Estimating Endogenous Ordered Response Panel Data Models with an Application to Income Gradient in Child Health*

Kajal Lahiri†[1] and Liu Yang‡[2]

[1]Department of Economics, University at Albany, SUNY, NY 12222, USA

[2]School of Economics, Nanjing University, Jiangsu 210093, P. R. China

Abstract

In this paper, a control function approach is developed to deal with endogeneity issue in estimating panel data models with ordered response. This methodology is characterized by a two-step procedure, requiring only standard regression to be implemented in each step. To demonstrate the usefulness of our method, we investigate income gradient in child health based on U.S. data from the Panel Study of Income Dynamics. We find empirical evidence that income gradient rises during early childhood, followed by a decline after age 12. Ignoring endogeneity of family income and unobserved individual heterogeneity would underestimate the true gradient considerably.

JEL Classifications: C01, C13, C33, C35, C36, I10

Key words: Control function, Panel data, Endogeneity, Ordered response, Income gradient, Child health.

*We are grateful to Colin Cameron as the discussant of our paper and also to Anirban Basu, Partha Deb, Donna Gilleskie and Robin Lumsdaine for providing comments at the Annual Health Econometrics Workshop (AHEW) in Knoxville, September 26-28, 2019. We also benefited from discussions with Jeffery Wooldridge, and Jingya Song and Jinman Pang for numerous help with the data. Helpful suggestions from an anonymous referee are also acknowledged. This research was supported by the National Institute on Minority Health and Health Disparities, National Institutes of Health (grant number 1 P20 MD003373) and the National Natural Science Foundation of China (grant number 71603115). The content is solely the responsibility of the authors and does not represent the official views of the sponsoring agencies.

†Corresponding author. Tel.: +1 518 442 4758. E-mail address: klahiri@albany.edu.

‡Tel.: +86 18710075220. E-mail address: lyang2@nju.edu.edu.

1 Introduction

Endogeneity might arise for a variety of reasons in studying the causal relationship between two or more economic variables. It results in inconsistency of standard estimators that are justified in the absence of endogeneity and reverse causality. Fortunately, econometricians have proposed a wide range of solutions to the problem of endogeneity. If the same individuals are observed repeatedly over time, we can overcome the potential endogeneity problem created by the omission of individual heterogeneity. For instance, in linear panel data models, one can rely on a number of data transformations, e.g. within or first-difference, to get rid of the time-invariant unobserved effects and estimate the structural parameters by ordinary least squares. When endogeneity is caused by the correlation between included regressors and time-varying omitted variables, eliminating the individual heterogeneity alone would not work. In these circumstances, we need additional information to identify the parameters of interest. Among them, the exclusion restriction may be the most commonly used if we are able to find a set of valid instruments that are highly correlated with endogenous regressors but have no direct influence on the dependent variable. This is equivalent to expanding the system with more equations. The structural equation represents the functional relationship between the dependent variable and exogenous or endogenous regressors, whereas the reduced form equations write endogenous regressors as functions of all exogenous variables, including instruments. Recently, Klein and Vella (2010) and Lewbel (2012) exploited heteroscedasticity of the error terms for identification without the exclusion restriction. However, their approaches cannot be used when the dependent variable is discrete or of substantially limited range. Furthermore, they only considered cross-sectional design without controlling for individual heterogeneity.

In this paper, we estimate a parametric ordered response model based on panel data. Many economic variables are observed discrete with a natural order. Examples include educational attainment (“high school dropout”,“complete high school but not college”, and “complete college or a higher degree”), preference towards a commodity (“strongly dislike”, “neutral”, and “strongly like”), subjective rating of health (”excellent, ”very good”, ”good”, ”fair”,

and ”poor”), and the like. The typical estimation method in these contexts is maximum likelihood if cross-sectional data is available; see Maddala (1983) for a textbook treatment. This approach can be easily extended to panel data models where individual heterogeneity is present. Random effects model is used when individual heterogeneity is independent of all regressors, and estimation follows by integrating the unobserved effects out of the likelihood, provided its marginal distribution is specified correctly. Wooldridge (2010) provided more details on this model. However, independence assumption is too strong to be useful in many cases. In models with correlated individual effects, the fixed effects model is often used that allows for a general form of dependence. To this end, Das and van Soest (1999) decomposed the ordered response model into a series of models with binary outcomes, each of which is estimated by conditional maximum likelihood, as suggested by Anderson (1970) and Chamberlain (1980). These conditional logit estimators can be combined to yield the final estimator by using minimum distance. As in the binary panel data models with fixed effects, it is generally hard to find a sufficient statistic to be conditioned on for the probit case; see Greene and Hensher (2010). Another shortcoming of fixed effects specification is that only the parameter associated with an explanatory variable can be identified, and thus, it precludes computation of marginal effect, which is more informative in nonlinear regression models. In this paper, we consider the so-called “correlated random effects” model. The distinguishing feature of this approach is that it explicitly models the dependence between heterogeneity and regressors. Within this framework, estimation and inference can be undertaken in a straightforward manner. See Cameron and Trivedi (2005) for applications of this specification in other nonlinear panel data models.

When the included regressors are correlated with time-variant omitted variables, controlling for individual heterogeneity alone cannot eliminate endogeneity completely. Papke and Wooldridge (2008) described a control function approach by finding a set of valid instruments. The control function works since it captures the correlation between endogenous variables and unobservable error term. Holding the variation generated by the control function constant, the remaining part of the error term becomes uncorrelated with all regressors, and the structural parameters can be estimated consistently by a standard procedure. This is quite similar, in spirit, to the Heckman’s two-step estimator of a sample selection model. It

is well known that the OLS estimator of the structural equation using the observed sample is inconsistent. Heckman (1979) augmented this equation with the inverse Mills ratio which serves as a control function. Once the ratio is controlled for, all regressors in the structural equation become exogenous and thus OLS estimator restores its consistency.

The strength of the control function approach is threefold. First, it does not impose a stringent distributional assumption on the entire system compared with the joint maximum likelihood alternative. In particular, the distribution of the reduced form error is not restricted. Second, the functional form of the system can be made rather flexible. The endogenous regresssors may appear in the structural equation in various nonlinear ways. For example, they may interact with the exogenous regressors. In this setting, the traditional two stage least squares with endogenous regressors replaced by their first stage fitted values might fail, see Wooldridge (2010) for further discussion. Finally, this approach suggests a simple strategy to test the endogeneity of the regressors, as will become clear shortly. Moreover, the average marginal effect, which is more meaningful than coefficients in nonlinear econometric models, is much easier to estimate with control function.

In general, this methodology can be applicable to virtually all nonlinear panel data models. For example, Papke and Wooldridge (2008) considered this approach in fractional response panel data models, where the bounded nature of the dependent variable is recognized. Giles and Murtazashvili (2013) employed this approach to estimate a dynamic binary response panel data model with contemporaneous endogenous regressors. The current paper uses the control function approach to estimate ordered response panel data models.

To demonstrate the usefulness of this methodology, we revisit the relationship between family income and child health, known as income gradient in the health literature. The approach is implemented using the Child Development Supplement of the Panel Study of Income Dynamics, which is a panel data of moderate size. In this setting, one may have two sources of endogeneity. First, some unobserved family characteristics, like housing conditions and safety of neighborhoods, might affect both child health and included explanatory variables. Most existing studies using panel data simply pool all observations together without taking care of the unobserved individual heterogeneity. We contribute to the literature by correcting for this type of endogeneity within the correlated random effects framework. Sec-

ond, regressors may also be correlated with some time-varying determinants of child health. For brevity, we consider the case where family income is the only endogenous regressor with respect to child health. Our empirical results reveal an increasing trend for income gradient at early childhood if both types of endogeneity are accounted for. Ignoring endogeneity would underestimate income gradient dramatically for all age groups.

The remaining paper is organized as follows. In Section 2, we outline the basic model structure and derive the two-step control function estimator. This approach is applied to study the income gradient in child health using the Child Development Supplement data in Section 3. Section 4 concludes this paper with further remarks and discussions for future research. All mathematical proofs on the asymptotic properties of the proposed procedure are contained in the appendix.

2 Model Structure and Estimation Strategy

We denote a generic cross-sectional unit by i, and a particular time period by t. In this section, the number of periods T_i for individual i is fixed, and asymptotic analysis is conducted by letting the number of units grow without bound. To save notation, only balanced scenario is considered, i.e. $T_i = T$ for each i.

Suppose there exists a latent variable Y_{it}^*, which is related to the covariates in the following way:

$$Y_{it}^* = X_{1it}\beta_1 + Y_{2it}\beta_2 + c_i + \varepsilon_{it}, \tag{1}$$

where X_{1it} is $1 \times K$, Y_{2it} is a scalar, and $(\beta_1', \beta_2)'$ are unknown parameters. The presence of c_i in (1) enables us to control for unobserved individual heterogeneity, which is possibly correlated with other covariates in (1). Here, ε_{it} varies across i as well as across t, and it may or may not be serially correlated. Y_{2it} is likely to be correlated with both c_i and ε_{it}, and thus could be endogenous even when c_i is conditioned on. We are merely concerned with single endogenous regressor, but extension to multiple endogenous regressors is rather straightfor-

ward so long as more valid instrumental variables are available. The observed response $Y_{it} = j$ if $\eta_{j-1} < Y_{it}^* \leq \eta_j$ for $j = 1, ..., J$, where η's are thresholds satisfying $\eta_0 \equiv -\infty$, $\eta_J \equiv \infty$, and $\eta_{j-1} < \eta_j$. We summarize all unknown thresholds in a vector $\eta \equiv (\eta_1, ..., \eta_{J-1})'$.

In order for the control function method to work, we rely on the following exclusion restriction. Specifically, let $Z_{it} \equiv (X_{1it}, X_{2it})$ be a $1 \times M$ vector, where X_{2it} is $1 \times L$ ($L \geq 1$) and $M = K + L$. To model the dependence between c_i and other regressors within the correlated random effects framework, we follow the approach taken by Chamberlain (1980) and Mundlak (1978), and express c_i as

$$c_i = \bar{Z}_i \theta + u_i, \tag{2}$$

where $\bar{Z}_i$ is the time average of Z_{it}. After replacing c_i in (1) with (2), we get

$$Y_{it}^* = X_{1it}\beta_1 + Y_{2it}\beta_2 + \bar{Z}_i\theta + r_{it}, \tag{3}$$

and $r_{it} \equiv u_i + \varepsilon_{it}$. When all regressors in (3) are independent of r_{it}, $(\beta_1', \beta_2, \theta', \eta')'$ can be estimated by maximum likelihood if we are willing to assume r_{it} follows a specific distribution. However, endogeneity of Y_{2it} invalidates this procedure.

To identify the structural parameters of interest in the presence of endogenous Y_{2it}, we write the reduced form equation, relating Y_{2it} to the set of exogenous variables in all time periods $Z_i \equiv \{Z_{i1}, ..., Z_{iT}\}$, as

$$Y_{2it} = Z_{it}\gamma + \bar{Z}_i\lambda + v_{it}. \tag{4}$$

Again, we have used Chamberlain's approach to derive (4). Endogeneity of Y_{2it} in (3) comes from the correlation between r_{it} and v_{it}. Formally, the condition

$$D(r_{it}|Y_{2it}, Z_i) = D(r_{it}|v_{it}) \sim N(v_{it}\rho, \sigma_r^2) \tag{5}$$

is imposed, where $D(r_{it}|Y_{2it}, Z_i)$ is the conditional distribution of r_{it} given Y_{2it} and Z_i, $\rho \equiv Cov(r_{it}, v_{it})/Var(v_{it})$, and $\sigma_r^2 \equiv Var(r_{it}) - Var(v_{it})\rho^2$. Therefore, the structural error r_{it} is

correlated with Y_{2it} only through its dependence on v_{it}. Homoscedasticity assumption in (5), i.e. the conditional variance of r_{it} given v_{it} is constant, is purely to ease the exposition. Extension to heteroskedacity of known form is performed straightforwardly.

