THE UNIVERSITY OF CHICAGO

Two-Stage Estimators for Unbalanced Panel Data Under Endogenous Selection with Weak Instrumental Variables Condition

By

Ziyu Song

May 21, 2024

A paper submitted in partial fulfillment of the requirements for the
Master of Arts degree in Social Sciences with a Concentration in Economics

Faculty Advisor: Jeffery R. Russell
Preceptor: Oscar Galvez-Soriano

Abstract

This paper proposes the novel two-stage estimators for unbalanced panel data that incorporate endogenous sample selection. Contrary to conventional methods, we rely on a weaker assumption which not requires sample selection to be random and ignorable. The first stage uses a bilateral truncated selection equation where we accommodate a weak instrumental variable (IV). This is due to the recognition that original selection is impacted on unobservable heterogeneity, so that unidentifiable endogenous selections problem exits, which causes optimal IV extremely difficult to acquire. The second stage develops Wooldridge's (2019) model to account for the correlation between unobservable heterogeneity in the main model with covariates and the heterogeneity in the selection mechanism. Simulations demonstrate the efficacy of our estimators in a variety of contexts. And an empirical application in financial panel data concludes a more accurate result than previous literature, also proves the advancement of our method.

Keywords: Unbalanced Panel Data, Endogenous Selection, Weak IV

1 Introduction

Panel data are widely used in economic and financial research due to its larger datasets with more variability and the ability to control for individual heterogeneity. However, panel datasets may exhibit bias due to sample selection problems (Baltagi and Song 2006) [13]. The panel data becomes unbalanced if the selection problem causes samples missing in the dataset, and the unbalanced panel is possibly typical when the sample size and time period are large enough. In empirical research, unbalanced panels are more frequently encountered than their balanced counterparts.

When utilizing a panel dataset with missing values, employing the typical approach of discarding any observations with missing information can lead to inefficient use of data and potentially unrepresentative results. Therefore, specialized estimation techniques are crucial for analyzing unbalanced panels. Researchers often base their studies on the assumption that data missingness is either random or non-random, with the former being more prevalent due to its reduced complexity in methodological handling. Popular methods for estimating random unbalanced panels, such as those proposed by Wooldridge (1995 [14]) and 2019 [6]), include strategies that allow unobserved heterogeneity to be correlated with observed covariates and sample selection in unbalanced panels. These methods are regarded as effective due to their robust approach to parameter estimation in unbalanced panels, making them valuable tools for empirical research.

However, assuming random unbalance is a very strong assumption for panel data. Additionally, another reason why most panels are unbalanced in practice is that missing observations may be created deliberately, indicating that the reason for unbalance is nonrandom. Nonrandomly missing data can also occur due to various self-selection rules. This issue is common in cross-sectional studies, but it is exacerbated in panel surveys. If missing observations in a panel dataset are not missing at random, many widely applied unbalanced panel estimators may be inconsistent (Nijman and Verbeek 1992 [15]), and inference based on the balanced subpanel is also inefficient (Baltagi and Song 2006 [13]). Consequently, specific methods are required for nonrandom unbalanced panels.

Since Hausman and Wise (1979) [16] introduced their seminal two-period attrition model—the first and most notable research in this field—many scholars have worked within the same framework. They relaxed Hausman and Wise's assumptions and expanded them in various directions (Ridder 1990 [24], 1992 [29]; Nijman and Verbeek 1992 [15]; Van Den Berg et al. 1994 [26]; Rochina-Barrachina 1999 [17]). The selection rule was set as an indicator function of a linear panel data model with additive heterogeneity; if the linear panel data model within the indicator function satisfies a condition such as being greater than zero, the selection indicator is set to 1, and this sample is selected. Zable (1992) [30] further developed this by using the mean of regressors in the linear selection model to explain the additive heterogeneity. Additionally, endogenous variables were also incorporated as causes of unbalance in the partial-population selection bias model by Moffitt, Fitzgerald, and Gottschalk (1999) [18]. Beyond linear-type models, Tobit-type models are also commonly used (Kyriazidou 1997 [19]; Lee 2001 [28]; Honore and Kyriazidou 2000 [31]). Other popular methods for selection equations include the Markov decision process (Sasaki 2015 [27]), standard bias-correction procedures (Lee and Han 2018 [8]), and LSDV bias approximations (Bruno 2005 [12]). Building on these various endogenous selection rules, scholars have presented their theoretical studies for unbalanced panels such as two-stage estimation with endogenous variables (Vella and Verbeek 1999 [33]), semiparametric first-difference (Lee 2001 [28]), semiparametric varying coefficient (Malikov et al. 2016 [1]) estimators, and CRE with IV (Joshi and Wooldridge 2019 [21]). Additionally, some scholars have explored new estimation methods using specific empirical datasets (Dustmann et al. 2007 [23]; Yang et al. 2023 [22]).

Sample selection rules play a crucial role in developing these methods (Wooldridge 2010 [20]). Verbeek and Nijman (1996) [25] explained that selection is ignorable if it does not affect the joint distribution of dependent and independent variables; otherwise, it is non-ignorable, also known as endogenous selection. It is always essential to distinguish between ignorable and non-ignorable missingness because, for both types, conventional estimators using the full unbalanced panel or the maximal balanced subset can be inconsistent if the missingness is endogenous and correlated with the regression error (Lee and Han 2018 [8]). Consequently, it is necessary to consider the mechanism causing the missing observations to obtain consistent estimates of the parameters of interest.

In this work, we introduce a two-stage estimation process for unbalanced panel data under conditions of endogenous selection. Building upon the foundational assumptions proposed by Wooldridge (2019) [6], we adopt a weaker assumption wherein the original selection is nonrandom, thereby deriving a newly randomized selection from our established selection equation. Our methodology stands out from previous approaches due to its computational efficiency and the flexibility afforded by these relaxed assumptions. Another manifestation of these relaxed assumptions is the use of a weak instrumental variable (IV). In the first stage, we tackle the challenge of identifying an ideal IV for endogenous selection in the presence of individual effects by permitting the use of a weak IV. The potential biases introduced by such a weak IV are mitigated through the implementation of a bilateral truncated selection model, which differs from the commonly used single-side selection model in the previous literature. The bilateral truncated selection model offers a balance between reducing sample size and ensuring reliable estimation; this tradeoff is governed by the determination of the truncation's upper boundary.

In the second stage, we build upon the main model structures proposed by Chamberlain

(1982) [7] and Wooldridge (2019), where the outcome is influenced by unobserved heterogeneity within the main model, and this heterogeneity is correlated with covariates and selections. Additionally, we introduce another layer of unobserved heterogeneity within the first-stage selection model. This consists of an individual effect that directly impacts selection and also correlates with the heterogeneity in the main model. Consequently, the relationship now encompasses both individual effects from selection and covariates correlated with the main model's heterogeneity. This approach not only corrects the influence from selection to main heterogeneity but also enhances the accuracy of estimates and the robustness of inferences, making it suitable for complex scenarios with multiple unobserved factors.

Furthermore, our estimator is designed to accommodate both linear (with a POLS type estimator) and nonlinear models (with a M-estimator), thereby offering versatility in modeling choices. Typically, we allow the use of semiparametric method if we lose the prior information of relationship format between weak IV and selection, and between individual effect in selection and main heterogeneity.

This paper is structured as follows: Section 2 details the first-stage strategy of our selection model, incorporating our foundational assumptions and the formulation of selection equations. Section 3 introduces the second-stage model and estimators, accommodating both linear and non-linear approaches. Section 4 presents simulation experiments that compare the efficacy of traditional methods with our novel approaches across various scenarios. Finally, Section 5 discusses an empirical application of our new estimator within financial panel data, analyzing the impact of corruption on financial development across different countries.

