THREE ESSAYS IN COMPLEX SAMPLES

by

Iraj Rahmani

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

ECONOMICS

2012

ABSTRACT

THREE ESSAYS IN COMPLEX SAMPLES

by

Iraj Rahmani

The samples used in econometric studies are not always sets of randomly drawn observations from the populations of interest. In many studies sampling has a complex design involving clustering and stratification. In stratification, the population is divided into subpopulations or strata based on exogenous or endogenous variables and then a random sample of unit observations or clusters is drawn from each stratum. Clusters are contiguous groups of units existing within a stratum. Reducing the cost of sampling or operational convenience might be reasons for applying stratification and clustering. On the other hand, particular interest in a small subpopulation may cause oversampling that justifies non-random sampling scheme.

This dissertation consists of three essays addressing estimation and inference in cross section and panel data models with non-random samples. In general, ignoring sampling design could produce inconsistent estimators and also inconsistent estimators for their standard errors. In the first essay, a multi-stage sampling design including standard stratification and clustering stages at first and variable probability sampling in the final stage is considered. The problem is studied under M-estimators framework. Under a set of regularity conditions, the usual weighting estimators are consistent and have asymptotic normal distributions. In cases that stratifications in the first or the second or in the both stages are exogenous, dropping the corresponding weights are allowed; we still have consistent estimators.

The second essay contributes to the subject of non-random sampling by studying efficiency in panel data models when data set comes from stratified samples. The goal in this chapter is to obtain more efficient estimators by considering correlation within panels in models with stratified structure. We do not try to find the efficiency bound in this kind of models. Our attempt is to increase efficiency in compare with pooled models that ignore correlations within panels. The

paper takes into account correlation within each panel and in each stratum under a GMM based framework. Theoretical development and Monte Carlo study show that by considering correlation within the panels in each stratum and adding them together with appropriate weights, finding more efficient estimators is possible. Like generalized estimating equations (GEE), we are able to consider the specific form for correlation for panels in each stratum. Monte Carlo results confirm that the new GMM estimators that is called weighted and unweighted GLS are more efficient than their competitors OLS and weighted OLS that simply overlook the correlation within the panels. In case of endogenous stratification, weighted GLS and in case of exogenous stratification unweighted GLS is doing better than the rest. For a specific sample size, this efficiency gain depends on what form is chosen for correlation and how strong or weak it is. We applied results to study determinants of inequality in the U.S. and estimation results show that efficiency gain in compare with POLS or weighted POLS is substantial.

The subject of the third essay is model selection problem. In complex samples involving stratification and clustering, the assumption that observations are distributed independently and identically is not held anymore and therefore the Vuong's (1989) model selection tests are not applicable directly. In order to generalize Vuong's results to estimators other than MLE, we study the problem under M- estimator framework that contains many estimators including but not limited to linear and non-linear least squares, MLE, and QMLE. The theoretical results show that for two nonnested competing models, the asymptotic property of the *weighted* tests statistics are not a function of the competing estimators but observations and has normal distribution. An interesting finding is that even in case of exogenous stratification, we cannot drop weights in the tests statistics since for nonnested tests both competing models should be misspecified under the null. We also apply results in two empirical studies.

To my late father,
my mother,
and my brothers, Behzad, and Reza, and my sister, Maryam

ACKNOWLEDGEMENTS

It would not have been possible to write this doctoral thesis without the help and support of many great people around me, to only some of whom it is possible to give particular mention here.

It is difficult to overstate my gratitude to my Ph.D. supervisor, Professor Jeffery Wooldridge. With his enthusiasm, his inspiration, and his unsurpassed knowledge of econometrics and statistics, he helped me to move in a right direction. I would have been lost without him. Throughout my thesis-writing period, he provided encouragement, sound advice, good teaching and many good ideas.

I would like to thank Professors Peter Schmidt, and Tim Vogelsang in the Department of Economics, and Tapabrata Maiti in the Department of Statistic and Probability for their encouragements and supports. It was great honor to have these great scholars as my committee members. I wish to thank Emma Iglesias, who was a member of my committee, for her help and support.

Also I would like to thank Professor Hassan Mohammadi at Illinois State University and Professor Kambiz H. Kiani at Shahid Beheshti Universty (the National University of Iran) who encourage me to continue my education in Iran and the U.S., and for their kind assistance with writing letters and wise advice.

My friendship with other Ph.D. students was very fruitful and I learnt many great lessons from them. I would like to thank all of them, particularly Dr. Do Won Kwak whom I learnt many things in the Stata Programming.

I received many helps from former and current Graduate Secretaries, Jennifer Carducci, and Lori Jean Nichols, and also Margaret Lynch, the Office Manager, the Department of Economics and would like to thank them for kindness and sincere assistance.

I owe thanks as well to Leila Ardestani, for the continuous support and encouragement that I received from her.

I wish to thank my wonderful and very kind brothers and sister for giving me their uncondi-

tional love and support. My brothers Behzad and Reza and my sister Maryam who have always been fountains of sincere friendship and love. I am so thankful for having them in all stages of my life. Without their help and support, I could not stay so many years far from home without worrying about any issue.

Lastly, and most importantly, I wish to thank my father Rahman Rahmani and my wonderful mother Roohangiz Saadati. They were my first teachers who thought me the most important elements of life; love, friendship, and forgiveness. It was a great unfortunate experience of losing my father almost at the end of my first year in the Ph.D. program. The pain is still fresh and his place in my heart will never be filled. He was the pillar of my life, the best friend, and a source of wisdom and great advice. In his absence, my dear mother did her best to keep me on the road sound and firm. I cannot find the suitable words to express my gratitude to her. To my parents and brothers and sister, I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

ASYMPTOTIC INFERENCE OF M-ESTIMATOR FORM MULTI-STAGE SAMPLES WITH VARIABLE PROBABILITY IN FINAL STAGE

1.1 Introduction

In economic analyses are often assumes that the observations come from simple random sampling. It means that a set of independent and identically distributed (i.i.d.) observations is available. However in reality many data sets used in economics and other branches of the social sciences are come form stratified sampling schemes that produces nonrandom samples. When a data set comes from nonrandom sampling schemes, i.i.d. assumption is not valid anymore and therefore one might need to make inference about the econometric model more carefully.

The goal in this chapter is to examine asymptotic properties of M-estimators when the observations come from several levels of stratifications.

Three well known stratified sampling schemes are standard stratified sampling (SS sampling), multinomial sampling, and variable probability sampling (VP sampling). In SS sampling, the population is divided into several subpopulation based on factors like income, race, gender, education, area of residence, etc. Then a random sample is taken within each subpopulation or stratum independently. The result is a sample of independent but not identically distributed observations. It should be emphasized that unlike simple random sampling, in SS sampling the proportions of observations within strata do not reflect population proportions as they would if the sample were selected randomly from the population.

Multinomial sampling scheme is similar to SS sampling. The difference is that in multinomial sampling first a stratum is chosen randomly and then samples randomly from the stratum. Although this kind of sampling is not common in practice but theoretically it is easier to deal with because it produces i.i.d. observations.

In VP sampling which also is known as Bernoulli sampling in the literature, first an observation is drawn randomly from the population, then its stratum is determined by the researcher. After determining its stratum, it will be kept in the sample with specific probability that is set by the researcher also. If the observation is not chosen then it will be returned to the population and its values are not recorded.

SS sampling scheme is often used when observations from each stratum are easily identified before sampling. Variable probability sampling scheme is more suitable when stratum of an observation is known only after sampling. For example determining a family's income bracket is difficult before sampling and therefore VP sampling is used.

In general, stratification can be based on dependent variable or variables, explanatory variables or both. Dividing the population of interest in terms of explanatory variables is called exogenous stratification. Stratification is endogenous if we define subpopulations with respect to dependent variables. Whether stratification is exogenous, or endogenous is determined only after defining an econometric model. In other words, determining a specific model comes first and then discussions about appropriate sampling schemes start.

Reviewing the literature shows that the subject have been studied by both statisticians and econometricians. In summery, stratification based on exogenous variables does not produce serious problems; one can ignore it and still obtains consistent estimates for parameters of the population. In this line of research we can mention DuMouchel and Duncan (1983) that confirms the above statement in a linear model for SS sampling. Manski and McFadden (1981) show it is true in maximum likelihood estimation where data set comes form multinomial sampling. Wooldridge (1999, 2001) shows same result is true for VP, and SS sampling when we consider the case in framework of M-estimators.

In practice combination of these methods of sampling are commonly used also. For instance the Panel Study of Income Dynamics (PSID) involves stratification and clustering. Bhattacharya (2005) describes a multi-stage sampling in which SS sampling is used in first level to choose some clusters in each stratum by simple random sampling, and then from each sampled cluster a few observations are chosen again by simple random sampling. In this scheme clusters are defined as contiguous groups of units existing within a stratum. For example in rural areas villages can be considered as clusters, and in urban areas, they are blocks or neighborhoods and in both examples unit observations are households.

In his paper, Bhattacharya (2005) drives asymptotic properties of estimators when data set comes from surveys whose designs involve stratification and clustering in GMM framework. In a set up similar to Bhattachary's multistage sampling, Wooldridge (2008) drives asymptotic variance of estimators in linear models.

The goal in this chapter, as mentioned already, is to investigate asymptotic properties of estimators when data set comes from multi-stage sampling. It is closely related to Bhattacharya (2005) sampling scheme with one distinction. We add variable probability sampling in final stage and then develop M-estimator framework for asymptotic inference to evaluate data from surveys with multi level of stratification and clustering structure. The set up is general enough to contain linear and non-linear models as well as maximum likelihood ones. This kind of sampling design is used in many surveys in practice, particularly those that involve phone interviews.

As an example of big scale survey that has a structure very similar to the sampling scheme considered in this study, we can name National Survey of Families and Households (NSFH) .The NSFH is a complex survey sample that involved five sampling stages. In the first stage of this national multistage sampling design, 100 primary sampling units were drawn from a list of all countries in the nation that had been stratified into two groups. In first stratum, 18 self -representing areas composed of the largest metropolitan areas that make up 36 % of the nation's population and second stratum contains the rest of the country. From the the first stratum, 36 primary sampling units were drawn with certainty. The second stratum that make up 64 % of the nation was divided into 32 strata, and two primary sampling units were drawn from each stratum using probability proportional to size sampling.

In the second stage an average of 17 block groups or enumeration districts from each primary sampling unit is randomly selected. Within each of these district, a list of 45 or more households was selected. These households were given a short screening interview to allow oversampling of certain interested groups like African American, cohabiting couples etc. Members of these groups in the cluster were selected with certainty, and others were selected at a lower rate. In the final stage, an adult from each household was randomly chosen as the eligible respondent. At the end from 45 or more households in each district or cluster 20 of them were included in the sample. Substitutions were not allowed. in this study, the sample size was 13007 primary respondents. The survey contains 1700 clusters, with an average of 7.6 respondents per cluster. In this study we have many clusters with small size. For more detail see Johnson and Elliott (1998).

The rest of paper is organized as follows. The next section presents the population optimization problem and basic framework. Sampling scheme and sample objective function are explained in section 3. In section 4 consistency and asymptotic normality of weighted M-estimator under multi-stage sampling is discussed. We introduce theories that summarize conditions needed to have consistent weighted estimators with asymptotic normal distribution. Also in section 4 we study estimating of asymptotic variances of M-estimators. In section 5, estimation under exogenous stratification is discussed. Under exogenous stratification in our model where more than one level of stratification exist, three cases are distinguishable. However we only consider the two first cases. In section 5, four theorems are presented to cover consistency and asymptotic distribution of M-estimators under exogenous stratification. In section 6, four examples are presented. In section 7, two-step M-estimator is discussed. In section 8, the last section, the main findings of the paper are reviewed.

1.2 The Population Optimization Problem

Our goal is to estimate a $P \times 1$ vector of parameter θ that minimize the population problem

$$\min_{\theta \in \Theta} \mathbb{E}\left[q\left(\mathbf{W}, \theta\right)\right] \tag{1.1}$$

where $\mathbb{E}[.]$ denotes the expectation with respect to the true distribution of $\mathbf{W}$, and $\theta \in \Theta$ and Θ is the parameter space that is a subset of Euclidean space $\mathbb{R}^P$. The objective function in the population is denoted as $q(\mathbf{W}, \theta)$ that is a function of $\mathbf{W}$ and θ. $\mathbf{W}$ is an $M \times 1$ random vector taking values in $\mathscr{W}$, where $\mathscr{W}$ is a subset of $\mathbb{R}^M$. We assume that there exists a unique solution $\theta_\circ \in \Theta$, that minimize population problem (1.1).

In cases that $q(.)$ is a correctly specified model, $\theta_\circ$ is the true parameter that uniquely minimize (1.1). However, in misspecified cases where $q(.)$ is not a correct model, there is no true value of θ, i.e. $\theta_\circ$. In these cases, it is standard to assume $\theta_\circ$ is the unique solution to (1.1).

We are usually interested in explaining a $K \times 1$ random vector $\mathbf{Y}$ conditional on a $L \times 1$ vector of explanatory variable $\mathbf{X}$ such as $\mathbb{E}(\mathbf{Y}|\mathbf{X})$. Here $K + L = M$, and $(\mathbf{X}, \mathbf{Y}) = \mathbf{W}$. Random vectors $\mathbf{X}$ and $\mathbf{Y}$ belong to subsets $\mathscr{X}$ and $\mathscr{Y}$ respectively, where $\mathscr{X} \subset \mathbb{R}^K$, $\mathscr{Y} \subset \mathbb{R}^L$ and union of $\mathscr{X}$ and $\mathscr{Y}$, denoted by $\mathscr{X} \cup \mathscr{Y}$ is $\mathscr{W}$. The framework is general enough to cover panel data models with large cross section dimension and small time periods T.

1.3 The Sampling Scheme

The sample design is a combination of standard stratification, clustering and variable probability sampling. First, according with SS sampling, the population is divided into S first stage strata that are non-overlapping and exhaustive. In this stage, stratification can be based on a variable or variables like the area of residence or race that allows us to divide the population easily. Each stratum s contains a mass of C_s clusters. For example these clusters in rural areas are villages, and in urban areas, they are blocks or neighborhoods. In next step N_s clusters with replacement are drawn randomly from each stratum s. Since in this study we require some sort of large-sample approximation, the assumption of with replacement is not important if the number of clusters samples, N_s, is "large". Each sampled cluster c from stratum s contains a finite population of M_{sc} households or units of observations. An observation (household) is selected by random from sampled cluster c in stratum s. In next stage the selected household is classified according to interested non-overlapping and exhaustive strata based on, for example, the level of income. The household is retained into the sample with some probability that is set by the practitioner. As it mentioned already, sampling in the second stage is called variable probability. The process is repeated for K (a constant and small number) of unit observations for each sampled cluster c in stratum s and a sample of K_{sc} households is obtained where $1 \leqslant K_{sc} \leqslant K$.

In practice a fixed and large number of clusters N_s are sampled randomly within each stratum s, and then within each sampled cluster, a small and fixed number of households are sampled randomly.

We can summarize sampling design as follows

i The population is divided into S non-overlapping and exhaustive first stage strata based on criteria like area of residence, race, age etc.

ii In stratum s, C_s clusters exist.

iii For each stratum s randomly draw N_s clusters with replcement.

iv Each sampled cluster c_s from stratum s contains a finite population of M_{sc} units (for example households).

v A household is selected by random from sampled cluster c_s in stratum s.

vi The household is classified according to interested non-overlapping and exhaustive strata (for example income level).

vii The household is retained into the sample with some probability that depends on interested stratum and is determined by the researcher.

viii The process is followed for K household in each sampled cluster c_s in stratum s and a sample of K_{sc} households is obtained.

Considering structure of most surveys in practice and the same as Bhattacharya (2005), two assumptions are made to study the asymptotic inference of the model. First assumption is that the number of clusters N goes to infinity with numbers of household staying fixed and finite within each cluster. The second assumption is that the clusters are independent within a stratum but household level variables are correlated within each cluster. Therefore for a given stratum s, clusters are independently but not identically distributed.

Under sample scheme, clusters are chosen by simple random sample within each stratum s independently. In second step unit observations (households) are chosen by variable probability. Therefore the sample optimization problem is

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{j=1}^{J} \sum_{m=1}^{K} v_{sc} p_j^{-1} r_{jm} z_{jm} q\left(\mathbf{W}_{scm}, \theta\right) \tag{1.2}$$

Here $N = N_1 + N_2 + \ldots + N_j$ and $v_{sc} = \frac{C_s}{\left(\frac{N_s}{N}\right)} \cdot \frac{M_{sc}}{K_{sc}}$. In the sample problem (1.2), r is an indicator variable that takes value one if $\mathbf{W}$ is in stratum j and zero otherwise. z is also an indicator variable that takes value one if $\mathbf{W}$ is kept in the sample and zero if not and therefore $P(z = 1) = p$. In order to study asymptotic properties of M-estimator, we also assume that the ratio of sampled clusters in each stratum s to total sampled clusters N or $\frac{N_s}{N}$ is constant and therefore $\sum_{j=1}^{J} a_s = 1$. We need this assumption in order to limit the range of fluctuations of weights v_{sc}.

If we re-index clusters from $i = 1, \cdots, N$, and define new indicator variable y_{is} such that y_{is} equals one if cluster i is from stratum s or $i \in s$ and zero otherwise, then the optimum problem is

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{m=1}^{K} y_{is} v_{is} p_j^{-1} r_{jm} z_{jm} q\left(\mathbf{W}_{ism}, \theta\right) \tag{1.3}$$

The weighted M-estimator $\hat{\theta}_w$ minimizes (1.2) over the parameter space Θ. v_{sc} and p_j^{-1} are weights corresponding to first level (SS sampling), and second level of stratifications (VP sampling) respectively. The inner summation in (1.2) is over all potential observations, which would appear in a random sample. The sample objective function weights each sampled observation unit (households for example) by product of the two weights corresponding to two level of stratifications i.e. $v_{sc} \cdot p_j^{-1}$. Note that all sampled observations from a same sampled cluster get same weights.

1.4 Estimation under Multi-stage Sampling

1.4.1 Consistency

In order to study the consistency of the weighted M-estimator defined by equation (1.2) is assumed that the parameter vector $\theta_\circ$ uniquely solves the population problem (1.1)

$$\min_{\theta \in \Theta} \mathbb{E}\left[q\left(\mathbf{W}, \theta\right)\right]$$

Moreover we need to show that uniform convergence in probability is hold. It is assumed that function $q(\cdot)$ satisfies some regularity conditions. We summarize these requirements for consistency of the weighted M-estimator in following theorem.

Theorem 1.4.1. *Let $\mathbf{W} \in \mathscr{W}$ be a random vector where $\mathscr{W} \subset \mathbb{R}^M$, and $\Theta \subset \mathbb{R}^P$, and $q : \mathscr{W} \times \Theta \to \mathbb{R}$ a real valued function. if*

1. *Θ is a compact set.*
2. *$v_{sc} \cdot p_j^{-1} > 0$ for all clusters and strata $s = 1, \ldots, S$, $j = 1, \ldots, J$, $c = 1, \ldots, N$.*
3. *For each $\theta \in \Theta$, $q(\mathbf{W}, \theta)$ is Borel measurable on $\mathscr{W}$.*
4. *For each $\mathbf{w} \in \mathscr{W}$, $q(\mathbf{w}, \theta)$ is continuous on Θ.*
5. *$|q(\mathbf{w}, \theta)| \leq b(\mathbf{w})$, where b is an arbitrary nonnegative function on $\mathscr{W}$ such that $\mathbb{E}\left[b(\mathbf{w})\right] < \infty$.*
6. *$\theta_\circ$ uniquely solves the population problem.*

Then uniform weak law of large numbers holds, and $\hat{\theta}_w \xrightarrow{P} \theta_\circ$ as $N \longrightarrow \infty$.

Proof. For each cluster i in stratum s define

$$g(\mathbf{W}_m, \theta) = \sum_{j=1}^{J} \sum_{m=1}^{K} v p_j^{-1} r_{jm} z_{jm} q(\mathbf{W}_m, \theta) \tag{1.4}$$

In this function weight v is a random variable since the number of observations from the final stage K_{is} is random. In fact v is a function of z_{jm}, indicator variable that shows if the randomly drawn observation from final stage is kept in the sample or discarded. Therefore we can consider $v \cdot z_{jm}$ as an indicator variable

$$v \cdot z_{jm} = \begin{cases} v \quad \text{if} \qquad z_{jm} = 1, \\ 0 \qquad \text{otherwise} \end{cases}$$

And its probability distribution function is

$$f(v \cdot z_{jm}) = \begin{cases} p_j & \text{if } v \cdot z_{jm} = v, \\ 1 - p_j & \text{if } v \cdot z_{jm} = 0. \end{cases}$$

Therefore the expected value of (1.4) is

$$\mathbb{E}\left[g\left(\mathbf{W}_m, \theta\right)\right] = \sum_{j=1}^{J} \sum_{m=1}^{K} p_j^{-1} \mathbb{E}\left[v r_{jm} z_{jm} q(\mathbf{W}_m, \theta)\right] \tag{1.5}$$

Since $v \cdot z_{jm}$ is independent of r_{jm}, the right hand of (1.21) is equal to

$$= \sum_{j=1}^{J} \sum_{m=1}^{K} p_j^{-1} \mathbb{E}(v z_{jm}) \mathbb{E}\left[r_{jm} q(\mathbf{W_m}, \theta)\right] = \sum_{j=1}^{J} \sum_{m=1}^{K_{sc}} p_j^{-1} p_j v \mathbb{E}\left[r_{jm} q(\mathbf{W}_m, \theta)\right]$$

which can be simplified to

$$= \mathbb{E}\left[\sum_{m=1}^{K_{sc}} \sum_{j=1}^{J} v r_{jm} q(\mathbf{W}_m, \theta)\right] = \mathbb{E}\left[\sum_{m=1}^{K_{sc}} v q(\mathbf{W}_m, \theta)\right]$$

Last equality holds because$\sum_{j=1}^{J} r_{jm} = 1$. Therefore the expected value of (1.21) is equal to

$$M_{is} \cdot \mathbb{E}\left[q(\mathbf{W}, \theta)\right] \tag{1.6}$$

M_{is} is the population number of observation in cluster i and stratum s and hence it is constant and it does not effect estimation and inference. By assumption (6) of the Theorem(1.3.1) $\theta_{\circ}$ solves the population problem (1.1) uniquely and so is the unique solution for (1.6). Next we need to show that (1.4) satisfies the uniform law of large numbers for each stratum s. By assumption (3) of Theorem(1.3.1), $q(\cdot)$ is a continuous function on Θ for each $\mathbf{W} \in \mathscr{W}$, and therefore $g(\cdot)$ defined by (1.4) has same property. $g(\cdot)$ is bounded also, because $|g(\mathbf{W}, \theta)| = |\sum_{j=1}^{J} \sum_{m=1}^{K} p_j^{-1} r_{jm} z_{jm} q(\mathbf{W}, \theta)| \leq C \cdot |q(\mathbf{W}, \theta)| \leq C \cdot b(\mathbf{W})$ by assumption (5) where $C = \max(p_1^{-1}, p_2^{-1}, \ldots, p_J^{-1})$. This complete the proof. □

1.4.2 Asymptotic Normality of the Weighted M-Estimator

In order to show that the weighted M-estimator is asymptotically normally distributed, conditions mentioned for consistency in Theorem (1.3.1) is not enough and additional assumptions are needed. Theorem (1.3.2) lists these new assumptions that imply asymptotic normality of the weighted M-estimator.

