A control function approach to estimate panel data binary response model

Amaresh K. Tiwari

School of Economics and Business Administration, University of Tartu, Tartu, Estonia

ABSTRACT

We propose a new control function (CF) method to estimate a binary response model in a triangular system with multiple unobserved heterogeneities. The CFs are the expected values of the heterogeneity terms in the reduced form equations conditional on the histories of the endogenous and the exogenous variables. The method requires weaker restrictions compared to CF methods with similar imposed structures. If the support of endogenous regressors is large, average partial effects are point-identified even when instruments are discrete. Bounds are provided when the support assumption is violated. An application and Monte Carlo experiments compare several alternative methods with ours.

KEYWORDS

Average partial effects; child labor; point & partial identification; triangular system; unobserved heterogeneities

JEL CLASSIFICATION

C13; C18; C33; J13

1. Introduction

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

Arellano and Bonhomme (2011) provide an overview, and categorize, of some of the methods developed to estimate the quantities of interest. These include the fixed effect (FE) approach that treat heterogeneity or individual effects as parameters to be estimated, where several approaches have been proposed to correct for bias due the incidental parameter problem. Wooldridge (2019), points out that the FE approach, although 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. Second, the recent bias adjustments methods require the assumptions of stationarity and weak dependence; in some cases, the strong assumption of serial independence (conditional on the heterogeneity) is maintained. However, in empirical work dealing with linear models, it has been found that idiosyncratic errors exhibit serial dependence. Also, “the requirement of stationarity is strong and has substantive restrictions as it rules out staples in empirical work such as including separate year effects, which can be estimated very precisely given a large cross section.”

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

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

Some of the papers that have developed control function method to study binary or fractional response outcomes are Rivers and Vuong (1988), Blundell and Powell (2004) (BP), Papke and Wooldridge (2008) (PW), Rothe (2009) and Semykina and Wooldridge (2018) (SW). A partial list of papers that have studied nonparametric control function estimation of nonseparable, including binary response, models are Altonji and Matzkin (2005) (AM), Florens et al. (2008), Imbens and Newey (2009) (IN) and HW, where the focus is on estimating heterogeneous effect of endogenous treatment.

To our knowledge, the papers that allow certain unobserved heterogeneities to be correlated with the exogenous variables while developing a control function method are: PW, Fernández-Val and Vella (2011) (FV), HW, Kim and Petrin (2017) and SW. While PW and SW use the framework of correlated random effects to account for correlation between individual specific effects and the exogenous variables, FV consider fixed effect estimation in both the stages, where the control variable is based on estimates of the fixed effects in the reduced form equation. FV’s method although more general, requires large $T$ to correct the bias due to incidental parameters problem. HW develop a generalized version of differencing – which differences out the fixed effects – to identify local average responses in a nonseparable and binary response models. However, identification in HW requires that individuals do not experience a change in covariates over time; this requirement that the support of all regressors overlap over time could be hard to satisfy – e.g., it rules out time trends and time dummies. Kim and Petrin (2017) exploit restrictions in the conditional moment of unobserved heterogeneity given instruments to develop “generalized control function.” We allow for the correlation between the unobserved individual effects and the instruments in a manner similar to PW’s and SW’s.

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

The exceptions to our knowledge are PW, FV, HW and SW, who consider panel data where multiple heterogeneities constitute of time invariant random effects and idiosyncratic errors. While the imposed structures in PW and SW are similar to ours, they make the traditional control function assumption, and so their control function is scalar, whereas our control function is vector valued, whose dimension depend on the dimension of unobserved heterogeneity and the number of endogenous variables. HW’s specification of the triangular system does not nest ours and FV’s fixed effects method requires long panels.

We propose that the expected values of the heterogeneity terms of the reduced form equations conditional on the history of endogenous variables, $X_i \equiv (x_{i1}, ..., x_{iT})$, and the same of the exogenous variables, $Z_i \equiv (z_{i1}, ..., z_{iT})$ be used as control functions. The proposed control
functions are identified when the distributions of the heterogeneity terms are specified. We argue that (1) for triangular systems with setups similar to ours, these control functions imply a weaker restriction than the commonly made control function assumptions, and (2) the traditionally used control functions may not provide consistent estimates in a panel data setting such as ours.

Our method, while being simple, makes a number of contributions to the literature. First, we allow for multiple heterogeneities, albeit with restrictions, in the triangular system, where most papers, adopting the control function approach to handle endogeneity, do not. Second, when the support of the endogenous variables is large, ASF or the APEs are point-identified even when the instruments have a small support. We exploit panel data with repeated observations of the same unit for the purpose of point-identification when support requirement is met. Sharp bounds on the ASF and the APEs are provided when the support assumption is not satisfied. Third, the method accounts for multiple endogenous variables, all of which are determined simultaneously, whereas most papers on control function consider a single endogenous variable. Finally, our model retains the attractive features of PW’s, where no assumptions are made on the serial dependence among the outcome variable.

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

The rest of the article is organized as follows. In section 2, we introduce the model and discuss identification and estimation of structural measures of interests for a discrete response model in a triangular system with random effects. In section 3, we discuss the results of the Monte Carlo experiments, where we compare our estimator with some of the existing methods for panel data binary response model with imposed structures similar to ours. Section 4 contains the application of the proposed estimator to study income and wealth effects on the incidence of child labor. And finally in section 5 we conclude. Proofs of the theorems and lemmas are provided in Appendix A.

Due to space constraint, the following have been put in the Supporting Information Appendix: generalized estimating equation (GEE) estimation of probit conditional mean function (Supporting Information Appendix 1), extension of the random effect model in the main text to allow for random coefficients (Supporting Information Appendix 2), large sample properties of the estimator (Supporting Information Appendix 3), other technical details (Supporting Information Appendix 4).

2. Model specification and identification and estimation of structural measures

Consider the following binary choice model in a triangular setup:

\[ y_{it} = \mathbb{1}\{y_{it}^* = (w_{it}', x_{it}')\phi + \theta_i + \xi_{it} > 0\}, \]  

(2.1)

where \( \mathbb{1}\{.\} \) is an indicator function that takes value 1 if the argument in the parenthesis holds true and 0 otherwise. In (2.1), \( \theta_i \) is the unobserved time invariant individual effect and \( \xi_{it} \) is the idiosyncratic error component. The variables, \( x_{it} \), are endogenous in the sense that \( \xi_{it} \perp x_{it} | \theta_i \); whereas most papers studying panel data binary choice model assume that \( \xi_{it} \perp x_{it} | \theta_i \). We assume that each of the endogenous variables are continuous and have a large support. The dimension of \( x_{it} \) is \( d_x \) and the dimension of the exogenous variables, \( w_{it} \), is \( d_w \).

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

\[ x_{it} = \pi z_{it} + \alpha_i + \epsilon_{it}. \]  

(2.2)
In (2.2), \( \pi \) has a row dimension of \( d_x \), \( \pi_i \equiv (x_{1i}, ..., x_{id_x})' \) is the \((d_x \times 1)\) vector of unobserved time invariant individual effects, \( e_{it} \equiv (e_{i1}, ..., e_{id_x})' \) is the \((d_x \times 1)\) vector of idiosyncratic error terms, and \( z_{it} \equiv (w'_{it}, \tilde{\omega}_{it})' \) is of dimension \( d_z \). The dimension of the vector of instruments, \( \tilde{\omega}_{it} \), is greater than or equal to the dimension of \( x_i \). Such exclusion restriction, where \( \tilde{\omega}_{it} \) appears in the reduced form but not in the structural, are justified on economic grounds.

Since the exogenous variables, \( w_{it} \), have no bearing on the identification results obtained in the article, to ease notations we suppress it in the binary response model in the rest of the article. All assumptions and results are to be understood as conditional on \( w_{it} \). Second, in the rest of the article, except when needed, we will drop the individual subscript, \( i \).

While we refer (2.2) as reduced form equation, it is possible that the triangular system in (2.1) and (2.2) is in fact fully simultaneous (see e.g., Blundell and Powell 2004). However, even if a simultaneous system is not triangular, the triangular representation, such as the above, can be easily derived if the simultaneous equations involving \( x_i \) is triangular. For the sake of exposition, we limit the analysis to fixed coefficients with random effects; a straightforward extension of the method to allow for random coefficients is discussed in Appendix 2 of the Online Supplemental Appendix.

We first define some notations. Let \( X \equiv (x_1, ..., x_T) \), a \((d_x \times T)\) matrix, denote the history of the endogenous variables, \( x \); let \( Z \equiv (z_1, ..., z_T) \) of dimension \((d_z \times T)\), denote the history of the exogenous variables, \( z \); similarly, \( \zeta \equiv (\zeta_1, ..., \zeta_T)' \) is a vector containing the realizations of idiosyncratic shocks in the structural equation, and \( \epsilon \equiv (\epsilon_1, ..., \epsilon_T) \) is a \((d_x \times T)\) matrix containing the realizations of idiosyncratic shocks in the reduced form equations.

The first assumptions toward identifying the structural measures of interest such as ASF and APE are:

AS 1. \( \zeta, \epsilon \perp \perp Z, \theta, x \).

AS 2. (a) \( \theta, \zeta | X, Z, \alpha \sim \theta, \zeta | e, Z, x \sim \theta, \zeta | e, x \) where \( e = X - E(X|Z, x) \),
(b) \( \theta, \zeta_t \perp \epsilon_{t-1} | x, \epsilon_t \).

AS 3.

\[ x | Z \sim N\left[ E(x|Z), \Lambda_{xx} \right] \text{ and } \epsilon_t \sim N\left[ 0, \Sigma_{\epsilon} \right], \]

where \( E(x|Z) = \pi \tilde{z} \) could be either Chamberlain’s or Mundlak’s specification for correlated random effects.

Assumptions AS 1 and AS 2, which serve to account for unobserved confounders such as the unobserved heterogeneities that are fixed at least in short panels and to eliminate the confounding influences of observed and unobserved confounders, are weaker than the identifying assumptions for the traditional control function method such as in BP and Rothe (2009). In the traditional control function method, (a) \( Z \) is assumed independent of all heterogeneity terms, \( \theta, \zeta_t, x, \epsilon \) and (b) it is assumed that \( \theta + \zeta_t \perp \perp X | x + \epsilon_t = v_t \); such an assumption also implies that heterogeneity in each of the \( d_x \) reduced form equations is scalar, whereas one would like to allow for additional heterogeneities such as individual effects and/or random coefficients. We allow \( Z \) to be correlated with \( \theta \) and \( x \), and in part (a) of AS 2, assume that conditional on the history of reduced form error terms, \( \epsilon \) and \( x, Z \), and thereby \( X \), is independent of the structural error terms \( \theta \) and \( \zeta_t \). The assumption in AS 2 (b), where only contemporaneous errors are correlated, has been made to ease exposition, and can be dropped.

To see why Assumption AS 2 (a) is justified or realistic in empirical settings, consider the empirical example in BP, in which they estimate the causal effect of “other” household income on work participation decision by men without college education. The triangular model in BP augmented with individual effects is given as:

\[ y_t = 1 \{ y_t^* = x_t \varphi + z_t \varphi_x + \theta + \zeta_t > 0 \} \]

\[ x_t = z_t \pi_1 + z_{2t} \pi_{21} + z_{22} \pi_{22} + \alpha + \epsilon_t, \]
where $y_t'$ is the number of hours worked in a week by the man in the house; $x_t$, which is weekly “other” household income and which includes the income of the spouse, is endogenous; $z_{1t}$ is a set of strictly exogenous variables that includes various observable social demographic variables; $z_{21t}$ is a set of strictly exogenous variables that includes household characteristics, for example, the education level of the spouse; and the instrument, $z_{22t}$, is the weekly welfare benefit entitlement variable, which is excluded from the structural Equation (2.3). This entitlement variable measures the transfer income the family would receive if neither spouse was working.

The above triangular representation can be obtained by augmenting with individual effects the simultaneous equation model considered in Blundell and Smith (1994):

$$ y_t' = x_t \varphi + z_{1t} \varphi_z + \theta + \zeta_t $$  \hspace{1cm} (2.5)

$$ x_t = y_t' \beta_y + z_{21t} \beta_{z1} + z_{22t} \beta_{z2} + \gamma + \zeta_t. $$  \hspace{1cm} (2.6)

Since the structural equation, (2.5) is derived using a wage equation (see also BP), the individual effect, $\theta = f(\mu, \omega)$, could be a composite of unobserved “taste” for work, $\mu$, and unobserved ability/productivity, $\omega$; whereas $\gamma$ in Equation (2.6) could represent household’s or spouse’s unobserved productivity (see Blundell et al. 2007). Substituting $x_t \varphi + z_{1t} \varphi_z + \theta + \zeta_t$ for $y_t'$ in Equation (2.6), we get the reduced form in Equation (2.4), where

$$ \pi_1 = \frac{\varphi \beta_y}{1 - \varphi \beta_y}, \pi_{21} = \frac{\beta_{z1}}{1 - \varphi \beta_y}, \pi_{22} = \frac{\beta_{z2}}{1 - \varphi \beta_y}, \alpha = \frac{\theta \beta_y + \gamma}{1 - \varphi \beta_y}, \text{ and } \epsilon_t = \frac{\zeta_t \beta_y + \zeta_t}{1 - \varphi \beta_y}. $$

Let $z_t = \{z_{1t}, z_{21t}, z_{22t}\}$. First, given what the unobserved heterogeneities, $\theta, \gamma, \text{ and } \alpha = g(\theta, \gamma)$, are, it is quite likely that $z_t$, which includes the education level of the couple and the welfare benefits they receive, is correlated with them. Second, if, as in HW, $\zeta_t$ in Equation (2.5) represents new private information revealed to the household, which affects both $y_t'$ and $x_t$, then (a) $\zeta_t = f(\zeta_t)$ in Equation (2.6) and (b) even after conditioning on individual effects, $x_t$ and $\zeta_t$ would be dependent.

For the example above and in general, given $Z$, the only source of dependence between $X$ and $(\theta, \zeta_t)$ is through the relationship between $(x, \epsilon_1, ..., \epsilon_T)$ and $(\theta, \zeta_t)$. Therefore, given $Z$, conditioning on $(x, \epsilon_1, ..., \epsilon_T)$ eliminates this source of dependency. Since $\epsilon_t$ and $\zeta_t$ are by Assumption AS 1 independent of $Z$, it can be shown that Assumption AS 2 (a) boils down to $\theta \perp \perp Z | x$. In other words, what we are assuming is that no information about $Z$ is contained in $\theta$ over and above that contained in $x$. This assumption, in effect, is similar to the assumption of strict exogeneity in panel data models: once the endogeneity of $x_t$ has been addressed by conditioning on $(x, \epsilon_1, ..., \epsilon_T)$, given time-invariant heterogeneity, $x, Z$ has no extraneous influence on $y_t'$.

Now, as the unobserved conditioning variables, $x$ and $\epsilon_t$, cannot be identified separately, for identifying structural measures our method then requires that we be able to recover the conditional distribution of $\alpha$ given $X$ and $Z$ so that the control functions, which are based on $E(x|X, Z)$, can be estimated. However, we do not know of any semiparametric or nonparametric estimator, and it is outside the scope of this article to develop one, where the distribution of the or the expectation of individual effects or random coefficients conditional on $X$ and $Z$ are estimated for a system of regressions. Although, with parametric specification of the error components as in Assumption AS 3, this conditional distribution is obtained readily.

Bjørn (2004) proposed a step-wise maximum likelihood method for estimating the systems of regression equations, where the distributions of error components are specified as normal. Given Assumption AS 3, where the conditional distribution of $x$ given $Z$ and the marginal distribution of $\epsilon_t$ are both normal, the tail, $a = x - E(x|Z) = x - \bar{z} \bar{z}$, is distributed normally with conditional mean zero and variance $\Lambda_{x2}$. We can, therefore, write the reduced form in (2.2) as

$$ x_t = \pi z_t + \bar{z} \bar{z} + a + \epsilon_t, $$  \hspace{1cm} (2.7)

which can be estimated using the method in Bjørn (2004)
Some recent papers that have employed Mundlak’s specification for correlated random effects are listed in SW. Although we have assumed the error term to be normally distributed, as we discuss here, violation of assumed normality of reduced form errors is unlikely to have a bearing on the estimates. First, the coefficients in Biørn (2004) are estimated by the method of generalized least squares (GLS), which does not require normality of the errors, \( \alpha \) and \( \epsilon_t \). Second, Biørn shows that the ML estimates of the covariance matrices, \( \Sigma_{\epsilon t} \) and \( \Lambda_{22} \), for a moderately large \( N \) are approximately same as those that are obtained when the distributions of \( \alpha \) and \( \epsilon_t \) are unknown. Third, estimating the reduced form equation augmented with \( z = T^{-1} \sum_{t=1}^{T} z_t \) – to account for the correlation between \( \alpha \) and \( Z \) – by GLS yields fixed-effects (FE) estimates of \( \pi \) for time varying \( z_t \) (see Wooldridge 2019).

For scalar \( x \), Baltagi et al. (2010) allow for heteroscedastic \( a \) and serial correlation among \( \epsilon_t \). Thus, when \( d_x = 1 \), the assumptions that \( a \) is completely independent of \( Z \) and that \( \epsilon_t' \)s are i.i.d. can be weakened to allow for nonspherical error components. However, since we want to account for the endogeneity of multiple endogenous regressors, we will stick to Assumption AS 3, and estimate the first-stage parameters, \( \Theta_t = \{ \pi, \bar{\pi}, \Sigma_{\epsilon t}, \Lambda_{22} \} \), of the reduced form Equation (2.7) using Biørn’s step-wise likelihood method, which is briefly described in Appendix 4 of the Online Supplemental Appendix.

