hat{\alpha}_{LAND}$	-0.0885**		-0.00610	
	(0.0437)		(0.0145)	
$\hat{\alpha}_{ASSET}$	5.839***		-1.380***	
	(1.593)		(0.335)	
$\hat{\epsilon}_{INCOME}$	0.0624**		0.00166	
	(0.0254)		(0.00771)	
$\hat{\epsilon}_{LAND}$	-0.185***		0.00358	
	(0.0678)		(0.0152)	
$\hat{\epsilon}_{ASSET}$	2.798**		-1.059***	
	(1.277)		(0.409)	

Total number of Index children: 1264; Total number of observations: 2528
Total number of observations with $DWORK = 1$: 883
Total number of observations with $DSCHOOL = 1$: 2348

All the specifications include time dummy, district dummies, and the interaction of time and district dummies.
The heteroscedastic specification for $DWORK$ includes age, household size, asset index, mother's education, midday meal, $\hat{\alpha}_{INCOME}$, $\hat{\alpha}_{LAND}$, and some district dummies. The heteroscedastic specification for $DSCHOOL$ includes age, landholding, $\hat{\alpha}_{INCOME}$, and some district dummies.
Standard errors are in parentheses.
Significance levels : $*$: 10% $**$: 5% $***$: 1%

TABLE 8

APE's of Farm Asset & Midday Meal on Child's Decision to Attend School for different Social Groups

	Scheduled Caste/Tribe	**Other Backward Class**	**Others**
Farm Asset Index	0.34***	0.2924***	0.1447
	(0.0925)	(0.0906)	(0.0968)
Midday Meal	0.7701***	0.6623***	0.3278*
	(0.1698)	(0.1695)	(0.1845)

Standard errors in parentheses.
Significance levels : $*$: 10% $**$: 5% $***$: 1%

Supplementary Appendix for Nonlinear Panel Data Model with Continuous Endogenous Regressors and General Instruments

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

APPENDIX A: Maximum Likelihood Estimation of the Reduced form Equations

In this section we briefly describe stepwise maximum likelihood procedure[1] in Biørn (2004) that we employ to estimate the reduced form system of equation

$$\boldsymbol{x}_{it} = \mathbf{Z}'_{it}\boldsymbol{\delta} + \tilde{\boldsymbol{\alpha}}_i + \boldsymbol{\epsilon}_{it}. \tag{A.1}$$

While Biørn deals with unbalanced panel, here we assume that our panel is balanced. Let N be the total number of individuals. Let $\boldsymbol{x}_{i(T)} = (\boldsymbol{x}'_{i1}, \ldots \boldsymbol{x}'_{ip})'$, $\mathbf{Z}_{i(T)} = (\mathbf{Z}'_{i1}, \ldots \mathbf{Z}'_{iT})'$ and $\boldsymbol{\epsilon}_{i(T)} = (\boldsymbol{\epsilon}'_{i1}, \ldots \boldsymbol{\epsilon}'_{iT})'$ and write the model as

$$\boldsymbol{x}_{i(T)} = \mathbf{Z}'_{i(T)}\boldsymbol{\delta} + (e_p \otimes \tilde{\boldsymbol{\alpha}}_i) + \boldsymbol{\epsilon}_{i(T)} = \mathbf{Z}'_{i(T)}\boldsymbol{\delta} + \boldsymbol{u}_{i(T)}, \tag{A.2}$$

$$\mathrm{E}(\boldsymbol{u}_{i(T)}\boldsymbol{u}'_{i(T)}) = I_T \otimes \Sigma_{\epsilon\epsilon} + E_T \otimes \Lambda_{\alpha\alpha} = K_T \otimes \Sigma_{\epsilon\epsilon} + J_T \otimes \Sigma_{(T)} = \Omega_{u(T)} \tag{A.3}$$

where

$$\Sigma_{(T)} = \Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha}, \tag{A.4}$$

and 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 $\Omega_{u(T)}$.

[1] The STATA routine "xtsur" implements Biørn's stepwise MLE for system of regressions.

A.1 GLS estimation

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

$$Q = \sum_{i=1}^{N} \boldsymbol{u}'_{i(T)}[K_T \otimes \Sigma^{-1}_{\epsilon\epsilon}]\boldsymbol{u}_{i(T)} + \sum_{i=1}^{N} \boldsymbol{u}'_{i(T)}[J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}]\boldsymbol{u}_{i(T)}. \tag{A.5}$$

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

$$\hat{\boldsymbol{\delta}}_{GLS} = \left[\sum_{i=1}^{N} \mathbf{Z}'_{i(T)}[K_T \otimes \Sigma^{-1}_{\epsilon\epsilon} + J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}]\mathbf{Z}_{i(T)}\right]^{-1} \times \left[\sum_{i=1}^{N} \mathbf{Z}'_{i(T)}[K_T \otimes \Sigma^{-1}_{\epsilon\epsilon} + J_T \otimes (\Sigma_{\epsilon\epsilon} + T\Lambda_{\alpha\alpha})^{-1}]\boldsymbol{x}_{i(T)}\right]. \tag{A.6}$$

A.1.1 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., $\tilde{\boldsymbol{\alpha}}_i \sim \text{IIN}(0, \Lambda_{\alpha\alpha})$ and $\boldsymbol{\epsilon}_{it} \sim \text{IIN}(0, \Sigma_{\epsilon\epsilon})$, then $\boldsymbol{u}_{i(T)} = (e_T \otimes \tilde{\boldsymbol{\alpha}}_i) + \boldsymbol{\epsilon}_{i(T)} \sim \text{IIN}(0_{mT,1}, \Omega_{u(T)})$. 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,

$$\mathcal{L}_i = \frac{-mT}{2}\ln(2\pi) - \frac{1}{2}\ln|\Omega_{u(T)}| - \frac{1}{2}Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}), \tag{A.7}$$

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_i = \frac{-mNT}{2}\ln(2\pi) - \frac{1}{2}N\ln|\Omega_{u(T)}| - \frac{1}{2}\sum_{i=1}^{N} Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}), \tag{A.8}$$

where

$$Q_{i(T)}(\boldsymbol{\delta}, \Sigma_{\epsilon\epsilon}, \Lambda_{\alpha\alpha}) = [\boldsymbol{x}_{i(T)} - \mathbf{Z}'_{i(T)}\boldsymbol{\delta}]'[K_T \otimes \Sigma^{-1}_{\epsilon\epsilon} + J_T \otimes (\Sigma_{\epsilon\epsilon} + p\Lambda_{\alpha\alpha})^{-1}][\boldsymbol{x}_{i(T)} - \mathbf{Z}'_{i(T)}\boldsymbol{\delta}], \tag{A.9}$$