Theorem 1.4.2. *In addition to the conditions of Theorem(1.3.1), if*

7. $\theta_\circ$ *is in the interior of* Θ *or* $\theta_\circ \in int(\Theta)$.

8. $\mathbf{s}(\mathbf{W},\theta)$ *the score of the objective function is continuously differentiable on* $int(\Theta)$.

9. *Each element of Hessian matrix,* $\mathbf{H}(\mathbf{W},\theta)$ *is bounded in absolute value by a function* $b(\mathbf{W})$, *where* $\mathbb{E}[b(\mathbf{W})] < \infty$.

10. $\mathbf{A}_w = \mathbb{E}\left[\nabla_\theta^2 q(\mathbf{W},\theta)\right]$ *is nonsingular.*

11. $\mathbb{E}[\mathbf{s}(\mathbf{W},\theta_\circ)] = 0$ *and each element of* $\mathbf{s}(\mathbf{W},\theta)$ *has finite second moment. Then*

$$\sqrt{N}\left(\hat{\theta}_w - \theta_\circ\right) \xrightarrow{d} Normal\left(\mathbf{0}, \mathbf{A}_w^{-1}\mathbf{B}_w\mathbf{A}_w^{-1}\right) \tag{1.7}$$

Here $\mathbf{B}_w$ is

$$\begin{aligned}
\mathbf{B}_w = &\sum_{s=1}^{S}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} v^2 p_j^{-2} r_{jm} z_{jm} \nabla_\theta q(\mathbf{W},\theta_\circ)\nabla_\theta q(\mathbf{W},\theta_\circ)'\right] \\
&+ \sum_{s=1}^{S}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v^2 p_j^{-1} p_{j'}^{-1} r_{jm} r_{j't} z_{jm} z_{j't} \nabla_\theta q(\mathbf{W},\theta_\circ)\nabla_\theta q(\mathbf{W},\theta_\circ)'\right] \\
&- \sum_{s=1}^{S}\mathbb{E}\left[\left(\sum_{j=1}^{J}\sum_{m=1}^{K} v\cdot p_j^{-1} r_{jm} z_{jm} \nabla_\theta q(\mathbf{W},\theta_\circ)\right)\cdot\left(\sum_{j=1}^{J}\sum_{m=1}^{K} v\cdot p_j^{-1} r_{jm} z_{jm} \nabla_\theta q(\mathbf{W},\theta_\circ)\right)'\right]
\end{aligned} \tag{1.8}$$

Proof. Score of objective function in each stratum s is

$$\mathbf{s}_{cs}(\theta) = \nabla_\theta g_{cs}(\mathbf{W}_m,\theta)' = \sum_{j=1}^{J}\sum_{m=1}^{K} v\cdot p_j^{-1} r_{jm} z_{jm} \nabla_\theta q(\mathbf{W}_{mcs},\theta)' \tag{1.9}$$

Because clusters are independent sequence in each stratum s by assumption, we can apply the central limit theorem for the sampled clusters within each stratum. Therefore

$$N_s^{-1/2}\sum_{s=1}^{N_s}\left[\mathbf{s}_{cs}(\theta_\circ)-\mathbb{E}\left(\mathbf{s}_s(\theta_\circ)\right)\right]\xrightarrow{d} Normal\left(\mathbf{0},\mathbf{B}_s\right) \tag{1.10}$$

In (1.10) $\mathbb{E}\left[\mathbf{s}_s(\theta_\circ)\right]=\mathbb{E}\left[\nabla_\theta g(\mathbf{W},\theta_\circ)'|\mathbf{W}\in\mathscr{W}_s\right]$. $\mathbf{B}_s$ is the variance of score function in stratum s and is equal to

$$\begin{aligned}
\mathbf{B}_s &= var\left[\mathbf{s}_{cs}(\theta_\circ)\right] = var\left[\nabla_\theta g_{cs}(\mathbf{W}_m,\theta_\circ)\right]\\
&= var\left[\sum_{j=1}^{J}\sum_{m=1}^{K} v p_j^{-1} r_{jm} z_{jm}\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)\right]\\
&= \mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} v^2 p_j^{-2} r_{jm} z_{jm}\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)'\right] \qquad (1.11)\\
&+ \mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v^2 p_j^{-1} p_j'^{-1} r_{jm} r_{j't} z_{jm} z_{j't}\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)'\right]\\
&- \mathbb{E}\left[\left(\sum_{j=1}^{J}\sum_{m=1}^{K} v p_j^{-1} r_{jm} z_{jm}\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)\right)\right]\cdot\mathbb{E}\left[\left(\sum_{j=1}^{J}\sum_{m=1}^{K} v p_j^{-1} r_{jm} z_{jm}\nabla_\theta q(\mathbf{W}_{mcs},\theta_\circ)\right)'\right]
\end{aligned}$$

Variance of score function consists of three terms. The first term in (1.11) is simply the variance of score if a simple random sample is in hand. In other words, the first part is correct variance if *i.i.d* observations are available. Second and third terms in (1.11) are added due to the sample design. The second term measures the cluster effect and accounts for correlation within clusters. This term is positive in most cases and it is substantial if the degree of correlation between the observations inside a single cluster is high and/or K the number of observations sampled from each cluster increases. The third part captures the stratum effect. It is negative and therefore reduces the size of variance.

We also obtain the following important equality by using(1.6) in Theorem(1.3.1)

$$\sum_{s=1}^{S}\mathbb{E}\left[\nabla_\theta g_{cs}(\mathbf{W}_m,\theta_\circ)\right]=\sum_{s=1}^{S}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} v_{cs} p_j^{-1} r_{jm} z_{jm}\nabla_\theta q(\mathbf{W}_{csm},\theta_\circ)\right]\equiv\mathbf{0} \tag{1.12}$$

Using (1.12) the score of the objective function, multiplied by $\sqrt{N}$ can be written as

$$N^{-1/2}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\nabla_\theta g_{cs}(\mathbf{W}_m,\theta_\circ)=N^{-1/2}\sum_{s=1}^{S}\left[\sum_{c=1}^{N_s}\nabla_\theta g_{cs}(\mathbf{W}_m,\theta_\circ)-\mathbb{E}\left[\nabla_\theta g_{cs}(\mathbf{W}_m,\theta)\right]\right] \tag{1.13}$$

because the sampled clusters across strata and are also independent by assumption, then (1.13) has asymptotic normal distribution with mean zero and variance equal to $\mathbf{A}_w^{-1}\mathbf{B}_w\mathbf{A}_w^{-1}$. □

1.4.3 Estimating the Asymptotic Variance

Obtaining consistent estimation of the asymptotic variance of $\sqrt{N}(\hat{\theta}_w - \theta_\circ)$ is fairly straightforward. First, we need to have a consistent estimation of Hessian matrix $\mathbf{A}_w$. It is second-order partial derivative of (1.4) sum over all strata

$$\mathbf{A}_w = \sum_{s=1}^{S} \mathbb{E}\left[\nabla_\theta^2 g_{cs}(\mathbf{W}_m, \theta_\circ)\right] = \sum_{s=1}^{J} \mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} v p_j^{-1} r_{jm} z_{jm} \nabla_\theta^2 q(\mathbf{W}, \theta_\circ)\right] \tag{1.14}$$

By lemma (4.3) in Newey and McFadden (1994) and under the assumptions of Theorem (1.3.2) consistent estimator of $\mathbf{A}_w$ is

$$\hat{\mathbf{A}}_w = N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is} p_j^{-1} y_{is} r_{jm} z_{jm} \nabla_\theta^2 q(\mathbf{w}_{ism}, \hat{\theta}_w)$$

As Wooldridge (2010), we assume that the elements of $\nabla_\theta q(\mathbf{W},\theta)\nabla_\theta q(\mathbf{W},\theta)'$ are bounded in absolute value by a function with finite expectation in order to have consistent estimation of $\mathbf{B}_w$. Then, a consistent estimator of $\mathbf{B}_w$ is

$$\begin{aligned}
\hat{\mathbf{B}}_w = & N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is}^2 p_j^{-2} y_{is} r_{jm} z_{jm} \nabla_\theta q(\hat{\theta}_w)\nabla_\theta q(\hat{\theta}_w)' \\
& + N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v_{is}^2 p_j^{-1} p_{j'}^{-1} y_{is} r_{jm} r_{j't} z_{jm} z_{j't} \nabla_\theta q(\hat{\theta}_w)\nabla_\theta q(\hat{\theta}_w)' \\
& - \sum_{s=1}^{S}\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is} p_j^{-1} y_{is} r_{jm} z_{jm} \nabla_\theta q(\hat{\theta}_w)\right]\cdot\left[\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is} p_j^{-1} y_{is} r_{jm} z_{jm} \nabla_\theta q(\hat{\theta}_w)\right]'
\end{aligned}$$

Here $\nabla_\theta q(\hat{\theta}_w) \equiv \nabla_\theta q(\mathbf{w}_{ism}, \hat{\theta}_w)$.

Therefore the estimate of asymptotic variance of $\hat{\theta}_w$ is

$$\widehat{Avar(\hat{\theta}_w)} = \hat{\mathbf{A}}_w^{-1}\hat{\mathbf{B}}_w\hat{\mathbf{A}}_w^{-1}/N \tag{1.15}$$

The diagonal elements of (1.15) are the asymptotic variances of estimated parameters.

1.5 Estimation under Exogenous stratification

Partitioning $\mathbf{w}$ as $(\mathbf{x},\mathbf{y})$ and then dividing the population of interest purely based on $\mathbf{x}$ in a model that is made to explain distribution of $\mathbf{Y}$ given $\mathbf{x}$, $\mathbb{E}\left[\mathbf{Y}|\mathbf{X}=\mathbf{x}\right]$ is called exogenous stratification. In multi-stage sampling, exogenous stratification can be applied in any stages of sampling.

In the sampling scheme described in section 2, there are two levels of stratification. In first level, standard sampling and in second level, variable probability sampling are used. Stratification in each level can be endogenous or exogenous and therefore three possibilities can be distinguished when at least we have one level of exogenous stratification. In case one both levels of stratification are exogenous. In case two, the first level of stratification is exogenous but is endogenous in second level. Alternatively in case three, first level of stratification is endogenous and second level is exogenous. Since case three is very unlikely to be used in practice, we limit our studies to cases one and two.

1.5.1 Consistency of the Unweighted M-Estimator

Assume $\mathbf{W}$ is partitioned as $(\mathbf{X},\mathbf{Y})$, then in exogenous stratification population problem is

$$\min_{\theta\in\Theta}\mathbb{E}\left[q(\mathbf{W},\theta)|\mathbf{X}\right] \tag{1.16}$$

Our analysis of weighted estimator in previous section can be applied with or without exogenous stratification. However weighting observations in exogenous case is not necessary anymore and an unweighted estimator is also consistent.

1.5.1.1 Consistency of the Unweighted M-Estimator: Case One

As mentioned above, in case one both level of stratifications are exogenous. The unweighted estimator solves the sample objective function

$$\min_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{m=1}^{K}y_{is}r_{jm}z_{jm}q(\mathbf{w}_{ism},\theta) \tag{1.17}$$

Objective function (1.17) is same as (1.3) without the weights $v_{is} \cdot p_j^{-1}$. The following theorem states conditions for consistency of unweighted estimator.

Theorem 1.5.1. *Assume that first five conditions in Theorem (1.3.1) hold. Add new two following conditions*

6. *Stratification in first and second levels are based on exogenous variables* $\mathbf{x}$*. It means that stratification is a deterministic function of* $\mathbf{x}$ *in both levels.*

7. *For all* $\mathbf{x}$*,* $\theta_\circ$ *solves* $\min_{\theta \in \Theta} \mathbb{E}\left[q(\mathbf{W}, \theta)|\mathbf{X}\right]$*, and* $\theta_\circ$ *uniquely minimizes*
$$\sum_{j=1}^{J} \sum_{m=1}^{K} p_j \mathbb{E}\left[r_{jm} q(\mathbf{W}_m, \theta)\right] = \sum_{j=1}^{J} \sum_{m=1}^{K} p_j \mathbb{E}\left[r_{jm} \mathbb{E}\left[q(\mathbf{W}_m, \theta)\right] |\mathbf{X}\right] \tag{1.18}$$

Then uniform weak law of large numbers holds and $\hat{\theta}_u \longrightarrow \theta_\circ$ *in probability as* $N \to \infty$*.*

Proof. We need to show that θ is the unique solution to
$$\mathbb{E}\left[\sum_{j=1}^{J} \sum_{m=1}^{K} r_{jm} z_{jm} q(\mathbf{W}_m, \theta)\right] \tag{1.19}$$
By assumption (6) in Theorem (1.4.1), r_{jm} is a function of $\mathbf{x}$. Also z_{jm} is independent of $\mathbf{w}$ and consequently of $\mathbf{x}$. Therefore

$$\mathbb{E}\left[r_{jm} z_{jm} q(\mathbf{W}_m, \theta)|\mathbf{X}\right] = r_{jm} \mathbb{E}(z_{jm}|\mathbf{X}) \mathbb{E}\left[q(\mathbf{W}_m, \theta)|\mathbf{X}\right] = r_{jm} p_j \mathbb{E}\left[q(\mathbf{W}_m, \theta)|\mathbf{X}\right] \tag{1.20}$$

By assumption (7) in Theorem (1.4.1), $r_{jm} p_j \mathbb{E}\left[q(\mathbf{W}_m, \theta)|\mathbf{X}\right]$ is minimized at $\theta_\circ$, but perhaps not uniquely. By iterated expectation we have $\mathbb{E}[q(\mathbf{W}_m, \theta)] = \mathbb{E}\left[\mathbb{E}[q(\mathbf{W}_m, \theta)|\mathbf{X}]\right]$ and therefore $\theta_\circ$ is a solution to (1.19). Then the expectation of (1.19) is same as (1.18), and by assumption $\theta_\circ$ is unique solution to (1.18). □

1.5.1.2 Consistency of Unweighted M-estimator: Case Two

When first level of stratification is exogenous while the second level is endogenous, a logical analogy is that we can drop the weight associated to first level of stratification i.e. v_{is} but need to keep the weight associated to the second level of stratifican i.e. p_j^{-1} in order to have consistent estimator. Next theorem confirms the truth of this analogy under specific conditions.

Theorem 1.5.2. *Assume that first five conditions in Theorem (1.3.1) hold. Add new two following conditions*

6. *Stratification in first level is a deterministic function of* $\mathbf{x}$.

7. $\theta_\circ$ *is the unique solution to* $\mathbb{E}\left[q(\mathbf{W},\theta)|\mathbf{x}\in\mathbf{X}\right]$ *for all s.*

Then uniform law of large numbers hold and $\hat{\theta}_u \xrightarrow{p} \theta_\circ$ *as* $N\to\infty$.

Proof. The expected value of cluster i in stratum s is

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} p_j^{-1} r_{jm}z_{jm}q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right] = \sum_{j=1}^{J}\sum_{m=1}^{K} p_j^{-1}\mathbb{E}\left[r_{jm}z_{jm}q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right]\\
&=\sum_{j=1}^{J}\sum_{m=1}^{K} p_j^{-1}\mathbb{E}\left[z_{jm}|\mathbf{x}\in\mathbf{X}_s\right]\cdot\mathbb{E}\left[r_{jm}q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right] = \sum_{j=1}^{J}\sum_{m=1}^{K}\mathbb{E}\left[r_{jm}q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right]\\
&=\mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} r_{jm}q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right] = \mathbb{E}\left[\sum_{m=1}^{K_{sc}} q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right]\\
&=\sum_{m=1}^{K_{sc}}\left[q(\mathbf{W}_m,\theta)|\mathbf{x}\in\mathbf{X}_s\right] \qquad (1.21)
\end{aligned}$$

By assumption (7) in Theorem (1.4.2) $\theta_\circ$ is unique solution for $\mathbb{E}\left[q(\mathbf{W},\theta)|\mathbf{x}\in\mathbf{X}_s\right]$ and so is unique solution for last equality in (1.21). We also need to show that the uniform law of large numbers holds for each s which is similar to the argument as in Theorem (1.3.1). □

1.5.2 Asymptotic Normality of the Unweighted M-Estimator

According to previous section, asymptotic normality results for the unweighted estimator when stratification is based on $\mathbf{x}$ in both levels or just in first level are represented in frame of the following two theorems.

Theorem 1.5.3. *In addition to the condition of Theorem (1.4.1) if*

8. $\theta_\circ$ *is in the interior of* Θ *or* $\theta_\circ \in int(\Theta)$.

9. *For all* $\mathbf{w}\in\mathscr{W}$, $\nabla_\theta q(\mathbf{w},\cdot)$ *the score of objective function is continuously differentiable on* $int(\Theta)$.

10. *Each element of Hessian matrix,* $\mathbf{H}(\mathbf{W},\theta)$ *is bounded in absolute value by an arbitrary function* $b(\mathbf{w})$*, where* $\mathbb{E}[b(\mathbf{w})] < \infty$*.*

11. *For all* $\mathbf{x}$*,* $\mathbb{E}[\nabla_\theta q(\mathbf{W},\theta_\circ)|\mathbf{X}=\mathbf{x}] = \mathbf{0}$*, and all elements of* $\nabla_\theta q(\mathbf{W},\theta)$ *has finite second moment.*

12. $\mathbf{A}_u = \sum_{s=1}^{S} \mathbb{E}\left[\sum_{j=1}^{j}\sum_{m=1}^{K} r_{jm}z_{jm}\nabla^2_\theta q(\mathbf{W_m},\theta_\circ)|\mathbf{X}=\mathbf{x}\right]$ *is nonsingular.*

Then

$$\sqrt{N}(\hat{\theta}_u - \theta_\circ) \xrightarrow{d} Normal\left(\mathbf{0}\,,\, \mathbf{A}_u^{-1}\mathbf{B}_u\mathbf{A}_u^{-1}\right) \tag{1.22}$$

where

$$\begin{aligned}\mathbf{B}_u = &\sum_{s=1}^{S}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} p_j r_{jm}\nabla_\theta q(\mathbf{W},\theta_\circ)\nabla_\theta q(\mathbf{W},\theta_\circ)'|\mathbf{X}=\mathbf{x}\right]\\ &+\sum_{s=1}^{S}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} p_j p_{j'} r_{jm} r_{j't}\nabla_\theta q(\mathbf{W},\theta_\circ)\nabla_\theta q(\mathbf{W},\theta_\circ)'|\mathbf{X}=\mathbf{x}\right]\end{aligned} \tag{1.23}$$

for all $\mathbf{x}$*.*

Proof. In this case, stratifications in both levels are exogenous and the score of the objective function in each stratum s is $\mathbf{s}(\mathbf{W}_m,\theta) = \nabla_\theta g(\mathbf{W}_m,\theta) = \sum_{j=1}^{j}\sum_{m=1}^{K} r_{jm}z_{jm}\nabla_\theta q(\mathbf{W},\theta)$. Under assumption (11) in Theorem (1.5.1), the expected value of the score is

$$\mathbb{E}[\mathbf{s}(\mathbf{W}_m,\theta_\circ)|\mathbf{x}] = \mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} r_{jm}z_{jm}\nabla_\theta q(\mathbf{W}_m,\theta_\circ)|\mathbf{X}=\mathbf{x}\right] = \mathbf{0} \tag{1.24}$$

Then by applying central limit theorem for independent clusters within each stratum, asymptotic distribution of the score in stratum s is

$$N_s^{-1/2}\sum_{i=1}^{N_s}[\mathbf{s}_{is}(\mathbf{W}_m,\theta_\circ)|\mathbf{X}=\mathbf{x}] \xrightarrow{d} Normal(\mathbf{0}\,,\,\mathbf{B}_s^u) \tag{1.25}$$

$\mathbf{B_s}^u$, represents the variance of the score function in stratum s under exogenous stratification. It is equal to

$$\begin{aligned}
\mathbf{B}_s^u =& var\left[\mathbf{s}(\mathbf{W}_m, \theta_\circ)|\mathbf{X}=\mathbf{x}\right] = var\left[\nabla_\theta g(\mathbf{W}_m, \theta_\circ)|\mathbf{X}=\mathbf{x}\right] \\
=& var\left[\sum_{j=1}^{J}\sum_{m=1}^{K} r_{jm}z_{jm}\nabla_\theta q(\mathbf{W}, \theta_\circ)|\mathbf{X}=\mathbf{x}\right] \\
=& \mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} r_{jm}z_{jm}\nabla_\theta q(\mathbf{W}, \theta_\circ)\nabla_\theta q(\mathbf{W}, \theta_\circ)'|\mathbf{X}=\mathbf{x}\right] \\
&+ \mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} r_{jm}r_{j't}z_{jm}z_{j't}\nabla_\theta q(\mathbf{W}, \theta_\circ)\nabla_\theta q(\mathbf{W}, \theta_\circ)'|\mathbf{X}=\mathbf{x}\right] \qquad (1.26)
\end{aligned}$$

for all $\mathbf{x}$. Independency of z's from $\mathbf{W}$ and each other, and also indepdency of clusters between strata leads us to $\sum_{s=1}^{S}\mathbf{B}_s^u$ which is the score of the objective function $\mathbf{B}_u$ in (1.23) and this complete the proof. □

It is interesting to note that under assumption (11), Theorem (1.5.1) in exogenous stratification, the effect of stratification is vanished as comparing (1.23) and (1.8) show this point.

