ESSAYS ON HETEROGENEITY IN ECONOMETRIC MODELS

By

Shengwu Shang

A DISSERTATION

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

Economics–Doctor of Philosophy

2013

ABSTRACT

ESSAYS ON HETEROGENEITY IN ECONOMETRIC MODELS

By

Shengwu Shang

The dissertation consists of three parts and the theme is to deal with heterogeneity in econometrics models for positive response variables. The first part studies the models with multiplicative heterogeneity for cross sectional data; the multiplicative heterogeneity can be transformed from the log linear model with additive heterogeneity. We introduce the notion of Average Partial Effect (APE) and Conditional APE (CAPE); the estimators and their asymptotic distribution are proposed. In order to catch the positivity of the unknown conditional expectation function of the unobserved heterogeneity, we borrow the idea of power series approximation of unknown function in Newey (1993, 1994) and develop an "exponential sieves" estimator for CAPE suggested in Wooldridge (1992a).

The second part of the dissertation pertains to extending results for CAPE in chapter 1 for panel data sets. First, Using the models in Wooldridge (1999), We compare three main estimation methods for positive response variable– FE method for log linear model (LFE), Poisson Quasi-Maximum Likelihood (PQML) and Generalized Method of Moment (GMM) – by Monte Carlo Simulation and real life data set. It is not surprising that LFE estimator is not consistent when PQML is; however, we do find circumstance where both LFE and PQML estimators are consistent plus LFE is more efficient. With this regard, we introduce GMM to improve the efficiency of PQML estimator as well as keeping the consistency; this way also finds a solution to the problem raised in Wooldridge (1999). From the simulation results, we find that GMM can reduce the standard error of PQML estimator by almost a half. Second, an "exponential sieves" estimator for CAPE is proposed under panel data setting; the result automatically extends the results in Ai and Norton (2008) from cross sectional setting to panel data models. Third, We also apply the GMM to a US domestic

airlines data set and the result shows that GMM improves the efficiency by 10% compared with PQML.

The third part investigates the effect of spatial correlation for fractional response variable. By a MEAP data of Michigan in 2009/2010 school year, we investigate again the effect of school financing reform on school performance which is studied by Papke (2005, 2008), Pake and Wooldridge (2008); we use both level math test pass rate (linear case)and its log odds ratio (nonlinear) as dependent variable to run OLS and GLS regression; Conley (1999)'s spatial dependence corrected standard errors are calculated and find that the statistical significance for some regressors hinges on the choice of cut off points ; however there do exist other factors whose statistical significance is robust to the choice. This way we shed some light on how to pick the right window size. Moreover, by transforming LOR back to level rate, we find the spending effect estimated from linear model is about $4 \sim 6\%$ higher than from nonlinear one.

This thesis is dedicated to my mom, Jinxiang Mao, and my dad, Chuchuan Shang.

ACKNOWLEDGEMENTS

Understanding econometrics, in my opinion, is like peeling an onion. Each time you peel away one layer you discover that another awaits. Each time you think you reach some understanding of intuition, estimation and application, you later discover that much more needs to be understood. I fear that I will never reach the core of the onion. But this is what makes the subject of econometrics so exciting.

I would never been able to finish my dissertation without guidance, suggestions and support from several people. First, my deepest appreciation goes to my advisor, Professor Jeffrey M. Wooldridge. Albeit lots of challenges and struggles on my studies at Michigan State University, I have never regretted working with him. He accepted me as one of his students when I felt lost and hopeless four years ago and helped me get through the most difficult time in my life. He provided me his perspective and helpful comments all the time and his contributions to this dissertation were invaluable. I could never have reached the depths of this dissertation without his mentorship and insightful advice. In my mind, he is a genuine scientist; He sets an example of a world-class researcher and teacher for me to model myself after in my career life.

I also want to express my deep appreciation to Professor Peter Schmidt for his guidance in my research. I did not take any courses from him, which is really regretful; but his office door was always open for me whenever I had questions. His comments such as “econometric research is to tell a story” really were eye opening for me. He amazed me with his thorough understanding of theories and ability to expressing them in simple words. In addition, I cannot say enough thanks to Professor Steve Woodbury, who was always sharp on comments and had an open mind, providing formidable advice that greatly improved this dissertation. My special gratitude goes to Professor Lijian Yang, who has been a big brother for me and ready for help throughout my entire graduate career. Beside them, I would like to sincerely thank the following faculty and staff members at MSU: Leslie E. Papke, Jack Meyer, Byron Brown, Timothy Vogelsang, Richard Baillie, Todd Elder, Tony Doblas Madrid, Kun Ho Kim, Dorothy Pathak, Pramod Pathak, Ed Mahoney, Ashton Shortridge, Andrew O. Finley, Michael Frazier, R. V. Ramamoorthi and Margaret Lynch. I have been very fortunate to meet them at MSU and they all helped me at various stages of my doctoral study.

While I cannot mention everyone, I would like to thank my fellow graduate students, particularly Jeff Brown, Quentin Brummet, Paul Burkander, Dooyeon Cho, Sanders Chang, Simon Chang, Hon Foong Cheah, Yu-Wei Chu, Myoung-Jin Keay, Do Won Kwak, Jinyoung Lee, Cuicui Lu, Tan Lu, Kritkorn Nawakitphaitoon, Ilya Rahkovsky, Seunghwa Rho, Monthien Satimanon, Valentin Verdier, Wei-Siang Wang, Yali Wang. We had numerous discussions and I appreciate their thoughtful comments and enjoyable conversation.

I thank my parents Jinxiang Mao and Chuchuan Shang for their unwavering love and support throughout this endeavor; without their encouragement, I could have never accomplished this. I would like to express my gratitude to my elder brother Shengwen Shang , elder sister Shenghua Shang and younger brother Shenggang Shang who took my part of the responsibility for my parents, which made it possible for me to stay single-minded on my studies for so long. Last but not least, I would like to thank my wife who was always there cheering me up and stood by me through the good and bad times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

ON ESTIMATING PARTIAL EFFECTS AFTER RETRANSFORMATION

1.1 Introduction

Strictly positive response variables are very common in economics and other social sciences. Just a few examples include prices, populations, and firm sales. If $Y > 0$ is the response that we would like to explain, the most common approach for modeling Y is to use a linear model for its natural log, $\log(Y)$, and then estimate the linear model using an appropriate technique – usually ordinary least squares (OLS) or instrumental variables (IV).

There are at least two reasons modeling $\log(Y)$ may not be sufficient. First, one might wish to predict Y, not $\log(Y)$. When the prediction of Y is based on its expected value conditional on a vector of covariates – say, $\mathbf{X}$ – in general there is no way to recover $E(Y|\mathbf{X})$ from $E[\log(Y)|\mathbf{X}]$. The difficulty in predicting Y given a model for $\log(Y)$ has long been recognized, and solutions are available under varying levels of assumptions. Duan (1983) covers the case where an additive error in the model for $\log(Y)$ is assumed independent of $\mathbf{X}$, and this method is covered even in some introductory econometrics texts [for example, Wooldridge (2009, Chapter 6)]. Under distributional assumptions, such as assuming Y given $\mathbf{X}$ follows a lognormal distribution, parametric heteroskedasticity in $\text{Var}[\log(Y)|\mathbf{X}]$ is easily allowed.

More recently, Ai and Norton (2008) provide a semiparametric approach that produces consistent predictions under weak assumptions, although the approach they use allows the possibility that predictions of Y can be negative. Wooldridge (1992) proposes direct estimation of $E(Y|\mathbf{X})$ via quasi-likelihood methods using flexible functional forms that ensure nonnegative predictions. Of course, nonparametric methods, of the type covered in Li and Racine (2007), can be used, too.

Related to the prediction issue is the calculation of partial effects. Wooldridge (1992) makes a case for basing partial effects on $E(Y|\mathbf{X})$ (in cases where the elements of $\mathbf{X}$ are appropriately "exogenous"). Such an approach underlies the work by Ai and Norton (2008), who begin with a linear model for $\log(Y)$ but then employ nonparametric methods to recover an estimate $E(Y|\mathbf{X})$ – without placing further restrictions on $D(Y|\mathbf{X})$, the conditional distribution of Y given $\mathbf{X}$. But basing partial effects on $E(Y|\mathbf{X})$ (or even some other feature of the conditional distribution, such as the median) is not the only possibility. Lately, the notion of an average partial effect (APE) has become important in applied econometrics; see, for example, Wooldridge (2005, 2010). The APE is closely tied to Blundell and Powell's (2003) average structural function (ASF), which is defined by averaging out unobservables from a "structural" model. One potentially important implication of the ASF approach has largely gone unnoticed: the ASF approach can provide very different partial effects than that based on $E(Y|\mathbf{X})$. A heteroskedastic probit example is provided by Wooldridge (2005): the partial effects based on $E(Y|\mathbf{X}) = P(Y = 1|\mathbf{X})$ and those based on the ASF need not even have the same sign, let alone similar magnitudes.

One reason the APE/ASF concept is appealing is because it can be easily applied to cases where explanatory variables should be treated as endogenous. Consequently, the focus on the ASF has led to a widely applicable class of control function estimators in nonlinear models – see Blundell and Powell (2003, 2004). Further, in a broad class of models, the sign of an APE and an underlying parameter on the covariate of interest are the same. In this paper we highlight another useful feature of the APE approach: it provides justification for Duan-type retransformation estimators even when Duan's key assumption – independence between the underlying error U and the covariates $\mathbf{X}$ – is violated.

We also consider the notion of a "conditional" average partial effect (CAPE) (for example, Wooldridge, 2004, 2005), which is considered more generally as the "local average response" in Altonji and Matzkin (2005). Interestingly, general consideration of CAPEs leads to essentially the same estimation problem described in Ai and Norton (2008), except that our approach here allows for endogenous explanatory variables. Plus, when on restricts the nature of the condititioning set, then simple strategies are available that do not require complicated nonparametric estimation.

After presenting the model and definitions of partial effects in Section 2, Section 3 discusses estimation of APEs – which, in the current seeting, turns out to be nothing more than extending Duan's (1983) "smearing" estimate to more general settings. We consider the estimation of CAPEs in Section 4, and Section 5 contains a brief conclusion. Technical derivations are contained in an appendix.

1.2 The Model and Partial Effects

The setup we consider is a standard linear model with $\log(Y)$ as the dependent variable. Initially assume that, in the population,

$$\log(Y) = \mathbf{X}\boldsymbol{\beta} + U \tag{1.2.1}$$

$$E(U|\mathbf{X}) = 0, \tag{1.2.2}$$

where $\mathbf{X}$ is a $1 \times K$ vector of covariates with first element $\mathbf{X}$ unity. We could consider nonlinear regression functions in place of $\mathbf{X}\boldsymbol{\beta}$, but retransformation methods are almost always applied when the transformed variable follows a linear-in-parameters model. As in Duan (1983), we could also consider other strictly monotonic transformations of Y but the natural logarithm is by far the most popular.

Given equation (1.2.1), we can write

$$Y = \exp(\mathbf{X}\boldsymbol{\beta} + U) = \exp(\mathbf{X}\boldsymbol{\beta})\exp(U). \tag{1.2.3}$$

Following the discussion in the introduction, we base the partial effects of interest on this equation because we are interested in how the X_j affect Y, not $\log(Y)$.

Ai and Norton (2008) focus on the conditional mean $E(Y|\mathbf{X})$, which can be written generally as

$$E(Y|\mathbf{X}) = \exp(\mathbf{X}\boldsymbol{\beta})E[\exp(U)|\mathbf{X}] \equiv \exp(\mathbf{X}\boldsymbol{\beta})r(\mathbf{X}) \tag{1.2.4}$$

where $r(\mathbf{X}) \equiv E[\exp(U)|\mathbf{X}]$. If X_j is a continuous variable then the partial effect on $\mu(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ is

$$\frac{\partial\mu(\mathbf{x})}{\partial x_j} = \frac{\partial E(Y|\mathbf{X} = \mathbf{x})}{\partial x_j} = \beta_j \exp(\mathbf{x}\boldsymbol{\beta})r(\mathbf{x}) + \exp(\mathbf{x}\boldsymbol{\beta})\frac{\partial r(\mathbf{x})}{\partial x_j}. \tag{1.2.5}$$

To estimate this partial effect we need to estimate $r(\cdot)$ in addition to $\boldsymbol{\beta}$ – the problem considered by Ai and Norton (2008). We reconsider this problem from a different perspective in Section 4.

Equation (1.2.5) clearly shows that the partial effect of x_j on $E(Y|\mathbf{X} = \mathbf{x})$ need not even have the same sign as β_j, and the magnitude of the partial effect can depend on $r(\cdot)$ in a rather complicated way. In the special case $E[\exp(U)|\mathbf{X}] = \exp(\mathbf{X}\boldsymbol{\delta})$, $E(Y|\mathbf{X}) = \exp[\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\delta})]$, and so the partial effects of x_j on $E(Y|\mathbf{X} = \mathbf{x})$ is the same sign as $\beta_j + \delta_j$.

There is another definition of partial effects that is easier to summarize and, as it turns out, also easier to estimate. Following Blundell and Powell (2003), the average structural function (ASF) is defined as

$$ASF(\mathbf{x}) = E[\exp(\mathbf{x}\boldsymbol{\beta})\exp(U)] = \exp(\mathbf{x}\boldsymbol{\beta})E[\exp(U)] \equiv \eta \exp(\mathbf{x}\boldsymbol{\beta}), \tag{1.2.6}$$

where

$$\eta \equiv E[\exp(U)].$$

The definition of the ASF is related to the notion of a "Marshallian structural function" defined in Heckman (2001). In defining the ASF it is important to see that the covariates are held fixed at specified values, with U averaged out. Once the ASF is obtained, we can see how this function changes as the x_j change. As discussed in Wooldridge (2005), the ASF is closely tied to the notion of an average partial effect (APE). From (1.2.3), the partial effect of X_j on Y is

$$\frac{\partial Y}{\partial X_j} = \beta_j \exp(\mathbf{X}\boldsymbol{\beta}) \exp(U).$$

To get the APE we average U out of the partial effect for given covariate values $\mathbf{x}$. In other words, the APE at $\mathbf{x}$ is

$$APE_j(\mathbf{x}) = E\left[\beta_j \exp(\mathbf{x}\boldsymbol{\beta}) \exp(U)\right] = \eta[\beta_j \exp(\mathbf{x}\boldsymbol{\beta})],$$

and this is easily seen to be the partial derivative of the ASF. A similar argument works if we use discrete changes in x_j rather than a calculus approximation.

An attractive feature of the ASF is that its definition does not require us to take a stand on possible dependence between U and $\mathbf{X}$. The definition is unchanged even if U and $\mathbf{X}$ are correlated. When U and $\mathbf{X}$ are independent, $\mu(\mathbf{x}) = ASF(\mathbf{x})$, but generally these quantities differ when $D(U|\mathbf{X})$ depends on $\mathbf{X}$ – even if $\mathbf{X}$ is exogenous in the sense of assumption (1.2.2). The potential difference between average partial effects and partial effects based on $E(Y|\mathbf{X} = \mathbf{x})$ has been pointed out by Wooldridge (2005), who uses a probit model with heteroskedasticity to illustrate that when U and $\mathbf{X}$ are not independent, the APEs and partial effects based on $E(Y|\mathbf{X})$ will be different – perhaps very different. Unfortunately, it does not seem possible to resolve the issue of how one *should* compute partial effects. The choice between $E(Y|\mathbf{X} = \mathbf{x})$ and $ASF(\mathbf{x})$ is essentially one of preference. The main contribution of the current paper is to show that it is easy to estimate the ASF in the retransformation context without taking a stand on $D(U|\mathbf{X})$.

One shortcoming with APEs is that the heterogeneity is averaged across the entire population whereas the covariates are set at specific values. Altonji and Matzkin (2005) argue that finding partial effects conditional on specific outcomes $\mathbf{x}$ is generally more useful – what they call a "local average response" (LAR). Wooldridge (2005) also discusses partial effects average over a subset of the population, based on observable characteristics.

As a general statement, suppose

$$Y = g(\mathbf{X}, U),$$

so that, for a continuous covariate, the partial effect at $\mathbf{X} = \mathbf{x}$ is $\partial g(\mathbf{x}, U)/\partial x_j$. Now suppose we wish to compute the expected value of this partial effect not across the entire distribution of U, but for the subset of the population with $\mathbf{X} = \mathbf{x}$. Then we can define a conditional average partial effect (CAPE) as

$$CAPE_j(\mathbf{x}) = E\left[\left.\frac{\partial g(\mathbf{x}, U)}{\partial x_j}\right| \mathbf{X} = \mathbf{x}\right]$$

In the exponential case, the CAPE can be expressed as

$$\begin{aligned} CAPE_j(\mathbf{x}) &= \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) E[\exp(U)|\mathbf{X} = \mathbf{x}] \\ &\equiv \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) r(\mathbf{x}) \end{aligned}$$

Note that this is different than basing partial effects on $E(Y|\mathbf{X} = \mathbf{x})$, yet we need to estimate the same function, $r(\mathbf{x})$. From an interpretation standpoint CAPE has the convenient feature that it is the same sign as β_j; we simply need to compute the scale factor, $r(\mathbf{x})$, that multiplies $\beta_j \exp(\mathbf{x}\boldsymbol{\beta})$. The function $APE_j(\mathbf{x})$ replaces $r(\mathbf{x})$ with the mean value $\eta = E[r(\mathbf{X})] = E[\exp(U)]$.

As with the APE, the CAPE makes sense even if $\mathbf{X}$ includes endogenous elements. In fact, the CAPE includes as a special case average treatment effects for various subpopulations when treatment assignment is endogenous. (With binary treatments we would use changes, not derivatives.)

1.3 Estimating the APEs

The expression for $ASF(\mathbf{x})$ makes it clear that a $\sqrt{N}$-consistent estimator of $ASF(\mathbf{x})$ is available if $\sqrt{N}$-consistent estimators of $\boldsymbol{\beta}$ and η are available. By contrast, the dependence of $\mu(\mathbf{x})$ on the nonparametric function $r(\mathbf{x})$ means that partial effects based on $E(Y|\mathbf{X}=\mathbf{x})$ are not generally estimable at the $\sqrt{N}$-rate (because nonparametric rates of convergence are slower – often much slower – than $\sqrt{N}$). Thus, we can estimate the ASF much more precisely than $E(Y|\mathbf{X}=\mathbf{x})$.

Because $\mu(\mathbf{x})$ and $ASF(\mathbf{x})$ both depend on $\boldsymbol{\beta}$, we first need to consistently estimate $\boldsymbol{\beta}$. It is very common to use OLS as the estimator of $\boldsymbol{\beta}$, even though it may not be asymptotically efficient under $E(U|\mathbf{X})=0$. (For example, a weighted least squares estimator that attempts to exploit nonconstant $\text{Var}(U|\mathbf{X})$ could be more efficient.)

As is well known, the assumption $E(U|\mathbf{X})=0$ does not substantively restrict $r(\mathbf{x}) \equiv E[\exp(U)|\mathbf{X}=\mathbf{x}]$: it can be virtually any positive function of $\mathbf{x}$. Of course, if we suitably restrict $D(U|\mathbf{X})$ then we can usually find $r(\mathbf{x})$. A useful assumption is that U and $\mathbf{X}$ are independent, a case considered by Duan (1983); see also Wooldridge (2009, Section 6.4). Then

$$r(\mathbf{X}) = E[\exp(U)|\mathbf{X}] = E[\exp(U)] = \eta, \tag{1.3.1}$$

and it follows that

$$E(Y|\mathbf{X}=\mathbf{x}) = \eta \exp(\mathbf{x}\boldsymbol{\beta}) = ASF(\mathbf{x}). \tag{1.3.2}$$

[If we specify the distribution of U, then we can sometimes write η in terms of higher moments of U. For example, if $U \sim Normal(0,\sigma^2)$ then $\eta = \exp(\sigma^2/2)$; see Wooldridge (2009, Section 6.4).]

In the case considered by Duan (1983) – where U and $\mathbf{X}$ are assumed to be independent – estimation of η is straightforward. First, by the law of large numbers,

$$N^{-1}\sum_{i=1}^{N}\exp(U_i) \xrightarrow{p} \eta. \qquad (1.3.3)$$

The average is not an estimator because we do not observe the U_i. Instead, given a random sample $\{(\mathbf{X}_i, Y_i) : i = 1, ..., N\}$, obtain $\hat{\boldsymbol{\beta}}$ from the OLS regression

$$\log(Y_i) \text{ on } \mathbf{X}_i,\ i = 1, ..., N \qquad (1.3.4)$$

and then let $\hat{U}_i = \log(Y_i) - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ be the OLS residuals. Because $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ it is not suprising that

$$\widehat{\eta} = N^{-1}\sum_{i=1}^{N}\exp(\hat{U}_i) \qquad (1.3.5)$$

is generally consistent for η. Wooldridge (2010, Lemma 12.1) contains a general result that implies consistency under weak regularity conditions. Because U and $\mathbf{X}$ are independent, we estimate $\mu(\mathbf{x})$ and $ASF(\mathbf{x})$ in exactly the same way:

$$\hat{\mu}(\mathbf{x}) = \hat{\eta}\exp(\mathbf{x}\hat{\boldsymbol{\beta}}) = \widehat{ASF}(\mathbf{x}). \qquad (1.3.6)$$

Wooldridge (2010, Problem 12.17) can be used to show $\sqrt{N}(\hat{\eta} - \eta)$ has a limiting normal distribution and to find its asymptotic variance. Below we consider the problem of estimating the joint asymptotic variance under very weak assumptions.

For estimating the ASF, there is an important point about Duan's estimator in (1.3.6): it is a consistent estimator of $\eta = E[\exp(U)]$ even when U and $\mathbf{X}$ are dependent. In fact, the most we need to assume is

$$E(\mathbf{X}'U) = \mathbf{0} \qquad (1.3.7)$$

as this is sufficient for OLS to consistently estimate $\boldsymbol{\beta}$. Duan (1983) was interested in recovering $E(Y|\mathbf{X})$, which is why he assumed independence between U and $\mathbf{X}$; see also Abrevaya

(2002), who obtained the asymptotic variance of the predictions under the assumption of independence. But when we view Duan's estimator as estimating the scale factor that appears in the ASF, the estimator is generally consistent provided $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$. In settings where we intend $\mathbf{X}$ to be exogenous in (1.2.1), it suffices to assume (1.3.7) – in which case the OLS estimator from (1.3.3) is consistent but not necessarily unbiased.

In Section 2 we mentioned how the definition of the ASF is unchanged regardless of dependence between U and $\mathbf{X}$. As can be seen from equation (1.2.6), we consistently esimate $ASF(\mathbf{x}) = \eta \exp(\mathbf{x}\boldsymbol{\beta})$ provided we have consistent estimators of η and $\boldsymbol{\beta}$, and (1.3.5) shows we just need a consistent estimator of $\boldsymbol{\beta}$ to consisently estimate η. Having some elements of $\mathbf{X}$ endogenous in (1.2.1) in the sense that $Cov(\mathbf{X}, U) \neq \mathbf{0}$ causes no problems for estimating $ASF(\mathbf{x})$ provided we have suitable instrumental variables. In particular, suppose we have a $1 \times L$ vector satisfying

$$\begin{aligned} E(\mathbf{Z}'U) &= \mathbf{0} \\ rank\ E(\mathbf{Z}'\mathbf{X}) &= K \\ rank\ E(\mathbf{Z}'\mathbf{Z}) &= L \end{aligned} \tag{1.3.8}$$

Under these assumptions, the 2SLS estimator (as well as other generalized method of moments estimators) is consistent for $\boldsymbol{\beta}$; see, for example, Wooldridge (2010, Chapter 5). Then, we can let $\widehat{\boldsymbol{\beta}}$ be the 2SLS estimator from

$$\log(Y_i) = \mathbf{X}_i\boldsymbol{\beta} + U_i \tag{1.3.9}$$

using IVs $\mathbf{Z}_i$. The $\widehat{U}_i$ are now the 2SLS residuals, and $\widehat{\eta}$ is still computed from (1.3.5). Notice that it would be meaningless in this context to base partial effects on $E(Y|\mathbf{X})$ or $E(Y|\mathbf{X}, \mathbf{Z})$, whereas the ASF can have a causal interpretation. We summarize the above as the following theorem.

Theorem 1.3.1. *As assumptions in equation* (1.3.8) *and* $Cov(\mathbf{X}, U) \neq \mathbf{0}$*,* $\widehat{\boldsymbol{\beta}}$ *be the 2SLS* *estimator from equation* (1.3.9) *using IVs* $\mathbf{Z}_i$ *and* $\widehat{\eta}$ *is defined in equation* (1.3.5)*, then*

$$\sqrt{N}\begin{pmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\eta}-\eta\end{pmatrix}\overset{d}{\to}N(\mathbf{0},\boldsymbol{\Omega})$$

Where $\boldsymbol{\Omega}\equiv E\begin{pmatrix}\mathbf{S}_i\\ Q_i\end{pmatrix}*\begin{pmatrix}\mathbf{S}_i' & Q_i'\end{pmatrix}$

PROOF: See the appendix.

1.4 Estimating Conditional APEs

Estimation of CAPEs is more difficult due to the need to estimate the function $r(\mathbf{x})$, which is the same problem faced by Ai and Norton (2008). The motivation for their general approach is straightforward. If we could observe the U_i then we could use nonparametric regression of $\exp(U_i)$ on $\mathbf{X}_i$. Because we do not know $\boldsymbol{\beta}$ we replace U_i with the OLS residuals $\widehat{U}_i$ and use $\exp(\widehat{U}_i)$ as the dependent variable in a nonparametric regression.

Because of technical complications, Ai and Norton propose linear series estimation, where $\exp(\widehat{U}_i)$ is regressed on various functions of $\mathbf{X}_i$. As in all nonparametric contexts, the rate of convergence of $\widehat{r}(\cdot)$ to $r(\cdot)$ is slower than $\sqrt{N}$, and much slower when the dimension of $\mathbf{X}$ is large. See Ai and Norton (2008) for details.

From equation (1.2.4), we have:

$$r(\mathbf{x})\equiv E\left(\frac{Y}{\exp(\mathbf{X}\boldsymbol{\beta})}\middle|\mathbf{X}=\mathbf{x}\right)\qquad(1.4.1)$$

From now on, the focus is how to estimate $r(\mathbf{x})$. In the literature, we do have at least two ways to estimate $r(\mathbf{x})$: one way is the traditional parametric method; we assume $r(\mathbf{x})=g(\mathbf{x}\boldsymbol{\alpha})$, where $g(.)$ is a real function satisfying certain conditions. The remained work is to estimate the parameter $\boldsymbol{\alpha}$. If we assume $g(.)$ is a linear function of $\mathbf{x}\boldsymbol{\alpha}$, $\boldsymbol{\alpha}$ can be estimated by the classic ordinary least square; if $g(.)$ is nonlinear, then nonlinear least

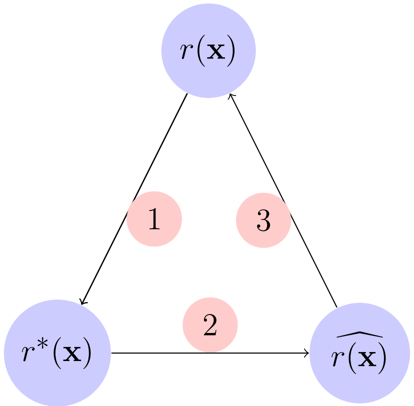

Figure 1.1: Picture of Sieve Estimator: For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

squared can be applied. Details can be referred to Wooldridge (2010). The other way is noparametric method which does not impose any parametric assumptions on $r(\mathbf{x})$. As for the comparison of those two methods, refer to a recent paper by Ackerberg et al. (2011). Here the nonparametric assumption is used, and we will propose a sieve estimator for $r(\mathbf{x})$; the idea of sieves can be vividly explained by the above picture Figure1.1.

In Figure1.1, $r(\mathbf{x})$ is unknown and is our target, we approximate it with $r^*(\mathbf{x})$; we know more terms, the number of which is denoted as M, can lead to more close approximations. However,with more terms to approximate, the estimating bias from $r(\mathbf{x})$ to $\widehat{r(x)}$ is increasing. The triangle picture shows that we need to choose a balanced M. This can also be seen from the order of the mean square error in theorem 2. Theoretically, as long as $r(\mathbf{x})$ is smooth enough, we can always find a reasonable M. Newey (1994) and Ai and Norton (2008) suggest a sample based method to decide M-cross validation.

Note that $r(\mathbf{x}) > 0$, we cannot guarantee it if we follow the method in Ai and Norton (2008). At the end of his paper, Wooldridge (1992a) suggests “exponential sieves”to replace linear ones to span the unknown space composed of positive functions. Similarly, Hirano et al.(2003) propose a “logit series” estimator for the probability distribution functions. Inspired by previous work, we develop an “exponential sieve” estimator.

Let $\|\mathbf{B}\| = [\text{trace}(\mathbf{B'B})]^{1/2}$ be the Euclidian norm of a matrix $\mathbf{B}$; $\zeta(M) = \sup_{\mathbf{x}\in\Xi} \|\mathbf{G}^M(\mathbf{x})\|$; the power series $\mathbf{G}^M(\mathbf{x})$ we use here are the same as in Hirano et al (2003).First, we list some assumptions:

Assumption 1: $(Y_1, \mathbf{X}_1), \dots, (Y_N, \mathbf{X}_N)$ are i.i.d and $\text{Var}(Y|\mathbf{X})$ is bounded and bounded away from zero; $\text{Var}\left(\frac{Y}{\exp(\mathbf{X}\boldsymbol{\beta})} - r(\mathbf{X})\right)^4$ is bounded.

Assumption 2:(i) the smallest eigenvalue of $E[\mathbf{G}^M(\mathbf{X})'\mathbf{G}^M(\mathbf{X})]$ is bounded away from zero uniformly in M;(ii) there is a sequence of constants $\zeta(M)$ and $M(N)$ such that $\zeta(M)^2M/N \longrightarrow 0$ as $N \longrightarrow \infty$

These assumptions are as usual as in the literature; since we use power series here,$\zeta_0(M) = CM$, later on we use this equivalence lots of times in the derivation.

Assumption 3: If $f : \mathbb{R}^K \longrightarrow \mathbb{R}$ is s times continuously differentiable and $\mathbf{g}^M(\mathbf{x}) = [1, \mathbf{x}, ..., \mathbf{x}^n]', M = (n+1)$, Note that $\mathbf{g}^M(\mathbf{x})$ has powers in $\mathbf{x}$ at least up to n; there is a M-vector $\boldsymbol{\gamma}_M$ such that for $\mathbf{G}^M(\mathbf{x}) = \mathbf{A}_M\mathbf{g}^M(\mathbf{x})$, and on the compact set $\Xi \subseteq \mathbb{R}^K$,

$$\sup_{\mathbf{x}\in\Xi} \left| f(\mathbf{x}) - \mathbf{G}^M(\mathbf{x})\gamma_M \right| < C_1 n^{-s} \leq C_2 M^{-s}, \tag{1.4.2}$$

This assumption is used as a fact in Hirano et al (2003), while Newey (1997) puts it as assumption. To ensure that the approximation of $r(\mathbf{x})$ is positive, we first approximate the log of $r(\mathbf{x})$:

$$\sup_{\mathbf{x}\in\Xi} \left| \log(r(\mathbf{x})) - \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M \right| < CM^{-s}, \tag{1.4.3}$$

So the exponential sieves estimator of $r(\mathbf{x})$ is $\widehat{r}(\mathbf{x}) = \exp(\mathbf{G}^M(\mathbf{x})\widehat{\boldsymbol{\pi}}_M)$, where M is fixed and

$$\widehat{\boldsymbol{\pi}}_M = \arg\min_{\boldsymbol{\pi}} \sum_{i=1}^{N} \left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}) \right)^2 \tag{1.4.4}$$

For $N \longrightarrow \infty$, we have $\|\hat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}_M^*\| \xrightarrow{p} 0$, where

$$\boldsymbol{\pi}_M^* = \arg\min_{\boldsymbol{\pi}} E \left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}) \right)^2 \tag{1.4.5}$$

Lemma 1: Suppose that:

1. the support Ξ of $\mathbf{x}$ is a compact set of $\mathbb{R}^K$
2. $r(\mathbf{x})$ is s times continuously differentiable, that is, $r(\mathbf{x}) \in C^s$,with $s \geq 4$
3. $r(\mathbf{x})$ is bounded away from zero
4. the density of $\mathbf{X}$ is bounded away from zero on Ξ

Then:

$$\left| r(\mathbf{x}) - \exp(\mathbf{G}_M(\mathbf{x})\boldsymbol{\pi}_M^*) \right| = O(CM^{-s}),$$

PROOF: See appendix.

Lemma 1 is corresponding to the step 1 in the picture; as long as the M is large enough, the deviation term will vanish. With a higher M, the converge rate is faster.

Lemma 2: Suppose that same four conditions as in Lemma 1 hold. In addition, suppose that: (5) $M(N)$ is a sequence of values of M satisfying $M(N) \to \infty$, and $\zeta(M(N))^4/N \to 0$. Then

$$\|\widehat{\boldsymbol{\pi}}_{M(N)} - \boldsymbol{\pi}^*_{M(N)}\| = O_p\left(\sqrt{\frac{M(N)}{N}}\right)$$

PROOF: See appendix.

Theorem 1.4.1. *Given all the assumptions in Lemma 1 and 2, then*

$$\int [r(\mathbf{x}) - \widehat{r}(\mathbf{x})]^2 dF(\mathbf{x}) = O(M/N + M^{-2s})$$

PROOF: Note that,

$$\begin{aligned}
& \int [r(\mathbf{x}) - \widehat{r(\mathbf{x})}]^2 dF(\mathbf{x}) \\
= & \int [r(\mathbf{x}) - \exp(\mathbf{G}^M(\mathbf{x}_i)\boldsymbol{\pi}^*_M) + \exp(\mathbf{G}^M(\mathbf{x}_i)\boldsymbol{\pi}^*_M) - \exp(\mathbf{G}^M(\mathbf{x}_i)\widehat{\boldsymbol{\pi}}_M)]^2 dF(\mathbf{x}) \\
\leq & \int [r(\mathbf{x}) - \exp(\mathbf{G}^M(\mathbf{x}_i)\boldsymbol{\pi}^*_M)]^2 + [\exp(\mathbf{G}^M(\mathbf{x}_i)\boldsymbol{\pi}^*_M) - \exp(\mathbf{G}^M(\mathbf{x}_i)\widehat{\boldsymbol{\pi}}_M)]^2 dF(\mathbf{x}) \\
= & \; O(M/N + M^{-2s})
\end{aligned}$$

The last equality follows from Lemma 1 and 2. From this theorem, we can see the two steps in the picture and the tradeoff of increase the M

Let

$$\begin{aligned}
\boldsymbol{\Sigma}_M & \equiv E[\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i)\right)^2 \exp(2\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)], \\
\mathbf{Q}_M & \equiv E[\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(2\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)], \\
V_M(\mathbf{x}) & \equiv \mathbf{G}^M(\mathbf{x})\mathbf{Q}_M^{-1}\boldsymbol{\Sigma}_M\mathbf{Q}_M^{-1}(\mathbf{x})\mathbf{G}^M(\mathbf{x})'\exp(2\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M).
\end{aligned}$$

Lemma 3: Suppose that the same four conditions as in Lemma 1 hold, then

$$\sqrt{N}V_M^{-1/2}(\mathbf{x})\left(\widehat{r(\mathbf{x})} - r(\mathbf{x})\right) \xrightarrow{d} N(0,1) \tag{1.4.6}$$

Note that $V_M^{-1/2}(\mathbf{x}) \leq O(M^{-1/2})$, so this result gives an upper bound $O_p((M/N)^{1/2})$ converge rate, which is lower than $O_p((1/N)^{1/2})$; Ai and Norton (2008) find the similar result for linear case. This result also coincides many other results in semi/no parametric literature.

The structure of the APE estimator is interesting: it is a scale factor times the estimator of the parameter of interest. The ideal is originally from the so-called average structure function in Blundell and Powell (1994), and Papke and Wooldridge (2008) advance further to estimate APE of parameter of interest in a panel data probit model with and without endogenous explanatory variables.

From section 2, the CAPE can be expressed as

$$\begin{aligned} CAPE_j(\mathbf{x}) &= \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) E[\exp(U)|\mathbf{X}=\mathbf{x}] \\ &\equiv \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) r(\mathbf{x}) \end{aligned}$$

So the estimator of $CAPE$ follows as :

$$\widehat{CAPE_j}(\mathbf{x}) = \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}})\widehat{r(\mathbf{x})} = \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) \exp(\mathbf{G}^M(\mathbf{x})\widehat{\boldsymbol{\pi}}_M)$$

Theorem 1.4.2. *Given $N^{(1/2)}M^{-(s+1)} \to 0$ as $N \to \infty$ and $\beta_j \neq 0$, Assumptions 1-3 and all the assumptions in Lemma 1 and 2, then*

$$\sqrt{N} V_M^{-1/2}(\mathbf{x}) \left(\widehat{CAPE_j}(\mathbf{x}) - CAPE_j(\mathbf{x}) \right) \xrightarrow{d} N(0, V)$$

Where $V = \beta_j^2 \exp(2\mathbf{x}\boldsymbol{\beta})$

PROOF: See appendix.