2 Selection

2.1 The Assumptions

This subsection introduces the basic assumptions of our estimation methods, which are weaker than those typically used in previous literature. We consider a standard unbalanced panel data model, where c_i represents additive heterogeneity.

$$y_{it} = x_{it}\beta + c_i + u_{it} \tag{1}$$

Where y_{it} is element of a $(N \times T) \times 1$ matrix Y, x_{it} is element of a $(N \times T) \times k$ matrix X, c_i is element of a $(N \times T) \times 1$ matrix C, u_{it} is element of a $(N \times T) \times 1$ matrix U. $i \in \{1, 2, ..., N\}$ and $t \in \{1, 2, ..., T\}$. Wooldridge (2019)[6] define a selection indicator s_{it}, s_{it} is a dummy of a random selection, if $\{x_{it}, y_{it}\}$ is complete, $s_{it} = 1$; otherwise $s_{it} = 0$. Meanwhile, he assumes:

$$E[u_{it}|x_i, c_i, s_i] = 0, E[y_{it}|x_i, c_i, s_i] = E[y_{it}|x_i, c_i] \tag{2}$$

These indicate selection is independent to error term and conditional distribution of y_{it}, so the selection is random and ignorable. Note that ignorability means selection not influences the joint distribution of y_{it} , otherwise, we have non-ignorability (also called endogenous selection). Ignorable and random selection is easier to estimate but it is a quite strong assumption. Non-ignorable selection is more common in economic and financial emperical research, such as research on financial development indicator may encounter unbalanced

panel when important economic events happen like crises. If the assumptions in formula 2 is too strong to achieve, that is

$$E[u_{it}|x_i, c_i, s_i] \neq 0, E[y_{it}|x_i, c_i, s_i] \neq E[y_{it}|x_i, c_i] \tag{3}$$

Given the first inequation and the basic assumption of panel data that $u_{it} \perp\!\!\!\perp \{x_i, c_i\}$, we can summarize this selection is likely correlated with the error term. In other words, s_{it} is correlated something out of the model. Let s_{it} be the identical selection. Suppose the observable factor is the random instrumental variable g_{it}, which is the observable reasons for the selection, specificly, $g_{it} = \{g_{it}^1, g_{it}^2\}$. $g_{it}^1 \in G_1$ is excluded IV, $g_{it}^1 \not\perp\!\!\!\perp u_{it}$. In previous example, g_{it}^1 could be crisis dummy. The $g_{it}^2 \in G_2$ is included IV. Where G_1 and G_2 are $(N \times T) \times d_1$ and $(N \times T) \times d_2$ matrices [1], the dimension of endogenous selected variables should equal to $d_1 + d_2$.

Since the panel data encompasses a large number of individuals (large N), we allow s_{it} to not only be correlated with the IV, but also with some unobserved individual effect a_i. Including this individual effect in the selection process is a common practice in the literature, as evidenced by previous scholars such as Kyriazidou (1997) [19] and Dustmann and Rochina-Barrachina (2007) [23]. Consequently, the selection equation for s_{it} is given by:

$$s_{it} = E(s_{it}|g_{it}) + m_{it}^* \tag{4}$$

where $m_{it}^* = a_i + e_{it}$ represents the component of s_{it} that cannot be explained by g_{it}. Subsequently, we present a Wooldridge-type assumption

$$E[u_{it}|x_i, c_i, m_i^*] = 0, E[y_{it}|x_i, c_i, m_i^*] = E[y_{it}|x_i, c_i] \tag{5}$$

Wooldridge (2019)'s assumption necessitates that s_{it} be independent and random. However, our formula 5 requires only that the residuals of s_{it} be independent and random, which is a less stringent condition, thereby constituting a weaker assumption.

Given the selection equation in formula 4, we can express the model of interest 1

$$E(y_{it}|x_{it}, c_i, a_i, g_{it}, s_{it} = 1) = x_{it}\beta + c_i + h(a_i, g_{it}) \tag{6}$$

Where

$$h(a_i, g_{it}) = E(u_{it}|a_i, g_{it}) \tag{7}$$

We can estimate the parameter of interest by directly regressing using formula 6. However, the function $h(a_i, g_{it})$ is unknown, and additional assumptions are required. The inclusion of a_i complicates the formatting of $h(a_i, g_{it})$, especially in the case of short panels where T is limited. To circumvent this challenge, we will not directly utilize 6; instead, we demonstrate that a suitable substitute exists.

$$E(y_{it}|x_{it}, c_i, a_i, g_{it}, s_{it} = 1) = E(y_{it}|x_{it}, c_i, m_{it} = 1) \tag{8}$$

[1] Usually, G_2 includes some subsets of X or selection of these subsets. It is also possible that $G_2 = \emptyset$.

Where m_{it} is a new selection dummy, $m_{it} = f(m_{it}^*)$ as functional transformation of m_{it}^*, and

$$E(y_{it}|x_{it}, c_i, m_{it} = 1) = x_{it}\beta + c_i + E(u_{it}|m_{it}) \tag{9}$$

Continuously, it is easy to prove that m_{it} is a random selection indicator which satisfies Woodridge-type assumption as well (proof in Appendix A.1)

$$E[u_{it}|x_i, c_i, m_i] = 0, E[y_{it}|x_i, c_i, m_i] = E[y_{it}|x_i, c_i] \tag{10}$$

Which implies $m_{it} \perp\!\!\!\perp u_{it}$, and we can simplify 9 to

$$E(y_{it}|x_{it}, c_i, m_{it} = 1) = x_{it}\beta + c_i \tag{11}$$

That is much easier to estimate if we have this new dummy m_{it}.

We require a random selection from the non-ignorable selection model, as posited in formula 5. This approach of separating a random selection from a non-random selection is analogous to the strategy employed by Malikov, Kumbhakar, and Sun (2016) [1]. The subsequent subsection will introduce the design of the new selection variable m_{it}.

2.2 Bilateral Truncated Selection

It is important to note that identifying the optimal instrumental variable (IV) for endogenous selection influenced by unknown individual effects is more challenging than in typical cases. Specifically, when individual effects are unknown and correlated with s_{it}, isolating purely exogenous variations becomes an increasingly complex task. Consequently, the IV may not sufficiently explain the variation in the endogenous variable[2]. Previous scholars have recognized this issue in empirical studies (Pokropek 2016 [10], Zawadzki et al. 2023 [9]) and have also demonstrated it theoretically (Cui et al. 2020 [32]). Therefore, a more flexible IV condition is necessary. We allow g_{it} to be a weak IV,[3] and if a weak IV is tolerable, then any IV stronger than a weak IV is also acceptable[4]. If some endogenous selections cannot be explained by the IV, we refer to these samples as Unidentifiable Endogenous Selections (UES), which also contribute to the weakness of the IV. The presence of unidentifiable endogenous selections leads to biased and inconsistent estimations, as we demonstrate in Appendix A.2.

Previous scholars typically employ a single-side indicator function to model the selection indicator, as seen in works such as Kyriazidou (1997) [5], Semykina and Wooldridge (2013) [3], Malikov and Kumbhakar (2014) [2], and Liu and Yu (2022) [4], where the selection indicator is represented by $1(h(\cdot) > 0)$, with $h(\cdot)$ being a function of selection. This type of selection rule, while generally correct, becomes risky under nonrandom selection as it does not address the issue of unidentifiable endogenous selection. To overcome this, we propose

[2]For instance, consider using bank failures as an IV. If there are unknown individual effects in selection, such as a manager's personal networking relationships or hidden reserves, the bank may still operate (data exists) despite observed failures. Thus, the IV does not capture the endogeneity for these samples.

[3]Note that this does not imply that any weak variable can be randomly chosen as an IV. The IV is weak primarily because of the existence of individual effects.

[4]While a weak IV is tolerable, a stronger IV is still preferred.

using a bilateral truncation indicator function to model a new selection rule. Our objective is to classify samples with unidentifiable endogenous selection as missing, specifically setting $m_{it} = 0$. As demonstrated in Appendix A.2, if we define b_{it} as the bias, the selection indicator becomes $f(m^*_{it} + b_{it})$, with $b_{it} = 0$ indicating the absence of unidentifiable endogenous selection issues. The distribution of $m^*_{it} + b_{it}$ shows that samples under unidentifiable endogenous selection deviate significantly from normal samples, effectively becoming outliers.

