CONTROL FUNCTION METHODS IN APPLIED ECONOMETRICS

By

Riju Joshi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2018

ABSTRACT

CONTROL FUNCTION METHODS IN APPLIED ECONOMETRICS

By

Riju Joshi

This dissertation considers estimation and inference in three econometric models containing issues commonly encountered with observational data. Fundamental issues of self-selection, endogeneity, missing observations are pervasive in observational data. Moreover, often, the observations in a dataset are rarely statistically independent and have complex dependence structures. These issues can have a significant effect on the causal effect analysis and pose serious limitations on the popular methodologies that either maintain restrictive assumptions and/or require complicated and computationally tedious solutions.The dissertation aims to apply control function method as the primary tool to design estimation procedures under relaxed distributional and functional form assumptions. I describe computationally simple solutions to these issues to obtain more precise results. These estimation procedures are obtained under relaxed distributional and functional form assumptions allowing a researcher to incorporate more variability (or heterogeneity).

Chapter 1: Specification Tests in Unbalanced panels with Endogeneity *(joint work with Jeffrey M Wooldridge)*

This chapter develops specification tests for unbalanced panels with endogenous explanatory variables. We obtain a general equivalence results for the Random Effects 2SLS and Pooled 2SLS in an unbalanced panel. This algebraic result serves as the foundation to the fully-robust regression based Hausman Test to compare RE2SLS and FE2SLS estimators in form of a Variable Addition Test. In addition, we also obtain an equivalence result for Control Function estimators and FE2SLS estimators in an unbalanced panel. The results helps us to obtain regression-based fully robust specification test to check the correlation between the explana-

tory variables and the unobserved idiosyncratic errors. The test compares FE estimators with FE2SLS estimators

Chapter 2: Control Function Sieve Estimation of Endogenous Switching Models with Endogeneity *(joint work with Jeffrey M Wooldridge)*

In this chapter, we propose a sieve estimation procedure for estimating average treatment effects with a binary treatment in the framework of endogenous switching models. We consider a generalized model for the reduced form of the treatment variable that allows for the heterogeneity in terms of a distribution- free, conditional-heteroskedastic error term. We derive a simple, two-step estimation method that uses control function methods to correct for the endogeneity of the treatment assignment. consider the effect of attending a catholic high school on student math test scores.

Chapter 3:Control Function Estimation of Spatial Error Models with Endogeneity

This chapter considers estimation of linear regression models that allows some covariates to be endogenous when the data is suspected to exhibit spatial dependence. For example a hedonic price model in the housing markets not only has endogenous covariates such as schooling quality that are of prime interest but also often has spatially correlated neighborhood variables that are difficult to fully incorporate explicitly. These omitted spatially correlated neighborhood variables induce spatial correlation in the errors in the model. This paper uses control function method to control for endogeneity and incorporates the spatial dependence of data to achieve more precise results. I describe an estimation strategy that first divides the observations into groups based on the distance between them and then imposes control function assumptions to model the endogeneity *within* each group. A computationally simple two-step estimation procedure is suggested for a parametric estimation strategy where a *GLS-type* estimation is proposed that accounts for only the *within-group* correlations while ignoring the *across-group*

correlations. Results from the Monte Carlo simulation studies show that we obtain noticeable efficiency gains through this estimation procedure.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to the chair of my dissertation committee, Jeff Wooldridge, for his guidance and for his endless patience. He has always been supportive of me and my research and his kind words of encouragement really helped me grow as a researcher. I feel truly fortunate to have him as mentor. I would also like to thank Peter Schmidt, Kyooil Kim, and Saweda Liverpool-Tasie for serving on my committee and providing valuable feedback and assistance.

I also appreciate the comments of seminar participants at Michigan State University, Association of Public Policy Analysis and Management Fall 2016 Research Conference and North American Summer Meetings of the Econometric Society, and the 27th Annual Meeting of the Midwest Econometrics Group. I am specially thankful to the participants at the The Third MEG Mentoring Workshop for Junior Female Economists.

I am grateful for the financial support I received from the Graduate School and the Department of Economics at Michigan State University, including the Delia Koo Global Student Scholarship. I am especially thankful to Siddharth Chandra for providing me with financial support several times during my time in graduate school. I also appreciate the support and advice that Lori Jean Nichols and Todd Elder gave me as I navigated the graduate program and job market.

I am truly indebted to my friend Muzna for all the emotional support, companionship and caring she provided that helped me through the difficult times. I am also grateful to my dearest friends Annie, Pallavi, Alyssa, Akanksha, Danielle, Walter, Meenakshi, Udita, Ashesh and Kelsie for making my graduate experience so memorable.

I am truly grateful to my parents Prakash and Usha and my sister Richa for their selfless love and faith in me. They always encouraged and helped me at every stage of my personal and academic life. Finally, I am grateful to my husband and love of my life, Pathikrit for always

believing in me. He has always been there for me and without his love and support I would be lost. He is truly my rock.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

SPECIFICATION TESTS IN UNBALANCED PANELS WITH ENDOGENEITY

1.1 Introduction

Panel data has become very popular in contemporary empirical work specially in social and behavioral sciences. Hsiao (1985,1986), Klevmarken (1989) and Baltagi (2001) attribute this popularity to the ability of panel data to capture dynamics through its time dimension providing more variability in the data. Panel data allows the observations to have heterogeneity and this allows researchers to model complicated behavioral patterns. However, finding or constructing a balanced panel data is extremely rare. In most cases, observations on each time period for all cross-sectional units are not available and we have an unbalanced or incomplete panel dataset. This is particularly common when the cross-sectional unit is a firm, household or a person. For instance, an unbalanced panel might be a result of the survey design, as in the case of a rotating panel, where equally sized sets of sample units are brought in and out of the sample in some specified pattern. In other cases, incomplete panels might arise due to cross-sectional units dropping out leading to the problem of attrition.

Specification and estimation issues of econometric models for unbalanced panels have primarily focused on testing for the presence of selection bias and estimating models if sample selection is present. Nijman and Verbeek (1992) develop a simple test to check for sample selection bias in random effects framework and Wooldridge (2010) extends this for fixed effects. Wooldridge (1995) develops variable addition tests for selection bias and these tests are further extended for models with endogenous variables in Semykina and Wooldridge (2010). Hsiao et al. (2008) propose limited information test to check for selection issues. In addition to testing for sample selection, a number of studies have addressed the issue of treatment of unbalanced panels in the presence of selection bias. Numerous parametric and semi-parametric solutions

to correct for sample selection bias have been proposed in econometrics literature. Wooldridge (1995), Semykina and Wooldridge (2010), Kyriazidou (1997), Rochina-Barrachina (1999) are just a few notable examples that consider the estimation in unbalanced panels for both linear and non-linear models with exogenous and endogenous covariates.

In the absence of selection bias, we can extend the models and the estimation methods for balanced panels to their unbalanced counterpart. Specification testing for drawing comparisons between different estimation methods can also be extended for unbalanced panels. Specification tests such as Hausman Test to compare fixed effects and random effects have been developed for unbalanced panels with exogenous sampling. However, as is the case with balanced panels, the traditional Hausman test maintains the assumption that the conditional variances of the composite error terms[1] have the random effects structure. This becomes one of the key limitation of the traditional Hausman Test as the failure of this assumption distorts the asymptotic distribution of the test statistic.

Wooldridge (2010) explains why we need to develop a test statistic that is robust to the violation of the random effects structure of the composite errors. A comparison between the fixed effects and random effects estimator essentially boils down to testing the correlation between unobserved heterogeneity and covariates. This is captured in the conditional first moment assumptions of the model. The assumption that the composite errors have the random effects structure, on the other hand, is an assumption about the second moments. Traditional Hausman test is concerned with verifying the validity of the former while also maintaining the latter. Failure of the second moment assumptions have serious consequences for the test statistic which is primarily concerned with the first moment. In fact, in this case it causes the test statistic to have a non-standard asymptotic distribution. Thus a non-robust Hausman test statistic which tests conditional mean specifications has no systematic power against the violation of the conditional variance specifications.

[1] Unobserved heterogeniety and idiosyncratic errors are clubbed together to form composite errors

The limitations of the traditional Hausman specification test when used as a pretest of random effect specification are also studied in Guggenberger (2010) . In particular, it is shown both theoretically and through Monte-Carlo simulations that the asymptotic size of the t-statistic that is based on either the random effect or fixed effect specification based on the outcome of Hausman pretest, is severely distorted.

For balanced panels, this issue is addressed by the fully-robust regression-based Hausman Test. The fundamental idea behind regression-based tests in panel data is given by Correlated Random Effects models due to Mundlak (1978). For unbalanced panels with exogenous explanatory variables, correlated random effects models are developed in Wooldridge (2016) that subsequently lead to a simple fully-robust Hausman specification tests to compare Fixed Effects and Random Effects estimators. Models with individual specific slopes are also considered and correlated random effect assumption is used to develop tests for correlation between the selection and heterogeneous slopes.

This paper extends Wooldridge (2016) for the unbalanced panels where some elements of the time-varying explanatory variables are allowed to be correlated with the unobserved idiosyncratic shocks. In particular, Correlated Random Effects models are developed for unbalanced panels with endogeneity and simple specification tests are suggested to compare Fixed Effects 2SLS (FE2SLS) and Random Effects 2SLS (RE2SLS) estimators. Wooldridge (2016) obtains an algebraic equivalence result where Fixed Effects estimator is computed as a Pooled OLS estimator (P2SLS) of the model by adding time averages of the covariates (averaged across the unbalanced panel) as additional explanatory variables. We obtain a similar result for the case when some of the covariates are allowed to be endogenous.

Regression based Hausman test to check for correlation between the instruments and individual heterogeneity in unbalanced panel data models begins with modeling of the unobserved heterogeneity in terms of the time averages of the instruments. While this seems to be a natural extension of the balanced panel models, Wooldridge (2016) explains how CRE models in the

unbalanced panels differ from their balanced counterpart. The unbalanced nature of the panel is reflected in the time averages of the instruments as the number of time periods for which observations are available differs across different cross-sectional units. Thus time averages of the instruments are defined only for the full set of observations. In addition, unlike in the balanced case, the time averages of aggregate time variables are also included because we average different time periods for different cross-sectional units. In other words, in unbalanced panels, the unobserved heterogeneity is modeled in terms of the instruments *and* the selection.

This paper uses Mundlak (1978) assumption to model the unobserved heterogeneity, and we get an algebraic result that the FE2SLS estimator for the coefficient on the covariates is obtained by doing a P2SLS or RE2SLS on the augmented model. This algebraic result showing the equivalences of the estimators serves as the building block for the specification test that compares FE2SLS and RE2SLS estimators. It provides a way to obtain a regression based Hausman test that is fully-robust to the second moment conditions of the composite errors.

We also consider a test to check the endogeneity of some explanatory variables using the control function approach. In addition to being a bit unwieldy, the traditional Hausman test for checking the endogenous explanatory variables suffers from shortcomings like giving a wrong degrees of freedom and often gives a negative χ^2 test statistic. Moreover, it is also not robust to heteroscadasticity. To address this issue, we adopt the control function approach and obtain an equivalence result. More specifically, we show that adding the residuals from the reduced form equation to the original model yields FE2SLS estimators. This result is then used to develop a simple regression based fully robust Hausman Test for endogeneity of the explanatory variables.

The paper is structured as follows. Section 1.2 introduces the general model for unbalanced panels and the assumptions maintained in this paper. Section 1.3 specifies the key estimation methods for unbalanced panels with endogeneity, namely FE2SLS and RE2SLS. Section 1.4 obtains algebraic equivalences between different estimators. In Section 1.5, we develop a simple fully-robust regression based Hausman specification test to compare Random Effects 2SLS

and Fixed Effects 2SLS estimators. In Section 1.6, we consider the control function approach to detect the possible endogeneity of explanatory variables in unbalanced panel data model. In Section 1.7, we briefly talk about an empirical strategy that could be followed as protocol for approaching endogeneity issues in a linear model with unbalanced panels. We illustrates this strategy and our theoretical findings with an empirical application in Section 1.8. More specifically, we study the effects of spending on student performance in Michigan schools. Section 1.9 concludes the paper.

1.2 Model

We begin by assuming that a random sample is drawn from an underlying population that consists of a large number of units for whom data on T time periods are potentially observable.

In our model, for an individual i, at time period t, y_{it} denote the potentially observed outcome variable, $\mathbf{x}_{it}$ is a $1 \times K$ vector of potentially observed time-variant covariates and $\mathbf{w}_i$ denotes the set of time-invariant variables (that contains unity). In addition to the potentially observable variables, we also draw unobservables for each i and c_i denotes the unobserved heterogeneity associated with each i. The idiosyncratic errors are denoted by u_{it}.

The standard linear model with additive heterogeneity is given as:

Assumption 1.2.1

$$y_{it} = \mathrm{x}_{it}\boldsymbol{\beta} + \mathrm{w}_i\boldsymbol{\delta} + c_i + u_{it} \tag{1.1}$$

We believe that some elements of $\mathbf{x}_{it}$ are correlated with u_{it}, or even u_{ir} with $r \neq t$. To deal with this endogeneity issue we have a set of $1 \times L$ possible instrumental variables $\mathbf{z}_{it}$ with $L \geq K$. This set of instruments $\mathbf{z}_{it}$ not only includes the excluded exogenous variables but also all the elements of $\mathbf{x}_{it}$ that are exogenous.

To allow for an unbalanced panel, we introduce a binary selection indicator s_{it} which is defined as

Assumption 1.2.2

$$s_{it} = \begin{cases} 1 & \textit{if and only if } (\mathrm{x}_{it}, y_{it}, \mathrm{z}_{it}) \textit{ is fully observed} \\ 0, & \textit{otherwise} \end{cases} \tag{1.2}$$

In other words, $\mathbf{s}_i \equiv \{s_{i1}, s_{i2}, ..., s_{iT}\}$ is the series of selection indicator for each i. This implies that $s_{it} = 1$ if time period t for unit i can be used in estimation. The number of time periods for which a unit i is observed is denoted by T_i which is simply equal to $\sum_{r=1}^{T} s_{ir}$. Our panel data can be concisely represented as a vector of a randomly drawn sample across the cross section dimension i , with fixed time periods T: $\{(s_{it}, \mathbf{x}_{it}, y_{it}, \mathbf{z}_{it}); c_i\}$.

Since we allow for endogeneity of some of the explanatory variables $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$, our assumptions in this paper primarily structure the relationship between the instruments $\mathbf{z}_i \equiv (\mathbf{z}_{i1}, ..., \mathbf{z}_{iT})$, selection indicators $\mathbf{s}_i$, unobserved individual heterogeneity c_i and the idiosyncratic errors $\mathbf{u}_i = (u_{i1}, ..., u_{iT})$.

1.3 Estimation Methods

We consider two main estimation methods to estimate our key parameter of interest $\boldsymbol{\beta}$: Fixed Effects 2SLS and Random Effects 2SLS.

1.3.1 Fixed Effects 2SLS (FE2SLS)

First, we consider the case where unobserved heterogeneity c_i is allowed to be correlated with the history of instruments $\mathbf{z}_i$. As with the balanced panel analysis, FE2SLS approach for the unbalanced panel data transforms (1.1) to eliminate the unobserved effect c_i. In unbalanced panels, the fixed effects transformation, also called the within transformation is obtained by first multiplying equation (1.1) by s_{it}:

$$s_{it} y_{it} = s_{it} \mathbf{x}_{it} \boldsymbol{\beta} + s_{it} \mathrm{w}_i \boldsymbol{\delta} + s_{it} c_i + s_{it} u_{it} \tag{1.3}$$

Averaging this equation across t for each i gives us the time averaged equation:

$$\bar{y}_i = \bar{\mathrm{x}}_i\boldsymbol{\beta} + \mathrm{w}_i\boldsymbol{\delta} + c_i + \bar{u}_i \tag{1.4}$$

The time averages are given by $\bar{y}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ and $\bar{\mathrm{x}}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}$. Note that the time averages for $y_{it}, \mathbf{x}_{it}$, (and also $\mathbf{z}_{it}$) are computed only for periods when data exists on the full set of variables. (1.4) is of interest in its own right because its Pooled 2SLS estimation using $\bar{\mathrm{z}}_i$ as instruments gives us the Between 2SLS estimator of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}_{B2SLS}$

Time demeaning (1.3) using (1.4) we get:

$$s_{it}(y_{it} - \bar{y}_i) = s_{it}(\mathbf{x}_{it} - \bar{\mathrm{x}}_i)\boldsymbol{\beta} + s_{it}(u_{it} - \bar{u}_i) \tag{1.5}$$

We denote the time demeaned variables as $\ddot{\mathrm{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathrm{x}}_i)$ and $\ddot{y}_{it} = (y_{it} - \bar{y}_i)$.

The FE2SLS estimator is obtained by estimating (1.5) by Pooled OLS using $\ddot{\mathrm{z}}_{it} \equiv (\mathbf{z}_{it} - \bar{\mathrm{z}}_i)$ as instruments:

$$\hat{\boldsymbol{\beta}}_{FE2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}_{it}'\ddot{\mathrm{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}_{it}'\ddot{\mathrm{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}_{it}'\ddot{\mathrm{x}}_{it}\right)\right]^{-1} \times \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}_{it}'\ddot{\mathrm{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}_{it}'\ddot{\mathrm{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}_{it}'\ddot{y}_{it}\right)\right]$$

The key assumptions sufficient for the consistency of FE2SLS estimator on the unbalanced panel can be stated as:

Assumption 1.3.1 *For all* $i = 1, ..., N$

- <u>*FE2SLS.1*</u> $\mathbb{E}[u_{it}|\mathrm{z}_i, c_i, \boldsymbol{s}_i, \mathrm{w}_i] = 0$
- <u>*FE2SLS.2*</u> $\sum_{t=1}^{T}\mathbb{E}[s_{it}\ddot{\mathrm{z}}_{it}'\ddot{\mathrm{x}}_{it}]$ *is full rank and* $\sum_{t=1}^{T}\mathbb{E}[s_{it}\ddot{\mathrm{z}}_{it}'\ddot{\mathrm{z}}_{it}]$ *is full rank*

Assumption FE2SLS.1 implies two things. First is the strict exogeneity of selection and instruments with respect to the idiosyncratic errors. Second, we allow for correlation between c_i and vector of instruments $\mathbf{z}_i$. We also allow for selection s_{it} at time t to be correlated with

$(\mathbf{z}_i, c_i)$. Assumption FE2SLS.2 is the appropriate rank condition that ensures invertibility of matrices in an unbalanced panel data. It naturally implies that any time-invariant variables are dropped out of our the fixed effects analysis. Under the assumptions stated above, FE2SLS on the unbalanced panel is consistent and is asymptotically normal.

1.3.2 Random Effects

Fixed effects estimation methods suffer from a few limitations that arise due to the demeaning of the variables. As illustrated in (1.5), time-invariant observables are eliminated in the process of eliminating c_i. In addition, because of time-demeaning, any observation with $T_i = 1$ also drops out. In addition, much of the variation in the data is also removed in the demeaning process. Random Effects estimation provides a remedy to these issues by imposing additional assumptions.

Assumption 1.3.2 *For all* $i = 1, ..., N$

- *RE2SLS 1.* $\mathbb{E}[c_i|\mathrm{w}_i, \mathrm{z}_i, \boldsymbol{s}_i] = \mathbb{E}[c_i] = 0$

$\mathbb{E}[c_i] = 0$ can be assumed whenever we include an intercept in our model. Random Effects transformation becomes straightforward if we add the assumption that:

Assumption 1.3.3 *For all* $i = 1, ..., N$

- *RE2SLS 2.* $\mathbb{E}[\boldsymbol{u}_i \boldsymbol{u}_i'|\mathrm{w}_i, \mathrm{z}_i, c_i, \boldsymbol{s}_i] = \sigma_u^2 \mathbf{I}_T$ *and* $\mathbb{E}[c_i|\mathrm{w}_i, \mathrm{z}_i, \boldsymbol{s}_i] = \sigma_c^2$

Analogous to the case of balanced panel data, we get a straightforward Generalized Least Square(GLS) transformation when we define

$$\theta_i = 1 - \left[\frac{\sigma_u^2}{(\sigma_u^2 + T_i \sigma_c^2)}\right]^{\frac{1}{2}} \tag{1.6}$$

where θ_i is viewed as a random variable which is a function of T_i. In addition, T_i is also exogenous since $\mathbb{E}[u_{it}|\mathrm{w}_i, \mathbf{z}_i, c_i, \mathbf{s}_i] = 0$ and $\mathbb{E}[c_i|\mathrm{w}_i, \mathbf{z}_i, \mathbf{s}_i] = \mathbb{E}[c_i] = 0$. This implies that $\mathbb{E}[c_i + u_{it}|\mathrm{w}_i, \mathbf{z}_i, T_i] = 0$.

A Pooled 2SLS on the selected sample of $(y_{it} - \theta_i \bar{y}_i)$ on $(\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)$ and $(1-\theta_i)\mathbf{w}_i$ using $(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)$ as instruments will give us RE2SLS estimator: $\hat{\boldsymbol{\beta}}_{RE2SLS}$. Not surprisingly, just as in the balanced panel data case, the consistency of our estimator will not be affected if we use an incorrect variance-covariance structure. We would just calculate the fully robust variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{RE2SLS}$. However, this is true only under the assumption of strict exogeneity of instruments and selection with respect to $c_i + u_{it}$. As Wooldridge (2010) notes, this is a very restrictive assumption.

1.4 An Algebraic Equivalence Result

Wooldridge (2016) obtains a general equivalence result for the Random Effects and Pooled OLS in unbalanced panels. We extend the equivalence result further for the RE2SLS and Pooled 2SLS in an unbalanced panel. As we will see below, the equivalence result requires no assumption except that the appropriate matrices are invertible in the sample.

Consider the following regression model:

$$(y_{it} - \theta_i \bar{y}_i) = (\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (1-\theta_i)\bar{\mathbf{z}}_i \boldsymbol{\xi} + (1-\theta_i)\mathbf{w}_i \boldsymbol{\delta} + u_{it} \tag{1.7}$$

where we follow the notation of Section 2. The Pooled 2SLS estimator for $\boldsymbol{\beta}$ for selected sample (i.e for $s_{it} = 1$) using $(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)$ as instruments for $(\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)$ is given as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{P2SLS} = [(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)'\ddot{\mathbf{z}}_{it})(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'(\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i))]^{-1} \times \\ [(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)'\ddot{\mathbf{z}}_{it})(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'(y_{it} - \theta_i \bar{y}_i))]\end{aligned} \tag{1.8}$$

where as before, we define $\ddot{\mathbf{z}}_{it} = (\mathbf{z}_{it} - \bar{\mathbf{z}}_i)$. This expression is obtained by following the simple idea of the Frisch-Waugh-Lovell theorem and extending it for the instrumental variables estimation. (See Section 1.9.1)

Theorem 1 *If* $(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{z}_{it})$ *and* $[(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\boldsymbol{x}_{it} - \theta_i\bar{\boldsymbol{x}}_i)'\ddot{z}_{it})(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{z}_{it})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}(\boldsymbol{x}_{it} - \theta_i\bar{\boldsymbol{x}}_i))]$ *are nonsingular matrices, then the Pooled 2SLS estimator obtained above is algebraically equivalent to the Fixed Effect 2SLS estimator on the unbalanced panel:*

$$\hat{\boldsymbol{\beta}}_{P2SLS} = \hat{\boldsymbol{\beta}}_{FE2SLS}$$

Proof. For $\theta_i = 0, \forall i$,

$$\hat{\boldsymbol{\beta}}_{P2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}'_{it}\ddot{z}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\mathbf{x}_{it}\right)\right]^{-1} \times \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}'_{it}\ddot{z}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}y_{it}\right)\right]$$

Noting that $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}'_{it}\ddot{z}_{it}\right)$ simplifies to $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathrm{x}'_{it}\ddot{z}_{it}\right)$, we get the expression for $\hat{\boldsymbol{\beta}}_{FE2SLS}$

For $0 \leq \theta_i < 1$, first consider $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)'\ddot{z}_{it}\right)$.

$$\begin{aligned}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)'\ddot{z}_{it}\right) &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}'_{it}\ddot{z}_{it} - \theta_i\bar{\mathbf{x}}'_i\ddot{z}_{it})\right) \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}'_{it}\ddot{z}_{it}\right) - \left(\sum_{i=1}^{N}\theta_i\bar{\mathbf{x}}'_i\sum_{t=1}^{T} s_{it}\ddot{z}_{it}\right)\end{aligned}$$

Note that $\sum_{t=1}^{T} s_{it}\ddot{z}_{it} = 0$. So we get

$$\begin{aligned}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)'\ddot{z}_{it}\right) &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}'_{it}\ddot{z}_{it}\right) \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}'_{it}\ddot{z}_{it}\right)\end{aligned}$$

Similarly, we get:

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}(y_{it}-\theta_i\bar{y}_i)\right) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}y_{it}\right)$$
$$= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{y}_{it}\right)$$

Substituting these expressions, our 2SLS coefficient becomes:

$$\hat{\boldsymbol{\beta}}_{P2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}'_{it}\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{\mathbf{x}}_{it}\right)\right]^{-1} \times$$
$$\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}'_{it}\ddot{z}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{z}'_{it}\ddot{y}_{it}\right)\right]$$

$$= \hat{\boldsymbol{\beta}}_{FE2SLS}$$

■

1.5 Regression based fully-robust Hausman Test to compare RE2SLS and FE2SLS

Specification test in this section is concerned with comparing the FE2SLS and RE2SLS estimators. In other words, we want to test that heterogeneity is mean independent of the instruments and selection in all time periods. As we mentioned before, the traditional Hausman test statistic to compare RE2SLS and FE2SLS suffers from several limitations. An elegant way to deal with this is provided by the regression-based fully robust Hausman test. We apply the principle of the Wu-Hausman endogeneity test with additionally using Mundlak (1978) device to modify our regression model.

A comparison of of FE2SLS and RE2SLS estimators in unbalanced panels is essentially a test of correlation between unobserved heterogeneity and instruments. In other words, Hausman

test can be interpreted as a test of

$$\mathbb{E}[c_i|\mathbf{z}_i,\mathbf{s}_i] = \mathbb{E}[c_i] \tag{1.9}$$

In particular, our model is

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + c_i + u_{it} \tag{1.10}$$

for $i = 1,2,...,N$ and $t = 1,2,...,T$. Recall that $\mathbf{w}_i$ contains unity.

We specify a correlated random effects structure for c_i due to Mundlak (1978):

$$c_i = \xi_0 + \bar{\mathrm{z}}_{\boldsymbol{i}}\boldsymbol{\xi} + a_i, \tag{1.11}$$

where $Cov[a_i,\mathrm{z}_i] = 0$, $\mathbb{E}[a_i] = 0$

Substituting for c_i in our regression model, we get the usual Mundlak equation:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + \bar{\mathrm{z}}_{\boldsymbol{i}}\boldsymbol{\xi} + a_i + u_{it} \tag{1.12}$$

where ξ_0 is absorbed into the intercept in w_i. Now regression based fully-robust Hausman test is simply an application of Theorem 1. If we estimate the above equation by RE2SLS or Pooled 2SLS using $(\mathbf{z}_{it},\bar{\mathrm{z}}_i)$ as instruments then we know from the Theorem 1 that the resulting estimator of $\boldsymbol{\beta}$ is the FE2SLS estimator. The regression based Hausman test is simply a Wald test of

$$H_0 : \boldsymbol{\xi} = 0$$

To obtain a fully robust test, we estimate (1.12) by pooled 2SLS and use cluster-robust inference.

Testing for Correlation between Selection and Idiosyncratic Errors

The test developed above helps us to rule out estimation of parameters by RE2SLS method. However, FE2SLS estimation methods are consistent only if selection is strictly exogenous with respect to the idiosyncratic errors conditional on the unobserved effect. Thus, ruling out the possibility of relation between selection (s_{it}) and idiosyncratic errors (u_{it}), conditional on

c_i is imperative before we proceed. Further, in our context, it will be erroneous to use Heckman (1976) test (extended to the panel data context in Wooldridge (1995)) because it requires the exogenous components of x_{it} to be observable in all time periods. We do not impose such restriction in our model.

Nijman and Verbeek (1992) suggest a simple test for testing of selection bias in the context of random effects which also works in the fixed effects estimation. We simply add the lagged value of the selection indication : $s_{i,t-1}$ to the original model and check for its significance. We can also add T_i and check for its significance using a t-test. This test is extended to the fixed effects framework in Wooldridge (2010). We add $s_{i,t-1}$ in our model and estimate the model by fixed effect using $s_{it} = 1$. A simple robust t-test of the coefficient on $s_{i,t-1}$ tests the null hypothesis that selection in the previous period is not significant. Another alternative which is useful in the attrition problems is to add the lead value of the selection indicator: $s_{i,t+1}$.

Once we rule out selection bias in our data, we can now rely on Fixed Effects estimation methods to estimate the coefficients of our model. We would also like to check for the endogeneity of our explanatory variables. Next section develops simple regression based fully-robust tests for endogeneity.

1.6 Robust Hausman Test to compare FE vs FE2SLS

Previous sections focus on developing the test for checking if the instruments are correlated with the individual heterogeneity. A natural addition to those tests would be to check whether or not the explanatory variables are correlated with the idiosyncratic shocks. Traditional Hausman Test for checking endogeneity of explanatory variables in panel data suffers from several shortcomings (Baum (2006)). It often generates negative χ^2 test statistic which makes the test infeasible. In addition, often the degrees of freedom are wrongly calculated and this leads to degeneracies. And lastly, the traditional Hausman test statistic is not robust to heteroscadasticity

and the robust versions of the test are not readily available in most softwares. [2]

In this section, we use control function approach to check the endogeneity of explanatory variables. Control function approach deals with endogenous explanatory variables by formalizing the correlation between the endogenous explanatory variables and unobservables of idiosyncratic shocks. In the cross-section data analysis with linear endogenous explanatory variables, it has been shown that control function approach leads to the same estimators as that of 2SLS. We will show in this section that this algebraic result also holds in unbalanced panels. In the context of panel data, the control function approach would yield FE2SLS estimators. Through this result, we are able to obtain a regression-based fully robust specification test to compare fixed effects estimators with fixed effect 2SLS estimators.

1.6.1 Model

The model is defined in terms of the following assumptions:

Assumption 1.6.1 *For all $i = 1,...,N$ the first assumption specifies a linear equation for the outcome variable:*

$$y_{it1} = \mathrm{x}_{it}\boldsymbol{\beta} + \mathrm{y}_{it2}\boldsymbol{\alpha} + c_{i1} + u_{it} \tag{1.13}$$

y_{it1} denotes the potentially observed outcome variable. In this section we slightly modify the notations and explicitly denote the $1 \times K_{y_2}$ vector of endogenous variables by $\mathbf{y}_{it2}$. In other words, while $\mathbf{x}_{it}$ denote the $1 \times K_x$ vector of potentially observed explanatory variables that are not correlated with the idiosyncratic errors u_{it}, some elements of $\mathbf{y}_{it2}$ are allowed to be correlated with u_{it}. We introduce the $1 \times K_w = 1 \times (K_x + K_z)$ vector $\mathbf{w}_{it}$ as the full set of instruments that also contain exogenous variables, i.e $\mathbf{w}_{it} = \{\mathbf{x}_{it}, \mathbf{z}_{it}\}$, where the $1 \times K_z$ $\mathbf{z}_{it}$ serve as instruments for $\mathbf{y}_{it2}$. The reduced form equation for y_{it2} is given by:

[2] We could always set up the whole model as a GMM problem.

Assumption 1.6.2 *We assume that the reduced form of the endogenous variable is a set of linear equations:*

$$\boldsymbol{y}_{it2} = \mathrm{w}_{it}\boldsymbol{\gamma} + c_{i2} + \boldsymbol{v}_{it} \tag{1.14}$$

We introduce the selection indicator as:

Assumption 1.6.3

$$s_{it} = \begin{cases} 1 & \textit{if and only if } (\mathrm{w}_{it}, y_{it}) \textit{ is fully observed} \\ 0, & \textit{otherwise} \end{cases} \tag{1.15}$$

As before, the time averages are computed only for periods when data exists on the full set of variables.

To obtain regression based fully-robust Hausman test checking for the endogeneity of y_{it2}, we first estimate the reduced form (1.14) by Fixed Effects using the selection sample, i.e for $s_{it} = 1$. We obtain residuals from this regression, and denote them by $\hat{\mathrm{v}}_{it}$. Next, we augment equation (1.13) with $\hat{\mathrm{v}}_{it}$ and obtain:

$$y_{it1} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{y}_{it2}\boldsymbol{\alpha} + \hat{\mathrm{v}}_{it}\boldsymbol{\rho} + error_{it} \tag{1.16}$$

(1.16)is called the control function equation and serves as the primary equation for obtaining the Hausman test. The $error_{it}$ term comprises of both individual heterogeneity and idiosyncratic error. Our key algebraic result is stated in Theorem 2:

Theorem 2 *Estimate the augmented equation* (1.16) *by Fixed Effects using the selected sample and let* $\tilde{\boldsymbol{\beta}}_{FE(aug.)}$ *and* $\tilde{\boldsymbol{\alpha}}_{FE(aug.)}$ *denote the Fixed Effect estimators of the augmented equation. Then,*

$$\tilde{\boldsymbol{\beta}}_{FE(aug.)} = \hat{\boldsymbol{\beta}}_{FE2SLS} \textit{ and } \tilde{\boldsymbol{\alpha}}_{FE(aug.)} = \hat{\boldsymbol{\alpha}}_{FE2SLS}$$

The proof of the result is shown in the Section 1.10.2. This result essentially gives us an elegant way to obtain the regression based test to check the possible endogeneity of the

explanatory variables. Following the foundations of Section 4, regression based fully-robust Hausman Test is given by a simple Wald test of

$$H_o : \boldsymbol{\rho} = 0$$

using robust standard errors.

1.7 A Strategy for an Applied Econometrician

In this paper, we have suggested two kinds of specification tests. One is to test the endogeneity of the explanatory variables with the time-varying idiosyncratic shocks (Section 1.6) and the other is to test the endogeneity of the instruments with the time-constant unobserved individual effects (Section 1.5). An empirical econometrician would begin with testing the endogeneity of the explanatory variables in her model. This would essentially mean that we would compare FE estimator with FE2SLS estimator using the regression based fully-robust Hausman Test that essentially takes the form of a Variable Addition Test (VAT) given in Section 1.5. A failure to reject the null implies that we can consider our explanatory variables to be exogenous with respect to the time-varying unobserved idiosyncratic shocks and go ahead with the usual Random Effects and Fixed Effects analysis. We could further compare the RE and FE estimators using a Hausman Test.

A rejection of the null implies that we need to account for the endogeneity of our explanatory variables. This entails using instruments and we have two key estimation methods: RE2SLS and FE2SLS. Now, it is possible that our instruments are exogenous not only with respect to the idiosyncratic errors but also with respect to the unobserved time-invariant individual heterogeneity. If FE2SLS estimates seem to be imprecise, then we can use the VAT version of Hausman Test described in Section 6 to compare RE2SLS and FE2SLS estimators. This would help us to determine which of the estimation methods fit our model the best.