The asymptotic results when stratification is exogenous just in first level is very similar to case one. Next theorem summarizes main conditions and results.

Theorem 1.5.4. *Same conditions as Theorem (1.5.1) last two ones that are replaced with following*

11. $\mathbb{E}\left[\mathbf{s}(\mathbf{W}, \theta_\circ)|\mathbf{x}\in\mathbf{X}\right] = \mathbf{0}$*, in other words we assume the score of the objective function under exogenous in first stage is zero. Also we assume that elements of* $\mathbf{s}(\mathbf{W}, \theta)$ *have finite second moment.*

12. $\mathbf{A}_{\bar{u}} = \sum_{s=1}^{S}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} p_j^{-1} r_{jm}z_{jm}\nabla_\theta^2 q(\mathbf{W}, \theta_\circ)|\mathbf{x}\in\mathbf{X}\right]$ *is nonsingular.*

Then

$$\sqrt{N}(\hat{\theta}_{\bar{u}} - \theta_\circ \xrightarrow{d} Normal(\mathbf{0}\,,\, \mathbf{A}_{\bar{u}}^{-1}\mathbf{B}_{\bar{u}}\mathbf{A}_{\bar{u}}^{-1})$$

where

$$\mathbf{B}_{\bar{u}} = \sum_{s=1}^{S} \mathbb{E}\left[\sum_{j=1}^{J}\sum_{m=1}^{K} p_j^{-2} r_{jm} z_{jm} \nabla_{\theta} q(\mathbf{W},\theta_{\circ}) \nabla_{\theta} q(\mathbf{W},\theta_{\circ})' | \mathbf{X} = \mathbf{x}\right]$$
$$+ \sum_{s=1}^{S} \mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} p_j^{-1} p_{j'}^{-1} r_{jm} r_{j't} z_{jm} z_{j't} \nabla_{\theta} q(\mathbf{W},\theta_{\circ}) \nabla_{\theta} q(\mathbf{W},\theta_{\circ})' | \mathbf{X} = \mathbf{x}\right]$$

for all $\mathbf{x}$.

Proof. It is similar to the Theorem (1.5.1). We just need to weight observations by p_j^{-1} that corresponding with VP sampling in second stage. Like previous case, stratification effect due to SS sampling in first stage is zero. □

1.6 Examples

This section contains some examples that illustrate theoretical results. It also covers some special cases.

Example 1.

As the first example consider a simple liner model

$$y = \mathbf{x}\beta + u \tag{1.27}$$

Here $\mathbf{x}$ is a $1 \times K$ vector of exogenous variables and β is the $K \times 1$ vector of parameters of interest. Assuming noncorrelationo between exogenous x's and error term u, $\mathbb{E}(\mathbf{x}'u) = \mathbf{0}$, the weighted estimator provides consistent estimates of β.

The sample optimization problem is

$$\min_{\theta\in\Theta} \frac{1}{N} \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc} p_j^{-1} r_{jm} z_{jm} (y_{scm} - \mathbf{x}_{scm}\beta)^2 \tag{1.28}$$

First order condition is

$$\frac{1}{N} \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc} p_j^{-1} r_{jm} z_{jm} \mathbf{x}'_{scm} (y - \mathbf{x}_{scm}\hat{\beta}) = \mathbf{0}$$

or

$$\frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}p_j^{-1}r_{jm}z_{jm}\mathbf{x}'_{scm}\hat{u}_{scm} = \mathbf{0}$$

where $\hat{u}_{scm} = y_{scm} - \mathbf{x}_{scm}\hat{\beta}$. In this linear model and under multi-stage sampling scheme, a consistent estimators of asymptotic variances of $\hat{\beta}$'s are obtained by applying Theorem (1.3.2) where consistent estimators of $\mathbf{A}_w$, and $\mathbf{B}_w$ are

$$\hat{\mathbf{A}}_w = \frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}p_j^{-1}r_{jm}z_{jm}\mathbf{x}'_{scm}\mathbf{x}_{scm}$$

and

$$\begin{aligned}\hat{\mathbf{B}}_w = & \frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}^2p_j^{-1}r_{jm}z_{jm}\mathbf{x}'_{scm}\mathbf{x}_{scm} \\ & + \frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v_{sc}^2p_j^{-1}p_{j'}^{-1}r_{jm}r_{j't}z_{jm}z_{j't}\hat{u}_{scm}\hat{u}_{sct}\mathbf{x}'_{scm}\mathbf{x}_{sct} \\ & - \sum_{s=1}^{S}\frac{1}{N}\left[\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}p_j^{-1}r_{jm}z_{jm}\hat{u}_{scm}\mathbf{x}'_{scm}\right]\cdot\left[\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}p_j^{-1}r_{jm}z_{jm}\hat{u}_{scm}\mathbf{x}'_{scm}\right]'\end{aligned}$$

Example 2.

As the second example, consider binary models like logit or probit. In binary response models of the form

$$P(y=1|\mathbf{x}) = G(\mathbf{x}\beta) \equiv p(\mathbf{x})$$

where $\mathbf{x}$ is $1\times K$, β is $K\times 1$, we take the first element of $\mathbf{x}$ to be unity. Also we assume $0 < G(\mathbf{x}\beta < 1$ for all $\mathbf{x}$ and β. The log-likelihood for observation i is

$$l_i(\beta) = y_i\log\left[G(\mathbf{x_i}\beta)\right] + (1-y_i)\left[1-G(\mathbf{x_i}\beta)\right]$$

The weighted estimator in this case simply is the weighted maximum likelihood that gives observation i in cluster c in stratum s corresponding weight that is $v_{sc}\cdot p_j^{-1}$.

In this example, consistent estimator of $\mathbf{A}_w$ and $\mathbf{B}_w$ according to Theorem (1.3.2) are

$$\hat{\mathbf{A}}_w = \frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}p_j^{-1}r_{jm}z_{jm}g_{scm}^2\mathbf{x}'_{scm}\mathbf{x}_{scm}/\hat{\xi}_{scm} \quad (1.29)$$

and

$$
\begin{aligned}
\hat{\mathbf{B}}_w =& \frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc}^2 p_j^{-2} r_{jm} z_{jm} g_{scm} \mathbf{x}_{scm}' \mathbf{x}_{scm} \\
&+ \frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v_{sc}^2 p_j^{-1} p_{j'}^{-1} r_{jm} r_{j't} z_{jm} z_{j't} g_{scm} g_{sct} \mathbf{x}_{scm}' \mathbf{x}_{sct} \\
&- \sum_{s=1}^{S}\frac{1}{N}\left[\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc} p_j^{-1} r_{jm} z_{jm} g_{scm} \mathbf{x}_{scm}'\right]\cdot\left[\sum_{c=1}^{N_s}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{sc} p_j^{-1} r_{jm} z_{jm} g_{scm} \mathbf{x}_{scm}'\right]'
\end{aligned}
$$

Here $g(z) = \dfrac{dG(z)}{dz}$ and $\hat{\xi}_{scm} = \hat{G}_{scm}(1-\hat{G}_{scm})$.

Example 3.

Example 3 is a special case when p_j is set equal 1. In other words, we eliminate last level of stratification or VP sampling. In this case our results in section 3 change to:

$$\hat{\mathbf{A}}_w = N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{m=1}^{K} v_{is} y_{is} \nabla_\theta^2 q(\mathbf{w}_{ism},\hat{\theta})$$

And estimation of $\mathbf{B}_w$ is

$$
\begin{aligned}
\hat{\mathbf{B}}_w =& N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{m=1}^{K} v_{is}^2 y_{is} \nabla_\theta q(\mathbf{w}_{ism},\hat{\theta}) \nabla_\theta q(\mathbf{w}_{ism},\hat{\theta})' \\
&+ N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v_{is}^2 y_{is} \nabla_\theta q(\mathbf{w}_{ism},\hat{\theta}) \nabla_\theta q(\mathbf{w}_{ism},\hat{\theta})' \\
&- \sum_{s=1}^{S}\frac{1}{N}\left(\sum_{i=1}^{N}\sum_{m=1}^{K} v_{is} y_{is} \nabla_\theta q(\mathbf{w}_{ism},\hat{\theta})\right)\cdot\left(\sum_{i=1}^{N}\sum_{m=1}^{K} v_{is} y_{is} \nabla_\theta q(\mathbf{w}_{ism},\hat{\theta})\right)'
\end{aligned}
$$

These results are similar to Bhattacharya's (2005) ones. Also Wooldridge (2008) obtains same results in case of linear model estimated by least squares.

Example 4.

Consider a case without first level of stratification and clusters that contains just one unit of observation. Then our results will change to

$$\hat{\mathbf{A}}_w = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{J} p_j^{-1} r_{ij} z_{ij} \nabla_\theta^2 q(\mathbf{w},\hat{\theta})$$

And

$$\hat{\mathbf{B}}_w = N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{J} p_j^{-2} r_{ij} z_{ij} \nabla_\theta q(\mathbf{w}_i, \hat{\theta}) \nabla_\theta q(\mathbf{w}_i, \hat{\theta})$$

These are same results as Wooldridge (1999) in studying variable probability sampling case.

1.7 Two-Step M-Estimator

Consider a panel data model for a random draw *i* from the population

$$\mathbb{E}(\mathbf{Y}_i | \mathbf{X}_i = \mathbf{x}_i) = \mathbf{m}(\mathbf{x}_i, \theta_\circ) \tag{1.30}$$

where $\mathbf{y}_i$ is a $T \times 1$ vector on the dependent variable and $\mathbf{m}(\mathbf{x}_i, \theta)$ is a $T \times 1$ of conditional mean functions. Here we assume that explanatory variables are strictly exogenous. Stratification is normally done on variables on first period. A consistent, asymptotically normal estimator is obtained by applying pooled weighted M-estimator discussed in previous sections. The estimator of asymptotic variance of $\hat{\theta}_w$ is obtained from (1.15), where $\nabla_\theta q(\mathbf{w}_{ism}, \hat{\theta}_w)$ is the $P \times T$ matrix. Arbitrary serial correlation and heteroskedasticity are allowed in calculation of the estimator (1.15).

Under assumption (1.30), where conditional mean is correctly specified, can we do more in context of stratified samples? This is the question that we will answer in the next chapter. In general, under (1.30) we can use generalized least squares (GLS) methods to obtain more efficient estimators of the parameters appearing in a set of conditional mean functions. To obtain more efficient estimators we usually need $\hat{\theta}_w$ from the first step.

Let $\Omega(x_i, \gamma)$ be a model for the $T \times T$ conditional variance matrix $Var(\mathbf{Y}_i | \mathbf{X}_i)$. If this model is correctly specified, in general, we can obtain consistent estimator of the true parameters in the variance matrix, $\gamma_\circ$. In most application, we obtain an estimation of γ from a first step, for example by using residuals from an initial weighted M-estimator, discussed in this paper. Given, $\hat{\gamma}$, and assuming that conditional variance matrix is nonsingular for all *i*, we can estimate $\theta_\circ$ by solving

$$\min_{\theta} \sum_{i=1}^{N} \left[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \theta)\right]' \left[\Omega(\mathbf{x}_i, \hat{\gamma})\right]^{-1} \left[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \theta)\right]. \tag{1.31}$$

Wooldridge (2010) calls the solution to (1.31) weighted multivariate nonlinear least squares (WMNLS) estimator.

Interestingly, even if the chosen model for the conditional variance $\Omega(\mathbf{x}_i, \gamma)$ is misspesified, WMNLS estimator might produce a more efficient estimator of $\theta_\circ$ than an estimator that ignores variances and covariances at least under (1.30). In most cases, a misspecified model of variance matrix captures key features of the conditional second moments. This is the key insight in the generalized estimating equation (GEE) literature, which is typically applied to panel data models. In GEE literature, the conditional variance matrix $\Omega(\mathbf{x}_i, \gamma)$ is called working variance matrix, which is allowed, and in many cases is known, to be misspecified. In next chapter we investigate the problem of efficient estimator in panel data models where simple random sampling is not a correct assumption in more detail.

1.8 Conclusion

Many data sets in economics studies and other branches of social sciences are not *i.i.d* observations but come from multi-stage stratification and clustered surveys. These surveys usually produce data that are not random. Then statistical inference could be faulty if we overlook sampling design.

In this chapter I examine statistical inference in multi-stage sampling designs in framework of M-estimators. The results show that neglecting sampling scheme causes overestimating or underestimating of the variances. Applying weighted M-estimator gives consistent and normally distributed estimators regardless of stratification type; exogenous or endogenous. However under exogenous stratification, unweighted estimators are consistent. The results show that variance of the weighted estimator consists of three parts. Part one measures variance under "*i.i.d*" assumption. Parts two and three take into account clustering and stratification effects. Clustering effects are usually positive, while stratification effect is negative. These two rarely offset each other and therefore overlooking these two parts is potentially problematic.

An interesting question that arise is the possibility of having more efficient estimator under stratified samples. Under simple random sampling, and assuming that the conditional mean is correctly specified, we can apply generalized least squares methods to obtain efficiency gain. We follow up this possibility in panel data models with stratified sampling schemes in next chapter.

Chapter 2

ASYMPTOTIC EFFICIENCY IN THE PANEL DATA MODELS WITH STRATIFIED SAMPLING

2.1 Introduction

Finding more efficient estimators helps researchers to increase the precision of their statistical inferences. Efficiency usually comes at a price. It requires stronger assumptions needed for consistency. Here we assume that explanatory variables are strictly exogenous. However in panel data studies, this assumption is violated in models with lagged dependent variables and perhaps in models without lagged dependent variables. Fixed effects (FE) and random effects (RE) are two well known linear methods used in empirical studies that require strict exogeneity of the estimators. In RE approach, the serial correlation in the composite error is exploited in a generalized least squares (GLS) framework. In GLS procedure we also need to add assumptions on conditional variance matrix of the error term.

The issue of efficiency in context of stratified sampling has been the subject of interest already. Among others, Cosslett (1981a, 1081b, and 1993), and Imbens (1992) examines efficiency for discrete choice models. Imbens and Lancaster (1996) develop a new estimator for this estimation problem and show that it achieves the semi-parametric efficiency bound for this case. Recently Tripathi (2011) develops efficient empirical likelihood-based inference for moment restriction models when data are collected by stratified sampling schemes. In this chapter, we study efficiency in panel data models with stratified data. The main idea is to utilize information within panels similar to GLS method in order to gain efficiency. It should be emphasized that we are only approximate the efficient estimator in the sample and try to obtain more efficient estimates compare with just pooled estimators that ignore correlations within panels. In other words, our goal in this chapter is not to find the efficiency bound.

The paper is organized as follows. The next section presents the model and conditional moment restriction that we need as an assumption to be held in the population. Section 3 introduces the sampling scheme, sampling objective function and relevant probabilities. In section 4 we first show that conditional moment restrictions is held in the sample. Then we discuss about efficient estimators by referring to some well known works in the literature and how one should apply them in the context of stratified samples. In the same section we drive a function that minimize the asymptotic variance. In section 6 we do a Monte Carlo experiment with the normal linear case and look at the results of applying new estimators in case of exogenous and endogenous stratification. Section 7 shows application of the method on PSID data. In the last section the main finding of the paper will be summarized and ends with some concluding remarks. All proofs and tables are contained in Appendices.

2.2 The Moment Conditions

Let $\mathbf{W}_i$ be a $M \times 1$ random vector taking values in $\mathscr{W} \subset \mathbb{R}^M$, where $\mathbb{R}^M$ is an M-dimensional Euclidean space. Some feature of the distribution of $\mathbf{W}$ is function of a $P \times 1$ parameter vector θ that is an element of the parameter space Θ where $\Theta \subset \mathbb{R}^P$. Now consider the class of estimators such that a zero conditional moment restriction in the population is satisfied:

$$\mathbb{E}\left[\mathbf{r}\left(\mathbf{W}, \theta_\circ\right)|\mathbf{W}_2\right] = \mathbf{0} \text{ for all } \mathbf{W}_2 \in \mathscr{W}_2 \tag{2.1}$$

Here $\mathbf{r}\left(\mathbf{W}, \theta\right)$ is a $L \times 1$ vector of functions, $\theta_\circ$ satisfies the conditional moment assumptions and $\mathbf{W}_2 \in \mathbb{R}^K$ is a sub-vector of $\mathbf{W} \in \mathbb{R}^M$. For instance $\mathbf{r}\left(\mathbf{W}, \theta\right)$ can be a vector of residuals and $\mathbf{W}_2$ a vector of instrumental variables. We need standard regularity conditions such as continuity and differentiability of $\mathbf{r}\left(\mathbf{W}, \theta\right)$ on the interior of Θ.

2.3 Sampling Scheme

The analysis of asymptotic behaviors of an estimator becomes more complicated when the data set comes from non-random sampling schemes like stratified samples. One important source of the complexity is the difference between the population distribution on the one hand and sample distribution on the other hand. However, in simple random sampling these two distributions are the same.

In multinomial sampling, stratum $\mathscr{W}_j$ is a subset of $\mathscr{W}$ for $j = 1, \cdots, J$. Let Q_s be the probability of a randomly drawn observation lying in $\mathscr{W}_j$ i.e.

$$Q_j = P\left(\mathbf{W} \in \mathscr{W}_j\right) \tag{2.2}$$

And let S be the stratum indicator that shows from which stratum an observation was drawn. In a multinomial scheme, first the stratum indicator s_i where $s_i \in \{1, 2, \cdots, J\}$ is chosen randomly with probability H_j. It means

$$H_j = P(S_i = j) \tag{2.3}$$

In the second step, observation $\mathbf{W}_i$ is randomly drawn from the stratum which the indicator $s_i = j$. This leads to the sample objective function

$$\sum_{i=1}^{N} \sum_{j=1}^{J} 1\left[S_i = j\right] \frac{Q_j}{H_j} \mathbf{r}\left(\mathbf{V}_i, \theta\right) \tag{2.4}$$

Unlike random sampling where all the observations are equally weighted no matter which subpopulation or stratum they belong, in multinomial sampling scheme observations depend on their stratum have different weights. The objective function in (2.4) weights observation i by $\frac{Q_j}{H_j}$ if it comes from stratum j. So if all observations are weighted equally or if $Q_s = H_s$ for all s then there is no gain of stratified sampling over random sampling.

To emphasize the difference between distribution of observations in population and in the sample under stratified sampling scheme, random vectors in population and in the sample are represented with $\mathbf{W}$ and $\mathbf{V}$ respectively.