Note that, theorem 3 requires that at least one element of $\boldsymbol{\beta}$ is not zero: if $\boldsymbol{\beta}$ is zero completely, the model does not make any practical use; but it is possible that some elements of $\boldsymbol{\beta}$ are zero. If there is some element in $\boldsymbol{\beta}$ is zero, without loss of generality, let's put it as β_j=0, then:

Corollary: Given all the assumptions in theorem 3 except β_j=0, then

$$\sqrt{N} \left(\widehat{CAPE_j}(\mathbf{x}) - CAPE_j(\mathbf{x}) \right) \xrightarrow{d} N(0, W)$$

Where $CAPE_j(\mathbf{x}) \equiv \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) r(\mathbf{x})$; as for the detailed form of W, refer to proof in appendix.

PROOF: See appendix.

This is an interesting result: if some element of $\boldsymbol{\beta}$ is zero, then the estimator for the corresponding $CAPE$ is $\sqrt{N}$ consistent, which is faster than the rates of other CAPEs of nozero $\boldsymbol{\beta}$. From theorem 3, we can see that under $\sqrt{N}V_M^{-1/2}(\mathbf{x})$, $\widehat{CAPE_j}(\mathbf{x})$ degenerates to zero when $\beta_j = 0$; so it is not surprising that we need faster rate to get a non-degenerating asymptotic distribution. There are several such examples in Ferguson (1996), e.g., refer to example 3 in Chapter 7; also, theorem 5.4 in Lee (2004).

1.5 Application to Treatment Effects

As an example of where we might want to apply retransformation after IV estimation, consider the modern treatment effect literature. Let D denote a binary treatment and denote the strictly positive countfactual outcomes as $Y(0)$ and $Y(1)$. We observe D and

$$Y = (1 - D)Y(0) + DY(1).$$

Assume we have covariates $\mathbf{W}$ such that

$$\begin{aligned} Y(0) &= \exp(\alpha_0 + \mathbf{W}\boldsymbol{\beta}_0 + U) \\ Y(1) &= \exp(\alpha_1 + \mathbf{W}\boldsymbol{\beta}_1 + U) \end{aligned}$$

where, for simplicity, we assume only one source of heterogeneity, U, and $E(U) = 0$. Also, assume that U and $\mathbf{W}$ are independent. then the average treatment effect, as a function of $\mathbf{w}$, defined by

$$\tau_{ate}(\mathbf{w}) = E[Y(1) - Y(0)|\mathbf{W} = \mathbf{w}],$$

is easily seen to be

$$\tau_{ate}(\mathbf{w}) = \eta[\exp(\alpha_1 + \mathbf{w}\boldsymbol{\beta}_1) - \exp(\alpha_0 + \mathbf{w}\boldsymbol{\beta}_0)]$$

where $\eta = E[\exp(U)]$. Furthermore, in terms of the observed outcome Y, $\tau_{ate}(\mathbf{w})$ can be written in terms of its ASF, $ASF(d, \mathbf{w})$. To see how, write

$$Y = Y(0)^{(1-D)}Y(1)^D$$

and so

$$Y = \exp(\alpha_0 + \gamma D + \mathbf{W}\boldsymbol{\beta}_0 + D \cdot \mathbf{W}\boldsymbol{\delta} + U),$$

where $\gamma = \alpha_1 - \alpha_0$ and $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$. So the ASF for Y is

$$\begin{aligned} ASF(d, \mathbf{w}) &= \exp(\alpha_0 + \gamma d + \mathbf{w}\boldsymbol{\beta}_0 + d \cdot \mathbf{w}\boldsymbol{\delta})E[\exp(U)] \\ &= \eta \exp(\alpha_0 + \gamma d + \mathbf{w}\boldsymbol{\beta}_0 + d \cdot \mathbf{w}\boldsymbol{\delta}) \end{aligned}$$

If we evaluate the ASF at $d = 1$ and $d = 0$ and difference we get

$$\begin{aligned} ASF(1, \mathbf{w}) - ASF(0, \mathbf{w}) &= \eta[\exp(\alpha_0 + \gamma + \mathbf{w}\boldsymbol{\beta}_0 + \mathbf{w}\boldsymbol{\delta}) - \exp(\alpha_0 + \mathbf{w}\boldsymbol{\beta}_0)] \\ &= \eta[\exp(\alpha_1 + \mathbf{w}\boldsymbol{\beta}_1) - \exp(\alpha_0 + \mathbf{w}\boldsymbol{\beta}_0)] \\ &= \tau_{ate}(\mathbf{w}) \end{aligned}$$

We can now apply a simple 2SLS strategy if we assume

$$E(U|\mathbf{Z}) = 0$$

for IVs $\mathbf{Z}$ that include $\mathbf{W}$ and at least one element not in $\mathbf{W}$ that predicts treatment status, D. Write

$$\log(Y_i) = \alpha_0 + \gamma D_i + \mathbf{W}_i\boldsymbol{\beta}_0 + D_i \cdot \mathbf{W}_i\boldsymbol{\delta} + U_i$$

and use instruments, say, $(1, \mathbf{Z}_i, \mathbf{Z}_i \otimes \mathbf{W}_i)$, or one can be selective with the interactions. Or, as described by Wooldridge (2010, Chapter 21), probit or logit fitted values can be used as IVs, in which case the list would look like

$$(1, \hat{\mathbf{G}}_i, \mathbf{W}_i, \hat{\mathbf{G}}_i \cdot \mathbf{W}_i),$$

where the $\hat{\mathbf{G}}_i$ are the fitted probabilities from a binary response model of D_i on $\mathbf{Z}_i$.

Given the 2SLS residuals $\hat{U}_i$ we compute $\hat{\eta}$ exactly as in equation (1.3.5), and then the estimated ASF is

$$\widehat{ASF}(d, \mathbf{w}) = \hat{\eta} \exp(\hat{\alpha}_0 + \hat{\gamma} d + \mathbf{w}\hat{\boldsymbol{\beta}}_0 + d \cdot \mathbf{w}\hat{\boldsymbol{\delta}})$$

For any $\mathbf{w}$ we estimate $\hat{\tau}_{ate}(\mathbf{w}) = \widehat{ASF}(1, \mathbf{w}) - \widehat{ASF}(0, \mathbf{w})$, and the unconditional average treatment effect, $\tau_{ate} = E[Y(1) - Y(0)]$, is consistently estimated as

$$\begin{aligned} \hat{\tau}_{ate} &= N^{-1} \sum_{i=1}^{N} [\widehat{ASF}(1, \mathbf{W}_i) - \widehat{ASF}(0, \mathbf{W}_i)] \\ &= N^{-1} \sum_{i=1}^{N} \hat{\eta} [\exp(\hat{\alpha}_1 + \mathbf{W}_i \hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{W}_i \hat{\boldsymbol{\beta}}_0)] \end{aligned}$$

The average treatment effect on the treated is

$$\begin{aligned} \tau_{att}(\mathbf{w}) &= E[Y(1) - Y(0) | D = 1, \mathbf{W} = \mathbf{w}] \\ &= [\exp(\alpha_1 + \mathbf{w}\boldsymbol{\beta}_1) - \exp(\alpha_0 + \mathbf{w}\boldsymbol{\beta}_0)] E[\exp(U) | D = 1, \mathbf{W} = \mathbf{w}] \\ &\equiv [\exp(\alpha_1 + \mathbf{w}\boldsymbol{\beta}_1) - \exp(\alpha_0 + \mathbf{w}\boldsymbol{\beta}_0)] r(1, \mathbf{w}), \end{aligned}$$

where

$$r(d, \mathbf{w}) = E[\exp(U) | D = d, \mathbf{W} = \mathbf{w}].$$

Now we use nonparametrics of $\exp(\hat{U}_i)$ on $\mathbf{W}_i$ for $D_i = 1$.

$$\hat{\tau}_{att}(\mathbf{w}) = [\exp(\hat{\alpha}_1 + \mathbf{w}\hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{w}\hat{\boldsymbol{\beta}}_0)] \hat{r}(1, \mathbf{w})$$

Often reported is

$$\tau_{att} = E[Y(1) - Y(0)|D = 1]$$

and a consistent estimator is

$$N_1^{-1} \sum_{i=1}^{N} D_i[\exp(\hat{\alpha}_1 + \mathbf{W}_i \hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{W}_i \hat{\boldsymbol{\beta}}_0)]\hat{r}(1, \mathbf{W}_i).$$

1.6 Concluding Remarks

We have shown that a common retransformation method due to Duan (1983) can be used generally to estimate the parameters in the average structural function under weak assumptions on the dependence between the error and the covariates. In a standard regression setting, even a zero correlation assumption suffices. Further, the method easily applies when instrumental variables are needed to consistently estimate the coefficients in the log-linear model.

Our derivation of the joint asymptotic distribution of the parameters indexing the ASF holds under weak assumptions – much weaker than the independence assumption used in the standard Duan (1983) setting where errors and explanatory variables are independent.

In the case where $\mathbf{X}$ does not contain what we traditionally think of as endogenous variables – so that the dependence of $D(U|\mathbf{X})$ comes through moments other than $E(U|\mathbf{X})$ – one might question whether partial effects defined through the ASF are more useful than those obtained from $E(Y|\mathbf{X} = \mathbf{x})$. After all, for prediction purposes we prefer $E(Y|\mathbf{X} = \mathbf{x})$ to $ASF(\mathbf{x})$ because the former is the minimum mean square predictor of Y. But if we are interested in getting the best predictor of Y then we might just model $E(Y|\mathbf{X})$ – such as a flexible exponential function – and estimate it directly. Of course, we would get partial effects directly, too.

We also considered estimation of conditional APEs, which generally differ from the APEs when U and $\mathbf{X}$ are dependent (even though they may be mean independent). Here, we are led to an estimation problem essentially the same as Ai and Norton (2008). However, the calculation of partial effects differs, and the CAPEs have the same signs is the coefficients on the log-linear model. Some of this extends immediately to other transformations, such as $\log[Y/(1-Y)]$ when $0 < Y < 1$. More work is needed in treatment effect examples with more than one unobservable. Heckman switching regression. “Simple” solution is to make distributional assumptions.

Chapter 2

ON THE USE OF EXPONENTIAL VERSUS LOG-LINEAR MODELS FOR PANEL DATA

2.1 Introduction

As we all know, more and more data are collected over time for the same cross section units with the advance of technology; a big chunk of them are to deal with positive response variables, just name a few, the observed wage rate for workers over 5 years, number of patents applied for the firms in the last several years and etc. The immediate extension of the results in chapter one to the panel data setting would be that model the logarithm transformation of positive response variable in linear functional form of covariates still and assume, at a minimum, the error in each time period was assumed to be uncorrelated with the explanatory variables in the same time period. Basically, we just stack the observations of all time periods for each cross section unit and repeat the analysis in chapter one again; as we can see, everything just follows with a bigger sample size in terms of application. However, we know that assumption is too strong for certain panel data applications.In fact, a primary motivation for using panel data is to solve the omitted variables problem, and error terms in each time period are hardly uncorrelated most of the time. So a more interesting extension would be modeling logarithm transformation of positive response variable in a "modern" panel data setting, which explicitly contain a time-constant unobserved effect and treat it as random variable, drawn from the population along with the observed explained and explanatory variables. We can refer to the "linear panel data model (LPD)" in Wooldridge (2010) for more details.

While, the problem of modeling the logarithm transformation of positive response variable in a LPD still exists and it can become even much worse because that unobserved time invariant heterogeneity can either cause sever correlation among the error terms, or itself is correlated with covariates; even with it removed as in the standard Fixed Effect (FE) transformation, the problem can hardly be solved. For example, Blackbrun (2007) compares Poisson Quasi-Maximum Likelihood (PQML) method to wage equation and first difference to log wage equation; he finds that the latter overestimates the union coverage effect on wage by almost 14% using panel data of 1989 and 1993 NLSY. We will also see this from an analysis of a data set about the American domestic airline market from year 1997 to 2000. The results of usual FE method estimation are in Table B.1. Based on the results, we would say the price elasticity of demand of the market is over one in magnitude. We know this number is way too big and does not make any sense intuitively. For example, Park et. al.(2007) finds that most city pairs airline routes for 12 main US carriers are inelastic in short run. We will say more about it in the application section. This prompts our question of the estimation method itself. Considering the retransformation problem we have done in chapter 1, it is natural to ask the question: Why should we use the logarithmic transformation of the response variables instead of modeling them directly?

Several attempts are made in the literature. Most of the work uses nonlinear panel data models and relies on the method of conditional maximum likelihood (CML) , where a sufficient statistic (the sum of the explained variable across time) is conditioned on to remove the unobserved effect. Examples include Chamberlain (1980, 1984) for binary responses and Hausman et al.(1984) for count data. Wooldridge (1999) investigates robustness properties of these CMLEs to misspecification of the initially specified joint distribution and has shown CMLEs to be consistent when only the conditional mean in the unobserved effects multiplicative panel data model is correctly specified, which means that the consistency is robust to arbitrary patterns of serial correlation. Moreover, the results hold not only for binary or count variables, but also for any nonnegative variables. So it is not surprising that the estimation method is widely used in empirical applications, especially after Simcoe (2008) writes a STATA code titled "*xtpqml* "[1]. It sheds new light on how to estimate

[1] the code is updated to *xtpoisson* in STATA12

the conditional mean parameters consistently for positive response variable without taking logarithmic transformation in panel data models.

So, this chapter starts with Wooldridge (1999). First, we maintain the conditional mean specification condition (equation (3.1) in Wooldridge (1999)) and show what consequence of logarithm transformation can cause for the usual FE estimation in LPM by Mont Carlo simulation; Second, we extend the "exponential sieve" estimator in chapter 1 to the panel data model- estimate the average partial effect (APE) of the respective interesting explanatory variables; Third, we use "exponential sieve" estimator to construct the optimal IV as in Newey (1993, 1994) to improve the the efficiency for the GMM estimator suggested in Wooldridge (1999). Section 2 introduces model; under which the simulations are constructed to compare PQML and LFE estimators in section 3. A "exponential sieve" estimator of APE in the panel data model is formulated in section 4 .Section 5 discusses the GMM method for models with conditional mean and variance functions specified. The GMM can be understood as double PQML: one is for original scale of dependent variable; the other is for squared scale. We also propose the optimal instrument variables (OIV) estimator used in Newey (1993, 1994). An empirical application to airfare data is provided in section 6, and some remarks are contained in section 7.

2.2 Model 1

The model we are considering is the following:

$$Y_{it} = \exp(\mathbf{X}_{it}\boldsymbol{\beta})C_i V_{it}, \tag{2.2.1}$$

where C_i is the unobserved heterogeneity and V_{it} is the multiplicative idiosyncratic error term; we assume both C_i and V_{it} are positive. The immediate question is how to estimate β. It is tempting to take log transformation for both sides of equation (2.2.1):

$$\log(Y_{it}) = \mathbf{X}_{it}\boldsymbol{\beta} + \log(C_i) + \log(V_{it}), \tag{2.2.2}$$

This is typical linear panel data model as in Wooldridge (2010) if the relevant assumptions are made. Justifications for this log transformation include to deal with a dependent variable badly skewed to the right, and to interpret β as semi elasticity of Y with respect to $\mathbf{X}$, for example in Manning (1998), Ai and Norton (2000). As suggested in Wooldridge (2010), it is natural to do the usual FE transformation to remove the heterogeneity:

$$\log(\ddot{Y}_{it}) = \ddot{\mathbf{X}}_{it}\boldsymbol{\beta} + \log(\ddot{V}_{it}), \tag{2.2.3}$$

Where, $\ddot{\mathbf{X}}_{it} = \mathbf{X}_{it} - T^{-1}\sum_{t=1}^{T}\mathbf{X}_{it}$, similarly for $\log(\ddot{Y}_{it})$ and $\log(\ddot{V}_{it})$. So pooled ordinary least square can be used to estimate $\boldsymbol{\beta}$, which is exactly the FE method in Wooldridge (2010); we denote this method as LFE. Whether LFE is consistent or not depends on the assumptions, one of which is that $\ddot{\mathbf{X}}_{it}$ and $\log(\ddot{V}_{it})$ are uncorrelated. Instead, we assume:

$$E(V_{it}|\mathbf{X}_i, C_i) = 1, \tag{2.2.4}$$

Note that, we can easily get:

$$E(Y_{it}|\mathbf{X}_i, C_i) = \exp(\mathbf{X}_{it}\boldsymbol{\beta} + \log(C_i)), \tag{2.2.5}$$

So the correct conditional mean specification condition as in Wooldridge (1999) is satisfied; We can use Poission Qausi Maximum Likelihood (PQML) to estimate $\boldsymbol{\beta}$ consistently. It interesting to compare LFE estimator of $\boldsymbol{\beta}$ with PQML under equation (2.2.1) and (2.2.4). It is not difficult to know that LFE estimator of $\boldsymbol{\beta}$ is not consistent since equation (2.2.4) can not guarantee that error terms in equation (2.2.3), $\log(\ddot{V}_{it})$, are uncorrelated to the covariates, $\ddot{\mathbf{X}}_{it}$. Simulations to compare LFE with PQML are in the following section.

2.3 PQML VS. LFE: a Simulation Approach

In this section, we compare PQML with LFE by 2 Monte Carlo Simulations. The key point here is how to generate V_{it} which are positive and satisfy equation (2.2.4). In the first simulation, we specify that V_{it} has Gamma distribution and find that PQML estimator is consistent while LFE is not; while in the second simulation, we specify that V_{it} has log normal distribution; we find both are consistent and LFE is more efficient than PQML.

2.3.1 Simulation 1

We follow the model in section 2, i. e., equation (2.2.3). The data generating process is specified as the following:

- $i = 1, 2, ..., N$; $t = 1, 2, ..., T$,
- $\mathbf{X}_{it} \overset{i.i.d}{\sim} \mathbf{N}(0, 1)$,
- $C_i = \exp(\overline{\mathbf{X}}_i + \mathbf{N}(0, 1))$, where $\overline{\mathbf{X}}_i = T^{-1} \sum_{t=1}^{T} \mathbf{X}_{it}$
- $\beta = .1$,
- $V_{it} \overset{i.i.d}{\sim} \text{Gamma}(\alpha, \gamma), \text{where } \gamma = 1/\alpha = \exp(a * \mathbf{X}_{it}^2 + b * \mathbf{X}_{it})$,
- $T = 5, N =$500 if it is not specified,
- Number of simulations =1000.

We have two reasons to specify the Gamma distribution of V_{it} that way: 1. it guarantees $V_{it}' s$ are positive; 2. It makes that the condition of equation (2.2.4) is satisfied since $E(V_{it}|\mathbf{X}_i, C_i) = \alpha * \gamma = 1$, $\forall a$, b. As for the values of a and b, they will be specified in the specific setting.

Table B.2 and Table B.3 show the results for PQML and LFE methods; Note the true value of β is .1. The PQML estimates are very close to it for all the values of a and b. However, for LFE, the bias is big; such as $a = .01, b = .1$, $a = .05, b = .1$, and $a = .1, b = .1$. These strengthen the findings in Blackburn (2007); they tell us that we should be very careful to use LFE to deal with positive response variables. On the other hand, as we can see from equation (2.2.3) that the correlation between $\ddot{\mathbf{X}}$ and $\log(\ddot{V}_{it})$ biases the LFE estimator; so whenever the value of b is relatively high compared to a, the correlation between $\mathbf{X}$ and

$\log(V_{it})$, which is denoted as $\rho_{x,\, lv}$, is high, and forces the LFE away from its target. We consider an extreme case here: let $a = .1$, $b = 0$ and do the simulations for three different N values: 500, 1000 and 2000; and the results are in table titled “special case”. We can see that all three LFE estimates beat PQML in terms of bias and efficiency. The bias of PQML decreases with creasing of sample size, but for small sample, it may not perform as well as LFE does.

2.3.2 Simulation 2

Compared with simulation 1, simulation 2 is the same as simulation 1 except for the distribution of V_{it}. Here, V_{it} has a log-normal distribution. The data generating process is specified as the following:

- $i = 1, 2, ..., N$; $t = 1, 2, ..., T$,
- $\mathbf{X}_{it} \overset{i.i.d}{\sim} \mathbf{N}(0, 1)$,
- $C_i = \exp(\overline{\mathbf{X}}_i + \mathbf{N}(0, 1))$,
- $\beta = .1$,
- $V_{it} = \exp(a * \mathbf{X}_{it}^2 + b * \mathbf{X}_{it} * z_{it})$, where $z_{it} \overset{i.i.d}{\sim} \mathbf{N}(0, 1)$ if not specified.
- $T = 5$, N =500 or 1000,
- Number of simulations =1000.

For the distribution of V_{it}, the positive issue is obviously satisfied; but the values of a and b cannot be arbitrary since equation (2.2.4) is required to be satisfied. By algebra, we need have $b^2 = -2a$.

Table B.4 shows that LFE estimator is as good as PQML in terms of consistency; moreover, PQML estimator is more sensitive to the values of a and b. On the other hand, the standard errors of LFE estimator is only half of PQML. This result is very interesting since this adds one more advantage to the log transformation-more efficient.

2.3.3 Simulation 3

Compared with simulation 1, simulation 2 is the same as simulation 1 except for the distribution of V_{it}. Here, V_{it} has a log-normal distribution. The data generating process is specified as the following:

- $i = 1, 2, ..., N$; $t = 1, 2, ..., T$,
- $\mathbf{X}_{it} \overset{i.i.d}{\sim} \mathbf{N}(0, 1)$,
- $C_i = \exp(\overline{\mathbf{X}}_i + \mathbf{N}(0, 1))$,
- $\beta = .1$,
- $V_{it} = \exp(a{*}\mathbf{X}_{it}^2{+}b{*}\mathbf{X}_{it}{*}z_{it})$, where $\mathbf{Z} = \begin{pmatrix} z_{i1} \\ z_{i2} \\ z_{i3} \\ z_{i4} \\ z_{i5} \end{pmatrix} \sim \mathbf{N}(0, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$
  .
- $T = 5$, N =500 and $\rho = \{-.95, -.5, -.1, .1, .5, .95\}$,
- Number of simulations =1000.

The focus of this simulation is setting of V_{it}: note that equation (2.2.4) only specifies relationship between V_{it}, $\mathbf{X}_i$ and ci, it does not exclude any correlation among error terms of all the time periods for a certain cross section unit. The variance covariance matrix of $\mathbf{Z}$, Σ is the popular exchangeable working matrix in generalized estimating equation literature. The reason that name is called is every observation in an individual is equally correlated with every other observation in that individual. The degree of correlation is measured by the intraclass correlation coefficient. Here we can understand that $\text{Corr}(z_{it}, z_{it+j}) = \rho^j$, $j = 0, 1, 2, \cdots, T-1$ and $|\rho| < 1$. All the estimation results are in Table B.6; overall, both methods perform well in terms of bias and efficiency. However, we do see some fluctuation when ρ varies in its ranges: from Figure B.3, we can see that the biases of LFE are fluctuating around zero while all PQML estimators display downward bias; but there is no monotonic relationship between the bias and the ρ. When it comes to standard errors, the story is different: Figure B.4 shows that the standard errors of both estimators are increasing with crease of ρ; and LFE always beats PQML. We repeat the analysis for $N = 1000$, and similar patterns apply. So the serial correlation can cause problem to PQML, especially when sample size is not that big.

2.4 Estimating APE: Exponential Sieve Estimator

After Wooldridge (1999) proves that PQML is appropriate for any non-negative dependent variable-not just count data that follow a Poisson distribution and the estimator is robust to arbitrary patterns of serial correlation, the method is widely used in all kinds of empirical works, especially after Simcoe (2008) writes a STATA command *xtpqml*. However, since most of the time, the model is nonlinear in observed explanatory variables with multiplicative unobserved heterogeneity, it is difficulty to interpret the estimates of parameters, which are the popular so called semi-elasticities in log linear models; this is one of the key drawbacks of the method compared with OLS/IV to the log linear models. Inspired by what we have done in chapter 1, we propose a nonparametric estimation method for APE in model as in equation (2.2.1). This way, we extend the analysis in chapter 1 from cross section to panel data settings.

2.4.1 Estimating APE

Refer to equation (2.2.1), let U_{it}=C_iV_{it} be the whole error term. For the simplicity of denotation, we drop i for the time being. Similarly to what we have done in Chapter One, we define APE as follows:

$$APE_j(\mathbf{x_t}) = E\left[\beta_j \exp(\mathbf{x_t}\boldsymbol{\beta})U_t\right] = \eta[\beta_j \exp(\mathbf{x_t}\boldsymbol{\beta})], \tag{2.4.1}$$

Here the idea is the same as in Chapter One: the expectation in equation (2.4.1) is taken only with respect to U_t. In order to estimate $APE_j(\mathbf{x_t})$, the key is to estimate η; from equation (2.2.5), we have,

$$E\left(\left.\frac{Y_{it}}{\exp(\mathbf{X}_{it}\boldsymbol{\beta})}\right|\mathbf{X}_i\right) = E(U_{it}|\mathbf{X}_i), \tag{2.4.2}$$

So,

$$\eta \equiv E(U_{it}) = E\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\boldsymbol{\beta})}\right),$$

Then the straight forward estimator for η is:

$$\hat{\eta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\boldsymbol{\beta}})}, \tag{2.4.3}$$

Where $\hat{\beta}$ is the PQML estimator of β in equation (2.2.1). Similarly to Theorem 1 in Chapter one, we have the following:

Theorem 2.4.1. *Let* $\widehat{\boldsymbol{\beta}}$ *be the PQML estimator from equation* (2.2.1) *with condition in equation* (2.2.5)*and* $\widehat{\eta}$ *is defined in equation* (2.4.3)*, then*

$$\sqrt{N}\begin{pmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\eta}-\eta\end{pmatrix} \xrightarrow{d} \mathbf{N}(\mathbf{0},\boldsymbol{\Omega})$$

Where $\boldsymbol{\Omega} \equiv E\begin{pmatrix}\mathbf{S}_i\\ Q_i\end{pmatrix} \times \begin{pmatrix}\mathbf{S}_i' & Q_i'\end{pmatrix}$

PROOF: See appendix.

With delta method, it is natural to have the asymptotic result for the device in equation (2.4.1):

Corollary 2.4.2.

$$\sqrt{N}\left(\widehat{APE}_j(\mathbf{x_t}) - APE_j(\mathbf{x_t})\right) \xrightarrow{d} \mathbf{N}(0, \mathbf{\Theta'\Omega\Theta})$$

Where

$$\widehat{APE}_j(\mathbf{x_t}) = \hat{\eta}\widehat{\beta}_j \exp(\mathbf{X}_{it}\hat{\boldsymbol{\beta}})$$

$$\mathbf{\Theta} = \mathbf{\nabla}_{\boldsymbol{\theta}}\{\eta[\beta_j \exp(\mathbf{x_t}\boldsymbol{\beta})\},\ \boldsymbol{\theta} = \left(\boldsymbol{\beta}',\ \eta\right)'$$

2.4.2 Estimating CAPE

First, we assume:

$$D(C_i|\mathbf{X}_i) = D(C_i|\overline{\mathbf{X}}_i), \tag{2.4.4}$$

where $\overline{\mathbf{X}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{X}_{it}$, which is the same as in the simulations in the previous section, represents the time-averaged $\mathbf{X}_{it}$ over the various panel periods. This assumption can date back to Mundlak (1978) and Chamberlain (1980, 1982, 1984), and more recently appears in Papke and Wooldridge (2008). Of course, we don't need the normal distribution assumption as they do.

Combine equations (2.2.1), (2.2.4) and (2.4.4), and we get the following by iterated conditional expectation:

$$\begin{aligned} E(Y_{it}|\mathbf{X}_i) &= E[(Y_{it}|\mathbf{X}_i, C_i)|\mathbf{X}_i] \\ &= \exp(\mathbf{X}_{it}\beta)E(C_i|\mathbf{X}_i) \\ &= \exp(\mathbf{X}_{it}\beta)E(C_i|\overline{\mathbf{X}}_i) \qquad (2.4.5) \end{aligned}$$

So, we define:

$$E\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\beta)}\middle|\mathbf{X}_i\right) = E(C_i|\overline{\mathbf{X}}_i) \equiv r(\overline{\mathbf{X}}_i), \tag{2.4.6}$$

Here, the $r(\overline{\mathbf{X}}_i)$ is our target to estimate and assume it has the same properties as in chapter 1. The only difference is that the independent variable is $\overline{\mathbf{X}}_i$ instead of $\mathbf{X}_i$, and variables in the left hand side of above equation has one more layer of subscript, t, other than i. So the exponential sieves estimator of $r(\overline{\mathbf{X}}_i)$ is $\hat{r}(\overline{\mathbf{X}}_i) = \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\hat{\boldsymbol{\pi}}_M)$, where

$$\hat{\boldsymbol{\pi}}_M = \arg\min_{\boldsymbol{\pi}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\boldsymbol{\beta}})} - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2 \tag{2.4.7}$$

For $N \longrightarrow \infty$, we have $\|\hat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}^*_M\| \xrightarrow{p} 0$, where

$$\boldsymbol{\pi}^*_M = \arg\min_{\boldsymbol{\pi}} E\sum_{t=1}^{T}\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\boldsymbol{\beta})} - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2 \tag{2.4.8}$$

Lemma 2.4.3. *Suppose that:*

1. *the support Ξ of $\overline{\mathbf{X}}$ is a compact set of $\mathbb{R}^J$*

2. *$r(\mathbf{X})$ is s times continuously differentiable, that is, $r(\mathbf{X}) \in C^s$, with $s \geq 4$*

3. *$r(\mathbf{X})$ is bounded away from zero*

4. *the density of $\overline{\mathbf{X}}$ is bounded away from zero on Ξ*

Then:

$$\left|r(\overline{\mathbf{X}}) - \exp(\mathbf{G}_M(\overline{\mathbf{X}})\boldsymbol{\pi}^*_M)\right| = O_p(CM^{-s}), \tag{2.4.9}$$

PROOF: See appendix.

Lemma 2.4.4. *Suppose that same three conditions as in Lemma 1 hold. In addition, suppose that: (iv) $M(N)$ is a sequence of values of M satisfying $M(N) \to \infty$, and $\zeta(M(N))^4/N \to 0$. Then*

$$\|\hat{\boldsymbol{\pi}}_{M(N)} - \boldsymbol{\pi}^*_{M(N)}\| = O_p\left(\sqrt{\frac{M(N)}{N}}\right) \tag{2.4.10}$$

PROOF: See appendix.

These two lemmas are almost the same as lemma 1 and 2 in chapter 1 expect here is for $\overline{\mathbf{X}}_i$.

Theorem 2.4.5. *Given all the assumptions in Lemma 2.4.3 and Lemma 2.4.4, then*

$$\int [r(\mathbf{X}) - \hat{r}(\mathbf{X})]^2 dF(\mathbf{X}) = O(M/N + M^{-2s}) \tag{2.4.11}$$

PROOF: the proof is the same as Theorem 2 in chapter 1.

From this theorem, we can see that the same kind of trade off argument as in chapter 1 happens here: we know more terms, which means higher value of M, can lead to more close approximations. However, with more terms to approximate, the estimating bias from $r(\mathbf{X})$ to $\hat{r}(\mathbf{X})$ is increasing. Refer to chapter 1 for detailed illustration.

The next question is how to estimate the CAPE. First, we need to define the CAPE in the new device; usually, we need to find the functional form of $E(Y_{it}|\mathbf{X}_{it})$. From equation (2.2.1):

$$\frac{\partial Y_{it}}{\partial x_{ij}} = \beta_j \exp(\mathbf{X}_{it}\boldsymbol{\beta})C_i V_{it},$$

By iterated expectation and combine equation (2.4.5) and equation (2.4.6)

$$E\left(\left.\frac{\partial Y_{it}}{\partial x_{ij}}\right|\mathbf{X}_i\right) = \beta_j \exp(\mathbf{X}_{it}\boldsymbol{\beta})r(\overline{\mathbf{X}}_i)$$

So

$$CAPE_j(\mathbf{x}) \equiv E\left(\left.\frac{\partial Y_{it}}{\partial x_{ij}}\right|\mathbf{X}_i = \mathbf{x}\right) = \beta_j \exp(\mathbf{x}_t\boldsymbol{\beta})r(\overline{\mathbf{x}}),$$

Where, $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, $\overline{\mathbf{x}} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_t$. The natural estimator for CAPE is:

$$\widehat{CAPE}_j(\mathbf{x}) = \widehat{\beta}_j \exp(\mathbf{x}_t\widehat{\boldsymbol{\beta}})\widehat{r}(\overline{\mathbf{x}}),$$

Since $\widehat{\boldsymbol{\beta}}$ can be easily obtained from PQML method, the next will focus on how to estimate $r(\overline{\mathbf{x}})$.

Let

$$\begin{aligned}
\boldsymbol{\Sigma}_M &\equiv E\left[\sum_{t=1}^{T}\mathbf{G}^M(\overline{\mathbf{X}}_i)'\mathbf{G}^M(\overline{\mathbf{X}}_i)\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\boldsymbol{\beta})} - r(\overline{\mathbf{X}}_i)\right)^2 \exp(2\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}_M^*)\right], \\
\mathbf{Q}_M &\equiv E[\mathbf{G}^M(\overline{\mathbf{X}}_i)'\mathbf{G}^M(\overline{\mathbf{X}}_i)\text{exp}(2\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}_M^*)], \\
V_M(\bar{\mathbf{x}}) &\equiv \mathbf{G}^M(\bar{\mathbf{x}})\mathbf{Q}_M^{-1}\boldsymbol{\Sigma}_M\mathbf{Q}_M^{-1}(\bar{\mathbf{x}})\mathbf{G}^M(\bar{\mathbf{x}})'\text{exp}(2\mathbf{G}^M(\bar{\mathbf{x}})\boldsymbol{\pi}_M^*).
\end{aligned}$$

Lemma 2.4.6. *Suppose that the same four conditions as in Lemma 2.4.3 hold, then*

$$\sqrt{N}V_M^{-1/2}(\overline{\mathbf{x}})\left(\widehat{r}(\overline{\mathbf{x}}) - r(\overline{\mathbf{x}})\right) \xrightarrow{d} \mathbf{N}(0,1)$$

Theorem 2.4.7. *Given $N^{(1/2)}M^{-(s+1)} \to 0$ as $N \to \infty$ and $\beta_j \neq 0$, assumptions 1-3 and all the assumptions in Lemma 2.4.3 and Lemma 2.4.4, then*

$$\sqrt{N}V_M^{-1/2}(\mathbf{x})\left(\widehat{CAPE}_j(\mathbf{x}) - CAPE_j(\mathbf{x})\right) \xrightarrow{d} \mathbf{N}(0,V)$$

Where $V = \beta_j^2\exp(2\mathbf{x}\boldsymbol{\beta})$

PROOF: See appendix.

Note that, theorem 2.4.7 requires that at least one element of $\boldsymbol{\beta}$ is not zero: if $\boldsymbol{\beta}$ is zero completely, the model does not make any practical use; but it is possible that some elements of β are zero. If there is some element in $\boldsymbol{\beta}$ is zero, without loss of generality, let's put it as β_j=0, then:

Corollary 2.4.8. *Given all the assumptions in theorem 2.4.7 except β_j=0, then*

$$\sqrt{N}\left(\widehat{CAPE}_j(\mathbf{x}) - CAPE_j(\mathbf{x})\right) \overset{d}{\to} N(0, W)$$

Where $CAPE_j(\mathbf{x}) \equiv \beta_j \exp(\mathbf{x}\boldsymbol{\beta})r(\overline{\mathbf{x}})$ as in equation ; as for the detailed form of W, refer to proof in appendix.

PROOF: See appendix.