1.8 Empirical Illustration

To illustrate the methods above, we consider the problem of estimating the effects of spending on student performance. We use the data on standardized test scores from 1992 to 1998 at Michigan schools to determine the effects of spending on math test outcomes for fourth graders. Papke (2005) studies this using school-level data and a linear functional form. Papke and Wooldridge (2008) extend the analysis by recognizing the fractional nature of the pass rates. Specifically, they use fractional response models for the district level panel data. Both find non-trivial effects of spending on test pass rates. Since we deal with linear models in our paper, our analysis is closer to Papke (2005).

1.8.1 Background

Funding for K-12 schools in Michigan dramatically changed in 1994 from local, property-tax based system to a statewide system supported primarily through a higher sales tax. The primary goal of this policy change was to equalize spending and this was reflected in the rise of per-pupil spending. Papke (2005) studies the effect of this policy change on student performance. The data used comes from annual Michigan School Reports (MSRs). The outcome variable of study is the percentage of students passing the Michigan Educational Assessment Program (MEAP) math test for 4th graders: *math4*. The key explanatory variable is log of average per pupil expenditure: *log(avgrexpp)* which serves as a measure of per-pupil spending.

The data used in Papke (2005) is an unbalanced panel data. In this empirical illustration, we revisit this problem taking the note of the incomplete nature of the panel. In addition, we use Stata 14 and have fully robust standard errors as 'clustering' is available for all the regressions. [3]

[3]Previous versions of Stata do not allow to compute fully robust standard errors for example in RE2SLS. However, we could bootstrap to obtain proper standard errors.

1.8.2 Results

Since our purposes are primarily illustrative, we focus on the simple specification:

$$math4_{it} = \theta_t + \beta_1 log(avgrexpp_{it}) + \beta_2 lunch_{it} + \beta_3 log(enroll_{it}) + c_{i1} + u_{it} \tag{1.17}$$

where i indexes school and t indexes year. In this specification, θ_t is captured by adding time dummies. The covariate vector is given as $\mathbf{x}_{it} = [log(avgrexpp_{it}), lunch_{it}, log(enroll_{it})]$. Papke (2005) argues that $log(avgrexpp_{it})$ could be endogenous as spending could be correlated with the idiosyncratic shocks u_{it}. She uses the district foundation grant $log(found_{it})$ as instruments. Thus, we have a vector of instruments : $\mathbf{z}_{it} = [log(found_{it}), lunch_{it}, log(enroll_{it})]$, where $[lunch_{it}, log(enroll_{it})]$ serve as their own instruments.

A simple Nijman-Verbeek (1992) test verifies that the selection is not correlated with the idiosyncratic shocks. Specifically, we add the lagged value of the selection indicator to our model and found it insignificant. This allows us to apply our tests to this empirical problem.

We begin by conducting a test to check the endogeneity of the explanatory variables. As mentioned before, Papke (2005) argues that the primary variable of interest *log(avgrexpp)* is endogenous in the sense that it is correlated with the time varying idiosyncratic erros. She verifies this claim using a fully robust Hausman test that compares the Pooled OLS and Pooled 2SLS estimators. In this paper, we further verify the endogeneity of *log(avgrexpp)* using the control function approach as described in Section 6. More specifically, we begin by estimating the reduced form equation:

$$log(avgrexpp_{it}) = \phi_t + \pi_1 lunch_{it} + \pi_2 log(enroll_{it}) + \pi_3 log(found_{it}) + c_{i2} + v_{it2} \tag{1.18}$$

using fixed effects. The residuals from this regression are denoted by $\hat{v}_{it2}$. The control function equation would be equation (1.17) augmented with $\hat{v}_{it2}$:

$$math4_{it} = \theta_t + \beta_1 log(avgrexpp_{it}) + \beta_2 lunch_{it} + \beta_3 log(enroll_{it}) + \rho \hat{v}_{it} + c_{1i} + error_{it} \tag{1.19}$$

We then estimate the above equation using fixed effects. The results are given in Column (1) of Table 1. We see that the equivalence result holds and the estimates are equal to the FE2SLS

estimates given in Column (2). To check for the endogeneity of $log(avrgexpp)$, we check the significance of the estimate of the coefficient on $\hat{v}$. We see that it is significantly different from zero thus we can conclude that average spending per pupil is endogenous. This is the rejection at 10 percent level. If we think of it as rejection, then we can go ahead and do a RE2SLS and FE2SLS analysis. However, since this is only at 10 percent level, then we can also only look at FE and RE. This is done in the Column (2) and Column (3) of Table 1.1. We compare FE and RE estimates through the regression based Hausman test for unbalanced panels given in Wooldridge (2016). The regression model augumented with the Mundlak's device is estimated using Random Effects and the standard errors are robust. The results are given in Columm (4). To test for the correlation of the explanatory variables with the individual heterogeneity, we check the joint significance of $\{\overline{lunch_i}, \overline{log(enroll)}_i, \overline{log(avgrexpp)}_i, \overline{y96}, \overline{y97}, \overline{y98}\}$. The $\boldsymbol{\chi}_2$ and the p values of the test are given in Table 2. We can clearly infer that null of no correlation between the explanatory variables and unobserved individual time-invariant heterogeneity is rejected, validating the FE estimation method.

Next, we do a RE2SLS and FE2SLS analysis. The results are given in Columns (5) and (6) of Table.1 Both give a statistically significant estimate of the coefficient on $log(avgrexpp)$. The results verify that the effects of spending on student performance are non-trivial. This is consistent with the results obtained in Papke (2005) and Papke and Wooldridge (2008). We find that the RE2SLS estimates are quite different from the FE2SLS estimates and this motivates us to test for the correlation of the instrument with the individual heterogeneity. We use the Mundlak (1978) device and apply the regression based fully-robust Hausman Test developed in Section 5. This also allows us to verify our equivalence results. Recall that we model the individual heterogeneity as: $\mathbb{E}[c_{1i}|\mathbf{z}_i] = \xi_0 + \bar{\mathbf{z}}_i\boldsymbol{\xi}$ and add it to our model. Our estimating equation becomes:

$$\begin{aligned} math4_{it} = \quad & \theta_t + \beta_1 log(avgrexpp)_{it} + \beta_2 lunch_{it} + \beta_3 log(enroll)_{it} \\ & + \xi_1 \overline{lunch_i} + \xi_2 \overline{log(enroll)}_i + \xi_3 \overline{log(found)}_i + \eta_{it} \end{aligned} \tag{1.20}$$

Aggregate time dummies should be added in the specification. In other words, $\bar{\boldsymbol{z}}_{\boldsymbol{i}}$ include the averages of the year dummies also. This is an important aspect in which our analysis differs due to the unbalanced nature of the panel. Since different individuals have different T_i, we are averaging over different time periods for different i. Thus the time averages of the aggregate time variables changes across i.

We estimate the equation(1.20) by RE2SLS and the results are given in Column (7) of Table 1.1. The estimates verify our equivalence result for the unbalanced panels with endogeneity.

To check for the correlation of our instrument $log(found)$ with the unobserved heterogeneity c_i, we check the joint significance of the coefficients on $\bar{\mathbf{z}}_i$. This translate into checking the joint significance of the coefficients on $\{\overline{lunch_i}, \overline{log(enroll)}_i, \overline{log(found)}_i, \overline{y96}, \overline{y97}, \overline{y98}\}$. The $\boldsymbol{\chi}_2$ and the p values of the test are given in Table 1.2. We find that the variables are jointly significant that illustrates a non-zero correlation between the instruments and the time-invariant unobserved individual-specific heterogeneity. This validates FE2SLS to be an appropriate estimation procedure.

1.9 Technical Details

1.9.1 Derivation for $\hat{\beta}_{P2SLS}$

To obtain the expression for $\hat{\boldsymbol{\beta}}_{P2SLS}$ in section 1.3, we use Frisch-Waugh-Lovell theorem for instrument variables. Consider

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \varepsilon_i$$

where X_{2i} is exogenous with respect to ε_i and X_{1i} is endogenous with respect to ε_i. To deal with this endogeneity problem, we have instruments Z_i. Then Firsch-Waugh-Lovell Theorem states that the 2SLS estimator of β_1 can be estimated as:

- First, regress Z_i on X_{2i} and obtain the residuals R_i
- Next, regress Y_i on X_{1i} using R_i as instruments.

We follow the similar procedure in the following steps:

Step 1 : First we run a Pooled OLS as:

$$(\mathbf{z}_{it} - \theta_i\bar{\mathbf{z}}_i) = (1-\theta_i)\bar{\mathbf{z}}_i\boldsymbol{\alpha}_1 + (1-\theta_i)\bar{\mathbf{w}}_i\boldsymbol{\alpha}_2 + error \text{ using } s_{it} = 1$$

Theorem 3 *Consider the regression:*

Table 1.1: Empricial Illustration Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
*math*4	CF	FE	RE	CRE	RE2sls	FE2sls	CRE2sls
$log(avgexp)$	47.00**	5.084	7.256***	5.084	19.47***	47.00*	47.00*
	(23.32)	(3.46)	(1.80)	(3.46)	(2.69)	(25.03)	(25.04)
lunch	-0.005	0.02	-0.37***	0.02	-0.38***	-0.005	-0.005
	(0.05)	(0.0433)	(0.01)	(0.04)	(0.01)	(0.05)	(0.05)
$log(enrol)$	6.483	-3.174	-1.814**	-3.174	-0.523	6.483	6.483
	(5.63)	(2.22)	(0.84)	(2.22)	(0.89)	(6.16)	(6.16)
*y*96	-2.558	1.598***	1.274***	1.598***	0.0966	-2.558	-2.55
	(2.367)	(0.549)	(0.473)	(0.549)	(0.519)	(2.534)	(2.535)
*y*97	-6.63**	-1.43**	-1.51***	-1.43**	-3.03***	-6.63**	-6.63**
	(2.984)	(0.610)	(0.503)	(0.610)	(0.568)	(3.18)	(3.18)
*y*98	6.111*	11.72***	11.74***	11.72***	10.05***	6.11*	6.11*
	(3.23)	(0.66)	(0.53)	(0.66)	(0.61)	(3.44)	(3.44)
$\overline{log(avgexp)}$				3.137			
				(4.082)			
$\overline{lunch}$				-0.43***			-0.45***
				(0.04)			(0.04)
$\overline{log(enrol)}$				1.42			-4.51
				(2.388)			(4.18)
$\overline{y96}$				-4.97			-2.202
				(4.21)			(4.57)
$\overline{y97}$				2.89			2.436
				(4.34)			(4.57)
$\overline{y98}$				-			-
				15.61***			16.24***
				(4.38)			(4.73)
$\hat{v}$	-42.48*						
	(23.52)						
$log(found)$							-20.40
							(17.10)
Constant	-361.8	38.61	26.98	25.04	-80.65***	-361.8	-136.9*
	(222.2)	(35.03)	(17.05)	(19.81)	(24.73)	(238.9)	(73.77)
Obs	5,913	5,913	5,913	5,913	5,913	5,913	5,913
# Schools	1,643	1,643	1,643	1,643	1,643	1,643	1,643

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.2: Specification Tests

Specification Tests	*Chi-Squared*	*p-value*	*Degrees of Freedom*
FE vs RE	109.53	0.0	6
FE2SLS vs RE2SLS	105.67	0.0	6

$$(z_{it} - \theta_i \bar{z}_i) = (1-\theta_i)\bar{z}_i \boldsymbol{\alpha_1} + (1-\theta_i)\bar{w}_i \boldsymbol{\alpha_2} + error.$$

Pooled OLS estimators using $s_{it} = 1$ *will yield* $\widehat{\boldsymbol{\alpha}_1} = \boldsymbol{I}_L$ *(Identity matrix) and* $\widehat{\boldsymbol{\alpha}_2} = \mathbf{0}$ *(zero matrix).*

Proof. First run a pooled regression of $(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)$ on $(1-\theta_i)\bar{\mathbf{z}}_i$ for $s_{it} = 1$. The coefficient will be

$$\begin{aligned}
&= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2 \bar{\mathbf{z}}_i' \bar{\mathbf{z}}_i\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2 \bar{\mathbf{z}}_i' (\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)\right] \\
&= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2 \bar{\mathbf{z}}_i' \bar{\mathbf{z}}_i\right]^{-1} \left[\sum_{i=1}^{N} (1-\theta_i)\bar{\mathbf{z}}_i' \sum_{t=1}^{T} s_{it}(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)\right] \\
&= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2 \bar{\mathbf{z}}_i' \bar{\mathbf{z}}_i\right]^{-1} \left[\sum_{i=1}^{N} (1-\theta_i)\bar{\mathbf{z}}_i' (T_i \bar{\mathbf{z}}_i - T_i \theta_i \bar{\mathbf{z}}_i)\right] \\
&= \left[\sum_{i=1}^{N} (1-\theta_i)^2 \bar{\mathbf{z}}_i' \bar{\mathbf{z}}_i\right]^{-1} \left[\sum_{i=1}^{N} (1-\theta_i)^2 \bar{\mathbf{z}}_i' \bar{\mathbf{z}}_i\right] \\
&= \boldsymbol{I}
\end{aligned}$$

The residuals from this regression would be equal to

$$\begin{aligned}
&(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_{it}) - (\bar{\mathbf{z}}_i - \theta_i \bar{\mathbf{z}}_i) \\
&= (\mathbf{z}_{it} - \bar{\mathbf{z}}_i) \\
&\equiv \ddot{\mathrm{z}}_{it}
\end{aligned}$$

Next, we run a POLS on the regression of

$$(1-\theta_i)\mathbf{w}_i \text{ on } (1-\theta_i)\bar{\mathbf{z}}_i \text{ for } s_{it} = 1$$

As it is clear, both the coefficient and the residuals from this regression would only depend on i.

Coefficient:

$$= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2\bar{\mathbf{z}}_i'\mathbf{w}_i\right]$$
$$= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\mathbf{w}_i\right]$$

Residuals:

$$= (1-\theta_i)\mathbf{w}_i - (1-\theta_i)\bar{\mathbf{z}}_i\left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\mathbf{w}_i\right]$$

$$\equiv \tilde{\boldsymbol{e}}_i(depends\,only\,on\,i)$$

To obtain $\hat{\boldsymbol{\alpha}}_2$, run a POLS of $\ddot{\mathrm{z}}_{it} = (\mathbf{z}_{it} - \bar{\mathbf{z}}_i)$ on $\tilde{\boldsymbol{e}}_i$ for $s_{it} = 1$. (Frisch-Waugh-Lovell theorem). We get

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_2 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{\boldsymbol{e}}_i'\tilde{\boldsymbol{e}}_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{\boldsymbol{e}}_i'\ddot{\mathrm{z}}_{it}\right] \\
&= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{\boldsymbol{e}}_i'\tilde{\boldsymbol{e}}_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{\boldsymbol{e}}_i'(\mathbf{z}_{it} - \bar{\mathbf{z}}_i)\right] \\
&= \left[\sum_{i=1}^{N} T_i\tilde{\boldsymbol{e}}_i'\tilde{\boldsymbol{e}}_i\right]^{-1}\left[\sum_{i=1}^{N}\tilde{\boldsymbol{e}}_i'\Big(\sum_{t=1}^{T} s_{it}\mathbf{z}_{it} - T_i\bar{\mathbf{z}}_i\Big)\right] \\
&= \left[\sum_{i=1}^{N} T_i\tilde{\boldsymbol{e}}_i'\tilde{\boldsymbol{e}}_i\right]^{-1}\left[\sum_{i=1}^{N}\tilde{\boldsymbol{e}}_i'\big(T_i\bar{\mathbf{z}}_i - T_i\bar{\mathbf{z}}_i\big)\right] \\
&= \mathbf{0}
\end{aligned}$$

Finally, since $\hat{\boldsymbol{\alpha}}_2 = \mathbf{0}$, to obtain $\hat{\boldsymbol{\alpha}}_1$ we simply do a POLS of $(\mathbf{z}_{it} - \theta_i\bar{\mathbf{z}}_i)$ on $(1-\theta_i)\bar{\mathbf{z}}_i$ for $s_{it} = 1$. This was shown above to be equal to $\boldsymbol{I}_L$. ■

Using Theorem 3, we obtain the residuals from the regression in Step 1 as $(\mathbf{z}_{it} - \bar{\mathbf{z}}_i) = \ddot{\mathrm{z}}_{it}$

Step 2 : Regress

$$(y_{it} - \theta_i\bar{y}_i) \text{ on } (\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)$$

for $s_{it} = 1$ using $(\mathbf{z}_{it} - \bar{\mathbf{z}}_i) = \ddot{\mathrm{z}}_{it}$ as instruments. The resulting 2SLS coefficient will be

$$\hat{\boldsymbol{\beta}}_{P2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)'\ddot{\mathrm{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}'_{it}\ddot{\mathrm{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}'_{it}(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)\right)\right]^{-1} \times$$
$$\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)'\ddot{\mathrm{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}'_{it}\ddot{\mathrm{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{z}}'_{it}(y_{it} - \theta_i\bar{y}_i)\right)\right]$$

1.9.2 *Proof of Theorem 2*

We have :

$$y_{it1} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{y}_{it2}\boldsymbol{\alpha} + c_{i1} + u_{it} \tag{1.21}$$

and the reduced form equation:

$$\mathbf{y}_{it2} = \mathbf{w}_{it}\boldsymbol{\gamma} + c_{i2} + \mathbf{v}_{it} \tag{1.22}$$

The first step is to estimate the reduced form equation using Fixed Effects. We first time demean equation (1.22):

$$\ddot{\mathrm{y}}_{it2} = \ddot{\mathrm{w}}_{it}\boldsymbol{\gamma} + \ddot{\mathrm{v}}_{it} \tag{1.23}$$

and do a simple POLS using the selected sample, i.e for $s_{it} = 1$. Denote the FE residuals as $\hat{\ddot{\mathrm{v}}}_{it}$. The control function equation would be obtained by augmenting equation (1.21) with $\hat{\ddot{\mathrm{v}}}_{it}$:

$$y_{it1} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{y}_{it2}\boldsymbol{\alpha} + \hat{\ddot{\mathrm{v}}}_{it}\boldsymbol{\rho} + error_{it} \tag{1.24}$$

We will show that the FE estimators of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in equation (1.22) would be identical to the FE2SLS estimators from equation (1.21).

Proof:

To obtain the FE estimators of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in augmented equation (1.24), let $\mathbf{x}_{it1} \equiv (\mathbf{x}_{it}, \mathbf{y}_{it2})$, $\mathbf{x}_{it2} \equiv \hat{\ddot{\mathrm{v}}}_{it}$ and $\boldsymbol{\beta}_1 \equiv \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix}$. Our equation becomes:

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \mathbf{x}_{it2}\boldsymbol{\rho} + \eta_{it} \tag{1.25}$$

The error term η_{it} in equation (1.25) contains both individual heterogeneity term and the idiosyncratic error. FE transformation is given as:

$$\ddot{y}_{it1} = \ddot{x}_{it1}\boldsymbol{\beta}_1 + \ddot{x}_{it2}\boldsymbol{\rho} + \ddot{\boldsymbol{\eta}}_{it} \tag{1.26}$$

FE estimator for $\boldsymbol{\beta}_1$ is obtained by doing a POLS on equation (26) using $s_{it} = 1$. To obtain an expression for $\boldsymbol{\beta}_1$, we will use Frisch-Waugh-Lovell theorem:

- *Step 1*: Do a POLS of $\ddot{x}_{it1}$ on $\ddot{x}_{it2}$ for $s_{it} = 1$. Let the estimator for the coefficient be denoted by $\hat{\boldsymbol{\pi}}$:

$$\hat{\boldsymbol{\pi}} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}'_{it2}\ddot{x}_{it2}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}'_{it2}\ddot{x}_{it1}\right]$$

Consider

$$\begin{aligned}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}'_{it2}\ddot{x}_{it1}\right] &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}x'_{it2}\ddot{x}_{it1}\right] \\ &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}(\ddot{x}_{it}, \ddot{y}_{it2})\right] \text{ since } \mathbf{x}_{it1} \equiv (\mathbf{x}_{it}, \mathbf{y}_{it2}),\ \mathbf{x}_{it2} \equiv \hat{\ddot{v}}_{it} \\ &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{x}_{it}, \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{y}_{it2}\right]\end{aligned}$$

Note that since $\hat{\ddot{v}}_{it}$ are FE residuals from the reduced form equation (1.23), by construction we will have $\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{w}'_{it}\hat{\ddot{v}}_{it} \equiv \mathbf{0}$,

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{w}_{it} = \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}(\ddot{x}_{it}, \ddot{z}_{it}) &= \mathbf{0} \\ \implies \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{x}_{it}, \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{z}_{it}\right] &= (\mathbf{0}, \mathbf{0})\end{aligned}$$

This implies that $\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{x}_{it} = \mathbf{0}$. So we get

$$\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{x}_{it}, \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{y}_{it2}\right] = \left[\mathbf{0}, \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{v}}'_{it}\ddot{y}_{it2}\right]$$

$\Longrightarrow$

$$\begin{aligned}
\hat{\boldsymbol{\pi}} &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}'_{it2}\ddot{\mathrm{x}}_{it2}\right]^{-1}\left[\mathbf{0}, \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{\mathrm{v}}}'_{it}\ddot{\mathrm{y}}_{it2}\right] \\
&= \left(\mathbf{0}, \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}'_{it2}\ddot{\mathrm{x}}_{it2}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{\mathrm{v}}}'_{it}\ddot{\mathrm{y}}_{it2}\right]\right) \\
&= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}'_{it2}\ddot{\mathrm{x}}_{it2}\right]^{-1}\left[\mathbf{0}, \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{\ddot{\mathrm{v}}}'_{it}\ddot{\mathrm{y}}_{it2}\right] \\
&= \left(\mathbf{0}, \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}'_{it2}\ddot{\mathrm{x}}_{it2}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}'_{it2}\ddot{\mathrm{y}}_{it2}\right]\right) \text{ since } \mathbf{x}_{it2} \equiv \hat{\ddot{\mathrm{v}}}_{it}
\end{aligned}$$

Now, note that $[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathrm{x}}'_{it2}\ddot{\mathrm{x}}_{it2}]^{-1}[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}'_{it2}\ddot{\mathrm{y}}_{it2}]$ is nothing but the FE estimator when $\mathbf{y}_{it2}$ is regressed on $\mathbf{x}_{it2} \equiv \hat{\ddot{\mathrm{v}}}_{it}$ for $s_{it} = 1$.

This by construction would be equal to identity matrix: $\mathbf{I}$. This implies

$$\hat{\boldsymbol{\pi}} = [\mathbf{0}, \mathbf{I}]$$

The residuals from regressing $\ddot{\mathrm{x}}_{it1}$ on $\ddot{\mathrm{x}}_{it2}$ would be:

$$\begin{aligned}
\hat{\mathrm{r}}_{it} &= \ddot{\mathrm{x}}_{it1} - \ddot{\mathrm{x}}_{it2}[\mathbf{0}, \mathbf{I}] \\
&= [\ddot{\mathrm{x}}_{it}, \ddot{\mathrm{y}}_{it2}] - [\mathbf{0}, \ddot{\mathrm{x}}_{it2}] \\
&= [\ddot{\mathrm{x}}_{it}, (\ddot{\mathrm{y}}_{it2} - \ddot{\mathrm{x}}_{it2})]
\end{aligned}$$

Now, note that $(\ddot{\mathrm{y}}_{it2} - \ddot{\mathrm{x}}_{it2})$ is nothing but the predicted values of $\ddot{\mathrm{y}}_{it2}$ from equation (23):

$$\ddot{\mathrm{y}}_{it2} = \hat{\ddot{\mathrm{y}}}_{it2} + \ddot{\mathrm{x}}_{it2}$$

- *Step 2* Do a POLS of $\ddot{y}_{it1}$ on $\hat{\mathrm{r}}_{it}$ for $s_{it} = 1$. This would be a regression of $\ddot{\mathrm{y}}_{it1}$ on $\ddot{\mathrm{x}}_{it2}$ and the predicted values from the reduced form equation: $\hat{\ddot{\mathrm{y}}}_{it2}$. This precisely the FE2SLS method.

1.10 Concluding Remarks

Literature on unbalanced panels can be broadly classified into two broad sections. The first section focuses on the problems related to the detection of selection bias and estimation methods to correct for this bias, if it is not ruled out. The second section looks at the cases when selection bias does not exist. When data is missing at random, then estimation methods of the balanced panel case can be extended. A similar adaptation can be done for specification tests. However, just like in the balanced case, the limitations of these tests would also hold in the unbalanced case. While methods to counter these have been developed for the balanced panels, for unbalanced panels, limited work has been done to formally account for these limitations. This paper hopes to contribute to the existing literature by attempting to address this issue.

The preliminary linear model considered in this paper has many possible extensions for future research. We are interested in developing correlated random effects models for linear unbalanced panel data models with individual specific slopes. This would give us a way to obtain a specification test for testing for heterogeneous slopes. Moreover, in this paper we have assumed that conditional on the observables, selection mechanism is exogenous to the idiosyncratic shocks. This assumption seldom holds in most unbalanced panels. Thus analysis for specifications for unbalanced panels where the exogenous sampling does not hold is another area in which we would like to further extend our analysis.

CHAPTER 2

CONTROL FUNCTION SIEVE ESTIMATION OF ENDOGENOUS SWITCHING MODELS WITH ENDOGENEITY

2.1 Introduction

Evaluating the causal effects of a program or a policy intervention is one of the most important questions in econometric analysis. In the case of discrete treatments, endogenous switching models provide a powerful framework to capture the causal effects. Switching regression models have been extensively used to estimate structural shifts: to capture the parameter variation where each possible state of parameter vector is called a *regime*. Endogenous switching models have been used as the primary econometric technique in labour economics to study wage differentials between public and private sectors(Adamchik and Bedi(1983)), union and non-union members (Lee(1978)). It has also been used in modeling housing demand and modeling of markets in disequilibrium (Thorst(1977)).

Endogenous switching models have been traditionally estimated using joint maximum likelihood estimation. This requires a full specification of the joint distribution of the unobservables. This approach not only places restrictive assumptions on the model but is also computationally challenging. Murtazashvili and Wooldridge(2016) use control function methods to obtain computationally simple estimation of switching models under both coefficient homogeneity and heterogeneity. They however still maintain distributional assumptions and homoskedastic errors.

In this paper, we generalize the endogenous switching models by allowing a more flexible reduced form for the treatment variable. This allows us to incorporate more heterogeneity in our model. To allow for heterogeneity in the treatment variable model, we allow the errors in the reduced form to have conditional heteroskedasticity. It allows the unobservables to contribute to the dependent variable in a heterogeneous manner. One can argue that we can also incorporate

heterogeneity in an econometric model by allowing for individual-specific slopes in the reduced form. While this specification of the treatment variable model will allow the model to be more structural, this also imposes a specific structure to the heteroskedasticity. We would prefer if our reduced form has a general form of heteroskedasticity. An attractive feature of allowing heteroskedasticity to be of a general form is that this allows the unconditional distribution of the error term to be of an unknown form, that further relaxes the distributional assumptions on the errors.

The endogeneity of the treatment variable or the switching indicator is modeled through control function methods. In particular, we model the endogeneity in terms of the relationship between the errors of the primary equation for the outcome variable (*'structural' errors*) and the errors of the reduced form equation for the treatment variable (*'reduced-form' errors*). Since we are allowing the *'reduced-form' errors* to have heteroskedasticity that is conditional on the explanatory variables, the relationship between the two error terms is also modeled in terms of the explanatory variables. This is done by using ***Conditional Linear Projections*** (CLPs). CLPs are just conditional counterparts to the usual linear projections that are popular in modeling the conditional expectations, specially of the auxiliary terms of an econometric model. The key assumption is that the relationship between the *'structural' errors* and the *'reduced-form' errors* that is reflected in terms of the conditional expectation of the *'structural' errors* can be represented in terms of a linear projection *conditional* on explanatory variables. *Conditional linear projections restrict linearity only in terms of the errors while allowing for the dependence on the explanatory variables to be of an unknown functional form*. Moreover, when we consider the model which allows for individual-specific slopes in the outcome equation, we have an added component contributing the endogeneity: endogeneity arising due to the correlation between the treatment and idiosyncratic gain to the treatment. This correlation is again in terms of the explanatory variables because of conditional heteroskedasticity. However, as we will see, conditional linear projection allow us to model this endogeneity as well.

The estimation is done in two steps. The correction terms for the endogeneity are obtained in the first step estimation of the reduced form model for the treatment variable. An estimating equation that accounts for the endogeneity through these correction terms is estimated in the second step. Since the heteroskedasticity of the *reduced-form* error as well as the correction terms in the estimating equation are of an unknown functional form, we estimate both the step using the semi-nonparametric methods. In particular, we use *method of sieves* in both the steps to obtain the estimation procedure. However, we also illustrate how one can impose the functional assumptions in the model making it fully parametric and thus estimate both the step through standard parametric methods. Thus we obtain both parametric and semi-nonparametric estimation procedures.

An attractive feature of endogenous switching model is that it serves as an *umbrella model* for other econometric models. Specifically, the estimation strategy can be modified to estimate the parameters in sample selection models and models with a binary endogenous variable. The estimation procedures in this paper are obtained for the model with homogeneous treatment effects and heteroskedastic reduced form model for the treatment variable and is extended to three models. The first model incorporates heterogeneous treatment effects through individual-specific slopes in the outcome equation. We further extend the estimation strategy to the models with binary endogenous variable and sample selection models.

We also give a detailed large sample properties of the estimators. Large sample properties of the parametric estimators are obtained by the straightforward GMM types treatment of the two-step estimation. To obtain the large sample estimation theory of the sieve estimators, we refer to Hahn, Liao and Ridder(2018). Hahn, Liao and Ridder(2018) provide a general unified mathematical framework investigating the asymptotic properties of such sieve two-step estimators. Since our sieve estimation procedure neatly fits into this framework, our estimator follows these results. They also show the numerical equivalence result for the parametric variances and sieve variances. We use this equivalence result to obtain the expressions for variances of the

two-step sieve estimators.

The paper is organized as follows. In section 2.2, we begin with the traditional model of endogenous switching models with constant coefficients. We describe the assumptions of conditional heteroskedasticity in the reduced form and obtain the estimating equation using the control function approach that uses conditional linear predictors. In section 2.3, we obtain both parametric and sieve two-step estimation procedures. Section 2.4 describes the asymptotics of our estimators and obtain the results of consistency, asymptotic normality and expressions for the asymptotic variances using the numerical equivalence result. Section 2.5 extends our analysis to other econometric models. In particular, we consider endogenous switching models with individual-specific slopes in the outcome equation, models with a binary endogenous explanatory variable and sample selection models. Section 2.6 illustrate our estimation procedures to an empirical application. Section 2.7 concludes the paper.

2.2 Constant Coefficients Endogenous Switching Regression

2.2.1 Model

Suppose that we are interested in evaluating the causal effect of a program on N individuals indexed by $i = 1, 2, ..., N$. The treatment status of an individual i is denoted by a binary variable y_{2i} that takes the value 1 if she is treated and 0 otherwise. As we will see later, y_{2i} could denote a binary endogenous explanatory variable or the *selection indicator* in sample selection models. In the framework of switching models, y_{2i} becomes the endogenous switching indicator for two *regimes*:

$$y_{2i} = \begin{cases} 0 & \text{for } \textit{Regime 0} \\ 1 & \text{for } \textit{Regime 1} \end{cases} \tag{2.1}$$

Consistent with the treatment literature, we postulate the existence of two potential outcomes, one in each *regime* denoted by $\{y_{1i}^{(0)}, y_{1i}^{(1)}\}$ where the superscript denotes the regime.

Our first assumption states that these counter-factual outcomes are linear in the parameters:

Assumption 2.2.1 *For each* $i = 1, 2, ..., N$,

$$y_{1i}^{(0)} = \mathrm{x}_{1i}\boldsymbol{\gamma}^{(0)} + u_i^{(0)}$$
$$y_{1i}^{(1)} = \mathrm{x}_{1i}\boldsymbol{\gamma}^{(1)} + u_i^{(1)}$$

where x_{1i} is a $1 \times K_{x_1}$ vector of exogenous explanatory variables that includes unity. x_{1i} also allows all functional forms (like log, squares) that are linear in parameters. $\{u_i^{(0)}, u_i^{(1)}\}$ are the unobservables. We allow for different coefficients in each *regime*, however these coefficients are not individual-specific. Note that our observed outcome variable give by y_{1i} can be expressed as:

$$y_{1i} = (1 - y_{2i})y_{1i}^{(0)} + y_{2i}y_{1i}^{(1)} \tag{2.2}$$

Substituting for the counter-factuals, we get our primary equation for the switching regression model with constant coefficients $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$ as:

$$y_{1i} = (1 - y_{2i})\mathrm{x}_{1i}\boldsymbol{\gamma}^{(0)} + y_{2i}\mathrm{x}_{1i}\boldsymbol{\gamma}^{(1)} + (1 - y_{2i})u_i^{(0)} + y_{2i}u_i^{(1)} \tag{2.3}$$

Changing the notations slightly, we re-write (2.3) as:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + v_{0i} + y_{2i}v_{1i} \tag{2.4}$$

where: $\boldsymbol{\beta}_0 \equiv \boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\beta}_1 \equiv \boldsymbol{\gamma}^{(1)} - \boldsymbol{\gamma}^{(0)}$, $v_{0i} \equiv u_i^{(0)}$, $v_{1i} \equiv u_i^{(1)} - u_i^{(0)}$. The average treatment effect is captured by the parameter $\boldsymbol{\beta}_1$.

The switching indicator y_{2i} is a binary variable modeled as:

$$y_{2i} = \mathbf{1}[\mathrm{x}_i\boldsymbol{\beta}_2 - v_{2i} \geq 0] \tag{2.5}$$

where x_i is a $1 \times K_x$ vector of explanatory variables with $K_x > K_{x_1}$. We allow $\mathrm{x}_{1i} \subset \mathrm{x}_i$ [1]

[1] As we will see, this would serve as an exclusion restriction to ensure proper estimation.

Heteroskedastic Errors: The key feature of our model is that we allow for the heterogeneity in the treatment assignment by allowing for a heteroskedastic error variance in the treatment model.

Assumption 2.2.2 We incorporate multiplicative heteroskedasticity by assuming*:*

$$v_{2i} = \sigma_2(\mathrm{x}_i) \cdot u_{2i} \ , \ u_{2i} \perp\!\!\!\perp \mathrm{x}_i \ , \ u_{2i} \sim \mathcal{N}(0,1) \tag{2.6}$$

where $\mathcal{N}$ denotes Normal Distribution and $\sigma_2(\mathrm{x}_i)$ is a function of all the explanatory variables. As we will see, the form of heteroskedasticity is specified in the parametric estimation procedure and is allowed to be of an unknown function in the semi-nonparametric estimation procedure.

Before we describe the method to obtain the estimating equations, consider the interpretation of the error structure in the model of the treatment variable. Substitution of equation (2.6) in the reduced form equation for the treatment variable implies:

$$y_{2i} = \mathbf{1}[\mathrm{x}_i \boldsymbol{\beta}_2 - \sigma_2(\mathrm{x}_i) u_{2i} \geq 0] \tag{2.7}$$

If we interpret u_{2i} to be an unobserved variable that effects the probability of an individual i to be in the treated group, then equation (2.7) suggests that the contribution of this unobserved variable depends on the individual's other covariates that effects his/her selection into the program in a flexible way(as reflected in the unspecified form of the functional form of the heteroskedasticity). For instance, consider the standard example of studying the effect of a job training program on wages and assume that u_{2i} is a measure of unobserved ability. A heteroskedastic error structure in the treatment equation implies that the contribution of the unobserved ability on the probability of being selected or participating in the job training program depends on the individual's socio-economic factors in a flexible way that is captured by $\sigma_2(\mathrm{x}_i)$.