2.4 Efficient estimation under moment restrictions

2.4.1 Moment restrictions in the sample

To study efficiency in panel data models when data set comes form stratified samples and under conditional expectation assumption (2.1), one first needs to evaluate conditional expectation of the sample objective function in equation (2.4). To this end, first for each observation i, define

$$q(S,\mathbf{V},\theta) = \sum_{j=1}^{J} 1[S=j]\frac{Q_j}{H_j}\mathbf{r}(\mathbf{V},\theta) \tag{2.5}$$

$q(\cdot)$ is a function of random variable S, an indicator variable representing stratum of observation i, and random vector $\mathbf{V}$. This function also depends on the sampling weight of each observation i, $\frac{Q_j}{H_j}$, that are assumed to be known. We want to show that the expected value of function $q(\cdot)$ given $\mathbf{V}_2$ and evaluated in true parameter value $\theta_\circ$ is zero.

$$\mathbb{E}[q(S,\mathbf{V},\theta_\circ)|\mathbf{V}_2] = \sum_{j=1}^{J}\mathbb{E}\left\{1[S=j]\frac{Q_j}{H_j}\mathbf{r}(\mathbf{V},\theta_o)|\mathbf{V}_2\right\} = \mathbf{0} \tag{2.6}$$

Using definition of expected value and assuming that $\mathbf{V}$ is a continuous random vector, expected value of (2.5) is

$$\sum_{j=1}^{J}\int_{\mathbf{v}\in\mathscr{W}} 1[s=j]\frac{Q_j}{H_j}\mathbf{r}(\mathbf{v},\theta_o)\cdot g(s,\mathbf{v}|\mathbf{v}_2)\,d\mathbf{v} \tag{2.7}$$

Equation (2.7) shows that we need to find the conditional *sampling* density of S and $\mathbf{V}$ given $\mathbf{V}_2$, or $g(s,\mathbf{v}|\mathbf{v}_2)$. Imbens and Lancaster (1996) show that this conditional density function is

$$g(s,\mathbf{v}|\mathbf{v}_2) = \frac{f(\mathbf{v}|\mathbf{v}_2,\theta)\frac{H_s}{Q_s}}{\sum_{j=1}^{J}\frac{H_j}{Q_j}R(j,\mathbf{v}_2,\theta)} \tag{2.8}$$

Equation (2.8) represent conditional sampling density of S and $\mathbf{V}$ given $\mathbf{V}_2$ in terms of conditional density of $\mathbf{V}$ given $\mathbf{V}_2$ in the population, sampling weight $\frac{H_s}{Q_s}$, and $R(s,\mathbf{v}_2,\theta)$. Here $R(s,\mathbf{v}_2,\theta)$ is defined to be the probability that a random drawn observation is in stratum s given $\mathbf{V}_2$. It is a known function of $s,\mathbf{v}_2$, and θ. Also it is important to note that since we assume the strata are not overlapping, the conditional sampling density of S and $\mathbf{V}$ given $\mathbf{V}_2$ is the same as the conditional density of $\mathbf{V}$ given $\mathbf{V}_2$ i.e. $g(s, \mathbf{v}|\mathbf{v}_2) = g(\mathbf{v}|\mathbf{v}_2)$. By substituting (2.8) in (2.7) we have

$$\sum_{j=1}^{J} \int_{\mathbf{v} \in \mathscr{W}} 1[s = j] \frac{Q_j}{H_j} \mathbf{r}(\mathbf{v}, \theta) \cdot \frac{f(\mathbf{v}|\mathbf{v}_2, \theta) \dfrac{H_j}{Q_j}}{\sum\limits_{j=1}^{J} \dfrac{H_j}{Q_j} R(j, \mathbf{v}_2, \theta)} d\mathbf{v}$$

$$= \sum_{j=1}^{J} \int_{\mathbf{v} \in \mathscr{W}} 1[s = j] \mathbf{r}(\mathbf{v}, \theta) \cdot \frac{f(\mathbf{v}|\mathbf{v}_2, \theta)}{\sum\limits_{j=1}^{J} R(j, \mathbf{v}_2, \theta)} d\mathbf{v} \tag{2.9}$$

Since $\mathscr{W}_1, \mathscr{W}_2, \cdots, \mathscr{W}_J$ are mutually disjoint and the union set of this disjoints subpopulations, $\bigcup_{j=1}^{J} \mathscr{W}_j$, covers whole population, saying that stratum of observation i is j or $S_i = j$ is equivalent to say that observation i belongs to subpopulation j or $\left[\mathbf{v}_i \in \mathscr{W}_j\right]$. So we can exchange $1[s = j]$ with $1\left[\mathbf{w}_i \in \mathscr{W}_j\right]$ in expression (2.9) which gives us

$$= \sum_{j=1}^{J} \int_{\mathbf{v} \in \mathscr{W}} 1\left[\mathbf{v} \in \mathscr{W}_j\right] \mathbf{r}(\mathbf{v}, \theta) \cdot \frac{f(\mathbf{v}|\mathbf{v}_2, \theta)}{\sum\limits_{j=1}^{J} R(j, \mathbf{v}_2, \theta)} d\mathbf{v} \tag{2.10}$$

In expression (2.10), $\sum_{j=1}^{J} R(j, \mathbf{v}_2, \theta)$ is constant and $1\left[v \in \mathscr{W}_j\right]$ just defines the limits of integration and therefore (2.10) can be rewritten as

$$= \frac{1}{\sum\limits_{j=1}^{J} R(j, \mathbf{v}_2, \theta)} \sum_{j=1}^{J} \int_{\mathbf{v} \in \mathscr{W}_j} \mathbf{r}(\mathbf{v}, \theta) \cdot f(\mathbf{v}|\mathbf{v}_2, \theta) d\mathbf{v}$$

$$= \eta \int_{\mathbf{v} \in \mathscr{W}} \mathbf{r}(\mathbf{v}, \theta) \cdot f(\mathbf{v}|\mathbf{v}_2, \theta) d\mathbf{v} \tag{2.11}$$

Here $\eta = \dfrac{1}{\sum\limits_{j=1}^{J} R(j, \mathbf{v}_2, \theta)}$ is a constant and equation (2.11) by definition is the conditional expectation of $\mathbf{r}(\cdot)$ or

$$\int_{\mathbf{v} \in \mathscr{W}} \mathbf{r}(\mathbf{v}, \theta) \cdot f(\mathbf{v}|\mathbf{v}_2, \theta) d\mathbf{v} = \mathbf{E}[\mathbf{r}(\mathbf{V}, \theta) | \mathbf{V}_2] \tag{2.12}$$

By assumption (2.1), equation (2.12) evaluated in true parameter value $\theta_\circ$, is equal to zero. Hence we show that although multinomial sampling changes the distribution of observations in the sample but zero conditional mean assumption is still held. we summarize the above finding in the following lemma.

Lemma 2.4.1. *If zero conditional moment (2.1) evaluated in true parameter value $\boldsymbol{\theta}_\circ$ holds in the population, then under multinomial stratification sampling scheme, its analog in the sample (2.6) evaluated in $\boldsymbol{\theta}_\circ$ is zero also.*

The result is valid under standard stratified and variable probability sampling schemes too. Imbens and Lancaster (1996) show that these three common types of stratification can be analyzed in a unified manner. They show that regardless of the actual sampling scheme efficient inference should be identical for both standard stratified sampling and multinomial sampling. And variable probability sampling model is just a re-parametrization of the multinomial sampling scheme and therefore the inference should be identical for both models.

2.4.2 Efficient estimation

The result in previous section opens door to apply the well known results developed by Chamberlain (1987), and Newey and McFadden (1994) to find the smallest asymptotic variance under zero conditional mean assumption (2.1). To find such a solution let

$$\Omega(\mathbf{W}_2, \theta_\circ) = \mathbb{E}\left[\mathbf{r}(\mathbf{W}, \boldsymbol{\theta}_\circ)\mathbf{r}(\mathbf{W}, \boldsymbol{\theta}_\circ)' |\mathbf{W}_2\right] = Var[\mathbf{r}(\mathbf{W}, \boldsymbol{\theta}_\circ)|\mathbf{W}_2] \tag{2.13}$$

be the $T \times T$ conditional variance of $\mathbf{r}(\mathbf{W}, \boldsymbol{\theta}_\circ)$ given $\mathbf{W}_2$, in the population, and define

$$G(\mathbf{W}_2, \theta_\circ) = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\mathbf{r}(\mathbf{W}, \boldsymbol{\theta}_\circ)|\mathbf{W}_2] \tag{2.14}$$

be the $T \times P$ conditional mean of gradient in the population. Then it can be shown that

$$\mathbf{Z}^*(\mathbf{W}_2, \theta_\circ) = \Omega(\mathbf{W}_2, \theta_\circ)^{-1}\mathbf{G}(\mathbf{W}_2, \theta_\circ) \tag{2.15}$$

is the function that minimize the asymptotic variance. This function is $T \times P$ and the efficient method of moments estimator solves

$$\mathbb{E}\left[\mathbf{Z}^*(\mathbf{W}_2, \theta_\circ)'\mathbf{r}(\mathbf{W}, \boldsymbol{\theta}_\circ)\right] = \mathbf{0} \tag{2.16}$$

Since stratification changes the distribution of observations in the sample we need first to evaluate conditional variance of the sample objective function $q(S,\mathbf{V},\theta)$. In the first appendix, we show that this variance is equal to

$$\mathbb{E}\left[q(S,\mathbf{V},\theta_\circ)\,q(S,\mathbf{V},\theta_\circ)'\,|\mathbf{V}_2\right] = \sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,\mathbf{r}(\mathbf{V},\theta_\circ)'\,|\mathbf{V}_2, S=j\right] \tag{2.17}$$

We can write the right hand side of equation (2.17) in terms of the conditional variance of $\mathbf{r}(\mathbf{V},\theta)$ in each stratum and so (2.17) can be rewritten as

$$\begin{aligned} var\left[q(S,\mathbf{V},\theta_\circ)\,|\mathbf{V}_2\right] = {} & \sum_{j=1}^{J}\frac{Q_j}{H_j} var\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,|\mathbf{V}_2, S=j\right] \\ & + \sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,|\mathbf{V}_2, S=j\right]\mathbb{E}\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,|\mathbf{V}_2, S=j\right]' \end{aligned} \tag{2.18}$$

Expression (2.18) show that sampling conditional variance of $\mathbf{r}(\mathbf{V},\theta)$ is equal to the sum of conditional weighted variances in strata plus the sum of conditional weighted squares of means in strata. To see the effect of stratification, it is useful to compare it with random sample case, where each observation in population has same weight or in other words $Q_j = H_j$ for all j, and assume conditional expected value in each stratum is equal to conditional expected value in the population which is zero by assumption. Then equation (2.18) reduces to sum of conditional variances in strata.

There are two interesting cases that need attention. First case is when strata are function of exogenous variables $\mathbf{V}_2$. Then the stratification is exogenous. It causes the second term in right hand side of (2.18) to be zero, because

$$\mathbb{E}\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,|\mathbf{V}_2, S=j\right] = \mathbb{E}\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,|\mathbf{V}_2\right] = \mathbf{0}$$

and equation (2.18) simplifies to

$$var\left[q(S,\mathbf{V},\theta_\circ)\,|\mathbf{V}_2\right] = \sum_{j=1}^{J}\frac{Q_j}{H_j}\left\{var\left[\mathbf{r}(\mathbf{V},\theta_\circ)\,|\mathbf{V}_2\right]\right\} = \Omega(\mathbf{V}_2)\sum_{j=1}^{J}\frac{Q_j}{H_j} \tag{2.19}$$

and since $\sum_{j=1}^{J}\frac{Q_j}{H_j}$ is constant, it does not affect the variance, and therefore conditional variance of $q(S,\mathbf{V},\theta)$ is equal to conditional variance of $\mathbf{r}(\mathbf{V},\theta)$ in the population.

The second case occurs when despite changes of the variances between strata the structure of correlation remains constant. As an example consider cases like $AR(p)$ or $MA(q)$. If variance-covariance matrix remains same despite stratification then by Equation(2.17) the sample objective function $q(S, \mathbf{V}, \theta)$ has same variance as $\mathbf{r}(\mathbf{V}, \theta)$ in the population. Actually in the next section we assume that the variance-covariance matrix does not change by stratification and then check the simulation results for this case by assuming that the correlation follows $AR(1)$ process.

We also need to check the score of the objective function. In the first appendix we also show that the conditional expected value of the sample gradient vector is

$$\mathbb{E}\left[\nabla_\theta q(S, \mathbf{V}, \theta) \,|\mathbf{V}_2\right] = \mathbb{E}\left[\nabla_\theta \mathbf{r}(\mathbf{V}, \theta) \,|\mathbf{V}_2\right]. \tag{2.20}$$

The right hand of (2.20) is the conditional expected value of the population Jacobian matrix. It leads us to optimal instruments matrix that is a $T \times P$ matrix

$$\mathbf{Z}^*(\mathbf{V}_2) \equiv \{var[q(S, \mathbf{V}, \theta_\circ) \,|\mathbf{V}_2]\}^{-1} E\left[\nabla_\theta q(S, \mathbf{V}, \theta_\circ) \,|\mathbf{V}_2\right] \tag{2.21}$$

Therefore the efficient method of moments estimator GMM solves the sample moment conditions

$$\sum_{i=1}^{N} \mathbf{Z}^*(\mathbf{V}_2)' q\left(\widehat{\theta}\right) = \mathbf{0} \tag{2.22}$$

This is a $T \times P$ matrix.

2.5 Examples

In this section some specific examples are covered that illustrate the theoretical results.

Example 1.

As the first example consider linear model

$$\mathbf{Y}_i = \mathbf{X}_i \theta + \mathbf{U}_i \tag{2.23}$$

where $\mathbf{Y}$ is a $T \times 1$ vector of dependent variables, $\mathbf{X}$ is a $T \times P$ matrix of control variables, θ is a $P \times 1$ vector of parameters and finally $\mathbf{U}$ is a $T \times 1$ vector of error terms. In this example

$\mathbf{r}(\mathbf{X}_i, \mathbf{Y}_i, \theta) = \mathbf{U}_i = \mathbf{Y}_i - \mathbf{X}_i\theta$. We assume strict exogeneity assumption

$$\mathbb{E}(\mathbf{U}|\mathbf{X}) = \mathbf{0} \tag{2.24}$$

and add assumption that variance-covariance function in the population is function of control variables $\mathbf{X}$

$$\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}\right] = \Omega(\mathbf{X}) \tag{2.25}$$

Under multinomial sampling scheme we have

$$q(\mathbf{X}, \mathbf{Y}, S, \theta) = \sum_{j=1}^{J} 1[S=j]\frac{Q_j}{H_j}\mathbf{U} \tag{2.26}$$

and by equations(2.12) and (2.17) conditional expected value and conditional variance of this function are

$$\mathbb{E}[q(\mathbf{X}, \mathbf{Y}, S, \theta)|\mathbf{X}] = \mathbf{0} \tag{2.27}$$

and

$$var[q(\mathbf{X}, \mathbf{Y}, S, \theta)|\mathbf{X}] = \sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}, S=j\right] \tag{2.28}$$

respectively. From (2.28), it is clear that variance matrix is a function of $\mathbf{X}$ and strata. Also conditional expected value of gradient vector in this simple linear model is $\mathbf{x}$ according to (2.20). Therefore optimal choice of instrument according to (2.21) is

$$\left\{\sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}, S=j\right]\right\}^{-1}\mathbf{x} \tag{2.29}$$

And GMM solution that produces efficient estimators solves

$$\sum_{i=1}^{N}\mathbf{x}_i'\left\{\sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}, S=j\right]\right\}^{-1}\sum_{j=1}^{J}1[S=j]\frac{Q_j}{H_j}\mathbf{U}_i = \mathbf{0} \tag{2.30}$$

Computational version of (2.30) can be written as

$$\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}_{ij}'\left\{\sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}, S=j\right]\right\}^{-1}\mathbf{u}_{ij} = \mathbf{0} \tag{2.31}$$

Estimation of $\theta_\circ$ are obtained by solving equation (2.31). These parameters estimations are

$$\hat{\theta}_{wGMM}=\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}'_{ij}\hat{\Omega}^{-1}(\mathbf{x},j)\mathbf{x}_{ij}\right)^{-1}\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}'_{ij}\hat{\Omega}^{-1}(\mathbf{x},j)\mathbf{y}_{ij}\right) \tag{2.32}$$

that looks like a GLS estimator. In (2.32) N_j is the sample size in stratum j and $\hat{\Omega}^{-1}(\mathbf{x},j)$ is an estimation of variance matrix $var\left[q\left(\mathbf{X},\mathbf{Y},S,\theta\right)|\mathbf{X}=\mathbf{x}\right]=\sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}=\mathbf{x},S=j\right]$. To have a clear idea about equation (2.32), as an example, assume that variance matrix (2.28) is a function of gender. In this case we need to obtain $\hat{\Omega}^{-1}(female_i,j)$ and $\hat{\Omega}^{-1}(male_i,j)$ for each stratum j; $j\in\{1,2,\ldots,S\}$. In cases that conditional variance matrix is a function of continuous explanatory variable x_{ij}, one possible solution is to divide it into some interval and then estimate variance matrix for each interval in each stratum j separately. Of course if we know the functional form of relationship between the variance matrix and the explanatory variable x_i in each stratum j, we can improve the efficiency in our model by incorporating this knowledge in the estimation process. For example if we are interested in relationship between saving and income in different states and a theory provides an specific form for the variance matrix that relates changes in second moments of saving to changes in income and other explanatory variables then we are in a situation like weighted least squares that provides more efficient estimators relative to OLS. Note that in weighted least square the reason for weighting observations is to solve heteroskedastisity problem while in models with stratified or complex sampling design we need weights even in homoskedastic cases.

we can summarize the above procedure to find a GMM estimator in panel data models with stratified structure in few simple steps as follows:

1. Obtain a consistant estimation of θ.
2. Obtain residuals $\hat{u}_{ijt}$
3. Estimate $\Omega_j(\mathbf{X})=\mathbb{E}[\mathbf{U}\mathbf{U}'|\mathbf{X},S=j]$ for each stratum$j$. Call them $\hat{\Omega}_j(\mathbf{X})$.
4. Form $\hat{\Omega}(\mathbf{X},S)$ by adding weighted $\hat{\Omega}_j(\mathbf{X})$.
5. By substitute $\hat{\Omega}(\mathbf{X},S)$ in equation (2.32), we obtain $\hat{\theta}_{wGMM}$ which we hope it is more efficent than a pooled estimator.

Obtaining a consistent estimator of θ should not be difficult. Any computer package that allows users to estimate surveys panel data can be used to do the first step.

Example 2.

The second example considers a nonlinear model. Assume

$$\mathbb{E}\left[Y_t|\mathbf{X}_t\right] = m\left(\mathbf{X}_t, \beta_\circ\right), \quad t = 1, \cdots, T \tag{2.33}$$

Here $\{(\mathbf{X}_t, Y_t) : 1, 2, \cdots, T\}$ is the time series observations for a random draw from the cross section population and assumption (2.33) simply means that parametric model for $\mathbb{E}\left[Y_t|\mathbf{X}_t\right]$ has been correctly specified. For example if Y is a count variable, a Poisson QMLE can be used. In this case and in general for $Y \geq 0$ and unbounded from above, the most common conditional mean function is the exponential

$$m\left(\mathbf{X}_t, \beta\right) = \exp\left(\mathbf{X}_t\beta\right) \tag{2.34}$$

where $\mathbf{X}_t$ is $1 \times K$ and contains unity as its first element, and β is $K \times 1$. If we impose the Generalized Linear Models (GLM) assumption then

$$var\left(Y_t|\mathbf{X}_t\right) = \sigma_\circ^2 m\left(\mathbf{X}_t, \beta_\circ\right) = \sigma_\circ^2 \exp\left(\mathbf{X}_t\beta\right), \quad t = 1, 2, \cdots, T \tag{2.35}$$

In this model $\mathbf{r}\left(\mathbf{X}, \mathbf{Y}, \beta\right) = \mathbf{Y} - \exp\left(\mathbf{X}\beta\right) = \mathbf{U}$ and $\mathbf{U}$ is $T \times 1$ vector with elements $U_t = Y_t - \exp\left(X_t\beta\right)$ for $t = 1, 2, \cdots, T$. By multinomial stratified sampling and according to (4.1), sample objective function is

$$q\left(\mathbf{X}, \mathbf{Y}, S, \beta\right) = \sum_{j=1}^{J} 1\left[S = j\right] \frac{Q_j}{H_j}\mathbf{U} \tag{2.36}$$

Its conditional expected value is zero as it shown in general case, and its conditional variance is

$$var\left[q\left(\mathbf{X}, \mathbf{Y}, S, \beta\right)|\mathbf{X}\right] = \sum_{j=1}^{J} \frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X}, S = j\right] \tag{2.37}$$

by (2.17). Conditional expected value of gradient of sample objective function is

$$\mathbb{E}\left[\nabla_{\beta} q(\mathbf{X},\mathbf{Y},S,\beta)\,|\mathbf{X}\right] = \begin{pmatrix} -\mathbf{X}_1 \exp(\mathbf{X}_1\beta) \\ -\mathbf{X}_2 \exp(\mathbf{X}_2\beta) \\ \vdots \\ -\mathbf{X}_T \exp(\mathbf{X}_T\beta) \end{pmatrix}_{T\times K} = \mathbf{R}(\mathbf{X}) \tag{2.38}$$

Then optimal choice of instruments is given by

$$\left\{\sum_{j=1}^{J} \frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X},S=j\right]\right\}^{-1}\mathbf{R}(\mathbf{X}) \tag{2.39}$$

And finally GMM estimators are obtained by solving

$$\sum_{i=1}^{N}\mathbf{R}(\mathbf{X})'\left\{\sum_{j=1}^{J} \frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X},S=j\right]\right\}^{-1}\sum_{j=1}^{J} 1\left[S=j\right]\frac{Q_j}{H_j}\mathbf{U}_i = \mathbf{0} \tag{2.40}$$

Here, one way to approach is to model $\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X},S=j\right]$ similar to the first example hoping to obtain more efficient estimators. However, we can choose a hypothesized structure for the within-panel correlation like generalized estimating equations (GEE) literature. The main idea in GEE approach in panel data models is that under strict exogenity assumption (2.1), even a misspecified model for the conditional variance (2.17) that nevertheless captures key features of the conditional second moments might lead to a more efficient estimator of $\theta_{\circ}$ than an estimator that ignores variances and covariances. Identity matrix is the simplest form of the correlation within the panels that assumes independency or in other words no correlation within panels. Exchangeable correlation matrix is a simple extension to this structure. This matrix looks like

$$\Lambda(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & & \vdots \\ \vdots & & \ddots & \alpha \\ \alpha & & \alpha & 1 \end{pmatrix} \tag{2.41}$$

Here parameter α is a scalar that shows common correlation among observations within the panels.

For an example consider a health study in which the panels are clinics and the observations within the panels i.e. clinics are patients.

If observations within the panels have a natural order it is more reasonable to assume the autoregressive structure for within the panels correlation. In health study case for instance, one can consider that the panels represent patients who are measured over time. In this case an autoregressive process can be a good model for dependency of a patient's health conditions over time. In section 5 we consider the autoregressive structure implied by the $AR(1)$ for the correlation matrix to study a linear model. There are several ways in which we might hypothesize the within-panel correlation. To see more options and examples see Hardin and Hilbe (2003).