### 2.1. Identification of structural coefficients

Now, by Assumptions AS 1 and AS 2 the dependence of \((\theta, \zeta)\) on \( X, Z \) and \( \alpha \) is characterized by \( \alpha \) and \( \epsilon_t \). If \( \alpha \) and \( \epsilon_t \) could be identified, we could augment the structural equation with \( \alpha \) and \( \epsilon_t \), and estimate the coefficients, \( \varphi \). Since \( \alpha \) and \( \epsilon_t \) are not identified separately, the traditional control function approach assumes that the composite error, \( v_t = \alpha + \epsilon_t \), which are estimated as the residuals of the reduced form equations, is independent of \( Z \) and that conditional on \( v_t \), \( X \) is independent of \( \theta + \zeta \). Such an assumption, as we discuss in detail, could quite likely be violated.

In Theorem 1, we show that by estimating the modified structural Equation (2.9), which is augmented with the additional control variables, \( \hat{\alpha}(X, Z) \equiv E(\alpha | X, Z) \) and \( \hat{\epsilon}_t(X, Z) \equiv E(\epsilon_t | X, Z) \), the structural coefficients \( \varphi \) can be estimated consistently. The modified structural Equation (2.9) is derived based on Lemma 1, and the control variables are identified in Lemma 2.

In Lemma 1, we show that:

**Lemma 1.** If (i) Assumptions AS 1 and AS 2 hold and (ii) \( E(\theta | \alpha) \) and \( E(\zeta_t | \epsilon_t) \) are linear in \( \alpha \) and \( \epsilon_t \), respectively, so that \( E(\theta | \alpha) + E(\zeta_t | \epsilon_t) = \varphi_x \alpha + \varphi_{\alpha} \epsilon_t \), then, \( E(\theta + \zeta_t | X, Z) \) depends on \((X, Z)\) only through \( \hat{\alpha}(X, Z) \) and \( \hat{\epsilon}_t(X, Z) \).

**Proof of Lemma 1.** Now,

\[
E(\theta + \zeta_t | X, Z) = E(E(\theta + \zeta_t | X, Z, \alpha) | X, Z) = E(E(\theta + \zeta_t | \alpha, \epsilon_t) | X, Z)
\]

\[
= E(E(\theta | \alpha) + E(\zeta_t | \epsilon_t) | X, Z) = \varphi_x E(\alpha | X, Z) + \varphi_{\alpha} E(\epsilon_t | X, Z),
\]

(2.8)

where the first equality is due to the law of iterated expectations, the second is due to Assumptions AS 2, the third due to AS 1 and the fourth due assumption (ii) in the Lemma.

In Lemma 2, we show that:

**Lemma 2.** Let \( x_t = \pi z_t + \bar{\pi} z + a + \epsilon_t \), \( t \in \{1, \ldots, T\} \), where \( z = \frac{1}{T} \sum_{t=1}^{T} z_t \), and let AS 3 hold, then \( \alpha = \bar{\pi} z + a \), given \( X \) and \( Z \), is distributed with conditional mean

\[a^{1}\text{In Lemma 2, part (b), in the Appendix we derive the conditional distribution of } \alpha \text{ given } X \text{ and } Z \text{ for the estimator in Baltagi et al. (2010), where } x \text{ and } \epsilon_t \text{ are both scalar, } x \text{ is heteroscedastic and the distribution of } \epsilon_t \text{ is nonspherical.}
\[ E(\alpha|X, Z) \equiv \tilde{\alpha}(X, Z, \Theta_1) = \bar{\pi}z + E(\alpha|X, Z) = \bar{\pi}z + \Omega \Sigma_{\epsilon}^{-1} \sum_{t=1}^{T} (x_t - \pi z_t - \bar{\pi}z), \]

where \( \Omega = [T \Sigma_{\epsilon}^{-1} + \Lambda_{xx}]^{-1} \) is the conditional variance of \( \alpha \) given \( X \) and \( Z \); \( \Lambda_{xx} \) and \( \Sigma_{\epsilon} \) being the covariance matrices of \( \alpha \) and \( \epsilon_t \), respectively.

**Proof of Lemma 2.** Given in Appendix A.

Conditional mean of \( \epsilon_t \) given \( X \) and \( Z \) is then given by

\[ \hat{\epsilon}_t(X, Z, \Theta_1) = x_t - \pi z_t - E(\alpha|X, Z) = x_t - \pi z_t - \hat{\alpha}(X, Z, \Theta_1) = v_t - \hat{\epsilon}_t(X, Z, \Theta_1). \]

**Remark 1.** The expected posterior estimates, \( \hat{\alpha} \), of \( \alpha \) in Lemma 2, however, is the empirical Bayes or the James and Stein (1961) shrinkage estimator of \( \alpha \) (see Efron 2010). The empirical Bayes estimation has gained certain popularity in economics. In education economics, it is employed as a procedure to calculate teacher value added and often as a way to make imprecise estimates more reliable (see Guarino et al. 2015, and the references therein).

Now, we can write \( \hat{\alpha} \) in Lemma 2 as

\[ \hat{\alpha} = \bar{\pi}z + E(\alpha|X, Z) = \bar{\pi}z + \left[ \Sigma_{\epsilon}^{-1} + \frac{1}{T} \Lambda_{xx}^{-1} \right]^{-1} \bar{\pi}z + \left[ \Sigma_{\epsilon}^{-1} + \frac{1}{T} \Lambda_{xx}^{-1} \right]^{-1} \sum_{t=1}^{T} (x_t - \pi z_t - \bar{\pi}z). \]

With the reduced form equation being specified as in Equation (2.7) and given Assumption AS 3, it can be verified that, for a given \( \Theta_1 \), the MLE of \( \alpha \) is \( \frac{1}{T} \sum_{t=1}^{T} (x_t - \pi z_t - \bar{\pi}z) \). Since for small \( T \) the MLE of \( \alpha \) is less reliable, the shrinkage factor of the empirical Bayes estimator shrinks the MLE of \( \alpha \) toward its mean, \( \bar{\pi}z \). Given consistent estimates of reduced-form parameters, \( \Theta_1 \), the empirical Bayes estimate, \( \alpha \), of \( \alpha \) is the minimum mean squared error predictor of \( \alpha \) under normality, and therefore, a justified estimator of \( \alpha \).

For large \( T \), since the estimates of \( \pi \) are the FE estimates of \( \pi \), it can be shown that \( \hat{\alpha} \) consistently estimates the fixed effects, \( \alpha \). With the FE estimates of \( \alpha \) given by \( \hat{\alpha}_{FE} = \frac{1}{T} \sum_{t=1}^{T} (x_t - \pi z_t) \), we can write \( \hat{\alpha} \) as

\[ \hat{\alpha} = \bar{\pi}z + \left[ \Sigma_{\epsilon}^{-1} + \frac{1}{T} \Lambda_{xx}^{-1} \right]^{-1} \Sigma_{\epsilon}^{-1}(\hat{\alpha}_{FE} - \bar{\pi}z). \]

Assuming \( N \) is large to have consistently estimated the reduced form parameters, since \( \hat{\alpha}_{FE} \) converges in probability to \( \alpha \) and \( \left[ \Sigma_{\epsilon}^{-1} + \frac{1}{T} \Lambda_{xx}^{-1} \right]^{-1} \Sigma_{\epsilon}^{-1} \) to an identity matrix as \( T \to \infty \), by continuous mapping theorem it can be shown that \( \hat{\alpha} \to \alpha \), and consequently \( \hat{\epsilon}_t \to 0 \).

Given Lemma 1, we have \( E(y_t'|X, Z) = x_t'\varphi + E(\theta + \zeta_t|X, Z) = x_t'\varphi + \varphi_z \hat{\alpha} + \varphi_z \hat{\epsilon}_t \). We can then write Equation (2.1) written as

\[ y_t = 1\{X_t'\Theta_2 + \eta_t > 0\}, \quad (2.9) \]

where \( \Theta_2 \equiv (\varphi_2', \varphi_2', \varphi_z')' \), \( X \equiv (x_t, \hat{\alpha}_t, \hat{\epsilon}_t)' \) and \( \eta_t = \theta + \zeta_t - E(\theta + \zeta_t|X, Z) \). The two vectors, \( \varphi_z \) and \( \varphi_z \), when estimated give us a test of exogeneity of \( x_t \). Although \( \eta_t \) by construction is mean

---

2For notational convenience, we use \( \hat{\alpha}(X, Z, \Theta_1) \), \( \tilde{\alpha}(X, Z) \) and \( \hat{\alpha} \) interchangeably; the same for \( \hat{\epsilon}_t(X, Z, \Theta_1) \), \( \tilde{\epsilon}(X, Z) \) and \( \hat{\epsilon}_t \).

3When there is a single endogenous regressor, so that the reduced form has a single equation, then one can employ the estimation method in Gu and Koekker (2017), who, for longitudinal data, have developed a nonparametric estimation method to estimate the empirical Bayes estimates of the individual effects, \( \alpha_t \), and the distribution of \( \alpha_t \). Since the posterior mean \( \hat{\alpha} \) is estimated nonparametrically, the large sample properties of the structural coefficients will have to be worked out anew.
independent of \( X_t \), for estimation of binary response model in (2.9), the strong condition of complete independence is required (Manski 1988), or

\[ \text{AS 4. } \eta_t \perp \Gamma X_t. \]

Since \( \varepsilon_t \) and \( \tilde{\alpha} \), both, are of dimension \( d_x \), the dimension of \( X_t \) is \( 3d_x \). The identification conditions for \( \Theta_2 \) in (2.9) to be identified when \( \eta_t \) is assumed to follow a known distribution are: (a) \( \eta_t \) be independent of \( X_t \) and (b) \( \text{rank}(E(X_t X_t')) = 3d_x \). In Theorem 1 we show that condition (b) is satisfied.

**Theorem 1.** If (i) \( \text{rank}(E(x_t x_t')) = d_x \); (ii) \( \text{rank}(\Pi) = d_x \), where \( \Pi = (\pi, \tilde{\pi}) \); (iii) \( \text{rank}(E(z_t z_t')(z_t z_t')) = k \), where \( k = \text{dim}(z_t z_t') \); and (iv) if AS 3 holds so that the covariance matrices of \( \varepsilon_t \) and \( \tilde{\alpha} \) are of full rank, then \( \text{rank}(E(X_t X_t')) = 3d_x \).

**Proof of Theorem 1.** Given in Appendix A.

Condition (ii) is the rank condition in IN, and it underscores the necessity of exclusion restriction for identification. Conditions (i) to (iii) in Theorem 1 are standard conditions for identification of \( \varphi \) in the traditional control function methods, where the control function is the composite error, \( v_t = \alpha + \varepsilon_t = x_t - \pi z_t \). Our conditioning variables, however, are \( \varepsilon_t \) and \( \tilde{\alpha} \), which are also functions of \( \Lambda_{\alpha} \) and \( \Sigma_{\alpha} \). Positive definiteness of \( \Lambda_{\alpha} \) and \( \Sigma_{\alpha} \) in condition (iv) helps establish the statement of the Theorem to be true.

**Appendix 1** of the Online Supplemental Appendix discusses how one can use the method of generalized estimating equation (GEE), which can account for heteroscedasticity and serial dependence in the response outcome, to estimate \( \Theta_2 \). Following Theorem 1, since the components of \( x_t \) are continuous, with scale and location normalization, \( \Theta_2 \) can be estimated by semiparametric methods without specifying the distribution of \( \eta_t \) (see Horowitz 2009, for a review of identification results for semiparametric binary choice models).

Now, we have demonstrated that, with \( \tilde{\alpha} \) and \( \tilde{\varepsilon}_t \) identified in Lemma 2, AS 1, AS 2, condition (ii) of Lemma 1, and AS 4 can help us identify \( \varphi \). Given \( \tilde{\alpha} \) and \( \tilde{\varepsilon}_t \), the same, however, can be achieved through the following assumption:

**ACF 1.** (a) \( \zeta, \theta \mid X, Z, \tilde{\alpha} \sim \zeta, \theta \mid V, Z, \tilde{\alpha} \sim \zeta, \theta \mid V, \tilde{\alpha} \), where \( V \equiv (v_1, ..., v_T) = X - \pi Z \) and \( \tilde{\alpha} = E(\tilde{\alpha} | X, Z) \).

(b) \( \theta, \zeta_t \perp V_{-t} | v_t, \tilde{\alpha} \).

In part (a), the assumption is that the dependence of \( (\theta, \zeta) \) on \( X \) and \( Z \) is completely characterized by \( V \) and \( \tilde{\alpha} \). Since \( \tilde{\varepsilon}_t = v_t - \tilde{\alpha} \), there is one-to-one mapping between \( (\tilde{\varepsilon}_t, \tilde{\alpha}) \) and \( (v_t, \tilde{\alpha}) \), and therefore, the conditioning \( \sigma \)-algebra, \( \sigma(\tilde{\varepsilon}_t, \tilde{\alpha}) \), is same as the \( \sigma \)-algebra, \( \sigma(v_t, \tilde{\alpha}) \). By parts (a) and (b), therefore, \( X \) is independent of \( (\zeta_t, \theta) \) given \( (\tilde{\varepsilon}_t, \tilde{\alpha}) \). In other words, in Assumption ACF 1 we are proposing \( \tilde{\varepsilon}_t \) and \( \tilde{\alpha} \) as control functions for panel data. If we further assume that \( E(\zeta_t + \theta | \tilde{\varepsilon}_t, \tilde{\alpha}) \) is linear in \( \tilde{\varepsilon}_t \) and \( \tilde{\alpha} \), and let \( \eta_t = \zeta_t + \theta - E(\zeta_t + \theta | \tilde{\varepsilon}_t, \tilde{\alpha}) \), we obtain Equation (2.9).

Given AS 4, as shown in Theorem 1, we can then estimate the structural coefficients, \( \varphi \).
We now compare the proposed control functions with the traditional control functions, and argue for the appropriateness of the proposed control functions in the context of panel data. Now, we have pointed out that conditioning on the proposed control functions, \( \hat{e}_t \) and \( \hat{\alpha} \), is equivalent to conditioning on the traditional control function, \( \nu_t(x_t, z_t) = x_t - \pi z_t \), and additionally on individual specific information as summarized by \( \hat{\alpha}(X,Z) \). That is, in assuming that \((\zeta_t, \theta) \perp X|\hat{e}_t, \hat{\alpha}\), we are saying that no information about \( X \) is contained in \((\zeta_t, \theta)\) over and above that contained in \((\hat{e}_t, \hat{\alpha})\) or equivalently in \((\nu_t, \hat{\alpha})\). This, as we discuss in Remark 2 and Remark 3, may not hold true if only \( \nu_t \) is assumed to be the control function.

**Remark 2.** When \( Z \perp (\theta, x, \zeta_t, \epsilon_t) \), then, the requirement of the traditional control function method that \( Z \) be independent of \((\nu_t, \zeta_t + \theta)\) is violated and \( \zeta_t + \theta \perp Z|\nu_t \) does not hold generally. In which case, Assumption ACF 1 seems plausible as \( \zeta_t + \theta \) is mean independent of \((X,Z)\) given \((\hat{\alpha}, \hat{e}_t)\). This assumption is related to the dependence assumptions in AM, Bester and Hansen (2009) (BH) and HW, where the distribution of unobserved effects depends on the observed variables only through certain function of the observed variables. These functions, as BH argue, may be viewed as sufficient statistic. AM assume that \((\zeta_t, \theta)\) is independent of \( X \) given certain summary statistics such as the mean, \( T^{-1} \sum_{t=1}^{T} x_t \), or index functions of summary statistics, while in BH these functions of observed variables are assumed to be unrestricted index functions. In our case, the control function, \((\hat{e}_t, \hat{\alpha})\), is motivated by the result that under certain restrictions, the mean of \( \theta + \zeta_t \) given the histories, \((X,Z)\), of the endogenous and the exogenous variables, depends on \((X,Z)\) only through \( \hat{\alpha} \) and \( \hat{\alpha} \). Moreover, as \((\hat{\alpha}, \hat{e}_t)\) consistently estimates \((x_t, \epsilon_t)\) when \( T \) is large (see Remark 1), it implies that for large \( T \), Assumptions ACF 1 and AS 2 are asymptotically equivalent.

**Remark 3.** When \( Z \perp (\theta, x, \zeta_t, \epsilon_t) \), as in the traditional control function approach, then \((\zeta_t, \theta) \perp Z|V \) holds true. Since \( V \) is invertible in \( X \) when \( Z \) is given, we have \((\zeta_t, \theta)|X,Z \sim (\zeta_t, \theta)|V, Z \sim (\zeta_t, \theta)|V \). In panel data, therefore, when \( Z \) is independent of the unobserved heterogeneities, \( V \) should be employed as the control function.