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 (A.6). To solve *subproblem*(B) Biørn derives expressions for the derivatives of $\mathcal{L}$ with respect to $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$, which, for balanced panel, yields a closed form solution for $\Sigma_{\epsilon\epsilon}$ and $\Lambda_{\alpha\alpha}$. The complete stepwise algorithm for solving jointly subproblems (A) and (B) then consists in switching between (A.6) and minimizing (A.8) 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 an individual i with respect to $\boldsymbol{\delta}$, vech$(\Sigma_{\epsilon\epsilon})$ and vech$(\Lambda_{\alpha\alpha})$ are

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

$$\frac{\partial \mathcal{L}_i}{\partial \text{vec}(\Lambda_{\alpha\alpha})} = -\frac{1}{2}\text{vec}\left[T\Sigma_{(T)}^{-1} - T\Sigma_{(T)}^{-1}B_{ui(T)}\Sigma_{(T)}^{-1}\right],$$

and

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

where $W_{ui(T)}$ and $B_{ui(T)}$ respectively are

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

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

The second order conditions are:

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

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

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

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

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

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

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

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

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

APPENDIX B: Asymptotic Covariance Matrix for Structural Parameters

Obtaining second stage structural parameters with first stage consistent estimates $\hat{\Theta}_1$ is asymptotically equivalent to estimating the subsequent stage parameters with the true value of Θ_1^*. But 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. In this section we derive the asymptotic covariance matrix of the coefficients.

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 we write the first and second step estimation as an MM estimation based on the following population moment conditions:

$$\text{E}(\mathcal{L}_{i\Theta_1}) = E\frac{\partial \ln L_i(\Theta_1)}{\partial \Theta_1} = 0 \tag{B.1}$$

$$\text{E}(H_{i\Theta_2}(\Theta_1, \Theta_2)) = 0 \tag{B.2}$$

and where $L_i(\Theta_1)$ is the likelihood function for individual i for the first step system of reduced form equations and $\text{E}(H_{i\Theta_2}(\Theta_1, \Theta_2))$ 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

$$\frac{1}{N}\mathcal{L}_{\Theta_1}(\hat{\Theta}_1) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathcal{L}_i(\hat{\Theta}_1)}{\partial \Theta_1} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln L_i(\hat{\Theta}_1)}{\partial \Theta_1} \tag{B.3}$$

where $\mathcal{L}_i(\Theta_1)$ and the first order conditions with respect to Θ_1 are given in appendix A. $\Theta_1 = \{\boldsymbol{\delta}', \text{vech}(\Lambda_{\alpha\alpha})', \text{vech}(\Sigma_{\epsilon\epsilon})'\}'$ and N is the total number of individuals/firms.

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

$$\frac{1}{N} H_{\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2) = \frac{1}{N} \sum_{i=1}^{N} H_{i\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2). \tag{B.4}$$

We have shown that the structural equations augmented with the control functions $\hat{\boldsymbol{\alpha}}_i(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_1)$ and $\hat{\boldsymbol{\epsilon}}_{it}(\boldsymbol{X}_i, \mathcal{Z}_i, \Theta_1)$ 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. Newey has derived the asymptotic distribution of the second step estimates of a two step sequential estimator.

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

$$\frac{1}{N} \begin{bmatrix} \mathcal{L}_{\Theta_1}(\hat{\Theta}_1) \\ H_{\Theta_2}(\hat{\Theta}_1, \hat{\Theta}_2) \end{bmatrix} = 0. \tag{B.5}$$

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

$$\frac{1}{N} \begin{bmatrix} \mathcal{L}_{\Theta_1\Theta_1} & 0 \\ H_{\Theta_2\Theta_1} & H_{\Theta_2\Theta_2} \end{bmatrix} \begin{bmatrix} \sqrt{N}(\hat{\Theta}_1 - \Theta_1^*) \\ \sqrt{N}(\hat{\Theta}_2 - \Theta_2^*)) \end{bmatrix} = -\frac{1}{\sqrt{N}} \begin{bmatrix} \mathcal{L}_{\Theta_1} \\ H_{\Theta_2}. \end{bmatrix} \tag{B.6}$$

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) $B_{\Theta\Theta_N}$ 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 $B_*^{-1}A_*B_*^{-1\prime}$.

$$\sqrt{N}(\hat{\Theta}-\Theta^*) \overset{a}{\sim} \mathrm{N}[(0),(B_*^{-1}A_*B_*^{-1\prime})] \tag{B.7}$$

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

$$\begin{aligned} V_2^* = \mathbb{H}_{\Theta_2\Theta_2}^{-1}V_{HH}\mathbb{H}_{\Theta_2\Theta_2}^{-1\prime} + \mathbb{H}_{\Theta_2\Theta_2}^{-1}\mathbb{H}_{\Theta_2\Theta_1}\{\mathbb{L}_{\Theta_1\Theta_1}^{-1}V_{LL}\mathbb{L}_{\Theta_1\Theta_1}^{-1\prime}\}\mathbb{H}_{\Theta_2\Theta_1}'\mathbb{H}_{\Theta_2\Theta_2}^{-1\prime} \\ -\mathbb{H}_{\Theta_2\Theta_2}^{-1}\{\mathbb{H}_{\Theta_2\Theta_1}\mathbb{L}_{\Theta_1\Theta_1}^{-1}V_{LH} + V_{HL}\mathbb{L}_{\Theta_1\Theta_1}^{-1\prime}\mathbb{H}_{\Theta_2\Theta_1}'\}\mathbb{H}_{\Theta_2\Theta_2}^{-1\prime}. \end{aligned} \tag{B.8}$$

To estimate V_2^*, sample analog of the B_*, B_N given in (B.6), 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 $V_{LH_N} = \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{i\Theta_1}(\hat{\Theta}_1)H_{i\Theta_2}(\hat{\Theta}_1,\hat{\Theta}_2)'$. The first and the second order conditions for Biørn's MLE estimator to estimate Θ_1, to compute the sample analog of $\mathbb{L}_{\Theta_1\Theta_1}$, and to compute A_N are provided in appendix A of the supplementary appendix.