2.2.2 Estimating Equation

First note that y_{2i} is a function of (x_i, v_{2i}). Thus, to obtain the estimating equation, we first write:

$$\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathbb{E}[v_{0i}|\mathrm{x}_i, v_{2i}] + y_{2i}\mathbb{E}[v_{1i}|\mathrm{x}_i, v_{2i}] \tag{2.8}$$

$\mathbb{E}[v_{0i}|\mathrm{x}_i, v_{2i}]$ and $\mathbb{E}[v_{1i}|\mathrm{x}_i, v_{2i}]$ are the *correction terms* that correct for the bias due to the endogenous switching.

It is clear from (2.8) that we need to obtain expressions for the correction terms. We will obtain these expressions in terms of *Conditional Linear Projections* (or predictions).

Conditional Linear Projections(CLPs)

Linear projections are a popular tool in econometric models to approximate the conditional expectations. In our model, we want to approximate the relationships between the unobservables in the outcome equation and treatment equation conditional on x_i as is reflected in the *correction terms*. In addition, since we would allow the conditional heteroskedasticity to be of an unknown form in the semi-nonparametric estimation, we would like to impose linearity only with respect to v_{2i} leaving the functional form with respect to x_i unspecified. Thus, for our purposes, we will the use *conditional* linear projections. Denoted by $\mathbb{L}[v_{ji}|v_{2i}; \mathrm{x}_i]$, ***Conditional Linear Projection*** is defined as the linear projection of v_{ji} on v_{2i} *conditional* on x_i, for $j = \{0, 1\}$. The theory of CLPs was first developed in Hansen and Richard(1987) in which they essentially extended the conventional Hilbert Space analysis to the conditional framework. Wooldridge(1999) uses the concept of CLP to develop the orthogonality conditions to obtain a distribution free estimation of nonlinear panel data models. In the context of this paper, CLP form the key assumption:

Assumption 2.2.3 The conditional expectations are assumed to be linear in v_{2i} conditional on

x_i:

$$\mathbb{E}[v_{0i}|v_{2i}, x_i] = \mathbb{L}[v_{0i}|v_{2i}; x_i] \equiv \left(\frac{\sigma_{02}(x_i)}{\sigma_2^2(x_i)}\right) v_{2i} \tag{2.9}$$

$$\mathbb{E}[v_{1i}|v_{2i}, x_i] = \mathbb{L}[v_{1i}|v_{2i}; x_i] \equiv \left(\frac{\sigma_{12}(x_i)}{\sigma_2^2(x_i)}\right) v_{2i} \tag{2.10}$$

where: $\sigma_{j2}(x_i) \equiv Cov(v_{ji}, v_{2i}|x_i)$, $j = \{0,1\}$ *denote the conditional variances between the error terms in the outcome equation and the error term in the reduced form equation for the treatment equation.*

Assumption (2.2.3) imposes linearity of $\mathbb{E}[v_{ji}|v_{2i}, x_i], j = \{0,1\}$ but only with respect to v_{2i}. Since we are using linear projections conditional on x_i, we still allow $\mathbb{E}[v_{ji}|v_{2i}, x_i]$ to be a flexible non-linear function of x_i. To give some context to our assumption, first denote the linear projections in an unconditional case as $\mathbb{L}_u[v_{ji}|v_{2i}, x_i]$ defined as the linear projection of v_{ji} on v_{2i} and x_i, for $j = \{0,1\}$. We can use the unconditional linear projections to model the relationships between the unobservables in the case where we impose the stronger assumption of (v_{0i}, v_{1i}, v_{2i}) being jointly independent of x_i. In this case, Assumption 2.2.3 translate into $\mathbb{E}[v_{ji}|v_{2i}, x_i] = \mathbb{E}[v_{ji}|v_{2i}, x_i] = \mathbb{L}_u[v_{ji}|v_{2i}]$. Furthermore, if we relax the independence assumption but impose linearity on x_i, Assumption 2.3 becomes $\mathbb{E}[v_{ji}|v_{2i}, x_i] = \mathbb{L}_u[v_{ji}|v_{2i}, x_i]$.

Substituting Assumption (2.2.3) in equation (2.8), we get

$$\mathbb{E}[y_{1i}|x_i, v_{2i}] = x_{1i}\boldsymbol{\beta}_0 + y_{2i}x_{1i}\boldsymbol{\beta}_1 + \left(\frac{\sigma_{02}(x_i)}{\sigma_2^2(x_i)}\right) v_{2i} + y_{2i}\left(\frac{\sigma_{12}(x_i)}{\sigma_2^2(x_i)}\right) v_{2i} \tag{2.11}$$

Next, note that $\mathbb{E}[u_{2i}|x_i, y_{2i}]$ are the ***Generalized Errors*** from the reduced form model for the treatment variable. In our context,

$$\begin{aligned}\mathbb{E}[u_{2i}|x_i, y_{2i}] &= \left[y_{2i}\lambda\left(\frac{x_i\boldsymbol{\beta}_2}{\sigma_2(x_i)}\right) - (1-y_{2i})\lambda\left(-\frac{x_i\boldsymbol{\beta}_2}{\sigma_2(x_i)}\right)\right] \\ &\equiv h(y_{2i}, x_i)\end{aligned}$$

$\lambda(.)$ is the Inverse Mills Ratio and $h(y_{2i}, \mathrm{x}_i)$ denote the ***generalized residuals*** from the first stage estimation.

$$\begin{aligned} \implies \mathbb{E}[v_{2i}|\mathrm{x}_i, y_{2i}] &= \sigma_2(\mathrm{x}_i)\mathbb{E}[u_{2i}|\mathrm{x}_i, y_{2i}] \\ &= \sigma_2(\mathrm{x}_i)h(y_{2i}, \mathrm{x}_i) \end{aligned}$$

We also have: $\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i}] = \mathbb{E}[\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}]|\mathrm{x}_i, y_{2i}]$. So we get:

$$\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \left(\frac{\sigma_{02}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right)\mathbb{E}[v_{2i}|\mathrm{x}_i, y_{2i}] + y_{2i}\left(\frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right)\mathbb{E}[v_{2i}|\mathrm{x}_i, y_{2i}]$$

$\implies$

$$\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \underbrace{\left(\frac{\sigma_{02}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)}_{\equiv \mathfrak{g}_0(\mathrm{x}_i)} h(y_{2i}, \mathrm{x}_i) + y_{2i}\underbrace{\left(\frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)}_{\equiv \mathfrak{g}_1(\mathrm{x}_i)} h(y_{2i}, \mathrm{x}_i)$$

We get our estimating equation as:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathfrak{g}_0(\mathrm{x}_i)h(y_{2i}, \mathrm{x}_i) + y_{2i}\mathfrak{g}_1(\mathrm{x}_i)h(y_{2i}, \mathrm{x}_i) + \eta_i \tag{2.12}$$

where $\mathbb{E}[\eta_i|\mathrm{x}_i, y_{2i}] = 0$ by construction.

2.3 Estimation Strategy

For estimation:

$$\begin{aligned} y_{2i} &= \mathbf{1}[\mathrm{x}_i\boldsymbol{\beta}_2 - \sigma_2(\mathrm{x}_i)u_{2i} \geq 0] \\ y_{1i} &= \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathfrak{g}_0(\mathrm{x}_i)h(y_{2i}, \mathrm{x}_i) + y_{2i}\mathfrak{g}_1(\mathrm{x}_i)h(y_{2i}, \mathrm{x}_i) + \eta_i \end{aligned}$$

These two equations suggest a two-step estimation procedure. In the first step we estimate the reduced form model of the treatment variable or the switching indicator. In the second step, we will plug-in the estimates from the first stage and will estimate equation (2.12). Since we

have terms of unknown functional form given by $\{\sigma_2(\mathrm{x}_i), \mathfrak{g}_0(\mathrm{x}_i), \mathfrak{g}_1(\mathrm{x}_i)\}$, we would estimate both the step through semi-nonparametric procedures, specifically through *method of sieves*. However, one can easily impose both the distributional and functional form assumptions and estimate both the steps through standard parametric procedures. In this section we describe both parametric and sieve estimation of our model.

2.3.1 Parametric Estimation

To estimate the model parametrically, we will first specify the functional form of the heteroskedasticity in the errors in the treatment model. More specifically, we assume:

Assumption 2.3.1 *For all* $i = 1, ..., N$

$$\sigma_2(\mathrm{x}_i) \equiv exp(\mathrm{x}_i \boldsymbol{\Pi}_2) \tag{2.13}$$

This implies $Var[v_{2i}|\mathrm{x}_i] = [exp(\mathrm{x}_i \boldsymbol{\Pi}_2)]^2$. This will allow us to estimate the reduced form model for the treatment using heteroskedastic probit.

First Step: The true parameters to be estimated in the first stage can be denoted by $\boldsymbol{\theta}_{02} \equiv \{\boldsymbol{\beta}_{02}, \boldsymbol{\Pi}_{02}\}$ that belong to a finite dimensional parameter space $\boldsymbol{\Theta}_2$. Denoting, $\Phi(.)$ as Normal cdf, our first stage parametric estimators $\hat{\boldsymbol{\theta}}_2 \equiv (\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\Pi}}_2)$ solve:

$$\hat{\boldsymbol{\theta}}_2 = arg \max_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2} \sum_{i=1}^{N} \left[y_{2i} \log[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{exp(\mathrm{x}_i \boldsymbol{\Pi}_2)}\right)] + (1 - y_{2i}) \log[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{exp(\mathrm{x}_i \boldsymbol{\Pi}_2)}\right)] \right] \tag{2.14}$$

In practice, we would run the standard *het probit* command in Stata. This would give us estimators for $\boldsymbol{\beta}_2$ and $\boldsymbol{\Pi}_2$ which would subsequently give us an estimator of $\lambda(.)$ and $h(\mathrm{x}_i, y_{2i})$. Denote these estimators as $\hat{\lambda}_i \equiv \lambda\left(\frac{\mathrm{x}_{2i}\hat{\boldsymbol{\beta}}_2}{\hat{\sigma}_2(\mathrm{x}_i)}\right)$, $\hat{h}_i \equiv h(\mathrm{x}_i, y_{2i})$

Second Step: Plugging in the estimator from the first stage estimation in equation (2.12):

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathfrak{g}_0(\mathrm{x}_i)\hat{h}(y_{2i}, \mathrm{x}_i) + y_{2i}\mathfrak{g}_1(\mathrm{x}_i)\hat{h}(y_{2i}, \mathrm{x}_i) + \eta_{1i} \tag{2.15}$$

In the second stage, we further specify the functional forms for the terms $\mathfrak{g}_0(\mathrm{x}_i)$ and $\mathfrak{g}_1(\mathrm{x}_i)$. The simplest assumption is that these terms are linear:

Assumption 2.3.2 *For all* $i = 1, ..., N$

$$\mathfrak{g}_0(\mathrm{x}_i) \equiv \mathrm{x}_i \boldsymbol{\Omega}_{02} \tag{2.16}$$

$$\mathfrak{g}_1(\mathrm{x}_i) \equiv \mathrm{x}_i \boldsymbol{\Omega}_{12} \tag{2.17}$$

where $\{\boldsymbol{\Omega}_{02}, \boldsymbol{\Omega}_{12}\}$ *are parameters.*

Since we can include interactions, squares, logarithms and other functional forms of the explanatory variables in x_i, the above assumptions impose linearity only in terms of the parameters. We denote the full list of true parameters to be estimated in the second stage as $\boldsymbol{\theta}_{01} \equiv \{\boldsymbol{\beta}_{00}, \boldsymbol{\beta}_{01}, \boldsymbol{\Omega}_{0,01}, \boldsymbol{\Omega}_{0,12}\}$ that belong to the finite dimensional parameter space $\boldsymbol{\Theta}_1$. In the second step, we can estimate the parameters by running a simple least squares. More specifically, the estimator $\hat{\boldsymbol{\theta}}_1$ solves:

$$\hat{\boldsymbol{\theta}}_1 = arg \min_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \sum_{i=1}^{N} \left[y_{1i} - \mathrm{x}_{1i} \boldsymbol{\beta}_0 - y_{2i} \mathrm{x}_{1i} \boldsymbol{\beta}_1 - \hat{h}_i \mathrm{x}_i \boldsymbol{\Omega}_{02} - y_{2i} \hat{h}_i \mathrm{x}_i \boldsymbol{\Omega}_{12} \right]^2 \tag{2.18}$$

2.3.2 Sieve Estimation

The primary feature of our model is that we do not have any assumptions on the functional forms for the heteroskedasticity of the error terms. We estimate the unknown functions using the sieve estimation procedures. We need to impose some regularity conditions on the unknown functions to be able to estimate them. More specifically, we need to specify the smoothness of the unknown functions to be estimated. Define $\mathfrak{h} \equiv [\sigma_2(\mathrm{x}_i), \mathfrak{g}_0(\mathrm{x}_i), \mathfrak{g}_1(\mathrm{x}_i)]$ to be the collection of all the unknown functions in our model. We assume that $\mathfrak{h}$ belongs to a **Hölder class**. We give the technical definition of the Hölder Class in the Section 2.7.

Hölder Class of functions are the most popular in the non-parametric estimation procedures in econometrics because they can be easily estimated by *linear sieves*. A sieve is called a *finite-dimensional linear sieve* if it is a linear span of finitely many basis functions. Power series, Fourier series, splines, B-splines and wavelets are some of the most popular linear sieves used in sieve estimation.

The estimation in the semi-nonparametric setting also takes place in two steps. Since in this case, we do not impose any functional form assumptions on $\{\sigma_2(\mathrm{x}_i), \mathfrak{g}_0(\mathrm{x}_i), \mathfrak{g}_1(\mathrm{x}_i)\}$, our parameters to be estimated no longer lie in the finite dimensional parameter space $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$. Now that we are optimizing on an infinite dimensional space, we will use sieve estimation methods to obtain the estimators in both the stages. The preliminary step in both the first and second stage estimation will be to define the basis functions and the sieve spaces.

First Step: In the first stage, our true parameters to be estimated are $\{\boldsymbol{\beta}_{20}, \sigma_{20}(\mathrm{x}_i)\}$ that lie in the infinite dimensional parameter space $\mathscr{A}_2$. For sieve estimation, we define the finite dimensional sieve space as:

$$\mathscr{A}_{2N} = \mathscr{B}_2 \mathscr{X} \left\{ \sigma_2(.) = exp(\mathtt{S}_{K_{\sigma,N}}(\mathrm{x}_i)\boldsymbol{\Pi}_N) : \boldsymbol{\Pi}_N \in \mathbb{R}^{K_{\sigma,N}} \right\} \tag{2.19}$$

where $\mathscr{X}$ denotes the tensor product. $\mathtt{S}_{K_{\sigma,N}}(\mathrm{x}_i) \equiv \{\mathtt{s}_1(.), ..., \mathtt{s}_{K_{\sigma,N}}(.)\}$ is $1 \times K_{\sigma,N}$ vector of basis functions. Thus we get: $\boldsymbol{\alpha}_{2N} \equiv (\boldsymbol{\beta}_2, \sigma_{2N}) \in \mathscr{A}_{2N}$ which is our finite dimensional sieve space. Next, denoting, $\Phi(.)$ as Normal cdf, our estimators are:

$$\hat{\boldsymbol{\alpha}}_{2N} = arg \max_{\boldsymbol{\alpha}_{2N} \in \mathscr{A}_{2N}} \sum_{i=1}^{N} y_{2i} \log[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{exp(\mathtt{S}_{K_{\sigma,N}}(\mathrm{x}_i)\boldsymbol{\Pi}_N)}\right)] + (1 - y_{2i}) \log[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{exp(\mathtt{S}_{K_{\sigma,N}}(\mathrm{x}_i)\boldsymbol{\Pi}_N)}\right)] \tag{2.20}$$

This would give us estimators for $\boldsymbol{\beta}_{20}$ and $\sigma^2_{20}(\mathrm{x}_i)$ which would subsequently give us an estimator of $\lambda(.)$ and $h(\mathrm{x}_i, y_{2i})$. Denote the first stage estimators as $\hat{\boldsymbol{\beta}}_2$, $\hat{\sigma}^2_2(\mathrm{x}_i)$, $\hat{\lambda}_i \equiv \lambda\left(\frac{\mathrm{x}_{2i}\hat{\boldsymbol{\beta}}_2}{\hat{\sigma}_2(\mathrm{x}_i)}\right)$, $\hat{h}_i \equiv \hat{h}(\mathrm{x}_i, y_{2i})$.

Second Step: For the second stage we have the sieve space for $\{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \mathfrak{g}_0(\mathrm{x}_i), \mathfrak{g}_1(\mathrm{x}_i)\}$

$$\mathscr{A}_{1N} = \mathscr{B}_1 \mathscr{X} \left\{ \mathfrak{g}_0(.) = \mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02} : \boldsymbol{\Omega}_{N,02} \in \mathbb{R}^{K_{\mathrm{g}_0,N}} \right\}$$
$$\mathscr{X} \left\{ \mathfrak{g}_1(.) = \mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12} : \boldsymbol{\Omega}_{N,12} \in \mathbb{R}^{K_{\mathrm{g}_1,N}} \right\} = \mathscr{B}_1 \mathscr{X} \mathscr{G}_{0N} \mathscr{X} \mathscr{G}_{1N} \tag{2.21}$$

where $\mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i) \equiv \{\mathrm{g}_{0,1}(.), ..., \mathrm{g}_{0,K_{\mathrm{g}_0,N}}(.)\}$ is $1 \times K_{\mathrm{g}_0,N}$ vector of basis functions and $\mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i) \equiv \{\mathrm{g}_{1,1}(.), ..., \mathrm{g}_{1,K_{\mathrm{g}_1,N}}(.)\}$ is $1 \times K_{\mathrm{g}_1,N}$ vector of basis functions. Plugging in

equation (12):

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02}\hat{h}_i + y_{2i}\mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12}\hat{h}_i + \eta_{1i} \tag{2.22}$$

As suggested by equation (21), in the second stage the estimator $\hat{\boldsymbol{\alpha}}_{1N} \equiv \{\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \hat{\mathfrak{g}}_0(\mathrm{x}_i), \hat{\mathfrak{g}}_1(\mathrm{x}_i)\} \in \mathscr{A}_{1N}$ where $\{\hat{\mathfrak{g}}_0(\mathrm{x}_i), \hat{\mathfrak{g}}_1(\mathrm{x}_i)\} \equiv \{\mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i)\widehat{\boldsymbol{\Omega}}_{N,02}, \mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)\widehat{\boldsymbol{\Omega}}_{N,12}\}$ solve the following optimization problem:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{1N} = arg \min_{\boldsymbol{\alpha}_{1N}\in\mathscr{A}_{1N}} \sum_{i=1}^{N} [y_{1i} - \mathrm{x}_{1i}\boldsymbol{\beta}_0 - y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 \\ - \mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02}\hat{h}_i - y_{2i}\mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12}\hat{h}_i]^2\end{aligned} \tag{2.23}$$

In practice, this amounts to running regression y_{1i} on x_{1i}, $y_{2i}\mathrm{x}_{1i}$, $\hat{h}_i\mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i)$, $\hat{h}_i y_{2i}\mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)$.

2.4 Asymptotics

This section describes the asymptotic properties of our two-step estimators, both in parametric and sieve estimation. An important consideration in a two-step estimation procedure is whether and how the estimation error of the first step estimators affects the asymptotic variance of the second step estimators. In the parametric case, the methods of adjusting the asymptotic variance of the second step estimator follows the standard procedures given in Wooldridge(2002). In the sieve case, we apply the general results of Hahn,Liao and Ridder(2018) wherein the statistical properties of sieve two step estimators are derived. We further apply the numerical equivalence results of Hahn,Liao and Ridder(2018) due to which we can treat the two-step sieve estimation as if it were a standard two-step parametric estimation and thus conduct a practical inference on the parameters.

2.4.1 Asymptotic Properties of the Parametric Estimators

In the parametric setting, the consistency of the parameters follows under finite moment conditions. Valid inference of the parameters in the second step should take into account the inclusion of the generalized residuals obtained in the first step. More specifically, we can obtain the asymptotic distribution and the variance analytically by formulating the two-step estimation method in an one-step method of moment framework. The asymptotic distribution of our parametric estimators is given as:

$$\sqrt{N}\begin{pmatrix}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{01})\\(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_{02})\end{pmatrix}\rightarrow_d \mathscr{N}(\mathbf{0},\mathbb{V})\,,\ \mathbb{V}\equiv\mathbb{A}^{-1}\mathbb{B}\mathbb{A}'^{-1} \tag{2.24}$$

We derive the analytical expression for the asymptotic variance of the parametric estimator in the Appendix. Alternatively, we can also run a bootstrap routine.

2.4.2 Asymptotic properties of the Sieve Estimators

In this section, we would specify the asymptotics of our two-step sieve M estimators and the assumptions needed for these properties to hold. There is a rich literature on the the asymptotic properties of 2 step semi-nonparametric estimators. Two-step sieve M estimators have been studied in great detail in Hahn, Liao and Ridder (2018). They derive the asymptotic properties such as consistency and the asymptotic distribution of the estimators. Our estimation problem is very well behaved and our model falls neatly into their framework. Thus all the calculations and properties are easily adapted here and we map their results by verifying the conditions.

Assumption 2.4.1 We assume that for *i = 1,2,...,N:* $\{y_{1i}, y_{2i}, \mathrm{x}_i\}$ *is i.i.d*

Assumption 2.4.2 *In the first step, we assume that* $\boldsymbol{\alpha}_{20}\equiv\{\boldsymbol{\beta}_{20},\sigma^2_{20}(\mathrm{x}_i)\}\in\mathscr{A}_2$ *is the identified as an unique solution to sup*$_{\boldsymbol{\alpha}_2\in\mathscr{A}_2}\mathbb{E}[Q_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_2)]$ *where*

$$Q_2(Z_{2i},\boldsymbol{\alpha}_2)\equiv y_{2i}\log[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right)]+(1-y_{2i})\log[1-\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right)] \tag{2.25}$$

We estimate $\boldsymbol{\alpha}_{20} \in \mathscr{A}_2$ by $\hat{\boldsymbol{\alpha}}_{2N} \in \mathscr{A}_{2N}$ where $\mathscr{A}_{2N}$ is a finite dimensional sieve space defined as:

$$\begin{aligned}\mathscr{A}_{2N} &= \mathscr{B}_2 \mathscr{X} \left\{ \sigma_2(.) = exp(\mathrm{S}_{K_{\sigma,N}}(\mathrm{x}_i)\boldsymbol{\Pi}_N) : \boldsymbol{\Pi}_N \in \mathbb{R}^{K_N^{\sigma}} \right\} \\ &= \mathscr{B}_2 \mathscr{X} \mathscr{S}_{2N}\end{aligned}$$

where the $dim(\mathscr{A}_{2N}) = dim(\mathscr{B}_2) + dim(\mathscr{S}_{2N}) = K_x + K_{\sigma,N} \equiv K_2$.

Assumption 2.4.3 *Sieve Spaces in the First Step*

(i) *The sieve spaces $\mathscr{A}_{2n}$ are compact under $||.||_{\mathscr{A}_2}$*

(ii) $\mathscr{A}_{2n} \subseteq \mathscr{A}_{2\,n+1} \subseteq ... \subseteq \mathscr{A}_2 \forall n \geq 1$

(iii) $\exists \pi_n \boldsymbol{\alpha}_{20} \in \mathscr{A}_{2n} \ni ||\pi_n \boldsymbol{\alpha}_{20} - \boldsymbol{\alpha}_{20}||_{\mathscr{A}_2} \to 0$ *as* $n \to \infty$

$\hat{\boldsymbol{\alpha}}_{2N}$ is defined as:

$$\frac{1}{N}\sum_{i=1}^{N} Q_2(\mathbf{Z}_{2i}, \hat{\boldsymbol{\alpha}}_{2N}) \geq sup_{\boldsymbol{\alpha}_{2N} \in \mathscr{A}_{2N}} \frac{1}{N}\sum_{i=1}^{N} Q_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{2N}) - O_p(\varepsilon_{2N}^2) \tag{2.26}$$

where ε_{2N} is the magnitude of the optimization error. Define $\mathbf{Z}_{2i} \equiv (y_{2i}, \mathbf{x}_i)$.

Let $\vec{\boldsymbol{\beta}}_{10} \equiv (\boldsymbol{\beta}_{00}, \boldsymbol{\beta}_{10})'$, $\mathfrak{g}_0(\mathrm{x}_i) \equiv (\mathfrak{g}_{00}(\mathrm{x}_i), \mathfrak{g}_{10}(\mathrm{x}_i))$. Further define $\vec{\mathbf{x}}_{1i} \equiv (\mathbf{x}_{1i}, y_{2i}\mathbf{x}_{1i})$, $\vec{\boldsymbol{h}}_i \equiv \vec{\boldsymbol{h}}(\boldsymbol{\alpha}_{02}; \mathbf{Z}_{2i}) \equiv [h(\boldsymbol{\alpha}_{02}; \mathbf{Z}_{2i}), y_{2i}h(\boldsymbol{\alpha}_{02}; \mathbf{Z}_{2i})]'$

Assumption 2.4.4 *In the second stage, we assume that $\boldsymbol{\alpha}_{10} \equiv (\vec{\boldsymbol{\beta}}_{10}, \mathfrak{g}_0) \in \mathscr{A}_1$ is the unique solution to $sup_{\boldsymbol{\alpha}_1 \in \mathscr{A}_1} \mathbb{E}[Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_{20})]$ where*

$$Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_{20}) \equiv -\frac{[y_{1i} - (\vec{\mathrm{x}}_{1i}\vec{\boldsymbol{\beta}}_1 + \vec{\mathfrak{g}}_0(\mathrm{x}_i)\vec{\boldsymbol{h}}_i)]^2}{2} \tag{2.27}$$

We estimate $\boldsymbol{\alpha}_{10} \in \mathscr{A}_1$ by $\hat{\boldsymbol{\alpha}}_{1N} \in \mathscr{A}_{1N}$ where $\mathscr{A}_{1N}$ is a finite dimensional sieve space defined as

$$\begin{aligned}\mathscr{A}_{1N} &= \mathscr{B}_1 \mathscr{X} \left\{ \mathfrak{g}_0(.) = \mathfrak{G}_{0,K_{\mathfrak{g}_0,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02} : \boldsymbol{\Omega}_{N,02} \in \mathbb{R}^{K_{\mathfrak{g}_0,N}} \right\} \\ &\quad \mathscr{X} \left\{ \mathfrak{g}_1(.) = \mathfrak{G}_{1,K_{\mathfrak{g}_1,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12} : \boldsymbol{\Omega}_{N,12} \in \mathbb{R}^{K_{\mathfrak{g}_1,N}} \right\} \\ &= \mathscr{B}_1 \mathscr{X} \mathscr{G}_N\end{aligned}$$

where the $dim(\mathscr{A}_{1N}) = dim(\mathscr{B}_1) + dim(\mathscr{G}_N) = K_{x_1} + K_{\mathfrak{g}_0,N} + K_{\mathfrak{g}_1,N} \equiv K_1$

Assumption 2.4.5 *Sieve Spaces in the Second Step*

(i) *The sieve spaces $\mathscr{A}_{1n}$ are compact under $||.||_{\mathscr{A}_1}$*

(ii) $\mathscr{A}_{1n} \subseteq \mathscr{A}_{1\,n+1} \subseteq ... \subseteq \mathscr{A}_1 \forall n \geq 1$

(iii) $\exists \pi_n \boldsymbol{\alpha}_{10} \in \mathscr{A}_{1n} \ni ||\pi_n \boldsymbol{\alpha}_{10} - \boldsymbol{\alpha}_{10}||_{\mathscr{A}_1} \to 0$ *as* $n \to \infty$

$\hat{\boldsymbol{\alpha}}_{1N}$ is defined as:

$$\frac{1}{N}\sum_{i=1}^{N} Q_1(\mathbf{Z}_{1i}, \hat{\boldsymbol{\alpha}}_{1N}, \hat{\boldsymbol{\alpha}}_{2N}) \geq sup_{\boldsymbol{\alpha}_{1N} \in \mathscr{A}_{1N}} \frac{1}{N}\sum_{i=1}^{N} Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{1N}, \hat{\boldsymbol{\alpha}}_{2N}) - O_p(\varepsilon_{1N}^2) \qquad (2.28)$$

where ε_{1N} is the magnitude of the optimization error.

2.4.2.1 Consistency

The following theorem provide the consistency result of our two step sieve estimators.

Theorem 4 *(a) Under Assumptions 2.4.1, 2.4.2 and 2.4.3 the first stage estimator $\hat{\boldsymbol{\alpha}}_{2N}$ is consistent for $\boldsymbol{\alpha}_{20}$ under the pseudo metric $||.||_{\mathscr{A}_2}$ defined on $\mathscr{A}_2$. The convergence rate is defined as δ_{2N}^*.*

(b) Under Assumptions 2.4.1, 2.4.4 and 2.4.5 the second step estimator $\hat{\boldsymbol{\alpha}}_{1N}$ is consistent for $\boldsymbol{\alpha}_{10}$ under the pseudo metric $||.||_{\mathscr{A}_1}$ defined on $\mathscr{A}_1$. The convergence rate is defined as δ_{1N}^*.

Given the convergence rates, we can define the *shrinking neighborhoods* and assume that

- $\hat{\boldsymbol{\alpha}}_{2N}$ belongs to a *shrinking neighborhood* of $\boldsymbol{\alpha}_{20}$ with probability approaching 1 defined as:

$$\mathscr{N}_{2N} \equiv \{\boldsymbol{\alpha}_{2N} \in \mathscr{A}_{2N} : ||\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}||_{\mathscr{A}_2} \leq \delta_{2N}\}$$

  where $\delta_{2N} \equiv \delta_{2N}^* log(log(N)) = o(1)$.

- $\hat{\boldsymbol{\alpha}}_{1N}$ belongs to a shrinking neighborhood of $\boldsymbol{\alpha}_{10}$ with probability approaching 1 defined as:

$$\mathscr{N}_{1N} \equiv \{\boldsymbol{\alpha}_{1N} \in \mathscr{A}_{1N} : ||\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}||_{\mathscr{A}_1} \leq \delta_{1N}\}$$

  where $\delta_{1N} \equiv \delta_{1N}^* log(log(N)) = o(1)$

The assumptions for the consistency of the first step sieve estimator correspond to the theoretical conditions in Theorm 3.1 in Chen(3.1). Since the likelihood function in the first step is a CDF, it satisfies he continuous and thus is sufficient for consistency.

The consistency of the second step sieve estimator takes into account that the first step sieve estimator is consistent. Since our first step sieve estimator is essentially an interaction term in the second step estimating equation, we have a fairly straightforward least square criterion function in the second step that satisfies the continuity and uniform convergence conditions.

2.4.2.2 Asymptotic Normality

The asymptotic normality results follow the general asymptotic theory developed by Hahn,Liao and Ridder(2018). We denote $\hat{\boldsymbol{\alpha}}_N \equiv \{\hat{\boldsymbol{\alpha}}_{1N}, \hat{\boldsymbol{\alpha}}_{2N}\}$ which is an estimator of $\boldsymbol{\alpha}_0 \equiv \{\boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{20}\} \in \mathscr{A} \equiv \mathscr{A}_1 \mathscr{X} \mathscr{A}_2$ We are interested in the asymptotic properties of a linear functional $\rho(\hat{\boldsymbol{\alpha}})$ which is an estimator of $\rho(\boldsymbol{\alpha}_0)$. We defined the *shrinking neighborhoods* for our first stage and second stage estimators as $\mathscr{N}_{2N}, \mathscr{N}_{1N}$ respectively. Based on these definitions, we can assume that $\hat{\boldsymbol{\alpha}}_N$ belongs to the *shrinking neighborhood* defined as: $\mathscr{N}_N \equiv \{(\boldsymbol{\alpha}_{1N}, \boldsymbol{\alpha}_{2N}) : \boldsymbol{\alpha}_{1N} \in \mathscr{N}_{1N}$ and $\boldsymbol{\alpha}_{2N} \in \mathscr{N}_{2N}\}$, with probability approaching one.