Given that the unobserved heterogeneities, \((\theta, x)\), which represent unobserved, time-invariant attributes, such as preferences, technologies or abilities, influence the choice of \( x_t \) in each time period, it is not only with \( x_t \) that the errors, \( \theta + \zeta_t \), are correlated, but generally with the entire history, \( X \), of the endogenous variable. Moreover, since \( x_t \) is endogenous, due to potential feedback from \( y_t \) to \( x_t \), for \( s > t \), it is likely that the optimal choice of \( x_t \) depends on \( \zeta_t \) from the past, or more generally \( \zeta_t \) from other time periods, which is likely to make \( \zeta_t \) and \( x_t \) from other time periods dependent.\(^6\) If only \( \nu_t \) is employed as the control function, as it has been traditionally, then, \( X \) may not be conditionally independent of \((\zeta_t, \theta)\). Therefore, there will exist some partial correlation between \( y_t \) and \( x_t \) from the other time periods if the dependency between the structural errors, \((\theta, \zeta_t)\), and the history, \( X \), is not accounted for. Employing only \( \nu_t \) as the control function places a strong restriction on the dependence between \((\theta, \zeta_t)\) and \( X \), which, as shown in Proposition 1, is unlikely to hold.\(^7\)

**Proposition 1.** Let \( Z \perp (\theta, x, \zeta_t, \epsilon_t) \). When \((\theta, \zeta_t)\) and \((x_t, \epsilon_t)\) are correlated, then, \( \theta + \zeta_t \perp X|V \) whereas \( \theta + \zeta_t \not\perp X|\nu_t \), where \( \nu_t = x_t + \epsilon_t = x_t + \pi z_t \) and \( V \equiv (\nu_1, \ldots, \nu_T)\).

\(^6\)For example, in the study of child labor in section 4, the endogenous variables, household income, amount of land owned by a household and index of household ownership of productive farm assets, in each period will also depend on unobserved household characteristics such as parents’ abilities, quality of land or possibly other omitted variables fixed at the household level. Moreover, apart from contemporaneous shocks, \((\zeta_t, \epsilon_t)\), that affect the current choices of both \( x_t \) and \( y_t \), there may be feedback from lagged values of \( \zeta \) or \( y_t \) to \( x_t \). In the study of the child labor, for example, current choices of labor supply can impact the future choices of endogenous variables aforementioned. It is, thus, possible that \((\theta, \zeta_t)\) and \( X \), as in the considered example of child labor, could be dependent.

\(^7\)See also section 3, Fig. C.2, where through numerical experiments we demonstrate the claims made in this Remark and Proposition 1.
Proof of Proposition 1. Given in Appendix A.

Here, we would like to point that \((\hat{z}, \hat{e}_t)\) can be employed in semiparametric methods in BP or in Rothe (2009) to estimate \(\varphi\) and measures like the ASF semiparametrically. BP extend the the matching estimator for the single-index models with exogenous variables to allow for control functions for handling endogeneity, whereas Rothe develops a semiparametric maximum likelihood (SML) method for binary response model to account for endogeneity using control functions. These semiparametric methods do not require one to specify the conditional distribution of \(\theta + \zeta_t\) given \(\hat{e}_t\) and \(\hat{z}\). The methods, however, do require that, \(z = \tilde{z}\), contains an instrument that is continuous. If all instruments are discrete, the “rank condition” in BP and condition (ii) of Theorem 1 in Rothe, necessary for identification, are violated. We do not pursue semiparametric estimation of binary choice models with the control functions developed in this article any further. Semiparametric estimation and the large sample properties of the estimates are left for future research.

Wooldridge (2015) in providing an overview of control function (CF) methods writes, “in evaluating the scope of an estimation method, it is important to understand how it works in familiar settings, including cases when it is not necessarily needed.” While noting that the standard CF method for cross sectional data when applied to linear models gives coefficients for the endogenous variables that are equal to the standard two-stage least squares (2SLS), Wooldridge points out certain advantages of the CF approach, such as providing a robust, regression based Hausman test of exogeneity, compared to the 2SLS approach. Before ending this subsection, we, therefore, show that using \(z\) and \(\hat{e}_t\) as additional covariates in linear panel data models is equivalent to estimating the models by a certain 2SLS method.

Now, in linear panel data models,

\[
y_t = x_t' \varphi + \theta + \zeta_t,
\]

when instruments are correlated with time invariant heterogeneity, fixed effect two-stage least squares (FE2SLS), which employs time-variant instruments that are deviations from the group-mean, \(\tilde{z}_t = z_t - \bar{z}\), is employed. An alternative approach (see Wooldridge 2010, chapter 11), is to write \(\theta\) in Equation (2.10) as \(\theta = E(\theta|Z) + \tau\), specify \(E(\theta|Z)\) as in Mundlak, and estimate the model by pooled 2SLS using \((\hat{z}_t, \tilde{z}_t)\) as instruments.

By Assumptions AS 1, AS 2, AS 3 and condition (ii) of Lemma 1, we have \(E(\theta|Z) = E(E(\theta|x)|Z) = E(\varphi_x z|Z) = \varphi_x \tilde{z}\). Thus, we can write the model in Equation (2.10) as

\[
y_t = x_t' \varphi + \varphi_x \tilde{z} + \tau + \zeta_t,
\]

where the heterogeneity term, \(\tau + \zeta_t\) and \(X\) are dependent even though \(\tau + \zeta_t\) is mean independent of \(Z\). Thus, we can estimate of \(\varphi\) in (2.11) by using instrument variables. Also, by Assumptions ACF 1 and, with a slight abuse of notations, letting \(E(\zeta_t + \theta|\hat{e}_t, \hat{z}) = \varphi_x \hat{z} + \varphi_x \hat{e}_t\), the linear model in Equation (2.10) as can be written as

\[
y_t = x_t' \varphi + \varphi_x \hat{z} + \varphi_x \hat{e}_t + \eta_t,
\]

where \(\eta_t = \theta + \zeta_t - E(\theta + \zeta_t|\hat{e}_t, \hat{z})\).

Theorem 2. Let the estimate of \(\varphi\) in (2.11) by pooled 2SLS using \((\hat{z}_t, \tilde{z}_t)\) as additional instruments be denoted by \(\hat{\varphi}_{IV}\) and let the estimate of the same obtained from estimating (2.12) by pooling the data be denoted by \(\hat{\varphi}_{CF}\), then \(\hat{\varphi}_{CF} = \hat{\varphi}_{IV}\).

Proof of Theorem 2. Given in Appendix A.

---

10Since conditioning on \(\hat{e}_t\) and \(\hat{z}\) is equivalent to conditioning on \(v_t = x_t - \pi z_t\) and \(\hat{z}\), and if identification requires that conditional on the control variables, \(\hat{e}_t\) and \(\hat{z}\), \(v_t\) and \(\hat{z}\) the vector \(x_t\) contains at least one, \(x_t'\), continuously distributed component with non-zero coefficient, then it would be necessary that \(z_t\) contains a continuously distributed regressor.
Proof of Lemma 3.

Estimating (2.11) by pooled 2SLS using \((\tilde{z}_t, \tilde{z})\) as additional instruments is similar to estimating by the error component two-stage least squares (EC2SLS) method proposed in Baltagi (1981), with the difference that the time-varying instruments are allowed to be correlated to the individual specific unobserved heterogeneity.

2.2. Identification of ASF and APEs

With \(\hat{a} \) and \(\hat{e}_t\) as control functions, we have

\[
\Pr(y_t = 1|X, \hat{a}, \hat{e}_t) = \int \{-(\zeta_t + \theta) < x'_t \phi \} dF(\zeta_t + \theta|X, \hat{a}, \hat{e}_t) = \int \{-(\zeta_t + \theta) < x'_t \phi \} dF(\zeta_t + \theta|\hat{a}, \hat{e}_t) = F(x'_t \phi; \hat{a}, \hat{e}_t),
\]

where \(F(x'_t \phi; \hat{a}, \hat{e}_t)\) is the conditional CDF of \(\zeta_t + \theta\) given \((\hat{a}, \hat{e}_t)\) evaluated at \(x'_t \phi\).

Given a particular value \(x\) of \(x_t\), averaging \(F(x'_t \phi; \hat{a}, \hat{e}_t)\) over \((\hat{a}, \hat{e}_t)\), we get the ASF:

\[
G(\hat{a}) = \int F(x'_t \phi; \hat{a}, \hat{e}_t) dF(\hat{a}, \hat{e}_t),
\]

\[
= \int \left[ \int 1\{x'_t \phi + \theta + \zeta_t > 0\} dF(\theta + \zeta|\hat{a}, \hat{e}_t) \right] dF(\hat{a}, \hat{e}_t)
= E_{\theta + \zeta}(1\{x'_t \phi + \theta + \zeta > 0\}).
\]

(2.13)

The APE of changing a variable, say \(\hat{x}_k\), from \(\hat{x}_k\) to \(\hat{x}_k + \Delta_k\) can be obtained as

\[
\frac{\Delta G(\hat{a})}{\Delta_k} = \frac{G(\hat{x}_k, \hat{x}_k + \Delta_k) - G(\hat{x}_k)}{\Delta_k}.
\]

(2.14)

To point-identify the ASF, \(G(\hat{a})\), it is required that \(F(x'_t \phi = x'_t \phi; \hat{a} = \hat{a}, \hat{e}_t = \hat{e}_t)\) be evaluated at all values of \((\hat{a}, \hat{e})\) in the support of the unconditional distribution of \((\hat{a}, \hat{e})\). This requires that the support of the conditional distribution of \((\hat{a}, \hat{e})\) conditional on \(x_t = \hat{x}\) be equal to the support of the unconditional distribution. It ensures that for any group of individuals defined in terms of \((\hat{a}, \hat{e}_t)\), at least some experience \(x_t = \hat{x}\). This is analogous to the overlap condition in the program evaluation literature, where treatment is discrete.

For many triangular systems that employ the control function, \(v_{i_t}\), or \(\{F(v_{i_1, t})...F(v_{i_s, t})\}' - where \(F(v_{i_1, t})\), the CDF of \(v_{i_1, t}\) is equal to \(F(x_{i_1, t}|z_{i_t})\), the CDF of \(x_{i_1, t}\) given \(z_{i_t}\) – the requirement of common support necessitates that along with the rank condition (see Theorem 1) the set of instruments, \(z_{i_t}\), contains a continuous instrument with large support (this is discussed in IN, Florens et al. and in Blundell et al. (2003)). In Lemma 3, we show that when the instruments have small support – that is, when instruments are binary, discrete or continuous but without large support – the support requirement for \(G(\hat{a})\) to be point-identified by the “partial-mean” formulation in Equation (2.13) is satisfied if \(x\) has a large support.

Lemma 3. If the endogenous variables, \(x\), has large a support, then under AS 3, the support of the conditional distribution of \(\hat{a}(X, Z, \Theta_1)\) and \(\hat{e}_t(X, Z, \Theta_1)\), conditional on \(x_t = \hat{x}\), is same as the support of their marginal distribution.

Proof of Lemma 3. Given in Appendix A.

In our approach, the control functions, \(\hat{e}_t\) and \(\hat{a}\), are smooth, unbounded functions of \(x_t\)'s, \(t \in \{1, ..., T\}\). Therefore, when \(x\) is continuous with a large support, because the \(x_t\)'s, \(s \neq t\), are unrestricted the ranges of \(\hat{a}\) and \(\hat{e}_t = x_t - \pi z_{i_t} - \hat{a}\) conditional on \(x_t = \hat{x}\) do not depend on \(x_t\). Since the result does not rely on any kind of restriction on \(z_{i_t}\)'s support, our method circumvents
the need to have a continuous instrument with large support to point-identify the ASF and/or the APEs when some of the \( x \)'s have large supports.

Before proceeding further, we note that these results could be useful for computing the quantile structural function (QSF), as in IN, for the kind of triangular setups considered in Blundell et al. (2003), where the structural equation is nonseparable in errors, but additively separable in the reduced form. When the nonseparable structural and reduced form equations are strictly increasing in their respective scalar errors, D’Haultfoeuille and Février (2015) and Torgovitsky (2015) show that the structural function, \( y_t = g(x_t, \varepsilon_t) \), is point-identified with discrete instruments. The key observation in both the papers is that establishing sufficient conditions under which the fixed point problem admits a unique solution, they are able to avoid the partial-mean formulation in Equation (2.13), or in Equation (6) of IN’s, to identify the QSF, which in their case is same as the structural function. Since \( g(x_t, \varepsilon_t) \) is required to be strictly monotonic in \( \varepsilon_t \), these methods, however, are not suitable for identification in discrete choice models.

When the support condition in Lemma 3 is not satisfied, one can establish bounds on the ASF and the APE’s. Let \( A \) be the unconditional support of \( (\hat{\alpha}, \hat{\varepsilon}_t) \) and \( A(\hat{x}) \equiv \{ \hat{\alpha}, \hat{\varepsilon}_t : f(\hat{\alpha}, \hat{\varepsilon}_t | \hat{x}) > 0 \} \) be the support of \( (\hat{\alpha}, \hat{\varepsilon}_t) \) conditional on \( \hat{x} \). When the support of \( x \) is bounded and the instruments have a small support then \( A \neq A(\hat{x}) \). Now, let

\[
\tilde{G}(\hat{x}) = \int_{A(\hat{x})} F(\hat{x}' \varphi; \hat{\alpha}, \hat{\varepsilon}_t)dF(\hat{\alpha}, \hat{\varepsilon}_t) \tag{2.15}
\]

be the identified object and let \( P(\hat{x}) = \int_{A \setminus A(\hat{x})} dF(\hat{\alpha}, \hat{\varepsilon}_t) \). Since

\[
G(x_t) = \tilde{G}(\hat{x}) + \int_{A \setminus A(\hat{x})} F(\hat{x}' \varphi; \hat{\alpha}, \hat{\varepsilon}_t)dF(\hat{\alpha}, \hat{\varepsilon}_t)
\]

and since \( 0 \leq F(\hat{x}' \varphi; \hat{\alpha}, \hat{\varepsilon}_t) \leq 1 \), the above equation implies that \( G(\hat{x}) \in [\tilde{G}(\hat{x}), \tilde{G}(\hat{x}) + P(\hat{x})] \); that is, \( G(\hat{x}) \) is set-identified. The bounds on \( G(\hat{x}) \) are sharp since there are no restrictions on \( E(y_t | x_t = \hat{x}, \hat{\alpha}, \hat{\varepsilon}_t) = F(\hat{x}' \varphi; \hat{\alpha}, \hat{\varepsilon}_t) \) imposed by the data.

It follows then that the APEs are also set-identified when support requirement in Lemma 3 is not met. To derive bounds for the APE of changing \( x_k \) from \( \bar{x}_k \) to \( \bar{x}_k + \Delta_k \), let us first denote \( (\hat{x}_{-k}, \bar{x}_k + \Delta_k) \) by \( \hat{x}_{Ak} \). Since the ASF, \( G(\hat{x}_{Ak}) \), at \( \hat{x}_{Ak} \) is partially identified, where \( G(\hat{x}_{Ak}) \in [\tilde{G}(\hat{x}_{Ak}), \tilde{G}(\hat{x}_{Ak}) + P(\hat{x}_{Ak})] \), the APE of \( x_k \) at \( \hat{x} \), \( \Delta G(\hat{x})/\Delta_k \), lies in the interval,

\[
\left[ \frac{\tilde{G}(\hat{x}_{Ak}) - \tilde{G}(\hat{x}) - P(\hat{x})}{\Delta_k}, \frac{\tilde{G}(\hat{x}_{Ak}) + P(\hat{x}_{Ak}) - \tilde{G}(\hat{x})}{\Delta_k} \right], \tag{2.16}
\]

where the sharpness of the bounds on APE derives from that of the bounds on ASF.

Once we have the consistent estimates, \( \Theta_2 \), of \( \Theta_2 \), to estimate the bounds on APE of a variable, \( x_k \), we first, as in IN, estimate the support of the estimates, \( \hat{\alpha}_{it}, \hat{\varepsilon}_{it} \), given \( \hat{x} \) as

\[
\hat{A}(\hat{x}) = \{ \hat{\alpha}_{it}, \hat{\varepsilon}_{it} : \hat{f}(\hat{\alpha}_{it}, \hat{\varepsilon}_{it} | \hat{x}) \geq \delta(\hat{p}), (\hat{\alpha}_{it}, \hat{\varepsilon}_{it}) \in A \},
\]

where \( \hat{f}(\hat{\alpha}_{it}, \hat{\varepsilon}_{it} | \hat{x}) \), which is the estimate of the conditional density of \( (\hat{\alpha}_{it}, \hat{\varepsilon}_{it}) \) given \( \hat{x} \), is obtained by employing the method of estimating the conditional density function in Hall et al. (2004).\(^{11}\) \( \hat{A} \) is an estimator of the support, \( A \), containing all \( (\hat{\alpha}_{it}, \hat{\varepsilon}_{it}) \). In the above, \( \delta(\hat{p}) \) is the trimming parameter and, as discussed in Cadre et al. (2013), is obtained as a solution to the following equation:

\(^{11}\)R’s “np” package developed by Hayfield and Racine (2008) implements the method. The package’s “npcdens” function computes kernel conditional density estimates of \( p \) variables conditional on \( q \) variables.
\[
\int_{\{j \geq \delta\}} f(\hat{\mathbf{z}}_i, \hat{e}_{it} | \mathbf{x}) = \tilde{p}, \text{where } \tilde{p} \text{ is a fixed probability level.}
\]

When \(\tilde{p}\) is close to 1, the upper level set, \(\{\hat{\mathbf{z}}_i, \hat{e}_{it} : f(\hat{\mathbf{z}}_i, \hat{e}_{it} | \mathbf{x}) \geq \delta(\tilde{p})\}\), is close to the support of the conditional distribution.