In what follows, we assume that the second stage structural estimation involves estimating a binary response model. The results can be straightforwardly adapted to estimate the covariance matrix for other nonlinear models. For binary response model the score function pertaining to the minimand in (2.19) of the main text is given by

$$\begin{aligned} H_{i\Theta_2}(\Theta_1,\Theta_2) &= -\nabla_{\Theta_2}\mathbf{m}_i(\boldsymbol{X}_i,\mathcal{Z}_i,\Theta_2)'[\mathbf{V}(\boldsymbol{X}_i,\mathcal{Z}_i,\tilde{\Upsilon})]^{-1}[\mathbf{y}_i-\mathbf{m}_i(\boldsymbol{X}_i,\mathcal{Z}_i,\Theta_2)] \\ &= -\nabla_{\Theta_2}\mathbf{m}_i(\Theta_1,\Theta_2)'\tilde{\mathbf{V}}^{-1}\mathbf{u}_i, \end{aligned} \tag{B.9}$$

where $\mathbf{m}_i(\Theta_1,\Theta_2) \equiv \mathbf{m}_i(\boldsymbol{X}_i,\mathcal{Z}_i,\Theta_2)$ is the T vector with t^{th} element $\mathbf{m}(\mathcal{W}_{it},\Theta_2) = \Phi\left(\frac{\mathcal{X}_{it}'\boldsymbol{\varphi}+\Sigma_\alpha\hat{\boldsymbol{\alpha}}_i+\Sigma_\epsilon\hat{\boldsymbol{\epsilon}}_{it}}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}\right) \equiv \mathbf{m}_{it}(\Theta_1,\Theta_2)$ and $\tilde{\mathbf{V}} \equiv \mathbf{V}(\boldsymbol{X}_i,\mathcal{Z}_i,\tilde{\Upsilon})$. Now

$$\begin{aligned} \nabla_{\Theta_{2s}}\mathbf{m}_{it}(\Theta_1,\Theta_2) &= \phi\left(\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}\right)\frac{1}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}\mathbb{X}_{it}' \text{ and} \\ \nabla_{\Theta_{2h}}\mathbf{m}_{it}(\Theta_1,\Theta_2) &= -\phi\left(\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}\right)\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}h_{\Theta_{2h}}'(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}), \end{aligned}$$

where $\mathbb{X}_{it} = \{\mathcal{X}_{it}',\hat{\boldsymbol{\alpha}}_i'(\Theta_1),\hat{\boldsymbol{\epsilon}}_{it}'(\Theta_1)\}'$, $\Theta_{2s} = \{\boldsymbol{\varphi}',\bar{\Sigma}_{\theta\alpha}',\tilde{\Sigma}_{\zeta\epsilon}'\}'$, and Θ_{2h} is such that $W_{it}'\Theta_{2h} = h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it})$. That is, $\Theta_2 = \{\Theta_{2s}',\Theta_{2h}'\}'$ and

$$\nabla_{\Theta_2}\mathbf{m}_{it}(\Theta_1,\Theta_2) = \phi\left(\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(W_{it}'\Theta_{2h})}\right)\frac{1}{\exp(W_{it}'\Theta_{2h})}\begin{bmatrix}\mathbb{X}_{it} & -(\mathbb{X}_{it}'\Theta_{2s})W_{it}\end{bmatrix}',$$

which is a row vector with dimension that of Θ_2. Wooldridge (2002) and Wooldridge (2003) show (see Problem 12.11) that $\mathbb{H}_{\Theta_2\Theta_2}$ of B^* is given by

$$\mathbb{H}_{\Theta_2\Theta_2} = \mathrm{E}[H_{i\Theta_2\Theta_2}(\Theta_1,\Theta_2)] = \mathrm{E}[\nabla_{\Theta_2}\mathbf{m}_i(\Theta_1,\Theta_2)'\tilde{\mathbf{V}}^{-1}\nabla_{\Theta_2}\mathbf{m}_i(\Theta_1,\Theta_2)], \tag{B.10}$$

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 $\hat{\mathbf{V}}=\mathbf{V}(\boldsymbol{X}_i,\mathcal{Z}_i,\hat{\Upsilon})=\mathbf{V}(\boldsymbol{X}_i,\mathcal{Z}_i,\hat{\Theta}_2,\hat{\rho})$.

Computation[2] of $H_{\Theta_2\Theta_1}=\sum_{i=1}^{N}H_{i\Theta_2\Theta_1}=\sum_{i=1}^{N}\frac{\partial H_{i\Theta_2}(\Theta_1,\Theta_2)}{\partial\Theta_1'}$ needed to obtain sample analog of $\mathbb{H}_{\Theta_2\Theta_1}$ can, however, be challenging because Θ_1 enters the second stage of the sequential estimator through $\hat{\boldsymbol{\alpha}}_i(\Theta_1)$ and $\tilde{\Sigma}_{\epsilon\epsilon}^{-1}\hat{\boldsymbol{\epsilon}}_{it}(\Theta_1)$. To obtain $\frac{\partial H_{i\Theta_2}(\Theta_1,\Theta_2)}{\partial\Theta_1'}$ consider the following

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

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

$$\mathbb{H}_{\Theta_2\Theta_1}=\text{E}[H_{i\Theta_2\Theta_1}(\Theta_1,\Theta_2)]=\text{E}[\nabla_{\Theta_2}\mathbf{m}_i(\Theta_1,\Theta_2)'\tilde{\mathbf{V}}^{-1}\nabla_{\Theta_1}\mathbf{m}_i(\Theta_1,\Theta_2)],\tag{B.11}$$

which can be approximated as

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

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

$$\nabla_{\Theta_1}\mathbf{m}_{it}(\Theta_1,\Theta_2)=\phi\left(\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(h)}\right)\frac{1}{\exp(h)}\left(\Theta_{2s}'\frac{\partial\mathbb{X}_{it}}{\partial\Theta_1'}-\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(h)}(\Theta_{2h}'\frac{\partial W_{it}}{\partial\Theta_1'})\right),\tag{B.12}$$

[2]In the MLE framework,

$$H_{i\Theta_{2s}}(\Theta_1,\Theta_2)=\sum_{t=1}^{T}\frac{(y_{it}-\Phi(.))\phi(.)\mathbb{X}_{it}'}{\Phi(.)(1-\Phi(.))}\text{ and }H_{i\Theta_{2h}}(\Theta_1,\Theta_2)=-\sum_{t=1}^{T}\frac{(y_{it}-\Phi(.))\phi(.)\mathbb{X}_{it}'\Theta_{2s}W_{it}'}{\Phi(.)(1-\Phi(.))\exp(W_{it}'\Theta_{2h})}.$$