An important implication of (5) is that Z_i may not be exogenous in (3), i.e., a subset of Z_i could be correlated with r_{it}, although the correlation vanishes when v_{it} is controlled for. In addition, (5) rules out the possibility that Y_{2it} is discrete or of substantially limited range.[1] To see why this is the case, note that joint independence between (r_{it}, v_{it}) and Z_i clearly implies the first equality in (5). When Y_{2it} is discrete, the possible values v_{it} could take are determined by the value of Z_i. This means that v_{it} and Z_i must be dependent. Actually, this is a drawback associated with nearly all control function applications, including those that are flexible in functional forms. See Blundell and Powell (2003, 2004), Papke and Wooldridge (2008), and Rothe (2009) for more details. Nevertheless, the current framework allows for the categorical instruments in Z_i.

Chesher and Smolinski (2012) examined the identification issue of ordered probit regression with potentially endogenous covariates that are discrete, leaving the mechanism that generates those covariates unspecified, and they concluded that the model of this type is set, not point, identified. When Y_{2it} is not continuous, the conventional method to (point) identify and estimate the structural parameters is to impose a stringent distributional assumption on Y_{it} and Y_{2it} given Z_i. All parameters involved are estimated by maximum likelihood accordingly. The frequentist version of this procedure was adopted by Rivers and Vuong (1988), while Munkin and Trivedi (2008) proposed a Bayesian solution to this problem. In spite of its estimation efficiency in large samples, likelihood-based methods are computationally intensive. Kawakatsu and Largey (2009) developed an EM algorithm to facilitate computation of these models. Besides, likelihood-based methods are not robust to even a slight deviation from the hypothesized distribution. Recently, Lewbel and Dong (2015) devised a simple estimator in a binary response model allowing for a discrete endogenous regressor. This methodology is computationally feasible and thus empirically attractive. However, it does require the existence of a so-called "very special" regressor with a large support constraint, and is not designed for ordered response models with multiple categories.

[1] In the same vein, endogenous binary treatment effects are not identified as well.

After controlling for v_{it}, Y_{2it} is no longer endogenous. Alternatively, (5) can be written as

$$r_{it} = v_{it}\rho + \zeta_{it}, \tag{6}$$

where $\zeta_{it} \sim N(0, \sigma_r^2)$. Putting (6) in place of r_{it} in (3), we obtain

$$Y_{it}^* = X_{1it}\beta_1 + Y_{2it}\beta_2 + \bar{Z}_i\theta + v_{it}\rho + \zeta_{it}. \tag{7}$$

A point worth noting here is that nonlinear functions of the endogenous covariate Y_{2it}, such as Y_{2it}^2 and the interaction terms between Y_{2it}^2 and X_{1it}, are permitted in (7). It follows from (5) and (6) that ζ_{it}, by construction, is independent of v_{it} and Z_i. Because v_{it} and Y_{2it} are one-to-one functions of each other given Z_i by (4), ζ_{it} is also independent of $X_{1i} \equiv (X_{1i1}, ..., X_{1iT})$, Y_{2it}, as well as any nonlinear function of (Y_{2it}, X_{1i}). This is in sharp constrast with the traditional two-stage least square procedure (TSLS), where both Y_{2it} and its nonlinear functions have to be instrumented properly. In the control function approach, the additional nonlinear functions of endogenous covariates do not need extra treatment since we have imposed a more stringent independence assumption between ζ_{it} and (v_{it}, Z_i). As pointed out by Wooldridge (2010), there exists cases where the control function estimator is inconsistent while the TSLS estimator is. On the other hand, the control function estimator is generally more efficient than the TSLS counterpart provided the independence assumption is true.

Now, (7) takes the form of a conventional ordered probit model and can be estimated routinely. Specifically, the log-likelihood function $l_i(\alpha)$ for unit i is defined as

$$\sum_{t=1}^{T}\sum_{j=1}^{J} I(Y_{it} = j) log(\Phi(\frac{\eta_j - X_{1it}\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho}{\sigma_r}) - \Phi(\frac{\eta_{j-1} - X_{1it}\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho}{\sigma_r})), \tag{8}$$

where α is the vector containing all parameters appearing in (8), and $\Phi(\cdot)$ is the standard normal distribution function.

To motivate (8), we observe that

$$\begin{aligned} & P(Y_{it}=j|Z_i,Y_{2it}) = P(\eta_{j-1} < Y_{it}^* \leq \eta_j|Z_i,Y_{2it}) \\ = \; & P(\eta_{j-1} - X_{1it}\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho < \zeta_{it} \leq \eta_j - X_{1it}\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho|Z_i,Y_{2it}). \end{aligned}$$

Equation (8) results from the independence between ζ_{it} and (Z_i, Y_{2it}). However, this is not the conditional probability $P(Y_{it}=j|Z_i,Y_{2i})$ where $Y_{2i} \equiv (Y_{2i1},...,Y_{2iT})'$ if Y_{2is} $(s \neq t)$ affects ζ_{it}. Moreover, $P(Y_{it}=j, Y_{is}=q|Z_i,Y_{2i})$ may not be equal to $P(Y_{it}=j|Z_i,Y_{2i})P(Y_{is}=q|Z_i,Y_{2i})$ unless $\{\zeta_{it} : t=1,...,T\}$ are serially independent, which is questionable because they may include a common component inherent in (r_{it}, v_{it}). As a consequence, (8) is not the fully specified log-likelihood of observed ordered responses for unit i. Instead, it is partially correctly specified. Despite the fact that (8) is not completely correct, maximizing it still yields an estimator with desired asymptotic properties, as confirmed in the appendix. Wooldridge (2010) called $l_N(\alpha) \equiv \sum_{i=1}^N l_i(\alpha)$ as partial log-likelihood in a panel data context. For the sake of identification, α needs to be restricted. For example, σ_r can be set to 1, i.e, ζ_{it} follows the standard normal distribution, since $(\beta_1', \beta_2, \theta', \rho, \eta')'$ are identified up to a scale factor.[2] The full set of identification conditions, besides the normalization $\sigma_r = 1$, is given in the appendix.

The partial maximum likelihood estimator $\hat{\alpha}$ maximizes $l_N(\alpha)$ over the allowed parameter space. This is nothing but the usual ordered probit regression after adding $\bar{Z}_i$ and v_{it} as additional regressors and then pooling all NT observations together. To make this procedure operational, the unknown error v_{it} in (8) must be estimated first. This is done by estimating (4) using pooled OLS or feasible GLS to get the residual $\hat{v}_{it}$. Under relatively weak conditions, maximizing $l_N(\alpha)$ with v_{it} replaced by $\hat{v}_{it}$ would lead to a $\sqrt{N}$-consistent estimator $\hat{\alpha}$ for all identified parameters. Given that both steps are easily performed in most of the standard statistical packages, this procedure is computationally appealing. However, the default asymptotic variance of $\hat{\alpha}$ cannot be used since the first stage estimation uncertainty of $\hat{v}_{it}$ must be addressed adequately. We adopt the approach of Newey and McFadden (1994), which fits

[2] An alternative identification strategy is setting $Var(r_{it}) = 1$, under which $\sigma_r = \sqrt{1 - Var(v_{it})\rho^2}$. Since $Var(v_{it})$ can be identified in (4) and the scaled parameters, say, ρ/σ_r, are also identified in (8), the structural parameters of original interest can be identified as well. We refer to Giles and Murtazashvili (2013) and Papke and Wooldridge (2008) to appreciate how it works.

a two-step estimation problem of this type into the framework of generalized method of moments. Consistency and asymptotic normality are established by verifying a set of regularity conditions laid by them, as detailed in the appendix.

In a nonlinear econometric model, the estimates of unknown parameters are often less informative than the true marginal effect. Suppose the objective of the current study is to measure the effect of Y_{2it} on the probability of the lowest possible outcome $Y_{it} = 1$. Instead of focusing on the parameter β_2, we are more interested in the partial derivative of the conditional probability of $Y_{it} = 1$ with respect to Y_{2it}. The question is what should be conditioned on if Y_{2it} is endogenous. The choice of variables in the conditioning set is essentially determined by the aim of the study. If Y_{2it} is exogenous given c_i in (1), $P(Y_{it} = 1|X_{1it}, Y_{2it}, c_i)$ may be of particular interest. Otherwise, $P(Y_{it} = 1|X_{1it}, Y_{2it}, c_i)$ does not tell anything meaningful and thereby are not structurally important. Based on an omitted variable formulation, Papke and Wooldridge (2008) suggested computing $P(Y_{it} = 1|Z_i, Y_{2it})$, which is $\Phi(\eta_1 - X_{1it}\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho)$ by (8). Suppose we are mainly concerned with the effect of Y_{2it} when X_{1it} takes a specific value, say, $X_{1it} = x_1^o$. The average partial effect, which is widely accepted to quantify the impact of Y_{2it} in practice, can be obtained as

$$-\frac{\beta_2}{T}E(\sum_{t=1}^{T}\phi(\eta_1 - x_1^o\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho)), \tag{9}$$

where the expectation is taken over the distribution of $(Y_{2it}, \bar{Z}_i, v_{it})$ across i. Under quite general conditions, (9) can be consistently estimated by

$$-\frac{\hat{\beta}_2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi(\hat{\eta}_1 - x_1^o\hat{\beta}_1 - Y_{2it}\hat{\beta}_2 - \bar{Z}_i\hat{\theta} - v_{it}\hat{\rho}), \tag{10}$$

where the notation $\hat{}$ means parameter estimate, and $\phi(\cdot)$ is the standard normal density function. It is difficult, if not infeasible, to derive the asymptotic variance of (10). Instead, panel data bootstrap is used in our empirical illustration, as recommended by Papke and Wooldridge (2008).

3 Empirical Application: Income Gradient in Child Health

Understanding the relationship between family income and child health, and the pathways through which the former affects the latter enables the policy makers to reduce, if not eliminate, the income-related health disparity in a cost effective manner. The benefit of doing so could be substantial due to the far reaching implications of child health on the socioeconomic status and well-being when children get into their adulthood. Moreover, a part of the intergenerational transmission of socioeconomic status may be attributed to the impact of family income on child health. The positive association between family income and child health has been well documented in the health economics literature, see Case et al. (2002), Chen et al. (2017), Condliffe and Link (2008), Currie et al. (2003), Currie et al. (2007), Fernald et al. (2012), Khanam et al. (2009) and Swaminathan et al. (2019), just to mention a few. Fletcher and Wolfe (2014) provided a comprehensive overview of this topic. Children in the rich family are generally more healthy than their poor counterparts, often known as the income gradient in child health. This relationship is fairly robust to the choice of the estimation sample. When it comes to the magnitude and other nuances of the gradient, empirical evidence has not been unambiguous. The pioneering work by Case et al. (2002) using U.S. data found that the gradient becomes stronger as children age, and analyzed the sources of the steepening gradient in terms of incidence and accumulation of health shocks. However, some recent studies using data from Germany (Reinhold and Jürges (2012)), and Australia (Khanam et al. (2009)) appear to contradict the steepening gradient hypothesis when more explanatory variables are considered.