Therefore, we design the selection rule as follows:

$$m_{it} = 1(0 < \frac{m^*_{it} - \mu_{m^*}}{\sigma_{m^*}} < z_c) \tag{12}$$

It employs a bilateral truncated method. In cases where the instrumental variable (IV) is binary, we can directly identify a constant z_c through observation. For more common scenarios where the IV is not binary, to clarify the boundary conditions, we specify that

$$z_c = |\Phi^{-1}[Pr[(s_{ues,it} \in S) \cap (s_{it} = 1)]]| \tag{13}$$

z_c is z score of standard normal distribution. Where, $Pr[(s_{ues,it} \in S) \cap (s_{it} = 1)]$ is the probability that UES appears in right tail. Suppose that S includes UES and S samples locate at right tail are two independent event, then we have:

$$Pr[(s_{ues,it} \in S) \cap (s_{it} = 1)] = Pr(s_{ues,it} \in S) * Pr(s_{it} = 1) \tag{14}$$

And

$$Pr(s_{ues,it} \in S) = \frac{A'A}{e_2'e_2 - \frac{(e_2'I)^2}{NT}} \tag{15}$$

It represents the probability that s_{it} is UES. e_2 is $(N \times T) \times 1$ matrix of residual of regress s_{it} on g^2_{it}. And

$$Pr(s_{it} = 1) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \tag{16}$$

It is the probability that UES located at right tail. The bilateral truncation method is based on the probability that a sample is located on the right side of the standard normal distribution. For a detailed explanation and proof of this approach, see Appendix A.3 which discusses formula 13. In this method, we truncate the two tails of the standardized distribution of m^*_{it}. It is important to note that $m_{it} = 0$ represents three types of samples: originally missing samples (where $s_{it} = 0$), samples affected by endogenous selection, and unidentifiable endogenous samples. Therefore, the subset $\{(y_{it}, x_{it})|m_{it} = 1\}$ is an exogenous subset of $\{(y_{it}, x_{it})|s_{it} = 1\}$.

We use partial R^2 of G_1, $R^2_{G_1}$, as measurement of weakness of IV (Shea 1997 [38]). Furthermore, in scenarios where an exactly weak IV is used—characterized by $R^2_{G_1}$ close to 0—it is necessary that $Pr(s_{it} = 1) > 0.5$. This requirement ensures that at least half of the data are non-missing. If this condition is not met, $z_c < 0$, rendering the bilateral truncation ineffective.

If there exists partial endogenous selection, this implies that the selection for some variables is random, while for others, it is not. Consider the case where the selection of y_{it} is a

random selection, denoted as s_{it}^y, and the selection of x_{it} is an endogenous selection, denoted as s_{it}^x. The final selection indicator would then be given by $m_{it}^x s_{it}^y$, where m_{it}^x is calculated from s_{it}^x as previously described[5].

2.3 Expectation of Selection

If we still use additive heterogeneity in selection equation, we have

$$s_{it} = E(s_{it}|g_{it}) + a_i + e_{it} \tag{17}$$

The expectation form could be linear model or other dummy variable model as needed.

However, the choice of g_{it} plays an important role in our selection rule. The growth of weakness of IV may cause more sample missing. As the suggestion of Han and Lee (2018)[8], we may consider to try the lag of covariates. Let z_{it} be the balanced subset of x_{it}, and model $E(s_{it}|z_{it-1})$ to test if z_{it-1} could be a IV for selection. If we also consider previous selection's influence to current selection, we modify the conditional expectation to $E(s_{it}|g_{it}, s_{it-1})$. We can directly model the conditional expectation once we get g_{it} and have reliable prior assumption of relationship between g_{it} and s_{it}. But if we have no prior assumption, we need a semiparametric model with $\eta(g_{it})$ as non-parametric term, that is

$$s_{it} = \eta(g_{it}) + a_i + m_{it}^* \tag{18}$$

The non-parametric term allows all possible relationship between IV and selection. Applying Boneva et al. (2015) [11] semiparametric estimation method, let $\eta(g_{it}) = \sum_{k=1}^{K} \alpha_k \mu_k(g)$. [6] $\mu_k(g)$ is nonparametric component functions(with k components), α_k is the coefficient, the estimator is

$$\hat{\alpha} = \left(\frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \pi(g_{it}) \hat{\mu}(g_{it}) \hat{\mu}(g_{it})^T \right)^{-1} \left(\frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \pi(g_{it}) \hat{\mu}(g_{it}) s_{it} \right) \tag{19}$$

where $\pi(.)$ is a weighting function, $\hat{\mu}(.)$ is estimator of $\mu(.)$ through weighted Nadaraya–Watson. This semiparametric approach to modeling an endogenous selection with a weak instrumental variable can enhance model flexibility and robustness (also improves the partial R^2) by not constraining the relationship to a specific functional form and by being more accommodating of data complexity and nonlinearity. In conclusion, this selection equation can lead to more reliable inference even when strong instruments are not available.

[5] This methodology is also applicable in the case of full endogenous selection, provided that multiple different selection equations are required. In such cases, the final selection indicator would be the product $\prod_{j=1}^{k+1} m_{it}^j$, where $j \in \{y, x^1, ..., x^k\}$.

[6] Boneva et al. (2015)'s original methods use $\eta_i(g_{it}) = \sum_{k=1}^{K} \alpha_{ik} \mu_k(g)$, we slightly modify it because we don't need the coefficient change cross i

3 Estimations

3.1 Linear Model

Given the selection indicator in last section, if we allow individual effect a_i correlated to g_i, that is to continue Chamberlain (1982) and Wooldridge (2019)'s methods, we have

$$E(a_i|g_{it}) = \bar{g}_i\xi_g \tag{20}$$

ξ_g is coefficient parameter. Or by formula 17 and without this assumption

$$E(a_i|g_{it}, s_{it}) = s_{it} - E(s_{it}|g_{it}) \tag{21}$$

Similarly, if we allow both individual effect in selection and x_{it} correlated to heterogeneity in main model, that is

$$E(c_i|x_i, a_i) = \bar{x}_i\xi_x + E(a_i|g_i)\gamma_a \tag{22}$$

where ξ_x and γ_a are coefficient parameters.

Then, we include them in our main models. Consider the simple linear nonrandom unbalanced panel by Wooldridge-type formula with Mundlak device, that is

$$E(y_{it}|c_i, x_{it}, m_{it} = 1) = m_{it}x_{it}\beta + m_{it}E(c_i|x_i, a_i) \tag{23}$$

Or in other format,

$$m_{it}[y_{it} - \theta_i\bar{y_i}] = m_{it}[x_{it} - \theta_i\bar{x_i}]\beta + m_{it}(1-\theta_i)\bar{x}_i\xi_x + m_{it}(1-\theta_i)\bar{g}_i\xi_g\gamma_a + m_{it}[u_{it} - \theta_i\bar{u_i}] \tag{24}$$

where [7]

$$\theta_i = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T_i(\sigma_{\bar{x}_i}^2\rho_x^2 + \sigma_{\bar{g}_i}^2\rho_{\bar{g}_i}^2\gamma_{a_i}^2)}\right)^{1/2} \tag{25}$$

σ^2 is the variance of each variable, see Appendix A.3.1 for details of θ_i formula. Apply Pooled OLS, define the quasi-time demeaning format $\tilde{x}_{it} = x_{it} - \theta_i\bar{x}_i$ and $\tilde{y}_{it} = y_{it} - \theta_i\bar{y}_i$, then the estimator of interest parameter β is

$$\hat{\beta} = \left(\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{x}_{it}\hat{m}_{it}\tilde{x}_{it}\right)^{-1}\left(\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{x}_{it}\hat{m}_{it}\tilde{y}_{it}\right) \tag{26}$$

where $\hat{m}_{it}$ is estimator of m_{it}. The estimator $\hat{\beta}$ is consistent, which is proved in Appendix A.4. The estimator is FE estimator if $\theta_i = 1$; it is RE estimator if $0 < \theta_i < 1$; POLS estimator if $\theta_i = 0$.

The asymptotic distribution of this estimator is

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma_u'(X'\hat{M}X)^{-1}\sigma_u) \tag{27}$$

where X is $(N \times T) \times k$ (k regressors) of x_{it}, $\hat{M}$ is $(N \times T) \times (N \times T)$ matrix of $\hat{m}_{it}$.

[7] There could be a term of time constant variable z_i like Mundlak and Wooldridge's original formulas

Although this estimator is consistent for short panel (large N, fix T), if we have long panel with large T, we need consider the time influence in each series. We can modify to a dynamic model by adding the lag of y_{it} as covariates, but it is likely that the lag is correlated to error, so we need to use GMM type methods to replace POLS.

