Three Essays on Applied Econometrics

by

Jinwen Xu

B.A., Shanghai Jiao Tong University, 2005
M.A., The University of British Columbia, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Economics)

The University Of British Columbia
(Vancouver)

August 2014

Abstract

This dissertation consists of three chapters. Chapter 1 investigates how returns to education are related to occupation choices. Specifically, I investigate the returns to attending a two-year college and a four-year college and how these returns to education differ from a blue-collar occupation to a white-collar occupation. To address the endogenous education and occupation choices, I use a finite mixture model. I show how the finite mixture model can be nonparametrically identified by using test scores and variations in wages across occupations over time. Using data taken from the National Longitudinal Survey of Youth (NLSY) 1979, I estimate a parametrically specified model and find that returns to education are occupation specific. Specifically, a two-year college attendance enhances blue-collar wages by 24% and white-collar wages by 17% while a four-year college attendance increases blue-collar wages by 23% and white-collar wages by 30%. Chapter 2 and Chapter 3 study how to perform econometric analysis with complex survey data, which is widely used in large scale surveys. Although it is attractive in terms of sampling costs, it introduces complication in statistical analysis, when compared with the simple random sampling method. In Chapter 2, I study the properties of M-estimators when they are used with complex survey data. To undo the over- and under-representation effects of the complex survey design, it is typically necessary to use the survey weight in M-estimation. I establish the consistency and asymptotic normality of the weighted M-estimators. I also discuss how to estimate the asymptotic covariance matrix of the M-estimators. Further, I demonstrate serious consequences of ignoring the survey design in M-estimation and inference based on it. In Chapter 3, I consider specification testing with complex survey data. Specifically, I modify the standard m-testing framework to propose a new method to test if a given model is correct for a subpopulation. The proposed test has advantages over the standard m-testing, taking account of likely heterogeneity of subpopulation distributions. All of the three chapters deal with heterogeneity of subpopulation distributions, whether or not the subpopulation identity is known (Chapter 2 and Chapter 3) and unknown (Chapter 1).

Preface

Chapter 2 and Chapter 3 are joint works with Professor Shinichi Sakata. As part of the training and under the guidance of Professor Sakata, I was involved in all stages of the research and other aspects of the analysis.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere gratitude to all those who supported me during the PhD program at the University of British Columbia. Without their help, I would not be able to complete this dissertation. First, I would like to thank my thesis supervisor Professor Hiroyuki Kasahara for his exceptional guidance throughout the research process. Without him, I would not be able to go through the most difficult days when my previous thesis supervisor left UBC and when I met great frustrations in my research. I also would like to thank my thesis committee members Professor W. Craig Riddell and Professor Thomas Lemieux. Professor W. Craig Riddell opened the door towards academic research for me. It was his course ”Policy Evaluation and Research Design” that inspired my passion for economics and motivated me to pursue the doctoral degree in economics. He provided not only great help in research but also invaluable emotional support during my PhD training. Professor Thomas Lemieux was always there to answer my questions and concerns. His constructive advice and feedback greatly improved the quality of my work. I am very grateful for his precious support when I felt helpless and depressed under the great pressure of job searching. I thank Professor Shinichi Sakata for his meticulous supervision of my research when he was at UBC. I have learnt a lot about academic writing from working together with him on the second and third chapters of my dissertation. Another person I would like to extend my sincere appreciation to is Professor Lang Wu. Discussions with him were of great help deepening my understanding of sophisticated statistical models. His encouragement relieved my anxiety and helped to keep my chin up. I would also like to thank Professor Henry Siu for his kind help as the Graduate Director when I faced difficulties in both research and personal life in my third year in the PhD program. Big thanks to

Professor Kevin Song, Professor Vadim Marmer, Professor Paul Shrimpf, Professor David Green, Professor Kevin Milligan, Professor Nicole Fortin, Professor Florian Hoffmann and all the other attendants in the econometrics lunch seminars and the empirical lunch seminars for their helpful suggestions and feedback to my research and presentations . I thank the Canadian Labour Market and Skills Researcher Network (CLSRN) for the financial support. Thanks to Maureen Chin and Kathy Meredith for their diligently taking care of all of the administrative aspects of the PhD and the job market. Special thanks to Nishant Chadha, Xiaodan Gao and Haimin Zhang for sharing laughters and tears together, and witnessing each other becoming a maturer and better self. Lastly, I would like to thank my parents for their unconditional love, for their endless support, for always making me feel I am home on the phone calls even though we are whole Pacific Ocean away.

Chapter 1

Returns to Education and Occupation Choices

1.1 Introduction

The association between education and earnings is perhaps the most well-documented and studied subject in social science. Much recent work by economists has investigated the extent to which this correlation is causal in nature (see Card, 2001, for a recent survey of this literature, also see Heckman *et al.* , 2006a). However, how returns to education are related to the choice of occupations has received less attention.

It is plausible that returns to education are occupation specific. Returns to education are found to be lower in secondary sector occupations (Blaug, 1985, Dickens and Lang, 1985) and in occupations which do not require the education that one obtains(Duncan and Homan, 1981,Sicherman, 1991). In this chapter, I examine how returns to attending a two-year college and a four-year college differ from those of a blue-collar occupation to those of a white-collar occupation, using the National Longitudinal Survey of Youth (NLSY) 1979. Intuitively, the wage premium for high school graduates attending a two-year college may be higher in a blue-collar occupation such as that of a machinist than it may be in a white-collar occupation such as that of a manager while the wage premium for high school graduates attending a four-year college may be higher in a white-collar occupation than it may be in a blue-collar occupation.

The main complication of estimating the occupation-specific returns to education comes from the endogenous education and occupation choice. As in Roys model (Roy, 1951), individuals are endowed with different abilities to work in a blue-collar occupation or a white-collar occupation. They tend to work in the occupation in which they have a comparative advantage. Moreover, occupation abilities can also influence the education choice. For example, individuals who know that they are more likely to work in a white-collar occupation are more likely to attend a four-year college, which would increase the white-collar wages they would earn more so than attending a two-year college would. In addition, individuals vary in their education psychic costs. Those with lower education psychic costs may obtain more education than may those with higher education psychic costs (Willis and Rosen, 1979; Willis, 1986; Carneiro *et al.* , 2003). While the occupation abilities and the education psychic costs are known to individuals making education and occupation decisions, these abilities and costs are unobserved by the econometrician. In the presence of self-selection in both education and occupation, the Ordinary Least Squares (OLS) estimates of occupation-specific returns to education are biased. One traditional way of dealing with the endogeneity issue in the returns to education literature is to use compelling instruments for education such as institutional rules or natural experiments (see Card, 2001, for a survey of papers using IV approach in this literature). However, the standard IV approach is hard to implement here because it is difficult to find good instruments for both education and occupation choices.

I address the issue of endogeneity in education and occupation by explicitly modelling the sequential education and occupation choices. The unobserved occupation abilities and education psychic costs are specified with a flexible multinomial distribution in a finite mixture model. Departing from previous papers that use a finite mixture model to tackle the endogeneity issue in the education literature, I achieve nonparametric identification of the finite mixture model without imposing parametric assumptions on the joint distribution of wages, education, and occupation choices. Based on Kasahara and Shimotsu (2009) and Kasahara and Shimotsu (2012), I rigorously show how to nonparametrically identify the occupation abilities using the variations in wages across occupations over time. Since the information from the panel data alone is not enough to identify the unobserved education psychic costs, I bring in additional data. Specifically, I use scores from four tests (math skills, verbal skills, coding speed, and mechanic comprehension) conducted by the Armed Force Vocational Aptitude Battery(ASVAB), together with the Rotter Locus of Control test score and the Rosenberg Self-Esteem Scale. I show that conditional on occupation abilities and education psychic costs the education psychic costs can be nonparametrically identified under the assumption that the test scores do not directly affect wages, education, or occupation choices. My identification strategy allows the unobserved occupation abilities and education psychic costs to be freely correlated. Carneiro *et al.* (2003), Hansen *et al.* (2004), Heckman *et al.* (2006b), and Cunha and Heckman (2008) also use test scores to identify their mixture model. However, they assume the unobserved variables to be mutually independent. Cunha *et al.* (2010) relax the strong independence assumption, but their identification replies on the assumption that distributions are bounded complete. My identification strategy does not require this strong rank condition.

While I show that the finite mixture model can be nonparametrically identified, estimating the high-dimensional model nonparametrically is nearly impossible given the relatively small sample size of NLSY 1979. Therefore, I impose some parametric forms in wage, education, occupation, and test scores to facilitate the estimation. I find that attendance of a two-year college enhances blue-collar wages by 24% and white-collar wages by 17%. Therefore, attendance of a two-year college helps accumulate more blue-collar skills than it does white-collar skills. The reverse holds true for attendance of a four-year college, which increases blue-collar wages by 23% and white-collar wages by 30%.

This chapter is the first to quantify the occupation-specific returns to attending a two-year and a four-year college. Although many papers have estimated returns to a two-year college and a four-year college (Kane and Rouse, 1995; Grubb, 1997; Light and Strayer, 2004; Marcotte *et al.* , 2005), these papers assumed that returns to education are homogeneous across occupations. The occupation-specific returns to education suggest that analyzing the potential impact of an education policy, such as tuition subsidy, requires consideration of individuals possible occupation choices when these individuals have finished school because returns to education depend on their subsequent occupation choices.

Moreover, this paper helps us understand the choice made between attendance of a two-year and that of a four-year college. I find that individuals make their post-secondary education choices based on both occupation abilities and education psychic costs. The idea that individuals invest in education based on their occupation abilities was first raised in a seminal paper by Willis and Rosen (1979). Willis and Rosen studied the choice made by high school graduates between entering the labour market or attending a college; they suggest that individuals who are more suitable to the college labour market are more likely to attend a college. Keane and Wolpin (1997) extend Willis and Rosen (1979) by taking into account the sequential choices of education and occupation. They studied how individuals with different occupation abilities make year-by-year decisions as to whether to further their education. My paper departs from that of Keane and Wolpin by bringing additional data, the test scores, to achieve nonparametric identification of the education psychic costs. I find that the education psychic costs play an important role in post-secondary decisions. This is consistent with the findings in Carneiro *et al.* (2003). Carneiro *et al.* (2003) extend the model of Willis and Rosen (1979) to account for the education psychic costs and use the ASVAB scores to identify the education psychic costs. They find that individuals decide whether to attend college or not taking into account education psychic costs. In addition, I show that without considering the selection based on the unobserved education psychic costs, the returns to attending a two-year college are biased upward.

The rest of the paper proceeds as follows. Section 2 describes the data. Section 3 discusses the empirical specifications. Section 4 shows the nonparametric identification of the finite mixture model. Section 5 reports the empirical results, and Section 6 concludes.

1.2 Data

This paper uses data taken from the NLSY79. The NLSY79 is a U.S. national survey of 12686 young men and women who were 14-22 years old in 1979. It consists of a core random sample of civilian youths, a supplemental sample of minority and economically disadvantaged youths and a sample of youths in the military. The analysis is based on the 2439 male respondents in the core random sample. The individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis. I use the observations from 1979 to 1994.

The NLSY79 collects information on individuals' education attainment and the type of post-secondary education individuals in which were enrolled. I assign individuals to three educational categories: high school graduates, two-year college attendants and four-year college attendants.[1] High school graduates are those who are reported to have completed at least 12 years of education and have never attended either a two-year college or a four-year college. Two-year college attendants are those who are reported to have enrolled in a two-year college and have never attended a four-year college. Four-year college attendants are the ones who are reported to have enrolled in a four-year college. I distinguish two-year college and four-year college education because a two-year college provides more technical and vocational programs while a four-year college offers more academic and professional programs.

The NLSY79 asks individuals about their occupations and the associated hourly payment in each survey year. I assign individuals to a blue-collar occupation and a white-collar occupation[2] according to the occupation they work the most during the survey year based on one-digit census codes. Blue-collar occupations are (1) craftsmen, foremen, and kindred; (2)operatives and kindred; (3) laborers, except farm; (4) farm laborers and foremen; and (5) service workers. White-collar occupations are (1) professional, technical, and kindred; (2) managers, officials, and proprietors; (3) sales workers; (4) formers and farm managers; and (5) clerical and kindred.

One advantage of the NLSY79 is that many of the respondents were in school when they were first interviewed. Therefore, information about their first jobs are available. Such information about initial conditions is especially useful because it is important to take into account the persistent shocks in wages and occupation choices as pointed out by Hoffmann (2011).

[1] Although the main analysis of this paper is based on these three education groups, I examine the occupation-specific returns to a bachelor's degree because usually college graduates earn more than college dropouts (Jaeger and Page, 1996). However, I do not investigate the occupation-specific returns to an associate degree because the sample size of associate degree earners are too small to give any reasonable estimates.

[2] Although a finer aggregation is possible, I focus on two occupation categories to emphasize the importance of the role of occupational choices in returns to education.

To identify the individual unobserved occupation abilities and education psychic costs, I use ASVAB, which was administrated in 1979, to construct four test scores: math skill, verbal skill, coding speed, and mechanic comprehension. The higher scores indicates higher skills. In addition, I use the Rotter Locus of Control Scale, which was administered in 1979, and the Rosenberg Self-Esteem Scale ,which was administered in 1980. The Rotter Locus of Control Scale measures whether individuals believe that events in their life derive primarily from their own actions. It is normalized to the case that a higher score indicates higher degree of control individuals feel they possess over their life. The Rosenberg Self-Esteem Scale measures perceptions of self worth. A higher score indicates higher self-esteem.[3]

[3] All measures are standardize to mean zero and variance one.

Table 1.1: Discriptive Statistics

	Overall			High School			2-yr College			4-yr College		
Variables	Obs	Mean	S.D	Obs	Mean	S.D	Obs	Mean	S.D	Obs	Mean	S.D
2-yr college attendant	934	0.175	0.380	318	0.000	0.000	163	1.000	0.000	453	0.000	0.000
4-yr college attendant	934	0.485	0.500	318	0.000	0.000	163	0.000	0.000	453	1.000	0.000
Highest grade completed	934	13.943	2.249	318	11.880	0.567	163	13.074	1.034	453	15.700	1.864
Age in 1979	934	17.217	2.082	318	16.447	1.628	163	17.362	2.033	453	17.706	2.223
Initial job(white collar)	934	0.394	0.489	318	0.110	0.313	163	0.264	0.442	453	0.640	0.480
Initial wage	934	11.603	15.738	318	9.879	22.438	163	10.795	5.004	453	13.103	12.024
Initial wage(blue collar)	934	10.249	17.129	318	10.060	23.754	163	10.910	5.234	453	10.089	4.535
Initial wage(white collar)	934	13.684	13.068	318	8.410	3.431	163	10.472	4.340	453	14.797	14.374
Mother education	934	12.269	2.086	318	11.355	1.835	163	12.202	1.919	453	12.934	2.066
Father education	934	12.726	3.038	318	11.226	2.469	163	12.748	2.604	453	13.770	3.111
Number of siblings	934	2.767	1.748	318	3.110	1.850	163	2.755	1.757	453	2.530	1.631
Broken family at age 14	934	0.127	0.334	318	0.151	0.359	163	0.147	0.355	453	0.104	0.305
South at age 14	934	0.239	0.427	318	0.239	0.427	163	0.221	0.416	453	0.245	0.431
Urban at age 14	934	0.730	0.444	318	0.619	0.486	163	0.791	0.408	453	0.786	0.411

Table 1.1 presents the sample summary statistics by the three education groups: high school graduates, two-year college attendants, and four-year college attendants.[4] The sample consists of 934 individuals, of which 34% are high school graduates, 17.5% are two-year college attendants, and 48.5% are 4-year college attendants. On average, the high school graduates complete 11.9 years of schooling. The two-year college attendants finish 13.1 years of school. The complete years of schooling of the two-year attendants suggests that a large fraction of the two-year attendants do not graduate[5]. The four-year college attendants complete 15.7 years of schooling, which suggests that a large proportion of the four-year college attendants obtain a bachelor's degree[6]. The comparison of the fraction of individuals working in a white-collar occupation as their first jobs[7] across the three education groups suggests that the probability of the initial job in a white-collar occupation increases with education: around 11% of the high school graduates, 26.4% of the two-year college attendants, and 64% of the four-year college attendants work initially in a white-collar occupation. The average wages associated with the first jobs as presented in table 1.1 suggest that the higher the education, the higher the wages: on average, the high school graduates earn $9.88, the two-year college attendants earn $10.80, and the four-year college attendants earn $13.10. Further, I look at the blue-collar wages and white-collar wages associated with individuals' first jobs. For those who initially work in a white-collar occupation, I find that higher education are associated with higher wages: the high school graduates earn around $8.41, the two-year attendants earn around $10.47, and the four-year attendants earn around $14.80. However, the relationship between wages and education is different for those who initially work in a blue-collar occupation: the two-year college attendants earn the most among the three education groups, and the high school graduates and the four-year college attendants earn almost the same. The average blue-collar wages of the high school graduates are $10.06, those of the two-year college attendants are $10.91, and those of the four-year college attendants are $10.09. Table 1.1

[4] Table A.1 gives the summary statistics for two-year college dropouts, those with an associate degree, four-year college dropouts, and those with a bachelor's degree.

[5] Around 75% of the two-year attendants do not have an associated degree.

[6] Around 70% of the four-year college attendants obtain a bachelor's degree.

[7] I look at individuals' first jobs to get rid of the impact of work experience on the probability of working in a white-collar occupation.

also shows that the three education groups have quite different family background. Individuals whose parents have more education, who have fewer siblings, grew up in a two-parent family, and live in an urban area at age 14 tend to obtain more education.

Table 1.2 presents the average test scores of the six tests across education groups and occupation groups. Table 1.2a shows that individuals who initially work in a white-collar occupation perform better than those who initially work in a blue-collar occupation in all the six test scores, and therefore, the six test scores may be informative about individuals' occupation abilities. Table 1.2b shows the six test scores increase with education. Further, table 1.2c and table 1.2d show that the six test scores increase with education when conditional on initial occupations. The positive correlation between education attainment and the test scores suggest that the six test scores may be informative about individuals' education psychic costs.

1.3 Empirical Specification

In this section I specify the wage regression in which returns to education are occupation-specific, explicitly model how individuals make their subsequent post-secondary education choices and occupation choices based on their unobserved occupation abilities and education psychic costs, and present the test scores regression specification, which is essential for the identification of the finite mixture model.

To control for the selection in education and occupation, I specify the joint distribution of occupation abilities and education psychic costs by a multinomial distribution in a finite mixture model. A finite mixture model assumes that the overall population consists of M types of people. Each type shares the same occupation abilities and education psychic costs, and different types are different in occupation abilities and/or education psychic costs. I assume that the unobserved types affect the intercepts of the wage regression, the test scores regression, and the expectations on utility in the choice of postsecondary education and occupations. The superscript m in the following equations represents the mth type-specific parameters. The finite mixture model is discussed in more detail in Section 1.3.2.

Table 1.2: Test Scores by Education and Initial Occupation

(a) By Intial Occupation

	Blue Collar			White Collar		
Variable	Obs	Mean	S.D.	Obs	Mean	S.D.
Math skill	566	-0.323	0.961	368	0.497	0.844
Verbal skill	566	-0.279	1.066	368	0.428	0.700
Coding speed	566	-0.231	0.966	368	0.356	0.947
Mechanical	566	-0.065	1.044	368	0.101	0.920
Locus of control	566	-0.099	0.982	368	0.153	1.009
Self-esteem	566	-0.110	1.000	368	0.169	0.978

(b) By Education

	High School			2-yr College			4-yr College		
Variable	Obs	Mean	S.D.	Obs	Mean	S.D.	Obs	Mean	S.D.
Math skill	318	-0.684	0.814	163	-0.236	0.870	453	0.565	0.812
Verbal skill	318	-0.672	1.107	163	-0.068	0.867	453	0.497	0.607
Coding speed	318	-0.430	0.911	163	-0.099	0.955	453	0.338	0.953
Mechanical	318	-0.248	1.096	163	0.049	0.969	453	0.156	0.903
Locus of control	318	-0.256	0.956	163	-0.017	1.026	453	0.186	0.982
Self-esteem	318	-0.325	0.947	163	0.058	0.943	453	0.207	0.999

(c) By Education, Blue-Collar

	High School			2-yr College			4-yr College		
Variable	Obs	Mean	S.D.	Obs	Mean	S.D.	Obs	Mean	S.D.
Math skill	283	-0.708	0.801	120	-0.255	0.925	163	0.294	0.910
Verbal skill	283	-0.734	1.105	120	-0.071	0.855	163	0.359	0.705
Coding speed	283	-0.476	0.916	120	-0.126	0.965	163	0.116	0.934
Mechanical	283	-0.257	1.097	120	0.063	0.959	163	0.294	0.910
Locus of control	283	-0.263	0.971	120	0.006	1.001	163	0.108	0.943
Self-esteem	283	-0.348	0.943	120	0.071	0.969	163	0.171	1.024

(d) By Education, White-Collar

	High School			2-yr College			4-yr College		
Variable	Obs	Mean	S.D.	Obs	Mean	S.D.	Obs	Mean	S.D.
Math skill	35	-0.495	0.899	43	-0.181	0.698	290	0.718	0.708
Verbal skill	35	-0.178	1.004	43	-0.061	0.908	290	0.574	0.531
Coding speed	35	-0.061	0.783	43	-0.022	0.933	290	0.462	0.942
Mechanical	35	-0.178	1.108	43	0.011	1.009	290	0.718	0.708
Locus of control	35	-0.204	0.834	43	-0.079	1.101	290	0.23	1.003
Self-esteem	35	-0.139	0.975	43	0.021	0.877	290	0.228	0.986

1.3.1 A Model for Postsecondary Education, Occupation Choice and Wages

The Model for Wages

Different from the conventional Mincer-type wage specification, I allow returns to attending a two-year and a four-year college to depend on the occupation choice. The log wage, W_{it}, for individual i at time t is as follows,

$$W_{it} = \alpha^m_{W,1} + \alpha^m_{W,2} O_{it} + \beta_1 2\text{YR}_i + \beta_2 4\text{YR}_i + \beta_3 2\text{YR}_i O_{it} + \beta_4 4\text{YR}_i O_{it} + X'_{it}\beta_5 + \varepsilon_{W,it}, \tag{1.1}$$

where 2YR$_i$ is a dummy variable which equals 1 if individual i is a two-year college attendant, 4YR$_i$ is a dummy variable which equals 1 if individual i is a four-year college attendant, and O_{it} is a dummy variable which equals 1 if individual i works in a white-collar occupation at time t. The occupation-specific work experience and its squared terms are collected into X_{it}. Since different occupations reward the occupation-specific work experience differently, X_{it} also includes the interaction terms of the occupation-specific work experience and the occupation dummy variable O_{it}, and the interaction terms of the occupation-specific work experience squared and O_{it}.

The returns to attending a two-year college and a four-year college in a blue-collar occupation are represented by β_1 and β_2 respectively, and the returns to attending a two-year college and a four-year college in a white-collar occupation are denoted by $\beta_1 + \beta_3$ and $\beta_2 + \beta_4$ respectively. Since a two-year college focuses on technical and vocational programs while a four-year college provides academic and vocational programs, we would expect the returns to attending a two-year college to be higher in a blue-collar occupation than a white-collar occupation, i.e. $\beta_3 < 0$, and the returns to attending a four-year college to be higher in a white-collar occupation than a blue-collar occupation, i.e. $\beta_4 > 0$.

The relationship between wages and innate occupation abilities are captured by $\alpha^m_{W,1}$ and $\alpha^m_{W,2}$, which are specific to type m. A large value of $\alpha^m_{W,1}$ means type m has a high blue-collar ability, and a large value of $\alpha^m_{W,1} + \alpha^m_{W,2}$ implies type m has a high white-collar ability. In other words, if type m has a comparative advantage in a white-collar occupation than a blue-collar occupation, we would expect $\alpha^m_{W,2} > 0$.

I assume that productivity shocks $\varepsilon_{W,it}$ follow a first-order Markov process[8]:

$$\varepsilon_{W,it} = \rho \varepsilon_{W,it-1} + \zeta_{it},$$

where $\varepsilon_{W,i1} \overset{iid}{\sim} N(0, \sigma_{W,1})$ and $\zeta_{it}|\varepsilon_{W,it-1} \overset{iid}{\sim} N(0, \sigma_{W,2})$.

The Model for Occupation Choices

In each period, individuals choose to work in either in a blue-collar or a white-collar occupation to maximize life-time income. Let $I_{O,it}$ denote the latent utility associated with a white-collar occupation relative to a blue-collar occupation at time t:

$$I_{O,it} = \alpha_O^m + \lambda_1 2\text{YR}_i + \lambda_2 4\text{YR}_i + \lambda_3 O_{it-1} + X_{it}'\lambda_4 + \varepsilon_{O,it}, \qquad (1.2)$$

where $\varepsilon_{O,it} \overset{iid}{\sim} N(0,1)$. Since the latent utility, I_{it}, depends on wages, all the regressors in Equation (1.1) are included. In addition, the occupation choice at time $t-1$ may affect the occupation choice at time t because job switching costs may prevent individuals from moving from one occupation to another. Such a relationship between the occupation choices at time $t-1$ and time t are captured by the dummy variable, O_{it-1}, which equals to 1 if the job at time $t-1$ is a white-collar occupation.

The type-specific intercept, α_O^m, reflects that the latent utility, $I_{O,it}$, depends on occupation abilities. In other words, holding everything else the same, an individual with a comparative advantage in a white-collar occupation is more likely to work in a white-collar occupation than an individual with a comparative advantage in a blue-collar occupation.

[8] I assume the same productivity shock for a blue-collar and a white-collar occupation. It is because occupation choice in current period is influenced by current wages in blue- and white-collar occupations. The current occupation-specific wages depend on the blue- and white-collar productivity shocks in the last period. However, we only observe the wage associated with the last period occupation an individual worked in. The wage associated with the other occupation is unobserved. So if we consider occupation-specific productivity shocks, we have to integrate out the unobserved productivity shock associated the other occupation. This significantly increase the computation burden.

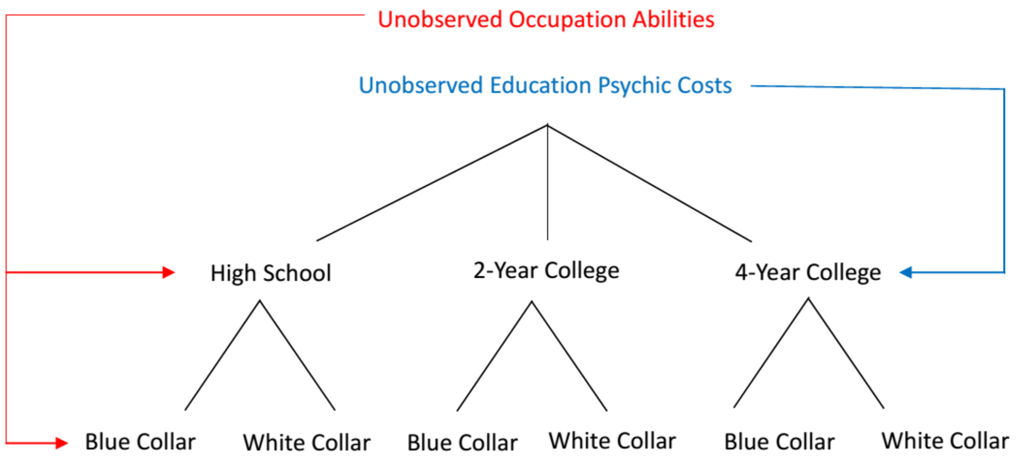

Figure 1.1: Sequential Education and Occupation Choices

As illustrated in figure 1.1, occupation abilities drive both wages and occupation choices. Therefore, the occupation choice in Equation (1.1) is endogenous and OLS estimates of occupation-specific returns to education are biased in general.

The Model for The Education Choice

A high school graduate faces three options: attending a two-year college, attending a four-year college, and entering the labour market without pursuing more education. She makes the postsecondary education decision to maximize the life-time utility. Let $I_{S,ij}$ [9]represent the net benefit associated with education level j ($j \in \{1,2,3\}$) relative to the benefit associated with education level 1:

$$I_{S,ij} = \begin{cases} \varepsilon_{S,ij} & \text{if } j = 1 \\ \alpha^m_{S,j} + Z'_{S,i}\delta_j + \varepsilon_{S,ij} & \text{if } j = 2,3 \end{cases} \tag{1.3}$$

where $\{\varepsilon_{S,ij}\}_{j=1}^3$ are mutually independent and follows type I extreme value distribution while $Z_{S,i}$ includes family background variables. The intercept, $\alpha^m_S = (\alpha_{S,2}, \alpha_{S,3})'$, is different across types. The reasons are twofold. First, future occupations and wages depend on occupation abilities. For instance, an individual with a comparative advantage in a white-collar occupation would expect herself to be more likely to work in a white-collar occupation and tend to attend a four-year college because a four-year college helps accumulating more white-collar skills than blue-collar skills. Second, education psychic costs also play an important role. For example, an individual with a comparative advantage in a white-collar occupation may choose to attend a two-year college rather than a four-year college if her psychic costs to attend a four-year college are high, although a four-year college enhances white-collar skills more than a two-year college.

As shown in figure 1.1, occupation abilities are related to both wages and the postsecondary education choice. Education psychic costs, which affect the education choice, may be correlated with occupation abilities and cause the correlation between wages and the education choices as well. Hence, the education dummy variables are endogenous in Equation (1.1) and the OLS estimates of the occupation-specific returns to education are biased in general.

[9] $I_{S,i1}$ is normalized to 0.

To sum up, the education dummy variables and the occupation choice in Equation (1.1) are endogenous because the unobserved types connects wages, occupation choices, and the postsecondary education choice. I address the endogeneity issue using a finite mixture model in which the distribution of types are specified by a flexible multinomial distribution. Since same types of individuals have the same occupation abilities and education psychic costs, the variations in education and occupation choices within type, holding the observables constant, are purely random. Once the finite mixture model is nonparametrically identified, we can get unbiased and consistent estimates of the occupation-specific returns to attending a two-year and a four-year college.

The Model for The Six Test Scores

As I will discuss in details in Section 1.4, To achieve nonparametric identification of the finite mixture model, I bring in additional information. Specifically, I use four test scores conducted from ASVAB. They are tests for math skills, verbal skills, coding speed, and mechanic comprehension. I also use the Rotter Locus of Control, and the Rosenberg Self-Esteem Scale.

In the following specification for test scores, I take into account the possibility that the test scores are influenced by the education level at the date of the tests. Since the tests were administered to all respondents in the sample in year 1979 and 1980, when they were between 14 and 22 years of age and many had finished their schooling, the tests may not be fully informative about the occupation abilities and education psychic costs (Hansen *et al.* , 2004;Heckman *et al.* , 2006b). Let $Q_{i,r}$ denote the test score in test r:

$$Q_{ir} = \alpha^{m}_{Q,r} + \theta_{r,1}2\text{YR}_{ir} + \theta_{r,2}4\text{YR}_{ir} + Z'_{i,r}\theta_{r,3} + \varepsilon_{Q,ir},\ for\ r = 1,\ldots,6, \qquad (1.4)$$

where 2YR_{ir} is a dummy variable, which equals 1 if individual i was a two-year college attendant at the time test r was administrated, and 4YR_{ir} is a dummy variable, which equals 1 if individual i was a four-year college attendant at the time test r was administrated. Other observables, which influence the test score r, such as family background variables and the age when test r was administrated, are collected in $Z_{i,r}$.

The intercept $\alpha^m_{Q,r}$ is subpopulation-specific, because the test scores reflect the occupation abilities and education psychic costs. For example, mechanic comprehension is important to a blue-collar occupation. The Rotter Locus of Control which measures people's belief in their ability to control life may be important to a management job. Math and verbal skills can reflect education psychic costs.

I assume that the test scores are mutually independent conditional on occupation abilities, education psychic costs, and the observables, i.e. $\varepsilon_{Q,ir} \perp\!\!\!\perp \varepsilon_{Q,ir'}$ for $r \neq r'$ and $\varepsilon_{Q,ir} \sim N(0, \sigma_{Q,r})$ for $r \in \{1, \ldots, 6\}$. Further, I assume that the test scores do not *directly* affect wages, occupation and education choices once conditional on occupation abilities, education psychic costs, and the observables, i.e. $\varepsilon_{Q,ir} \perp\!\!\!\perp \varepsilon_{W,it}$, $\varepsilon_{Q,ir} \perp\!\!\!\perp \varepsilon_{O,it}$, and $\varepsilon_{Q,ir} \perp\!\!\!\perp \varepsilon_{S,ij}$. These two assumptions are the key for the nonparametric identification of the finite mixture model, which will be discussed in Section 1.4.

1.3.2 A Finite Mixture Model

In the finite mixture model, the conditional joint distribution of wages $\{W_{it}\}^T_{t=1}$, occupations $\{O_{it}\}^T_{t=1}$, education S_i, and tests $\{Q_{ir}\}^6_{r=1}$ in the overall population is a weighted average of type-specific conditional joint distribution. The weight π^m is the proportion of type m. Formally,

$$\begin{aligned} & f(\{W_{it}, O_{it}\}^T_{t=1}, S_i, \{Q_{ir}\}^6_{r=1} | \{X_{it}\}^T_{t=1}, Z_{S,i}, \{Z_{ir}\}^6_{r=1}) \qquad (1.5) \\ = & \sum_{m=1}^{M} \pi^m f^m(\{W_{it}, O_{it}\}^T_{t=1}, S_i, \{Q_{ir}\}^6_{r=1} | \{X_{it}\}^T_{t=1}, Z_{S,i}, \{Z_{ir}\}^6_{r=1}), \end{aligned}$$

where $\{X_{it}\}^T_{t=1}$, $Z_{S,i}$, and $\{Z_{ir}\}^6_{r=1}$ are observables in Equation (1.1), Equation (1.2), Equation (1.3), and Equation (1.4). With the assumptions (i) test scores do not directly affect wages, occupations, and education conditional on type, (ii) the error terms in test scores are mutually independent, (iii) the error terms in wage follows a first order Markov process, (iv) the occupation choice is only affected by the previous occupation, not the whole occupation history, and (v) the regressors and the error terms in Equation (1.1), Equation (1.2), Equation (1.3), and Equation (1.4) are independent, I simplify the type-specific conditional joint distribution of wages, occupations, education, and test scores, and express the population conditional joint distribution as follows[10]:

$$
\begin{aligned}
& f(\{W_{it}, O_{it}\}_{t=1}^{T}, S_i, \{Q_{ir}\}_{r=1}^{6} | \{X_{it}\}_{t=1}^{T}, Z_{S,i}, \{Z_{ir}\}_{r=1}^{6}) \qquad (1.6)\\
= & \sum_{m=1}^{M} \pi^m f^m(W_{i1}|O_{i1}, S_i) \prod_{t=2}^{T} f^m(W_{it}|O_{it}, S_i, X_{it}, W_{it-1}, O_{it-1}, X_{it-1})\\
& \times f^m(O_{i1}|S_i) \prod_{t=2}^{T} f^m(O_{it}|O_{it-1}, S_i, X_{it}) f^m(S_i|Z_{S,i}) \prod_{r=1}^{6} f^m(Q_{ir}|Z_{ir}).
\end{aligned}
$$

In Section 1.4, I rigorously show how this finite mixture model is nonparametrically identified, i.e. how to recover the unknowns, which are on the right hand side of Equation (1.6), from the observed restriction, which is on the left hand side of Equation (1.6). Many papers that use a finite mixture model in the returns to education literature do not show the nonparametric identification. In other words, their finite mixture models may rely on restrictive parametric assumptions, which can lead to biased estimates of occupation-specific returns.

Once the nonparametric identification of the finite mixture model is established, I use the Maximum Likelihood Estimator (MLE) to estimate the occupation-specific returns to attending a two-year and a four-year college. Although the finite mixture model can be nonparametrically identified, estimating a high dimensional nonparametric statistical model requires very heavy computation and is nearly impossible given the relatively small sample size of NLSY 1979. Therefore, I estimate a statistical model with parametric assumptions in Section 1.3.1. Let $Y_i = (\{W_{it}, O_{it}, X_{it}\}_{t=1}^{T}, S_i, Z_i, \{Q_{ir}, Z_{ir}\}_{r=1}^{6})$. The log-likelihood contribution for a particular individual is as follow:

$$
\begin{aligned}
& L(Y_i; \alpha_W, \alpha_O, \alpha_S, \alpha_R, \beta, \lambda, \delta, \theta, \sigma_W, \sigma_Q, \rho) \qquad (1.7)\\
= & log(\sum_{m=1}^{M} \pi^m \mathscr{L}_W^m(Y_i; \alpha_W, \beta, \sigma_W, \rho) \mathscr{L}_O^m(Y_i; \alpha_O, \lambda) \mathscr{L}_S^m(Y_i; \alpha_S, \delta) \mathscr{L}_Q^m(Y_i; \alpha_Q, \theta, \sigma_Q)).
\end{aligned}
$$

where $\alpha_W = \{\alpha_{W,1}^m, \alpha_{W,2}^m\}_{m=1}^{M}$, $\alpha_O = \{\alpha_O^m\}_{m=1}^{M}$, $\alpha_S = \{\{\alpha_{S,j}\}_{j=2}^{3}\}_{m=1}^{M}$, $\alpha_R = \{\{\alpha_{Q,r}\}_{r=1}^{6}\}_{m=1}^{M}$, $\beta = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$, $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$, $\sigma_W = \{\sigma_{W,1}, \sigma_{W,2}\}$, and $\sigma_Q = \{\sigma_{Q,r}\}_{r=1}^{6}$.

[10] Please refer to Appendix A.2 for more details about how these assumptions simplify the type-specific conditional joint distribution of wages, occupations, education, and test scores.

Note the likelihood contribution of a particular individual who belongs to subpopulation m consists of four pieces:

$\mathscr{L}_W^m(Y_i; \alpha_W, \beta, \sigma_W, \rho)$–the likelihood contribution of wages;

$\mathscr{L}_O^m(Y_i; \alpha_O, \lambda)$–the likelihood contribution of occupation;

$\mathscr{L}_S^m(Y_i; \alpha_S, \delta)$–the likelihood contribution of education;

$\mathscr{L}_Q^m(Y_i; \alpha_Q, \theta, \sigma_Q)$–the likelihood contribution of test scores.

The detailed expressions for each of the four likelihood contributions are collected in Appendix A.3.

1.4 Nonparametric Identification of The Finite Mixture Model

In this section, I discuss the nonparametric identification of the finite mixture model using the results in Kasahara and Shimotsu (2009) and Kasahara and Shimotsu (2012). Nonparametric identification means that the proportion of types, and type-specific joint distributions of wages, occupations, education, and test scores, which is unknown, can be recovered from the observed empirical population joint distribution of wages, occupations, education, and test scores. I use two sources of information to achieve the nonparametric identification: variations of wages across occupations over time and test scores. I show that the wage history is helpful to identify the occupation abilities. Yet, test scores are essential to identify the education psychic costs.

1.4.1 Nonparametric Identification of the Occupation Abilities

I use variations in wages across occupations over time to identify the occupation abilities. Intuitively, individuals with a comparative advantage in a white-collar occupation may have high white-collar wages, and hence, the fraction of individuals with high white-collar wages can be informative about the fraction of individuals with a comparative advantage in a white-collar occupation. In other words, the fractions of individuals with high white-collar wages over time impose restrictions on the unknowns type probabilities and type-specific distributions.

Three elements are the important determinants of identification: (1) the time-dimension of panel data, (2) the variation in the occupation-specific work experience, and (3) the heterogeneity in wages and occupational choices of individuals with different occupation abilities conditional on the occupation-specific work experience. The number of observed restrictions depend on the first two elements. The third element says that variations in wages are informative about the occupation abilities.

Let's start with a simple case in which wage and occupation distribution functions are stationary and there is no serial correlation.

Proposition 1.1. *Suppose Assumption A.1 and Assumption A.2 hold. With $T \geq 3$, π^m, $f^m(S_i|Z_i)$, $f^m(W_{i1}|O_{i1},S_i)$,$f^m(O_{i1}|S_i)$, $f^m(W_{it}|O_{it},S_i,X_{it},W_{it-1},O_{it-1},X_{it-1})$, and $f^m(O_{it}|S_i,X_{it},O_{it-1})$ for $t \geq 2$ can be identified up to M types.*

The assumptions and the proof of Proposition 1.1 are collected in Appendix A.4. The number of types M that can be identified depends on the number of values $\{X_{it}\}_{t=1}^{T}$ can take and its changes over time. The key insight is that each different value of $\{X_{it}\}_{t=1}^{T}$ imposes different restrictions on the type probabilities and type-specific distributions.

The assumption that current wage and occupational choice are not influenced by the lagged values is restrictive. The productivity shocks in the wage equation can be serially correlated and occupation in the last period can affect the occupation searching cost in the current period. The next proposition relaxes this strong assumption by allowing current wage and occupation depend on those in the last period.

Proposition 1.2. *Suppose Assumption A.3 and Assumption A.4 hold, and assume $T \geq 6$. Then π^m, $f^m(S_i|Z_i)$, $f^m(W_{i1}|O_{i1},S_i)$, $f^m(O_{i1}|S_i)$, $f^m(W_{it}|O_{it},S_i,X_{it},W_{it-1},O_{it-1},X_{it-1})$, and $f^m(O_{it}|S_i,X_{it},O_{it-1})$ for $t \geq 2$ can be nonparametrically identified up to M types.*

The assumptions and the proof of Proposition 1.2 are in Appendix A.4. If there is longer dependence in either wage or occupational choice, a longer panel is required. For example, suppose current wage is affected by wage two periods before, then at least 9-period observations are needed for identification.

The education psychic costs cannot be nonparametrically identified with panel data. The reason is that postsecondary education is one-period choice in my model, so there is no information over time that can distinguish the unobserved noises and unobserved education psychic costs in the education equations. Although Keane and Wolpin (1997) consider year-by-year schooling decisions, education is not an option for each time period due to the fact that the probability of going back to school after working is very low. Therefore, education psychic costs are not nonparametrically identified in Keane and Wolpin (1997).

1.4.2 Nonparametric Identification of the Education Psychic Costs

In order to nonparametrically identify the education psychic costs, I use six test scores. They are math, verbal, coding, mechanical tests in ASVAB, the Rotter Locus of Control, and the Rosenberg Self-Esteem Scale. The nonparametric identification using test scores are intuitive. For example, individuals with low education psychic costs may have good math and verbal test scores. So the fraction of individuals with good test scores in math and verbal is informative about the fraction of individuals with low education psychic costs.

Assume that the test scores do not directly affect postsecondary choices conditional on type and some observables. In other words, test scores do not affect postsecondary education application and admission once type and other observables are known. In addition, assume that there are three test scores which are independent from each other conditional on type and some observables. These two assumptions lead to the nonparametric identification of education psychic costs.

Proposition 1.3. *With access to three test scores (Q_1, Q_2, Q_3), Suppose Assumption A.5 holds. Then π^m, $f^m(S_i|Z_i)$, and $\{f^m(Q_{ir}|S_{ir}, Z_{ir})\}_{r=1}^3$ can be nonparametrically identified up to M types.*

Assumption A.5, and the proof of Proposition 1.3 are collected in Appendix A.4.

The nonparametric identification of the education psychic costs using test scores does not require the six test scores to be perfect proxies. In other words, the nonparametric identification does not need the assumption that education does not affect test scores as it does when test scores are used as proxies. Such assumption is restrictive because some respondents already finished schooling when the test were administrated. The finding that education does influence the ASVAB scores, Rotter

Locus of Control, and Rosenberg Self-Esteem Scale in Heckman *et al.* (2006b) further show the importance to relax such assumption.

Not only can test scores identify the education psychic costs, they can identify the occupation abilities as well. For instance, the Rotter Locus of Control which measures people's belief in their ability to control life may be important to a management job. So the fraction of individuals who perform well in this test is informative about the fraction of individuals with a comparative advantage in a white-collar occupation. It implies that the nonparametric identification of the finite mixture model can be achieve without the panel data, although the additional information from the variations in wages across occupations over time are helpful to increase the efficiency. Different from previous papers such as Keane and Wolpin (1997), Belzil and Hansen (2002), and Belzil and Hansen (2007), which reply on a long panel data to identify a finite mixture model, test scores allow me to apply a finite mixture model to data with limited periods of observations.

Assume that the test scores do not directly affect wages, education and occupation choices conditional on type and some observables. Further, assume that there are three test scores which are independent from each other conditional on type and some observables. These two assumptions lead to the nonparametric identification of both occupation abilities and education psychic costs.

Proposition 1.4. *With access to three test scores (Q_1, Q_2, Q_3), Suppose Assumption A.6 holds. Then π^m, $f^m(S_i|Z_i)$, $\{f^m(Q_{ir}|S_{ir}, Z_{ir})\}_{r=1}^2$, $f^m(W_{i1}|O_{i1}, S_i)$, $f^m(O_{i1}|S_i)$, $f^m(W_{it}|O_{it}, S_i, X_{it}, W_{it-1}, O_{it-1}, X_{it-1})$, and $f^m(O_{it}|S_i, X_{it}, O_{it-1})$ for $t \geq 2$ can be nonparametrically identified up to M types.*

Assumption A.6 and the proof of Proposition 1.4 are collected in Appendix A.4.

The exclusion condition that test scores do not directly affect wage, education and occupation choices conditional on type and some observables is the key asumption to nonparametrically identify occupation abilities and education psychic costs using test scores. Intuitively, the exclusion of test scores from occupation and wage means that once employers know an individual's type, addition information about test scores would not influence the their decision on hiring and salary. The exclusion of test scores from education means that test scores do not affect postsecondary education application and admission once type is known. The assumption that test scores are exclusive from wage, education, and occupation is different from the exclusion condition in the IV approach and Heckman's two-step. While the exclusion variable in the IV approach and Heckman's two-step must not be correlated with the unobserved type, the exclusion variable in the finite mixture model has to be correlated with the unobserved type.

1.5 Empirical Results

EM algorithm (Dempster *et al.* , 1977) is applied in this paper to facilitate the computation of finding the maximum likelihood estimates. As is well known, direct maximization of the likelihood function based on Newton-Raphson type algorithm is difficult for a finite mixture model because of the possibility of many local maxima. EM algorithm is a method for finding maximum likelihood estimates by iterating an expectation (E) step and a maximization (M) step. In E step, the expectation of the log-likelihood is calculated given the estimated the type proportions and parameters in the previous iteration. In M step, parameters are updated by maximizing the expected log-likelihood found in E step. The details about the E step and M step are discussed in Appendix A.5. Each iteration increases the value of log-likelihood and it stops when convergence is achieved. The corresponding estimates upon convergence are either a local maximum or saddle points. EM algorithm is found to be sensitive to initial parameters. I choose initial parameters following the approach suggested by Heckman and Singer (1984) and a detailed discussion is in Appendix A.6.

The empirical results are presented below. I assume that the overall population consists of four types. There are two occupation abilities type: type 1 and type 2 have the same occupation abilities, and so do type 3 and type 4. Within each occupation ability type, I consider two types of education psychic costs: type 1 and type 2 are different in education psychic costs, although they have the same occupation abilities. Similarly, type 3 and type 4 have different education psychic costs but share the same occupation abilities.

1.5.1 Occupation-specific Returns to Education, Education and Occupation Choices

As illustrated in Section 1.4, the nonparametric identification of the finite mixture model heavily relies on the informativeness of the test scores about the unobserved types. If test scores are reflective about unobserved occupation abilities and education psychic costs, we would expect different types to have different test scores.