A potential limitation of the aforementioned studies is that the positive correlation between family income and child health observed in the data may not have a causal interpretation. The evidence on the positive association cab be explained by the fact that income and health are determined simultaneously. A natural consequence of simultaneity is the endogeneity of family income in the child health equation, leading to the failure of the standard estimation procedures. By focusing on children, we do not need to worry about this too much since it is reasonable to assume that child health has little effect on family income

in most developed countries. This argument, however, does not rule out the possibility that other family characteristics, whether observable or not, may affect both family income and child health. If the analysts fail to incorporate these information into their regression models, endogeneity will emerge as a result of omitted variables. One of the examples of these "third factors" is parents' education. While higher education typically leads to higher income, more educated parents are also able to take care of their children better simply because they have more health-related knowledge. Fortunately, most household surveys contain education variables, at least for one of the parents. Therefore, controlling for education is a regular strategy to mitigate, if not eliminate, the omitted variable bias. Unlike parents' education, many other confounding factors are unobservable or not available in a particular survey. In this circumstance, it requires additional information for identification purpose. If the same child is observed repeatedly over time, we can control for the unobservable individual heterogeneity within a fixed effects framework. However, when family income is suspected to be correlated with unobservable time-varying factors, we may need some instrumental variables that are highly correlated with income. In order to address the potential endogeneity, we take the occupational characteristics (white/blue color jobs) and working hours of both parents as four separate instruments for family income. It is reasonable to postulate (and we test later) that these variables have no direct effect on child health after controlling for family income, yet they are highly correlated with family income.

Our analysis is based on information gathered from the Child Development Supplement (CDS), which has a panel structure with three waves. CDS, as an important supplement to the main Panel Study of Income Dynamics (PSID), contains a rich collection of information on parents and their children aged 0-19. These include, but are not limited to, reliable assessments of the cognitive, behavioral, and health status of a number of children in the family obtained from a variety of sources. There are 3,563 children surveyed in the first wave (1997). Among them, only 2,907 and 1,506 remained in the second (2002) and third (2007) waves respectively. To illustrate the methodology of Section 2, a balanced panel is considered, which only includes those appearing in all three waves. In each wave, the primary caregiver was asked to report the health information of his/her children. In line with most of the previous studies, we use primary-caregiver-rated health as the measure of child health. The primary

caregiver would assess the child health according to the 1 to 5 ordinal scale, with 1 coded as excellent, 2 = very good, 3 = good, 4 = fair and 5 coded as poor. Most of the children in our sample are relatively healthy and none of them fall into the worst category (5), as is evident from Table 1. The family income is obtained from main PSID data, which can be linked with CDS via the family identifier. Since it is the long-run average income that determines health investments and health, following the literature, we take logarithm of "permanent" income as our family income variable (in 2007 constant dollars). Yearly moving averages of annual income ending in the year prior to the interview over the relevant sample period was used as the measure of permanent income. The information regarding the instrumental variables is stored in the main PSID data. The survey also contains a rich set of control variables, such as gender, race, smoking status of parents, health insurance, and so on. To facilitate the analysis, we also restrict our attention to a smaller sub-sample, which results from the original balanced panel by dropping individuals with missing values on any variable we use. The final sample size is 2,496. Table 1 summarizes several descriptive statistics of our sample, and they all look reasonable. Condliffe and Link (2008) reported some of the statistics using PSID data, but using only the 2002 panel. Further details on our 3-wave panel data can be found in Chatterji et al (2013).

Figure 1 plots the probability of excellent health status against the family income for four age groups, as predicted by separate univariate probit regressions in the absence of any control. As family income increases, child health improves rapidly for all age groups, but the gain slows down considerably after family income reaches around $400,000$, showing that the effect is strongest on the poorest. In addition, we observe substantial heterogeneity in the steepening of income gradient for different age groups. This is an important issue in health economics because the steepening of gradient implies as children age, the detrimental effects of low family income accumulate over time. The gradient seems to be flattening for children in their teens - this is similar to the evidence presented in Fletcher and Wolf (2014). We are interested in testing these patterns formally after controlling for a set of other explanatory variables, unobserved individual heterogeneity and endogeneity of family income.

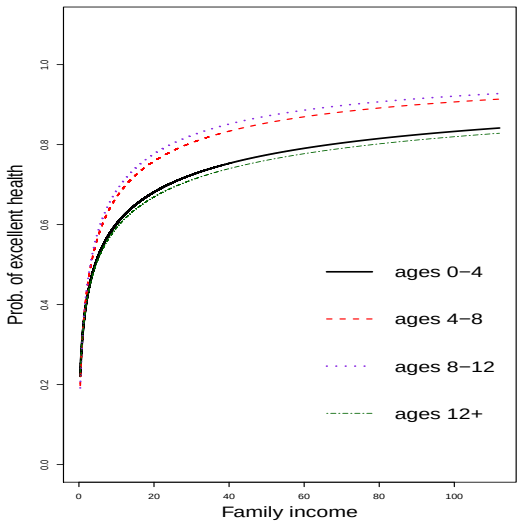
To implement the two-step procedure of Section 2, extra care is needed when some exogenous variables are time-invariant, such as gender and race of children. Without further

Table 1: Descriptive statistics

variables	mean	std.	min.	max.
primary-caregiver-rated health				
-"excellent"	0.581	0.493	0	1
-"very good"	0.300	0.458	0	1
-"good"	0.099	0.299	0	1
-"fair"	0.019	0.136	0	1
family income	8.539	7.479	0.352	112.007
child's age	9.212	4.789	0.100	19.010
mother's age at child birth	28.350	5.632	14.580	44.330
father's working hours	2106	658.417	0	4853
mother's working hours	1219	793.068	0	4187
mother's education	14.270	2.060	6	17
mother's health (1=excellent or very good)	0.645	0.478	0	1
father's health (1=excellent or very good)	0.668	0.471	0	1
mother's occupation (1=white collar)	0.255	0.436	0	1
father's occupation (1=white collar)	0.248	0.432	0	1
number of children in the family	2.204	0.997	0	9
child's gender (1=male)	0.538	0.499	0	1
child's race (1=black)	0.258	0.438	0	1
had food stamp	0.089	0.285	0	1
had food stamp during pregnancy	0.131	0.337	0	1
metropolitan area	0.436	0.496	0	1
smoking in the family unit	0.296	0.457	0	1
housing (1=very clean or somewhat clean)	0.484	0.500	0	1
environment (1=safe)	0.803	0.398	0	1
had health insurance	0.897	0.304	0	1
had chronic condition(s)	0.485	0.500	0	1
birth weight$\leq$5.5lb	0.059	0.235	0	1
N		2,496		

Notes: Family income is measured in the unit of 10,000 dollars per year. White collar refers to professional, technical, and kindred workers, managers and administrators (except farm), as well as sales workers. Metropolitan area is defined as one with 1 million population or more. The child had chronic condition if a medical professional has ever told the parents that their child has any of the following diseases: epilepsy, asthma, ear infections, diabetes, anemia, elevated lead in the blood, orthopedic impairment, developmental problems, allergies, and other health problems. All statistics are based on the balanced panel we use for estimation.

Figure 1: Probability of excellent health v.s. family income for four age groups



assumptions, we cannot isolate their partial effects on child health from that via the individual heterogeneity c_i. Nevertheless, they are always included in the regression model as long as there is no multi-colinearity amongst them.[3] As a result, the estimated coefficients of these variables cannot be interpreted as the partial effects on child health. Instead, they measure the overall effects that are sum of the effects on Y_{it}^* and c_i. Fortunately, we are not primarily interested in the partial effects of these variables. We concentrate on the partial effect of family income, which changes over time. To explore the possibility that income gradient may vary with the age of the child, we supplement the health equation with interaction terms between log-income and three age group dummies (group 4-8, 8-12, 12+, with group 0-4 being the baseline). Recall that higher Y_{it}^* indicates worse health. Thus, we expect that the coefficient of log-income to be negative.

We first conduct a preliminary regression experiment with results summarized in column (6) of Table 2. We employ the specification that is extensively adopted, but with a broader set of covariates, to examine the dynamics of income gradient and compare our estimates with the existing literature. Specifically, in column (6), we ignore the possible endogeneity of family income, and regress child health on all explanatory variables by pooled ordered probit approach, which has been the prevalent specification in previous studies involving panel data.

[3]Within the correlated random effects framework, we merely control for time averages of those exogenous variables that vary over time, namely, chronic condition, father's working hours, mother's working hours, food stamp, number of children, health insurance, safe environment, clean house, mother's health, and father's health.

Note that the time average of the exogeneous variables $\bar{Z}_i$ in equation (7) are not included, and thereby the correlation between c_i and Z_i is also overlooked in this scenario. The result shows that only the interaction terms between age groups and log-income are significant at the conventional 5% level. A number of control variables including parents' education, use of food stamps, parents' smoking, living in a metropolital area, child having chronic health conditions, and race come up significantly with expected signs.

Next, we examine what happens when the regression is augmented by the time average $\bar{Z}_i$ and when log-income is instrumented by parents' occupational choices and working hours. In the lower panel of Table 2, we report the joint significance test of X_{2it} in the first stage regression (4).[4] Since this is a linear regression by pooling all observations across individuals and over time, heteroskedasticity and autocorrelation within a cross-sectional unit have to be tackled for the purpose of drawing valid inference. Following the robust approach in Wooldridge (2010), we construct the Wald statistic using the heteroskedasticity and autocorrelation robust asymptotic covariance matrix of the pooled OLS estimator. If the proposed instruments were uncorrelated with the endogenous covariate, the Wald statistic would follow a chi-squred distribution with L (i.e. dimension of X_{2it}) degrees of freedom. Fortunately, p-values in the table, almost zero in magnitude, indicate that our instruments are strongly correlated with the log-income regardless of which covariates are controlled for. The high χ^2 statistics rule out any potential concerns related to weak instruments.

Regarding instrument validity, there have been some epidemiological research suggesting that parent's hours of work and quality of occupations may directly affect child health, cf. Nicholson et al. (2012). So it is very important to test the assumed exogeneity of our instruments. Case et al. (2002) and Condliffe and Link (2008) used occupational characteristics as IVs. Certainly, the participation of both parents in paid work improves children's well-being because paid employment will generate more family income to afford high quality day and health care. After all, parental employment is the best protection for children against the

[4]In the first stage, the dependent variable is the log of family (permanent) income. The covariates include: race, gender, number of children, whether child has chronic condition, whether child has low birth weight, whether family is in metropolitan area, whether house is clean, whether environment is safe, whether somebody smokes in the family, whether child has health insurance, mother's age at child birth, food stamp recipient, food stamp during pregnancy, mother's education, father's and mother's working hours, father's and mother's occupational choices (blue/white colar jobs), as well as the time average of those variables that vary over time. The complete first stage regression results are available upon request.