3.2 Non-Linear Model

Nonlinear models are often considered in unbalanced panel data due to their ability to handle more complex relationships among variables that linear models might oversimplify, especially when data exhibits non-linear patterns. Suppose the interest distribution of y_{it} conditional on covariates, heterogenity and selection is

$$D(y_{it}|x_{it}, c_i, m_{it}) \tag{28}$$

Given the selection is ignorable, that is

$$D(y_{it}|x_{it}, c_i, m_{it}) = D(y_{it}|x_{it}, c_i) \tag{29}$$

Next, consider the first stage of regression. Since s_{it} includes a random effect a_i, and a_i could be a function of g_{it} by Chamberlain and Wooldridge like before, we set the density of a_i to be

$$D(a_i|g_{it}) \tag{30}$$

Similarly, the random effect in the main model c_i, also has distribution conditional on x_{it}. Since we allow the random effect a_i correlated to selection, and selection may correlated to c_i in previous literature, we directly let a_i correlated to c_i in our case. That is

$$D(c_i|x_{it}, a_i) \tag{31}$$

Now we need to find the conditional distribution of c_i on both x_{it} and g_{it} because we only observe them two. The c_i is conditional on a_i, but the a_i is conditional on g_{it}, then get

$$D(c_i|g_{it}, x_{it}) = \int_{R^A} D(c_i|x_{it}, a_i)D(a_i|g_{it})da_i \tag{32}$$

Where A is dimension of a_i. This formula indicates the relationship that g_{it} impacts on a_i, $\{g_{it}, a_i\}$ and x_{it} impact on c_i.

Meanwhile, modify the conditional distribution of y_{it}, the distribution of y_{it} conditional on x_{it} and g_{it} is

$$f(y_t|x_{it}, g_{it}) = \int_{R^C} D(y_t|c_i, x_{it})D(c_i|g_{it}, x_{it})dc \tag{33}$$

Where C is the dimension of c_i. This is similar to Wooldridge (2019)'s formula. Then plug formula 32 into, we have

$$f(y_t|x_{it}, g_{it}) = \int_{R^C}\int_{R^A} D(y_t|c_i, x_{it})D(c_i|x_{it}, a_i)D(a_i|g_{it})dadc \tag{34}$$

The $f(y_t|x_{it}, g_{it})$ is final format we interest, it conditional only on given variables $\{g_{it}, x_{it}\}$ rather than unobserveable random effect.

In order to achieve the estimator of $f(y_t|x_{it}, g_{it})$, the likelihood function is

$$l(\beta; y_{it}, x_{it}, g_{it}) = \sum_{i=1}^{N}\sum_{t=1}^{T} m_{it} log[f(y_{it}|x_{it}, g_{it})] \tag{35}$$

Then, the M-estimator is

$$\hat{\beta}_M = argmax_\beta l(\beta; y_{it}, x_{it}, g_{it}) \tag{36}$$

Continuously define the Jacobian and Hessian as following.

$$J_{it}(\beta) = \nabla_\beta \log[f(y_{it}|x_{it}, g_{it}; \beta)]' \tag{37}$$

$$H_{it}(\beta) = \nabla^2_\beta \log[f(y_{it}|x_{it}, g_{it}; \beta)] \tag{38}$$

Then we can use them to show the asymptotic property of this M-estimator.

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta) \tag{39}$$

Where V is variance-covriance matrix,

$$V_\beta = (-\mathbf{E}[\mathbf{MH}])^{-1} \mathbf{Var}(\mathbf{MJ}) (-\mathbf{E}[\mathbf{MH}])^{-1}) \tag{40}$$

Where H is matrix of H_{it}, J is matrix of J_{it}. The performance of this M-estimator is shown in following section.

4 Simulation

This section will use simulation experiments to illuminate the performance of our new estimation methods. For linear model,we generated a panel data samples x_{it} and y_{it}, where $y_{it} = x_{it}\beta + c_i + u_{it}$, β is set to 0.5, u_{it} is random sampled from standard normal. The endogenous selection is generated as a dummy variable strongly correlated to u_{it}. Then, let $g2$ be a dummy IV of s_{it}, $corr(s_{it}, g2) < 0.5$, and partial $R^2 = 0.4$.

To compare our method with previous method, we firstly apply Wooldridge (2019)'s method, and plot the asymptotic distribution of estimator with our new estimator in Fig 1a. The graph shows the new estimator is centered around the true value of 0.5, tails drop off faster, more peaked and narrow than previous estimator. So if we naively ignore the endogenous selection problem, the estimator become more biased, less sufficient and less reliable.

Next, we consider the advantage of bilateral tails cutting. We design two m_{it}, the one is our two-sides tails cutting in formula 12, the other is single-side tail cutting, that is $m = 1(0 < \frac{m^*_{it} - \mu_{m^*}}{\sigma_{m^*}})$. Fig1b shows the difference of estimators with these two m_{it}. The distribution with double cutting are more unbiased and narrower. One the contrary, single-side cutting method has a fatter or longer right tail, more susceptible to positive outliers or extreme values. But, as shown in distribution, bilateral tails cutting usually has more

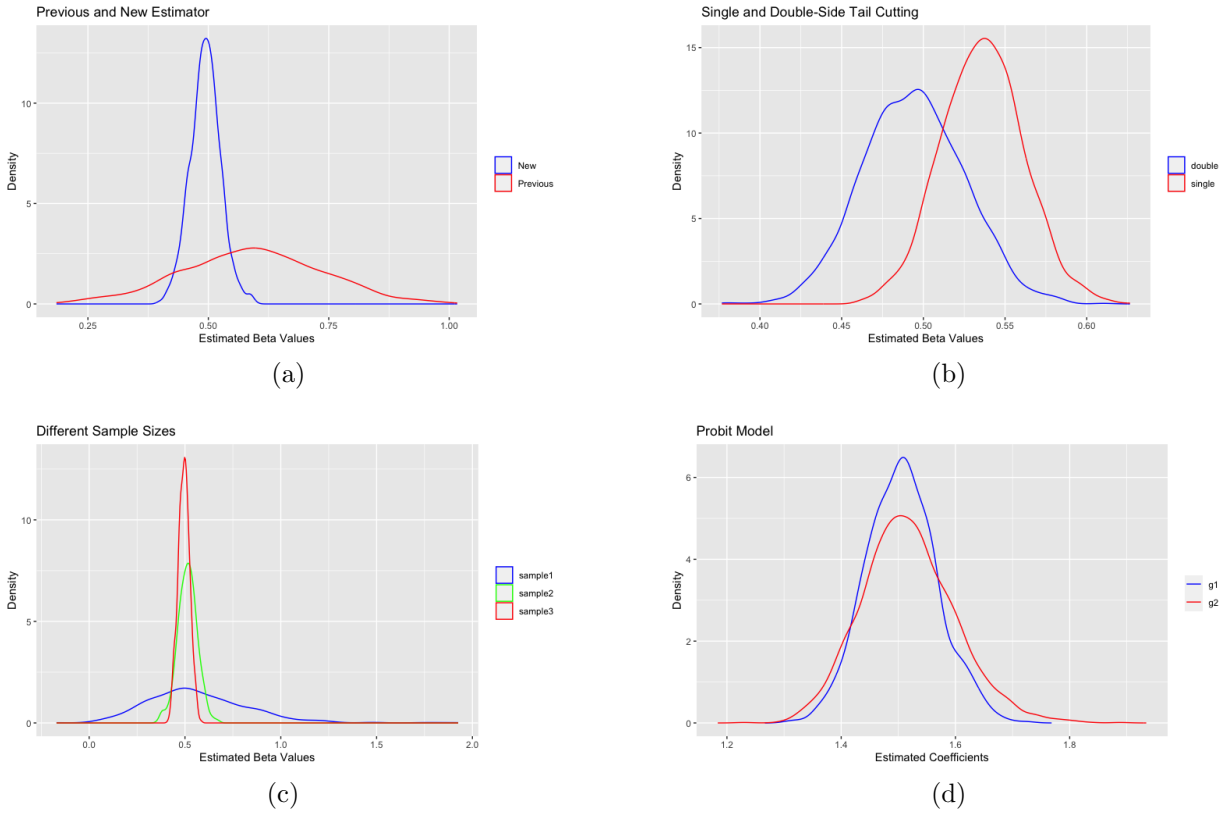

Figure 1: Estimators' Simulation Asymptotic Distributions

samples losing (smaller samples size) so that converges slower.

Fig 1c provides a visual reinforcement of the influence of sample size. We set three different sample size: sample1 $\{N, T\} = \{50, 5\}$; sample2 $\{N, T\} = \{100, 10\}$; sample3 $\{N, T\} = \{150, 15\}$. All three samples have peaks around the same point of β, which suggests that the estimator is unbiased and consistent across different sample sizes. But the larger sample size yields a tighter distribution of estimated beta values, indicating reduced variance and increased precision of the estimates. This enhances the reliability of statistical inferences, making larger sample sizes preferable for robust estimation.