By assuming correlation matrix (2.41), and adding GLM assumption (2.35), the variance of sample objective function reduces to

$$var\left[q\left(\mathbf{X},\mathbf{Y},S,\beta\right)|\mathbf{X}\right]=\mathbf{m}\left(\mathbf{X}\right)^{\frac{1}{2}}\Lambda\left(\alpha\right)\mathbf{m}\left(\mathbf{X}\right)^{\frac{1}{2}}\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sigma_j\right)\tag{2.42}$$

where $\mathbf{m}\left(\mathbf{X}\right)$ is

$$\begin{pmatrix} m\left(\mathbf{X}_1,\beta\right) & 0 & \cdots & 0 \\ 0 & m\left(\mathbf{X}_2,\beta\right) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & & 0 & m\left(\mathbf{X}_T,\beta\right) \end{pmatrix}\tag{2.43}$$

and by dropping $\sum_{j=1}^{J}\frac{Q_j}{H_j}\sigma_j$ in (2.42) equation (2.40) changes to (2.44) in the sample

$$\sum_{i=1}^{N}\mathbf{R}\left(\mathbf{x}\right)'\mathbf{m}\left(\mathbf{x}\right)^{\frac{-1}{2}}\Lambda\left(\alpha\right)^{-1}\mathbf{m}\left(\mathbf{x}\right)^{\frac{-1}{2}}\sum_{j=1}^{J}1\left[S=j\right]\frac{Q_j}{H_j}\mathbf{u}_i=\mathbf{0}\tag{2.44}$$

Equation (2.44) can be represented as

$$\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N}\mathbf{R}\left(\mathbf{x}\right)'\mathbf{m}\left(\mathbf{x}\right)^{\frac{-1}{2}}\Lambda\left(\alpha\right)^{-1}\mathbf{m}\left(\mathbf{x}\right)^{\frac{-1}{2}}\mathbf{u}_{ij}=\mathbf{0}\tag{2.45}$$

2.6 The normal linear model: A Monte Carlo investigation

It would be insightful to have a Monte Carlo analysis of a number of examples of stratified sampling in the normal linear model. We consider the simple following model

$$\mathbf{Y}_i = \mathbf{X}_i\theta + \mathbf{U}_i \qquad \mathbf{U}|\mathbf{X} = \mathbf{x} \sim \mathbf{N}\left(\mathbf{0}, \sigma^2\Omega\right), \text{ and } \quad x_{it} \sim N(0,1) \text{ for } i = 1, \cdots, N \tag{2.46}$$

In this simple two-variable linear regression model $\mathbf{Y}_i$ is a $T \times 1$ vector of dependent variable, $\mathbf{X}_i$ is a $T \times 2$ matrix of explanatory variables where the first column is a constant term. $\mathbf{U}_i$ is a $T \times 1$ vector of error terms. The vector of parameters θ has two elements; intercept $\theta_\circ$ and slope θ_1. We set zero and one as true values of the intercept and slope in population respectively. We also assume that the only control variable in the model X_{it} has the normal standard distribution and error term $\mathbf{U}$ has the normal distribution with mean zero and variance $\sigma^2\Omega$, where Ω has a first order autoregression $AR(1)$ structure with parameters ρ and $\sigma^2 = 1$. Under these assumptions variance-covariance matrix is

$$\mathbb{E}\left[\mathbf{U}_i\mathbf{U}_i'|\mathbf{X}_i\right] = \mathbb{E}\left[\mathbf{U}_i\mathbf{U}_i'\right] = \sigma^2\Omega = \begin{pmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{T-1} & & \rho & 1 \end{pmatrix} \tag{2.47}$$

Three strata are considered

$\mathscr{W}_1 = \mathscr{X} \times (-\infty, -0.25)$ and $\mathscr{W}_2 = \mathscr{X} \times (-0.25, 1.5)$ and $\mathscr{W}_3 = \mathscr{X} \times (1.5, \infty)$

That are endogenous. We also consider three exogenous strata that are

$\mathscr{W}_1 = (-\infty, -0.25) \times \mathscr{Y}$ and $\mathscr{W}_2 = (-0.25, 1.5) \times \mathscr{Y}$ and $\mathscr{W}_3 = (1.5, \infty) \times \mathscr{Y}$

In all cases the strata are defined by dividing the population into subpopulations in the first period $t = 1$. $\mathscr{W} = (\mathscr{X}, \mathscr{Y})$ is the population space where $\mathscr{W} \subset \Re^2$. In this example population weights Q_1, Q_2, and Q_3 are known. These weights are obtained from normal distribution with mean zero and variance two in endogenous case and standard normal distribution in exogenous one. The H_s or sampling probabilities are equal to $\frac{N_s}{N}$ for $s \in \{1,2,3\}$. Here N_s is the number of observations from stratum s and N is the sum of the total number of observations in the sample.

We estimate parameters and their asymptotic variances by estimators developed in this paper which we call them weighted GLS and un-weighted GLS and compare them with OLS and weighted pooled OLS.

In this exercise sample objective function is (2.26) in the first example and its variance is equal to

$$\begin{aligned}
var\left[q\left(\mathbf{X},\mathbf{Y},S,\theta\right)|\mathbf{X}\right] &= \sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{U}\mathbf{U}'|\mathbf{X},S=j\right] \\
&= \sum_{j=1}^{J}\frac{Q_j}{H_j}\left(\sigma_1^2\Omega+\sigma_2^2\Omega+\sigma_3^2\Omega\right) \\
&= \left(\sum_{j=1}^{3}\frac{Q_j}{H_j}\sigma_j^2\right)\Omega \qquad (2.48)
\end{aligned}$$

The term $\left(\sum_{j=1}^{3}\frac{Q_j}{H_j}\sigma_j^2\right)$ in (2.48) is a constant and has no effect in estimating the parameters so we drop it for simplicity. Therefore, with these simplifications, $var\left[q\left(\mathbf{X},\mathbf{Y},S,\theta\right)|\mathbf{X}\right]=\Omega$ is the variance matrix in the population which is not a function of control variables $\mathbf{X}$. Of course, by changing assumptions, we can obtain different estimations for the variance. In this example, we consider the simplest case by assuming strong assumptions to make estimation easy to execute.

Weighted GLS estimation of $\theta_\circ$ and θ_1 are obtained by solving equation (2.30) in the first example. These parameters estimations are

$$\hat{\theta}_{wGLS}=\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}_{ij}'\Omega^{-1}\mathbf{x}_{ij}\right)^{-1}\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}_{ij}'\Omega^{-1}\mathbf{y}_{ij}\right) \qquad (2.49)$$

that looks like GLS estimator. In (2.49) N_j is the sample size in stratum j.

To have a good judgement of how much gain a practitioner obtains by using this estimator a comparison between estimators developed in this paper and two other estimators is done by using simulation. The comparison is between weighted GLS estimator equation (2.49) and unweighted GLS estimator which is exactly same as (2.49) but drops the weights $\frac{Q_j}{H_j}$ for all $j's$, and weighted pooled OLS that ignores the correlation over time for each cross section observation i

$$\hat{\theta}_{wOLS}=\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}_{ij}'\mathbf{x}_{ij}\right)^{-1}\left(\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\mathbf{x}_{ij}'\mathbf{y}_{ij}\right) \qquad (2.50)$$

And usual OLS assuming homoscedasticity. We also estimate feasible version of weighted and unweighted GLS estimators and call them weighted and unweighted FGLS estimators respectively. So in total six estimators are evaluated in the simulation.

Also we look at the variance of these estimators to evaluate their efficiency. An appropriate estimator of asymptotic variance of $\hat{\theta}_{wGLS}$ is

$$\widehat{Avar\left(\hat{\theta}_{wGLS}\right)} = \hat{\mathbf{A}}_w^{-1}\hat{\mathbf{B}}_w\hat{\mathbf{A}}_w^{-1} \tag{2.51}$$

Where $\hat{\mathbf{A}}_w = \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} \mathbf{x}'_{ij}\Omega^{-1}\mathbf{x}_{ij}$ and $\hat{\mathbf{B}}_w$ is more complicated

$$\begin{aligned} \hat{\mathbf{B}}_w &= \sum_{j=1}^{J} \frac{Q_j^2}{H_j^2} \sum_{i=1}^{N_j} \mathbf{x}'_{ij}\Omega^{-1}\hat{\mathbf{u}}_{ij}\hat{\mathbf{u}}'_{ij}\Omega^{-1}\mathbf{x}_{ij} \\ &\quad - \sum_{j=1}^{J} \frac{Q_j^2}{H_j^2} \left(\frac{1}{N_j}\sum_{i=1}^{N_j} \mathbf{x}'_{ij}\Omega^{-1}\hat{\mathbf{u}}_{ij}\right)\left(\frac{1}{N_j}\sum_{i=1}^{N_j} \mathbf{x}'_{ij}\Omega^{-1}\hat{\mathbf{u}}_{ij}\right)' \end{aligned} \tag{2.52}$$

If weights are dropped from (2.51) we obtain the estimator of asymptotic variance of unweighted GLS; $\widehat{Avar\left(\hat{\theta}_{uwGLS}\right)}$. And if the variance-covariance matrix Ω is dropped from (2.51) we have estimator of asymptotic variance of weighted pooled OLS in hand.

As it mentioned already, in this exercise we consider variance-covariance matrix with AR(1) structure in the population. As equation (2.48) shows if we assume ρ does not change by changing the stratum then there is only one parameter that we need to estimate i.e. ρ. In practice it is possible to estimate different ρ for different stratum too but for simplicity, in this example, we assume ρ in each stratum is equal to the value of ρ in population. Therefore, it is a constant parameter and not a function of strata. However, there is a point that we should keep in mind. we call our estimators GLS (weighted or unweighted) whenever we use the value of ρ in population, since the equation (2.52) is very similar to well known GLS estimators. However, this naming may cause some confusion, since we do not know the true value of ρ in each stratum. We just “assume” that they are equal to the value of ρ in the population of interest. Now, it should be clear why in tables B.1 to B.16, the estimated parameters from GLS and FGLS are very close in most cases.

We consider four values for correlation parameter . It changes from no correlation $\rho = 0$ to high degree of correlation $\rho = .9$. In between $\rho = 0.1$ and $\rho = 0.5$ are considered. It helps us to see the effect of correlation magnitude on the efficiency gain in this simple exercise.

Tables B.1 to B.4 and B.5 to B.8 in the second appendix summarize the results for cases $T = 2$, and $T = 5$ respectively when stratification is exogenous. In these tables means and their standard errors, and mean squared errors for the intercept and slope for six estimators are reported. When $\rho = 0$, POLS and unweighted GLS (uwGLS) are almost identical as we expected. Under exogenous stratification it has been shown that ignoring stratification does not cause any problem. See for example Manski and McFadden (1981), DuMouchel and Duncan (1983) and Wooldridge (1999, 2001). This is clearly seen in the tables. In both $T = 2$, and $T = 5$ cases OLS that ignores stratification and unweighted GLS and its feasible counterpart are superior to weighted POLS and weighted GLS and their feasible versions; they are consistent and more efficient.

The interesting point, that is actually the main issue of this paper, appears when correlation increases. Now as ρ increases un-weighted GLS and its feasible version which takes correlation into account by considering it in estimation process, shows its efficiency over OLS and weighted POLS that simply ignore correlation. This is especially correct about estimation of slope rather than intercept that does not change very much except in a high degree of correlation.

As an illustration consider exogenous stratification and $T = 2$ and $\rho = 0.9$ in table B.4. In this case the mean of the slope is almost the same in both POLS and un-weighted FGLS, but the latter estimator mean squared error is .0151 compare to .0355 in the former. It shows about 57 percent reduction that is substantial. In case T=5 the cutback is even more and it is about 62 percent that is presented in table B.8.

When ρ decreases to 0.5, the improvement in efficiency is still considerable. In this case and $T = 2$ in table B.3, mean of standard deviation for slope decreases from .0355 for OLS to .0302 for un-weighted FGLS. The mean of standard deviation diminishes about 18 percent when T is 5 (table B.7). Meanwhile the mean of standard deviation for intercept does not show any changes at all when $T = 2$ but it shows a sign of improvement as T increases to 5 albeit not too much.

The simulation results show that in the case of exogenous stratification in a panel data model, un-weighted GLS and its feasible counterpart that consider the structure of variance-covariance matrix in estimation is better than OLS and weighted POLS that simply ignore correlations within each cross-section observation. Another interesting observation is that even weighted GLS in the case of exogenous stratification is getting better in terms of reduction of bias and smaller variances when the degree of correlation increases. It is definitely superior to weighted POLS and its estimation of slope has much smaller variance comparing with POLS when ρ exceeds .5 (Tables B3, B4, B7 and B8).

Also the results show that as correlation increases the difference between standard deviation of mean-presented in parentheses- and mean of standard deviation of intercept increases for OLS estimator that a sign that variance of OLS estimator is inconsistent and the inconsistency raises along with . However the inconsistency for the variance of slope is much lesser and does not show significant variation alongside the change in correlation between observations.

The main challenge is when stratification is based on the endogenous variable. In this case un-weighted estimators are generally inconsistent. Tables B.9 to B.12 and B.13 to B.16 summarize the results when stratification is based on the endogenous variable for cases $T = 2$ and $T = 5$ respectively. As it is expected OLS and un-weighted GLS and its feasible un-weighted counterpart all produce inconsistent estimations for both the slope and intercept. The interesting point is that this inconsistency shrinks for slope but enlarges for intercept by increase in the correlation parameter for both estimators. Moreover OLS gives inconsistent estimation of the variances too although it is reduced by increase in ρ. It can be seen by comparing the standard deviation of mean presented in parentheses and mean of standard deviations.

Results presented in the tables B.9 to B.16 show that in case of endogenous stratification weighted estimators- weighted POLS, weighted GLS and weighted FGLS- are consistent and are almost same in low values of correlation parameter $\rho = 0$ and $\rho = 0.1$. The difference between these estimators are more remarkable when correlation parameter ρ starts growing. For example in table B.11, when $T = 2$ and $\rho = 0.5$, while weighted POLS and GLS are both consistent estimators for the slope, the latter has standard deviation of mean equal to .0365 comparing to the former which is equal to .0414. In case of $T = 5$ this difference is even more considerable (look at table B.15).

Superiority of weighted GLS or its feasible equivalent to other estimators is unambiguously clear if we increase ρ to 0.9. In this case and when T=2, mean of standard deviation of the slope is .0186 which is less than half of the same value for its closest competitor i.e. weighted POLS that is about .0408 (Table B.12). The difference between weighted GLS and the rest of the estimators is even more dramatic when T=5. In Table B.16, we can compare the efficiency of weighted GLS and weighted POLS. Here the mean of the standard deviation of the slope for weighted GLS is just 36 percent of the same value for weighted POLS (.0098 verses .0272). Of course this big advantage of weighted GLS verses weighted POLS are just substantial for the estimation of the slope not the intercept.

In another set of Monte Carlo experiments, we relaxed the assumption that the correlation matrix is same for all strata and estimate the matrix for each stratum. The results are even better; we have estimators with smaller variances although in most cases the variances of the old and the new ones are very close. We also repeated the experiment by changing the covariance matrix structure to the random effect model. The results show that weighted and unweighted GLS estimator are efficient estimators in the endogenous and exogenous stratification respectively. In order to make the appendices shorter we do not report the related tables and results.

Overall the simulations show the way to some tentative conclusions. First, finding more efficient estimators in panel data models with stratified sampling structure and under appropriate assumptions is possible. Depending on whether the stratification is based on exogenous or endogenous variable, the GMM estimators developed in this paper, i.e. unweighed or weighted GLS, outperform OLS or weighted POLS which do not consider the correlations over time within each panel *i*.

Second, this superiority is positively related to the level of correlation of a cross section observation through time. In low level of correlation there is no big advantage of using GMM over OLS or weighted POLS. This is changed when correlation parameter is get bigger.

Also for the same sample size the efficiency gain depends on what the structure of correlation is or what kind of structure is chosen in case of GEE models. The simulation results show that correlation matrix structure affect the amount of reduction in the variances of the estimators.

2.7 Determinants of Family income in the U.S: An Empirical Application

In this section we analyze the determinants of family income and sources that cause family incomes varies across households. We estimate a simple linear model that considers total family income a function of family characteristics like education of head of family, age of the head, gender of the head, marital status of the head and so on. The model is estimated with different methods. These methods are pooled OLS, weighted POLS, feasible GLS and weighted feasible GLS methods developed in this paper to compare the efficiency gain if there exist any.

The source of the data set used in this exercise is the 2003-2009 Panel study of Income Dynamics (PSID). The PSID is a complex longitudinal panel survey that have collected data from the same families and their descendants in United States since 1968. Data has been collected on a wide range of economic, social, demographic, psychological and health factors over the life course and across generations. The sample size has grown from roughly 4,800 families in 1968, to about 7,400 by 2005, and to more than 8,690 families and 24,385 individuals as of 2009 (Heeringa, et al., 2011) . As of 2009, the PSID has information on over 70,000 individuals collected over the past four decades.

the core sample of individuals and their families in PSID is rooted in two distinct samples. The Survey Research Center designed a nationally representative sample that known as the SRC sample. The second sample known as the Survey of Economic Opportunity or SEO sample, drawn mainly from lower income families. An oversample of low-income families was included to provide adequate sample size for investigating poverty related issues. Roughly 18,000 individuals living in 4800 households were members of the original 1968 sample. In 1997, PSID Immigrant Supplement added 511 immigrant families to the core sample to obtain more complete picture from the population and to enhance representativeness.

Individuals in PSID fall in two categories; sample and non-sample persons. By definition a sample person is someone who is either a resident of a PSID original sample family in 1968, or an offspring born to or adopted by a sample individual who is actively engage in the study at the time. The definition of sample persons slightly relaxed in 1994 and allowed a child born to or adopted by a sample person who was not participating in the study to be considered as a sample person. According to Heeringa, et al. (2011a), from 24,385 individuals distributed in 8,690 families in 2009, 17,471 are PSID sample persons and 6,914 are non-sample spouses and family members.

Longitudinal weights are calculated at the beginning of a four year (two wave) cycle. The last cycle began in 2007, and therefore the 2009 weights are just "carry-over" weights. Weights need to adjust for attrition and also changes in family size that happens because of marriage, divorce, death, and other additions of new members. The longitudinal family weight in PSID is the average of the positive individual weights for sample person and zero value weights for non-sample persons in the family. For example if a PSID sample person with an individual longitudinal weight of 100 has spouse who is a PSID non-sample person with assigned weight equal to 0, then the family weight for this two-person family is 50. For more detail on the construction of the PSID longitudinal family and individual weights see Heeringa, et al. (2011a, b).

To study the relationship between income and family characteristics covariates, a simple linear model is considered where dependent variable is total family income or $tfinc$, which is the sum of taxable income of the family head and his wife and other members of the family last year plus social security income of the head, his wife and other members of the family unit. This variable can take negative values that indicate net losses occur as a result of business or farm activities. The model is represented as

$$tfinc_{it} = \mathbf{X}_{it}\beta + v_{it} \tag{2.53}$$

where $\mathbf{X}$ is a vector of family characteristics. Here $v_{it} \equiv c_i + u_{it}$, $t = 1,\ldots,T$ are the composite error, c_i, $i = 1,\ldots,N$ are unobserved heterogeneity, and idiosyncratic errors are u_{it} . Parameters of interest is represented by vector β. The vector $\mathbf{X}$ include the total family wealth ($twealth$), the head's age, age square ($age2$) and age cube ($age3$), $health$ condition of the head, marital status, education level, and employment status of the head, family size ($fsize$), persons less than 6 years of age in the family ($aychild6$), the head's father and mother education levels, race and gender of the head, and number of persons less than 18 years of age (*nchild*) in the family as well as year dummy variables and intercept. The variable *twealth* is constructed as sum of seven asset types, net of debt value plus value of home equity. We also added interaction terms between education level of the head (*edu_hs*) and his *age* and between *edu_hs* and the head's employment status, *unemployed* to the model. Tables B.17 and B.18 provide variables description and summary statistics respectively.

The panel in this empirical study consists of 4 waves (7 years) starting 2003 and ending 2009. The 2003 longitudinal family weights are used. After dropping all observations with missing values and strata with just one panel, the final data set is a balanced panel, contains of 15,672 observations or 3,918 panels distributed between 33 strata.

To estimate family income equation (2.53), seven methods are used. These seven methods are pooled OLS, and weighted pooled OLS that ignore the serial correlation problem, and feasible versions of generalized least squares (GLS) that consider two forms for the serial correlation. The first form is a first-order autoregression $AR(1)$, and in the second form the random effect structure is estimated for unconditional variance matrix of error term v_{it}. the remaining three methods are weighted FGLS discussed in this paper. Beside $AR(1)$ and the random effects, we estimate unrestricted variance matrix of error term i.e. $\hat{\Omega} = N^{-1}\sum_{i=1}^{N}\hat{\hat{\mathbf{v}}}_i\hat{\hat{\mathbf{v}}}_i'$ where the $\hat{\hat{\mathbf{v}}}_i$ is a 4×1 of the pooled OLS residuals. We call these three methods wFGLS_ar1, wFGLS_re and wFGLS_un respectively. We hope to obtain efficiency gain by using the latter estimators to estimate total family income equation (2.53).

The estimation results are presented in Table B.19. Robust standard errors are listed in parentheses. In wFGLS_ar1, wFGLS_re and wFGLS_un standard errors are calculated using equations (2.51) and (2.52). In Table B.19, $\hat{\lambda}$ is a consistent estimation of λ, and λ is

$$\lambda = 1 - \{1/[1 + T(\sigma_c^2/\sigma_u^2)]\}^{1/2} \tag{2.54}$$

where σ_c^2, and σ_u^2 are the variance of c_i, and the variance of u_{it}, respectively. If $\hat{\lambda}$ is close to unity, the random effects (RE) and fixed effects (FE) estimates tend to be close.

We just estimate one variance matrix for all strata, same as the Monte Carlo study case represented in last section. The results show that almost all coefficients have expected signs regardless of the method of estimation. However, depending on which method we use for estimation, their magnitudes widely differ in many cases. For example, in terms of absolute value, the coefficient on *edu_hs* estimated by weighted pooled OLS is -31.082, and the same coefficient drops to just -2.636 when the model is estimated by FGLS_re and rise to -9.913 in FGLS_ar1 case (columns 2, 3 and 5 in Table B.19). These substantial changes in the size of most coefficients are mainly due to weighting. A simple comparison between unweighted FGLS methods in columns 3 and 5 with their weighted counterparts in columns 4 and 6 in Table B.19, shows substantial effect of weighting on size of the coefficients. For instance, consider again coefficient on *edu_hs*. It is about 10 times bigger if the family income equation is estimated by wFGLS_re rather than FGLS_re. Same coefficient is almost 3 times bigger if wFGLS_ar1 is used for estimating the same model instead of FGLS_ar1. As another example consider coefficient on *health*. The size of the coefficient falls almost 50% when the model is estimated by FGLS instead of wFGLS.