The above results are the analog extensions as in Chapter 1. But we do have a special result for panel data case. From equation (2.4.5), we know that $E(Y_{it}|\mathbf{X}_i)$ only depends on $\mathbf{X}_{it}$ and $\overline{\mathbf{X}}_i$ with assumption of equation (2.4.4); so we do the following manipulation:

$$\begin{aligned} E(Y_{it}|\mathbf{X}_{it}, \overline{\mathbf{X}}_i) &= E[E(Y_{it}|\mathbf{X}_i)|\mathbf{X}_{it}, \overline{\mathbf{X}}_i] \\ &= E[\exp(\mathbf{X}_{it}\boldsymbol{\beta})r(\overline{\mathbf{X}}_i)|\mathbf{X}_{it}, \overline{\mathbf{X}}_i] \\ &= \exp(\mathbf{X}_{it}\boldsymbol{\beta})r(\overline{\mathbf{X}}_i) \qquad (2.4.12) \end{aligned}$$

Following the same derivation as in Blundell and Powell (2003), we have:

$$\begin{aligned} ASF(\mathbf{X}_t) &= \int E(Y_{it}|\mathbf{X}_{it} = \mathbf{X}_t, \overline{\mathbf{X}}_i)dF_{\overline{\mathbf{X}}_i} \\ &= \exp(\mathbf{X}_t\boldsymbol{\beta})\int r(\overline{\mathbf{X}}_i)dF_{\overline{\mathbf{X}}_i} \qquad (2.4.13) \end{aligned}$$

If we let $E(Y_{it}|\mathbf{X}_{it}, \overline{\mathbf{X}}_i) = H(\mathbf{X}_{it}, \overline{\mathbf{X}}_i)$, then the RHS of equation (2.4.13) is $\int H(\mathbf{X}_{it}, \overline{\mathbf{X}}_i)dF_{\overline{\mathbf{X}}_i}$, which is exactly the so called average structural function in Blundell and Powell (2003); as in chapter 1, we define APE basing on ASF. Denote the average partial effect with respect to $\mathbf{X}_t$ as:

$$\begin{aligned} \boldsymbol{\lambda} &\equiv \frac{\partial ASF(\mathbf{X}_t)}{\partial \mathbf{X}_t} \\ &= \frac{\partial \exp(\mathbf{X}_t\boldsymbol{\beta})}{\partial \mathbf{X}_t}\int r(\overline{\mathbf{X}}_i)dF_{\overline{\mathbf{X}}_i} \\ &= \boldsymbol{\beta}\exp(\mathbf{X}_t\boldsymbol{\beta})\int r(\overline{\mathbf{X}}_i)dF_{\overline{\mathbf{X}}_i} \qquad (2.4.14) \end{aligned}$$

The corresponding estimator is:

$$\widehat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\beta}}_{pqml} \exp(\mathbf{X}_t \hat{\boldsymbol{\beta}}_{pqml}) \left(N^{-1} \sum_{i=1}^{N} \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i) \hat{\boldsymbol{\pi}}_M) \right) \tag{2.4.15}$$

The structure of $\boldsymbol{\lambda}$ is interesting: it only depends on $\mathbf{X}_t$; the effect of cross sectional heterogeneity is coorperated as a scalar factor. Note that it is between APE and $CAPE(\mathbf{x})$: the former measures the overall effect of covariates and heterogeneity; the latter considers both cross section and time dimensions. While $\boldsymbol{\lambda}$ considers the effect of covariates on both dimensions, but only overall effect of heterogeneity. Papke and Wooldridge (2008) get a similar device in a panel data probit model with and without endogenous explanatory variables. To obtain asymptotic result for $\widehat{\boldsymbol{\lambda}}$, we go further following Papke and Wooldridge (2008):

$$\widehat{\boldsymbol{\tau}} = \hat{\boldsymbol{\beta}}_{pqml} \left((NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \exp(\mathbf{X}_{it} \hat{\boldsymbol{\beta}}_{pqml} + \mathbf{G}^M(\overline{\mathbf{X}}_i) \hat{\boldsymbol{\pi}}_M) \right)$$

Note that $\widehat{\boldsymbol{\lambda}}$ can be considered as a special case of $\widehat{\boldsymbol{\tau}}$:

$$\begin{aligned} & \hat{\boldsymbol{\beta}}_{pqml} \left((NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \exp(\mathbf{X}_{it} \hat{\boldsymbol{\beta}}_{pqml} + \mathbf{G}^M(\overline{\mathbf{X}}_i) \hat{\boldsymbol{\pi}}_M) \right) \\ = \; & \hat{\boldsymbol{\beta}}_{pqml} \exp(\mathbf{X}_t \hat{\boldsymbol{\beta}}_{pqml}) \left(N^{-1} \sum_{i=1}^{N} \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i) \hat{\boldsymbol{\pi}}_M) \right) \end{aligned}$$

Where $\mathbf{X}_{it} = (\mathbf{X}_t, \underbrace{0, \ \cdots, \ 0}_{T-1})$. So as long as we obtain asymptotic results for $\widehat{\boldsymbol{\tau}}$, the result for $\widehat{\boldsymbol{\lambda}}$ can be easily achieved:

Theorem 2.4.9. *Given $N^{(1/2)}M^{-s} \to 0$ as $N \to \infty$, and all the assumptions in Lemma 2.4.3 and 2.4.4 , then*

$$\sqrt{N}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{V}) \tag{2.4.16}$$

PROOF: See appendix.

The structure of the variance is a little complicated, the derivations and estimation are in the appendix.

Note that, it follows from equation (2.4.6):

$$E\left(r(\overline{\mathbf{X}}_i)\right) \equiv E\left(E(C_i|\overline{\mathbf{X}}_i)\right) = E\left[E\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\beta)}\middle|\mathbf{X}_i\right)\right] = E\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\beta)}\right)$$

Then there is an "naive" estimator of APE automatically:

$$\hat{\boldsymbol{\beta}}_{pqml}\exp(\mathbf{X}_t\hat{\boldsymbol{\beta}}_{pqml})\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\boldsymbol{\beta}}_{pqml})}$$

With all these solved, we can propose the following estimating procedure:

Step 1: do the PQMLE of Y_{it} on $\mathbf{X}_{it}$, and denote the estimator of β as $\widehat{\beta}_{pqml}$;

Step 2: use power series to approximate $r(\overline{\mathbf{X}}_i)$, denote as $\hat{r}(\overline{\mathbf{X}}_i)$;

Step 3: estimate the APE (or CAPE) of parameter of interest and corresponding standard error.

2.5 More Efficient Estimator: GMM

As we find that in simulation 2 in section 3, when both PQML and LFE estimators are consistent, the latter is more efficient than the former. Wooldridge (1999) suggests GMM method to improve the efficiency of PQML estimator. The key step of GMM is to derive a moment condition from equation (2.2.3).

2.5.1 Model 2

As we all known, we need a class to compare efficiency; the model we are considering here as follows:

$$\mathrm{E}(Y_{it}|\mathbf{X}_i,\phi_i,\psi_i) = \phi_i\mu(\mathbf{X}_{it},\boldsymbol{\beta}_0), \tag{2.5.1}$$

$$\mathrm{Var}(Y_{it}|\mathbf{X}_i,\phi_i,\psi_i) = \psi_i^2[E(Y_{it}|\mathbf{X}_i,\phi_i,\psi_i)]^2 = \psi_i^2[\phi_i\mu(\mathbf{X}_{it},\boldsymbol{\beta}_0)]^2 \tag{2.5.2}$$

$$\mathrm{Cov}(Y_{it}, Y_{ir}|\mathbf{X}_i, \phi_i, \psi_i) = 0, t \neq r \tag{2.5.3}$$

This model is initiated in Wooldridge (1999) who put it as a question. As we can see that the function $\mu(.,.)$ can be any positive functional form, such as exponential in section 2. Compared with model 1, we have one more equation to specify the relationship between the conditional variance and the square of conditional mean. Note that ψ_i is also a random variable, so it is the extension of constant coefficient of variation model.

2.5.2 GMM

From the above two equations, we can easily get the following equation:

$$E(Y_{it}^2|\mathbf{X}_i,\ \phi_i,\ \psi_i) = (1+\psi_i^2)[E(Y_{it}|\mathbf{X}_i, \phi_i)]^2 = (1+\psi_i^2)\phi_i^2[\mu(\mathbf{X}_{it}, \boldsymbol{\beta}_0)]^2 \tag{2.5.4}$$

if we put the Y_{it}^2 as the the Y_{it}, $(1+\psi_i^2)\phi_i^2$ as the $\exp(C_i)$ and $[\mu(\mathbf{X}_{it},\ \boldsymbol{\beta}_0)]^2$ as $\exp(\mathbf{X}_{it}\boldsymbol{\beta})$ in the equation (2.2.5), we can apply the PQML method to Y_{it}^2. From here, we can derive the moment conditions. We define:

$$\sum_{t=1}^{T}(Y_{it})^j = n_{ji}, \quad j = 1,\ 2$$

$$p_{jt}(\mathbf{X}_i,\ \boldsymbol{\beta}_0) \equiv \frac{\mu^j(\mathbf{X}_{it}, \boldsymbol{\beta}_0)}{\sum_{r=1}^{T}\mu^j(\mathbf{X}_{ir}, \boldsymbol{\beta}_0)}, \quad j = 1,\ 2$$

$$\mathbf{u}_{ji}(\boldsymbol{\beta}) \equiv (\mathbf{Y}_i)^j - \mathbf{p}_j(\mathbf{X}_i, \boldsymbol{\beta})n_{jt}, \quad j = 1,\ 2$$

where, $\mathbf{p}_j(\mathbf{X}_i,\ \boldsymbol{\beta}) \equiv [p_{j1}(\mathbf{X}_i,\ \boldsymbol{\beta}),\ ...,\ p_{jT}(\mathbf{X}_i,\ \boldsymbol{\beta})]'$, $\mathbf{Y}_i = [Y_{i1},\ ...,\ Y_{iT}]'$

By iterated expectation, we have:

$$E(\mathbf{u}_{ji}(\boldsymbol{\beta}_0)|\mathbf{X}_i,\ \phi_i,\ \psi_i) = E((\mathbf{Y}_i)^j|\mathbf{X}_i,\ \phi_i,\ \psi_i) - \mathbf{p}_j(\mathbf{X}_i,\ \boldsymbol{\beta})E(n_{ji}|\mathbf{X}_i,\ \phi_i,\ \psi_i) = \mathbf{0}, \quad j = 1,\ 2 \tag{2.5.5}$$

So,

$$E(\mathbf{D}_j(\mathbf{X}_i,\ \boldsymbol{\beta})'\mathbf{u}_{ji}(\beta_0)) = \mathbf{0}, \quad j = 1,\ 2 \tag{2.5.6}$$

Where $\mathbf{D}_j(\mathbf{X}_i,\, \boldsymbol{\beta})$ is any appropriate function of $\mathbf{X}_i$.

Wooldridge (1999) suggests the following:

$$\mathbf{D}_j(\mathbf{X}_i, \boldsymbol{\beta}) = [\mathbf{W}_j(\mathbf{X}_i, \boldsymbol{\beta})\nabla_{\boldsymbol{\beta}}\mathbf{p}_j(\mathbf{X}_i, \boldsymbol{\beta})|\nabla_{\boldsymbol{\beta}}\mathbf{p}_j(\mathbf{X}_i, \boldsymbol{\beta})]. \tag{2.5.7}$$

where, $\mathbf{W}_j(\mathbf{X}_i,\, \boldsymbol{\beta}) \equiv [\text{diag}\{p_{j1}(\mathbf{X}_i,\, \boldsymbol{\beta}),\, ...,\, p_{jT}(\mathbf{X}_i,\, \boldsymbol{\beta})\}]^{-1}$

If we define:

$$\mathbf{D}(\mathbf{X}_i, \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{D}_1(\mathbf{X}_i, \boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2(\mathbf{X}_i, \boldsymbol{\beta}) \end{bmatrix}$$

$$\mathbf{u}_i(\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{u}_{1i}(\boldsymbol{\beta}) \\ \mathbf{u}_{2i}(\boldsymbol{\beta}) \end{bmatrix}$$

the two moment conditions can be combined as:

$$E(\mathbf{D}(\mathbf{X}_i)'\mathbf{u}_i(\boldsymbol{\beta}_0)) = \mathbf{0} \tag{2.5.8}$$

GMM follows easily:

$$\hat{\boldsymbol{\beta}}_{gmm} = \arg\min \left(\sum_{i=1}^{N} \widetilde{\mathbf{D}}_i'\mathbf{u}_i(\boldsymbol{\beta})\right)' \left(\sum_{i=1}^{N} \widetilde{\mathbf{D}}_i'\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i'\widetilde{\mathbf{D}}_i\right)^{-1} \left(\sum_{i=1}^{N} \widetilde{\mathbf{D}}_i'\mathbf{u}_i(\boldsymbol{\beta})\right) \tag{2.5.9}$$

Where $\widetilde{\mathbf{D}}_i = \mathbf{D}(\mathbf{X}_i,\, \hat{\boldsymbol{\beta}}_{pqml})$, $\widetilde{\mathbf{u}}_i = \mathbf{u}_i(\hat{\boldsymbol{\beta}}_{pqml})$.

As pointed out by Wooldridge (1999), GMM with only $\mathbf{D}_1(\mathbf{X}_i,\, \boldsymbol{\beta})$ and $\mathbf{u}_{1i}(\boldsymbol{\beta})$ is identical to the PQML.

Theorem 2.5.1. *As consistent estimators for $\boldsymbol{\beta}_0$ in model 2, GMM is more efficient than PQML.*

PROOF: From Wooldridge (2002, section 8.3.3), we know:

$$A\,\text{Var}(\hat{\boldsymbol{\beta}}_{gmm}) = \mathbf{C}'^{-1}\mathbf{C}$$

$$A\,\mathrm{Var}(\hat{\boldsymbol{\beta}}_{pqml}) = \mathbf{C}_1'\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1$$

Where,

$$\boldsymbol{\Lambda} = E\left(\mathbf{D}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{D}_i\right)$$

$$\mathbf{C} = E\left(\mathbf{D}_i'\boldsymbol{\nabla}_{\boldsymbol{\beta}}\mathbf{u}_i(\boldsymbol{\beta})\right)$$

$\boldsymbol{\Lambda}_1$ and $\mathbf{C}_1$ are similar to $\boldsymbol{\Lambda}$ and $\mathbf{C}$ with $\mathbf{D}_i$ and $\mathbf{u}_i$ replaced by $\mathbf{D}_{1i}$ and $\mathbf{u}_{1i}$ respectively. From White (1984), the matrix $\mathbf{C}_1'\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1 - \mathbf{C}'^{-1}\mathbf{C}$ is p.s.d.; then the result follows immediately.

2.5.3 Simulation 3

The data generating process is specified as the following:

- $\mathbf{X}_{it} \overset{i.i.d}{\sim} \mathbf{N}(0,1)$,
- $C_i = \overline{\mathbf{X}}_i + \mathbf{N}(0,1)$,
- $\beta = .1$,
- $V_{it} = \exp(-.125 * \overline{\mathbf{X}}_i^2 + .5 * \overline{\mathbf{X}}_i * z_{it})$, where $z_{it} \overset{i.i.d}{\sim} \mathbf{N}(0,1)$.
- $Y_{it} = \exp(\mathbf{X}_{it}\beta + C_i)V_{it}$,
- T=5, Number of simulations =5000 .

Here, V_{it} has the similar form as in simulation 1 and 2; however, since V_{it} has to satisfy equation (2.2.2), we use $\overline{\mathbf{X}}_i$ instead of $\mathbf{X}_{it}$. We only use $a = -.125$, $b = .5$ here since other values have the similar results.

Table B.8 has the results for all the three methods. All the three estimators are consistent. But the differences of standard errors are big. Compared with PQML with GMM, standard error of the latter is about a half of the former; so the extra moment condition does matter here and GMM is the right direction to improve the efficiency. However, the standard errors of LFE are smaller than GMM. Theoretically, there are cases where LFE is not consistent while GMM and PQML are; but we fail to find a simulation .

2.5.4 Optimal IV Estimator

Considering Model 2, we can derive the so-called optimal instrumental variables (OIV) estimator in Newey (1990, 1993) to improve efficiency. For the moment condition, $\mathbf{u}_{1i}(\boldsymbol{\beta}) \equiv \mathbf{Y_i} - \mathbf{p_1}(\mathbf{X_i}, \boldsymbol{\beta}) n_{\mathbf{1i}}$, we first need to find its variance structure. By the conditions equations (2.5.1)-(2.5.3)from model 2, we can get:

$$\text{Var}(\mathbf{u}_{1i}|\mathbf{X}_i) = E(\mathbf{u}_{1i} * \mathbf{u}'_{1i}|\mathbf{X}_i) = E[\psi_i^2 \phi_i^2 |\mathbf{X}_i] * \mathbf{\Omega}_i \tag{2.5.10}$$

Where,

$$\mathbf{\Omega}_i(r,s) = \begin{cases} \left[1 - \frac{2*\mu(\mathbf{X}_{is},\boldsymbol{\beta})}{\sum_{t=1}^T \mu(\mathbf{X}_{it},\boldsymbol{\beta})} + \frac{\sum_{t=1}^T \mu(\mathbf{X}_{it},\boldsymbol{\beta})^2}{(\sum_{t=1}^T \mu(\mathbf{X}_{it},\boldsymbol{\beta}))^2}\right] \mu(\mathbf{X}_{is},\boldsymbol{\beta})^2, & \text{if} \;\; r = s; \\ \\ \left[\frac{\sum_{t=1}^T \mu(\mathbf{X}_{it},\boldsymbol{\beta})^2}{(\sum_{t=1}^T \mu(\mathbf{X}_{it},\boldsymbol{\beta}))^2} - \frac{\mu(\mathbf{X}_{is},\beta)+\mu(\mathbf{X}_{ir},\boldsymbol{\beta})}{\sum_{t=1}^T \mu(\mathbf{X}_{it},\boldsymbol{\beta})}\right] \mu(\mathbf{X}_{is},\boldsymbol{\beta})\mu(\mathbf{X}_{ir},\boldsymbol{\beta}), & \text{if} \;\; r \neq s. \end{cases}$$

In fact, if we use matrix algebra, we can do the following:

$$\mathbf{u}_{1i}(\beta) = \left[I_T - \begin{pmatrix} p_{i1} & p_{i1} & \cdots & p_{i1} \\ p_{i2} & p_{i2} & \cdots & p_{i2} \\ \vdots & \vdots & \cdots & \vdots \\ p_{iT} & p_{iT} & \cdots & p_{iT} \end{pmatrix} \right] \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iT} \end{pmatrix}$$

so,we can get the following:

$$\text{Var}(\mathbf{u}_{1i}(\boldsymbol{\beta})|\mathbf{X}_i,\ \phi_i,\ \psi_i) = \mathbf{A}\,\text{Var}\left(\begin{matrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iT} \end{matrix} \middle| \ \mathbf{X}_i,\ \phi_i,\ \psi_i \right) \mathbf{A}' = \psi_i^2 \phi_i^2 \mathbf{A}\mathbf{B}\mathbf{A}'$$

where,

$$\mathbf{A} = \left[I_T - \begin{pmatrix} p_{i1} & p_{i1} & \cdots & p_{i1} \\ p_{i2} & p_{i2} & \cdots & p_{i2} \\ \vdots & \vdots & \cdots & \vdots \\ p_{iT} & p_{iT} & \cdots & p_{iT} \end{pmatrix} \right]$$

$$\mathbf{B} = \begin{pmatrix} \mu(\mathbf{X}_{i1}, \boldsymbol{\beta})^2 & 0 & \cdots & 0 \\ 0 & \mu(\mathbf{X}_{i2}, \boldsymbol{\beta})^2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \mu(\mathbf{X}_{iT}, \boldsymbol{\beta})^2 \end{pmatrix}$$

So,

$$\mathrm{Var}(\mathbf{u}_{1i}(\boldsymbol{\beta})|\mathbf{X}_i) = E(\psi_i^2 \phi_i^2 | \mathbf{X}_i)\mathbf{ABA}'$$

Note that,

$$\mathbf{\Omega}_i = \mathbf{ABA}'$$

The remained part is to find $E(\psi_i^2 \phi_i^2 | \mathbf{X}_i)$; note that:

$$\mathrm{Var}(\mathbf{u}_{1it}|\mathbf{X}_i) = E[\psi_i^2 \phi_i^2 | \mathbf{X}_i] \left[1 - \frac{2 * \mu(\mathbf{X}_{it}, \boldsymbol{\beta})}{\sum_{t=1}^T \mu(\mathbf{X}_{it}, \boldsymbol{\beta})} + \frac{\sum_{t=1}^T \mu(\mathbf{X}_{it}, \boldsymbol{\beta})^2}{(\sum_{t=1}^T \mu(\mathbf{X}_{it}, \boldsymbol{\beta}))^2} \right] \mu(\mathbf{X}_{it}, \boldsymbol{\beta})^2,$$

So,

$$E[\psi_i^2 \phi_i^2 | \mathbf{X}_i] = E\left(\left. \frac{u_{1it}^2}{\left[1 - \frac{2 * \mu(\mathbf{X}_{it}, \boldsymbol{\beta})}{\sum_{t=1}^T \mu(\mathbf{X}_{it}, \boldsymbol{\beta})} + \frac{\sum_{t=1}^T \mu(\mathbf{X}_{it}, \boldsymbol{\beta})^2}{(\sum_{t=1}^T \mu(\mathbf{X}_{it}, \boldsymbol{\beta}))^2} \right] \mu(\mathbf{X}_{it}, \boldsymbol{\beta})^2} \right| \mathbf{X}_i \right),$$

If we assume that:

$$E[\psi_i^2 \phi_i^2 | \mathbf{X}_i] = E[(1 + \psi_i^2)\phi_i^2 | \overline{\mathbf{X}_i}] = h(\overline{\mathbf{X}}_i) \tag{2.5.11}$$

Here, function $h(.)$ is a unknown function. As we know that $h(.)$ is the conditional expectation of positive random variables, so it is positive. Hence the main problem is still how to catch the positivity. Newey (1994) uses the truncation, which works in the large sample case. Since the truncation threshold is very arbitrary, then it may not work well in finite sample case. Next, we try to find the $\mathbf{D}(\mathbf{X}_i)$ as in Newey (1990, 1993):

$$\begin{aligned}
\mathbf{D}(\mathbf{X}_i) &\equiv E(\nabla_{\boldsymbol{\beta}} u_{1i}(\boldsymbol{\beta})|\mathbf{X}_i) \\
&= E(\nabla_{\boldsymbol{\beta}}[\mathbf{Y_i} - \mathbf{p_1}(\mathbf{X_i}, \boldsymbol{\beta})\mathbf{n_{1i}}]|\mathbf{X}_i) \\
&= -\nabla_{\boldsymbol{\beta}}[\mathbf{p}_1(\mathbf{X}_i, \boldsymbol{\beta})]E(\mathbf{n}_{1i}|\mathbf{X}_i) \\
&= -\nabla_{\boldsymbol{\beta}}[\mathbf{p}_1(\mathbf{X}_i, \boldsymbol{\beta})]\sum_{t=1}^{T}\mu(\mathbf{X}_{it}, \boldsymbol{\beta})E(\phi_i|\mathbf{X}_i) \\
&= -\nabla_{\beta}[\mathbf{p}_1(\mathbf{X}_i, \boldsymbol{\beta})]\sum_{t=1}^{T}\mu(\mathbf{X}_{it}, \boldsymbol{\beta})r(\overline{\mathbf{X}}_i)
\end{aligned}$$

Here again we will use "exponential sieve" to estimate $h(\overline{\mathbf{X}}_i)$ and $r(\overline{\mathbf{X}}_i)$ as in section 3; note that the method shows up two times for OIV estimator. We use cross validation as in Newey (1993) to choose the number of terms M.

We propose the following estimating method:

Step 1: do the PQML of Y_{it} on $\mathbf{X}_{it}$, and denote the estimator of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{pqml}$; also define, $\hat{u}_{1it} = Y_{it} - n_{1i}p_{it}(\widehat{\boldsymbol{\beta}}_{pqml})$;

Step 2: use power series to approximate $h(\overline{\mathbf{X}}_i)$, denote as $\hat{h}(\overline{\mathbf{X}}_i)$; define $\widetilde{\mathbf{B}}(\mathbf{X}_i) = \hat{\mathbf{D}}(\mathbf{X}_i) * (\hat{\boldsymbol{\Omega}}_i)^{-1}/\hat{h}(\overline{\mathbf{X}}_i)$;

Step 3: do the following GMM,

$$\hat{\boldsymbol{\beta}}_{gmm} = \arg\min\left(\sum_{i=1}^{N}\widetilde{\mathbf{B}}_i'\mathbf{u}_{1i}(\boldsymbol{\beta})\right)'\left(\sum_{i=1}^{N}\widetilde{\mathbf{B}}_i'\widetilde{\mathbf{B}}_i\right)^{-1}\left(\sum_{i=1}^{N}\widetilde{\mathbf{B}}_i'\mathbf{u}_{1i}(\boldsymbol{\beta})\right) \qquad (2.5.12)$$

We use a simulation to end this section:

Simulation 4

We follow the model 2 in section 5, i.e, equations (2.5.1)-(2.5.3). The data generating process is specified as the following:

- $i = 1, 2, ..., N$; $t = 1, 2, ..., T$,

- $\mathbf{X}_{it} \overset{i.i.d}{\sim} \mathbf{N}(0,1)$,
- $C_i = \exp(\overline{\mathbf{X}}_i + \mathbf{N}(0,1))$, where $\overline{\mathbf{X}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{X}_{it}$
- $\beta = .1$,
- $V_{it} \overset{i.i.d}{\sim}$Gamma$(\alpha, \gamma)$, where $\gamma = 1/\alpha = \exp(-.125 * \overline{\mathbf{X}}_i^2 + .5 * \overline{\mathbf{X}}_i * z_{it})$, where $z_{it} \overset{i.i.d}{\sim} \mathbf{N}(0,1)$,
- $T = 5$, N =500 if it is not specified,
- $Y_{it} = \exp(\mathbf{X}_{it}\beta)C_iV_{it}$,
- T=5, Number of simulations =5000 .

The estimation results are in Table B.9; for the LFE, PQML and GMM, estimates are almost the same as Table B.8: all of them are very close to the true value of parameter; in terms of efficiency, LFE is best and PQML is the worst. Now with the new estimation method, OIV, we can see that it performs very well considering the standard errors, which are close to those of LFE. Finally, we do find a method that can compete with LEF.

2.6 Empirical Application

In this section we illustrate GMM estimator as well as PQML by estimating APE of price on demand for airline market. The market is defined the same as in Park et. al. (2007), which is a trip between origin and destination cities. For each route, i,e, an pair of cities, the measurement is taken everyday. In order to wash out the daily fluctuations, like weekend and holidays, the yearly average is taken from year 1997 to 2000. Of course, for some of the routes, they do not have records for these consecutive four years and have to be dropped; but the drop rate is less than 10%-we have over 1000 routes. So the data set is a typical balanced panel data set. The description of the key variables as follows:

For each route,the variable *passen* measures the average number of daily passengers for the year; $lfare$ is log transformation of average one-way fare in US dollars; *concen* is the market share of the biggest carrier along the route. Refer to Table B.10 for details.

The Table B.11 shows the estimation results for the three methods. The GMM method reduce the standard errors by about 10% compared with PQML. The estimates of PQML and GMM are close to each other. As for the practice purpose, from equation (2.4.12) in section 4, we know the APE for $lfare$

$$\widehat{\beta}_j \exp(\mathbf{x_t}\hat{\boldsymbol{\beta}}_{pqml})\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\boldsymbol{\beta}}_{pqml})} \approx 512 \tag{2.6.1}$$

$$\hat{\beta}_j * N^{-1}\sum_{i=1}^{N}\exp(\mathbf{x}_t\hat{\boldsymbol{\beta}}_{pqml} + \mathbf{G}^M(\overline{\mathbf{X}}_i)\hat{\boldsymbol{\pi}}_M) \approx 507 \tag{2.6.2}$$

$$\hat{\beta}_j * \exp(\mathbf{x}_t\hat{\boldsymbol{\beta}}_{pqml} + \mathbf{G}^M(\overline{\mathbf{x}})\hat{\boldsymbol{\pi}}_M) \approx 421 \tag{2.6.3}$$

Where $\mathbf{x}_t$ is evaluated at the median values of $lfare$ and *concen* and year dummy is 1998; $\overline{\mathbf{x}}$ is evaluated at at mean values of $lfare$ and *concen*. For the interest variable $lfare$, a 1% price rise in air tickets will exclude about 500 passengers overall while about 400 at the specific setting. Compare those result with LFE: a 1% price rise in air tickets will exclude about 6 passengers overall (here we evaluate *passen* at its mean). By calculating the corresponding standard errors, all the estimates are statistically significant at 5% level. From Figure B.5, we can see that *passen* is skewed to the right and is not symmetric even after taking logarithmic transformation. In terms of partial effect of airfare on market demand, the advantage of equation 2.6.3 is that it can give arbitrary partial effect of airfare on demand for any value of $\mathbf{x}$. For example, we can consider the APE of airfare on market demand for the route which has the highest demand of 8497: in Year 1998, a 1% price rise in air tickets will exclude about 545 passengers from the route; this way, the price elasticity of demand is about 6%.

2.7 Conclusions

We summarize the three usual estimation methods for positive response variables. By simulation, we list the disadvantages and advantages for each of them-basically the trade off between consistency and efficiency. As long as the conditional mean function is correctly specified,the PQML is robust to any distribution of the error terms. On the other hand, it suffers for less preciseness for error with certain distribution. If we can assume more on the distribution, we can use GMM to reduce the standard errors of PQML estimator. By the application of the airfare data set, we show that it works. As for the estimation of APE, it keeps the property of distribution free and amplify the positivity of the conditional mean function of the heterogeneity.

Many issues can be studied in the future research. For one, whether it is possible to develop a test to distinguish PQML and LFE, like the Hausman test for RE and FE estimators. Another interest problem is what the consequence if the conditional mean function is misspecified.

Chapter 3

A SPATIAL ANALYSIS OF SPENDING EFFECT ON MEAP

3.1 Introduction

As we all know, public schools of the K-12 educational system in the U.S. are financed mostly by local revenue, primarily by taxes levied on property. One of the disadvantages of this policy is that this can potentially lead to economic inequality across school districts within a state since, as is often argued, demand for (and affordability of) a good education increases with parental income and educational attainment. Take Michigan as an example, in year 1992, the per pupil expenditure in a rich school district (the name is Bloomfield Hills School District, DCODE 63080)could reach as more than 9 times as it in a poor school district (the name is Ionia Township S/D #2, DCODE 34360).[1] It is under this kind of background, in 1994, Michigan initiated a school finance reform which is called Proposal A, aimed at equalization of school finances among school districts within state boundaries. A great body of research has been done for its impacts. Papke(2005, 2008)use panel data sets, either school level or school district level, and find that there is statistically significantly positive relationship between student performance which is measured in the pass rate of math test for fourth graders and finance expenditure with linear regression models. Moreover, as pointed out in Papke (2008), the magnitude of the effect of initially high-performing districts are lower than the initially low-performing ones [2], which suggests clearly the nonlinearity of the relationship. Under this circumstance, Papke and Wooldridge (2008) extend the analysis further to a nonlinear model with a panel data set of school district but with less

[1] We get this from the data set used in Papke and Wooldridge (2008); In terms of household income, Chakrabarti and Roy (2012) find that the median income in a rich school district is more than three times of it in a poor one.

[2] and Roy (2011) even claims that the finance reform may have had negative effect on student's performance in the highest spending districts.

time span compared with Papke (2008). One of the challenges in Papke and Wooldridge (2008) is to deal with the unobserved heterogeneity in probit setting; with strictly exogenous explanatory variables, they propose a conditional normal assumption following Mundlak (1978) and Chamberlain (1980) device. The usual shortcoming for nonlinear models, to make inference about the average partial effect, is elegantly overcome. More recently with even boarder data coverage, Chakrabarti and Roy (2012) study the impact of Proposal A on spatial segregation of housing market and find that there is continued high demand for residence in highest-spending districts, suggesting the importance of neighborhood peer effects ("local" social capital). However, none of them accounts for the spatial dependence in the unobserved cross sectional fixed effect, either building level or school district level. This is the first reason that the problem is revisited.

On the other hand, as pointed out in Papke (2008), using a data set with time span of 10 years, Papke (2008) finds that although spending inequality was reduced in the years of immediately after Proposal A, equalization has slowed considerably since year 2000. Chakrabarti and Roy (2012) also notice that there is continued high demand for residence in the highest-spending communities, implying that even a comprehensive government aid program can fail to make a large impact on residential segregation. So considering the data set we use is in Year 2010, it is interesting to investigate the effect Proposal A after 15 years of implementation, especially with spatial effects controlled.

Spatial econometrics can be understood as a parallel extension of time series with time index replaced by space. In applied literature, issues relating to geographic proximity, transportation, spillover effects, etc., are important. Indeed, in recent years the spatial analysis in economics is booming: refer to Case (1991), Anselin and Florax (1995), Kelejian and Prucha (1999), Anselin (2010) and literatures therein. Modeling spatial interactions that arises in spatially referenced data is traditionally done by incorporating the spatial dependence into the covariance structure via an autoregressive model. For example, Wall (2004) analyzed the SAT scores of all 48 contiguous states of America for the year of 1999 by two mostly used models in spatial statistics: conditional autoregressive model (CAR) and simultaneously autoregressive model (SAR). Both of models cooperate spatial dependence in the covariance structure as a function a neighbor matrix and often a fixed unknown spatial correlation parameter; refer to the paper and Banerjee et.al.(2004) for more details. Obviously, that approach is not robust to the misspecfication of the covariance functional form; what's more, as it states in Conley (1999), whenever there are errors in the measurement of spatial dependence, which is happening very often in application, we cannot estimate parameters of interest consistently without assuming distributions of the errors; as we all know, the distribution assumptions are most of the time naive in reality. Conley and Molinari (2007) show how poorly MLE could perform when the distribution is misspecified; while the method in Conley (1999) works well. Moreover, he extends the spatial dependence to a broad economics system, which is general dependence within a cross section and not necessarily related to geographical features; for example, he creates a notion of "economics distance" among the some countries in the world in Conley and Ligon (2002); the physical distance definitely will contribute to it but it also has some type of "border effect" as in Engle and Roger (1996). On the other hand, he also pioneers a nonparametric approach for covariance structure estimation; the basic idea of that estimator, as it is pointed out in Keller and Shiue (2007), it is spatial version of Newey-West (1987) type heteroskedasticity and autocorrelation robust covariance estimator. The method is nicely used in many areas of economics, such as development, international economics and labor, etc; refer to Conley and Ligon (2002), Conley and Topa (2002), Conley and Dupor (2003).

However, Conley (1999) does not elaborate about how to choose the cutoff points other than the asymptotic condition of one third order of the sample size(or the width of sample region). By simulation, Conley and Molinari (2007) argue that the choice is relevant; so it might be hard for an empirical researcher to apply the method. Hence, the purpose of this paper is twofold. First, we use the Michigan MEAP data of year 2010 to retrieve the studies in Papke (2005, 2008) and in a nonlinear regression model; to investigate the effect of spatial dependence specified by Conley (1999), we propose two ways to choose cutoff points and show how the spatial dependence corrected standard errors are related to the choice. Second, in order to get the average partial effect (APE) as in Papke and Wooldridge (2008), an average structure function (ASF) device in Blundell and Powell (2003) is used to get the estimates of the spending effect as well as other ones. The remainder of this paper is organized as follows. Section 2 introduces the linear and nonlinear models; data description and related issues are contained in section 3. Section 4 presents the linear regression with all kinds of standard error calculations. Section 5 provides the results of nonlinear model and how the APE is estimated correspondingly. Conclusions and discussions are in Section 6.

3.2 Model

Let (Y_i, $\mathbf{X}_i$) be a sequence of observations for area A_i; first following Papke (2005, 2008), the linear model is considered:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + C_i, \quad E(C_i|\mathbf{X}_i) = 0 \tag{3.2.1}$$

where Y_i is scalar, $\mathbf{X}_i$ is $1 \times K$ vector and $\boldsymbol{\beta}$ is $K \times 1$ vector of unknown parameters; for area A_i: Y_i and $\mathbf{X}_i$ can be understood as $Y(A_i)$ and $\mathbf{X}(A_i)$ respectively; but for simpler denotation, we just put them as Y_i and $\mathbf{X}_i$. C_i or $C(A_i)$ is the unobserved spatial heterogeneity for each area A_i. If there are total areas of N, then $i = 1, 2, \cdots, N$.

Under general conditions [e.g., Wooldridge (2010, Chapter 4)], the OLS estimator of $\boldsymbol{\beta}$ based on N observations looks like:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{3.2.2}$$

Where $\mathbf{Y} \equiv [Y_1, Y_2, \cdots, Y_N]'$ is $N \times 1$, $\mathbf{X} \equiv [\mathbf{X}_1', \mathbf{X}_2', \cdots, \mathbf{X}_N']'$ is $N \times K$. With little algebra [refer to Wooldridge (2010, Chapter 4)], we can get:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i'C_i\right), \tag{3.2.3}$$

So with general conditions, $\hat{\boldsymbol{\beta}}$ is weakly consistent for $\boldsymbol{\beta}$ and asymptotically normally distributed; note that the asymptotic distribution really hinges on the second term in equation (3.2.3).

If we assume the following conditions:

$$Var(C_i|\mathbf{X}_i) = \sigma^2,\ i = 1,\ 2,\ \cdots,\ N; \tag{3.2.4}$$

$$\{\mathbf{X}_i' C_i\colon\ i = 1,\ 2,\ \cdots,\ N\}\ \text{ is an uncorrelated sequence.} \tag{3.2.5}$$

Condition (3.2.4) is the appropriate homoskedasticity assumption and (3.2.5) is no correlation requirement for the sequence; it can be also extended to be independent. Then the covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Let $\hat{\sigma}^2 = SSR/(N-K)$ be the usual square of the standard error of the regression; the usual standard error of jth OLS estimator $\hat{\beta}_j$ is the square root of the jth diagonal element of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. That is also the 'standard error' printed out by all regression packages.

Without homoskedasticity [condition (3.2.4)] but with no correlation [condition (3.2.5)], we can allow arbitrary heteroskedasticity of conditional variance C on $\mathbf{X}$ by Huber (1976) and White (1980) robust standard errors, which are square root of diagonal terms of matrix $(\mathbf{X}'\mathbf{X})^{-1}(\sum_{i=1}^{N}\mathbf{X}_i'\hat{C}_i^2\mathbf{X}_i)(\mathbf{X}'\mathbf{X})^{-1}$. Note that $\frac{1}{N}(\sum_{i=1}^{N}\mathbf{X}_i'\hat{C}_i^2\mathbf{X}_i)$ is a consistent estimator of $Var\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i'C_i\right)$, the second term of equation (3.2.3). Algebra details are referred to Wooldridge (2010, Chapter 4).