For the nonlinear model, we use example of probit response function. Let

$$P(y_{it} = 1|x_{it}, c_i) = \Phi(x_{it}\beta + c_i) \tag{41}$$

Then, assume $D(a_i|g_it)$ is linear normal, with

$$E(a_i|g_{it}) = (\bar{g}_i - \mu_g)\rho_g \tag{42}$$

This is same as Wooldridge (2019)'s assumption. s_{it} is still the selection equation in formula 20, also has normal density. Also normal density $D(c_i|x_it, a_i)$, with

$$E(c_i|x_{it}, a_i) = (\bar{x}_i - \mu_x)\rho_x + \eta(a_i) \tag{43}$$

where $(\bar{x}_i + \mu_x)\theta_x$ is Wooldridge-type method as well, but $\eta(a_i)$ is unknown function of a_i. It could be non-parametric model, but we set as linear model to reduce computational burden (with the coefficient γ_{a_i}). Let the conditional variance be

$$Var(c_i|x_{it}, a_i) = \sigma^2_{\bar{x}_i}\rho^2_x + \sigma^2_{\bar{g}_i}\rho^2_{\bar{g}_i}\gamma^2_{a_i} \tag{44}$$

Then we have

$$P(y_{it} = 1|x_{it}, g_{it}, m_{it} = 1) = \Phi\left[\frac{x_{it}\beta + E(c_i|x_{it}, a_i)}{\left(1 + \sigma^2_{\bar{x}_i}\rho^2_x + \sigma^2_{\bar{g}_i}\rho^2_{\bar{g}_i}\gamma^2_{a_i}\right)^{\frac{1}{2}}}\right] \tag{45}$$

We set specific values for each parameters, and simulate to show the asymptotic distribution of $\hat{\beta}$ (with true $\beta = 1.5$) in Fig 1d. Note that we generate another weak IV $g1$ here, g1 is stronger than g2. The asymptotic distributions of the estimators for g1 and g2 indicate that both provide decently consistent estimations of the coefficients, as evidenced by their peakedness at the true value. Estimator g1 demonstrates a slightly superior performance with a tighter distribution, suggesting a more precise estimation compared to the broader spread of g2's estimator.

5 Application

Corruption is a pervasive issue in financial markets, commonly perceived as detrimental to a country's development. But some empirical research proposed that corruption may improve economic or financial development (Ahlin and Pang 2008 [36]), typically in developing countries (Song et al. 2021 [37]). Conversely, other research indicates that reducing corruption benefits advanced economies (Schneider et al. 2022 [35]). Although these findings suggesting potential benefits of corruption cannot be outright dismissed as inaccurate, the prevailing view supports the notion that corruption generally harms economic development, regardless of a country's economic status. That is because when we use financial panel across countries, the unobserved individual effect may exist and impact on our estimation [8].

Our estimator is more safety than others because we control the bias from IV in endogenous selection, so that we can control the unknown individual effect better. We will use an unbalanced financial panel [9] of 140 country (from 2000 to 2014) to study how controlling of corruption influence financial development. Specifically, we want to know how controlling of corruption influence different sectors in financial development. We use IMF Financial Development indicator as dependent variable. It includes 7 variables, that is $Y = [FD, FIA, FID, FIE, FMA, FMD, FME]$ [10]. The independent matrices are $X_1 =$

[8]For instance, special political system is one of the unobserved heterogeneity. In the developing country Russia, the unique political system characterized by strong centralized authority and substantial state control over the economy significantly influences financial development. This system enables corruption to facilitate access to exclusive contracts and state funding, allowing those with government connections to thrive. As a result, corruption can boost financial development by enabling quicker execution of large-scale projects.

[9]Data source is [34]

[10]F: financial, I: institution, M: market, D:depth, A: access, E:efficiency.

$[corr, lnGDP, comp]$ [11], $X_2 = [corr$[12]$, dep, corr \times dep]$, dep is dummy for developed countries [13], this X_2 is prepared for another regression model to compare the effect between country types.

In order to estimate for first stage, set IV matrix $G = [FM, FI, Crisis.dummy, GDP.per]$ [14]. Then, we get the $R^2 = 7\%$ form regress S on G, and select the samples for $m_{it} \in [0, z\ score = 1.28]$. We study how corr influence FD by regressing Y on X_1, the results are in table 1 and left of table 2. Also study the difference between developed and developing countries by regressing Y on X_2 in right part of table 2. The rest results in Appendix A.5.

Table 1: Regression results for Y on X_1 with samples of developed(left) and developing (right)countries separately

	Developed Countries (corr)				Developing Countries (corr)			
Indicator	Coeff.	se	z-value	p-value	Coeff.	se	z-value	p-value
FD	0.0415	0.0180	2.3021	0.0213	0.0039	0.0029	1.3521	0.1763
FIA	-0.0102	0.0254	-0.4029	0.687	-0.0007	0.0059	-0.1214	0.9034
FIE	0.0534	0.0218	2.4478	0.0144	0.0044	0.0048	0.9127	0.3614
FID	0.0397	0.0246	1.6169	0.1059	0.0054	0.0026	2.1149	0.0344
FMA	0.0566	0.0295	1.9205	0.0548	0.0072	0.0036	2.0131	0.0441
FMD	0.0615	0.0362	1.6981	0.0895	0.0054	0.0045	1.1966	0.2315
FME	0.0261	0.0619	0.4211	0.6737	0.0058	0.0076	0.7645	0.4446

Table 2: Regression results for Y on X_1 across all countries (left), and Y on X_2 across all countries (right)

	All Countries(corr)				$corr \times dep$			
Indicator	Coeff.	se	z-value	p-value	Coeff.	se	z-value	p-value
FD	0.0082	0.0029	2.8132	0.0049	0.0473	0.0186	2.5502	0.0108
FIA	0.0046	0.0056	0.8300	0.4065	-0.0169	0.0355	-0.4754	0.6345
FIE	0.0108	0.0045	2.3818	0.0172	0.0525	0.0281	1.8665	0.0620
FID	0.0092	0.0029	3.1900	0.0014	0.0304	0.0184	1.6509	0.0988
FMA	0.0108	0.0037	2.8820	0.0040	0.0504	0.0245	2.0577	0.0396
FMD	0.0134	0.0048	2.7876	0.0053	0.0584	0.0301	1.9442	0.0519
FME	0.0171	0.0081	2.1171	0.0343	0.0562	0.0506	1.1113	0.2665

The findings from our results challenge prior assumptions that controlling corruption does not benefit financial development in developing countries. Contrary to earlier views,

[11] corr: control of corruption, lnGDP: ln(GDP), comp: competition indicators, $comp = [Boone, Lerner, Concentration\ Ratios]$

[12] The World Bank world wide governance indicators: control of corruption

[13] $dep_i = 1$ if i is developed country, otherwise i is developing country.

[14] Crisis.dummy: dummy for economic crisis for i and t. GDP.per: GDP per capital

the results show that improved corruption control significantly enhances financial development, including the depth and efficiency of financial institutions and market access. But not to financial institutions access. It makes sense because entrenched social and economic inequalities can persist and impact on institution access even as corruption controls improve in many developing countries. This indicates that, not only for advanced economics, effective governance and corruption reduction are crucial for creating robust financial systems in developing nations, thus supporting economic growth and stability. The positive coefficients of $corr \times dep$ suggest that as corruption control improves, its positive impact on financial market efficiency is stronger in developed countries than in developing ones, which verifies previous scholars' results. Additionally, the traditional unbalanced panel estimator applied to this dataset yielded results that demonstrate smaller coefficients and larger p-values compared to our estimator, which also proves the advantage of our estimator.

6 Conclusion

We propose a novel two-stage estimator for unbalanced panel data that addresses endogenous sample selection without relying on the assumption of randomness and ignorability in selection. We base on Wooldridge (2019) but relax the fundamental assumption, we assume the residual of selection is random rather than selection itself. In the first stage, we allow unobserved heterogenity exist in selection equation. Our approach utilizes a bilateral truncated selection model, allow a more flexible instrumental variable condition to manage the complications arising from unobservable heterogeneity, which often hampers the identification of endogenous selection issues and complicates the acquisition of an good IV. Any IV, even a weak IV, is tolerable, because the truncation will correct the IV bias.