Table 2: Results from ordered probit regression

	accounting for endogeneity					ignoring endogeneity
variables	(1)	(2)	(3)	(4)	(5)	(6)
resid	0.337***	0.336***	0.284**	0.324**	0.352**	
	(0.118)	(0.129)	(0.132)	(0.162)	(0.177)	
ln(income)	-0.612***	-0.539***	-0.472***	-0.461***	-0.416**	-0.059
	(0.111)	(0.123)	(0.126)	(0.165)	(0.201)	(0.064)
ln(income)*age4-8	-0.061*	-0.070**	-0.091**	-0.099**	-0.109***	-0.121***
	(0.036)	(0.036)	(0.037)	(0.039)	(0.039)	(0.041)
ln(income)*age8-12	-0.073*	-0.088**	-0.111***	-0.116***	-0.130***	-0.153***
	(0.038)	(0.038)	(0.040)	(0.043)	(0.044)	(0.043)
ln(income)*age12+	0.009	-0.007	-0.034	-0.036	-0.057	-0.079**
	(0.036)	(0.036)	(0.037)	(0.040)	(0.041)	(0.040)
male(=1)		0.047	-0.009	-0.006	-0.004	0.002
		(0.061)	(0.060)	(0.060)	(0.060)	(0.048)
black(=1)		0.188**	0.252**	0.312***	0.250***	0.358***
		(0.087)	(0.088)	(0.092)	(0.094)	(0.060)
chronic			0.202***	0.210***	0.215***	0.504***
			(0.067)	(0.067)	(0.068)	(0.049)
low birth weight			0.006	0.014	0.005	0.006
			(0.137)	(0.137)	(0.133)	(0.101)
no. of children				0.037	0.038	0.029
				(0.037)	(0.037)	(0.025)
metropolitan				-0.054	-0.051	-0.116**
				(0.073)	(0.071)	(0.051)
smoking in family				0.129*	0.094	0.153***
				(0.071)	(0.068)	(0.055)
clean house				0.078	0.076	-0.070
				(0.067)	(0.069)	(0.049)
safe envrn.				0.017	0.018	-0.082
				(0.072)	(0.074)	(0.062)
food stamp at preg				0.053	0.007	0.080
				(0.109)	(0.106)	(0.081)
food stamp				-0.231*	-0.238*	-0.247***
				(0.124)	(0.128)	(0.095)
health insurance				-0.031	-0.026	-0.024
				(0.094)	(0.096)	(0.078)
mother's age at birth					0.007	-0.001
					(0.007)	(0.005)
mother's edu					0.007	-0.013
					(0.020)	(0.014)
mother's health(good=1)					0.012	-0.268***
					(0.069)	(0.055)
father's health(good=1)					-0.103	-0.105*
					(0.073)	(0.055)
η_1	-0.494***	-0.327*	0.139	0.137	0.159	-0.279
	(0.159)	(0.191)	(0.188)	(0.263)	(0.350)	(0.228)
η_2	0.516***	0.693***	1.204***	1.214***	1.257***	0.795***
	(0.160)	(0.193)	(0.190)	(0.263)	(0.350)	(0.229)
η_3	1.437***	1.620***	2.182***	2.204***	2.268***	1.779***
	(0.173)	(0.207)	(0.205)	(0.277)	(0.359)	(0.235)
time averages	Yes	Yes	Yes	Yes	Yes	No
1st stage χ^2	360.514***	325.766***	320.512***	253.864***	184.497***	
N	2,496	2,496	2,496	2,496	2,496	2,496

Notes: (η_1, η_2, η_3) are threshold values in equation (8). Standard errors are in parenthesis. Asterisks indicate significance levels: $*p \leq 0.1$, $**p \leq 0.05$, $***p \leq 0.01$.

adverse health effects of low income. However, there can be trade-off between family time for child development and parents' hours and quality of work. Some have hypothesized that children whose health is most vulnerable to intergenerationally transmitted disadvantages are precisely those with low family incomes, longer combined parental work hours and poorer quality of jobs. Most studies have, however, focused on the effect of mother's hours and quality of work on children's behavioral health and obesity, see Courtemanche et al. (2017). Hsin and Felfe (2014) found maternal work to have no effect on time in activities that positively affect children's healthy development. In a comprehensive meta analysis of 69 studies, Lucas-Thompson et al. (2010) showed that early maternal employment per se is rarely associated with children's later outcomes including health. These diverse considerations prompted us to test the exogeneity of our instruments.

We use the logic of Sargan test that any set of the instruments, under the null that they are all truly exogenous, should be absent in equation (8) once the correction term v_{it} is controlled for. This requires us to use only one instrument to just-identify the model and test if the other three instruments are significant in (8). Based on the benchmark specification in column (5) of Table 2, the p-values of four Wald tests,[5] each of which has 3 degrees of freedom, are given by 0.433 (when mother's working hours is the only instrument for family income), 0.229 (when father's working hours is the only instrument for family income), 0.221 (when mother's occupational choice is the only instrument for family income), and 0.598 (when father's occupational choice is the only instrument for family income), providing strong evidence in support of the exogeneity of our instruments. These results are consistent with the secular rise in women's overall workforce participation and other factors have also played a role in enabling more women to stay in the labor force after pregnancy. Norms and infrastructure with respect to how families approach work and child rearing have shifted such that women no longer drop out of the labor force upon becoming a mother, see Dave and Young (2019). The 1980's witnessed the emergence of flexible work schedules, and employment based child care benefits, making it easier for women with children to continue to work. In contemporary American society, the presumed trade-off between maternal work and quality

[5]To accommodate arbitrary heteroskedasticity and autocorrelation, we use the asymptotic covariance matrix of the second stage estimator in the appendix to construct the Wald statistics.

child care is possibly no longer exist, supporting our exogeneity test results. [6]

The second stage pooled ordered probit results are summarized in columns (1)-(5) of Table 2, which are produced by successively adding more covariates in the health equation. The log income and its interactions with age groups (except age 12+) are highly significant across all specifications. However, the impact of family income on child health deteriorates from -0.612 to -0.416 as more variables are added, indicating the existance of multiple transmission channels that translate income into better health. With these channels being fixed, income effect is mediated to some extent, as expected. As column (5) shows, the income (in addition to its age group interactions) still remains highly significant and is much larger than that in column (6), even with all controls in place. We also report the standard errors obtained by adjusting for the estimation uncertainty in computing residuals v_{it} in columns (1)-(5), as detailed in the appendix. Not surprisingly, we find that taking care of the extra uncertainty due to the use of the generated regressor makes most of the coefficients less precise in column (5).

It is clear from equation (5) that after controlling for the correlated individual effects, the only source of endogeneity stems from non-zero parameter ρ. As shown in the first row of Table 2, the significant coefficients of residual (resid) in columns (1) - (5) imply that family income is indeed endogenous. The positive coefficient for resid implies that unobserved variables (neighborhood quality, for example) inadvertently excluded from the specification affect both family income and child health positively. Hence, the estimation results, by overlooking endogeneity, would have been misleading. In particular, by comparing columns (5) and (6), we find the income coefficient to be statistically significant, and larger in size when the endogeneity issue is properly dealt with. This agrees with the evidence provided by Kuehnle (2014) using a probit model on British data that after correcting for endogeneity, the income gradient in child health rises considerably. A similar result was also found by

[6] Alternatively, one can use any set of three instruments to over-identify the model and check if the last one is present in (8) via the robust t test. The p-values are given by 0.204 (for mother's working hours), 0.746 (for father's working hours), 0.950 (for mother's occupational choice), and 0.125 (for father's occupational choice). Exogeneity is overwhelmingly confirmed once again. Following the extant literature, we also tried few distal variables like years of education and smoking status of grandparents as additional instruments, cf. Doyle et al. (2007). But in our analysis, these variables did not pass the instrument validity test. Kuehule (2104) and Wei and Fenny (2019) have used local area unemployment rates in this context. We used state unemployment rates as IV, but this variable had very little correlation with family income, and was subsequently dropped. The county-level labor market variables will possibly work well in this context.

Papke and Wooldrige (2008) on the effect of local area income on test scores of schools in their fractional response model using control function. The qualitative patterns of income gradient are not affected much by endogeneity. Increasing gradient is more likely to occur during early childhood. Initially, family income becomes more and more vital for health as child grows up and the income effect peaks for group 8-12. After age 12, the size of the gradient starts to decline. This finding is in line with Figure 1 and Fletcher and Wolf (2014) as well. After controlling for endogeneity in column (5), compared to the specification in column (6) without allowing for endogeneity, parents health, smoking in family, and living in an metropolitan area lost their statistical significance. However, black, chronic health, and food stamp continued to be statistically significant even after allowing for endogeneity of family income. It appears the effect of the remaining variables, most notably that of parents' health and education have been absorbed in the income variable, in addition to their effects via the individual heterogeneity. The list of all such variables that were included in our estimates in columns (1) through (5) in table 2 are reported in footnote 4. Among these, time averages of chronic conditions, clean house, mother's health, and father's working hours were significant in column (5) at the 5% level. Thus, we successfully identified the effect of mother's health in the presence of a significant income coefficient. As Kuehnle (2014) has pointed out, there have been problems in delineating its separate effect in the presence of family income. We also note that the strong income gradient withstands some "third facotr" explanations in terms of parental education and health, low birth weight, etc.

To quantify the difference between columns (5) and (6), we cannot directly compare the point estimates in Table 2. Karaca-Mandic et al. (2012) argued that the estimated coefficients in nonlinear econometric models can change dramatically due to the normalization of the error variance induced by an omitted seemingly irrelevant heterogeneity. In contrast, the marginal effect is relatively stable. Inspired by this insight, we compare the marginal effect of family income on child health. Specifically, we consider the effect on the probability that the child is in excellent health. Our framework can also be used to evaluate the marginal effects at other points of the health distribution, cf. Davillas et al. (2019). To see how income gradient evolves dynamically, we consider the marginal effects of family income across age groups. In general, the marginal effects in a nonlinear econometric model depend on the

values of all variables. To obtain a single summary, we calculate the average marginal effect of family income. For group 0-4, the income gradient is calculated as

$$-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi(m_{it}^{0-4})\frac{\hat{\beta}_{0-4}}{income_{it}},$$

where m_{it}^{0-4} is defined as

$$\hat{\eta}_1 - ln(income)_{it}\hat{\beta}_{0-4} - X_{1it}\hat{\beta} - \bar{Z}_i\hat{\theta} - v_{it}\hat{\rho},$$

$\hat{\beta}_{0-4}$ is the estimated coefficient of log-income, and X_{1it} denotes all exogenous explanatory variables. For group 4-8, the income gradient is

$$-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi(m_{it}^{4-8})\frac{\hat{\beta}_{0-4}+\hat{\beta}_{4-8}}{income_{it}},$$

where m_{it}^{4-8} is defined as

$$\hat{\eta}_1 - ln(income)_{it}(\hat{\beta}_{0-4}+\hat{\beta}_{4-8}) - X_{1it}\hat{\beta} - \bar{Z}_i\hat{\theta} - v_{it}\hat{\rho}.$$

Here, $\hat{\beta}_{4-8}$ is the estimated coefficient of the interaction term between log-income and group 4-8. The average gradient for other groups can be derived analogously. The estimated age-gradient profiles are summarized in Table 3, and the 95% confidence intervals are constructed by cluster bootstrap (resampling across cross-sectional units). We find that ignoring endogeneity underestimates the impact of family income on the probability of reporting excellent health by $2.2-2.5$ percentage points. For example, when the income level is raised by $\$10,000$ per year, the primary caregiver is, on average, 2.9% more likely to rate the health of his/her child as excellent according to column '0-4' in Table 3. The gradient is only 0.4% when endogeneity is ignored. Taken together, in our empirical illustration, correcting for endogeneity makes a big difference on the effect of family income on child health. The estimates suggest a positive and increasing gradient till about age 12, but a flat gradient in the teens. Also, given that most of the transmission variables like parents' education are time invariant in the sample, the varying income gradient as children age produces a convincing

evidence that family income casually affects child health, cf. Fletcher and Wolf (2014).