Riesz Representer for the First Step: Suppose that for all $\boldsymbol{\alpha}_{2N} \in \mathscr{N}_{2N}$, we can approximate $Q_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{2N}) - Q_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})$ by $\Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}]$ such that $\Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}]$ is linear in $\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}$. Since $\boldsymbol{\alpha}_{20}$ is a unique maximizer of $\mathbb{E}[Q_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_2)]$ on $\mathscr{A}_2$, we can let:

$$-\frac{\partial \mathbb{E}[\Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20} + \tau[\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}])[\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}]]}{\partial \tau}\bigg|_{\tau=0} \equiv ||\boldsymbol{\alpha}_{2N} - \boldsymbol{\alpha}_{20}||^2_{\mathscr{A}_2}$$

which defines a norm on $\mathscr{N}_{2N}$. Let $\mathscr{H}_2$ be the closed linear span of $\mathscr{N}_{2N} - \{\boldsymbol{\alpha}_{20}\}$ under $||.||_{\mathscr{A}_2}$ which is a Hilbert Space under $||.||_{\mathscr{A}_2}$ with the corresponding inner product $\langle .,. \rangle$ defined as:

$$\langle \mathtt{a}_{2N}, \mathtt{b}_{2N} \rangle_{\mathscr{A}_2} = -\frac{\partial \mathbb{E}\left[\Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20} + \tau \mathtt{b}_{2N})[\mathtt{a}_{2N}]\right]}{\partial \tau}\bigg|_{\tau=0}$$

for any $\mathtt{a}_{1N}, \mathtt{b}_{2N} \in \mathscr{H}_2$ Typically, we will have

$$\Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\mathtt{a}_{2N}] = \frac{\partial Q_2(\mathbf{Z}_{2N}, \boldsymbol{\alpha}_{20} + \tau \mathtt{a}_{2N})}{\partial \tau}\bigg| \text{ and}$$
$$\langle \mathtt{a}_{2N}, \mathtt{b}_{2N} \rangle_{\mathscr{A}_2} = -\mathbb{E}\left[\frac{\partial \Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20} + \tau \mathtt{b}_{2N})[\mathtt{a}_{2N}]}{\partial \tau}\bigg|_{\tau=0}\right] \quad (2.29)$$

if the derivative exists and we can interchange the derivative and expectation. We assume that there is a linear functional $\partial_2 \rho(\boldsymbol{\alpha}_0)[.] : \mathscr{H}_2 \to \mathbb{R}$ such that

$$\partial_2 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_2] = \frac{\partial \rho(\boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{20} + \tau \mathtt{a}_2)}{\partial \tau}\bigg|_{\tau=0} \text{ for all } \mathtt{a}_2 \in \mathscr{H}_2 \quad (2.30)$$

Let $\boldsymbol{\alpha}_{20,N}$ denote the projection of $\boldsymbol{\alpha}_{20}$ on $\mathscr{A}_{2N}$ under the norm $||.||_{\mathscr{A}_2}$. Let $\mathscr{H}_{2N}$ denote the Hilbert space generated by $\mathscr{N}_{2N} - \{\boldsymbol{\alpha}_{20,N}\}$. Then $dim(\mathscr{H}_{2N}) = dim(\mathscr{A}_{2N}) < \infty$. By *Riesz Representation Theorem*, there exists a *sieve Riesz Representer* $\mathfrak{a}^*_{2_N} \in \mathscr{H}_{2N}$ such that

$$\partial_2 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_2] = \langle \mathfrak{a}^*_{2_N}, \mathtt{a}_2 \rangle_{\mathscr{A}_2} \forall \mathtt{a}_2 \in \mathscr{H}_{2N} \text{ and } ||\mathfrak{a}^*_{2_N}||^2_{\mathscr{A}_2} = \sup_{0 \neq \mathtt{a}_2 \in \mathscr{H}_{2N}} \frac{\left|\partial_2 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_2]\right|^2}{||\mathtt{a}_2||^2_{\mathscr{A}_2}}.$$

Riesz Representer for the Second Step: Suppose that for all $\boldsymbol{\alpha}_{1N} \in \mathscr{N}_{1N}$, we can approximate $Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{1N}, \boldsymbol{\alpha}_{20}) - Q_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{20})$ by $\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{20})[\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}]$ such that

$\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{20})[\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}]$ is linear in $\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}$. Since $\boldsymbol{\alpha}_{10}$ is a unique maximizer of $\mathbb{E}[Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_{20})]$ on $\mathscr{A}_1$, we can let:

$$-\frac{\partial \mathbb{E}[\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10} + \tau[\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}], \boldsymbol{\alpha}_{20})[\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}]]}{\partial \tau}\bigg|_{\tau=0} \equiv ||\boldsymbol{\alpha}_{1N} - \boldsymbol{\alpha}_{10}||^2_{\mathscr{A}_1}$$

which defines a norm on $\mathscr{N}_{1N}$. Let $\mathscr{H}_1$ be the closed linear span of $\mathscr{N}_{1N} - \{\boldsymbol{\alpha}_{10}\}$ under $||.||_{\mathscr{A}_1}$ which is a Hilbert Space under $||.||_{\mathscr{A}_1}$ with the corresponding inner product $\langle .,. \rangle_{\mathscr{A}_1}$ defined as:

$$\langle \mathtt{a}_{1N}, \mathtt{b}_{1N} \rangle_{\mathscr{A}_1} = -\frac{\partial \mathbb{E}[\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10} + \tau \mathtt{b}_{1N}, \boldsymbol{\alpha}_{20})[\mathtt{a}_{1N}]]}{\partial \tau}\bigg|_{\tau=0}$$

for any $\mathtt{a}_{1N}, \mathtt{b}_{1N} \in \mathscr{H}_1$ Typically, we will have

$$\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_0)[\mathtt{a}_{1N}] = \frac{\partial Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10} + \tau \mathtt{a}_{1N}, \boldsymbol{\alpha}_{20})}{\partial \tau}\bigg|$$

$$\text{and } \langle \mathtt{a}_{1N}, \mathtt{b}_{1N} \rangle_{\mathscr{A}_1} = -\mathbb{E}\left[\frac{\partial \Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10} + \tau \mathtt{b}_{1N}, \boldsymbol{\alpha}_{20})[\mathtt{a}_{1N}]}{\partial \tau}\bigg|_{\tau=0}\right] \quad (2.31)$$

if the derivative exists and we can interchange the derivative and expectation. We assume that there is a linear functional $\partial_1 \rho(\boldsymbol{\alpha}_0)[.] : \mathscr{H}_1 \to \mathbb{R}$ such that

$$\partial_1 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_1] = \frac{\partial \rho(\boldsymbol{\alpha}_{10} + \tau \mathtt{a}_1, \boldsymbol{\alpha}_{20})}{\partial \tau}\bigg|_{\tau=0} \text{ for all } \mathtt{a}_1 \in \mathscr{H}_1 \quad (2.32)$$

Let $\boldsymbol{\alpha}_{10,N}$ denote the projection of $\boldsymbol{\alpha}_{10}$ on $\mathscr{A}_{1N}$ under the norm $||.||_{\mathscr{A}_1}$. Let $\mathscr{H}_{1N}$ denote the Hilbert space generated by $\mathscr{N}_{1N} - \{\boldsymbol{\alpha}_{10,N}\}$. Then $dim(\mathscr{H}_{1N}) = dim(\mathscr{A}_{1N}) < \infty$. By *Riesz Representation Theorem*, there exists a *sieve Riesz Representer* $\mathfrak{a}^*_{1_N} \in \mathscr{H}_{1N}$ such that

$$\partial_1 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_1] = \langle \mathfrak{a}^*_{1_N}, \mathtt{a}_1 \rangle_{\mathscr{A}_1} \forall \mathtt{a}_1 \in \mathscr{H}_{1N} \text{ and } ||\mathfrak{a}^*_{1_N}||^2_{\mathscr{A}_1} = \sup_{0 \neq \mathtt{a}_1 \in \mathscr{H}_{1N}} \frac{\left|\partial_1 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_1]\right|^2}{||\mathtt{a}_1||^2_{\mathscr{A}_1}}.$$

Let $\mathscr{H} = \mathscr{H}_1 \mathscr{X} \mathscr{H}_2$. For any $\mathtt{a} = (\mathtt{a}_1, \mathtt{a}_2) \in \mathscr{H}$, we denote

$$\partial_\alpha \rho(\boldsymbol{\alpha}_0)[\mathtt{a}] = \partial_1 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_1] + \partial_2 \rho(\boldsymbol{\alpha}_0)[\mathtt{a}_2] \quad (2.33)$$

To evaluate the effect of first step sieve estimation on the asymptotic variance of the second step sieve estimator, we define

$$\mathtt{F}_1(\boldsymbol{\alpha}_0)[\mathtt{a}_1] = \frac{\partial \mathbb{E}[Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{10} + \tau_1 \mathtt{a}_1, \boldsymbol{\alpha}_{20})]}{\partial \tau_1}\bigg|_{\tau_1=0} \text{ for any } \mathtt{a}_1 \in \mathscr{H}_1 \quad (2.34)$$

and

$$\mathtt{F}(\boldsymbol{\alpha}_0)[\mathtt{a}_1,\mathtt{a}_2] = \left.\frac{\partial[\mathtt{F}_1(\boldsymbol{\alpha}_{10},\boldsymbol{\alpha}_{20}+\tau_2\mathtt{a}_2)]}{\partial\tau_2}\right|_{\tau_2=0,} \text{ for any } \mathtt{a}_2 \in \mathscr{H}_2 \tag{2.35}$$

Assume that $\mathtt{F}(\boldsymbol{\alpha}_0)[.,.]$ is a bilinear functional on $\mathscr{H}$. Given the Riesz Representer $\mathfrak{a}^*_{1_N}$, define $\mathfrak{a}^*_{\mathtt{F}_N} \in \mathscr{H}_{2N}$ as:

$$\mathtt{F}(\boldsymbol{\alpha}_0)[\mathtt{a}_{2N},\mathfrak{a}^*_{1_N}] = \left\langle \mathtt{a}_{2N},\mathfrak{a}^*_{\mathtt{F}_N}\right\rangle_{\mathscr{A}_2} \text{ for any } \mathtt{a}_{2N} \in \mathscr{H}_{2N}$$

Finally, we define:

$$||\mathscr{V}^*_N||^2 \equiv Var\left[\frac{\sum_{i=1}^N\left[\Delta_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_{20})[\mathtt{a}^*_{2_N}]+\Delta_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_{20})[\mathfrak{a}^*_{\mathtt{F}_N}]+\Delta_1(\mathbf{Z}_{1i},\boldsymbol{\alpha}_0)[\mathtt{a}^*_{1_N}]\right]}{N^{1/2}}\right]$$

Theorem 5 *Under the assumptions of Theorem 3.1 in Hahn, Liao and Ridder(2018), the two step estimator satisfies:*

$$\frac{\sqrt{N}\left[\rho(\hat{\boldsymbol{\alpha}}_{1N},\hat{\boldsymbol{\alpha}}_{2N})-\rho(\boldsymbol{\alpha}_{10,N},\boldsymbol{\alpha}_{20,N})\right]}{||\mathscr{V}^*_N||} \to_d \mathscr{N}(0,1) \tag{2.36}$$

$$\frac{\sqrt{N}\left[\rho(\hat{\boldsymbol{\alpha}}_{1N},\hat{\boldsymbol{\alpha}}_{2N})-\rho(\boldsymbol{\alpha}_{10},\boldsymbol{\alpha}_{20})\right]}{||\mathscr{V}^*_N||} \to_d \mathscr{N}(0,1) \tag{2.37}$$

2.4.2.3 Consistent Variance Estimator

To obtain a method of inference, we now need a consistent estimator of $||\mathscr{V}^*_N||$. In this section we obtain the sample analog of $||\mathscr{V}^*_N||$ which would be a consistent estimator. First assume that our data is *iid* and both the criterion functions are twice pathwise differentiable with respect to $\boldsymbol{\alpha}_{2N}$ and $(\boldsymbol{\alpha}_{1N},\boldsymbol{\alpha}_{2N})$ in $\mathscr{N}_N$. Next we define

$$\Delta_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_{2N})[\mathtt{a}_{2N}] \equiv \left.\frac{\partial Q_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_{2N}+\tau\mathtt{a}_{2N})}{\partial\tau}\right|_{\tau=0}$$

$$\text{and } \mathbb{H}_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_{2N})[\mathtt{a}_{2N},\mathtt{b}_{2N}] \equiv \left.\frac{\partial\Delta_2(\mathbf{Z}_{2i},\boldsymbol{\alpha}_{2N}+\tau\mathtt{a}_{2N})[\mathtt{b}_{2N}]}{\partial\tau}\right|_{\tau=0} \tag{2.38}$$

for all $(\mathtt{a}_{2N}, \mathtt{b}_{2N}) \in \mathscr{H}_{2N}$.

Similarly, define

$$\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{1N}, \boldsymbol{\alpha}_{2N})[\mathtt{a}_{2N}] \equiv \left.\frac{\partial Q_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{1N} + \tau \mathtt{a}_{1N}, \boldsymbol{\alpha}_{2N})}{\partial \tau}\right|_{\tau=0}$$

$$\text{and } \mathbb{H}_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{1N}, \boldsymbol{\alpha}_{2N})[\mathtt{a}_{2N}, \mathtt{b}_{2N}] \equiv \left.\frac{\partial \Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{1N} + \tau \mathtt{a}_{1N}, \boldsymbol{\alpha}_{2N})[\mathtt{b}_{2N}]}{\partial \tau}\right|_{\tau=0} \tag{2.39}$$

for all $(\mathtt{a}_{1N}, \mathtt{b}_{1N}) \in \mathscr{H}_{1N}$.

Empirical Riesz representer $\hat{\mathtt{a}}^*_{2_N}$ is defined as:

$$\partial \rho_2(\hat{\boldsymbol{\alpha}}_N)[\mathtt{a}_{2N}] = \left\langle \mathtt{a}_{2N}, \hat{\mathtt{a}}^*_{2_N} \right\rangle_{N,\mathscr{A}_2} \text{ for all } \mathtt{a}_{2N} \in \mathscr{H}_{2N} \text{ for any } \mathtt{a}_{2N} \in \mathscr{H}_{2N}. \tag{2.40}$$

where

$$\left\langle \mathtt{a}_{2N}, \mathtt{b}_{2N} \right\rangle_{N,\mathscr{A}_2} \equiv -\frac{1}{N} \sum_{i=1}^{N} \mathbb{H}_2(\mathbf{Z}_{2i}, \hat{\boldsymbol{\alpha}}_{2N})[\mathtt{a}_{2N}, \mathtt{b}_{2N}] \tag{2.41}$$

Similarly, we define *Empirical Riesz representer* $\hat{\mathtt{a}}^*_{1_N}$ as:

$$\partial_1 \rho(\hat{\boldsymbol{\alpha}}_N)[\mathtt{a}_{1N}] = \left\langle \mathtt{a}_{1N}, \hat{\mathtt{a}}^*_{1_N} \right\rangle_{N,\mathscr{A}_1} \text{ for all } \mathtt{a}_{1N} \in \mathscr{H}_{1N} \text{ for any } \mathtt{a}_{1N} \in \mathscr{H}_{1N}. \tag{2.42}$$

where

$$\left\langle \mathtt{a}_{1N}, \mathtt{b}_{1N} \right\rangle_{N,\mathscr{A}_1} \equiv -\frac{1}{N} \sum_{i=1}^{N} \mathbb{H}_2(\mathbf{Z}_{1i}, \hat{\boldsymbol{\alpha}}_N)[\mathtt{a}_{1N}, \mathtt{b}_{1N}] \tag{2.43}$$

and we define $\hat{\mathtt{a}}^*_{\mathtt{F}_N}$ as:

$$\mathtt{F}_N(\hat{\boldsymbol{\alpha}}_N)[\hat{\mathtt{a}}^*_{1N}, \mathtt{a}_{2N}] = \left\langle \hat{\mathtt{a}}^*_{\mathtt{F}_N}, \mathtt{a}_{2N} \right\rangle_{N,\mathscr{A}_2} \text{ for all } \mathtt{a}_{2N} \in \mathscr{H}_{2N} \tag{2.44}$$

where

$$\mathtt{F}_N(\hat{\boldsymbol{\alpha}}_N)[\hat{\mathtt{a}}^*_{1_N}, \mathtt{a}_{2N}] = \frac{1}{N} \sum_{i=1}^{N} \left.\frac{\partial \Delta_1(\mathbf{Z}_{1i}, \hat{\boldsymbol{\alpha}}_{1N}, \hat{\boldsymbol{\alpha}}_{2N} + \tau \mathtt{a}_{2N})[\hat{\mathtt{a}}^*_{1_N}]}{\partial \tau}\right|_{\tau=0} \tag{2.45}$$

A simple sample analog of $||\mathscr{V}^*_N||$ is given as:

$$||\widehat{\mathscr{V}^*_N}||^2 = \frac{1}{N} \sum_{i=1}^{N} \left| \Delta_2(\mathbf{Z}_{2i}, \hat{\boldsymbol{\alpha}}_{2N})[\hat{\mathtt{a}}^*_{2_N} + \hat{\mathtt{a}}^*_{\mathtt{F}_N}] + \Delta_1(\mathbf{Z}_{1i}, \hat{\boldsymbol{\alpha}}_{1N}, \hat{\boldsymbol{\alpha}}_{2N})[\hat{\mathtt{a}}^*_{1_N}] \right|^2 \tag{2.46}$$

Theorem 6 *Hahn Liao and Ridder(2018) establish the following results in Theorem 4.1:*

$$\left| \frac{||\widehat{\mathscr{V}_N^*}||}{||\mathscr{V}_N^*||} - 1 \right| = o_p(1) \tag{2.47}$$

Therefore,

$$\frac{\sqrt{N}[\rho(\hat{\boldsymbol{\alpha}}_{1N}, \hat{\boldsymbol{\alpha}}_{2N}) - \rho(\boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{20})]}{||\widehat{\mathscr{V}_N^*}||} \rightarrow_d \mathscr{N}(0,1) \tag{2.48}$$

2.4.2.4 Explicit Expressions in our model

Let

$$\left\langle \mathtt{a}_{2N}, \mathtt{b}_{2N} \right\rangle_{\mathscr{A}_2} = \mathscr{R}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\mathtt{a}_{2N}, \mathtt{b}_{2N}] \equiv \mathbb{E}[-\mathbb{H}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\mathtt{a}_{2N}, \mathtt{b}_{2N}]] \text{ for all } \mathtt{a}_{2N}, \mathtt{b}_{2N} \in \mathscr{H}_2$$

$$\left\langle \mathtt{a}_{1N}, \mathtt{b}_{1N} \right\rangle_{\mathscr{A}_1} = \mathscr{R}_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{0})[\mathtt{a}_{1N}, \mathtt{b}_{1N}] \equiv \mathbb{E}[-\mathbb{H}_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_{0})[\mathtt{a}_{1N}, \mathtt{b}_{1N}]] \text{ for all } \mathtt{a}_{1N}, \mathtt{b}_{1N} \in \mathscr{H}_1$$

In the first step, define $\bar{\mathtt{S}} \equiv \begin{pmatrix} \vec{\mathbf{1}}_{K_x} \\ \mathtt{S}'_{2N}(.) \end{pmatrix}$ where $\vec{\mathbf{1}}_{K_x}$ is a $K_x \times 1$ vector of ones and $\mathtt{S}(.)$ are the $1 \times K_N^{\sigma}$ vector of basis functions in $\mathscr{S}_{2N}$. Now for the first step, we will obtain the following expressions:

$$\Delta_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\bar{\mathtt{S}}] = \begin{pmatrix} \mathbf{x}_i' \dot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}) \\ -\mathtt{S}(.)_i' \mathbf{x}_i \boldsymbol{\beta}_2 \dot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}) \end{pmatrix}$$

$$\dot{\boldsymbol{q}}_{2i}(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}) = \frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right)\right]} \phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right) \frac{1}{\sigma_{20}^2(\mathbf{x}_i)}$$

$$\mathbb{E}(-\mathbb{H}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\bar{\mathtt{S}}, \bar{\mathtt{S}}]) = \begin{pmatrix} \mathbb{E}[\mathbf{x}_i' \mathbf{x}_i (\ddot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}))] & \mathbb{E}[-\mathbf{x}_i' \mathtt{S}(.)_i' (\mathbf{x}_i \boldsymbol{\beta}_2 \ddot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}))] \\ \mathbb{E}[-\mathtt{S}(.)_i' \mathbf{x}_i (\mathbf{x}_i \boldsymbol{\beta}_2 \ddot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}))] & \mathbb{E}[-\mathtt{S}(.)_i' \mathtt{S}(.)((\mathbf{x}_i \boldsymbol{\beta}_2)^2 \ddot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}))] \end{pmatrix}$$

$$\ddot{\boldsymbol{q}}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20}) = \frac{\left[\phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right)\right]^2}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_{20}^2(\mathbf{x}_i)}\right)\right][\sigma_{20}^2(\mathbf{x}_i)]^2}$$

$$\mathfrak{a}_{2N}^* = \bar{\mathtt{S}}[\mathscr{R}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\bar{\mathtt{S}}, \bar{\mathtt{S}}]]^{-1} \partial_2 \rho(\boldsymbol{\alpha}_0)[\bar{\mathtt{S}}]$$

In the second step, define $\bar{\mathtt{G}} \equiv \begin{pmatrix} \vec{\mathbf{1}}_{K_{x_1}} \\ \mathtt{G}'_N(.) \end{pmatrix}$ where $\vec{\mathbf{1}}_{K_{x_1}}$ is a $K_x \times 1$ vector of ones and $\mathtt{G}(.)$ are the $1 \times K_N^g$ vector of basis functions in $\mathscr{G}_N$. Further define $\partial \vec{\boldsymbol{h}}_i \equiv \frac{\partial \vec{\boldsymbol{h}}_i}{\partial \boldsymbol{\alpha}_{20}}$ Now for the second step, we will obtain the following expressions:

$$\Delta_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_0)[\bar{\mathtt{G}}] = \begin{pmatrix} \vec{\mathbf{x}}'_{1i}\eta_i \\ \mathtt{G}(.)'_i \vec{\boldsymbol{h}}_i \eta_i \end{pmatrix}$$

$$\mathbb{E}\left(-\mathbb{H}_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_0)[\bar{\mathtt{G}}, \bar{\mathtt{G}}]\right) = \begin{pmatrix} \mathbb{E}[\vec{\mathbf{x}}'_{1i}\vec{\mathbf{x}}_{1i}] & \mathbb{E}[-\vec{\mathbf{x}}'_i \mathtt{G}(.)'_i \vec{\boldsymbol{h}}_i] \\ \mathbb{E}[-\mathtt{G}(.)'_i \vec{\mathbf{x}}_i \vec{\boldsymbol{h}}_i] & \mathbb{E}[-\mathtt{G}(.)'_i \mathtt{G}(.)(\vec{\boldsymbol{h}}_i)^2] \end{pmatrix}$$

$$\mathtt{F}(\boldsymbol{\alpha}_0)[\bar{\mathtt{S}}, \bar{\mathtt{G}}] = \begin{pmatrix} \vec{\mathbf{x}}'_{i1} \mathtt{G}(.) \partial \vec{\boldsymbol{h}}_i \bar{\mathtt{S}} & \mathtt{G}(.)' \mathtt{G}(.) \partial \vec{\boldsymbol{h}}_i \bar{\mathtt{S}} \end{pmatrix}$$

$$\mathfrak{a}^*_{1_N} = \bar{\mathtt{G}}[\mathscr{R}_1(\mathbf{Z}_{1i}, \boldsymbol{\alpha}_0)[\bar{\mathtt{G}}, \bar{\mathtt{G}}]]^{-1} \partial_1 \rho(\boldsymbol{\alpha}_0)[\bar{\mathtt{G}}]$$

$$\mathfrak{a}^*_{\mathtt{F}_N} = \bar{\mathtt{S}}[\mathscr{R}_2(\mathbf{Z}_{2i}, \boldsymbol{\alpha}_{20})[\bar{\mathtt{S}}, \bar{\mathtt{S}}]]^{-1} \mathtt{F}(\boldsymbol{\alpha}_0)[\bar{\mathtt{S}}, \mathfrak{a}^*_{1_N}]$$

2.4.3 Numerical Equivalence

It has been well established in the theoretical econometrics literature that the in many cases of semi-nonparametric estimation, one can obtain the semi-nonparametric variances using the standard formulas derived in the parametric estimation. These numerical equivalence also hold true in the estimation strategies that involve two step estimation procedures. This greatly simplifies the estimation of the semiparametric asymptotic variance that takes account of the first step estimation. In this section, we will use this numerical equivalence result to derive the expression of the asymptotic variances of our sieve estimators.

Recall that in our model, at each step we have unknown functions forms: $\sigma_2(\mathrm{x}_i)$ in the first stage and $\{\mathfrak{g}_0(\mathrm{x}_i), \mathfrak{g}_1(\mathrm{x}_i)\}$ in the second stage. In the sieve estimation procedure, we construct the sieve spaces that are defined in section 2.2. Suppose that we make the incorrect Assumption that the unknown functions in both the steps take the parametric form. In particular, suppose

that for the misspecification, we have the following models for the functions:

$$\sigma_{20}(.) = \mathtt{S}(.)\mathbf{\Pi}_{\sigma_2}$$
$$\mathfrak{g}_0(.) = \mathtt{G}_0(.)\mathbf{\Omega}_{\mathfrak{g}_0}$$
$$\mathfrak{g}_1(.) = \mathtt{G}_1(.)\mathbf{\Omega}_{\mathfrak{g}_1}$$

where the terms are defined as in section 2.2 except that we have now suppressed the superscripts for notational simplicity. The most important thing under this misspecification is that the dimensions of these terms in the parametric approximation is equal to the number of basis functions in the sieve estimation. Under these misspecification, the criterion functions become:

$$\mathtt{Q}_1(\mathrm{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \mathtt{G}_0(.)\mathbf{\Omega}_{\mathfrak{g}_0}, \mathtt{G}_1(.)\mathbf{\Omega}_{\mathfrak{g}_1}; \mathtt{S}(.)\mathbf{\Pi}_{\sigma_2}) \equiv \mathtt{Q}_1$$
$$\mathtt{Q}_2(\mathrm{x}_i, \boldsymbol{\beta}_2, \mathtt{S}(.)\mathbf{\Pi}_{\sigma_2}) \equiv \mathtt{Q}_2$$

Define

$$\boldsymbol{\beta}_{\mathtt{Q}_1} \equiv \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \mathbf{\Omega}_{\mathfrak{g}_0}, \mathbf{\Omega}_{\mathfrak{g}_1}\}$$
$$\boldsymbol{\beta}_{\mathtt{Q}_1} \equiv \{\boldsymbol{\beta}_2, \mathbf{\Pi}_{\sigma_2}\}$$

To solve for the asymptotic variances, we cast the optimization problem into the GMM framework:

$$\begin{pmatrix} \frac{\partial \mathtt{Q}_1}{\partial \widehat{\boldsymbol{\beta}}_{\mathtt{Q}_1}} \\ \frac{\partial \mathtt{Q}_2}{\partial \widehat{\boldsymbol{\beta}}_{\mathtt{Q}_2}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \tag{2.49}$$

The asymptotic result is as follows:

$$\sqrt{N}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathtt{Q}_1} - \boldsymbol{\beta}_{\mathtt{Q}_1} \\ \widehat{\boldsymbol{\beta}}_{\mathtt{Q}_2} - \boldsymbol{\beta}_{\mathtt{Q}_2} \end{pmatrix} \rightarrow_d \mathscr{N}(\mathbf{0}, \mathtt{V}) \tag{2.50}$$

where:

$$
\begin{aligned}
\mathtt{V} &\equiv \mathtt{A}^{-1}\mathtt{B}\mathtt{A}^{-1} \\
\mathtt{A} &\equiv \mathbb{E}\begin{pmatrix} \frac{\partial^2 \mathtt{Q}_1}{\partial \boldsymbol{\beta}_{\mathtt{Q}_1}\partial \boldsymbol{\beta}_{\mathtt{Q}_1}} & \frac{\partial^2 \mathtt{Q}_1}{\partial \boldsymbol{\beta}_{\mathtt{Q}_1}\partial \boldsymbol{\beta}_{\mathtt{Q}_2}} \\ \frac{\partial^2 \mathtt{Q}_2}{\partial \boldsymbol{\beta}_{\mathtt{Q}_2}\partial \boldsymbol{\beta}_{\mathtt{Q}_1}} & \frac{\partial^2 \mathtt{Q}_2}{\partial \boldsymbol{\beta}_{\mathtt{Q}_2}\partial \boldsymbol{\beta}_{\mathtt{Q}_2}} \end{pmatrix} \equiv \begin{pmatrix} \mathtt{Q}_{11} & \mathtt{Q}_{12} \\ \mathtt{Q}_{21} & \mathtt{Q}_{22} \end{pmatrix} \\
\mathtt{B} &\equiv \mathbb{E}\begin{pmatrix} \frac{\partial \mathtt{Q}_1}{\partial \boldsymbol{\beta}_{\mathtt{Q}_1}}[\frac{\partial \mathtt{Q}_1}{\partial \boldsymbol{\beta}_{\mathtt{Q}_1}}]' & \frac{\partial \mathtt{Q}_1}{\partial \boldsymbol{\beta}_{\mathtt{Q}_1}}[\frac{\partial \mathtt{Q}_2}{\partial \boldsymbol{\beta}_{\mathtt{Q}_2}}]' \\ \frac{\partial \mathtt{Q}_2}{\partial \boldsymbol{\beta}_{\mathtt{Q}_2}}[\frac{\partial \mathtt{Q}_1}{\partial \boldsymbol{\beta}_{\mathtt{Q}_1}}]' & \frac{\partial \mathtt{Q}_2}{\partial \boldsymbol{\beta}_{\mathtt{Q}_2}}[\frac{\partial \mathtt{Q}_2}{\partial \boldsymbol{\beta}_{\mathtt{Q}_2}}]' \end{pmatrix} \equiv \begin{pmatrix} \mathtt{B}_{11} & \mathtt{B}_{12} \\ \mathtt{B}_{21} & \mathtt{B}_{22} \end{pmatrix}
\end{aligned}
$$

Using the results from the partitioned matrix, we can further extract the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathtt{Q}_1} - \boldsymbol{\beta}_{\mathtt{Q}_1})$ as:

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\mathtt{Q}_1} - \boldsymbol{\beta}_{\mathtt{Q}_1}) \to_d \mathcal{N}(\mathbf{0}, \mathtt{V}_{\boldsymbol{\beta}_{\mathtt{Q}_1}}) \tag{2.51}$$

where

$$
\begin{aligned}
\mathtt{V}_{\boldsymbol{\beta}_{\mathtt{Q}_1}} &\equiv \mathtt{A}_{11}\mathtt{B}_{11}\mathtt{A}_{11}' + \mathtt{A}_{12}\mathtt{B}_{21}\mathtt{A}_{11}' + \mathtt{A}_{11}\mathtt{B}_{12}\mathtt{A}_{21}' + \mathtt{A}_{12}\mathtt{B}_{22}\mathtt{A}_{11}' \\
\mathtt{A}_{11} &\equiv \left(\mathtt{Q}_{11} - \mathtt{Q}_{12}\mathtt{Q}_{22}^{-1}\mathtt{Q}_{21}\right)^{-1} \\
\mathtt{A}_{12} &\equiv -\mathtt{Q}_{11}\mathtt{Q}_{12}\left(\mathtt{Q}_{22} - \mathtt{Q}_{21}\mathtt{Q}_{11}^{-1}\mathtt{Q}_{12}\right)^{-1} \\
\mathtt{A}_{21} &\equiv -\mathtt{Q}_{22}\mathtt{Q}_{21}\left(\mathtt{Q}_{11} - \mathtt{Q}_{12}\mathtt{Q}_{22}^{-1}\mathtt{Q}_{21}\right)^{-1}
\end{aligned}
$$

The estimators for these expressions are simply the sample analogue of the population terms. We denote the estimator for $\mathtt{V}_{\boldsymbol{\beta}_{\mathtt{Q}_1}}$ as $\hat{\mathtt{V}}_{\boldsymbol{\beta}_{\mathtt{Q}_1}}$ The numerical equivalence result implies that

$$||\widehat{\mathscr{V}}_N^*||^2 = \hat{\mathtt{V}}_{\boldsymbol{\beta}_{\mathtt{Q}_1}} \tag{2.52}$$

2.5 Empirical Illustration

We illustrate our methods using the subset of the data on student performance and catholic school attendance from Altonji, Elder and Taber(2005). We begin with the the model

$$math12_i = \alpha_1 + \mathbf{x}_{1i}\boldsymbol{\beta}_0 + \delta cathhs_i + cathhs_i\mathbf{x}_{1i}\boldsymbol{\beta}_1 + v_{0i} + cathhs_i v_{1i} \tag{2.53}$$

where $\mathbf{x}_{1i}$ includes mother's education, father's education and the log of family income. Our primary parameter of interest is δ which the coefficient on *cathhs*. The instruments for *cathhs*, which is a binary indicator for attending a Catholic high school, is distance from nearest Catholic high school divided into five bins. Thus the switching indicator *cathhs* modeled as:

$$catths_i = \mathbb{1}[\alpha_2 + \mathbf{x}_i\boldsymbol{\beta}_2 - v_{2i} > 0] \tag{2.54}$$

where $\mathbf{x}_i$ contains $\mathbf{x}_{1i}$ and four distance dummy variables.

Preliminary Results[2]: According to our estimation procedure, we do the estimation in two steps. In the first step, we estimate the binary response model for the *cathhs* and obtain the generalised residuals denoted by $\widehat{gr}$. In the second step, we run the estimating equation that contains the corrections terms accounting for the endogeneity of *cathhs*. The results of the second step estimation are given in Table 1.

The four columns of Table 1 contain the second step estimation results from four estimation procedures. In the first column, we assume that the model for the switching indicator *cathhs* is homoskedastic and the model for the outcome variable *math*12 contains constant coefficients. Thus we run a simple *probit* command in the first step, obtain the generalized residuals and the run a simple *OLS* of

$$math12_i \text{ on } 1, cathhs_i, \mathbf{x}_{1i}, cathhs_i \times (\mathbf{x}_{1i} - \overline{\mathbf{x}_{1i}}), \widehat{gr2}, cathhs_i \times \widehat{gr2} \tag{2.55}$$

[2]All the estimation is done in Stata 13 and the standard errors are based on 1000 bootstrap replications.

where centering $\mathbf{x}_{1i}$ ensures that the coefficient on *cathhs* gives the average treatment effect. The estimates in the first column are similar to those obtained in Wooldridge(2015). The average treatment effect obtained in population is negative and is not statistically different from zero.

In the second model we again assume that the model for the switching indicator *cathhs* is homoskedastic, so the first step is again *probit*. However, now we allow for random coefficients on all the explanatory variables. So in the second stage, we run *OLS* of

$$math12_i \text{ on } 1, cathhs_i, \mathbf{x}_{1i}, cathhs_i \times (\mathbf{x}_{1i} - \overline{\mathbf{x}_{1i}}), \widehat{gr2}, cathhs_i \times \widehat{gr2}, \\ \widehat{gr2} \times (\mathbf{x}_{1i} - \overline{\mathbf{x}_{1i}}), cathhs \times \widehat{gr2} \times (\mathbf{x}_{1i} - \overline{\mathbf{x}_{1i}}) \quad (2.56)$$

The results of the second step estimation are given in column 2 of Table 1 and replicate the results obtained in Wooldridge(2015): the average treatment effect in the enitre population is essentially zero.

The third model now introduced heteroskedasticity in the reduced form model for the switching indicator *cathhs*. As a result, our estimating equation in the second step now takes the form of Equation(12). We estimate the model by both parametric and sieve estimation in both steps. Column 3 of Table 1 gives the results from the parametric estimation. The first step was a simple *hetprobit* of *cathhs* on $\{1, \mathbf{x}_i\}$ with the heteroskedastic function being a function of $\mathbf{x}_i$. In the second step, the unknown functions are assumed to take the form as given in Assumption 3.2. As before, center the explanatory variables to make the coefficient on *cathhs* consistent for the average treatment effect. We see that the average treatment is again not significantly different from zero.

Finally, in column 4 of Table 1, we estimate both the first step estimation and the second step estimation through sieve estimation. In the first step, the sieve space for the heteroskedastic variance function $\sigma_2^2(\mathbf{x}_i)$ is given as:

$$\mathscr{S}_{2N} = \left\{ exp(\mathtt{S}_{K_{\sigma},N}(\mathrm{x}_i)\mathbf{\Pi}_N) : \mathbf{\Pi}_N \in \mathbb{R}^{K_N^{\sigma}} \right\} \quad (2.57)$$

where $\mathtt{S}_{K_{\sigma,N}}(\mathrm{x}_i)$ is a vector of basis functions that contains the elements of the second order polynomials of vector $\mathbf{x}_i$. Considering this sieve space, we estimate the first stage model using a flexible *het probit* to obtain the estimates for the generalized residuals. In the second step, we define the sieve spaces for the unknown functions $\mathfrak{g}_0(), \mathfrak{g}_1(.)$ as follows:

$$\mathscr{G}_{0N} = \left\{ \mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02} : \boldsymbol{\Omega}_{N,02} \in \mathbb{R}^{K_{\mathrm{g}_0,N}} \right\}, \mathscr{G}_{1N} = \left\{ \mathtt{G}_{1,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12} : \boldsymbol{\Omega}_{N,12} \in \mathbb{R}^{K_{\mathrm{g}_1,N}} \right\} \tag{2.58}$$

where both $\mathtt{G}_{0,K_{\mathrm{g}_0,N}}(\mathrm{x}_i), \mathtt{G}_{0,K_{\mathrm{g}_1,N}}(\mathrm{x}_i)$ are each vector of basis functions that contains the elements of the second order polynomials of vector $\mathbf{x}_{1i}$. After defining the sieve spaces for the unknown functions, we run a flexible *OLS* in the second step estimation with the explanatory variables appropiately demeaned. The results from the second step estimation are given in the column 4 of Table 1. We once again find that the average treatment of attending a catholic high school does not have a significant effect on the scores.