Since $\text{E}(y_{it}-\Phi(.)|\boldsymbol{X},\mathcal{Z})=0$, it can be shown that

$$\text{E}(H_{i\Theta_{2s},\Theta_1'}(\Theta_1,\Theta_2)|\boldsymbol{X},\mathcal{Z})=-\sum_{t=1}^{T}\frac{\phi(.)^2}{\Phi(.)(1-\Phi(.))\exp(W_{it}'\Theta_{2h})}\mathbb{X}_{it}(\Theta_{2s}'\frac{\partial\mathbb{X}_{it}}{\partial\Theta_1'}-\mathbb{X}_{it}'\Theta_{2s}\Theta_{2h}'\frac{\partial W_{it}}{\partial\Theta_1'})$$

and

$$\text{E}(H_{i\Theta_{2h},\Theta_1'}(\Theta_1,\Theta_2)|\boldsymbol{X},\mathcal{Z})=\sum_{t=1}^{T}\frac{\phi(.)^2}{\Phi(.)(1-\Phi(.))}\frac{\mathbb{X}_{it}'\Theta_{2s}}{\exp(W_{it}'\Theta_{2h})}W_{it}(\Theta_{2s}'\frac{\partial\mathbb{X}_{it}}{\partial\Theta_1'}-\mathbb{X}_{it}'\Theta_{2s}\Theta_{2h}'\frac{\partial W_{it}}{\partial\Theta_1'}).$$

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

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

Since $\mathcal{X}_{it}$ above is not a function of Θ_1, $\frac{\partial \mathcal{X}_{it}}{\partial \Theta_1'} = \mathbf{0}_{\mathcal{X}}$, where $\mathbf{0}_{\mathcal{X}}$ is a null matrix with row dimension that of column vector $\mathcal{X}_{it}$ and column dimension that of column vector Θ_1. In section C of this supplementary appendix we derive the derivative of $\hat{\boldsymbol{\alpha}}_i(\Theta_1)$ and $\hat{\boldsymbol{\epsilon}}_{it}(\Theta_1)$ with respect to $\Theta_1 = \{\boldsymbol{\delta}', \text{vec}(\Lambda_{\alpha\alpha})', \text{vec}(\Sigma_{\epsilon\epsilon})'\}'$. We show that

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{\alpha}}_i}{\partial \boldsymbol{\delta}'} &= \mathbb{O}_{\mathbf{Z}i}' - \frac{1}{U_{dr}^2} \sum_{t=1}^{T} \Big[U_{nr} U_{nr}' - U_{dr} F_{dr} \Big] \Sigma_{\epsilon\epsilon}^{-1} \mathbf{Z}_{it}', \\
\frac{\partial \hat{\boldsymbol{\epsilon}}_{it}}{\partial \boldsymbol{\delta}'} &= -\mathbf{Z}_{it}' + \frac{1}{U_{dr}^2} \sum_{t=1}^{T} \Big[U_{nr} U_{nr}' - U_{dr} F_{dr} \Big] \Sigma_{\epsilon\epsilon}^{-1} \mathbf{Z}_{it}', \\
\frac{\partial \hat{\boldsymbol{\alpha}}_i}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} &= \frac{1}{2U_{dr}^2} [U_{dr} F_{nr} - U_{nr} \text{vec}(F_{dr})'] (\Lambda_{\alpha\alpha}^{-1} \otimes \Lambda_{\alpha\alpha}^{-1}), \\
\frac{\partial \hat{\boldsymbol{\alpha}}_i}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'} &= \frac{1}{2U_{dr}^2} \sum_{t=1}^{T} \Big[U_{dr} (-\mathbf{r}_{it}' \otimes F_{dr} - F_{dr} \otimes \mathbf{r}_{it}' + F_{nr}) \\
&\quad - U_{nr} \text{vec}(-U_{nr} \mathbf{r}_{it}' - \mathbf{r}_{it} U_{nr}' + F_{dr})' \Big] (\Sigma_{\epsilon\epsilon}^{-1} \otimes \Sigma_{\epsilon\epsilon}^{-1}),
\end{aligned}$$

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

where

$$\begin{aligned}
&U_{nr} = \textstyle\int I_m \tilde{\boldsymbol{\alpha}} \exp(-\frac{1}{2} r(\Theta_1, \tilde{\boldsymbol{\alpha}})) \phi(\tilde{\boldsymbol{\alpha}}) d\tilde{\boldsymbol{\alpha}}, \quad F_{nr} = \textstyle\int I_m \tilde{\boldsymbol{\alpha}} \text{vec}(\tilde{\boldsymbol{\alpha}} \tilde{\boldsymbol{\alpha}}')' \exp(-\frac{1}{2} r(\Theta_1, \tilde{\boldsymbol{\alpha}})) \phi(\tilde{\boldsymbol{\alpha}}) d\tilde{\boldsymbol{\alpha}}, \\
&U_{dr} = \textstyle\int \exp(-\frac{1}{2} r(\Theta_1, \tilde{\boldsymbol{\alpha}})) \phi(\tilde{\boldsymbol{\alpha}}) d\tilde{\boldsymbol{\alpha}}, \quad F_{dr} = \textstyle\int \tilde{\boldsymbol{\alpha}} \tilde{\boldsymbol{\alpha}}' \exp(-\frac{1}{2} r(\Theta_1, \tilde{\boldsymbol{\alpha}})) \phi(\tilde{\boldsymbol{\alpha}}) d\tilde{\boldsymbol{\alpha}}, \\
&\mathbf{r}_{it} = \boldsymbol{x}_{it} - \mathbf{Z}_i \boldsymbol{\delta}, \text{ and } r(\Theta_1, \tilde{\boldsymbol{\alpha}}) = \textstyle\sum_{t=1}^{T} (\boldsymbol{x}_t - \mathbf{Z}_t' \boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}})' \Sigma_{\epsilon\epsilon}^{-1} (\boldsymbol{x}_t - \mathbf{Z}_t' \boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}}).
\end{aligned}$$

$\mathbb{O}_{\mathbf{Z}i} = \text{diag}((0_z', \bar{\boldsymbol{z}}_i')', \ldots, (0_z', \bar{\boldsymbol{z}}_i')')$, where $0_z'$ is a vector of zeros of having the dimension of $\boldsymbol{z}_{it}$. Numerical integration technique, discussed in appendix D of this supplementary appendix, can be used to compute $\hat{U}_{nr}$, $\hat{U}_{dr}$, $\hat{F}_{nr}$, and $\hat{F}_{dr}$ at the estimated value $\hat{\Theta}_1$ to obtain the error adjusted standard errors of the structural estimates.

In Lemma 1 in the main text we showed that $\hat{U}_{nr}(\hat{\Theta}_1)$ and $\hat{U}_{dr}(\hat{\Theta}_1)$ converge almost surely to $U_{nr}(\Theta_1^*)$, $U_{dr}(\Theta_1^*)$. By application of Lemma 1 it can be also shown that $\hat{F}_{nr}(\hat{\Theta}_1)$, and $\hat{F}_{dr}(\hat{\Theta}_1)$ converge almost surely to $F_{nr}(\Theta_1^*)$, and $F_{dr}(\Theta_1^*)$ respectively. This would imply that $H_{i\Theta_2\Theta_1}(\hat{\Theta}_1, \hat{\Theta}_2)$ converge almost surly to $H_{i\Theta_2\Theta_1}(\Theta_1^*, \Theta_2^*)$, and 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 $\text{E}(H_{i\Theta_2\Theta_1}(\Theta_1^*, \Theta_2^*)) = \mathbb{H}_{\Theta_2\Theta_1}$.