For those two types of standard errors, they allow no correlation among the sequence $\{\mathbf{X}_i'C_i : i = 1,\ 2,\ \cdots,\ N\}$. To relax that assumption, we can try two directions. One is to divide the whole N observations into G groups: $\mathbf{W}_g = (\mathbf{X}_i',\ \cdots,\ \mathbf{X}_j')'$, $\hat{\mathbf{v}}_g = (\hat{C}_i,\ \cdots,\ \hat{C}_j)'$, where $g = 1, \cdots, G$; then the group corrected standard errors of elements of $\hat{\boldsymbol{\beta}}$ are square root of diagonal terms of matrix $\left(\sum_{g=1}^{G}\mathbf{W}_g'\mathbf{W}_g\right)^{-1}\left(\sum_{g=1}^{N}\mathbf{W}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbf{W}_g\right)\left(\sum_{g=1}^{G}\mathbf{W}_g'\mathbf{W}_g\right)^{-1}$, which allow arbitrary correlation within each group while independence is assumed across groups; that is similar to the idea for cluster corrected robust standard errors in Wooldridge (2010, Chapter 20); and Wang et. al (2013) also adopt that idea: split the whole sample into many groups, one of which has two observations with arbitrary dependence; and a bivariate Probit method is proposed. Note that when $G = 1$, then the cluster corrected standard errors are H-W standard errors.

Conley (1999) goes that way further: he divides the sequence $\{\mathbf{X}_i' C_i : i = 1, 2, \cdots, N\}$ into two groups with one group is lagged the other in two non-opposing directions and allow arbitrary correlation between the two groups; with different lag, we can get two different groups and corresponding correlation between them-that is, the sequence is regrouped with times of the number of lags; then final estimator of $Var\left(\sum_{i=1}^{N} \mathbf{X}_i' C_i\right)$, which is denoted as $\hat{\mathbf{V}}$, is the weighted sum of all the covariances. Let $\hat{C}_i$ be the OLS residual in equation (3.2.1) , that is:

$$\begin{aligned}\hat{\mathbf{V}} &= \sum_{i=0}^{L_1(N)} \sum_{j=0}^{L_2(N)} W(i,j) \sum_{s=i+1}^{N} \sum_{t=j+1}^{N} (\mathbf{X}_{s-i}' \hat{C}_{s-i} \hat{C}_{t-j} \mathbf{X}_{t-j} + \mathbf{X}_{t-j}' \hat{C}_{t-j} \hat{C}_{s-i} \mathbf{X}_{s-i}) \\ &- \sum_{s=1}^{N} \sum_{t=1}^{N} (\mathbf{X}_s' \hat{C}_s \hat{C}_t \mathbf{X}_t) \qquad (3.2.6)\end{aligned}$$

Where $W(i,j)$ are the weights for covariance with one direction lagged of i and the other j; Without lags, that is $i = j = 0$, then $W(i,j) = 1$. $L_1(N)$ and $L_2(N)$ are the cutoff points of the two direction and both of them converge to ∞ with order of $L_1(N) = o(N^{1/3})$ and $L_2(N) = o(N^{1/3})$ as $N \longrightarrow \infty$. Intuitively, $\hat{\mathbf{V}}$ can be understood as the Newey and West (1987) estimator of $Var\left(\sum_{i=1}^{N} \mathbf{X}_i' C_i\right)$ in two non-opposing directions summed up with proper weights; so it is basically a nonparametric estimator. Hence, Conley's estimator of variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{V}} (\mathbf{X}'\mathbf{X})^{-1}$.

As for the nonlinear model, we follow Papke and Wooldridge (1996):

$$Y_i = G(\mathbf{X}_i \boldsymbol{\beta} + C_i), \qquad (3.2.7)$$

Where $G(\cdot)$ is any continuous function with range of $(0,1)$ in the real line; for example, we can have $G(x) = \frac{\exp(x)}{1+\exp(x)}$ or $G(x) = \Phi(x)$, where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{2\pi} \exp(-t^2/2) dt$. So it is natural to have:

$$G^{-1}(Y_i) = \mathbf{X}_i \boldsymbol{\beta} + C_i, \qquad (3.2.8)$$

The equation (3.2.8) is very appealing since its right hand side has the linear form as equation (3.2.1) does. For example, if $G(\cdot)$ takes the logistic function form, then equation (3.2.8) turns into:

$$\log\left(\frac{Y_i}{1-Y_i}\right) = \mathbf{X}_i\boldsymbol{\beta} + C_i, \tag{3.2.9}$$

This is the popular linear model for log-odds ratio. As in Papke and Wooldridge (1996), with usual assumption like, $E(C_i|\mathbf{X}_i) = 0$, we can get $E\left[\log\left(\frac{Y_i}{1-Y_i}\right)\middle|\mathbf{X}_i\right] = \mathbf{X}_i\boldsymbol{\beta}$. So we can model transformed response variable instead of original one. This is the advantage of equation (3.2.7); but we know the responses should be strictly in the interval of $(0, 1)$ since 0 and 1 will explode the transform. Note that the model in 3.2.9 is different from the model specification $E(Y_i|\mathbf{X}_i, C_i) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta}+C_i)}{1+\exp(\mathbf{X}_i\boldsymbol{\beta}+C_i)}$, in which case the responses can be 0 and 1. Another issue is the average partial effect, which is hard to be estimated in most nonlinear models. But since we have equation (3.2.7), from which we can get:

$$\frac{\partial Y_i}{\partial x_{ij}} = g(\mathbf{X}_i\boldsymbol{\beta} + C_i)\beta_j, \tag{3.2.10}$$

where $g(x) = \frac{\partial G(x)}{\partial x}$. The rest of the paper will keep models (3.2.1) and (3.2.8), and assume $E(C_i|\mathbf{X}_i) = 0$ and necessary rank condition, then OLS estimators of $\boldsymbol{\beta}$ in both models are consistent, we will check standard errors under different conditional covariance structures of $\hat{\boldsymbol{\beta}}$ in (3.2.2) and the corresponding APE estimates.

3.3 Data

3.3.1 Data Characteristics and Sources

All the data are from Michigan Department of Education: expenditure data are from Bulletin 1014 ; enrollment and free and reduced price lunch data are from the Center for Educational Performance and Information (CEPI); the test results data are from Michigan Educational Assessment Program (MEAP). The data resource is the same as in Papke (2005, 2008) and Papke and Wooldridge (2008), but there is some change for the name of the program since we use the data of Year 2010. Papke (2005) uses building level data but we use school district level ones, which are the same as in Papke (2008). Other than the reasons explained in Papke and Wooldridge (2008), we pick the school districts for the convenience of spatial analysis. As for the dependent variable, $math4$, the pass rate of math test for forth graders in each school district of year 2010, which is defined the same in Papke (2005, 2008). The variable of interest, average expenditure per pupil in each school district, is collected differently than in Papke (2005, 2008); instead, we follow the definition in Papke and Wooldridge (2008)-take the average per pupil real expenditure in the last four years: $avgexp = (avgexp + avgexp_{-1} + avgexp_{-2} + avgexp_{-3})/4$. Since both papers find significant effect of the previous expending on school performance, it is meaningful to collect expenditure variables this way-the reason we take last 4 years is that we are considering the test pass rates for forth graders. The real dollars are calculated in the price of year 2010 Midwest Urban Price index from Bureau of Labor Statistics. For the purpose of comparison, we also use the average per pupil expenditure of year 2010. Two other explanatory variables are the same in previous papers except in year 2010: *enroll* is the student enrollment for each school district in academic year 2009/10 and *lunch* is percentage of students who are eligible for free lunch or reduced price lunch program. *scdist* is the distance in kilometers for each school district to its nearest 2-year or 4-year college. Inspired by Kane and Rouse (1995), we are trying to investigate the effect the higher education has on the K-12 system, which is not done in previous studies. Intuitively, the schools in a school district which is closer to a college should have a higher chance to perform well, like it is easier for a forth grader to find a tutor since the college students are around; or the students can take advantages of the facilities in the colleges. On the other hand, most locations of colleges are in the areas with higher social economic status, which has positive externality to public schools. The summary statistics for these variables are in Table C.1.

There are 551 school districts in shape file which is from Michigan Geographic Data Library. Combined with MEAP data, there are only 518 school districts because of data missing. With help of ArcMAP, we plot the the map of the all the school districts in Figure

C.1. As for the colleges in Michigan, I download all the 131 Michigan 2-year and 4-year colleges from National Center for Education Statistics website, and plot into the map(Figure C.2). If we plot the test pass rates into the map (Figure C.3), we can find some areas are clustered with high rates while others with low; that is, the there is possibility that high rates are congested together; so do the lower rates. That is the what the spatial dependence stems from. When add college data in and get Figure C.4, we do find that some colleges are in or close to the areas with high math test pass rates; while other colleges are in the areas with low test rates. How to measure the spatial dependence among those 518 school districts is the main part of application. Since the shape of each district is not regular, and the area sizes of them vary from $3.7km^2$ to $3317.7km^2$, we should treat them as area data instead of point ones. The next subsection will explain how to set up the spatial dependence among these available 518 school districts.

3.3.2 Spatial Dependence Measurement

Since we treat each school district as an irregular lattice, the physical distance between the centroids of any two districts cannot be a perfect measure for spatial dependence among them; for example, for some districts, their centroids are outside of the polygons. So distance based spatial weighting matrix, which is perfect for points data, is not that appropriate here; on the other hand, the dependence which is measured only by contiguity for any two areas ignores the size variations among those school districts. Keller and Shiue (2007) apply both methods and have excellent descriptions about them. Since Conley (1999) initiates a measurement for spatial dependence, which is robust to measurement error (Conley, 2007), we will adopt their method. First, we get the projected latitude and longitude coordinates for the centroid of each school districts[3]; then we use the smallest distance among all those 518 centroids to construct little squares in the whole state; whenever a centroid is in a square, the

[3] the reason we use projected lattitude/longitude instead of original ones is that we want to use coordinates in a plane rather a sphere

coordinates of the upright corner the lattice are the new coordinates for each school district. A picture is worth one thousand words: we choose part of the 518 school districts and make a picture to explain the process: as in figure C.5, we have 96 school districts and we plot the 96 centroids into the map (Figure C.6); then we divide the whole area into 70 × 70 little squares in Figure C.7 with unit of about 2.5 kilometers, which is about the smallest distance among those 96 points. Now, we know the new coordinates for each centroid are decided by the little square where it lies and those are called Conley coordinates for each school district in Figure C.8. From there, we can see that one of the advantage of this measurement is that as long as a centroid is in a square, it will have the same coordinates no matter where it stay in the square; this way, it can cover the damage of measurement error while the traditional ways cannot. Thereafter, we define the spatial weighting matrix, $W = (w_{ij})$, this way:

$$w_{ij} = \begin{cases} (1 - \frac{|i|}{L_1})(1 - \frac{|j|}{L_2}), & \text{for } |i| < L_1,\ |j| < L_2 \\ 0, & \text{else} \end{cases}$$

We can see that the weight function is a Barlett window in each direction. The choice of L_1 and L_2 is based on theory and practice: In theory, it should not be bigger than 1/3 root of the sample region; among those 518 school districts, the sample region is about 1000 horizontally and vertically; so we can take as about 10, which is about 10 kilometers since we know each unit of the coordinate is about one kilometer. While from practice, we know we should have a distance longer than that. We will explain more about that in the analysis. For the coordinates beyond L_1 and L_2, we would not consider the dependence any more. Of course, we can set L_1 and L_2 as very large and so that we will include all the sample districts. the The missing data definitely will not cause big problems for the matrix defined this way, while the $0-1$ measure of spatial dependence based on contiguity will: some element would be one were data for all 551 districts available. This is another advantage compared with conventional method. With this spatial weighting matrix, we will use the Conley (1999) estimator described in section 2 to account for dependence among school districts.

3.4 Linear Model

As in Papke (2005, 2008), we run the following regression:

$$\begin{aligned} math4_i &= \beta_0 + \beta_1 \log(avgexp)_i + \beta_2 \log(enroll)_i + \beta_3 lunch_i + \beta_4 \log(scdist)_i + C_i, \\ math4_i &= \beta_0 + \beta_1 \log(exp10)_i + \beta_2 \log(enroll)_i + \beta_3 lunch_i + \beta_4 \log(scdist)_i + C_i, \\ & \quad (4.1a, 4.1b) \end{aligned}$$

The only difference of those two models [equations (4.1a, 4.1b)] is the first explanatory variable: we replace *avgexp* in the first regression with *exp*10. From the description, we know that some differences do exist between them and we wonder how those difference matter for the test rate. The reason we take logarithms transformation for explanatory variables *avgexp* and *scdist* is that their distributions are skewed: It can be easily seen from the histograms (Figure C.9); and Manning (1998) claims that logarithms transformation is a good way to model those kinds of variables. Such analysis also applies to *enroll*.

3.4.1 Ordinary Least Square (OLS)

Here, we first assume that usual assumptions of linear regression as in Wooldrdige (2010, Chapter4) hold; the OLS results are in the first two columns of Table C.2. Note that the usual standard errors are calculated under assumption of homoskedasticity as well as other assumptions for consistency. As we all know, the magnitudes of standard errors are crucial to the inference: they will decide whether the estimates of parameters are significant or not at a given significance level. From Table C.2, we can see that all the explanatory variables except *scdist* are significant at 5% level: the effect of student expenditure as well as enrollment on school performance is positive and lunch program is negative; interestingly, the effect of a school district's distance to its nearest college is positive which means the longer the better and it is against usual intuition. Fortunately, the effect is not significant at 5% level or even at 10%; and we see the same pattern for the second regression and the size of the effect of the last four years' expenditure on school performance is slightly bigger than year 2010. However, if we abandon the homoskedasticity assumption and allow arbitrary functional form of the conditional variance of the error term C_i on explanatory variables and we will get the the White standard errors; then the positive effective of student enrollment is not significant at 5% level any more; so does for the second regression. Further, if we calculate the robust standard errors corrected for the intermediate school districts (ISD), even the positive effect of year 2010's expenditure is not significant at 5% level. Considering Michigan geographical feature, we divide the whole sample into two clusters: the Upper peninsula and lower; if we correct for the peninsula clusters, neither the last four years' average expenditure nor year 2010's expenditure has the significant positive effect on school performance. Now, we know it is important to have a reasonable method to calculate standard errors.

Now we calculate Conley (1999) standard errors; Based on the description in section 2 and 3, we can get the coordinates for each centroid of 518 school districts. The key step is how to choose the cutoff points. Since the coordinates range from 0 to 758 horizontally, and from 0 to 880 vertically, from theory the cutoff points should not bigger than 10; while we think the number should be larger; we will design two schemes to see how the significance level changes with respect to its corresponding cutoff points. First, we have 11 pairs of the cutoff points, which increases by 50 or 100 by each step; and results are in Table C.3 and C.4. We can see that the overall trend of the standard error magnitudes is decreasing with increasing of the window size; but we do see some local fluctuations: for example in Table C.3, for the variable $exp10$, the Conley standard errors of the estimator of its parameter are increasing for the first three pairs of cutoff points. At 5% level, the estimators of β_1 and β_3 are all statistically significant under all the cutoff points; but for β_2, its estimator is not significant at 5% level if the first three pairs are chosen. In Table C.4, the similar pattern holds for the regression with $exp10$ replaced by $avgexp$, while the corresponding errors are a little inflated. Secondly, we have 7 pairs of points, which are the the 5%, 15%, 25%, 50%, 75%, 95% and 100% percentiles of the coordinates respectively; and results are in Table C.5 and C.6. Note that, with each of the pair, we can make that many of the 518 sample covered in the weight calculation. Surprisingly, the similar story happens as the first scheme; and there is no difference of signifiance between only 5% samples are considered for positive weights and all samples are included for β_1 and β_3 at 5% level; while we do see the difference for β_2: more sample considered for positive weights more significant. Considering each unit of the coordinates is equal to about 2 kilometers, it is not that hard to have a reasonable choice; there is a catch here: if the decreasing trend of the Conley standard errors holds, we make them as small as possible: for example, I make a pair of 10^6, even the β_4 can be statistically significant at 5% level; but we know that number does not make any sense considering none of the coordinates is greater than 900.

3.4.2 Generalized Least Square (GLS)

The matrix form of equation (3.2.1) [or equations (4.1a, 4.1b)]:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{C} \tag{3.4.1}$$

Where $\mathbf{Y}$ and $\mathbf{X}$ are as in equation (3.2.2), $\mathbf{C} = [C_1, C_2, \cdots, C_N]'$.

If we assume the Spatial Autoregressive (SAR) structure in the error term $\mathbf{C}$:

$$\mathbf{C} = \rho\mathbf{W}\mathbf{C} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0},\ I_N\sigma^2) \tag{3.4.2}$$

Where ρ is the unknown spatial correlation parameter and $\mathbf{W}$ is the weighting matrix; that is a conventional assumption in the spatial econometrics literature. Based on theoretical results of maximum likelihood estimator (MLE) and QMLE in Lee (2004), STATA has the build-in function of "spreg ml" to implement MLE or QMLE given a weighting matrix.

The first interesting result is that estimates of covariate parameters $\boldsymbol{\beta}$ are close while the estimates of spatial correlation parameter ρ are different for different weighting matrix. From Table C.17 and Table C.18, we can see that the magnitude of ρ estimates under weighting matrix of contiguity is only one third of it under weighting matrix of inverse distance; if we investigate the weighting matrix further, we find that contiguity weighting matrix is much smaller than the inverse distance weighting matrix element by element; the usual summary statistics of the correlation matrix of error terms are in Table C.19 and Table C.20.

One strong assumption of MLE or QMLE for SAR model is the homoskedasticity; without it, MLE (or QMLE) are not consistent. That result has been explored thoroughly in the literature , such as Arraiz at. el. (2010), Kelejian and Prucha (2010), Lin and Lee (2010). They all come up with some new methods to cover heteroskedasticity. On the other hand, the SAR structure might be misspecified. Here we combine the idea of SAR structure and Conley standard errors, and a new form of standard errors is introduced. From equation (3.4.2):

$$\mathbf{C} = (I_N - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$$

Plug into queation (3.4.1) :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + (I_N - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \tag{3.4.3}$$

With the homoskedasticity, GLS in euqation (3.4.3) is the same as MLE with $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, I_N\sigma^2)$ (or QMLE without normality asumption); that is aslo what “spreg ml” reports in STATA. Moreover, GLS is also equivalent to the OLS in the following model:

$$(I_N - \rho\mathbf{W})\mathbf{Y} = (\mathbf{I_N} - \boldsymbol{\rho}\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.4.4}$$

Since ρ is unkown, by the idea of quasi-GLS, we can replace ρ with its GLS estimator, $\widehat{\rho}$, in equation (3.4.3). Then we apply Conley's method to the following model:

$$\mathbf{Y}^{*} = \mathbf{X}^{*}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.4.5}$$

Where $\mathbf{Y}^{*} = (\mathbf{I_N} - \widehat{\boldsymbol{\rho}}\mathbf{W})\mathbf{Y}$, $\mathbf{X}^{*} = (\mathbf{I_N} - \widehat{\boldsymbol{\rho}}\mathbf{W})\mathbf{X}^{*}$.

Note that OLS in equation (3.4.5) wtih Conley standard errors has at least two advantanges: first, it allows hetroskedastictiy by H-W standard errors and robust to condition in equation (3.2.5); second, it is robust to the misspecification of SAR struction. Note that even if the SAR structure is wrong, this transformation adds one more possible structure for the error terms; and we know that Quasi-GLS is still consistent as long as $E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0$, which means strictly exogenous covariates. From the results in Table C.13, the heteroskedasticity does make a difference for magnitude of standard errors of expenditure and enrollment; the similar patterns for the Conley standard errors as in OLS case. Since the efficiency gain from Conley's method only works for high cutoff points in which case correlation is weak, we can see that the SAR structure is not a good choice for error term.

3.5 Nonlinear Model

3.5.1 Regression of Log Odds Ratio

We follow the model in equation (3.2.7) with the following form :

$$Y_i = \frac{\exp(\mathbf{X}_i\boldsymbol{\alpha} + e_i)}{1 + \exp(\mathbf{X}_i\boldsymbol{\alpha} + e_i)}, \tag{3.5.1}$$

One of the advantages of the above equation (3.5.1), compared with equation (3.2.1), is that logistic function of its right hand side is strictly in the range of (0,1), which will make prediction in the interval; however, it comes at the expense of the difficulty of estimating parameters and causal effect. Fortunately, the equation (3.5.1) can be easily transformed as follows:

$$\begin{aligned}
\log\left(\frac{math4}{1 - math4}\right) &= \alpha_0 + \alpha_1 \log(avgexp)_i + \alpha_2 \log(enroll)_i + \alpha_3 lunch_i + \alpha_4 \log(scdist)_i + e_i, \\
\log\left(\frac{math4}{1 - math4}\right) &= \alpha_0 + \alpha_1 \log(exp10)_i + \alpha_2 \log(enroll)_i + \alpha_3 lunch_i + \alpha_4 \log(scdist)_i + e_i, \\
&\quad (4.5.1a,\ 4.5.1b)
\end{aligned}$$

The left hand side of the above two equations (4.5.1a, b) are the logarithm transformation of relative risk of passing the MEAP math test for each school district; it is popularly called as "log odds ratio" in literature. The right hand sides are exactly the same as in equations (4.1a, b). Avoiding abusing of notation, we use $\boldsymbol{\alpha}$ for the parameters of interests; more important is that interpretations of the parameters of interest are different: the $\boldsymbol{\beta}$ in equations (4.1a, b) are the average partial effect (APE) of covariates on pass rate, while α's in equations (5.1a, b) are APE of covariates on log ratio of each school's pass rate to its failure rate. Of course, we can retrieve the first APE and it is in the next subsection.

First, we regress the log odds ratio under four types of assumptions for the conditional covariance of e_i on the regressors and the estimates are in the first column of Table C.7; as in previous section, we consider two sets of expenditure variable: *avgexp* and $exp10$. The signs of the estimates are as expected: the effect of expenditure and enroll on the odds ratio is positive; although the *lscdist* shows some positive effect, it is not statistically significant at even 10% level, let alone 5%. we remember those results show up in the level pass rate regression in the previous section. However, the significance picture is different. Under iid assumption, increasing *avgexp* by 1% will increase the odds ratio by about 5% and about 4% for $exp10$; and the effects are statistically significant at 5% level. If we correct for arbitrary correlation among e_is or between e_is and explanatory variables, the significance claim does not hold any more; nor do the corrected clusters of ISD and peninsula cases. That gives us motivation to investigate further about the assumptions for the conditional covariance structure. On the contrary, the effect of enrollment on pass rate ratio is robust to the assumptions: if we increase the enrollment by 5%, then the odds ratio will increase by 9% with either *avgexp* or $exp10$ controlled; and the effect is statistically significant at 5% level. The same pattern holds for *lunch*: every 1% point increase in the percentage of students in free lunch or reduced price lunch for a school district, it will lead to reduce the odds ratio by almost 3%; and the effect is statistically significant at 5% level.

Then, we consider the Conley (1999) standard errors. As in the level pass rate case, it is important to find an appropriate window size for the weights; we apply the same scheme. For the first 11 pairs of cutoff points and corresponding results are in Table C.8 and C.9, the global trend of the magnitude of standard errors for each variable is decreasing with increase of cutoff points; while for some variable, like log($exp10$), there is some local distortion, and for other variables, like *lunch* and log(*enroll*), the decreasing trend is strict. Given at 5% level, the statistical significance of OLS estimators of *lunch* and log(*enroll*) is robust to the choice of cutoff points: all of them are significant; while for log(*scdist*), the OLS estimator of its parameter is never significant at 5% level even at 1000 of the window size. So we are very confident to say that there positive effect of *enroll* on the odds ratio of math test pass rate, while negative for *lunch* at 5% significant level; and *scdist* is not significant factor to explain the ratio. As for log($exp10$), its effect only becomes statistically significant at 5% level when we set the window at size of 1000; the similar story is told when log($exp10$) is replaced by log(*avgexp*), whose positive effect on the odds ratio is changed to be significant at the cutoff points of 600 and 700 and over. When the percentiles are specified for the cutoff points, judging from the results in Table C.10 and C.11 the same thing happens for *lunch*, log(*enroll*) and log(*scdist*): the statistical significance of their effects on the odds ratio at 5% level holds the same situation as first scheme of cutoff points. For log($exp10$), it effect is not statistically significant at 5% level even though all the points are included in the weights to correct for spatial dependence; the case of log(*avgexp*) is slightly different: when 95% or above of the sample are included for weights, its effect is significant. The same comment shows up as for the level test pass rate: as long as we keep the cutoff size large enough, all the variables are statistically significant factors to explain the ratio although the magnitude of the window does not make any practical sense. In a word, if we try to explain the odds ratio of the test instead of rates themselves, only two factors are statistically significant at 5% level: *lunch* and log(*enroll*); the former keeps its significance while the latter is changed-it replace the expenditure variables as in level pass rates case.

3.5.2 Estimate the APE for Level Rates

As we all know, economists are interested in what a model would suggest for a policy as well as estimating parameters in the model itself. After we go to the estimation of parameters, we notice that one of disadvantages for the nonlinear model as in equation (3.5.1) is that we cannot estimate APE as smoothly as in linear case, where the estimates of parameters themselves speak everything; Plus, there is another issue-the difference between APE and PAE, refer to Wooldridge (2005, 2010) for more detailed discussion about this. In this paper, we define the APE with respect to x_j evaluated at $\mathbf{X}$ as follows:

$$APE_j\left(\mathbf{X}\right) = \alpha_j \int_{-\infty}^{+\infty} \frac{\exp(\mathbf{X}\boldsymbol{\alpha}+e)}{\left(1+\exp(\mathbf{X}\boldsymbol{\alpha}+e)\right)^2} f(e)de, \tag{3.5.2}$$

Note that equation (3.5.2) follows the idea of average structure function(ASF) in Blundell and Powell (2003); one of the advantages of ASF is that it can specify partial effect with respect to individual change of $\mathbf{X}$, not just average. Refer to equations (5.1a, b), $\mathbf{X}\boldsymbol{\alpha} = \alpha_0 + \alpha_1 \log(avgexp) + \alpha_2 \log(enroll) + \alpha_3 lunch + \alpha_4 \log(scdist)$; for example, if we want to know the expenditure effects on school performance, then APE with respect to $\log(avgexp)$ is of our interest; by equation (3.5.2), $APE_1\left(\mathbf{X}\right) = \alpha_1 \int_{-\infty}^{+\infty} \frac{\exp(\mathbf{X}\boldsymbol{\alpha}+e)}{(1+\exp(\mathbf{X}\boldsymbol{\alpha}+e))^2} f(e)de$. The notion of that is the partial effect defined in equation (3.2.10) with the heterogeneity averaged out over its whole population. Let $\hat{\boldsymbol{\alpha}}$ be parameter estimator and $\hat{e}_i$ be the residuals from log odds ratio OLS regression, it is straight forward to estimate the APE this way:

$$\widehat{APE}_j\left(\mathbf{X}\right) = \hat{\alpha}_j \frac{1}{N} \sum_{i=1}^{N} \frac{\exp(\mathbf{X}\hat{\boldsymbol{\alpha}}+\hat{e}_i)}{\left(1+\exp(\mathbf{X}\hat{\boldsymbol{\alpha}}+\hat{e}_i)\right)^2}, \tag{3.5.3}$$

The idea behind this method is the method of moment; it can also be understood as Duan's(1983) swearing estimator. But we know, we do not need to assume the independence of $\mathbf{X}_i$ and e_i, which can be an attractive feature. As for the asymptotic variance of $\widehat{APE}_j\left(\mathbf{X}\right)$, it can be obtained by delta method, which is conveniently implemented using the method of moments approach in Newey and McFadden (1994). Bootstrapping methods can also be readily applied.

Continuing the *avgexp* example, we can estimate the APE at the mean value of all $\mathbf{X}$ as .1151 with standard error of .0663. The results for other explanatory variables evaluated at mean and 25%, 50%, 75%, 95% percentiles are in Table C.12. We can see that the magnitudes of the APEs do correlate to the values at which the $\mathbf{X}$ are evaluated; take the *avgexp* example, the trend of its APE is decreasing with higher percentile of $\mathbf{X}$ and the difference of APE between 95% and 25% is about 2%: the average expenditure in the lower end of school districts kicks in at a higher rate than in the higher ones. This coincides the results in Papke (2008) who divides the sample into two groups by the median of average expenditure; thanks to the nonlinear model, we do not need to split the data and we can investigate it in an even finer setting: technically, we can get the APE for any value of $\mathbf{X}$, which definitely is attractive to the practitioners. We can also find the similar trend for APEs of lunch program and enrollment albeit the change is smaller. As for inference, the APEs of enrollment and lunch are statistically significant at 5% level when *avgexp* is used to control for expenditure, whose significance level is 10%; when the expenditure of year 2010 is used, the significance story for enrollment and lunch is the same while the expenditure is not statistically significant at 10% level any longer. This is another difference compared with line model: if there is a temporally lagged effect for the expenditure, the linear model cannot catch it while nonlinear does. To dig into the question deeper, we compare the estimates of linear and nonlinear models and put the results in figure C.10-C.12. First, for the average expenditure, the linear estimate is about $4 \sim 6\%$ higher than nonlinear ones, with the gap wider for the higher spending school districts. But for enrollment, the trend is reversed: the effect of enrollment estimated in the nonlinear model is higher than in the linear one, although the difference is about half percentage and the gap shrinks with more students enrolled. Positive effect of enrollment coincides with the "peer effect" in Lin (2010). And for free lunch program, its effects on school performance both in linear model and nonlinear are almost the same.

3.6 Conclusions

We have investigated the effects of Proposal A in Michigan on school performance in a broader setting compared with the previous literature, such as the nonlinear model, the interaction between K12 public schools and colleges; To control the spatial dependence among all school districts, we adopt the method in Conley (1999) and explore the effect of different cutoff points. Considering the statistical significance varies with standard errors, we find that the OLS estimates in the linear level test rates are statistically significant at 5% level given a reasonable window size; however, the picture is different for the nonlinear model, which is the linear in the log odds ratio of the level test rates: even for a very large cutoff points, the spending effect is not statistically significant at 5% level any more. What's more, after transformed back from log odds ratio into level rates, both the magnitudes and statistical significance of APEs are changed; one of the advantages of the nonlinear model is that it can catch the effect variation with specific part of population. Also, the difference between some estimates from the linear and nonlinear models is not negligible, this raises a question about issues of specifications of regression functional form.

As for the future work, the interaction between public schools and charter schools, as in Imberman(2011), is an interesting extension. Also, to make use of panel data is also a promising direction.

APPENDICES

Appendix A

PROOF OF THEOREM 1: Note that here we prove the 2SLS case; and OLS is special case. Abrevaya (2002) mentions how the delta method can be used for the Duan estimator, but considers the case where U, $\mathbf{X}$ are independent, which is not necessarily in our case

Let

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta)',\ \widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\eta})'.$$

$$\begin{aligned}
\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= \sqrt{N}\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\eta} - \eta \end{pmatrix} = \begin{pmatrix} N^{-1/2}\sum_{i=1}^{N} \mathbf{A}^{-1}\mathbf{Z}_i' U_i \\ N^{-1/2}\sum_{i=1}^{N} (\mathbf{P}\mathbf{S}_i + \exp(U_i) - \eta) \end{pmatrix} + o_p(1) \\
&\equiv N^{-1/2}\sum_{i=1}^{N} \begin{pmatrix} \mathbf{S}_i \\ Q_i \end{pmatrix} + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A} &\equiv (\mathbf{C}'\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{D}^{-1} \\
\mathbf{C} &\equiv E(\mathbf{Z}_i'\mathbf{X}_i),\ \mathbf{D} \equiv E(\mathbf{Z}_i'\mathbf{Z}_i),\ \mathbf{P} \equiv E(\mathbf{X}_i \exp(U_i)) \\
\mathbf{S}_i &\equiv \mathbf{A}^{-1}\mathbf{Z}_i' U_i \\
Q_i &\equiv \mathbf{P}\mathbf{S}_i + \exp(U_i) - \eta
\end{aligned}$$

So by central limit theorem, this finishes the proof. The above process is similar to the solution to question 12.17 in Wooldridge (2010). For details, refer to Wooldridge (2011).

PROOF OF LEMMA 1: From (1.4.3), and by monotonicity of logarithmic function,

$$\begin{aligned}
& \exp(-CM^{-s} + \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) \\
< \; & r(\mathbf{x}) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) \qquad \text{(A.1)} \\
< \; & \exp(CM^{-s} + \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M),
\end{aligned}$$

By the mean value theorem applied to the lower and upper bound:

$$\begin{aligned}
& \exp(-CM^{-s} + \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) \\
= \; & -CM^{-s}\exp(\xi_1), \quad \xi_1 \in [-CM^{-s} + \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M, \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M]
\end{aligned}$$

$$\begin{aligned}
& \exp(CM^{-s} + \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) \\
= \; & CM^{-s}\exp(\xi_2), \quad \xi_2 \in [\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M, CM^{-s} + \mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M]
\end{aligned}$$

So, for the $\boldsymbol{\pi}_M$ that satisfies (A.1), we have:

$$\sup_{\mathbf{x}\in\Xi} \left| r(\mathbf{x}) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M) \right| < CM^{-s},$$

So,

$$E\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M)\right)^2 \leq CM^{-2s}$$

Note that:

$$\begin{aligned}
& E\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi})\right)^2 \\
= \; & E\left[\mathrm{Var}\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)}\middle| \mathbf{X}_i\right)\right] + E\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi})\right)^2
\end{aligned}$$

Considering equation (1.4.5) and the first term in the above equation is constant w.r.t $\boldsymbol{\pi}$, we get

$$\boldsymbol{\pi}_M^* = \arg\min_{\boldsymbol{\pi}} E\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi})\right)^2$$

So,

$$E\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)\right)^2 \leq E\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi})\right)^2 \leq CM^{-2s}$$

Hence,

$$\left|r(\mathbf{x}) - \exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M)\right| = O(CM^{-s})$$

PROOF OF LEMMA 2:

this proof drives heavily from proof of Lemma 2 in Hirano et.al.(2003) In the sequel we write M for $M(N)$.By definition of $\mathbf{G}^M(\mathbf{x})$,

$$\widehat{S}_M = \frac{1}{N}\sum_{i=1}^{N} \mathbf{G}^M(\mathbf{X}_i)\mathbf{G}^M(\mathbf{X}_i)'$$

has expecatation equal to I_M.By Newey(1997), it satisfies

$$\left\|\widehat{S}_M - I_M\right\| = O_p\left(\zeta(M)\sqrt{\frac{M}{N}}\right),$$

which converges to zero in probability by condition (5). Hence the probability that the smallest eigenvalue of $\widehat{S}_M$ is larger than 1/2 goes to one. Let

$$L_N(\boldsymbol{\pi}) = -\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi})\right)^2$$

Next, we will show that

$$\frac{1}{N}\frac{\partial L_N}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi}^*_M) = O_p\left(\sqrt{\frac{M}{N}}\right), \tag{A.2}$$

Consider

$$
\begin{aligned}
& E\left\|\frac{1}{N}\frac{\partial L_N}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi}_M^*)\right\|^2 \\
= \; & \frac{1}{N} trE\left[\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})}-\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\right)^2\exp(2\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i))\right] \\
= \; & \frac{1}{N} trE\left\{\left[\mathrm{Var}\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)}\middle|\mathbf{X}_i\right)+\left(r(\mathbf{X}_i)-\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\right)^2+o_p(1)\right]\right. \\
& \left.\exp(2\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\right\} \\
\leq \; & \frac{C}{N} trE\left[\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\right] \\
\leq \; & \frac{CM}{N}
\end{aligned}
$$

and Markov inequality implies (A.2). Next, let

$$
\eta = \inf_{\mathbf{x}\in\Xi, M}\left(2\exp(\mathbf{G}^M(\mathbf{x}\right)\boldsymbol{\pi}_M^*)-r(\mathbf{x}))\exp(2\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M^*))
$$

,which by assumptions and Lemma 1 is positive. For any $\varepsilon > 0$, choose C such that for N large enough

$$
P\left(\left\|\frac{1}{N}\frac{\partial L_N}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi}_M^*)\right\| < \eta C\sqrt{\frac{M}{N}}\right) \geq 1-\frac{\varepsilon}{2} \tag{A.3}
$$

Note that,

$$
\begin{aligned}
& \sup_{\mathbf{x}\in\Xi, |\boldsymbol{\pi}-\boldsymbol{\pi}^*|<\eta C\sqrt{\frac{M}{N}}}\left|\exp(2\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi})-\exp(2\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M^*)\right| \\
\leq \; & \sup_{\mathbf{x}\in\Xi, |\boldsymbol{\pi}-\boldsymbol{\pi}^*|<\eta C\sqrt{\frac{M}{N}}} |CG^M(\mathbf{x})(\boldsymbol{\pi}-\boldsymbol{\pi}^*)| \\
\leq \; & \zeta(M)C\sqrt{\frac{M}{N}}
\end{aligned}
$$

which goes to zero, so that for large enough N

$$\inf_{\mathbf{x}\in\Xi, \|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|<\eta C\sqrt{\frac{M}{N}}} \Big(2\exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}) - r(\mathbf{x}))\exp(2\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi})\Big) \geq 4\eta$$

Choose N large enough so that this inequality holds. that (A.3) holds with probability at least $1-\varepsilon/2$, and the the probability that the smallest eigenvalue of $\widehat{S}_M$ is larger than $1/2$ is at least $1-\varepsilon/2$. Then the probability that both of these hold is at least $1-\varepsilon$, then for every $\boldsymbol{\pi}$ with $\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$, a second order expansion gives

$$\frac{1}{N}L_N(\boldsymbol{\pi}) = \frac{1}{N}L_N(\boldsymbol{\pi}^*_M) + \frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)(\boldsymbol{\pi}-\boldsymbol{\pi}^*) + \frac{1}{2N}(\boldsymbol{\pi}-\boldsymbol{\pi}^*)'\frac{\partial^2 L_N}{\partial\boldsymbol{\pi}\partial\boldsymbol{\pi}'}(\bar{\boldsymbol{\pi}})(\boldsymbol{\pi}-\boldsymbol{\pi}^*), \quad \text{(A.4)}$$

where $\|\bar{\boldsymbol{\pi}}-\boldsymbol{\pi}^*\| \leq \|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$. We have

$$\begin{aligned}
&\frac{1}{2N}\frac{\partial^2 L_N}{\partial\boldsymbol{\pi}\partial\boldsymbol{\pi}'}(\bar{\boldsymbol{\pi}}) \\
= \;& -\frac{1}{2N}\sum_{i=1}^{N}\left(2\exp(\mathbf{G}^M(\mathbf{X}_i)\bar{\boldsymbol{\pi}}) - \frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})}\right)\exp(\mathbf{G}^M(\mathbf{X}_i)\bar{\boldsymbol{\pi}})\mathbf{G}^M(\mathbf{X}_i))'\mathbf{G}^M(\mathbf{X}_i)) \\
= \;& -\frac{1}{2N}E\left[\Big(2\exp(\mathbf{G}^M(\mathbf{X}_i)\bar{\boldsymbol{\pi}}) - r(\mathbf{X}_i)\Big)\exp(\mathbf{G}^M(\mathbf{X}_i)\bar{\boldsymbol{\pi}})\mathbf{G}^M(\mathbf{X}_i))'\mathbf{G}^M(\mathbf{X}_i))\right] + o_p(1) \\
\leq \;& -2\eta\widehat{S}_M + o_p(1)
\end{aligned}$$

with its eigenvalues bounded away from zero in absolute value by η. Then, rearranging (A.4) and using the triangle inequality, with probability greater than $1-\varepsilon$, for$\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$,

$$\begin{aligned}
\frac{1}{N}L_N(\boldsymbol{\pi}) - \frac{1}{N}L_N(\boldsymbol{\pi}^*_M) &\leq \frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)(\boldsymbol{\pi}-\boldsymbol{\pi}^*) - \eta\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|^2 + o_p(1) \\
&\leq \left\|\frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)\right\| \|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| - \eta\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|^2 + o_p(1) \\
&= \left(\left\|\frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)\right\| - \eta\sqrt{\frac{M}{N}}\right)\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| + o_p(1) \\
&< 0
\end{aligned}$$

That is, we have with probability greater than $1-\varepsilon$, $\frac{1}{N}L_N(\boldsymbol{\pi}) < \frac{1}{N}L_N(\boldsymbol{\pi}^*_M)$ for all $\boldsymbol{\pi}$ with $\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$ Since $L_N(\boldsymbol{\pi})$ is continuous, it has a maximum on the compact set $\boldsymbol{\pi} : \|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| \leq \sqrt{\frac{M}{N}}$. By the last inequality, this maximum must occur for some $\widehat{\boldsymbol{\pi}}_M$ with $\|\widehat{\boldsymbol{\pi}}_M-\boldsymbol{\pi}^*\| < \sqrt{\frac{M}{N}}$. Hence the first order conditions are satisfied at $\widehat{\boldsymbol{\pi}}_M$ and by concavity of $L_N(\boldsymbol{\pi})$, $\widehat{\boldsymbol{\pi}}_M$ maximize $L_N(\boldsymbol{\pi})$ over all of $\mathbf{G}^M$. Because the probability of this is greater than $1-\varepsilon$ with ε arbitrary, we conclude that $\widehat{\boldsymbol{\pi}}_M$ exists and satisfies the first order conditions with probability approaching one, and that $\|\widehat{\boldsymbol{\pi}}_{M(N)}-\boldsymbol{\pi}^*_{M(N)}\| = O_p\left(\sqrt{\frac{M(N)}{N}}\right)$.