The big effect of weighting on estimation should not view unusual. Since PSID purposely oversample low income family and in our model income is the dependent variable, OLS using the stratified sample does not consistently estimate the parameters of the total family income because the stratification is endogenous. This is true for unweighted FGLS estimators i.e. FGLS_ar1 and FGLS_re also. The pooled OLS standard errors are smaller than the weighted pooled OLS ones as we expected. In chapter one we showed that by ignoring stratification, the pooled OLS tends to underestimate standard errors. The standard errors are even smaller in the other two unweighted FGLS estimators as it was expected. Therefore, despite smaller standard errors of the unweighted estimators, the main competition is between the weighted pooled OLS on the one hand and the weighted FGLS methods on the other hand that reflects in columns 2, 4, 6 and 7 in Table B.19.

The main idea in this chapter was to increase efficiency in panel data models with stratified data by considering serial correlation in each panel. Under correct conditional mean specification, even a wrong working correlation matrix that captures key features of the conditional second moments might lead to a more efficient estimator. Comparison between the weighted pooled OLS and the weighted FGLS estimators in Table B.19 shows that standard errors are smaller almost in all cases for latter estimators indeed. The only exceptions are coefficients on father and mother education levels $fedu\_hs$, $medu\_hs$. Reduction in standard errors are considerable. For example, standard error on $twealth$ reduces about 33 by using wFGLS to estimate the family income equation. Standard errors of the rest of coefficients drops between 4% (coefficients on *age* and *nchild*), and about 35% (coefficients on *unemployed*, and *unem.edu_hs*). Three consistent estimators i.e. wFGLS_ar1, wFGLS_re, and wFGLS_un are very stable in estimating almost all coefficients, but it seems that efficiency gain is higher in case of wFGLS_re, and wFGLS_un compare to wFGLS_ar1.

2.8 Conclusion

Efficiency in panel data models where data set comes from stratified sample schemes is investigated in this paper. We start from some conditional moments in the population and then based on works done by Chamberlain (1987) and Newey and McFadden (1994) propose a GMM estimator that takes into account dependency structure within the panels. The result is an efficient GMM estimator that is computationally simple to implement. By estimating covariance matrix for each stratum or even estimating same covariance matrix for all strata we are able to improve efficiency.

Monte Carlo simulation results show that the new estimators that we called them weighted and unweighted GLS (and FGLS) in general do better in compare with ordinary least square or weighted and unweighted pooled OLS that simply overlook dependency in the data. In case of endogenous stratification weighted GLS is the efficient estimator among all, and in case of exogenous stratification dropping weight and using unweighted GLS produce best performance as we expect. Of course the gains of new estimators are smaller when we have weaker correlation structure in the panel. Monte Carlo experiments show that the structure of correlation matrix has affects on efficiency gain.

Also simulation results suggest that by increasing T, the importance and effects of endogenous stratification is reduced. A convincing explanation is that by increasing T, the weight of first period diminishes that makes the sample get closer to simple random sampling. Another interesting finding is that by increasing the degree of correlation ρ, inconsistency declines that can attributed to decrease in degree of freedom movements of observations.

We apply the method to estimate a simple linear model using PSID data. Although PSID has very complex structure including multi-stage stratification, by considering very simple form for the working variance matrix the new estimators decrease standard errors on most coefficients in the model.

Chapter 3
MODEL SELECTION TESTS IN COMPLEX SAMPLES

3.1 Introduction

Using the Kullback-Leibler Information Criterion to measure the closeness of a model to the truth, Vuong (1989) developed a classical approach to model selection. He proposes simple likelihood ratio based statistics for testing the null hypothesis that the competing models are equally close to the true data generating process against the alternative hypothesis that one model is closer. In his approach both, one, or neither of the two competing models is misspecified. He assumes that observations are independent and identically distribute (i.i.d.). All of his tests are based on likelihood ratio principle, and consequently he drives asymptotic distribution of the likelihood ratio statistics that covers both nested, overlapping and non-nested models.

While Vuong's tests are based on i.i.d. assumption, in practice in most large surveys, such as the Current Population Survey (CPS), the Panel survey of Income Dynamics (PSID) and National Survey of Families and households (NSFH) that require stratified and clustered samples, simple random sampling and therefore i.i.d. observations is not the right assumption. In other words, a non-random sampling scheme like Standard Stratified (SS) sampling or Variable Probability (VP) sampling, or complex survey design like CPS does not produce a set of independent, identically distributed random variables. Clearly, the i.i.d. assumption is one of the limitation of the Vuong's model selection tests in case of complex samples. This assumption is restrictive when considering time series data too. Rivers and Vuong (2002) along with Findley (1990, 1991) and Findley and Wei (1993) relax this assumption for time series cases like ARMA models and some dynamic regression models.

Also Vuong's model selection tests cannot be used to differentiate between two econometric models defined by moment conditions, or more generally, between two competing models that are incompletely specified. The second limitation in applying the Vuong's tests happens because they are based on the likelihood function. These tests require that competing models belong to some parametric family of distributions and therefore they must be completely parametrized.

While maximum likelihood method is a widespread method of estimation in econometric studies, there are other common methods of estimations that are used by researchers. Techniques like least absolute deviation, nonlinear least squared, generalized method of moments (GMM), or other extremum estimators are used by researchers for different reasons. This is the third limitation of Vuong's tests.

This paper contributes to the subject by extending Vuong's model selection tests for competing models with stratified multistage cluster sampling. Many data sets used in microeconometrics research are collected by surveys like CPS or PSID that have complex multi stage sampling structure and violate the i.i.d. assumption needed in Vuong's tests. Also, In order to generalize Vuong's results to cases other than MLE, we study the problem in M-estimators framework. Many econometrics estimators are M-estimators including but not limited to linear and nonlinear regression, conditional maximum likelihood including discrete response models.

The paper is organized as follows. In section 3.2, we define two nonnested competing models. In section 3.3, we consider basic framework under standard stratified sampling. We start with standard stratified sampling because it is widely used in practice to divide the population of interest into subpopulations or strata and it gives us a base to extend the results to more complex sampling designs. Section 3.4 introduces tests statistics under SS and VP sampling and also multi stage sampling scheme. Also in section 3.4, I show that the test statistics has normal distribution asymptotically. In section 3.5, we extend the model selection test to panel data models with standard stratification design. An interesting problem is if we need to weight the test statistics when stratification is exogenous. I discuss this point in section 3.6. Section 3.7, shows applications of the tests in two empirical examples. Section 3.8 summarizes the results and conclude.

3.2 The Nonnested Competing Models

Consider the population minimization problem

$$\min_{\theta \in \Theta} \mathbb{E}\left[q(\mathbf{W}, \theta)\right] \tag{3.1}$$

where scalar $q(.)$ denotes an objective function depending on $\mathbf{W}$ and θ and $\mathbf{W}$ is an $M \times 1$ random vector taking values in $\mathscr{W} \subset \mathbb{R}^M$. Data generating process depends on θ which is a $P \times 1$ parameter vector and it belongs to parameter space Θ, and Θ is a subset of Euclidean space $\mathbb{R}^P$ or in other words $\Theta \subset \mathbb{R}^P$. We assume that there is a unique value that minimize population problem (3.1) on parameter space Θ at $\theta_\circ$ called true parameter value that generates the data.

In many applications, the vector $\mathbf{W}$ is partitioned into $\mathbf{W} = (\mathbf{X}, \mathbf{Y})$ where $\mathbf{X}$ and $\mathbf{Y}$ are respectively K and L dimensional vectors with $L + K = M$. We are often interested in some aspect of the conditional distribution of $\mathbf{W}$ given $\mathbf{X}$, such as $\mathbb{E}(\mathbf{Y}|\mathbf{X})$.

Now as Vuong (1989) consider two competing objective functions $q_1(\mathbf{W}, \theta_2)$ and $q_2(\mathbf{W}, \theta_2)$. These two competing functions are nonnested in the sense that neither can be represented as a special case of the other. It is important to have a clear idea about nonnested models. Vuong considers two sets of conditional models $\mathbf{F}_\theta = \{f(y|\mathbf{x}, \theta); \theta \in \Theta\}$ and $G_\gamma = \{g(y|\mathbf{x}, \gamma); \gamma \in \Gamma\}$ and then defines two models nonnested if and only if

$$\mathbf{F}_\theta \cap \mathbf{G}_\gamma = \emptyset \tag{3.2}$$

This definition is more suitable for MLE cases where we have full distribution assumptions about the endogenous variables given the exogenous variables for the two competing models. For more general cases as Wooldridge (2010) we consider the following definition

$$P\left[q_1(\mathbf{W}, \theta_1^*) \neq q_2(\mathbf{W}, \theta_2^*)\right] > 0 \tag{3.3}$$

It means that the two function $q_1(., \theta_1^*)$ and $q_2(., \theta_2^*)$ evaluated at the psuedo-true values θ_1^* and θ_2^* must differ for a nontrivial set of outcomes on $\mathbf{W}_i$ if they are nonnested. By this definition nested models are ruled out as well as other forms of degeneracies.

As the first example assume we have a random variable Y and would like to model $\mathbb{E}(Y|\mathbf{X})$ as a function of the explanatory variables $\mathbf{X}$, a $K \times 1$ vector. W specify two competing models; a linear $q_{i1}(\theta_1) = (Y_i - \mathbf{X}_i\theta_1)^2$ and a nonlinear $q_{i2}(\theta_2) = (Y_i - \exp(\mathbf{X}_i\theta_2))^2$. These models are nonnested if the mean of Y_i given $\mathbf{X}_i$ depends on the nonconstant elements in $\mathbf{X}_i$. Yet if the mean function is independent of $\mathbf{X}_i$, or in other words $\mathbb{E}(Y_i|\mathbf{X}_i) = \mathbb{E}(Y_i)$, then the two models are linear with same constant means. In this case two models are nested and the limiting standard normal distribution for Vuong's type statistic breaks down.

3.3 Basic Framework under Standard Stratified Samples

The population problem is $\min_{\theta\in\Theta} \mathbb{E}\left[q(\mathbf{W},\theta)\right]$ and we assume $\theta_\circ$ uniquely solves the problem. Let $q_1(\mathbf{W},\theta_1^*)$ and $q_2(\mathbf{W},\theta_2^*)$ be the two competing models where both may be misspecified. The null hypothesis is

$$\mathbf{H}_0 : \mathbb{E}\left[q_{i1}(\mathbf{W}_i,\theta_1^*)\right] = \mathbb{E}\left[q_{i2}(\mathbf{W}_i,\theta_2^*)\right] \tag{3.4}$$

Depending on what method we use to estimate these two competing models, the alternative hypothesis is

$$\mathbf{H}_{A_{q_1}} : \mathbb{E}\left[q_{i1}(\mathbf{W}_i,\theta_1^*)\right] > \mathbb{E}\left[q_{i2}(\mathbf{W}_i,\theta_2^*)\right] \tag{3.5}$$

or

$$\mathbf{H}_{A_{q_2}} : \mathbb{E}\left[q_{i1}(\mathbf{W}_i,\theta_1^*)\right] < \mathbb{E}\left[q_{i2}(\mathbf{W}_i,\theta_2^*)\right] \tag{3.6}$$

For example if the competing estimators are QMLEs, then the alternative $\mathbf{H}_{A_{q_1}}$ means $q_1(.)$ is better than $q_2(.)$ because its value of the likelihood function is bigger than the other.

To test the null (3.4) against alternative (3.5) or (3.6), in context of complex samples, suppose the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ solve the sample objective function with complex design that involves stratification and clustering. In this section we first consider sample objective function under standard stratified sampling scheme and then consider other types of sampling design. In standard stratified sampling the population of interest is divided into J nonempty, mutually exclusive, and exhaustive strata and then a random sample of size N_j is drawn from stratum j, where $j = 1,\dots,J$. Then for each j, we have random sample $\{\mathbf{W}_{ij} : i = 1,2,\dots,N_j\}$. See Wooldridge (2001). Therefore sample objective function is

$$\sum_{j=1}^{J} Q_j \left(\frac{1}{N_j} \sum_{i=1}^{N_j} q(\mathbf{W}_{ij}, \theta) \right) \tag{3.7}$$

Equation (3.7) can be rephrased as

$$\frac{1}{N} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} q(\mathbf{W}_{ij}, \theta) \tag{3.8}$$

where Q_j is the population frequencies or in other words the probability that a randomly drawn observation from the population falls into stratum j and $H_j \equiv \frac{N_j}{N}$ is the fraction of observations in stratum j. As (3.8) shows in standard stratified sampling observation i in stratum j is weighted by $\frac{Q_j}{H_j}$.

We also assume that $\hat{\theta}_1$ and $\hat{\theta}_2$ converge to θ_1^* and θ_2^* respectively. They are referred to as pseudo true value and are not necessary equal to true value $\theta_\circ$ and therefore the both models may be misspecified.

In order to construct Vuong type test, we need following lemma that shows by assuming $\sqrt{N}$-consistency of $\hat{\theta}_g$ for θ_g for $g = 1,2$ we can find a test statistic that its asymptotic distribution is not affected by the two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.

Lemma 3.3.1. *If $\hat{\theta}_1$ and $\hat{\theta}_2$ are $\sqrt{N}$-consistent estimators for θ_1^* and θ_2^* then*

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\theta}_g) = \frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \theta_g^*) + o_p(1) \tag{3.9}$$

for $g = 1,2$.

Proof. Assuming that $q(.)$ is a differentiable function in respect to θ_g, from a Taylor expansion of $\sum_{i=1}^{J} q(\mathbf{W}_{ij}, \hat{\theta}_g)$ and then dividing both side by N_j we obtain

$$\frac{1}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\theta}_g) \approx \frac{1}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \theta_g^*) + \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)(\hat{\theta}_g - \theta_g^*) \tag{3.10}$$

Multiplied by Q_j and then sum over j, (3.10) can be written as

$$\begin{aligned}\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\theta}_g) \approx &\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \theta_g^*)\\ &+\Big\{\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} \nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)\Big\}(\hat{\theta}_g - \theta_g^*)\end{aligned} \tag{3.11}$$

Finally if we times both side by $\sqrt{N}$ we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\theta}_g) \approx &\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \theta_g^*)\\ &+\Big\{\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} \nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)\Big\}\cdot\sqrt{N}(\hat{\theta}_g - \theta_g^*)\end{aligned} \tag{3.12}$$

In the second term in the right hand side of (3.12)

$$\begin{aligned}plim\sum_{j=1}^{J} Q_j\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)\right) =&plim\left\{\frac{1}{N}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)\right\}\\ =&\mathbb{E}\left[\nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)\right] = \mathbf{0}\end{aligned} \tag{3.13}$$

See Wooldridge (2010). Therefore

$$\sum_{j=1}^{J} Q_j\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta q_g(\mathbf{W}_{ij}, \theta_g^*)\right) = o_p(1) \tag{3.14}$$

and since by assumption $\hat{\theta}_g$ is $\sqrt{N}$-consistent, $\sqrt{N}(\hat{\theta}_g - \theta_g^*) = O_p(1)$. Therefore the second term product in (3.12) is $o_p(1)$ and it can be written as

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\theta}_g) \approx \frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \theta_g^*) + o_p(1) \tag{3.15}$$

This complete the proof. □

Note that the right hand side of equation (3.9) in Lemma 3.2.1 is just a function of random vector $\mathbf{W}_{ij}$. Now we are ready to set up tests statistics similar to Vunge's tests with asymptotic normal distribution under the null hypothesis that the two nonnested competing models are fit equally well.

3.4 Tests Statistics

3.4.1 The Test Statistic under Standard Stratified Sampling

In this section we construct tests statistics that allow us to discriminate between two competing models. Let $q_{ij1}(\mathbf{W}_{ij},\theta_1)-q_{ij2}(\mathbf{W}_{ij},\theta_2)\equiv r_{ij}(\mathbf{W}_{ij},\theta_1,\theta_2)$. Then by Lemma 3.2.1 we have

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)=\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}r_{ij}(\mathbf{W}_{ij},\theta_1^*,\theta_2^*)+o_p(1) \tag{3.16}$$

The following theorem shows that (3.16) under some conditions has asymptotic normal distribution.

Theorem 3.4.1. *For $g\in\{1,2\}$ assume that*

1. $\{\mathbf{W}_{ij}:i=1,2,\ldots,N_j, j=1,\ldots,J\}$ *follows the standard stratified sample scheme.*

2. $N_j\to\infty$ *for each j.*

3. Θ_g *is a compact subset of* $\mathbb{R}^P$.

4. *The objective function* $\mathbb{E}\left[q_g(.,\theta_g)\right]$ *has unique solution on* Θ_g *at* θ_g^*.

5. θ_g^* *is an interior point of* Θ_g.

6. *For each* $\mathbf{w}\in\mathscr{W}$, $q_g(\mathbf{w},.)$ *is continuous on* Θ.

7. $q_g(\mathbf{w},.)$ *is twice continuously differentiable on* Θ.

8. $\mathbb{E}\left[\nabla_\theta q(\mathbf{W},\theta_g^*)q(\mathbf{W},\theta_g^*)'\right]<\infty$ *and* $\mathbb{E}\left[\nabla_\theta q(\mathbf{W},\theta_g^*)\right]=\mathbf{0}$

9. *For all θ, $|\partial^2 q_g(\mathbf{w},\theta_g)/\partial\theta_{gk}\partial\theta_{gm}| \leq b(\mathbf{w})$, all k and m, where $\mathbb{E}[b(w)] < \infty$.*

then

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2) - \sqrt{N}\cdot\mathbb{E}\left[r(\mathbf{W},\theta_1^*,\theta_2^*)\right] \xrightarrow{d} N(0,\eta^2). \tag{3.17}$$

where

$$\eta^2 = \sum_{j=1}^{J}\frac{Q_j^2}{H_j} var[r(\mathbf{W},\theta_1^*,\theta_2^*)|\mathbf{W}\in\mathscr{W}_j] \tag{3.18}$$

Proof. The proof is essentially same as Theorem 3.3 in Vuong (1989) and Theorems 3.1, and 3.2 in Wooldridge (2001). The first assumption shows the diverge from i.i.d. observations assumption in the Vuong model. For the asymptotic analysis, we need second assumption to be sure that the number of observations N_j in each stratum j goes to infinity. The regularity assumptions 2 to 6 are similar to those of Vuong (1989) and we need assumption 8, and 9 since we extend likelihood function to more general one i.e. $q(.)$ function. Also these same regularity assumptions ensures that $\hat{\theta}_g$ is consistent and has normal distribution asymptotically. See Wooldridge (2010). □

Now we have test statistic necessary to choose between two competing models. The null hypothesis is

$$\mathbf{H}_0 : \mathbb{E}\left[q_{i1}(\mathbf{W}_i,\theta_1^*)\right] = \mathbb{E}\left[q_{i2}(\mathbf{W}_i,\theta_2^*)\right] \tag{3.19}$$

against

$$\mathbf{H}_A : \mathbb{E}\left[q_{i1}(\mathbf{W}_i,\theta_1^*)\right] > \mathbb{E}\left[q_{i2}(\mathbf{W}_i,\theta_2^*)\right] \tag{3.20}$$

Under (3.19), $\mathbb{E}\left[r(\mathbf{W},\theta_1^*,\theta_2^*)\right] = 0$ and (3.17) can be written as

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2) \xrightarrow{d} N(0,\eta^2). \tag{3.21}$$

A consistent estimator of η^2 is

$$\begin{aligned}\hat{\eta}^2 \equiv& \sum_{j=1}^{J}\frac{Q_j^2}{H_j}\left\{\frac{1}{N_j}\sum_{i=1}^{N_j}\left[r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2) - \bar{r}_j(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)\right]^2\right\}\\ =&\frac{1}{N}\sum_{j=1}^{J}\frac{Q_j^2}{H_j^2}\sum_{i=1}^{N_j}\left[r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2) - \bar{r}_j(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)\right]^2\end{aligned} \tag{3.22}$$

Here $\bar{r}_j(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)=\frac{1}{N_j}\sum_{i=1}^{N_j} r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)$. Therefore Voung type model selection statistic is

$$\frac{\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)}{\left\{\sum_{j=1}^{J}\frac{Q_j^2}{H_j}\left\{\frac{1}{N_j}\sum_{i=1}^{N_j}\left[r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)-\bar{r}_j(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)\right]^2\right\}\right\}^{1/2}}\xrightarrow{d} N(0,1) \tag{3.23}$$

or

$$\frac{\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)}{\left\{\frac{1}{N}\sum_{j=1}^{J}\frac{Q_j^2}{H_j^2}\sum_{i=1}^{N_j}\left[r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)-\bar{r}_j(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)\right]^2\right\}^{1/2}}\xrightarrow{d} N(0,1) \tag{3.24}$$

3.4.2 The Test Statistic under Variable Probability Sampling

When observations in the strata are difficult to identify prior to sampling, or when collecting information on the variable determining stratification is cheap relative to the cost of collecting the remaining information variable probability sampling is convenient. In variable probability sampling or VP sampling in short, an observation is first drawn at random from the population. If the observation fall into stratum j, it is kept with probability p_j. For example if we need to define stratification in terms of individual incomes, we might draw randomly a person from the population, determine his income class, and then keep him in the sample with a probability that depend on his income class and is set by the researcher.

In variable probability samples, under the null hypothesis that two competing models are equally fit i.e. (3.19), the test statistic is

$$\frac{\frac{1}{\sqrt{N}}\sum_{1=1}^{N}\sum_{j=1}^{J}p_j^{-1}r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)}{\left\{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}p_j^{-2}\left[r_{ij}(\mathbf{W}_{ij},\hat{\theta}_1,\hat{\theta}_2)\right]^2\right\}^{1/2}}\xrightarrow{d} N(0,1) \tag{3.25}$$

that is very similar with what we obtained for standard stratified samples under the null in last section. We just need to replace weights $\frac{Q_j}{H_j}$ with p_j in (3.24). Here we need the sampling probabilities $p_1, p_2, \ldots, p_j$ be all strictly positive. The rest of the assumptions needed to hold this result are same as Theorem 3.3.1. For more details see Wooldridge (1999).