Finally, we would like to state that though we have provided analytical expression for the covariance matrix, V_2^*, and the estimated covariance matrix for the specifications in the application (Section 3) of the proposed model is based on the analytical expression for V_2^*, we do not, however, recommend to follow this approach in practice. Since the expressions for $\mathbb{L}_{\Theta_1\Theta_1}$ and $\mathbb{H}_{\Theta_2\Theta_1}$ are cumbersome to compute, we suggest that bootstrapping procedure be employed to approximate the variance of the estimated coefficient. Moreover, these expressions are likely to be different when a different estimator for the first stage reduced form is required.

APPENDIX C: Derivative of the Control Functions with respect to Θ_1

First consider the derivative of $\hat{\boldsymbol{\alpha}}_i = \text{diag}(\bar{\mathbf{z}}_i, \ldots, \bar{\mathbf{z}}_i)'\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i = \bar{\boldsymbol{Z}}_i'\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i$ with respect to $\text{vec}(\Lambda_{\alpha\alpha})$. We have

$$\begin{aligned}
&\frac{\partial(\bar{\boldsymbol{Z}}_i'\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i)}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} = \frac{\partial \hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} \\
&= \frac{\partial}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}\left[\frac{\int \tilde{\boldsymbol{\alpha}} \exp(-\frac{1}{2}r(\Theta_1, \tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int \exp(-\frac{1}{2}r(\Theta_1, \tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}\right] = \frac{\partial}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}\left[\frac{\int f_{nr}(., \tilde{\boldsymbol{\alpha}})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int f_{dr}(., \tilde{\boldsymbol{\alpha}})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}\right] \\
&= \frac{[\int f_{nr}(., \tilde{\boldsymbol{\alpha}})\frac{\partial\phi(\tilde{\boldsymbol{\alpha}})}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}d\tilde{\boldsymbol{\alpha}}][\int f_{dr}(., \tilde{\boldsymbol{\alpha}})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}] - [\int f_{nr}(., \tilde{\boldsymbol{\alpha}})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}][\int f_{dr}(., \tilde{\boldsymbol{\alpha}})\frac{\partial\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}]}{[\int f_{dr}(., \tilde{\boldsymbol{\alpha}})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}]^2},
\end{aligned} \tag{C.1}$$

where $r(\Theta_1, \tilde{\boldsymbol{\alpha}}) = \sum_{t=1}^{T}(\boldsymbol{x}_t - \mathbf{Z}_t'\boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{x}_t - \mathbf{Z}_t'\boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}})$.

Since $\phi(\tilde{\boldsymbol{\alpha}}) = \frac{1}{(2\pi)^{m/2}|\Lambda_{\alpha\alpha}|^{1/2}}\exp(-\frac{1}{2}\tilde{\boldsymbol{\alpha}}'\Lambda_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}})$ we have

$$\begin{aligned}
\frac{\partial\phi(\tilde{\boldsymbol{\alpha}})}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} &= \frac{-\exp(-\frac{1}{2}\tilde{\boldsymbol{\alpha}}'\Lambda_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}})}{2(2\pi)^{m/2}|\Lambda_{\alpha\alpha}|^{3/2}}\frac{\partial|\Lambda_{\alpha\alpha}|}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} + \frac{\exp(-\frac{1}{2}\tilde{\boldsymbol{\alpha}}'\Lambda_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}})}{(2\pi)^{m/2}|\Lambda_{\alpha\alpha}|^{1/2}}\frac{\partial(-\frac{1}{2}\tilde{\boldsymbol{\alpha}}'\Lambda_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}})}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} \\
&= -\frac{1}{2}\phi(\tilde{\boldsymbol{\alpha}})\left(\frac{1}{|\Lambda_{\alpha\alpha}|}\frac{\partial|\Lambda_{\alpha\alpha}|}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} + \frac{\partial(\tilde{\boldsymbol{\alpha}}'\Lambda_{\alpha\alpha}^{-1}\tilde{\boldsymbol{\alpha}})}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}\right) \\
&= -\frac{1}{2}\phi(\tilde{\boldsymbol{\alpha}})\left(\text{vec}(\Lambda_{\alpha\alpha}^{-1})' + \text{vec}(-(\Lambda_{\alpha\alpha}^{-1})'\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'(\Lambda_{\alpha\alpha}^{-1})')'\right)\frac{\partial \text{vec}(d(\Lambda_{\alpha\alpha}))}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} \\
&= -\frac{1}{2}\phi(\tilde{\boldsymbol{\alpha}})\left(\text{vec}(\Lambda_{\alpha\alpha}^{-1})' + \text{vec}(-(\Lambda_{\alpha\alpha}^{-1})'\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'(\Lambda_{\alpha\alpha}^{-1})')'\right).
\end{aligned} \tag{C.2}$$

Given (C.2), (C.1) can be simplified as

$$\frac{\partial(\bar{\boldsymbol{Z}}_i'\boldsymbol{\varrho}+\hat{\tilde{\boldsymbol{\alpha}}}_i)}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} = -\frac{1}{2U_{dr}^2}\Big[[U_{nr}\text{vec}(\Lambda_{\alpha\alpha}^{-1})' - F_{nr}(\Lambda_{\alpha\alpha}^{-1}\otimes\Lambda_{\alpha\alpha}^{-1})']U_{dr} \\ - U_{nr}[U_{dr}\text{vec}(\Lambda_{\alpha\alpha}^{-1})' - F_{dr}(\Lambda_{\alpha\alpha}^{-1}\otimes\Lambda_{\alpha\alpha}^{-1})']\Big] \\ = \frac{1}{2U_{dr}^2}[U_{dr}F_{nr} - U_{nr}\text{vec}(F_{dr})'](\Lambda_{\alpha\alpha}^{-1}\otimes\Lambda_{\alpha\alpha}^{-1})', \tag{C.3}$$

where

$$U_{nr} = \int I_m\tilde{\boldsymbol{\alpha}}\exp(-\frac{1}{2}r(\Theta_1,\tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}, \quad F_{nr} = \int I_m\tilde{\boldsymbol{\alpha}}\text{vec}(\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}')'\exp(-\frac{1}{2}r(\Theta_1,\tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}$$
$$U_{dr} = \int \exp(-\frac{1}{2}r(\Theta_1,\tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}, \quad F_{dr} = \int \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'\exp(-\frac{1}{2}r(\Theta_1,\tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}. \tag{C.4}$$