Table 2.1: Empirical Illustration Results

Explanatory Variables	(1) Hom_Const	(2) Hom_Rand	(3) Het_para	(4) Het_Sieve
$cathhs$	-0.953	0.277	0.847	0.482
	(1.747)	(2.054)	(2.133)	(2.200)
$mthed$	0.709***	0.619***	0.599***	0.653***
	(0.0642)	(0.0801)	(0.0788)	(0.0811)
$fthed$	0.876***	0.871***	0.877***	0.886***
	(0.0583)	(0.0725)	(0.0716)	(0.0731)
$lfaminc$	1.858***	1.821***	1.856***	1.776***
	(0.149)	(0.179)	(0.177)	(0.184)
$cathhs \times (mthed - \overline{mthed})$	-0.0851	0.149	-0.677	-0.357
	(0.262)	(0.972)	(0.882)	(0.972)
$cathhs \times (fthed - \overline{fthed})$	0.184	-0.541	-0.143	0.387
	(0.238)	(0.891)	(0.894)	(0.931)
$cathhs \times (lfaminc - \overline{lfaminc})$	-0.691	-1.968	-2.445	-4.435*
	(0.634)	(2.092)	(2.200)	(2.457)
Observations	7,444	7,444	7,444	7,444
R-squared	0.185	0.186	0.186	0.189

Bootstrapped Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.1 (cont'd)

Explanatory Variables	(1) Hom_Const	(2) Hom_Rand	(3) Het_para	(4) Het_Sieve
$\widehat{gr}$	-1.523*	-0.592	-0.603	-0.652
	(0.797)	(0.895)	(0.924)	(0.914)
$cathhs \times \widehat{gr}$	3.308**	1.744	1.417	1.750
	(1.306)	(1.489)	(1.540)	(1.584)
$\widehat{gr} \times (mthed - \overline{mthed})$		-0.915*	-1.121**	11.32**
		(0.484)	(0.448)	(5.436)
$\widehat{gr} \times (fthed - \overline{fthed})$		-0.0478	-0.00513	2.846
		(0.453)	(0.440)	(5.285)
$\widehat{gr} \times (lfaminc - \overline{lfaminc})$		-0.519	-0.158	-21.22**
		(1.191)	(1.106)	(9.967)
$\widehat{gr} \times (mthed - \overline{mthed})^2$				-0.457***
				(0.165)
$\widehat{gr} \times (mthed \times fthed - \overline{mthed \times fthed})$				0.140
				(0.191)
Observations	7,444	7,444	7,444	7,444
R-squared	0.185	0.186	0.186	0.189

Bootstrapped Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.1 (cont'd)

Explanatory Variables	(1) Hom_Const	(2) Hom_Rand	(3) Het_para	(4) Het_Sieve
$\widehat{gr} \times (mthed \times lfaminc - \overline{mthed \times lfaminc})$				-0.140
				(0.494)
$\widehat{gr} \times (fthed - \overline{fthed})^2$				-0.317**
				(0.139)
$\widehat{gr} \times (fthed \times lfaminc - \overline{fthed \times lfaminc})$				0.381
				(0.507)
$\widehat{gr} \times (lfaminc - \overline{lfaminc})^2$				0.847
				(0.534)
$cathhs \times \widehat{gr} \times (mthed - \overline{mthed})$		0.826	1.554**	-12.08**
		(0.774)	(0.715)	(5.883)
$cathhs \times \widehat{gr} \times (fthed - \overline{fthed})$		0.536	0.255	-4.156
		(0.703)	(0.701)	(5.831)
$cathhs \times \widehat{gr} \times (lfaminc - \overline{lfaminc})$		1.251	1.184	22.28**
		(1.646)	(1.667)	(10.66)
Observations	7,444	7,444	7,444	7,444
R-squared	0.185	0.186	0.186	0.189

Bootstrapped Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.1 (cont'd)

Explanatory Variables	(1) Hom_Const	(2) Hom_Rand	(3) Het_para	(4) Het_Sieve
$cathhs \times \widehat{gr} \times (motheduc - \overline{motheduc})^2$				0.502***
				(0.181)
$cathhs \times \widehat{gr} \times (motheduc \times fatheduc - \overline{motheduc \times fatheduc})$				-0.137
				(0.211)
$cathhs \times \widehat{gr} \times (motheduc \times lfaminc - \overline{motheduc \times lfaminc})$				0.110
				(0.547)
$cathhs \times \widehat{gr} \times (fatheduc - \overline{fatheduc})^2$				0.273*
				(0.150)
$cathhs \times \widehat{gr} \times (fatheduc \times lfaminc - \overline{fatheduc \times lfaminc})$				-0.159
				(0.576)
$cathhs \times \widehat{gr} \times (lfaminc - \overline{lfaminc})^2$				-0.898
				(0.591)
intercept	11.18***	12.87***	12.69***	12.71***
	(1.381)	(1.649)	(1.653)	(1.706)
Observations	7,444	7,444	7,444	7,444
R-squared	0.185	0.186	0.186	0.189

Bootstrapped Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We also compute the average treatment effect on the untreated and the average treatment effect on the treated using the method given in Wooldridge(2015) for the heteroskedastic sieve estimation by running separate regressions. The results are given in Table 2. The estimates gives us a deeper insight into the reason leading to average treatment effect in the population being not statistically different from zero. As one can see, the ATE on the Treated is positive and significantly different from zero. However, the ATE on the Untreated is not significantly different from zero. In this dataset, though the ATE on the Untreated is numerically small, the fact that the only 452 observations are treated compared to 6992 untreated observations makes the average treatment effect in the population not significantly different from zero.

Table 2.2: Treatment Effects

	Observed Coefficient	**Bootstrap Std. Error**	**Z**	**P >z**	**Normal Based [95 % confidence interval]**	
ATE on the Treated	2.774772	1.319732	2.10	0.036	.1881439	5.3614
ATE on the Untreated	.2969452	2.088351	0.14	0.887	-3.796147	4.390037
ATE	.4473989	1.97412	0.23	0.821	-3.421804	4.316602

Treated Observations = 452
Untreated Observations = 6992

2.6 Technical Details

2.6.1 Asymptotic Variance for the Parametric Estimators

The asymptotic variance in the parametric setting can be obtained by using the M estimation methods. We will first define the notations in the parametric setting. We have the following models:

- $\mathbf{w}_i \equiv \{\mathrm{x}_i, y_{2i}\mathrm{x}_{1i}, h(y_{2i}, \mathrm{x}_i)\mathrm{x}_i, y_{2i}h(y_{2i}, \mathrm{x}_i)\mathrm{x}_i\}$

- $\boldsymbol{\theta}_1 \equiv \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}_{02}, \boldsymbol{\Omega}_{12}\}$
- $\boldsymbol{\theta}_2 \equiv \{\boldsymbol{\beta}_2, \boldsymbol{\Pi}\}$
- $\boldsymbol{\theta} \equiv \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$

Next, we will define the estimation problem in the M estimation framework:

$$q(\mathbf{w}_1, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2) \equiv \begin{pmatrix} q_1(\mathbf{w}_1, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ q_2(\mathbf{w}_1, \boldsymbol{\theta}_2) \\ q_3(\mathbf{w}_1, \boldsymbol{\theta}_2) \end{pmatrix} \tag{2.59}$$

where:

$$q_1(\mathbf{w}_1, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2) \equiv -\left[y_{1i} - \mathrm{x}_{1i}\boldsymbol{\beta}_0 - y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 - \hat{h}_i \mathrm{x}_i \boldsymbol{\Omega}_{01} - y_{2i}\hat{h}_i \mathrm{x}_i \boldsymbol{\Omega}_{01}\right] \mathbf{w}_i' \equiv -\eta_i \mathbf{w}_i' \tag{2.60}$$

$$q_2(\mathbf{w}_1, \boldsymbol{\theta}_2) \equiv \frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]} \phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \frac{\mathrm{x}_i'}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \tag{2.61}$$

$$q_3(\mathbf{w}_1, \boldsymbol{\theta}_2) \equiv \frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]} \phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \tilde{x}_i \mathrm{x}_i' \tag{2.62}$$

where $\tilde{x}_i \equiv [\mathrm{x}_i\boldsymbol{\beta}_2][-1][exp(\mathrm{x}_i\boldsymbol{\Pi})]^{-1}$

The vector $q(.)$ denotes the first order conditions for our estimation criterion functions that implies:

$$\mathbb{E}\begin{pmatrix} q_1(\mathbf{w}_i, \boldsymbol{\theta}_{01}; \boldsymbol{\theta}_{02}) \\ q_2(\mathbf{w}_i, \boldsymbol{\theta}_{02}) \\ q_3(\mathbf{w}_i, \boldsymbol{\theta}_{02}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{2.63}$$

The asymptotic distribution of our parametric estimators is given as:

$$\sqrt{N}\begin{pmatrix} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01}) \\ (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{02}) \end{pmatrix} \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbb{V}) \tag{2.64}$$

where:

$$\mathbb{V} \equiv \mathbb{A}^{-1}\mathbb{B}\mathbb{A}'^{-1}$$
$$\mathbb{A} \equiv \mathbb{E}\left[\frac{\partial q(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}}\right]$$
$$\mathbb{B} \equiv \mathbb{E}[q(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2) q(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)']$$

In addition, we define $\sigma_2^2(\mathrm{x}_i, \Pi) \equiv Var(v_{i2}|\mathrm{x}_i)$ In the rest of this subsection, we will define the terms for $\mathbb{A}, \mathbb{B}$.

$$\frac{\partial q(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial q_1(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} & \frac{\partial q_1(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\beta}_2} & \frac{\partial q_1(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\Pi}} \\ \frac{\partial q_2(\mathbf{w}_i, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} & \frac{\partial q_2(\mathbf{w}_i, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\beta}_2} & \frac{\partial q_2(\mathbf{w}_i, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\Pi}} \\ \frac{\partial q_3(\mathbf{w}_i, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} & \frac{\partial q_3(\mathbf{w}_i, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\beta}_2} & \frac{\partial q_3(\mathbf{w}_i, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\Pi}} \end{pmatrix}$$

Next,

$$\frac{\partial q_1(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} = \mathbf{w}_i' \mathbf{w}_i$$
$$\frac{\partial q_1(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} = \mathbf{w}_i' \mathbf{0} \quad \mathbf{0} \quad \lambda^{\mathrm{d}}\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \frac{\mathrm{x}_i'}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \mathrm{x}_i \boldsymbol{\Omega}_{02} \quad y_{2i} \lambda^{\mathrm{d}}\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \frac{\mathrm{x}_i'}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \mathrm{x}_i \boldsymbol{\Omega}_{12}$$
$$\frac{\partial q_1(\mathbf{w}_i, \boldsymbol{\theta}_1; \boldsymbol{\theta}_2)}{\partial \boldsymbol{\Pi}} = \mathbf{w}_i' \mathbf{0} \quad \mathbf{0} \quad \lambda^{\mathrm{d}}\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \tilde{x}_i \mathrm{x}_i \boldsymbol{\Omega}_{02} \quad y_{2i} \lambda^{\mathrm{d}}\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \tilde{x}_i \mathrm{x}_i \boldsymbol{\Omega}_{12}$$

$$\frac{\partial q_2(\mathbf{w}_i,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} = \mathbf{0}$$

$$\frac{\partial q_2(\mathbf{w}_i,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\beta}_2} = \left[-\frac{\left[\phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]^2 \frac{\mathrm{x}_i'\mathrm{x}_i}{\sigma_2^2(\mathrm{x}_i,\boldsymbol{\Pi})}}{\left[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]\left[1-\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]} + \tilde{q}_2(\mathbf{w}_i,\boldsymbol{\theta}_2) \right]$$

$$\frac{\partial q_2(\mathbf{w}_i,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\Pi}} = \left[-\frac{\left[\phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]^2 \frac{\tilde{x}_i\mathrm{x}_i\mathrm{x}_i'}{\sigma_2^2(\mathrm{x}_i,\boldsymbol{\Pi})}}{\left[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]\left[1-\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]} + \tilde{q}_2(\mathbf{w}_i,\boldsymbol{\theta}_2) \right]$$

$$\frac{\partial q_3(\mathbf{w}_i,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} = \mathbf{0}$$

$$\frac{\partial q_3(\mathbf{w}_i,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} = \left[-\frac{\left[\phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]^2 \frac{\tilde{x}_i\mathrm{x}_i\mathrm{x}_i'}{\sigma_2^2(\mathrm{x}_i,\boldsymbol{\Pi})}}{\left[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]\left[1-\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]} + \tilde{q}_3(\mathbf{w}_i,\boldsymbol{\theta}_2) \right]$$

$$\frac{\partial q_3(\mathbf{w}_i,\boldsymbol{\theta}_2)}{\partial \boldsymbol{\Pi}} = \left[-\frac{\left[\phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]^2 \frac{\tilde{x}_i^2\mathrm{x}_i\mathrm{x}_i'}{\sigma_2^2(\mathrm{x}_i,\boldsymbol{\Pi})}}{\left[\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]\left[1-\Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)\right]} + \tilde{q}_3(\mathbf{w}_i,\boldsymbol{\theta}_2) \right]$$

where $\lambda^{\mathrm{d}}(.)$ denotes the first derivative of the inverse mills ratio. In addition, $\mathbb{E}[\tilde{q}_2(\mathbf{w}_i,\boldsymbol{\theta}_2)]$ and $\mathbb{E}[\tilde{q}_3(\mathbf{w}_i,\boldsymbol{\theta}_2)]$ are functions of $\mathbf{w}_i,\boldsymbol{\theta}_2$ that will become equal to zero when we will take expectations because they contain $[y_{2i} - \Phi\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i,\boldsymbol{\Pi})}\right)]$

Next we define $\mathbb{B} = \mathbb{E}[q(\mathbf{w}_i,\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)q(\mathbf{w}_i,\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)']$

$$= \mathbb{E}\left[\begin{pmatrix} q_1(\mathbf{w}_i,\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) \\ q_2(\mathbf{w}_i,\boldsymbol{\theta}_2) \\ q_3(\mathbf{w}_i,\boldsymbol{\theta}_2) \end{pmatrix} \begin{pmatrix} q_1(\mathbf{w}_i,\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)' & q_2(\mathbf{w}_i,\boldsymbol{\theta}_2)' & q_3(\mathbf{w}_i,\boldsymbol{\theta}_2)' \end{pmatrix} \right]$$

$$= \mathbb{E}\left[\begin{pmatrix} q_1q_1' & q_1q_2' & q_1q_3' \\ q_2q_1' & q_2q_2' & q_2q_3' \\ q_3q_1' & q_3q_2' & q_3q_3' \end{pmatrix} \right]$$

Note that we are the suppressing the arguments of the function for expositional simplicity

$$
\begin{aligned}
q_1 q_1' &= \eta_i \mathbf{w}_i' \mathbf{w}_i \\
q_1 q_2' &= -\eta_i \mathbf{w}_i' \mathrm{x}_i \frac{\phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]} \\
q_1 q_3' &= -\eta_i \mathbf{w}_i' \mathrm{x}_i \tilde{x}_i \phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]} \\
q_2 q_1' &= \frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]} \phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \frac{\mathrm{x}_i'}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \mathbf{w}_i \eta_i \\
q_2 q_2' &= \left[\frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]}\right]^2 \left[\frac{\phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right]^2 \mathrm{x}_i' \mathrm{x}_i \\
q_2 q_3' &= \left[\frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]}\right]^2 \frac{\left[\phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]^2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \tilde{x}_i \mathrm{x}_i' \mathrm{x}_i \\
q_3 q_1' &= \left[\frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]}\right] \phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right) \tilde{x}_i \mathrm{x}_i' \mathbf{w}_i \eta_i \\
q_3 q_2' &= \left[\frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]}\right]^2 \frac{\left[\phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]^2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})} \tilde{x}_i \mathrm{x}_i' \mathrm{x}_i \\
q_3 q_3' &= \left[\frac{y_{2i} - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)}{\left[\Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]\left[1 - \Phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]}\right]^2 \left[\phi\left(\frac{\mathrm{x}_i \boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i, \boldsymbol{\Pi})}\right)\right]^2 \tilde{x}_i^2 \mathrm{x}_i' \mathrm{x}_i
\end{aligned}
$$

2.6.2 Holder class

Assume that $\mathfrak{X}$ is the support of x and is compact.[3] Denote $\mathrm{x} \equiv (x_1, ..., x_K)$, where the dimension $dim(\mathrm{x}) = K$,. Let $0 < \gamma \leq 1$. Further define $|\mathrm{x}|_e \equiv \sum_{k=1}^{K} (x_k^2)^{1/2}$ to be the Euclidean Norm. Next,

[3] We suppress the subscript i to simplify the notations.

we define $\mathfrak{f}: \mathfrak{X} \to \mathscr{R}$ to be a real valued function.

- *Hölder Condition*: $\mathfrak{f}$ are said to satisfy a Hölder condition with exponent γ if

$$\exists c > 0 \text{ such that } \forall \{\mathrm{x}, \mathbf{y}\} \in \mathfrak{X}, |\mathfrak{f}(\mathrm{x}) - \mathfrak{f}(\mathbf{y})| \leq c|\mathrm{x} - \mathbf{y}|_e^{\gamma}$$

Next, let $\alpha \equiv (\alpha_1, ..., \alpha_K), [\alpha] \equiv \alpha_1 + ... + \alpha_K$. A *Differential Operator* is defined as $\mathfrak{D}^{\alpha} \equiv \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} ... \partial x_K^{\alpha_K}}$. Next, let m be a non-negative integer and let $p = m + \gamma$.

- $\mathfrak{f}$ is said to be *p- smooth* if *(a)* $\mathfrak{f}$ is m times continuously differentiable, *(b)*, and *(c)* $\mathfrak{D}^{\alpha}\mathfrak{f}$ satisfies Hölder condition for exponent γ for all α with $[\alpha] = m$

- *Hölder Ball*: Define Hölder class to be the class of *p -smooth* functions $\mathfrak{f}$, denoted by $\Lambda^p(\mathfrak{X})$. Define $\mathfrak{C}^m(\mathfrak{X})$ to be a space of *m- times continuously differentiable functions* $\mathfrak{f}$. Hölder ball with smoothness $p = m + \gamma$ is defined as:

$$\Lambda_c^p(\mathfrak{X}) = \Bigg[\mathfrak{f} \in \mathfrak{C}^m(\mathfrak{X}) : sup_{[\alpha] \leq m} \, sup_{\mathrm{x} \in \mathfrak{X}} |\mathfrak{D}^{\alpha}\mathfrak{f}(\mathrm{x})| \leq c; \\ sup_{[\alpha]=m} \, sup_{\mathrm{x}, \mathbf{y} \in \mathfrak{X}, \mathrm{x} \neq \mathbf{y}} \frac{|\mathfrak{D}^{\alpha}\mathfrak{f}(\mathrm{x}) - \mathfrak{D}^{\alpha}\mathfrak{f}(\mathrm{x})|}{|\mathrm{x} - \mathbf{y}|_e^{\gamma}} \leq c \Bigg]$$

The key assumption in our model is that we need to assume that the unknown functions that need to be estimated belong to the Hölder class. Formally stating, we assume that

$$\mathfrak{h} \in \Lambda_c^p(\mathfrak{X}) \tag{2.65}$$

2.7 Application of the Estimation Strategy to other econometric models

The estimation strategy of this paper can be applied to the other econometric models as well. In this paper, we describe how our estimation strategy can be extended to three econometric models. The first model that we consider incorporates individual specific slope in the outcome equation along with heteroskedastic reduced form model for the treatment variable. The second model that we consider is a standard sample selection model, with heteroskedastic errors.

Finally, we consider the econometric model in which the primary equation has a binary endogenous variable, that is has a heteroskedastic reduced form. We show how our estimation strategy, both parameteric and sieve estimation can be extended for the estimation of parameters of these models.

2.7.1 Heterogeneous Coefficients Model

The first model that we consider is an endogenous switching model with heterogeneous coefficients in the primary equation for the outcome variable. As before, we also allow the reduced form equation of the treatment variable to have conditional heteroskedastic errors. Allowing individual specific slopes in the primary outcome equation allows us to capture heterogeneity in the treatment effects. However, individual-specific slopes also become another source that contributes to the endogeneity of the treatment. As we will see, conditional linear predictors allow us to model this endogeneity in a similar way as before. This allows to incorporate substantial amount of heterogeneity in our econometric model.

Model: We define the primary outcome equation for each $i = 1, ..., N$ as:

Assumption 2.7.1

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{b}_{0i} + y_{2i}\mathrm{x}_{1i}\boldsymbol{b}_{1i} + v_{0i} + y_{2i}v_{1i} \tag{2.66}$$

Where $\boldsymbol{b}_{gi} \equiv \begin{pmatrix} b_{1,gi} & \dots & b_{K_1,gi} \end{pmatrix}'$ denote the individual specific slopes in each regime $g \in \{0,1\}$. The primary outcome equation is derived in the similar manner as before using the counterfactual framework but with individual specific slopes:

$$\begin{aligned} y_{1i}^{(0)} &= \mathrm{x}_{1i}\boldsymbol{\gamma}_i^{(0)} + u_i^{(0)} \\ y_{1i}^{(1)} &= \mathrm{x}_{1i}\boldsymbol{\gamma}_i^{(1)} + u_i^{(1)} \\ y_{1i} &= (1 - y_{2i})y_{1i}^{(0)} + y_{2i}y_{1i}^{(1)} \end{aligned}$$

In the random coefficient framework, our primary parameters of interest are the average population effects denoted by $\boldsymbol{\beta}_g \equiv \mathbb{E}[\mathbf{b}_{gi}]$. Thus we have $\boldsymbol{\beta}_g = \mathbf{b}_{gi} + \boldsymbol{d}_{gi}$ where $\boldsymbol{d}_{gi} \equiv$

$\begin{pmatrix} d_{1,gi} & \dots & d_{K_1,gi} \end{pmatrix}'$ with $\mathbb{E}[\boldsymbol{d}_{gi}] = 0$ for $g = \{0,1\}$ by construction. Substituting, the outcome equation (43) becomes:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\boldsymbol{d}_{0i} + v_{0i} + y_{2i}\mathrm{x}_{1i}\boldsymbol{d}_{1i} + y_{2i}v_{1i} \tag{2.67}$$

The switching indicator y_{2i} is a binary variable modeled as before in Assumption 2.2.

Estimation Equation: To obtain the estimating equation, we first write:

$$\begin{aligned}\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\mathbb{E}[\boldsymbol{d}_{0i}|\mathrm{x}_i, v_{2i}] \\ + y_{2i}\mathrm{x}_{1i}\mathbb{E}[\boldsymbol{d}_{1i}|\mathrm{x}_i, v_{2i}] + \mathbb{E}[v_{0i}|\mathrm{x}_i, v_{2i}] + y_{2i}\mathbb{E}[v_{1i}|\mathrm{x}_i, v_{2i}]\end{aligned} \tag{2.68}$$

The expressions for $\mathbb{E}[v_{0i}|\mathrm{x}_i, v_{2i}]$ and $\mathbb{E}[v_{1i}|\mathrm{x}_i, u_{2i}]$ can be obtain from Assumption 2.2.3. In this estimating equation, we have two additional terms. These additional terms stem from the correlation of the heterogeneous coefficients and the endogenous treatment. These terms are $\mathbb{E}[\boldsymbol{d}_{0i}|\mathrm{x}_i, v_{2i}]$ and $\mathbb{E}[\boldsymbol{d}_{1i}|\mathrm{x}_i, v_{2i}]$. To obtain the appropriate expressions for these additonal *correction terms,* we extend the CLP approximation to

Assumption 2.7.2 *For $i = 1, \dots N$ and for $g = \{0,1\}$,*

$$\mathbb{E}[\boldsymbol{d}_{gi}|v_{2i}, \mathrm{x}_i] = \mathbb{L}[\boldsymbol{d}_{gi}|v_{2i}, \mathrm{x}_i] = \left(\frac{\boldsymbol{\sigma}_{\boldsymbol{d}g2}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right) v_{2i} \tag{2.69}$$

where

$$\boldsymbol{\sigma}_{\boldsymbol{d}g2}(\mathrm{x}_i) \equiv \begin{pmatrix} \sigma_{d_{g1}2}(\mathrm{x}_i) & \sigma_{d_{g2}2}(\mathrm{x}_i) & \dots & \sigma_{d_{gK_1}2}(\mathrm{x}_i) \end{pmatrix}' \quad \text{and} \quad \sigma_{d_{gj}2}(\mathrm{x}_i) \equiv Cov(d_{gi,j}, v_{2i}|\mathrm{x}_i) \tag{2.70}$$

for $g = \{0,1\}$ and $j = 1, 2, \dots, K_1$. Substituting the expressions for the *correction terms* we get:

$$\begin{aligned}\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_0 \\ + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\left(\frac{\boldsymbol{\sigma}_{\boldsymbol{d}_0 2}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right) v_{2i} + y_{2i}\mathrm{x}_{1i}\left(\frac{\boldsymbol{\sigma}_{\boldsymbol{d}_1 2}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right) v_{2i} + \left(\frac{\sigma_{02}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right) v_{2i} \\ + y_{2i}\left(\frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}\right) v_{2i}\end{aligned} \tag{2.71}$$

Next,using the similar calculations as in section 2.3, we get:

$$\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\left(\frac{\boldsymbol{\sigma}_{\boldsymbol{d}_0 2}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)h(y_{2i},\mathrm{x}_i) + y_{2i}\mathrm{x}_{1i}\left(\frac{\boldsymbol{\sigma}_{\boldsymbol{d}_1 2}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)h(y_{2i},\mathrm{x}_i) + \left(\frac{\sigma_{02}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)h(y_{2i},\mathrm{x}_i) + y_{2i}\left(\frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)h(y_{2i},\mathrm{x}_i)$$

where $h(y_{2i},\mathrm{x}_i)$ is the expression for generalized residuals as defined previously. Next, define:

$$\mathfrak{f}_{\boldsymbol{d}_g}(\mathrm{x}_i) \equiv \begin{pmatrix} \mathfrak{f}_{\boldsymbol{d}_{g1}}(\mathrm{x}_i) \\ \dots \\ \mathfrak{f}_{\boldsymbol{d}_{gK_1}}(\mathrm{x}_i) \end{pmatrix} \equiv \begin{pmatrix} \frac{\sigma_{d_{g1}2}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)} \\ \vdots \\ \frac{\sigma_{d_{gK_1}2}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)} \end{pmatrix} \text{ for } g = \{0,1\}$$

$$\mathfrak{f}_{vg} \equiv \left(\frac{\sigma_{g2}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right) \text{ for } g = \{0,1\}$$

Using this notation, we get our final estimating equation as:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\mathfrak{f}_{\boldsymbol{d}_0}(\mathrm{x}_i)h(y_{2i},\mathrm{x}_i) + y_{2i}\mathrm{x}_{1i}\mathfrak{f}_{\boldsymbol{d}_1}(\mathrm{x}_i)h(y_{2i},\mathrm{x}_i) + \mathfrak{f}_{v_0}(\mathrm{x}_i)h(y_{2i},\mathrm{x}_i) + y_{2i}\mathfrak{f}_{v_1}(\mathrm{x}_i)h(y_{2i},\mathrm{x}_i) + \eta_i \quad (2.72)$$

where $\mathbb{E}[\eta_i|\mathrm{x}_i, y_{2i}] = 0$ by construction.

Parametric Estimation: The estimation will be done in two steps as before with the first step being a *hetprobit* with Assumption 3.1. We will obtain the estimator for $h(y_{2i},\mathrm{x}_i)$. In the second step, we will have to specify the functional forms for the additional terms $\mathfrak{f}_{\boldsymbol{d}_0}(\mathrm{x}_i)$ and $\mathfrak{f}_{\boldsymbol{d}_1}(\mathrm{x}_i)$. This in turn means that we would be specifying the functional forms for $\mathfrak{f}_{\boldsymbol{d}_{gj}}(\mathrm{x}_i)$ for each $g = \{0,1\}$ and for all $j = \{1,...,K_1\}$. The simplest functional form is a linear in parameter specification:

Assumption 2.7.3 *For all $i = 1,...,N$, for each $g = \{0,1\}$, for all $j = 1,...,K_1$ assume*

$$\mathfrak{f}_{\boldsymbol{d}_{gj}}(\mathrm{x}_i) = \mathrm{x}_i\boldsymbol{\omega}_{d_{gj}} \quad (2.73)$$

where each $\boldsymbol{\omega}_{d_{gj}}$ is a $K \times 1$ vector of coefficients. This implies:

$$\mathfrak{f}_{\boldsymbol{d}_g}(\mathrm{x}_i) = \mathrm{X}_i \boldsymbol{\Omega}_{d_g}, \text{ where } \mathrm{X}_i \equiv \begin{pmatrix} \mathrm{x}_i & \mathbf{0} & \dots & \dots \\ \mathbf{0} & \mathrm{x}_i & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathrm{x}_i \end{pmatrix} \text{ and } \boldsymbol{\Omega}_{d_g} \equiv \begin{pmatrix} \boldsymbol{\omega}_{d_{g1}} \\ \boldsymbol{\omega}_{d_{g2}} \\ \vdots \\ \boldsymbol{\omega}_{d_{gK_1}} \end{pmatrix}$$

Substituting the functional forms of the additional terms with the terms defined in Assumption 3.1 and the estimators from the first step, we have the following parametric estimating equation:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\mathrm{X}_i\boldsymbol{\Omega}_{d_0}\hat{h}_i + y_{2i}\mathrm{x}_{1i}\mathrm{X}_i\boldsymbol{\Omega}_{d_1}\hat{h}_i + \hat{h}_i\mathrm{x}_i\boldsymbol{\Omega}_{01} + y_{2i}\hat{h}_i\mathrm{x}_i\boldsymbol{\Omega}_{01} + \eta_i \tag{2.74}$$

In the second step, equation (51) is then estimated using least squares.

Sieve Estimation: The first step of sieve estimation also follows the same procedure as in section 2.2 with defining the sieve space as $\mathscr{A}_{2N}$ and then estimating the result MLE. In the second stage, we extend the sieve space to accommodate the additional functional forms arising due to the presence of individual specific slopes.The sieve spaces for $\{\mathfrak{f}_{v_0}(\mathrm{x}_i), \mathfrak{f}_{v_1}(\mathrm{x}_i)\}$ would be similar to those defined for $\{\mathfrak{g}_0(\mathrm{x}_i), \mathfrak{g}_1(\mathrm{x}_i)\}$. The sieve spaces for $\{\mathfrak{f}_{\boldsymbol{d}_0}(\mathrm{x}_i), \mathfrak{f}_{\boldsymbol{d}_1}(\mathrm{x}_i)\}$ would also be straightforward, but a little more notationally involved.

In particular, first define sieve spaces for each component of $\mathfrak{f}_{\boldsymbol{d}_g}(\mathrm{x}_i)$:

$$\left\{ \mathfrak{f}_{\boldsymbol{d}_{gj}}(\mathrm{x}_i) = \mathtt{D}_{gj}(\mathrm{x}_i)\boldsymbol{\omega}_{N,d_{gj}} \equiv \mathtt{D}_g^{K_N^{\mathtt{d}_{gj}}}(\mathrm{x}_i)\boldsymbol{\omega}_{N,d_{gj}} : \boldsymbol{\omega}_{N,d_{gj}} \in \mathbb{R}^{K_N^{\mathtt{d}_{gj}}} \right\} \tag{2.75}$$

for each $j = \{1, ..., K_1\}$ and for $g = \{0, 1\}$. $\mathtt{D}_g^{K_N^{\mathtt{d}_{gj}}}(.)$ is a $1 \times K_N^{\mathtt{d}_{gj}}$ vector of basis functions. Next, let $K_N^{\boldsymbol{d}_g} \equiv (K_N^{\mathtt{d}_{g1}} + K_N^{\mathtt{d}_{g2}} + ... + K_N^{\mathtt{d}_{gK_1}})$ and define:

$$\mathtt{D}_g(\mathrm{x}_i) \equiv \begin{pmatrix} \mathtt{D}_{g1}(\mathrm{x}_i) & \mathbf{0} & \dots & \dots \\ \mathbf{0} & \mathtt{D}_{g2}(\mathrm{x}_i) & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathtt{D}_{gK_1}(\mathrm{x}_i) \end{pmatrix}_{(K_1 \times K_N^{\boldsymbol{d}_g})} \text{ and } \boldsymbol{\Omega}_{N,d_g} \equiv \begin{pmatrix} \boldsymbol{\omega}_{N,d_{g1}} \\ \boldsymbol{\omega}_{N,d_{g2}} \\ \vdots \\ \boldsymbol{\omega}_{N,d_{gK_1}} \end{pmatrix}_{(K_N^{\boldsymbol{d}_g} \times 1)} \tag{2.76}$$

We have the sieve space for $\mathfrak{f}_{\boldsymbol{d}_g}(\mathrm{x}_i)$ for $g = \{0,1\}$,

$$\left\{ \mathfrak{f}_{\boldsymbol{d}_g}(\mathrm{x}_i) = \mathtt{D}_g(\mathrm{x}_i)\boldsymbol{\Omega}_{N,d_g} : \boldsymbol{\Omega}_{N,d_g} \in \mathbb{R}^{K_N^{\boldsymbol{d}_g}} \right\} \tag{2.77}$$

In particular, our sieve space in the second stage for the parameters $\{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \mathfrak{f}_{\boldsymbol{d}_0}(\mathrm{x}_i), \mathfrak{f}_{\boldsymbol{d}_1}(\mathrm{x}_i), \mathfrak{f}_{v_0}(\mathrm{x}_i), \mathfrak{f}_{v_1}(\mathrm{x}_i)\}$ is now defined as:

$$\begin{aligned} \mathscr{A}_{1N} = \mathscr{B}_1 \mathscr{X} & \left\{ \mathfrak{f}_{v_0}(.) = \mathtt{G}_0^{K_N^{v_0}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02} : \boldsymbol{\Omega}_{N,02} \in \mathbb{R}^{K_N^{v_0}} \right\} \mathscr{X} \\ & \left\{ \mathfrak{f}_{v_1}(.) = \mathtt{G}_1^{K_N^{v_1}}(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12} : \boldsymbol{\Omega}_{N,12} \in \mathbb{R}^{K_N^{v_1}} \right\} \\ \mathscr{X} & \left\{ \mathfrak{f}_{\boldsymbol{d}_0}(\mathrm{x}_i) = \mathtt{D}_0(\mathrm{x}_i)\boldsymbol{\Omega}_{N,d_0} : \boldsymbol{\Omega}_{N,d_0} \in \mathbb{R}^{K_N^{\boldsymbol{d}_1}} \right\} \mathscr{X} \\ & \left\{ \mathfrak{f}_{\boldsymbol{d}_1}(\mathrm{x}_i) = \mathtt{D}_1(\mathrm{x}_i)\boldsymbol{\Omega}_{N,d_1} : \boldsymbol{\Omega}_{N,d_1} \in \mathbb{R}^{K_N^{\boldsymbol{d}_1}} \right\} \end{aligned} \tag{2.78}$$

where $\mathtt{G}_0^{K_N^{v_0}}(\mathrm{x}_i) \equiv \left\{ \mathrm{g}_0^1(.), ..., \mathrm{g}_0^{K_N^{v_0}}(.) \right\}$ is a $1 \times K_N^{v_0}$ vector of basis functions , $\mathtt{G}_1^{K_N^{v_1}}(\mathrm{x}_i) \equiv \left\{ \mathrm{g}_1^1(.), ..., \mathrm{g}_0^{K_N^{v_1}}(.) \right\}$ is a $1 \times K_N^{v_1}$ vector of basis functions. Plugging in equation(49) and suppressing the superscripts:

$$\begin{aligned} y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_0 + y_{2i}\mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathrm{x}_{1i}\mathtt{D}_0(\mathrm{x}_i)\boldsymbol{\Omega}_{N,d_0}\hat{h}_i + y_{2i}\mathrm{x}_{1i}\mathtt{D}_1(\mathrm{x}_i)\boldsymbol{\Omega}_{N,d_1}\hat{h}_i \\ + \mathtt{G}_0(\mathrm{x}_i)\boldsymbol{\Omega}_{N,02}\hat{h}_i + y_{2i}\mathtt{G}_1(\mathrm{x}_i)\boldsymbol{\Omega}_{N,12}\hat{h}_i + \eta_i \end{aligned} \tag{2.79}$$

As suggested by equation (56), in the second step, we obtain the estimators by running regression y_{1i} on $\mathrm{x}_{1i}, y_{2i}\mathrm{x}_{1i}, \mathrm{x}_{1i}\mathtt{D}_0(\mathrm{x}_i)\hat{h}_i, y_{2i}\mathrm{x}_{1i}\mathtt{D}_1(\mathrm{x}_i)\hat{h}_i, \mathtt{G}_0(\mathrm{x}_i)\hat{h}_i, y_{2i}\mathtt{G}_1(\mathrm{x}_i)\hat{h}_i$

2.7.2 Binary Endogenous Variable

The next model that we consider are the models with a binary endogenous variable. Adaptation of the estimating strategy helps to incorporate heterogeneity in the reduced form by allowing heteroskedastic errors.