A new selection indicator m_{it} is applied for second stage. By developing Wooldridge's (2019) model further, we account for the correlation between the unobservable heterogeneity in the main model with covariates and the heterogeneity in the selection mechanism. And developing both linear estimator and non-linear M-estimator for second stage.

Simulations demonstrate the robustness, consistency, and efficacy of our estimators in various contexts, highlighting their superiority over traditional methods in dealing with different sample sizes and IV conditions. We applied our methodology to a financial panel encompassing 140 countries to examine the impact of corruption control on financial development. Our estimator, which effectively controls for unobserved heterogeneity and selection bias, confirms the intuitive conclusion that controlling corruption benefits financial development in both developed and developing countries. However, the impact is more pronounced in developed countries. This finding contrasts with previous studies that suggested corruption control does not benefit financial development in developing countries. Our results not only strengthen the theoretical underpinnings of our estimator but also offer crucial empirical evidence that can inform policy-making in developing regions.

A Appendix

A.1 Formula 10

Proof. Basing on the assumption in Formula 5, since u_{it} is independent of m_i given x_i and c_i, and $m_{it} = f(m^*_{it})$ is a binary variable determined by m^*_{it} but does not carry additional information about u_{it}, it follows that $E[u_{it}|x_i, c_i, m_i] = 0$.

Similarly, given that m_{it} is an indicator function of m^*_{it} and partitions the sample space defined by m^*_{it} without adding new information about y_{it} beyond what is captured by x_i and c_i, we have:

$$E[y_{it}|x_i, c_i, m_i] = E[y_{it}|x_i, c_i, m^*_i] = E[y_{it}|x_i, c_i].$$

□

A.2 Inconsistency and Bias by Unidentifiable Endogenous Selection

Proof. Consider the case of a linear model in unbalanced panel data, we apply Wooldridge (1995 & 2019) method to estimate it, regress

$$f(m^*_{it})[y_{it} - \bar{y_i}] = f(m^*_{it})[x_{it} - \bar{x_i}]\beta + f(m^*_{it})[u_{it} - \bar{u_i}]$$

Where $f(m^*_{it})$ is the selection indicator, rewrite it in s_{it}

$$f(s_{it} - E(s_{it}|g_{it}))[y_{it} - \bar{y_i}] = f(s_{it} - E(s_{it}|g_{it}))[x_{it} - \bar{x_i}]\beta + f(s_{it} - E(s_{it}|g_{it}))[u_{it} - \bar{u_i}]$$

If there is unidentifiable endogenous selection, it means s_{it} is significant influenced by unidentifiable outside factors, then $E(s_{it}|g_{it})$ not provides unbiased estimation for s_{it}. Let the bias be $b(u_{it})$, a function of u_{it}, $b(u_{it}) \neq 0$, since bias is strongly related u_{it}. Then we have

$$m^*_{it} = s_{it} - E(s_{it}|g_{it}) + b(u_{it})$$

The basic assumption in Formula 5 is violated. The selection f(.) is still endogenous, and a endogenous selection will cause inconsistent estimation (this is proved by Lee and Han 2018[8]). Meanwhile, the regression model becomes

$$f(s_{it}-E(s_{it}|g_{it})+b(u_{it}))[y_{it}-\bar{y_i}] = f(s_{it}-E(s_{it}|g_{it})+b(u_{it}))[x_{it}-\bar{x_i}]\beta+f(s_{it}-E(s_{it}|g_{it})+b(u_{it}))[u_{it}-\bar{u_i}]$$

Let $\tilde{x_{it}} = f(s_{it} - E(s_{it}|g_{it}) + b(u_{it}))[x_{it} - \bar{x_i}]$, easily find $\tilde{x_{it}} \not\perp u_{it}$. Thus, the estimation will be biased and inconsistent. □

A.3 The Upper Boundary in Formula 12

Proof. WTS z_c always capture the Unidentifiable Endogenous samples we want to drop. We need: (1) know the Prob that the sample is Unidentifiable Endogenous sample; (2) Know the location of these samples we want to drop in distribution

Let S be $(N \times T) \times 1$ matrix of s_{it}, G be $(d_1 + d_2) \times (N \times T)$ matrix of G_1 and G_2, α be $k \times 1$ matrix of slopes, A be $(N \times T) \times 1$ matrix of unknown individual effect a_i, m^* be $(N \times T) \times 1$ matrix of residuals. The original selection model is (suppose linear)

$$S = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \alpha + A + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{46}$$

But we don't know A, we only have

$$S = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \alpha^* + \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \tag{47}$$

Where α^* is a new slope matrix different from α. The estimation of α^* may not be consistent, but if we let $\alpha^* = \alpha$, we have

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} + A \tag{48}$$

Regress G_1 on G_2 get the residual as e_3

$$m_2 = e_3'\gamma + \epsilon \tag{49}$$
$$e_2 = e_3'\gamma - A + \epsilon \tag{50}$$

We only have the R^2 form regression of m, rather than e. Denote it as $R^2_{G_1}$, which is partial R^2 of IV G_1.

$$1 - R^2_{G_1} = \frac{(\epsilon - A)'(\epsilon - A)}{SST} \tag{51}$$

Given that A is independent with ϵ,

$$\frac{A'A}{SST} = 1 - R^2_{G_1} - \frac{\epsilon'\epsilon}{SST} \tag{52}$$

Since the IV may not explain S well, there exist Unidentifiable Endogenous Samples (UES), i.e. the samples in S that endogeneity in these S cannot be explained by our IV. Define S_{ues} to be the set of these UES samples. Our goal is to drop these UES, but we don't know which samples are S_{ues} at this moment.

Set $P(S_{ues} \in S)$ as a probability, it reflects the proportion of the variability in the S that cannot be explained by the IV G in model 47. It is probability that the sample in S and it is UES.

$$P(S_{ues} \in S) = \frac{A'A}{SST} \tag{53}$$

$\frac{A'A}{TSS}$ is the proportion of the total variability in S that is explained by A given G_1 which is likely to be weak. We can suppose that $var(\epsilon) = 1$, so that $\frac{\epsilon'\epsilon}{SST}$ becomes $\frac{1}{SST}$.

Next, need to know the location of S_{ues} (in distribution), we locate S_{ues} through the distribution of its corresponding residual m^*_{ues}. Still consider the regression model 47

$$S = G'\alpha^* + m^{**} \tag{54}$$

Where m^{**} is a new residual matrix, it includes both normal residuals and m^*_{ues}. But m^* only includes normal residuals. The estimator of α^* is

$$\hat{\alpha}^* = \hat{\alpha} + b \tag{55}$$

Where b is Omitted Variable Bias, because we omit A. The m^* is

$$m^* = S - \hat{S} \tag{56}$$

The distance (difference) between m^* in formula 47 and m^{**} in formula 54 is

$$|m^{**} - m^*| = |S - G'\hat{\alpha} - S + G'\hat{\alpha} - G'bias| = |G'bias| \tag{57}$$

Since the bias always not 0, that means these m^*_{ues} are outliers, always far away (has a non-zero distance) to ordinary m^*. So the m^*_{ues} locate at the outermost tails of standard normal distribution of m^*.

The probability that m^* located in right tail is same as the probability that $s_{it} = 1$, that is

$$P(s_{it} = 1) = \frac{1}{NT} \sum \sum s_{it} \tag{58}$$

Then, the probability that the sample is UES and located at right tail is

$$P[(m^*_{ues} \in m^*) \cap (s_{it} = 1)] = P[(S_{ues} \in S) \cap (s_{it} = 1)] = P(S_{ues} \in S) * P(s_{it} = 1) \tag{59}$$

Therefore, we need to cut z_c-length the rightmost tail.

$$z_c = |\Phi^{-1}[P[(m^*_{ues} \in m^*) \cap (s_{it} = 1)]]| \tag{60}$$

z_c is the z-score of standard normal.