PROOF OF LEMMA 3:

$$\begin{aligned}
& \sqrt{N}V_M^{-1/2}\left(\widehat{r(\mathbf{x})}-r(\mathbf{x})\right) \\
= & \sqrt{N}V_M^{-1/2}\left(\widehat{r(\mathbf{x})}-\exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M)\right)+\sqrt{N}V_M^{-1/2}\left(\exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M)-r(\mathbf{x})\right) \\
\equiv & T1+T2, \qquad \text{(A.5)}
\end{aligned}$$

By mean value theorem:

$$\begin{aligned}
T1 &= V_M^{-1/2}\sqrt{N}\left(\exp(\mathbf{G}^M(\mathbf{x})\widehat{\boldsymbol{\pi}}_M)-\exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M)\right) \\
&= V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\check{\boldsymbol{\pi}}_M)\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_M-\boldsymbol{\pi}^*_M\right), \qquad \text{(A.6)}
\end{aligned}$$

From equation (1.4.4), we know:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})}-\exp(\mathbf{G}^M(\mathbf{X}_i)\widehat{\boldsymbol{\pi}}_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\widehat{\boldsymbol{\pi}}_M)=\mathbf{0}$$

Taylor expansion around $\boldsymbol{\pi}^*_M$ for the left hand side of the above equation:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M) +$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - 2\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\left(\widehat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}^*_M\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \frac{Y_i}{\exp(\mathbf{X}_i\beta)}\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M) +$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M) +$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M) +$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\left(\widehat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}^*_M\right)$$

$$-\frac{1}{N}\sum_{i=1}^{N}\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\left(\widehat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}^*_M\right)$$

$$= 0$$

So

$$\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}^*_M\right) =$$

$$\Big\{\frac{1}{N}\sum_{i=1}^{N}\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M) -$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\}^{-1}$$

$$\Big\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \frac{Y_i}{\exp(\mathbf{X}_i\beta)}\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M) +$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M) +$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}^*_M)\Big\}$$

$$\equiv \{\mathbf{E} - \mathbf{F}\}^{-1}\{\mathbf{G} + \mathbf{H} + \mathbf{J}\}$$

So from equation (A.6),

$$
\begin{aligned}
T1 &= V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M)\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}_M^*\right) \\
&= \mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\breve{\boldsymbol{\pi}}_M)\{\mathbf{E}-\mathbf{F}\}^{-1}\{\mathbf{G}+\mathbf{H}+\mathbf{J}\},
\end{aligned}
$$

Let

$$
\begin{aligned}
\boldsymbol{\Sigma}_M &\equiv E[\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i)\right)^2\exp(2\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)], \\
\mathbf{Q}_M &\equiv E[\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(2\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)], \\
V_M(\mathbf{x}) &\equiv \mathbf{G}^M(\mathbf{x})\mathbf{Q}_M^{-1}\boldsymbol{\Sigma}_M\mathbf{Q}_M^{-1}(\mathbf{x})\mathbf{G}^M(\mathbf{x})'\exp(2\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}_M^*).
\end{aligned}
$$

so

$$
T1 = V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\breve{\boldsymbol{\pi}}_M)\{\mathbf{E}-\mathbf{F}\}^{-1}\{\mathbf{G}+\mathbf{H}+\mathbf{J}\}
$$

Note that

$$
\begin{aligned}
\|\mathbf{E}\| &= \left\|\frac{1}{N}\sum_{i=1}^N \mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(2\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right\| \leq O_p(M^2) \\
\|\mathbf{F}\| &\leq \left\|\frac{1}{N}\sum_{i=1}^N \left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right)\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right\| \\
&\quad \left\|\sum_{i=1}^N \mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\right\| \leq O_p(M^{2-s}) \\
\|\mathbf{G}\| &\leq \left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \left(\frac{Y_i}{\exp(\mathbf{X}_i\widehat{\beta})} - \frac{Y_i}{\exp(\mathbf{X}_i\beta)}\right)\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\right\| \left\|\sum_{i=1}^N \mathbf{G}^M(\mathbf{X}_i)'\right\| \leq O_p(M) \\
\|\mathbf{J}\| &\leq \left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \Big(r(\mathbf{X}_i) - \exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\Big)\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\right\| \left\|\sum_{i=1}^N \mathbf{G}^M(\mathbf{X}_i)'\right\| \\
&\leq O_p(N^{1/2}M^{1-s})
\end{aligned}
$$

Since $s > 2$, so $\|\mathbf{F}\| \xrightarrow{p} 0$ as $M \longrightarrow \infty$. Hence $\{\mathbf{E} - \mathbf{F}\}^{-1}$ is equivalent to $\mathbf{E}^{-1}$ as $M \longrightarrow \infty$; while

$$\begin{aligned}
\|\mathbf{E}^{-1}\mathbf{G}\| &\leq \|\mathbf{E}^{-1}\|\|\mathbf{G}\| \leq O_p(M^{-2})O_p(M) = O_p(M^{-1}) \\
\|\mathbf{E}^{-1}\mathbf{J}\| &\leq \|\mathbf{E}^{-1}\|\|\mathbf{J}\| \leq O_p(M^{-2})O_p(N^{1/2}M^{-s}) = O_p(N^{1/2}M^{-(s+2)})
\end{aligned}$$

Here, we assume $N^{1/2}M^{-(s+1)} \longrightarrow 0$ as $N \longrightarrow \infty$; so

$$\begin{aligned}
& T1 \\
= \; & V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M)\{\mathbf{E}\}^{-1}\{\mathbf{H}\} + o_p(1) \\
= \; & V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M) \\
& \left\{\frac{1}{N}\sum_{i=1}^{N}\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\mathbf{G}^M(\mathbf{X}_i)'\mathbf{G}^M(\mathbf{X}_i)\exp(\mathbf{G}^M(\mathbf{X}_i)\widetilde{\boldsymbol{\pi}}_M)\right\}^{-1} \\
& \left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\right\} + o_p(1) \\
= \; & V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M)\mathbf{Q}^{-1} \times \\
& \left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i)\right)\mathbf{G}^M(\mathbf{X}_i)'\exp(\mathbf{G}^M(\mathbf{X}_i)\boldsymbol{\pi}_M^*)\right\} + o_p(1)
\end{aligned}$$

Next, Let

$$\begin{aligned}
\boldsymbol{\epsilon} &= [\epsilon_1, \cdots, \epsilon_N]', \\
\epsilon_i &= \frac{Y_i}{\exp(\mathbf{X}_i\beta)} - r(\mathbf{X}_i), \\
\mathbf{O}(\widetilde{\boldsymbol{\pi}}_M) &= \frac{1}{\sqrt{N}}[\exp(\mathbf{G}^M(\mathbf{X}_1)\widetilde{\boldsymbol{\pi}}_M)\mathbf{G}^M(\mathbf{X}_1)', \ldots, \exp(\mathbf{G}^M(\mathbf{X}_N)\widetilde{\boldsymbol{\pi}}_M)\mathbf{G}^M(\mathbf{X}_N)'], \\
Z_{iN} &= V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M)\{\mathbf{O}(\widetilde{\boldsymbol{\pi}}_M)\mathbf{O}(\widetilde{\boldsymbol{\pi}}_M)'\}^{-1}\mathbf{O}(\boldsymbol{\pi}_M^*)_i\epsilon_i/\sqrt{N}
\end{aligned}$$

so that

$$\sum_{i=1}^{N} Z_{iN} = V_M^{-1/2}\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M)\{\mathbf{O}(\widetilde{\boldsymbol{\pi}}_M)\mathbf{O}(\widetilde{\boldsymbol{\pi}}_M)'\}^{-1}\mathbf{O}(\boldsymbol{\pi}^*_M)\boldsymbol{\epsilon}/\sqrt{N}$$

Note that for each N, $Z_{iN}(i = 1, \cdots, N)$ is i.i.d. Also, $E[Z_{iN}] = 0,\sum_{i=1}^{N} E[Z_{iN}^2] = 1$; and $\forall \varepsilon > 0$

$$\begin{aligned} & NE[1(|Z_{iN}| > \varepsilon)Z_{iN}^2] = N\varepsilon^2 E[1(|Z_{iN}/\varepsilon| > 1)(Z_{iN}/\varepsilon)^2] \\ \leq \; & N\varepsilon^2 E[(Z_{iN}/\varepsilon)^4] \\ \leq \; & N\varepsilon^2 \|V_M^{-2}\| \|\mathbf{G}^M(\mathbf{x})\exp(\mathbf{G}^M(\mathbf{x})\bar{\boldsymbol{\pi}}_M)\|^2 E[\|\mathbf{O}(\boldsymbol{\pi}^*_M)\|^2 E[\epsilon_i^4|\mathbf{x}_i]]/(N^2\varepsilon^4) \\ \leq \; & C\zeta_0(M)^2 M/N \longrightarrow 0 \end{aligned}$$

Then by Lindbergh-Feller central limit theorem,$\sum_{i=1}^{N} Z_{iN} \overset{d}{\to} N(0,1)$,i.e.

$$T1 \overset{d}{\to} N(0,1)$$

As for the second term, T_2, in equation(A.5)

$$\begin{aligned} \left|\sqrt{N}V_M^{-1/2}\left(\exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M) - r(\mathbf{x})\right)\right| \; & \leq \; \left|\sqrt{N}V_M^{-1/2}\right| \left|\left(\exp(\mathbf{G}^M(\mathbf{x})\boldsymbol{\pi}^*_M) - r(\mathbf{x})\right)\right| \\ & \leq \; O(\sqrt{N}M^{-(2+s)}) \longrightarrow 0 \end{aligned}$$

So,

$$\sqrt{N}V_M^{-1/2}\left(\widehat{r(\mathbf{x})} - r(\mathbf{x})\right) \overset{d}{\to} N(0,1)$$

QED.

PROOF OF THEOREM 3:

Note that

$$\begin{aligned}
& \widehat{CAPE_j}(\mathbf{x}) - CAPE_j(\mathbf{x}) \\
= \; & \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) \exp(\mathbf{G}^M(\mathbf{x})\widehat{\boldsymbol{\pi}}_M) - \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) r(\mathbf{x}) \\
= \; & \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) \exp(\mathbf{G}^M(\mathbf{x})\widehat{\boldsymbol{\pi}}_M) - \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) r(\mathbf{x}) \\
& + \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) r(\mathbf{x}) - \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) r(\mathbf{x}) \\
= \; & \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) \left(\widehat{r(\mathbf{x})} - r(\mathbf{x})\right) + r(\mathbf{x}) \left(\widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) - \beta_j \exp(\mathbf{x}\boldsymbol{\beta})\right)
\end{aligned}$$

From result in theorem 1, we know $\sqrt{N}(\widehat{\beta}_j - \beta_j) = O_p(1)$; so by delta method

$$\sqrt{N} \left(\widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) - \beta_j \exp(\mathbf{x}\boldsymbol{\beta})\right) = O_p(1);$$

so

$$\left\| \sqrt{N} V_M^{-1/2} r(\mathbf{x}) \left(\widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) - \beta_j \exp(\mathbf{x}\boldsymbol{\beta})\right) \right\| \leq C \left\| V_M^{-1/2} \right\| \leq C M^{-1/2} \longrightarrow 0$$

While From Lemma 3 and Slutsky theorem

$$\sqrt{N} V_M^{-1/2} \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) \left(\widehat{r(\mathbf{x})} - r(\mathbf{x})\right) \xrightarrow{d} \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) N(0,1)$$

Hence

$$\sqrt{N} V_M^{-1/2} \left(\widehat{CAPE_j}(\mathbf{x}) - CAPE_j(\mathbf{x})\right) \xrightarrow{d} \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) N(0,1)$$

QED.

PROOF OF COROLLARY: Note that,

$$\begin{aligned}
& \sqrt{N} \left(\widehat{CAPE_j}(\mathbf{x}) - CAPE_j(\mathbf{x})\right) \\
= \; & \sqrt{N} \widehat{\beta}_j \exp(\mathbf{x}\widehat{\boldsymbol{\beta}}) \widehat{r(\mathbf{x})}
\end{aligned}$$

From theorem 1, we know that $\sqrt{N}\widehat{\beta}_j = \sqrt{N}(\widehat{\beta}_j - \beta_j) \xrightarrow{d} N(0, \ell\boldsymbol{\Omega}\ell')$, where $\ell = (0, \cdots, 1, 0, \cdots, 0)$. So:

$$\sqrt{N} \left(\widehat{CAPE_j}(\mathbf{x}) - CAPE_j(\mathbf{x})\right) \xrightarrow{d} N(0, r^2(\mathbf{x}) \exp(2\mathbf{x}\boldsymbol{\beta}) \ell\boldsymbol{\Omega}\ell')$$

QED.

Appendix B

Tables and Figures

Table B.1: Estimation results: xtreg

Variable	**Coefficient**	**(Std. Err.)**
lfare	-1.163	(0.023)
concen	0.145	(0.040)
y98	0.045	(0.006)
y99	0.104	(0.006)
y00	0.197	(0.006)
Intercept	11.769	(0.116)

Table B.2: Simulation results where V_{it} has Gamma distribution

	$a = 0$			$a = .01$		
	$b = .01$	$b = .05$	$b = .1$	$b = .01$	$b = .05$	$b = .1$
$\hat{\beta}_{lfe}$	.0934	.0675	.0350	.0939	.0666	.0315
	(.0289)*	(.0283)	(.0284)	(.0290)	(.0230)	(.0294)
$\hat{\beta}_{pqml}$	.1001	.0988	.0967	.1011	.1002	.0994
	(.0415)	(.0404)	(.0421)	(.0410)	(.0411)	(.0439)
$se(\hat{\beta}_{lfe})$	.0286	.0287	.0291	.0292	.0294	.0296
$se(\hat{\beta}_{pqml})$	.0388	.0388	.0397	.0390	.0393	.0405
$\rho_{\mathbf{X},V} **$	.0008	-.0004	-.0003	.0000	-.0004	-.0014
$\rho_{\mathbf{X},lv} **$	-.0057	-.0250	-.0450	-.0046	-.0263	-.0523
$mean(lv)$	-.5781	-.5767	-.5800	-.5836	-.5855	-.5884
$sd(lv)$	1.2816	1.2839	1.2914	1.2919	1.2945	1.3021

* Monte Carlo Standard Deviations in parentheses

** $\rho_{\mathbf{X},V} = \mathrm{Corr}(\mathbf{X}, V)$, $\rho_{\mathbf{X},lv} = \mathrm{Corr}(\mathbf{X}, lv)$,$lv = \log(V)$

Table B.3: Simulation results where V_{it} has Gamma distribution(Continued)

	$a = .05$			$a = .1$		
	$b = .01$	$b = .05$	$b = .1$	$b = .01$	$b = .05$	$b = .1$
$\hat{\beta}_{lfe}$	.0921	.0630	.0222	.0922	.0495	.0032
	(.0327)	(.0319)	(.0318)	(.0365)	(.0376)	(.0373)
$\hat{\beta}_{pqml}$	.0997	.1023	.0998	.0996	.0972	.1006
	(.0449)	(.0454)	(.0454)	(.0484)	(.0471)	(.0516)
$se(\hat{\beta}_{lfe})$	.0320	.0321	.0327	.0374	.0380	.0387
$se(\hat{\beta}_{pqml})$	.0414	.0421	.0427	.0448	.0453	.0467
$\rho_{\mathbf{X},V}$	-.0000	.0004	.0004	-.0001	-.0011	.0002
$\rho_{\mathbf{X},lv}$	-.0057	-.0277	-.0575	-.0062	-.0350	-.0672
$mean(lv)$	-.6145	-.6138	.6169	-.6572	-.6592	-.6604
$sd(lv)$	1.3358	1.3398	1.3492	1.4140	1.4197	1.4309

Table B.4: Special case

	a=.1,b=0		
N	500	1000	2000
$\hat{\beta}_{lfe}$	.0997	.1004	.1006
	(.0372)	(.0266)	(.0185)
$\hat{\beta}_{pqml}$	.1017	.1021	.1011
	(.0461)	(.0334)	(.0248)
$se(\hat{\beta}_{lfe})$	.0375	.0266	.0189
$se(\hat{\beta}_{pqml})$	.0445	.0329	.0240
$\rho_{x,v}$	.0001	.0005	.0006
$\rho_{x,lv}$	-.0003	.0009	.0004
$mean(lv)$	-.6574	-.6558	-.6558
$sd(lv)$	1.4120	1.4126	1.4134

Table B.5: Simulation results where V_{it} has log-normal distribution

	a=-.125, b=.5			a=-.5, b=1		
N	500	1000	2000	500	1000	2000
$\hat{\beta}_{lfe}$	0.10001	0.09990	0.09991	0.09827	0.10144	0.10127
	(.01998)	(.01411)	(.01015)	(.04807)	(.03315)	(.02410)
$\hat{\beta}_{pqml}$	0.09837	0.09655	0.09834	0.07994	0.09138	0.09515
	(.04476)	(.03261)	(.02412)	(.13131)	(.11863)	(.09854)
$se(\hat{\beta}_{lfe})$	0.01969	0.01396	0.00987	0.04838	0.03424	0.02418
$se(\hat{\beta}_{pqml})$	0.03758	0.02778	0.02084	0.08555	0.07361	0.05789
$\rho_{\mathbf{X},V}$	-0.0012	-0.0010	-0.0005	-0.0004	0.00025	0.00051
$\rho_{\mathbf{X},lv}$	0.00003	0.0003	-0.0003	-0.0011	0.00088	0.00074
$mean(lv)$	-0.1249	-0.1254	-0.1251	-0.5004	-0.5001	-0.4998
$sd(lv)$	0.52993	0.5305	0.53047	1.2255	1.2254	1.2243

Table B.6: $V_{it} = \exp(a * x_{it}^2 + b * x_{it} * z_{it})$ with $N = 500$

	a=-.5, b=1, N=500					
$\rho*$	-0.95	-0.5	-0.1	0.1	0.5	0.95
$\hat{\beta}_{lfe}$	0.10010	0.09669	0.09978	0.10090	0.09768	0.10261
	(.04456)	(.04563)	(.04795)	(.04927)	(.05510)	(.06163)
$\hat{\beta}_{pqml}$	0.07839	0.08340	0.08476	0.08303	0.08036	0.08497
	(.10740)	(.13876)	(.12536)	(.13958)	(.12433)	(.13743)
$se(\hat{\beta}_{lfe})$	0.04448	0.04558	0.04751	0.04909	0.05334	0.06217
$se(\hat{\beta}_{pqml})$	0.06785	0.08034	0.08202	0.08652	0.08887	0.09637
$\rho_{x,v}$	-0.00018	-0.00247	0.00054	0.00014	-0.00177	0.00190
$\rho_{x,lv}$	-0.00005	-0.00268	-0.00007	0.00087	-0.00163	0.00235
$mean(lv)$	-0.49941	-0.49949	-0.49859	-0.49969	-0.49972	-0.49948
$sd(lv)$	1.22230	1.22259	1.22251	1.22435	1.22324	1.22320

$^*z_i \sim N(I_5, \Sigma)$, $\rho = \text{Corr}(z_{it}, z_{it+1})$

Table B.7: $V_{it} = \exp(a * x_{it}^2 + b * x_{it} * z_{it})$ with $N = 1000$

	a=-.5, b=1, N=1000					
ρ	-0.95	-0.5	-0.1	0.1	0.5	0.95
$\hat{\beta}_{lfe}$	0.09984	0.09959	0.09845	0.10023	0.10137	0.10145
	0.03032	0.03281	0.03553	0.03549	0.03643	0.04594
$\hat{\beta}_{pqml}$	0.08465	0.08509	0.08018	0.08802	0.08665	0.08614
	0.10493	0.09423	0.10190	0.12003	0.11121	0.11349
$se(\hat{\beta}_{lfe})$	0.03152	0.03231	0.03370	0.03469	0.03764	0.04395
$se(\hat{\beta}_{pqml})$	0.06084	0.06522	0.06652	0.07327	0.07415	0.08087
$\rho_{x,v}$	-0.00129	0.00053	-0.00198	-0.00088	0.00101	-0.00066
$\rho_{x,lv}$	-0.00039	-0.00024	-0.00124	0.00013	0.00090	0.00087
$mean(lv)$	-0.50118	-0.49880	-0.49996	-0.49942	-0.49900	-0.50027
$sd(lv)$	1.22425	1.22294	1.22432	1.22375	1.22204	1.22413

Table B.8: Simulation results with $V_{it} = \exp(-.125\overline{\mathbf{X}}_i^2 + .5\overline{\mathbf{X}}_i * z_{it})$

N	2000	1000	500	250	100
$\hat{\beta}_{pqml}$	0.0998	0.1001	0.1001	0.0998	0.1001
	(0.0070)	(0.0095)	(0.0132)	(0.0183)	(0.0278)
$\hat{\beta}_{lfe}$	.10004	.09997	.09989	.09995	.09991
	(0.0025)	(0.0035)	(0.0050)	(0.0071)	(0.0112)
$\hat{\beta}_{gmm}$	0.1001	0.1003	0.1001	0.1002	0.1003
	(0.0042)	(0.0058)	(0.0079)	(0.0107)	(0.0168)
$se(\hat{\beta}_{pqml})$	0.0074	0.0102	0.0142	0.0205	0.0384
$se(\hat{\beta}_{lfe})$	0.0025	0.0035	0.0050	0.0070	0.0110
$se(\hat{\beta}_{gmm})$	0.0044	0.0057	0.0073	0.0094	0.0136

Table B.9: Simulation results for four estimators

N	2000	1000	500	250	100
$\hat{\beta}_{pqml}$	0.0999	0.1002	0.1003	0.0995	0.999
	(0.0068)	(0.0093)	(0.0146)	(0.0179)	(0.0265)
$\hat{\beta}_{lfe}$	.10002	.09989	.09998	.09992	.09995
	(0.0027)	(0.0038)	(0.0052)	(0.0069)	(0.0120)
$\hat{\beta}_{gmm}$	0.9999	0.1002	0.9998	0.1003	0.1004
	(0.0044)	(0.0055)	(0.0080)	(0.0110)	(0.0172)
$\hat{\beta}_{oiv}$	0.1001	0.9999	0.1001	0.9998	0.1006
	(0.0026)	(0.0038)	(0.0066)	(0.0077)	(0.0118)
$se(\hat{\beta}_{pqml})$	0.0077	0.0110	0.0139	0.0207	0.0379
$se(\hat{\beta}_{lfe})$	0.0028	0.0039	0.0052	0.0073	0.0109
$se(\hat{\beta}_{gmm})$	0.0045	0.0060	0.0071	0.0090	0.0140
$se(\hat{\beta}_{oiv})$	0.0024	0.0034	0.0053	0.0069	0.0106

Table B.10: Summary statistics

	Obs	Mean	Std. Dev.	Min	Max
passen	4596	636.8242	812	2	8497
lfare	4596	5.095601	0.4363999	3.610918	6.257668
concen	4596	0.6101149	0.196435	0.1605	1

Table B.11: Dependent variable, *passen*

	$lfare$	$concen$	$y88$	$y89$	$y00$
$\hat{\beta}_{pqml}$	-0.8658	-0.1289	0.0427	0.1093	0.1899
$\hat{\beta}_{lfe}$	-1.1632	0.1455	0.0454	0.1038	0.1970
$\hat{\beta}_{gmm}$	-0.8515	-0.1450	0.0431	0.1081	0.1911
$se(\hat{\beta}_{pqml})$	0.0366	0.0544	0.0037	0.0054	0.0085
$se(\hat{\beta}_{lfe})$	0.1101	0.0890	0.0049	0.0063	0.0101
$se(\hat{\beta}_{gmm})$	0.0336	0.0538	0.0035	0.0049	0.0069