Table 3: The average derivatives of probability of excellent health for four age groups

age group	0-4	4-8	8-12	12+
ignoring endogeneity				
average derivative	0.004	0.012	0.014	0.010
	(-0.005,0.013)	(0.004,0.021)	(0.006,0.023)	(0.002,0.017)
accounting for endogeneity				
average derivative	0.029	0.036	0.037	0.032
	(0.001,0.056)	(0.009,0.062)	(0.011,0.063)	(0.006,0.059)

Notes: The 95% confidence intervals are constructed by cluster bootstrap (resampling across cross-sectional units) with 20,000 bootstrap repititions.

4 Conclusion

We propose a control function approach to estimating ordered response panel data models in the presence of endogeneity. Like Papke and Wooldridge (2008), our procedure is computationally attractive in that all regressions involved are quite standard. In addition, the asymptotic distribution of this two-step estimator is derived although the statistical inference can also be carried out by boostrap. This approach is illustrated by reexamining the income gradient in child health using the data from the Panel Study of Income Dynamics. The empirical finding implies that correcting for endogeneity would lead investigators to draw a different conclusion as to the impact of family income on child health and its dynamics as children grow up. The effects are significantly underestimated without correcting for endogeneity. One caveat of our analysis is that if the vulnerable mothers in terms of health, education, family income, etc. are less aware of their children's true health status, and report accordingly, the estimated income gradient will even be steeper that we find after correcting for endogeneity. Using German Socio-Economic Panel data, Sandner and Jungmnann (2016) found that the concordance between maternal ratings and children's true health decreases in mothers with multiple risk burdens.

Compared with the joint maximum likelihood approach, our method relaxes the distributional assumption on the error term of the reduced-form equation, and thereby reduces the risk of misspecification. However, it does require the conditional independence between structural error and exogenous variables, which means that discrete endogenous regressors are ruled out in our framework. How to extend the current procedure to handle discrete endogenous regressors without imposing strong restrictions is still an open question. in addition, as discussed in section 3, the structural parameters associated with time-invariant explanatory variables cannot be identified in the correlated random effects framework. This is not an issue in our empirical example since the primary interest is centered on the partial effect of family income, which is time-varying. There may be situations where we are interested in the effects of time-invariant observables, which may or may not be endogenous in the structural equation. The approaches proposed by Hausman and Taylor (1981) and Chatterji et al. (2014) might be possible alternatives to pursue after recognizing the ordered nature of the dependent variable. We leave these unresolved issues as topics for further research.

Mathematical Appendix

In this appendix, we attempt to show the two-step estimator $\hat{\alpha}$ defined as the maximizer of $l_N(\alpha)$ over a compact parameter space Θ is consistent and asymptotically normally distributed even if v_{it} in $l_N(\alpha)$ is replaced by the first step residual $\hat{v}_{it}$. Throughout this section, we assume that the second moments of all random variables are finite so that the quantities being studied are well-defined. Decompose α into two sub-vectors: $\alpha_1 \equiv (\beta_1', \beta_2, \theta', \rho, \eta')'$ and $\alpha_2 \equiv (\gamma', \lambda')'$. Hence, α_1 contains parameters estimated at the second stage, whereas those estimated at the first stage are included in α_2. Given α_2 and the normalization $\sigma_r = 1$, the partial derivative of $l_i(\alpha_1, \alpha_2)$ with respect to α_1 is

$$\begin{aligned}
\frac{\partial l_i(\alpha_1,\alpha_2)}{\partial \beta_1} &= \sum_{t=1}^{T}\sum_{j=1}^{J} I(Y_{it}=j)\frac{(\phi(h_{it}^{j-1})-\phi(h_{it}^{j}))X_{1it}'}{\Phi(h_{it}^{j})-\Phi(h_{it}^{j-1})}, \\
\frac{\partial l_i(\alpha_1,\alpha_2)}{\partial \beta_2} &= \sum_{t=1}^{T}\sum_{j=1}^{J} I(Y_{it}=j)\frac{(\phi(h_{it}^{j-1})-\phi(h_{it}^{j}))Y_{2it}}{\Phi(h_{it}^{j})-\Phi(h_{it}^{j-1})}, \\
\frac{\partial l_i(\alpha_1,\alpha_2)}{\partial \theta} &= \sum_{t=1}^{T}\sum_{j=1}^{J} I(Y_{it}=j)\frac{(\phi(h_{it}^{j-1})-\phi(h_{it}^{j}))\bar{Z}_i'}{\Phi(h_{it}^{j})-\Phi(h_{it}^{j-1})}, \\
\frac{\partial l_i(\alpha_1,\alpha_2)}{\partial \rho} &= \sum_{t=1}^{T}\sum_{j=1}^{J} I(Y_{it}=j)\frac{(\phi(h_{it}^{j-1})-\phi(h_{it}^{j}))v_{it}}{\Phi(h_{it}^{j})-\Phi(h_{it}^{j-1})}, \\
\frac{\partial l_i(\alpha_1,\alpha_2)}{\partial \eta_j} &= \sum_{t=1}^{T} I(Y_{it}=j)\frac{\phi(h_{it}^{j})}{\Phi(h_{it}^{j})-\Phi(h_{it}^{j-1})} - I(Y_{it}=j+1)\frac{\phi(h_{it}^{j})}{\Phi(h_{it}^{j+1})-\Phi(h_{it}^{j})}, \quad (11)
\end{aligned}$$

where $h_{it}^{j} \equiv \eta_j - X_{1it}\beta_1 - Y_{2it}\beta_2 - \bar{Z}_i\theta - v_{it}\rho$, and $v_{it} \equiv Y_{2it} - Z_{it}\gamma - \bar{Z}_i\lambda$.

Suppose we obtain $\hat{v}_{it}$ by pooled OLS, which minimizes the sum of pooled squared errors $\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{2it} - Z_{it}\gamma - \bar{Z}_i\lambda)^2$. Let $sse_i(\alpha_2) \equiv \sum_{t=1}^{T}(Y_{2it} - Z_{it}\gamma - \bar{Z}_i\lambda)^2$. The derivative of $sse_i(\alpha_2)$ with respect to α_2 is

$$\begin{aligned}
\frac{\partial sse_i(\alpha_2)}{\partial \gamma} &= -2\sum_{t=1}^{T}(Y_{2it} - Z_{it}\gamma - \bar{Z}_i\lambda)Z_{it}', \\
\frac{\partial sse_i(\alpha_2)}{\partial \lambda} &= -2\sum_{t=1}^{T}(Y_{2it} - Z_{it}\gamma - \bar{Z}_i\lambda)\bar{Z}_i'. \quad (12)
\end{aligned}$$

Define $g_i(\alpha_1,\alpha_2)$ to be $(\partial l_i(\alpha_1,\alpha_2)'/\partial\alpha_1, \partial sse_i(\alpha_2)'/\partial\alpha_2)'$. Assume there exists a true value $\alpha^* \in \Theta$. We proceed by showing $E(g_i(\alpha_1^*,\alpha_2^*)) = 0$.

If Z_i is strictly exogenous in (4), $\alpha_2^* \equiv (\gamma^{*'}, \lambda^{*'})'$ satisfies

$$E((Y_{2it} - Z_{it}\gamma^* - \bar{Z}_i\lambda^*)(Z_{it}, \bar{Z}_i)') = 0 \tag{13}$$

for each t. Therefore, $E(\partial sse_i(\alpha_2^*)/\partial\alpha_2) = 0$. For $\partial l_i(\alpha_1,\alpha_2)/\partial\alpha_1$, we only consider $\partial l_i(\alpha_1,\alpha_2)/\partial\beta_2$, which is a scalar.

$$\begin{aligned}
E(\frac{\partial l_i(\alpha_1,\alpha_2)}{\partial\beta_2}) &= \sum_{t=1}^{T}\sum_{j=1}^{J} E(I(Y_{it}=j)\frac{(\phi(h_{it}^{j-1}) - \phi(h_{it}^{j}))Y_{2it}}{\Phi(h_{it}^{j}) - \Phi(h_{it}^{j-1})}) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J} E(E(I(Y_{it}=j)|Z_i, Y_{2it})\frac{(\phi(h_{it}^{j-1}) - \phi(h_{it}^{j}))Y_{2it}}{\Phi(h_{it}^{j}) - \Phi(h_{it}^{j-1})}) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J} E(P(Y_{it}=j|Z_i, Y_{2it})\frac{(\phi(h_{it}^{j-1}) - \phi(h_{it}^{j}))Y_{2it}}{\Phi(h_{it}^{j}) - \Phi(h_{it}^{j-1})}) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J} E((\Phi(h_{it}^{*j}) - \Phi(h_{it}^{*j-1}))\frac{(\phi(h_{it}^{j-1}) - \phi(h_{it}^{j}))Y_{2it}}{\Phi(h_{it}^{j}) - \Phi(h_{it}^{j-1})}),
\end{aligned} \tag{14}$$

where $h_{it}^{*j} \equiv \eta_j^* - X_{1it}\beta_1^* - Y_{2it}\beta_2^* - \bar{Z}_i\theta^* - v_{it}^*\rho^*$, and $v_{it}^* \equiv Y_{2it} - Z_{it}\gamma^* - \bar{Z}_i\lambda^*$. The second equality of (14) is due to the law of iterated expectations, and the last equality is true since $\alpha_1^* \equiv (\beta_1^{*'}, \beta_2^*, \theta^{*'}, \rho^*, \eta^{*'})'$ is the true parameters. Evaluated at (α_1^*,α_2^*), (14) reduces to

$$E(\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial\beta_2}) = \sum_{t=1}^{T}\sum_{j=1}^{J} E((\phi(h_{it}^{*j-1}) - \phi(h_{it}^{*j}))Y_{2it}). \tag{15}$$

For each t, let the marginal distribution of (Z_i, Y_{2it}) is $f_t^*(Z_i, Y_{2it})$ if it exists. The joint distribution of (Y_{it}, Z_i, Y_{2it}) is the product of $f_t^*(Z_i, Y_{2it})$ and the conditional distribution of Y_{it} given (Z_i, Y_{2it}), i.e.

$$p_t^*(Y_{it}, Z_i, Y_{2it}) \equiv f_t^*(Z_i, Y_{2it})\prod_{j=1}^{J}(\Phi(h_{it}^{*j}) - \Phi(h_{it}^{*j-1}))^{I(Y_{it}=j)}. \tag{16}$$