To estimate \(\delta(\tilde{p})\), let

\[
\hat{H}(\gamma) = \frac{1}{N T} \sum_{i,t} 1 \{ f(\hat{\mathbf{z}}_i, \hat{e}_{it} | \mathbf{x}) \leq \gamma \} \text{ be the estimate of } H(\gamma) = \Pr(f(\hat{\mathbf{z}}_i, \hat{e}_{it} | \mathbf{x}) \leq \gamma),
\]

and let the estimate of \((1 - \tilde{p})\)-quantile of the law of \(f(\hat{\mathbf{z}}_i, \hat{e}_{it} | \mathbf{x})\) be \(\gamma(\tilde{p}) = \inf\{\gamma \in \mathbb{R} : \hat{H}(\gamma) \geq 1 - \tilde{p}\}\). \(\gamma(\tilde{p})\) can be computed by considering the order statistic induced by the sample: \(\hat{f}(\hat{\mathbf{z}}_1, \hat{e}_{11,t} | \mathbf{x}), ..., \hat{f}(\hat{\mathbf{z}}_N, \hat{e}_{N,T,t} | \mathbf{x})\). Cadre et al. (2013) note that whenever \(H(\gamma)\) is continuous at \(\gamma(\tilde{p})\), then \(\delta(\tilde{p}) = \gamma(\tilde{p})\). Following IN, for the application in section 4, we set \(\tilde{p} = 0.975\).

Given the estimate \(\hat{A}(\mathbf{x})\), we can estimate \(\hat{G}(.)\) and \(P(.)\) in (2.15) at \(\mathbf{x}\) as

\[
\hat{G}(\mathbf{x}) = \frac{1}{N T} \sum_{i,t} \Phi(\mathbf{x}'\hat{\varphi} + \hat{\varphi}_2 \hat{\mathbf{z}}_i + \hat{\varphi}_3 \hat{e}_{it}) 1[\{\hat{\mathbf{z}}_i, \hat{e}_{it}\} \in \hat{A}(\mathbf{x})] \text{ and } \hat{P}(\mathbf{x})
\]

\[
= \frac{1}{N T} \sum_{i,t} 1[\{\hat{\mathbf{z}}_i, \hat{e}_{it}\} \not\in \hat{A}(\mathbf{x})], \text{respectively.} \tag{2.17}
\]

Now that we can estimate \(\hat{G}(.)\) and \(P(.)\) at any \(\mathbf{x}\), the bounds on APE in (2.16), too, can be computed.

In Appendix 3 of the Online Supplemental Appendix, we derive the asymptotic covariance matrix of the second-stage coefficient estimates when the first stage estimation involves estimating a system of regression using the method in Biørn (2004). Given the covariance matrix of the second-stage coefficient estimates, we also derive the confidence intervals (CIs) proposed in Imbens and Manski (2004) for the partially identified APEs.

However, first, because the expressions needed to compute the covariance matrices might be computationally involved, and second, because new expressions for the covariance matrix of the second-stage coefficient estimates will have to be derived when a different estimator for the first stage reduced form is employed, we suggest that bootstrapping procedure be employed to approximate the variance of the estimated coefficient. To obtain bootstrap standard errors for control function methods, both parts of the estimation are included for every bootstrap sample (see Wooldridge 2015), where resampling, as in PW, can be done at the level of cross-sectional unit.

### 3. Monte Carlo experiments

In this section, we discuss the results of the Monte Carlo (MC) experiments, which we conduct to analyze the finite sample behavior of our model and compare the estimates of APEs from ours and alternative estimators to the true measures of the APEs. Since we want to compare the performance of our estimator to the performances of alternative estimators with setups similar to ours, such as that of PW’s, which has a single endogenous regressor, we first conduct the simulation exercise with one endogenous variable, \(x\). And since our method allows for multiple endogenous regressors, we also experiment with two endogenous regressors, \((x_1, x_2)\).

In the first simulation exercise, we consider the following data generating process (DGP):

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

\[
x_{it} = \pi z_{it} + \xi_t + \epsilon_{it}, i = 1, ..., n, t = 1, ..., 5, \tag{3.2}
\]

and where \(z_{it}\) is the instrument. We assume that \(\varphi = -1\) and that \(\pi = 1.5\). We allow the individual specific effects \(\xi_i\) and \(\theta_i\) to be correlated with the vector of instruments, \(Z_i = (z_{i1}, ..., z_{i5})'\). The
$z_t$'s are i.i.d and marginally distributed as $N[0, \sigma_z^2]$, where $\sigma_z = 5$. The variables, $Z_t$, $x_t$ and $\theta_t$ are drawn from the following distribution: $(Z_t', x_t', \theta_t') \sim N[0, \Sigma_{x\theta\theta}]$, where $\sigma_x = 3$, $\sigma_\theta = 4$, $\rho_{xx} = 0.4$, $\rho_{x\theta} = 0.2$ and $\rho_{\theta\theta} = 0.5$. The above choice of correlation coefficients ensures that, conditional on $x_t$, the conditional correlation between $z_t$ and $\theta_t$, $\rho_{z\theta|x} = \rho_{z\theta} - \rho_{x\theta} \rho_{\theta\theta} = 0$, which, in this case, also implies that conditional on $x_t$, $\theta_t \perp \!\!\!\perp Z_t | x_t$.

In accordance with Assumption AS 1, we assume that $(\xi_t, \epsilon_t) \perp (Z_t', x_t, \theta_t)'$ and draw $(\zeta_t, \epsilon_t)$ from $N[0, \Sigma_{\xi\epsilon}]$, where the elements $\sigma^2_{\xi}, \sigma^2_{\epsilon}$, and $\rho_{\epsilon\xi}$ of $\Sigma_{\xi\epsilon}$ are assumed as $\sigma^2_{\xi} = \sigma^2_{\epsilon} = 1$ and $\rho_{\epsilon\xi} = 0.75$. The DGP assumptions, $\theta_t \perp \!\!\!\perp Z_t | x_t$ and $(Z_t', x_t, \theta_t)' \perp (\xi_t, \epsilon_t)$, together satisfy Assumption 2 (a),\footnote{Given the DGP assumptions, it can be verified that $\xi_t \perp \!\!\!\perp Z_t | x_t, \epsilon_t$ and $\theta_t \perp \!\!\!\perp Z_t | x_t, \epsilon_t$; these two then imply that $\theta_t, \zeta_t \perp \!\!\!\perp x_t$ or equivalently $\theta_t, \zeta_t \perp \!\!\!\perp X_t | x_t, \epsilon_t$.} and since $(\zeta_t, \epsilon_t)$ are i.i.d., they also satisfy Assumption 2 (b).

From this DGP, we generate $(Z_t', x_t, \theta_t)'$ and $(\zeta_t, \epsilon_t)$ of varying size, $n$, with $t$ fixed at $t = 5$. We then discretized $z_t$ to take value 1 if $z_t > 0$ and 0 otherwise. Having generated $(Z_t', x_t, \theta_t)'$ and $(\zeta_t, \epsilon_t)$, we generate $x_t$ according to (3.2) and then $y_t$ according to (3.1).

We have showed that by estimating

$$y_{it} = 1\{\varphi x_{it} + \varphi_x \hat{x}_i + \epsilon_{it} + \eta_{it} > 0\}, \quad (3.3)$$

where $\hat{x}_i$ and $\epsilon_{it}$ are the control variables, as a probit model we can obtain consistent estimates of the ASF and the APE. The control variables, $\hat{x}_i$ and $\epsilon_{it}$, are obtained from the estimates of the reduced form Equation (3.2), augmented with $\pi z_i = \pi T^{-1} \sum_{t=1}^T z_{it}$, which is estimated as a random effect model by the method of MLE. The APE at $\hat{x}$ after estimating Equation (3.3) could be obtained by averaging

$$\frac{1}{\Delta x} \Phi \left( \frac{\varphi x_{it} + \varphi_x \hat{x}_i + \epsilon_{it}}{\sigma_{\eta}} \right) \quad (3.4)$$

over $\hat{x}$ and $\hat{\epsilon}$ at $x_{it} = \hat{x} + \Delta x$ and $x_{it} = \hat{x}$ and taking the difference. In the above, $\sigma_{\eta}^2$ is the variance of $\eta_{it}$ in Equation (3.3) and $\Phi$ is the standard cumulative normal density function.

Now, while in practice the heterogeneity terms, $(\theta_t, \zeta_t)$ and $(x_t, \epsilon_t)$, are unobserved, in MC experiments we do know what these values are. By averaging $1\{x_{it} \varphi + \varphi_x \hat{x}_i + \epsilon_{it} > 0\}$ over $(\theta_t, \zeta_t)$ at $x_{it} = \hat{x}$ to obtain $G(\hat{x})$ and the same at $x_{it} = \hat{x} + \Delta x$ to obtain $G(\hat{x} + \Delta x)$, we could compute the true measure of APE, $\frac{\partial G(x_t)}{\partial x}$, at $x_t = \hat{x}$ by computing $\frac{G(\hat{x} + \Delta x) - G(\hat{x})}{\Delta x}$. For the exercise, we chose $\hat{x} = 1$ and $\Delta x = 0.05$. Since we average over realizations of $(\theta_t, \zeta_t)$, there is some variability in the values of $\frac{\partial G(x)}{\partial x}$ over the replications; the average over the replications for every sample size is reported in the tables containing the results. For notational convenience we will denote the true APE by $\frac{\partial G(x)}{\partial x}$. Estimates of APE from any of the model considered in this section will be denoted by $\frac{\partial G(x)}{\partial x}$.

One of the alternative estimators, which has its setup similar to ours is the method proposed by PW. To address the issue of endogeneity, PW also propose a two-step control function method. They first assume that $\theta_t = E(\theta_t | Z_t) + \tau_i = \pi_0 \bar{z}_i + \tau_i$ and $x_t = E(x_t | Z_t) + a_t = \bar{z}_t \hat{x}_i + a_t$, where $\bar{z}_t = T^{-1} \sum_{t=1}^T z_{it}$. Given the assumptions, they write the triangular system in (3.1) and (3.2) as

$$y_{it} = 1\{\varphi x_{it} + \pi_0 \bar{z}_i + \tau_i + \epsilon_{it} > 0\} \quad (3.5)$$

$$x_{it} = \pi \bar{z}_i + \bar{z}_t \hat{x}_i + v_{PWit}, \quad (3.6)$$

where $v_{PWit} = a_i + \epsilon_{it}$. They then make the control function assumption that $\tau_i + \epsilon_{it} \perp \!\!\!\!\perp x_{it}$; this allows them to estimate the APE at $x_t = \hat{x}$ by averaging
over $\pi_0 \tilde{z}$ and $v_{pw}$ at $x_{it} = \tilde{x} + \Delta x$ and $x_{it} = \tilde{x}$ taking the difference. In the above, $\rho$ is population regression coefficient of $\tau_i + \zeta_{it}$ on $v_{PWit}$, and where $v_{PWit}$ is obtained as residuals after estimating (3.6) in the first stage. The conditional distribution of $\tau_i + \zeta_{it}$ given $v_{PWit}$ is assumed to follow a normal distribution with variance $\sigma_{pw}^2$. If their method gives consistent estimates of APE, then it must be that the above measure is equal to $\frac{\partial G(x_i)}{\partial x}$.

In correlated random effects (CRE) probit and in conditional logit (CL) models in Chamberlain (1984), $x_{it}$ is assumed to be independent of the idiosyncratic term, $\zeta_{it}$. While in CRE probit model $E(\theta_i|X_i)$ is specified, in the CL model the distribution of $\theta_i$ is left unspecified. Assuming that $\theta_i = \pi_0 \tilde{x}_i + \tau_i$, where $\pi_0 \tilde{x}_i$ is the specification for $E(\theta_i|X_i)$, the structural equation for the CRE probit model is given by

$$y_{it} = 1\{\varphi x_{it} + \pi_0 \tilde{x}_i + \tau_i + \zeta_{it} > 0\},$$

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

$$\frac{1}{\Delta x} \left[ \Phi \left( \frac{\varphi x_{it} + \pi_0 \tilde{x}_i}{\sigma_{CRE}} \right) dF(\pi_0 \tilde{x}) - \Phi \left( \frac{\varphi x_{it} + \pi_0 \tilde{x}_i}{\sigma_{CRE}} \right) dF(\pi_0 \tilde{x}) \right] = \frac{\partial G(x_i)}{\partial x},$$

where the LHS is the measure of APE of $x$ at $x_{it}$ pertaining to the CRE probit model.

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

$$\frac{1}{\Delta x} \left[ \Lambda(x_{it} + \Delta x, \theta_i) dF(\theta) - \Lambda(x_{it}, \theta_i) dF(\theta) \right],$$

where $\Lambda(x_{it}, \theta_i) = \Pr(y_{it} = 1|x_{it}, \theta_i) = \frac{\exp \left( \varphi x_{it} + \theta_i \right)}{1 + \exp \left( \varphi x_{it} + \theta_i \right)}$. Once we have estimated $\varphi$ by estimating the CL model, we can estimate the APE by averaging $\Lambda(x_{it} + \Delta x, \theta_i)$ and $\Lambda(x_{it}, \theta_i)$ over $\theta_i$ and taking the difference.

Table B.1 provides the results for various sample size, $n$, with $m = 2000$ Monte Carlo replications. In the Table and in Fig. C.1 we compare the performance of our method, which we term CRECF method, to the alternative estimators considered above.

In Fig. C.1, we plot the densities of $m = 2000$ MC estimates of $\partial G(x_{it})/\partial x - \partial G(x_{it})/\partial x$, at $x_{it} = 1$, obtained for four estimation methods for different sample sizes. It can be seen from the figure that for each of the alternative estimators, the APE of $x$ is estimated with a bias, which persists as the sample size grows larger. Thus, even as the variance of $\partial G(x_{it})/\partial x - \partial G(x_{it})/\partial x$ for each of the alternative methods decreases, the root mean square error (RMSE) for alternative methods in Table B.1 decreases quite slowly.

Since the CL and CRE probit models do not account for the endogeneity of $x_{it}$, with respect to the transitory errors, $\xi_{it}$, the methods can give biased results. Unexpectedly, however, the method proposed by PW, which tries to accounts for correlation between $Z_i$ and $\theta_i$, and the correlation of $x_{it}$ with both $\theta_i$ and $\xi_{it}$ gives the least satisfactory results. This suggests that under a more general DGP, as in our MC experiments, their control function assumption that $\tau_i + \zeta_{it} \perp X_i|v_{PWit}$, where both $\tau_i + \zeta_{it}$ and $v_{PWit}$ are assumed to be independent of $Z_i$, is, as discussed in Remark 3 and Proposition 1, likely to get violated.

---

13The acronym derives from fact that the control functions are based on correlated random effects in the reduced form equations.
To validate the claims made in Remark 3, we conduct some MC experiments to compare the APE when \( V_{PW,i} = \{v_{PW,1}, ..., v_{PW,n}\} \) is used as a control function as against when only \( v_{PW,i} \) is used a control function. In Fig. C.2 below, we have plotted the density of difference between the estimated and the true APEs, \( \partial G(x_i)/\partial x - \partial G(x_i^0)/\partial x \), where the estimated APEs are the APEs that are obtained by varying the control functions and the instruments in PW’s model. As can be seen in figure, when the instrument is continuous with a large support, employing \( V_{PW,i} \) yields consistent estimates of the APE, whereas if only \( v_{PW,i} \) is employed as control function, then, we get estimates that are biased. When the same instrument is discretized to take value 1 and 0, we get biased estimates when either \( V_{PW,i} \) or \( v_{PW,i} \) is employed as control function. This is because when the instrument is discrete, even though \( V_{PW,i} \) is the appropriate control function, the APE is not point but partially identified.

That the APEs when using the traditional control function, as in PW’s, may not be point identified when the instrument, \( z_{it} \), is binary even when \( x_{it} \) has a large support can be seen in Fig. C.3(c), which has the plots of the level sets of the kernel estimates of the joint density of \( (x, v_{PWit}) \). The figure suggests that the common support requirement for point identification may be satisfied only over a small range of \( x \) values in PW’s model. Whereas from Fig. C.3(a,b), we can deduce that the support of the conditional distribution of \( (\hat{e}_{it}, \hat{z}_i) \) given \( x \) is almost the same for large ranges of \( x \).

As our method allows for multiple endogenous regressors, we also conduct a simulation exercise with two endogenous regressors. The two instruments, \( z_{it} = (z_{1it}, z_{2it}) \), are i.i.d and marginally distributed as \( N[0, \sigma_z^2] \), where \( \sigma_z = 5, \sigma_z = 2 \) and \( \rho_{z1z2} = 0.25 \). The instruments, \( Z_i = \{z_{1i}, ..., z_{5i}\} \) and the individual effects are correlated, and follow the joint distribution: \( (Z_0, x_1, x_2, \theta_i)^T \sim N[0, \Sigma] \), where \( \sigma_1 = 6, \sigma_2 = 2, \sigma_4 = 4, \rho_{z1z2} = 0.2, \rho_{z1z1} = 0.3, \rho_{z2z2} = 0.25, \rho_{z2z4} = 0.3, \rho_{z2z2} = 0.5, \rho_{z2\theta} = 0.1, \rho_{z4\theta} = 0.15, \rho_{z2\theta} = 0.25 \). The above choice of correlation coefficients ensures that, conditional on \( x_i = (x_{z1}, x_{z2}) \), the conditional correlation between \( z_{it} \) and \( \theta_i \) is 0. Having generated the data, we then discretize the instruments to take values 0 and 1: \( z_{1it} \) takes value 1 if it is non-negative and 0 otherwise, while \( z_{2it} \) takes value 1 if it is greater than or equal 1 and 0 otherwise. The idiosyncratic error terms \( (\varepsilon_{it}, \varepsilon_{it}, \varepsilon_{it}) \) are drawn from \( N[0, \Sigma] \), where the elements of \( \Sigma \) are assumed as \( \sigma_i = \sigma_1 = \sigma_2 = 1, \rho_{\varepsilon_1} = 0.75, \rho_{\varepsilon_2} = 0.25 \) and \( \rho_{\varepsilon_1\varepsilon_2} = 0.5 \).