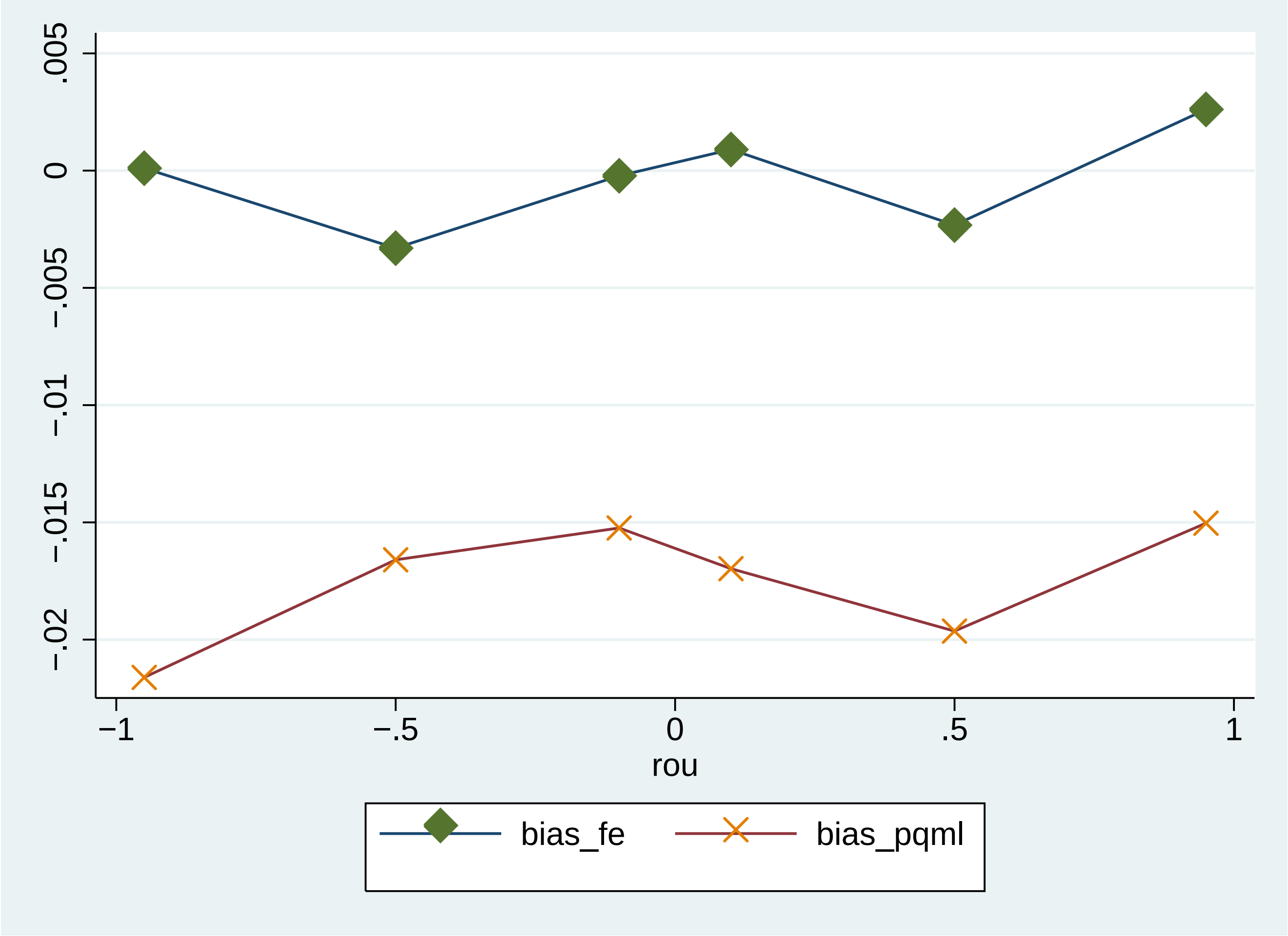

Figure B.1: Bias of LFE and PQML with change of ρ

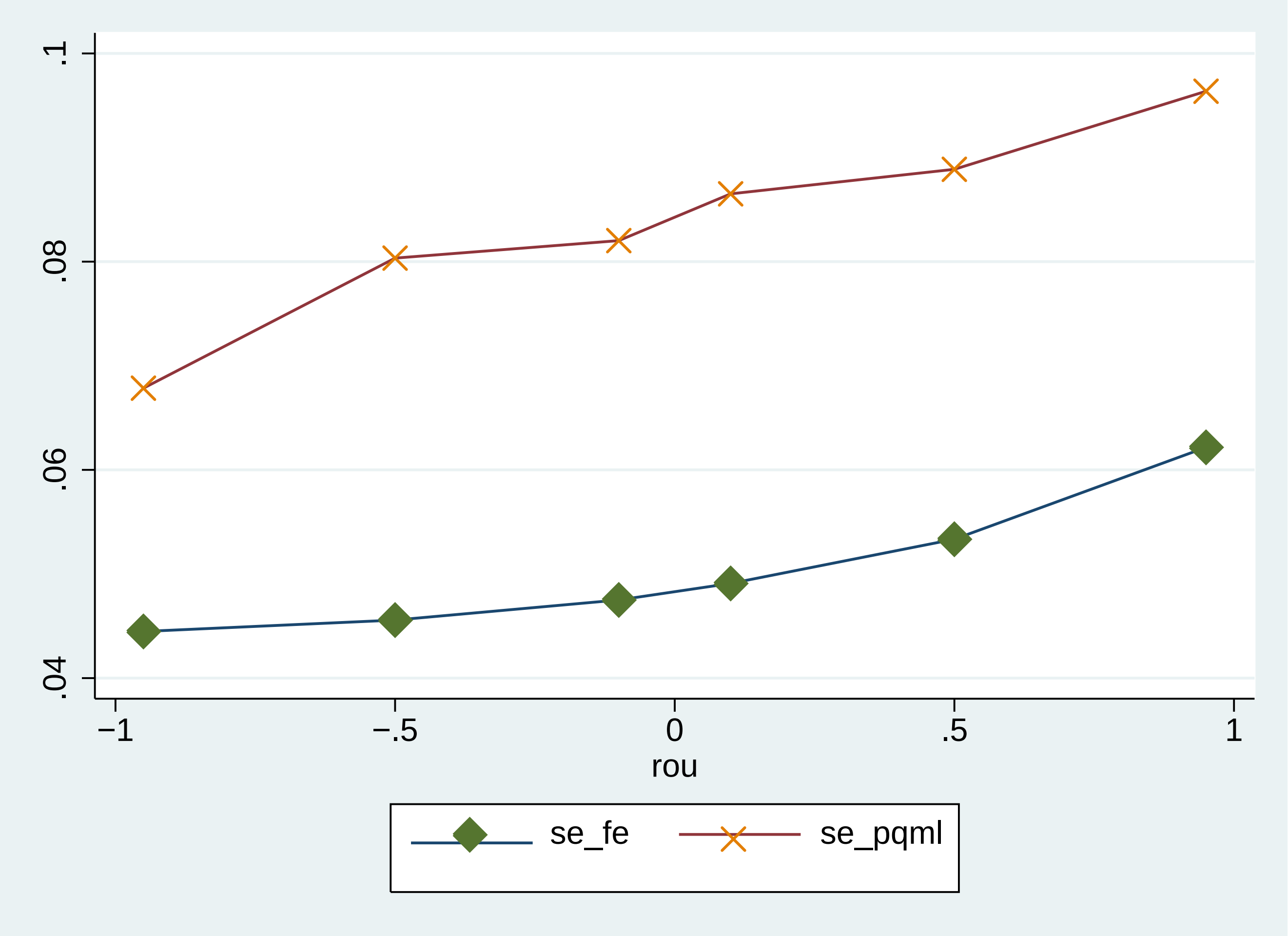

Figure B.2: Std. error of LFE and PQML with change of ρ

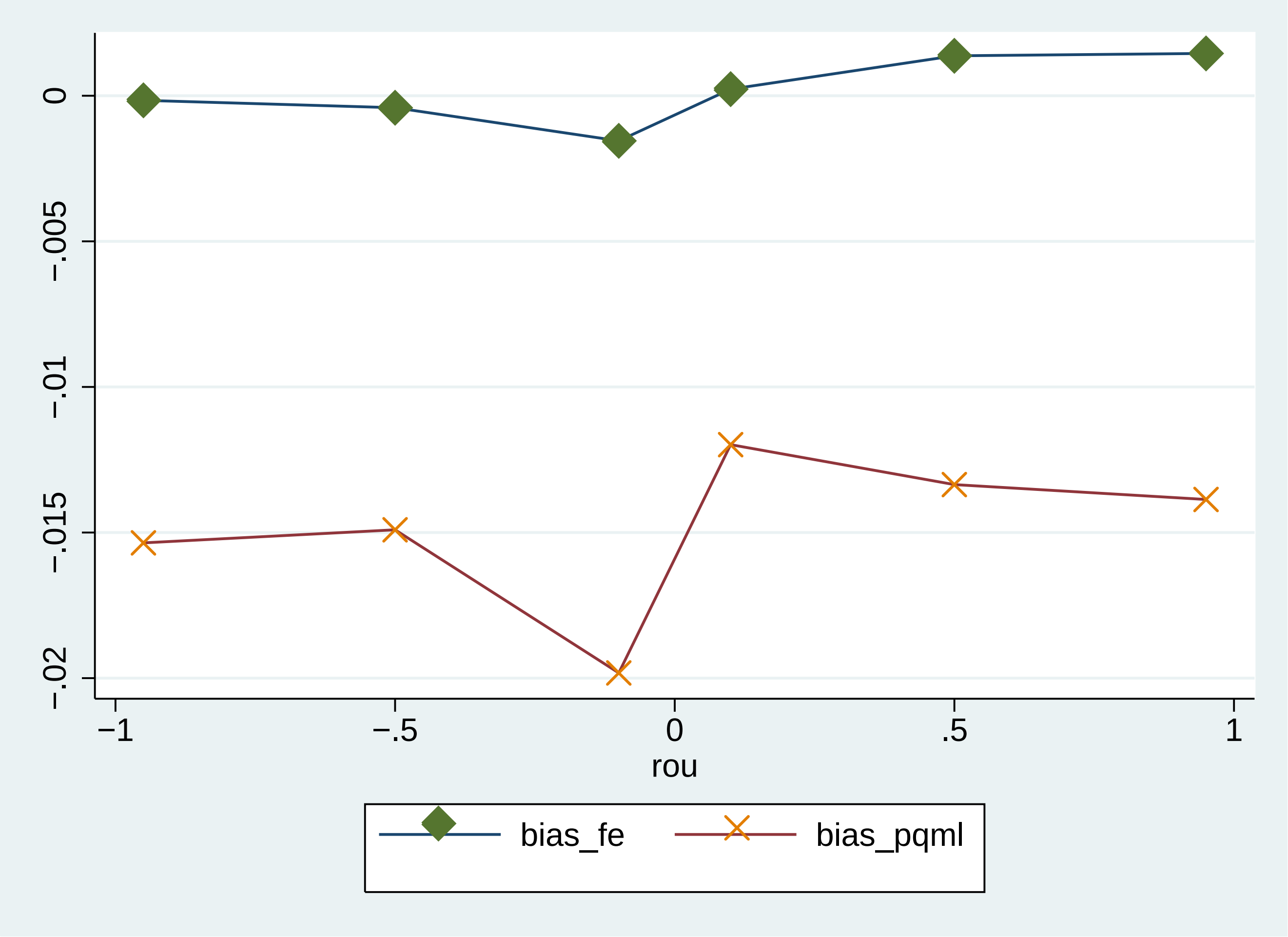

Figure B.3: Bias of LFE and PQML with change of ρ, N=1000

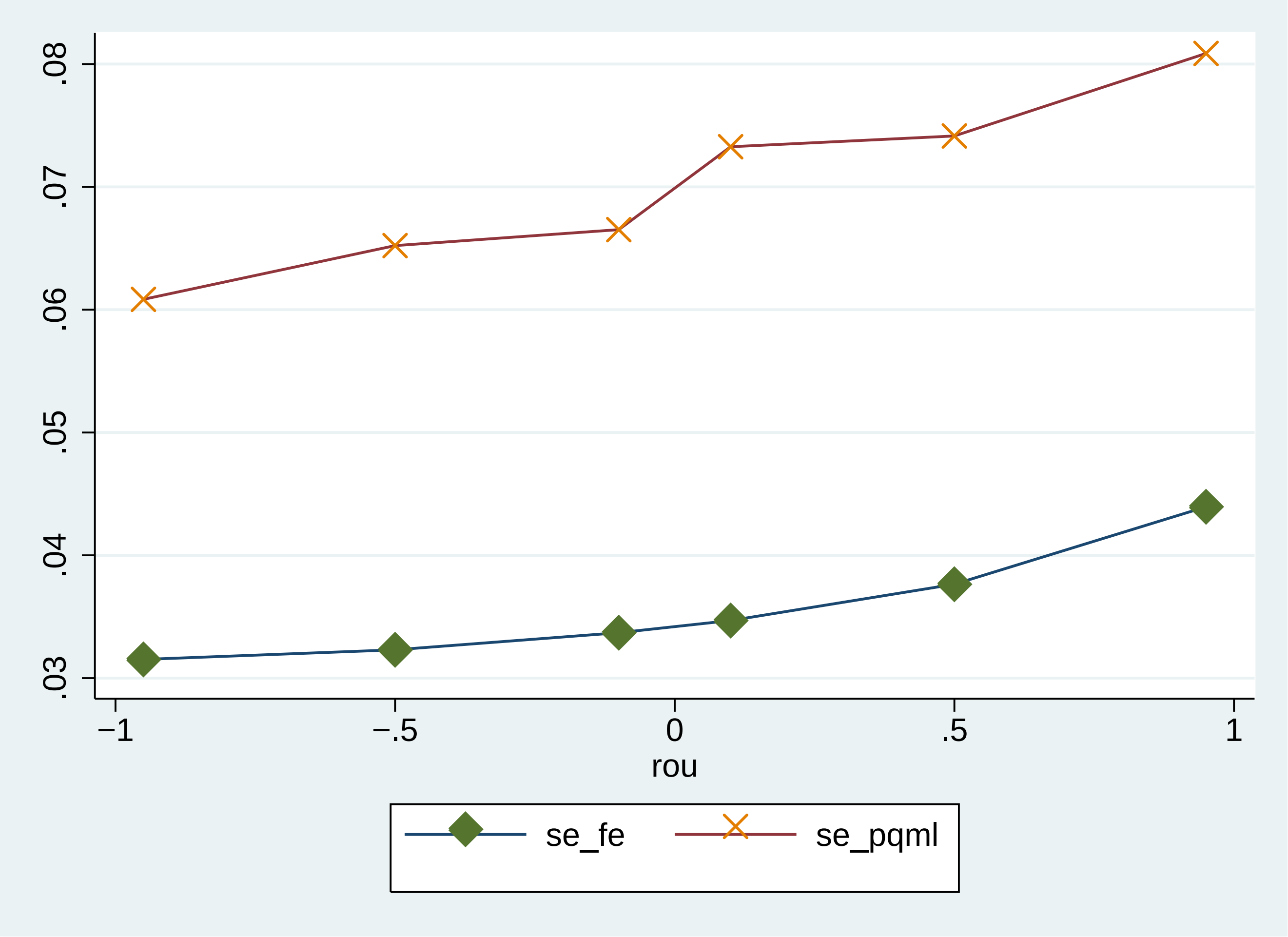

Figure B.4: Std. error of LFE and PQML with change of ρ, N=1000

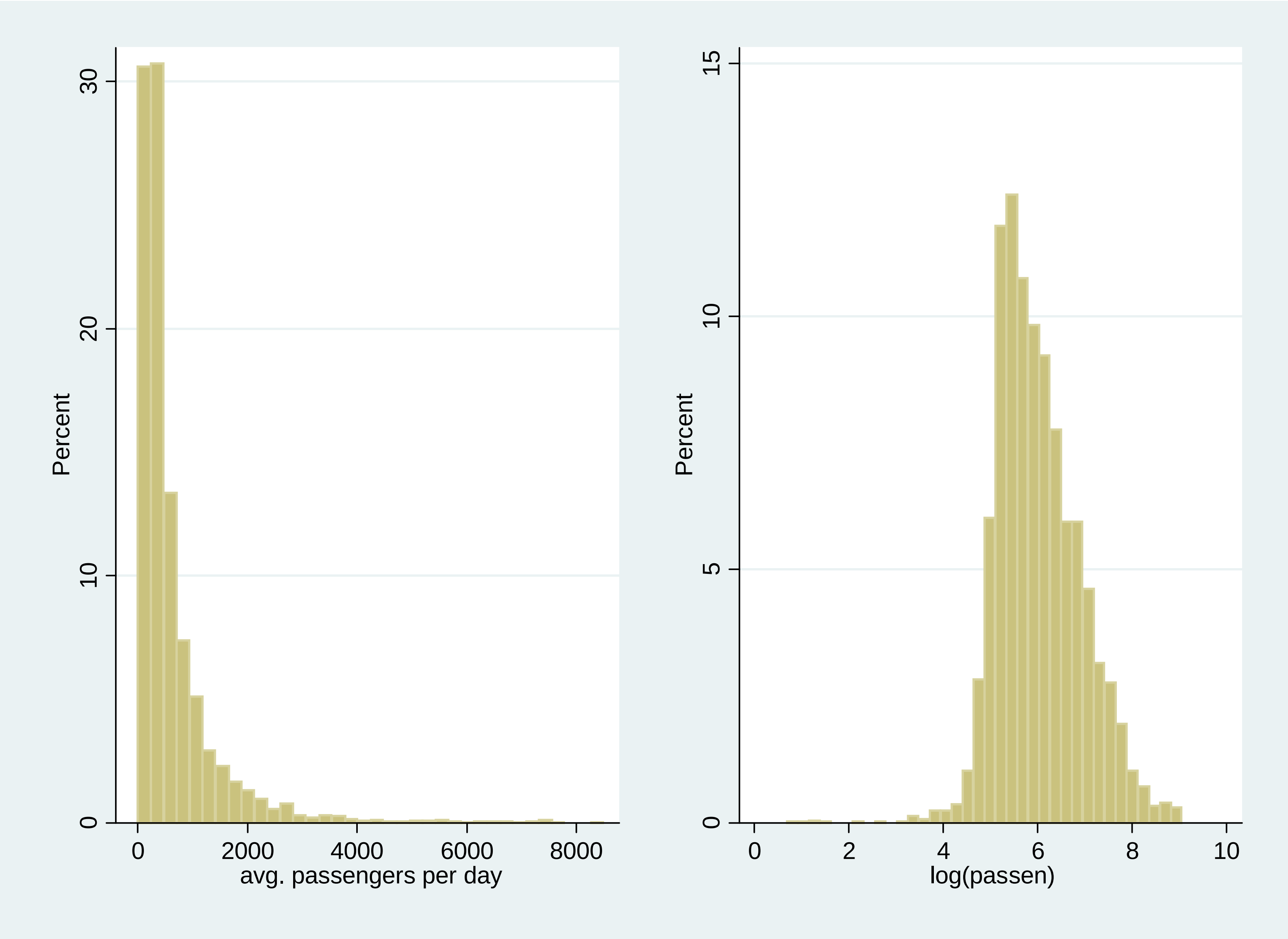

Figure B.5: Histogram of Passengers

Proofs

PROOF OF Theorem 2.4.1

$$\begin{aligned}\sqrt{N}\begin{pmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\eta}-\eta\end{pmatrix} &= \begin{pmatrix}N^{-1/2}\sum_{i=1}^{N}\mathbf{A}^{-1}\mathbf{V}_i\\ N^{-1/2}\sum_{i=1}^{N}(\mathbf{PS}_i+\mathbf{U}_i-\eta)\end{pmatrix}+o_p(1)\\ &\equiv N^{-1/2}\sum_{i=1}^{N}\begin{pmatrix}\mathbf{S}_i\\ Q_i\end{pmatrix}+o_p(1)\end{aligned}$$

Where

$$\mathbf{V}_i = \mathbf{Y}_i - \mathbf{p}(\mathbf{X}_i,\ \boldsymbol{\beta})n_i$$

$$\mathbf{A} = E(n_i\nabla_{\boldsymbol{\beta}}\mathbf{p}(\mathbf{X}_i,\ \boldsymbol{\beta})'\mathbf{W}(\mathbf{X}_i,\ \boldsymbol{\beta})\boldsymbol{\nabla}_{\boldsymbol{\beta}}\mathbf{p}(\mathbf{X}_i,\ \boldsymbol{\beta}))$$

$$\mathbf{p}(\mathbf{X}_i,\ \boldsymbol{\beta}) = \left[\frac{\exp(\mathbf{X}_{i1}\boldsymbol{\beta})}{\sum_{t=1}^{T}\exp(\mathbf{X}_{it}\boldsymbol{\beta})},\ \cdots,\ \frac{\exp(\mathbf{X}_{iT}\boldsymbol{\beta})}{\sum_{t=1}^{T}\exp(\mathbf{X}_{it}\boldsymbol{\beta})}\right]'$$

$$\mathbf{W}(\mathbf{X}_i,\ \boldsymbol{\beta}) = \left[\text{diag}\ \left(\frac{\exp(\mathbf{X}_{i1}\boldsymbol{\beta})}{\sum_{t=1}^{T}\exp(\mathbf{X}_{it}\boldsymbol{\beta})},\ \cdots,\ \frac{\exp(\mathbf{X}_{iT}\boldsymbol{\beta})}{\sum_{t=1}^{T}\exp(\mathbf{X}_{it}\boldsymbol{\beta})}\right)\right]^{-1}$$

$$\mathbf{P} = E(\mathbf{X}_{it}'\mathbf{U}_i),\ \ Q_i = \mathbf{PS}_i + \mathbf{U}_i - \eta$$

PROOF OF LEMMA 2.4.3: From equation (1.4.3), and by monotonicity of logarithmic function,

$$\begin{aligned}&\exp(-CK^{-s}+\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)-\exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)\\ <\ & r(\mathbf{X}))-\exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)\\ <\ & \exp(CK^{-s}+\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)-\exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M),\end{aligned}$$

By the mean value theorem applied to the lower and upper bound:

$$\begin{aligned}&\exp(-CK^{-s} + \mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M) - \exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)\\ = & \ -CK^{-s}\exp(\xi_1), \quad \xi_1 \in [-CK^{-s} + \mathbf{G}^M(\mathbf{X}),\ \mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M]\end{aligned}$$

$$\begin{aligned}&\exp(CK^{-s} + \mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M) - \exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)\\ = & \ CK^{-s}\exp(\xi_2), \quad \xi_2 \in [\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M,\ CK^{-s} + \mathbf{G}^M(\mathbf{X})]\end{aligned}$$

So, for the $\boldsymbol{\pi}_M$ that satisfies equation (1.4.3), we have:

$$\sup_{\mathbf{X}\in\Xi} |r(\mathbf{X}) - \exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M)| < CK^{-s},$$

So,

$$E\left(r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}_M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}_M)\right)^2 \le CK^{-2s}$$

Note that:

$$\begin{aligned}&E\sum_{t=1}^{T}\left(\frac{Y_{it}}{\exp(\mathbf{x}_{it}\beta)} - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2\\ = & \sum_{t=1}^{T}\mathrm{Var}\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\beta)}\middle|\mathbf{X}_i\right) + TE\left(r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2\end{aligned}$$

So,

$$\boldsymbol{\pi}_M = \arg\min_{\boldsymbol{\pi}} E\left(r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2$$

So,

$$E\left(r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}_M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M)\right)^2 \le E\left(r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}_M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2 \le CK^{-2s}$$

So, similarly,

$$\left|r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}_M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M\right| = O_p(CK^{-s})$$

PROOF OF LEMMA 2.4.4: this proof drives heavily from proof of Lemma 2 in Hirano et.al.(2003) In the sequel we write M for $M(N)$. By definition of $\mathbf{G}^M(\mathbf{X})$,

$$\hat{S}_M = \frac{1}{N}\sum_{i=1}^{N}\mathbf{G}^M(\overline{\mathbf{X}}_i)\mathbf{G}^M(\overline{\mathbf{X}}_i)'$$

has expecatation equal to I_M. By Newey (1997), it satisfies

$$\left\|\hat{S}_M - I_M\right\| = O_p\left(\zeta(M)\sqrt{\frac{M}{N}}\right),$$

which converges to zero in probability by condition (iv). Hence the probability that the smallest eigenvalue of $\hat{S}_M$ is larger than 1/2 goes to one. Let

$$L_N(\boldsymbol{\pi}) = -\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\beta})} - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi})\right)^2$$

Next, we will show that

$$\frac{1}{N}\frac{\partial L_N}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi}^*_M) = O_p\left(\sqrt{\frac{M}{N}}\right), \tag{B.1}$$

Consider

$$\begin{aligned}
& E\left\|\frac{1}{N}\frac{\partial L_N}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)\right\|^2 \\
= \; & \frac{1}{N}\text{tr } E\left[\sum_{t=1}^{T}\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\beta})} - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M)\right)^2 \exp(2\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M)\mathbf{G}^M(\overline{\mathbf{X}}_i)'\mathbf{G}^M(\overline{\mathbf{X}}_i)\right] \\
= \; & \frac{T}{N}\text{tr } E\left\{\left[\text{Var}\left(\frac{Y_{it}}{\exp(\mathbf{X}_{it}\beta)}|\mathbf{X}_i\right) + \left(r(\overline{\mathbf{X}}_i) - \exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M)\right)^2 + o_p(1)\right]\right. \\
& \left.\exp(2\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M)\mathbf{G}^M(\overline{\mathbf{X}}_i)'\mathbf{G}^M(\overline{\mathbf{X}}_i)\right\} \\
\leq \; & \frac{C}{N}\text{tr } E\left[\mathbf{G}^M(\overline{\mathbf{X}}_i)'\mathbf{G}^M(\overline{\mathbf{X}}_i)\right] \\
\leq \; & \frac{CK}{N}
\end{aligned}$$

and Markov inequality implies (B.1). Next, let

$$\eta = \inf_{\mathbf{X}\in\Xi, M} \left(2\exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M^*) - r^*(\mathbf{X}))\exp(2\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M^*)\right)$$

,which by assumptions and Lemma 1 is positive. For any $\varepsilon > 0$, choose C such that for N large enough

$$P\left(\left\|\frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}_M^*)\right\| < \eta C\sqrt{\frac{M}{N}}\right) \geq 1 - \frac{\varepsilon}{2} \quad \text{(B.2)}$$

Note that,

$$\begin{aligned} & \sup_{\mathbf{X}\in\Xi, |\boldsymbol{\pi}-\boldsymbol{\pi}^*|<\eta C\sqrt{\frac{M}{N}}} \left|\exp(2\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}) - \exp(2\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}_M^*)\right| \\ \leq & \sup_{\mathbf{X}\in\Xi, |\boldsymbol{\pi}-\boldsymbol{\pi}^*|<\eta C\sqrt{\frac{M}{N}}} |CK(\mathbf{X})(\boldsymbol{\pi}-\boldsymbol{\pi}^*)| \\ \leq & \zeta(M)C\sqrt{\frac{M}{N}} \end{aligned}$$

which goes to zero,so that for large enough N

$$\inf_{\mathbf{X}\in\Xi, \|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|<\eta C\sqrt{\frac{M}{N}}} \left(2\exp(\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi}) - r^*(\mathbf{X}))\exp(2\mathbf{G}^M(\mathbf{X})\boldsymbol{\pi})\right) \geq 4\eta$$

Choose N large enough so that this inequality holds. that (B.2) holds with probability at least $1-\varepsilon/2$, and the the probability that the smallest eigenvalue of $\hat{S}_M$ is larger than $1/2$ is at least $1-\varepsilon/2$. Then the probability that both of these hold is at least $1-\varepsilon$, then for every $\boldsymbol{\pi}$ with $\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$, a second order expansion gives

$$\frac{1}{N}L_N(\boldsymbol{\pi}) = \frac{1}{N}L_N(\boldsymbol{\pi}_M^*) + \frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}_M^*)(\boldsymbol{\pi}-\boldsymbol{\pi}^*) + \frac{1}{2N}(\boldsymbol{\pi}-\boldsymbol{\pi}^*)'\frac{\partial^2 L_N}{\partial\boldsymbol{\pi}\partial\boldsymbol{\pi}'}(\overline{\boldsymbol{\pi}})(\boldsymbol{\pi}-\boldsymbol{\pi}^*) \quad \text{(B.3)}$$

where $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\| \leq \|\boldsymbol{\pi} - \boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$. We have

$$\begin{aligned}
&\frac{1}{2N}\frac{\partial^2 L_N}{\partial\boldsymbol{\pi}\partial\boldsymbol{\pi}'}(\overline{\boldsymbol{\pi}}) \\
= & -\frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(2\exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\overline{\boldsymbol{\pi}}) - \frac{Y_{it}}{\exp(\mathbf{X}_{it}\hat{\beta})}\right)\exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\overline{\boldsymbol{\pi}})\mathbf{G}^M(\overline{\mathbf{X}}_i))'\mathbf{G}^M(\overline{\mathbf{X}}_i)) \\
= & -\frac{1}{2N}E\left[\left(2\exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\overline{\boldsymbol{\pi}}) - r^*(\overline{\mathbf{X}}_i)\right)\exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\overline{\boldsymbol{\pi}})\mathbf{G}^M(\overline{\mathbf{X}}_i))'\mathbf{G}^M(\overline{\mathbf{X}}_i))\right] + o_p(1) \\
\leq & -2\eta\hat{S}_M + o_p(1)
\end{aligned}$$

with its eigenvalues bounded away from zero in absolute value by η. Then, rearranging (B.3) and using the triangle inequality, with probability greater than $1-\varepsilon$, for$\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$,

$$\begin{aligned}
\frac{1}{N}L_N(\boldsymbol{\pi}) - \frac{1}{N}L_N(\boldsymbol{\pi}^*_M) &\leq \frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)(\boldsymbol{\pi} - \boldsymbol{\pi}^*) - \eta\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|^2 + o_p(1) \\
&\leq \left\|\frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)\right\|\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\| - \eta\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|^2 + o_p(1) \\
&= \left(\left\|\frac{1}{N}\frac{\partial L_N}{\partial\boldsymbol{\pi}}(\boldsymbol{\pi}^*_M)\right\| - \eta\sqrt{\frac{M}{N}}\right)\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\| + o_p(1) \\
&< 0
\end{aligned}$$

That is, we have with probability greater than $1-\varepsilon$, $\frac{1}{N}L_N(\boldsymbol{\pi}) < \frac{1}{N}L_N(\boldsymbol{\pi}^*_M)$ for all $\boldsymbol{\pi}$ with $\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\| = \sqrt{\frac{M}{N}}$ Since $L_N(\boldsymbol{\pi})$ is continuous, it has a maximum on the compact set $\boldsymbol{\pi} : \|\boldsymbol{\pi} - \boldsymbol{\pi}^*\| \leq \sqrt{\frac{M}{N}}$. By the last inequality, this maximum must occur for some $\hat{\boldsymbol{\pi}}_M$ with $\|\hat{\boldsymbol{\pi}}_M - \boldsymbol{\pi}^*\| < \sqrt{\frac{M}{N}}$. Hence the first order conditions are satisfied at $\hat{\boldsymbol{\pi}}_M$ and by concavity of $L_N(\boldsymbol{\pi})$, $\hat{\boldsymbol{\pi}}_M$ maximize $L_N(\boldsymbol{\pi})$ over all of $\mathbf{G}^M$. Because the probability of this is greater than $1-\varepsilon$ with ε arbitrary, we conclude that $\hat{\boldsymbol{\pi}}_M$ exists and satisfies the first order conditions with probability approaching one, and that $\|\hat{\boldsymbol{\pi}}_{M(N)} - \boldsymbol{\pi}^*_{M(N)}\| = O_p\left(\sqrt{\frac{M(N)}{N}}\right)$.

PROOF OF THEOREM 2.4.9:

let

$$
\begin{aligned}
\omega_{it} &= (\mathbf{X}_{it}, \mathbf{G}^M(\overline{\mathbf{X}}_i)), \\
\theta &= \begin{pmatrix} \beta \\ \boldsymbol{\pi}^*_M \end{pmatrix} \\
j(w_{it}, \theta) &= T^{-1} \sum_{t=1}^{T} \exp(\mathbf{X}_{it}\beta + \mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M)\beta
\end{aligned}
$$

So,

$$
\sqrt{N}(\hat{\tau} - \tau) = \sqrt{N}(\hat{\tau} - Ej(w_{it}, \theta)) + \sqrt{N}(Ej(w_{it}, \theta) - \tau) \equiv T1 + T2
$$

Note that,

$$
\begin{aligned}
T1 \;&=\; \sqrt{N}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(\exp\Big(\mathbf{X}_{it}\hat{\beta}_{pqml}+\mathbf{G}^{M}(\overline{\mathbf{X}}_i)\hat{\boldsymbol{\pi}}_M\Big)\hat{\beta}_{pqml} \\
&\quad -\exp\Big(\mathbf{X}_{it}\beta+\mathbf{G}^{M}(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^{*}_{M}\Big)\beta\Big) \\
&+\; \sqrt{N}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(\exp\Big(\mathbf{X}_{it}\beta+\mathbf{G}^{M}(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^{*}_{M})\Big)\beta - Ej(w_{it},\theta)\Big) \\
&=\; \sqrt{N}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(\exp(w_{it}\hat{\theta})\hat{\beta}_{pqml}-\exp(w_{it}\theta)\beta\Big) \\
&+\; \sqrt{N}\left((N)^{-1}\sum_{i=1}^{N}T^{-1}\sum_{t=1}^{T}\exp(w_{it}\theta)-ET^{-1}\sum_{t=1}^{T}\exp(w_{it}\theta)\right)\beta \\
&=\; N^{-1}\sum_{i=1}^{N}T^{-1}\sum_{t=1}^{T}\exp(w_{it}\theta)\sqrt{N}(\hat{\beta}_{pqml}-\beta) \\
&+\; N^{-1}\sum_{i=1}^{N}\nabla_{\theta}\left((T)^{-1}\sum_{t=1}^{T}\exp(w_{it}\theta)\right)(\theta)\left(\hat{\theta}-\theta\right) \\
&+\; \sqrt{N}\left(N^{-1}\sum_{i=1}^{N}j(w_{it},\theta)-Ej(w_{it},\theta)\right)+o_p(1) \\
&=\; N^{-1}\sum_{i=1}^{N}j(w_{it},\theta)\sqrt{N}(\hat{\beta}_{pqml}-\beta) \\
&+\; N^{-1}\sum_{i=1}^{N}\nabla_{\theta}\left(j(w_{it},\theta)\right)(\theta)\left(\hat{\theta}-\theta\right) \\
&+\; \sqrt{N}\left(N^{-1}\sum_{i=1}^{N}j(w_{it},\theta)-Ej(w_{it},\theta)\right)+o_p(1)
\end{aligned}
$$

While, $\left\|N^{-1}\sum_{i=1}^{N}\nabla_{\theta}\left(j(w_{it},\theta)\right)(\theta)\right\| = O_p\left(\zeta(M)\right), \left\|\hat{\theta}-\theta\right\| = O_p\left(\sqrt{\frac{M(N)}{N}}\right)$ So the second term of the above equation is of order $O_p\left(\zeta(M)\sqrt{\frac{M(N)}{N}}\right)$, which vanishes by assumptions in Lemma 2.

On the other hand,

$$T2 \equiv \sqrt{N}(Ej(w_{it},\theta)-\tau) = \sqrt{N}\left(T^{-1}\sum_{t=1}^{T} E\exp(\mathbf{X}_{it}\beta)(\exp(\mathbf{G}^M(\overline{\mathbf{X}}_i)\boldsymbol{\pi}^*_M) - r(\overline{\mathbf{X}}_i))\right)\beta$$

From Lemma 1, we know the term in the parasynthesis is of order $O_p(CK^{-s})$; as long as $N^{(1/2)}M^{-s} \to 0$, it can be ignored too.

Hence,

$$\sqrt{N}(\hat{\tau}-\tau) = N^{-1}\sum_{i=1}^{N} j(w_{it},\theta)\sqrt{N}(\hat{\beta}_{pqml}-\beta)+\sqrt{N}\left(N^{-1}\sum_{i=1}^{N} j(w_{it},\theta) - Ej(w_{it},\theta)\right)+o_p(1)$$

From Wooldridge (1999), the first term follows as:

$$\sqrt{N}(\hat{\beta}_{pqml}-\beta) = N^{-1/2}\sum_{i=1}^{N} \nabla^2_{\beta}\mathbf{p}_1(\mathbf{X}_i,\beta)W_1(\mathbf{X}_i,\beta)(\mathbf{Y}_i - \mathbf{p}_1(\mathbf{X}_i,\beta)n_{1i})$$

And the second term:

$$\sqrt{N}\left(N^{-1}\sum_{i=1}^{N} j(w_{it},\theta) - Ej(w_{it},\theta)\right) = N^{-1/2}\sum_{i=1}^{N}(j(w_{it},\theta) - Ej(w_{it},\theta))$$

Therefore:

$$\begin{aligned}\sqrt{N}(\hat{\tau}-\tau) &= N^{-1/2}\sum_{i=1}^{N}\{Ej(w_{it},\theta)\nabla^2_{\beta}\mathbf{p}_1(\mathbf{X}_i,\beta)W_1(\mathbf{X}_i,\beta)(\mathbf{Y}_i - \mathbf{p}_1(\mathbf{X}_i,\beta)n_{1i}) \\ &+ (j(w_{it},\theta) - Ej(w_{it},\theta))\} + o_p(1)\end{aligned}$$

We follow that:

$$\sqrt{N}(\hat{\tau}-\tau) \Rightarrow \mathbf{N}(0,V)$$

Where,

$$V = \text{Var}\Big(Ej(w_{it},\theta)\nabla^2_\beta \mathbf{p}_1(\mathbf{X}_i,\beta)W_1(\mathbf{X}_i,\beta)(\mathbf{Y}_i - \mathbf{p}_1(\mathbf{X}_i,\beta)n_{1i}) + (j(w_{it},\theta) - Ej(w_{it},\theta))\Big)$$

Q.E.D.

As for the estimation of V is straight forward:

let,

$$\begin{aligned}\hat{V}_{i1} &= \left(N^{-1}\sum_{i=1}^{N} j(w_{it},\hat{\theta})\right)\nabla^2_\beta \mathbf{p}_1(\mathbf{X}_i,\hat{\beta}_{pqml})W_1(\mathbf{X}_i,\hat{\beta}_{pqml})(\mathbf{Y}_i - \mathbf{p}_1(\mathbf{X}_i,\hat{\beta}_{pqml})n_{1i}) \\ &+ \left(j(w_{it},\hat{\theta}) - N^{-1}\sum_{i=1}^{N} j(w_{it},\hat{\theta})\right)\end{aligned}$$

then,

$$\hat{V} = N^{-1}\sum_{i=1}^{N}\hat{V}_{i1}\hat{V}'_{i1}$$

Note, for the denotations here, please refer to section 2 and 3.

GMM Simulation Setup:

- we do the following setting up:

$$\sum_{t=1}^{T} y_{it} = n_i$$

$$\sum_{t=1}^{T} y_{it}^2 = n_{i2}$$

$$p_t(x_i,\beta) \equiv \frac{\exp(\beta x_{it})}{\sum_{r=1}^{T}\exp(\beta x_{ir})}$$

$$p_{t2}(x_i,\beta) \equiv \frac{\exp(2\beta x_{it})}{\sum_{r=1}^{T}\exp(2\beta x_{ir})}$$

$$\mathbf{p}(x_i,\beta) \equiv [p_1(x_i,\beta),\ ...,\ p_T(x_i,\beta)]',$$

$$\mathbf{p}_2(x_i,\beta) \equiv [p_{12}(x_i,\beta),\ ...,\ p_{T2}(x_i,\beta)]',$$

$$u_{1i}(\beta) \equiv Y_i - p(x_i,\beta)n_i, where \ \ Y_i = [Y_{i1},\ \cdots,\ Y_{iT}]'$$

$$u_{2i}(\beta) \equiv Y_{i2} - p_2(x_i,\beta)n_{i2}, where, Y_{i2} = [Y_{i1}^2,\ \cdots,\ Y_{iT}^2]'$$

$$D_1(x_i,\beta) = \left[x_{i1} - \frac{\sum_{r=1}^T x_{ir}\exp(\beta x_{ir})}{\sum_{r=1}^T \exp(\beta x_{ir})},\ ...,\ x_{iT} - \frac{\sum_{r=1}^T x_{ir}\exp(\beta x_{ir})}{\sum_{r=1}^T \exp(\beta x_{ir})}\right]',$$

$$D_2(x_i,\beta) = \left[2x_{i1} - \frac{\sum_{r=1}^T 2x_{ir}\exp(2\beta x_{ir})}{\sum_{r=1}^T \exp(2\beta x_{ir})},\ ...,\ 2x_{iT} - \frac{\sum_{r=1}^T 2x_{ir}\exp(2\beta x_{ir})}{\sum_{r=1}^T \exp(2\beta x_{ir})}\right]',$$

$$\begin{aligned} D_3(x_i,\beta) &= n_i\left[p_1(x_i,\beta)\left(x_{i1} - \frac{\sum_{r=1}^T x_{ir}\exp(\beta x_{ir})}{\sum_{r=1}^T \exp(\beta x_{ir})}\right),\ \cdots,\right. \\ &\quad \left. p_T(x_i,\beta)\left(x_{iT} - \frac{\sum_{r=1}^T x_{ir}\exp(\beta x_{ir})}{\sum_{r=1}^T \exp(\beta x_{ir})}\right)\right] \\ &= \left[\frac{n_i x_{i1}\exp(\beta x_{i1})}{\sum_{r=1}^T \exp(\beta x_{ir})} - \frac{n_i\exp(\beta x_{i1})\sum_{r=1}^T x_{ir}\exp(\beta x_{ir})}{(\sum_{r=1}^T \exp(\beta x_{ir}))^2},\ \cdots,\right. \\ &\quad \left. \frac{n_i x_{iT}\exp(\beta x_{iT})}{\sum_{r=1}^T \exp(\beta x_{ir})} - \frac{n_i\exp(\beta x_{iT})\sum_{r=1}^T x_{ir}\exp(\beta x_{ir})}{(\sum_{r=1}^T \exp(\beta x_{ir}))^2}\right] \end{aligned}$$

$$\begin{aligned} D_i(x_i,\beta) &= \begin{bmatrix} D_1(x_i,\beta) & \mathbf{0} \\ \mathbf{0} & D_2(x_i,\beta) \end{bmatrix} \\ u_i(\beta) &= \begin{bmatrix} u_{1i}(\beta) \\ u_{2i}(\beta) \end{bmatrix} \end{aligned}$$

- Step 1: PQML

$$\hat{\beta}_{pqml} = \arg\max \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it} \log(p_t(x_i,\beta))$$

$$se(\hat{\beta}_{pqml}) = \left(\left(\sum_{i=1}^{N} \widetilde{D}_3 \widetilde{D}_1 \right)^{-1} \left(\sum_{i=1}^{N} \widetilde{D}_1' \widetilde{u}_{i1} \widetilde{u}_{1i}' \widetilde{D}_1 \right) \left(\sum_{i=1}^{N} \widetilde{D}_3 \widetilde{D}_1 \right)^{-1} \right)^{1/2}$$

Where,

$$\widetilde{D}_1 = D_1(x_i, \hat{\beta}_{pqml}), \widetilde{D}_3 = D_3(x_i, \hat{\beta}_{pqml}), \widetilde{u}_{1i} = u_{1i}(\hat{\beta}_{pqml})$$

- Step 2: GMM

$$\hat{\beta}_{gmm} = \arg\min \left(N^{-1} \sum_{i=1}^{N} \widetilde{D}_i' u_i(\beta) \right)' \left(N^{-1} \sum_{i=1}^{N} \widetilde{D}_i' \widetilde{u}_i \widetilde{u}_i' \widetilde{D}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \widetilde{D}_i' u_i(\beta) \right)$$

$$se(\hat{\beta}_{gmm}) = \left(\left(\sum_{i=1}^{N} \widetilde{D}_i' \nabla_{\hat{\beta}} u_i \right)' \left(\sum_{i=1}^{N} \widetilde{D}_i' \widetilde{u}_i \widetilde{u}_i' \widetilde{D}_i \right)^{-1} \left(\sum_{i=1}^{N} \widetilde{D}_i' \nabla_{\hat{\beta}} u_i \right) \right)^{-1/2}$$

Where,

$$\widetilde{D}_i = D_i(x_i, \hat{\beta}_{pqml}), \widetilde{u}_i = u_i(\hat{\beta}_{pqml})$$

$$
\begin{aligned}
& \nabla_{\hat{\beta}} u_i \\
= & \begin{bmatrix} \nabla_{\beta} u_{1i}(\hat{\beta}_{gmm}) \\ \\ \nabla_{\beta} u_{2i}(\hat{\beta}_{gmm}) \end{bmatrix} \\
= & \Bigg[\left(\frac{x_{i1} \exp(\hat{\beta}_{gmm} x_{i1})}{\sum_{r=1}^{T} \exp(\hat{\beta}_{gmm} x_{ir})} - \frac{\exp(\beta x_{i1}) \sum_{r=1}^{T} x_{ir} \exp(\hat{\beta}_{gmm} x_{ir})}{(\sum_{r=1}^{T} \exp(\hat{\beta}_{gmm} x_{ir}))^2} \right) n_i, \\
& \qquad \cdots, \\
& \left(\frac{x_{iT} \exp(\hat{\beta}_{gmm} x_{iT})}{\sum_{r=1}^{T} \exp(\hat{\beta}_{gmm} x_{ir})} - \frac{\exp(\beta x_{iT}) \sum_{r=1}^{T} x_{ir} \exp(\hat{\beta}_{gmm} x_{ir})}{(\sum_{r=1}^{T} \exp(\hat{\beta}_{gmm} x_{ir}))^2} \right) n_i \\
& \\
& \left(\frac{2x_{i1} \exp(2\hat{\beta}_{gmm} x_{i1})}{\sum_{r=1}^{T} \exp(2\hat{\beta}_{gmm} 2x_{ir})} - \frac{\exp(2\hat{\beta}_{gmm} x_{i1}) \sum_{r=1}^{T} 2x_{ir} \exp(2\hat{\beta}_{gmm} x_{ir})}{(\sum_{r=1}^{T} \exp(2\hat{\beta}_{gmm} x_{ir}))^2} \right) n_{2i}, \\
& \qquad \cdots, \\
& \left(\frac{2x_{iT} \exp(2\hat{\beta}_{gmm} x_{iT})}{\sum_{r=1}^{T} \exp(2\hat{\beta}_{gmm} 2x_{ir})} - \frac{\exp(2\hat{\beta}_{gmm} x_{iT}) \sum_{r=1}^{T} 2x_{ir} \exp(2\hat{\beta}_{gmm} x_{ir})}{(\sum_{r=1}^{T} \exp(2\hat{\beta}_{gmm} x_{ir}))^2} \right) n_{2i} \Bigg]
\end{aligned}
$$