Table 1.3: Estimated Test Scores Parameters (Equation (1.4))

	Math	Verbal	Coding	Mechanic	Self-control	Self-esteem
Constant						
Type 1	-0.021	-1.122 ***	-1.315 ***	-2.430 ***	-1.081 ***	-1.548 ***
	(0.315)	(0.285)	(0.404)	(0.367)	(-0.446)	(0.474)
Deviation of type 2 from type 1	-1.271 ***	-0.868 ***	-0.742 ***	-0.734 ***	-0.380 ***	-0.140 ***
	(0.074)	(0.078)	(0.113)	(0.099)	(-0.122)	(0.120)
Deviation of type 3 from type 1	-2.052 ***	-2.187 ***	-1.445 ***	-1.661 ***	-0.588 ***	-0.535 **
	(0.096)	(0.076)	(0.104)	(0.110)	(-0.128)	(0.129)
Deviation of type 4 from type 1	-0.622 ***	-0.271 ***	-0.528 ***	-0.226 ***	-0.410 ***	-0.257 ***
	(0.067)	(0.076)	(0.089)	(0.086)	(-0.100)	(0.096)
2-year college	0.019 ***	0.038 ***	0.019	0.030	0.005	0.005
	(0.014)	(0.014)	(0.017)	(0.017)	(-0.021)	(0.022)
4-year college	0.041 ***	0.024 ***	0.018 ***	0.017 **	0.006 **	0.020 **
	(0.009)	(0.010)	(0.012)	(0.012)	(-0.015)	(0.014)
Mother education	0.012 **	-0.027 ***	-0.012	-0.010 *	-0.010	-0.033 *
	(0.013)	(0.013)	(0.019)	(0.018)	(-0.020)	(0.020)
Father education	-0.064 ***	-0.090 ***	0.115 ***	-0.092	-0.025	0.003
	(0.068)	(0.069)	(0.102)	(0.093)	(-0.100)	(0.113)
#siblings	-0.034	0.033 ***	0.042	-0.146	0.048	0.003
	(0.055)	(0.056)	(0.074)	(0.067)	(-0.082)	(0.083)
Broken family at age 14	0.008 *	0.013	-0.005	-0.186 *	-0.035	0.029
	(0.051)	(0.053)	(0.075)	(0.070)	(-0.087)	(0.079)
South at age 14	-0.012	0.044 *	0.067	0.134 **	0.072	0.083
	(0.013)	(0.013)	(0.019)	(0.017)	(-0.023)	(0.024)
Urban at age 14	0.278	0.168	0.129	-0.052 ***	0.224	0.104
	(0.076)	(0.082)	(0.110)	(0.097)	(-0.15)	(0.156)
Age	0.542 **	0.305 ***	0.210 ***	-0.286 ***	0.145 ***	0.210 ***
	(0.066)	(0.070)	(0.091)	(0.086)	(-0.121)	(0.122)

Dependent variable: math, verbal, coding speed, mechanical comprehension, locus of control, and self-esteem scores. [11]

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 1.3 reports the estimated parameters in Equation (1.4). It shows that the test scores vary across the four types and confirms that test scores are helpful to nonparametric identification of the finite mixture model. In addition, table 1.3 suggests that it is important to take into account the impact of education on test scores. A two-year college attendance significantly increases math and verbal test scores and a four-year college attendance significantly improves all the six test scores. The finding that education improves test scores implies that the test scores are not perfect proxies and using them as proxies would not help addressing the endogeneity issue to give unbiased estimates of occupation-specific returns to attending a two-year college and a four-year college.

Table 1.4 reports the estimated parameters in Equation (1.1). It shows that returns to education are occupation specific. The return to attending a two-year college is significantly higher in a blue-collar occupation than a white-collar occupation. A two-year college attendance increases blue-collar hourly payment by 24% and white-collar hourly payment by 17%. Regarding the returns to attending a four-year college, a four-year college attendance significantly increases more white-collar hourly wages than blue-collar wages. A four-year college attendant's hourly wage is 23% higher in a blue-collar occupation and 30% higher in a white-collar occupation than a high school graduate. Comparing a two-year college and a four-year college, these two kinds of postsecondary education institutions increase blue-collar wages similarly while a four-year college attendance is significantly more helpful to enhancing white-collar wages than a two-year college attendance does.

Converting the returns to attending a two-year college and a four-year college into annual returns, the corresponding annual return[12] to two-year college education is 20% in a blue-collar occupation, and 14% in a white-collar occupation. The corresponding annual return to four-year college education is 6% in blue-collar occupation, and 8% in white-collar occupation.

Among the people with post-secondary education in the sample, 27% are two-year college attendants and 73% are four-year college attendants. Hence, on average one year post-secondary education increases blue-collar wages by 10%

[12] According to Table 1.1, two-year college attendants have 1.20 year more schooling than high school graduates and four-year college attendants have 3.82 year more schooling than high school graduates on average.

Table 1.4: Estimated Wage Parameters (Equation (1.1))

	Blue Collar	White Collar
Constants		
Type 1 and 2	6.932 ***	7.044 ***
	(0.021)	(0.033)
Type 3 and 4	6.485 ***	6.457 ***
	(0.021)	(0.038)
2-year college	0.243 ***	0.170 ***
	(0.024)	(0.036)
4-year college	0.231 ***	0.296 ***
	(0.022)	(0.031)
Blue-collar experience	0.078 ***	0.033 ***
	(0.008)	(0.009)
Blue-collar experience squared	-0.285 ***	-0.076
	(0.069)	(0.075)
White-collar experience	0.040 ***	0.079 ***
	(0.012)	(0.008)
White-collar experience squared	-0.044	-0.191 **
	(0.169)	(0.084)
Hypothesis testing		p-value
blue-collar return=white-collar return, 2-year college		0.016
blue-collar return=white-collar return, 4-year college		0.018
2-year college return=4-year college return, blue-collar		0.311
2-year college return=4-year college return, white-collar		0.000

Dependent variable: log hourly salary

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

$(20\% \times 27\% + 6\% \times 73\%)$ and white-collar wages by $10\%(14\% \times 27\% + 8\% \times 73\%)$. Keane and Wolpin (1997) find that the annual return to postsecondary education is 2.4% in a blue-collar occupation and 7% in a white-collar occupation. The annual return to postsecondary education in a white-collar occupation in this paper is similar to that reported in Keane and Wolpin (1997). However, the annual return to postsecondary education in a blue-collar occupation is higher in this paper than in Keane and Wolpin (1997) where they do not distinguish a two-year college and a four-year college.

The type-specific constants reported in table 1.4 suggest that individuals are endowed with different occupation abilities. Among the four types of individuals,

Table 1.5: Estimated Average Partial Effects, Occupation Choice

	Initial job	Subsequent jobs
Constant		
Type 1 and 2	-0.246 ***	-0.137 **
	(0.057)	(0.071)
Deviation of type 3 and 4 from type 1 and 2	-0.162 ***	-0.056 **
	(0.063)	(0.027)
2-year college	0.159 ***	0.042 **
	(0.044)	(0.019)
4-year college	0.503 ***	0.122 ***
	(0.028)	(0.046)
Blue-collar experience		-0.044
		(0.193)
Blue-collar experience squared		0.262 **
		(0.128)
White-collar experience		0.054
		(0.120)
White-collar experience squared		-0.266 **
		(0.137)
White-collar job in the last period		0.287 ***
		(0.079)

The average partial effect is calculated as the average of the partial effect of each individual.

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

type 1 and type 2 share the same occupation abilities, but are different in the education psychic costs. Type 3 and type 4 have the same occupation abilities, but different education psychic costs. Although the occupation abilities and the education psychic costs may be correlated, the education psychic costs do not directly affect wages as assumed. So type 1 and type 2 earn the same, and so do type 3 and type 4. As reported in table 1.4, type 1 and type 2 earn more in a white-collar occupation than a blue-collar occupation. In other words, type 1 and type 2 have a comparative advantage in a white-collar occupation. On the other hand, type 3 and type 4 are similarly productive in a blue-collar and a white-collar occupation, because they earn similar wages in a white-collar and a blue-collar occupation.

Table 1.5 shows the estimated average partial effects in the occupation choice. Column 1 in table 1.5 reports the estimated average partial effects in making the occupation choice in the first job. It indicates that individuals with a comparative advantage in a white-collar occupation (type 1 and type 2) are 16% more likely to choose to work in white-collar jobs than those with a comparative advantage in a blue-collar occupation (type 3 and type 4). Moreover, education increases the probability of being employed in a white-collar occupation. Comparing to high school graduates, two-year college attendants are 16% more likely to work in a white-collar occupation and four-year college attendants are 50% more likely to work in a white-collar occupation.

Column 2 in table 1.5 reports the estimated average partial effects in making the occupation choice in the sequential jobs. It shows that individuals with a comparative advantage in a white-collar occupation (type 1 and type 2) are 6% more likely to work in white-collar jobs than those with a comparative advantage in a blue-collar occupation (type 3 and type 4) in the sequential jobs. Education has smaller influence on occupation in subsequent jobs than initial jobs. Attending a two-year colleges and a four-year college increase the probability of being employed by a white-collar occupation by 4% and 12% respectively. One important factor which affects the occupation choice is the occupation in the previous period. An individual who worked in a white-collar occupation in the previous period is 29% more likely to work in a white-collar occupation in the current period than an individual who worked in a blue-collar occupation in the previous period does.[13]

Regarding the postsecondary education choice, if individuals consider their future occupations when making their education decisions, we would expect individuals with a comparative advantage in a white-collar occupation (type 1 and type 2) to be more likely to attend a four-year college than individuals earn similarly in a blue-collar occupation and a white-collar occupation (type 3 and type 4). Table 1.6 reports the average partial effects in the postsecondary education choice. It shows that type 1 and type 2 are 53% more likely to attend a four-year college than type 3 and type 4. This finding confirms that the occupation abilities affect the education

[13] I have consider the case that individuals may have different occupation tastes. For example, those who enjoy working outdoors may prefer a construction worker position to an economist position. To copy with the potential heterogeneity in occupation taste, I estimate a finite mixture model with 8 types. Specifically, I consider two occupation abilities types. Within each occupation abilities type, there are two education psychic costs type. In addition, I look at two occupation taste types for individuals with the same occupation abilities and education psychic costs. I do not find that individuals with the same occupation abilities and education psychic costs behave differently in the occupation choices.

Table 1.6: Estimated Average Partial Effects, Educational Choice

	2-year College	4-year College
Constant		
Type 1	-0.028	-0.293
	(0.179)	(0.283)
Deviation of type 2 from type 1	0.083	-0.416 ***
	(0.828)	(0.11)
Deviation of type 3 from type 1	0.092	-0.525 ***
	(0.112)	(0.116)
Deviation of type 4 from type 1	0.094	-0.204 ***
	(0.792)	(0.085)
Mother education	-0.004	0.038
	(0.645)	(0.078)
Father education	0.004	0.031
	(0.735)	(0.087)
#siblings	-0.001	-0.021
	(0.103)	(0.086)
Broken family at age 14	0.032	-0.066
	(0.063)	(0.126)
South at age 14	-0.013	0.077 **
	(0.059)	(0.033)
Urban at age 14	0.059	0.044
	(0.070)	(0.043)

The average partial effect is calculated as the average of the partial effect of each individual.

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

choice. Further, the results in table 1.6 suggest that individuals take into account the education psychic costs when making their education decisions. Although type 1 and type 2 share the same occupation abilities, type 1 is 42% more likely to attend a four-year college than type 2 is, which indicates that type 1 has lower psychic costs to attend a four-year college than type 2 does. Similarly, type 4 are more likely to attend a four-year college than type 3 does, which suggests that type 4 has lower psychic costs to attend a four-year college than type 3 does. Regarding the decision to attend a two-year college, the similarity across the four types in the decision to attend a two-year college suggests no self-selection in attending a two-year college.

To sum up, individuals self select into different education groups based on their occupation abilities and education psychic costs. Their occupation choices are also influenced by their occupation abilities. Failure to address the endogenous education

Table 1.7: Occupation-Specific Returns

	(1)	(2)	(3)	(4)
2-year college × blue-collar	0.224 ***	0.188 ***	0.277 ***	0.243 ***
	(0.024)	(0.023)	(0.024)	(0.024)
2-year college × white-collar	0.136 ***	0.105 ***	0.192 ***	0.170 ***
	(0.034)	(0.031)	(0.035)	(0.036)
4-year college × blue-collar	0.246 ***	0.172 ***	0.237 ***	0.231 ***
	(0.022)	(0.023)	(0.022)	(0.022)
4-year college × white-collar	0.295 ***	0.227 ***	0.293 ***	0.296 ***
	(0.029)	(0.027)	(0.030)	(0.031)

Column (1): OLS estimates of the occupation-specific returns to education

Column (2): OLS estimates of the occupation-specific returns to education when six test scores are included as proxies for occupation abilities

Column (3): estimates of the occupation-specific returns to education when controlling for occupation abilities only

Column (4): estimates of the occupation-specific returns to education when controlling for both occupation abilities and education psychic costs

All the regressors in Equation (1.1) are included.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

and occupation choices can result in biased estimates of occupation-specific returns. Due to the complication of the self-selection problem here, it is hard to tell the direction of the possible bias. I compare the estimates of occupation-specific returns to education when controlling or not controlling for occupation abilities and/or education psychic costs in table 1.7. Column (4) presents the estimates of occupation-specific returns to attending a two-year college and a four-year college controlling both the occupation abilities and education psychic costs (the same estimates as those reported in table 1.4). Column (1) gives the OLS estimates of the occupation-specific returns to education. The OLS estimates of the occupation-specific returns to attending a two-year college are slightly lower than those in column (4) and the OLS estimates of the occupation-specific returns to attending a four-year college are comparable to those in column (3). Column (2) in table 1.7 shows the OLS estimates of the occupation-specific returns to education when six test scores are included as proxies for occupation abilities and education psychic costs. Using test scores as proxies requires that education does not affect test scores. However, the results in table 1.3 suggest that education helps to improve the performance in all the six tests. The estimated returns to attending a two-year college and a four-year college in column (2) are around 6 percentage points lower than those reported in column (4). Column (3) in table 1.7 gives the estimates of the occupation-specific returns to education only controlling for the occupation abilities[14]. The estimates of the occupation-specific returns to attending a two-year college in column (3) are larger than both the OLS estimates in column (1) and those in column (4). The estimates of the occupation-specific returns to attending a four-year college in column (3) are comparable to those in column (1) and column (4). The results in table 1.7 suggest that the possible biases are of different direction and cancel out each other although education and occupation choices are endogenous.

1.5.2 Conditional Independence of Wages and Test Scores

One of the key assumptions of the nonparametric identification of the finite mixture model is that conditional on type and some other observables the six test scores do not directly affect wages. I test the validation of this conditional independence assumption by including the six test scores one by one into the wage equation (Equation (1.1)). The idea is that assuming the finite mixture model is nonparametrically identified using the other test scores, the coefficient on test score r, which is included in the wage equation, should be zero when test score r and wages are conditionally independent.

[14] Here, I estimate a finite mixture model in which there are two occupation abilities types and everyone has the same education psychic costs. Since the variations in wages across occupations over time are sufficient to nonparametrically identify occupation abilities, I do not use test scores as an additional source of nonparametric identification

Table 1.8: Testing Conditional Independence of Wages and Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
2-year college × blue-collar	0.203 ***	0.275 ***	0.250 ***	0.235 ***	0.243 ***	0.240 ***
	(0.024)	(0.025)	(0.025)	(0.024)	(0.024)	(0.024)
2-year college × white-collar	0.123 ***	0.198 ***	0.176 ***	0.163 ***	0.170 ***	0.165 ***
	(0.037)	(0.034)	(0.035)	(0.037)	(0.036)	(0.036)
4-year college × blue-collar	0.185 ***	0.291 ***	0.239 ***	0.205 ***	0.231 ***	0.227 ***
	(0.024)	(0.023)	(0.023)	(0.022)	(0.022)	(0.022)
4-year college × white-collar	0.241 ***	0.357 ***	0.305 ***	0.267 ***	0.296 ***	0.291 ***
	(0.034)	(0.031)	(0.032)	(0.032)	(0.031)	(0.031)
Math	0.027 ***					
	(0.010)					
Verbal		-0.044 ***				
		(0.008)				
Coding			-0.009			
			(0.008)			
Mechanic				0.025 ***		
				(0.007)		
Locus of control					0.0003	
					(0.008)	
Self-esteem						0.008
						(0.008)

All the regressors in Equation (1.1) are included.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 1.8 presents the estimated coefficients on the six test scores. The estimated coefficients on coding speed, Rotter locus of control, and Rosenberg self-esteem scale are not significantly different from zero. According to Proposition 1.4, the finite mixture model can be nonparametrically identified because we have three tests satisfied the conditional independence assumption. However, we reject the hypothesis that the conditional independence assumption holds for math skill, verbal skill, and mechanical comprehension. The reason of the finding that math, verbal, mechanical comprehension scores affect wages is that I only consider a small number of types (two occupation abilities types and two education psychic costs types). It is possible that there are heterogeneity in occupation abilities and education psychic costs within each of the four types and math skill, verbal skill, and mechanical comprehension scores are informative about these within type heterogeneity. I check the sensitivity of the estimates in two ways. First, I include the math, verbal, and mechanical comprehension scores together into the wage equation and check whether the estimates of the occupation-specific returns to education are different from those reported in table 1.4. Table 1.9 shows that the estimates of the occupation-specific returns to education are around 2 to 5 percentage points smaller than those in table 1.4. Second, I increase the number of types from 4 (two types of occupation abilities and two types of education psychic costs) to 6 (three types of occupation abilities and two types of education psychic costs) and the corresponding estimates in the wage equation are presented in table 1.10. Table 1.10 shows that the estimates of occupation-specific returns to education are around 1 to 4 percentage points smaller than those reported in table 1.4.

1.5.3 The Occupation-Specific Returns to A Bachelor's Degree

The wage gap between college dropouts and college graduates are documented in the literature (Jaeger and Page, 1996). It is interesting to examine the occupation-specific returns to college graduate besides the occupation-specific returns to college attendants. Due to the small sample size of the two-year college graduates, I focus on investigating the occupation-specific returns for those obtained a bachelor's degree.

Among the four-year college attendants in my sample, around 70% obtained

Table 1.9: Estimated Wage Parameters, with Three Test Scores in the Wage Equation

	Blue Collar	White Collar
Constants		
Type 1 and 2	6.914 ***	7.048 ***
	(0.021)	(0.032)
Type 3 and 4	6.461 ***	6.443 ***
	(0.023)	(0.038)
2-year college	0.222 ***	0.121 ***
	(0.023)	(0.034)
4-year college	0.228 ***	0.273 ***
	(0.024)	(0.032)
Blue-collar experience	0.075 ***	0.034 ***
	(0.008)	(0.009)
Blue-collar experience squared	-0.278 ***	-0.125 **
	(0.069)	(0.074)
White-collar experience	0.050 ***	0.080 ***
	(0.012)	(0.009)
White-collar experience squared	-0.158	-0.205 ***
	(0.161)	(0.084)
Test Scores		
Math	0.085 ***	
	(0.013)	
Verbal	-0.137 ***	
	(0.012)	
Mechanical comprehension	0.068 ***	
	(0.009)	

Dependent variable: log hourly salary

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

The coefficients on the three test scores are restricted to be the same across occupations.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

a bachelor's degree. A simple comparison of the first year wage of the four-year college dropouts and the four-year college graduates shows that the four-year college dropouts and the four-year college graduates earn similarly in a blue-collar occupation, yet the four-year college graduates earn 30% more than the four-year college dropouts in a white-collar occupation[15].

Table 1.11 presents the estimated parameters of the wage equation (Equation

[15] For more summary statistics of the college dropouts and college attendants, please refer to table A.1.

Table 1.10: Estimated Wage Parameters in the Wage Equation (6 types)

	Blue Collar	White Collar
Constants		
Type 1 and 2	6.998 ***	7.116 ***
	(0.027)	(0.034)
Type 3 and 4	6.501 ***	6.439 ***
	(0.026)	(0.041)
Type 5 and 6	6.609 ***	6.679 ***
	(0.021)	(0.037)
2-year college	0.235 ***	0.126 ***
	(0.023)	(0.036)
4-year college	0.238 ***	0.283 ***
	(0.021)	(0.032)
Blue-collar experience	0.076 ***	0.03 ***
	(0.008)	(0.009)
Blue-collar experience squared	-0.284 ***	-0.063
	(0.07)	(0.078)
White-collar experience	0.042 ***	0.078 ***
	(0.011)	(0.008)
White-collar experience squared	-0.075	-0.211 ***
	(0.162)	(0.086)

Dependent variable: log hourly salary

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4 as well as type 5 and type 6.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4 as well as type 5 and type 6.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

(1.1)) using the sample where the four-year college dropouts are eliminated. It shows that a bachelor's degree increases blue-collar wages by 26% and white-collar wages by 33% for a high school graduate. Comparing to the returns to attending a four-year college as reported in table 1.4, ie. 23% and 30% respectively for a blue-collar occupation and a white-collar occupatio, the estimated occupation-specific returns to a bachelor's degree are not much higher.

1.5.4 The Expected Returns to Education

Individuals make their education choices taking into account their future occupations. Yet, they do not know exactly their occupations because of the uncertainty in the labour market. Therefore, their education choices are based on the expected returns to attending a two-year and a four-year college. Below, I calculated the expected

Table 1.11: The Occupation-Specific Returns to A Bachelor's Degree

	Blue Collar	White Collar
Constants		
Type 1 and 2	6.914 ***	6.983 ***
	(0.024)	(0.035)
Type 3 and 4	6.465 ***	6.425 ***
	(0.023)	(0.042)
2-year college	0.257 ***	0.191 ***
	(0.025)	(0.036)
Bachelor's degree	0.257 ***	0.328 ***
	(0.035)	(0.033)
Blue-collar experience	0.084 ***	0.047 ***
	(0.009)	(0.01)
Blue-collar experience squared	-0.326 ***	-0.162 **
	(0.076)	(0.08)
White-collar experience	0.038 ***	0.083 ***
	(0.014)	(0.01)
White-collar experience squared	-0.049	-0.214 **
	(0.189)	(0.094)

Dependent variable: log hourly salary

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

The four-year college dropouts are eliminated from the sample.

Standard errors are in parenthesis.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

returns to education by simulating a sample of 10000 observations.

Panel A of table 1.12 shows that the expected returns to attending a two-year college are around 23% for type 1 and type 2 (individuals with a comparative advantage in a white-collar occupation) over time[16], and they are around 22% for type 3 and type 4 (individuals with a comparative advantage in a blue-collar occupation) over time. Returns to attending a two-year college are similar to all types, which explains that individuals do not select to attend a two-year college based on their occupation abilities as suggested by the results in table 1.6. Regarding a four-year college, the expected returns to attending a four-year college for type 1 and type 2 are 34% in the first year and increase to 40% nine years later. Returns

[16] Since the occupation choice depend on the occupation abilities and not influenced by education psychic costs directly, the expected returns to attending a two-year college are the same for individuals with same occupation abilities and different education psychic costs. That is to say that type 1 and type 3 have the same expected returns to attending a two-year college, and type 2 and type 4 gain the same in earnings from attending a two-year college

Table 1.12: Expected Returns to Education

	2-Year College		4-Year College	
	Type 1 and 2	Type 3 and 4	Type 1 and 2	Type 3 and 4
Panel A: Total				
1st year	0.238	0.225	0.337	0.251
5th year	0.225	0.212	0.349	0.255
10th year	0.239	0.22	0.397	0.287
Panel B: Occupation-Specific Skills Accumulation				
1st year	0.216	0.229	0.277	0.264
5th year	0.206	0.221	0.283	0.271
10th year	0.201	0.221	0.288	0.275
Panel C: Better Occupation Match				
1st year	0.022	-0.004	0.06	-0.013
5th year	0.019	-0.009	0.065	-0.016
10th year	0.038	-0.001	0.109	0.011

Calculation is based on the simulation of 10000 observations.

Panel A: total expected returns to attending a two-year college and a four-year college

Panel B: expected returns to education from enhancing the occupation-specific skills

Panel C: expected returns to education from increasing the probability of being employed in a white-collar occupation

Total expected returns to education (Panel A) is the sum of the expected returns from enhancing the occupation-specific skills (Panel B) and the expected returns from increasing the probability of being employed in a white-collar occupation (Panel C).

Type 1 and type 2 have the same occupation abilities. So do type 3 and type 4.

Type 1 and type 2 have different education psychic costs. So do type 3 and type 4.

to attending a four-year college for type 3 and type 4 are 25% in the first year and increase to 29% in the tenth year. Returns to attending a four-year college are around 9 percentage points higher for type 1 and type 2 than type 3 and type 4 in the first year and the difference increases to 11 percentage points in the 10th year. Therefore, individuals with a comparative advantage in a white-collar occupation (type 1 and type 2) are more likely to attend a four-year college than those with a comparative advantage in a blue-collar occupation (type 3 and type 4) as shown in table 1.6. Comparing returns to attending a two-year and a four-year college, returns to attending a four-year college are 10 percentage points higher than returns to a two-year college in the beginning and the discrepancy is enlarged to 16 percentage points after nine years for type 1 and type 2. For type 3 and type 4, returns to attending a four-year college are 3 percentage points higher than returns to attending a two-year college in the beginning and the difference increases to 7 percentage points after nine years. The difference between returns to attending a two-year college and a four-year college echoes the findings in Belzil and Hansen (2002) that returns can be education-level-specific.

The expected returns to attending a two-year and a four-year college are different across types of people with different occupation abilities. The relationship between returns to education and innate abilities are well documented in the literature (Belzil and Hansen, 2007; Carneiro *et al.* , 2003). The reason of the correlation between the expected returns to education and the occupation abilities is that the expected returns to education are related to the probability of working in a white-collar occupation, which depend on occupation abilities. The expected returns to education and the probability of working in a white-collar occupation can be related in two ways. First, education helps accumulating white-collar and blue-collar skills differently. For example, attending a four-year college increases white-collar skills more than blue-collar skills. Therefore, individuals with a comparative advantage in a white-collar occupation, who are more likely to work in a white-collar occupation, have higher returns to attending a four-year college on average. Second, education enhances the probability of working in a white-collar occupation. For individuals with a comparative advantage in a white-collar occupation, the reward to their occupation abilities are higher in a white-collar occupation than a blue-collar occupation. Attending a four-year college education increases the probability of working in a white-collar occupation and leads to a high reward to their occupation abilities. I decompose the returns to education into these two parts: enhancing occupation-specific skills and increasing the probability of working in a white-collar occupation.

Panel B of table 1.12 shows the part of returns to education from enhancing occupation-specific skills and panel C of table table 1.12 presents the part of returns to education from increasing the probability of being employed in a white-collar occupation. Let's first look at the decomposition of the expected returns to attending a two-year college. For individuals with a comparative advantage in a white-collar occupation (type 1 and type 2), a two-year college attendance increases wages by 22% from enhancing the occupation-specific skills. The part of expected returns from increasing the probability of working in a white-collar occupation is 2% at the beginning and 4% at the end. As type 1 and type 2 become more likely to work in a white-collar occupation, their latter part of the expected returns to attending a two-year college increases. For individuals with a comparative advantage in a blue-collar occupation (type 3 and type 4), the expected returns from enhancing occupation-specific skills are around 22% over the ten years, which is comparable to those of type 1 and type 2 in magnitude. The part of expected returns from increasing the probability of being employed in a white-collar occupation is almost close to zero over time. This is because type 3 and type 4 are rewarded similarly to their occupation abilities in both occupations. Next, let's look at the decomposition of the expected returns to attending a four-year college. For type 1 and type 2, the part of expected returns attending a four-year college due to occupation-specific skills accumulation is around 28%. The part of expected returns from increasing the probability of working in a white-collar occupation increases from 6% to 14% with the probability of working in a white-collar occupation increasing from 71% to 83% over time. After ten years in the labour market, 65% of the total expected returns to attending a four-year college education are from its impact on occupation-specific skills accumulation for type 1 and type 2. For type 3 and type 4. The part of expected returns to attending a four-year college due to occupation-specific skills accumulation is around 26%, which is comparable to that for type 1 and type 2 in magnitude. The part of expected returns to attending a four-year college from its influence on occupation affiliation is close to zero because type 3 and type 4 are rewarded similarly to their occupation abilities in a blue-collar and a white-collar occupation.

The increasing expected returns to attending a four-year college for individuals with a comparative advantage in a white-collar occupation (type 1 and type 2) imply a faster wage growth rate of four-year college attendants than high school graduates. This finding is consistent with Willis and Rosen (1979) where they find that a college attendant's wage grows faster than a high school graduates. This chapter suggests that one important reason for the faster wage growth rate of four-year college attendants is that they switch to the occupation they have a comparative advantage of over time.

1.5.5 Test Scores and Returns to Education

As shown in table 1.3, test scores are informative about individuals occupation abilities and education psychic costs. Once the type-specific joint distributions of the six test scores are identified, we can get the probabilities of types conditional on the six test scores using Bayes' rule. In other words, we are able to tell which type an individual is most likely to be given her six test scores and demographic information. Further, we can infer her expect returns to attending a two-year college and a four-year college.

I simulate the six test scores for 10,000 high school graduates, whose parents are high school graduates, who have three siblings, were raised in a two-parent family, lived in the northern urban area of U.S. at age 14, and took the six test scores at age 18. For simplicity, I divide the six test scores into two groups: cognitive tests (math skills, verbal skills, coding speed, and mechanical comprehension) and noncognitive tests (the Rotter Locus of Control and the Rosenberg Self-Esteem Scale). Then I calculate the average scores, $\hat{Q}_c$ and $\hat{Q}_{nc}$, for the two groups. For further simplicity, each of the two average test scores are partitioned into four parts. The proportion of each type conditional on test scores are presented in table 1.13. For example, look at an individual with high school graduates parents, 3 siblings, raised in a two parents family, lived in the North urban area of U.S. at age 14, and took the cognitive and noncognitive tests at age 18. If all her test scores are at the 10th percentile, her net average scores are in the cell of row 1 and column 1 and she is 0.1% likely to be type 1, 61.5% likely to be type 2, 29.8% likely to be type 3, and 8.6% likely to be type 4. If all her test scores are at the 50th percentile, her net average scores are in the cell of row 2 and column 3 and she is 15% likely to be type 1, 0.1% likely to be type 2, 32.5% likely to be type 3, and 52.3% likely to be type 4. If all her test scores are at the 90th percentile, her net average scores are in the cell of row 4 and column 4 and she is 90.9% likely to be type 1, 0% likely to be type 2, 0% likely to be type 3, and 9.1% likely to be type 4. Once the conditional proportion of types is known, her returns to education can be calculated accordingly. Table 1.14 shows the expected returns to attending a two-year college and a four-year college for such an individual with test scores at the 10th, 50th, and 90th percentile. Since returns to two-year college are similar to all types, the expected returns to

Table 1.13: Posterior Probabilities of Types Conditional on Test Scores

(a) Type 1

	$\bar{Q}_{nc} \leq -0.5$	$-0.5 < \bar{Q}_{nc} \leq 0$	$0 < \bar{Q}_{nc} \leq 0.5$	$\bar{Q}_{nc} \geq 0.5$
$\bar{Q}_c \leq -0.5$	0.001	0	0.013	0.007
$-0.5 < \bar{Q}_c \leq 0$	0.065	0.127	0.150	0.191
$0 < \bar{Q}_c \leq 0.5$	0.328	0.459	0.565	0.680
$\bar{Q}_c \geq 0.5$	0.782	0.777	0.873	0.909

(b) Type 2

	$\bar{Q}_{nc} \leq -0.5$	$-0.5 < \bar{Q}_{nc} \leq 0$	$0 < \bar{Q}_{nc} \leq 0.5$	$\bar{Q}_{nc} \geq 0.5$
$\bar{Q}_c \leq -0.5$	0.615	0.54	0.445	0.357
$-0.5 < \bar{Q}_c \leq 0$	0.000	0.001	0.001	0.002
$0 < \bar{Q}_c \leq 0.5$	0.000	0.000	0.000	0.000
$\bar{Q}_c \geq 0.5$	0.000	0.000	0.000	0.000

(c) Type 3

	$\bar{Q}_{nc} \leq -0.5$	$-0.5 < \bar{Q}_{nc} \leq 0$	$0 < \bar{Q}_{nc} \leq 0.5$	$\bar{Q}_{nc} \geq 0.5$
$\bar{Q}_c \leq -0.5$	0.298	0.353	0.432	0.529
$-0.5 < \bar{Q}_c \leq 0$	0.313	0.343	0.325	0.327
$0 < \bar{Q}_c \leq 0.5$	0.056	0.043	0.033	0.031
$\bar{Q}_c \geq 0.5$	0.000	0.006	0.000	0.000

(d) Type 4

	$\bar{Q}_{nc} \leq -0.5$	$-0.5 < \bar{Q}_{nc} \leq 0$	$0 < \bar{Q}_{nc} \leq 0.5$	$\bar{Q}_{nc} \geq 0.5$
$\bar{Q}_c \leq -0.5$	0.086	0.106	0.111	0.107
$-0.5 < \bar{Q}_c \leq 0$	0.623	0.529	0.523	0.481
$0 < \bar{Q}_c \leq 0.5$	0.616	0.498	0.402	0.289
$\bar{Q}_c \geq 0.5$	0.218	0.217	0.127	0.091

Calculation is based on the simulation of 10000 high school graduates whose parents are high school graduates, who have three siblings, were raised in a two-parent family, lived in the northern urban area of U.S. at age 14, and took the six test scores at age 18.

$\bar{Q}_c$ is the average (standardized) math, verbal, coding speed, mechanical comprehension scores.

$\bar{Q}_nc$ is the average (standardized) Rotter Locus of Control and Rosenberg Self-Esteem Scale.

Each cell shows the probability of belonging to a specific type given that $\bar{Q}_c$ and $\bar{Q}_{nc}$ fall in a specific region.

Table 1.14: Expected Returns to Education, By Test Scores

	2-Year College			4-Year College		
	10th Percentile	50th Percentile	90th Percentile	10th Percentile	50th Percentile	90th Percentile
1st year	0.230	0.232	0.238	0.280	0.293	0.333
5th year	0.217	0.219	0.225	0.286	0.300	0.342
10th year	0.225	0.228	0.236	0.323	0.339	0.388

Calculation is based on the simulation of 10000 observations.

two-year college are almost the same for different test scores. Regarding a four-year college, the expected returns to attending a four-year college increase as test scores increase. The reason is that high test scores imply a high probability of a comparative advantage in a white-collar occupation, and a comparative advantage with a white-collar occupation are associated with a high returns to attending a four-year college.

1.6 Conclusion

In this paper, I examine the returns to attending a two-year college and a four-year college and how the returns to education differ from those of a white-collar occupation to those of a blue-collar occupation. Despite a vast literature on returns to education, the existing research on how returns to attending a two-year college and a four-year college depend on the occupation choice is limited. The reason for this limitation is that it is difficult to estimate the occupation-specific returns in the presence of endogenous education and occupation choices. On the one hand, individuals are endowed with different abilities to work in a blue-collar occupation or a white-collar occupation. They tend to work in the occupation in which they have a comparative advantage. Moreover, they are more likely to choose the type of postsecondary education that intensively accumulates the skills needed in the occupations they would like to work in when they finish schooling. Therefore, occupation abilities drive wages, education, and occupation. On the other hand, individuals vary in their education psychic costs, which may be correlated with occupation abilities. While the occupation abilities and the education psychic costs are known to the individuals making education and occupation decisions, these abilities and costs are unobserved by the econometrician, thus leading to the missing variable problem. The instrumental variables (IV) approach, conventionally used to deal with the endogeneity issue in the returns to education literature, is difficult to implement here simply because good instruments for both education and occupation are difficult to find.

I address the endogeneity issue in education and occupation by explicitly modeling the sequential education and occupation choices, specifying the unobserved occupation abilities and education psychic costs with a flexible multinomial distribution in a finite mixture model. I show how to nonparametrically identify the occupation abilities using the variations in wages across occupations over time. However, the information from the panel data alone is not enough to identify the education psychic costs. In order to achieve nonparametric identification of the education psychic costs, I use test scores such those of the ASVAB, the Rotter Locus of Control, and the Rosenberg Self-Esteem Scale. I show that conditional on occupation abilities and education psychic costs the education psychic costs can be nonparametrically identified under the assumption that the test scores do not directly affect wages, education, or occupation choices.

Using data taken from the National Longitudinal Survey of Youth (NLSY) 1979, I estimate a parametrically specified finite mixture model for joint wages, education, occupation, and test scores and find that returns to education are occupation-specific. Specifically, I find that attendance of a two-year college enhances blue-collar wages by 24% and white-collar wages by 17% while attendance of a four-year college increases blue-collar wages by 23% and white-collar wages by 30%.

Chapter 2

M-Estimation with Complex Survey Data

2.1 Introduction

The complex survey design, which is also known as the stratified multistage clustered sample design, is widely used in large scale surveys. For example, the Monthly Current Population Survey (CPS, US), the Panel Study of Income Dynamics (PSID, US), the Labour Force Survey(LFS, Canada), and the Survey of Labour and Income Dynamics(SLID, Canada) employ the complex survey design.

In complex survey sampling, the population is partitioned into mutually exclusive and exhaustive subpopulations called strata. Each stratum is then partitioned into primary sampling units (psu), and each of the psu's is partitioned into secondary sampling units (ssu). In the m-stage cluster sampling, this process is repeated m times to form finer and finer partitions. The sampling units created by the last stage of partitioning is called the ultimate sampling unit (usu). Though an usu is typically a group of individuals (e.g., individuals in a contiguous dwelling segment), it is possible that an usu is an individual. The number of stages for the recursive partitioning may differ from a stratum to another.

The complex survey method makes use of the recursive structure of sampling units described above for randomly selecting individuals into the sample in each stratum. The selection of individuals starts with a few draws of psu's, which is followed by a few draws of ssu's within the selected psu's, which is further followed by a few draws of tertiary sampling units within the selected ssu's, and so on. Once usu's are selected in this manner, every individual in the usu is selected into the sample. The selection of sampling units is statistically independent across strata.

While the complex survey method is attractive in terms of the sampling cost, it introduces some complication in statistical analysis using it, when we compare it with the simple random sampling method. First, the number of observations in each stratum is random. Second, conditional on the random number of observations, the observations are dependent in a complicated manner, because of the recursive selection of sampling units. It is therefore important to study how such features of complex survey data should be incorporated in estimation and statistical inferences in econometric analysis. For this reason, this chapter investigates the properties of M-estimators when used with complex survey data.

In the survey sampling literature, there are two prevailing views about the population. The principal difference between these two views is the randomness which is used to give stochastic structure to the inference. One is called a design base, which assumes that each individual in the population carries a constant (i.e., nonrandom) character vector. Following Neyman (1934)'s seminal paper, the design-based approach regards the probability ascribed by the sampling design to the various subsets of the finite population as the primary source of randomness. For example, Krewski and Rao (1981), Binder (1983), and Sakata (2000) are design-based. The other is called a model base. It assumes that each individual draws a characteristic vector from a distribution (superpopulation). For example, Fuller (1984), Hung (1990), Chamblessa and Boyle (1985), and Breckling *et al.* (1994) are model-based. Cassel *et al.* (1977) and Sarndal *et al.* (1992) examine both design-based and model-based inference in detail. Sarndal *et al.* (1978) provide a comprehensive comparison of the design-based and model-based approaches. In this paper, we provide a unified framework, for which both a design base and a model base are special cases. In such framework, we study the behavior of the M-estimators, which includes widely used least squares estimator and maximum likelihood estimator. We consider how statistical inferences should be made based on the M-estimation. Specifically, we consider two stages of correction (relative to simple random sampling) to account for the features of complex survey data. First, we follow Horvitz and Thompson (1952) to deal with the fact that individuals are usually under- or over-represented in complex survey data. Horvitz and Thompson (1952) proposed to estimate the total and mean of a superpopulation in a stratified sample by weighting the sample using inclusion probability, the probability of becoming part of the sample. We show that our proposed weighted M-estimator is unbiased, where the sample is weighted by inclusion probability. Second, standard errors of the estimates with complex survey data are different from those with simple random sampling data. Clustering increases the standard errors and stratification reduces the standard errors. These two effects do not cancel out in general. We derive the asymptotic distribution of the estimates and compute standard errors which are robust to the sample-design effects. Our analysis employs the standard asymptotic framework in the complex survey literature, in which the number of strata grows to infinity [1].

The rest of this chapter is organized as follows. Section 2.2 describes our assumptions about the data generating process and introduces the weighted M-estimator. Sections 2.3 and 2.4 establish consistency and asymptotic normality of the weighted M-estimator, respectively. Section 2.5 then discusses how to estimate the asymptotic covariance matrix of the weighted M-estimator.

2.2 *M*-Estimators

Like many studies in the literature of estimation with complex survey data, this chapter employs the asymptotics along which the number of strata grows to infinity. Consider a sequence of populations, $\{\mathscr{P}_L\}_{L\in\mathbb{N}}$, such that $\mathscr{P}_L$ consists of N_L individuals, each of whom carries numerical values for a set of v characteristics, X. The population $\mathscr{P}_L$ is split into L strata. Let N_{Lh} denote the number of individuals in the hth stratum in population $\mathscr{P}_L$. The value of X for the kth individual in the hth

[1] Chen and Rao (2007) discuss large sample properties of statistical inferences in i.i.d. case taking into account both the characteristics of the finite population and the sampling design employed. They show that the law of large numbers which assumes that both sample size and population size increase to infinity does not hold in the context of finite population. However, asymptotic theory for independent but not identically distributed (i.n.i.d) observations in such context, which is essential for complex survey data, is unavailable and not straightforward to derive. We will study large sample properties of the proposed weighted M-estimator with correction for finite population size in our future work.

stratum in population $\mathscr{P}_L$, which is denoted $\tilde{X}_{hk}$, is drawn from a v-variate distribution P_{hk}, where the draws of $\tilde{X}_{hk}$ are independent across individuals and strata. In this setup, the product measure $\otimes_{h=1}^{L} \otimes_{k=1}^{N_{Lh}} P_{hk}$ can be viewed as the superpopulation distribution of X's behind the population $\mathscr{P}_L$. Note that P_{hk} does not depend on L. We assume:

Assumption 2.1. *The double array, $\tilde{X} \equiv \{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$, of $v \times 1$ random vectors on a probability space $(\Omega, \mathscr{F}, P)$ is independently (but not identically) distributed. For each $L \in \mathbb{N}$, the population $\mathscr{P}_L$ consists of L stratum, and for each $h \in \{1,2,\ldots,L\}$, the hth stratum contains N_{Lh} individuals ($N_{Lh} \in \mathbb{N}$). For each $(k,h,L) \in \mathbb{K} \equiv \{(\tilde{k},\tilde{h},\tilde{L}) \in \mathbb{N}^3 : \tilde{k} \leq N_{Lh}, \tilde{h} \leq \tilde{L}\}$, the kth individual in the hth stratum in the Lth population $\mathscr{P}_L$ carries $\tilde{X}_{hk}$. The array $\{N_{Lh} : (h,L) \in \mathbb{H}\}$ satisfies that*

$$\sup\{N_{Lh}/(N_L/L) : (h,L) \in \mathbb{H}\} < \infty, \tag{2.1}$$

where $N_L \equiv \sum_{h=1}^{L} N_{Lh}$, $L \in \mathbb{N}$.

In Assumption 2.1, we impose independence on $\tilde{X}$ for simplicity. Even if we assume that $\tilde{X}$ is weakly dependent in each stratum instead, the results of this chapter would stay essentially the same. Because N_L/L is the average stratum size, (2.1) requires that none of the stratum asymptotically dominates the others in size, reflecting that the sizes of strata are typically comparable to each other in practice.

A sample design is a mechanism to select individuals in the population into the sample in each of the strata. Suppose there are n_{Lh} psu's selected in stratum h and n_{Lhi} ultimate sampling units selected in the ith selected psu in stratum h of $\mathscr{P}_L$. Let C_{Lhk} denote the (random) number of times individual k in stratum h is selected into the data set under population $\mathscr{P}_L$. We assume that the sample design satisfies:

Assumption 2.2. *(a) For each $L \in \mathbb{N}$, the L collections of nonnegative integer-valued random variables,*

$$\{C_{L1k} : k \in \{1,\ldots,N_{L1}\}\}, \ldots, \{C_{LLk} : k \in \{1,\ldots,N_{LL}\}\}$$

which are defined on $(\Omega, \mathscr{F}, P)$, are independent.

(b) *The collection of random variables* $\{C_{Lhk} : (k,h,L) \in \mathbb{K}\}$ *is independent from* $\{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$.

(c) *There exists* $\bar{C} \in \mathbb{N}$ *such that for each* $(k,h,L) \in \mathbb{K}$, *the support of* C_{Lhk} *is contained in the interval* $[0,\bar{C}]$.

(d) *It holds that*

$$\bar{p}_l \equiv \inf\big\{P[C_{Lhk} > 0]/N_{Lh}{}^{-1} : (k,h,L) \in \mathbb{K}\big\} > 0$$

and that

$$\bar{p}_u \equiv \sup\big\{P[C_{Lhk} > 0]/N_{Lh}{}^{-1} : (k,h,L) \in \mathbb{K}\big\} < \infty.$$

Assumption 2.2(a) means that the sampling is independent across strata, as is the case virtually in all complex survey design. Assumption 2.2(b) reflects the fact that selection of individuals into the data set does not depend on the values of X drawn by the individuals. Assumption 2.2(c) requires that there be a maximum number of times an individual can be possibly selected into the sample, as is the case in typical sample designs (in fact, many of the designs allows an individual to be included in the sample at most once). Assumption 2.2(d) specifies that individuals' probabilities of selection into the sample proportional and the stratum population size are in inverse proportion in our asymptotics, loosely speaking.

The mean number of times the kth individual in stratum h in population $\mathscr{P}_L$ is selected into the sample is equal to $\mathrm{E}[C_{Lhk}]$. We can view $\mathrm{E}[C_{Lhk}]$ as showing how well each individual is represented in the sample design; the higher it is, the better the individual is represented. The reciprocal of $\mathrm{E}[C_{Lhk}]$, $w_{Lhk} \equiv \mathrm{E}[C_{Lhk}]^{-1}$, is called the survey weight of the kth individual in stratum h under the Lth population. The survey designer can compute the survey weight, precisely knowing the distribution of C_{Lhk}. We let W_{Lhij} denote the survey weight of the jth selected individual in the ith selected psu in stratum h under the Lth population $\mathscr{P}_L$. As Proposition 2.1 shows below, the survey weight can be used for unbiasedly estimating the population mean of variables in each stratum. In this chapter, the mixture of $P_{h1}, \ldots, P_{hN_{Lh}}$ with equal weights $1/N_{Lh}$ is called the distribution of X in the hth stratum in the Lth population and denoted $\bar{P}_{Lh}$. Also, the mixture of $\bar{P}_{Lh}, \ldots, \bar{P}_{LL}$ with the corresponding weights

$N_{L1}/N_L, \dots, N_{LL}/N_L$ is called the distribution of X in the entire Lth population and denoted $\bar{P}_L$.

Proposition 2.1. *Suppose that Assumptions 2.1 and 2.2 hold. Also, let* $\{\phi_{Lh} : (h,L) \in \mathbb{H} \equiv \{(\tilde{h},\tilde{L}) \in \mathbb{N}^2 : \tilde{h} \le \tilde{L}\}\}$ *be an array of Borel-measurable functions from* $\mathbb{R}^v$ *to* $\mathbb{R}$ *such that for each* $k \in \mathbb{N}$ *and each* $(h,L) \in \mathbb{H}$, $\int |\phi_{Lh}(x)| P_{hk}(dx) < \infty$. *Then:*

(a) For each $(h,L) \in \mathbb{H}$,

$$\mathrm{E}\left[N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}) \,\middle|\, \tilde{X}\right] = N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} \phi_{Lh}(\tilde{X}_{hk}).$$

(b) For each $(h,L) \in \mathbb{H}$,

$$\mathrm{E}\left[N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij})\right] = \int \phi_{Lh}(x) \bar{P}_{Lh}(dx).$$

(c) For each $L \in \mathbb{N}$,

$$\mathrm{E}\left[N_L^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij})\right] = \sum_{h=1}^{L} (N_{Lh}/N_L) \int \phi_{Lh}(x)\, \bar{P}_{Lh}(dx).$$

(d) If, in addition, $\phi_{L1} = \phi_{L2} = \cdots = \phi_{LL}$ *for each* $L \in \mathbb{N}$, *it holds that for each* $L \in \mathbb{N}$,

$$\mathrm{E}\left[N_L^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij})\right] = \int \phi_{L1}(x)\, \bar{P}_L(dx).$$

In Proposition 2.1, (i) shows that the sample average weighted by W_{Lhij}/N_{Lh} corrects the over- and under-representation of individuals and unbiasedly estimate the population mean (given the individuals' draws of X from the superpopulation). This property of the weighted average is called the design unbiasedness. The design unbiasedness immediately implies the unbiasedness of the weighted average, as (ii) claims. Given the unbiasedness of the weighted average, we can also estimate superpopulation means unbiasedly, taking the average of the unbiased stratum mean estimator weighted by N_{Lh}/N_L, as (iii) and (iv) states.