With \( Z_i \) and the error terms in place, we next generate \( x_{1it}, x_{2it} \) and \( y_{it} \) according to:

\[
\begin{align*}
 x_{1it} &= -1z_{1it} + 0.05z_{2it} + x_{1i} + \varepsilon_{1it} \\
 x_{2it} &= 0.025z_{1it} + 0.75z_{2it} + x_{2i} + \varepsilon_{2it} \\
 y_{it} &= 1\{-1x_{1it} + 0.5x_{2it} + \theta_i + \zeta_{it} > 0\}.
\end{align*}
\]

We compute the APEs at \( \bar{x}_1 = 0.5, \bar{x}_2 = 1 \) and chose \( \Delta x_1 = 0.05 \) and \( \Delta x_2 = 0.1 \). Table B.2 provides the results for various sample size, \( n \), with \( m = 2000 \) Monte Carlo replications. In the Table, we compare the performance of our method to the CRE probit and the conditional logit models. Since PW consider only a single endogenous regressor, their model is not considered in these simulations. With lowest RMSE for every sample size, our method outperforms the CRE probit and the conditional logit models.

The results, therefore, imply that assuming \( (\varepsilon_{it}, \hat{z}_i) \) as control function for identifying the ASF and APE, may not be restrictive, and that the developed method can yield consistent result.

To conclude, this finite sample study establishes the following:

1. Our method performs well with sample sizes frequently encountered in practice.
2. It performs better than the alternative estimators with setups similar to ours.
3. Employing \( V_{PW} \), instead of \( v_{PW} \), as a control function yielded consistent APEs when the instrument had a large support. This suggest that when \( Z \) is independent of the error terms
and the instruments, $z_t$, have a large support, then employing $V$ as a control function could yield consistent APEs.

4. Implications of ownership of land and farm assets on child labor

4.1. Introduction

Child labor is a pressing concern in all developing countries. According to International Labour Office’s 2016 estimates, worldwide, 152 million children in the age group of 5 to 17 years age group are victims of child labor; 62 million of which are in the Asia-Pacific region. Conditions of child labor can vary. Many children work in hazardous industries that take a toll on their health. Moreover, when children work, they forego education and 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) (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 (Basu et al. 1998 & BR), along with imperfection in labor and land market (Basu et al. 2010; Bhalotra and Heady 2003; Dumas 2006) and capital market (BR) to be the major causes of child labor. BR show that child labor increases when endowments of parents are low, and that when capital market imperfections exist and parents cannot borrow, child labor becomes inefficiently high.

Basu et al. (2010) (BDD) point out that papers like Bhalotra and Heady (2003) (BHy) and Dumas (2006) show that in some developing countries the amount of work the children of a household do increases with the amount of land possessed by the household. Since land is usually strongly correlated with a household’s income, this finding seems to challenge the presumption that child labor involves the poorest households. They argue that these perverse findings are a facet of labor and land market imperfections, and that in developing countries, poor households in order to escape poverty want to send their children to work but are unable to do so because they have no access to labor markets close to their home. In such a situation, if the household comes to acquire some wealth, say land, its children, if only to escape penury, will start working. However, if the household’s land ownership continues to rise, then beyond a point the household will be well-off enough and it will not want to make its children work.

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

Cockburn and Dostie (2007) (CD) in their analysis of child labor in Ethiopia find that in presence of labor market imperfections, all assets need not be child labor enhancing. They find that certain productive assets that enable an increase in the total family income may not necessarily increase child labor. They show that assets such as oxen and plows that are operated by adults decrease child labor. To test this hypothesis, in our empirical specification we include an index of productive farm assets.
Now, while land and labor market imperfections may exist in developing countries, the extent of imperfection may not be uniform across all countries, or regions within a country. Hence, the relationship between child labor and different kinds of assets, such as landholding or agrarian assets, is an empirical question. The question is important because policy implications could be different under different relationships between various kinds of assets and child labor. For example, if one were to confirm the findings in BHy and BDD, then if monetary transfers are used to increase landholding or land redistribution is done in favor of the poor, child labor may in fact increase. On the other hand, when monetary transfers are used to increase agrarian assets, then there is an inverse relationship between agrarian assets and child labor holds, such transfers could reduce the incidence of child labor.

In our data, we find nonagricultural income to be much higher than agricultural income (Table B.4). This suggests that land is not the only source of income as in BHy and BDD. BDD assume that land is the only source of income and derive a regression equation where household income is left out. Since nonagricultural income constitutes a major portion of total household income, we also control for household income.

We also find that, over time, land size distribution has become more unequal (Table B.4), which indicates that land market exists in the regions from where the data has been collected. Now, if land market exists, even if imperfect, then it is unlikely that land owned by households will be exogenous to a household’s labor supply as in BHy and BDD, where land is mainly inherited, but endogenously determined along with household’s, including children’s, labor supply decisions. However, endogeneity could also arise due to omitted variables. To account for the endogeneity of landholding along with that of productive assets and household income, we employ the method developed in the article.

4.2. Data and empirical model

4.2.1. Data

We conduct our empirical analysis at the level of the child using two waves, 2006–2007 and 2009–2010, of the data from Young Lives Study (YLS), a panel study from six districts of the state of United 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. Finally, excluding children for whom relevant information was missing, we were left with 2458 children. Table B.3 and Table B.4 have the relevant descriptive statistics.

Children were asked how much time they spent in the reference period (a typical day in the last week) doing (a) wage labor, (b) non-wage labor or (c) domestic chores. If the answer was positive number of hours for any one of the activities, then, the binary variable $D_{WORK}$ was assigned value 1, 0 otherwise. The major component of work (not reported here) is due to domestic chores. While both domestic and non-domestic work registered increase over the years, the increase in the proportion of children doing non-domestic work was higher.

In Table B.4, we find that while land ownership has become more unequal, the average size of land owned increased over the years. Farming Asset Index, which too increased over the years, was 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.

---

14In 2014 the north-western portion of the then Andhra Pradesh was separated to form the new state of Telangana.

15Wage labor involves activities for pay, work done for money outside of household, or work done for someone not a part of 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.

16Farming assets constitute of agriculture tools, carts, pesticide pumps, plows, water pumps, threshers, tractors and other farm equipments.
4.2.2. Empirical model

Let $y_{it}$ be the binary variable that takes value 1 if the parents of the child $i$ decide that the child works and 0 otherwise. The decision is modeled as in Equation (2.1), where $y_{it}^{a}$ is amount of time devoted to work by child $i$ in period $t$. The set of endogenous variables, $x_{it}$, include income ($INCOME_{it}$) of the household to which the child $i$ belongs, size of the land holdings ($LAND_{it}$), and the index of productive farm assets ($ASSET_{it}$).

To address the issues of endogeneity and heterogeneity, we employ the two-step control function methodology developed in the article, where the control functions,

$$\tilde{z}'_{i} = (\tilde{z}_{INCOME_{it}}, \tilde{z}_{LAND_{it}}, \tilde{z}_{ASSET_{it}})$$

and

$$\tilde{\epsilon}'_{it} = (\tilde{\epsilon}_{INCOME_{it}}, \tilde{\epsilon}_{LAND_{it}}, \tilde{\epsilon}_{ASSET_{it}}),$$

are obtained from the estimates of the first stage reduced form Equations (2.7). After augmenting the structural Equation (2.1) with the control functions, we get the modified structural Equation (2.9), which we estimate as a probit model.17

To identify the impact of the endogenous variables on the parents’ decision to make their children work, we employ the following instruments: (1) NREGS, which is the total sanctioned amount at the mandal (region) level at the beginning of financial year (in 2008–2009 prices) to support an employment guarantee scheme; (2) CASTE, caste (social group) of the child; and (3) a set of four indicator variables that capture the level of infrastructural development in the household’s locality/settlement.

The National Rural Employment Guarantee Scheme (NREGS) was initiated in 2006 by the Government of India with the objective to alleviate rural poverty. NREGS legally entitles rural households to 100 days of employment in unskilled manual labor (on public work projects) at a prefixed wage. Now, it can be seen in Table B.4 that over the years, 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. (2016) claiming NREGS to be a valid instrument for income, argue that since fund sanctioned at the beginning of the financial year is not be affected by current demand for work, the funds sanctioned is exogenous and more funds imply more work opportunity in NREGS, which can have a positive effect on household income. Also, the total fund allocation to NREGS increased during the period 2007–2010. However, this increase was not uniform across the 15 mandals.18

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.19 Fairing

---

17For many children, as we know, the optimal choice of $y_{it}^{a}$ is the corner solution, $y_{it}^{a} = 0$. For corner solution outcomes, we are interested in features of the distribution such as $\int Pr(y_{it}^{a} > 0|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it})dF(\tilde{z}, \tilde{\epsilon})$ and $\int E(y_{it}^{a}|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it})dF(\tilde{z}, \tilde{\epsilon})$, where $X_{it} = (x_{it}, w_{it})'$ and $E(y_{it}^{a}|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it}) = Pr(y_{it}^{a} = 0|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it})0 + Pr(y_{it}^{a} > 0|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it})E(y_{it}^{a}|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it}, y_{it}^{a} > 0)$.

18Due to lack of space, in this application we study only $\int Pr(y_{it}^{a} > 0|x_{it}, \tilde{z}_{it}, \tilde{\epsilon}_{it})dF(\tilde{z}, \tilde{\epsilon})$.

19Data on the sanctioned funds at the mandal level was obtained from the Andhra Pradesh Government’s website on NREGS (http://nrega.ap.gov.in/).

In fact, this stratification was so endemic that the constitution of India aggregated these castes into a schedule of the constitution and provided them with affirmative action cover in both education and public sector employment. This constitutional initiative was viewed as a key component of attaining the goal of raising the social and economic status of the SC/STs to the levels of the non-SC/STs.
better than SC/ST’s are those belonging to the “Other Backward Classes” (OBC). 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 belongs to group labeled as “Others” (OT).

We claim that CASTE is a valid instrument for landholding because, although average wealth and income are evidently distributed along caste lines, we do not find a significant variation in child labor or school enrollment across caste or social group to which the child belongs (Table B.5). In other words, no social group is inherently disposed to make their children work or send them to school. This could be because rising awareness, overtime, about returns from education persuades families of all castes to send their children to school. We find support for the assertion in the literature too. Hnatkovska et al. (2012) find significant convergence in the education attainment levels and occupation choice of SC/ST’s and non-SC/ST’s between 1983 and 2004–2005; moreover, the convergence in education level has been highest for the youngest cohort. Second, time-invariant ethnicity variable such as caste cannot be correlated with unobserved time-invariant heterogeneity such as parents’ or children’s abilities and land quality.

Our assertion that the preferences of parents regarding child labor and schooling does not differ systemically across social groups is 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?” 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 a set of four dummy variables, which indicate (1) if drinkable water is provided in the locality, (2) if the services of a national bank are provided in the locality, (3) if private hospitals exist in the locality and (4) if access to the locality is via an engineered road. As in BHy, these variables, which indicate the level of infrastructure development, are employed to instrument the index of productive farm assets.

4.3. Discussion of results

The results of the first stage reduced form equations in Table B.6 suggest that our instruments are good predictors of the endogenous variables. First, we find that an increase in the amount sanctioned for NREGS projects increases the household income. Second, CASTE does, on an average, correctly predict the economic (income, land holding and assets) status of households. Finally, the dummy variables indicating the level of infrastructure development are positively correlated with the index of productive farm assets.

Before we begin to discuss the result of the second-stage estimation in Table B.7, we state a few points regarding the estimation. (a) The only exogenous explanatory variables in our

---

20The Government of India classifies, a classification based on social and economic conditions, some of its citizen as Other Backward Classes (OBC). The OBC list is dynamic (castes and communities can be added or removed) and is supposed to change from time to time depending on social, educational and economic factors. In the constitution, OBC’s are described as “socially and educationally backward classes,” and government is enjoined to ensure their social and educational development.
parsimonious specification are the age and the sex of the children. (b) The specification includes district dummies, a time dummy and the interaction of the two to account for the fact that the districts to which children belong may have different economic growth trajectories as well as trends related to work and education. The time dummy allows us to control for changes in demand and supply of work over time. (c) Since the support assumption for point identification of the APEs is not met, we estimate the bounds on the APEs and the 95% confidence interval (CI95%) for the partially identified APEs. (d) For the continuous variables, the bounds on the APE of a variable were computed by increasing the variable by one standard deviation from its mean, where the mean and the SD of the variable are from the 2010 data. For age, the bounds on APE were computed by increasing the mean age in 2010 by 1 year. (e) The standard errors of the coefficients were estimated using the analytical expression of the covariance matrix derived in Appendix 3 of the Online Supplemental Appendix.

We begin by comparing the results from Chamberlain’s CRE probit model with the estimates obtained from applying the method developed in this article. The significance of estimated coefficients of the control functions suggests that income, land size and productive farm assets are endogenously determined along with household’s labor supply, including that of the child’s, decisions. When income and wealth are not instrumented, as in the CRE probit, considering the discussion in the article, we get an incorrect sign for the coefficient on income. Moreover, the result of CRE probit suggests that ownership of land and farm assets do not affect child labor, which, given the many recent evidences, is unlikely in a developing country. The results, thus, make clear the importance of accounting for endogeneity of income, landholding and farm asset.

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

Since the upper and lower bounds of the APE of productive farm assets are high and since the CI95% is only marginally bigger than the bound, it seems that ownership of farm assets leads to a significantly high reduction in children’s participation in work. Dumas, BHy and CD argue that an increase in asset holding that increases the marginal productivity of labor induces two opposite effects on labor. While the income effect of increased wealth tends to reduce the labor time, the substitution effect, due to the absence of labor market, provides incentives for work and tends to increase children’s labor time. Our results suggest that the wealth effect of farm assets, which are not likely to be operated by children, dominate to reduce children’s labor time. Second, 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. Finally, we find that older children and boys are more likely to work.

21Although we do not report here, we did not find that nonlinear terms of income, land and productive assets to be significant. We had also included four education related dummy variables, two for the father and two for the mother. The dummy variables for the mother, for example, indicated (1) if the mother had some schooling and (2) if the mother had attended secondary or post-secondary school. The education dummies, although substantially affecting household income, did not seem to affect child labor propensity. This suggests that parents’ education level has had no independent impact except through income.

22In 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, which validates using infrastructure variables as instruments for farming assets.
5. Concluding remarks

The objective of the article has been to develop a method to estimate structural measures of interest such as the APEs for panel data binary response model in a triangular system while accounting for multiple unobserved heterogeneities. The unobserved heterogeneity terms constitute of time invariant random effects/coefficients and idiosyncratic errors. We propose that the expected values – conditional on the histories of the endogenous variables, \( X_t \equiv (x'_{t1}, \ldots, x'_{tT})' \), and the exogenous variables, \( Z_t \equiv (z'_{t1}, \ldots, z'_{tT})' \) – of the heterogeneity terms be used as control functions (CF).

The proposed method makes a number of interesting contributions to the literature. First, among the class of triangular system with imposed structures similar to ours, the proposed CF method requires weaker restrictions than the traditional control function methods. Second, when instruments have a small support, the CFs, which exploit panel data, help in point-identifying structural measures such as the APEs when the endogenous variables have a large support. Bounds on the structural measures are provided when the support assumption is not satisfied. Third, the method allows for multiple endogenous variables, all of which are determined simultaneously. Finally, in an equivalence result we showed that for linear panel data models, the resulting estimates are equivalent to the ones that are obtained when the structural model is estimated by a certain two-stage least squares. Also, Monte Carlo experiments show that compared to alternative panel data binary choice models similar to ours, our method performs better.

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

Finally, we would like to note that (i) extension of the methodology for estimating dynamic binary choice models and (ii) identification and estimation the proposed control functions without making distributional assumptions about the heterogeneity terms of the reduced form equations would be important contributions to the literature.

Appendix A

Proofs

Lemma 2.

(a) Let \( X \equiv (x_1, \ldots, x_T)' \) and \( Z \equiv (z_1, \ldots, z_T) \). If \( x_t \) is specified as

\[
x_t = \pi z + \eta + a + \epsilon_t, t \in \{1, \ldots, T\}, \tag{A.1}
\]

where \( \eta = \frac{1}{T} \sum_{t=1}^T z_t \) and \( a \sim N(0, \Lambda) \), and \( a \) is i.i.d., then

\[
E(x_t | X, Z) = \pi z + E(a | X, Z) = \pi z + \Omega \Sigma_{\epsilon_t}^{-1} \sum_{t=1}^T (x_t - \pi z_t - \bar{\eta}),
\]

where \( \Omega = [T \Sigma_{\epsilon_t}^{-1} + \Lambda]^{-1} \) is the conditional variance of \( a \) given \( X \) and \( Z \).