- Step 3: OIV

$$
\hat{\beta}_{oiv} = \arg\min \left(N^{-1} \sum_{i=1}^{N} \widetilde{B}_i' u_{1i}(\beta) \right)' \left(N^{-1} \sum_{i=1}^{N} \widetilde{B}_i' \widetilde{B}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \widetilde{B}_i' u_{1i}(\beta) \right)
$$

$$
se(\hat{\beta}_{gmm}) = \left(\left(\sum_{i=1}^{N} \widetilde{B}_i' \nabla_{\hat{\beta}} u_{1i} \right)' \left(\sum_{i=1}^{N} \widetilde{B}_i' \widetilde{B}_i \right)^{-1} \left(\sum_{i=1}^{N} \widetilde{B}_i' \nabla_{\hat{\beta}} u_{1i} \right) \right)^{-1/2}
$$

Where,

$$
\begin{aligned}
\widetilde{u}_i &= u_{1i}(\hat{\beta}_{pqml}) \\
\widetilde{B}(\mathbf{X}_i) &= \hat{D}(\mathbf{X}_i) * (\hat{\Omega}_i)^{-1} / \hat{g}(\overline{\mathbf{X}}_i) \\
\hat{D}_i &= -\nabla_{\beta}[\mathbf{p}_1(\mathbf{X}_i, \hat{\beta}_{pqml})] \sum_{t=1}^{T} \exp(\hat{\beta}_{pqml} \mathbf{X}_{it}) \hat{r}(\overline{\mathbf{X}}_i)
\end{aligned}
$$

$$
\begin{aligned}
& \nabla_{\hat{\beta}} u_{1i} \\
\equiv \;& \nabla_{\beta} u_{1i}(\hat{\beta}_{oiv}) \\
= \;& \left[\left(\frac{x_{i1} \exp(\hat{\beta}_{oiv} x_{i1})}{\sum_{r=1}^{T} \exp(\hat{\beta}_{oiv} x_{ir})} - \frac{\exp(\hat{\beta}_{oiv} x_{i1}) \sum_{r=1}^{T} x_{ir} \exp(\hat{\beta}_{oiv} x_{ir})}{(\sum_{r=1}^{T} \exp(\hat{\beta}_{oiv} x_{ir}))^2} \right) \exp(\mathbf{G}^M(\mathbf{X}_i)\hat{\boldsymbol{\pi}}_M), \right. \\
& \qquad \cdots, \\
& \left. \left(\frac{x_{iT} \exp(\hat{\beta}_{oiv} x_{iT})}{\sum_{r=1}^{T} \exp(\hat{\beta}_{oiv} x_{ir})} - \frac{\exp(\hat{\beta}_{oiv} x_{iT}) \sum_{r=1}^{T} x_{ir} \exp(\hat{\beta}_{oiv} x_{ir})}{(\sum_{r=1}^{T} \exp(\hat{\beta}_{oiv} x_{ir}))^2} \right) \exp(\mathbf{G}^M(\mathbf{X}_i)\hat{\boldsymbol{\pi}}_M) \right]
\end{aligned}
$$

Appendix C

Tables

Table C.1: Summary Statistics

	Mean	Median	Standard deviation	Minimum	Maximum	Sample size
math4	37.73	38.05	15.107	3.1	81.3	518
avgexp	9037	8646	1644.950	7258	28611	518
exp10	9251	8800	1862.571	6890	30379	518
enroll	2769.4	1596.5	4473.652	64	75263	518
lunch	39.087	37.792	16.049	5.993	87.815	518
scdist	20.303	16.935	16.250	.134	80.952	518

Table C.2: OLS Regression, dependent variable=math4

	Coef. Estimate	Usual Std. Err.	H-W Std. Err.	ISD cluster Std. Err.
log(avgexp)	0.1595	0.0405*	0.0558*	0.0671*
lunch	-0.0058	0.0004*	0.0004*	0.0005*
log(enroll)	0.0147	0.0068*	0.0076	0.0086
log(scdist)	0.0030	0.0068	0.0068	0.0080
constant	-0.9639	0.3744*	0.4973	0.6036
log(exp10)	0.1289	0.0373*	0.0552*	0.0699
lunch	-0.0058	0.0004*	0.0004*	0.0005*
log(enroll)	0.0145	0.0068*	0.0077	0.0087
log(scdist)	0.0017	0.0068	0.0068	0.0080
constant	-0.6848	0.3451*	0.4895	0.6259

* Significant at 5% level

Table C.3: OLS Regression with Conley S.E., dependent variable=math4

		log(exp10)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.1289	0.0145	-0.00576	0.0017	-0.6848
H-W	Std. Err.	0.0552*	0.0077	0.00038	0.0068	0.4895
Conley	Std. Err.					
cut1	cut2					
50	100	0.0606*	0.0086	0.00044*	0.0080	0.5507
100	150	0.0636*	0.0085	0.00044*	0.0079	0.5743
150	200	0.0640*	0.0077	0.00042*	0.0075	0.5737
200	250	0.0633*	0.0072*	0.00037*	0.0075	0.5696
350	400	0.0586*	0.0062*	0.00032*	0.0069	0.5322
400	500	0.0552*	0.0063*	0.00030*	0.0062	0.4981
500	600	0.0517*	0.0060*	0.00027*	0.0057	0.4682
600	700	0.0482*	0.0054*	0.00025*	0.0054	0.4366
700	800	0.0455*	0.0050*	0.00024*	0.0051	0.4134
800	900	0.0430*	0.0046*	0.00022*	0.0049	0.3926
1000	1000	0.0404*	0.0042*	0.00021*	0.0048	0.3709

Table C.4: OLS Regression with Conley S.E., dependent variable=math4

		log(avgexp)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.1595	0.0147	-0.00579	0.0030	-0.9639
White	Std. Err.	0.0558	0.0076	0.00036	0.0068	0.4973
Conley	Std. Err.					
cut1	cut2					
50	100	0.0616*	0.0086	0.00042*	0.0080	0.5640
100	150	0.0638*	0.0084	0.00042*	0.0080	0.5802
150	200	0.0634*	0.0076	0.00040*	0.0075	0.5710
200	250	0.0625*	0.0071*	0.00035*	0.0075	0.5655
350	400	0.0591*	0.0061*	0.00029*	0.0070	0.5412
400	500	0.0559*	0.0062*	0.00028*	0.0063	0.5112
500	600	0.0529*	0.0059*	0.00025*	0.0058	0.4856*
600	700	0.0497*	0.0053*	0.00023*	0.0055	0.4569*
700	800	0.0474*	0.0049*	0.00022*	0.0052	0.4371*
800	900	0.0452*	0.0046*	0.00021*	0.0050	0.4178*
1000	1000	0.0430*	0.0041*	0.00019*	0.0048	0.3991*

Table C.5: OLS Regression with Conley S.E., dependent variable=math4

		log(exp10)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.1289	0.0145	-0.00576	0.0017	-0.6848
H-W	Std. Err.	0.0552	0.0077	0.00038	0.0068	0.4895
Conley	Std. Err.					
cut1	cut2					
242	21	0.0553*	.00758	.00044*	0.0073	0.4942
364	41	0.0570*	.00760	.00049*	0.0078	0.5102
435	88	0.0620*	0.0080	.00050*	0.0089	0.5602
546.5	167.5	0.0602*	0.0072*	.00045*	0.0093	0.5541
649	284	0.0587*	0.0059*	.00033*	0.0088	0.5454
719	637	0.0491*	0.0052*	.00025*	0.0057	0.4506
758	880	0.0437*	0.0047*	.00023*	0.0050	0.3976

Table C.6: OLS Regression with Conley S.E., dependent variable=math4

		log(avgexp)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.1595	0.0147	-0.00579	0.0030	-0.9639
H-W	Std. Err.	0.0558	0.0076	0.00036	0.0068	0.4973
Conley	Std. Err.					
cut1	cut2					
242	21	0.0561*	.00755	.00042*	0.0073	0.5037
364	41	0.0578*	.00760	.00046*	0.0077	0.5226
435	88	0.0629*	0.0080	.00047*	0.0089	0.5740
546.5	167.5	0.0617*	0.0072*	.00042*	0.0093	0.5747
649	284	0.0609*	0.0059*	.00030*	0.0089	0.5713
719	637	0.0515*	0.0051*	.00023*	0.0058	0.4783*
758	880	0.0457*	0.0047*	.00021*	0.0051	0.4218*

Table C.7: OLS Regression, dependent variable=log($\frac{math4}{1-math4}$)

	Coef. Estimate	usual Std. Err	H-W Std. Err	ISD cluster Std. Err.	Peninsula cluster S.E.
log(avgexp)	0.5277	0.1976*	0.2946	0.3505	0.0937
log(enroll)	0.0919	0.0332*	0.0387*	0.0446*	0.0021*
lunch	-0.0275	0.0018*	0.0018*	0.0023*	0.0007*
log(scdist)	0.0235	0.0332	0.0328	0.0398	0.0166
constant	-5.0362	1.8271*	2.6355	3.1478	0.8732
log(exp10)	0.4152	0.1815*	0.2811	0.3554	0.0631
log(enroll)	0.0911	0.0333*	0.0389*	0.0450*	0.0000*
lunch	-0.0273	0.0018*	0.0019*	0.0024*	0.0007*
log(scdist)	0.0192	0.0331	0.0324	0.0396	0.0167
constant	-4.0088	1.6817*	2.5017	3.1720	0.5806

Table C.8: OLS Regression with Conley S.E., dependent variable=log($\frac{math4}{1-math4}$)

		log(exp10)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.4152	0.0911	-.0273	0.0192	-4.0088
H-W	Std. Err.	0.2811	0.0389	.00193	0.0324	2.5017
Conley	Std. Err.					
cut1	cut2					
50	100	0.3094	0.0432*	.00224*	0.0396	2.8198
100	150	0.3235	0.0430*	.00225*	0.0407	2.9325
150	200	0.3244	0.0397*	.00216*	0.0397	2.9270
200	250	0.3209	0.0376*	.00192*	0.0400	2.9141
350	400	0.2907	0.0331*	.00159*	0.0373	2.6737
400	500	0.2737	0.0335*	.00152*	0.0333	2.5026
500	600	0.2596	0.0319*	.00135*	0.0309	2.3823
600	700	0.2443	0.0287*	.00124*	0.0291	2.2454
700	800	0.2320	0.0263*	.00116*	0.0278	2.1394
800	900	0.2208	0.0245*	.00109*	0.0267	2.0409*
1000	1000	0.2079*	0.0222*	.00101*	0.0257	1.9330*

Table C.9: OLS Regression with Conley S.E., dependent variable=log($\frac{math4}{1-math4}$)

		log(avgexp)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.5277	0.0919	-.0275	0.0235	-5.0362
H-W	Std. Err.	0.2946	0.0387	.00187	0.0328	2.6355
Conley	Std. Err.					
cut1	cut2					
50	100	0.3248	0.0431*	.00218*	0.0401	2.9776
100	150	0.3345	0.0428*	.00217*	0.0411	3.0533
150	200	0.3323	0.0394*	.00208*	0.0400	3.0183
200	250	0.3292	0.0373*	.00184*	0.0405	3.0113
350	400	0.3037	0.0330*	.00147*	0.0381	2.8254
400	500	0.2856	0.0334*	.00141*	0.0342	2.6533
500	600	0.2720	0.0319*	.00125*	0.0317	2.5345*
600	700	0.2576*	0.0287*	.00116*	0.0298	2.4021*
700	800	0.2468*	0.0263*	.00109*	0.0285	2.3058*
800	900	0.2361*	0.0245*	.00102*	0.0273	2.2103*
1000	1000	0.2250*	0.0223*	.00095*	0.0263	2.1153*

Table C.10: OLS Regression with Conley S.E., dependent variable=log($\frac{math4}{1-math4}$)

		log(exp10)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.4152	0.0911	-0.0273	0.0192	-4.0088
H-W	Std. Err.	0.2811	0.0389	0.0019	0.0324	2.5017
Conley	Std. Err.					
cut1	cut2					
242	21	0.2787	0.0389*	0.0022*	0.0368	2.5073
364	41	0.2855	0.0387*	.00236*	0.0400	2.5739
435	88	0.3076	0.0412*	.00237*	0.0473	2.8016
546.5	167.5	0.3007	0.0374*	0.0021*	0.0506	2.8079
649	284	0.3026	0.0314*	0.0016*	0.0479	2.8503
719	637	0.2478	0.0278*	0.0012*	0.0308	2.3079
758	880	0.2238	0.0250*	0.0011*	0.0268	2.0654

Table C.11: OLS Regression with Conley S.E. in Nonlinear Model

		log(avgexp)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.5277	0.0919	-0.0275	0.0235	-5.0362
H-W	Std. Err.	0.2946	0.0387	0.0019	0.0328	2.6355
Conley	Std. Err.					
cut1	cut2					
242	21	0.2952	0.0389*	0.0021*	0.0369	2.6643
364	41	0.3024	0.0387*	0.0023*	0.0399	2.7459
435	88	0.3230	0.0413*	0.0022*	0.0472	2.9710
546.5	167.5	0.3192	0.0376*	0.0020*	0.0511	3.0140
649	284	0.3274	0.0317*	0.0015*	0.0487	3.1091
719	637	0.2665*	0.0279*	0.0011*	0.0315	2.5120*
758	880	0.2386*	0.0251*	0.0010*	0.0275	2.2302*

Table C.12: APEs with Bootstrap S.E. in Nonlinear Model

	APE Evaluated at				
	Mean	25%	50%	75%	95%
log(avgexp)	0.1151*	0.1193*	0.1156*	0.1100*	0.0982*
	(.0663)	(.0683)	(.0663)	(.0636)	(.0590)
log(enroll)	0.0200**	0.0208**	0.0201**	0.0191**	0.0171**
	(.0083)	(.0085)	(.0084)	(.0082)	(.0078)
lunch	-0.0060**	-0.0062**	-0.0060**	-0.0057**	-0.0051**
	(.0004)	(.0004)	(.0004)	(.0004)	(.0004)
log(scdist)	0.0051	0.0053	0.0052	0.0049	0.0044
	(.0071)	(.0073)	(.0072)	(.0068)	(.0063)
log(exp10)	0.0905	0.0940	0.0910	0.0863	0.0767
	(.0631)	(.0652)	(.0632)	(.0603)	(.0553)
log(enroll)	0.0199**	0.0206**	0.0200**	0.0189**	0.0168**
	(.0084)	(.0086)	(.0084)	(.0082)	(.0077)
lunch	-0.0060**	-0.0062**	-0.0060**	-0.0057**	-0.0051**
	(.0004)	(.0004)	(.0004)	(.0004)	(.0004)
log(scdist)	0.0042	0.0043	0.0042	0.0040	0.0035
	(.0070)	(.0072)	(.0070)	(.0067)	(.0061)

* Significant at 10% level
** Significant at 5% level
Bootstraps standard errors are in parenthesis

Figures

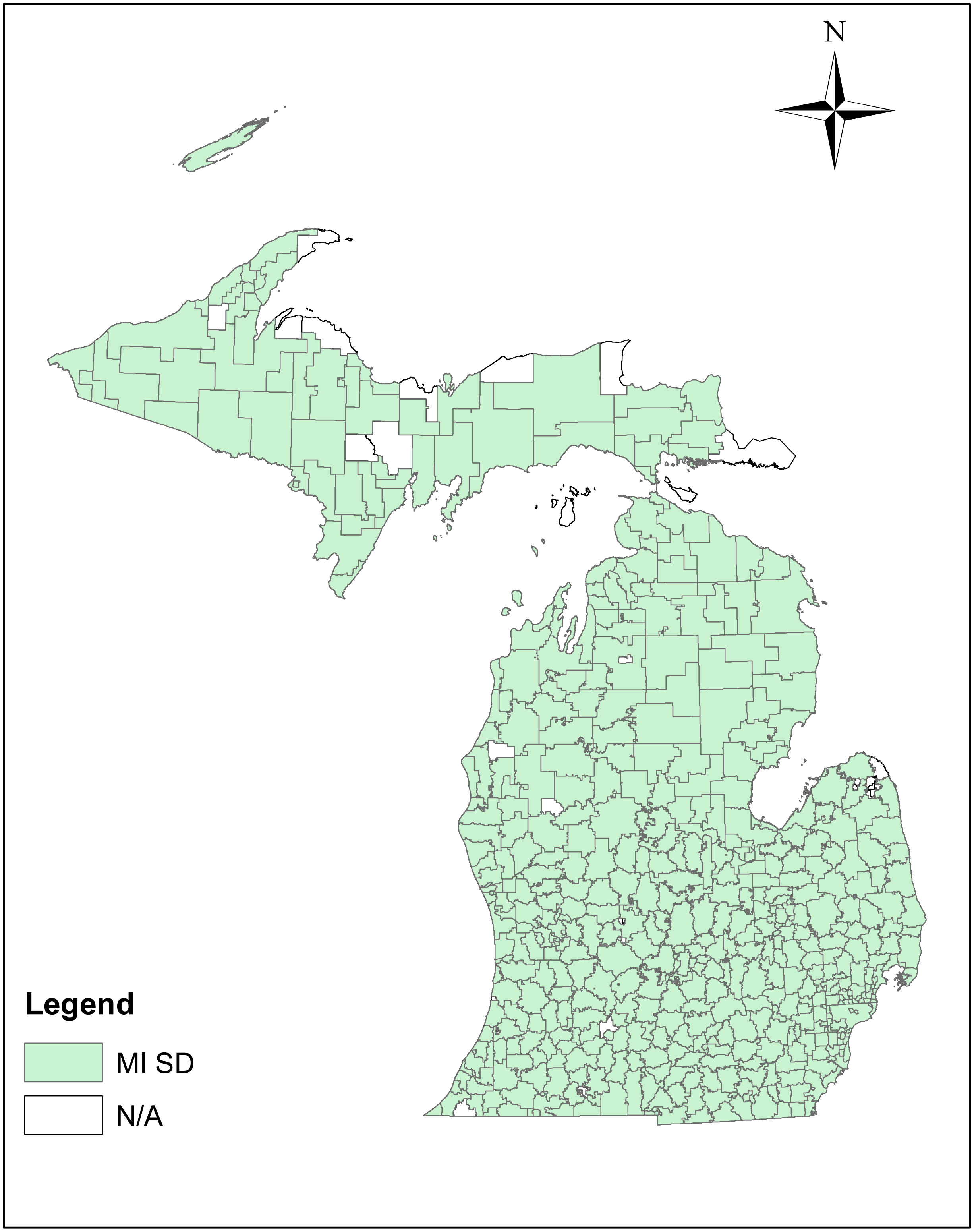

Figure C.1: All school districts of Michigan in 2010: For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

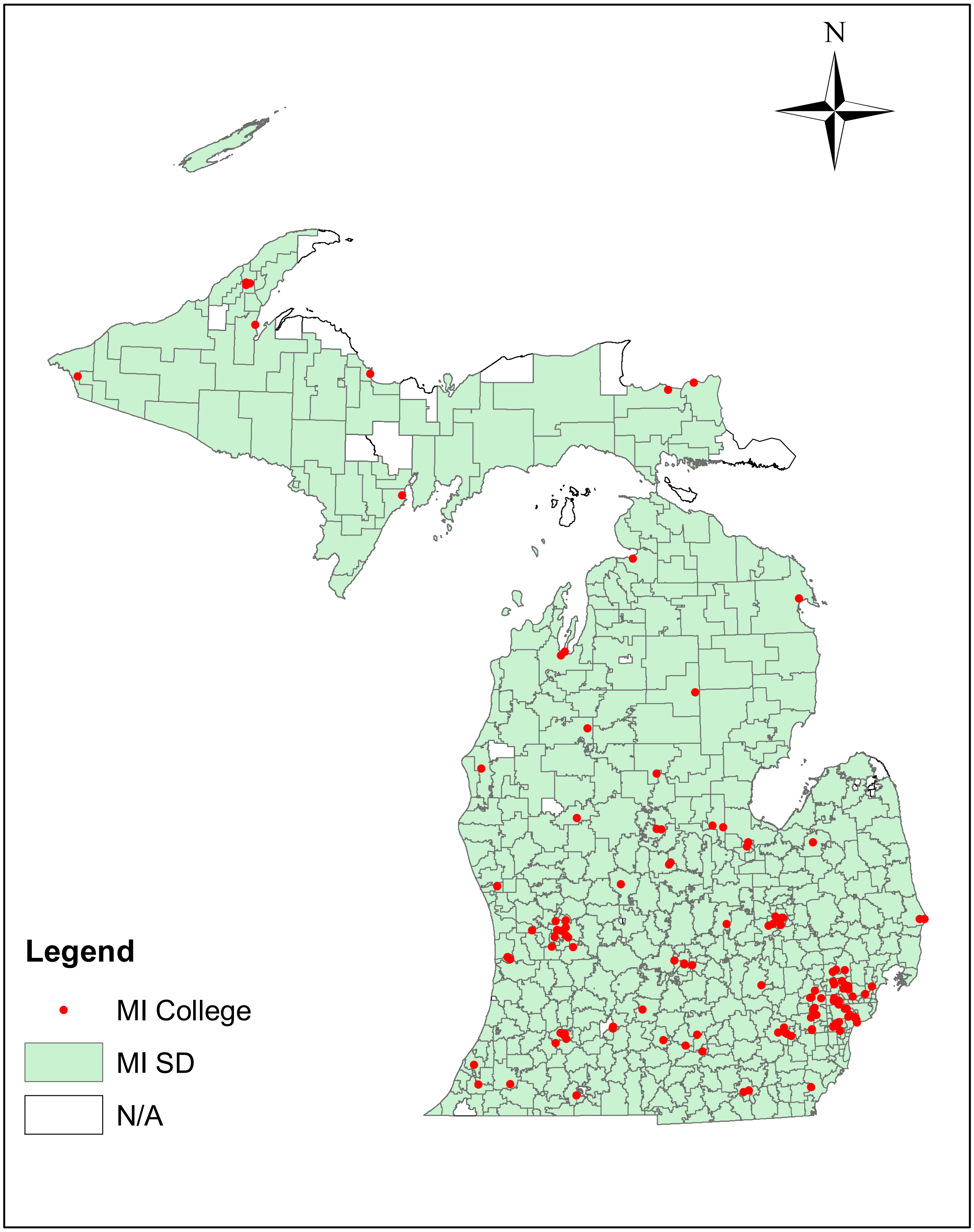

Figure C.2: All Colleges of Michigan in 2010.

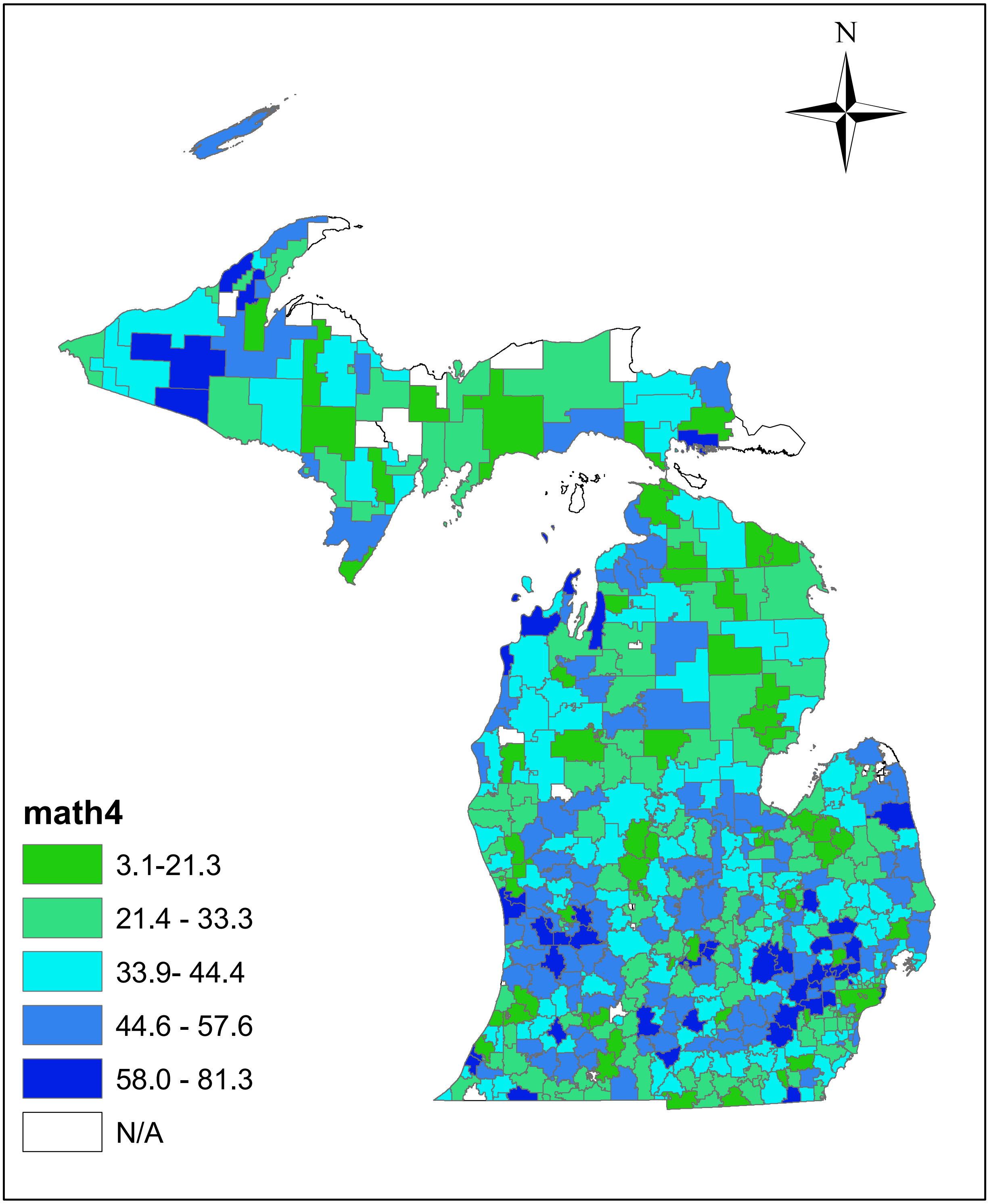

Figure C.3: MEAP math pass rate for 4th graders of Michigan in 2010.

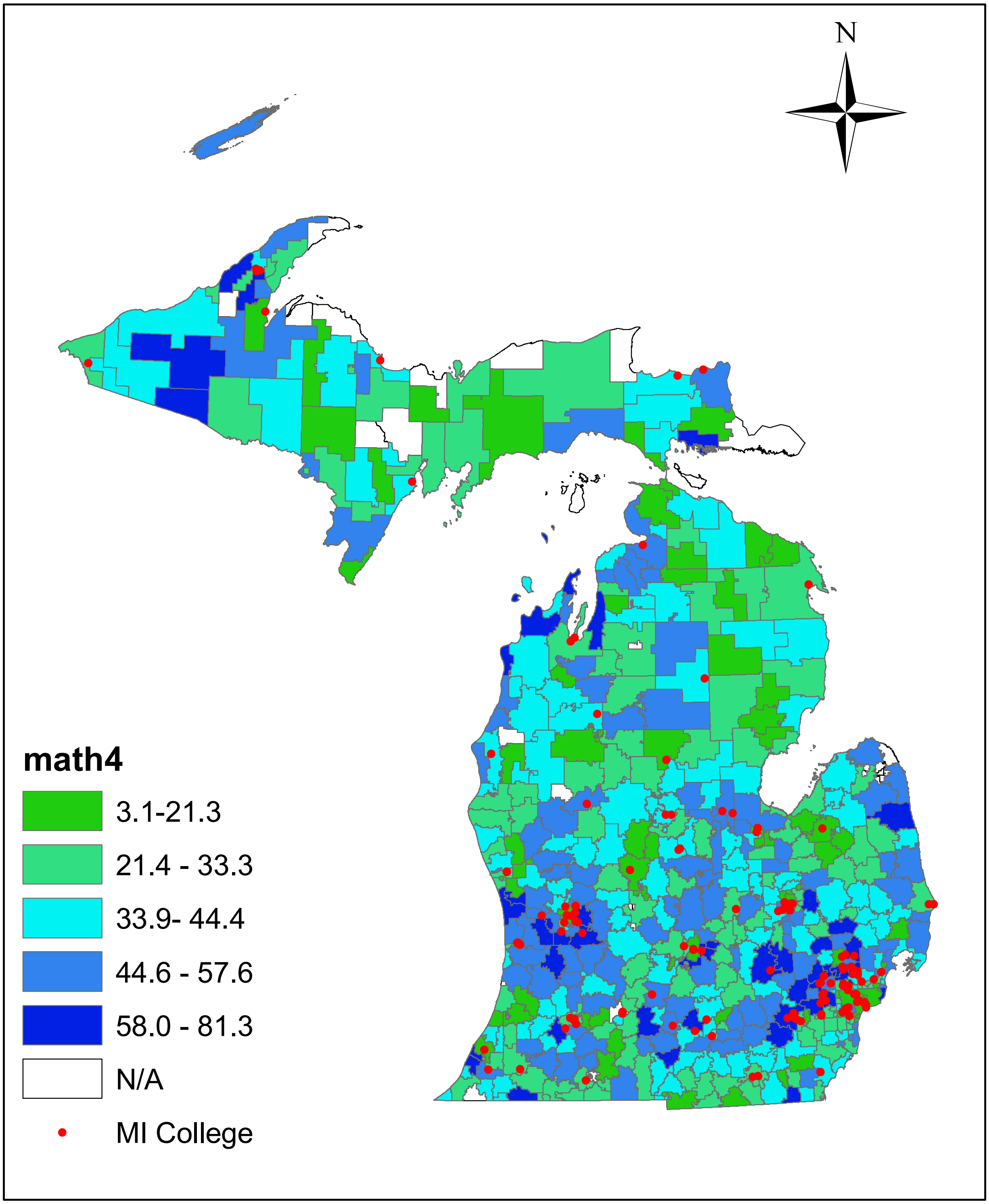

Figure C.4: Colleges and math pass rate for 4th graders of Michigan in 2010.

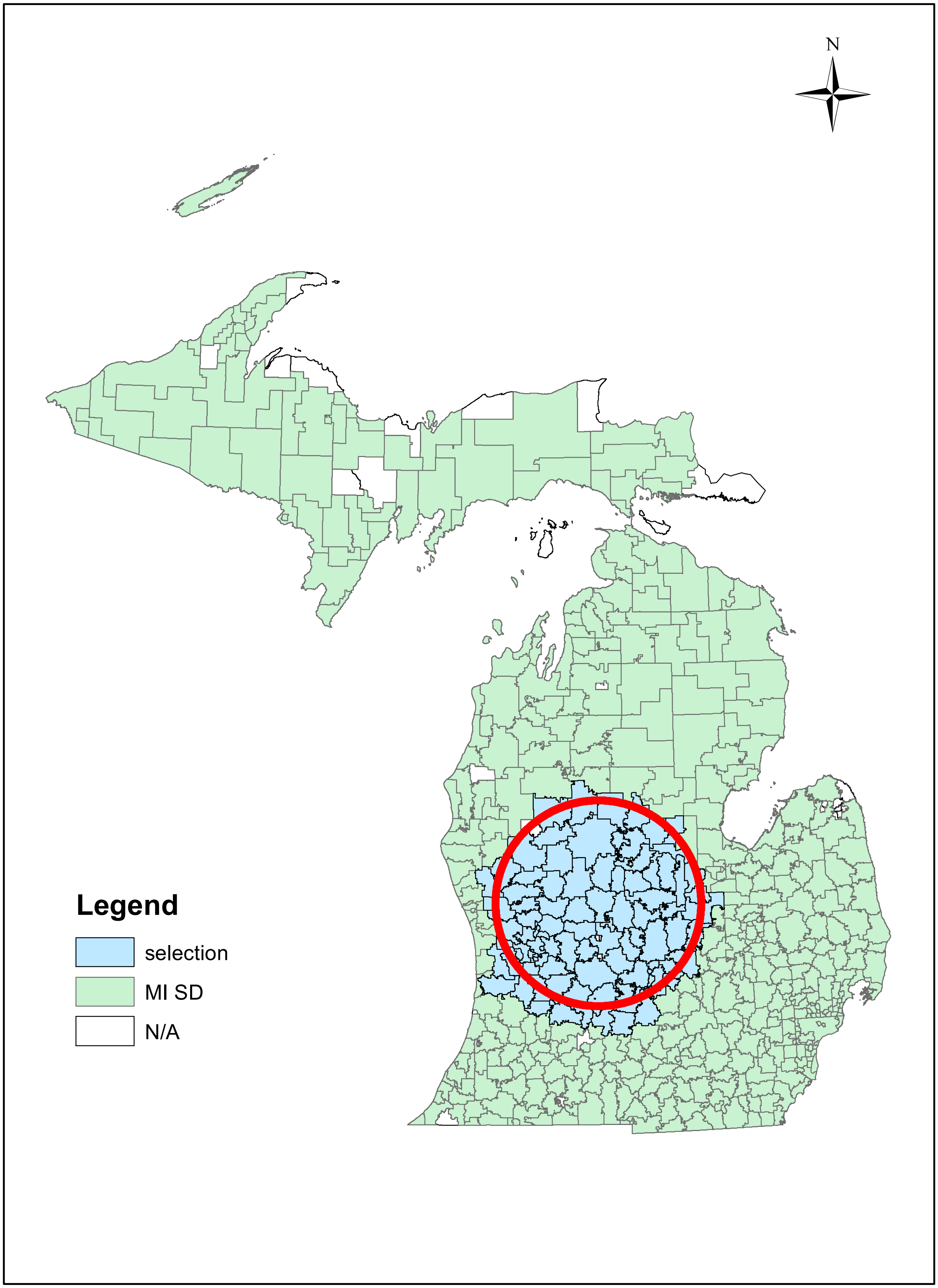

Figure C.5: Selection of 96 School Districts.

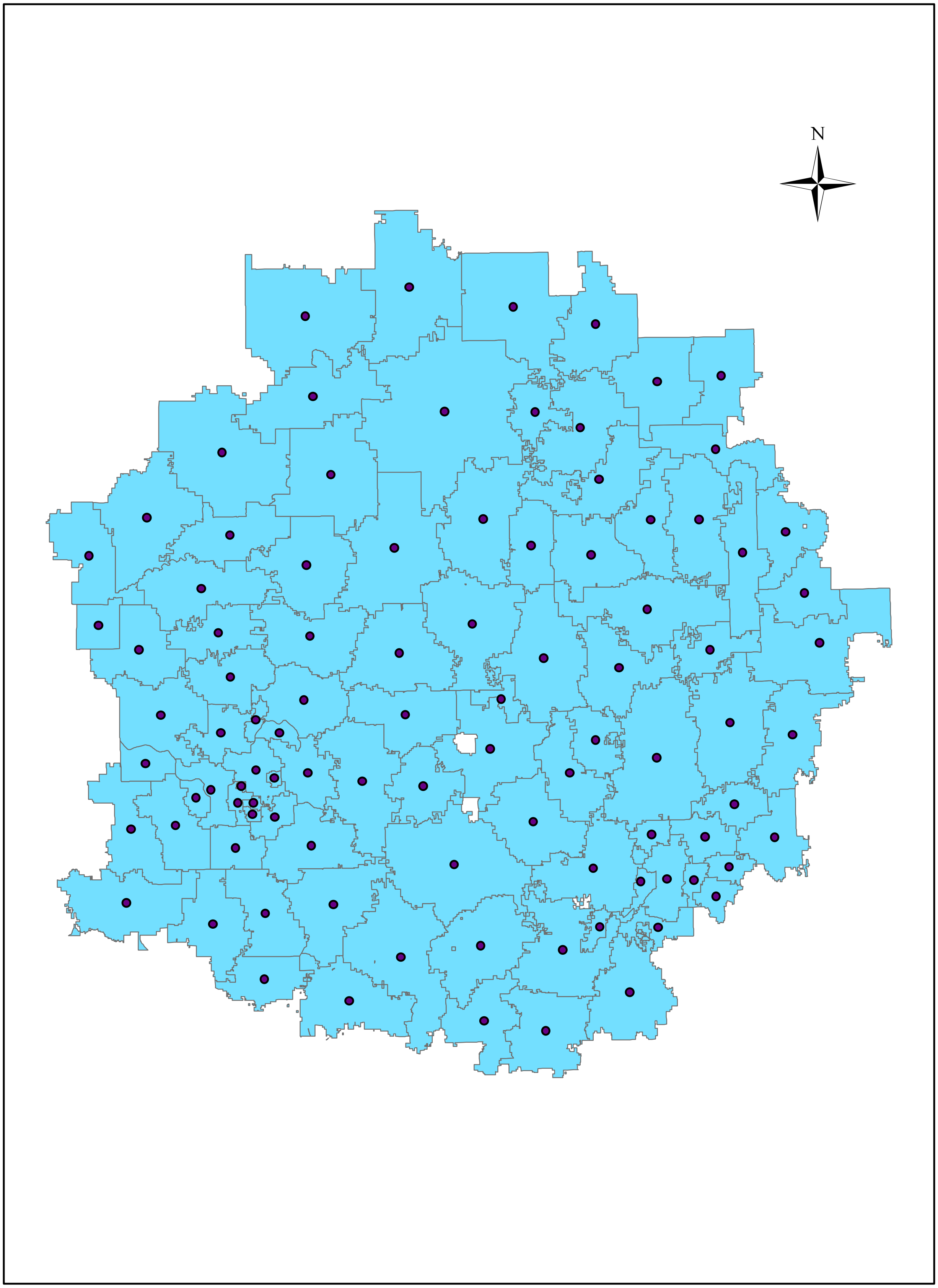

Figure C.6: Selection of 96 School Districts with centroids.