We now introduce the parameter of interest. Assume:

Assumption 2.3. *The set Θ is nonempty compact subset of $\mathbb{R}^p$ ($p \in \mathbb{N}$). The function $q : \mathbb{R}^v \times \Theta \to \mathbb{R}$ is Borel-measurable, and for each $x \in \mathbb{R}^v$, $q(x, \cdot) : \Theta \to \mathbb{R}$ is continuous.*

Our parameter of interest is characterized as the maximizer of $\bar{Q}_L : \Theta \to \mathbb{R}$ defined by

$$\bar{Q}_L(\theta) \equiv \int q(x,\theta)\bar{P}_L(dx) = \sum_{h=1}^{L} \frac{N_{Lh}}{N_L} \int q(x,\theta)\bar{P}_{Lh}(dx), \quad \theta \in \Theta.$$

For each $\theta \in \Theta$, $N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} q(X_{Lhij}, \theta)$ estimates $\int q(x,\theta)\bar{P}_{Lh}(dx)$ unbiasedly. A natural estimator of the parameter is a (weighted) M-estimator constructed based on this fact. For each $L \in \mathbb{N}$, define a function $Q_L : \Omega \times \Theta \to \mathbb{R}$ by

$$Q_L(\omega,\theta) \equiv \frac{1}{N_L} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\, q(X_{Lhij}(\omega), \theta),$$

and write $Q_L(\theta) \equiv Q_L(\cdot,\theta)$, $\theta \in \Theta$. The M-estimator is the maximizer of Q_L on Θ. We can easily verify the existence of the M-estimator by using the standard result on the existence of the measurable maximum (Gallant and White, 1988, Lemma 2.1).

Theorem 2.2. *Suppose that Assumptions 2.1–2.3 hold. Then for each $L \in \mathbb{N}$, there exists a Θ-valued random vector $\hat{\theta}_L$ such that $Q_L(\hat{\theta}_L) = \sup_{\theta\in\Theta} Q_L(\theta)$.*

2.3 Consistency

In studying the large sample behavior of the M-estimators in our setup, the following result is useful.

Proposition 2.1. *Suppose that Assumptions 2.1 and 2.2 hold. Let Γ a set, a a nonnegative real number, and $\{\phi_{Lh} : (h,L) \in \mathbb{H}, \}$ an array of measurable function from $\mathbb{R}^v \times \Gamma$ to $\mathbb{R}$ such that for each $\gamma \in \Gamma$ and each $(h,L) \in \mathbb{H}$, $\phi_{Lh}(\cdot,\gamma) : \mathbb{R}^v \to \mathbb{R}$ is*

Borel measurable. If

$$\sup_{\gamma\in\Gamma}\sup_{(h,L)\in\mathbb{H}}\int|\phi_{Lh}(x,\gamma)|^a P_{kh}(dx)<\infty,$$

then it holds that

$$\left\{\left\|N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\phi_{Lh}(X_{Lhij},\gamma)\right\|_a : (h,L)\in\mathbb{H},\, \gamma\in\Gamma\right\}$$

is uniformly $\mathscr{L}_a$-bounded.

Proposition 2.1 says that if the stratum mean of superpopulation is uniformly $\mathscr{L}_a$-bounded, its unbiased estimator is also $\mathscr{L}_a$-bounded.

For compactly stating the assumptions we employ to establish the consistency and asymptotic normality results, we introduce a terminology.

Definition 1. *Given Assumption 2.1, let Γ be a finite-dimensional Euclidean space and $\{\phi_h\}_{h\in\mathbb{N}}$ a sequence of measurable functions from $(\mathbb{R}^v\times\Gamma,\mathscr{B}^v\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$. We say that $\{\phi_h\}$ is* LB(a) on Γ*, where $a\in[1,\infty)$, if the following conditions are satisfied.*

- *(a) For each $\gamma\in\Gamma$, $\{\phi_h(\tilde{X}_{hk},\gamma) : (k,h)\in\mathbb{N}^2\}$ is uniformly $\mathscr{L}_a$-bounded.*
- *(b) There exist a continuous function $g:\mathbb{R}\to[0,\infty)$ and a sequence of Borel-measurable functions, $\{d_h:\mathbb{R}^v\to[0,\infty)\}_{h\in\mathbb{N}}$ such that $g(y)\downarrow 0$ as $y\downarrow 0$,* $\sup_{(k,h)\in\mathbb{N}^2}\int d_h(x)P_{hk}(dx)<\infty$*, and for each $(\gamma_1,\gamma_2)\in\Gamma^2$ and each $x\in\mathbb{R}^v$, $|\phi_h(x,\gamma_2)-\phi_h(x,\gamma_1)|\le d_h(x)g(|\gamma_2-\gamma_1|)$.*

In Definition 1, the term “LB” comes from “Lipschitz and bounded”.

In establishing our consistency result, we use the following concept.

Definition 2. *Given Assumption 2.1 and 2.2, let Γ be a finite-dimensional Euclidean space and $\{\Phi_{Lh} : (h,L)\in\mathbb{H}\}$ an array of measurable functions from $(\Omega\times\Gamma,\mathscr{F}\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$. We say that $\{\Phi_{Lh}\}$ is* SLB(a) on Γ*, where $a\in[1,\infty)$, if the following conditions are satisfied (“S” stands for “stratum-wise”).*

- *(a) For each $\gamma\in\Gamma$, $\{\Phi_{Lh}(\cdot,\gamma) : (h,L)\in\mathbb{H}\}$ is uniformly $\mathscr{L}_a$-bounded.*

(b) There exists a continuous function $g : \mathbb{R} \to [0,\infty)$ *such that* $g(y) \to 0$ *as* $y \downarrow 0$*, and an array of nonnegative random variables* $\{D_{Lh} : (h,L) \in \mathbb{H}\}$ *on* $(\Omega, \mathscr{F}, P)$ *that satisfies that*

$$\sup_{L\in\mathbb{N}} L^{-1} \sum_{h=1}^{L} \mathrm{E}[D_{Lh}] < \infty,$$

and for each $(\gamma_1,\gamma_2) \in \Gamma^2$ *and each* $(h,L) \in \mathbb{H}$*,* $|\Phi_{Lh}(\cdot,\gamma_2) - \Phi_{Lh}(\cdot,\gamma_1)| \leq D_{Lh}\, g(|\gamma_2 - \gamma_1|)$*.*

The SLB property is useful in our analysis, being closely related to the LB property used in describing some of the assumptions imposed in the main text.

Lemma 2.2. *Suppose that Assumptions 2.1 and 2.2 hold. Let* Γ *be a finite-dimensional Euclidean space, a a positive real number, and* $\{\phi_h\}_{h\in\mathbb{N}}$ *a sequence of measurable functions from* $(\mathbb{R}^v \times \Gamma, \mathscr{B}^v \otimes \mathscr{B}(\Gamma))$ *to* $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$ *that is LB(a) on* Γ*. Then the array of functions from* $(\mathbb{R}^v \times \Gamma, \mathscr{B}^v \otimes \mathscr{B}(\Gamma))$ *to* $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$*,* $\{\Phi_{Lh}\}_{L\in\mathbb{N}}$*, defined by*

$$\Phi_{Lh}(\omega,\gamma) \equiv N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\, \phi_h(X_{Lhij}(\omega),\gamma), \quad \omega \in \Omega,\, \gamma \in \Gamma,\, (h,L) \in \mathbb{H}$$

is SLB(a) on Γ*.*

In establishing the consistency result, we require:

Assumption 2.4. *For some real number* $\delta > 0$*, the function q is* $LB(1+\delta)$ *on* Θ*.*

Assumption 2.4 along with the preceding assumptions are sufficient for the uniform convergence of $\{Q_L(\theta) - \bar{Q}_L(\theta)\}_{L\in\mathbb{N}}$ in probability to zero over $\theta \in \Theta$. For each $L \in \mathbb{N}$, let Θ_L^* denote the set of maxima of $\bar{Q}_L$ on Θ. To turn uniform convergence of the estimation objective function into the consistency of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$, we need identifiability of $\{\Theta_L^*\}_{L\in\mathbb{N}}$, namely:

Assumption 2.5. *For any real number* $\varepsilon > 0$

$$\liminf_{L\to\infty} \inf\{\bar{Q}_L(\theta) - \bar{Q}_L^* : d(\theta,\Theta_L^*) \geq \varepsilon,\, \theta \in \Theta\} > 0,$$

where $\bar{Q}_L^* \equiv \sup_{\theta\in\Theta} \bar{Q}_L(\theta)$, and d is the Euclidean metric on Θ, so that

$$d(\theta,\Theta_L^*) = \inf_{\theta^*\in\Theta_L^*} d(\theta,\theta^*), \quad \theta\in\Theta, L\in\mathbb{N}.$$

Assumption 2.5 rules out the possibility that as L goes to infinity, the $\bar{Q}_L(\theta)$ function gets flatter around maxima.

We are now ready state our consistency result.

Theorem 2.3. *Suppose that Assumption 2.1–2.5 hold. Then $\{d(\hat{\theta}_L,\Theta_L^*)\}_{L\in\mathbb{N}}$ converges in probability to zero.*

An interesting special case of our setup is the case in which the model is correctly specified for each stratum. For each $(h,L)\in\mathbb{H}$, define the function $\bar{Q}_{Lh}:\Theta\to\mathbb{R}$ by

$$\bar{Q}_{Lh}(\theta) \equiv \int q(x,\theta)\bar{P}_{Lh}(dx).$$

We mean by the stratum-wise correct specification:

Assumption 2.6. *There exists $\theta_0\in\Theta$ such that for each $(h,L)\in\mathbb{H}$, $\bar{Q}_{Lh}$ is maximized at θ_0 over Θ. Also, Θ_L^* is a singleton for almost all $L\in\mathbb{N}$.*

When all strata share the same true parameter value, this parameter value is also the true parameter value of the superpopulation, i.e. $\theta_L^* = \theta_0$. As $\hat{\theta}_L$ is consistent for θ_L^* (Theorem 2.3), it is also consistent for θ_0.

Corollary 2.4. *Suppose that Assumption 2.1–2.6 hold. Then $\{d(\hat{\theta}_L,\theta_0)\}_{L\in\mathbb{N}}$ converges in probability to zero.*

2.4 Asymptotic Normality

We apply the standard linearization approach with smooth (generalized) scores to achieve the asymptotic normality.

Assumption 2.7. *For almost all $L\in\mathbb{N}$, Θ_L^* is a singleton, and there exists a sequence $\{\theta_L^*\in\Theta_L^*\}_{L\in\mathbb{N}}$ and a compact set $\Theta_0\subset\operatorname{int}\Theta$, to which $\{\theta_L^*\}$ is uniformly interior.*

Assumption 2.7 rules out the case that $\{\theta_L^*\}$ is on the boundary of Θ, so that $\nabla\bar{Q}_L^* = 0$.

Assumption 2.8. *(a) For each $x \in \mathbb{R}^v$, $q(x,\cdot)$ is twice continuously differentiable on* $\operatorname{int}\Theta$.

(b) For some real number $\delta > 0$, $\nabla^2 q$ is LB$(1+\delta)$ on Θ_0.

(c) The sequence $\{A_L^* \equiv A_L(\theta_L^*)\}_{L\in\mathbb{N}}$ is asymptotically uniformly nonsingular, where for each $L \in \mathbb{N}$, $A_L : \operatorname{int}\Theta \to \mathbb{R}^{p\times p}$ is defined by

$$A_L(\theta) \equiv \int \nabla^2 q(x,\theta)\bar{P}_L(dx), \quad \theta \in \operatorname{int}\Theta.$$

(d) For some real number $\delta > 0$,

$$\sup_{\theta\in\Theta_0} \sup_{(k,h)\in\mathbb{N}^2} \int |\nabla q(x,\theta)|^{2+\delta} P_{hk}(dx) < \infty.$$

(e) For each $(k,h) \in \mathbb{N}^2$ and each $\theta \in \Theta_0$,

$$\nabla \int q(x,\theta)\, P_{hk}(dx) = \int \nabla q(x,\theta)\, P_{hk}(dx).$$

(f) The sequence $\{B_L^* \equiv B_L(\theta_L^*)\}_{L\in\mathbb{N}}$ is asymptotically uniformly nonsingular, where for each $\theta \in \Theta$, $B_L(\theta) \equiv L^{-1}\sum_{h=1}^{L} B_{Lh}(\theta)$; for each $\theta \in \Theta$ and each $(h,L) \in \mathbb{H}$, $B_{Lh}(\theta) \equiv \operatorname{var}[S_{Lh}(\cdot,\theta)]$ and

$$\begin{aligned} &S_{Lh}(\omega,\theta) \\ &\equiv (L/N_L)\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\nabla q(X_{lhij}(\omega),\theta) \\ &= (LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\nabla q(X_{lhij}(\omega),\theta), \quad \omega \in \Omega,\ \theta \in \operatorname{int}\Theta. \end{aligned}$$

Assumption 2.8 allows us to derive the standard asymptotic linear representation of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ and establish the asymptotic normality of $\hat{\theta}_L$ based on the asymptotic normality of the score evaluated at θ_L^*. Most *M*-estimators used in econometrics satisfy the condition (a), though it rules out some estimators such as Koenker and Bassett (1978) quantile regression estimator. The conditions (b), (d), and (e) are mild, requiring that the gradient and Hessian of q does not have too fat tails under the distributions in $\{P_{hk} : (k,h) \in \mathbb{N}^2\}$. The asymptotic uniform nonsingularity of $\{A^*_L\}_{L\in\mathbb{N}}$ and $\{B^*_L\}$ imposed in (c) and (f) are also easily satisfied, as long as $\bar{P}_L$ does not change too much as L grows to infinity.

We now state the asymptotic normality result.

Theorem 2.1. *Suppose that Assumption 2.1–2.5, 2.7, and 2.8 hold. Then the sequence* $\{D^*_L \equiv A^{*-1}_L B^*_L A^{*-1}_L\}_{L\in\mathbb{N}}$ *is bounded and uniformly nonsingular. Also, it holds that* $D^{*-1/2}_L L^{1/2}(\hat{\theta}_L - \theta^*_L) \overset{A}{\sim} \mathrm{N}(0, I_p)$ *as* $L \to \infty$.

2.5 Estimation of the Asymptotic Covariance Matrix

Consistent estimation of D^*_L in Theorem 2.1 can be performed by consistently estimating A^*_L and B^*_L. The proof of the asymptotic normality shows that $\{\nabla^2 Q_L\}_{L\in\mathbb{N}}$ is uniformly consistent for $\{A_L\}_{L\in\mathbb{N}}$ over Θ_0. Let $\{\hat{A}_L : \Omega \to \mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$ be a sequence of random matrices such that $\hat{A}_L = \nabla^2 Q_L(\hat{\theta}_L)$ whenever $\hat{\theta}_L \in \operatorname{int}\Theta$. Given the consistency of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ for $\{\theta^*_L\}_{L\in\mathbb{N}}$ and the above mentioned uniform consistency of $\{\nabla^2 Q_L\}_{L\in\mathbb{N}}$ for $\{A_L\}$, it is straightforward to verify consistency of $\{\hat{A}_L\}_{L\in\mathbb{N}}$ for $\{A^*_L\}_{L\in\mathbb{N}}$.

We next consider estimation of B^*_L. For a wide range of multistage cluster sample designs, the literature offers design-unbiased estimators of the covariance matrix of the total estimator, where the design unbiasedness means the conditional unbiasedness given $\tilde{X}$. (Wolter, 1985, pp. 11–16), for example, lists many of such estimators. We here assume that such an estimator is available for estimation of the variance of $\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\nabla q(X_{lhij}, \theta)$ in each stratum and each $\theta \in \Theta_0$.

Assumption 2.9. *An array of measurable functions from* $(\Omega\times\Theta, \mathscr{F}\otimes\mathscr{B}(\Theta))$ *to* $(\mathbb{R}^{p\times p}, \mathscr{B}^{p\times p})$, $\{\tilde{K}_{Lh} : \Omega\times\Theta \to \mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$, *satisfies that for each* $\theta \in \operatorname{int}\Theta$ *and each* $(h,L) \in \mathbb{H}$,

$$\mathrm{E}[\tilde{K}_{Lh}(\cdot,\theta)\,|\,\tilde{X}] = \operatorname{var}\left[\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\nabla q(X_{lhij}(\omega),\theta)\right],$$

and for some real number $\delta > 0$, $\{N^{-2}_{Lh}\tilde{K}_{Lh} : (h,L) \in \mathbb{H}\}$ *is SLB*$(1+\delta)$ *on* Θ_0.

The requirement that $\{N^{-2}_{Lh}\tilde{K}_{Lh} : (h,L) \in \mathbb{H}\}$ be SLB$(1+\delta)$ is satisfied by the prevailing estimators of the covariance matrix of the total estimators, given the assumptions already imposed in this chapter.

Define $\{\tilde{B}_{Lh} : \Omega \times \Theta \to \mathbb{R}^{p\times p} : (h,L) \in \mathbb{H}\}$ by

$$\begin{aligned}\tilde{B}_{Lh}(\omega,\theta) &\equiv (L/N_L)^2 \tilde{K}_{Lh}(\omega,\theta) \qquad (2.2)\\ &= (LN_{Lh}/N_L)^2 N_{Lh}^{-2} \tilde{K}_{Lh}(\omega,\theta), \quad \omega \in \Omega,\, \theta \in \Theta, (h,L) \in \mathbb{H}.\end{aligned}$$

Under Assumption 2.9, $(L/N_L)^2\tilde{K}_{Lh}(\cdot,\theta)$ is a design-unbiased estimator of the (conditional) covariance matrix of $S_{Lh}(\cdot,\theta)$ for each $\theta \in \Theta_0$, i.e.,

$$\mathrm{E}[(L/N_L)^2\tilde{K}_{Lh}(\cdot,\theta)\,|\,\tilde{X}] = \mathrm{var}[S_{Lh}(\cdot,\theta)\,|\,\tilde{X}], \quad \theta \in \Theta_0,\, (h,L) \in \mathbb{H}.$$

Nevertheless, its unconditional mean is

$$\begin{aligned}\mathrm{E}[\tilde{B}_{Lh}(\cdot,\theta)] &= \mathrm{E}\big[\mathrm{var}[S_{Lh}(\cdot,\theta)\,|\,\tilde{X}]\big]\\ &= \mathrm{var}[S_{Lh}(\cdot,\theta)] - \mathrm{var}\big[\mathrm{E}[S_{Lh}(\cdot,\theta)\,|\,\tilde{X}]\big], \quad \theta \in \Theta_0,\, (h,L) \in \mathbb{H},\end{aligned}$$

where the first equality follows by the law of iterated expectations, and the second equality follows from the fact that the mean conditional variance and the variance of the conditional mean sums up to the unconditional mean. Because

$$\mathrm{E}[S_{Lh}(\cdot,\theta)\,|\,\tilde{X}] = (LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{k=1}^{N_{Lh}} \nabla q(\tilde{X}_{hk},\theta), \quad \theta \in \Theta_0,\, (h,L) \in \mathbb{H}$$

by Proposition 2.1, it holds that $\mathrm{E}[\tilde{B}_{Lh}(\cdot,\theta)]$ is biased for $\mathrm{var}[S_{Lh}(\cdot,\theta)]$ by

$$\mathrm{var}\big[\mathrm{E}[S_{Lh}(\cdot,\theta)\,|\,\tilde{X}]\big] = (LN_{Lh}/N_L)^2 N_{Lh}^{-2}\sum_{k=1}^{N_{Lh}} \mathrm{var}\big[\nabla q(\tilde{X}_{hk},\theta)\big],\ \theta \in \Theta_0,\, (h,L) \in \mathbb{H}. \qquad (2.3)$$

This bias is zero if $\nabla q(\tilde{X}_{hk},\theta)$ is degenerate (as is the case in the finite population framework). If $\nabla q(\tilde{X}_{hk},\theta)$ is not degenerate, it is negative definite. Under our current assumption, however, $\{LN_{Lh}/N_L : (h,L) \in \mathbb{H}\}$ is bounded (Assumption 2.1), and it holds that

$$\begin{aligned}\sup\Big\{\big|\operatorname{var}[\nabla q(\tilde{X}_{hk},\theta)]\big| : \theta\in\Theta_0,\,(k,h)\in\mathbb{N}^2\Big\}\\ \le \sup\Big\{\big|\mathrm{E}[\nabla q(\tilde{X}_{hk},\theta)\nabla' q(\tilde{X}_{hk},\theta)]\big| : \theta\in\Theta_0,\,(k,h)\in\mathbb{N}^2\Big\}\\ +\sup\Big\{\big|\mathrm{E}[\nabla q(\tilde{X}_{hk},\theta)]\,\mathrm{E}[\nabla' q(\tilde{X}_{hk},\theta)]'\big| : \theta\in\Theta_0,\,(k,h)\in\mathbb{N}^2\Big\}<\infty\end{aligned} \tag{2.4}$$

(Assumption 2.8), the size of the bias is asymptotically governed by N_{Lh}. In sum, the bias of $\tilde{B}_{Lh}(\cdot,\theta)$ in estimation of $\operatorname{var}[S_{Lh}(\cdot,\theta)]$ is zero or asymptotically negligible uniformly in all strata if:

Assumption 2.10. *(a) All random vectors of $\tilde{X}$ are degenerate; or*

(b) it holds that $\liminf_{L\to\infty}\inf_{h\in\{1,2,\ldots,L\}} N_{Lh}=\infty$.

Intuitively, in finite population framework, there is no random draws from the superpopulation. Hence, conditional variance equals unconditional variance. Given (b), stratum size of the population is large enough. So that the population converges to superpopulation in each stratum. Hence, the conditional variance converges to the unconditional variance.

We now define $\{\tilde{B}_L : \Omega\times\Theta\to\mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$ by

$$\tilde{B}_L(\omega,\theta)\equiv L^{-1}\sum_{h=1}^{L}\tilde{B}_{Lh}(\omega,\theta),\quad \omega\in\Omega,\,\theta\in\Theta.$$

Application of a uniform law of large numbers to the triangle array $\{\tilde{B}_{Lh} : (h,L)\in\mathbb{H}\}$ establishes the uniform consistency of $\{\tilde{B}_L\}$ for $\{B_L\}$ over Θ_0, which leads to the following result along with the consistency of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ for $\{\theta^*_L\}_{L\in\mathbb{N}}$.

Theorem 2.1. *Under Assumptions 2.1–2.5 and 2.7–2.10, $\{\hat{B}_L\equiv\tilde{B}_L(\cdot,\hat{\theta}_L)\}_{L\in\mathbb{N}}$ is consistent for $\{B_L\}_{L\in\mathbb{N}}$.*

Chapter 3

m-Testing of Stratum-Wise Model Specification in Complex Survey Data

3.1 Introduction

An economic model is often estimated by using complex survey data obtained from a population. The estimated model is then sometimes used to analyze phenomena in a subpopulation, which is typically a stratum in the complex survey design or a union of strata.

In order for stratum-specific analysis using a model estimated for the entire population to be valid, it is sufficient that (a) the model is correctly specified for each stratum, and (b) all strata share the same true parameter value. We refer to the combination of (a) and (b) as *stratum-wise correct specification of the model*. It is worthwhile to test the specification of the model in strata, before conducting stratum-specific analyses. For example, the Current Population Survey (CPS), one of the principal source of data on U.S. labour market, is collected using complex survey sampling method. It provides comprehensive information about individuals' employment status and their demographic characteristics, and is used by state and local government, which is usually a stratum or a union of a small number of strata in the CPS, for planning and budgeting purposes and to determine the need for the local employment and training services. Due to the features of complex survey sampling of the CPS [1], it is not possible to estimate a model for only one or a few strata, and the estimation requires using the sample of the whole population or a large number of strata (Please refer to Sakata and Xu (2010) for a discussion about estimation with complex survey data). The application of the estimated model to local labour markets requests that the estimated model also holds for local labour markets.

In this chapter, we consider how to test the stratum-wise specification of a model within the *m*-testing framework, which was pioneered by Newey (1985) and Tauchen (1985) and extended by White (1987). Our approach is closely related to the one taken in Sakata (2009), in which the data are assumed to be collected by simple random sampling in each of many subpopulations.

Suppose that there are L strata in the population, and n_h primary sampling units (psu) are drawn from stratum $h \in \{1, 2, \ldots, L\}$. In the ith psu in stratum h, n_{hi} observations are drawn according to the sample design (n_{hi} is in general random). Let X_{hij} denote the jth observation of $\mathbb{R}^v$-valued variable X from the ith psu in stratum h, and W_{hij} the survey weight for the observation.

Also, suppose that a model with a parameter space $\Theta \subset \mathbb{R}^p$ is given. The model may be designed to capture some conditional probability, conditional expectation, conditional variance, or something else. Also, suppose that it is known that when the given model is correctly specified for stratum h, for each π in a set $\Pi \subset \mathbb{R}^r$,

$$\int m_h(x, \theta_{0h}, \pi) \bar{P}_h(dx) = 0,$$

where $m_h : \mathbb{R}^v \times \Theta \times \Pi \to \mathbb{R}^q$ is a function known to the researcher, θ_{0h} is the true parameter value for stratum h (the parameter π is usually referred to as a nuisance parameter), and $\bar{P}_h$ is the probability distribution of X in stratum h. By using the

[1] The CPS a monthly survey which interviews a representative sample of the civilian noninstitutional population 16 years and older. It splits the U.S. population into 792 strata, where each stratum is a subregion of a state or in some cases an entire state. On average 2.5 primary sample units (PSU) are randomly selected for each stratum. Within the PSUs, the CPS directly samples ultimate sampling units (USU), a geographically compact group of approximately four addresses.

survey weights W_{hij}'s, we can rewrite the above equality as

$$\mathrm{E}\left[\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{hij} m_h(X_{hij}, \theta_{0h}, \pi)\right] = 0, \quad h = 1, 2, \ldots, L$$

(see Proposition 3.1(iv) below). Under the stratum-wise correct specification, the single true parameter θ_0 shared by all strata satisfies that for each π in a set $\Pi \subset \mathbb{R}^r$,

$$\mathrm{E}\left[\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{hij} m_h(X_{hij}, \theta_0, \pi)\right] = 0, \quad h = 1, 2, \ldots, L. \tag{3.1}$$

In many applications, θ_0 can be accurately estimated by an estimator $\hat{\theta}$ using the entire sample under the stratum-wise correct specification.

Just for the sake of illustration, suppose that Π is a singleton whose only element is π^*. Then a usual m-test would employ the statistic $L\hat{M}'\hat{C}\hat{M}$, where

$$\hat{M} \equiv \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{hij} \hat{m}_{hij},$$

$$\hat{m}_{hij} \equiv m_h(X_{hij}, \hat{\theta}, \pi^*),$$

and $\hat{C}$ is a suitably chosen $q \times q$ weighting matrix, which is typically the inverse of an estimated covariance matrix of $\hat{M}$. Because $\hat{M}$ approximates

$$\bar{M}^* \equiv \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} \mathrm{E}[W_{hij} m_h(X_{hij}, \theta^*, \pi^*)]$$

the m-test would detect the violation of the null hypothesis only through the deviation of $\bar{M}^*$ from the origin. When all strata have the same distribution, (3.1) is equivalent to zeroness of $\bar{M}^*$. When the stratum have heterogeneous distributions, on the other hand, it is possible that $\bar{M}^*$ is nearly equal to zero, even when (3.1) is grossly violated in some stratum. It is therefore preferable to design a test that directly checks (3.1), instead of the zeroness of $\bar{M}^*$ in out current setup.

A natural way to formulate a test of (3.1) within the standard m-testing framework is to consider moment conditions for each stratum separately, by taking

$$\hat{m}^{\dagger}_{hij} \equiv (\underbrace{0,0,\ldots,0}_{(h-1)q},\hat{m}'_{hij},\underbrace{0,0,\ldots,0}_{(L-h)q})'$$

for the moment function of the jth observation in the ith psu in the hth stratum. One might find such approach related to the growing literature on use of many moment conditions: Koenker and Machado (1999) Carrasco and Florens (2000), Donald *et al.* (2003), Doran and Schmidt (2006), Han and Phillips (2006), Carrasco *et al.* (2007), and Anatolyev (2008), just to mention a few chapters in the area. Nevertheless, it is important to note that all elements of $\hat{m}^{\dagger}_{hij}$ are zeros, except for the q elements in $\hat{m}_{hij}$. Such moment functions are ruled out by the regularity conditions in the existing literature at our best knowledge. There are two major problems in the approach using $\hat{m}^{\dagger}$. First, the usual normal approximation to the distribution of $\hat{M}^{\dagger} \equiv \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\hat{m}^{\dagger}_{hij}$ would not work well in typical complex survey designs, in which only a few psu's are selected in each stratum. Second, the weighting matrix $\hat{C}^{\dagger}$ would have a serious problem, because the estimated asymptotic covariance matrix of $\hat{M}^{\dagger}$ is either nearly or exactly singular given the small number of psu's per stratum. Thus, the behavior of $\hat{M}^{\dagger\prime}\hat{C}^{\dagger}\hat{M}^{\dagger}$ would be very different from what the standard asymptotics suggests. The use of the usual m-test based on the moment conditions separately set up for each stratum is not appealing for this reason.

In this chapter, we propose an implementation of m-testing that overcomes the above-mentioned difficulties, being hinted by Sakata (2009). In our approach, we *unbiasedly* estimate a quadratic form of the moment vector in question for each stratum, where the quadratic form has a positive definite weighting matrix, and then take the average of the estimated quadratic form over the strata to obtain a statistic. We consider both the situation in which the nuisance parameter π is estimated (i.e., picked based on data) and the situation in which a researcher desires to base the test on the whole moment conditions indexed by the nuisance parameter. Our test rejects the null hypothesis when the statistic described above is largely positive, loosely speaking.

The rest of the chapter is organized as follows. In Section 3.2, we formalize our problem setup. In Section 3.3, we then propose a method to test the null hypothesis and derive its large sample properties in the case in which an estimated nuisance parameter is used. In Section 3.4, we remove the estimation of the nuisance parameter from the setup and consider the test based on the whole moment conditions indexed by the nuisance parameter. The proofs of the theorems are collected in the Appendices.

We employ the following convention and symbols throughout this chapter. Limits are taken along the sequence of numbers of stratum (denoted L) growing to infinity, unless otherwise indicated. For each matrix A, $|A|$ denotes the Frobenius norm of A, i.e., $|A| \equiv \sqrt{\mathrm{tr}(A'A)}$, and A^+ the Moore-Penrose (MP) inverse of A. By applying the MP inverse in division by scalars, we rule that division by zero equals zero. We use the MP inverses of random matrices instead of the regular inverses to avoid technical problems caused by the singularity of the random matrices that occurs with a small probability. The reader can safely replace the MP inverses with the regular inverses, when applying the formulas in practice. Also, for each random matrix Z and positive real number a, $\|Z\|_a$ denotes the $\mathscr{L}_a$-norm of $|Z|$.

3.2 Problem Setup

We now formalize the problem setup described in Section 3.1. Like many studies in the literature of estimation with complex survey data, this chapter employs the asymptotics along which the number of strata grows to infinity. Consider a sequence of populations, $\{\mathscr{P}_L\}_{L\in\mathbb{N}}$, such that $\mathscr{P}_L$ consists of N_L individuals, each of whom carries numerical values for a set of v characteristics, X. The population $\mathscr{P}_L$ is split into L strata. Let N_{Lh} denote the number of individuals in the hth stratum in population $\mathscr{P}_L$. The value of X for the kth individual in the hth stratum in population $\mathscr{P}_L$, which is denoted $\tilde{X}_{hk}$, is drawn from a v-variate distribution P_{hk}, where the draws of $\tilde{X}_{hk}$ are independent across individuals and strata. In this setup, the product measure $\otimes_{h=1}^{L} \otimes_{k=1}^{N_{Lh}} P_{hk}$ can be viewed as the superpopulation distribution of X's behind the population $\mathscr{P}_L$. Note that P_{hk} does not depend on L. We assume:

Assumption 3.1. *The double array, $\tilde{X} \equiv \{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$, of $v \times 1$ random vectors on a probability space $(\Omega, \mathscr{F}, P)$ is independently (but not identically) distributed. For each $L \in \mathbb{N}$, the population $\mathscr{P}_L$ consists of L stratum, and for each $h \in \{1,2,\ldots,L\}$, the hth stratum contains N_{Lh} individuals ($N_{Lh} \in \mathbb{N}$). For each*

$(k,h,L) \in \mathbb{K} \equiv \{(\tilde{k},\tilde{h},\tilde{L}) \in \mathbb{N}^3 : \tilde{k} \le N_{Lh}, \tilde{h} \le \tilde{L}\}$, the kth individual in the hth stratum in the Lth population $\mathscr{P}_L$ carries $\tilde{X}_{hk}$. The array $\{N_{Lh} : (h,L) \in \mathbb{H}\}$ satisfies that

$$\sup\{N_{Lh}/(N_L/L) : (h,L) \in \mathbb{H}\} < \infty, \tag{3.2}$$

where $N_L \equiv \sum_{h=1}^{L} N_{Lh}$, $L \in \mathbb{N}$.

In Assumption 3.1, we impose independence on $\tilde{X}$ for simplicity. Even if we assume that $\tilde{X}$ is weakly dependent in each stratum instead, the results of this chapter would stay essentially the same. Because N_L/L is the average stratum size, (3.2) requires that none of the stratum asymptotically dominates the others in size, reflecting that the sizes of strata are typically comparable to each other in practice.

A sample design is a mechanism to select individuals in the population into the sample in each of the strata. Suppose there are n_{Lh} psu's selected in stratum h and n_{Lhi} ultimate sampling units selected in the ith selected psu in stratum h of $\mathscr{P}_L$. Let C_{Lhk} denote the (random) number of times individual k in stratum h is selected into the data set under population $\mathscr{P}_L$. We assume that the sample design satisfies:

Assumption 3.2. *(a) For each $L \in \mathbb{N}$, the L collections of nonnegative integer-valued random variables,*

$$\big\{C_{L1k} : k \in \{1,\dots,N_{L1}\}\big\}, \dots, \big\{C_{LLk} : k \in \{1,\dots,N_{LL}\}\big\}$$

which are defined on $(\Omega,\mathscr{F},P)$, are independent.

(b) The collection of random variables $\{C_{Lhk} : (k,h,L) \in \mathbb{K}\}$ is independent from $\{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$.

(c) There exists $\bar{C} \in \mathbb{N}$ such that for each $(k,h,L) \in \mathbb{K}$, the support of C_{Lhk} is contained in the interval $[0,\bar{C}]$.

(d) It holds that

$$\bar{p}_l \equiv \inf\big\{P[C_{Lhk} > 0]/{N_{Lh}}^{-1} : (k,h,L) \in \mathbb{K}\big\} > 0$$

and that

$$\bar{p}_u \equiv \sup\bigl\{P[C_{Lhk} > 0]/{N_{Lh}}^{-1} : (k,h,L) \in \mathbb{K}\bigr\} < \infty.$$

Assumption 3.2(a) means that the sampling is independent across strata, as is the case virtually in all complex survey design. Assumption 3.2(b) reflects the fact that selection of individuals into the data set does not depend on the values of X drawn by the individuals (this effectively requires that the variables in X constitutes the sampling frames are fixed be constant under distribution P_{hk} for each $(k,h) \in \mathbb{N}^2$). Assumption 3.2(c) requires that there be a maximum number of times an individual can be possibly selected into the sample, as is the case in typical sample designs (in fact, many of the designs allows an individual to be included in the sample at most once). Assumption 3.2(d) specifies that individuals' probabilities of selection into the sample proportional and the stratum population size are in inverse proportion in our asymptotics, loosely speaking.

The mean number of times the kth individual in stratum h in population $\mathscr{P}_L$ is selected into the sample is equal to $\mathrm{E}[C_{Lhk}]$. We can view $\mathrm{E}[C_{Lhk}]$ as indicating how well each individual is represented in the sample design; the higher it is, the better the individual is represented. The reciprocal of $\mathrm{E}[C_{Lhk}]$, $w_{Lhk} \equiv \mathrm{E}[C_{Lhk}]^{-1}$, is called the survey weight of the kth individual in stratum h under the Lth population. The survey designer can compute the survey weight, precisely knowing the distribution of C_{Lhk}. We let W_{Lhij} denote the survey weight of the jth selected individual in the ith selected psu in stratum h under the Lth population $\mathscr{P}_L$. As Proposition 3.1 shows below, the survey weight can be used for unbiasedly estimating the population mean of variables in each stratum. In this chapter, the mixture of $P_{h1}, \ldots, P_{hN_{Lh}}$ with equal weights $1/N_{Lh}$ is called the distribution of X in the hth stratum in the Lth population and denoted $\bar{P}_{Lh}$. Also, the mixture of $\bar{P}_{Lh}, \ldots, \bar{P}_{LL}$ with the corresponding weights $N_{L1}/N_L, \ldots, N_{LL}/N_L$ is called the distribution of X in the entire Lth population and denoted $\bar{P}_L$.

Proposition 3.1. *Suppose that Assumptions 3.1 and 3.2 hold. Also, let* $\{\phi_{Lh} : (h,L) \in \mathbb{H} \equiv \{(\tilde{h},\tilde{L}) \in \mathbb{N}^2 : \tilde{h} \le \tilde{L}\}\}$ *be an array of Borel-measurable functions from* $\mathbb{R}^v$ *to* $\mathbb{R}$ *such that for each* $k \in \mathbb{N}$ *and each* $(h,L) \in \mathbb{H}$, $\int |\phi_{Lh}(x)| P_{hk}(dx) < \infty$. *Then:*

(a) For each $(h,L) \in \mathbb{H}$*,*

$$\mathrm{E}\left[N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\phi_{Lh}(X_{Lhij})\,\middle|\,\tilde{X}\right] = N_{Lh}\sum_{k=1}^{N_{Lh}}\phi_{Lh}(\tilde{X}_{hk}).$$

(b) For each $(h,L) \in \mathbb{H}$*,*

$$\mathrm{E}\left[N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\phi_{Lh}(X_{Lhij})\right] = \int \phi_{Lh}(x)\bar{P}_{Lh}(dx).$$

(c) For each $L \in \mathbb{N}$*,*

$$\mathrm{E}\left[N_{L}^{-1}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\phi_{Lh}(X_{Lhij})\right] = \sum_{h=1}^{L}(N_{Lh}/N_L)\int \phi_{Lh}(x)\,\bar{P}_{Lh}(dx).$$

(d) If, in addition, $\phi_{L1} = \phi_{L2} = \cdots = \phi_{LL}$ *for each* $L \in \mathbb{N}$*, it holds that for each* $L \in \mathbb{N}$*,*

$$\mathrm{E}\left[N_{L}^{-1}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\phi_{Lh}(X_{Lhij})\right] = \int \phi_{L1}(x)\,\bar{P}_{L}(dx).$$

In Proposition 3.1, (i) shows that the sample average weighted by W_{Lhij}/N_{Lh} corrects the over- and under-representation of individuals and unbiasedly estimate the stratum population mean (given the individuals' draws of X from the superpopulation). This property of the weighted average is called the design unbiasedness. The design unbiasedness immediately implies the unbiasedness of the weighted average for the stratum mean, as (ii) claims. Given the unbiasedness of the weighted average, we can also estimate superpopulation means unbiasedly, taking the average of the unbiased stratum mean estimator weighted by N_{Lh}/N_L, as (iii) and (iv) states.

In our large-L asymptotics, we often check the moment conditions for averages weighted by the survey weights. The result result is handy in such tasks.

Proposition 3.2. *Suppose that Assumptions 3.1 and 3.2 hold. Let* Γ *a set, a a nonnegative real number, and* $\{\phi_{Lh} : (h,L) \in \mathbb{H}, \}$ *an array of measurable function from* $\mathbb{R}^v \times \Gamma$ *to* $\mathbb{R}$ *such that for each* $\gamma \in \Gamma$ *and each* $(h,L) \in \mathbb{H}$*,* $\phi_{Lh}(\cdot,\gamma) : \mathbb{R}^v \to \mathbb{R}$ *is*

Borel measurable. If

$$\sup_{\gamma\in\Gamma}\sup_{(h,L)\in\mathbb{H}}\int|\phi_{Lh}(x,\gamma)|^a P_{kh}(dx)<\infty,$$

then it holds that

$$\left\{\left\|N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\phi_{Lh}(X_{Lhij},\gamma)\right\|_a : (h,L)\in\mathbb{H},\,\gamma\in\Gamma\right\}$$

is uniformly $\mathscr{L}_a$-bounded.

By this proposition, we see that the $\mathscr{L}_a$-boundedness of a function of $\tilde{X}$ uniform both in individuals and the index γ nicely translates into the uniform $\mathscr{L}_a$-boundedness of its unbiased estimator under Assumptions 3.1 and 3.2.

Following Section 3.1, we now assume that the stratum-wise correct specification implies a set of moment conditions that there exists θ_0 in the parameter space Θ such that for each $h\in\mathbb{N}$ and for each π in a space Π, $\bar{m}_{Lh}(\theta,\pi)=0$, where

$$\bar{m}_{Lh}(\theta,\pi)\equiv\int m_h(x,\theta,\pi)\,\bar{P}_{Lh}(dx),\quad (\theta,\pi)\in\Theta\times\Pi.$$

and the sets Θ and Π and the function $m_h:\mathbb{R}^v\times\Pi\times\Omega\to\mathbb{R}^q$ satisfy:

Assumption 3.3. *The sets Θ and Π are nonempty Borel-measurable subsets of the p- and r-dimensional Euclidean spaces ($p<\infty$, $r<\infty$), respectively, and $\{m_h:\mathbb{R}^v\times\Theta\times\Pi\to\mathbb{R}^q\}_{h\in\mathbb{N}}$ is a sequence of functions measurable-$(\mathscr{B}^v\otimes\mathscr{B}(\Theta)\otimes\mathscr{B}(\Pi))/\mathscr{B}^q$ such that for each $(h,k)\in\mathbb{N}^2$ and each $(\theta,\pi)\in\Theta\times\Pi$, $\int|m_h(x,\theta,\pi)|\,P_{hk}(dx)<\infty$.*

Like the usual m-tests, the tests proposed in this chapter requires an estimator $\hat{\theta}_L$ that is consistent for the true parameter θ_0 under the stratum-wise correct specification. The estimator may be an M-estimator, a GMM-estimator, or others, as long as it satisfies the next assumption regardless of whether or not the specification is stratum-wise correct.

Assumption 3.4. *The set Θ_0 is a compact subset of Θ. A Θ-valued estimator $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ on $(\Omega,\mathscr{F},P)$ is consistent for a sequence $\{\theta_L^*\in\Theta_L\}_{L\in\mathbb{N}}$, i.e., $\{|\hat{\theta}_L-\theta_L^*|\}_{L\in\mathbb{N}}$ converges to zero in probability-P.*

Under the stratum-wise correct specification, the "pseudo-true parameter" θ_L^* coincides with θ_0, so that it holds that

$$\text{for each } \pi \in \Pi \text{ and each } (h,L) \in \mathbb{H}, \quad \bar{m}_{Lh}(\theta_L^*, \pi) = 0. \tag{3.3}$$

The test we develop in this chapter rejects stratum-wise correct specification if data exhibit evidence against (3.3).

A few examples of situations that fulfill the requirements in Assumptions 3.3 and 3.4 follow. e.g., Assume that $\{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$ is uniformly $\mathscr{L}_{2+2\delta}$-bounded for some positive real number δ. Let Y denote the first characteristic in X, and Z the remaining $v-1$ characteristics in X ($v \geq 2$). Partition $\tilde{X}_{hk}$ as $\tilde{X}_{hk} = (\tilde{Y}_{hk}, \tilde{Z}'_{hk})'$, where $\tilde{Y}_{hk}$ is a random variable, and $\tilde{Z}_{hk}$ is a $(v-1) \times 1$ random vector, and set $\Theta = \mathbb{R}^{v-1}$. We say that the linear regression of Y on Z is stratum-wise correct for the conditional mean of Y given Z, if there exists $\theta_0 \in \Theta$ such that the conditional mean of $Y - Z'\theta_0$ given Z is zero under each of the stratum distributions $\bar{P}_{Lh}$ ($(h,L) \in \mathbb{H}$). By using the law of large numbers of (Sen, 1970, Theorem 3) and Proposition 3.2, we can easily verify that the weighted least squares estimator $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ in regression of Y_{Lhij} on Z_{Lhij} using the survey weight is consistent for

$$\left\{\theta_L^* \equiv \left(\int zz'\, \bar{P}_L(dy,dz)\right)^{-1} \int zy\, \bar{P}_L(dy,dz)\right\}_{L\in\mathbb{N}},$$

provided that $\{\int zz'\, \bar{P}_L(dy,dz)\}_{L\in\mathbb{N}}$ is uniformly nonsingular. Also, it holds that $\theta_L^* = \theta_0$ for each $L \in \mathbb{N}$ under the stratum-wise correct specification.

Let g be a Borel-measurable function from $\mathbb{R}^{v-1}$ to $\mathbb{R}^q$ such that $\int g(z)'g(z)\, P_{hk}(dy,dz) < \infty$ for each $(h,k) \in \mathbb{N}^2$. Then the stratum-wise correct specification of the linear regression implies that

$$\int g(z)\,(y - z'\theta_L^*)\, \bar{P}_{Lh}(dy,dz) = 0, \quad (h,L) \in \mathbb{H}. \tag{3.4}$$

Though this condition involves no nuisance parameter, we can create one artificially to put this example squarely in the framework described above. Set $\Pi = \mathbb{R}$ and

$$m_h(x,\theta,\pi) = g(z)\,(y - z'\theta), \quad x = (y,z) \in \mathbb{R} \times \mathbb{R}^{v-1}, (\theta,\pi) \in \Theta \times \Pi, h \in \mathbb{N}.$$

Then (3.4) is written in the form of (3.3). e.g., In the setup of Example 3.2, assume in addition that $\int (z'z)^3 P_{hk}(dy,dz) < \infty$ for each $(h,k) \in \mathbb{N}^2$. Also, let $\Pi = \mathbb{R}^{v-1}$. Then the stratum-wise correct specification implies that

$$\int (z'\pi)^j (y - z'\theta_L^*) \bar{P}_{Lh}(dy,dz) = 0, \quad \pi \in \Pi, \, j \in \{2,3\}, \, (h,L) \in \mathbb{H}. \tag{3.5}$$

If we set

$$m_h(x,\theta,\pi) = (z'\pi, \, (z'\pi)^2)' \, (y - z'\theta), \quad x = (y, z') \in \mathbb{R} \times \mathbb{R}^{v-1}, \, (\theta,\pi) \in \Theta \times \Pi, \, h \in \mathbb{N}.$$

Then (3.5) is written in the form of (3.3). e.g., In the setup of Example 3.2, let $\Phi : \mathbb{R}^{v-1} \to \mathbb{R}^{v-1}$ be a bounded, Borel-measurable, bounded one-to-one function and Π a nonempty compact subset of $\mathbb{R}^{v-1}$ with a positive Lebesgue measure. Then the stratum-wise correct specification is equivalent to that there exists $\theta_0 \in \Theta$ such that

$$\int \exp(\Phi(z)'\pi) \, (y - z'\theta_0) \, \bar{P}_{Lh}(dy,dz) = 0, \quad \pi \in \Pi, \, (h,L) \in \mathbb{H}, \tag{3.6}$$

as can be verified using (Bierens, 1990, Theorem 1). If we set

$$m_h(x,\theta,\pi) = \exp(\Phi(z)'\pi) \, (y - z'\theta), \quad x = (y,z')' \in \mathbb{R} \times \mathbb{R}^{v-1}, \, (\theta,\pi) \in \Theta \times \Pi, \, h \in \mathbb{N},$$

(3.6) is written in the form of (3.3).