3.4.3 Tests Statistics under Multi-Stage Sampling

Clustering and stratification are main features of survey data. For example National Survey of Families and Households (NSFH), is a complex survey sample. It has multistage design that involves clustering, stratification and variable probability sampling. Clusters are groups of families, households or individuals positioned or occurring a relatively close association. For example in a school, students in each class are form a cluster. In rural areas villages, and in urban areas, neighborhoods are clusters.

The sampling design considered here is closely related to Bhattacharya (2005). In the first stage, the population of interest is divided into S subpopulations or strata. They are exhaustive and mutually exclusive. Within stratum s, there are C_s clusters. In the next step N_s clusters are drawn randomly. Since the asymptotic analysis is based on number of clusters going to infinity, we assume that in each stratum a large number of clusters is sampled. Units (for example households) within each cluster allow for arbitrary correlations. Each sampled cluster c in stratum s contains a finite population of M_{sc} units (for example households) of observations. Finally, for each sampled cluster c in stratum s, randomly sample K_{sc} households with replacement. Sample objective function is

$$\frac{1}{N}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc} q_g\left(\mathbf{W}_{scm}, \theta_g\right) \tag{3.26}$$

for $g = 1,2$. Here $N = N_1 + N_2 + \ldots + N_S$ is the total number of clusters sampled and $v_{sc} = \frac{C_s}{\left(\frac{N_s}{N}\right)}\frac{M_{sc}}{K_{sc}}$ is weight associated with observations $m = 1,\ldots,K_{sc}$ within cluster c within stratum

s. We assume $\dfrac{N_s}{N}$ converges to a_s where a_s is fixed and $0 < a_s < 1$. By this assumption, weights v_{sc} be constant.

By same reasoning as section 3.3 we can show that asymptotic distribution of the following statistic is not affected by estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.

$$\frac{1}{\sqrt{N}}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\cdot r_{scm}\left(\mathbf{W}_{scm},\hat{\theta}_1,\hat{\theta}_2\right) \tag{3.27}$$

Here $r_{scm} = q_{scm1}\left(\mathbf{W}_{scm},\hat{\theta}_1\right) - q_{scm2}\left(\mathbf{W}_{scm},\hat{\theta}_2\right)$ is the difference between the two objective functions for each unit m, in cluster c, in stratum s. Also we can show under the null hypothesis that both competing models equally fit well (3.27) has asymptotic normal distribution

$$\frac{1}{\sqrt{N}}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\cdot r_{scm}\left(\mathbf{W}_{scm},\hat{\theta}_1,\hat{\theta}_2\right) \xrightarrow{d} N(0,\xi^2) \tag{3.28}$$

Because of correlation within clusters, the variance of (3.27), ξ^2 is more complicated than η^2 in 3.4.1. A consistent estimator of ξ^2 is

$$\begin{aligned}\hat{\xi}^2 = &\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}^2 r_{scm}^2\left(\hat{\theta}_1,\hat{\theta}_2\right)\\ &+\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\sum_{m'\neq m}^{K_{sc}} v_{sc}^2 r_{scm}\left(\hat{\theta}_1,\hat{\theta}_2\right) r_{scm'}\left(\hat{\theta}_1,\hat{\theta}_2\right)\\ &-\sum_{s=1}^{S}\frac{1}{N_s}\left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc} r_{scm}\left(\hat{\theta}_1,\hat{\theta}_2\right)\right)^2\end{aligned} \tag{3.29}$$

The first term in (3.29) is a correct estimate of the variance under simple random sampling. Under non-random sampling, it is not true anymore and we need to add the other two terms that are estimations of clustering and stratification effects respectively. In general, in most cases, correlation between unit observation (for example families) in each cluster is positive and therefore the second term appears with a positive sign. On the other hand because of stratification, more homogenous observations are sampled in each stratum that decreases the variance and hence it enters in the formula with negative sign. Therefore, ignoring clustering effect (the second term) causes underestimating the true variance while overlooking stratification effect, we overestimate it.

Extending Bhattacharya's (2005) model to more complex sampling designs, in chapter one, we investigate a sampling design with variable probability sampling in the final stage. The framework resemble complex surveys like NSFH and other routine phone surveys in practice. In this case an appropriate statistic for choosing between two competing models is very similar to (3.29) as follows

$$\begin{aligned}\hat{\xi}^2 = {} & N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is}^2 p_j^{-2} y_{is}\tau_{jm}\tau_{j't}z_{jm}z_{j't}r_{isjm}^2\left(\hat{\theta}_1,\hat{\theta}_2\right) \\ & +N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{m=1}^{K}\sum_{t\neq m}^{K} v_{is}^2 p_j^{-1} p_{j'}^{-1} y_{is}\tau_{jm}\tau_{j't}z_{jm}z_{j't}r_{isjm}\left(\hat{\theta}_1,\hat{\theta}_2\right) r_{isjm'}\left(\hat{\theta}_1,\hat{\theta}_2\right) \\ & -\sum_{s=1}^{S}\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is} p_j^{-1} y_{is}\tau_{jm}z_{jm}r_{isjm}\left(\hat{\theta}_1,\hat{\theta}_2\right)\right] \\ & \times\left[\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{m=1}^{K} v_{is} p_j^{-1} y_{is}\tau_{jm}z_{jm}r_{isjm}\left(\hat{\theta}_1,\hat{\theta}_2\right)\right] \qquad (3.30)\end{aligned}$$

Here in (3.30), v_{is} are weights exactly as (3.29) corresponding to first level of stratification. p_j are weights corresponding to variable probability sampling. Indicator variable τ_{jm} takes value one if observation $\mathbf{W}$ in the second level of stratification (variable probability sampling) is in stratum j and zero otherwise. Indicator variable z_{jm} corresponds to the second level of stratifiaction too. It take value one if $\mathbf{W}$ is kept in the sample and zero otherwise and therefore $P(z=1)=p$. y_{is} is an indicator variable also. It is equal to one if cluster i is in stratum s.

3.5 Model Selection Tests in Panel Data Models

Model selection tests in panel data models with complex sampling designs are similar to the tests in the cross section cases. When $D(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$ is fully specified, the Vuong's approach is directly applicable using MLE. In less restrictive cases when we do not have a complete densities- like partial or pooled MLEs or other M-estimators- we need to account for the time series dependence properly. Assume, for each t, $q_{t1}(\mathbf{W}_t,\theta_1)$ and $q_{t2}(\mathbf{W}_t,\theta_2)$ are competing models of the conditional density in each time period. Here, the same null hypothesis, (3.19), still means the models fit equally well but it is the weakest sense. The convergence result in equation (3.17) still holds under the null. Under assumption (3.3), the models are nonnested and the variance η^2 is positive. In estimating η^2, the serial dependence in $\{q_{it1}(\theta_1^*)-q_{it2}(\theta_2^*)\}$ is a new extra term that must be added in calculations. Let $\hat{r}_{ijt} = q_{ijt1}(\hat{\theta}_1) - q_{ijt2}(\hat{\theta}_2)$ denote the difference in estimated functions for each t, and stratum j, and $\bar{r}_{jt} = N_j^{-1} \sum_{i=1}^{N_j} \hat{r}_{ijt}$. Then, in case of standard stratification, a consistent estimate for η^2 is

$$\hat{\eta}^2 = \frac{1}{N} \sum_{j=1}^{J} \frac{Q_j^2}{H_j^2} \sum_{i=1}^{N_j} \mathbf{1}_T' \mathbf{D}_{ij} \mathbf{D}_{ij}' \mathbf{1}_T \tag{3.31}$$

where $\mathbf{1}_T$ is the $T \times 1$ vector of ones and $\mathbf{D}_{ij}$ is a $T \times 1$ vector defined as

$$\begin{bmatrix} \hat{r}_{ij1} - \bar{r}_{j1} \\ \hat{r}_{ij2} - \bar{r}_{j2} \\ \vdots \\ \hat{r}_{ijT} - \bar{r}_{jT} \end{bmatrix} \tag{3.32}$$

Therefore model selection test in a panel data model with standard stratification design is

$$\frac{\frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} \sum_{t=1}^{T} \hat{r}_{ijt}}{\left\{ \frac{1}{N} \sum_{j=1}^{J} \frac{Q_j^2}{H_j^2} \sum_{i=1}^{N_j} \mathbf{1}_T' \mathbf{U}_{ij} \mathbf{1}_T \right\}^{1/2}} \tag{3.33}$$

Here $\mathbf{U}_{ij}$ is an upper triangular matrix, obtained from $\mathbf{D}_{ij}\mathbf{D}_{ij}'$ by changing values of entries below its diagonal to zero[1]. Test statistic (3.33) has standard normal distribution. Note that in variance estimator (3.31) the mean difference $\bar{r}_{jt}$ varies across t and j but is same across i. If we replace hypothesis (3.19) with the stronger one, $\mathbb{E}\left[q_{it1}(\theta_1^*)\right] = \mathbb{E}\left[q_{it2}(\theta_2^*)\right]$ for $t = 1, \ldots, T$, then we can replace $\bar{r}_{jt}$ with the average of $\hat{r}_{ijt}$ across i and t, $\bar{r}_j$. Here the mean difference $\bar{r}_j$ is just a function of strata.

3.6 Tests Statistics and Exogenous Stratification

It is known that when the population of interest is divided into subpopulations or strata by exogenous variables unweighted estimators are consistent and even more efficient than weighted ones and it does not cause any real problems. However model selection tests are a different matter.

[1] Since $\mathbf{D}_{ij}\mathbf{D}_{ij}'$ is a symmetric matrix, $\mathbf{U}_{ij}$ could be a lower triangular matrix.

Usually, we are interested in cases that a model for some feature of the distribution of Y given $\mathbf{X}$ is correctly specified. Then in correctly specified model, $\theta_\circ$ solves

$$\min_{\theta \in \Theta} \mathbb{E}\left[q(\mathbf{W},\theta)|\mathbf{X}\right] \tag{3.34}$$

for all $\mathbf{x} \in \mathscr{X}$. For example assume we are performing nonlinear least squares on a correctly specified parametric model of $\mathbb{E}(Y|\mathbf{X})$, then in this case $\mathbf{W} = (Y,\mathbf{X})$. In other words our objective function is

$$q(\mathbf{W},\theta) = [Y - m(\mathbf{X},\theta)]^2/2 \tag{3.35}$$

and $\theta_\circ$ is the *true* parameter vector such that

$$\mathbb{E}\left(Y|\mathbf{X} = \mathbf{x},\theta_\circ\right) = m(\mathbf{x},\theta_\circ) \tag{3.36}$$

for all $\mathbf{x}$. Then $\theta_\circ$ solves $\min_{\theta \in \Theta} \mathbb{E}\left[q(\mathbf{Y},\theta)|\mathbf{x}\right] = \mathbb{E}\{[Y - m(\mathbf{X},\theta)]^2/2|\mathbf{x}\}$ for all $\mathbf{x}$. It means that $\theta_\circ$ minimizes $\mathbb{E}\left[q(Y,\theta)|\mathbf{x} \in \mathscr{X}_j\right]$ for each j.

However when the underlying model is misspecified in the sense that $\theta_\circ$, the soultion to (3.1) does not solve (3.34) for each $\mathbf{x}$, the unweighted estimator is not consistent for $\theta_\circ$ while weighted estimator is consistent for $\theta_\circ$.

In model selection tests when the goal is to choose between two nonnested competing models, the null hypothesis (3.19) will only hold if both models are misspesified. If one model were correctly specified, then equality in (3.19) will change to strict inequality in favor of the correctly specified model, assuming the objective functions are not same. For example suppose that in our example above there are two competing models for $\mathbb{E}(Y|\mathbf{x})$, i.e. $m_1(\mathbf{x},\theta_1)$ and $m_2(\mathbf{x},\theta_2)$ that are both misspecified. In this case θ_g^*, $g = 1,2$ does not solve

$$\min_{\theta_g \in \Theta_g} \mathbb{E}\left[q_g(\mathbf{W},\theta_g)|\mathbf{X} = \mathbf{x}\right] \tag{3.37}$$

for all $\mathbf{x} \in \mathscr{X}$, and therefore unweighted estimator is inconsistent for θ_g^*. On the other hand weighted estimator delivers consistent estimator for θ_g^*. Since in the model selection test we need consistent estimators for θ_g^* for $g = 1,2$, we need weight observations appropriately even in case of exogenous stratification.

3.7 Empirical Examples

For illustration purpose, the date set *nhanes*2 provided by Stata is used to contrast two competing models. The data set *nhanes*2 has complex sampling scheme including clustering and stratification. We are interested in modeling the risk of heart attack as a function of variables like *age*, *sex*, *race*, *weight*, and *height* [2]. The dependent variable is *heartatk* that is a binary variable. It is equal to one if the observation has experienced heart attack, and zero otherwise. Two competing models are probit and Bernoulli with contemporary log-log link, estimated by GLM. By ignoring sampling design, the quasi-log likelihood evaluated at relevant estimates for probit and Bernoulli models are -555.028, and -556.665 respectively obtained from 4238 observations. The statistic $\hat{\eta}^2$ turns out to be 6.433, and therefore unweighted Vuong'e test statistic is equal to 0.645. On the other hand, if we consider sampling scheme the values obtained for quasi-log likelihood are -559.393 and -561.889 respectively and our estimation for weighted η^2 is 11.599. Therefore weighted Vuong's statistic in (3.24) is .733. Both tests are in favor of the probit model, and although the weighted Vuong's test is bigger, using a standard normal test at 5% the difference is not statistically significant.

As a second example, consider the determinants of family income in the United States discussed in chapter 2, section 2.7 using panel data set obtained from PSID. We are interested in choosing between two competing models wFGLS-ar1 (model 1) and wFGLS-re (model 2), where in the first model a random effect structure is considered for dependency within panels while in the second one we model this dependecy as AR(1). The null hypothesis is

$$\mathbf{H}_\circ : \mathbb{E}[U_1^2] = \mathbb{E}[U_2^2] \tag{3.38}$$

against alternative

$$\mathbf{H}_A : \mathbb{E}[U_1^2] > \mathbb{E}[U_2^2] \tag{3.39}$$

[2] The complete set of covariates considered in this example are: *houssize*, *age*, *agesq*, *sex*, *height*, *weight*, *iron*, *diabets*, *sizeplace*, *vitaminc*, *zinc*, *copper*, *female*, *black*, *race*, *orace*, *region*1, *region*2, *region*3, *rural*, *highbp*, *highlead*, and *healthstat*. for more information about the data set see http://www.stata-press.com/data/r10/svy.html.

The proper test statistic in this case is (3.33) and its value is about .93 which although is in favor of second model (wFGLS-re), but we cannot reject null hypothesis in favor of alternative at 5% confidence interval.

3.8 Conclusion

In many applied econometric studies researchers are forced to choose between competing models that seems equally well in fitting the data. Model selection tests are suitable tools to distinguish "better" model or models. However Vounge (1989) model selection tests are not readily applicable in cases that data sets come from complex sampling design. In this paper Vounge type tests purpose for the cases that data is not a set of i.i.d. observations due to stratification and clustering. The results show that the test statistics have normal distribution and have to be weighted. An interesting finding is that even in case of exogenous stratification we cannot drop the weights since for nonnested models by null assumption two competing models are misspecified. The tests are applicable for panel data models with complex samples designs but we need to account for time series dependence properly. One advantage of the model selection tests is that they can be obtained easily in empirical studies.

APPENDICES

Appendix A

PROOFS

In first appendix we show that the conditional variance of sample objective function is (2.17).

Proof. Starting point is definition of variance.

$$
\begin{aligned}
& var\left\{\sum_{j=1}^{J} 1\left[S=j\right]\frac{Q_j}{H_j}\mathbf{r}\left(\mathbf{V},\theta_\circ\right)|\mathbf{V}_2\right\} \\
& \quad =\mathbb{E}\left\{\sum_{j=1}^{J} 1\left[S=j\right]\frac{Q_j^2}{H_j^2}\mathbf{r}\left(\mathbf{V},\theta_\circ\right)\mathbf{r}\left(\mathbf{V},\theta_\circ\right)'|\mathbf{V}_2\right\} \\
& \quad =\sum_{j=1}^{J}\int_{\mathbf{v}\in\mathscr{W}} 1\left[S=j\right]\frac{Q_j^2}{H_j^2}\mathbf{r}\left(\mathbf{v},\theta_\circ\right)\mathbf{r}\left(\mathbf{v},\theta_\circ\right)' g\left(s,\mathbf{v}|\mathbf{v}_2\right)d\mathbf{v}
\end{aligned} \tag{A.1}
$$

since stratification is not overlapping we can substitute $1\left[S=j\right]$ with $1\left[\mathbf{v}\in\mathscr{W}_j\right]$. Therefore

$$
\begin{aligned}
& var\left\{\sum_{j=1}^{J} 1\left[S=j\right]\frac{Q_j}{H_j}\mathbf{r}\left(\mathbf{V},\theta_\circ\right)|\mathbf{V}_2\right\} \\
& \quad =\sum_{j=1}^{J}\int_{\mathbf{v}\in\mathscr{W}} 1\left[\mathbf{v}\in\mathscr{W}_j\right]\frac{Q_j^2}{H_j^2}\mathbf{r}\left(\mathbf{v},\theta_\circ\right)\mathbf{r}\left(\mathbf{v},\theta_\circ\right)' g\left(s,\mathbf{v}|\mathbf{v}_2\right)d\mathbf{v} \\
& \quad =\sum_{j=1}^{J}\int_{\mathbf{v}\in\mathscr{W}} 1\left[\mathbf{v}\in\mathscr{W}_j\right]\frac{Q_j^2}{H_j^2}\mathbf{r}\left(\mathbf{v},\theta_\circ\right)\mathbf{r}\left(\mathbf{v},\theta_\circ\right)' \frac{f\left(\mathbf{v}|\mathbf{v}_2,\theta\right)\frac{H_j}{Q_j}}{\sum_{j=1}^{J}\frac{H_j}{Q_j}R\left(j,\mathbf{v}_2,\theta\right)}d\mathbf{v} \\
& \quad =\frac{1}{\sum_{j=1}^{J}\frac{H_j}{Q_j}R\left(j,\mathbf{v}_2,\theta\right)}\sum_{j=1}^{J}\int_{\mathbf{v}\in\mathscr{W}_j}\frac{Q_j}{H_j}\mathbf{r}\left(\mathbf{v},\theta_\circ\right)\mathbf{r}\left(\mathbf{v},\theta_o\right)' f\left(\mathbf{v}|\mathbf{v}_2,\theta\right)d\mathbf{v} \\
& \quad =\eta\sum_{j=1}^{J}\frac{Q_j}{H_j}\int_{\mathbf{v}\in\mathscr{W}_j}\mathbf{r}\left(\mathbf{v},\theta_\circ\right)\mathbf{r}\left(\mathbf{v},\theta_o\right)' f\left(\mathbf{v}|\mathbf{v}_2,\theta\right)d\mathbf{v} \\
& \quad =\ \eta\cdot\sum_{j=1}^{J}\frac{Q_j}{H_j}\mathbb{E}\left[\mathbf{r}\left(\mathbf{V},\theta_\circ\right)\mathbf{r}\left(\mathbf{V},\theta_\circ\right)'|\mathbf{V}_2,S=j\right]
\end{aligned} \tag{A.2}
$$

By dropping the constant term η in (A.2), we obtain equation (2.17). This complete the proof. □

To show that $\mathbf{R}_{\circ}(\mathbf{V}_2)$ the conditional expectation of gradient in sample is same as gradient of the objective function in population we start from definition:

$$\begin{aligned}
\mathbf{R}_{\circ}(\mathbf{V}_2) = & \sum_{j=1}^{J} \mathbb{E}\left\{1[S=j]\frac{Q_j}{H_j}\nabla_{\theta}\mathbf{r}(\mathbf{V},\theta_{\circ})|\mathbf{V}_2\right\} \\
= & \sum_{j=1}^{J} \int_{\mathbf{v}\in\mathscr{W}} 1[s=j]\frac{Q_j}{H_j}\nabla_{\theta}\mathbf{r}(\mathbf{v},\theta_{\circ})g(s,\mathbf{v}|\mathbf{v}_2)d\mathbf{v} \\
= & \frac{1}{\sum_{j=1}^{J}\frac{H_j}{Q_j}R(j,\mathbf{v}_2,\theta)}\mathbb{E}[\nabla_{\theta}\mathbf{r}(\mathbf{V},\theta_{\circ})|\mathbf{V}_2] \\
= & \eta\cdot\mathbb{E}[\nabla_{\theta}\mathbf{r}(\mathbf{V},\theta_{\circ})|\mathbf{V}_2] \qquad \text{(A.3)}
\end{aligned}$$

Appendix B

TABLES[1]

Table B.1: Exogenous Stratification with $\rho = 0.0$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	.9999	.9991	.9999	.9991	.9999	.9991
	(.0353)	(.0431)	(.0353)	(.0431)	(.0354)	(.0432)
$s_{\hat{\beta}_1}$	.0355	.0443	.0353	.0443	.0353	.0443
$\hat{\beta}_\circ$	.0014	.0008	.0014	.0008	.0014	.0008
	(.0426)	(.0471)	(.0426)	(.0471)	(.0426)	(.0471)
$s_{\hat{\beta}_\circ}$	.0422	.0473	.0420	.0473	.0420	.0473

$\hat{\rho} = .0016(.0610)$ in feasible cases.