Also, from (C.3) we can conclude that

$$\frac{\partial\hat{\boldsymbol{\epsilon}}_{it}}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} = \frac{-\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Lambda_{\alpha\alpha})'} = \frac{-1}{2U_{dr}^2}[U_{dr}F_{nr} - U_{nr}\text{vec}(F_{dr})'](\Lambda_{\alpha\alpha}^{-1}\otimes\Lambda_{\alpha\alpha}^{-1})'. \tag{C.5}$$

Now consider the derivative of $\hat{\boldsymbol{\alpha}}_i = \bar{\boldsymbol{Z}}_i'\boldsymbol{\varrho}+\hat{\tilde{\boldsymbol{\alpha}}}_i$ with respect to $\text{vec}(\Sigma_{\epsilon\epsilon})$. We have

$$\frac{\partial(\bar{\boldsymbol{Z}}_i'\boldsymbol{\varrho}+\hat{\tilde{\boldsymbol{\alpha}}}_i)}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{\partial}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'}\left[\frac{\int\tilde{\boldsymbol{\alpha}}\exp(-\frac{1}{2}\sum_{t=1}^T\boldsymbol{\epsilon}_{it}'\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int\exp(-\frac{1}{2}\sum_{t=1}^T\boldsymbol{\epsilon}_{it}'\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}\right]$$
$$= -\frac{1}{2}\left[\frac{\int\tilde{\boldsymbol{\alpha}}\psi(\tilde{\boldsymbol{\alpha}})\frac{\partial\sum_{t=1}^T\boldsymbol{\epsilon}_{it}'\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it}}{\partial\text{vec}(\Sigma_{\epsilon\epsilon})'}d\tilde{\boldsymbol{\alpha}}\int\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}} - \int\tilde{\boldsymbol{\alpha}}\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}\int\psi(\tilde{\boldsymbol{\alpha}})\frac{\partial\sum_{t=1}^T\boldsymbol{\epsilon}_{it}'\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it}}{\partial\text{vec}(\Sigma_{\epsilon\epsilon})'}d\tilde{\boldsymbol{\alpha}}}{(\int\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}})^2}\right],$$

where $\psi(\tilde{\boldsymbol{\alpha}}) = \exp(-\frac{1}{2}r(\Theta_1,\tilde{\boldsymbol{\alpha}})\phi(\tilde{\boldsymbol{\alpha}})$. With $\frac{\partial\sum_{t=1}^T\boldsymbol{\epsilon}_{it}'\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it}}{\partial\text{vec}(\Sigma_{\epsilon\epsilon})'} = \sum_{t=1}^T\text{vec}(-(\Sigma_{\epsilon\epsilon}^{-1})'\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}'(\Sigma_{\epsilon\epsilon}^{-1})')'$ the above can be written as

$$\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{1}{2(\int\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}})^2}\sum_{t=1}^T\Big[\int\tilde{\boldsymbol{\alpha}}\psi(\tilde{\boldsymbol{\alpha}})\text{vec}((\Sigma_{\epsilon\epsilon}^{-1})'\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}'(\Sigma_{\epsilon\epsilon}^{-1})')'d\tilde{\boldsymbol{\alpha}}\int\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}} \\ - \int\psi(\tilde{\boldsymbol{\alpha}})\text{vec}((\Sigma_{\epsilon\epsilon}^{-1})'\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}'(\Sigma_{\epsilon\epsilon}^{-1})')'d\tilde{\boldsymbol{\alpha}}\int\tilde{\boldsymbol{\alpha}}\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}\Big]$$
$$= \frac{1}{2U_{dr}^2}\sum_{t=1}^T\Big[\int\tilde{\boldsymbol{\alpha}}\psi(\tilde{\boldsymbol{\alpha}})\text{vec}(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}')'(\Sigma_{\epsilon\epsilon}^{-1}\otimes\Sigma_{\epsilon\epsilon}^{-1})'d\tilde{\boldsymbol{\alpha}}U_{dr} \\ - U_{nr}\int\psi(\tilde{\boldsymbol{\alpha}})\text{vec}(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}')'(\Sigma_{\epsilon\epsilon}^{-1}\otimes\Sigma_{\epsilon\epsilon}^{-1})'d\tilde{\boldsymbol{\alpha}}\Big]$$
$$= \frac{1}{2U_{dr}^2}\sum_{t=1}^T\Big[\int(U_{dr}\tilde{\boldsymbol{\alpha}}\text{vec}(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}')' - U_{nr}\text{vec}(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}')')\psi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}\Big](\Sigma_{\epsilon\epsilon}^{-1}\otimes\Sigma_{\epsilon\epsilon}^{-1})' \tag{C.6}$$

To simply (C.6) further, write $\boldsymbol{\epsilon}_{it}$ as $\boldsymbol{\epsilon}_{it} = \boldsymbol{x}_{it} - \mathbf{Z}'_{it}\boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}} = \mathbf{r}_{it} - \tilde{\boldsymbol{\alpha}}$, where $\mathbf{r}_{it} = \boldsymbol{x}_{it} - \mathbf{Z}'_{it}\boldsymbol{\delta}$. Then $\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}'_{it} = \mathbf{r}_{it}\mathbf{r}'_{it} - \tilde{\boldsymbol{\alpha}}\mathbf{r}'_{it} - \mathbf{r}_{it}\tilde{\boldsymbol{\alpha}}' + \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'$, and (C.6) after some simplification can be written as

$$\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'} = \frac{1}{2U_{dr}^2}\sum_{t=1}^{T}\Big[U_{dr}(-\mathbf{r}'_{it}\otimes F_{dr} - F_{dr}\otimes\mathbf{r}'_{it} + F_{nr}) \\ - U_{nr}\text{vec}(-U_{nr}\mathbf{r}'_{it} - \mathbf{r}_{it}U'_{nr} + F_{dr})'\Big](\Sigma_{\epsilon\epsilon}^{-1}\otimes\Sigma_{\epsilon\epsilon}^{-1})', \tag{C.7}$$

where U_{nr}, U_{dr}, F_{nr}, and F_{dr} have been defined in (C.4).

Also, since $\frac{\partial\hat{\tilde{\boldsymbol{\epsilon}}}_{it}}{\partial \text{vec}(\tilde{\Sigma}_{\epsilon\epsilon})'} = \frac{-\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\tilde{\Sigma}_{\epsilon\epsilon})'}$, the derivative of $\hat{\boldsymbol{\epsilon}}_i(t)$ with respect to $\text{vec}(\tilde{\Sigma}_{\epsilon\epsilon})$ can be obtained from (C.7). We note here that $\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Lambda_{\alpha\alpha})'}$ and $\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial \text{vec}(\Sigma_{\epsilon\epsilon})'}$, for an individual i, are constant for all time periods.