For any other $\alpha \in \Theta$, we can define $p_t(Y_{it}, Z_i, Y_{2it})$ similarly by replacing α^* in (16) with α. It is easy to verify that $p_t(Y_{it}, Z_i, Y_{2it})$ is also a valid distribution. Define the Kullback-Leibler

information criterion (KLIC) as

$$\psi_t(p_t^*, p_t) \equiv E(log(\frac{p_t^*(Y_{it}, Z_i, Y_{2it})}{p_t(Y_{it}, Z_i, Y_{2it})})), \tag{17}$$

where $E(\cdot)$ is the expectational operator with respect to the joint distribution of (Y_{it}, Z_i, Y_{2it}). An important result about KLIC is that $\psi_t(p_t^*, p_t) \geq 0$ if $p_t^*(Y_{it}, Z_i, Y_{2it})$ is the true distribution. In other words, $E(log(p_t^*(Y_{it}, Z_i, Y_{2it}))) \geq E(log(p_t(Y_{it}, Z_i, Y_{2it})))$. As a result, α^* maximizes $E(log(p_t(Y_{it}, Z_i, Y_{2it})))$ over Θ. Since

$$E(log(p_t(Y_{it}, Z_i, Y_{2it}))) = E(log(f_t^*(Z_i, Y_{2it}))) + E(\sum_{j=1}^{J} I(Y_{it} = j) log(\Phi(h_{it}^j) - \Phi(h_{it}^{j-1}))), \tag{18}$$

α^* also maximizes $E(\sum_{j=1}^{J} I(Y_{it} = j) log(\Phi(h_{it}^j) - \Phi(h_{it}^{j-1})))$. By the law of iterated expectations,

$$E(\sum_{j=1}^{J} I(Y_{it} = j) log(\Phi(h_{it}^j) - \Phi(h_{it}^{j-1}))) = \sum_{j=1}^{J} E((\Phi(h_{it}^{*j}) - \Phi(h_{it}^{*j-1})) log(\Phi(h_{it}^j) - \Phi(h_{it}^{j-1}))).$$

A necessary condition for the maximization implies that

$$\sum_{j=1}^{J} \frac{\partial E((\Phi(h_{it}^{*j}) - \Phi(h_{it}^{*j-1})) log(\Phi(h_{it}^{*j}) - \Phi(h_{it}^{*j-1})))}{\partial \alpha} = 0 \tag{19}$$

if α^* lies in the interior of Θ.

To justify the interchange of expectation and derivative in (19), we use the Generalized Mean-Value Theorem, which is stated in Glasserman (1991). Again, we only consider the derivative with respect to β_2. Without loss of generality, $(\Phi(h_{it}^{*j}) - \Phi(h_{it}^{*j-1}))$ can be ignored since it is bounded between 0 and 1 and is independent of α. Look at the univariate real-valued function $d_{it}^*(\beta_2)$, defined as

$$log(\Phi(\eta_j^* - X_{1it}\beta_1^* - Y_{2it}\beta_2 - \bar{Z}_i\theta^* - v_{it}^*\rho^*) - \Phi(\eta_{j-1}^* - X_{1it}\beta_1^* - Y_{2it}\beta_2 - \bar{Z}_i\theta^* - v_{it}^*\rho^*)). \tag{20}$$

(20) is differentiable for any $\beta_2 \in R$. In particular, (20) is differentiable on a bounded closed interval $[a, b]$, where $a \leq \inf_{\alpha \in \Theta} \beta_2$ and $b \geq \sup_{\alpha \in \Theta} \beta_2$. By the Generalized Mean-Value

Theorem, we have

$$|\frac{d_{it}^*(\beta_2^*+h)-d_{it}^*(\beta_2^*)}{h}| \leq \sup_{\beta_2\in[a,b]}|\frac{dd_{it}^*(\beta_2)}{d\beta_2}|, \tag{21}$$

where $\beta_2^*+h \in [a,b]$. $E(\sup_{\beta_2\in[a,b]}|dd_{it}^*(\beta_2)/d\beta_2|)$ is shown to be finite by the argument in (33). It then follows from dominated convergence that

$$\begin{aligned}
&\frac{\partial E((\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))log(\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1})))}{\partial\beta_2} \\
= &\lim_{h\to 0} E(\frac{(\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))(d_{it}^*(\beta_2^*+h)-d_{it}^*(\beta_2^*))}{h}) \\
= &E((\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))\lim_{h\to 0}\frac{d_{it}^*(\beta_2^*+h)-d_{it}^*(\beta_2^*)}{h}) \\
= &E((\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))\frac{dd_{it}^*(\beta_2^*)}{d\beta_2}) \\
= &E((\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))\frac{\partial log(\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))}{\partial\beta_2}).
\end{aligned} \tag{22}$$

The second equality in (22) is true because

$$E((\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1}))\sup_{\beta_2\in[a,b]}|\frac{dd_{it}^*(\beta_2)}{d\beta_2}|) \leq E(\sup_{\beta_2\in[a,b]}|\frac{dd_{it}^*(\beta_2)}{d\beta_2}|) < \infty. \tag{23}$$

Since (22) holds for any j, we obtain the moment condition

$$\sum_{j=1}^{J} E((\phi(h_{it}^{*j-1})-\phi(h_{it}^{*j}))Y_{2it}) = 0. \tag{24}$$

Taking summation of (24) over t, we have the desired result $E(\partial l_i(\alpha_1^*,\alpha_2^*)/\partial\beta_2) = 0$. To summarize, $E(g_i(\alpha_1^*,\alpha_2^*)) = 0$.

Now, we show (α_1^*,α_2^*) is uniquely determined and thus is identified. To identify α_2^*, the following two conditions are sufficient:

$$E(v_{it}|Z_i) = 0 \text{ and } rank\ E([I_T, j_T(j_T'j_T)^{-1}j_T']Z_i) = 2M, \tag{25}$$

where I_T is T-dimensional identity matrix, and j_T is $T\times 1$ vector of ones. (25) implies Z_i is

strictly exogenous in (4) and the columns of $[I_T, j_T(j_T'j_T)^{-1}j_T']Z_i$ must be linearly independent. Importantly, this rules out time-invariant elements in Z_i.

Identification of α_1^* requires more effort. Under conditions in (25), α_2^* is identified and thus can be taken to be known. Since $P(Y_{it}=j|Z_i,Y_{2it})$ is specified as $\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1})$, it is necessary to verify that there is no other α_1, coupled with α_2^*, in Θ such that $P(Y_{it}=j|Z_i,Y_{2it})=\Phi(\tilde{h}_{it}^{j})-\Phi(\tilde{h}_{it}^{j-1})$. Here $\tilde{h}_{it}^{j}$ is defined as h_{it}^{j} with v_{it} replaced by v_{it}^*. Consider $j=1$ first. $P(Y_{it}=1|Z_i,Y_{2it})=\Phi(h_{it}^{*1})$. Suppose there exists another α_1 such that $\Phi(h_{it}^{*1})=\Phi(\tilde{h}_{it}^{1})$ almost surely and at least one element of α_1 differs from the corresponding element of α_1^*. We must have

$$\eta_1^*-\eta_1=X_{1it}(\beta_1^*-\beta_1)+Y_{2it}(\beta_2^*-\beta_2)+\bar{Z}_i(\theta^*-\theta)+v_{it}^*(\rho^*-\rho) \tag{26}$$

almost surely. Because the left hand side of (26) is a constant, the variance of right hand side must be zero, i.e. $((\beta_1^*-\beta_1)',\beta_2^*-\beta_2,(\theta^*-\theta)',\rho^*-\rho)VarQ_{it}((\beta_1^*-\beta_1)',\beta_2^*-\beta_2,(\theta^*-\theta)',\rho^*-\rho)'=0$, where $Q_{it}\equiv(X_{1it},Y_{2it},\bar{Z}_i,v_{it}^*)$. If $VarQ_{it}$ is nonsingular, then $((\beta_1^*-\beta_1)',\beta_2^*-\beta_2,(\theta^*-\theta)',\rho^*-\rho)=0$, which implies $\eta_1^*=\eta_1$ by (26).

For $j=2$, $P(Y_{it}=2|Z_i,Y_{2it})=\Phi(h_{it}^{*2})-\Phi(h_{it}^{*1})$. By the preceding argument, $\Phi(h_{it}^{*1})=\Phi(\tilde{h}_{it}^{1})$ and both α^* and α share the same coefficients of Q_{it}. Hence,

$$\eta_2^*-\eta_2=X_{1it}(\beta_1^*-\beta_1)+Y_{2it}(\beta_2^*-\beta_2)+\bar{Z}_i(\theta^*-\theta)+v_{it}^*(\rho^*-\rho)=0. \tag{27}$$

Iterating the process until $j=J$, we have $\alpha=\alpha^*$. To put it differently, α^* is the unique vector in Θ such that $P(Y_{it}=j|Z_i,Y_{2it})=\Phi(h_{it}^{*j})-\Phi(h_{it}^{*j-1})$ for any j. The key restriction for identification of α_1 is the nonsingularity of $VarQ_{it}$, which is true if and only if there is no linear relationship between the variables in Q_{it}. In particular, we require X_{2it} and $\bar{X}_{2i}$ in (4) to be partially correlated with Y_{2it} once X_{1it} and $\bar{X}_{1i}$ are controlled for. To summarize, conditions in (25) and nonsingularity of $VarQ_{it}$ for each t are sufficient for identification of α^* in the system.

However, identification is not sufficient for uniqueness of α_1^* as the solution to $E(\partial l_i(\alpha_1,\alpha_2^*)/\partial\alpha_1)=0$. Nevertheless, Lemma 2.2 in Newey and McFadden (1994) establishes the equivalence of identification and uniqueness of α_1^* as the maximizer of

$E(l_i(\alpha_1, \alpha_2^*))$. Pratt (1981) proved that $l_i(\alpha_1, \alpha_2^*)$, as a function of α_1, is strictly concave. For $\tau \in (0,1)$, let α_1^1 and α_1^2 be two distinct vectors lying in a convex set containing Θ. It follows from the definition of strict concavity that

$$l_i(\tau\alpha_1^1 + (1-\tau)\alpha_1^2, \alpha_2^*) - \tau l_i(\alpha_1^1, \alpha_2^*) - (1-\tau) l_i(\alpha_1^2, \alpha_2^*) > 0. \tag{28}$$

Problem 19 of Section 18.2 in Royden and Fitzpatrick (2010) implies

$$E(l_i(\tau\alpha_1^1 + (1-\tau)\alpha_1^2, \alpha_2^*) - \tau l_i(\alpha_1^1, \alpha_2^*) - (1-\tau) l_i(\alpha_1^2, \alpha_2^*)) > 0. \tag{29}$$

This demonstrates that $E(l_i(\alpha_1, \alpha_2^*))$ is strictly concave in α_1, so α_1^* uniquely solves $E(\partial l_i(\alpha_1, \alpha_2^*)/\partial\alpha_1) = 0$. Or, we can interpret α^* as the unique solution to

$$\min_{\alpha\in\Theta} E'(g_i(\alpha_1, \alpha_2)) I_R E(g_i(\alpha_1, \alpha_2)), \tag{30}$$

where I_R is the identity matrix of dimension $R \equiv K_1 + 3M + J + 1$. Since the number of parameters in α is equal to the number of moment restrictions, selection of weighting matrix does not matter. We choose I_R as the weighting matrix for simplicity. $\hat{\alpha}$ solves the sample counterpart of (30)

$$\min_{\alpha\in\Theta} (\sum_{i=1}^N g_i(\alpha_1, \alpha_2))' I_R (\sum_{i=1}^N g_i(\alpha_1, \alpha_2)). \tag{31}$$

Since I_R is positive definite, (31) is equivalent to solving for the root of the equation $\sum_{i=1}^N g_i(\alpha_1, \alpha_2) = 0$, which is approximately the two-step estimator $\hat{\alpha}$ in large samples. Suppose $\{(Z_i, Y_i, Y_{2i}) : i = 1, 2, ...\}$ is independently and identically distributed. It is straightforward to check that conditions of Theorem 2.6 in Newey and McFadden (1994), except the uniform integrability of $g_i(\alpha_1, \alpha_2)$, are satisfied. We are going to argue that the uniform integrability also holds.