3.3 A Test with Estimated Nuisance Parameters

Let $\{\pi_L^*\}_{L \in \mathbb{N}}$ be an arbitrary sequence in Π. Then the condition (3.3) apparently implies that

$$\text{for each } (h,L) \in \mathbb{H}, \quad \bar{m}_{Lh}(\theta_L^*, \pi_L^*) = 0, \tag{3.7}$$

which further implies that

$$\text{for each } L \in \mathbb{N}, \quad \sum_{h=1}^{L} \beta_{Lh} \bar{m}_{Lh}(\theta_L^*, \pi_L^*) = 0, \tag{3.8}$$

where β_{Lh}'s are positive real numbers to weight strata. If we take the usual *m*-testing approach, (3.8) would be the basis for our test. In Example 3.2, for example one might set $\pi_L^* = \theta_L^*$ and obtain a variant of Ramsey (1969) regression specification error test (RESET).

Nevertheless, the usual implementation of the m-testing based on (3.8) can be problematic for our purpose, as we discussed in Section 3.1. To overcome the difficulty, we here formulate a test based on the average of the quadratic form of $\bar{m}_{Lh}^* \equiv \bar{m}_{Lh}(\theta_L^*, \pi_L^*)$, taken over the L strata with some weights β_{Lh}, where

$$\bar{m}_{Lh}(\theta, \pi) \equiv \int m_h(x, \theta, \pi)\, \bar{P}_{Lh}(dx), \quad (\theta, \pi) \in \Theta \times \Pi,\, (h, L) \in \mathbb{H},$$

and the quadratic form has a positive definite weighting matrix. We assume that

Assumption 3.5. *The array of positive real numbers, $\{\beta_{Lh} : (h, L) \in \mathbb{H}\}$, satisfies that for each $L \in \mathbb{N}$, $\sum_{h=1}^{L} \beta_{Lh} = 1$, and that*

$$\limsup_{L \to \infty} L \sup\big\{\beta_{Lh} : h \in \{1, 2, \dots, L\}\big\} < \infty.$$

For example, one may set N_{Lh}/N_L or $N_{Lh}^2 / \sum_{h'=1}^{L} N_{Lh'}^2$ to β_{Lh}. It is straightforward to verify that the required conditions are satisfied by these choices under (3.2) in Assumption 3.1.

Let $\mathbb{S}^q$ denote the set of all $q \times q$ positive semidefinite symmetric matrices, and $\Lambda_L \in \mathbb{S}^q$ a positive definite matrix possibly dependent on L. Note that for each $(h, L) \in \mathbb{H}$, $\bar{m}_{Lh}^{*\prime} \Lambda_L \bar{m}_{Lh}^* \geq 0$; and $\bar{m}_{Lh}^{*\prime} \Lambda_L \bar{m}_{Lh}^* = 0$ if and only if $\bar{m}_{Lh}^* = 0$. Define $\alpha_L : \Theta \times \Pi \times \mathbb{S}^q \to \mathbb{R}$ by

$$\begin{aligned}\alpha_L(\theta, \pi, \Lambda) &\equiv \sum_{h=1}^{L} \beta_{Lh} \bar{m}_{Lh}(\theta, \pi)' \Lambda \bar{m}_{Lh}(\theta, \pi) \\ &= \sum_{h=1}^{L} \beta_{Lh} \mathrm{tr}(\Lambda \bar{m}_{Lh}(\theta, \pi) \bar{m}_{Lh}(\theta, \pi)') \quad (\theta, \pi, \Lambda) \in \Theta \times \Pi \times \mathbb{S}^q, \quad L \in \mathbb{N}.\end{aligned} \tag{3.9}$$

Then $\alpha_L^* \equiv \alpha_L(\theta_L^*, \pi_L^*, \Lambda_L)$ is nonnegative for each $L \in \mathbb{N}$; and $\alpha_L^* = 0$ for a given $L \in \mathbb{N}$ if and only if $\bar{m}_{Lh} = 0$ for each $h \in \{1, 2, \dots, L\}$. Thus, it holds that $\alpha_L^* = 0$ for each $L \in \mathbb{N}$ if and only if (3.7) holds. We base our test of stratum-wise correct specification on this simple fact. Namely, our test rejects the stratum-wise correct specification if an estimate of α_L^* is positive and far from zero.

Before formulating an estimator of α_L^*, we introduce two terminologies, which let us below state our assumptions compactly.

Definition 3. *Given Assumption 3.1, let Γ be a finite-dimensional Euclidean space and $\{\phi_h\}_{h\in\mathbb{N}}$ a sequence of measurable functions from $(\mathbb{R}^v \times \Gamma, \mathscr{B}^v \otimes \mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$. We say that $\{\phi_h\}$ is* LB(a) on Γ*, where $a \in [1,\infty)$, if the following conditions are satisfied.*

(a) For each $\gamma \in \Gamma$, $\{\phi_h(\tilde{X}_{hk},\gamma) : (k,h) \in \mathbb{N}^2\}$ is uniformly $\mathscr{L}_a$-bounded.

(b) There exist a continuous function $g : \mathbb{R} \to [0,\infty)$ and a sequence of Borel-measurable functions, $\{d_h : \mathbb{R}^v \to [0,\infty)\}_{h\in\mathbb{N}}$ such that $g(y) \downarrow 0$ as $y \downarrow 0$, $\sup_{(k,h)\in\mathbb{N}^2} \int d_h(x) P_{hk}(dx) < \infty$, and for each $(\gamma_1,\gamma_2) \in \Gamma^2$ and each $x \in \mathbb{R}^v$, $|\phi_h(x,\gamma_2) - \phi_h(x,\gamma_1)| \le d_h(x) g(|\gamma_2 - \gamma_1|)$.

Definition 4. *Given Assumption 3.1 and 3.2, let Γ be a finite-dimensional Euclidean space and $\{\Phi_{Lh} : (h,L) \in \mathbb{H}\}$ an array of measurable functions from $(\Omega\times\Gamma, \mathscr{F}\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$. We say that $\{\Phi_{Lh}\}$ is* SLB(a) on Γ*, where $a \in [1,\infty)$, if the following conditions are satisfied ("S" stands for "stratum-wise").*

(a) For each $\gamma \in \Gamma$, $\{\Phi_{Lh}(\cdot,\gamma) : (h,L) \in \mathbb{H}\}$ is uniformly $\mathscr{L}_a$-bounded.

(b) There exists a continuous function $g : \mathbb{R} \to [0,\infty)$ such that $g(y) \to 0$ as $y \downarrow 0$, and an array of nonnegative random variables $\{D_{Lh} : (h,L) \in \mathbb{H}\}$ on $(\Omega,\mathscr{F},P)$ that satisfies that

$$\sup_{L\in\mathbb{N}} L^{-1} \sum_{h=1}^{L} \mathrm{E}[D_{Lh}] < \infty,$$

and for each $(\gamma_1,\gamma_2) \in \Gamma^2$ and each $(h,L) \in \mathbb{H}$, $|\Phi_{Lh}(\cdot,\gamma_2) - \Phi_{Lh}(\cdot,\gamma_1)| \le D_{Lh}\, g(|\gamma_2 - \gamma_1|)$.

The term "LB" comes from "Lipschitz and bounded". The SLB property is useful in our analysis, being closely related to the LB property used in describing some of the assumptions imposed in the main text.

Lemma 3.1. *Suppose that Assumptions 3.1 and 3.2 hold. Let Γ be a finite-dimensional Euclidean space, a a positive real number, and $\{\phi_h\}_{h\in\mathbb{N}}$ a sequence of measurable functions from $(\mathbb{R}^v \times \Gamma, \mathscr{B}^v \otimes \mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$ that is LB(a) on Γ. Then the array of functions from $(\mathbb{R}^v \times \Gamma, \mathscr{B}^v \otimes \mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$, $\{\Phi_{Lh}\}_{L\in\mathbb{N}}$, defined by*

$$\Phi_{Lh}(\omega,\gamma) \equiv N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\, \phi_h(X_{Lhij}(\omega),\gamma), \quad \omega\in\Omega,\, \gamma\in\Gamma,\, (h,L)\in\mathbb{H}$$

is SLB(a) on Γ.

We now formulate an estimator of α_L^* and establishing its consistency. Assume:

Assumption 3.6. *(a) The sequence $\{\theta_L^*\}_{(h,L)\in\mathbb{H}}$ is uniformly interior to Θ_0, a compact subset of Θ.*

(b) The Π-valued estimator $\{\hat{\pi}_L\}_{L\in\mathbb{N}}$ on $(\Omega,\mathscr{F},P)$ is consistent for a sequence $\{\pi_L^* \in \Pi\}_{L\in\mathbb{N}}$ that is uniformly interior to Π_0, a compact subset of Π.

(c) The sequence $\{m_h\}_{h\in\mathbb{N}}$ is LB($2+2\delta$) on $\Theta_0\times\Pi_0$ for some real number $\delta>0$.

(d) The sequence $\{\Lambda_L \in \mathbb{S}^q\}_{L\in\mathbb{N}}$ is bounded.

Assumptions 3.6 is mild. For example, we consider how the required conditions can be satisfied in Example 3.2. For each $(k,h)\in\mathbb{N}^2$, let $\tilde{Y}_{hk}$ denote the first component of $\tilde{X}_{hk}$, and $\tilde{Z}_{hk}$ a vector consisting of the second to last component of $\tilde{X}_{hk}$. Assume that $\{\tilde{Y}_{kh} : (k,h)\in\mathbb{N}^2\}$ and $\{\tilde{Z}_{kh} : (k,h)\in\mathbb{N}^2\}$ are uniformly $\mathscr{L}_{4+4\delta}$- and $\mathscr{L}_{12+12\delta}$-bounded, respectively, where δ is some positive real number. Set $\pi_L^* = \theta_L^*$ and $\hat{\pi}_L = \hat{\theta}_L$. Then $\{\theta_L^*\}_{L\in\mathbb{N}}$ is uniformly bounded, so that we can choose for Θ_0 a compact subset of $\mathbb{R}^{v-1}$, to which $\{\theta_L^*\}_{L\in\mathbb{N}}$ is uniformly interior. We then set $\Pi_0=\Theta_0$. Assumptions 3.6(a) and (b) are now satisfied. We can also easily verify that Assumptions 3.6(c) is fulfilled. Of course, the uniform $\mathscr{L}_{12+12\delta}$-boundedness of $\{\tilde{Y}_{hk}\}$ may look too restrictive in some applications. In such a case, we should choose a different $\{m_h : h\in\mathbb{N}\}$ that results in less stringent moment requirements.

In formulating an estimator of α_L^*, we first consider how to estimate $\alpha_L(\theta,\pi,\Lambda)$ with fixed $\theta\in\Theta_0$, $\pi\in\Pi_0$, and $\Lambda\in\mathbb{S}^q$. Note that $\alpha_L(\theta,\pi,\Lambda)$ is a weighted average of

$$\mathrm{tr}(\Lambda\,\bar{m}_{Lh}(\theta,\pi)\,\bar{m}_{Lh}(\theta,\pi)') \tag{3.10}$$

taken over the L strata, with weights β_{Ln}. If we can estimate (3.10) with an asymptotically negligible bias for each $h \in \{1,\dots,L\}$, plugging the asymptotically unbiased estimator into (3.9) generates an estimator that converges to $\alpha_L(\theta,\pi,W)$ in probability-P by the law of large numbers for independent but not identically distributed processes.

For each $(h,L) \in \mathbb{H}$, define $\tilde{m}_{Lh} : \Omega\times\Theta\times\Pi \to \mathbb{R}^q$ by

$$\tilde{m}_{Lh}(\omega,\theta,\pi) \equiv {N_{Lh}}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\, m_h(X_{Lhij}(\omega),\theta,\pi), \quad (\omega,\theta,\pi)\in\Omega\times\Theta\times\Pi.$$

Then $\tilde{m}_{Lh}(\cdot,\theta,\pi)$ is an unbiased estimator of $\bar{m}_{Lh}(\theta,\pi)$ by Proposition 3.1. Nevertheless, $\tilde{m}_{Lh}(\cdot,\theta,\pi)\tilde{m}_{Lh}(\cdot,\theta,\pi)'$ is biased for $\bar{m}_{Lh}(\theta,\pi)\bar{m}_{Lh}(\theta,\pi)'$ by $\mathrm{var}[\tilde{m}_{Lh}(\cdot,\theta,\pi)]$. To correct the bias, we need to estimate it. For a wide range of multistage cluster sample designs, the literature offers design-unbiased estimators of the covariance matrix of the stratum total estimator $N_{Lh}\tilde{m}_{Lh}(\cdot,\theta,\pi) = \sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij} m_h(X_{Lhij},\theta,\pi)$, where the design unbiasedness means the unbiasedness for the conditional covariance matrix of $N_{Lh}\tilde{m}_{Lh}(\cdot,\theta,\pi)$ given $\tilde{X}$. (Wolter, 1985, pp. 11–16), for example, lists many of such estimators. We here assume the availability of such an estimator in each stratum and each $(\theta,\pi)\in\Theta\times\Pi$.

Assumption 3.7. *The array* $\{\tilde{\Sigma}_{Lh} : \Omega\times\Theta\times\Pi \to \mathbb{S}^q\}_{(h,L)\in\mathbb{H}}$ *satisfies that for each* $(\theta,\pi)\in\Theta\times\Pi$ *and each* $(h,L)\in\mathbb{H}$, $\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)$ *is measurable-*$\mathscr{F}/\mathscr{B}(\mathbb{S}^q)$*. Also, the array* $\{N_{Lh}^{-2}\tilde{\Sigma}_{Lh}\}$ *is SLB(*$1+\delta$*) on* $\Theta_0\times\Pi_0$ *for some real number* $\delta > 0$*. Further, it holds that*

$$\mathrm{E}[\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}] = \mathrm{var}[N_{Lh}\tilde{m}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}].$$

Given the estimator $\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)$, a design-unbiased estimator of $\mathrm{var}[\tilde{m}_{Lh}(\cdot,\theta,\pi)]$ is $N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)$. Our estimator $\breve{\alpha}_L(\cdot,\theta,\pi,\Lambda) : \Omega\to\mathbb{R}$ of $\alpha_L(\theta,\pi,\Lambda)$ is obtained by replacing $\bar{m}_{Lh}(\theta,\pi)\bar{m}_{Lh}(\theta,\pi)'$ with $\tilde{m}_h(\cdot,\theta,\pi)\tilde{m}_h'(\cdot,\theta,\pi) - N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)$ in

(3.9):

$$\check{\alpha}_L(\omega,\theta,\pi,\Lambda) \equiv \sum_{h=1}^{L} \beta_{Lh}\,\mathrm{tr}\big(\Lambda\big(\tilde{m}_{Lh}(\omega,\theta,\pi)\tilde{m}_{Lh}(\omega,\theta,\pi)' \\ - N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\omega,\theta,\pi)\big)\big), \quad \theta \in \Theta, \Lambda \in \mathbb{S}^q, L \in \mathbb{N}.$$

Nevertheless, this estimator is not exactly unbiased, because the unconditional mean of $N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)$ is

$$\begin{aligned} \mathrm{E}[N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)] &= \mathrm{E}[\,\mathrm{var}[\tilde{m}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}]\,] \\ &= \mathrm{var}[\tilde{m}_{Lh}(\cdot,\theta,\pi)] \\ &\quad - \mathrm{var}\big[\,\mathrm{E}[\tilde{m}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}]\,\big], \quad (\theta,\pi) \in \Theta\times\Pi, (h,L) \in \mathbb{H}, \end{aligned}$$

where the first equality follows by the law of iterated expectations, and the second equality follows from the fact that the mean conditional variance and the variance of the conditional mean sums up to the unconditional mean. Because

$$\mathrm{E}[\tilde{m}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}] = N_{Lh}^{-1}\sum_{k=1}^{N_{Lh}} m_h(\tilde{X}_{hk},\theta,\pi), \quad (\theta,\pi) \in \Theta\times\Pi, (h,L) \in \mathbb{H}$$

by Proposition 3.1, $N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)$ is biased for $\mathrm{var}[\tilde{m}_{Lh}(\cdot,\theta,\pi)]$ by

$$\mathrm{var}\big[\,\mathrm{E}[\tilde{m}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}]\,\big] = N_{Lh}^{-2}\sum_{k=1}^{N_{Lh}} \mathrm{var}[m_h(\tilde{X}_{hk},\theta,\pi)], \ (\theta,\pi) \in \Theta\times\Pi, (h,L) \in \mathbb{H}.$$

Thus, $\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)$ is biased for $\alpha_L(\theta,\pi,\Lambda)$ by

$$\begin{aligned} \mathrm{E}[\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)] - \alpha_L(\theta,\pi,\Lambda) &= -\sum_{h=1}^{L}\beta_{Lh}\mathrm{tr}\Big(\Lambda\,\mathrm{var}\big[\mathrm{E}[\tilde{m}_{Lh}(\cdot,\theta,\pi)\,|\,\tilde{X}]\big]\Big) \\ = -\sum_{h=1}^{L}\beta_{Lh}N_{Lh}^{-2}&\sum_{k=1}^{N_{Lh}}\mathrm{tr}\Big(\Lambda\,\mathrm{var}[m_h(\tilde{X}_{hk},\theta,\pi)\Big), \quad (\theta,\pi) \in \Theta\times\Pi, L \in \mathbb{N}. \end{aligned}$$

The bias is zero if $m_h(\tilde{X}_{hk},\theta,\pi)$ is degenerate (as is the case in the finite population setup). If $m_h(\tilde{X}_{hk},\theta)$ is not degenerate, it holds under our current assumption that

$$\begin{aligned}
&\sup\Big\{\big|\operatorname{var}[m_h(\tilde{X}_{hk},\theta,\pi)]\big| : (\theta,\pi)\in\Theta_0\times\Pi_0,\,(k,h)\in\mathbb{N}^2\Big\}\\
&\leq \sup\Big\{\big|\mathrm{E}[m_h(\tilde{X}_{hk},\theta,\pi)m_h(\tilde{X}_{hk},\theta,\pi)]\big| : (\theta,\pi)\in\Theta_0\times\Pi_0,\,(k,h)\in\mathbb{N}^2\Big\}\\
&+\sup\Big\{\big|\mathrm{E}[m_h(\tilde{X}_{hk},\theta,\pi)]\,\mathrm{E}[m_h(\tilde{X}_{hk},\theta,\pi)]'\big| : (\theta,\pi)\in\Theta_0\times\Pi_0,\,(k,h)\in\mathbb{N}^2\Big\}<\infty,
\end{aligned}\tag{3.11}$$

so that the size of the bias of $\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)$ for $\alpha_L(\theta,\pi,\Lambda)$ is $O(1/\min\{N_{Lh} : h\in\{1,2,\ldots,L\}\})$ uniformly in $(\theta,\pi,\Lambda)\in\Theta_0\times\Pi_0\times\mathbb{\Lambda}$, where $\mathbb{\Lambda}$ is any compact subset of $\mathbb{S}^q$. Thus, we can conclude that the bias is zero or asymptotically negligible if:

Assumption 3.8. *(a) All random vectors in $\tilde{X}$ are degenerate; or*

(b) it holds that $\liminf_{L\to\infty}\min_{h\in\{1,2,\ldots,L\}}N_{Lh}=\infty$.

Applying a suitable uniform law of large numbers to the array

$$\begin{aligned}
&\big\{\beta_{Lh}\operatorname{tr}\big(\Lambda\big(\tilde{m}_{Lh}(\omega,\theta,\pi)\tilde{m}_{Lh}(\omega,\theta,\pi)'-N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\omega,\theta,\pi)\big)\big) : (h,L)\in\mathbb{H}\big\},\\
&\qquad(\theta,\pi)\in\Theta_0\times\Pi_0,\,\Lambda\in\mathbb{S}^q
\end{aligned}$$

after demeaning it establishes that $\{\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)-\mathrm{E}[\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)]\}_{L\in\mathbb{N}}$ converges to zero in probability-P uniformly in $(\theta,\pi,\Lambda)\in\Theta_0\times\Pi_0\times\mathbb{\Lambda}$, where $\mathbb{\Lambda}$ is a compact subset of $\mathbb{S}^q$, to which $\{\Lambda_L\}_{L\in\mathbb{N}}$ is uniformly interior. Thus, if Assumption 3.8 holds, $\{\check{\alpha}_L\}_{L\in\mathbb{N}}$ is uniformly consistent for $\{\alpha_L\}_{L\in\mathbb{N}}$ on $\Theta_0\times\Pi_0\times\mathbb{\Lambda}$. In estimation of α_L^*, we use $\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)$, replacing θ with the estimator $\hat{\theta}_L$ of θ_L^*, π with an estimator $\hat{\pi}_L$ of π_L^*, and Λ with Λ_L to obtain an estimator of α_L^*. We can also allow for use of a data-dependent weighting matrix $\hat{\Lambda}_L$ instead of Λ_L, provided:

Assumption 3.9. *The sequence of random matrices, $\{\hat{\Lambda}_L : \Omega\to\mathbb{S}^q\}_{L\in\mathbb{N}}$, satisfies that $\{|\hat{\Lambda}_L-\Lambda_L|\}_{L\in\mathbb{N}}$ converges to zero in probability-P.*

Our estimator $\{\hat{\alpha}_L\equiv\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\hat{\Lambda}_L)\}_{L\in\mathbb{N}}$ is consistent for $\{\alpha_L^*\}_{L\in\mathbb{N}}$, given the consistency of $\{(\hat{\theta}_L,\hat{\pi}_L,\hat{\Lambda}_L)\}_{L\in\mathbb{N}}$ for $\{(\theta_L^*,\pi_L^*,\Lambda_L)\}_{L\in\mathbb{N}}$ and the uniform consistency of $\{\check{\alpha}_L\}_{L\in\mathbb{N}}$ for $\{\alpha_L\}_{L\in\mathbb{N}}$ over $\Theta_0\times\Pi_0\times\mathbb{\Lambda}$.

Theorem 3.2. *Under Assumptions 3.1–3.9, $\{\alpha_L^*\}_{L\in\mathbb{N}}$ is bounded, and $\{\hat{\alpha}_L - \alpha_L^*\}_{L\in\mathbb{N}}$ converges to zero in probability-P.*

For the weighting matrix, one may desire to take the inverse of $\sum_{h=1}^{L} \beta_{Lh}\mathrm{var}[\tilde{m}_{Lh}^*]$ for the weighting matrix, where $\tilde{m}_{Lh}^* \equiv \tilde{m}_{Lh}(\cdot,\theta^*,\pi^*)$, because it allows the test to take into account the "average" noisiness of $\tilde{m}_{Lh}^*$. Given that $\sum_{h=1}^{L} \beta_{Lh}N_{Lh}^{-2}\mathrm{var}[\tilde{m}_{Lh}^*]$ is unknown, we would use the inverse of $\sum_{h=1}^{L} \beta_{Lh}N_{Lh}^{-2}\hat{\Sigma}_{Lh}$, for the weighting matrix, where $\hat{\Sigma}_{Lh} \equiv \tilde{\Sigma}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)$. The estimated weighting matrix can be shown to be consistent for the desired one under our current assumptions.

Remark. It would certainly be more desirable to use a stratum-specific weighting matrix to reflect the noisiness of $\tilde{m}_{Lh}^*$ in each stratum h. Nevertheless, such group-specific weighting matrix cannot be estimated accurately. We thus require that the same weighting matrix be used in every stratum.

We can also derive the large-L distribution of $\{\hat{\alpha}_L\}_{L\in\mathbb{N}}$ under the stratum-wise correct specification, imposing a few additional assumptions. Let ∇, ∇_θ, and ∇_π denote the gradient operator with respect to $(\theta',\pi')'$, only θ, and only π, respectively.

Assumption 3.10. *(a) There exists a bounded sequence of $p\times p$ matrices, $\{J_L^*\}_{L\in\mathbb{N}}$, and a sequence of Borel-measurable functions from $\mathbb{R}^v\times\Theta$ to $\mathbb{R}^p$, $\{\psi_h\}_{h\in\mathbb{N}}$, such that $\{\psi_h\}_{h\in\mathbb{N}}$ is LB($2+2\delta$) on Θ_0, $\{\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\psi_{Lhij}^* : (h,L)\in\mathbb{H}\}$ is a zero-mean array, where $\psi_{Lhij}^* \equiv \psi_h(X_{Lhij},\theta_L^*)$, and*

$$L^{1/2}(\hat{\theta}_L-\theta_L^*) = J_L^* L^{-1/2}\sum_{h=1}^{L}\frac{N_{Lh}}{N_L/L}{N_{Lh}}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\psi_{Lhij}^* + \mathrm{o}_P(1). \quad (3.12)$$

(b) $\hat{\pi}_L-\pi_L^* = \mathrm{O}_P(L^{-1/2})$.

(c) For each $h\in\mathbb{N}$ and each $x\in\mathbb{R}^v$, $m_h(x,\cdot,\cdot):\Theta\times\Pi\to\mathbb{R}^v$ is continuously differentiable, and $\{\nabla m_h:\mathbb{R}^v\times\Theta\times\Pi\to\mathbb{R}^{p\times q}\}_{h\in\mathbb{N}}$ is LB($2+2\delta$) on $\Theta_0\times\Pi_0$ for some real number $\delta>0$.

(d) For each $(h,L)\in\mathbb{H}$ and each $x\in\mathbb{R}^v$, each element of $\tilde{\Sigma}_{Lh}(x,\cdot,\cdot):\Theta\times\Pi\to\mathbb{R}^v$ is continuously differentiable, and $\{N_{Lh}^{-2}\nabla(\mathrm{vech}(\tilde{\Sigma}_{Lh})) : (h,L)\in\mathbb{H}\}$ is SLB($1+\delta$) on $\Theta_0\times\Pi_0$ for some real number $\delta>0$.

(e) The sequence $\{m_h\}_{h\in\mathbb{N}}$ is LB($4+4\delta$) on $\Theta_0\times\Pi_0$ for some real number $\delta>0$.

(f) The array $\{N_{Lh}^{-2}\tilde{\Sigma}_{Lh}\}_{(h,L)\in\mathbb{H}}$ is SLB($2+2\delta$) on $\Theta_0\times\Pi_0$ for some real number $\delta>0$.

(g) The array $\{\mathrm{var}[\xi^*_{Lh}]:(h,L)\in\mathbb{H}\}$ is uniformly positive, where

$$\begin{aligned}\xi^*_{Lh}\equiv{}&L\beta_{Lh}\mathrm{tr}\Big(\Lambda_L\Big(\tilde{m}^*_{Lh}\tilde{m}^*_{Lh}{}'-N_{Lh}^{-2}\tilde{\Sigma}^*_{Lh}\Big)\Big)\\&+G^*_{1,L}{}'J^*_L(LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\psi^*_{Lhij},\quad(h,L)\in\mathbb{H},\end{aligned}$$

$\tilde{m}^*_{Lh}\equiv\tilde{m}_{Lh}(\cdot,\theta^*_L,\pi^*_L)$, $\tilde{\Sigma}^*_{Lh}\equiv\tilde{\Sigma}(\cdot,\theta^*_L,\pi^*_L)$, $G^*_{1L}\equiv G_{1L}(\theta^*_L,\pi^*_L,\Lambda_L)$, and

$$\begin{aligned}G_{1L}(\theta,\pi,\Lambda)&\equiv\mathrm{E}[\nabla_\theta\tilde{\alpha}_L(\cdot,\theta,\pi,\Lambda)]\\&=\sum_{h=1}^{L}\beta_{Lh}\Big(2N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\mathrm{E}\big[W_{Lhij}\nabla_\theta m_h(X_{Lhij},\theta,\pi)\Lambda\tilde{m}_{Lh}(\cdot,\theta,\pi)\big]\\&\qquad-N_{Lh}^{-2}\mathrm{E}\big[\nabla_\theta(\mathrm{vec}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi))\big]\,\mathrm{vec}\Lambda\Big),\\&\qquad(\theta,\pi)\in\Theta\times\Pi,\,\Lambda\in\mathbb{S}^q,\,L\in\mathbb{N}.\end{aligned}$$

(h) If all random vectors in $\tilde{X}$ are not degenerate, it holds that

$$\liminf_{L\to\infty}\frac{\min_{h\in\{1,2,\ldots,L\}}N_{Lh}}{L^{1/2}}=\infty.$$

Condition (a) of Assumption 3.10 requires that $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ admits the asymptotic linear representation. Conditions (b), (c), and (e) are mild like the moment conditions in Assumption 3.6, though moment requirements are tightened. In the RESET approach in Example 3.2 setting $\pi^*_L=\theta^*_L$ and $\hat{\pi}_L=\hat{\theta}_L$, for instance, the uniform $\mathscr{L}_{4+4\delta}$-boundedness of $\{\tilde{Y}_{hk}:(k,h)\in\mathbb{N}^2\}$ and the uniform $\mathscr{L}_{24+24\delta}$-boundedness of $\{\tilde{Z}_{hk}:(k,h)\in\mathbb{N}^2\}$ with some positive real constant δ are sufficient for conditions (a)–(c) and (e), where we set $\psi_h(x,\theta)=z(y-z'\theta)$, where $x=(y,z')'\in\mathbb{R}\times\mathbb{R}^{v-1}$, and $J^*_L=\int zz'\bar{P}_L(dy,dz)$. Under the same conditions, conditions (d) and (f) are also satisfied, when a typical estimator is used for $\tilde{\Sigma}_{Lh}$. The uniformly positiveness of $\{\mathrm{var}[\xi^*_{Lh}]:(h,L)\in\mathbb{H}\}$ imposed in (g) is innocuous, though it is a high-level assumption. Finally, condition (h) adds the strengthen the uniform divergence of the stratum sizes imposed in Assumption 3.8. It ensures that the bias of $\tilde{\Sigma}_{Lh}$ is asymptotically negligible in our derivation of the asymptotic normality result. It is again consistent with the common view that the stratum population size is large enough that the most characteristics of $\bar{P}_{Lh}$ can be well captured by those of the stratum population.

We are now ready to state the asymptotic normality of $\hat{\alpha}_L$ under the stratum-wise correct specification.

Theorem 3.3. *Suppose that Assumptions 3.1–3.10 hold. If (3.3) holds (i.e., when the model is stratum-wise correctly specified),*

$$L^{1/2}\hat{\alpha}_L = L^{-1/2}\sum_{h=1}^{L}\xi^*_{Lh} + \mathrm{o}_P(1) \quad \textit{and} \tag{3.13}$$

$V_L^{-1/2}L^{1/2}\hat{\alpha}_L \overset{A}{\sim} \mathrm{N}(0,1)$, *where* $V_L \equiv L^{-1}\sum_{h=1}^{L}\mathrm{var}[\xi^*_{Lh}]$, $L \in \mathbb{N}$.

To formulate a useful test of the stratum-wise correct specification, we need an estimator of V_L to standardize $\hat{\alpha}_L$. Given the definition of $\{\xi^*_{Lh}\}$ in Assumption 3.10(g), estimation of V_L requires estimation of J^*_L. An estimator of J^*_L is typically constructed based on the formula of J^*_L, which is sometimes only valid under the stratum-wise correct specification. To reflect this reality, we assume:

Assumption 3.11. *There exists a bounded sequence of constant $p \times p$ matrices $\{\bar{J}_L\}_{L\in\mathbb{N}}$ such that $\hat{J}_L - \bar{J}_L \to 0$ in probability-P.*

Under the stratum-wise correct specification, it should hold that $\bar{J}_L = J^*_L$, while $\bar{J}_L$ may not coincide with J^*_L under the alternatives. In the example discussed above, we can set $\hat{J}_L = N_L^{-1}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}Z_{Lhij}Z'_{Lhij}$ and $\bar{J}_L = \int zz'\,\bar{P}_L(dy,dz)$ (which coincides with J^*_L) for each $L \in \mathbb{N}$, to satisfy Assumption 3.11.

To obtain an estimator of V_L, we approximate ξ^*_{Lh} by

$$\begin{aligned}\hat{\xi}_{Lh} \equiv{}& L\beta_{Lh}\mathrm{tr}\Bigg(\hat{\Lambda}_L\Big(\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)' - N_{Lh}^{-2}\hat{\Sigma}_{Lh}\Big)\\ &+ \hat{G}'_{1,L}\hat{J}_L(LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\hat{\psi}_{Lhij}, \quad (h,L)\in\mathbb{H}, \quad \text{where}\end{aligned}$$

$$\hat{G}_{1L} \equiv \tilde{G}_{1L}(\cdot, \theta, \pi),$$

$$\begin{aligned}\tilde{G}_{1L}(\omega, \theta, \pi) &\equiv \nabla_\theta \check{\alpha}_L(\omega, \theta, \pi_L) \\ &= \sum_{h=1}^{L} \beta_{Lh} \Bigg(2N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij}(\omega) \nabla_\theta m_h(X_{Lhij}(\omega), \theta, \pi) \Lambda \tilde{m}_{Lh}(\omega, \theta, \pi) \\ &\qquad - N_{Lh}^{-2} \nabla_\theta (\mathrm{vec} \tilde{\Sigma}_{Lh}(\omega, \theta, \pi))\, \mathrm{vec} \Lambda \Bigg), \\ &\qquad\qquad (\theta, \pi) \in \Theta \times \Pi,\ \Lambda \in \mathbb{S}^q,\ L \in \mathbb{N}, \text{ and}\end{aligned}$$

$$\hat{\psi}_{Lhij} \equiv \psi_{Lh}(X_{Lhij}, \hat{\theta}_L, \hat{\pi}_L), \quad j \in \{1, \ldots, n_{hi}\}, i \in \{1, \ldots, n_h\}, (h, L) \in \mathbb{H}.$$

With this approximation, we now estimate V_L by

$$\hat{V}_L \equiv L^{-1} \sum_{h=1}^{L} \hat{\xi}_{Lh}^2, \quad L \in \mathbb{N}.$$

Our test statistic is thus

$$\mathscr{T}_L \equiv \frac{L^{1/2} \hat{\alpha}_L}{\hat{V}_L^{1/2}}, \quad L \in \mathbb{N}.$$

The large-L behavior of this statistic is described in the next theorem. Let $\bar{\xi}_{Lh}$ denote the expression obtained by replacing J_L^* with $\bar{J}_L$ in the definition of ξ_{Lh}^*.

Theorem 3.4. *(a) Under Assumptions 3.1–3.9, 3.10(b)–(f), and 3.11, the sequence*

$$\left\{ \bar{V}_L \equiv L^{-1} \sum_{h=1}^{L} \mathrm{var}[\bar{\xi}_{Lh}] + L^{-1} \sum_{h=1}^{L} (L\beta_{Lh})^2 (\bar{m}_{Lh}^{*\prime} \Lambda_L \bar{m}_{Lh}^*)^2 \right\}_{L \in \mathbb{N}}$$

is bounded, and $\{\hat{V}_L - \bar{V}_L\}_{L \in \mathbb{N}}$ converges in probability-P to zero.

(b) Suppose that Assumptions 3.1–3.11 hold. If (3.3) holds, and $\bar{J}_L = J_L^*$ for each $L \in \mathbb{N}$ (i.e., when the model is stratum-wise correctly specified), then $\mathscr{T}_L \overset{A}{\sim} \mathrm{N}(0,1)$.

(c) Suppose that Assumptions 3.1–3.9, 3.10(b)–(f), (h), and 3.11 hold. If $\{\alpha_L^*\}_{L \in \mathbb{N}}$ is uniformly positive, then for each $c \in \mathbb{R}$, $P[\mathscr{T}_L > c] \to 1$.

Theorem 3.4 shows that we can perform a level-p test of the stratum-wise correct specification by rejecting the null hypothesis when $\mathscr{T}_L$ is greater than the $(1-p)$-quantile of the standard normal distribution. It also shows that the test has a power approaching 1, if $\{\alpha_L^*\}$ is uniformly positive. When $\{\Lambda_L\}_{L\in\mathbb{N}}$ is uniformly positive definite, the imposed uniform positivity of $\{\alpha_L^*\}$ can be interpreted as requiring that the average of the squared length of $\bar{m}_{Lh}^*$ taken over all strata in the population does not shrink, as the second result in the next proposition states. The uniform positive definiteness of $\{\Lambda_L\}$ is a mild requirement.

Proposition 3.5. *Suppose that Assumptions 3.1–3.5 and 3.6(d) hold.*

(a) If $\{\alpha_L^*\}_{L\in\mathbb{N}}$ is uniformly positive, so is $\{\sum_{h=1}^L \beta_{Lh}\,|\bar{m}_{Lh}^*|^2\}_{L\in\mathbb{N}}$.

(b) If, in addition, $\{\Lambda_L\}_{L\in\mathbb{N}}$ is uniformly positive definite, the converse of (a) holds.

3.4 Tests without Estimation of Nuisance Parameters

The specification test of Bierens (1990) is an m-test based the moment condition described in Example 3.2. In practice, it is not clear what value we should choose for the nuisance parameter π. A way to overcome this difficulty is to calculate $\mathscr{T}_L$ for each of the possible values of the nuisance parameter and summarize the results in the form of a scalar statistic. For each $\pi\in\Pi$, let $\mathscr{T}_L(\pi)$ denote statistic $\mathscr{T}_L$ obtained by taking π for $\hat{\pi}_L$ and $\hat{\Lambda}_L(\pi)$ for $\hat{\Lambda}_L$, where $\hat{\Lambda}_L(\pi)$ is a weight matrix that is possibly dependent on π. In this section, we consider tests based on statistics that can be written as functions φ of the random function $\pi\mapsto\mathscr{T}_L(\pi)$ on Π such as $\sup_{\pi\in\Pi}\mathscr{T}_L(\pi)$.

We continue imposing the conditions employed in Section 3.3 that are not directly related to π_L^*, though we now require $\Pi=\Pi_0$. That is, we impose Assumptions 3.1–3.5, 3.7, 3.8, and 3.11 with no changes, while we modify Assumptions 3.6, 3.9, and 3.10 to accommodate the current approach. The modified conditions of Assumption 3.6 and 3.9 are:

Assumption 3.12. *(a) Assumptions 3.6(a) and (c) hold, with $\Pi_0=\Pi$ nonempty and compact.*

(b) *The sequence* $\{\Lambda_L\}_{L\in\mathbb{N}}$ *of uniformly equicontinuous functions from* Π *to* $\mathbb{S}^q$ *satisfies that* $\sup\{|\Lambda_L(\pi)| : \pi \in \Pi, L \in \mathbb{N}\} < \infty$.

(c) *The sequence* $\{\hat{\Lambda}_L : \Omega \times \Pi \to \mathbb{S}^q\}_{L\in\mathbb{N}}$ *of functions measurable-*$\mathscr{F} \otimes \mathscr{B}(\Pi)/\mathscr{B}(\mathbb{S}^q)$ *satisfies that for each* $\omega \in \Omega$*,* $\hat{\Lambda}_L(\omega,\cdot) : \Pi \to \mathbb{S}^q$ *is continuous and that* $\sup_{\pi\in\Pi} |\hat{\Lambda}_L(\pi) - \Lambda_L(\pi)| \to 0$ *in probability-P.*

Under this assumption combined with Assumptions 3.1–3.4, $\{\hat{\alpha}_L(\pi) \equiv \breve{\alpha}_L(\cdot, \theta_L^*, \pi, \hat{\Lambda}_L(\pi)\}_{L\in\mathbb{N}}$ is consistent for $\{\alpha_L^*(\pi) \equiv \alpha_L(\theta_L^*, \pi, \Lambda_L(\pi))\}_{L\in\mathbb{N}}$ uniformly in $\pi \in \Pi$, corresponding to the result of Theorem 3.2 in Section 3.3.

Theorem 3.1. *Under Assumptions 3.1–3.5, 3.7, 3.8, 3.11, and 3.12,* $\{\alpha_L^*(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$ *is bounded, and* $\sup_{\pi\in\Pi} |\hat{\alpha}_L(\pi) - \alpha_L^*(\pi)| \to 0$ *in probability-P.*

We now state the modified version of Assumption 3.10. For convenience, we introduce the LBP and SLBP properties, which are slightly stronger versions of the LB and LBP properties.

Definition 5. *Given Assumption 3.1, let* Γ *be a finite-dimensional Euclidean space and* $\{\phi_h\}_{h\in\mathbb{N}}$ *a sequence of measurable functions from* $(\mathbb{R}^v \times \Gamma, \mathscr{B}^v \otimes \mathscr{B}(\Gamma))$ *to* $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$*. We say that* $\{\phi_h\}$ *is* LBP(a) on Γ*, where* $a \in [1,\infty)$*, if it is LB(a) fulfilling the conditions required in Definition 3 with h that satisfies that* $h(y)/y^s \to 0$ *as* $y \downarrow 0$ *for some real number* $s > 0$.

Definition 6. *Given Assumption 3.1 and 3.2, let* Γ *be a finite-dimensional Euclidean space and* $\{\Phi_{Lh} : (h,L) \in \mathbb{H}\}$ *an array of measurable functions from* $(\Omega \times \Gamma, \mathscr{F} \otimes \mathscr{B}(\Gamma))$ *to* $(\mathbb{R}^{l_1\times l_2}, \mathscr{B}^{l_1\times l_2})$*. We say that* $\{\Phi_{Lh}\}$ *is* SLBP(a) on Γ*, where* $a \in [1,\infty)$*, if it is SLB(a) fulfilling the conditions required in Definition 4 with h that satisfies that* $h(y)/y^s \to 0$ *as* $y \downarrow 0$ *for some real number* $s > 0$.

Remark. The P in the terms “LBP” and “SLBP” stands for the requirement that h is dominated by a power function in the neighborhood of the origin.

The SLBP property is useful in our analysis, being closely related to the LBP property in the manner parallel to the relationship between the LB and SLB proprieties.

Lemma 3.2. *The assertion of Lemma 3.1 holds even if LB and SLB are replaced with LBP and SLBP, receptively, in the lemma.*

Write

$$\bar{m}^*_{Lh}(\pi) \equiv \bar{m}_{Lh}(\theta^*_L, \pi), \quad \pi \in \Pi, i \in \mathbb{I}_g, (h,L) \in \mathbb{H},$$

$$\tilde{\Sigma}^*_{Lh}(\pi) \equiv \tilde{\Sigma}_{Lh}(\cdot, \theta^*_L, \pi), \quad \pi \in \Pi, (h,L) \in \mathbb{H},$$

$$\tilde{m}^*_{Lh}(\pi) \equiv \tilde{m}_{Lh}(\omega, \theta^*_L, \pi), \quad \pi \in \Pi, (h,L) \in \mathbb{H},$$

$$G^*_{1L}(\pi) \equiv G_{1L}(\theta^*_L, \pi, \Lambda_L(\pi)), \quad \pi \in \Pi, L \in \mathbb{N} \text{ and}$$

$$\xi^*_{Lh}(\pi) = L\beta_{Lh}\text{tr}\left(\Lambda_L\left(\tilde{m}^*_{Lh}(\pi)\tilde{m}^*_{Lh}(\pi)' - N_{Lh}^{-2}\tilde{\Sigma}^*_{Lh}(\pi)\right)\right)$$
$$+ G^*_{1,L}(\pi)' J^*_L(LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\psi^*_{Lhij}, \quad (h,L) \in \mathbb{H},$$

The modified assumption is:

Assumption 3.13. *(a) Assumptions 3.10(a), (c)–(f), and (h) hold, with LB and SLB replaced by LBP and SLBP, respectively.*

(b) The array $\{\text{var}[\xi^*_{Lh}(\pi)] : \pi \in \Pi, (h,L) \in \mathbb{H}\}$ is uniformly positive.

Imposing Assumption 3.13 along with Assumptions 3.1–3.5, 3.7, 3.8, 3.11, and 3.12, we derive the large-L distribution of $\{\pi \mapsto L^{1/2}\hat{\alpha}_L(\pi)\}_{L\in\mathbb{N}}$, a sequence of random functions from Π to $\mathbb{R}$, under the null. In so doing, we employ a functional central limit theorem derived from (Pollard, 1990, Theorem 10.6), which requires an additional assumption. Let

$$V_L(\pi) \equiv L^{-1}\sum_{h=1}^{L}\text{var}[\xi^*_{Lh}(\pi)], \quad \pi \in \Pi, L \in \mathbb{N}.$$

The additional assumption is:

Assumption 3.14. *There exists a function $K : \Pi^2 \to \mathbb{R}$ such that for each $(\pi_1, \pi_2) \in \Pi^2$,*

$$K_L(\pi_1,\pi_2) \equiv L^{-1}\sum_{h=1}^{L}\text{cov}[V_L(\pi_1)^{-1/2}\xi^*_{Lh}(\pi_1), V_L(\pi_2)^{-1/2}\xi^*_{Lh}(\pi_2)] \to K(\pi_1,\pi_2).$$

Note that $K_L(\pi_1,\pi_2)$ is the coefficient of correlation between $L^{-1}\sum_{h=1}^{L}\xi_{Lh}^{*}(\pi_1)$ and $L^{-1}\sum_{h=1}^{L}\xi_{Lh}^{*}(\pi_2)$. Assumption 3.14 requires that the correlation coefficient converges as L grows large.

We now state the asymptotic Gaussianity of $\{\pi\mapsto L^{1/2}\hat{\alpha}_L(\pi)\}_{L\in\mathbb{N}}$ on Π under the stratum-wise correct specification.

Theorem 3.3. *Suppose that Assumptions 3.1–3.5, 3.7, 3.11, 3.12, and 3.13(a) hold. If (3.3) holds (i.e., when the model is stratum-wise correctly specified), it holds that*

$$\sup_{\pi\in\Pi}\left|L^{1/2}\hat{\alpha}_L(\pi)-L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^{*}(\pi)\right|\to 0 \textit{ in probability-}P. \tag{3.14}$$

If, in addition, Assumptions 3.13(b), and 3.14 hold, the sequence of random functions from Π to $\mathbb{R}$

$$\{\pi\mapsto V_L(\pi)^{-1/2}L^{1/2}\hat{\alpha}_L(\pi)\}_{L\in\mathbb{N}} \tag{3.15}$$

converges in distribution to the zero-mean Gaussian process with covariance kernel K concentrated on $\mathsf{U}(\Pi)$, the set of all uniformly continuous $\mathbb{R}$-valued functions on Π.

Write

$$\hat{\xi}_{Lh}(\pi)\equiv L\beta_{Lh}\mathrm{tr}\Bigg(\hat{\Lambda}_L(\pi)\Big(\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\pi)\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\pi)'-N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\hat{\theta}_L,\pi)\Big)$$
$$+\hat{G}_{1,L}(\pi)'\hat{J}_L(LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\hat{\psi}_{Lhij},\quad (h,L)\in\mathbb{H},\quad \text{where}$$

$$\hat{G}_{1L}(\pi)\equiv\nabla_{\theta}\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi),\quad L\in\mathbb{N}.$$

We then estimate $V_L(\pi)$ by

$$\hat{V}_L(\pi)\equiv L^{-1}\sum_{h=1}^{L}\hat{\xi}_{Lh}^{2}(\pi),\quad \pi\in\Pi, L\in\mathbb{N}.$$

Our test is based on the random function

$$\pi\mapsto\mathscr{T}_L(\pi)\equiv\frac{L^{1/2}\hat{\alpha}_L(\pi)}{\hat{V}_L(\pi)^{1/2}},\quad L\in\mathbb{N}.$$

The large-L behavior of $\{\pi \mapsto \hat{V}_L(\pi)\}_{L\in\mathbb{N}}$ and $\{\pi \mapsto \mathscr{T}_L(\pi)\}_{L\in\mathbb{N}}$ are described in the next theorem. Let $\bar{\xi}_{Lh}(\pi)$ denote the expression obtained by replacing J_L^* with $\bar{J}_L$ in the definition of $\xi_{Lh}^*(\pi)$.