Model: The model comprises of a primary equation for the outcome variable y_{1i} that is assumed to be linear in parameters:

Assumption 2.7.4 *For all* $i = 1, ..., N$

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + v_{1i} \tag{2.80}$$

where y_{2i} is a binary endogenous variable. x_{1i} are the exogenous variables. We also have a linear reduced form model for y_{2i} given as:

Assumption 2.7.5

$$y_{2i} = \mathbf{1}[\mathrm{x}_i\boldsymbol{\beta}_2 - v_{2i} < 0], \tag{2.81}$$

where x_i is a vector of all the exogenous variables that includes x_{1i}. The formal version for the exogeneity is given as:

Assumption 2.7.6

$$\mathbb{E}[v_{1i}|\mathrm{x}_i] = 0 \tag{2.82}$$

The next assumption defines the heteroskedastic structure to the errors in the reduced form:

$$v_{2i} \equiv \sigma_2(\mathrm{x}_i)u_{2i} \tag{2.83}$$

$$u_{2i} \perp\!\!\!\perp \mathrm{x}_i,\ u_{2i} \sim \mathcal{N}(0,1) \tag{2.84}$$

Finally, we use conditional linear predictors to model the endogeneity of y_{2i} through control function methods:

Assumption 2.7.7

$$\mathbb{E}[v_{1i}|v_{2i};\mathrm{x}_i] = \mathbb{L}[v_{1i}|v_{2i};\mathrm{x}_i] = \frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)}v_{2i} \tag{2.85}$$

This Assumption helps us to obtain the correction term that we need to obtain the estimating equation.

Estimating Equation: To obtain the estimating equation, we follow the similar algorithm as given in section 2.2:

$$\begin{aligned}\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i}] &= \mathbb{E}[\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}]|\mathrm{x}_i, y_{2i}] = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + \mathbb{E}[\mathbb{E}[v_{1i}|\mathrm{x}_i, v_{2i}]|\mathrm{x}_i, y_{2i}] \\ &= \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + \underbrace{\left(\frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)}_{\equiv \mathfrak{g}(\mathrm{x}_i)} h(\mathrm{x}_i, y_{2i}) \\ \mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i}] &= \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + \mathfrak{g}(\mathrm{x}_i)h(\mathrm{x}_i, y_{2i})\end{aligned}$$

where $h(\mathrm{x}_i, y_{2i})$ is the generalized residuals from the first stage estimation of the reduced form model for y_{2i}. Thus we get our estimating equation as:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + \mathfrak{g}(\mathrm{x}_i)h(y_{2i}, \mathrm{x}_i) + \eta_i \tag{2.86}$$

with $\mathbb{E}[\eta_i|\mathrm{x}_i, y_{2i}] = 0$ by construction. Estimation of equation (2.63) is analogous to the two step estimation described in the previous sections. We can either impose the parametric Assumptions by specifying the functional forms of the $\sigma_2(\mathrm{x}_i)$ and $\mathfrak{g}(\mathrm{x}_i)$ or can perform the sieve estimation in both the steps.

Parametric Estimation: In the first step we define again specify the functional form for $\sigma_2(\mathrm{x}_i)$ as in Assumption 3.3 and will run the *hetprobit* and obtain the estimator for the generalized residuals. The second step involves specifying the functional form for the $\mathfrak{g}(\mathrm{x}_i)$:

Assumption 2.7.8 *Assume that for all $i = 1, ..., N$,*

$$\mathfrak{g}(\mathrm{x}_i) = \mathrm{x}_i\boldsymbol{\Omega}_{12} \tag{2.87}$$

Next we just substitute the functional form and the estimators from the first step and obtain:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + \mathrm{x}_i\boldsymbol{\Omega}_{12}\hat{h}_i + \eta_i \tag{2.88}$$

that is solved by regressing y_{1i} on x_{1i}, y_{2i}, $\hat{h}_i\mathrm{x}_i$.

Sieve Estimation: The firs step in sieve estimation again begins with define the sieve space $\mathscr{A}_{2N}$ as done previously and running the resulting *hetprobit* In the second step, the sieve

space is defined as $\mathscr{A}_{1N} = \mathscr{B}_1 \times \left\{ \mathfrak{g}(\mathrm{x}_i) = \mathtt{G}^{K_N^{\mathfrak{g}}}(\mathrm{x}_i)\boldsymbol{\Omega}_{12,N} : \boldsymbol{\Omega}_{12,N} \in \mathbb{R}^{K_N^{\mathfrak{g}}} \right\}$ where $\mathtt{G}^{K_N^{\mathfrak{g}}}(\mathrm{x}_i) \equiv \{\mathrm{g}^1(.), ..., \mathrm{g}^{K_N^{\mathfrak{g}}}(.)\}$ is $1 \times K_N^{\mathfrak{g}}$ vector of basis functions. Substituting the first stage estimators and the basis functions, we get

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + y_{2i}\beta_y + \mathtt{G}^{K_N^{\mathfrak{g}}}(\mathrm{x}_i)\boldsymbol{\Omega}_{12}\hat{h}_i + \eta_i \tag{2.89}$$

and the parameters are obtained by regressing y_{1i} on x_{1i}, y_{2i}, $\hat{h}_i\mathtt{G}^{K_N^{\mathfrak{g}}}(\mathrm{x}_i)$.

2.7.3 Sample Selection Model

Finally we extend our estimation strategy to sample selection models with heteroskedasticity. Sample selection models are one of the leading applications for the semi and nonparameteric estimation techniques. However, as Vella(1998) notes, heteroskedasticity in sample selection models still remains an unexplored issue. Since heteroskedasticity in sample selection models manifests itself in the model for the selection indicator that is usually non-linear, it creates inconsistency. On the other hand, identifying and correcting the issues arising due to heteroskedasticity require restrictive distribution assumptions (Vella(1998)).

Heteroskedastic sample selection models have been previously addressed by Donald(1995) and Chen and Khan(2003). Donald(1995) permits heteroskedasticity of an unknown form and maintains the assumption of bivariate normality of the errors. He suggests a two-step estimation procedure that allows a non-parametric estimation of the model. However, as we will see, his method becomes complicated due to necessity of a trimming rule. Chen and Khan(2003) propose a three-step estimator that corrects for heteroskedastic sample selection bias. Their estimation procedure involves nonparametric estimation of propensity scores in the first step and nonparametric quantile estimation of the conditional interquartile range of the outcome equation dependent variable for the selected observations in the second step. The first two steps yield a reweighted outcome equation that is partially linear that is analogous Donald's(1995) and is thus estimated using the appropriate methods. While innovative in nature, their three-step estimation

strategy requires multiple smoothing parameters, including those for the quantile estimation in the second step and a trimming function in the third step.

The estimation procedure suggested in this section, addresses these issues by proposing a simple two-step estimation that corrects for the sample selection bias and is straightforward to apply. We incorporates heteroskedasticity in the sample selection models that could either be given a functional form in the parametric estimation or could be of an unknown form in the sieve estimation. As before, we model the endogeniety arising in the model due to the sample selection bias using the conditional linear predictors.

Model: The model comprises of two latent variables: a continuous variable y_{1i}^* that yields the outcome variable and a latent variable y_{2i}^* that yields the binary selection indicator. Both latent variables are assumed to be linear in parameters:

Assumption 2.7.9 *For all* $i = 1, ..., N$

$$y_{1i}^* = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + v_{1i} \tag{2.90}$$

$$y_{2i}^* = \mathrm{x}_i\boldsymbol{\beta}_1 - v_{2i},\ \mathrm{x}_{1i} \subset \mathrm{x}_i \tag{2.91}$$

$\Longrightarrow$

$$y_{1i} = y_{2i} \bullet y_{1i}^* \tag{2.92}$$

$$y_{2i} = \mathbf{1}[\mathrm{x}_i\boldsymbol{\beta}_2 - v_{2i} < 0] \tag{2.93}$$

In other words,

$$y_{2i} = \begin{cases} 1 & \text{if and only if } \{\mathrm{x}_i, y_{1i}\}, \text{ is fully observed} \\ 0, & \text{otherwise} \end{cases} \tag{2.94}$$

We incorporate conditional heteroskedasticity of the multiplicative form:

Assumption 2.7.10

$$v_{2i} \equiv \sigma_2(\mathrm{x}_i)u_{2i},\ u_{2i} \perp\!\!\!\perp \mathrm{x}_i,\ u_{2i} \sim \mathcal{N}(0,1) \tag{2.95}$$

The endogeneity of sample selection is reflected in the relationship between the latent errors of the outcome equation and the selection equation. The relationship is modeled using conditional linear predictors that restricts the linearity only in terms of v_{2i} which allowing x_i to enter in an unspecified functional form.

Assumption 2.7.11

$$\mathbb{E}[v_{1i}|v_{2i};\mathrm{x}_i] = \mathbb{L}[v_{1i}|v_{2i};\mathrm{x}_i] = \frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2^2(\mathrm{x}_i)} v_{2i} \tag{2.96}$$

where y_{2i} is now the selection indicator. The nonrandom selection is reflected in equation (53).

Estimating Equation: To obtain the estimating equation, first note that since y_{1i} is observed only when $y_{2i} = 1$, our key equation in sample selection models is $\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i} = 1]$. Thus our algorithm changes slightly to:

$$\begin{aligned}
\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i} = 1] &= \mathbb{E}[\mathbb{E}[y_{1i}|\mathrm{x}_i, v_{2i}]|\mathrm{x}_i, y_{2i} = 1] \\
&= \mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathbb{E}[\mathbb{E}[v_{1i}|\mathrm{x}_i, v_{2i}]|\mathrm{x}_i, y_{2i} = 1] \\
&= \mathrm{x}_{1i}\boldsymbol{\beta}_1 + \underbrace{\left(\frac{\sigma_{12}(\mathrm{x}_i)}{\sigma_2(\mathrm{x}_i)}\right)}_{\equiv \mathfrak{g}(\mathrm{x}_i)} \lambda\left(\frac{\mathrm{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right) \\
\mathbb{E}[y_{1i}|\mathrm{x}_i, y_{2i} = 1] &= \mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathfrak{g}(\mathrm{x}_i)\lambda\left(\frac{\mathrm{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right)
\end{aligned}$$

with $\lambda(.)$ being the Inverse Mills Ratio.

Thus we get our estimating equation as:

$$y_{1i} = \mathrm{x}_{1i}\boldsymbol{\beta}_1 + \mathfrak{g}(\mathrm{x}_i)\lambda\left(\frac{\mathrm{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right) + \eta_i \tag{2.97}$$

with $\mathbb{E}[\eta_i|\mathrm{x}_i, y_{2i} = 1] = 0$ by construction.

Estimating Procedure: The estimating strategy is exactly as before for both parametric and sieve estimation with two main differences. First, in the first step we need to obtain only the estimator for $\lambda\left(\frac{\mathrm{x}_i\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right) \equiv \hat{\lambda}_i$ and not the full generalized residual. Secondly, in the second stage, we will run the regression only using the observations with $y_{2i} = 1$ for obvious reasons. To give

some context to our estimation strategy, we compare it with that proposed by Donald(1995). To obtain the estimation procedure, Donald(1995) divides equation (74) by $\lambda_i \equiv \lambda\left(\frac{\mathrm{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2(\mathrm{x}_i)}\right)$ that yields:

$$\frac{y_{1i}}{\lambda_i} = \frac{\mathrm{x}_i}{\lambda_i}\boldsymbol{\beta}_1 + \mathfrak{g}(\mathrm{x}_i) + \frac{\eta_i}{\lambda_i}$$

Replacing λ_i with $\hat{\lambda}_i$, yields a partially linear model that is then and then estimated using a differencing procedure analogous to Robinson(1988). In Donald(1995) the first stage estimation is done using nonparametric methods and λ_i is constructed by inverting the standard normal density function that is only defined if the arguments of the density function are between zero and one. In addition, $\hat{\lambda}_i$ the enters the estimating equation in the denominator. As Donald(1995) notes, this requires a trimming function that complicates the estimation. In the semi-nonparametric estimation procedure suggested in this paper requires no such methods and directly estimates the unknown function in the estimating equation using sieve methods.

2.8 Concluding Remarks

This paper proposes a two-step sieve control function estimation of endogenous switching models. We allow the probit reduced form model for the binary switching indicator to incorporate heterogeniety by allowing for a conditional heteroskedastic error term. Allowing the heteroskedastic function to be of an unknown form, we also relax the distributional assumption in the reduced form. Conditional Linear Projections help us to obtain the corrections terms in the estimating equation in the environment of conditional heteroskedasticity and unspecified distributional assumptions. Though the focus of the paper is the sieve estimation, we also suggest how one can still incorporate an heteroskedastic reduced form and estimate the model parametrically. We also extend the model to the endogenous switching models with individual specific slopes in the outcome equation, linear models with a binary endogenous variable and sample selection models. The extension to the sample selection model serves as an important

contribution to the limited literature of sample selection models with heteroskedasticity.

CHAPTER 3

CONTROL FUNCTION ESTIMATION OF SPATIAL ERROR MODELS WITH ENDOGENEITY

3.1 Introduction

Cross-sectional dependence creates an interesting and challenging environment for estimation and inference in applied econometrics. In the context of time-series data, we have several tractable procedures for estimation with correlated observations because of the uni-directional nature of dependence that renders a natural ordering to the data: outcomes in the past can affect the outcomes in the future but not vice-versa. In the cross-sectional context, on the other hand, the lack of such natural ordering and structure restricts the use of general forms of dependence that are routinely allowed in time-series data.

Allowing the observations to exhibit spatial dependence is an important way to model cross-sectional dependence among economic agents. In the framework of spatial dependence, the individuals are interdependent due to their locations in an Euclidean space and their proximity with each other other. This proximity is measured as the *economic distance* between the individuals as described in Conley (1999) that can be defined within the framework of the specific empirical application.

In this paper we consider the $\mathbb{R}^2$ Euclidean space for the sake of exposition. Empirical questions in the field of regional and urban economics, agricultural and environmental economics, industrial organization and as well as public health and epidemiology- all are concerned with data that exhibits significant spatial dependence due to the geographical location of the observations. Specifically, consider the hedonic price models in the housing markets. Hedonic price models are regression models where the price of a commodity is regressed on its attributes. In housing markets, the price of a house depends not only on its own attributes such as floor plan etc but also on the *location* of the house. The spatial dependence is exhibited in the un-

observed (or poorly measured) neighborhood variables. These spatial omitted neighborhood variables enter the unobservables of the regression model inducing spatial dependence in the errors. This serves as the key motivation behind **Spatial Error Models (SEM)** that provide one framework to capture spatial dependence in the data. Spatial dependence can also be modeled in the framework of **Spatial Lag Models** were the outcome variables are assumed to exhibit spatial correlation.

In this paper, we consider the **Spatial Error Models**. Thus, we posit that the spatial dependence in the data is captured in the unobservables of the models. This entails having a non-spherical variance-covariance matrix of the errors. The spatial dependence is accounted for in estimation of SEMs by incorporating the correlations between the observations in the estimation procedure. This entails accounting for all the pairwise correlations of the spatial data in a framework similar to *Generalized Least Squares (GLS)* estimation. However, in a large sample we have a huge error variance-covariance matrix and this makes the efficient *GLS-type* estimation procedure very cumbersome to implement.

In this paper, we contribute to the literature by obtaining an estimation procedure for linear regression models with spatially correlated errors and endogenous variables. Our estimation procedure achieves efficiency gains by taking account of the spatial correlations between the observations. We provide a computationally simple estimation procedure by dividing the data into groups based on the distances between them and then accounting for only the correlation between the observations *within* a group while ignoring the correlations between the observations *between* the groups. The intuition is based on The First Law of Geography, according to Tobler (1970): *"everything is related to everything else, but near things are more related than distant things"*. Since correlations *within* a group accounts for the most correlation in the data, we get significant efficiency gains while simultaneously avoiding the tedious calculations of the traditional *GLS* estimation with the variance-covariance matrix for the entire data. This is also motivated by the nature in which spatial data is collected: data is often collected from different

geographical regions leading to a natural *clustering* or *division* of the observations. However the estimation is different from the estimation procedures with clustered data because with spatial data, the correlations within a group are the function of the distance between the observations in a group. In addition, unlike the estimation procedures with clustered data, we do not impose the independence assumptions between different groups.

Lu and Wooldridge (2017) consider estimation of linear econometric models with spatial data in which they describe an estimation procedure by first dividing the observations into groups and then using only the *within-group* information while ignoring the *across-group* correlations. For the linear models, they obtain a *Quasi-GLS* estimation procedure that uses a *tapered* error covariance matrix as opposed to using the full error covariance matrix used in the traditional *GLS* estimation. Estimation for the non-linear models can be done in the similar way to obtain *Generalized Estimating Equations* that account for the spatial correlations in the data (Lu 2013). Wang, Iglesias and Wooldridge (2013) also suggest a similar estimation procedure for spatial Probit models in which the observations are divided by pairwise groups and then bivariate normal distributions are specified within each group.

In Lu and Wooldridge (2017) framework of estimation of models with spatial data, all the covariates are assumed to be uncorrelated with the unobservables. However, in practice, explanatory variables are often correlated with the errors leading to endogeneity that causes inconsistencies of estimators. For example, consider the hedonic housing price models where one is interested in studying the causal relationship between school quality and house prices. Nguyen-Hoang and Yinger (2011) give a detailed account of the studies that have explored the impact of school quality on housing values. In this empirical question; the school quality, often measured by some indicator of overall school achievement scores; is correlated with the omitted neighborhood variables giving rise to the problem of endogeneity. In addition,the houses located near one another to have similar unobservable attributes and this motivates the spatial error framework of spatial dependence. This serves as an important motivation behind

the design of the econometric model of this paper.

In this paper, we consider linear models with spatially correlated data and allow some covariates to be endogenous and implement control function method to correct for the endogeneity. Our estimation strategy divides the observations into groups based on the distance between them and then incorporates the correlation structure of individuals *within* a group. The group structure is also conducive to obtaining additional instruments to correct for the endogeneity. Specifically, we recognize that for each individual within a group, the exogenous variables of his/her within-group neighbors are also eligible instruments for that individual. We use these additional instruments in the traditional 2SLS estimation to obtain what we call **Grouped Two Stage Least Square** estimator. We also describe how the groupwise spatial dependence can be incorporated with these additional instruments in an a Generalized Instrument Variable framework to obtain a **Spatial Generalized Instrument Variable** estimator. Finally we describe a two step control function estimation method in which we explicitly model the endogeneity through a control function assumption that is imposed for each group. This control function assumption also incorporates the *groupwise* spatial dependence and gives us a **Spatial Control Function** estimator.

Spatial econometrics has experienced some major advancements in the asymptotic theory research. Asymptotic theory for estimation with dependent processes establishes laws of large numbers and central limit theorems by imposing some structure to the cross-sectional dependence and regularity conditions. In practice, this entails clearly defining the nature of spatial dependence and then either deriving or applying the asymptotic results. Conley (1999) establishes law of large numbers and central limit theorems for stationary mixing processes. Jenish and Prucha (2009) consider the spatial asymptotics for more general form of dependence including nonstationary mixing processes. Jenish and Prucha (2012) derive the law of large numbers and central limit theorems for *near-epoch* dependent random fields. Under this general framework, they also establish the consistency and asymptotic normality results for the *GMM* estimators. In

this paper, the asymptotic results are described under the framework of *near-epoch* dependence to give it the most general treatment.

There are two framework under which spatial asymptotic theory can be developed. Under *increasing-domain* asymptotics, we assume that the number of observations increase and the minimum distance between the observations is bounded below by a positive constant. In contrast, we have the *fixed-domain* or *infill* aymptotics, wherein, the observations become increasingly dense in a fixed and bounded region. While there is a vast literature establishing several asymptotic results of consistency and normality in the *increasing-domain* framework (Mardia and Marshall (1984), Cressie and Lahiri (1993), Lee (2004), Conley (1999)) ; under the *fixed-domain* framework, as the interactions between the observations increase with the sample size, it is known that Maximum Likelihood estimators are inconsistent (Lee (2004)). We obtain the asymptotic properties of the estimators under the *increasing-domain* framework.

For the asymptotic properties in our context, we assume that as the number of observations increases, the number of groups increases while the group size remains fixed. We further assume that the minimum distance between the observations and the groups is bounded below by a positive constant. The asymptotic properties of the **Spatial Control Function** are obtained by collecting the estimating equations and describing the two-step estimation procedure as a *one-step* estimation procedure. In addition, since we have a two-step estimation, we also obtain the asymptotic variance of the **Spatial Control Function** that corrects for the first-step estimation.

We also obtain consistent variance-covariance estimators that are made robust to the misspecification of the spatial correlation structure as well as the correlation between the observations in different groups. This accounts for *across-group* dependence between the observations that we ignore in the estimation procedure making our inference robust. We use the *HAC* estimator defined by Lu and Wooldridge (2017) that considers all the observations/groups within a fixed radius of a particular observation/group.

The paper is structured as follows. In section 2, we define a linear regression model with the

spatially dependent errors and endogenous covariates. In section 3, we obtain the estimating equations by implementing control function methods. In section 4, we describe our estimation procedure that takes account of the *within-group* correlations. In section 5, we obtain the asymptotic properties of our estimator under *near-epoch* dependence. We also provide consistent variance estimators. In section 6, we describe the design and the results of the Monte Carlo simulation studies to illustrate the small sample properties of our estimators. Finally, in section 7, we conclude and suggest possible avenues for future research.

3.2 Model

Let $\mathscr{S}$ be a two dimensional Euclidean space in which the population resides. Let s_i represent a location in $\mathscr{S}$ for $i = 1,2,....$ Denote the distance between locations $\{s_i\} \in \mathscr{S}$ and location $\{s_j\} \in \mathscr{S}$ by d_{ij}. The data points sampled at a location $s_i \in \mathscr{S}$ are denoted by $\{y_{1s_i}, y_{2s_i}, \mathbf{z}_{s_i}\}$ where $i = 1,2,...,N$. In addition, we have $\{u_{s_i}, v_{s_i}\}$ as the underlying unobservables. To make the notations simple, we will denote the index s_i by i.

3.2.1 A Linear Regression Model

The *primary* equation for our outcome variable is assumed to be linear in parameters:

Assumption 3.2.1 *We have a linear in parameter regression model for the outcome variable:*

$$y_{1i} = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 + \tau y_{2i} + u_i \tag{3.1}$$

where y_{2i} is the scalar endogenous variable and $\boldsymbol{x}_{1i}$ is a $1 \times K_{x_1}$ vector of exogenous explanatory variables with the first element being equal to unity. The vector of coefficients on $\boldsymbol{x}_{1i}$ is denoted by $\boldsymbol{\beta}_1$ which is $K_{x_1} \times 1$. τ is a scalar coefficient on the scalar endogenous variable y_{2i}.

We defined a linear in parameter reduced form model for the endogenous variable: The reduced form of the scalar endogenous variable is also assumed to be linear in parameters:

Assumption 3.2.2 *For* $i = 1, 2, ..., N$

$$y_{2i} = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2 + v_i \tag{3.2}$$

where $\boldsymbol{x}_{2i}$ *is a* $1 \times K_{x_2}$ *vector of exogenous explanatory variables that also include* x_{1i}*. We assume that* $K_{x_2} \geq K_{x_1}$ *that serves as an* ***exclusion condition needed for identification****. The vector of coefficients on* x_{2i} *is denoted by* $\boldsymbol{\beta}_2$ *which is* $K_{x_2} \times 1$*.*

In matrix form, we write our model as:

$$Y_{1N} = \boldsymbol{X}_{1N}\boldsymbol{\beta}_1 + \tau Y_{2N} + U_N \tag{3.3}$$

$$Y_{2N} = \boldsymbol{X}_{2N}\boldsymbol{\beta}_2 + V_N \tag{3.4}$$

where

$$Y_{1N} \equiv \begin{pmatrix} y_{11} & \dots & y_{1N} \end{pmatrix}' \; ; \; \boldsymbol{X}_{1N} \equiv \begin{pmatrix} \boldsymbol{x}_{11} & \dots & \boldsymbol{x}_{1N} \end{pmatrix}' \; ; \; Y_{2N} \equiv \begin{pmatrix} y_{21} & \dots & y_{2N} \end{pmatrix}' \tag{3.5}$$

$$U_N \equiv \begin{pmatrix} u_1 & \dots & u_N \end{pmatrix}' \; ; \; V_N \equiv \begin{pmatrix} v_1 & \dots & v_N \end{pmatrix}' \; ; \; \boldsymbol{X}_{2N} \equiv \begin{pmatrix} \boldsymbol{x}_{21} & \dots & \boldsymbol{x}_{2N} \end{pmatrix}' \tag{3.6}$$

We assume that the strict exogeneity of the instruments hold. The assumption for the exogenous variables is formally stated as below:

Assumption 3.2.3 *Strict Exogeneity Condition:*

$$\mathbb{E}\left[\begin{pmatrix} U_N \\ V_N \end{pmatrix} \middle| \boldsymbol{X}_{2N} \right] = \mathbb{E}\begin{pmatrix} U_N \\ V_N \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \tag{3.7}$$

3.2.2 Spatially Correlated Errors

In this paper, the spatial nature of the data is captured in the variance-covariance matrix of the error terms. In other words, we consider **Spatial Error Models (SEMs)**. As Dubin (1988) notes, SEMs are an integral part of the urban economics and they play a special role in housing

hedonic regression analysis. In housing markets, the omitted variables will have high degree of dependence because houses that are near each other will have similar *"neighborhood variables"* (Dubin (1988)). Dubin (1988) recognizes crime rates, quality of public schools , race as common *neighborhood variables* that are either unobservables and/or are difficult to accurately measure. Thus, often these variables that exhibit spatial dependence enter the error terms of the regression model motivating a SEM analyses.

Glass, Kenegalieva and Sickles (2012) also use SEM analysis the context of state vehicle usage in the US. They make two economic arguments for incorporating SEM. First, in the context of state vehicle usage in the US, they argue SEMs gives a fuller representation of spatial dependence than models which do not include a spatial autocorrelation term. Second, they illustrate how SEMs allow researchers to permit Wald tests of whole sets of coefficients against one another to ascertain if models which are estimated using disaggregated data contain more information than the aggregate model.

In our model, we have two types of errors for each individual. Each individual has a *primary* error u_i which is the unobservable term in the *primary* equation and the a *reduced-form* error v_i which is the unobservable term in the reduced form equation . Thus we focus on the variance-covariance matrix of $\begin{pmatrix} U_N \\ V_N \end{pmatrix}$. However, since we also have endogeneity in our model, the variance-covariance matrix of the error terms also needs to capture that. The following assumption formalizes the spatial nature of the data and the endogeneity of the model:

Assumption 3.2.4 *The variance-covariance matrix is given as:*

$$\mathbb{V}\left[\begin{pmatrix} U_N \\ V_N \end{pmatrix} \middle| \boldsymbol{X}_{2N}, D, \rho, \lambda\right] = \mathbb{V}\left[\begin{pmatrix} U_N \\ V_N \end{pmatrix} \middle| D, \rho, \lambda\right] = \Lambda_N(D, \rho, \lambda) \tag{3.8}$$

where D contains all the pairwise distance between the observations in the data, ρ is the vector of variance-covariance parameters and λ is the spatial parameter.

The essential idea that we would like to formalize through Assumption **4** is twofold. First, we specify that the unobservables of all the individuals exhibit spatial correlations. In other words, for all $i, j = 1, 2, ..., N$, u_i is spatially correlated with u_j and v_i is spatially correlated with v_j. Second, we postulate that the *primary* error of an individual i is spatially correlated with the both the *reduced-form* error of the individual i *and* the *reduced-form* error of the individual j. In other words, for all $i, j = 1, 2, ..., N$, u_i is spatially correlated with v_j.

We can write the error variance-covariance matrix as:

$$\Lambda_N(D, \rho, \lambda) = \begin{bmatrix} \Lambda_{UU,N}(D, \rho, \lambda) & \Lambda_{UV,N}(D, \rho, \lambda) \\ \Lambda_{UV,N}(D, \rho, \lambda) & \Lambda_{VV,N}(D, \rho, \lambda) \end{bmatrix} \tag{3.9}$$

where $\Lambda_{UU,N}(D, \rho, \lambda) \equiv \mathbb{V}[U_N]$,$\Lambda_{VV,N}(D, \rho, \lambda) \equiv \mathbb{V}[V_N]$ and $\Lambda_{UV,N}(D, \rho, \lambda) \equiv \mathbb{C}ov[U_N, V_N]$. In our model, we express the endogeneity of y_{2i} through the relationship between the errors of the *primary* equation and the errors of the *reduced-form* equation. Formally, this relationship is expressed in the control function assumption that we state below. The components of the control function assumption, however, are defined in the variance-covariance matrix denoted by $\Lambda_{UV,N}(\lambda)$.

3.3 Estimating Equation

Our estimating equation needs to take account of the endogeneity in our spatially correlated data. To deal with the endogeneity in the spatially correlated data of our model, we divide our data set into G groups and then use the information on these observations *within* a group. The essential idea that we want to capture is that the internal correlation between these observations in a group are more important than the external correlation. We divide the total number of observations into G groups, each containing L_g number of observations. For simplicity assume that there are same number of observations in each group; i.e $L_1 = L_2 =, ..., = L_G \equiv L$.

Writing the model at the *group* level, we have:

For $g = 1, ..., G$:

$$Y_{1g} = \boldsymbol{X}_{1g}\boldsymbol{\beta}_1 + \tau Y_{2g} + U_g \tag{3.10}$$

with

$$Y_{1g} \equiv \begin{pmatrix} y_{1g_1} & \dots & y_{1g_L} \end{pmatrix}' \; ; \; \boldsymbol{X}_{1g} \equiv \begin{pmatrix} \boldsymbol{x}_{1g_1} & \dots & \boldsymbol{x}_{1g_L} \end{pmatrix}'$$
$$Y_{2g} \equiv \begin{pmatrix} y_{2g_1} & \dots & y_{2g_L} \end{pmatrix}' \; ; \; U_g \equiv \begin{pmatrix} u_{g_1} & \dots & u_{g_L} \end{pmatrix}'$$

Next, note that since y_{2i} is a function of $\boldsymbol{Z}_{2g}, V_g$ where $\boldsymbol{Z}_{2g}$ is the vector of all the exogenous variables within the group g:

$$\mathbb{E}[Y_{1g}|\boldsymbol{Z}_{2g}, V_g] = \boldsymbol{X}_{1g}\boldsymbol{\beta}_1 + \tau Y_{2g} + \mathbb{E}[U_g|\boldsymbol{Z}_{2g}, V_g] \tag{3.11}$$

As we will describe in section 3.2, $\boldsymbol{Z}_{2g}$ can contain more elements than $\boldsymbol{X}_{2g} \equiv \begin{pmatrix} \boldsymbol{x}_{2g_1} & \dots & \boldsymbol{x}_{2g_L} \end{pmatrix}'$.

If we do not have endogeneity, we will have $\mathbb{E}[U_g|\boldsymbol{Z}_{2g}, V_g] = 0$. This is the framework studied in Lu and Wooldridge (2017). However, ruling out endogenous covariates in a regression analysis is often not justified by economic theory. In practice, explanatory variables are often correlated the errors due to the common encountered issues of omitted variables and/or measurement errors which lead to inconsistencies in parameter estimation. This calls for a more general treatment of linear regression models that allow for both spatially dependent errors *and* endogenous covariates.

In this paper, endogeneity of y_{2i} is captured in the relationship between the *primary* errors and the *reduced-form* errors. This relationship is captured by $\mathbb{E}[U_g|\boldsymbol{Z}_{2g}, V_g]$ which is also termed as the *correction term*. Obtaining an expression for the *correction term* and including it in the estimating equation enables us to obtain the consistent estimators for the parameters. The control function approach proceeds by the imposing some structure on the *correction terms*. In our context, since the unobservables also exhibit spatial dependence, the control function assumption that we would impose allows us to incorporate this spatial dependence.

3.3.1 Control Function Assumption

Control function approach models the endogeneity of the covariates by explicitly specifying the endogeneity in the estimation. In our model, the control function method corrects for the endogeneity of y_2 by modeling the relationship between the errors of the *primary* equation and the error of the *reduced-form* equation. Since we have divided the data into groups, we specify the control function assumption for each group. This also incorporates the spatial correlation between the errors of each group. Thus the components of the control function assumption are given by the variance-covariance matrix of each group.

To obtain the expression for the control function assumption, define the group specific variance-covariance matrix of the unobservables as:

$$\mathbb{V}\left[\begin{pmatrix} U_g \\ V_g \end{pmatrix} \middle| \boldsymbol{Z}_{2g}, D_g, \lambda\right] = \begin{bmatrix} \Lambda_{UU,g}(\lambda) & \Lambda_{UV,g}(\lambda) \\ \Lambda_{UV,g}(\lambda) & \Lambda_{VV,g}(\lambda) \end{bmatrix} \tag{3.12}$$

where D_g contains all the pairwise distance between the observations *within* a group g

Formally stating, our control function assumption is:

Assumption 3.3.1 *For each* $g = 1, 2, .., G$

$$\mathbb{E}[U_g|\boldsymbol{Z}_{2g}, V_g] = \mathbb{E}[U_g|V_g] = \left[\Lambda_{VV,g}\right]^{-1}\left[\Lambda_{UV,g}\right] V_g \tag{3.13}$$

Define $\Pi_g = \left[\Lambda_{VV,g}\right]^{-1}\left[\Lambda_{UV,g}\right]$ and using the vectorization operator we can write:

$$\mathbb{E}[U_g|V_g] = \left[\Lambda_{VV,g}\right]^{-1}\left[\Lambda_{UV,g}\right] V_g = \Pi_g V_g \tag{3.14}$$

$$= \left[V_g' \otimes \mathbb{I}_L\right]\left[vec(\Pi_g)\right] \tag{3.15}$$

Substituting in the equation 10 we obtain the estimating equation as:

$$Y_{1g} = \boldsymbol{X}_{1g}\boldsymbol{\beta}_1 + \tau Y_{2g} + \left[V_g' \otimes \mathbb{I}_L\right]\left[vec(\Pi_g)\right] + \eta_g \tag{3.16}$$

where $\mathbb{E}[\eta_g|\boldsymbol{Z}_{2g}]$.