(b) Suppose we have a single endogenous variable, \( x_t \). Let \( X \equiv (x_1, \ldots, x_T) \) and define \( Z \equiv (z_1, \ldots, z_T) \). Suppose \( x_t \) is given by

\[
x_t = \pi z + \eta + a + \epsilon_t, t = 1, \ldots, T, \tag{A.2}
\]
where \( \z = \frac{1}{T} \sum_{t=1}^{T} z_t \). If the errors, \( \epsilon \equiv (\epsilon_1, ..., \epsilon_T) \), are normally distributed with mean 0 and are nonspherical such that \( E(\epsilon^e \epsilon^e)' = \Omega_{\epsilon t} \), then

\[
E(\epsilon^e X, Z) = \hat{\epsilon}(X, Z) = (x_1 - \pi z_1 - \pi \z)' \alpha _1 + ... + (x_T - \pi z_T - \pi \z)' \omega _T,
\]

where \( (\omega _1, ..., \omega _T)' = \frac{\Omega_{\epsilon t}^{-T} \pi}{(\pi^T, \Omega_{\epsilon t}^{-1} + \pi^2)} \) and \( e_T \) is a vector of ones of dimension \( T \).

**Proof 1.**

(a) To obtain \( \hat{\epsilon}(X, Z) = E(\epsilon|X, Z) \), we first derive \( f(\epsilon|X, Z) \), the conditional density function of \( \epsilon \) given \( X \) and \( Z \). By Bayes’ rule we have

\[
f(\epsilon|X, Z) = \frac{f(X, Z|\epsilon)f(\epsilon)}{f(X, Z)},
\]

where the last equality is obtained because \( Z \) is independent of the residual individual effects, \( a_i \); that is, \( f(Z|a) = f(Z) \).

Since \( a_i \perp \epsilon_i \), \( \epsilon_i \) is i.i.d., \( a_i \sim N(0, \Lambda_{\epsilon t}) \) and \( \epsilon_i \sim N(0, \Sigma_\epsilon) \), then, given (A.1), it implies that \( X \), given \( Z \), is normally distributed with mean \( (\pi z_1 + \pi \z, ..., (\pi z_T + \pi \z))' \), and variance \( \Sigma = I_T \otimes \Sigma_\epsilon + E_T \otimes \Lambda_{\epsilon t} \), where \( I_T \) is an identity matrix of dimension \( T \) and \( E_T \) is a \( T \times T \) matrix of ones. That is, \( f(X|Z) \) in (A.3) is given by

\[
f(X|Z) = \frac{1}{\sqrt{(2\pi)^m T |\Sigma|}} \exp \left[ -\frac{1}{2} (R - \pi^{2} \Sigma^{-1} R) \right],
\]

where \( R = X - \begin{bmatrix} \pi z_1 + \pi \z \\ \vdots \\ \pi z_T + \pi \z \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ r_T \end{bmatrix} \)

and \( m = d_x \) is the dimension of \( x \). Since \( \operatorname{rank}(E_T) = 1 \), we can use example 5 in Miller (1981), which is on the inverse of sum of two Kronecker products, to obtain

\[
\Sigma^{-1} = I_T \otimes \Sigma_{\epsilon t}^{-1} - E_T \otimes [\Sigma_\epsilon + \operatorname{tr}(E_T) \Lambda_{\epsilon t}]^{-1} \Lambda_{\epsilon t} \Sigma_{\epsilon t}^{-1} \quad \text{and} \quad |\Sigma| = |\Sigma_{\epsilon t}|^{(T-1)} |\Sigma_\epsilon + \operatorname{tr}(E_T) \Lambda_{\epsilon t}| \quad \text{where} \quad \operatorname{tr}(E_T) = T,
\]

which allows us to write \( f(X|Z) \) as

\[
f(X|Z) = \frac{1}{\sqrt{(2\pi)^{mT} T^{T-1} |\Sigma_{\epsilon t}|}} \times \exp \left[ -\frac{1}{2} \left( R' [I_T \otimes \Sigma_{\epsilon t}^{-1}] R - \sum_{t=1}^{T} r_t' [\Sigma_\epsilon + T \Lambda_{\epsilon t}]^{-1} \Lambda_{\epsilon t} \Sigma_{\epsilon t}^{-1} \sum_{t=1}^{T} r_t \right) \right].
\]

Since \( f(X|Z, a) = f((\epsilon_1', ..., \epsilon_T')') \), \( \epsilon_i \)'s are i.i.d., \( \epsilon_i \sim N(0, \Sigma_\epsilon) \) and \( a \sim N(0, \Lambda_{\epsilon t}) \), \( f(X|Z, a)f(a) \) in (A.3) is

\[
f(X|Z, a)f(a) = \frac{1}{\sqrt{(2\pi)^{mT + m} T^{T-1} |\Lambda_{\epsilon t}|}} \times \exp \left[ -\frac{1}{2} \left( R - e_T \otimes a \right)' [I_T \otimes \Sigma_{\epsilon t}]^{-1} \left( R - e_T \otimes a \right) + a' \Lambda_{\epsilon t}^{-1} a \right],
\]

where \( R - e_T \otimes a = (\epsilon_1', ..., \epsilon_T')' \), \( e_T \) being vector of ones of dimension \( T \).

The following matrix results,

1. \( A_{m \times m} \otimes B_{n \times n}^{-1} = A_{m \times m}^{-1} \otimes B_{n \times n}^{-1} \),
2. \( (A_{p \times q} \otimes B_{r \times s}) (C_{q \times k} \otimes D_{s \times t}) = A_{p \times q} C_{q \times k} B_{r \times s} D_{s \times t} \) and
3. \( e_T^T I_T e_T = T \),

allow us to write the expression in the square parenthesis in (A.5) as

\[
R' [I_T \otimes \Sigma_{\epsilon t}^{-1}] R - a' \Sigma_{\epsilon t}^{-1} \sum_{t=1}^{T} r_t - \sum_{t=1}^{T} r_t^' \Sigma_{\epsilon t}^{-1} a + a' \left[ \Lambda_{\epsilon t}^{-1} + T \Sigma_{\epsilon t}^{-1} \right] a.
\]
Using the results in (A.4), (A.5) and (A.6) and the result that $|A^{-1}| = |A|^{-1}$, if $A$ is nonsingular, we get

$$f(a|X, Z) = \frac{f(X|Z, a)f(a)}{f(X|Z)} = \frac{1}{\sqrt{(2\pi)^m|\Sigma_e|}} \exp \left( -\frac{1}{2} \left[ a - \Omega \Sigma_e^{-1} \sum_{t=1}^{T} r_t \right] \Omega^{-1} \left[ a - \Omega \Sigma_e^{-1} \sum_{t=1}^{T} r_t \right] \right).$$

Let $[\Lambda_{z2}^{-1} + T\Sigma_e^{-1}] = \Omega^{-1}$, then, $\sum_{t=1}^{T} r_t [\Sigma_e + TA_{z2}]^{-1} \Lambda_{z2} \Sigma_e^{-1} \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} r_t \Sigma_e^{-1} \Omega \Omega^{-1} \sum_{t=1}^{T} r_t$ and $\sum_{t=1}^{T} r_t \Sigma_e^{-1}$, in (A.7), after a few matrix manipulations, can be written as

$$\sum_{t=1}^{T} r_t [\Sigma_e + TA_{z2}]^{-1} \Lambda_{z2} \Sigma_e^{-1} \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} r_t \Sigma_e^{-1} \Omega \Omega^{-1} \sum_{t=1}^{T} r_t \text{ and}$$

$$\sum_{t=1}^{T} r_t \Sigma_e^{-1} = \sum_{t=1}^{T} r_t \Sigma_e^{-1} \Omega \Omega^{-1}, \text{respectively.}$$

Given (A.8) and (A.9), we can write $f(a|X, Z)$ in (A.7) as

$$f(a|X, Z) = \frac{1}{\sqrt{(2\pi)^m|\Omega|}} \exp \left( -\frac{1}{2} \left[ a - \Omega \Sigma_e^{-1} \sum_{t=1}^{T} r_t \right] \Omega^{-1} \left[ a - \Omega \Sigma_e^{-1} \sum_{t=1}^{T} r_t \right] \right).$$

In other words, $a$, given $X$ and $Z$, is normally distributed with conditional mean

$$E(a|X, Z) = \hat{a}(X, Z) = \Omega \Sigma_e^{-1} \sum_{t=1}^{T} (x_t - \pi z_t - \hat{z}_t)$$

and conditional variance $\Omega = \Sigma_e [\Omega_e + TA_{z2}]^{-1} \Lambda_{z2}$.

(b) While discussing the restrictions imposed on the reduced form equation, we had stated that when $d_e = 1$, the assumption that $\epsilon$ and $\epsilon_t$ are completely independent of $Z$ can be weakened to allow for nonspherical error components. Suppose that $\epsilon, \epsilon_t, t = 1, ..., T$ are serially dependent such that $\epsilon \equiv (\epsilon_1, ..., \epsilon_T)'$ normally distributed with $E(\epsilon') = \Omega_e$ and $a$ is normally distributed and is heteroscedastic as in Baltagi et al. (2010).

To obtain $\hat{a}(X, Z) = E(a|X, Z)$, as in part (a), we first derive $f(a|X, Z)$. Using the fact that $Z \perp a$, by an application of Bayes' rule, as in part (a), Equation (A.3), we have $f(a|X, Z) = \frac{f(X|a, Z)f(a)}{f(X|Z)}$, where $f(a)$ is the normal density function of $a$.

Now, since in (A.2), $a \perp e$, $a \sim N(0, \sigma_a^2)$ and $\epsilon \sim N(0, \Omega_e)$, it implies that $X$, given $Z$, is normally distributed with mean $(\pi z_t + \hat{z}_t), ..., (\pi z_T + \hat{z}_T)'$, and variance $\Sigma = \Omega_e + \sigma_a^2 \epsilon T \epsilon_T$, where $\epsilon_T$ is a vector of ones of dimension $T$. That is,

$$f(X|Z) = \frac{1}{\sqrt{(2\pi)^{T/2} |\Sigma|}} \exp \left( -\frac{1}{2} X \Sigma^{-1} X \right), \text{ where } X = \begin{bmatrix} \pi z_1 + \hat{z}_1 \\ \vdots \\ \pi z_T + \hat{z}_T \end{bmatrix},$$

and where by Sherman–Morrison formula, $\Sigma^{-1} = \Omega^{-1} - \sigma_a^2 \epsilon_T \Omega_e^{-1} \epsilon_T$, and $|\Sigma| = |\Omega_e| (1 + \sigma_a^2 \epsilon_T \Omega_e^{-1} \epsilon_T)$.

Since $X$ given $(Z, a)$ has the same distribution as $\epsilon = R - ae_T$, we have

$$f(X|Z, a)f(a) = \frac{1}{\sqrt{2\pi|\Omega_e| \sigma_a^2}} \exp \left( -\frac{1}{2} \left[ (R - ae_T)' \Omega_e^{-1} (R - ae_T) + \sigma_a^2 \right] \right).$$

Finally, because $f(a|X, Z) = \frac{f(X|Z, a)f(a)}{f(X|Z)}$, using (A.10) and (A.11), it can be shown that a given $X$ and $Z$ is normally distributed with conditional mean

$$E(a|X, Z) = \hat{a}(X, Z) = (x_1 - \pi z_1 - \hat{z}_1) \omega_1 + ... + (x_T - \pi z_T - \hat{z}_T) \omega_T,$$

where $(\omega_1, ..., \omega_T)' = \frac{\Omega_e^{-1} \epsilon_T}{(\sigma_a^2 \epsilon_T \Omega_e^{-1} \epsilon_T + \sigma_z^2)}$, and conditional variance, $\sigma_a^2 \sigma_z^2 (\sigma_a^2 \epsilon_T \Omega_e^{-1} \epsilon_T + 1)^{-1}$. 

A. K. TIWARI
Proposition 1. Let \( Z \iff (\theta, \mathbf{z}, \zeta, \epsilon) \). When \( \theta \) and \( \mathbf{z} \) are correlated and so are \( \zeta \) and \( \epsilon \), then \( \theta + \zeta \iff X|V \) whereas \( \theta + \zeta \iff X|V \) so that \( v_1 = \mathbf{z} + \epsilon_1 = x_1 + \pi z_1 \) and \( V \equiv (v_1, \ldots, v_T) \).

Proof 2. Now, to show that \( r_i = \theta + \zeta_i \iff X|V \), we can show that
\[
E(f(r_i)|X, V) = E(f(r_i)|V),
\]
where \( f \) is real, bounded and measurable function (see proposition 2.3 in Constantinou and Dawid 2017).

Since \( X = \pi Z + V \), there is one-to-one mapping between \( (X, V) \) and \( (\pi Z, V) \), and therefore, the conditioning \( \sigma \)-algebra, \( \sigma(X, V) \), is same as the \( \sigma \)-algebra, \( \sigma(\pi Z, V) \). Hence,
\[
E(f(r_i)|X, V) = E(f(r_i)|\pi Z, V).
\]
(\ref{2.13})

Since \( r_i \) and \( V \) are independent of \( \pi Z \), we get
\[
E(f(r_i)|\pi Z, V) = E(f(r_i)|V),
\]
and therefore, we have
\[
E(f(r_i)|X, V) = E(f(r_i)|V),
\]
which is what we wanted to show.

When the control function is \( v_1 \), we have
\[
E(f(r_i)|X, v_1) = E(f(r_i)|X_{-1}, \mathbf{x}, v_1) = E(f(r_i)|X_{-1}, \pi z_1, v_1) = E(f(r_i)|X_{-1}, \pi z_1, v_1),
\]
where the second equality follows because \( \mathbf{x}_1 = \pi z_1 + v_1 \) and third by the same logic by which we get (\ref{2.13}). Because \( r_i = \theta + \zeta_i \) and \( v_1 = \mathbf{z} + \epsilon_1, s \neq t \), are correlated even after conditioning on \( v_1 \), so, conditional on \( v_1, r_i \) is correlated with \( \mathbf{x}_1 = \pi z_1 + v_1, s \neq t \); that is, \( E(f(r_i)|X_{-1}, \pi z_1, v_1) \neq E(f(r_i)|\pi z_1, v_1) = E(f(r_i)|v_1) \).

Theorem 1. If (i) \( \text{rank}(E(\mathbf{x}_t|x'_t)) = d_\beta \); (ii) \( \text{rank}(\Pi) = d_\alpha \), where \( \Pi = (\pi \quad \pi) \); (iii) \( \text{rank}(E((x'_t, x'_t')(x'_t, x'_t))) = k \)
where \( k = \dim((x'_t, x'_t)) \); and (iv) if Assumption AS 3 holds so that the covariance matrices of \( \mathbf{e}_t \) and \( \mathbf{a} \) are of full rank, then \( \text{rank}(E(\mathbf{x}_t|x'_t)) = 3d_\epsilon \).

Proof 3. Now, condition (i) of the theorem is the “rank condition” for the standard probit model when \( x_t \) is exogenous and the object of interest is \( \phi \) or marginal effects. This condition is assumed to hold true. Similarly, condition (ii) is the rank condition for the identification of the reduced form coefficients, \( \Pi = (\pi \quad \pi) \), which is also assumed to hold.

To begin with, without loss of generality assume that \( \mathbf{z} \) is uncorrelated with the individual effects \( \mathbf{a} \) so that \( \pi \mathbf{z} = 0 \). This implies that we can ignore \( \mathbf{z} \) in the reduced form Equation (A.1) and consider only the dimension of \( \mathbf{z}_t \), which is \( d_\epsilon \), in condition (iii) of the theorem, and that
\[
\Pi = \pi \mathbf{a} = \pi \mathbf{e}_t = x_t - \pi z_t - \pi \mathbf{a}, k = d_\epsilon \text{ and } \pi \mathbf{a} \text{ is a } 3d_\epsilon \times d_\epsilon \text{ matrix.}
\]

Since, \( e_t = v_t - \mathbf{a}_t \), then if \( E(\mathbf{x}_t|x'_t) \), where \( x'_t = (x'_t, e'_t, \mathbf{a}'_t)' \), were to be invertible (or equivalently have a rank of \( 3d_\epsilon \)) so, too, would
\[
E \begin{bmatrix} x_t \\ v_t \\ \mathbf{a} \end{bmatrix} \begin{bmatrix} x'_t \\ v'_t \\ \mathbf{a}'_t \end{bmatrix} \] be.

This is equivalent to stating that the columns of \( (x'_t, v'_t, \mathbf{a}') \) are linearly independent. Now, if the columns of \( (x'_t, v'_t, \mathbf{a}') \) are linearly independent, then every subset of its columns, too, is linearly independent. Thus, to show the statement of the theorem to be true, we can show that
\[
\text{rank}(E[(x'_t, v'_t, \mathbf{a}')x'_t]) = d_\epsilon, \tag{A.14}
\]
\[
\text{rank}(E[(x'_t, v'_t, \mathbf{a}')v'_t]) = d_\epsilon \tag{A.15}
\]
\[
\text{rank}(E[(x'_t, v'_t, \mathbf{a}')\mathbf{a}']) = d_\epsilon. \tag{A.16}
\]
as both \( v_t \) and \( \mathbf{a}'_t \) are vectors of dimension \( d_\epsilon \).