Theorem 3.4. *(a) Under Assumptions 3.1–3.3, 3.11, 3.12, and 3.13(a), the set*

$$\Big\{\bar{V}_L(\pi) \equiv L^{-1}\sum_{h=1}^{L} \operatorname{var}[(\bar{\xi}_{Lh}(\pi)] + L^{-1}\sum_{h=1}^{L}(L\beta_{Lh})^2(\bar{m}_{Lh}^*(\pi)'\Lambda_L(\pi)\bar{m}_{Lh}^*(\pi))^2$$

$$\pi \in \Pi, L \in \mathbb{N}\Big\}$$

is bounded, $\{\bar{V}_L : \Pi \to \mathbb{R}\}_{L\in\mathbb{N}}$ is uniformly equicontinuous, and $\{\hat{V}_L(\pi) - \bar{V}_L(\pi)\}_{L\in\mathbb{N}}$ converges in probability-P to zero uniformly in $\pi \in \Pi$.

(b) Suppose that Assumptions 3.1–3.5, 3.7, and 3.11–3.14 hold. If (3.3) holds, and $\bar{J}_L = J_L^*$ for each $L \in \mathbb{N}$ (i.e., when the model is stratum-wise correctly specified), then the sequence of random functions $\{\pi \mapsto \mathscr{T}_L(\pi)\}_{L\in\mathbb{N}}$ converges in distribution to the zero-mean Gaussian process with covariance kernel K concentrated on $\mathsf{U}(\Pi)$.

We now apply a real-valued map φ to the random function $\pi \mapsto \mathscr{T}_L(\pi)$, to obtain a scalar statistic suitable for use in testing our null hypothesis. Given any possible realization of the data, the set $\{\pi \in \Pi : \hat{V}_L(\pi) = 0\}$ is Borel-measurable for each $L \in \mathbb{N}$, because $\hat{V}_L(\pi)$ is continuous in π. Also, $\hat{\alpha}_L(\pi)$ is continuous in π. It follows that $\pi \mapsto \mathscr{T}_L(\pi)$ is a Borel-measurable functional on Π for each possible realization of the data. We thus pick a map φ defined on $\mathscr{M}(\Pi)$, the set of all Borel-measurable functions from Π to $\mathbb{R}$. The map φ given by

$$\varphi(f) \equiv \sup_{\pi\in\Pi} f(\pi), \quad f \in \mathscr{M}(\Pi) \tag{3.16}$$

is such a map. The one written as

$$\varphi(f) \equiv \int_{\Pi} \max\{f(\pi), 0\}\, d\pi, \quad f \in \mathscr{M}(\Pi), \tag{3.17}$$

where the integral is taken with respect to the Lebesgue measure, is another.

In (3.17), the effect of negative values of $f(\pi)$ is suppressed before the integral is taken. To appreciate benefits of this mechanism, recall that $\alpha_L^*(\pi)$ is known to be nonnegative for every $\pi \in \Pi$. A negatively large realized value of $\mathscr{T}_L(\pi)$ can be viewed as a reflection of an error in estimation of $\alpha_L^*(\pi)$. By suppressing the negative part of the integrand in (3.17), we prevent such estimation errors for some π's from canceling out the effect of positive values of $\mathscr{T}_L(\pi)$ for other π's.

Like most other maps discussed in similar contexts in the literature, the maps φ in (3.16) and (3.17) satisfy the conditions imposed in the next assumption.

Assumption 3.15. *The function φ from $\mathscr{M}(\Pi)$ endowed with the uniform metric to the one-dimensional Euclidean space is continuous. It is also monotonic in the sense that whenever a pair f_1 and f_2 in $\mathscr{M}(\Pi)$ satisfies that $f_1 \geq f_2$, it holds that $\varphi(f_1) \geq \varphi(f_2)$. Further, it satisfies that whenever a sequence $\{f_j \in \mathscr{M}(\Pi)\}_{j\in\mathbb{N}}$ satisfies that for each π in some Borel-measurable subset $\bar{\Pi}$ of Π with a nonzero Lebesgue measure, $\lim_{j\to\infty} f_j(\pi) = \infty$, it holds that $\lim_{j\to\infty} \varphi(f_j) = \infty$.*

Given such a map φ, our test should reject the null hypothesis if $\varphi(\mathscr{T}_L)$ exceeds a suitably chosen critical value.

To pick the critical value in the test, we use the result of Theorem 3.4(ii). Let η be a zero-mean Gaussian process with kernel K concentrated on $\mathsf{U}(\Pi)$. Then it follows from the theorem by the continuous mapping theorem (van der Vaart and Wellner, 1996, Lemma 1.3.6) that $\{\varphi(\mathscr{T}_L)\}_{L\in\mathbb{N}}$ converges in distribution to $\varphi(\eta)$ under the null. To utilize this fact in formulation of a test, we need to estimate K, on which the distribution of $\varphi(\eta)$ depends. A natural estimator of K is $\{\hat{K}_L\}_{L\in\mathbb{N}}$ defined by

$$\hat{K}_L(\pi_1,\pi_2) \equiv (\hat{V}_L(\pi_1)\hat{V}_L(\pi_2))^{-1/2}L^{-1}\sum_{h=1}^{L}\hat{\xi}_{Lh}(\pi_1)\hat{\xi}_{Lh}(\pi_2), \quad (\pi_1,\pi_2) \in \Pi_0^2, L \in \mathbb{N}.$$

Under the current assumptions, $\hat{K}_L(\pi_1,\pi_2)$ is consistent for $K(\pi_1,\pi_2)$ uniformly in $(\pi_1,\pi_2) \in \Pi^2$ if $\{\alpha_L^*(\pi)\}_{L\in\mathbb{N}}$ converges to zero uniformly in $\pi \in \Pi$, in particular, under the null hypothesis.

Having a consistent estimator $\hat{K}_L$ of K, it is easy to generate a zero-mean Gaussian process with covariance kernel $\hat{K}_L$, as Hansen (1996) explains. Let $\{\nu_h\}_{h\in\mathbb{N}}$ be an i.i.d. sequence of standard normal random variables independent from the data. It then holds that for each $L \in \mathbb{N}$,

$$\hat{\mathscr{T}}_L(\pi) \equiv \hat{V}_L(\pi)^{-1/2} L^{-1/2} \sum_{h=1}^{L} \hat{\xi}_{Lh}(\pi) \nu_h$$

is a zero-mean Gaussian process with covariance kernel $\hat{K}_L$ conditionally given the data. When for each $\pi \in \Pi$ and each $L \in \mathbb{N}$, $\alpha_L^*(\pi) = 0$, the conditional distribution of $\pi \mapsto \hat{\mathscr{T}}_L(\pi)$ weakly converges to a Gaussian process with covariance kernel K concentrated on $\mathsf{U}(\Pi)$ in probability-P, as one might expect from the uniform consistency of $\{\hat{K}_L\}$ for K. see Gine and Zinn (1990) for the concept of weak convergence in probability). It follows that we can take for the critical value in our test the $(1-p)$-quantile of the distribution of $\varphi(\mathscr{T}_L)$ conditional on the data. In addition, we can show that $\sup_{\pi \in \Pi} |\hat{\mathscr{T}}_L(\pi)| = \mathrm{O}_P(1)$, regardless of whether or not the null hypothesis is true. This means that the critical value would stay bounded as $L \to \infty$, even under alternatives.

In practice, the test described above can be conveniently implemented by using the p-value transformation explained in Hansen (1996). The next theorem provides information essential for the test in such form.

Theorem 3.5. *(a) Suppose that Assumptions 3.1–3.5, 3.7, and 3.11–3.15 hold. Also suppose that (3.3) holds, and $\bar{J}_L = J_L^*$ for each $L \in \mathbb{N}$ (these conditions hold if the model is stratum-wise correctly specified). Let $\hat{F}_L$ be the conditional distribution function of $\hat{\varphi}_L \equiv \varphi(\hat{\mathscr{T}}_L)$ given the data. Also, let F denote the distribution function of $\varphi_0 \equiv \varphi(\eta)$, where η is a zero-mean Gaussian process with covariance kernel K concentrated on $\mathsf{U}(\Pi)$. Write $\varphi_L \equiv \varphi(\mathscr{T}_L)$, $L \in \mathbb{N}$. Then $\{\hat{p}_L \equiv 1 - \hat{F}_L(\varphi_L)\}_{L \in \mathbb{N}}$ and $\{p_L \equiv 1 - F(\varphi_L)\}_{L \in \mathbb{N}}$ satisfy that $\hat{p}_L - p_L \to 0$ in probability-P. If, in addition, F is continuous and increasing on the support of φ_0, $\{\hat{p}_L\}_{L \in \mathbb{N}}$ is asymptotically distributed with the uniform distribution on $[0,1]$.*

(b) Suppose that Assumptions 3.1–3.5, 3.7, 3.11, 3.12, and 3.13(a) hold. If there exists a Borel-measurable subset $\bar{\Pi}$ of Π with a nonzero Lebesgue measure such that $\liminf_{L \to \infty} \inf_{\pi \in \bar{\Pi}} \alpha_L^*(\pi) > 0$, then $\hat{p}_L \to 0$ in probability-P.

Our test rejects the null hypothesis, if and only if $\hat{p}_L$ is lower than the specified level. Theorem 3.5 confirms that the test has asymptotically correct size under the null and a power approaching to one if for some $\bar{\Pi} \subset \Pi$ with a non-zero Lebesgue measure, $\inf_{\pi \in \bar{\Pi}} \alpha_L^*(\pi)$ is bounded away from zero for almost all $L \in \mathbb{N}$.

Bibliography

Anatolyev, Stanislav. 2008 (Sept.). *Inference in Regression Models with Many Regressors*. Working Paper 125. Center for Economic and Financial Research, New Economic School. → pages 61

Belzil, Christian, and Hansen, Jorgen. 2002. Unobserved ability and the return to schooling. *Econometrica*, **70**, 2075–2091. → pages 21, 38

Belzil, Christian, and Hansen, Jorgen. 2007. A structural analysis of correlated random coefficent wage regression model. *Journal of Econometrics*, **140**, 827–848. → pages 21, 38

Bierens, Herman J. 1990. A Consistent Conditional Moment Test of Functional Form. *Econometrica*, **58**(6), 1443–1458. → pages 68, 79

Binder, David A. 1983. On the Variance of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, **51**(3), 279–292. → pages 45

Blaug, Mark. 1985. Where are we now in the economics of education. *Economics of Education Review*, **4**(1), 17–28. → pages 1

Breckling, J. U., *et al.* . 1994. Maximum Likelihood Inference from Sample Survey Data. *International Statistical Review*, **62**(3), 349–363. → pages 45

Card, David. 2001. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, **69**(5), 1127–1160. → pages 1, 2

Carneiro, Pedro, Hansen, Karsten T., and Heckman, James J. 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review*, **44**, 361–422. → pages 2, 3, 4, 38

Carrasco, Marine, and Florens, Jean-Pierre. 2000. Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory*, **16**(6), 797–834. → pages 61

Carrasco, Marine, Chernov, Mikhail, and Florens, Jean-Pierre Ghysels, Eric. 2007. Efficient Estimation of General Dynamic Models with a Continuum of Moment Conditions. *Journal of Econometrics*, **140**(2), 529–573. → pages 61

Cassel, C. M., Sarndal, Carl-Erik, , and Wretman, Jan. 1977. *Foundations of inference in survey sampling*. New York: Wiley. → pages 45

Chamblessa, Lloyd E., and Boyle, Kerrie E. 1985. Maximum Likelihood Methods for Complex Sample Data: Logistic Regression And Discrete Proportional Hazards Models. *Communications in Statistics-Theory and Methods*, **14**(6), 1377–1392. → pages 45

Chen, Jiahua, and Rao, J. N. K. 2007. Asymptotic normality under twophase sampling designs. *Statist. Sinica*, 1047–1064. → pages 46

Cunha, Flavio, and Heckman, James J. 2008. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, **43**(4), 738–782. → pages 3

Cunha, Flavio, Heckman, James J., and Schennach, Susanne M. 2010. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, **78**(3), 883–931. → pages 3

Davidson, J. 1994. *Stochastic Limit Theory-An Introduction for Econometricians, Advanced Textbooks in Econometrics*. → pages 150, 152

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.Series B (Methodological)*, **39**(1), 746–773. → pages 22

Dickens, William T., and Lang, Kevin. 1985. A test of dual labor market theory. *The American Economic Review*, **75**(4), 792–805. → pages 1

Donald, Stephen G., Imbens, Guido W., and Newey, Whitney K. 2003. Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions. *Journal of Econometrics*, **117**(1), 55–93. → pages 61

Doran, Howard E., and Schmidt, Peter. 2006. GMM Estimators with Improved Finite Sample Properties Using Principal Components of the Weighting Matrix, with an Application to the Dynamic Panel Data Model. *Journal of Econometrics*, **133**(1), 387–409. → pages 61

Duncan, G., and Homan, S. 1981. The incidence and wage effects of overeducation. *Economics of Education Review*, **1**(1), 75–86. → pages 1

Folland, G. 1984. *Real analysis: Modern techniques and their applications*. → pages 111, 112, 122, 123

Fuller, Wayne A. 1984. Least Squares and Related Analyses for Complex Survey Designs. *Survey Methodology*, **10**(1), 97–118. → pages 45

Gallant, A. Ronald, and White, Halbert. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell. → pages 50, 112

Gine, Evarist, and Zinn, Joel. 1990. Bootstrapping General Empirical Measures. *The Annals of Probability*, **18**(2), 851–869. → pages 85

Grubb, W. Norton. 1997. The Returns to Education in the Sub-Baccalaureate Labor Market, 1984-1990. *Economics of Education Review*, **16**(3), 231–245. → pages 3

Han, Chirok, and Phillips, Peter C. B. 2006. GMM with Many Moment Conditions. *Econometrica*, **74**(1), 147–192. → pages 61

Hansen, Bruce E. 1996. Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis. *Econometrica*, **64**(2), 413–430. → pages 84, 85

Hansen, Karsten, Heckman, James J., and Mullen, Kathleen J. 2004. The Effects of Schooling and ability on Achievement test scores. *Journal of Econometrics*, **121**(1-2), 39–98. → pages 3, 15

Heckman, James J., and Singer, Burton. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320. → pages 22

Heckman, James J., Lochner, Lance J., and Todd, Petra E. 2006a. Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. **I**, 307–458. → pages 1

Heckman, James J., Stixrud, Jora, and Urzua, Sergio. 2006b. The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labour Economics*, **24**(3), 746–773. → pages 3, 15, 21

Hoffmann, Florian. 2011. An empirical model of life-cycle earnings and mobility dynamics. working paper. → pages 5

Horvitz, D. G., and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**(260), 663–685. → pages 46

Hung, Hsien-Ming. 1990. Nonlinear Regression Analysis for Complex Surveys. *Communications in Statistics-Theory and Methods*, **19**(9), 3447–3468. → pages 45

Jaeger, David A., and Page, Marianne E. 1996. Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education. *The Review of Economics and Statistics*, **78**(4), 733–740. → pages 5, 33

Kane, Thomas J., and Rouse, Cecilia Elena. 1995. Labor Market Returns to Two- and Four-Year College. *American Economic Review*, **85**(3), 600–614. → pages 3

Kasahara, Hiroyuki, and Shimotsu, Katsumi. 2009. Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, **77**, 135–175. → pages 2, 18

Kasahara, Hiroyuki, and Shimotsu, Katsumi. 2012. Nonparametric identification of multivariate mixtures. Discussion papers 2010-09, Graduate School of Economics, Hitotsubashi University. → pages 2, 18

Keane, Michael P., and Wolpin, Kenneth I. 1997. The career decisions of young men. *Journal of Political Economy*, **105**, 473–522. → pages 4, 20, 21, 26

Koenker, Roger, and Machado, José A. F. 1999. GMM Inference When the Number of Moment Conditions is Large. *Journal of Econometrics*, **93**(2), 327–344. → pages 61

Koenker, Roger W., and Bassett, Jr., Gilbert W. 1978. Regression Quantiles. *Econometrica*, **46**(1), 33–50. → pages 54

Krewski, D., and Rao, J. N. K. 1981. Inference from Stratified Samples: Properties of the Linearization, Jacknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, **9**(5), 1010–1019. → pages 45

Light, Audrey, and Strayer, Wayne. 2004. Who Receives the College Wage Premium? Assessing the Labor Market Returns to Degrees and College Transfer Patterns. *The Journal of Human Resources*, **39**(3), 411–482. → pages 3

Marcotte, Dave E., Bailey, Thomas, Borkoski, Carey, and Kienzl, Greg S. 2005. The Returns of a Community College Education: Evidence From the National Education Longitudinal Survey. *Educational Evaluation and Policy Analysis*, **27**(2), 157–175. → pages 3

Newey, Whitney K. 1985. Maximum Likelihood Specification Testing and Conditional Moment Tests. *Econometrica*, **53**(5), 1047–1070. → pages 59

Neyman, Jerzy. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**(4), 558–625. → pages 45

Pollard, David. 1990. *Empirical Processes: Theory and Applications*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 2. Hayward, California: Institute of Mathematical Statistics. → pages 81, 136

Ramsey, J. B. 1969. Tests for Specification Errors in Classical Linear Least-Squares Analysis. *Journal of the Royal Statistical Society, Series B*, **71**, 351–371. → pages 69

Roy, A. D. 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, **3**(2), 135–146. → pages 2

Sakata, Shinichi. 2000. *Quasi-Maximum Likelihood Estimation with Complex Survey Data*. Mimeo., University of Michigan. → pages 45

Sakata, Shinichi. 2009 (Sept.). *m-Testing of Model Specification in Many Groups*. Mimeo., University of British Columbia. → pages 59, 61

Sakata, Shinichi, and Xu, Jinwen. 2010 (Oct.). *M-Estimation with Complex Survey Data*. Mimeo., University of British Columbia. → pages 59

Sarndal, Carl-Erik, Thomsen, Ib, Hoem, Jan M., Lindley, D. V., Barndorff-Nielsen, O., and Dalenius, Tore. 1978. Design-Based and Model-Based Inference in Survey Sampling. *Scandinavian Journal of Statistics*, **5**(1), 27–52. → pages 45

Sarndal, Carl-Erik, Swensson, Bengt, and Wretman, Jan. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag. → pages 45

Sen, Pranab Kumar. 1970. On Some Convergence Properties of One-Sample Rank Order Statistics. *Annals of Mathematical Statistics*, **41**(6), 2140–2143. → pages 67

Sicherman, N. 1991. ”Overeducation” in the labor market. *Journal of Labor Economics*, **9**(2), 101–122. → pages 1

Tauchen, G. 1985. Diagnostic Testing and Evaluation of Maximum Likelihood Models. *Journal of Econometrics*, **30**(1/2), 415–444. → pages 59 van der Vaart, Aad W., and Wellner, Jon A. 1996. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York: Springer. → pages 84, 148, 151

White, H. 1984. *Asymptotic theory for econometricians*. → pages 117

White, Halbert. 1987. Specification Testing in Dynamic Models. *Chap. 1, pages 1–58 of:* Truman, F. Bewley (ed), *Advances in Econometrics—Fifth Wold Congress*. Econometric Society Monographs, vol. 1. New York: Cambridge University Press. → pages 59

White, Halbert. 1994. *Estimation, Inference and Specification Analysis*. → pages 117, 118

Willis, Robert J. 1986. Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions. **I**, 525–602. → pages 2

Willis, Robert J., and Rosen, Sherwin. 1979. Education and self-selection. *Journal of Political Economy*, **87**, S7–S36. → pages 2, 4, 39

Wolter, Kirk M. 1985. *Introduction to Variance Estimation*. New York: Springer. → pages 55, 72

Appendix A

A.1 Summary Statistics, College Dropouts vs. College Graduates

Table A.1: Discriptive Statistics (More Education Categories)

	2-yr Dropouts			Associate Degree			4-yr Dropouts			Bachelor's Degree		
Variables	Obs	Mean	S.D	Obs	Mean	S.D	Obs	Mean	S.D	Obs	Mean	S.D
Highest grade completed	119	12.697	0.859	40	14.200	0.608	131	13.573	1.336	307	16.619	1.180
Age in 1979	119	17.277	2.004	40	17.525	2.075	131	17.55	2.146	307	17.759	2.246
Initial job(white collar)	119	0.244	0.431	40	0.275	0.452	131	0.374	0.486	307	0.752	0.432
Initial wage	119	10.19	4.579	40	12.185	5.605	131	10.669	5.18	307	14.106	13.981
Initial wage(blue collar)	119	10.524	4.926	40	12.274	6.006	131	10.099	4.513	307	9.935	4.603
Initial wage(white collar)	119	9.155	3.128	40	11.95	4.629	131	11.624	6.067	307	15.478	15.669
Mother education	119	12.126	1.964	40	12.4	1.751	131	12.328	1.854	307	13.166	2.096
Father education	119	12.689	2.626	40	12.9	2.479	131	13.038	3.134	307	14.055	3.072
Number of siblings	119	2.798	1.754	40	2.45	1.694	131	2.748	1.729	307	2.43	1.596
Broken family at age 14	119	0.176	0.383	40	0.075	0.267	131	0.115	0.32	307	0.094	0.293
South at age 14	119	0.244	0.431	40	0.175	0.385	131	0.244	0.431	307	0.248	0.432
Urban at age 14	119	0.84	0.368	40	0.65	0.483	131	0.725	0.448	307	0.814	0.389

In table A.1, I divide two-year college attendants into those with and without an associate degree, and divide four-year attendants to those with and without a bachelor's degree. Among the two-year college attendants, 75% do not obtain an associate degree. The average schooling years of the two-year college dropouts are 12.70 years and those of the two-year college graduates are 14.2 years. Among the four-year college attendants, around 30% drop out of four-year college while the majority obtain a bachelor's degree. The average schooling years of the four-year college dropouts are 13.6 years and those of the four-year college graduates are around 16.6 years. Regarding the the first job after schooling, those who obtain an associate degree are slightly more likely to work in a white-collar occupation than those drop out of a two-year college. The probability of initially working in a white-collar position of the two-year college dropouts is 24.4% and that of individuals with an associate degree is 27.5%. Interestingly, although the four-year college dropouts have more schooling years than those with an associate degree, the former are more likely to work in a white-collar occupation entering the labour market than the latter. On average, around 37.4% of the four-year college dropouts initially work in a white-collar occupation. Those with a bachelor's degree are much more likely to start with a white-collar occupation than the others. Around 75.2% of those with a bachelor's degree work in a white-collar occupation as their first jobs. Regarding wages, the two-year college dropouts and the four-year dropouts earn almost the same. The average hourly payment for the two-year college dropouts and the four-year dropouts are $10.19 and $10.67 respectively. Those with an associate degree earn around $12.19 per hour for their first jobs. The hourly payment of those with an associate degree is higher than that of the two-year college dropouts and four-year college dropouts. Those with a bachelor's degree earn the most among the post-secondary attendants. The average hourly payment of those with a bachelor's degree is around $14.11. When we take a closer look at wages by separating individuals into two occupation groups, those initially work in a blue-collar occupation and those initially work in a white-collar occupation, the story is different. The two-year college dropouts, the four-year college dropouts, and those with a bachelor's degree earn around $10 per hour if their first job is blue-collar while those with an associate degree earn $12.27 per hour if their first job is blue-collar. For those whose first job is white-collar, the two-year college dropouts earn \$9.15 per hour. Those with an associate degree and the four-year college dropouts earn around \$12 per hour. Those with a bachelor's degree earn more than the two-year college dropouts, those with an associate degree, and the four-year college dropouts. The hourly payment of those with a bachelor's degree is \$15.48 per hour. Although this paper mainly examines the impact of attendance of a two-year college and a four-year college on wages, I also provide estimates of the occupation-specific wage gains to obtain a bachelor's degree for a high school graduate by eliminating the four-year college dropouts from the sample. I do not study the occupation-specific returns to an associate degree because the sample size of those with an associate degree is too small or reasonable results.

A.2 Simplification of the Type-Specific Joint Distribution

Below, I show how to simplify the type-specific joint distribution of wages, occupations, education, and the test scores.

$$
\begin{aligned}
& f^m(\{W_{it},O_{it}\}_{t=1}^{T},S_i,\{Q_{ir}\}_{r=1}^{6}|\{X_{it}\}_{t=1}^{T},Z_{S,i},\{Z_{ir}\}_{r=1}^{6}) \\
=& f^m(\{W_{it},O_{it}\}_{t=1}^{T},S_i|\{Q_{ir}\}_{r=1}^{6},\{X_{it}\}_{t=1}^{T},Z_{S,i},\{Z_{ir}\}_{r=1}^{6}) \\
& \times f^m(\{Q_{ir}\}_{r=1}^{6}|\{X_{it}\}_{t=1}^{T},Z_{S,i},\{Z_{ir}\}_{r=1}^{6}) \\
=& f^m(\{W_{it},O_{it}\}_{t=1}^{T},S_i|\{Q_{ir}\}_{r=1}^{6},\{X_{it}\}_{t=1}^{T},Z_{S,i})f^m(\{Q_{ir}\}_{r=1}^{6}|\{Z_{ir}\}_{r=1}^{6}) \\
=& f^m(\{W_{it},O_{it}\}_{t=1}^{T},S_i|\{Q_{ir}\}_{r=1}^{6},\{X_{it}\}_{t=1}^{T},Z_{S,i})\prod_{r=1}^{6}f^m(\{Q_{ir}\}_{r=1}^{6}|\{Z_{ir}\}_{r=1}^{6}) \\
=& f^m(\{W_{it},O_{it}\}_{t=1}^{T},S_i|\{Q_{ir}\}_{r=1}^{6},\{X_{it}\}_{t=1}^{T},Z_{S,i})\prod_{r=1}^{6}f^m(\{Q_{ir}\}_{r=1}^{6}|\{Z_{ir}\}_{r=1}^{6}) \\
=& f^m(W_{i1}|O_{i1},S_i)\prod_{t=2}^{T}f^m(W_{it}|O_{it},S_i,X_{it},W_{it-1},O_{it-1},X_{it-1}) \\
& \times f^m(O_{i1}|S_i)\prod_{t=2}^{T}f^m(O_{it}|O_{it-1},S_i,X_{it})f^m(S_i|Z_{S,i})\prod_{r=1}^{6}f^m(Q_{ir}|Z_{ir}).
\end{aligned}
$$

The first equality holds under the assumption that the six test scores do not directly affect wages, occupations, and education conditional on type. The second equality holds under the assumption that the regressors and the error terms in

Equation (1.1), Equation (1.2), Equation (1.3), and Equation (1.4) are independent. The third equality holds under the assumption that the error terms in test scores are mutually independent ($\varepsilon_{Q,ir} \perp\!\!\!\perp \varepsilon_{Q,ir'}$ for $r \neq r'$). The fourth equality holds under the assumptions that the error terms in wage follows a first order Markov process ($\varepsilon_{W,it} = \rho \varepsilon_{W,it-1} + \zeta_{it}$) and the occupation choice is only affected by the previous occupation, not the whole occupation history.

A.3 Likelihood Contributions

(a) The likelihood contribution of wages:

$$\mathscr{L}_W^m(Y_i; \alpha_W, \beta, \sigma_W, \rho) = \phi(\frac{W_{i1} - \mu_{W,i1}}{\sigma_{W,1}}) \prod_{t=2}^{T} \phi(\frac{W_{it} - \mu_{W,i2}}{\sigma_{W,2}}).$$

The wage density functions follow a normal distribution according to the assumptions in Equation 1.1. Specifically,

$$\mu_{W,i1} = \alpha_{W1}^m + \alpha_{W,2}^m O_{it} + \beta_1 2\text{YR}_i + \beta_2 4\text{YR}_i + \beta_3 2\text{YR}_i O_{it} + \beta_4 4\text{YR}_i O_{it} + X_{it}'\beta_5,$$

and

$$\begin{aligned}\mu_{W,i2} =& \alpha_{W1}^m + \alpha_{W,2}^m O_{it} + \beta_1 2\text{YR}_i + \beta_2 4\text{YR}_i + \beta_3 2\text{YR}_i O_{it} + \beta_4 4\text{YR}_i O_{it} + X_{it}'\beta_5 \\ & - \rho(W_{it-1} - (\alpha_{W1}^m + \alpha_{W,2}^m O_{it-1} + \beta_1 2\text{YR}_i + \beta_2 4\text{YR}_i + \beta_3 2\text{YR}_i O_{it-1} \\ & + \beta_4 4\text{YR}_i O_{it-1} + X_{it-1}'\beta_5))\end{aligned}$$

where $D_{O,ijt}$ is a dummy variable, which equals 1 if individual i works in occupation j at time t

(b) The likelihood contribution of occupations:

$$\begin{aligned}\mathscr{L}_O^m(Y_i; \alpha_O, \lambda) =& \Phi(\alpha_O^m + \lambda_1 2\text{YR}_i + \lambda_2 4\text{YR}_i) \\ & \times \prod_{t=2}^{T} \Phi(\alpha_O^m + \lambda_1 2\text{YR}_i + \lambda_2 4\text{YR}_i + \lambda_3 O_{it-1} + X_{it}'\lambda_4).\end{aligned}$$

(c) The likelihood contribution of education:

$$\mathscr{L}_S^m(Y_i;\alpha_S,\delta) = \frac{exp(\alpha_{S,j}^m + Z_i'\delta_j)}{1+\sum_{j'=2}^3 exp(\alpha_{S,j'}^m + Z_i'\delta_{j'})}.$$

(d) The likelihood contribution of test scores:

$$\mathscr{L}_Q^m(Y_i;\alpha_Q,\theta,\sigma_Q) = \prod_{r=1}^6 \phi(\frac{Q_{ir}-\mu_{Q,ir}}{\sigma_{Q,r}}).$$

The density functions of test scores follow a normal distribution according to the assumptions in Equation 1.4, and $\mu_{Q,ir} = \alpha_r^m + \theta_{r,1}2\text{YR}_{ir} + \theta_{r,2}4\text{YR}_{ir} + Z_{i,r}'\theta_{r,3}$.

A.4 Assumptions and Proofs of Propositions

A.4.1 Assumptions and proof of Proposition 1.1

Assumption A.1. *For m=1,...,M and $t \geq 2$,*

(a)

$$f_t^m(W_t|O_t,X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^m(W_t|O_t,X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}),$$

and

$$f_t^m(O_t|X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^m(O_t|X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}).$$

(b)

$$f^m(W_t|O_t,X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^m(W_t|O_t,X_t,S),$$

and

$$f^m(O_t|X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^m(O_t|X_t,S).$$

Assumption A.1 reduces the number of unknown type-specific distributions and the conditional type-specific joint distributions of wages, occupations, and education can be simplified as follows:

$$\begin{aligned}
&f^m(\{W_t,O_t\}_{t=1}^T,S|\{X_{it}\}_{t=1}^T,Z_S)\\
=&f^m(W_1|O_1,S)\prod_{t=2}^T f^m(W_t|O_t,S,X_t)\\
&\times f^m(O_1|S)\prod_{t=2}^T f^m(O_t|S,X_t)f^m(S|Z_S).
\end{aligned}$$

For the sake of clarity, assume the support of X_t (t=2,...,T) is discrete and known. Let $(\eta_{t,1},\eta_{t,2},\ldots,\eta_{t,M-1})$ be elements of $\mathscr{X}_t$ for t=1,...,T. Fix $S=s$ and define, for $(\eta_t,\eta_1,z_S)\in\mathscr{X}_t\times\mathscr{X}_1\times\mathscr{Z}_{\mathscr{S}}$,

$$\begin{aligned}
\lambda_{O,\eta_1}^{*m} &= P^m(O_1=1|(X_1,S)=(\eta_1,s)),\\
\lambda_{O,\eta_t}^{m} &= P^m(O_t=1|(X_t,S)=(\eta_t,s)),\\
\widetilde{\pi}_{z_S}^m &= \pi^m P^m(S=s|Z_S=z_S)
\end{aligned}$$

Construct a matrix of type-specific distribution functions and type probabilities as

$$L_t=\begin{pmatrix} 1 & \lambda_{O,\eta_{t,1}}^1 & \cdots & \lambda_{O,\eta_{t,M-1}}^1\\ \ldots & \cdots & \ddots & \cdots\\ 1 & \lambda_{O,\eta_{t,1}}^M & \cdots & \lambda_{O,\eta_{t,M-1}}^M \end{pmatrix},\ for\, t=2,\ldots,T$$

$$D_{\eta_1}^O=diag(\lambda_{O,\eta_1}^{*1},\ldots,\lambda_{O,\eta_1}^{*M}),\ and\ V_{z_S}=diag(\widetilde{\pi}_{z_S}^1,\ldots,\widetilde{\pi}_{z_S}^M).$$

The elements of L_t,$D_{\eta_1}^O$, and V_{z_S} are parameters of the underlying mixture model to be identified. Consider we have data for three time periods i.e. $T=3$. Fix $O_t=1$ for all t and define

$$F_{Z_S,X_1,X_2,X_3}^{O*}=\sum_{m=1}^M\widetilde{\pi}_{Z_S}^m\lambda_{O,X_1}^{*m}\lambda_{O,X_2}^m\lambda_{O,X_3}^m$$

.

Now fix $O_2=O_3=1$ and define

$$F_{Z_S,X_2,X_3}^O=\sum_{m=1}^M\widetilde{\pi}_{Z_S}^m\lambda_{O,X_2}^m\lambda_{O,X_3}^m.$$

Similarly, define the following functions

$$
\begin{aligned}
F^{O*}_{Z_S,X_1,X_2} &= \sum_{m=1}^{M} \widetilde{\pi}^m_{Z_S} \lambda^{*m}_{O,X_1} \lambda^{m}_{O,X_2}, \\
F^{O*}_{Z_S,X_1,X_3} &= \sum_{m=1}^{M} \widetilde{\pi}^m_{Z_S} \lambda^{*m}_{O,X_1} \lambda^{m}_{O,X_3}, \\
F^{O*}_{Z_S,X_1} &= \sum_{m=1}^{M} \widetilde{\pi}^m_{Z_S} \lambda^{*m}_{O,X_1}, \\
F^{O}_{Z_S,X_2} &= \sum_{m=1}^{M} \widetilde{\pi}^m_{Z_S} \lambda^{m}_{O,X_2}, \\
F^{O}_{Z_S,X_3} &= \sum_{m=1}^{M} \widetilde{\pi}^m_{Z_S} \lambda^{m}_{O,X_3}.
\end{aligned}
$$

Arrange these into two $M \times M$ matrices:

$$
P^O_{z_S} = \begin{pmatrix}
1 & F^O_{z_S,\eta_{3,1}} & \cdots & F^O_{z_S,\eta_{3,M-1}} \\
F^O_{z_S,\eta_{2,1}} & F^O_{z_S,\eta_{2,1},\eta_{3,1}} & \cdots & F^O_{z_S,\eta_{2,1},\eta_{3,M-1}} \\
\vdots & \vdots & \ddots & \vdots \\
F^O_{z_S,\eta_{2,M-1}} & F^O_{z_S,\eta_{2,M-1},\eta_{3,1}} & \cdots & F^O_{z_S,\eta_{2,M-1},\eta_{3,M-1}}
\end{pmatrix},
$$

and

$$
P^{O*}_{z_S,\eta_1} = \begin{pmatrix}
F^{O*}_{z_S,\eta_1} & F^{O*}_{z_S,\eta_1,\eta_{3,1}} & \cdots & F^{O*}_{z_S,\eta_1,\eta_{3,M-1}} \\
F^{O*}_{z_S,\eta_1,\eta_{2,1}} & F^{O*}_{z_S,\eta_1,\eta_{2,1},\eta_{3,1}} & \cdots & F^{O*}_{z_S,\eta_1,\eta_{2,1},\eta_{3,M-1}} \\
\vdots & \vdots & \ddots & \vdots \\
F^{O*}_{z_S,\eta_1,\eta_{2,M-1}} & F^{O*}_{z_S,\eta_1,\eta_{2,M-1},\eta_{3,1}} & \cdots & F^{O*}_{z_S,\eta_1,\eta_{2,M-1},\eta_{3,M-1}}
\end{pmatrix}.
$$

To achieve identification, further assume:

Assumption A.2. *There exist some $\{\eta_{t,1},\ldots,\eta_{t,M-1}\}_{t=2}^{T}$ such that $P^O_{z_S}$ is of full rank and that all the eigenvalues of $(P^O_{z_S})^{-1}P^{O*}_{z_S,\eta_1}$ take distinct values.*

Proof of Proposition 1.1. $P^O_{z_S}$ and $P^{O*}_{z_S,\eta_1}$ can be expressed as the follows:

$$
P^O_{z_S} = L_2' V_{z_S} L_3, \text{ and } P^{O*}_{z_S,\eta_1} = L_2' V_{z_S} D^O_{\eta_1} L_3.
$$

Because $P^O_{z_S}$ is full rank, it follows that L_2 and L_3 are full rank. We can construct a matrix $A_{z_S} = (P^O_{z_S})^{-1} P^{O*}_{z_S,\eta_1} = L_3^{-1} D^O_{\eta_1} L_3$. Because $A_{z_S} L_3^{-1} = L_3^{-1} D^O_{\eta_1}$ and the eigenvalues of A_{z_S} are distinct, the eigenvalues of A_{z_S} determines the elements of $D^O_{\eta_1}$.

Moreover, the right eigenvectors of A_{z_S} are the columns of L_3^{-1} up to multiplicative constants. Denote $L_3^{-1}K$ to be the right eigenvectors of A_{z_S} where K is some diagonal matrix. Now we can determine $V_{z_S}K$ from the first row of $P^O_{z_S} L_3^{-1} K$ because $P^O_{z_S} L_3^{-1} K = L_2' V_{z_S} K$ and the first row of L_2' is a vector of ones. Then L_2' is determined uniquely by $L_2' = (P^O_{z_S} L_3^{-1} K)(V_{z_S} K)^{-1}$. Similarly, by construct a matrix $B_{z_S} = (P^{O\prime}_{z_S})^{-1} (P^{O*\prime}_{z_S,\eta_1})$, we can uniquely determine L_3'.

We can determine V_{z_S} from the first row of $P^O_{z_S} L_3^{-1} K$ because $P^O_{z_S} L_3^{-1} K = L_2' V_{z_S} K$ and the first row of L_2' is a vector of ones. Till now we have identified $\{\widetilde{\pi}^m_{z_S}\}_{m=1}^M$, $\{\lambda^m_{O,\eta_1}\}_{m=1}^M$ and $\{\{\lambda^m_{O,\eta_{t,j}}\}_{j=1}^{M-1}\}_{m=1}^M$ for $t = 2,3$.

Next I show how to identify $D^O_{x_1}$ for any $x_1 \in \mathscr{X}_1$. Let's construct $P^{O*}_{z_S,x1}$ in the same way as $P^{O*}_{z_S,\eta_1}$. It follows that $D^O_{x_1} = (L_2' V_{x_1})^{-1} P^*_{O,x_1} L_3^{-1}$. So $\{\lambda^{*m}_{O,x_1}\}_{m=1}^M$ for any $x_1 \in \mathscr{X}_1$ is identified.

To identify $\{\lambda^m_{O,x_2}\}_{m=1}^M$ for any $x_2 \in \mathscr{X}_2$, construct the following matrices:

$$L^{x_2} = \begin{pmatrix} 1 & \lambda^1_{O,x_2} \\ \vdots & \vdots \\ 1 & \lambda^M_{O,x_2} \end{pmatrix},$$

and

$$P^{x_2} = \begin{pmatrix} 1 & F^O_{z,\eta_{3,1}} & \cdots & F^O_{z,\eta_{3,M-1}} \\ F^O_{z,x_2} & F^O_{z,x_2,\eta_{3,1}} & \cdots & F^O_{z,x_2,\eta_{3,M-1}} \end{pmatrix}.$$

P^{x_2} can be expressed as $P^{x_2} = (L^{x_2})' V_{z_S} L_3$. So $(L^{x_2})' = P^{x_2} (V_{z_S} L_3)^{-1}$. So $\{\lambda^m_{O,x_2}\}_{m=1}^M$ is identified. With similar approach $\{\lambda^m_{O,x_3}\}_{m=1}^M$ for any $x_3 \in X_3$ can also be identified.

To identify $V_{z_S'}$ for any $z_S' \in \mathscr{Z}_{\mathscr{S}}$, construct $P^O_{z_S'}$ by replacing z_S with z_S' in $P^O_{z_S}$. $P^O_{z_S'}$ can be expressed as $P^O_{z_S'} = L_2' V_{z_S'} L_3$. Then $V_{z_S'} = (L_2')^{-1} P^O_{z_S'} L_3^{-1}$ and $\{\widetilde{\pi}^m_{z_S'}\}_{m=1}^M$ for any $z_S' \in \mathscr{Z}_{\mathscr{S}}$ is identified. By integrating out S, we can get $\{\pi^m\}_{m=1}^M$ and $f^m(S|Z_S) = \widetilde{\pi}^m_{z_S'} / \pi^m$.

I have shown the identification of π^m, $f^m(S|Z_S)$,$f^m(O_1|X_1,S)$ and $f^m(O_t|X_t,S)$

for any $(\{X_t\}_{t=1}^3, S, Z_S) \in \prod_{t=1}^3 \mathscr{X}_t \times \mathscr{S} \times \mathscr{Z}_{\mathscr{S}}$. The rest is to show the identification of the type-specific wage marginal distributions. Define

$$\begin{aligned}
\lambda^{*m}_{W,(w_1,x_1)} &= f^m((W_1,O_1)=(w_1,1)|(X_1,S)=(\eta_1,s)),\\
D^W_{w,\eta_1} &= diag(\lambda^{*1}_{W,(w,\eta_1)},\ldots,\lambda^{*M}_{W,(w,\eta_1)}).
\end{aligned}$$

Fix $W_1 = w_1$, $O_t = 1$ for $t = 1,2,3$, and define the following functions:

$$\begin{aligned}
F^{W*}_{Z_S,X_1,X_2,X_3} &= \sum_{m=1}^M \widetilde{\pi}^m_{Z_S}\lambda^{*m}_{W,(w_1,X_1)}\lambda^m_{O,X_2}\lambda^m_{O,X_3},\\
F^{W*}_{Z_S,X_1,X_2} &= \sum_{m=1}^M \widetilde{\pi}^m_{Z_S}\lambda^{*m}_{W,(w_1,X_1)}\lambda^m_{O,X_2},\\
F^{W*}_{Z,X_1,X_3} &= \sum_{m=1}^M \widetilde{\pi}^m_{Z_S}\lambda^{*m}_{W,(w_1,X_1)}\lambda^m_{O,X_3},\\
F^{W*}_{Z,X_1} &= \sum_{m=1}^M \widetilde{\pi}^m_{Z_S}\lambda^{*m}_{W,(w_1,X_1)}.
\end{aligned}$$

Arrange these to an $M \times M$ matrix:

$$P^{W*}_{z_S,\eta_1} = \begin{pmatrix}
F^{W*}_{z_S,\eta_1} & F^{W*}_{z_S,\eta_1,\eta_{3,1}} & \cdots & F^{W*}_{z_S,\eta_1,\eta_{3,M-1}}\\
F^{W*}_{z_S,\eta_1,\eta_{2,1}} & F^{W*}_{z_S,\eta_1,\eta_{2,1},\eta_{3,1}} & \cdots & F^{W*}_{z_S,\eta_1,\eta_{2,1},\eta_{2,M-1}}\\
\vdots & \vdots & \ddots & \vdots\\
F^{W*}_{z_S,\eta_1,\eta_{2,M-1}} & F^{W*}_{z_S,\eta_1,\eta_{2,M-1},\eta_{3,1}} & \cdots & F^{W*}_{z_S,\eta_1,\eta_{2,M-1},\eta_{3,M-1}}
\end{pmatrix}.$$

$P^{W*}_{z_S,\eta_1} = L_2' V_{z_S} D^W_{w_1,\eta_1} L_3$. Then $D^W_{w_1,\eta_1} = (L_2' V_{z_S})^{-1} P^{W*}_{z_S,\eta_1} L_3^{-1}$ and $f^m(W_1, O_1|X_1, S)$ is identified. Further $f^m((W_1|O_1,X_1,S) = f^m(W_1,O_1|X_1),S)/f^m((O_1|X_1),S)$.