□

A.3.1 Proof of θ_i in formula 25

Proof. By Wooldridge (2010), the original θ_i is

$$\theta_i = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T_i Var(c_i)} \right)^{1/2}$$

Since we have the conditional expectation of c_i is

$$E(c_i|x_i, a_i) = \bar{x}_i \xi_x + \bar{g}_i \xi_g \gamma_a$$

Then the conditional variance become

$$\text{Var}(c_i|x_{it}, a_i) = \text{Var}(\bar{x}_i)\xi_x^2 + \text{Var}(\bar{g}_i)\xi_{\bar{g}_i}^2 \gamma_{a_i}^2 + 2\xi_x \xi_g \gamma_{a_i} Cov(\bar{x}_i, \bar{g}_i)$$

Suppose x and g independent,

$$Var(c_i|x_{it}, a_i) = \sigma^2_{\bar{x}_i}\xi^2_x + \sigma^2_{\bar{g}_i}\xi^2_{\bar{g}_i}\gamma^2_{a_i}$$

□

A.4 Proof of Consistency of Linear Estimator

Proof. Suppose the consistent estimator of residual we get from first stage regression is $\hat{m}^*_{it}$, then

$$\hat{m}_{it} = 1(0 < \frac{\hat{m}^*_{it} - \mu_{m^*}}{\sigma_{m^*}} < z_c)$$

$\hat{m}_{it}$ is consistent estimator of m_{it} since $\hat{m}^*_{it}$ is consistent.

Let Y be $(N \times T) \times 1$ matrix of y_{it}, X be $(N \times T) \times k$ (k regressors) of x_{it}, $\hat{M}$ be $(N \times T) \times (N \times T)$ matrix of $\hat{m}_{it}$. $\tilde{X} = QX$, $\tilde{Y} = QY$, where $Q = I - \Theta 1_T(1'_T 1_T)1'_T$, Θ is a $(N \times T) \times (N \times T)$ block diagonal matrix with θ_i. Basing on Wooldridge (2019)'s [6] proof, we only need to show the additional part (with new selection dummy) not impact on consistency of β, it can be written as:

$$\hat{\beta} = (\tilde{X}'\hat{M}\tilde{X})^{-1}(\tilde{X}'\hat{M}\tilde{Y})$$

Substitute $\tilde{Y}$ from the model, $\tilde{Y} = \tilde{X}\beta + \varepsilon$:

$$\hat{\beta} = (\tilde{X}'\hat{M}\tilde{X})^{-1}(\tilde{X}'\hat{M}(\tilde{X}\beta + \varepsilon))$$

Expand and simplify:

$$\hat{\beta} = \beta + (\tilde{X}'\hat{M}\tilde{X})^{-1}(\tilde{X}'\hat{M}\varepsilon)$$

To prove $\hat{\beta} \xrightarrow{p} \beta$, we need to show that:

$$(\tilde{X}'\hat{M}\tilde{X})^{-1}(\tilde{X}'\hat{M}\varepsilon) \xrightarrow{p} 0$$

Given previous mean zero assumption of U, and iid sampling and weak law of large numbers, we can conclude $\tilde{X}'\tilde{U} \xrightarrow{p} 0$. Meanwhile, $\tilde{X}'M\tilde{U}$ is a submatrix of $\tilde{X}'\tilde{U}$, $\hat{M} \xrightarrow{p} M$ so $\tilde{X}'\hat{M}\tilde{U}$ is also a submatrix of $\tilde{X}'\tilde{U}$. Thus, $\tilde{X}'\hat{M}\tilde{U} \xrightarrow{p} 0$ by Componentwise Convergence Theorem. Concluding that:

$$\hat{\beta} \xrightarrow{p} \beta$$

□

A.5 Additional Empirical Results

By the rest results, GDP consistently shows a significant positive impact on all financial development indicators, highlighting the crucial role of economic growth in strengthening the financial sector. In contrast, competition, measured by the Lerner and Boone indices, has mixed effects: the Lerner index is generally positive, suggesting that less competition leads to

greater efficiency and stability, while the Boone indicator often has a negative impact, implying that intense competition can destabilize financial markets by reducing profitability.These results illustrate the intricate interplay between economic growth, competitive pressures, and financial development, emphasizing the broader context within which measures to control corruption operate and influence the financial sector.

Table 3: Random Effects Regression Results for Various Financial Variables

Dependent Variable	Variable	Coefficient	Std. Error	p-value
FD	lnGDP	0.13367	0.00472	$< 2.2e-16$
	Lerner	0.03024	0.00821	0.0002305
	Bank.con	-0.00001015	0.00009609	0.9158711
	Boone	-0.01168	0.00313	0.0001902
FIA	lnGDP	0.22491	0.00735	$< 2.2e-16$
	Lerner	0.03266	0.01499	0.0294
	Bank.con	-0.00102	0.000173	$3.745e-09$
	Boone	0.000655	0.00568	0.9082
FIE	lnGDP	0.08351	0.00594	$< 2.2e-16$
	Lerner	0.09639	0.01506	$1.534e-10$
	Bank.con	-0.000262	0.000170	0.1230
	Boone	-0.01192	0.00562	0.0339
FID	lnGDP	0.11182	0.00546	$< 2.2e-16$
	Lerner	0.03531	0.00868	$4.795e-05$
	Bank.con	-0.0001405	0.0001027	0.1712
	Boone	-0.0001057	0.00332	0.9746
FMD	lnGDP	0.13338	0.00802	$< 2.2e-16$
	Lerner	0.06169	0.01540	$6.176e-05$
	Bank.con	-0.00015642	0.00017837	0.38052
	Boone	-0.01038	0.00585	0.07613
FME	lnGDP	0.05916	0.01283	$4.015e-06$
	Lerner	-0.06103	0.02849	0.03217
	Bank.con	0.00061936	0.00032531	0.05692
	Boone	-0.00504	0.01074	0.63876
FMA	lnGDP	0.08180	0.00742	$< 2.2e-16$
	Lerner	0.00717	0.01255	0.568
	Bank.con	0.00089415	0.00014735	$1.295e-09$
	Boone	-0.04425	0.00479	$< 2.2e-16$

Table 4: Random Effects Regression Results for Various Financial Variables

Dependent Variable	Variable	Coefficient	Std. Error	z-value	p-value
FD	Intercept	0.285	0.014	19.95	$< 2.2e-16$
	corr	0.004	0.003	1.39	0.164
	dep	0.371	0.048	7.73	0.000
	corr:dep	0.047	0.019	2.55	0.011
FIA	Intercept	0.286	0.021	13.37	$< 2.2e-16$
	corr	-0.001	0.006	-0.15	0.882
	dep	0.453	0.082	5.52	0.000
	corr:dep	-0.017	0.036	-0.48	0.635
FID	Intercept	0.200	0.016	12.58	$< 2.2e-16$
	corr	0.006	0.003	2.00	0.045
	dep	0.484	0.051	9.52	0.000
	corr:dep	0.030	0.018	1.65	0.099
FIE	Intercept	0.618	0.012	49.61	$< 2.2e-16$
	corr	0.005	0.005	0.96	0.337
	dep	0.049	0.059	0.83	0.404
	corr:dep	0.052	0.028	1.87	0.062
FMA	Intercept	0.195	0.024	8.14	0.000
	corr	0.007	0.004	1.94	0.053
	dep	0.327	0.073	4.50	0.000
	corr:dep	0.050	0.025	2.06	0.040
FMD	Intercept	0.179	0.018	9.87	$< 2.2e-16$
	corr	0.006	0.005	1.17	0.241
	dep	0.443	0.070	6.37	0.000
	corr:dep	0.058	0.030	1.94	0.052
FME	Intercept	0.194	0.026	7.47	0.000
	corr	0.006	0.008	0.74	0.458
	dep	0.372	0.110	3.37	0.0008
	corr:dep	0.056	0.051	1.11	0.267

References

[1] Malikov, Emir, Subal C. Kumbhakar, and Yiguo Sun. 2016. "Varying Coefficient Panel Data Model in the Presence of Endogenous Selectivity and Fixed Effects." *Journal of Econometrics* 190, no. 2: 233-251. https://doi.org/10.1016/j.jeconom.2015.06.007.

[2] Emir Malikov, Subal C. Kumbhakar, "A Generalized Panel Data Switching Regression Model," *Economics Letters*, Volume 124, Issue 3, 2014, Pages 353-357, ISSN 0165-1765, https://doi.org/10.1016/j.econlet.2014.06.022.

[3] Anastasia Semykina, Jeffrey M. Wooldridge. (2010). "Estimating Panel Data Models in the Presence of Endogeneity and Selection," *Journal of Econometrics*, Volume 157, Issue 2, Pages 375-380, ISSN 0304-4076, https://doi.org/10.1016/j.jeconom.2010.03.039.

[4] Ruixuan Liu, Zhengfei Yu. (2022). "Sample Selection Models with Monotone Control Functions," *Journal of Econometrics*, Volume 226, Issue 2, Pages 321-342, ISSN 0304-4076, https://doi.org/10.1016/j.jeconom.2021.01.010.