As before, we only consider $\partial l_i(\alpha_1, \alpha_2)/\partial\beta_2$ in (11). For any j in $\{2, ..., J-1\}$, $\Phi(h_{it}^j) - \Phi(h_{it}^{j-1}) = \int_{h_{it}^{j-1}}^{h_{it}^j} \phi(x)dx$ and $h_{it}^j - h_{it}^{j-1} = \eta_j - \eta_{j-1} > 0$. If a lower value of $\int_{h_{it}^{j-1}}^{h_{it}^j} \phi(x)dx$ is desirable, we hope $\eta_j - \eta_{j-1}$ could be as small as possible. Given the compactness of Θ,

there exist two values η_j^o and η_{j-1}^o in Θ with $\eta_j^o > \eta_{j-1}^o$ such that $\eta_j^o - \eta_{j-1}^o$ is the smallest. Similarly, we can find the maximum of $|\beta_1|$, $|\beta_2|$, $|\theta|$, $|\rho|$, $|\gamma|$ and $|\lambda|$ over Θ, denoted by β_1^o, β_2^o, θ^o, ρ^o, γ^o and λ^o, respectively.[7]

Let $h_{it}^{jo} \equiv |\eta_j^o + \eta_{j-1}^o|/2 + (\eta_j^o - \eta_{j-1}^o)/2 + |X_{1it}|\beta_1^o + |Y_{2it}|\beta_2^o + |\bar{Z}_i|\theta^o + v_{it}^o\rho^o$, $h_{it}^{j-1o} \equiv |\eta_j^o + \eta_{j-1}^o|/2 - (\eta_j^o - \eta_{j-1}^o)/2 + |X_{1it}|\beta_1^o + |Y_{2it}|\beta_2^o + |\bar{Z}_i|\theta^o + v_{it}^o\rho^o$, and $v_{it}^o \equiv |Y_{2it}| + |Z_{it}|\gamma^o + |\bar{Z}_i|\lambda^o$. Since $h_{it}^{jo} - h_{it}^{j-1o} = h_{it}^j - \eta^j + \eta_j^o - (h_{it}^{j-1} - \eta^{j-1} + \eta_{j-1}^o) = \eta_j^o - \eta_{j-1}^o$, and $|h_{it}^{j-1o} + (\eta_j^o - \eta_{j-1}^o)/2| \geq |h_{it}^{j-1} - \eta^{j-1} + \eta_{j-1}^o + (\eta_j^o - \eta_{j-1}^o)/2|$, it then follows from the symmetry of $\phi(\cdot)$ around 0 that

$$\Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o}) \leq \Phi(h_{it}^j - \eta^j + \eta_j^o) - \Phi(h_{it}^{j-1} - \eta^{j-1} + \eta_{j-1}^o). \tag{32}$$

Moreover, $\Phi(h_{it}^j - \eta^j + \eta_j^o) - \Phi(h_{it}^{j-1} - \eta^{j-1} + \eta_{j-1}^o) \leq \Phi(h_{it}^j) - \Phi(h_{it}^{j-1})$ because $\eta_j^o - \eta_{j-1}^o$ achieves the minimum. As a result, $0 < \Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o}) \leq \Phi(h_{it}^j) - \Phi(h_{it}^{j-1})$ for any $\alpha \in \Theta$. If $E(1/(\Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o}))^2) < \infty$, then

$$\begin{aligned}
E(|\frac{(\phi(h_{it}^{j-1}) - \phi(h_{it}^j))Y_{2it}}{\Phi(h_{it}^j) - \Phi(h_{it}^{j-1})}|) &= E(\frac{|\phi(h_{it}^{j-1}) - \phi(h_{it}^j)||Y_{2it}|}{\Phi(h_{it}^j) - \Phi(h_{it}^{j-1})}) \\
&\leq \frac{1}{\sqrt{2\pi}} E(\frac{|Y_{2it}|}{\Phi(h_{it}^j) - \Phi(h_{it}^{j-1})}) \\
&\leq \frac{1}{\sqrt{2\pi}} E(\frac{|Y_{2it}|}{\Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o})}) \\
&\leq \frac{1}{\sqrt{2\pi}} E^{1/2}(\frac{1}{(\Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o}))^2}) E^{1/2}(Y_{2it}^2),
\end{aligned} \tag{33}$$

where the first inequality is true because $0 < \phi(x) \leq 1/\sqrt{2\pi}$ for any $x \in R$, and the third one is Cauchy-Schwarz inequality. Since $E(Y_{2it}^2) < \infty$, the uniform integrability of $(\phi(h_{it}^{j-1}) - \phi(h_{it}^j))Y_{2it}/(\Phi(h_{it}^j) - \Phi(h_{it}^{j-1})$ follows.

To ensure $E(1/(\Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o}))^2) < \infty$, the finite second moment assumption we have made is not enough. Note that $\Phi(h_{it}^{jo}) - \Phi(h_{it}^{j-1o}) = \phi(\tilde{m}_{it}^j)(\eta_j^o - \eta_{j-1}^o)$, where $\tilde{m}_{it}^j$ is a

[7] $|\beta_1|$ is the vector generated by taking absolute value of each element in β_1.

mean value between h_{it}^{jo} and h_{it}^{j-1o}. We have

$$\begin{aligned} E(1/(\Phi(h_{it}^{jo})-\Phi(h_{it}^{j-1o}))^2) &= \frac{1}{(\eta_j^o-\eta_{j-1}^o)^2}E(\frac{1}{\phi^2(\tilde{m}_{it}^j)}) \\ &= \frac{2\pi}{(\eta_j^o-\eta_{j-1}^o)^2}E(exp(\tilde{m}_{it}^{j2})) \\ &\leq \frac{2\pi}{(\eta_j^o-\eta_{j-1}^o)^2}E(exp(|h_{it}^{j-1o}|+\eta_j^o-\eta_{j-1}^o)^2). \quad (34) \end{aligned}$$

A sufficient condition for the finiteness of (34) is that the squared random variables are exponentially integrable, i.e. $E(exp(Y_{2it})^2) < \infty$. This is much stronger than existence of the second moment because the latter is implied by the former by Jensen's inequality.

When $j = 1$, $\Phi(h_{it}^j) - \Phi(h_{it}^{j-1}) = \Phi(h_{it}^1)$. We can verify the uniform integrability of $-\phi(h_{it}^1)Y_{2it}/\Phi(h_{it}^1)$ as before. The same reasoning applies when $j = J$. Finally, the uniform integrability of $\partial l_i(\alpha_1,\alpha_2)/\partial\beta_2$ follows since each element in the sum is uniformly integrable.

Let us look at $\partial sse_i(\alpha_2)/\partial\gamma$. By (12),

$$\begin{aligned} |\frac{\partial sse_i(\alpha_2)}{\partial\gamma_1}| &= 2|\sum_{t=1}^T (Y_{2it}-Z_{it}\gamma-\bar{Z}_i\lambda)Z_{1it}| \\ &\leq 2\sum_{t=1}^T |(Y_{2it}-Z_{it}\gamma-\bar{Z}_i\lambda)Z_{1it}| \\ &\leq 2\sum_{t=1}^T (|Y_{2it}|+|Z_{it}||\gamma|+|\bar{Z}_i||\lambda|)|Z_{1it}| \\ &\leq 2\sum_{t=1}^T (|Y_{2it}|+|Z_{it}|\gamma^o+|\bar{Z}_i|\lambda^o)|Z_{1it}|, \quad (35) \end{aligned}$$

where γ_1 is the first element of γ, and Z_{1it} is the first element of Z_{it}. $\partial sse_i(\alpha_2)/\partial\gamma$ is uniformly integrable by the finiteness of the second moments. Given the uniform integrability of $g_i(\alpha_1,\alpha_2)$, consistency of $\hat{\alpha}$ is established by Theorem 2.6 in Newey and McFadden (1994).

To show asymptotic normality, we use Theorem 3.4 in Newey and McFadden (1994). Again, all regularity conditions hold for the current case,[8] and it follows that

$$\sqrt{N}(\hat{\alpha}-\alpha^*) \xrightarrow{d} N(0,V), \quad (36)$$

[8]The uniform boundedness is verified following the same way that we derived consistency. The details are available upon request.

where

$$V \equiv E^{-1}(\frac{\partial g_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha})E(g_i(\alpha_1^*,\alpha_2^*)g_i(\alpha_1^*,\alpha_2^*)')E^{-1}(\frac{\partial g_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha})'. \tag{37}$$

If we partition $g_i(\alpha_1^*,\alpha_2^*)$ into two sub-vectors $(\partial l_i(\alpha_1^*,\alpha_2^*)'/\partial\alpha_1, \partial sse_i(\alpha_2^*)'/\partial\alpha_2)'$, $E(\partial g_i(\alpha_1^*,\alpha_2^*)/\partial\alpha)$ can be written as

$$\begin{pmatrix} E(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_1'}) & E(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_2'}) \\ 0 & E(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial\alpha_2\partial\alpha_2'}) \end{pmatrix}. \tag{38}$$

The inverse of $E(\partial g_i(\alpha_1^*,\alpha_2^*)/\partial\alpha)$ is thus

$$\begin{pmatrix} E^{-1}(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_1'}) & -E^{-1}(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_1'})E(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_2'})E^{-1}(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial\alpha_2\partial\alpha_2'}) \\ 0 & E^{-1}(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial\alpha_2\partial\alpha_2'}). \end{pmatrix}. \tag{39}$$

Likewise,

$$E(g_i(\alpha_1^*,\alpha_2^*)g_i(\alpha_1^*,\alpha_2^*)') = \begin{pmatrix} E(\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1}\frac{\partial l_i(\alpha_1^*,\alpha_2^*)'}{\partial\alpha_1}) & E(\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1}\frac{\partial sse_i(\alpha_2^*)'}{\partial\alpha_2}) \\ E(\frac{\partial sse_i(\alpha_2^*)}{\partial\alpha_2}\frac{\partial l_i(\alpha_1^*,\alpha_2^*)'}{\partial\alpha_1}) & E(\frac{\partial sse_i(\alpha_2^*)}{\partial\alpha_2}\frac{\partial sse_i(\alpha_2^*)'}{\partial\alpha_2}). \end{pmatrix}. \tag{40}$$

Plugging (39) and (40) into (37), we get the asymptotic variance of $\sqrt{N}(\hat{\alpha}_1 - \alpha_1^*)$

$$V_1 \equiv E^{-1}(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_1'})ME^{-1}(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial\alpha_1\partial\alpha_1'}), \tag{41}$$

where

$$
\begin{aligned}
M \equiv \quad & E\left(\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1}\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1}'\right) \\
& -E\left(\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1}\frac{\partial sse_i(\alpha_2^*)}{\partial \alpha_2}'\right)E^{-1}\left(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial \alpha_2 \partial \alpha_2'}\right)E\left(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)'}{\partial \alpha_1 \partial \alpha_2'}\right) \\
& -E\left(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1 \partial \alpha_2'}\right)E^{-1}\left(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial \alpha_2 \partial \alpha_2'}\right)E\left(\frac{\partial sse_i(\alpha_2^*)}{\partial \alpha_2}\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1}'\right) \\
& +E\left(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1 \partial \alpha_2'}\right)E^{-1}\left(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial \alpha_2 \partial \alpha_2'}\right)E\left(\frac{\partial sse_i(\alpha_2^*)}{\partial \alpha_2}\frac{\partial sse_i(\alpha_2^*)}{\partial \alpha_2}'\right) \\
& E^{-1}\left(\frac{\partial^2 sse_i(\alpha_2^*)}{\partial \alpha_2 \partial \alpha_2'}\right)E\left(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)'}{\partial \alpha_1 \partial \alpha_2'}\right). \qquad (42)
\end{aligned}
$$