To identify $f^m(W_t|O_t,X_t,S)$ for $t = 2$, define

$$\lambda^m_{W,(w_2,\eta_2)} = f^m((W_2,O_2)=(w_2,1)|(X_2,S)=(\eta_2,s)).$$

Fix $W_2 = w_2$, $O_t = 1$ for $t = 2,3$, and define the following functions:

$$F^W_{Z_S,X_2,X_3} = \sum_{m=1}^M \widetilde{\pi}^m_{Z_S}\lambda^m_{W,(w_2,X_2)}\lambda^m_{O,X_3},$$

$$F^{W}_{Z_S,X_2} = \sum_{m=1}^{M} \widetilde{\pi}^{m}_{Z_S} \lambda^{m}_{W,(w_2,X_2)}.$$

Then construct the following matrices:

$$L^{w_2} = \begin{pmatrix} 1 & \lambda^{1}_{W,(w_2,\eta_2)} \\ \vdots & \vdots \\ 1 & \lambda^{M}_{W,(2,\eta_2)} \end{pmatrix},$$

and

$$P^{w_2} = \begin{pmatrix} 1 & F^{O}_{z_S,\eta_{3,1}} & \cdots & F^{O}_{z_S,\eta_{3,M-1}} \\ F^{W}_{z_S,x_2} & F^{W}_{z_S,\eta_2,\eta_{3,1}} & \cdots & F^{W}_{z_S,\eta_2,\eta_{3,M-1}} \end{pmatrix}.$$

P^{w_2} can be expressed as $P^{w_2} = (L^{w_2})' V_{z_S} L_3$. Then $(L^{w_2})' = P^{w_2}(V_{z_S}L_3)^{-1}$ and $\{f^m(W_2,O_2|X_2,S)\}_{m=1}^{M}$ is identified and $f^m(W_2|O_2,X_2,S) = f^m(W_2,O_2|X_2,S)/f^m(O_2|X_2,S)$ for $m = 1,\ldots,M$. With similar approach, $f^m(W_3|O_3,X_3,S)$ can be identified for $m = 1,\ldots,M$. This completes the proof of Proposition 1.1. □

A.4.2 Assumptions and proof of Proposition 1.2

Assumption A.3. *For m=1,...,M and t* ≥ 2*,*

(a)

$$f^{m}_{t}(W_t|O_t,X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^{m}(W_t|O_t,X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}),$$

and

$$f^{m}_{t}(O_t|X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^{m}(O_t|X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}).$$

(b)

$$f^{m}(W_t|O_t,X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^{m}(W_t|O_t,S,X_t,W_{t-1},O_{t-1},X_{t-1}),$$

and

$$f^{m}(O_t|X_t,S,\{W_\tau,O_\tau,X_\tau\}_{\tau=2}^{t-1}) = f^{m}(O_t|O_{t-1},S,X_t).$$

Under Assumption A.3, the conditional joint distribution of wages, occupations, and education can be simplified as follows:

$$\begin{aligned}
&f^m(\{W_t,O_t\}_{t=1}^T,S|\{X_t\}_{t=1}^T,Z_S)\\
=&f^m(W_1|O_1,S)\prod_{t=2}^T f^m(W_t|O_t,S,X_t,W_{t-1},O_{t-1},X_{t-1})\\
&\times f^m(O_1|S)\prod_{t=2}^T f^m(O_t|O_{t-1},S,X_t)f^m(S|Z_S).
\end{aligned}$$

The transition process of (W_t,O_t,X_t) becomes a stationary first-order Markov process. Define $Y_t=(W_t,O_t,X_t)$. The variation of Y_t affects both the type-specific conditional joint distribution at period t and that at period $t+1$. This makes it difficult to construct factorization equations as before. To solve this problem, we look at every other period. Fix Y_t to be $\bar{y}_t$ for odd t and define

$$\begin{aligned}
&\widetilde{\pi}^m_{\bar{y},z_S}=\pi^m f^m(\bar{y}_1,s|z_S),\\
&\lambda^m_{\bar{y}}(Y_t)=f^m(\bar{y}_{t+1}|Y_t,s)f^m(Y_t|\bar{y}_{t-1},s),\\
&\lambda^{*m}_{\bar{y}}(Y_T)=f^m(Y_T|\bar{y}_{T-1},s).
\end{aligned}$$

Let ξ_t be element of $\mathscr{Y}_t$ and define

$$L_{t,\bar{y}}=\begin{pmatrix}1 & \lambda^1_{\bar{y}}(\xi_{t,1}) & \dots & \lambda^1_{\bar{y}}(\xi_{t,M-1})\\ \vdots & \vdots & \ddots & \vdots\\ 1 & \lambda^M_{\bar{y}}(\xi_{t,1}) & \dots & \lambda^M_{\bar{y}}(\xi_{t,M-1})\end{pmatrix},$$

$$V_{\bar{y}}=diag(\widetilde{\pi}^1_{\bar{y},z_S},\dots,\widetilde{\pi}^M_{\bar{y},z_S}),\ and\ D^O_{Y_T|\bar{y}}=diag(\lambda^{*1}_{\bar{y}}(Y_T),\dots,\lambda^{*M}_{\bar{y}}(Y_T)).$$

Then construct

$$P^O_{\bar{y}}=L'_{2,\bar{y}}V_{\bar{y}}L_{4,\bar{y}},$$

$$P^{O*}_{\bar{y}}=L'_{2,\bar{y}}D^O_{Y_T|\bar{y}}V_{\bar{y}}L_{4,\bar{y}}.$$

Further, assume

Assumption A.4. *There exist some $\{\xi_{t,1},\dots,\xi_{t,M-1}\}_{t=1}^T$ such that $P^O_{\bar{y}}$ is of full rank and that all the eigenvalues of $(P^O_{\bar{y}})^{-1}P^{O*}_{\bar{y}}$ take distinct values.*

Proof of Proposition 1.2. Without loss of generality, set $T = 6$. Fix $(Y_1, Y_3, Y_5) = (y_1, y_2, y_5)$ and define

$$\begin{aligned}
F^{*O}_{Y_2,Y_4,Y_6} &= \sum_{m=1}^{M} \widetilde{\pi}_{\bar{y},Z_S} \lambda^{m}_{\bar{y}}(Y_2) \lambda^{m}_{\bar{y}}(Y_4) \lambda^{*m}_{\bar{y}}(Y_T), \\
F^{*O}_{Y_2,Y_6} &= \sum_{m=1}^{M} \widetilde{\pi}_{\bar{y},z_S} \lambda^{m}_{\bar{y}}(Y_2) \lambda^{*m}_{\bar{y}}(Y_T), \\
F^{*O}_{Y_6} &= \sum_{m=1}^{M} \widetilde{\pi}_{\bar{y},z_S} \lambda^{*m}_{\bar{y}}(Y_T), \\
F^{O}_{Y_2,Y_4} &= \sum_{m=1}^{M} \widetilde{\pi}_{\bar{y},z_S} \lambda^{m}_{\bar{y}}(Y_2) \lambda^{m}_{\bar{y}}(Y_4), \\
F^{O}_{Y_2} &= \sum_{m=1}^{M} \widetilde{\pi}_{,z_S} \lambda^{m}_{\bar{y}}(Y_2), \\
F^{O} &= \sum_{m=1}^{M} \widetilde{\pi}_{\bar{y},z_S}.
\end{aligned}$$

And construct matrices as follows:

$$P^{O}_{\bar{y}} = \begin{pmatrix} F^{O} & F^{O}_{\xi_{4,1}} & \cdots & F^{O}_{\xi_{4,M-1}} \\ F^{O}_{\xi_{2,1}} & F^{O}_{\xi_{2,1},\xi_{4,1}} & \cdots & F^{O}_{\xi_{2,1},\xi_{4,M-1}} \\ \vdots & \vdots & \ddots & \vdots \\ F^{O}_{\xi_{2,M-1}} & F^{O}_{\xi_{2,M-1},\xi_{4,1}} & \cdots & F^{O}_{\xi_{2,M-1},\xi_{4,M-1}} \end{pmatrix},$$

and

$$P^{O*}_{\bar{y}} = \begin{pmatrix} F^{O*}_{\xi_6} & F^{O*}_{\xi_{4,1},\xi_6} & \cdots & F^{O*}_{\xi_{4,M-1},\xi_6} \\ F^{O*}_{\xi_{2,1},\xi_6} & F^{O*}_{\xi_{2,1},\xi_{4,1},\xi_6} & \cdots & F^{O*}_{\xi_{2,1},\xi_{4,M-1},\xi_6} \\ \vdots & \vdots & \ddots & \vdots \\ F^{O*}_{\xi_{2,M-1},\xi_6} & F^{O*}_{\xi_{2,M-1},\xi_{4,1},\xi_6} & \cdots & F^{O*}_{\xi_{2,M-1},\xi_{4,M-1},\xi_6} \end{pmatrix}.$$

Then repeat the argument of the proof of Proposition 1.1 and we achieve the identification of $\widetilde{\pi}^{m}_{\bar{y},z_S}$, $\lambda^{m}_{\bar{y}}(\xi_t)$, and $\lambda^{*m}_{\bar{y}}(Y_T)$. Then integrate out the other elements and apply Bayes' rule, we can get π^m, $f^m(W_1|O_1, X_1, S)$, $f^m(O_1|X_1, S)$, $f^m(S|Z_S)$, $f^m(W_t|O_t, S, X_t, W_{t-1}, O_{t-1}, X_{t-1})$, and $f^m(O_t|O_{t-1}, S, X_t)$.

□

A.4.3 Assumptions and proof of Proposition 1.3

Denote the support of Q_1, Q_2, and Q_3 by $\mathscr{Q}_1$, $\mathscr{Q}_2$, and $\mathscr{Q}_3$ respectively. Partition $\mathscr{Q}_1$ into M mutually exclusive and exhaustive subsets and denote the partitions as $\triangle_{Q_1} = \{\delta^1_{Q_1}, \ldots, \delta^M_{Q_1}\}$. Similarly denote the partitions of $\mathscr{Q}_2$ as $\triangle_{Q_2} = \{\delta^1_{Q_2}, \ldots, \delta^M_{Q_2}\}$. Let $\triangle = \triangle_{Q_1} \times \triangle_{Q_2}$. Also partition $\mathscr{Q}_3$ into 2 mutually exclusive and exhaustive subsets as $\triangle_{Q_3} = \{\delta^1_{Q_3}, \delta^2_{Q_3}\}$.

Let's define

$$\begin{aligned}
&p^m_{Q_1} = (P^m(Q_1 \in \delta^1_{Q_1}|s,z_s), \ldots, P^m(Q_1 \in \delta^M_{Q_1})|s,z_s)',\\
&p^m_{Q_2} = (P^m(Q_2 \in \delta^1_{Q_2}|s,z_s), \ldots, P^m(Q_2 \in \delta^M_{Q_2})|s,z_s)',\\
&p^m_{Q_3}(h) = P^m(Q_3 \in \delta^h_{Q_3}|s,z_s),\\
&\widetilde{\pi}^m = \pi^m f^m(s,z_s).
\end{aligned}$$

Collect the type-specific distributions into following matrices

$$L_{Q_1} = (p^1_{Q_1}, \ldots, p^M_{Q_1}),$$

$$L_{Q_2} = (p^1_{Q_2}, \ldots, p^M_{Q_2}),$$
$$V = diag(\widetilde{\pi}^1, \ldots, \widetilde{\pi}^M),$$

and

$$D_h = diag(p^1_{Q_3}(h), \ldots, p^M_{Q_3}(h)).$$

Let $P_s(Q_1 \in \delta^m_{Q_1}, Q_2 \in \delta^{m'}_{Q_2})$ be the probability that $Q_1 \in \delta^m_{Q_1}$ and $Q_2 \in \delta^{m'}_{Q_2}$ for $S = s$ and $P_s(Q_1 \in \delta^m_{Q_1}, Q_2 \in \delta^{m'}_{Q_2}, Q_3 \in \delta^h_{Q_3})$ be the probability that $Q_1 \in \delta^m_{Q_1}$, $Q_2 \in \delta^{m'}_{Q_2}$, and $Q_3 \in \delta^h_{Q_3}$ for $S = s$. Define two $M \times M$ matrices as follows:

$$P_\triangle = \begin{pmatrix} P_s(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^1_{Q_2}) & \ldots & P_s(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^M_{Q_2}) \\ \vdots & \ldots & \vdots \\ P_s(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^1_{Q_2}) & \ldots & P_s(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^M_{Q_2}) \end{pmatrix},$$

$$P_{\triangle,h} = \begin{pmatrix} P_s(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^1_{Q_2}, Q_3 \in \delta^h_{Q_3}) & \dots & P_s(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^M_{Q_2}, Q_3 \in \delta^h_{Q_3}) \\ \vdots & \dots & \vdots \\ P_s(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^1_{Q_2}, Q_3 \in \delta^h_{Q_3}) & \dots & P_s(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^M_{Q_2}, Q_3 \in \delta^h_{Q_3}) \end{pmatrix}.$$

Assume:

Assumption A.5. *There exists a partition $\triangle \times \triangle_{Q_3}$ on the variables (Q_1, Q_2, Q_3) for which the matrix $P_\triangle$ is nonsingular and the eigenvalues of $P_{\triangle,h}P_\triangle^{-1}$ are distinct for partition level $h = 1$ of the variable Q_3.*

Proof of Proposition 1.3. $P_\triangle$ and $P^*_{\triangle,h}$ can be expressed as the follows:

$$P_\triangle = L_{Q_1}V(L'_{Q_2}), \text{ and } P_{\triangle,h} = L_{Q_1}D_hV(L'_{Q_2}).$$

Since $P_\triangle$ is nonsingular, both L_{Q_1} and L_{Q_2} are nonsingular. Construct $A_h = P_{\triangle,h}P_\triangle^{-1} = L_{Q_1}D_hL_{Q_1}^{-1}$, and we have $A_hL_{Q_1} = L_{Q_1}D_h$. The distinct eigenvalues of A_h determines the elements of D_h, and its eigenvectors determine the columns of L_{Q_1} uniquely up to a multiplicative constant. Then L_{Q_1} is uniquely determined since the elements of each column of L_{Q_1} must sum to one. Construct $B_h = (P'_{\triangle,h})(P'_\triangle)^{-1} = L_{Q_2}D_hL_{Q_2}^{-1}$, and L_{Q_2} is determined using the similar argument. Once L_{Q_1} and L_{Q_2} are determined, V is uniquely determined by $V = (L_{Q_1})^{-1}P_\triangle(L'_{Q_2})^{-1}$. Then $\{\pi^m\}_{m=1}^M$ is identified by integrating out S and Z_S, and $f^m(S|Z_S) = \widetilde{\pi}^m/(\pi^m f(Z_S))$.

For any $q_1 \in \mathscr{Q}_1$, denote $p_{q_1} = (P^1_{Q_1}(q_1), \dots, P^M_{Q_1}(q_1))$ and define $P_{q_1,\triangle Q_2} = p_{q_1}V(L_{Q_2})'$. Then $p_{q_1} = P_{q_1,\triangle Q_2}(V(L_{Q_2})')^{-1}$, and $\{P^M_{Q_1}(q_1)\}_{m=1}^M$ is identified. Define $P_{\triangle Q_1,q_2}$ and $P_{\triangle Q_1,q_3}$ analogously and apply the same argument, $\{P^M_{Q_2}(q_2), P^M_{Q_3}(q_3)\}_{m=1}^M$ are identified.

□

A.4.4 Assumptions and proof of Proposition 1.4

Denote $p^m_{O_t} = P^m(O_t = 1|(S, Z_S) = (s, z_S))$, and $D_{O_t} = diag(p^1_{O_t}, \dots, p^M_{O_t})$. Construct an $M \times M$ matrix

$$P_{\triangle,O_t} = \begin{pmatrix} P(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^1_{Q_2}, O_t = 1) & \dots & P(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^M_{Q_2}, O_t = 1) \\ \vdots & \dots & \vdots \\ P(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^1_{Q_2}, O_t = 1) & \dots & P(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^M_{Q_2}, O_t = 1) \end{pmatrix},$$

Assume

Assumption A.6. *The eigenvalues of $P_{\triangle,O_t}P_{\triangle}^{-1}$ are distinct.*

Proof of Proposition 1.4. The proof of the nonparametric identification of education psychic costs using test scores is already shown in the proof of Proposition 1.3. Below, I prove the nonparametric identification of occupation abilities using test scores.

Express $P_{\triangle,O_t}$ as $P_{\triangle,O_t} = L_{Q_1}D_{O_t}VL'_{Q_2}$. Replacing $P_{\triangle,h}$ in the proof of Proposition 1.3 by $P_{\triangle,O_t}$, and repeating the proof, π^m, $f^m(S|Z_S)$, and $f^m(O_t|X_t,S)$ are identified.

Next, denote $p^m_{W_t} = F^m((W_t,O_t) = (w_t,1)|(S,Z_S) = (s,z_S))$ and $D_{W_t} = diag(p^1_{W_t},\dots,p^M_{W_t})$. Let $P(Q_1 \in \delta^m_{Q_1}, Q_2 \in \delta^{m'}_{Q_2},(\omega_t,1))$ be the probability that $Q_1 \in \delta^m_{Q_1}$, $Q_2 \in \delta^{m'}_{Q_2}$, $W_t = \omega_t$, and $O_t = 1$ for $S = s$. The corresponding $M \times M$ matrix is

$$P_{\triangle,w_t} = \begin{pmatrix} P_S(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^1_{Q_2},(\omega_t,1)) & \dots & P_S(Q_1 \in \delta^1_{Q_1}, Q_2 \in \delta^M_{Q_2},(\omega_t,1)) \\ \vdots & \dots & \vdots \\ P_S(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^1_{Q_2},(\omega_t,1)) & \dots & P_S(Q_1 \in \delta^M_{Q_1}, Q_2 \in \delta^M_{Q_2},(\omega_t,1)) \end{pmatrix}.$$

$P_{\triangle,w_t} = L_{Q_1}D_{w_t}VL'_{Q_2}$. Then $D_{w_t} = L^{-1}_{Q_1}P_{\triangle,w_t}(VL'_{Q_2})^{-1}$, and $f^m(W_t,O_t|X_t,S)$ is identified. By Bayes' rule, $f^m(W_t|O_t,X_t,S) = f^m(W_t,O_t|X_t,S)/f^m(O_t|X_t,S)$.

□

A.5 EM Algorithm

Consider $(k+1)$th iteration. In E step, calculate the expectated log-likelihood ϕ based on the estimates from the kth iteration:

$$\phi^{(k)} = \sum_{i=1}^{n}\sum_{m=1}^{M}\mu_i^{m(k)}(log\pi^m + log\mathscr{L}^m_W + log\mathscr{L}^m_O + log\mathscr{L}^m_S + log\mathscr{L}^m_Q),$$

where

$$\mu_i^{m(k)} = \frac{\pi^{m(k)}\mathscr{L}^{m(k)}_W\mathscr{L}^{m(k)}_O\mathscr{L}^{m(k)}_S\mathscr{L}^{m(k)}_Q}{\sum_{m=1}^{M}\pi^{m(k)}\mathscr{L}^{m(k)}_W\mathscr{L}^{m(k)}_O\mathscr{L}^{m(k)}_S\mathscr{L}^{m(k)}_Q}.$$

In M step, compute the parameters by maximizing the expected log-likelihood ϕ:

$\pi^{m(k+1)}$ satisfies $\frac{\partial \phi^k}{\partial \pi^{m(k+1)}} = 0$. Correspondingly,

$$\pi^{m(k+1)} = \frac{\sum_{i=1}^{n} \mu_i \pi^{m(k)}}{n}.$$

$\beta_W^{m(k+1)}$ satisfies $\frac{\partial \phi^k}{\partial \beta_W^{m(k+1)}} = 0$. And it can be simplified to

$$\frac{\partial \sum_{i=1}^{n} \sum_{t=1}^{T} \mathscr{L}_W^{m(k+1)}}{\partial \beta_W^{m(k+1)}} = 0,$$

which is an OLS regression.

$\gamma_O^{m(k+1)}$ satisfies $\frac{\partial \phi^k}{\partial \gamma_O^{m(k+1)}} = 0$, and it can be simplified to

$$\frac{\partial \sum_{i=1}^{n} \sum_{t=1}^{T} \mathscr{L}_O^{m(k+1)}}{\partial \gamma_O^{m(k+1)}} = 0.$$

($\theta_R^{m(k+1)}$ satisfies $\frac{\partial \phi^k}{\partial \theta_R^{m(k+1)}} = 0$. And it can be simplified to

$$\frac{\partial \sum_{i=1}^{n} \sum_{r=1}^{R} \mathscr{L}_Q^{m(k+1)}}{\partial \theta_R^{m(k+1)}} = 0,$$

which is a probit.

$\delta_S^{m(k+1)}$ satisfies $\frac{\partial \phi^k}{\partial \delta_S^{m(k+1)}} = 0$, and it can be simplified to

$$\frac{\partial \sum_{i=1}^{n} \mathscr{L}_S^{m(k+1)}}{\partial \delta_S^{m(k+1)}} = 0,$$

which is a multinomial logit.

A.6 Choice of Initial Values

The strategy is to start with estimating the parameters in Equation (1.7) when the population is homogenous (M=1) and then add one more type at a time and re-estimate the parameters. Let $\mathscr{L}_i^m$ denote the likelihood for individual i and define

$$\mu^m = \sum_{i=1}^{n} (1 - \frac{\mathscr{L}_i^m}{\sum_{k=1}^{m-1} L_i^k \pi^k})$$

The estimation follows the algorithm as below:

(a) Set $M = 1$ and $\pi^1 = 1$. Choose initial values for α_W, α_O, α_S, α_R, β, λ, δ, θ, σ_W, σ_Q, and ρ in Equation (1.7).

(b) Given the current value of M, maximize the likelihood over α_W, α_O, α_S, α_R, β, λ, δ, θ, σ_W, σ_Q, ρ, and π^m.

(c) Evaluate μ^{M+1} for a grid of values of the type-specific parameters.

(d) Set the type-specific parameters to the values that yield the smallest value for μ^{M+1}.

(e) Maximize the likelihood. Increase the value of M by 1. Return to Step (b).

Appendix B

B.1 Proofs of the Results in Section 2.2

Proof of PROPOSITION 2.1: Because the second result follows from the first by the law of the iterated expectations, and the third and fourth results can be easily derived from the second result, we only prove the first result below. Let (h, L) be an arbitrary member in $\mathbb{H}$. Then we have that

$$N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}) = N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} C_{Lhk} \phi_{Lh}(\tilde{X}_{hk}). \tag{B.1}$$

For each $k \in \mathbb{N}$, C_{Lhk} has a bounded support by Assumption 2.2(c), and $\phi_{Lh}(\tilde{X}_{hk})$ has a finite absolute moment by hypothesis, so that $C_{Lhk}\phi_{Lh}(\tilde{X}_{hk})$ has a finite absolute moment. From this fact and the independence of C_{Lhk} and $\tilde{X}_{hk}$ (Assumption 2.2(b)), it follows by Fubini's theorem (Folland, 1984, Theorem 2.37, pp. 65–66) that for each $k \in \mathbb{N}$,

$$\mathrm{E}[C_{Lhk}\phi_{Lh}(\tilde{X}_{hk}) \,|\, \tilde{X}] = \mathrm{E}[C_{Lhk}]\, \phi_{Lh}(\tilde{X}_{hk}).$$

Also, w_{Lhk} is the reciprocal of $\mathrm{E}[C_{Lhk}]$ by definition. Taking the conditional expectation of both side in (B.1) given $\tilde{X}$ and applying these facts yields that

$$\begin{aligned} \mathrm{E}\left[N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}) \,\middle|\, \tilde{X}\right] &= N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \mathrm{E}[C_{Lhk}]\, \phi_{Lh}(\tilde{X}_{hk}) \\ &= N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} \phi_{Lh}(\tilde{X}_{hk}). \end{aligned}$$

Thus, the desired result follows. □

Proof of THEOREM 2.2: Let n be an arbitrary natural number. For each $x \in \mathbb{R}^v$, $q(x,\cdot) : \Theta \to \mathbb{R}$ is continuous on Θ, given Assumption 2.3. Thus, for each $\omega \in \Omega$, $Q_L(\omega,\cdot) : \Theta \to \mathbb{R}$ is a continuous function on the compact set Θ under Assumptions 2.1 and 2.2. The desired result therefore follows by (Gallant and White, 1988, Lemma 2.1). □

B.2 Proofs of the Results in Section 2.3

Proof of PROPOSITION 2.1: Because for each $(h,L) \in \mathbb{H}$ and each $\gamma \in \Gamma$,

$$N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij},\gamma) = N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} C_{Lhk} \phi_{Lh}(\tilde{X}_{hk},\gamma),$$

we have that for each $(h,L) \in \mathbb{H}$ and each $\gamma \in \Gamma$,

$$\left\| N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij},\gamma) \right\|_a \leq N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \left\| C_{Lhk} \phi_{Lh}(\tilde{X}_{hk},\gamma) \right\|_a.$$

Given the independence between $\{C_{Lhk} : (k,h,L) \in \mathbb{K}\}$ and $\{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$ (Assumption 2.2(b)), it follows by Fubini's theorem (Folland, 1984, Theorem 2.37, pp. 65–66) that for each $(h,L) \in \mathbb{H}$ and each $\gamma \in \Gamma$,

$$\left\| N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij},\gamma) \right\|_a \leq N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \| C_{Lhk} \|_a \cdot \| \phi_{Lh}(\tilde{X}_{hk},\gamma) \|_a.$$

Under Assumption 2.2(b)–(d), we have that for each $k \in \mathbb{N}$ and each $(h,L) \in \mathbb{H}$,

$$w_{Lhk} = \mathrm{E}[C_{Lhk}]^{-1} \leq \big(0 \cdot P[C_{Lhk} = 0] + 1 \cdot P[C_{Lhk} > 0]\big)^{-1} = P[C_{Lhk} > 0]^{-1} \leq (N_{Lh}^{-1} \bar{p}_l)^{-1}$$

and

$$\| C_{Lhk} \|_a \cdot \leq \bar{C} P[C_{Lhk} > 0] \leq N_{Lh}^{-1} \bar{C} \bar{p}_u.$$

Also, $\Delta \equiv \sup_{\gamma\in\Gamma}\sup_{(k,h,L)\in\mathbb{K}} \|\phi_{Lh}(\tilde{X}_{hk})\|_a < \infty$ by hypothesis. It follows that

$$\left\| N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\phi_{Lh}(X_{Lhij})\right\|_a \le (\bar{p}_u/\bar{p}_l)\bar{C}\cdot N_{Lh}^{-1}\sum_{k=1}^{N_{Lh}} w_{Lhk}\|\phi_{Lh}(\tilde{X}_{hk},\gamma)\|_a$$
$$\le (\bar{p}_u/\bar{p}_l)\bar{C}\Delta, \quad (h,L)\in\mathbb{H}.$$

Because the right-hand side of the above inequality depends on neither h nor L, the desired therefore result follows. □

Proof of LEMMA 2.2: Because $\{\phi_h\}_{h\in\mathbb{N}}$ satisfies condition (a) in Definition 1, $\{\Phi_{Lh}\}_{L\in\mathbb{N}}$ is uniformly L_a-bounded by Proposition 2.1, so that it satisfies condition (a) of Definition 2. Due to condition (b) in Definition 1, $\{\phi_h\}$ satisfies that for each $(\gamma_1,\gamma_2)\in\Gamma^2$ and each $(h,L)\in\mathbb{H}$,

$$|\Phi_{Lh}(\cdot,\gamma_2)-\Phi_{Lh}(\cdot,\gamma_1)| \le N_{Lh}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\left|\phi_h(X_{Lhij}(\omega),\theta_2)-\phi_h(X_{Lhij}(\omega),\theta_1)\right|$$
$$\le D_{Lh}\, g(|\theta_2-\theta_1|),$$

where

$$D_{Lh}\equiv N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)d_h(X_{Lhij}(\omega)).$$

Because $\sup_{(k,h)\in\mathbb{N}^2}\int d_h(x)\,P_{hk}(dx)<\infty$, it follows from Proposition 2.1 that $\{D_{Lh} : (h,L)\in\mathbb{H}\}$ is uniformly L_1 bounded, so that $\Delta\equiv\sup_{(h,L)\in\mathbb{H}}\mathrm{E}[D_{Lh}]<\infty$. Using this fact, we obtain that

$$\sup_{L\in\mathbb{N}} L^{-1}\sum_{h=1}^{L}\mathrm{E}[D_{Lh}]\le\Delta,$$

as required by condition (b) in Definition 2. Thus, $\{\Phi_{Lh} : (h,L)\in\mathbb{H}\}$ is SLB(a). □

In proving Theorem 2.3, we employ a uniform law of large numbers stated in the next lemma, in which the SLB property takes an important role.

Lemma B.1. *Given Assumptions 2.1 and 2.2, let Γ be a finite-dimensional Euclidean space, and $\{F_{Lh} : (h,L)\in\mathbb{H}\}$ an array of measurable functions from $(\Omega\times\Gamma,\mathscr{F}\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$ such that for each $\gamma\in\Gamma$ and each $L\in\mathbb{N}$, $F_{L1}(\cdot,\gamma)$, ..., $F_{LL}(\cdot,\gamma)$ are independent, and $\{F_{Lh} : (h,L)\in\mathbb{H}\}$ is SLB($1+\delta$) on Γ for some $\delta\in(0,\infty)$. Then $\{\gamma\mapsto L^{-1}\sum_{h=1}^{L}\mathrm{E}[F_{Lh}(\cdot,\gamma)] : \Gamma\to\mathbb{R}\}_{L\in\mathbb{N}}$ is uniformly bounded* *and uniformly equicontinuous, and* $\{|L^{-1}\sum_{h=1}^{L} F_{Lh}(\cdot,\gamma) - L^{-1}\sum_{h=1}^{L} \mathrm{E}[F_{Lh}(\cdot,\gamma)]|\}$ *converges in probability-P to zero uniformly in* $\gamma \in \Gamma$.

Proof of LEMMA B.1: The result can be established essentially in the same way as Lemma A.3 of **?**. □

We now prove Theorem 2.3.

Proof of THEOREM 2.3: To establish the desired result, we apply the standard consistency result, e.g., (**?**, Lemma 4.2). Given the compactness of Θ (Assumption 2.3) and the identifiability of $\{\Theta_L^*\}_{L\in\mathbb{N}}$ (Assumption 2.5), it suffices to show the uniform convergence of $\{Q_L(\cdot,\theta) - \bar{Q}_L(\theta)\}_{L\in\mathbb{N}}$ to zero over $\theta \in \Theta$ in probability-P

To prove the above-mentioned uniform convergence, define $\{F_{Lh} : (h,L) \in \mathbb{H}\}$, an array of functions from $\Omega \times \Theta$ to $\mathbb{R}$ by

$$F_{Lh}(\omega,\theta) \equiv (L/N_L)\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\, q(X_{Lhij}(\omega),\theta), \quad \omega \in \Omega,\, \theta \in \Theta,\, (h,L) \in \mathbb{H}.$$

Then Q_L can be written as

$$Q_L(\theta) = L^{-1}\sum_{h=1}^{L} F_{Lh}(\cdot,\theta), \quad L \in \mathbb{N}.$$

Also, by Proposition 2.1, it follows from the definition of Q_L that for each $\theta \in \Theta$, $\mathrm{E}[Q_L(\theta)] = \bar{Q}_L(\theta)$. Thus, if $\{F_{Lh} : (h,L) \in \mathbb{H}\}$ obeys the uniform law of large numbers on Θ, the desired uniform convergence of $\{Q_L(\cdot,\theta) - \bar{Q}_L(\theta)\}_{L\in\mathbb{N}}$ holds.

To verify that $\{F_{Lh} : (h,L) \in \mathbb{H}\}$ obeys the uniform law of large numbers, we use Lemma B.1, which states that if (A) for each $\theta \in \Theta$ and each $L \in \mathbb{N}$, $F_{L1}(\cdot,\theta)$, $\ldots$, $F_{LL}(\cdot,\theta)$ are independent, and (B) for some positive real number δ, $\{F_{Lh}\}$ is SLB$(1+\delta)$, then $\{F_{Lh} : (h,L) \in \mathbb{H}\}$ obeys the uniform law of large numbers. Under Assumptions 2.1 and 2.2(a), (b), (A) is clearly satisfied. For (B), rewrite $F_{Lh}(\cdot,\theta)$ as

$$F_{Lh}(\cdot,\theta) = (LN_{Lh}/N_L)\, N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\, q(X_{Lhij},\theta), \quad \theta \in \Theta,\, (h,L) \in \mathbb{H}.$$

Because q, which is common for every stratum, is LB$(1+\delta)$, the array

$$\Big\{N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\,q(X_{Lhij},\theta) : (h,L)\in\mathbb{H}\Big\}.$$

is SLB$(1+\delta)$ by Lemma 2.2. The array $\{LN_{Lh}/N_L = N_{Lh}/(N_L/N_{Lh}) : (h,L)\in\mathbb{H}\}$ is also bounded (Assumption 2.1). It is straight forward to verify that $\{F_{Lh} : (h,L)\in\mathbb{H}\}$ is SLB$(1+\delta)$ using these facts. Thus, $\{F_{Lh} : (h,L)\in\mathbb{H}\}$ obeys the uniform law of large numbers, and the desired result follows. □

B.3 Proofs of the Results in Section 2.4

In proving Theorem 2.1, we employ a central limited theorem stated in the next lemma.

Lemma B.1. *Let $\{U_{Lh} : (h,L)\in\mathbb{H}\}$ be an array of uniformly $\mathscr{L}_{2+\delta}$-bounded, zero-mean $v\times 1$ random vectors for some $\delta\in[0,\infty)$ such that for each $L\in\mathbb{N}$, $U_{L1},\ldots,U_{LL}$ are independent. Then:*

(a) $L^{-1/2}\sum_{h=1}^{L} U_{Lh} = \mathrm{O}_P(1)$.

(b) Suppose in addition that $\delta>0$, and $\{V_L \equiv L^{-1}\sum_{h=1}^{L}\operatorname{var}[U_{Lh}]\}_{L\in\mathbb{N}}$ is uniformly positive definite. Then $V_L^{-1/2}L^{-1/2}\sum_{h=1}^{L} U_{Lh} \overset{A}{\sim} \mathrm{N}(0,1)$.

Proof of LEMMA B.1: The result can be established essentially in the same way as Lemma A.4 of **?**. □

The following lemma is also useful in proving Theorem 2.1.

Lemma B.2. *Define $\{\tilde{A}_L : \Omega\times\operatorname{int}\Theta\to\mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$*

$$\tilde{A}_L(\omega,\theta)\equiv\nabla^2 Q_L(\cdot,\theta) = N_L^{-1}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\nabla^2 q(X_{Lhij}(\omega),\theta),\quad \omega\in\Omega, L\in\mathbb{N}.$$

Then $\{A_L : \operatorname{int}\Theta\to\mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$ is uniformly bounded and uniformly equicontinuous on Θ_0, and $\{\tilde{A}_L(\cdot,\theta)-A_L(\theta)\}_{L\in\mathbb{N}}$ converges in probability-P to zero uniformly in $\theta\in\Theta_0$.

Proof of LEMMA B.2: Note that for each $L \in \mathbb{N}$,

$$\tilde{A}_L(\cdot,\theta) = L^{-1} \sum_{h=1}^{L} (LN_{Lh}/N_L^{-1}) N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \nabla^2 q(X_{Lhij},\theta), \quad \theta \in \Theta_0.$$

Because $\nabla^2 q$ is LB$(1+\delta)$,

$$\left\{ (\omega,\theta) \mapsto N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij}(\omega) \nabla^2 q(X_{Lhij}(\omega),\theta) : (\Omega,\Theta_0) \to \mathbb{R}^{p\times p} : (h,L) \in \mathbb{H} \right\}$$

is SLB$(1+\delta)$ by Lemma 2.2. Also, $\{LN_{Lh}/N_L = N_{Lh}/(N_L/N_{Lh}) : (h,L) \in \mathbb{H}\}$ is bounded (Assumption 2.1). It is straight forward to verify that

$$\left\{ (\omega,\theta) \mapsto (LN_{Lh}/N_L^{-1}) N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij}(\omega) \nabla^2 q(X_{Lhij}(\omega),\theta) : (\Omega,\Theta_0) \to \mathbb{R}^{p\times p} \right.$$
$$\left. : (h,L) \in \mathbb{H} \right\} \tag{B.2}$$

is SLB$(1+\delta)$. Further, the array (B.2) is row-wise independent (i.e., independent across strata for each $L \in \mathbb{N}$). Thus, application of Lemma B.1 to the array (B.2) yields that $\{\theta \mapsto \mathrm{E}[\tilde{A}_L(\cdot,\theta)] : \Theta_0 \to \mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$ is uniformly bounded and uniformly equicontinuous, and

$$\tilde{A}_L(\cdot,\theta) - \mathrm{E}[\tilde{A}_L(\cdot,\theta)] \to 0 \text{ in probability-}P \text{ as } L \to \infty.$$

The desired result follows from this fact, because

$$\mathrm{E}[\tilde{A}_L(\cdot,\theta)] = A_L(\theta), \quad \theta \in \Theta_0, L \in \mathbb{N}$$

by Lemma B.1. □

We now prove Theorem 2.1.

Proof of THEOREM 2.1: By Proposition 2.1, it follows from Assumption 2.8(d) that

$$\left\{ N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \nabla q(X_{lhij},\theta^*) : (h,L) \in \mathbb{H} \right\}$$

is uniformly $L_{2+\delta}$-bounded. Because $\{LN_{Lh}/N_L = N_{Lh}/(N_L/N_{Lh}) : (h,L) \in \mathbb{H}\}$ is bounded under Assumption 2.1, it follows that $\{S^*_{Lh} : (h,L) \in \mathbb{H}\}$ is uniformly $L_{2+\delta}$-bounded, so that $\{B_{Lh} = \mathrm{var}[S^*_{Lh}] : (h,L) \in \mathbb{H}\}$ is bounded. Also, $\{A_L^{*\,-1}\}_{L\in\mathbb{N}}$ is bounded, due to the uniform positive definiteness of $\{A^*_L\}_{L\in\mathbb{N}}$ (Assumption 2.8). The first claim follows from these facts.

To verify the second claim, we use the standard linearization approach with smooth (generalized) scores. More concretely, we employ Theorem 6.10 of White (1994). Nevertheless, the setup of the theorem is slightly different from ours; in particular, the theorem requires that the score is continuously differentiable on the entire parameter space, while our setup does not require differentiability of q on the boundary of the parameter space. To fill the discrepancies between our problem setup and that of the theorem, we introduce $\{\tilde{\theta}_L : \Omega \to \Theta_0\}_{L\in\mathbb{N}}$ defined by

$$\tilde{\theta}_L \equiv \begin{cases} \hat{\theta} & \text{if } \hat{\theta} \in \Theta_0, \\ \theta^*_L & \text{otherwise.} \end{cases}$$

(White, 1994, Theorem 6.10) applies to $\{\tilde{\theta}_L\}_{L\in\mathbb{N}}$ well, as it lives in the compact space Θ_0 on which the score is continuously differentiable. Also, $\tilde{\theta}_L$ coincides with $\hat{\theta}_L$ with a probability approaching one as L grows to infinity, because $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ is consistent for $\{\theta^*_L\}_{L\in\mathbb{N}}$, which is uniformly interior to Θ_0. It follows that for each real number $\delta > 0$,

$$\begin{aligned} &P\Big[\big|L^{1/2}(D_L^{*\,-1/2}(\tilde{\theta}_L - \theta^*_L) - L^{1/2}D_L^{*\,-1/2}(\hat{\theta}_L - \theta^*_L)\big| < \delta\Big] \\ &\quad = P\Big[\big|D_L^{*\,-1/2}L^{1/2}(\tilde{\theta}_L - \hat{\theta}_L)\big| < \delta\Big] \\ &\quad \geq P[\tilde{\theta}_L = \hat{\theta}_L] \to 1 \text{ as } L \to \infty, \end{aligned}$$

i.e., $\tilde{\theta}_L - \hat{\theta}_L = \mathrm{o}_P(L^{-1/2})$ as $L \to \infty$. Thus, by the asymptotic equivalence lemma (White, 1984, Lemma 4.7), $\{L^{1/2}(\hat{\theta}_L - \theta^*_L)\}_{L\in\mathbb{N}}$ has the same asymptotic distributions as $\{L^{1/2}(\tilde{\theta}_L - \theta^*_L)\}_{L\in\mathbb{N}}$, if the latter is convergent in distribution. It thus suffices to prove that

$$D_L^{*\,-1/2}L^{1/2}(\tilde{\theta}_L - \theta^*_L) \overset{A}{\sim} \mathrm{N}(0,I) \text{ as } L \to \infty. \tag{B.3}$$

Define $\{\Psi_L : \Omega \times \Theta_0 \to \mathbb{R}^p\}_{L\in\mathbb{N}}$ by

$$\Psi_L(\omega,\theta) \equiv L^{-1} \sum_{h=1}^{L} S_{Lh}(\omega,\theta), \quad \omega \in \Omega,\, \theta \in \Theta_0,\, L \in \mathbb{N}.$$

Because $\Theta_0 \subset \operatorname{int}\Theta$, it holds that $\psi_L(\cdot,\tilde{\theta}_L) = 0$, whenever $\hat{\theta}_L \in \Theta_0$, satisfying the first order condition for the maximization of Q_L. Because $\{\tilde{\theta}_L\}_{L\in\mathbb{N}}$ is consistent for $\{\theta_L^*\}_{L\in\mathbb{N}}$, which is uniformly interior to Θ_0, we have that

$$P\Big[\big|L^{1/2}\Psi_L(\cdot,\tilde{\theta}_L)\big| < \delta\Big] \geq P[\hat{\theta}_L \in \Theta_0] \to 1 \text{ as } L \to \infty,$$

i.e., $L^{1/2}\Psi(\cdot,\tilde{\theta}_L) = \mathrm{o}_P(1)$ as $L \to \infty$. Given this property, it follows from (White, 1994, Theorem 6.10) that (B.3) holds if (A) $\{B_L^*\}_{L\in\mathbb{N}}$ is bounded and uniformly positive definite, (B) $B_L^{*-1/2} L^{1/2}\Psi_L(\cdot,\theta_L^*) \overset{A}{\sim} \mathrm{N}(0,I)$ as $L \to \infty$, (C) $\{A_L : \Theta_0 \to \mathbb{R}^{p\times p}\}_{L\in\mathbb{N}}$ is uniformly equicontinuous, (D) $\{\tilde{A}_L(\cdot,\theta) - A_L(\theta)\}_{L\in\mathbb{N}}$ converges in probability-P to zero uniformly in $\theta \in \Theta_0$ as $L \to \infty$ (note that $\tilde{A}_L(\cdot,\theta)$ is the Hessian of Q_L at θ), and (E) $\{A_L^*\}_{L\in\mathbb{N}}$ is uniformly nonsingular. Because we have verified above that $\{B_{Lh} : (h,L) \in \mathbb{H}\}$ is bounded, the average of it, $\{B_L\}_{L\in\mathbb{N}}$ is also bounded, and (A) holds. Also, Lemma B.2 has verified that the conditions (C) and (D) hold. Further, (E) holds by Assumption 2.8(c).

To verify the remaining condition, (B), define the array $\{U_{Lh} : (h,L) \in \mathbb{H}\}$ by

$$U_{Lh} \equiv S_{Lh}^* - \mathrm{E}[S_{Lh}^*], \quad (h,L) \in \mathbb{H}.$$

Now, by Proposition 2.1, we have that

$$\mathrm{E}[S_{Lh}^*] = (LN_{Lh}/N_L) \int \nabla q(x,\theta_L^*)\, \bar{P}_{Lh}(dx), \quad (h,L) \in \mathbb{H}.$$

Under Assumption 2.8(e), we can rewrite the right-hand side of this equation to obtain that

$$\mathrm{E}[S_{Lh}^*] = (LN_{Lh}/N_L)\nabla \int q(x,\theta_L^*)\, \bar{P}_{Lh}(dx).$$

It follows that for each $L \in \mathbb{N}$,

$$L^{-1} \sum_{h=1}^{L} \mathrm{E}[S_{Lh}^{*}] = \nabla \int q(x, \theta_L^*)\, \bar{P}_L(dx) = \nabla \bar{Q}_L(\theta_L^*) = 0,$$

where the last equality follows by the first order condition for maximization of $\bar{Q}_L$. We thus have that

$$\Psi_L(\cdot, \theta_L^*) = L^{-1} \sum_{h=1}^{L} S_{Lh}^{*} = L^{-1} \sum_{h=1}^{L} U_{Lh}, \quad L \in \mathbb{N}.$$

If the array $\{U_{Lh}\}$ obeys the central limit theorem, condition (B) is satisfied. Because $\{S_{Lh}(\cdot, \theta) :, (h, L) \in \mathbb{H},\, \theta \in \Theta_0\}$ is uniformly $L_{2+\delta}$-bounded, $\{U_{Lh} : (h, L) \in \mathbb{H}\}$ is uniformly $L_{2+\delta}$-bounded. Also, $U_{L1}, \ldots, U_{LL}$ are independent for each $L \in \mathbb{N}$. Further, $\{L^{-1} \sum_{h=1}^{L} \mathrm{var}[U_{Lh}] = L^{-1} \sum_{h=1}^{L} B_{Lh}\}_{L \in \mathbb{N}}$ is uniformly nonsingular by Assumption f. The condition (B) thus follows by Lemma B.1. □

B.4 Proofs of the Results in Section 2.5

Lemma B.1. *Under Assumptions 2.1–2.8, $\{\hat{A}_L\}_{L \in \mathbb{N}}$ is consistent for $\{A_L^*\}_{L \in \mathbb{N}}$.*

Proof of Lemma B.1: We have that for each $L \in \mathbb{N}$,

$$|\hat{A}_L - A_L^*| \leq |\hat{A}_L - A_L(\hat{\theta}_L)| + |A_L(\hat{\theta}_L) - A_L^*|. \tag{B.4}$$

It suffices to show that each of the two terms on the right-hand side of this inequality converges to zero in probability-P.

To show the convergence of the first term, let ε be an arbitrary positive real number. Then we have that

$$\begin{aligned} &P\big[\,|\hat{A}_L - A_L(\hat{\theta}_L)| \geq \varepsilon\,\big] \\ &= P\big[\,|\hat{A}_L - A_L(\hat{\theta}_L)| \geq \varepsilon \text{ and } \hat{\theta}_L \in \Theta_0\,\big] + P\big[\,|\hat{A}_L - A_L(\hat{\theta}_L)| \geq \varepsilon \text{ and } \hat{\theta}_L \notin \Theta_0\,\big] \\ &\leq P\Big[\sup_{\theta \in \Theta_0} |\tilde{A}_L(\theta) - A_L(\theta)| \geq \varepsilon\Big] + P\big[\hat{\theta}_L \notin \Theta_0\big] \end{aligned} \tag{B.5}$$

where $\tilde{A}_L$ is as in Lemma B.2. The first term on the right-hand of this inequality converges to zero by Lemma B.2. For the second term, there exists a real number $c > 0$ such that for each $L \in \mathbb{N}$, the open ball with radius c centered at θ_L^* is contained in Θ_0, because $\{\theta_L^*\}_{L\in\mathbb{N}}$ is uniformly interior to Θ_0. Thus, $P\big[\hat{\theta}_L \notin \Theta_0\big]$ is dominated by $P\big[d(\hat{\theta}_L, \theta_L^*) > c\big]$, which converges to zero by the consistency of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ for $\{\theta^*\}_{L\in\mathbb{N}}$. It follows that both terms on the right-hand side of (B.5) converge to zero. Because ε is an arbitrary positive real number, this verifies that the first term on the right-hand side of (B.4) converges to zero in probability-P.