3.3.2 Instruments

Within a group, the exogenous variables for an individual g_i are given by $\boldsymbol{x}_{2g_i}$. However, the exogenous variables of other individuals in a group also serve as instruments for the individual g_i. In other words, if we denote the $\boldsymbol{X}_{2g_{-i}}$ as the vector of exogenous covariates of other individuals in a group g, then the full set of instruments for g_i are given by $\boldsymbol{z}_{2g_i} \equiv [\boldsymbol{x}_{2g_i}, \boldsymbol{X}_{2g_{-i}}]$

For example, suppose that $L = 2$. In this case, $Y_{2g} = \begin{pmatrix} y_{2g_1} & y_{2g_2} \end{pmatrix}'$. Recall that for each individual g_i in group g, we have the vector of instruments given by $\boldsymbol{x}_{2g_i}$. Recognizing that within a group the exogenous variables of the ones neighbors can also serve as the instruments, we can obtain extra instruments. That is in this case for g_1 the instruments are given by $\boldsymbol{Z}_{2g_1} \equiv [\boldsymbol{x}_{2g_1}, \boldsymbol{x}_{2g_2}]$.

We can write the within-group reduced form model for the endogenous variable that incorporates all the exogenous variables that are found within a group:

$$Y_{2g} = \boldsymbol{Z}_{2g}\boldsymbol{\delta}_2 + V_g \tag{3.17}$$

where $\boldsymbol{Z}_{2g} \equiv \begin{pmatrix} \boldsymbol{Z}_{2g_1} & \dots & \boldsymbol{Z}_{2g_L} \end{pmatrix}'$.

3.4 Estimation Procedures

In this section, we describe three new estimation procedures that incorporate the *additional* instruments as well as the *within-group* spatial dependence that is facilitated by the *groupwise* division of the data. We begin by describing a two-step estimation of the **Spatial Control Function Estimator**. Next, we describe a **Grouped Two Stage Least Square** estimation procedure that uses only the *additional* instruments and ignores the spatial dependence in the data. Finally, we describe the **Spatial Generalized Instrumental Variable** estimator that incorporates both the *additional* instruments as well as the *within-group* spatial dependence of the observations.

3.4.1 Control Function Estimation

We write the reduced form equation and the estimating equation again for convenience:

$$Y_{1g} = \boldsymbol{X}_{1g}\boldsymbol{\beta}_1 + \tau Y_{2g} + \left[V_g' \otimes \mathbb{I}_L\right]\left[vec(\Pi_g)\right] + \eta_g$$

$$Y_{2g} = \boldsymbol{Z}_{2g}\boldsymbol{\delta}_2 + V_g$$

The equations above suggest a two-step estimation procedure. In the first step we estimate the reduced form model and obtain the residuals. In the second step we plug in the first step residuals in the estimating equation. Then we use the Quasi-GLS estimation procedure described in Lu and Wooldridge (2017).

3.4.1.1 First Step

In the first step, we can re-write:

$$Y_{2g} = \boldsymbol{Z}_{2g}\boldsymbol{\delta}_2 + V_g \tag{3.18}$$

Denote the true parameters of the first-step as $\boldsymbol{\delta}_{2*} \in \boldsymbol{\delta}_2^P$ where $\boldsymbol{\delta}_2^P$ is a finite dimensional parameter space. The estimator denoted by $\widehat{\boldsymbol{\delta}}_2$ is given as:

$$\widehat{\boldsymbol{\delta}}_2 = arg\ min_{\boldsymbol{\delta}_2 \in \boldsymbol{\delta}_2^P} \frac{1}{G}\sum_{g=1}^{G}\left[\left(Y_{2g} - \boldsymbol{Z}_{2g}\boldsymbol{\delta}_2\right)'\left(Y_{2g} - \boldsymbol{Z}_{2g}\boldsymbol{\delta}_2\right)\right] \tag{3.19}$$

Since we are estimating a linear model, we get:

$$\widehat{\boldsymbol{\delta}}_2 = \left[\sum_{g=1}^{G}\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g}\right]^{-1}\left[\sum_{g=1}^{G}\boldsymbol{Z}_{2g}'Y_{2g}\right] \tag{3.20}$$

In addition, this gives the first stage residuals as:

$$\widehat{V}_g = Y_{2g} - \boldsymbol{Z}_{2g}\widehat{\boldsymbol{\delta}}_2 \tag{3.21}$$

3.4.1.2 Second Step

In the second stage, we have the estimating equation as:

$$Y_{1g} = \boldsymbol{X}_{1g}\boldsymbol{\beta}_1 + \tau Y_{2g} + \left[V_g' \otimes \mathbb{I}_L\right]\left[vec(\Pi_g)\right] + \eta_g$$

To motivate the estimation procedure in the second-step, assume that we fully observe V_g. Re-write the estimating equation as:

$$Y_{1g} = \boldsymbol{X}_g \boldsymbol{\theta}_1 + \eta_g \tag{3.22}$$

where $\boldsymbol{X}_g$ is a vector that contains all the covariates of the estimating equation and $\boldsymbol{\theta}_1$ is a vector that contains all the parameters. We can estimate the parameters in equation (22) using a simple a OLS and we will obtain consistent estimators. However, note that the errors in the estimating equation are going to have spatial dependence. This suggests that we can obtain some efficiency gains if we can incorporate the spatial dependence in the estimation procedure.

To estimate the parameters in this spatial setting, we will implement the *Quasi-GLS* estimation procedure proposed by Lu and Wooldridge (2017). Denote the variance-covariance matrix of η_g by $\Lambda_{g\eta}(\lambda_\eta)$ where λ_η is the spatial parameter. Note that the group-specific variance-covariance matrix $\Lambda_{g\eta}(\lambda_\eta)$ is contained in the full variance-covariance matrix for η_N denoted by $\Lambda_{N\eta}(\lambda_\eta)$.

Traditional GLS estimation considers the entire error variance-covariance matrix $\Lambda_{N\eta}(\lambda_\eta)$. This entails accounting for all the pairwise correlations of the observations which is computationally very tedious. However, since we divide the observations into groups according to the distances between the observations we recognize that the *within-group* correlations account for most of the correlations in the data and ignoring the *across-group* correlations should not lead to much efficiency loss. This is precisely the motivation behind the *Quasi-GLS* estimation procedure in Lu and Wooldridge (2017). *Quasi-GLS* estimation procedure involves using a *tapered* error variance-covariance matrix which essentially implies that we extract the variance-covariance matrix for *within-group* errors denoted by $\Lambda_{g\eta}(\lambda_\eta)$ and the correlations between the errors *across-groups* are set to zero.

Formally, define a $N \times N$ *tapering* matrix $\mathscr{T}$ whose individual components $\mathscr{T}_{i,j} = 1$ if the observations i, j belong to the same group and $\mathscr{T}_{i,j} = 0$ otherwise for $i, j = 1, 2, .., N$. Define $\Lambda_\eta \equiv diag\left[\Lambda_{1\eta}, \Lambda_{2\eta}, ..., \Lambda_{G\eta}\right]$. In other words, Λ_η is a block diagonal matrix that contains

only *within-group* error variance-covariances. Note that $\Lambda_\eta \equiv \mathscr{T} \circ \Lambda_{N\eta}$ where $\circ$ denotes the Hadamard product. Now, while the traditional GLS estimation is weighted by $\Lambda_{N\eta}$ and *Quasi-GLS* estimation is weighted by Λ_η.

For example, consider the simplest case of $N = 4$. Further assume that observations $\{1,2\}$ are in group 1 and observations $\{3,4\}$ are in group 2. In this case,

$$\mathscr{T} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \Lambda_\eta = \begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{33} & \sigma_{34} \\ 0 & 0 & \sigma_{43} & \sigma_{44} \end{pmatrix} = \mathscr{T} \circ \Lambda_{N\eta}$$

where $\Lambda_{N\eta} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}$.

In our context, we call *Quasi-GLS* estimator as the *Spatial Control-Function Estimator (sp.CF)*. Denote the true parameters of the second-step as $\boldsymbol{\theta}_{1*} \in \boldsymbol{\Theta}_1$ where $\boldsymbol{\Theta}_1$ is a finite dimensional parameter space. The estimator is obtained as:

$$\widehat{\boldsymbol{\theta}}_{1(sp.CF)} = arg\ min_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \frac{1}{G} \sum_{g=1}^{G} \left[\left(Y_{1g} - \boldsymbol{X}_g \boldsymbol{\theta}_1 \right)' \Lambda_{g\eta}^{-1} \left(Y_{1g} - \boldsymbol{X}_g \boldsymbol{\theta}_1 \right) \right] \tag{3.23}$$

As it is clear, by using the *Quasi-GLS* estimation procedure to obtain our *Control-Function* estimator in the second-step, we are taking account of some of the correlations in the data by including $\Lambda_{g\eta}(\lambda_\eta)$. Thus we get efficiency gains because we do not completely ignore the spatial dependence in the data.

In practice, we first plug in $\widehat{V}_g$ as the consistent estimators for V_g that controls for the endogeneity of Y_{2g}. Further, to obtain a *feasible* version of the *Spatial Control-Function* estimator, we will need a consistent estimator for $\Lambda_{g\eta}(\lambda_\eta)$ that depends on the pairwise distances and spatial parameter λ_η. Let $\widehat{\Lambda}_{g\eta}(\hat{\lambda}_\eta)$ be the consistent estimator for $\Lambda_{g\eta}(\lambda_\eta)$. For notational convenience, $\boldsymbol{X}_g$ now denotes the vector of all the covariates with $\widehat{V}_g$ instead of V_g.

The second-step essentially entails using the *Feasible Quasi-GLS* procedure of Lu and Wooldridge (2017) to obtain what we call the *Feasible Spatial Control-Function (F.sp.CF)* estimator, denoted by $\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)}$. The estimator is obtained as:

$$\widehat{\boldsymbol{\theta}}_{1(sp.CF)} = arg\ min_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \frac{1}{G} \sum_{g=1}^{G} \left[\left(Y_{1g} - \boldsymbol{X}_g \boldsymbol{\theta}_1 \right)' \widehat{\Lambda}_{g\eta}^{-1} \left(Y_{1g} - \boldsymbol{X}_g \boldsymbol{\theta}_1 \right) \right] \tag{3.24}$$

Solving the optimization routine the explicit expression is obtained as:

$$\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)} = \left[\sum_{g=1}^{G} \left(\boldsymbol{X}_g' \widehat{\Lambda}_{g\eta}^{-1} \boldsymbol{X}_g \right) \right]^{-1} \left[\sum_{g=1}^{G} \left(\boldsymbol{X}_g' \widehat{\Lambda}_{g\eta}^{-1} Y_{1g} \right) \right] \tag{3.25}$$

3.4.2 Incorporating Extra Instruments in other Estimation Procedures

Grouping the observations into groups based on the distance between the observations allows us to obtain *additional* instruments for each group. We can incorporate these instruments in the traditional estimation procedures to get some efficiency gains.

Let $\boldsymbol{z}_{1i} \equiv \{\boldsymbol{x}_{1i}, y_{2i}\}$ and denote $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}_1, \tau]'$ for notational simplicity. The primary outcome equation can be re-written as:

$$y_{1i} = \boldsymbol{z}_{1i} \boldsymbol{\beta} + u_i \tag{3.26}$$

The groupwise notation is given as:

$$Y_{1g} = \boldsymbol{Z}_{1g} \boldsymbol{\beta} + U_g \tag{3.27}$$

where $\boldsymbol{Z}_{1g}$ is the groupwise notation for $\boldsymbol{z}_{1i}$.

The traditional 2SLS estimator is given as:

$$\widehat{\boldsymbol{\beta}}_{(2SLS)} = \left[\left(\sum_{i=1}^{N} \boldsymbol{z}_{1i}' \boldsymbol{x}_{2i} \right) \left(\sum_{i=1}^{N} \boldsymbol{x}_{2i}' \boldsymbol{x}_{2i} \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{x}_{2i}' \boldsymbol{z}_{1i} \right) \right]^{-1} \times \left[\left(\sum_{i=1}^{N} \boldsymbol{z}_{1i}' \boldsymbol{x}_{2i} \right) \left(\sum_{i=1}^{N} \boldsymbol{x}_{2i}' \boldsymbol{x}_{2i} \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{x}_{2i}' y_{1i} \right) \right] \tag{3.28}$$

3.4.2.1 Grouped 2SLS Estimation

When we group the data to estimate the model with new instruments, we obtain the expression for what we call *grouped-2SLS*.

$$\widehat{\boldsymbol{\beta}}_{(grpd.2SLS)} = \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\boldsymbol{Z}_{1g}\right)\right]^{-1} \times \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'Y_{1g}\right)\right] \tag{3.29}$$

3.4.2.2 Spatial Generalized Instrumental Variable Estimation

We can also incorporate the extra instruments as well as the spatial dependence in the framework of *Generalized Instrumental Variables (GIV)* estimation.

Recall that we denoted $Var[U_g] = \Lambda_{UU,g}$ as the *groupwise* variance-covariance matrix for the primary outcome equation. The instruments are denoted as $\boldsymbol{Z}_{2g}$. *GIV* works by incorporating the spatial dependence of the errors through a *GLS* type transformation of equation (27). Assuming that $\Lambda_{UU,g}$ to be a positive definite, we obtain the *GLS* transformation as:

$$\Lambda_{UU,g}^{-1/2} Y_{1g} = \Lambda_{UU,g}^{-1/2}\boldsymbol{Z}_{1g}\boldsymbol{\beta} + \Lambda_{UU,g}^{-1/2} U_g \tag{3.30}$$

where $\Lambda_{UU,g}^{-1/2}\Lambda_{UU,g}^{-1/2} = \Lambda_{UU,g}^{-1}$.

Next, we estimate (32) using $\Lambda_{UU,g}^{-1/2}\boldsymbol{Z}_{2g}$ as instruments. We call this estimator as *Spatial GIV*:

$$\widehat{\boldsymbol{\beta}}_{(sp.GIV)} = \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\Lambda_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\Lambda_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\Lambda_{UU,g}^{-1}\boldsymbol{Z}_{1g}\right)\right]^{-1} \times \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\Lambda_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\Lambda_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\Lambda_{UU,g}^{-1}Y_{1g}\right)\right] \tag{3.31}$$

In practice, we use a Feasible version of the estimator once we plug in an estimator for $\widehat{\Lambda}_{UU,g}^{-1}$

$$\widehat{\boldsymbol{\beta}}_{(F.sp.GIV)} = \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{1g}\right)\right]^{-1} \times \tag{3.32}$$

$$\left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}Y_{1g}\right)\right] \tag{3.33}$$

3.4.3 Estimation of the spatial parameter

We can obtain a consistent estimator for λ_η using a straightforward minimization algorithm. We can obtain residuals from a preliminary consistent estimator of $\boldsymbol{\theta}_1$. Recall that we denoted the variance-covariance matrix of η_N by $\Lambda_{N\eta}(D,\lambda_\eta)$. We specify the functional form of the individual components of the matrix as $\Lambda_{N\eta;(i,j)}(d_{ij},\lambda_\eta)$ with $\Lambda_{N\eta;(i,i)} = \sigma_\eta^2$.

A consistent estimator for the variance of η_i denoted by $\widehat{\sigma}_\eta$ is given as

$$\widehat{\sigma}_\eta = \frac{1}{N-K_x}\sum_{i=1}^{N}\tilde{\eta}_i^2 \tag{3.34}$$

A consistent estimator for λ_η can be obtained as:

$$\hat{\lambda}_\eta = arg\ min\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\tilde{\eta}_i\tilde{\eta}_j - \boldsymbol{\Lambda}_{N\eta;(i,j)}(d_{ij},\widehat{\sigma}_\eta,\lambda_\eta)\right)^2 \tag{3.35}$$

3.5 Asymptotics

In spatial statistics, there are two main frameworks for the asymptotic analysis (Lee (2004), Cressie (1993)): *increasing-domain* and *fixed-domain*. In *increasing-domain* framework, we assume that the minimum distance between the observations is bounded below by a positive constant and the asymptotics are based on a growing observation region. In *fixed-domain* or *infill*

framework, we assume that as the number of observations increase, the observational region becomes increasingly dense while the observational region is assumed to be fixed and bounded. It has been well established (Cressie (1993) , Stein (1999), Ripley (1988), Zhang (2004)) that general results under *fixed-domain* asympotics are not available because as the number of observations increase, the number of interactions between the observations also increase and there is no theoretical basis for the usual behavior of estimators.

There exist a large body of literature on the asymptotic analysis with spatial dependence under the *increasing-domain* framework. Mardia and Marshall (1984), Cressie and Lahiri (1993) give consistency and asymptotic normality results for the maximum likelihood and other likelihood estimators for regression models with spatially correlated errors. Lee (2004) investigates the asymptotic properties of the maximum likelihood and quasi-maximum likelihood estimator for the spatial autoregressive models. Conley (1996) obtains the asymptotic results for the generalized method of moments estimators with stationary spatial data.

In this paper, we obtain the asymptotic results for out estimator under the *increasing-domain* framework. We collect our estimating equations and describe the two-step estimation procedure as a *one-step* estimation procedure that enables us to derive the asymptotic properties in a concise manner. We choose the framework of Jenish and Prucha (2012) to implement the law of large numbers and central limit theorem for spatial *near-epoch* dependence.

Let $\mathfrak{w}_i = \{y_{1i}, y_{2i}, \boldsymbol{x}_{2i}, u_i, v_i\}$ be all the random variables in our model for $i = 1, 2, 3....$

Assumption 3.5.1 *$\forall N$, $\mathfrak{w}_i$ are located on an infinitely countable lattic $\mathscr{D} \in \mathbb{R}^d$ of possibly uneven placed locations. The space $\mathbb{R}^d$ is equipped with the metric $d_{(i,j)} = max_{1 \leq l \leq d} |i_l - j_l|$, where j_l is the l^{th} element of j, and all elements of $\mathscr{D}$ are located at distances of at least $d_* > 0$ from each other*

Next, we define the nature of dependence of the spatial processes of our model to obtain asymptotic properties of our estimators. Jenish and Prucha (2012) obtains results for law of large

numbers and central limit theorem under the general set of restrictions on near cross-sectional dependence. Specifically, the spatial processes are assumed to exhibit *near-epoch* dependence.

Near-epoch dependence for random fields is described as:

Assumption 3.5.2 ***(a)*** *For some random filed* $\varepsilon = \{\varepsilon_{i,N}; i \in \mathscr{T}_N, \mathscr{T}_N \subset \mathscr{D}, N \geq 1\}$ *with* $|\mathscr{T}_N| \to \infty$ *as* $N \to \infty$ *where* $||$ *denotes the cardinality of a set; for some element* $\mathfrak{w}_i^a$ *of* $\mathfrak{w}_i$*, the random field* $\{\{\mathfrak{w}_i^a\}_{i=2,..,N}\}_{N\geq 1}$ *is* $\mathscr{L}_2$*-near-epoch dependent (NED) on the random field* ε*, i.e, :*

$$||\mathfrak{w}_i^a - \mathbb{E}\left(\mathfrak{w}_i^a | \mathscr{F}_{i,N}(s)\right)||_2 \leq C\psi(s)\big) \tag{3.36}$$

where $\mathscr{F}_{i,N}(s) = \sigma(\varepsilon_{j,N}; j \in \mathscr{T}_N, d_{i,j} \leq s)$*. C is some positive constant, and* $\psi(s)$ *is a deterministic sequence called NED coefficients with* $\psi(s) \geq 0$ *and* $lim_{s\to\infty}\psi(s) = 0$

(b) ε *is* α*-mixing coefficients satisfying Assumption 3 of Jenish and Prucha (2012).*

(c) $\psi(s)$ *satisfies* $\sum_{r=1}^{d-1} \psi(r) < \infty$

This assumption implies that the random variables in $\mathfrak{w}_i$ can be arbitrarily well approximated by neighboring observations of an α-mixing field. This includes the case where $\mathfrak{w}_i$ is the α-mixing as is the case in this model. (Verdier 2016)

Under Assumption 1 and Assumption 2 described above and Assumption (4) in Jenish and Prucha (2012) we can apply the weak law of large numbers and central limit theorem described in Theorem 1 and Theorem 2 in Jenish and Prucha (2012).

3.5.1 Asymptotics for Feasible Spatial Control Function Estimator

For $g = 1, 2, ..., G$:

$$Y_{1g} = \boldsymbol{X}_g \boldsymbol{\theta}_1 + \eta_g \tag{3.37}$$

$$Y_{2g} = \boldsymbol{Z}_{2g} \boldsymbol{\delta}_2 + V_g \tag{3.38}$$

Re-writing the equations:

$$\mathtt{y}_g = \mathtt{w}_g \boldsymbol{\theta} + \mathtt{u}_g \tag{3.39}$$

where

$$\mathtt{y}_g = \begin{pmatrix} Y_{1g} \\ Y_{2g} \end{pmatrix}, \mathtt{w}_g = \begin{pmatrix} \boldsymbol{X}_g & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{2g} \end{pmatrix}, \mathtt{u}_g = \begin{pmatrix} \eta_g \\ V_g \end{pmatrix} \tag{3.40}$$

Next assumption imposes strict exogeneity:

Assumption 3.5.3 *For $i = 1, 2, ..., N$*

$$\mathbb{E}[\mathtt{u}_i | \mathtt{w}_1, \mathtt{w}_2, ...] = 0 \tag{3.41}$$

where $\mathtt{u}_i, \mathtt{w}_i$ is individual level notation. This implies that for any matrix $\boldsymbol{\Omega}_g$ of conformable dimension, $\mathbb{E}[\mathtt{w}_g' \boldsymbol{\Omega}_g^{-1} \mathtt{w}_g] = \mathbf{0}$ for all g.

In the context of this paper, we have

$$\boldsymbol{\Omega}_g = \begin{pmatrix} \Lambda_{g\eta} & \mathbf{0} \\ \mathbf{0} & \mathbb{I} \end{pmatrix} \tag{3.42}$$

The *Spatial Control-Function estimator* is given as:

$$\widehat{\boldsymbol{\theta}}_{(sp.CF)} = arg\ min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{G} \sum_{g=1}^{G} \left[\left(\mathtt{y}_g - \mathtt{w}_g \boldsymbol{\theta} \right)' \boldsymbol{\Omega}_g^{-1} \left(\mathtt{y}_g - \mathtt{w}_g \boldsymbol{\theta} \right) \right] \tag{3.43}$$

The first order conditions for the optimization routine are given as:

$$\mathbf{S}_G(\widehat{\boldsymbol{\theta}}_{(sp.CF)}) \equiv \frac{1}{G} \sum_{g=1}^{G} \mathtt{s}_g \left(\widehat{\boldsymbol{\theta}}_{sp.CF}, \mathtt{w}_g \right) \equiv \frac{1}{G} \sum_{g=1}^{G} \mathtt{w}_g^{-1} \boldsymbol{\Omega}_g^{-1} \left(\boldsymbol{y}_g - \mathtt{w}_g \widehat{\boldsymbol{\theta}}_{(sp.CF)} \right) \tag{3.44}$$

$$= \frac{1}{G} \sum_{g=1}^{G} \begin{pmatrix} \mathtt{s}_{g1} \left(\widehat{\boldsymbol{\theta}}_{1(sp.CF)}, \boldsymbol{Z}_{2g}; \widehat{\boldsymbol{\delta}}_2 \right) \\ \mathtt{s}_{g2} \left(\widehat{\boldsymbol{\delta}}_2, \boldsymbol{Z}_{2g} \right) \end{pmatrix} \tag{3.45}$$

$$\equiv \begin{pmatrix} \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{X}_g' \Lambda_{g\eta} \left(Y_{1g} - \boldsymbol{X}_g \widehat{\boldsymbol{\theta}}_{1(sp.CF)} \right) \\ \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{Z}_{2g}' \left(Y_{2g} - \boldsymbol{Z}_{2g} \widehat{\boldsymbol{\delta}}_2 \right) \end{pmatrix} \tag{3.46}$$

$$= \mathbf{0} \tag{3.47}$$

These conditions are in fact the sample analogue of the moment conditions:

$$\mathbb{E}\left[\mathtt{s}_g \left(\boldsymbol{\theta}_*, \mathtt{w}_g \right) \middle| \mathtt{W}_G \right] = \mathbf{0} \tag{3.48}$$

where $\boldsymbol{\theta}_* \equiv \{\boldsymbol{\theta}_{1*}, \boldsymbol{\delta}_{2*}\}$ denote the true parameters, $\mathtt{W}_G$ is the full data matrix and

$$\mathrm{s}_g\left(\boldsymbol{\theta}, \mathtt{w}_g\right) = \begin{pmatrix} \mathrm{s}_{g1}\left(\boldsymbol{\theta}_1, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_2\right) \\ \mathrm{s}_{g2}\left(\boldsymbol{\delta}_2, \boldsymbol{Z}_{2g}\right) \end{pmatrix} \tag{3.49}$$

$$= \begin{pmatrix} \boldsymbol{X}_g' \Lambda_g \eta \left(Y_{1g} - \boldsymbol{X}_g \boldsymbol{\theta}_1\right) \\ \boldsymbol{Z}_g' \left(Y_{2g} - \boldsymbol{Z}_g \boldsymbol{\delta}_2\right) \end{pmatrix} \tag{3.50}$$

Thus the estimator proposed in this paper can also be described in a *GMM* framework. The asymptotic theory developed by Jenish and Prucha (2012) for spatial *GMM* can easily be implemented here. In this paper, we also explicitly correct for the first-step estimation.

Next define

$$\mathbb{H}_g(\boldsymbol{\theta}_*, \mathtt{w}_g) \equiv \left(\frac{\partial \mathrm{s}_g\left(\boldsymbol{\theta}_*, \mathtt{w}_g\right)}{\partial \boldsymbol{\theta}'}\right) = -\mathtt{w}_g' \boldsymbol{\Omega}_g^{-1} \mathtt{w}_g \tag{3.51}$$

The next two assumptions impose boundedness conditions and appropriate rank conditions.

Assumption 3.5.4 $\mathbb{A} \equiv \lim_{G\to\infty} \frac{1}{G} \sum_{g=1}^{G} \mathbb{E}\left(-\mathbb{H}_g(\boldsymbol{\theta}_*, \mathtt{w}_g)\right)$ *exists and has full rank*

In addition, assume:

Assumption 3.5.5 $\mathbb{B} \equiv lim_{G\to\infty} \mathbb{E}\left\{\left[\frac{1}{G}\sum_{g=1}^{G} \mathrm{s}_g\left(\boldsymbol{\theta}_*, \mathtt{w}_g\right)\right]\left[\frac{1}{G}\sum_{h=1}^{G} \mathrm{s}_h\left(\boldsymbol{\theta}_*, \mathtt{w}_h\right)\right]'\right\}$ *exists and has full rank.*

Since *Quasi-GLS* groups the observations according to the distances between the individuals, it is essentially grouping "nearby" observations. Thus the groups will also be *near-epoch* dependent processes. In addition, NED is preserved under addition and multiplication (Verdier 2016, Davidson 1994). So Theorem 1 and Theorem 3 of Jenish and Prucha (2012) can be implemented.

In practice, we get the feasible version of the estimator as:

$$\widehat{\boldsymbol{\theta}}_{(F.sp.CF)} = arg\ min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{G} \sum_{g=1}^{G} \left[\left(\mathrm{y}_g - \mathtt{w}_g\boldsymbol{\theta}\right)' \widehat{\boldsymbol{\Omega}}_g^{-1} \left(\mathrm{y}_g - \mathtt{w}_g\boldsymbol{\theta}\right)\right] \tag{3.52}$$

Theorem 7 *Under **Assumptions 1-10**, $\widehat{\boldsymbol{\theta}}_{(F.sp.CF)}$ is consistent and*

$$\sqrt{G}\left(\widehat{\boldsymbol{\theta}}_{(F.sp.CF)}-\boldsymbol{\theta}_{*}\right)\rightarrow_{d}\mathscr{N}\left(\mathbf{0},\mathbb{A}^{-1}\mathbb{B}\mathbb{A}^{-1}\right) \tag{3.53}$$

Proof. Following the procedure in Lu (2013), we will first obtain the asymptotic properties for $\widehat{\boldsymbol{\theta}}_{(sp.CF)}$ and then show that $\widehat{\boldsymbol{\theta}}_{(sp.CF)}$ and $\widehat{\boldsymbol{\theta}}_{(F.sp.CF)}$ are asymptotically equivalent.

- **Consistency of $\widehat{\boldsymbol{\theta}}_{(sp.CF)}$:** For consistency, note that

$$\widehat{\boldsymbol{\theta}}_{(sp.CF)}=\boldsymbol{\theta}_{*}+\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{w}_{g}\right)^{-1}\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{u}_{g}\right) \tag{3.54}$$

  Because *NED*, is preserved under multiplication and addition (Theorems 17.8 and 17.9 in Davidson (1994)) $\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{w}_{g}$ and $\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{u}_{g}$ as well as $\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{w}_{g}\right)$ and $\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{u}_{g}\right)$ are also NED processes. Using Theorem 1 in Jenish and Prucha (2012), we get

$$\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{w}_{g}\right)\rightarrow_{p}\mathbb{A} \tag{3.55}$$

$$\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_{g}^{\prime}\boldsymbol{\Omega}_{g}^{-1}\mathtt{u}_{g}\right)\rightarrow_{p}\mathbf{0} \tag{3.56}$$

  The consistency result follows.

- **Asymptotic Normality of $\widehat{\boldsymbol{\theta}}_{(sp.CF)}$:** Doing a mean value expansion around $\boldsymbol{\theta}_{*}$

$$\mathbf{S}_{G}(\widehat{\boldsymbol{\theta}}_{(sp.CF)})=\mathbf{0} \tag{3.57}$$

$$\mathbf{0}=\frac{1}{G}\sum_{g=1}^{G}\mathrm{s}_{g}\left(\widehat{\boldsymbol{\theta}}_{(sp.CF)},\mathtt{w}_{g}\right) \tag{3.58}$$

$$\mathbf{0}=\frac{1}{G}\sum_{g=1}^{G}\mathrm{s}_{g}\left(\boldsymbol{\theta}_{*},\mathtt{w}_{g}\right)+\frac{1}{G}\sum_{g=1}^{G}\mathbb{H}_{g}\left(\ddot{\boldsymbol{\theta}},\mathtt{w}_{g}\right)\left(\widehat{\boldsymbol{\theta}}_{(sp.CF)}-\boldsymbol{\theta}_{*}\right) \tag{3.59}$$

$$\sqrt{G}\left(\widehat{\boldsymbol{\theta}}_{(sp.CF)}-\boldsymbol{\theta}_{*}\right)=\left[-\frac{1}{G}\sum_{g=1}^{G}\mathbb{H}_{g}\left(\ddot{\boldsymbol{\theta}},\mathtt{w}_{g}\right)\right]^{-1}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathrm{s}_{g}\left(\boldsymbol{\theta}_{*},\mathtt{w}_{g}\right)\right] \tag{3.60}$$

where $\ddot{\boldsymbol{\theta}}$ has elements between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_*$

Using Theorem 1 and Theorem 2 of Jenish and Prucha (2012),

$$\left[-\frac{1}{G}\sum_{g=1}^{G}\mathbb{H}_g\left(\ddot{\boldsymbol{\theta}},\mathtt{w}_g\right)\right]\to_p\mathbb{A} \tag{3.61}$$

and

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathtt{s}_g\left(\boldsymbol{\theta}_*,\mathtt{w}_g\right)\to_d\mathcal{N}\left(\mathbf{0},\mathbb{B}\right) \tag{3.62}$$

Bringing these two results together, we get

$$\sqrt{G}\left(\widehat{\boldsymbol{\theta}}_{(sp.CF)}-\boldsymbol{\theta}_*\right)\to_d\mathcal{N}\left(\mathbf{0},\mathbb{A}^{-1}\mathbb{B}\mathbb{A}^{-1}\right) \tag{3.63}$$

- **Asymptotic Equivalence of $\widehat{\boldsymbol{\theta}}_{(sp.CF)}$ and $\widehat{\boldsymbol{\theta}}_{(F.sp.CF)}$:** We have

$$\widehat{\boldsymbol{\theta}}_{CF}=\boldsymbol{\theta}_*+\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{w}_g\right)^{-1}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{u}_g\right)$$

$$\widehat{\boldsymbol{\theta}}_{FCF}=\boldsymbol{\theta}_*+\left(\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{w}_g\right)^{-1}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{u}_g\right)$$

Note that since $\widehat{\boldsymbol{\Omega}}_g\to_p\boldsymbol{\Omega}_g$, we get:

$$\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{w}_g\to_p\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{w}_g \tag{3.64}$$

$$\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{w}_g\to_p\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{w}_g \tag{3.65}$$

Since $\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{w}_g\to_p\mathbb{A}$, this implies $\frac{1}{G}\sum_{g=1}^{G}\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{w}_g\to_p\mathbb{A}$

Next,

$$\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{u}_g=\mathtt{w}_g'(\boldsymbol{\Omega}_g^{-1}+o_p(1))\mathtt{u}_g \tag{3.66}$$

$$=\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{u}_g+o_p(1) \tag{3.67}$$

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{u}_g=\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathtt{w}_g'\boldsymbol{\Omega}_g^{-1}\mathtt{u}_g+o_p(1) \tag{3.68}$$

Thus we will have $\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{u}_g\to_d\mathcal{N}(\mathbf{0},\mathbb{B})$ and the asymptotic result follows.

■

3.5.1.1 Adjusting for first-step estimation

We can further tease out the asymptotic properties of $\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)}$ from our GMM framework. The asymptotic variance of $\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)}$ would adjust for the first-step estimation. We have:

$$\begin{aligned}\mathbf{S}_G(\widehat{\boldsymbol{\theta}}_{(F.sp.CF)}) &= \begin{pmatrix} \mathbf{S}_{G1}(\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)}, \boldsymbol{Z}_{2g}; \widehat{\boldsymbol{\delta}}_2) \\ \mathbf{S}_{G2}(\widehat{\boldsymbol{\delta}}_2; \boldsymbol{Z}_{2g}) \end{pmatrix} \\ &\equiv \begin{pmatrix} \frac{1}{G}\sum_{g=1}^{G} \mathrm{s}_{g1}(\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)}, \boldsymbol{Z}_{2g}; \widehat{\boldsymbol{\delta}}_2) \\ \frac{1}{G}\sum_{g=1}^{G} \mathrm{s}_{g2}(\widehat{\boldsymbol{\delta}}_2; \boldsymbol{Z}_{2g}) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}\end{aligned}$$

Next, define

$$\mathbb{H}_{g1} \equiv \left(\frac{\partial \mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right)}{\partial \boldsymbol{\theta}_1'} \right) \tag{3.69}$$

$$\mathbb{H}_{g2} \equiv \left(\frac{\partial \mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right))}{\partial \boldsymbol{\delta}_2'} \right) \tag{3.70}$$

Doing the similar calculations as before we will get:

$$\sqrt{G}\left(\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)} - \boldsymbol{\theta}_{1*}\right) = [\mathbb{A}_1]^{-1} \left[\frac{1}{G} \sum_{g=1}^{G} \mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \widehat{\boldsymbol{\delta}}_2\right) \right] + o_p(1) \tag{3.71}$$

where $\mathbb{A}_1 \equiv \lim_{G\to\infty} \left\{ -\frac{1}{G}\sum_{g=1}^{G} \mathbb{E}\left[\mathbb{H}_{g1}\right] \right\}$.