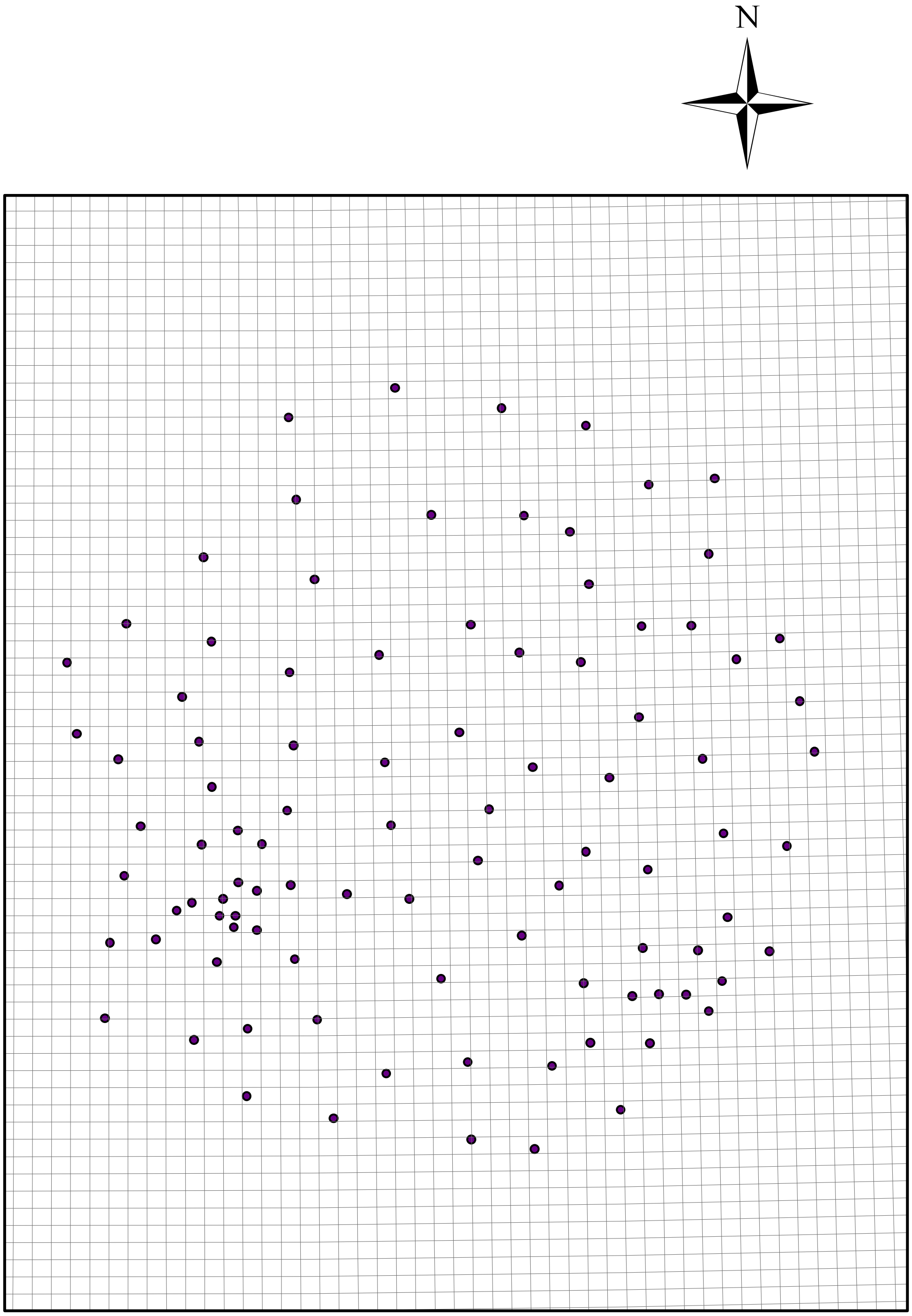

Figure C.7: Selection of 96 School Districts with centroids in grids.

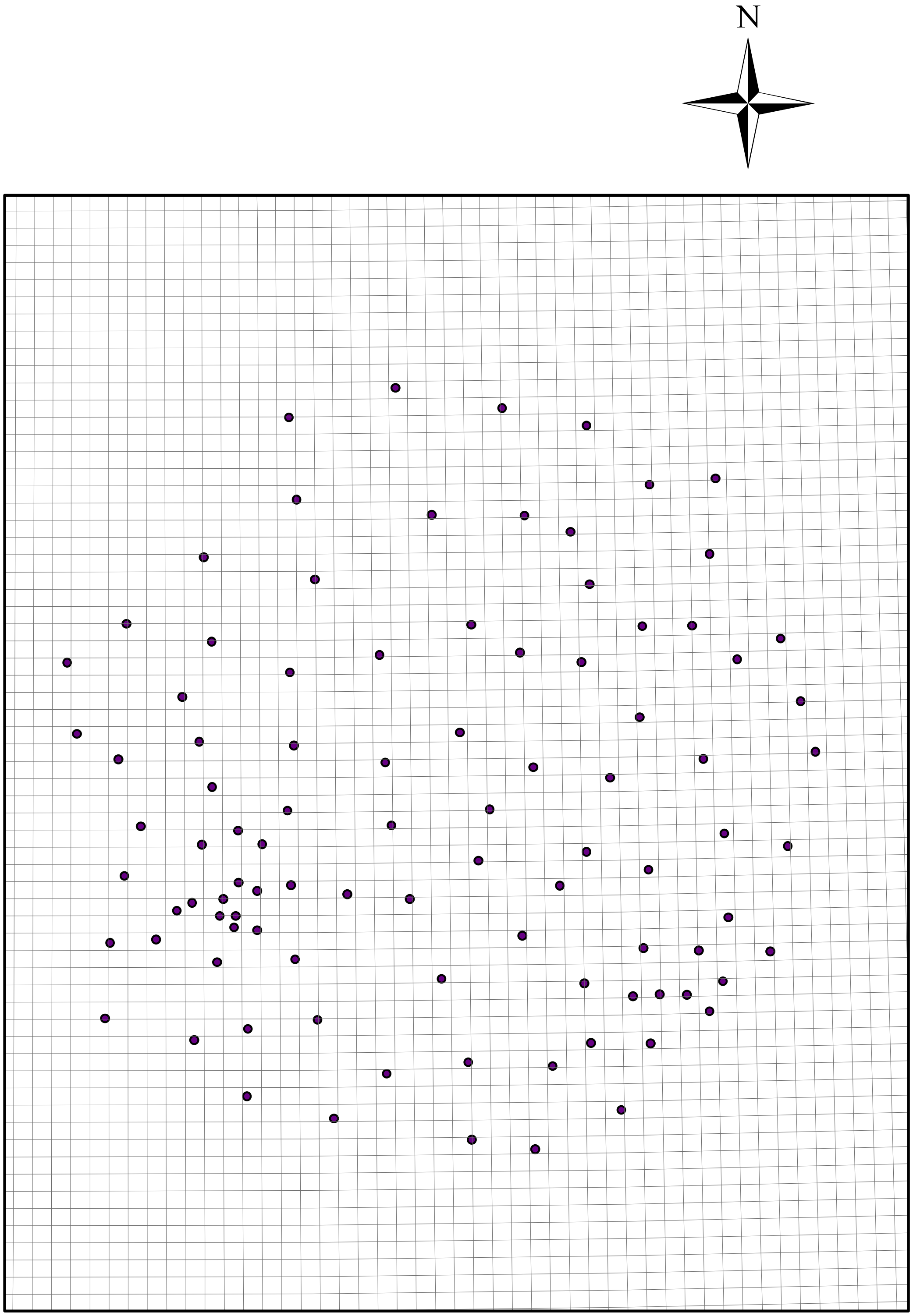

Figure C.8: Conley Coordinates of 96 School Districts.

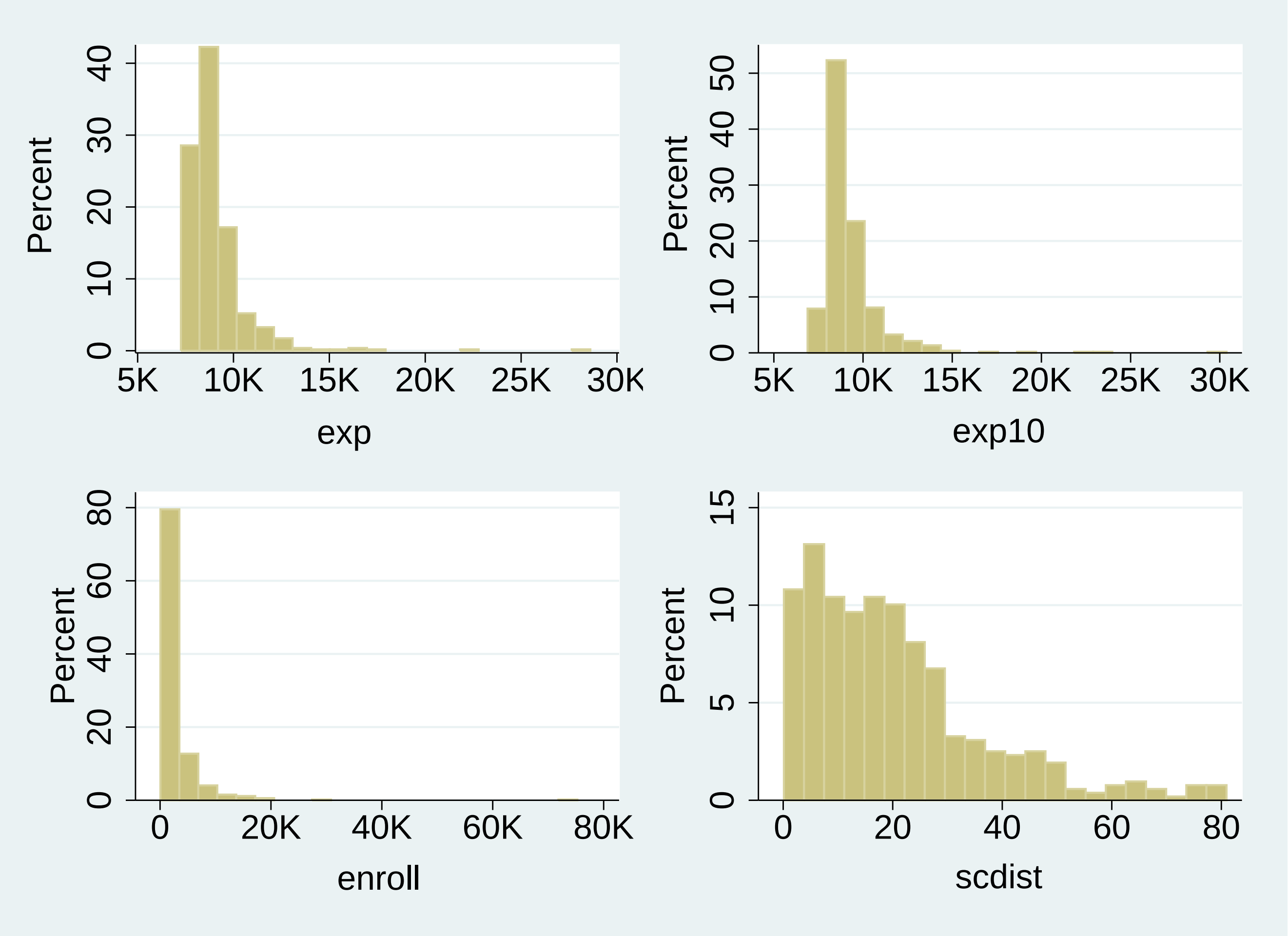

Figure C.9: Histogram for all covariates

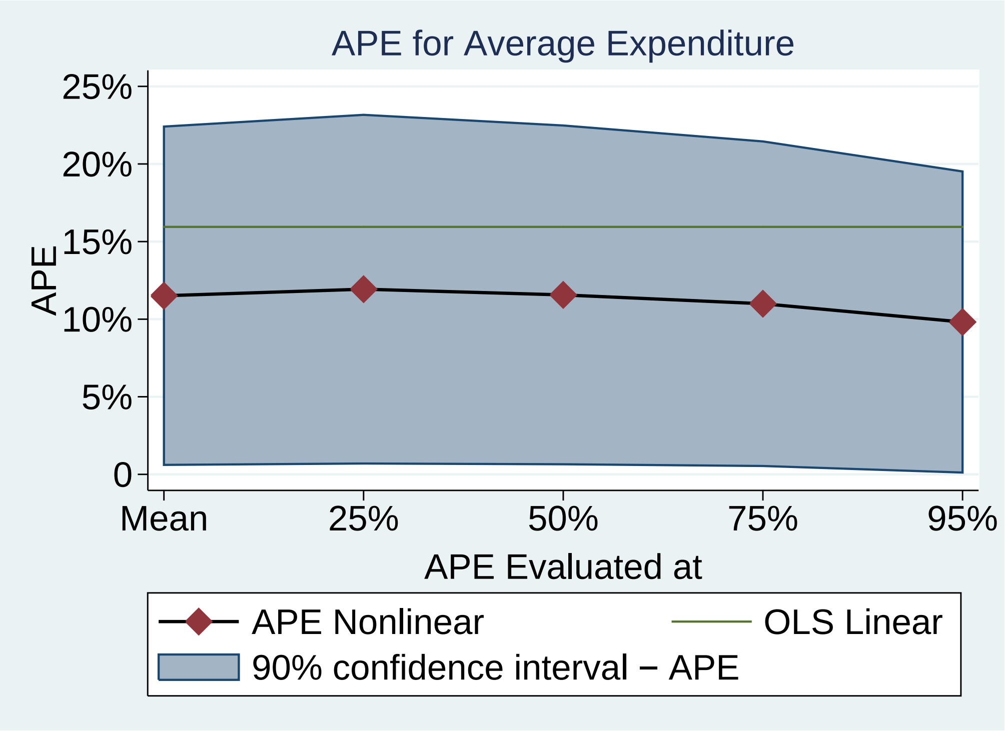

Figure C.10: APE w.r.t average expenditure

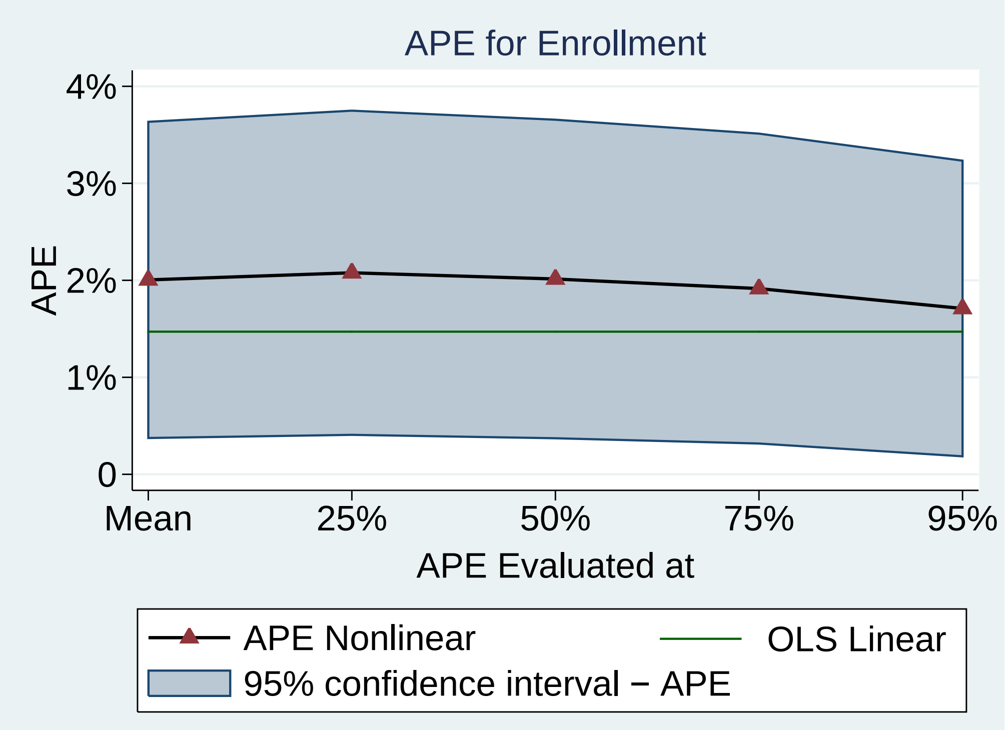

Figure C.11: APE w.r.t enroll

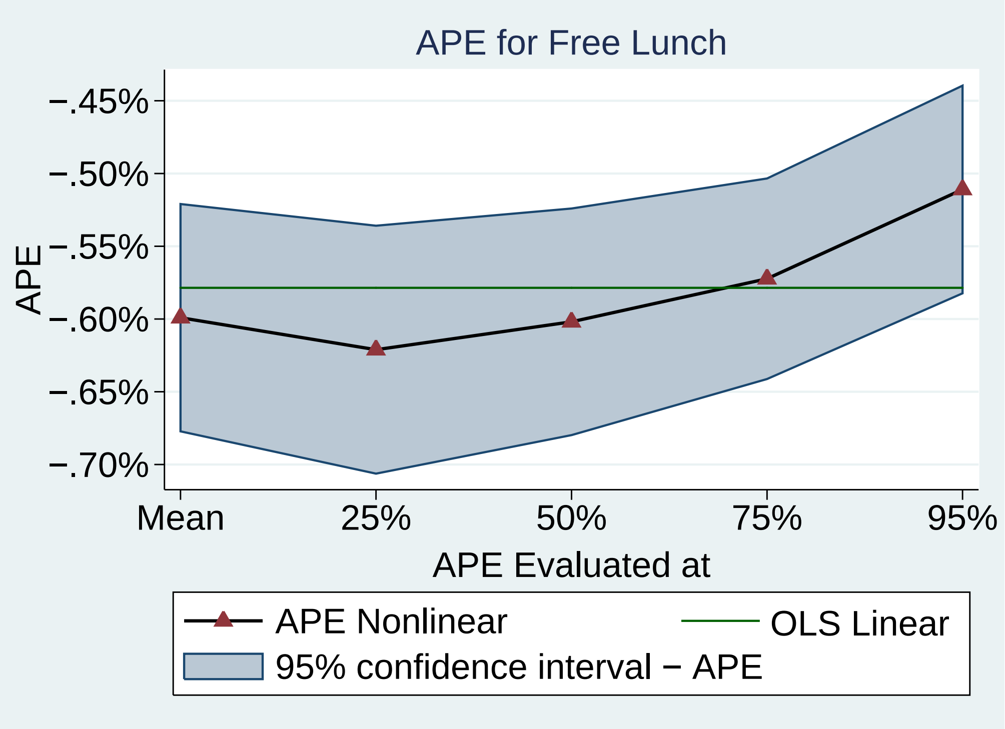

Figure C.12: APE w.r.t lunch

Table C.13: QGLS with Conley S.E., dependent variable=math4

		log(avgexp)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.1690	0.0171	-0.00589	0.00255	-1.0588
Usual	Std. Err.	0.0416*	0.00705*	0.00038*	0.00707	0.3862*
H-W	Std. Err.	0.0577*	0.00778*	0.00038*	0.00712	0.5137*
Conley	Std. Err.					
cut1	cut2					
50	100	0.0621*	0.0083	0.00041*	0.0083	0.5669
100	150	0.0634*	0.0079	0.00040*	0.0083	0.5739
150	200	0.0626*	0.0070	0.00038*	0.0079	0.5615
200	250	0.0615*	0.0065	0.00033*	0.0079	0.5550
350	400	0.0596*	0.0058*	0.00027*	0.0077	0.5469
400	500	0.0560*	0.0056*	0.00026*	0.0067	0.5131*
500	600	0.0534*	0.0053*	0.00024*	0.0063	0.4907*
600	700	0.0502*	0.0048*	0.00022*	0.0059	0.4614*
700	800	0.0479*	0.0044*	0.00021*	0.0057	0.4405*
800	900	0.0457*	0.0041*	0.0002*	0.0055	0.4205*
1000	1000	0.0433*	0.0037*	0.00018*	0.0054	0.4007*

Table C.14: QGLS with Conley S.E., dependent variable=math4 in Year 2010

		log(exp10)	log(enroll)	lunch	log(scdist)	constant
Coef.	Estimate	0.13496	0.0169	-0.00585	0.00142	-0.74907
Usual	Std. Err.	0.03828	0.00708	0.00039	0.00708	0.35662
H-W	Std. Err.	0.05759	0.00785	0.00039	0.00709	0.51117
Conley	Std. Err.					
cut1	cut2					
50	100	0.0616*	0.0083*	0.00043*	0.0083	0.5580
100	150	0.0638*	0.0080*	0.00043*	0.0084	0.5729
150	200	0.0639*	0.0071*	0.0004*	0.0080	0.5696
200	250	0.0631*	0.0066*	0.00035*	0.0080	0.5648
350	400	0.0584*	0.0056*	0.00031*	0.0074	0.5295
400	500	0.0556*	0.0057*	0.00029*	0.0066	0.5010
500	600	0.0522*	0.0054*	0.00026*	0.0062	0.4724
600	700	0.0486*	0.0048*	0.00024*	0.0059	0.4395
700	800	0.0458*	0.0044*	0.00023*	0.0057	0.4149
800	900	0.0433*	0.0041*	0.00021*	0.0055	0.3929
1000	1000	0.0406*	0.0037*	0.0002*	0.0053	0.3694*

Table C.15: SAR GLS, dependent variable=math4

Weight Matrix	Contiguity Weight		Inverse Dist Weight	
GLS Estimates	Coef.	Std. Err.	Coef.	Std. Err.
log(avgexp)	0.1602579	0.0420131*	0.1690079	0.0426937*
log(enroll)	0.015429	0.0069654*	0.0170857	0.0070499*
lunch	-0.005761	0.0004002*	-0.0058864	0.0003816*
log(scdist)	0.0034104	0.0072557	0.00255	0.0070406
constant	-0.9785768	0.3874754	-1.058793	0.3950128
ρ	0.2677935	0.0590989*	0.8456347	0.1394274*
σ^2	0.0133044	0.0008314*	0.013609	0.0008473*

Table C.16: SAR GLS, dependent variable=math4 in Year 2010

Weight Matrix	Contiguity Weight		Inverse Dist Weight	
GLS Estimates	Coef.	Std. Err.	Coef.	Std. Err.
log(exp10)	0.1245523	0.0387077*	0.1349613	0.0399863*
log(enroll)	0.0153031	0.0069908*	0.0168974	0.0070563*
lunch	-0.0057036	0.0004028*	-0.005847	0.0003847*
log(scdist)	0.0023392	0.0072584	0.0014194	0.0070466
constant	-0.6548259	0.3581062	-0.7490704	0.3715918
ρ	0.2638944	0.0592924*	0.8401031	0.1468289*
σ^2	0.0134163	0.0008383*	0.0137166	0.000854*

Table C.17: SAR GLS, dependent variable=math4, contiguity

math4	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
lexp	0.1602579	0.0420131	3.81	0	0.0779136	0.2426021
lenroll	0.015429	0.0069654	2.22	0.027	0.001777	0.029081
lunch	-0.005761	0.0004002	-14.4	0	-0.0065453	-0.0049766
lscdist	0.0034104	0.0072557	0.47	0.638	-0.0108106	0.0176313
cons	-0.9785768	0.3874754	-2.53	0.012	-1.738015	-0.219139
rho	0.2677935	0.0590989	4.53	0	0.1519618	0.3836253
sigma2	0.0133044	0.0008314	16	0	0.0116749	0.014934

Table C.18: SAR GLS, dependent variable=math4, inverse distance

math4	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
lexp	0.1690079	0.0426937	3.96	0	0.0853298	0.252686
lenroll	0.0170857	0.0070499	2.42	0.015	0.0032681	0.0309033
lunch	-0.0058864	0.0003816	-15.43	0	-0.0066343	-0.0051385
lscdist	0.00255	0.0070406	0.36	0.717	-0.0112493	0.0163492
cons	-1.058793	0.3950125	-2.68	0.007	-1.833003	-0.2845826
rho	0.8456345	0.1394278	6.07	0	0.572361	1.118908
sigma2	0.013609	0.0008473	16.06	0	0.0119483	0.0152696

Table C.19: Summary of Correlation

Overall Max		1		
Overall Min		3.33E-38		
Overall mean		0.00346437		
Overall variance		0.002071544		
Row mean				
	Percentiles	Smallest		
1%	0.0030197	2.85E-03		
5%	0.0031132	0.0028733		
10%	0.0031967	0.0029348	Obs	518
25%	0.0033227	0.0029758	Sum of Wgt.	518
50%	0.0034535		Mean	0.0034644
		Largest	Std. Dev.	0.0002449
75%	0.0035881	0.004259		
90%	0.0037387	0.0042819	Variance	6.00E-08
95%	0.0038485	0.0043556	Skewness	1.967204
99%	0.004069	0.0057957	Kurtosis	18.49351
Row min				
	Percentiles	Smallest		
1%	5.22E-38	3.33E-38		
5%	1.34E-36	3.33E-38		
10%	9.47E-36	4.15E-38	Obs	518
25%	9.24E-34	5.04E-38	Sum of Wgt.	518
50%	1.87E-29		Mean	1.35E-21
		Largest	Std. Dev.	1.11E-20
75%	1.49E-25	4.02E-20		
90%	8.06E-23	4.70E-20	Variance	1.24E-40
95%	1.47E-21	1.34E-19	Skewness	13.54964
99%	3.52E-20	1.92E-19	Kurtosis	209.841
Row variance				
	Percentiles	Smallest		
1%	0.0020325	2.03E-03		
5%	0.0020376	0.0020309		
10%	0.0020417	0.0020312	Obs	518
25%	0.002049	0.0020322	Sum of Wgt.	518
50%	0.002062		Mean	0.0020755
		Largest	Std. Dev.	0.0000389
75%	0.0020927	0.0022539		
90%	0.0021256	0.002289	Variance	1.51E-09
95%	0.0021453	0.0022999	Skewness	2.174566
99%	0.0022243	0.0023159	Kurtosis	10.42997

Table C.20: Summary of Correlation(div)

Overall Max(2nd largest)		1(.266828657)		
Overall Min		0.062464033		
Overall mean		0.079143268		
Overall variance		0.001712584		
Row mean				
	Percentiles	Smallest		
1%	0.0694387	0.0692969		
5%	0.0710834	0.0693381		
10%	0.0731629	0.0693386	Obs	518
25%	0.0765987	0.0693432	Sum of Wgt.	518
50%	0.0794727		Mean	0.0791433
		Largest	Std. Dev.	0.0042475
75%	0.0817451	0.0879126		
90%	0.0848623	0.0879339	Variance	0.000018
95%	0.0865209	0.0881153	Skewness	-0.2346403
99%	0.087731	0.0881707	Kurtosis	2.719421
Row min				
	Percentiles	Smallest		
1%	0.0625941	0.062464		
5%	0.06363	0.062464		
10%	0.063966	0.0624661	Obs	518
25%	0.0647366	0.0625628	Sum of Wgt.	518
50%	0.0656149		Mean	0.0658829
		Largest	Std. Dev.	0.0016574
75%	0.0668986	0.070315		
90%	0.0682221	0.0703266	Variance	2.75E-06
95%	0.0692942	0.070412	Skewness	0.5907457
99%	0.0701836	0.0704638	Kurtosis	2.952739
Row variance				
	Percentiles	Smallest		
1%	0.0016646	0.0016636		
5%	0.0016661	0.0016644		
10%	0.0016686	0.0016644	Obs	518
25%	0.0016759	0.0016645	Sum of Wgt.	518
50%	0.0016843		Mean	0.0016978
		Largest	Std. Dev.	0.0000361
75%	0.0017045	0.0018191		
90%	0.0017555	0.0018207	Variance	1.30E-09
95%	0.0017896	0.0018216	Skewness	1.809396
99%	0.0018151	0.0018218	Kurtosis	5.520294

BIBLIOGRAPHY

BIBLIOGRAPHY

Abrevaya, J., (2002). Computing marginal effects in the Box-Cox model, *Econometric Reviews* 21, 383-393.

Ackerberg, D., X. Chen, and J. Hahn (2012). A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators.*Review of Economics and Statistics*94, 481-498.

Ai, C., Chen, X., (2003). Efficient estimation of conditional moment restrictions models containing unknown functions. *Econometrica* 71, 1795-1843.

Ai, C., Norton, E. C., (2000). Standard errors for the retransformation problem with heteroscedasticity. *Journal of Health Economics*19, 697-718.

Ai, C., Norton, E. C., (2008). A semiparametric derivative estimator in log transformation models. *Econometrics Journal* 2, 538-553.

Altonji, J. G. and Martskin, R., (2005). Cross Section and Panel Data Estimators for Nonseperable Models with Endogenous Regressors, *Econometrica* 73, 1053-1102.

Anselin, L., Florax, R. (Eds.), 1995. New directions in Spatial Econometrics. Springer, Berlin.

Anselin, L., (2010). Thirty Years of Spatial Econometrics. *Papers in Regional Science* 89(1), 3-25.

Arraiz, I., Drukker, D.M., Kelejian, H. H., Prucha, I.R., (2010). A Spatial Cliff-Ord-Type Model with Heteroskedastic Innoviations: Small and Large Sample Rerults. *Journal of Regional Science* 50(2), 592-614.

Bajari, P., Chernozhukov, V., Hong, H. and Nekipelov, D., (2009). Identification and Efficient Semiparametric Estimation of aDynamic Discrete Game, working paper.

Banerjee, S., Carlin, B.P., Gelfand A.E., (2004). Hierarchical Modeling and Analysis for spatial data. Chapman and Hall/CRC Press, Boca Raton.

Berndt, E., Showalter, M., Wooldridge, J.M., (1993). An Empirical Investigation of the Box-Cox Model and a Nonlinear Least Squares Alternative, *Econometric Reviews* 12, 65-102.

Blackburn, M. L., (2007). Estimating Wage Differentials without Logarithms, *Labour Economics* 14, 73-98.

Blundell, R., Powell, J., (2003). Endogeneity in Nonparametric and Semiparametric Regression Models,in M. Dewatripont, L. P. Hansen and S. J. Turnsovsky (eds.) *Advances in Economics and Econometrics*, 312-357.

Blundell, R., Powell, J., (2004). Endogeneity in Semiparametric Binary Response Models, *Review of Economic Studies*, 6559.

Case, A., (1991). Spatial Patterns in Household Demand, *Econometrica* 59, 953-965.

Chakrabarti, R., Roy, J., (2012). Housing Markets and Residential Segregation Impacts of the Michigan School Finance Reform on Inter- and Intra-District Sorting, *Federal Reserve Bank of New York Staff Reports*, no.565.

Chamberlain, G., (1980). Analysis with qualitative data, *Review of Economic Studies* 47, 225-238.

Chamberlain, G., (1982). Multivariate regression models for panel data, *Journal of Econometrics* 18 5-46.

Chamberlain, G., (1984). Panel Data. in Handbook of Econometrics, Volume 2, ed. Z. Griliches and M. D. Intriligator. Amsterdam: North Holland, 1247-1318.

Chen, X., (2007). Large Sample Sieve Estimation of Semi-nonparametric Models, Heckman, James. J. and Leamer, Edward E., eds. *Handbook of Econometrics*, Vol. 6B, Chapter 76, North-Holland.

Conley, T. G., (1999). GMM estimation with cross sectional dependence, *Journal of Econometrics*92, 1-45.

Conley, T. G., Ligon, E.A., (2002). Economic distance, spillovers, and cross country comparisons, *Journal of Economic Growth* 7, 157-187.

Conley, T. G., Topa, G., (2002). Socio-economic distance and spatial patterns in unemployment, *Journal of Applied Econometrics* 17(4), 303-327.

Conley, T. G., Dupor, B., (2003). A spatial analysis of sectoral complementarity, *Journal of Political Economy* 111(2), 311-352.

Conley T. G., Molinari, F., (2007). Spatial correlation robust inference with errors in location or distance, *Journal of Econometrics*140 76-96

Duan, N.(1983). Smearing Estimate: A Nonparametric Restransformation Method, *Journal of American Statistical Association* 78, 605-610.

Engel, C., Rogers, J. H., (1996). How wide is the border? *The American Economic Review* 86 (5), 1112-1125.

Ferguson, T. S., (1996). A Course in Large Sample Theory, first edition. Chapman & Hall Press.

Hausman, J.A., Hall, B.H. and Griliches, Z., (1984). Econometric models for count data with an application to the patents-R&D relationship, *Econometrica* 52, 909-938.

Heckman, J. J., (2001). Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture, *Journal of Political Economy* 109, 673-748.

Hirano, K., Imbens, G. W. and Ridder, G., (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica* 71, 1161-1189.

Huber, P. J., (1967). The behavior of maximum likelihood estimates under non-standard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA.

Imberman, S. A., (2011). The Effect of Charter Schools on Achievement and Behavior of Public School Students. *Journal of Public Economics*, 95(7-8), 850-863.

Kane, T., Rouse, C. E., (1995). Labor Market Returns to Two- and Four-year College. *American Economic Review*85(3), 600-614.

Kelejian, H. H., Prucha, I.R., (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40 (2), 509-533.

Kelejian, H. H., Prucha, I.R., (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances, *Journal of Econometrics* 157, 53-67.

Keller, W., Shiue, C., (2007). The origin of spatial interaction, *Journal of Econometrics*140, 304-332

Lee, L., (2004). Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models, *Econometrica*72, 1899-1926.

Lin, X., (2010). Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables, *Journal of Labor Economics* 28(4), 825-860.

Lin, X., Lee, L.F., (2010). GMM estimation of spatial autoregressive models with unknown heteroskedasticity, *Journal of Econometrics* 157, 34-52.

Li, Q., Racine, J. S., (2007). Nonparametric Econoemtrics: Theory and Practice , Princeton and Oxford, Princeton University Press.

Mall, M. M., (2004). A Close Look at the Spatial Structure Implied by the CAR and SAR Models, *Journal of Statistical Planning and Inference* 145(1),121-133.

Manning, W. G., (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem, *Journal of Health Economics* 17(3), 283-295.

Mundlak, Y., (1978). On the pooling of time series and cross section data, *Econometrica*46, 69-85.

Mullahy, J., (1998). Much ado about two reconsidering retransformation and the two-part model in health econometrics, *Journal of Health Economics* 17, 247-81.

Newey, W. K., (1993). Efficient Estimation of Models with Conditional Moment Restrictions, G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics*, Volume 11: Econometrics.

Newey, W. K., (1994). Series estimation of regression functionals, *Econometric Theory* 10, 1-28.

Newey, W. K., (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics* 79, 147-68.

Newey, W. K., McFadden, D., (1994). Large Sample Estimation and Hypothesis Testing, in R.F. Engle and D. McFadden (eds.), Handbook of Econometrics, Volume 4. Amsterdam: North Holland, 2111-2245.

Papke, L. E., (2005). The effects of spending on test pass rates: Evidence from Michigan, *Journal of Public Economics* 89(5-6), 821-39.

Papke, L. E., (2008). The effects of changes in Michigan's school finance system, *Public Finance Review* 36(4), 456-74.

Papke, L.E., Wooldridge, J. M., (1996). Econometric Methods for Fractional Response variables with an Application to 401(K) Plan Participation rates, *Journal of Applied Econometrics* 11(1),619-32.

Papke, L. E., Wooldridge, J. M., (2008). Panel date methods for fractional response variables with an application to test pass rates, *Journal of Econometrics* 121,311-24.

Park, U.B., Sickles, R. C., and Simar, L., (2007). Semiparametric efficient estimation of dynamic panel data models, *Journal of Econometrics* 136, 281-301.

Roy, J., (2011). Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan, *Education Finance and Policy* 6(2), 137-167.

Simcoe, T., (2008). XTPQML: Stata module to estimate Fixed-effects Poisson (Quasi-ML) regression with robust standard errors. Repec online paper.

Wang, H., Iglesias, E. and Wooldridge, J. M., (2013). Partial Maximum Likelihood Estimation of a Spatial Bivariate Probit Model, *Journal of Econometrics*172(1), 77-89.

White, H., (1980). A heteroskedasticity-consistent covariance estimator and a direct test for heteroskedasticity, *Econometrica*48, 817-830.

Wooldridge, J. M. (1992a). A test for functional form against nonparametric alternatives, *Econometric Theory* 8, 452-475.

Wooldridge, J. M. (1992b). Some Alternatives to the Box-Cox Regression Model, *International Economic Review* 33, 935-955.

Wooldridge, J. M., (1997). Multiplicative Panel Data Models Without the Strict Erogeneity Assumption, *Econometric Theory* 13, 667-678.

Wooldridge, J. M., (1999). Distribution-free estimation of some nonlinear panel data models, *Journal of Econometrics* 90(1), 77-97.

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

Wooldridge, J. M., (2004). Estimating average partial effects under conditional moment independence assumptions , *CeMMAP working papers CWP03/04*, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Wooldridge, J. M., (2005). Unobserved heterogeneity and estimation of average partial effects. In: Andrews, D.W.K., Stock, J.H. (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge University Press, Cambridge, 27-55.

Wooldridge, J. M., (2009). Introductory Econometrics: A Modern Approach, fourth edition. Cincinnati, OH: South-Western College Publishing.

Wooldridge, J. M., (2010). Econometric Analysis of Cross Section and Panel Data, second edition. Cambridge, MA: MIT Press.

Wooldridge, J. M., (2011). Solutions Manual and Supplementary Materials for Econometric Analysis of Cross Section and Panel Data, second edition. Cambridge, MA: MIT Press.