We now turn to the second term on the right-hand side of (B.4). Pick a positive real number ε arbitrarily. Because $\{A_L\}_{L\in\mathbb{N}}$ is uniformly equicontinuous on Θ_0, there exists a real number $c > 0$ such that

$$\sup\big\{|A_L(\theta_1) - A_L^*(\theta_2)| : (\theta_1, \theta_2) \in \Theta_0, d(\theta_1, \theta_2) < c\big\} < \varepsilon.$$

With such a c, we have that

$$\begin{aligned} &P\big[\,|A_L(\hat{\theta}_L) - A_L^*| \geq \varepsilon\,\big] = P\big[\,|A_L(\hat{\theta}_L) - A_L(\theta_L^*)| \geq \varepsilon\,\big] \\ &\leq P\big[\,d(\hat{\theta}_L, \theta_L^*) \geq c \text{ and } \hat{\theta}_L \in \Theta_0\,\big] + P\big[\,\hat{\theta}_L \notin \Theta_0\,\big] \\ &\leq P\big[\,d(\hat{\theta}_L, \theta_L^*) \geq c\big] + P\big[\,\hat{\theta}_L \notin \Theta_0\,\big]. \end{aligned} \tag{B.6}$$

On the right-hand side of this inequality, the first term converges to zero by the consistency of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ for $\{\theta^*\}_{L\in\mathbb{N}}$, while the second term converges to zero as we have already verified above. Thus, the left-hand side of (B.6) converges to zero as $L \to \infty$. Because ε was chosen arbitrarily, this establishes that the second term on the right-hand side of (B.4) converges to zero in probability-P and completes the proof. □

Proof of THEOREM 2.1: Because $\{LN_{Lh}/N_L : (h,L) \in \mathbb{H}\}$ is bounded under Assumption 2.1, it is straightforward to verify that $\{\tilde{B}_{Lh} : (h,L) \in \mathbb{H}\}$ defined in (2.2) is SLB($1+\delta$) on the compact set Θ_0 under Assumption 2.9. Thus, application of Lemma B.1 to $\{\tilde{B}_{Lh} : (h,L) \in \mathbb{H}\}$ establishes that

$$\sup_{\theta\in\Theta_0} |\tilde{B}_L(\cdot,\theta) - \mathrm{E}[\tilde{B}_L(\cdot,\theta)]| = \sup_{\theta\in\Theta_0}\left|L^{-1}\sum_{h=1}^{L}\tilde{B}_{Lh}(\cdot,\theta) - L^{-1}\sum_{h=1}^{L}\mathrm{E}[\tilde{B}_{Lh}(\cdot,\theta)]\right| \to 0 \quad \text{in probability-}P \text{ as } L \to \infty. \tag{B.7}$$

If condition (a) of Assumption 2.10 holds, we have that $\mathrm{E}[\tilde{B}_L(\cdot,\theta)]$ coincides with $B_L(\theta)$, so that

$$\sup_{\theta\in\Theta_0}\left|\tilde{B}_L(\cdot,\theta)-B_L(\theta)\right| \to 0 \text{ in probability-}P \text{ as } L\to\infty.$$

If, instead, condition (b) holds, it follows from (2.3) that

$$\left|\mathrm{E}[\tilde{B}_L(\cdot,\theta)]-B_L(\theta)\right| \leq L^{-1}\sum_{h=1}^{L}(LN_{Lh}/N_L)^2 N_{Lh}^{-2}\sum_{k=1}^{N_{Lh}}\left|\mathrm{var}\left[\nabla q(\tilde{X}_{hk},\theta)\right]\right|.$$

Because

$$\Delta_1 \equiv \sup\left\{\left|\mathrm{var}[\nabla q(\tilde{X}_{hk},\theta)]\right| : \theta\in\Theta_0,\, (k,h)\in\mathbb{N}^2\right\} < \infty$$

(see (2.4)) and $\Delta_2 \equiv \sup\{LN_{Lh}/N_L : (h,L)\in\mathbb{H}\} < \infty$, it follows that

$$\sup_{\theta\in\Theta_0}\left|\mathrm{E}[\tilde{B}_L(\cdot,\theta)]-B_L(\theta)\right| \leq \Delta_1\Delta_2 L^{-1}\sum_{h=1}^{L}N_{Lh}^{-1} \leq \Delta_1\Delta_2\Big(\inf_{h\in\{1,\dots,L\}}N_{Lh}\Big)^{-1} \to 0 \\ \text{as } L\to\infty. \tag{B.8}$$

It follows from (B.7) and (B.8) that

$$\sup_{\theta\in\Theta_0}\left|\tilde{B}_L(\cdot,\theta)-B_L(\theta)\right| \leq \sup_{\theta\in\Theta_0}\left|\tilde{B}_L(\cdot,\theta)-\mathrm{E}[\tilde{B}_L(\cdot,\theta)]\right| + \sup_{\theta\in\Theta_0}\left|\mathrm{E}[\tilde{B}_L(\cdot,\theta)]-B_L(\theta)\right| \\ \to 0 \text{ in probability-}P \text{ as } L\to\infty.$$

Thus, $\{\tilde{B}_L\}_{L\in\mathbb{N}}$ is uniformly consistent for $\{B_L\}_{L\in\mathbb{N}}$ over Θ_0 under both of the conditions in Assumption 2.10. Given the uniform consistency of $\{\tilde{B}_L\}$ over Θ_0 and the consistency $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ for $\{\theta_L^*\}_{L\in\mathbb{N}}$ (which is uniformly interior to Θ_0), we can verify that $\{\hat{B}_L = \tilde{B}_L(\hat{\theta}_L)\}_{L\in\mathbb{N}}$ is consistent for $\{B_L^* = B_L(\theta_L^*)\}_{L\in\mathbb{N}}$ in the same way as we proved Lemma B.2. □

Appendix C

C.1 Proof of the Results in Section 3.2

Proof of PROPOSITION 3.1: Because the second result follows from the first by the law of the iterated expectations, and the third and fourth results can be easily derived from the second result, we only prove the first result below. Let (h, L) be an arbitrary member in $\mathbb{H}$. Then we have that

$$N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}) = N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} C_{Lhk} \phi_{Lh}(\tilde{X}_{hk}). \tag{C.1}$$

For each $k \in \mathbb{N}$, C_{Lhk} has a bounded support by Assumption 3.2(c), and $\phi_{Lh}(\tilde{X}_{hk})$ has a finite absolute moment by hypothesis, so that $C_{Lhk}\phi_{Lh}(\tilde{X}_{hk})$ has a finite absolute moment. From this fact and the independence of C_{Lhk} and $\tilde{X}_{hk}$ (Assumption 3.2(b)), it follows by Fubini's theorem (Folland, 1984, Theorem 2.37, pp. 65–66) that for each $k \in \mathbb{N}$,

$$\mathrm{E}[C_{Lhk}\phi_{Lh}(\tilde{X}_{hk}) \,|\, \tilde{X}] = \mathrm{E}[C_{Lhk}]\, \phi_{Lh}(\tilde{X}_{hk}).$$

Also, w_{Lhk} is the reciprocal of $\mathrm{E}[C_{Lhk}]$ by definition. Taking the conditional expectation of both side in (C.1) given $\tilde{X}$ and applying these facts yields that

$$\begin{aligned} \mathrm{E}\left[N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}) \,\middle|\, \tilde{X} \right] &= N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \mathrm{E}[C_{Lhk}]\, \phi_{Lh}(\tilde{X}_{hk}) \\ &= N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} \phi_{Lh}(\tilde{X}_{hk}). \end{aligned}$$

Thus, the desired result follows. □

Proof of PROPOSITION 3.2: Because for each $(h,L) \in \mathbb{H}$ and each $\gamma \in \Gamma$,

$$N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}, \gamma) = N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} C_{Lhk} \phi_{Lh}(\tilde{X}_{hk}, \gamma),$$

we have that for each $(h,L) \in \mathbb{H}$ and each $\gamma \in \Gamma$,

$$\left\| N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}, \gamma) \right\|_a \le N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \left\| C_{Lhk} \phi_{Lh}(\tilde{X}_{hk}, \gamma) \right\|_a.$$

Given the independence between $\{C_{Lhk} : (k,h,L) \in \mathbb{K}\}$ and $\{\tilde{X}_{hk} : (k,h) \in \mathbb{N}^2\}$ (Assumption 3.2(b)), it follows by Fubini's theorem (Folland, 1984, Theorem 2.37, pp. 65–66) that for each $(h,L) \in \mathbb{H}$ and each $\gamma \in \Gamma$,

$$\left\| N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}, \gamma) \right\|_a \le N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \|C_{Lhk}\|_a \cdot \|\phi_{Lh}(\tilde{X}_{hk}, \gamma)\|_a.$$

Under Assumption 3.2(b)–(d), we have that for each $k \in \mathbb{N}$ and each $(h,L) \in \mathbb{H}$,

$$w_{Lhk} = \mathrm{E}[C_{Lhk}]^{-1} \le \big(0 \cdot P[C_{Lhk} = 0] + 1 \cdot P[C_{Lhk} > 0]\big)^{-1} = P[C_{Lhk} > 0]^{-1} \le (N_{Lh}^{-1} \bar{p}_l)^{-1}$$

and

$$\|C_{Lhk}\|_a \cdot \le \bar{C} P[C_{Lhk} > 0] \le N_{Lh}^{-1} \bar{C} \bar{p}_u.$$

Also, $\Delta \equiv \sup_{\gamma \in \Gamma} \sup_{(k,h,L) \in \mathbb{K}} \|\phi_{Lh}(\tilde{X}_{hk})\|_a < \infty$ by hypothesis. It follows that

$$\begin{aligned} \left\| N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \phi_{Lh}(X_{Lhij}) \right\|_a &\le (\bar{p}_u / \bar{p}_l) \bar{C} \cdot N_{Lh}^{-1} \sum_{k=1}^{N_{Lh}} w_{Lhk} \|\phi_{Lh}(\tilde{X}_{hk}, \gamma)\|_a \\ &\le (\bar{p}_u / \bar{p}_l) \bar{C} \Delta, \quad (h,L) \in \mathbb{H}. \end{aligned}$$

Because the right-hand side of the above inequality depends on neither h nor L, the desired therefore result follows. □

C.2 Proof of the Results in Section 3.3

Proof of LEMMA 3.1: Suppose that Assumptions 3.1 and 3.2 hold. Let Γ be a finite-dimensional Euclidean space, a a positive real number, and $\{\phi_h\}_{h\in\mathbb{N}}$ a sequence of measurable functions from $(\mathbb{R}^v\times\Gamma,\mathscr{B}^v\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$ that is LB(a) on Γ. Then the array of functions from $(\mathbb{R}^v\times\Gamma,\mathscr{B}^v\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$, $\{\Phi_{Lh}\}_{L\in\mathbb{N}}$, defined by

$$\Phi_{Lh}(\omega,\gamma)\equiv N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}(\omega)\,\phi_h(X_{Lhij}(\omega),\gamma),\quad \omega\in\Omega,\,\gamma\in\Gamma,\,(h,L)\in\mathbb{H}$$

is SLB(a) on Γ. □

In proving Theorem 3.2, we employ a uniform law of large numbers stated in the next lemma, in which the SLB property takes an important role.

Lemma C.1. *Given Assumptions 3.1 and 3.2, let Γ be a finite-dimensional Euclidean space, and $\{F_{Lh}:(h,L)\in\mathbb{H}\}$ an array of measurable functions from $(\Omega\times\Gamma,\mathscr{F}\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$ such that for each $\gamma\in\Gamma$ and each $L\in\mathbb{N}$, $F_{L1}(\cdot,\gamma)$, ..., $F_{LL}(\cdot,\gamma)$ are independent, and $\{F_{Lh}:(h,L)\in\mathbb{H}\}$ is SLB($1+\delta$) on Γ for some $\delta\in(0,\infty)$. Then $\{\gamma\mapsto L^{-1}\sum_{h=1}^{L}\mathrm{E}[F_{Lh}(\cdot,\gamma)]:\Gamma\to\mathbb{R}\}_{L\in\mathbb{N}}$ is uniformly bounded and uniformly equicontinuous, and $\{|L^{-1}\sum_{h=1}^{L}F_{Lh}(\cdot,\gamma)-L^{-1}\sum_{h=1}^{L}\mathrm{E}[F_{Lh}(\cdot,\gamma)]|\}$ converges in probability-P to zero uniformly in $\gamma\in\Gamma$.*

Proof of LEMMA C.1: The result can be established essentially in the same way as Lemma A.3 of **?**. □

Also, the next result is useful in proving Theorem 3.2.

Lemma C.2. *Let r_1 and r_2 be positive real numbers and Γ a finite-dimensional Euclidean space. Write $r_0\equiv\min\{r_1,r_2\}$. Under Assumptions 3.1 and 3.2:*

(a) If measurable functions ϕ_1 and ϕ_2 from $(\mathbb{R}\times\Gamma,\mathscr{B}^v\otimes\mathscr{B}(\Gamma))$ to $(\mathbb{R}^{l_1\times l_2},\mathscr{B}^{l_1\times l_2})$ are LB(r_1) and LB(r_2) on Γ, respectively, then $\phi_1+\phi_2$ and $\phi_1\phi_2$ are LB(r_0) and LB($r_0/2$) on Γ, respectively.

(b) The claim of (a) holds even if each occurrence of LB in the claim is replaced with LBP.

(c) *If arrays* $\{F^1_{Lh} : (h,L) \in \mathbb{H}\}$ *and* $\{F^2_{Lh} : (h,L) \in \mathbb{H}\}$ *of measurable functions from* $(\Omega \times \Gamma, \mathscr{F} \otimes \mathscr{B}(\Gamma))$ *to* $(\mathbb{R}^{l_1 \times l_2}, \mathscr{B}^{l_1 \times l_2})$ *are SLB(*r_1*) and SLB(*r_2*) on* Γ*, respectively, then* $\{F^1_{Lh} + F^2_{Lh} : (h,L) \in \mathbb{H}\}$ *and* $\{F^1_{Lh} F^2_{Lh} : (h,L) \in \mathbb{H}\}$ *are SLB(*r_0*) and SLB(*$r_0/2$*) on* Γ*, respectively.*

(d) *The claim of (c) holds even if each occurrence of SLB in the claim is replaced with SLBP.*

Proof of LEMMA C.2: The proof is essentially the same as the one of (**?**, Lemma A.2). □

We now prove Theorem 3.2.

Proof of THEOREM 3.2: Because $\{\Lambda_L\}_{L\in\mathbb{N}}$ is bounded, there exists a compact subset $\mathbb{A}$ of $\mathbb{S}^q$, to which $\{\Lambda_L\}_{L\in\mathbb{N}}$ is uniformly interior. Because $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$, $\{\hat{\pi}_L\}_{L\in\mathbb{N}}$, and $\{\hat{\Lambda}_L\}_{L\in\mathbb{N}}$ are consistent for $\{\theta^*_L\}_{L\in\mathbb{N}}$, $\{\pi^*_L\}_{L\in\mathbb{N}}$, and $\{\Lambda_L\}_{L\in\mathbb{N}}$, respectively, it suffices to prove that $\{\check{\alpha}_L\}_{L\in\mathbb{N}}$ is consistent for $\{\alpha_L\}_{L\in\mathbb{N}}$ uniformly on $\Theta_0 \times \Pi_0 \times \mathbb{A}$ and that $\{\alpha_L\}$ is uniformly continuous on $\Theta_0 \times \Pi_0$.

As demonstrated in Section 3.3, $\check{\alpha}_L(\cdot,\theta,\pi)$ is unbiased for $\alpha_L(\theta,\pi)$ for each $(\theta,\pi) \in \Theta_0 \times \Pi_0$, if $\tilde{X}$ is degenerate. Suppose that $\tilde{X}$ is not degenerate. Let Δ denote the left-hand side of (3.11). Then we have that

$$0 \le \mathrm{tr}(\mathrm{var}[m_h(\tilde{X}_{hk},\theta,\pi)]) \le q\Delta, \quad (k,h) \in \mathbb{N}^2.$$

As demonstrated in Section 3.3, we also have that

$$\mathrm{E}[\check{\alpha}_L(\cdot,\theta,\pi)] - \alpha_L(\theta,\pi) = \sum_{h=1}^{L} \beta_{Lh} N_{Lh}^{-2} \sum_{k=1}^{N_{Lh}} \mathrm{tr}(\mathrm{var}[m_h(\tilde{X}_{hk},\theta,\pi)]), \quad (h,L) \in \mathbb{H}.$$

It follows that

$$\mathrm{E}[\check{\alpha}_L(\cdot,\theta,\pi)] - \alpha_L(\theta,\pi) \le q\Delta \sum_{h=1}^{L} \beta_{Lh} N_{Lh}^{-1}, \quad (\theta,\pi) \in \Theta_0 \times \Pi_0, L \in \mathbb{N}.$$

The right-hand side of this inequality, which does not depend on θ and π, converges to zero by Assumptions 3.5 and 3.8. It follows that

$$\sup_{(\theta,\pi)\in\Theta_0\times\Pi_0} \left|\mathrm{E}[\breve{\alpha}_L(\cdot,\theta,\pi)] - \alpha_L(\theta,\pi)\right| \to 0.$$

Thus, to establish the result of Theorem 3.2, it suffices to prove that

$$\sup_{(\theta,\pi)\in\Theta_0\times\Pi_0} \left|\breve{\alpha}_L(\cdot,\theta,\pi) - \mathrm{E}[\breve{\alpha}_L(\cdot,\theta,\pi)]\right| \to 0 \text{ in probability-}P \tag{C.2}$$

and $\{(\theta,\pi) \mapsto \mathrm{E}[\breve{\alpha}_L(\cdot,\theta,\pi)] : \Theta\times\Pi \to \mathbb{R}\}_{L\in\mathbb{N}}$ is uniformly continuous on $\Theta_0\times\Pi_0$.

Define an array $\{F_{Lh} : \Omega\times\Theta\times\Pi \to \mathbb{R}^{q\times q} : (h,L)\in\mathbb{H}\}$ by

$$F_{Lh}(\omega,\theta,\pi) \equiv L\beta_{Lh}\,\mathrm{tr}\Bigg(\Lambda\Bigg(\tilde{m}_{Lh}(\omega,\theta,\pi)\tilde{m}_{Lh}(\omega,\theta,\pi)' - N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\omega,\theta,\pi)\Bigg)\Bigg).$$

Under Assumptions 3.1 and 3.2, $F_{L1}(\cdot,\theta,\pi)$, ..., $F_{LL}(\cdot,\theta,\pi)$ are independent for each $(\theta,\pi)\in\Theta_0\times\Pi_0$ and each $L\in\mathbb{N}$. Because $\{L\beta_{Lh} : (h,L)\in\mathbb{H}\}$ is bounded (Assumption 3.5), it also follows by Lemma C.2 from Assumption 3.6(c) and 3.7 that $\{F_{Lh} : (h,L)\in\mathbb{H}\}$ is SLB$(1+\delta)$. Applying Lemma C.1 to $\{F_{Lh} : (h,L)\in\mathbb{H}\}$ establishes the desired result, because

$$\breve{\alpha}_L(\cdot,\theta,\pi) = L^{-1}\sum_{h=1}^{L} F_{Lh}(\cdot,\theta,\pi), \quad (\theta,\pi)\in\Theta_0\times\Pi_0,\,(h,L)\in\mathbb{H}.$$

□

We now turn to the proof of Theorem 3.3. Our proof uses the following double-array central limit theorem. To establish the asymptotic normality of $\{\hat{\alpha}_L\}_{L\in\mathbb{N}}$, we first derive the asymptotic distribution of $\{L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)\}_{L\in\mathbb{N}}$. and then show that $\{L^{1/2}\hat{\alpha}_L - L^{1/2}(\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)\}_{L\in\mathbb{N}}$ converges in probability-P to zero. In establishing the first result, we employ the following double-array central limit theorem.

Lemma C.3. *Let $\{U_{Lh} : (h,L)\in\mathbb{H}\}$ be an array of uniformly $\mathscr{L}_{2+\delta}$-bounded, zero-mean $v\times 1$ random vectors for some $\delta\in[0,\infty)$ such that for each $L\in\mathbb{N}$, U_{L1},..., U_{LL} are independent. Then:*

(a) $L^{-1/2}\sum_{h=1}^{L} U_{Lh} = \mathrm{O}_P(1)$.

(b) *Suppose in addition that $\delta > 0$, and $\{V_L \equiv L^{-1}\sum_{h=1}^{L} \mathrm{var}[U_{Lh}]\}_{L\in\mathbb{N}}$ is uniformly positive. Then $V_L^{-1/2}L^{-1/2}\sum_{h=1}^{L} U_{Lh} \overset{A}{\sim} \mathrm{N}(0,1)$.*

Proof of LEMMA C.3: See Lemma A.4 of **?**. □

Write

$$
\begin{aligned}
G_{2L}(\theta,\pi,\Lambda) &\equiv \mathrm{E}[\nabla_\pi \breve{\alpha}_L(\cdot,\theta,\pi,\Lambda)] \\
&= \sum_{h=1}^{L} \beta_{Lh}\Bigg(2N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} \mathrm{E}\big[W_{Lhij}\nabla_\pi m_h(X_{Lhij},\theta,\pi)\Lambda\tilde{m}_{Lh}(\cdot,\theta,\pi)\big] \\
&\qquad - N_{Lh}^{-2}\,\mathrm{E}\big[\nabla_\pi(\mathrm{vec}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi))\big]\,\mathrm{vec}\Lambda\Bigg), \\
G_L(\theta,\pi,\Lambda) &\equiv (G_{1L}(\theta,\pi,\Lambda)', G_{2L}(\theta,\pi,\Lambda)')', \quad G_L^* \equiv G_L(\theta^*,\pi_L^*,\Lambda_L), \\
&\qquad (\theta,\pi)\in\Theta\times\Pi,\ \Lambda\in\mathbb{S}^q,\ L\in\mathbb{N}.
\end{aligned}
$$

Lemma C.4. *(a) Suppose that Assumptions 3.1–3.3, 3.5, 3.6(a)–(c), 3.7, 3.8, and 3.10(c), (d) hold. Let $\mathbb{\Lambda}$ be an arbitrary nonempty compact subset of $\mathbb{S}^q$. Then $\{G_L : \Theta_0\times\Pi_0\times\mathbb{\Lambda}\to\mathbb{R}^{q+r}\}_{L\in\mathbb{N}}$ is uniformly bounded and uniformly equicontinuous, and $\{\nabla\breve{\alpha}_L(\cdot,\theta,\pi,\Lambda) - G_L(\theta,\pi,\Lambda)\}_{L\in\mathbb{N}}$ converges in probability-P to zero uniformly in $(\theta,\pi,\Lambda)\in\Theta_0\times\Pi_0\times\mathbb{\Lambda}$. If, in addition, (3.3) holds, then for each $L\in\mathbb{N}$ and each $\Lambda\in\mathbb{\Lambda}$, $G_{2L}^* \equiv G_2(\theta_L^*,\pi_L^*,\Lambda_L) = 0$.*

(b) *Suppose that Assumptions 3.1–3.8 and 3.10(a)–(f), (h) hold. If (3.3) holds, then*

$$
L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) = L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^* + \mathrm{o}_P(1) = \mathrm{O}_P(1). \tag{C.3}
$$

If, in addition, Assumption 3.10(g) holds, then $V_L^{-1/2}L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) \overset{A}{\sim}$ $\mathrm{N}(0,1)$.

Proof of LEMMA C.4: To prove (i), note that

$$
\nabla\breve{\alpha}_L(\cdot,\theta,\pi,\Lambda) = L^{-1}\sum_{h=1}^{L}\breve{g}_{Lh}(\cdot,\theta,\pi,\Lambda),
$$

where the array $\{\breve{g}_{Lh} : \Omega \times \Theta \times \Pi \times \mathbb{S}^q \to \mathbb{R}^{q+r} : (h,L) \in \mathbb{H}\}$ is defined by

$$\begin{aligned}\breve{g}_{Lh}(\cdot,\theta,\pi,\Lambda) =& L\beta_{Lh}\Bigg(2N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} W_{Lhij}\nabla_\theta m_h(X_{Lhij},\theta,\pi)\Lambda\tilde{m}_{Lh}(\cdot,\theta,\pi) \\ &- N_{Lh}^{-2}\nabla_\theta(\mathrm{vec}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi))\,\mathrm{vec}\Lambda\Bigg),\ (\theta,\pi) \in \Theta \times \Pi, \Lambda \in \mathbb{S}^q, L \in \mathbb{N}.\end{aligned}$$

By using Lemmas 3.1 and C.2, we can easily show that $\{\breve{g}_{Lh} : (h,L) \in \mathbb{H}\}$ is SLB($1+\delta$) on $\Theta_0 \times \Pi_0 \times \mathbb{A}$. Application of Lemma C.1 to $\{\breve{g}_{Lh}\}$ thus establishes the first result. For the second result, note that given each fixed $\Lambda \in \mathbb{S}^q$, each component of $\nabla\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)$ is bounded by an $\mathscr{L}_{1+\delta}$-bounded random variable in a neighborhood of (θ_L^*, π_L^*), so that

$$G_L^* = \mathrm{E}[\nabla\check{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L)] = \nabla\mathrm{E}[\check{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L)] = \nabla\alpha(\theta_L^*,\pi_L^*,\Lambda_L),\ \Lambda \in \mathbb{A}, L \in \mathbb{N}.$$

Because $\alpha(\theta_L^*,\pi,\Lambda_L) = 0$ for each $\pi \in \Pi$ under (C.3), it follows that

$$G_{2L}^* = \nabla_\pi\alpha(\theta_L^*,\pi_L^*,\Lambda_L) = 0, \quad L \in \mathbb{N}.$$

For (ii), note that there exists a real number $\varepsilon > 0$ such that for each $L \in \mathbb{N}$, the open ball in Θ with radius ε centered at θ_L^* is contained in $\mathrm{int}\,\Theta_0$, and the open ball in Π with radius ε centered at π_L^* is contained in $\mathrm{int}\,\Pi_0$, because $\{\theta_L^*\}$ and $\{\pi_L^*\}$ are uniformly interior to Θ_0 and Π_0, respectively. By the mean value theorem for random functions (**?**, Lemma 3), there exists sequences of random vectors $\{\ddot{\theta}_L : \Omega \to \Theta\}_{L\in\mathbb{N}}$ and $\{\ddot{\pi}_L : \Omega \to \Pi\}_{L\in\mathbb{N}}$ such that for each $L \in \mathbb{N}$, $\ddot{\theta}_L$ is on the line segment connecting $\hat{\theta}_L$ and θ_L^*, $\ddot{\pi}_L$ is on the line segment connecting $\hat{\pi}_L$ and π_L^*, and

$$\begin{aligned}&\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) - \check{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L) \\ &\quad = \nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L,\ddot{\pi}_L,\Lambda_L)'(\hat{\theta}_L - \theta_L^*) + \nabla_\pi\check{\alpha}_L(\cdot,\ddot{\theta}_L,\ddot{\pi}_L,\Lambda_L)'(\hat{\pi}_L - \pi_L^*),\end{aligned}$$

whenever $|\hat{\theta}_L - \theta_L^*| < \varepsilon$ and $|\hat{\pi}_L - \pi_L^*| < \varepsilon$ (where ε is as described above). By the result of (i) established above, $\{\check{G}_L(\cdot,\theta,\pi,\Lambda_L) - G_L(\theta,\pi,\Lambda_L)\}_{L\in\mathbb{N}}$ converges to zero uniformly in $(\theta,\pi) \in \Theta_0 \times \Pi_0$ in probability-P, and $\{G_L(\cdot,\cdot,\Lambda_L) : \Theta \times \Pi \to$

$\mathbb{R}^p\}_{L\in\mathbb{N}}$ is equicontinuous uniformly on $\Theta_0\times\Pi_0$. Also, we have that

$$|\ddot{\theta}_L-\theta_L^*|\leq|\hat{\theta}_L-\theta_L^*|=\mathrm{O}_P(L^{-1/2})$$

by Assumption 3.10(a) and Lemma C.3, and that

$$|\ddot{\pi}_L-\pi_L^*|\leq|\hat{\pi}_L-\pi_L^*|=\mathrm{O}_P(L^{-1/2})$$

by Assumption 3.10(b). It follows that $|\nabla\breve{\alpha}_L(\cdot,\ddot{\theta}_L,\ddot{\pi}_L,\Lambda_L)-G_L^*|=\mathrm{o}_P(1)$, and

$$\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)-\breve{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L)=G_{1L}^*(\hat{\theta}_L-\theta_L^*)+G_{2L}^*(\hat{\pi}_L-\pi_L^*)+\mathrm{o}_P(L^{-1/2}).$$

Applying Assumption 3.10(a) along with the second result of (i) of the current lemma in this equality establishes that

$$\begin{aligned}&\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)-\breve{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L)\\&=G_{1L}^*L^{-1}\sum_{h=1}^{L}(LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\psi_{Lhi}^*+\mathrm{o}_P(L^{-1/2}).\end{aligned}$$

Substituting this into the right-hand side of the identity:

$$\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)=(\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)-\breve{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L))+\breve{\alpha}_L(\cdot,\theta_L^*,\pi_L^*,\Lambda_L),\ L\in\mathbb{N},$$

applying the definition of $\breve{\alpha}_L$, and then rewriting the resulting expression using ξ_{Lh}^* yields the first equality in (C.3).

To show the second equality in (C.3), we rewrite the first equality in (C.3) as

$$L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)=L^{-1/2}\sum_{h=1}^{L}(\xi_{Lh}^*-\mathrm{E}[\xi_{Lh}^*])+L^{-1/2}\sum_{h=1}^{L}\mathrm{E}[\xi_{Lh}^*],\quad L\in\mathbb{N}.\quad\text{(C.4)}$$

Using Proposition 3.2 and Lemmas 3.1 and C.2, we can verify that $\{\xi_{Lh}^*\}$ is $\mathscr{L}_{2+2\delta}$-bounded array. It follows by Lemma C.3(a) that the first term on the right-hand side of (C.4) is $\mathrm{O}_P(1)$. Because $\{\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\psi_{Lhij}^*:(h,L)\in\mathbb{H}\}$ is a zero-mean array (Assumption 3.10(a)), the second term is equal to

$$L^{-1/2}\sum_{h=1}^{L} \mathrm{E}\left[L\beta_{Lh}\mathrm{tr}\left(\Lambda_L\left(\tilde{m}^*_{Lh}\tilde{m}^*_{Lh}{}' - N_{Lh}^{-2}\tilde{\Sigma}^*_{Lh}\right)\right)\right] = L^{1/2}\mathrm{E}[\check{\alpha}_L(\cdot,\theta^*_L,\pi^*_L)].$$

Note that $\mathrm{E}[\check{\alpha}_L(\cdot,\theta^*_L,\pi^*_L)]$ is the bias of $\check{\alpha}_L(\cdot,\theta^*_L,\pi^*_L)$ for α^*_L (which is zero by hypothesis). As discussed in Section 3.3, the bias of $\check{\alpha}_L(\cdot,\theta^*_L,\pi^*_L)$ for α^*_L $(=0)$ is $O(1/\min\{N_{L1},\dots,N_{LL}\})$, so that it is $o(1/L^{1/2})$ under Assumption 3.10 (h). Thus, the second term on the right-hand side of (C.4) is $o(1)$. This establishes the second equality in (C.3).

To verify the last claim of the current lemma, we substitute (C.3) into the right-hand side of (C.3) and multiply the resulting equality by $V_L^{-1/2}$ to obtain that

$$V_L^{-1/2}L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) = V_L^{-1/2}L^{-1/2}\sum_{h=1}^{L}(\xi^*_{Lh}-\mathrm{E}[\xi^*_{Lh}]) + \mathrm{o}_P(1). \tag{C.5}$$

By applying Lemma C.3(b) to $\mathscr{L}_{2+\delta}$ bounded, zero-mean array $\{\xi^*_{Lh}-\mathrm{E}[\xi^*_{Lh}]' : (h,L)\in\mathbb{H}\}$ and applying the asymptotic equivalence lemma delivers the desired result. □

To show the asymptotic equivalence between $\{L^{1/2}\hat{\alpha}_L\}_{L\in\mathbb{N}}$ and $\{L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)\}_{L\in\mathbb{N}}$, we use the following fact.

Lemma C.5. *Suppose that Assumptions 3.1–3.5, 3.6 (a)–(c), 3.7, and 3.10(a)–(f), (h) hold. If (3.3) holds, then*

$$\sum_{h=1}^{L}\beta_{Lh}\left(\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)' - N_{Lh}^{-2}\hat{\Sigma}_{Lh}\right) = \mathrm{O}_P(L^{-1/2}). \tag{C.6}$$

Proof of LEMMA C.5: For each $(i,j)\in\{1,2,\dots,q\}^2$, let E_{ij} denote the $q\times q$ matrix, all of whose elements are zeros except the (i,j)-element set equal to one. Then for each $i\in\{1,2,\dots,q\}$, E_{ii} is a symmetric matrix, and it holds that $\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,E_{ii})$ is equal to the ith diagonal element of the random matrix in question. It follows by Lemma C.4(ii) that each diagonal element of the random matrix is $\mathrm{O}_P(L^{-1/2})$. Also, for each distinct i and j in $\{1,2,\dots,q\}$, $E_{ij}+E_{ji}$ is a symmetric matrix, and it holds that $\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,E_{ij}+E_{ji})$ is equal to two times the (i,j)-element of the random matrix in question. Application Lemma C.4(ii), verifies that each off-diagonal element of the random matrix is $\mathrm{O}_P(L^{-1/2})$. □

We now show the asymptotic equivalence between $\{L^{1/2}\hat{\alpha}_L\}_{L\in\mathbb{N}}$ and $\{L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L)\}_{L\in\mathbb{N}}$.

Lemma C.6. *Suppose that 3.1–3.7, 3.9 and 3.10(a)–(f), (h) hold. If (3.3) holds, then* $L^{1/2}\hat{\alpha}_L - L^{1/2}\check{\alpha}(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) = \mathrm{o}_P(1)$.

Proof of LEMMA C.6: It follows from the definitions of $\{\hat{\alpha}_L\}_{L\in\mathbb{N}}$ and $\{\check{\alpha}_L\}_{L\in\mathbb{N}}$ that for each $L\in\mathbb{N}$,

$$
\begin{aligned}
&L^{1/2}(\hat{\alpha}_L - \check{\alpha}(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L))\\
&\qquad = \mathrm{tr}\Big((\hat{\Lambda}_L-\Lambda_L)L^{1/2}\sum_{h=1}^{L}\beta_{Lh}\Big(\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)\tilde{m}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L)' - N_{Lh}^{-2}\hat{\Sigma}_{Lh}\Big)\Big).
\end{aligned}
$$

The right-hand side of this equality converges in probability-P to zero by Assumption 3.9 and Lemma C.5. □

Proof of THEOREM 3.3: By Lemma C.6, we have that

$$
\begin{aligned}
L^{1/2}\hat{\alpha}_L &= L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) + L^{1/2}(\hat{\alpha}_L - \check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L))\\
&= L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\hat{\pi}_L,\Lambda_L) + \mathrm{o}_P(1). \qquad \text{(C.7)}
\end{aligned}
$$

The equality (3.13) follows from this equality. The asymptotic normality results also follows from (3.13) and Lemma C.4(ii) by the asymptotic equivalence lemma. □

For convenience in proving Theorem 3.4, define $\{\check{\xi}_{Lh} : \Omega\times\Theta\times\Pi\times\mathbb{S}^q\times\mathbb{R}^{p\times p}\times\mathbb{R}^q \to \mathbb{R} : (h,L)\in\mathbb{H}\}$ by

$$
\begin{aligned}
\check{\xi}_{Lh}(\cdot,\theta,\pi,\Lambda,J,G_1) &\equiv L\beta_{Lh}\mathrm{tr}\Big(\Lambda\Big(\tilde{m}_{Lh}(\cdot,\theta,\pi)\tilde{m}_{Lh}(\cdot,\theta,\pi)' - N_{Lh}^{-2}\tilde{\Sigma}_{Lh}(\cdot,\theta,\pi)\Big)\Big)\\
&\quad + G_1'J(LN_{Lh}/N_L)\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\psi_{Lh}(X_{Lhij},\theta),\\
&\qquad (\theta,\pi,\Lambda,J,G_1)\in\Theta\times\Pi\times\mathbb{S}^q\times\mathbb{R}^{p\times p}\times\mathbb{R}^q,\ (h,L)\in\mathbb{H},
\end{aligned}
$$

with which we have that for each $(h,L)\in\mathbb{H}$, $\xi^*_{Lh} = \check{\xi}_{Lh}(\cdot,\theta^*_L,\pi^*_L,\Lambda_L,J^*_L,G^*_{L1})$.

Proof of THEOREM 3.4: Because $\{\Lambda_L\}_{L\in\mathbb{N}}$ is bounded, there exists a compact subset $\mathbb{A}$ of $\mathbb{S}^q$ to which $\{\Lambda_L\}$ is uniformly interior. Analogously, there exists a compact subset $\mathbb{J}$ of $\mathbb{R}^{p\times p}$ and a compact subset $\mathbb{G}_1$ of $\mathbb{R}^{p\times 1}$ such that $\{J_L^*\}_{L\in\mathbb{N}}$ and $\{G_{1L}^*\}_{L\in\mathbb{N}}$ are uniformly interior to $\mathbb{J}$ and $\mathbb{G}_1$, respectively. Application of Lemma C.2 verifies that $\{\check{\xi}_{Lh}^2\}$ is SLB$(1+\delta)$ on $\Theta_0\times\Pi_0\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1$. It follows by Lemma C.1 that

$$\{(\theta,\pi,\Lambda,J,G_1)\mapsto \mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta,\pi,\Lambda,J,G_1)^2]:\Theta_0\times\Pi_0\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1\to\mathbb{R},\,(h,L)\in\mathbb{H}\}$$

is uniformly bounded and uniformly equicontinuous, and that

$$\sup_{(\theta,\pi,\Lambda,J,G_1)\in\Theta_0\times\Pi_0\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1}\left|L^{-1}\sum_{h=1}^{L}\check{\xi}_{Lh}(\cdot,\theta,\pi,\Lambda,J,G_1)^2-L^{-1}\sum_{h=1}^{L}\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta,\pi,\Lambda,J,G_1)^2]\right|$$
$$\to 0 \text{ in probability-}P.$$

Because

$$\hat{V}_L=L^{-1}\sum_{h=1}^{L}\check{\xi}_{Lh}(\cdot,\hat{\theta}_L,\hat{\pi}_L,\hat{\Lambda}_L,\hat{J}_L,\hat{G}_{1L})^2,\quad L\in\mathbb{N},$$

it follows from these facts and the consistency of $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$, $\{\hat{\pi}_L\}_{L\in\mathbb{N}}$, $\{\hat{\Lambda}_L\}_{L\in\mathbb{N}}$, $\{\hat{J}_L\}_{L\in\mathbb{N}}$, and $\{\hat{G}_{1L}\}_{L\in\mathbb{N}}$ for $\{\theta_L^*\}_{L\in\mathbb{N}}$, $\{\pi_L^*\}_{L\in\mathbb{N}}$, $\{\Lambda_L\}_{L\in\mathbb{N}}$, $\{\bar{J}_L\}_{L\in\mathbb{N}}$, and $\{G_{1L}^*\}_{L\in\mathbb{N}}$ that

$$\hat{V}_L-L^{-1}\sum_{h=1}^{L}\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta_L^*,\pi_L^*,\Lambda_L,\bar{J}_L,G_{1L}^*)^2]\to 0 \text{ in probability-}P.$$

Also, we have that for each $(h,L)\in\mathbb{H}$,

$$\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta_L^*,\pi_L^*,\Lambda_L,\bar{J}_L,G_{1L}^*)]=L\beta_{Lh}\bar{m}_L^{*\prime}\Lambda_L\bar{m}_L^*+L\beta_{Lh}\mathrm{tr}\Big(\Lambda_L N_{Lh}^{-2}\sum_{k=1}^{N_{Lh}}\mathrm{var}[m_h(\tilde{X}_{hk},\theta,\pi)]\Big),$$

where the second term on the right-hand side is convergent to zero uniformly in h (see the discussion on the bias of $\tilde{\Sigma}_L$ in Section 3.3). It follows that

$$\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta_L^*,\pi_L^*,\Lambda_L,\bar{J}_L,G_{1L}^*)^2]=\bar{V}_L+o(1).$$

This verifies claim (i).

For claim (ii), note that

$$\mathscr{T}_L - \frac{L^{1/2}\hat{\alpha}_L}{V_L^{1/2}} = (\hat{V}_L^{-1/2} - V_L^{-1/2}) L^{1/2}\hat{\alpha}_L, \quad L \in \mathbb{N}.$$

Because $\bar{V}_L = V_L$ by hypothesis, it follows from (i) established above and Assumption 3.10(g) that $\hat{V}_L^{-1/2} - V_L^{-1/2} = \mathrm{o}_P(1)$. Also, we know that $L^{1/2}\hat{\alpha}_L = \mathrm{O}_P(1)$ under (3.3), as $\{V_L^{-1/2} L^{1/2}\hat{\alpha}_L\}_{L\in\mathbb{N}}$ is convergent in distribution (Theorem 3.3), and $\{V_L = \bar{V}_L\}_{L\in\mathbb{N}}$ is bounded as shown above. It follows that $\mathscr{T}_L = L^{1/2}\hat{\alpha}_L / V_L^{1/2} + \mathrm{o}_P(1)$. The desired result follows from this equality by the asymptotic equivalence lemma and Theorem 3.3.

To prove (iii), recall that $\{\hat{V}_L\}_{L\in\mathbb{N}}$ is consistent for $\{\bar{V}_L\}_{L\in\mathbb{N}}$ that is bounded (Theorem 3.3). Also, note that for each $L \in \mathbb{N}$,

$$\bar{V}_L \geq L^{-1}\sum_{h=1}^{L}(L\beta_{Lh}\bar{m}_{Lh}^{*}{}'\Lambda_L\bar{m}_{Lh}^{*})^2 + o(1) \geq \left(L^{-1}\sum_{h=1}^{L}L\beta_{Lh}\bar{m}_{Lh}^{*}{}'\Lambda_L\bar{m}_{Lh}^{*}\right)^2 + o(1) = \alpha_L^{*2} + o(1),$$

where the second inequality follows by Jensen's inequality. It follows that when $\{\alpha_L^*\}_{L\in\mathbb{N}}$ is uniformly positive, so is $\{\bar{V}_L\}_{L\in\mathbb{N}}$.

Given the consistency of $\{\hat{V}_L\}$ for the bounded and uniformly positive sequence $\{\bar{V}_L\}$ and the consistency $\{\hat{\alpha}_L\}_{L\in\mathbb{N}}$ for the bounded sequence $\{\alpha_L\}_{L\in\mathbb{N}}$ (Theorem 3.2), we have that

$$L^{-1/2}\mathscr{T}_L - \bar{V}_L^{-1/2}\alpha_L^* = \bar{V}_L^{-1/2}\hat{\alpha}_L - \bar{V}_L^{1/2}\alpha_L^* \to 0 \text{ in probability-}P \tag{C.8}$$

(Theorems 3.2 and 3.4(i)). Let c be an arbitrary real number. Then we have that for each $L \in \mathbb{N}$,

$$\begin{aligned} P[\mathscr{T}_L > c] &= P[L^{-1/2}\mathscr{T}_L - \bar{V}_L^{-1/2}\alpha_L^* > L^{-1/2}c - \bar{V}_L^{-1/2}\alpha_L^*] \\ &\geq P[L^{-1/2}\mathscr{T}_L - \bar{V}_L^{-1/2}\alpha_L^* > L^{-1/2}c - \tau], \end{aligned}$$

where $\tau \equiv \inf\{\bar{V}_L^{-1/2}\alpha_L^* : L \in \mathbb{N}\} > 0$, because $\{\alpha_L^*\}_{L\in\mathbb{N}}$ is assumed to be uniformly positive, and $\{\bar{V}_L\}_{L\in\mathbb{N}}$ is bounded. Because $L^{-1/2}c < \tau/2$ for almost all $L \in \mathbb{N}$, it holds that for almost all $L \in \mathbb{N}$,

$$P[\mathscr{T}_L > c] \geq P[L^{-1/2}\mathscr{T}_L - \bar{V}_L^{-1/2}\alpha_L^* > -\tau/2] \geq P[|L^{-1/2}\mathscr{T}_L - \bar{V}_L^{-1/2}\alpha_L^*| < \tau/2].$$

Because the right-hand side of this equality converges to one by (C.8), the desired result follows. □

Proof of PROPOSITION 3.5: Because $\{\Lambda_L\}_{L\in\mathbb{N}}$ is bounded by Assumption 3.6(d), there exists a positive real number c_1 that is no smaller than the maximum eigenvalue of Λ_L for each $L \in \mathbb{N}$. With this c_1, we have that

$$\alpha_L^* = \sum_{h=1}^{L} \beta_{Lh} \bar{m}_{Lh}^{*}{}' \Lambda_L \bar{m}_{Lh}^* \leq c_1 \sum_{h=1}^{L} \beta_{Lh} |\,\mathrm{E}[m_{Lh}^*]\,|^2, \quad L \in \mathbb{N}.$$

Claim (a) therefore follows. When, in addition, $\{\Lambda_L\}$ is uniformly positive definite, there exists a real number $c_2 > 0$ that is no larger than the minimum eigenvalue of Λ_L for each $L \in \mathbb{N}$, so that

$$\alpha_L^* \geq c_2 \sum \beta_{Lh} |\,\mathrm{E}[m_{Lh}^*]\,|^2, \quad L \in \mathbb{N}.$$

Claim (b) therefore follows. □

Proof of the Results in Section 3.4

Proof of THEOREM 3.1: Under Assumption a, $\{m_h(\tilde{X}_{hk}, \theta, \pi) : (\theta, \pi) \in \Theta_0 \times \Pi_0,\, (k,h) \in \mathbb{N}^2\}$ is uniformly $\mathscr{L}_{2+2\delta}$-bounded for some real $\delta > 0$. By Proposition 3.2 it follows that $\{\tilde{m}_{Lh}(\cdot, \theta, \pi) : (\theta, \pi) \in \Theta_0 \times \Pi_0,\, (h, L) \in \mathbb{H}\}$ is uniformly $\mathscr{L}_{2+2\delta}$-bounded, and so is $\{\tilde{m}_{Lh}(\cdot, \theta_L^*, \pi) : \pi \in \Pi_0,\, (h, L) \in \mathbb{H}\}$. Thus, $\{\bar{m}_{Lh}(\theta_L^*, \pi) : \pi \in \Pi_0,\, (h, L) \in \mathbb{H}\}$ is bounded. It follows from this fact and Assumptions 3.12a, b that $\{\alpha_L^*(\pi) : \pi \in \Pi_0, L \in \mathbb{N}\}$ is bounded.

For the consistency result, define $c \equiv \sup\{|\Lambda_L(\pi)| : \pi \in \Pi, L \in \mathbb{N}\} + 1$. Then c is finite by Assumption 3.12(b). Write $\mathbb{A} \equiv \{\Lambda \in \mathbb{S}^q : |\Lambda| \leq c\}$, and let ε be an arbitrary positive real number. Then we have that

$$P\left[\sup_{\pi\in\Pi}\left|\hat{\alpha}_L(\pi)-\alpha_L^*(\pi)\right|\geq\varepsilon\right] \tag{C.9}$$

$$\leq P\left[\sup_{\pi\in\Pi}\left|\hat{\alpha}_L(\pi)-\alpha_L^*(\pi)\right|\geq\varepsilon,\ \hat{\theta}_L\in\Theta_0,\ \text{and}\ \sup_{\pi\in\Pi}|\hat{\Lambda}_L(\pi)|\leq c\right]$$

$$+P[\hat{\theta}_L\notin\Theta_0]+P\left[\sup_{\pi\in\Pi}|\hat{\Lambda}_L(\pi)|>c\right],\quad L\in\mathbb{N}. \tag{C.10}$$

The second and third terms on the right-hand side of this equality converge to zero by Assumptions 3.4 and 3.12(a), (c). It thus suffices to prove that the first term also converges to zero to establish the desired result.

Note that for each $\pi\in\Pi$ and each $L\in\mathbb{N}$,

$$\begin{aligned}\hat{\alpha}_L(\pi)-\alpha_L^*(\pi)&=\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi))-\alpha_L(\theta_L^*,\pi,\Lambda_L(\pi))\\&=\big(\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi))-\alpha_L(\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi))\big)+\big(\alpha_L(\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi))-\alpha_L(\theta_L^*,\pi,\Lambda_L(\pi))\big).\end{aligned}$$

It follows that if $\hat{\theta}_L\in\Theta_0$ and $\sup_{\pi\in\Pi}|\hat{\Lambda}_L(\pi)|\leq c$,

$$\begin{aligned}\sup_{\pi\in\Pi}|\hat{\alpha}_L(\pi)-\alpha_L^*(\pi)|\leq&\sup_{(\theta,\pi,\Lambda)\in\Theta_0\times\Pi\times\mathbb{A}}|\check{\alpha}_L(\cdot,\theta,\pi,\Lambda)-\alpha_L(\theta,\pi,\Lambda)|\\&+\sup_{\pi\in\Pi}|\alpha_L(\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi))-\alpha_L(\theta_L^*,\pi,\Lambda_L(\pi))|,\quad L\in\mathbb{N}.\end{aligned} \tag{C.11}$$

The first term on the right-hand side of this equality converge in probability-P to zero, because $\{\check{\alpha}_L\}_{L\in\mathbb{N}}$ is uniformly consistent for $\{\alpha_L\}_{L\in\mathbb{N}}$ on $\Theta_0\times\Pi_0\times\mathbb{A}$ as verified in the proof of Theorem 3.2. In the second term, $\{\alpha_L\}_{L\in\mathbb{N}}$ is uniformly continuous on $\Theta_0\times\Pi_0\times\mathbb{A}$, as established in the proof of Theorem 3.2. It follows by Assumptions 3.4 and 3.12(a), (c) that the second term converges to zero in probability-P. Thus, given that the inequality (C.11) holds with a probability approaching to one, the first term on the right-hand side of (C.10) converges to zero, and the desired result follows. □

In establishing the claims in Theorem 3.3, we use the following functional central limit theorem.