Finally, let us now consider the derivative of $\bar{\boldsymbol{Z}}'_i\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i$ and $\hat{\tilde{\boldsymbol{\epsilon}}}_{it}$ with respect to $\boldsymbol{\delta}'$. We have

$$\frac{\partial(\bar{\boldsymbol{Z}}'_i\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i)}{\partial\boldsymbol{\delta}'} = \frac{\partial\bar{\boldsymbol{Z}}'_i\boldsymbol{\varrho}}{\partial\boldsymbol{\delta}'} + \frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial\boldsymbol{\delta}'} = \mathbb{O}'_{\mathbf{Z}i} + \frac{\partial}{\partial\boldsymbol{\delta}'}\left[\frac{\int\tilde{\boldsymbol{\alpha}}\exp(-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}{\int\exp(-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it})\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}}\right]$$

$$= \mathbb{O}'_{\mathbf{Z}i} + \frac{1}{(\int\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}})^2}\sum_{t=1}^{T}\Big[\int\tilde{\boldsymbol{\alpha}}\exp(.)\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\mathbf{Z}'_{it}\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}\int\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}$$

$$- \int\tilde{\boldsymbol{\alpha}}\exp(.)\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}\int\exp(.)\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\mathbf{Z}'_{it}\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}}\Big], \tag{C.8}$$

where $\mathbb{O}_{\mathbf{Z}i} = \text{diag}((0'_z, \bar{z}'_i)', \ldots, (0'_z, \bar{z}'_i)')$, and $0'_z$ is a vector of zeros of having the dimension of z_{it}, which has been defined in Section 2 in the main text. To derive the result in (C.8) we used the fact that

$$\frac{\partial(\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\boldsymbol{\epsilon}_{it})}{\partial\boldsymbol{\delta}'} = 2\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\frac{\partial(\boldsymbol{\epsilon}_{it})}{\partial\boldsymbol{\delta}'} = -2\boldsymbol{\epsilon}'_{it}\Sigma_{\epsilon\epsilon}^{-1}\mathbf{Z}'_{it}.$$

With some of the results stated above it can be shown that

$$\frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial\boldsymbol{\delta}'} = \frac{1}{U_{dr}^2}\sum_{t=1}^{T}\Big[U_{nr}U'_{nr} - U_{dr}F_{dr}\Big]\Sigma_{\epsilon\epsilon}^{-1}\mathbf{Z}'_{it}.$$

Hence, we have

$$\frac{\partial(\bar{\boldsymbol{Z}}'_i\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i)}{\partial\boldsymbol{\delta}'} = \mathbb{O}'_{\mathbf{Z}i} - \frac{1}{U_{dr}^2}\sum_{t=1}^{T}\Big[U_{nr}U'_{nr} - U_{dr}F_{dr}\Big]\Sigma_{\epsilon\epsilon}^{-1}\mathbf{Z}'_{it}, \tag{C.9}$$

and

$$\frac{\partial\hat{\tilde{\boldsymbol{\epsilon}}}_{it}}{\partial\boldsymbol{\delta}'} = \frac{\partial(\boldsymbol{x}_{it} - \mathbf{Z}'_{it}\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'} - \frac{\partial\hat{\tilde{\boldsymbol{\alpha}}}_i}{\partial\boldsymbol{\delta}'} = -\mathbf{Z}'_{it} + \frac{1}{U_{dr}^2}\sum_{t=1}^{T}\Big[U_{nr}U'_{nr} - U_{dr}F_{dr}\Big]\Sigma_{\epsilon\epsilon}^{-1}\mathbf{Z}'_{it}. \tag{C.10}$$

From (C.9) and (C.10) we can see that while $\frac{\partial(\bar{\boldsymbol{Z}}'_i\boldsymbol{\varrho} + \hat{\tilde{\boldsymbol{\alpha}}}_i)}{\partial\boldsymbol{\delta}'}$ for an individual i remains the same for all time periods, $\frac{\partial\hat{\tilde{\boldsymbol{\epsilon}}}_{it}}{\partial\boldsymbol{\delta}'}$ varies with time.

C.1 Hypothesis Testing of Average Partial Effects

In section 2 we discussed the identification and estimation of the average partial effect (APE) of a variable w for a binary choice model. In order to draw inferences about the APE's we need to compute the standard errors of the estimated APE's. From equation (2.15) in the main text we know that estimated APE of w on the probability of $y_{it} = 1$, given $\mathcal{X}_{it} = \bar{\mathcal{X}}$, is given by

$$\begin{aligned}\frac{\partial \widehat{\Pr}(y_{it}=1|\bar{\mathcal{X}})}{\partial w} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\hat{\varphi}_w - \hat{h}'_w(.)(\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s})}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}\phi\left(\frac{\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s}}{\exp(h(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\epsilon}}_{it}))}\right) \\ &\equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} g_{wit}(\hat{\Theta}_2), \end{aligned} \tag{C.11}$$

where $\bar{\mathbb{X}}_{it} = \{\bar{\mathcal{X}}', \hat{\boldsymbol{\alpha}}'_i, \hat{\boldsymbol{\epsilon}}'_{it}\}'$ and $\Theta_{2s} = \{\boldsymbol{\varphi}', \Sigma'_\alpha, \Sigma'_\epsilon\}'$. Now, we know that by the linear approximation approach (delta method), the asymptotic variance of $\frac{\partial \widehat{\Pr}(y_{it}=1|\bar{\mathcal{X}})}{\partial w}$ can be estimated by computing

$$\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}'_2}\right]\hat{V}_2^*\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}'_2}\right]', \tag{C.12}$$

where $\Theta_2 = \{\Theta'_{2s}, \Theta'_{2h}\}'$, Θ_{2h} being the coefficients of the heteroscedastic specification, and $\hat{V}_2^*$ is the second stage error adjusted covariance matrix of Θ_2 estimated at $\hat{\Theta}_2$. $\frac{\partial g_{wit}(\hat{\Theta}_2)}{\partial \hat{\Theta}'_2}$ in (C.12) turns out to be

$$\begin{bmatrix} \frac{\phi(.)}{\exp(.)}[e_{ws} - \hat{h}'_w(.)\bar{\mathbb{X}}_{it} - \frac{1}{\exp^2(.)}(\hat{\varphi}_w - \hat{h}'_w(.)\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s})(\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s})\bar{\mathbb{X}}_{it}] \\ \frac{-\phi(.)}{\exp(.)}[e_{wh}\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s} + (1 - \frac{(\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s})^2}{\exp^3(.)})(\hat{\varphi}_w - \hat{h}'_w(.)\bar{\mathbb{X}}'_{it}\hat{\Theta}_{2s})\frac{\partial h(.)}{\partial \hat{\Theta}'_{2h}}] \end{bmatrix}', \tag{C.13}$$

where e_{ws} is a column vector having the dimension of Θ'_{2s} and with 1 at the position of φ_w in Θ_{2s} and zeros elsewhere and e_{wh} is a column vector having the dimension of Θ'_{2h} and with 1 at the position of $h'_w(.) = \gamma_w$ in Θ_{2h} and zeros elsewhere.