Since \( E[(x'_t, v'_t, \mathbf{a}')x'_t] \) in Equation (A.14) has \( 3d_\epsilon \) rows and \( d_\epsilon \) columns, \( \text{rank}(E[(x'_t, v'_t, \mathbf{a}')x'_t]) \leq d_\epsilon \). Consider \( x'_t \mathbf{x}'_t \) in Equation (A.14). Now, by condition (i) of the theorem, \( \text{rank}(E[x'_t|x'_t]) = d_\epsilon \), which implies that \( d_\epsilon \) columns of \( E[x'_t|\mathbf{x}'_t] \) are linearly independent. This then implies that the \( d_\epsilon \) columns of \( E[(x'_t, v'_t, \mathbf{a}')x'_t] \) are also linearly independent; that is, it implies that \( \text{rank}(E[(x'_t, v'_t, \mathbf{a}')x'_t]) \geq d_\epsilon \). Thus, we can conclude that \( \text{rank}(E[(x'_t, v'_t, \mathbf{a}')x'_t]) = d_\epsilon \).

Again, given that \( E[(x'_t, v'_t, \mathbf{a}')v'_t] \) in Equation (A.15) has \( 3d_\epsilon \) rows and \( d_\epsilon \) columns, \( \text{rank}(E[(x'_t, v'_t, \mathbf{a}')v'_t]) \leq d_\epsilon \). If we can show the \( \text{rank}(E[0_{d_\epsilon}, v'_t]) \) in Equation (A.15) is \( d_\epsilon \), then it would imply that the \( d_\epsilon \) columns of \( E[(x'_t, v'_t, \mathbf{a}')v'_t] \) are also linearly independent, implying that \( \text{rank}(E[(x'_t, v'_t, \mathbf{a}')v'_t]) \geq d_\epsilon \). Thus, we would be able to show that \( \text{rank}(E[(x'_t, v'_t, \mathbf{a}')v'_t]) = d_\epsilon \), as desired.

To show that \( E[0_{d_\epsilon}, v'_t] \) has a full column rank of \( d_\epsilon \) is equivalent to showing that \( v'_t c = (x'_t - x'_t \pi')c \neq 0 \) almost surely (a.s.) whenever \( c \neq 0 \), \( c \in R^{d_\epsilon} \) and \( x \neq \pi z_t \), a.s.
Now, by condition (i) of the theorem
\[ x'_t c \neq 0 \text{ a.s.} \] (A.17)

By condition (ii), according to which the rank of \( \pi' \) is \( d_\pi \)
\[ \pi' c = \epsilon_c \neq 0, \text{ where } \epsilon_c \text{ is of dimension, } d_\pi \times 1. \] (A.18)

By condition (iii) of the theorem and (A.18), we have
\[ z'_t \pi' c = z'_t \epsilon_c \neq 0 \text{ a.s.} \] (A.19)

From (A.17) to (A.19) we can conclude that \( E(b_0 w_t) \) has a full column rank of \( d_\pi \), thus we are able to establish (A.15).

Similarly, to show (A.16) to be true, we can show that \( \hat{x}'c \neq 0 \text{ a.s. whenever } c \neq 0, \text{ c } \in \mathbb{R}^d \). Now,
\[ \hat{x}'c = (\sum_{t=1}^{T} u_t) \Sigma^{-1} \sum_{t=1}^{T} u_t (\Sigma^{-1} + \Lambda^{-1})^{-1} c \]
\[ = (\sum_{t=1}^{T} u_t)c. \]

Because \( \Sigma_{\epsilon} \) and \( \Lambda_{z_2} \), the covariance matrices of \( \epsilon_t \) and \( \pi \), respectively, are symmetric positive definite matrices, \( \Sigma^{-1} [\Sigma^{-1} + \Lambda^{-1}]^{-1} \) is nonsingular. This implies that \( \Sigma^{-1} [\Sigma^{-1} + \Lambda^{-1}]^{-1} c = c_c \neq 0. \) Since by (A.17) and (A.19), \( (\hat{x}'_t - \hat{x}'_t')c = u'_t c \neq 0 \), when \( x_t \neq \pi x_t \text{ a.s.}, \) we, therefore, have
\[ \hat{x}'c = (\sum_{t=1}^{T} u_t)c \neq 0 \text{ a.s.}, \text{ thus establishing (A.16).} \]

Having established (A.14), (A.15) and (A.16), we can conclude that rank\( E(X_t X'_t) = 3d_\pi \).

**Theorem 2.** Let the linear structural model be
\[ y_t = x'_t \phi + \theta + \zeta_t, t \in \{1, \ldots, T\}. \]
Under Assumption AS 3 and condition (ii) of Lemma 1 we can write \( \theta = E(\theta | Z) + \tau = \phi_t \bar{z} + \tau \), which allows us to write the linear structural equation as
\[ y_t = x'_t \phi + \phi_t \bar{z} + \tau + \zeta_t. \] (A.21)

By Assumption ACF 1 and, with a slight abuse of notations, letting \( E(\zeta_t + \theta | \bar{z}, \pi) = \hat{x}' \phi \pi + \hat{\epsilon}' \phi \pi \), we can write \( \theta + \zeta_t \) in the structural equation as \( \theta + \zeta_t = E(\theta + \zeta_t | X, Z) + \eta_t = \hat{x}' \phi \pi + \hat{\epsilon}' \phi \pi + \eta_t \). This allows us to write the linear structural equation as
\[ y_t = x'_t \phi + \hat{x}' \phi \pi + \hat{\epsilon}' \phi \pi + \eta_t \] (A.22)
where \( \tau + \zeta_t \) is mean independent of \( Z \) and \( \eta_t = \theta + \zeta_t - E(\theta + \zeta_t | X, Z) \) is mean independent of \( X \) and \( Z \).

Given that \( X \) and \( \tau + \zeta_t \) are dependent, let \( \phi_{IV} \) denote the estimated coefficient of \( x_t \) when (A.21) is estimated by two stage least squares (2SLS) with instruments, \( \bar{z} = t \sum_{t=1}^{T} z_t, \bar{z}_t = z_t - \bar{z} \), and \( \hat{\pi} \). And let \( \phi_{IV} \) denote the estimate of \( \phi \) obtained from estimating (A.22) by pooling the data. In the following, we show that \( \phi_{IV} = \phi_{IV} \).

**Proof 4.** Without any loss of generality assume that there is only endogenous regressor, \( x_t \); the proof generalizes to multiple regressors. Given Assumption in AS 3, the reduced form is given by
\[ x_t = z'_t \pi + \pi'_t + a + \epsilon_t, \] (A.23)
where \( \pi + \pi = \pi + \epsilon_t \sim N(0, \sigma^2_\pi) \). Here, \( \bar{y}, \bar{x} \) and \( \bar{z} \) denote group means of \( y_t, x_t \) and \( z_t \), respectively.

Now, by Mundlak (1978) we know that \( \pi \) is equal to the within estimator, \( \pi_w \), of the reduced form Equation (A.23) and \( \pi = \pi_w - \pi_u \), where \( \pi_w \) is the between estimator of \( \pi \) in (A.23). Thus by Lemma 2, the control functions, \( \bar{x} \) and \( \bar{z}_t \), in (A.22) are, respectively, given by
\[ \bar{x} = \bar{z}'(\pi_b - \pi_u) + (1 - \lambda)(\bar{x} - z'_b \pi_b) \]
\[ \bar{z}_t = x_t - z'_t \pi_w - \bar{z}'(\pi_b - \pi_u) - (1 - \lambda)(\bar{x} - z'_b \pi_b), \] where \( \lambda = \frac{\sigma^2_\pi}{\sigma_x^2 + T\sigma_z^2} \).

It is convenient to write the panel model in (A.22) in vector form as
\[ Y = X \phi + \hat{A} \phi_x + \bar{E} \phi \pi + \eta, \] (A.24)
where \( Y \) is a NT \times 1 \text{ data matrix which has the observations on } y_{it} \text{ stacked in NT rows; similarly, } X \text{ has stacked observations on } x_{it}, \text{ } \hat{A} \text{ on } \hat{z}_t \text{ and } \bar{E} \text{ on } \epsilon_{it}. \text{ Now, given } \bar{x} \text{ and } \bar{z}_t, \text{ we can write } \hat{A} \text{ and } \bar{E}, \text{ respectively, as}
\[ \hat{A} = PZ \hat{\pi} + (1 - \lambda) PR \]
and \( \bar{E} = QR + \lambda PR \).
where \( R \) is a \( NT \times 1 \) matrix, which has the residuals of the reduced form Equation (A.23), \( r_\hat{\pi} = x_\hat{\pi} - Z\hat{\pi} - \hat{\pi}^{-1}(\hat{\pi}_0 - \hat{\pi}_w), \) stacked in \( NT \) rows, \( Z \) has stacked observations on \( z_\alpha \), \( P \) is a matrix that averages the observation across time for each individual, i.e., \( P = 1_T \otimes \bar{I}_T \) where \( I_T \) is a matrix of ones of dimension \( T \) and \( \bar{I}_T = I_T / T \), and \( Q = I_{NT} - P \) is a matrix that obtains the deviations from individual means. Given the above, we can write (A.24) as

\[
Y = X\phi + (PZ\pi + (1 - \lambda)PR)\phi_x + (QR + \lambda PR)\phi_e + \eta. \tag{A.25}
\]

Because \( P \) and \( Q \) are idempotent and \( P \) is orthogonal to \( Q \), premultiplying (A.25) through out by \( Q \), we get

\[
\hat{Y} = X\hat{\phi} + \hat{\eta}. \tag{A.26}
\]

where \( \hat{Y} \) is a \( NT \times 1 \) matrix that has \( \hat{y}_\alpha = y_\alpha - \hat{\eta}_\alpha \), the deviations of \( y_\alpha \) from the individual means, \( \hat{\eta}_\alpha \), (within transformation) stacked in \( NT \) rows, and \( \hat{R} = \hat{X} - \hat{Z}\hat{\pi}_w \) is the matrix obtained after within transforming the residuals, \( \hat{R} \), \( \hat{R} \), incidentally, are the residuals obtained after applying within transformation to the reduced form Equation (A.23) and estimating it as a fixed effect model.

Now, premultiplying (A.25) through out by \( P \), we get

\[
\hat{Y} = X\phi + (Z\pi + (1 - \lambda)R)\phi_x + \lambda R\phi_e + \eta, \tag{A.27}
\]

where \( \hat{Y} \) is a \( NT \times 1 \) matrix that has the individual means, \( \hat{y}_\alpha \), (between transformation) stacked in \( NT \) rows, and \( R = X - Z\pi_0 \) is the matrix obtained after between transforming the residuals, \( R \), \( R \), incidentally, are the residuals obtained after applying between transformation to the reduced form Equation (A.23) and estimating it by OLS.

Since the column space of the LHS variables, \([X, Z\pi + (1 - \lambda)R, \lambda R]\), in (A.27) is same as that of \([X, Z\pi, R]\), the projections of the two matrices will be the same. Therefore, the estimates of \( \phi, \phi_x \) and \( \phi_e \) in (A.27) will be the same from estimating the following equation:

\[
\hat{Y} = X\phi + Z\pi\phi_x + R\phi_e + \hat{\eta}. \tag{A.28}
\]

From (A.26) and (A.28), it is thus evident that estimating the following equation

\[
Y = X\phi + Z\pi\phi_x + [QR + PR]\phi_e + \eta, \tag{A.29}
\]

by OLS would yield the same result as estimating (A.25) by OLS.

By Frisch–Waugh–Lovell Theorem, then, \( \phi \) and \( \phi_x \) in (A.29) is estimated as

\[
\begin{bmatrix}
\hat{\phi}_{CF} \\
\hat{\phi}_x
\end{bmatrix} = \begin{bmatrix}
X^\prime Z \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} \left[ QM_R + PM_R \right] \begin{bmatrix}
X \\
\tilde{\pi} \hat{\pi}
\end{bmatrix} \begin{bmatrix}
X^\prime Z \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} \left[ QM_R + PM_R \right] Y,
\]

where \( \hat{\pi} \) is the estimate of \( \pi, M_R = I_{NT} - R[R^\prime R]^{-1}R^\prime \) and \( M_R = I_{NT} - R[R^\prime R]^{-1}R^\prime \). Since \( X \) and \( Y \) can be decomposed as \( X = \hat{X} + X \) and \( Y = \hat{Y} + Y \), respectively, we can write the above equation as

\[
\begin{bmatrix}
\hat{\phi}_{CF} \\
\hat{\phi}_x
\end{bmatrix} = \begin{bmatrix}
\tilde{\pi} \hat{\pi} \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} \left[ QM_R + PM_R \right] \begin{bmatrix}
\hat{X} + X \\
\hat{\pi} \hat{\pi}
\end{bmatrix} \begin{bmatrix}
\tilde{\pi} \hat{\pi} \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} \left[ QM_R + PM_R \right] \begin{bmatrix}
\hat{Y} + Y
\end{bmatrix}.
\]

Taking into account the orthogonality conditions in the above, the above simplifies to

\[
\begin{bmatrix}
\hat{\phi}_{CF} \\
\hat{\phi}_x
\end{bmatrix} = \begin{bmatrix}
\hat{X}^\prime M_R \hat{X} + \begin{bmatrix}
\tilde{\pi} \hat{\pi} \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} M_R \begin{bmatrix}
\hat{X} + X \\
\hat{\pi} \hat{\pi}
\end{bmatrix} \tilde{\pi} \hat{\pi}^\prime M_R \hat{Y} + \begin{bmatrix}
\tilde{\pi} \hat{\pi} \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} M_R \hat{Y} \hat{X} + \tilde{\pi} \hat{\pi}^\prime M_R \hat{Y} + \tilde{\pi} \hat{\pi}^\prime M_R \hat{Y}.
\]

(A.30)

Because \( \hat{R} = M_Z X, \) where \( M_Z = I_{NT} - Z[Z^\prime Z]^{-1}Z^\prime, \) it can be verified that \( \hat{X}^\prime M_R = X^\prime P_Z, \) where \( P_Z = Z[Z^\prime Z]^{-1}Z^\prime \) is the projection matrix. And because \( R = M_Z X, \) where \( M_Z = I_{NT} - Z[Z^\prime Z]^{-1}Z^\prime, \) it can be verified that \( X^\prime M_R = X^\prime P_Z, \) where \( P_Z = Z[Z^\prime Z]^{-1}Z^\prime \), and that \( \tilde{\pi} \hat{\pi}^\prime M_R = \tilde{\pi} \hat{\pi}^\prime P_Z \). Thus, we can write \( \hat{\phi}_{CF} \) and \( \hat{\phi}_x \) in (A.30) as

\[
\begin{bmatrix}
\hat{\phi}_{CF} \\
\hat{\phi}_x
\end{bmatrix} = \begin{bmatrix}
\hat{X}^\prime P_Z \\
\tilde{\pi} \hat{\pi}^\prime
\end{bmatrix} \begin{bmatrix}
\hat{X} + X \\
\hat{\pi} \hat{\pi}
\end{bmatrix} \begin{bmatrix}
P_Z \hat{X} + \begin{bmatrix}
\tilde{\pi} \hat{\pi} \\
\tilde{\pi} \tilde{\pi}^\prime
\end{bmatrix} P_Z \hat{Y}
\end{bmatrix},
\]

(A.31)

which is the same as the estimates \( \hat{\phi}_{IV} \) and \( \hat{\phi}_x \), when (A.21) is estimated by the 2SLS using \( z, z_i = z_i - \bar{z} \) and \( z_i^\prime \hat{\pi} \) as instruments.

If only the within transform, Equation (A.26), is estimated, then by a similar process as above, beginning with the Frisch–Waugh–Lovell Theorem, one obtains

\[
\hat{\phi}_{CF} = \begin{bmatrix}
\hat{X}^\prime P_Z \\
\tilde{\pi} \hat{\pi}^\prime
\end{bmatrix} \begin{bmatrix}
\hat{X}^\prime P_Z \\
\tilde{\pi} \hat{\pi}^\prime
\end{bmatrix} \hat{Y},
\]

(A.32)

which is same as the within estimate of \( \hat{\phi}_{IV} \), when (A.21) is estimated by fixed effect two-stage least squares (FE2SLS) that utilizes \( z_i = z_i - \bar{z} \) as instruments.

If only the between transform, Equation (A.28), is estimated, then, again, a similar process as the one used to derive (A.31), one finally obtains
which is the estimate of $\phi_{M}^{b}$ and $\phi_{M}^{b}$ when (A.21) is between transformed and estimated by two-stage least squares (2SLS) with $z$ and $\hat{z}$ as instruments. After some algebraic manipulations, we get

$$\phi_{M}^{b} = \left(\bar{X}'P_{Z}M_{Z\hat{z}}P_{Z}\bar{X}\right)^{-1}\bar{X}'P_{Z}M_{Z\hat{z}}\bar{Y},$$

where $M_{Z\hat{z}} = I - Z\hat{z}[\hat{z}'Z\hat{z}]^{-1}\hat{z}'Z'$.

**Lemma 3.** If the endogenous variables, $x$, has large a support, then under AS 3, the support of the conditional distribution of $\hat{z}(X, Z, \Theta_{1})$ and $\hat{e}(X, Z, \Theta_{1})$, conditional on $x = \bar{x}$, is same as the support of their marginal distribution.