Also, using the mean value expansion,

$$\frac{1}{G} \sum_{g=1}^{G} \mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_G; \widehat{\boldsymbol{\delta}}_2\right) = \frac{1}{G} \sum_{g=1}^{G} \mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_G; \boldsymbol{\delta}_{2*}\right) + \mathbb{F}\left(\widehat{\boldsymbol{\delta}}_2 - \boldsymbol{\delta}_{2*}\right) + o_p(1) \tag{3.72}$$

where

$$\mathbb{F} \equiv \lim_{G\to\infty} \left\{ -\frac{1}{G} \sum_{g=1}^{G} \mathbb{E}\left[\mathbb{H}_{g2}\right] \right\} \tag{3.73}$$

Next, we have a first order representation of $\sqrt{G}\left(\widehat{\boldsymbol{\delta}}_2 - \boldsymbol{\delta}_{2*}\right)$ as:

$$\sqrt{G}\left(\widehat{\boldsymbol{\delta}}_2 - \boldsymbol{\delta}_{2*}\right) = \frac{1}{G} \sum_{g=1}^{G} \mathrm{s}_g^{(2)}\left(\boldsymbol{\delta}_{2*}, \boldsymbol{Z}_{2g}\right) + o_p(1) \tag{3.74}$$

where $\mathrm{s}_g^{(2)}\left(\boldsymbol{\delta}_{2*}, \boldsymbol{Z}_{2g}\right)$ depends on the first-step estimation and $\mathrm{s}_{g2}\left(\boldsymbol{\delta}_2, \boldsymbol{Z}_{2g}\right)$ with $\mathbb{E}[\mathrm{s}_g^{(2)}\left(\boldsymbol{\delta}_{2*}, \boldsymbol{Z}_{2g}\right)] = \mathbf{0}$.

Now bringing together equations (70), (71) and (73), we have:

$$\sqrt{G}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1*}\right) = [\mathbb{A}_1]^{-1}\left[\frac{1}{G}\sum_{g=1}^{G}\mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_G; \boldsymbol{\delta}_{2*}\right) + \mathbb{F}\left[\frac{1}{G}\sum_{g=1}^{G}\mathrm{s}_g^{(2)}\left(\boldsymbol{\delta}_{2*}, \boldsymbol{Z}_{2g}\right)\right]\right] + o_p(1) \tag{3.75}$$

$$= [\mathbb{A}_1]^{-1}\left[\frac{1}{G}\sum_{g=1}^{G}\mathrm{s}_g^{(1)}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right)\right] + o_p(1) \tag{3.76}$$

where

$$\mathrm{s}_g^{(1)}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right) = \mathrm{s}_{g1}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right) + \mathbb{F}\mathrm{s}_g^{(2)}\left(\boldsymbol{\delta}_{2*}, \boldsymbol{Z}_{2g}\right) \tag{3.77}$$

Next define $\mathbb{B}_1 \equiv \lim_{G\to\infty}\left\{\frac{1}{G}\mathbb{E}\left[\sum_{g=1}^{G}\mathrm{s}_g^{(2)}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right)\sum_{g=1}^{G}\mathrm{s}_g^{(2)}\left(\boldsymbol{\theta}_{1*}, \boldsymbol{Z}_{2g}; \boldsymbol{\delta}_{2*}\right)'\right]\right\}$.

Using the law of large number and central limit theorem for the mixing sequence, we can follow the steps above to establish the following:

Theorem 8 *Under **Assumptions 1-10** ,*

$$\sqrt{G}\left(\widehat{\boldsymbol{\theta}}_{1(F.sp.CF)} - \boldsymbol{\theta}_{1*}\right) \to_d \mathcal{N}\left(\mathbf{0}, \mathbb{A}_1^{-1}\mathbb{B}_1\mathbb{A}_1^{-1}\right) \tag{3.78}$$

As Lu and Wooldridge (2017), these results allow for non-normal errors and for three kinds of mis-specifications in the error vairance-covariance matrix. It allows for the mis-specification that *Quasi-GLS* procedure makes by ignoring the correlations between individuals across the groups. Further, these results also allow for the mis-specification of the structure in the $\Lambda_{g\eta}$ as well as inconsistencies in the estimation of the spatial parameters.

3.5.1.2 A consistent estimator for variance robust to cross-sectional structure

To facilitates valid inference from the estimator described in the paper, we use the heteroskedasticity and autocorrelation (HAC) estimator suggested in Lu and Wooldridge (2017). Define

$$\tilde{\mathtt{u}}_g \equiv \mathtt{y}_g - \mathtt{w}_g\widehat{\boldsymbol{\theta}}_{(F.sp.CF)} \tag{3.79}$$

and let

$$\widehat{\mathtt{u}}_g = \mathtt{w}_g' \widehat{\boldsymbol{\Omega}}_g^{-1} \tilde{\mathtt{u}}_g \tag{3.80}$$

The kernel function given in Lu and Wooldridge (2017) denoted by $\mathscr{K}_{BC}(d_{g_i h_j})$; is the function of $i-th$ observation in group g denoted by g_i and $j-th$ observation in group g denoted by h_j is defined as

$$\mathscr{K}_{BC}(d_{ij}) = \begin{cases} 1 - d_{g_i h_j}/d_* \; d_{g_i h_j} \leq d_* \\ 0 \; d_{g_i h_j} > d_* \end{cases} \tag{3.81}$$

$$\widehat{Avar}\left(\widehat{\boldsymbol{\theta}}_{2FCF}\right)_{T.S.robust} = \left[\sum_{g=1}^{G}\left(\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{w}_g\right)\right]^{-1}\left[\sum_{g=1}^{G}\sum_{h=1}^{G}\mathscr{K}_{BC}(d_{gh})\widehat{\mathtt{u}}_g\widehat{\mathtt{u}}_h'\right]\left[\sum_{g=1}^{G}\left(\mathtt{w}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathtt{w}_g\right)\right]^{-1} \tag{3.82}$$

where $\mathscr{K}_{BC}(d_{gh})$ is the kernel matrix defined for group g and h.

3.5.2 Asymptotics for 2SLS, Grouped 2SLS and Spatial GIV

Once we establish the assumptions on the nature of spatial dependence, the results of Jenish and Prucha (2012) can be applied to obtain the asymptotic properties of the other estimators described in the paper. The arguments of the proof will be similar to those given in the previous section.

Re-writing the estimators :

$$\widehat{\boldsymbol{\beta}}_{(2SLS)} = \left[\left(\sum_{i=1}^{N}\boldsymbol{z}_{1i}'\boldsymbol{x}_{2i}\right)\left(\sum_{i=1}^{N}\boldsymbol{x}_{2i}'\boldsymbol{x}_{2i}\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{x}_{2i}'\boldsymbol{z}_{1i}\right)\right]^{-1} \times \left[\left(\sum_{i=1}^{N}\boldsymbol{z}_{1i}'\boldsymbol{x}_{2i}\right)\left(\sum_{i=1}^{N}\boldsymbol{x}_{2i}'\boldsymbol{x}_{2i}\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{x}_{2i}'y_{1i}\right)\right]$$

$$\widehat{\boldsymbol{\beta}}_{(grpd.2SLS)} = \left[\left(\sum_{g=1}^{G}\boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G}\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{1g}\right)\right]^{-1} \times \left[\left(\sum_{g=1}^{G}\boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{N}\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{N}\boldsymbol{Z}_{2g}'Y_{1g}\right)\right]$$

$$\widehat{\boldsymbol{\beta}}_{(F.sp.GIV)} = \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{1g}\right)\right]^{-1} \times \left[\left(\sum_{g=1}^{G} \boldsymbol{Z}_{1g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_{2g}'\widehat{\Lambda}_{UU,g}^{-1}Y_{1g}\right)\right]$$

Using the arguments described in the previous section, the estimators are consistent and asymptotically normal with asymptotic variance given as below:

For the traditional 2SLS estimator:

$$Avar[\sqrt{G}\left(\widehat{\boldsymbol{\beta}}_{(2SLS)} - \boldsymbol{\beta}\right)] = [\mathbb{A}_{2SLS}]^{-1}[\mathbb{B}_{2SLS}][\mathbb{A}_{2SLS}]^{-1} \tag{3.83}$$

where $\mathbb{A}_{2SLS} \equiv \left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{z}_{1i}'\boldsymbol{x}_{2i})\right]\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{x}_{2i}'\boldsymbol{x}_{2i})\right]\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{x}_{2i}'\boldsymbol{z}_{1i})\right]$
and
$\mathbb{B}_{2SLS} \equiv \left\{\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{z}_{1i}'\boldsymbol{x}_{2i})\right]\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{x}_{2i}'\boldsymbol{x}_{2i})\right]\right\}\left\{Var\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{x}_{2i}'u_i)\right]\right\}$
$\left\{\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{z}_{1i}'\boldsymbol{x}_{2i})\right]\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\boldsymbol{x}_{2i}'\boldsymbol{x}_{2i})\right]\right\}'$

For the *Grouped 2SLS* estimator:

$$Avar[\sqrt{G}\left(\widehat{\boldsymbol{\beta}}_{(grpd.2SLS)} - \boldsymbol{\beta}\right)] = [\mathbb{A}_{grpd.2SLS}]^{-1}[\mathbb{B}_{grpd.2SLS}][\mathbb{A}_{grpd.2SLS}]^{-1} \tag{3.84}$$

where

$$\mathbb{A}_{grpd.2SLS} \equiv \left[\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g})\right] \times \left[\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g})\right]\left[\lim_{N\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{1g})\right]$$

and

$$\mathbb{B}_{grpd.2SLS} \equiv \left\{\left[\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g})\right]\left[\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g})\right]\right\} \times \left\{Var\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{2g}'U_g)\right]\right\}\left\{\left[\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{1g}'\boldsymbol{Z}_{2g})\right]\left[\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbb{E}(\boldsymbol{Z}_{2g}'\boldsymbol{Z}_{2g})\right]\right\}'$$

Finally, for the *Feasible Spatial GIV* estimator:

$$Avar[\sqrt{G}\left(\widehat{\boldsymbol{\beta}}_{(F.sp.GIV)}-\boldsymbol{\beta}\right)]=[\mathbb{A}_{F.sp.GIV}]^{-1}[\mathbb{B}_{F.sp.GIV}][\mathbb{A}_{F.sp.GIV}]^{-1} \tag{3.85}$$

where $\mathbb{A}_{F.sp.GIV} \equiv \lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\left[\mathbb{E}(\boldsymbol{Z}'_{1g}\Lambda^{-1}_{UU,g}\boldsymbol{Z}_{2g})[\mathbb{E}(\boldsymbol{Z}'_{2g}\Lambda^{-1}_{UU,g}\boldsymbol{Z}_{2g})]^{-1}\mathbb{E}(\boldsymbol{Z}'_{2g}\Lambda^{-1}_{UU,g}\boldsymbol{Z}_{1g})\right]$,
$\mathbb{B}_{F.sp.GIV} \equiv Var\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\sum_{g=1}^{G}\boldsymbol{Z}'_{1g}\Lambda^{-1}_{UU,g}\boldsymbol{Z}_{2g}\right)\left(\sum_{i=1}^{N}\boldsymbol{Z}'_{2g}\Lambda^{-1}_{UU,g}\boldsymbol{Z}_{2g}\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{Z}'_{2g}\Lambda^{-1}_{UU,g}U_g\right)\right]$

3.6 Monte Carlo Simulations

In this section, we illustrate the properties of our estimator using Monte Carlo simulations. The simulation results show the finite sample properties of the estimator and we compare them with the traditional 2SLS estimator.

3.6.1 Data Generating Process

The total number of observations is given by N. The data is generated in a $\sqrt{N}\times\sqrt{N}$ lattice (Figure 1). In other words, each observation is located on the intersections of the square lattice. This generates a coordinate system for the entire data given as $\{(l_1,l_2): l_1,l_2 = 1,2,3,...,\sqrt{N}\}$ where (l_1,l_2) can be interpreted as the *"longitude"* and *"latitude"* for each observation. We consider $N = 400$ such that the observations are generated in a 20×20 lattice.

The pairwise distances between the observation *i* and *j* is calculated using these coordinates. We can use several distance measures such as Euclidean distance, Manhattan distance, maximum coordinate-wise distance among others. In this paper, we only consider the Euclidean distance for simplicity.

3.6.2 Model

The primary equation for the outcome variable is defined as:

$$y_{1i} = \beta_{11} + \tau y_{2i} + u_i + \mathcal{N}(0,1) \tag{3.86}$$

$$\{\beta_{11}, \tau\} \equiv \{1,1\} \tag{3.87}$$

where $u_i, i = \{1,2,...,N\}$ is a spatially correlated vector of unobservables.

We have three models for the endogenous variable y_{2i}: Three models for the endogenous variable y_{2i}:

$$\text{M1} : y_{2i} = 1 + 3 * x_{2i} + \rho u_i + \mathcal{N}(0,1) \tag{3.88}$$

$$\text{M2} : y_{2i} = 1 + 3 * x_{2i} + 2 * x_{(2,i+1)} + \rho u_i + \mathcal{N}(0,1) \tag{3.89}$$

$$\text{M3} : y_{2i} = 1 + 3 * x_{2i} + 3 * x_{(2,i+1)} + \rho u_i + \mathcal{N}(0,1) \tag{3.90}$$

where x_{2i} is an exogenous variable that is generated as spatially correlated random variable. $x_{(2,i+1)}$ is the exogenous variable (x) of nearest neighbor of the observation i. We include this in the specification of the reduced form model of y_{2i} to increase the strength of the *additional instruments* that we get when we divide the data into groups.

ρ is the level of endogeneity captured as the relationship between y_{2i} and u_i. To study the effect of increasing endogeneity on our estimation, we consider following different levels of ρ:

$$\rho = 1,3 \tag{3.91}$$

3.6.3 Spatial Correlations

To generate a spatially correlated random vector U_N, we use the *Negative Exponential* functional form of the spatial correlation structure. Specifically, each element of the variance-covariance matrix of a spatially correlated random variable is defined as

$$\Lambda_{u,N}(ij) = \sigma exp\left(-\frac{d_{ij}}{\lambda_u}\right);\ \ \sigma \equiv 1 \tag{3.92}$$

where λ_u is the spatial parameter that reflects how quickly the correlations decrease with the increasing distance.

A spatial correlated random vector is generated as:

$$\Lambda_{u,N}^{1/2}\mathcal{N}(0,1) \tag{3.93}$$

where $\Lambda_{u,N}^{1/2}$ is the Cholesky Decomposition and $\mathcal{N}(0,1)$ is a vector of *i.i.d* standard normal variables

To study the effect of increasing spatial dependence on our estimation, we consider following different levels of λ_u:

$$\lambda_u = 0.1, 0.5, 0.8 \tag{3.94}$$

The spatial parameter for X_{2N} is 0.5 and X_{2N} is generated in the same way using the Cholesky Decomposition and standard normal distribution.

3.6.4 Results

We estimate the model using *Tradional 2SLS, Grouped 2SLS, Spatial GIV and Spatial Control Function* estimation methods. For the *grouped* estimation, we divide the data set into groups with each group containing 2 observations. We report the robust standard errors generated in the way described in section 3.5.

Table 3.1 lists the results of model M1, Table 3.2 lists the results of model M2 and Table 3.3 lists the results of model M3. We list the Mean, Standard Deviation, Standard Errors made robust to spatial correlation, Bias, Root MSE, Rejection rate and Coverage Probability of 95 percent confidence intervals. The rejection rates and the confidence intervals are calculated for the 95 percent confidence levels. As we can see, Grouped 2SLS, Spatial GIV and Spatial CF estimators give standard deviations that are almost 40 percent less than the standard deviation of the Traditional 2SLS even though they use only the *nearest* neighbor.

Comparing the performance of *Grouped 2SLS, Spatial GIV and Spatial Cf*, we find that the results are comparable to each other. However, *Spatial GIV* performs better in some cases both in terms of point estimates and inference.

Table 3.1: M1 $y_{2i} = 1 + 3 * x_{2i} + \rho u_i + \mathcal{N}(0,1)$

ρ=1,λ=.1	Mean	S.D	S.E	Bias	RMSE	Cov
2SLS	0.99919	0.054815	0.049795	0.00081	0.054794	0.958
G2SLS	0.9995	0.05479	0.049792	0.000505	0.054764	0.959
Sp. GIV	0.99952	0.054769	0.049793	0.000476	0.054744	0.958
Sp. CF	0.99961	0.054997	0.056054	0.000388	0.054971	0.979
ρ=1,λ=.5	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0015	0.064765	0.05694	-0.0015	0.06475	0.946
G2SLS	1.0018	0.064777	0.056908	-0.00179	0.064769	0.947
Sp. GIV	1.0013	0.06395	0.05642	-0.00132	0.063932	0.946
Sp. CF	1.0014	0.064082	0.057404	-0.00136	0.064064	0.965
ρ=1,λ=.8	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99641	0.079801	0.067279	0.003591	0.079842	0.947
Grpd. 2SLS	0.99669	0.07972	0.067256	0.003313	0.079749	0.947
Sp. GIV	0.99745	0.074992	0.064722	0.002547	0.074998	0.948
Sp. CF	0.99748	0.075113	0.058463	0.002516	0.075117	0.941
ρ=3,λ=0.1	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99668	0.055837	0.049877	0.003319	0.055908	0.951
G2SLS	0.99753	0.05564	0.049824	0.002473	0.055667	0.956
Sp. GIV	0.99751	0.055652	0.04983	0.002487	0.05568	0.956
Sp. CF	0.99743	0.056075	0.057293	0.002572	0.056105	0.983
ρ=3,λ=0.5	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0012	0.064568	0.056942	-0.00123	0.064547	0.957
G2SLS	1.0022	0.064487	0.056824	-0.00216	0.064491	0.956
Sp. GIV	1.0023	0.063988	0.05646	-2.31E-03	0.063998	0.961
Sp. CF	1.0022	0.064128	0.0586	-0.00216	0.064132	0.976
ρ=3,λ=0.8	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99737	0.079359	0.067875	0.002632	0.079363	0.947
G2SLS	0.99835	0.079082	0.067655	0.001655	0.079059	0.949
Sp. GIV	0.99923	0.075195	0.065005	0.000767	0.075161	0.95
Sp. CF	0.99918	0.075495	0.060856	0.000818	0.075462	0.96

Table 3.2: M2 $y_{2i} = 1 + 3 * x_{2i} + 2 * x_{(2,i+1)} + \rho u_i + \mathcal{N}(0,1)$

ρ=1,λ=.1	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99896	0.039498	0.036398	0.001043	0.039492	0.958
G2SLS	0.99898	0.036695	0.033626	0.001017	0.03669	0.963
Sp. GIV	0.99898	0.036693	0.033627	0.001016	0.036689	0.963
Sp. CF	0.99909	0.036719	0.037769	0.000914	0.036712	0.98
ρ=1,λ=.5	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0019	0.047544	0.041376	-0.00187	0.047557	0.956
G2SLS	1.0022	0.045458	0.039112	-0.00217	0.045487	0.951
Sp. GIV	1.0022	0.045453	0.039126	-0.00219	0.045483	0.95
Sp. CF	1.0021	0.045482	0.039987	-0.00214	0.04551	0.961
ρ=1,λ=.8	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99961	0.058086	0.04803	0.000392	0.058058	0.931
G2SLS	0.99969	0.057516	0.046503	0.000309	0.057488	0.93
Sp. GIV	0.99974	0.057342	0.046516	0.000259	0.057314	0.932
Sp. CF	0.99969	0.057457	0.042024	0.000305	0.057429	0.932
ρ=3,λ=.1	Mean	SD	SE	Bias	RMSE	Coverage
2SLS	0.99834	0.037525	0.035985	0.001658	0.037543	0.961
Grpd. 2SLS	0.99881	0.03528	0.033241	0.001193	0.035283	0.961
Sp. GIV	0.99882	0.035274	0.033242	0.001183	0.035276	0.961
Sp. CF	0.99878	0.035336	0.038309	0.001218	0.035339	0.983
ρ=3,λ=.5	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99857	0.049543	0.04195	0.001427	0.049539	0.949
G2SLS	0.99874	0.047417	0.039778	0.001262	0.04741	0.951
Sp. GIV	0.99887	0.047432	0.039793	1.13E-03	0.047422	0.952
Sp. CF	0.9989	0.047558	0.040621	0.001096	0.047547	0.954
ρ=3,λ=0.8	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0031	0.058043	0.048244	-0.00312	0.058097	0.954
G2SLS	1.0033	0.056827	0.046631	-0.00332	0.056895	0.951
Sp. GIV	1.0036	0.056699	0.046655	-0.00364	0.056787	0.955
Sp. CF	1.0036	0.056764	0.043687	-0.00362	0.056851	0.952

Table 3.3: M3 $y_{2i} = 1 + 3 * x_{2i} + 3 * x_{(2,i+1)} + \rho u_i + \mathcal{N}(0,1)$

ρ=1,λ=0.1	Mean	Std. D	Std. E	Bias	RMSE	Coverage
2SLS	0.99794	0.046829	0.043653	0.002064	0.046851	0.963
G2SLS	0.99761	0.035071	0.032641	0.002393	0.035135	0.958
Sp. GIV	0.99761	0.035071	0.032641	0.002389	0.035135	0.958
Sp. CF	0.99761	0.035042	0.036407	0.002388	0.035106	0.973
ρ=1,λ=0.5	Mean	Std. D	SE	Bias	RMSE	Cov
2SLS	0.99786	0.051009	0.044939	0.002138	0.051028	0.96
G2SLS	0.9971	0.041219	0.035274	0.002899	0.0413	0.948
Sp. GIV	0.99713	0.041216	0.035272	0.002867	0.041295	0.948
Sp. CF	0.99717	0.041251	0.038399	0.002833	0.041328	0.969
ρ=1,λ=0.8	Mean	SD	SE	Bias	RMSE	Cov
2SLS	0.99773	0.052182	0.046249	0.002267	0.052205	0.958
G2SLS	1.0001	0.04499	0.038229	-0.0001	0.044968	0.956
Sp. GIV	1.0002	0.044975	0.038228	-0.00016	0.044953	0.956
Sp. CF	1.0002	0.044902	0.040531	-0.00017	0.04488	0.971
ρ=3,λ=0.1	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0019	0.047362	0.043963	-0.00186	0.047374	0.962
G2SLS	1.0013	0.034815	0.032403	-0.00133	0.034823	0.967
Sp. GIV	1.0013	0.034815	0.032403	-0.00134	0.034823	0.967
Sp. CF	1.0013	0.035018	0.037006	-0.00134	0.035026	0.982
ρ=3,λ=0.5	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0017	0.050229	0.04479	-0.00174	0.050234	0.962
G2SLS	1.0018	0.039554	0.035241	-0.00176	0.039573	0.959
Sp. GIV	1.0018	0.039535	0.035239	-1.85E-03	0.039558	0.958
Sp. CF	1.0018	0.03953	0.039045	-0.00177	0.03955	0.982
ρ=3,λ=0.8	Mean	SD	SE	Bias	RMSE	Cov
2SLS	1.0001	0.051082	0.04666	-0.00011	0.051057	0.958
G2SLS	0.99973	0.043154	0.038117	0.000273	0.043134	0.954
Sp. GIV	0.99994	0.043109	0.038117	6.26E-05	0.043088	0.955
Sp. CF	0.99999	0.042903	0.041549	1.29E-05	0.042881	0.974

3.6.5 Performance of Spatial GIV

Since the *Spatial GIV* estimator performs better as compared to the *SpatialCF* estimator under no additional assumptions, we conduct another set of Monte Carlo simulations to analyze it's properties. Specifically, the outcome variable $y_1 i$ is generated as (3.85). The model for the endogenous variable y_2 is give as $2 * x_{2i} + \rho u_i + \mathcal{N}(0,1)$, where ρ indicates the endogeneity which is allowed to be 1 and 3. To highlight the benefit of incorporating the within-group spatial correlation between the observation, we increase the group size to $L = 4$ and for simplicity, we do not include the *Spatial CF*. In addition, we only use one additional instrument. Finally, the spatial parameter for x is 0.2 and spatial parameter for u takes the value $0.1, 0.5, 2.0, 5.0$ The results are given in Table 3.4.

As we can see, when we increase the group size, the *Spatial GIV* estimator gives us very good efficiency gains even in the presence of high spatial correlation.

Table 3.4: Performance of Spatial GIV

ρ=1,λ=0.1	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9956	0.0783	0.072	0.004438	0.0783
G2SLS	0.9965	0.078	0.0719	0.003528	0.078
Sp. GIV	0.9965	0.0655	0.0624	0.003496	0.0655
ρ=1,λ=0.5	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9957	0.0781	0.072	0.00434	0.0781
G2SLS	0.9966	0.0777	0.0719	0.003421	0.0777
Sp. GIV	0.9967	0.0644	0.0617	0.003269	0.0644
ρ=1,λ=2.0	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9958	0.0776	0.0718	0.004207	0.0777
G2SLS	0.9967	0.0773	0.0717	0.003275	0.0773
Sp. GIV	0.9971	0.0607	0.0594	0.002886	0.0607
ρ=1,λ=5.0	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9956	0.0774	0.0715	0.00443	0.0774
2SLS	0.9964	0.0771	0.0713	0.003607	0.0771
Sp.GIV	0.9971	0.059	0.0584	0.002864	0.059
ρ=3,λ=0.1	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9941	0.0789	0.0727	0.005869	0.0791
G2SLS	0.9967	0.0781	0.0723	0.003284	0.0781
Sp. GIV	0.996	0.0657	0.0629	0.00397	0.0657
ρ=3,λ=0.5	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9943	0.0786	0.0727	0.005742	0.0788
G2SLS	0.9969	0.0776	0.0722	0.003136	0.0776
Sp. GIV	0.9964	0.0646	0.0621	0.003634	0.0646
ρ=3,λ=2.0	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9944	0.0782	0.0725	0.005576	0.0783
G2SLS	0.9971	0.0769	0.0718	0.002937	0.0769
Sp. GIV	0.9972	0.0609	0.0596	0.002784	0.0609
ρ=3,λ=5.0	Mean	Std. D	S.E	Bias	RMSE
2SLS	0.9943	0.0778	0.0721	0.005665	0.0779
Grpd. 2SLS	0.9966	0.0766	0.0714	0.003374	0.0766
Sp. GIV	0.9974	0.0592	0.0584	0.002618	0.0592

3.7 Conclusion and Future Research

In this paper, we propose a computationally simple estimation procedure to account for spatially correlated errors in linear econometric models with endogenous variables in a cross-section dataset. We describe how by dividing the observations into groups according to the distances between them and then taking account for only the spatial dependence of observation within a group gives us noticeable efficiency gains; even though we ignore the correlations between across-group observations. We suggest control-function approach to account for endogeneity in the model that gives us intuitive estimating equations. The estimating procedure in the paper provides an empirical researcher with a powerful tool that achieves more precise estimator while avoiding tedious calculations that are unavoidable when we attempt to take account of the entire correlation structure of all the observations.

The estimation strategy described in this paper can also be extended to incorporate non-parametric methods of estimation. For example, we can use *method of sieves* in the second-step of our estimation procedure. To explain the motivation of doing a sieve estimation in the second-step, consider the case where $L = 2$. In that case, $\mathbf{\Pi}_g$ is a deterministic function of the distance between the two observations in a group g denoted by d_g. We can allow $\mathbf{\Pi}_g$ to be a flexible function of d_g and this motivates a sieve approximation of $\mathbf{\Pi}_g$.

Further, when we have two observations in a group, sieve methods can also be implemented to estimate the spatial dependence without restricting the functional form for spatial correlation structure. Since the correlation between the two observations in each group depend on the spatial parameter λ_η and the distance between the two observations d_g, we can approximate the correlation function using the method of sieves. A formal description of the sieve methods in our estimation procedure and the asymptotic analysis is left for future research.

BIBLIOGRAPHY

BIBILIOGRAPHY

- Adamchik, V. & V. Bedi. (1983). Wage differentials between the public and the private sectors: Evidence from an economy in transition. Labour Economics 7: 203âĂŞ224.
- Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. Journal of Human resources, 40(4), 791-821.
- Arbia, G. (2006), Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence . Springer.
- Biorn, E. (1981). Estimating economic relations from incomplete cross-section/time-series data. Journal of Econometrics, 16(2), 221-236.
- Baltagi, B. H. (1985). Pooling cross-sections with unequal time-series lengths. Economics Letters, 18(2), 133-136.
- Baltagi, B. (2001). Econometric analysis of panel data, second edition. West Sussex, UK:Wiley
- Baltagi, B. H., & Chang, Y. J. (1994). Incomplete panels: A comparative study of alternative estimators for the unbalanced one-way error component regression model. Journal of Econometrics, 62(2), 67-89.
- Baltagi, B. H., Song, S. H., & Jung, B. C. (2002). A comparative study of alternative estimators for the unbalanced twoway error component regression model. The Econometrics Journal, 5(2), 480-493.
- Baltagi, B. H., & Song, S. H. (2006). Unbalanced panel data: A survey. Statistical Papers, 47(4), 493-523.
- Bell, K. P., and Bockstael, N. E. (2000). "Applying the generalized-moments estimation approach to spatial problems involving micro-level data." The review of economics and statistics, 82(1), 72-82.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. Handbook of econometrics, 6, 5549-5632.
- Chen, S., & Khan, S. (2003). Semiparametric estimation of a heteroskedastic sample selection model. Econometric Theory, 19(06), 1040-1064.
- Christakos, G. (1987), "On the Problem of Permissible Covariance and Variogram Models," Water Resources Research, 20, 251-265.

- Conley,T.G. (1999), "GMM Estimation with Cross Sectional Dependence" Journal of Econometrics, Volume 92, Issue 1, 1-45.
- Cressie, N.A.C. (1993), Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley & Sons Inc., a Wiley-Interscience Publication.
- Cressie, N and Lahiri, S.N (1993): "The Asymptotic Distribution of REML Estimators" Journal of Multivariate Analysis, 45, 217-233
- Donald, S. G. (1995). Two-step estimation of heteroskedastic sample selection models. Journal of Econometrics, 65(2), 347-380.
- Dubin, R.A. (1988), "Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms" The Review of Economics and Statistics , Vol. 70, No. 3, 466-474.
- Dubin, R. A. (1998). "Spatial autocorrelation: a primer". Journal of housing economics, 7(4), 304-327.
- Glass, A. J., Kenjegalieva, K., and Sickles, R. (2012). "The economic case for the spatial error model with an application to state vehicle usage in the US" working paper.
- Guggenberger(2010) :The Impact of a Hausman Pretest on the Size of a Hypothesis Test: the Panel Data Case. Journal of Econometrics 156(2), 337-343.
- Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the Econometric Society, 1251-1271.
- Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. Econometrica: Journal of the Econometric Society, 1377-1398.
- Hsiao, C., Shen, Y., Wang, B., & Weeks, G. (2008). Evaluating the effectiveness of Washington state repeated job search services on the employment rate of prime-age female welfare recipients. Journal of econometrics, 145(1), 98-108.
- Hahn, J., Liao, Z., & Ridder, G. (2018). Nonparametric two-step sieve M estimation and inference. Econometric Theory, 1-44.
- Hansen, L. P., & Richard, S. F. (1987). The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. Econometrica: Journal of the Econometric Society, 587-613.
- Kelejian, H.H., Prucha, I.R., 1999. "A generalized moments estimator for the autoregressive parameter in a spatial model." International Economic Review 40, 509âĂŞ533
- Kelejian, H.H., Prucha, I.R., 2001. "On the asymptotic distribution of the Moran I test statistic with applications." Journal of Econometrics 104, 219âĂŞ257.

- Kelejian, H.H. and Prucha I.R. (2007), "HAC estimation in a spatial framework," Journal of Econometrics, Volume 140, Issue 1, September 2007, Pages 131-154
- Kim, K. I. (2013). An Alternative Efficient Estimation of Average Treatment Effects. Journal of Market Economy, 42(3), 1-41.
- Kyriazidou, E. (1997). Estimation of a panel data sample selection model. Econometrica: Journal of the Econometric Society, 1335-1364.
- Lee, L. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. International Economic Review 19: 415âĂŞ433.
- Lee, L.F.(2004), "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models," Econometrica, Vol. 72, 1899-1925
- Lu, C., and Wooldridge, J. M. (2017). Quasi-generalized least squares regression estimation with spatial data. Economics Letters, 156, 138-141.
- Lu, C. (2013). Linear and nonlinear estimation with spatial data.
- Maddala, G. S. 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.
- Mardia, K.V. and R.J. Marshall (1984), "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression" Biometrika 71, 135-146.
- Mincer, J. and S. Polachek. 1974. Family investment in human capital: Earnings of women. Journal of Political Economy (Supplement) 82: S76âĂŞS108.
- Mundlak, Y. (1978). On the pooling of time series and cross section data. Econometrica: journal of the Econometric Society, 69-85.
- Murtazashvili, I., & Wooldridge, J. M. (2016). A control function approach to estimating switching regression models with endogenous explanatory variables and endogenous switching. Journal of Econometrics, 190(2), 252-266.
- Papke, L. E. (2005). The effects of spending on test pass rates: evidence from Michigan. Journal of Public Economics, 89(5), 821-839.
- Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. Journal of Econometrics, 145(1), 121-133.
- Rochina-Barrachina, M. E. (1999). A new estimator for panel data sample selection models. Annales d'Economie et de Statistique, 153-181.
- Semykina, A., & Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. Journal of Econometrics, 157(2), 375-380.

- Stein, M.L. (1999) "Interpolation of Spatial Data" New York: Springer-Verlag
- Thorst, R. 1977. Demand for housing: A model based on inter-related choices between owning and renting. Ph.D. dissertation, University of Florida.
- Tobler, W. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." Econ. Geog. Supplement 46, 234âĂŞ240.
- Vella, F. (1998). Estimating models with sample selection bias: a survey. Journal of Human Resources, 127-169.
- Verbeek, M., & Nijman, T. (1992). Testing for selectivity bias in panel data models. International Economic Review, 681-703.
- Wang, H., Iglesias, E. and Wooldrige, J.M. (2012). "Partial Maximum Likelihood Estimation of Spatial Probit Models," Journal of Econonometrics.
- Wansbeek, T., & Kapteyn, A. (1989). Estimation of the error-components model with incomplete panels. Journal of Econometrics, 41(3), 341-361.
- Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. Journal of econometrics, 68(1), 115-132.
- Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. Journal of Econometrics, 90(1), 77-97.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.
- Wooldridge, J.M. (2011), Econometric Analysis of Cross Section and Panel Data. MIT.
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. Journal of Human Resources, 50(2), 420-445.
- Wooldridge, J. M. (2016). Correlated random effects models with unbalanced panels. Manuscript (version July 2009) Michigan State University.