Table B.2: Exogenous Stratification with $\rho = 0.1$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	.9991	1.0006	.9991	1.0004	.9989	1.0002
	(.0354)	(.0446)	(.0352)	(.0444)	(.0354)	(.0445)
$s_{\hat{\beta}_1}$	.0355	.0446	.0351	.0443	.0350	.0443
$\hat{\beta}_\circ$	-.0006	.0011	-.0005	.0011	-.0005	.0011
	(.0449)	(.0492)	(.0449)	(.0492)	(.0449)	(.0492)
$s_{\hat{\beta}_\circ}$	.0422	.0495	.0440	.0495	.0440	.0495

$\hat{\rho} = .1021(.0597)$ in feasible cases.

[1] In tables B.1 to B.16 presented in this appendix, rows 2, and 4 are average values of estimated $\beta_\circ$ and β_1 obtained from 1000 simulated samples and the values in parenthesis are their standard deviation. rows 3, and 5 represent average values of estimated standard deviations of the estimators calculated by the formula discussed in chapter 2.

Table B.3: Exogenous Stratification with $\rho = 0.5$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	.9998	.9991	.9997	.9993	.9997	.9993
	(.0349)	(.0431)	(.0304)	(.0377)	(.0305)	(.0377)
$s_{\hat{\beta}_1}$	.0355	.0443	.0302	.0383	.0302	.0383
$\hat{\beta}_\circ$	.0014	.0007	.0014	.0007	.0014	.0007
	(.0513)	(.0575)	(.0513)	(.0575)	(.0513)	(.0575)
$s_{\hat{\beta}_\circ}$	.0422	.0578	.0509	.0578	.0509	.0578

$\hat{\rho} = .5008(.0527)$ in feasible cases.

Table B.4: Exogenous Stratification with $\rho = 0.9$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	.9999	.9995	.9997	.9997	.9997	.9997
	(.0351)	(.0442)	(.0153)	(.0193)	(.0154)	(.0193)
$s_{\hat{\beta}_1}$	.0355	.0442	.0151	.0192	.0151	.0193
$\hat{\beta}_\circ$	.0010	.0003	.0010	.0003	.0010	.0003
	(.0569)	(.0641)	(.0565)	(.0641)	(.0565)	(.0641)
$s_{\hat{\beta}_\circ}$	.0421	.0650	.0568	.0650	.0568	.0650

$\hat{\rho} = .8997(.0267)$ in feasible cases.

Table B.5: Exogenous Stratification with $\rho = 0.0$, 1000 replications

$T = 5$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0001	1.0003	1.0001	1.0003	1.0002	1.0003
	(.0232)	(.0257)	(.0232)	(.0257)	(.0232)	(.0257)
$s_{\hat{\beta}_1}$	.0244	.0271	.0242	.0262	.0242	.0262
$\hat{\beta}_\circ$	.0010	.0016	.0010	.0016	.0009	.0016
	(.0263)	(.0277)	(.0257)	(.0277)	(.0257)	(.0277)
$s_{\hat{\beta}_\circ}$	.0263	.0284	.0262	.0277	.0262	.0277

$\hat{\rho} = -.0003(.0303)$ in feasible cases.

Table B.6: Exogenous Stratification with $\rho = 0.1$, 1000 replications

			$T = 5$			
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0001	1.0003	1.0002	1.0003	1.0002	1.0003
	(.0232)	(.0257)	(.0230)	(.0256)	(.0231)	(.0256)
$s_{\hat{\beta}_1}$	.0244	.0271	.0240	.0260	.0240	.0260
$\hat{\beta}_\circ$	.0010	.0017	.0010	.0017	.0010	.0017
	(.0278)	(.0300)	(.0278)	(.0300)	(.0278)	(.0300)
$s_{\hat{\beta}_\circ}$	.0263	.0307	.0283	.0300	.0283	.0300

$\hat{\rho} = .0995(.0301)$ in feasible cases.

Table B.7: Exogenous Stratification with $\rho = 0.5$, 1000 replications

			$T = 5$			
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0000	1.0001	1.0002	1.0000	1.0004	1.0003
	(.0235)	(.0260)	(.0193)	(.0213)	(.0195)	(.0211)
$s_{\hat{\beta}_1}$	.0244	.0271	.0197	.0213	.0197	.0212
$\hat{\beta}_\circ$	.0014	.0023	.0011	.0019	-.0009	-.0002
	(.0382)	(.0411)	(.0377)	(.0407)	(.0371)	(.0403)
$s_{\hat{\beta}_\circ}$	.0263	.0423	.0383	.0406	.0382	.0405

$\hat{\rho} = .4995(.0258)$ in feasible cases.

Table B.8: Exogenous Stratification with $\rho = 0.9$, 1000 replications

			$T = 5$			
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	.9999	.9999	1.0000	.9999	1.0000	.9999
	(.0234)	(.0263)	(.0089)	(.0097)	(.0089)	(.0097)
$s_{\hat{\beta}_1}$	.0244	.0271	.0088	.0095	.0088	.0095
$\hat{\beta}_\circ$	.0008	.0015	.0004	.0010	.0004	.0010
	(.0532)	(.0577)	(.0524)	(.0569)	(.0524)	(.0569)
$s_{\hat{\beta}_\circ}$	.0263	.0586	.0531	.0564	.0531	.0564

$\hat{\rho} = .8991(.0126)$ in feasible cases.

Table B.9: Endogenous Stratification with $\rho = 0.0$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0710	.9996	1.0710	.9996	1.0709	.9997
	(.0381)	(.0412)	(.0381)	(.0412)	(.0381)	(.0411)
$s_{\hat{\beta}_1}$	.0413	.0417	.0376	.0417	.0375	.0416
$\hat{\beta}_\circ$	.1119	.0005	.1119	.0005	.1119	.0005
	(.0373)	(.0394)	(.0373)	(.0394)	(.0373)	(.0394)
$s_{\hat{\beta}_\circ}$	.0430	.0394	.0375	.0394	.0375	.0394

$\hat{\rho} = -.0012(.0562)$ in feasible cases.

Table B.10: Endogenous Stratification with $\rho = 0.1$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0695	.9997	1.0701	.9996	1.0701	.9997
	(.0382)	(.0412)	(.0379)	(.0411)	(.0379)	(.0410)
$s_{\hat{\beta}_1}$	.0412	.0417	.0374	.0415	.0373	.0414
$\hat{\beta}_\circ$	.1242	.0005	.1241	.0005	.1241	.0005
	(.0386)	(.0407)	(.0386)	(.0407)	(.0386)	(.0407)
$s_{\hat{\beta}_\circ}$	.0429	.0409	.0388	.0409	.0388	.0409

$\hat{\rho} = .0989(.0559)$ in feasible cases.

Table B.11: Endogenous Stratification with $\rho = 0.5$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0643	1.0001	1.0527	.9992	1.0529	.9993
	(.0389)	(.0414)	(.0334)	(.0365)	(.0335)	(.0365)
$s_{\hat{\beta}_1}$	.0413	.0414	.0329	.0364	.0329	.0363
$\hat{\beta}_\circ$	.1732	.0008	.1746	.0007	.1746	.0007
	(.0428)	(.0448)	(.0426)	(.0448)	(.0426)	(.0448)
$s_{\hat{\beta}_\circ}$	.0431	.0453	.0428	.0453	.0428	.0453

$\hat{\rho} = .4996(.0487)$ in feasible cases.

Table B.12: Endogenous Stratification with $\rho = 0.9$, 1000 replications

$T = 2$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0591	1.0007	1.0133	.9994	1.0133	.9994
	(.0404)	(.0419)	(.0174)	(.0193)	(.0175)	(.0193)
$s_{\hat{\beta}_1}$	.0419	.0408	.0170	.0186	.0170	.0186
$\hat{\beta}_\circ$	.2223	.0010	.2278	.0008	.2278	.0008
	(.0459)	(.0478)	(.0451)	(.0478)	(.0451)	(.0478)
$s_{\hat{\beta}_\circ}$	.0437	.0482	.0449	.0482	.0449	.0482

$\hat{\rho} = .9006(.0251)$ in feasible cases.

Table B.13: Endogenous Stratification with $\rho = 0.0$, 1000 replications

$T = 5$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0324	1.0001	1.0324	1.0001	1.0324	1.0001
	(.0240)	(.0261	(.0240)	(.0261)	(.0240)	(.0261)
$s_{\hat{\beta}_1}$	.0260	.0277	.0240	.0269	.0248	.0269
$\hat{\beta}_\circ$	.0469	-.0001	.0469	-.0001	.0469	-.0001
	(.0245)	(.0263)	(.0245)	(.0263)	(.0246)	(.0263)
$s_{\hat{\beta}_\circ}$	.0265	.0272	.0249	.0266	.0249	.0266

$\hat{\rho} = -.0003(.0280)$ in feasible cases.

Table B.14: Endogenous Stratification with $\rho = 0.1$, 1000 replications

$T = 5$						
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0321	1.0000	1.0317	1.0001	1.0317	1.0001
	(.0241)	(.0263)	(.0240)	(.0260)	(.0240)	(.0260)
$s_{\hat{\beta}_1}$	.0260	.0277	.0246	.0267	.0246	.0267
$\hat{\beta}_\circ$	.0523	-.0001	.0550	-.0001	.0550	-.0001
	(.0265)	(.0248)	(.0263)	(.0282)	(.0264)	(.0282)
$s_{\hat{\beta}_\circ}$	.0265	.0294	.0268	.0285	.0268	.0285

$\hat{\rho} = .0996(.0279)$ in feasible cases.

Table B.15: Endogenous Stratification with $\rho = 0.5$, 1000 replications

	$T = 5$					
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0300	.9997	1.0203	.9998	1.0203	.9998
	(.0241)	(.0264)	(.0202)	(.0217)	(.0202)	(.0217)
$s_{\hat{\beta}_1}$	.0261	.0276	.0202	.0219	.0202	.0218
$\hat{\beta}_\circ$	.0921	-.0002	.1019	-.0005	.1019	-.0004
	(.0357)	(.0383)	(.0341)	(.0367)	(.0341)	(.0367)
$s_{\hat{\beta}_\circ}$	.0266	.0394	.0347	.0370	.0347	.0370

$\hat{\rho} = .4991(.0241)$ in feasible cases.

Table B.16: Endogenous Stratification with $\rho = 0.9$, 1000 replications

	$T = 5$					
Average	POLS	wPOLS	uwGLS	wGLS	uwFGLS	wFGLS
$\hat{\beta}_1$	1.0252	.9997	1.0036	.9997	1.0036	.9997
	(.0245)	(.0259)	(.0093)	(.0099)	(.0093)	(.0099)
$s_{\hat{\beta}_1}$	.0267	.0272	.0091	.0098	.0091	.0098
$\hat{\beta}_\circ$	.1955	-.0011	.1980	-.0013	.1980	-.0013
	(.0442)	(.0476)	(.0426)	(.0460)	(.0426)	(.0460)
$s_{\hat{\beta}_\circ}$	.0271	.0486	.0432	.0463	.0432	.0463

$\hat{\rho} = .8993(.0121)$ in feasible cases.

Table B.17: Variables Descriptions

age	the actual age of Head
$aychild6$	1 if age of youngest person in the family is 6 or less
$black$	1 if Head is black
$female$	1 if Head is female
$fsize$	the actual number of persons in the family
$fweight3$	2003 core/immigrant family weight
edu_hs	1 if the highest level of Head's education is completed high school
$fedu\_hs$	1 if the highest level of Head's father education is completed high school
$medu\_hs$	1 if the highest level of Head's mother education is completed high school
$health$	1 if health condition of Head is good, very good or excellent
$married$	1 if Head is married
$nchild$	the actual number of persons currently in the family under 18 years of age
$unemployed$	1 if Head is unemployed
$tfinc$	total family money income last year
$twealth$	sum of values of seven asset types, net of debt value plus value of home equity

Table B.18: Summary Statistics

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
age	15,672	47.45	14.97	16	99
$aychild6$	15,672	0.197	0.398	0	1
$black$	15,672	0.282	0.450	0	1
$female$	15,672	0.250	0.433	0	1
$fsize$	15,672	2.638	1.414	1	10
$fweight3$	15,672	23.38	16.74	0	114.3
edu_hs	15,672	0.464	0.499	0	1
$fedu\_hs$	15,672	0.692	0.462	0	1
$medu\_hs$	15,672	0.702	0.457	0	1
$health$	15,672	0.854	0.353	0	1
$married$	15,672	0.566	0.496	0	1
$nchild$	15,672	0.787	1.128	0	8
$unemployed$	15,672	0.0501	0.218	0	1
$tfinc$	15,672	74.69	111.4	-99.26	6,317
$twealth$	15,672	310.0	1,201	-2,700	50,475

Table B.19: Determinants of Family Income in the U.S

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	POLS	wPOLS	FGLS_re	wFGLS_re	FGLS_ar1	wFGLS_ar1	wFGLS_un
twealth	0.036**	0.033**	0.011**	0.021**	0.013**	0.022**	0.019**
	(0.006)	(0.008)	(0.004)	(0.006)	(0.004)	(0.006)	(0.006)
edu_hs	-19.194**	-31.082**	-2.639	-25.548**	-9.913*	-27.669**	-26.751**
	(3.471)	(5.807)	(4.494)	(5.083)	(4.352)	(5.196)	(5.153)
age.edu_hs	-0.061	0.150	-0.501**	-0.005	-0.341**	0.036	0.008
	(0.081)	(0.119)	(0.100)	(0.101)	(0.096)	(0.103)	(0.102)
age	10.774**	12.950**	7.383**	11.821**	8.311**	11.980**	11.832**
	(0.748)	(1.286)	(0.932)	(1.233)	(0.901)	(1.240)	(1.234)
age2	-0.179**	-0.222**	-0.097**	-0.193**	-0.119**	-0.197**	-0.193**
	(0.016)	(0.026)	(0.017)	(0.024)	(0.017)	(0.024)	(0.024)
age3/1000	1.000**	1.000**	0.340**	1.000**	0.480**	1.000**	1.000**
	(0.097)	(0.156)	(0.100)	(0.137)	(0.100)	(0.139)	(0.137)
health	12.772**	13.861**	6.457**	10.803**	5.527**	10.619**	9.880**

Continued on next page

Table B.19 –continued from previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	POLS	wPOLS	FGLS_re	wFGLS_re	FGLS_ar1	wFGLS_ar1	wFGLS_un
	(1.219)	(2.039)	(1.068)	(1.509)	(1.058)	(1.503)	(1.421)
married	25.914**	26.845**	24.413**	29.093**	25.608**	29.349**	29.960**
	(1.788)	(3.022)	(2.168)	(2.870)	(2.129)	(2.865)	(2.861)
fsize	10.017**	12.830**	6.747**	10.396**	5.910**	9.983**	9.484**
	(1.019)	(1.530)	(1.010)	(1.241)	(1.049)	(1.280)	(1.178)
aychild6	-3.365	-5.642	-1.089	-3.201	-2.880	-4.425	-3.923
	(2.227)	(4.793)	(1.913)	(2.774)	(2.100)	(3.395)	(3.156)
unemployed	-22.596**	-27.979**	-7.613**	-16.387**	-6.692**	-15.962**	-14.189**
	(3.427)	(6.441)	(2.669)	(4.880)	(2.419)	(4.863)	(4.755)
unem.edu_hs	18.206**	24.091**	4.781	13.763*	4.742	14.353**	12.432*
	(3.786)	(7.295)	(3.042)	(5.552)	(2.737)	(5.517)	(5.338)
fedu_hs	-12.540**	-12.616**	-14.012**	-13.203**	-13.829**	-12.917**	-12.931**
	(1.793)	(3.344)	(3.157)	(3.503)	(3.204)	(3.474)	(3.498)
medu_hs	-7.334**	-8.722*	-9.418**	-9.820**	-9.070**	-9.663**	-9.671**
	(1.872)	(3.478)	(3.181)	(3.544)	(3.372)	(3.584)	(3.610)

Continued on next page

Table B.19 –continued from previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	POLS	wPOLS	FGLS_re	wFGLS_re	FGLS_ar1	wFGLS_ar1	wFGLS_un
black	-11.494**	-9.478**	-16.158**	-11.550**	-16.635**	-11.657**	-12.137**
	(0.985)	(2.063)	(1.538)	(1.879)	(1.460)	(1.898)	(1.912)
female	-7.389**	-6.304*	-12.720**	-8.815**	-12.680**	-8.666**	-9.010**
	(1.287)	(2.872)	(1.727)	(2.418)	(1.698)	(2.432)	(2.356)
nchild	-9.119**	-10.604**	-7.618**	-9.830**	-5.289**	-7.964**	-7.774**
	(1.263)	(1.991)	(1.172)	(1.645)	(1.464)	(2.056)	(1.917)
year05	2.015	2.537	2.994**	3.007*	2.913**	2.965*	3.058*
	(1.903)	(1.446)	(1.049)	(1.315)	(1.085)	(1.328)	(1.321)
year07	5.409**	6.344**	7.656**	7.615**	7.488**	7.520**	7.773**
	(2.031)	(1.910)	(1.190)	(1.597)	(1.302)	(1.631)	(1.618)
year09	11.822**	12.816**	12.772**	13.240**	12.721**	13.155**	13.280**
	(2.387)	(1.930)	(1.616)	(1.836)	(1.614)	(1.831)	(1.826)
Constant	-146.552**	-185.597**	-86.376**	-163.858**	-96.402**	-165.917**	-162.316**
	(10.626)	(18.300)	(15.406)	(18.229)	(14.094)	(18.201)	(18.251)

Continued on next page

Table B.19 –continued from previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	POLS	wPOLS	FGLS_re	wFGLS_re	FGLS_ar1	wFGLS_ar1	wFGLS_un
Observations	15,672	15,672	15,672	15,672	15,672	15,672	15,672
Number of fid		3,918	3,918	3,918	3,918	3,918	3,918
R-squared	0.286	0.282					
$\hat{\lambda}$			0.63	0.40			
$\hat{\rho}$					0.69	0.45	

Robust standard errors in parentheses

** $p<0.01$, * $p<0.05$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Bhattacharya, D. (2005): “Asymptotic Inference from Multi-Stage Samples” Journal of Econometrics, 126, 145-171.

[2] Cameron, A.C., Pravin, K.T. (2005): “Microeconometrics Methods and Applications” Cambridge University Press, New York, NY.

[3] Chamberlain, G. (1987): “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions” Journal of Econometrics, 34, 305-334.

[4] Cosslett, S.R. (1981a): “Efficient Estimation of Discrete Choice models” In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometrics Applications. MIT Press, Cambridge, MA.

[5] Cosslett, S.R. (1981b): “Maximum Likelihood Estimators for Choice-Based Samples” Econometrica 49, 1289-1316.

[6] Cosslett, S.R. (1993): “Estimation from Endogenously Stratified Samples” In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics, vol. 11, 1-43

[7] Findley, D. F. (1990): “Making Difficult Model Comparisons” mimeo, U.S. Bureau of the Census.

[8] Findley, D. F. (1991): “ Convergence of finite multistep predictors from incorrect models and its role in model selection” Note di Matematica XI, 145-55.

[9] Findley, D. F., Wei, C.Z. (1993): “ Moment bound for deriving time series CLT’s and model selection procedures” Statistica Sinica 3, 453-80.

[10] Hardin, J.H., Hilbe, J.M. (2003): “Generalized Estimating Equations” Chapman & Hall/CRC.

[11] Johnson D.R., Elliott L.A. (1998): “Sampling Design Effects: Do They Affect the Analyses of Data from the National Survey of Families and Households?” Journal of Marriage and Family, 60, 993-1001.

[12] Heeringa, S.G., Berglund, P.A., Khan, A. (2011): “Construction and Evaluation of the 2009 Longitudinal Individual and Family Weights” Panel Study of Income Dynamics Technical Report. Survey Research Center, University of Michigan, Ann Arbor.

[13] Heeringa, S.G., Berglund, P.A., Khan, A., Lee, S., Gouskova, E. (2011): “PSID Cross-sectional Individual Weights, 1997-2009” Panel Study of Income Dynamics Technical Report. Survey Research Center, University of Michigan, Ann Arbor.

[14] Hausman, J.A., Wise, D.A. (1981): “Stratification on an endogenous variable and estimation: The Gary income maintenance experiment” In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometrics Applications. MIT Press, Cambridge, MA, 365-391.

[15] Imbens, G. W. (1992): “An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling” Econometrica, 60, 1187-1214.

[16] Imbens, G. W., Lancaster, T. (1996): “Efficient Estimation and Stratified Sampling” Journal of Econometrics, 74, 289-318.

[17] Manski, C.F., Lerman, S. (1977): “The Estimation of Choice Probabilities from Choice-Based Samples” Econometrica, 45, 1977-1988.

[18] Manski, C.F., McFadden, D. (1981): “Alternative Estimators and Sample Desighns for Discrete Choice Analysis” In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometrics Applications. MIT Press, Cambridge, MA, 2-50.

[19] Newey, W.K., McFadden, D. (1994): “Large Sample Estimation and Hypothesis Testing” In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV, Amsterdam: North Holland, 2111-2245.

[20] Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Institute for Social Research, Survey Research Center, University of Michigan, Ann Arbor, MI (2012).

[21] Rivers, D., Vuong, Q. (2002): “Model Selection Tests for Nonlinear Dynamic Models” The Econometrics Journal, 5, 1-39

[22] Tripathi, G. (2011): “Moment-Based Inference with Stratified Data” Econometric Theory, 27,47-73.

[23] Vuong, Q. (1989): “Likelihood Ratio Tests for Model Selection and Non-Nested Hypothese” Econometrica, 57, 307-333

[24] Wooldridge, J.M. (1999): “Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples” Econometrica, 67 (6), 1385-1406.

[25] Wooldridge, J.M. (2001): “Asymptotic Properties of Weighted M-Estimators for Standard Stratifed Samples” Econometric Theory, 17, 451-470.

[26] Wooldridge, J.M. (2008): “Cluster and stratified sampling” Imbens / Wooldridge BEA/FTC Lectures, Lecture notes 7 & 8.

[27] Wooldridge, J.M. (2010): “Econometric Analysis of Cross-Section and Panel Data” (2nd ed.) MIT Press, Cambridge, MA.