**Proof 5.**

(a) We have shown that the expected value of $a = \bar{x} + a$ and $\epsilon_{t}$ given $Z, X$, where $a$ and $\epsilon_{t}$ are normally distributed with variances $\Lambda_{xz}$ and $\Sigma_{\epsilon \epsilon}$, respectively, are given

$$E(x|X, Z) = \bar{x} = \bar{x} + \alpha + \sum_{t=1}^{T} \Omega(x_{t} - \Pi Z_{t})$$

$$E(\epsilon_{t}|X, Z) = \hat{e}_{t} = x_{t} - \Pi Z_{t} - \sum_{t=1}^{T} \Omega(x_{t} - \Pi Z_{t})$$

where $\Pi = (\bar{x}, \hat{z})$, $Z_{t} = (z_{t}', \hat{z}_{t}')'$ and $\Omega = [T\Sigma_{\epsilon}^{-1} + \Lambda_{xz}^{-1}]^{-1}\Sigma_{\epsilon}^{-1}$.

Because the support of $x_{t}$ is $\mathbb{R}^{d_{x}}$ and $\Omega$ is a $d_{x} \times d_{x}$ nonsingular matrix,

$$\text{Supp}(\bar{x}) = \text{Supp}(\hat{e}_{t}) = \mathbb{R}^{d_{x}}$$

whether or not $z_{t}$ has a large support.

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

$$\text{Supp}(\hat{z}|x_{t} = \bar{x}) = \text{Supp}(\bar{x} + \Omega(x_{t} - \Pi Z_{t}) + \sum_{s \neq t} \Omega(x_{s} - \Pi Z_{s})) = \mathbb{R}^{d_{x}}$$

$$\text{Supp}(\hat{e}_{t}|x_{t} = \bar{x}) = \text{Supp}([I_{m} - \Omega](x - \Pi Z_{t}) - \sum_{s \neq t} \Omega(x_{s} - \Pi Z_{s})) = \mathbb{R}^{d_{x}}.$$

(b) When we have a single endogenous variable, $x_{t}$, given by

$$x_{t} = \Pi Z_{t} + \epsilon_{t}, t = 1, \ldots, T,$$

where the errors, $\epsilon \equiv (\epsilon_{1}, \ldots, \epsilon_{T})'$, are nonspherical such that $E(\epsilon \epsilon') = \Sigma_{\epsilon}$, an invertible $T \times T$ matrix and $a$ is normally distributed with variance $\sigma_{a}^{2}$, then, we showed that

$$\hat{a}(X, Z, \Theta_{1}) = (x_{1} - \Pi Z_{1})\omega_{1} + \ldots + (x_{T} - \Pi Z_{T})\omega_{T},$$

where $(\omega_{1}, \ldots, \omega_{T})' = \Sigma_{\epsilon}^{-1/2} \left(\frac{1}{\sum_{t=1}^{T} \omega_{t}}\right)$ and $e_{T}$ is a vector of ones of dimension $T$.

Given that $x_{t}'s$ have large supports, using a similar argument as in part (a), we get

$$\text{Supp}(\hat{a}|x_{t} = \bar{x}) = \text{RandSupp}(\hat{e}_{t}|x_{t} = \bar{x}) = \mathbb{R}.$$
Appendix B

Tables

Table B.1. Performance of the APE, $\frac{\partial G(x)}{\partial C_{22}} x = 1$, for alternative estimators.

| True APE | CRECF Method | Papke and Wooldridge | Chamberlain’s CRE probit | Chamberlain’s logit |
|----------|--------------|----------------------|--------------------------|--------------------|
| Mean RMSE | Mean RMSE | Mean RMSE | Mean RMSE | Mean RMSE |
| $N = 200$ | $-0.0931$ | $0.0445$ | $-0.0920$ | $0.0724$ | $-0.0354$ | $0.0561$ | $-0.0578$ | $0.0473$ | $-0.1070$ |
| $N = 500$ | $-0.0944$ | $0.0283$ | $-0.0932$ | $0.0654$ | $-0.0353$ | $0.0462$ | $-0.0578$ | $0.0310$ | $-0.1061$ |
| $N = 1000$ | $-0.0935$ | $0.0203$ | $-0.0936$ | $0.0616$ | $-0.0353$ | $0.0408$ | $-0.0579$ | $0.0238$ | $-0.1057$ |
| $N = 2000$ | $-0.0934$ | $0.0143$ | $-0.0936$ | $0.0597$ | $-0.0353$ | $0.0381$ | $-0.0579$ | $0.0189$ | $-0.1057$ |
| $N = 5000$ | $-0.0939$ | $0.0088$ | $-0.0936$ | $0.0592$ | $-0.0353$ | $0.0370$ | $-0.0579$ | $0.0144$ | $-0.1053$ |

RMSE is Root Mean Square Error and Mean is the mean value of $m = 2000$ APEs.

Table B.2. Performance of the APEs, $\frac{\partial G(x)}{\partial C_{22}} x_1 = 0.5$, $\frac{\partial G(x)}{\partial C_{22}} x_2 = 1$, for alternative estimators.

| True APE | CRECF Method | Chamberlain’s CRE probit | Chamberlain’s logit |
|----------|--------------|--------------------------|--------------------|
| Mean RMSE | Mean RMSE | Mean RMSE | Mean RMSE |
| $N = 500$ | $x_1$ | $-0.3328$ | $0.0538$ | $-0.3385$ | $0.0761$ | $-0.3874$ | $0.1114$ | $-0.4294$ |
| $x_2$ | $0.1655$ | $0.0411$ | $0.1817$ | $0.186$ | $0.3491$ | $0.2085$ | $0.372$ |
| $N = 1000$ | $x_1$ | $-0.3325$ | $0.0383$ | $-0.3395$ | $0.0659$ | $-0.3869$ | $0.103$ | $-0.428$ |
| $x_2$ | $0.1656$ | $0.0316$ | $0.1821$ | $0.1845$ | $0.349$ | $0.2069$ | $0.3715$ |
| $N = 2000$ | $x_1$ | $-0.331$ | $0.0282$ | $-0.3402$ | $0.0618$ | $-0.3879$ | $0.1009$ | $-0.4281$ |
| $x_2$ | $0.1662$ | $0.0247$ | $0.182$ | $0.1828$ | $0.3484$ | $0.2055$ | $0.3713$ |
| $N = 5000$ | $x_1$ | $-0.3316$ | $0.0192$ | $-0.3407$ | $0.0579$ | $-0.387$ | $0.098$ | $-0.4281$ |
| $x_2$ | $0.1657$ | $0.0205$ | $0.1813$ | $0.1832$ | $0.3487$ | $0.2059$ | $0.3714$ |

RMSE is Root Mean Square Error and Mean is the mean value of $m = 2000$ APEs.

Table B.3. 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 | 46.23 | 77.48 |
| 8 to 14 years | 22.88 | 26.85 | 49.73 | 11 to 17 years | 49.73 | 53.77 | 103.5 |
| 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 B.4. Descriptive statistics.

| Variable | 2007 | | 2010 | |
|----------|------|------|------|------|
| Child characteristics | Mean | Std. Dev. | Mean | Std. Dev. |
| Sex (Male = 1, Female = 0) | 0.52 | 0.50 | 0.52 | 0.50 |
| Age (yrs.) | 8.07 | 2.97 | 11.07 | 2.97 |
| Household characteristics | | | | |
| Parents participated in NREGS (Yes = 1 & No = 0) | 0.33 | 0.47 | 0.66 | 0.47 |
| Total number of days parents worked in NREGS | 9.21 | 21.44 | 36.00 | 48.10 |
| Land Owned (acre) | 2.32 | 3.42 | 3.86 | 43.53 |
| Farm Asset Index | -0.13 | 0.98 | 0.22 | 1.46 |
| Gini Coefficient for Land Owned | 0.62 | 0.74 | 0.62 | 0.74 |
| Total Income of Household (in Thousand ₹) | 20,787 | 35,813 | 29,013 | 62,225 |
| Annual nonagricultural income (₹) | 20,787 | 35,813 | 29,013 | 62,225 |
| Annual agricultural income (₹) | 5,060 | 23,319 | 9,936 | 42,746 |
| 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 |
| 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 |
| Community (Mandal) characteristics | | | | |
| Total NREGS amount sanctioned (in Million ₹) | 7.25 | 8.30 | 20.19 | 19.17 |

Total number of children/observations in each period: 2458.
### Table B.5. Descriptive statistics of some variables by caste.

| Year: 2007 | Scheduled Castes/Tribes | Other Backward Classes | Others |
|------------|-------------------------|------------------------|--------|
| Household income (in Thousand ₹) | 31.22 (33.94) | 31.64 (34.29) | 43.21 (48.59) |
| Land owned in acre | 1.58 (2.12) | 2.32 (3.51) | 3.08 (4.53) |
| Index of productive Farm asset | -0.22 (0.71) | -0.14 (1.02) | 0.04 (1.17) |
| School dummy D_SCHOOL = 1 | 0.90 (0.29) | 0.89 (0.32) | 0.96 (0.19) |
| Work dummy D_WORK = 1 | 0.33 (0.47) | 0.33 (0.47) | 0.29 (0.45) |

| Year: 2010 | Scheduled Castes/Tribes | Other Backward Classes | Others |
|------------|-------------------------|------------------------|--------|
| 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 (1.56) | 3.08 (4.53) |
| Index of productive Farm asset | 0.12 (1.16) | 0.29 (1.56) | 0.54 (1.89) |
| School dummy D_SCHOOL = 1 | 0.89 (0.31) | 0.87 (0.33) | 0.94 (0.23) |
| Work dummy D_WORK = 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 B.6. First stage reduced form estimates: joint estimation of income, land and farm assets equations.

| | Income | Landholding | Farm Asset |
|----------------|----------------|----------------|----------------|
| Total NREGS amount sanctioned (in Million ₹) | 0.047*** | -0.008 | -0.0003 |
| | (0.009) | (0.007) | (0.002) |
| Caste (SC/ST = 1, OBC = 2, OT = 3) | 9.220*** | 2.278*** | 0.171*** |
| | (1.217) | (0.726) | (0.030) |
| Drinkable Water in the Locality (Yes = 1 & No = 0) | 5.417 | -1.879 | 0.341*** |
| | (5.703) | (4.260) | (0.150) |
| National Bank in the Locality (Yes = 1 & No = 0) | -2.785 | 4.684** | 0.046 |
| | (3.099) | (2.315) | (0.082) |
| Engineered Road to the Locality (Yes = 1 & No = 0) | 0.159 | 2.413 | 0.183** |
| | (2.130) | (1.591) | (0.056) |
| Hospital in the Locality (Yes = 1 & No = 0) | -0.689 | -4.143*** | 0.056 |
| | (1.248) | (0.932) | (0.033) |
| Other Exogenous Variables of the Structural Equation: Age and Sex of the Child | Yes | Yes | Yes |

Biørn’s stepwise MLE was employed to obtain these estimates.

Significance levels: *: 10% **: 5% ***: 1% Standard errors (SE) in parentheses.

### Table B.7. Household income and wealth effect on incidence of child labor.

| CRE probit | Control function (CF) method |
|------------|-----------------------------|
| Coeff. | Coeff. | APE bounds | Coeff. |
| Income | 0.003*** | -0.0234*** | [-0.00532, -0.00451] | 0.005 |
| | (0.0008) | (0.0028) | (0.0007) | (0.003) |
| Landholding | 0.002 | 0.031*** | [0.00679, 0.00749] | 0.015** |
| | (0.002) | (0.007) | (0.00676, 0.00752) | (0.0065) |
| Farm Asset Index | -0.011 | -0.976*** | [-0.22077, -0.18734] | 1.512*** |
| | (0.0279) | (0.169) | (0.0057, 0.12552) | (0.129) |
| Age | 2.019*** | 0.402*** | [0.0757, 0.12533] | 0.0275*** |
| | (0.072) | (0.057) | (0.0755, 0.12552) | (0.003) |
| Sex | 0.644*** | 0.394*** | [0.07355, 0.12318] | 0.031*** |
| | (0.042) | (0.0473) | (0.07322, 0.12351) | (0.0075) |
| | | | | 0.882*** |
| | | | | (0.185) |

Total number of children: 2458. Total number of observations with positive outcome: 2128

Significance levels: *: 10% **: 5% ***: 1%, Standard errors (SE) in parentheses.
Figure C.1. Comparison with alternative estimators: density of $\frac{\partial G(\hat{x})}{\partial x} - \frac{\partial G(\hat{x})}{\partial x}$ at $x = 1$ for different sample size.
Figure C.2. Density of difference between the estimated and the true APEs, \( \partial G(x)/\partial x - \partial G(\bar{x})/\partial x \), at \( x = 1 \) where the estimated APEs, \( \partial G(x)/\partial x \), are obtained by varying the control functions and the instruments in PW’s model. Sample Size: \( N = 5000 \).

(a) Joint Density of \( \hat{\epsilon}_{it} \) and \( x \).  
(b) Joint Density of \( \hat{\alpha}_i \) and \( x \).  
(c) Joint Density of \( v_{it} \) and \( x \).

Figure C.3. Level curves of estimated joint density of \( x \) and various control functions.
Acknowledgments

I am grateful to the anonymous referee and the associate editor whose comments and suggestions have helped to substantially improve the article. Thanks are due to seminar participants at the The Bank of Estonia, the 10th Nordic Econometric Meeting (Stockholm), the Institute of Mathematics and Statistics (University of Tartu) and the Inaugural Baltic Economic Conference (Vilnius) for the same. I would especially like to thank Soham Sahoo for helping me with the data. All remaining errors are mine.

References

Afriyie, F., Mukhopadhyay, A., Sahoo, S. (2016). Female labor force participation and child education in India: Evidence from the national rural employment guarantee scheme. *IZA Journal of Labor & Development* 5(1):7. doi:10.1186/s40175-016-0053-y

Altonji, J. G., Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73(4):1053–1102. doi:10.1111/j.1468-0262.2005.00609.x

Arellano, M., Bonhomme, S. (2013). Binary choice panel data models with predetermined variables. *Journal of Econometrics* 115(1):125–157. doi:10.1016/S0304-4076(03)00095-2

Arellano, M., Bonhomme, S. (2011). Nonlinear panel data analysis. *Annual Review of Economics* 3(1):395–424. doi:10.1146/annurev-economics-111809-125139

Baland, J. M., Robinson, J. A. (2000). Is child labor inefficient? *Journal of Political Economy* 108(4):663–679. doi:10.1086/316097

Baltagi, B. (1981). Simultaneous equations with error components. *Journal of Econometrics* 17(2):189–200. doi:10.1016/0304-4076(81)90026-9

Baltagi, B. H., Song, S. H., Jung, B. C. (2010). Testing for heteroskedasticity and serial correlation in a random effects panel data model. *Journal of Econometrics* 154(2):122–124. doi:10.1016/j.jeconom.2009.04.009

Basu, K., Das, S., Dutta, B., Van, P. H. (1998). The economics of child labor. *American Economic Review* 88:412–427.

Basu, K., Das, S., Dutta, B. (2010). Child labor and household wealth: Theory and empirical evidence of an Inverted-U. *Journal of Development Economics* 91(1):8–14. doi:10.1016/j.jdeveco.2009.01.006

Bester, C. A., Hansen, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics* 27(2):235–250. doi:10.1198/jbes.2009.0017

Bhalotra, S., Heady, C. (2003). Child farm labor: The wealth paradox. *The World Bank Economic Review* 17(2):197–227. doi:10.1093/wber/ltg017

Biorn, E. (2004). Regression systems for unbalanced panel data: A stepwise maximum likelihood procedure. *Journal of Econometrics* 122:281–291.

Blundell, R., MaCurdy, T., Meghir, C. (2007). Chapter 69 labor supply models: Unobserved heterogeneity, nonparticipation and dynamics. *Handbook of Econometrics* 6:4667–4775.

Blundell, R., MaCurdy, T., Meghir, C., Powell, J., Smith, R. J. (1994). Coherency and estimation in simultaneous models with censored or qualitative dependent variables. *Journal of Econometrics* 64(1-2):355–373. doi:10.1016/0304-4076(94)90069-8

Blundell, R., MaCurdy, T., Meghir, C., Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont, M., Hansen, L., Turnovsky, S., eds., *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. 2, Cambridge: Cambridge University Press.

Blundell, R., Powell, J. (2004). Endogeneity in semiparametric binary response models. *Review of economic studies*, 71:655–679. doi:10.1111/j.1467-937X.2004.00299.x

Blundell, R.—Smith, R. J. (1994). Coherency and estimation in simultaneous models with censored or qualitative dependent variables. *Journal of Econometrics*, 64 (1-2):355–373. doi:10.1016/0304-4076(94)90069-8

Cadre, B., Pelletier, B., Pudlo, P. (2013). Estimation of density level sets with a given probability content. *Journal of Nonparametric Statistics*, 25 (1):261–272. doi:10.1080/10485252.2012.750319

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

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica* 78:159–168.

Cockburn, J., Dostie, B. (2007). Child work and schooling: The role of household asset profiles and poverty in rural Ethiopia. *Journal of African Economies* 16(4):519–563. doi:10.1093/jae/ejl045

Constantinou, P., Dawid, A. P. (2017). Extended conditional independence and applications in causal inference. *Annals of Statistics* 45(6):2618–2653.

D’Haultfoeuille, X., Fèverier, P. (2015). Identification of nonseparable triangular models with discrete instruments. *Econometrica* 83(3):1199–1210. doi:10.3982/ECTA10038

Dumas, C. (2006). Why do parents make their children work? A test of the poverty hypothesis in rural areas of Burkina Faso. *Oxford Economic Papers* 59(2):301–329. doi:10.1093/oep/gpl031