By application of delta method the variance of APE of w when w is a dummy variable can also be easily obtained. For lack of space we do not detail its estimation here.

APPENDIX D: Note on Numerical Integration

In order to obtain the structural estimates we have to compute the expected a Posteriori values of the time invariant individual effects given by:

$$\begin{aligned}\hat{\tilde{\boldsymbol{\alpha}}}(\boldsymbol{X}, \mathcal{Z}, \Theta_1) &= \frac{\int C\mathfrak{a}\exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{x}_t - \mathbf{Z}'_t\boldsymbol{\delta} - \tilde{\boldsymbol{\alpha}})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{x}_t - \mathbf{Z}'_t\boldsymbol{\delta} - C\mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}}{\int \exp(-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{x}_t - \mathbf{Z}'_t\boldsymbol{\delta} - C\mathfrak{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{x}_t - \mathbf{Z}'_t\boldsymbol{\delta} - C\mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}} \\ &= \frac{\int C\exp(-\frac{1}{2}r(\Theta_1,\mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}}{\int \exp(-\frac{1}{2}r(\Theta_1,\mathfrak{a}))\phi(\mathfrak{a})d\mathfrak{a}} = \frac{U_{nr}}{U_{dr}}, \end{aligned} \tag{D.1}$$

where $\tilde{\boldsymbol{\alpha}} = C\mathfrak{a}$, CC' being the Cholesky decomposition of the $(m \times m)$ covariance matrix $\Lambda_{\alpha\alpha}$, so that $d\tilde{\boldsymbol{\alpha}} = |C|d\mathfrak{a} = |\Lambda_{\alpha\alpha}|^{1/2}d\mathfrak{a}$, and $r(\Theta_1, \mathfrak{a}) = \sum_{t=1}^{T}(\boldsymbol{x}_t - \mathbf{Z}_t'\boldsymbol{\delta} - C\mathfrak{a})'\Sigma_{\epsilon\epsilon}^{-1}(\boldsymbol{x}_t - \mathbf{Z}_t'\boldsymbol{\delta} - C\mathfrak{a})$. And to obtain error adjusted covariance matrix in addition to U_{nr} and U_{dr} we have to estimate F_{nr} and F_{dr} given by

$$F_{nr} = \int I_m\tilde{\boldsymbol{\alpha}}\text{vec}(\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}')'\exp(-\frac{1}{2}r(\Theta_1, \tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}} \text{ and } F_{dr} = \int \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'\exp(-\frac{1}{2}r(\Theta_1, \tilde{\boldsymbol{\alpha}}))\phi(\tilde{\boldsymbol{\alpha}})d\tilde{\boldsymbol{\alpha}} \tag{D.2}$$

respectively.

Here we discuss how to compute U_{nr}, U_{dr}, F_{nr}, and F_{dr}. Take, for example, U_{nr}, which can be written as

$$\begin{aligned}\int C\mathfrak{a}\exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))\det(C)\phi(\mathfrak{a})d\mathfrak{a} &= \int C\mathfrak{a}\exp(-\frac{1}{2}r(\Theta_1, \mathfrak{a}))\det(C)\frac{1}{(2\pi)^{m/2}}e^{-\frac{\mathfrak{a}'\mathfrak{a}}{2}}d\mathfrak{a} \\ &= \int f(\mathfrak{a})\frac{1}{(2\pi)^{m/2}}e^{-\frac{\mathfrak{a}'\mathfrak{a}}{2}}d\mathfrak{a},\end{aligned}$$

where $\int f(\mathfrak{a})\frac{1}{(2\pi)^{m/2}}e^{-\frac{\mathfrak{a}'\mathfrak{a}}{2}}d\mathfrak{a} = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(\mathfrak{a})\frac{1}{(2\pi)^{m/2}}e^{-\frac{\mathfrak{a}'\mathfrak{a}}{2}}d\mathfrak{a}_1\ldots d\mathfrak{a}_m$.

A general treatment for numerically computing multidimensional integrals can be found in Krommer and Ueberhuber (1994). More recently Cools and Haegemans (1994) have developed integration rules for multidimensional integrals over infinite integration regions with a Gaussian weight function to evaluate integrals of the type stated above, and Genz and Keister (1996) have provided more efficient rules of the same. The integration rules consist of constructing $\mathfrak{N}$ weights, w_j, and points $\mathfrak{a}_j$, $\mathfrak{a}_j \in \mathbb{R}^m$, such that

$$Q(f) = \sum_{j=1}^{\mathfrak{N}} w_j f(\mathfrak{a}_j), \tag{D.3}$$

where $Q(f)$ approximates the integral $\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(\mathfrak{a})\frac{1}{(2\pi)^{m/2}}e^{-\frac{\mathfrak{a}'\mathfrak{a}}{2}}d\mathfrak{a}_1\ldots d\mathfrak{a}_m$. Fortran routines for computing $Q(f)$, developed in Genz and Keister (1996), can be obtained from Alan Genz's webpage. Heiss and Winschel (2008) develop multi-dimension integration rules on sparse grids which has the advantage over product rule extension of univariate quadrature in that it does not impose exponentially increasing computational costs with rising number of dimensions. STATA and Matlab codes for generating quadratures and weights on sparse grid for integration rule developed in Genz and Keister (1996) can be obtained from Frorian Heiss and Viktor Winschels web page, http://www.sparse-grids.de/.

References

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

Cools, R. and Haegemans, A. (1994). An Imbedded Family of Cubature Formulae for n-Dimensional Product Regions. *Journal of Computational and Applied Mathematics*, **51**, 251–260.

Genz, A. and Keister, B. (1996). Fully Symmetric Interpolatory Rules for Multiple Integrals over Infinite Regions with Gaussian Weight. *Journal of Computation and Applied Mathematics*, **71**, 299–309.

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

Krommer, A. R. and Ueberhuber, C. W. (1994). *Numerical Integration: on Advanced Computer Systems (Lecture Notes in Computer Science)*. Springer-Verlag, 1st edn.

Lutkepohl, H. (1996). *Handbook of Matrices*. Chichester: Wiley.

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.