If all terms except the first one are ignored, V_1 becomes

$$
E^{-1}\left(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1 \partial \alpha_1'}\right)E\left(\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1}\frac{\partial l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1}'\right)E^{-1}\left(\frac{\partial^2 l_i(\alpha_1^*,\alpha_2^*)}{\partial \alpha_1 \partial \alpha_1'}\right), \qquad (43)
$$

which is the usual robust asymptotic variance of the maximum likelihood estimator $\hat{\alpha}_1$ at the second stage. However, (43) is generally incorrect because it ignores the uncertainty of estimating $\hat{v}_{it}$. As argued by Newey and McFadden (1994), there are important special cases where (41) and (43) are equivalent. For example, if $E(\partial^2 l_i(\alpha_1^*,\alpha_2^*)/\partial\alpha_1\partial\alpha_2') = 0$, (41) reduces to (43). Unfortunately, this is not the case for the ordered probit model. By (11), it is easy to see that $E(\partial^2 l_i(\alpha_1^*,\alpha_2^*)/\partial\alpha_1\partial\alpha_2')$ is generally not zero, and the estimation uncertainty at the first stage must be taken into account. One important exception occurs when $\rho^* = 0$. This is true if and only if $Cov(r_{it}, v_{it}) = 0$, i.e. r_{it} and v_{it} are independent of each other by (5). As a consequence, Y_{2it} is exogenous and the usual maximum likelihood estimator of α_1^* is well-behaved. This implies that we can carry out the two-step procedure to get $\hat{\alpha}_1$ as before and use the standard t statistic to test the significance of $\hat{\rho}$. Specifically, we use the square root of the diagonal term in (43) corresponding to ρ as the denominator of the t statistic. Note that (43) is the variance matrix of sandwich form, which is robust to possible dynamic misspecification in partially specified panel data model. In this setting, the information equality, which states that $E^{-1}(-\partial^2 l_i(\alpha_1^*,\alpha_2^*)/\partial\alpha_1\partial\alpha_1') = E(\partial l_i(\alpha_1^*,\alpha_2^*)/\partial\alpha_1 \partial l_i(\alpha_1^*,\alpha_2^*)'/\partial\alpha_1)$, fails to hold, and (43) should be used instead of the simplified version $E(\partial l_i(\alpha_1^*,\alpha_2^*)/\partial\alpha_1 \partial l_i(\alpha_1^*,\alpha_2^*)'/\partial\alpha_1)$ that

is reported regularly in most statistical packages.

All unknown components in (41) can be estimated by their sample counterparts with the two-step estimator $\hat{\alpha}$ in place of α^*. For instance, $E(\partial^2 l_i(\alpha_1^*, \alpha_2^*)/\partial\alpha_1\partial\alpha_2')$ is estimated by $(\sum_{i=1}^{N} \partial^2 l_i(\hat{\alpha}_1, \hat{\alpha}_2)/\partial\alpha_1\partial\alpha_2')/N$. The consistency of these variance matrix estimators follows from Lemma 4.3 in Newey and McFadden (1994). Another convenient way to derive the approximate variance matrix is thought bootstrapping. Although there is no formal justification in the current case, we expect that the bootstrap is asymptotically valid given the smoothness of the objective function. Note that a bootstrap sample is generated by random sampling in the cross section dimension, i.e. drawing all time periods for a particular unit, if N, relative to T, is large enough.

References

Andersen, E. B. (1970), 'Asymptotic Properties of Conditional Maximum-Likelihood Estimators', *Journal of the Royal Statistical Society, Series B* **32**, 283–301.

Blundell, R. and Powell, J. L. (2003), Endogeneity in Nonparametric and Semiparametric Regression Models, *in* M. Dewatripont, L. P. Hansen and S. J. Turnovsky, eds, 'Advances in Economics and Econometrics', Cambridge University Press, pp. 312–357.

Blundell, R. and Powell, J. L. (2004), 'Endogeneity in Semiparametric Binary Response Models', *Review of Economic Studies* **71**, 655–679.

Cameron, A. C. and Trivedi, P. K. (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press.

Case, A., Lubotsky, D. and Paxson, C. (2002), 'Economic Status and Health in Childhood: The Origins of the Gradien', *The American Economic Review* **92**, 1308–1334.

Chamberlain, G. (1980), 'Analysis of Covariance with Qualitative Data', *Review of Economic Studies* **47**, 225–238.

Chatterji, P., Lahiri, K. and Song, J. (2013), 'The Dynamics of Income-related Health Inequality among American Children', *Health Economics* **22**, 623–629.

Chatterji, P., Lahiri, K. and Kim, D. (2014), 'Birthweight and Academic Achievement in Childhood', *Health Economics*, **23**, 1013–1035.

Chen, Y., Le, X. and Zhou, L. (2017), 'Does Raising Family Income Cause Better Child Health? Empirical Evidence from China', *Economic Development and Cultural Change* **65**, 495–520.

Chesher, A. and Smolinski, K. (2012), 'IV Models of Ordered Choice', *Journal of Econometrics* **166**, 33–48.

Courtemanche, C., Tchernis, R. and Zhou, X. (2017), Parental Work Hours and Childhood Obesity: Evidence Using Instrumental Variables Related to Sibling School Eligibility. NBER Working Paper No. 23376.

Condliffe, S. and Link, C. R. (2008), 'The Relationship between Economic Status and Child Health: Evidence from the United States', *The American Economic Review* **98**, 1605–1618.

Currie, A., Shields, M. A. and Price, S. W. (2007), 'The Child Health/Family Income Gradient: Evidence from England', *Journal of Health Economics* **26**, 213–232.

Currie, J. and Stabile, M. (2003), 'Socioeconomic Status and Child Health: Why is the Relationship Stronger for Older Children?', *The American Economic Review* **93**, 1813–1823.

Das, M. and van Soest, A. (1999), 'A Panel Data Model for Subjective Information on Household Income Growth', *Journal of Economic Behavior & Organization* **40**, 409–426.

Dave, D. M. and Yang, M. (2019), Maternal and Fetal Health Effects of Working During Pregnency. NBER Working Paper No. 26343.

Davillas, A., Jones, A. M. and Benezeval, M. (2019), The Income-Health Gradient: Evidence from Self-Reported health and Biomarkers in Understanding Society, *in* M. Tsionas, eds, 'Panel Data Econometrics', Elsevier Amsterdam, pp. 709–741.

Doyle, O., Harmon, C. and Walker, I. (2007), The Impact of Parental Income and Education on Child Health: Further Evidence for England. Warwick Economics Research Papers, No. 788.

Fernald, L. C. H., Kariger, P., Hildrobo, M. and Gertler, P. J. (2012), 'Socioeconomic Gradients in Child Development in Very Young Childhood: Evidence from India, Indonesia, Peru and Senegal', *Proceedings of the National Academy of Sciences* **109**, 17273–17280.

Fletcher, J. and Wolfe, B. L. (2014), 'Increasing our Understanding of the Health-Income Gradient in Children', *Health Economics* **23**, 473–486.

Giles, J. and Murtazashvili, I. (2013), 'A Control Function Approach to Estimating Dynamic Probit Models with Endogenous Regressors', *Journal of Econometric Methods* **2**, 69–87.

Glasserman, P. (1991), *Gradient Estimation via Perturbation Analysis*, Kluwer Academic.

Greene, W. H. and Hensher, D. A. (2010), *Modeling Ordered Choices: A Primer*, Cambridge University Press.

Hausman, J. A. and Taylor, W. E. (1981), 'Panel Data and Unobservable Individual Effects', *Econometrica* **49**, 1377–1398.

Heckman, J. J. (1979), 'Sample Selection Bias as a Specification Error', *Econometrica* **47**, 153-161.

Karaca-Mandic, P., Norton, E. C. and Dowd, B. (2012), 'Interaction Terms in Nonlinear Models', *Health Services Research* **47**, 255–274.

Kawakatsu, H. and Largey, A. G. (2009), 'EM Algorithms for Ordered Probit Models with Endogenous Regressors', *The Econometrics Journal* **12**, 164–186.

Khanam, R., Nghiem, H. S. and Connelly, L. B. (2009), 'Child Health and the Income Gradient: Evidence from Australia', *Journal of Health Economics* **28**, 805–817.

Klein, R. and Vella, F. (2010), 'Estimating a Class of Triangular Simultaneous Equations Models without Exclusion Restrictions', *Journal of Econometrics* **154**, 154–164.

Kuehnle, D. (2014), 'The Causal Effect of Family Income on Child Health in the UK', *Journal of Health Economics* **36**, 137-150.

Lewbel, A. (2012), 'Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models', *Journal of Business & Economic Statistics* **30**, 67–80.

Lewbel, A. and Dong, Y. (2015), 'A Simple Estimator for Binary Choice Models with Endogenous Regressors', *Econometric Reviews* **34**, 82–105.

Lucas-Thompson, R. G., Goldberg, W. A. and Prause, J. (2010), 'Maternal Work Early in the Lives of Children and its Distal Associations with Achievement and Behavior Problems: A Meta Analysis', *Psychological Bulletin* **136**, 915–942.

Maddala, G. S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

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

Munkin, M. K. and Trivedi, P. K. (2008), 'Bayesian Analysis of the Ordered Probit Model with Endogenous Selection', *Journal of Econometrics* **143**, 334–348.

Newey, W. K. and McFadden, D. (1994), Large Sample Estimation and Hypothesis Testing, *in* R. F. Engle and D. McFadden, eds, 'Handbook of Econometrics', North-Holland Amsterdam, pp. 2111–2245.

Nicholson, J. M., Strazdins, L., Brown, J. E. and Bittman, M. (2012), 'How Parents' Income, Time and Job Quality Affect Children's Health and Development', *Australian Journal of Social Issues* **47**, 505–525.

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

Pratt, J. W. (1981), 'Concavity of the Log Likelihood', *Journal of the American Statistical Association* **76**, 103–106.

Reinhold, S. and Jürges, H. (2012), 'Parental Income and Child Health in Germany', *Health Economics* **21**, 562–579.

Rivers, D. and Vuong, Q. H. (1988), 'Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models', *Journal of Econometrics* **39**, 347–366.

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

Royden, H. L. and Fitzpatrick, P. M. (2010), *Real Analysis*, Prentice Hall.

Sandner, M. and Jungmann, T. (2016), 'How Much can We Trust Maternal Ratings of Early Child Development in Disadvantaged Samples?', *Economics Letters* **141**, 73–76.

Swaminathan, H., Sharma, A. and Shah, N. G. (2019), 'Does the Relationship Between Income and Child Health Differ Across Income Groups? Evidence from India', *Economic Modelling* **79**, 57–73.

Wei, L. and Feeny, D. (2019), 'The Dynamics of the Gradient Between Child's health and Family Income: Evidence from Canada', *Social Science and Medicine* **222**, 182–189.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.