[5] Kyriazidou, Ekaterini. (1997). "Estimation of a Panel Data Sample Selection Model." *Econometrica* 65, no. 6: 1335–64. https://doi.org/10.2307/2171739.

[6] Jeffrey M. Wooldridge. (2019). "Correlated Random Effects Models with Unbalanced Panels," *Journal of Econometrics*, Volume 211, Issue 1, Pages 137-150, ISSN 0304-4076, https://doi.org/10.1016/j.jeconom.2018.12.010.

[7] Gary Chamberlain. (1982). "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, Volume 18, Issue 1, Pages 5-46, ISSN 0304-4076, https://doi.org/10.1016/0304-4076(82)90094-X.

[8] Lee, Goeun; Han, Chirok. (2018). "Bias Reduction by Imputation for Linear Panel Data Models with Nonrandom Missing," *Journal of Economic Theory and Econometrics*, Vol. 29, No. 1, pp. 1-25.

[9] Zawadzki, Roy S.; Grill, Joshua D.; Gillen, Daniel L.; and the Alzheimer's Disease Neuroimaging Initiative. (2023). "Frameworks for estimating causal effects in observational settings: comparing confounder adjustment and instrumental variables," *BMC Medical Research Methodology*, Vol. 23, No. 122. Available at: https://doi.org/10.1186/s12874-023-01936-2.

[10] Pokropek, Artur. (2016). "Introduction to instrumental variables and their application to large-scale assessment data," *Large-scale Assessments in Education*, Vol. 4, No. 4. Available at: https://doi.org/10.1186/s40536-016-0018-2.

[11] Boneva, Lena; Linton, Oliver; Vogt, Michael. (2015). "A semiparametric model for heterogeneous panel data with fixed effects," *Journal of Econometrics*, Volume 188, Issue 2, Pages 327-345, ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2015.03.003.

[12] Bruno, Giovanni SF. (2005). "Approximating the bias of the LSDV estimator for dynamic unbalanced panel data models," *Economics Letters*, Volume 87, Issue 3, Pages 361-366.

[13] Baltagi, Badi; Song, Seuck. (2006). "Unbalanced panel data: A survey," *Statistical Papers*, Springer, Volume 47, Issue 4, Pages 493-523, October.

[14] Wooldridge, Jeffery. (1995)."Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions." *Journal of Econometrics*, Volume 68, Issue 1, Pages 115-132.

[15] Nijman, Theo; Verbeek, Marno. (1992). "Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function." *Journal of Applied Econometrics*, Volume 7, Issue 3, Pages 243-257. http://www.jstor.org/stable/2285097.

[16] Hausman, Jerry A.; Wise, David A. (1979). "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, Volume 47, Issue 2, Pages 455-473. https://doi.org/10.2307/1914193.

[17] Rochina-Barrachina, María Engracia. (1999). "A New Estimator for Panel Data Sample Selection Models," *Annals of Economics and Statistics*, GENES, Issue 55-56, Pages 153-181.

[18] Moffit, Robert; Fitzgerald, John; Gottschalk, Peter. (1999). "Sample Attrition in Panel Data: The Role of Selection on Observables," *Annales d'Économie et de Statistique*, No. 55/56, Pages 129-152. https://doi.org/10.2307/20076194.

[19] Kyriazidou, Ekaterini. (1997). "Estimation of a Panel Data Sample Selection Model," *Econometrica*, Volume 65, No. 6, Pages 1335-1364. https://doi.org/10.2307/2171739.

[20] Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. Available at: http://www.jstor.org/stable/j.ctt5hhcfr.

[21] Joshi, Riju; Wooldridge, Jeffrey M. (2019). "Correlated Random Effects Models with Endogenous Explanatory Variables and Unbalanced Panels," *Annals of Economics and Statistics*, No. 134, Pages 243-268. https://doi.org/10.15609/annaeconstat2009.134.0243.

[22] Yang, Ye; Dogan, Osman; Taspinar, Suleyman. (2023). "Estimation of Matrix Exponential Unbalanced Panel Data Models with Fixed Effects: An Application to US Outward FDI Stock," *Journal of Business & Economic Statistics*, Vol. 42, Pages 1-59. https://doi.org/10.1080/07350015.2023.2200486.

[23] Dustmann, Christian; Rochina-Barrachina, María Engracia. (2007). "Selection correction in panel data models: An application to the estimation of females' wage equations," *Econometrics Journal*, Royal Economic Society, Vol. 10(2), Pages 263-293, July.

[24] Ridder, Geert. (1990). "Attrition in Multi-Wave Panel Data." In *Panel Data and Labor Market Studies*, eds. J. Hartog, G. Ridder, and J. Theeuwes. Amsterdam: North-Holland.

[25] Verbeek, Marno; Nijman, Theo. (1996). "Incomplete Panels and Selection Bias." In *The Econometrics of Panel Data*, eds. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

[26] Van der Berg, G.; Lindeboom, M.; Ridder, G. (1994). "Attrition in Longitudinal Panel Data and the Empirical Analysis of Dynamic Labour Market Behavior." *Journal of Applied Econometrics*, Volume 99, Pages 421-435.

[27] Sasaki, Yuya. (2015). "Heterogeneity and Selection in Dynamic Panel Data." *Journal of Econometrics*, Volume 188, No. 1, Pages 236-249. https://doi.org/10.1016/j.jeconom.2015.05.002.

[28] Lee, Myoung-jae. (2001). "First-Difference Estimator for Panel Censored-Selection Models." *Economics Letters*, Volume 70, No. 1, Pages 43-49. https://doi.org/10.1016/S0165-1765(00)00350-5.

[29] Ridder, Geert. (1992). "An Empirical Evaluation of Some Models for Non-Random Attrition in Panel Data." *Structural Change and Economic Dynamics*, Volume 3, No. 2, Pages 337-355. https://doi.org/10.1016/0954-349X(92)90011-T.

[30] Zabel, Jeffrey E. (1992). "Estimating fixed and random effects models with selectivity," *Economics Letters*, Elsevier, Volume 40, Issue 3, Pages 269-272, November.

[31] Honoré, Bo E.; Kyriazidou, Ekaterini. (2000). "Panel Data Discrete Choice Models with Lagged Dependent Variables." *Econometrica*, Volume 68, No. 4, Pages 839–874. http://www.jstor.org/stable/2999528.

[32] Cui, Guowei, Norkute, Milda, Sarafidis, Vasilis, and Yamagata, Takashi. 2020. "Two-Stage Instrumental Variable Estimation of Linear Panel Data Models with Interactive Effects." *ISER DP No. 1101, 2020*. Available at SSRN: https://ssrn.com/abstract=3692123 or http://dx.doi.org/10.2139/ssrn.3692123.

[33] Vella, Francis, and Marno Verbeek. 1999. "Two-step estimation of panel data models with censored endogenous variables and selection bias." *Journal of Econometrics* 90, no. 2: 239-263.

[34] Abdmoulah, Walid (2020), "Competition and financial institutions and markets development: A dynamic panel data analysis," Mendeley Data, V1, doi: 10.17632/x98zr48x5n.1.

[35] Krifa-Schneider, Hadjila, Matei, Iuliana, and Sattar, Abdul (2022). "FDI, Corruption and Financial Development Around the World: A Panel Non-Linear Approach." *Economic Modelling*, vol. 110, 105809. ISSN: 0264-9993. Available at: https://doi.org/10.1016/j.econmod.2022.105809.

[36] Ahlin, Christian and Pang, Jiaren (2008). "Are Financial Development and Corruption Control Substitutes in Promoting Growth?" *Journal of Development Economics*, vol. 86, no. 2, pp. 414-433. ISSN: 0304-3878. Available at: https://doi.org/10.1016/j.jdeveco.2007.07.002.

[37] Song, Chang-Qing, Chang, Chun-Ping, and Gong, Qiang (2021). "Economic Growth, Corruption, and Financial Development: Global Evidence." *Economic Modelling*, vol. 94, pp. 822-830. ISSN: 0264-9993. Available at: https://doi.org/10.1016/j.econmod.2020.02.022.

[38] Shea, John (1997). "Instrument Relevance in Multivariate Linear Models: A Simple Measure." *The Review of Economics and Statistics*, vol. 79, no. 2, pp. 348–352. doi: https://doi.org/10.1162/rest.1997.79.2.348.