Lemma C.7. *Given Assumptions 3.1 and 3.2, let* Γ *a nonempty compact subset of the r-dimensional Euclidean space, and* $\{U_{Lh} : (h,L) \in \mathbb{H}\}$*, a double array of measurable functions from* $(\Omega \times \Gamma, \mathscr{F}/\mathscr{B} \otimes \mathscr{B}(\Gamma))$ *to* $(\mathbb{R}, \mathscr{B})$ *such that* $\{U_{Lh}\}$ *is SLBP(*$2+\delta$*) on* Γ *for some* $\delta \in (0,\infty)$*, and for each* $(h,L) \in \mathbb{H}$ *and each* $\gamma \in \Gamma$*,* $\mathrm{E}[U_{Lh}(\cdot,\gamma)] = 0$*. Then:*

(a) The sequence $\{\sup_{\gamma\in\Gamma} |L^{-1/2} \sum_{h=1}^{L} U_{Lh}(\cdot,\gamma)\}_{L\in\mathbb{N}}$ *is uniformly* $\mathscr{L}_{2+\delta}$*-bounded; hence, it is* $\mathrm{O}_P(1)$*.*

(b)

$$\sup_{L\in\mathbb{N}} \mathrm{E}\Bigg[\sup\Bigg\{\Bigg|L^{-1/2}\sum_{h=1}^{L} U_{Lh}(\cdot,\gamma_2) - L^{-1/2}\sum_{h=1}^{L} U_{Lh}(\cdot,\gamma_1)\Bigg| \\ : |\gamma_1 - \gamma_2| < \kappa, (\gamma_1,\gamma_2) \in \Gamma^2\Bigg\}\Bigg] \to 0 \text{ as } \kappa \downarrow 0.$$

(c) The sequence of random functions $\{\gamma \mapsto L^{-1/2}\sum_{h=1}^{L} U_{Lh}(\cdot,\gamma)\}_{L\in\mathbb{N}}$ *is stochastically equicontinuous uniformly on* Γ*, i.e., for each pair of real numbers* $\varepsilon > 0$ *and* $\kappa > 0$*, there exists a real number* $\beta > 0$ *such that*

$$\limsup P\Bigg[\sup\Bigg\{\Bigg|L^{-1/2}\sum_{h=1}^{L} U_{Lh}(\cdot,\gamma_2) - L^{-1/2}\sum_{h=1}^{L} U_{Lh}(\cdot,\gamma_1)\Bigg| \\ : |\gamma_1 - \gamma_2| < \beta, (\gamma_1,\gamma_2) \in \Gamma^2\Bigg\} > \kappa\Bigg] < \varepsilon.$$

(d) Suppose in addition that there exists a function $K : \Gamma^2 \to \mathbb{R}$ *such that for each* $(\gamma_1,\gamma_2) \in \Gamma^2$

$$K_L(\gamma_1,\gamma_2) \equiv L^{-1}\sum_{h=1}^{L} \mathrm{cov}[U_{Lh}(\cdot,\gamma_1), U_{Lh}(\cdot,\gamma_2)] \to K(\gamma_1,\gamma_2).$$

Then the process $\{\gamma \mapsto L^{-1/2}\sum_{h=1}^{L} U_{Lh}(\cdot,\gamma)\}_{L\in\mathbb{N}}$ *converges in distribution to a zero-mean Gaussian process with covariance kernel K concentrated on* $\mathsf{U}(\Gamma)$*, the set of all uniformly continuous* $\mathbb{R}$*-valued functions.*

Proof of LEMMA C.7: The lemma can be proved by using (Pollard, 1990, Theorem 10.2). The detailed proof of Lemma C.7 is available from the author upon request. □

Proof of LEMMA 3.2: The result immediately follows from the proof of Lemma 3.1. □

To prove Theorem 3.3, we first derive the asymptotic linear representation of the sequence of random functions from Π to $\mathbb{R}$, $\{\pi \mapsto L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi))\}_{L\in\mathbb{N}}$ and then show that $\{\pi \mapsto L^{1/2}(\breve{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi)) - \hat{\alpha}_L(\pi))\}_{L\in\mathbb{N}}$ converges in probability-P to zero (function).

Lemma C.8. *Suppose that 3.1–3.5, 3.7, 3.12(a), (b), and 3.13(a) hold. If (3.3) holds, then:*

(a) $\sup_{\pi\in\Pi}|L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi)) - L^{-1/2}\sum_{h=1}^{L}\xi^*_{Lh}(\pi)| \to 0$ *in probability-P.*

(b) $\sup_{\pi\in\Pi}|L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi))| = \mathrm{O}_P(1)$.

(c) If in addition Assumption 3.13(b) and 3.14 hold, then the sequence of random functions

$$\left\{\pi \mapsto V_L(\pi)^{-1/2}L^{1/2}\breve{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi))\right\}_{L\in\mathbb{N}}$$

converges in distribution to a zero-mean Gaussian process with covariance kernel K concentrated on $\mathsf{U}(\Pi)$.

Proof of LEMMA C.8: To prove (i), note that there exists a real number $\varepsilon > 0$ such that for each $L\in\mathbb{N}$, the open ball in Θ with radius ε centered at θ^*_L is contained in $\mathrm{int}\,\Theta_0$. Fix $\pi\in\Pi$ arbitrarily. Then, by the mean value theorem for random functions (**?**, Lemma 3), there exists a sequence of random vectors $\{\ddot{\theta}_L(\pi):\Omega\to\Theta\}_{L\in\mathbb{N}}$ such that for each $L\in\mathbb{N}$, $\ddot{\theta}_L(\pi)$ is on the line segment connecting $\hat{\theta}_L$ and θ^*_L, and

$$\breve{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi)) - \breve{\alpha}_L(\cdot,\theta^*_L,\pi,\Lambda_L(\pi)) = \nabla_\theta\breve{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))'(\hat{\theta}_L-\theta^*_L)$$

whenever $|\hat{\theta}_L-\theta^*_L| < \varepsilon$. Subtracting $G^*_{1L}(\pi)(\hat{\theta}_L-\theta^*_L)$ from both sides of this equality and taking the absolute value of both sides of the resulting equality yields that

$$\begin{aligned}
&\big|\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi))-\check{\alpha}_L(\cdot,\theta_L^*,\pi,\Lambda_L(\pi))-G_{1L}^*(\pi)\,(\hat{\theta}_L-\theta_L^*)\big|\\
&\quad=\big|(\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))'-G_{1L}^*)(\pi)\,(\hat{\theta}_L-\theta_L^*)\big|\\
&\quad\leq|\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}^*(\pi)|\,|\hat{\theta}_L-\theta_L^*|,\quad \pi\in\Pi, L\in\mathbb{N},
\end{aligned}$$

so that

$$\begin{aligned}
&\sup_{\pi\in\Pi}\big|\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi))-\check{\alpha}_L(\cdot,\theta_L^*,\pi,\Lambda_L(\pi))-G_{1L}^*(\pi)\,(\hat{\theta}_L-\theta_L^*)\big|\\
&\quad\leq\sup_{\pi\in\Pi}|\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}^*(\pi)|\,|\hat{\theta}_L-\theta_L^*|,\quad L\in\mathbb{N},
\end{aligned}\tag{C.12}$$

whenever $|\hat{\theta}_L-\theta_L^*|<\varepsilon$. For the first factor on the right-hand side of the above inequality, we have that

$$\begin{aligned}
&|\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}^*(\pi)|\\
\leq&|\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}(\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))|\\
&+|G_{1L}(\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}(\theta_L^*,\pi,\Lambda_L(\pi))|,\quad \pi\in\Pi, L\in\mathbb{N}.
\end{aligned}$$

Let $\mathbb{A}$ be as in the proof of Theorem 3.1. Then we have that

$$\begin{aligned}
&\sup_{\pi\in\Pi}|\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}^*(\pi)|\\
&\quad\leq\sup_{\pi\in\Pi}|\nabla_\theta\check{\alpha}_L(\cdot,\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}(\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))|\\
&\qquad+\sup_{\pi\in\Pi}|G_{1L}(\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}(\theta_L^*,\pi,\Lambda_L(\pi))|\\
&\quad\leq\sup_{(\theta,\pi,W)\in\Theta_0\times\Pi\times\mathbb{A}}|\nabla_\theta\check{\alpha}_L(\cdot,\theta,\pi,W)-G_{1L}(\theta,\pi,W)|\\
&\qquad+\sup_{\pi\in\Pi}|G_{1L}(\ddot{\theta}_L(\pi),\pi,\Lambda_L(\pi))-G_{1L}(\theta_L^*,\pi,\Lambda_L(\pi))|,
\end{aligned}\tag{C.13}$$

whenever $|\hat{\theta}_L-\theta_L^*|<\varepsilon$. On the right-hand side of this inequality, the first term converges in probability-P to zero by Lemma C.4(i). The second term also converges to zero, because $\{L_{1L}\}$ is uniformly equicontinuous on $\Theta_0 \times \Pi \times \mathbb{\Lambda}$, and

$$\sup_{\pi \in \Pi} |\ddot{\theta}_L(\pi) - \theta_L^*| \leq |\hat{\theta}_L - \theta_L^*| \to 0 \text{ in probabiity-}P$$

by Assumption 3.4. Further, inequality (C.13) holds with a probability approaching one, as the probability that $|\hat{\theta}_L - \theta_L^*| < \varepsilon$ converges to one by Assumption 3.4. It follows that

$$\sup_{\pi \in \Pi} |\nabla_\theta \breve{\alpha}_L(\cdot, \ddot{\theta}_L(\pi), \pi, \Lambda_L(\pi)) - G_{1L}^*(\pi)| \to 0 \text{ in probability-}P.$$

Applying this result in inequality (C.12), which holds with a probability approaching one, establishes that

$$\begin{aligned} &\big|\breve{\alpha}_L(\cdot, \hat{\theta}_L, \pi, \Lambda_L(\pi)) - \breve{\alpha}_L(\cdot, \theta_L^*, \pi, \Lambda_L(\pi)) - G_{1L}^*(\pi)\,(\hat{\theta}_L - \theta_L^*)\big| \\ &= \mathrm{o}_P(|\hat{\theta}_L - \theta_L^*|) = \mathrm{o}_P(L^{-1/2}), \end{aligned}$$

where the second equality follows from Assumption 3.10(a) by Lemma C.3. Finally, multiplying both sides of this equality by $L^{1/2}$ and applying Assumption 3.10(a) to $(\hat{\theta}_L - \theta_L^*)$ on the left-hand side of the resulting inequality yields that

$$\sup_{\pi \in \Pi} \left| \begin{aligned} &L^{1/2}(\breve{\alpha}_L(\cdot, \hat{\theta}_L, \hat{\pi}_L, \Lambda_L(\pi)) - \breve{\alpha}_L(\cdot, \theta_L^*, \pi_L^*, \Lambda_L(\pi))) \\ &\qquad - G_{1L}^*(\pi) J_L^* L^{-1/2} \sum_{h=1}^{L} (LN_{Lh}/N_L) N_{Lh}^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{Lhij} \psi_{Lhij}^* \end{aligned} \right| = \mathrm{o}_P(1),$$

because the remainder term in the equality in Assumption 3.10(a) does not depend on π, and $\{G_{1L}^*(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$ is bounded. Given this equality, we have that

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\left|L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi)) - L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^*(\pi)\right| \\
&\quad = \sup_{\pi\in\Pi}\Bigg|L^{1/2}(\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi)) - \check{\alpha}_L(\cdot,\theta_L^*,\pi,\Lambda_L(\pi))) \\
&\qquad\quad + L^{-1/2}\check{\alpha}_L(\cdot,\theta_L^*,\pi,\Lambda_L(\pi)) - L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^*(\pi)\Bigg| \\
&\quad = \sup_{\pi\in\Pi}\Bigg|G_{1G}^*(\pi)J_L^* L^{-1/2}\sum_{h=1}^{L}(LN_{Lh}/N_L)N_{Lh}^{-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}W_{Lhij}\psi_{Lhij}^* \\
&\qquad\quad + L^{-1/2}\check{\alpha}_L(\cdot,\theta_L^*,\pi,\Lambda_L(\pi)) - L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^*(\pi)\Bigg| + o_P(1).
\end{aligned}
$$

We can easily verify that the first-term on the right-hand side of this equality is zero by using the definition of $\check{\alpha}_L$ and ξ_L^*. The result therefore follows.

For claim (ii), we apply Lemma C.2 to verify that $\{\xi_{Lh}^* : (h,L) \in \mathbb{H}\}$ is SLBP($2+2\delta$) on Π. The desired result follows from this fact and (i) by Lemma C.7(i).

We now turn to claim (iii). Because $\{V_L(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$ is uniformly positive by Assumption 3.13(b), $V_L(\pi)^{-1/2}$ is uniformly bounded. It follows from this fact and claim (i) of the current lemma that

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\left|V_L(\pi)^{-1/2}L^{1/2}\check{\alpha}_L(\cdot,\hat{\theta}_L,\pi,\Lambda_L(\pi)) - V_L(\pi)^{-1/2}L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^*(\pi)\right| \\
&\to 0 \text{ in probability-}P.
\end{aligned}
$$

It thus suffices to show that $\{V_L(\pi)^{-1/2}L^{-1/2}\sum_{h=1}^{L}\xi_{Lh}^*(\pi)\}_{L\in\mathbb{N}}$ converges in distribution to a zero-mean Gaussian process with covariance kernel K (see (**?**, Theorem 4.1)). We verify the desired by using Lemma C.7, by taking $\{\pi \mapsto V_L^{-1/2}(\pi)\xi_{Lh}^*(\pi)\}$ for $\{U_{Lh}\}$.

Because $\{\xi_{Lh}^* : (h,L) \in \mathbb{H}\}$ is SLBP($2+\delta$) on Π, and $\{V_L^{-1/2}(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$ is bounded, $\{\pi \mapsto V_L(\pi)^{-1/2}\xi_{Lh}^*(\pi) : (h,L) \in \mathbb{H}\}$ is SLBP($2+\delta$) on Π. In addition, the convergence of K_L to K is directly imposed in Assumption 3.14. Thus, conditions imposed in Lemma C.7 are satisfied in our current problem. The desired result therefore follows. □

Lemma C.9. *Suppose that 3.1–3.5, 3.7, 3.12(a), and 3.13(a) hold. If (3.3) holds, then*

$$\sum_{h=1}^{L} \beta_{Lh} \Big(\tilde{m}_{Lh}(\cdot, \hat{\theta}_L, \pi) \tilde{m}_{Lh}(\cdot, \hat{\theta}_L, \pi)' - N_{Lh}^{-2} \tilde{\Sigma}_{Lh}(\cdot, \hat{\theta}_L, \pi) \Big) = \mathrm{O}_P(L^{-1/2}). \quad \text{(C.14)}$$

Proof of LEMMA C.9: The result can be obtained by essentially repeating the argument in the proof of Lemma C.5, using Lemma C.8(ii) instead of Lemma C.4(ii). □

Lemma C.10. *Suppose that Assumptions 3.1–3.5, 3.7, 3.12, and 3.13(a) hold. If (3.3) holds, then* $\sup_{\pi\in\Pi} |L^{1/2}(\hat{\alpha}_L(\pi) - \check{\alpha}(\cdot, \hat{\theta}_L, \pi, \Lambda_L(\pi)))| = \mathrm{o}_P(1)$.

Proof of LEMMA C.10: By the definitions of $\{\hat{\alpha}_L\}_{L\in\mathbb{N}}$ and $\{\check{\alpha}_L\}_{L\in\mathbb{N}}$, we have that for each $\pi \in \Pi$ and each $L \in \mathbb{N}$,

$$\begin{aligned}
&L^{1/2}(\hat{\alpha}_L(\pi) - \check{\alpha}_L(\cdot, \hat{\theta}_L, \pi, \Lambda_L(\pi))) \\
&\quad = \mathrm{tr}\Bigg((\hat{\Lambda}_L(\pi) - \Lambda_L(\pi)) L^{1/2} \sum_{h=1}^{L} \beta_{Lh} \Big(\tilde{m}_{Lh}(\cdot, \hat{\theta}_L, \pi) \tilde{m}_{Lh}(\cdot, \hat{\theta}_L, \pi)' \\
&\qquad - N_{Lh}^{-2} \tilde{\Sigma}_{Lh}(\cdot, \hat{\theta}_L, \pi) \Big) \Bigg).
\end{aligned}$$

The right-hand side of this equality converges in probability-*P* to zero uniformly in $\pi \in \Pi$ by Lemma C.5 and Assumption 3.12(c). The desired result therefore follows. □

Now, Theorem 3.3 follows from Lemmas C.8 and C.10. We are also ready to prove Theorem 3.4.

Proof of THEOREM 3.3: The equality (3.14) immediately follows from Lemma C.8(i) and Lemma C.10. To establish the convergence of (3.15) in distribution, we use (3.14) together with the boundedness of $\{V_L(\pi)^{-1/2} : \pi \in \Pi, L \in \mathbb{N}\}$ to obtain that

$$\sup_{\pi\in\Pi} \Big| V_L^{-1/2}(\pi) L^{1/2} \hat{\alpha}_L(\pi) - V_L^{-1/2}(\pi) L^{1/2} \check{\alpha}(\cdot, \hat{\theta}_L, \hat{\pi}_L, \Lambda_L(\pi)) \Big| = \mathrm{o}_P(1).$$

The desired convergence in distribution follows from Lemma C.8(iii) by this fact (see (**?**, Theorem 4.1)). □

Proof of THEOREM 3.4: Let $\mathbb{A}$ be as in Theorem 3.1. Also, let $\mathbb{J}$ and $\mathbb{G}_1$ be compact subsets of $\mathbb{R}^{p\times p}$ and $\mathbb{R}^{p\times 1}$, respectively, such that the bounded arrays $\{\bar{J}_L\}_{L\in\mathbb{N}}$ and $\{G^*_{1L}(\pi) : \pi \in \Pi : L \in \mathbb{N}\}$ are uniformly interior to $\mathbb{J}$ and $\mathbb{G}_1$, respectively. Then

$$\{(\theta,\pi,\Lambda,J,G_1) \mapsto \mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta,\pi,W,J,G_1)^2] : \Theta_0 \times \Pi_0 \times \mathbb{A} \times \mathbb{J} \times \mathbb{G}_1 \to \mathbb{R},\, (h,L) \in \mathbb{H}\}$$

is uniformly bounded and uniformly equicontinuous, and it holds that

$$\sup_{(\theta,\pi,W,J,G_1)\in\Theta_0\times\Pi_0\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1} \Bigg| L^{-1}\sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot,\theta,\pi,W,J,G_1)^2 \\ - L^{-1}\sum_{h=1}^{L} \mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta,\pi,W,J,G_1)^2] \Bigg| \to 0 \text{ in probability-}P,$$

as verified in the proof of Theorem 3.4.

Note that

$$\hat{V}_L(\pi) = G^{-1}\sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot,\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi),\hat{J}_L,\hat{G}_{1L}(\pi))^2, \quad \pi \in \Pi, L \in \mathbb{N}, \quad \text{and}$$

$$\bar{V}_L(\pi) = L^{-1}\sum_{h=1}^{L} \mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta^*_L,\pi,\Lambda_L(\pi),\bar{J}_L,G^*_{1L}(\pi))^2, \quad \pi \in \Pi, L \in \mathbb{N}.$$

The claim (i) follows from these facts, given that $\{\hat{\theta}_L\}_{L\in\mathbb{N}}$ and $\{\hat{J}_L\}_{L\in\mathbb{N}}$ are consistent for $\{\theta^*_L\}_{L\in\mathbb{N}}$ and $\{\bar{J}_L\}_{L\in\mathbb{N}}$, respectively, and that $\{\hat{\Lambda}_L(\pi)\}_{L\in\mathbb{N}}$ and $\{\hat{G}_{1L}(\pi)\}_{L\in\mathbb{N}}$ are consistent for $\{\Lambda_L(\pi)\}_{L\in\mathbb{N}}$ and $\{G^*_{1L}(\pi)\}_{L\in\mathbb{N}}$, respectively, uniformly in $\pi \in \Pi$.

We now turn to claim (ii). By the definition of $\mathscr{T}_L$, we have that

$$\sup_{\pi\in\Pi}\left|\mathscr{T}_L(\pi) - \frac{L^{1/2}\hat{\alpha}_L(\pi)}{V_L^{1/2}(\pi)}\right| = \sup_{\pi\in\Pi}\left|(\hat{V}_L(\pi)^{-1/2} - V_L(\pi)^{-1/2})L^{1/2}\hat{\alpha}_L(\pi)\right| \\ \leq \sup_{\pi\in\Pi}|\hat{V}_L(\pi)^{-1/2} - V_L(\pi)^{-1/2}| \sup_{\pi\in\Pi}|G^{1/2}\hat{\alpha}_L(\pi)|, \quad L \in \mathbb{N}.$$

The first factor on the right-hand side converges in probability-P to zero by claim (i) of the current theorem and the assumption that $\bar{V}_L = V_L$ for each $L \in \mathbb{N}$. The second factor is $\mathrm{O}_P(1)$ by the second claim of Theorem 3.3 and the boundedness of $\{V_L(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$. Thus, we have that $\sup_{\pi\in\Pi} |\mathscr{T}_L(\pi) - L^{1/2}\hat{\alpha}_L(\pi)/V_L^{1/2}(\pi)| = \mathrm{o}_P(1)$. The result follows from this equality and Theorem 3.3 by (**?**, Theorem 4.1). □

To prove Theorem 3.5, we first prove a few lemmas.

Lemma C.11. *Suppose that Assumptions 3.1–3.5, 3.7, 3.11–3.14. Then* $\sup_{(\pi_1,\pi_2)\in\Pi^2} |\hat{K}_L(\pi_1,\pi_2) - \bar{K}_L(\pi_1,\pi_2)| \to 0$ *in probability-P, where* $\{\bar{K}_L : \Pi^2 \to \mathbb{R}\}_{L\in\mathbb{N}}$ *defined by*

$$\begin{aligned}\bar{K}_L(\pi_1,\pi_2) &\equiv (\bar{V}_L(\pi_1)\bar{V}_L(\pi_2))^{-1/2}\Bigg(L^{-1}\sum_{h=1}^{L}\mathrm{cov}[\bar{\xi}_L(\pi_1),\bar{\xi}_L(\pi_2)] \\ &\quad + L^{-1}\sum_{h=1}^{L}(L\beta_{Lh})^2\bar{m}^*_{Lh}(\pi_1)'\Lambda_L(\pi)\bar{m}^*_{Lh}(\pi_1)\bar{m}^*_{Lh}(\pi_2)'\Lambda_L(\pi)\bar{m}^*_{Lh}(\pi_2)\Bigg) + o(1), \\ &\quad (\pi_1,\pi_2)\in\Pi^2, \quad L \in \mathbb{N}. \end{aligned} \tag{C.15}$$

Proof of LEMMA C.11: Let $\mathbb{A}$, $\mathbb{J}$, and $\mathbb{G}_1$ be as in the proof of Theorem 3.3. By using Lemma C.2, we can show that

$$\begin{aligned}&\{(\theta,\pi_1,\pi_2,\Lambda_1,\Lambda_2,J,G_{1,1},G_{1,2}) \mapsto \check{\xi}_{Lh}(\cdot,\theta,\pi_1,\Lambda_1,J,G_{1,1})\check{\xi}_{Lh}(\cdot,\theta,\pi_2,\lambda_2,J,G_{1,2}) : \\ &\qquad \Theta_0\times\Pi\times\Pi\times\mathbb{A}\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1\times\mathbb{G}_1 \to \mathbb{R} : (h,L)\in\mathbb{H}\}\end{aligned}$$

is SLBP$(1+\delta)$ on $\Theta_0\times\Pi\times\Pi\times\mathbb{A}\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1\times\mathbb{G}_1$. It follows by Lemma C.1 that

$$\begin{aligned}&\{(\theta,\pi_1,\pi_2,\Lambda_1,\Lambda_2,J,G_{1,1},G_{1,2}) \\ &\mapsto \mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta,\pi_1,\Lambda_1,J,G_{1,1})\check{\xi}_{Lh}(\cdot,\theta,\pi_2,\Lambda_2,J,G_{1,2}) : (h,L)\in\mathbb{H}\}]\end{aligned}$$

is bounded and equicontinuous uniformly on $\Theta_0\times\Pi\times\Pi\times\mathbb{A}\times\mathbb{A}\times\mathbb{J}\times\mathbb{G}_1\times\mathbb{G}_1$, and that

$$\begin{aligned}
\sup_{(\theta,\pi_1,\pi_2,\Lambda_1,\Lambda_2,J,G_{1,1},G_{1,2})\in\Theta_0\times\Pi_0\times\mathbb{\Lambda}\times\mathbb{J}\times\mathbb{G}_1}&\Bigg|L^{-1}\sum_{h=1}^{L}\check{\xi}_{Lh}(\cdot,\theta,\pi_1,W_1,J,L_{1,1}) && \text{(C.16)}\\
&\times\check{\xi}_{Lh}(\cdot,\theta,\pi_2,W_2,J,L_{1,2})-L^{-1}\sum_{h=1}^{L}\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta,\pi_1,\Lambda_1,J,G_{1,1})\\
&\times\check{\xi}_{Lh}(\cdot,\theta,\pi_2,\Lambda_2,J,G_{1,2})]\Bigg|\\
&\to 0 \text{ in probability-}P. && \text{(C.17)}
\end{aligned}$$

Note that for each $(\pi_1,\pi_2)\in\Pi^2$ and each $L\in\mathbb{N}$,

$$\begin{aligned}
&\hat{K}_L(\pi_1,\pi_2)=(\hat{V}_L(\pi_1)\hat{V}_L(\pi_2))^{-1/2}\\
&\quad\times L^{-1}\sum_{h=1}^{L}\check{\xi}_{Lh}(\cdot,\hat{\theta}_L,\pi_1,\hat{\Lambda}_L(\pi_1),\hat{J}_L,\hat{G}_{1,L}(\pi_1))\check{\xi}_{Lh}(\cdot,\hat{\theta}_L,\pi_2,\hat{\Lambda}_L(\pi_2),\hat{J}_L,\hat{G}_{1,L}(\pi_2)).
\end{aligned}$$

Given (C.17), Assumption 3.13(b), and Theorem 3.4(i), it follows that $\{\hat{K}_L(\pi_1,\pi_2)\}_{L\in\mathbb{N}}$ is uniformly consistent for

$$\begin{aligned}
&\Bigg\{(\bar{V}_L(\pi_1)\bar{V}_L(\pi_2))^{-1/2}\\
&\times L^{-1}\sum_{h=1}^{L}\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta^*_L,\pi_1,\Lambda_L(\pi_1),J^*_L,G^*_{1,L}(\pi_1))\check{\xi}_{Lh}(\cdot,\theta^*_L,\pi_2,\Lambda_L(\pi_2),J^*_L,G^*_{1,L}(\pi_2))]\Bigg\}_{L\in\mathbb{N}},
\end{aligned}$$

which is uniformly equicontinuous in $(\pi_1,\pi_2)\in\Pi^2$. Because

$$\begin{aligned}
L^{-1}&\sum_{h=1}^{L}\mathrm{E}[\check{\xi}_{Lh}(\cdot,\theta^*_L,\pi_1,\Lambda_L(\pi_1),J^*_L,G^*_{1,L}(\pi_1))\check{\xi}_{Lh}(\cdot,\theta^*_L,\pi_2,\Lambda_L(\pi_2),J^*_L,G^*_{1,L}(\pi_2))]\\
&=L^{-1}\sum_{h=1}^{L}\mathrm{cov}[\xi^*_{Lh}(\pi_1),\xi^*_{Lh}(\pi_2)]\\
&\quad+L^{-1}\sum_{h=1}^{L}(L\beta_{Lh})^2\bar{m}^*_{Lh}(\pi_1)'\Lambda_L(\pi)\bar{m}^*_{Lh}(\pi_1)\\
&\qquad\times\bar{m}^*_{Lh}(\pi_2)'\Lambda_L(\pi)\bar{m}^*_{Lh}(\pi_2),\quad(\pi_1,\pi_2)\in\Pi^2,\,(h,L)\in\mathbb{H}, && \text{(C.18)}
\end{aligned}$$

the first result follows. □

The next lemma is necessary in establishing Lemma C.13 stated below, which is essential in proving Theorem 3.5.

Lemma C.12. *Suppose that Assumptions 3.1–3.5, 3.7, and 3.11–3.14. If (3.3) holds and $\bar{J}_L = J_L^*$ for each $L \in \mathbb{N}$, then $\{\pi \mapsto \hat{\mathscr{I}}_L(\pi)\}_{L\in\mathbb{N}}$ converges in distribution to a zero-mean Gaussian process in Π with covariance kernel K concentrated on $\mathsf{U}(\Pi)$.*

Proof of LEMMA C.12: Let $\mathbb{A}$, $\mathbb{J}$, and $\mathbb{G}_1$ be as in the proof of Theorem 3.4. Then the array of random functions $\{\check{\xi}_{Lh}\nu_h : (h,L) \in \mathbb{H}\}$ is stochastically equicontinuous on $\Theta_0 \times \Pi \times \mathbb{A} \times \mathbb{J} \times \mathbb{G}_1$ by Lemma C.7(iii). Pick a real number

$\varepsilon > 0$ arbitrarily. When for a real number $\beta > 0$, $|\hat{\theta}_L - \theta_L^*| < \beta/4$, $\sup_{\pi\in\Pi} |\hat{\Lambda}_L(\pi) - \Lambda_L(\pi)| < \beta/4$, $|\hat{J}_L - J_L^*| < \beta/4$, and $\sup_{\pi\in\Pi} |\hat{G}_{1L}(\pi) - G_{1L}^*(\pi)| < \beta/4$, we have that

$$(|\hat{\theta}_L - \theta_L^*|^2 + \sup_{\pi\in\Pi} |\hat{\Lambda}_L(\pi) - \Lambda_L(\pi)|^2 + |\hat{J}_L - J_L^*|^2 + \sup_{\pi\in\Pi} |\hat{G}_{1L}(\pi) - G_{1L}^*(\pi)|^2)^{1/2} < \beta.$$

Given this fact, for each real number $\beta > 0$,

$$\begin{aligned} P&\left[\sup_{\pi\in\Pi}\left|L^{-1/2}\sum_{h=1}^{L} \hat{\xi}_{Lh}(\pi)\,\nu_h - L^{-1/2}\sum_{h=1}^{L} \xi_{Lh}^*(\pi)\,\nu_h\right| > \varepsilon\right] \\ &= P\left[\sup_{\pi\in\Pi}\left|L^{-1/2}\sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot,\hat{\theta}_L,\pi,\hat{\Lambda}_L(\pi),\hat{J}_L,\hat{G}_{1L})\,\nu_h \right.\right. \\ &\qquad \left.\left. - L^{-1/2}\sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot,\theta_L^*,\pi,\Lambda_L(\pi),J_L^*,G_{1L}^*(\pi))\,\nu_h\right| > \varepsilon\right] \end{aligned} \tag{C.19}$$

is dominated by

$$\begin{aligned}
&P[\hat{\theta}_L \notin \Theta_0] + P[\hat{\Lambda}_L(\pi) \notin \mathbb{\Lambda} \text{ for some } \pi \in \Pi] + P[\hat{J}_L \notin \mathbb{J}] + P[\hat{G}_{1L}(\pi) \notin \mathbb{G}_1 \text{ for some } \pi \in \Pi] \\
&\quad + P[|\hat{\theta}_L - \theta_L^*| \geq b/4] + P\left[\sup_{\pi\in\Pi} |\hat{\Lambda}_L(\pi) - \Lambda_L(\pi)| \geq b/4\right] + P[|\hat{J}_L - J_L^*| \geq b/4] \\
&\quad + P\left[\sup_{\pi\in\Pi} |\hat{G}_{1L}(\pi) - G_{1L}^*(\pi)| \geq b/4\right] \\
&\quad + P\Bigg[\sup\Bigg\{\Bigg|L^{-1/2} \sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot, \theta_2, \pi, \Lambda_2, J_2, G_{1,2}) \nu_h \\
&\qquad\quad - L^{-1/2} \sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot, \theta_1, \pi, \Lambda_1, J_1, G_{1,1}) \nu_h\Bigg| : \\
&\qquad\quad (|\theta_2 - \theta_1|^2 + |W_2 - W_1|^2 + |J_2 - J_1|^2 + |G_2 - G_1|^2)^{1/2} < b, \\
&\qquad\quad (W_1, W_2) \in \mathbb{\Lambda}^2,\ (\theta_1, \theta_2) \in \Theta_0^2,\ (J_1, J_2) \in \mathbb{J}^2,\ (G_{1,2}, G_{1,1}) \in \mathbb{G}_1^2\Bigg\} > \varepsilon\Bigg], \\
&\qquad\qquad L \in \mathbb{N}.
\end{aligned}$$

In the above expression, we can choose b to make the last term arbitrarily small uniformly in $L \in \mathbb{N}$, while given the chosen b, all other terms converge to zero as $L \to \infty$. It follows that (C.19) converges to zero for each $\varepsilon > 0$, i.e.,

$$\sup_{\pi\in\Pi}\left|L^{-1/2} \sum_{h=1}^{L} \hat{\xi}_{Lh}(\pi)\, \nu_h - L^{-1/2} \sum_{h=1}^{L} \xi_{Lh}^*(\pi)\, \nu_h\right| = \mathrm{o}_P(1).$$

Further, $\{\sup_{\pi\in\Pi} |\hat{V}_L(\pi)|^{-1}\}_{L\in\mathbb{N}}$ is $\mathrm{O}_P(1)$, because $\{\hat{V}_L(\pi)\}_{L\in\mathbb{N}}$ is consistent for $\{V_L(\pi)\}_{L\in\mathbb{N}}$ uniformly in $\pi \in \Pi$, and $\{V_L(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$ is uniformly positive. It follows that

$$\sup_{\pi\in\Pi}\left|\hat{V}_L(\pi)^{-1/2} L^{-1/2} \sum_{h=1}^{L} \hat{\xi}_{Lh}(\pi)\, \nu_h - V_L(\pi)^{-1/2} L^{-1/2} \sum_{h=1}^{L} \xi_{Lh}^*(\pi)\, \nu_h\right| = \mathrm{o}_P(1). \tag{C.20}$$

Now, it is straightforward to verify that $\{V_L(\pi)^{-1/2} \xi_{Lh}^*(\pi) : \pi \in \Pi,\ (h,L) \in \mathbb{H}\}$ satisfies all conditions imposed on $\{U_{Lh}\}$ in Lemma C.7, by using Lemma C.2 and the independence between $\{\xi_{Lh}^*(\pi) : \pi \in \Pi,\ (h,L) \in \mathbb{H}\}$ together with the normality of ν_h. Thus, $\{\pi \mapsto V_L(\pi)^{-1/2} L^{-1/2} \sum_{h=1}^{L} \xi_{Lh}^*(\pi)\, \nu_h\}_{L\in\mathbb{N}}$ converges in distribution to the stated Gaussian limit. Given (C.20),

$$\left\{\pi \mapsto \hat{\mathscr{T}}_L = \hat{V}_L(\pi)^{-1/2} L^{-1/2} \sum_{h=1}^{L} \hat{\xi}_{Lh}(\pi)\, \nu_h \right\}_{L\in\mathbb{N}}$$

also converges in distribution to the same limit (**?**, Theorem 4.1), and the result follows. □

We now prove an important lemma about the conditional distribution of $\{\pi \mapsto \hat{\mathscr{T}}_L(\pi)\}_{L\in\mathbb{N}}$ given $\tilde{X}$ and $\{C_{Lhk} : (k,h,L) \in \mathbb{K}\}$, using the result on the corresponding unconditional distribution established in Lemma C.12.

Lemma C.13. *(a) Suppose that Assumptions 3.1–3.5, 3.7, 3.11, 3.12, and 3.13 hold. Then* $\sup_{\pi\in\Pi} |\hat{\mathscr{T}}_L(\pi)| = \mathrm{O}_P(1)$.

(b) Suppose that Assumptions 3.1–3.5, 3.7, 3.11–3.14 hold. If (3.3) holds and $\bar{J}_L = J_L^*$ *for each* $L \in \mathbb{N}$*, then the distribution of* $\{\pi \mapsto \hat{\mathscr{T}}_L(\pi)\}$ *conditional on* $\tilde{X}$ *and* $\{C_{Lhk} : (k,h,L) \in \mathbb{K}\}$ *weakly converges to a zero-mean Gaussian process with covariance kernel K concentrated on* $\mathsf{U}(\Pi)$ *in probability-P.*

Proof of LEMMA C.13: To prove claim (i), recall that

$$\hat{\xi}_{Lh}(\pi) = \check{\xi}_{Lh}(\cdot, \hat{\theta}_L, \pi, \hat{\Lambda}_L(\pi), \hat{J}_L, \hat{G}_{1L}(\pi)), \quad \pi \in \Pi, L \in \mathbb{N}.$$

Let $\mathbb{A}$, $\mathbb{J}$, and $\mathbb{G}_1$ be as in the proof of Theorem 3.3. In Lemma C.7(i), take $\Theta_0 \times \Pi \times \mathbb{A} \times \mathbb{J} \times \mathbb{G}_1$ for Γ and $\{\check{\xi}_{Lh}\nu_h : (h,L) \in \mathbb{H}\}$ for $\{U_{Lh} : (h,L) \in \mathbb{H}\}$. Then we can easily verify that all assumptions imposed in the lemma are satisfied, by using Lemma C.2. Thus,

$$\sup\left\{\left| L^{-1/2} \sum_{h=1}^{L} \check{\xi}_{Lh}(\cdot, \theta, \pi, \Lambda, J, G_1)\, \nu_h \right| \right.$$
$$\left. : (\theta, \pi, \Lambda, J, G_1) \in \Theta_0 \times \Pi \times \mathbb{A} \times \mathbb{J} \times \mathbb{G}_1 \right\} = \mathrm{O}_P(1).$$

Because the probability that $\hat{\theta}_L \in \Theta_0$, $\hat{J}_L \in \mathbb{J}$, and for each $\pi \in \Pi$, $\hat{\Lambda}_L(\pi)$ and $\hat{G}_{L1}(\pi)$

belong to $\mathbb{A}$ and $\mathbb{G}_1$, respectively, approaches one as $L \to \infty$, it follows that

$$\sup_{\pi\in\Pi}\left| L^{-1/2}\sum_{h=1}^{L}\hat{\xi}_{Lh}(\pi)\nu_h\right| = \mathrm{O}_P(1).$$

Further, $\{\hat{V}_L(\pi)\}_{L\in\mathbb{N}}$ is consistent for $\{\bar{V}_L(\pi)\}_{L\in\mathbb{N}}$ uniformly in π, which is positive uniformly in $\pi \in \Pi$ and $L \in \mathbb{N}$. Thus,

$$\sup_{\pi\in\Pi}|\hat{\mathscr{T}}_L(\pi)| = \sup_{\pi\in\Pi}\left|\hat{V}_L(\pi)^{-1/2}L^{-1/2}\sum_{h=1}^{L}\hat{\xi}_{Lh}(\pi)\nu_h\right| = \mathrm{O}_P(1).$$

We now turn to claim (ii). Our proof follows the strategy taken by **?**. Let $\ell^\infty(\Pi)$ denote the set of all bounded continuous functions on Π and $\mathrm{BL}_1(\ell^\infty(\Pi))$ the set of all real-valued Lipschitz functions on $\ell^\infty(\Pi)$ with a uniform norm and a Lipschitz constant both bounded by one. Then it suffices to show that

$$\sup_{\phi\in\mathrm{BL}_1(\ell^\infty(\Pi))}|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L))] - \mathrm{E}[\phi(\eta)]| \to 0 \text{ in probability-}P,$$

where η is a zero-mean Gaussian process with covariance kernel K concentrated on $\mathsf{U}(\Pi)$, and E_ν denotes the expectation taken with respect to $\nu \equiv \{\nu_h\}_{h\in\mathbb{N}}$ (see (van der Vaart and Wellner, 1996, pages 72–73)).

For each real number $b > 0$, Π can be covered by a finite number of open b-balls, because Π is compact. Let $M_\beta : \Pi \to \Pi$ be a function that returns a closest point among the centers of such b-balls. Because η is concentrated on $\mathsf{U}(\Pi)$, it holds that $\sup_{\pi\in\Pi}|\eta(M_b(\pi)) - \eta(\pi)| \to 0$ as $b \downarrow 0$ a.s. We thus have

$$\begin{aligned}
&\sup_{\phi\in\mathrm{BL}_1(\ell^\infty(\Pi))}\big|\mathrm{E}[\phi(\eta(M_b(\cdot)))] - \mathrm{E}[\phi(\eta)]\big| \le \mathrm{E}\left[\min\left\{2, \sup_{\pi\in\Pi}|\eta(M_b(\pi)) - \eta(\pi)|\right\}\right]\\
&\quad\le \mathrm{E}\big[\min\big\{2, \sup\{|\eta(\pi_2)) - \eta(\pi_1)| : |\pi_1 - \pi_2| < b,\, (\pi_1, \pi_2) \in \Pi^2\}\big\}\big]\\
&\quad\to 0 \text{ as } \beta \downarrow 0, \qquad\qquad (\mathrm{C}.21)
\end{aligned}$$

where the convergence follows by the dominated convergence theorem.

Let ε be an arbitrary positive real number. Then there exists a real number $b_1 > 0$ such that for each $b \in (0, b_1)$, the left-hand side of (C.21) is less than $\varepsilon/3$. It follows that for each $\beta \in (0, b_1)$,

$$\begin{aligned}
\sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} |\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)] - \mathrm{E}[\phi(\eta)]| &\leq \sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)] - \mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]\big| \\
&\quad + \sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))] - \mathrm{E}[\phi(\eta(M_b(\cdot)))]\big| \\
&\quad + \sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}[\phi(\eta(M_b(\cdot)))] - \mathrm{E}[\phi(\eta)]\big| \\
&\leq \sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)] - \mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]\big| \\
&\quad + \sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))] - \mathrm{E}[\phi(\eta(M_b(\cdot)))]\big| + \varepsilon/3 \quad L \in \mathbb{N}.
\end{aligned}$$

It further follows that for each $b \in (0, b_1)$,

$$\begin{aligned}
P\left[\sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} |\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)] - \mathrm{E}[\phi(\eta)]| > \varepsilon\right] \\
\leq P\left[\sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)] - \mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]\big| > \varepsilon/3\right] \\
+ P\left[\sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))] - \mathrm{E}[\phi(\eta(M_b(\cdot)))]\big| > \varepsilon/3\right], \quad L \in \mathbb{N}.
\end{aligned} \tag{C.22}$$

For the first term on the right-hand side of this inequality, we have that

$$\begin{aligned}
&\sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \big|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)] - \mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]\big| \\
&\quad \leq \sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} \mathrm{E}_\nu[|\phi(\hat{\mathscr{T}}_L) - \phi(\hat{\mathscr{T}}_L(M_b(\cdot)))|] \\
&\quad \leq \mathrm{E}_\nu\left[\sup_{\phi \in \mathrm{BL}_1(\ell^\infty(\Pi))} |\phi(\hat{\mathscr{T}}_L) - \phi(\hat{\mathscr{T}}_L(M_b(\cdot)))|\right] \\
&\quad \leq \mathrm{E}_\nu\left[\min\left\{2, \sup_{\pi \in \Pi} |\hat{\mathscr{T}}_L(M_b(\pi)) - \hat{\mathscr{T}}_L(\pi)|\right\}\right] \\
&\quad \leq \mathrm{E}_\nu\big[\min\big\{2, \sup\{|\hat{\mathscr{T}}_L(\pi_2) - \hat{\mathscr{T}}_L(\pi_1)| : |\pi_1 - \pi_2| < b, (\pi_1, \pi_2) \in \Pi^2\}\big\}\big].
\end{aligned}$$

Taking the expectation of both sides of this inequality and applying the law of iterated expectations to the right-hand side yields that

$$\begin{aligned}
&\mathrm{E}\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}\left|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)]-\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]\right|\right]\\
&\quad\le \mathrm{E}\big[\min\big\{2,\sup\{|\hat{\mathscr{T}}_L(\pi_2)-\hat{\mathscr{T}}_L(\pi_1)|:|\pi_1-\pi_2|<b,\,(\pi_1,\pi_2)\in\Pi^2\}\big\}\big]\\
&\quad\to \mathrm{E}\big[\min\big\{2,\sup\{|\eta(\pi_2)-\eta(\pi_1)|:|\pi_1-\pi_2|<b,\,(\pi_1,\pi_2)\in\Pi^2\}\big\}\big],
\end{aligned}$$

where the convergence follows by Lemma C.12. By applying the Markov inequality (Davidson, 1994, p. 132) to this result, we obtain that

$$\begin{aligned}
&\limsup P\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}\left|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)]-\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]\right|>\varepsilon/3\right]\\
&\quad\le (\varepsilon/3)^{-1}\,\mathrm{E}\big[\min\big\{2,\sup\{|\eta(\pi_2)-\eta(\pi_1)|:|\pi_1-\pi_2|<b,\,(\pi_1,\pi_2)\in\Pi^2\}\big\}\big].
\end{aligned}\tag{C.23}$$

Now, let κ be an arbitrary positive real number. As is the case in (C.21), the right-hand side of (C.23) can be made smaller than κ by setting b equal to a suitable value in $(0,b_1)$. Given such β, we have that

$$\begin{aligned}
&\mathrm{E}\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}\left|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]-\mathrm{E}[\phi(\eta(M_b(\cdot)))]\right|\right]\\
&\quad\le \mathrm{E}\left[\mathrm{E}_\nu\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}|\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))-\phi(\eta(M_b(\cdot)))|\right]\right]\\
&\quad\le \mathrm{E}\left[\min\left\{2,\sup_{\pi\in\Pi}|\hat{\mathscr{T}}_L(M_b(\pi))-\eta(M_b(\pi))|\right\}\right]\le\\
&\quad \mathrm{E}\left[\min\left\{2,\sup_{\pi\in\Pi}|\hat{\mathscr{T}}_L(\pi)-\eta(\pi)|\right\}\right]\to 0,
\end{aligned}$$

where the convergence follows by Lemma C.12. It follows by the Markov inequality (Davidson, 1994, p. 132) that

$$\limsup P\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}\left|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L(M_b(\cdot)))]-\mathrm{E}[\phi(\eta(M_b(\cdot)))]\right|>\varepsilon/3\right]=0.$$

Applying the above results in (C.22) yields that

$$\limsup P\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)]-\mathrm{E}[\phi(\eta)]|>\varepsilon\right]\leq\kappa.$$

Because κ is an arbitrary positive real number, it follows that

$$\limsup P\left[\sup_{\phi\in \mathrm{BL}_1(\ell^\infty(\Pi))}|\mathrm{E}_\nu[\phi(\hat{\mathscr{T}}_L)]-\mathrm{E}[\phi(\eta)]|>\varepsilon\right]\to 0.$$

As this holds for every $\varepsilon>0$, the second claim of the current lemma follows. □

We finally prove Theorem 3.5.

Proof of THEOREM 3.5: For claim (i), it follows from Theorem 3.4(ii) by subsequence theorem (van der Vaart and Wellner, 1996, Lemma 1.9.2(ii)) and the continuous mapping theorem (van der Vaart and Wellner, 1996, Lemma 1.3.6) that $\{\hat{\varphi}_L=\varphi(\hat{\mathscr{T}}_L)\}_{L\in\mathbb{N}}$ converges in distribution to $\varphi_0=\varphi(\eta)$ in probability. We thus have that $\sup_{x\in\mathbb{R}}|\hat{F}_L(x)-F(x)|\to 0$ in probability-P. The first result follows from this, because

$$|\hat{p}_L-p_L|=|\hat{F}_L(\varphi_L)-F(\varphi_L)|\leq\sup_{x\in\mathbb{R}}|\hat{F}_L(x)-F(x)|.$$

To prove the second result, note that if $\{p_L\}_{L\in\mathbb{N}}$ is convergent in distribution, $\{\hat{p}_L\}_{L\in\mathbb{N}}$ converges to the same limiting distribution as $\{p_L\}$ does, by the asymptotic equivalence lemma. It thus suffices to show that $\{p_L\}$ converges in distribution to the uniform distribution on $[0,1]$.

Recall that $\{\mathscr{T}_L\}_{L\in\mathbb{N}}$ converges in distribution to η if (**??**) holds (Theorem 3.4). Because F is assumed to be continuous and increasing on the support of φ_0, it follows by the continuous mapping theorem that $\{p_L=F(\varphi_L)=F(\varphi(\mathscr{T}_L))\}_{L\in\mathbb{N}}$ converges in distribution to $p_0=F(\varphi_0)=F(\varphi(\eta))$. The random variable p_0 is distributed uniformly on $[0,1]$, since for each $y\in(0,1)$,

$$P[F(p_0)\leq y]=P[p_0\leq F^{-1}(y)]=F(F^{-1}(y))=y,$$

where for each $y\in(0,1)$, $F^{-1}(y)\equiv\inf\{x\in\mathbb{R}:F(x)>y\}$. This completes the proof of claim of (i).

We now turn to claim (ii). Pick real numbers $\varepsilon > 0$ and $b > 0$ arbitrarily. Then, it suffices to show that for almost all $L \in \mathbb{N}$, $P[\hat{p}_L > \varepsilon] \leq \beta$. To establish this inequality, let $\bar{F}_L$ denote the unconditional distribution function of $\hat{\varphi}_L = \varphi(\hat{\mathscr{T}}_L)$ for each $L \in \mathbb{N}$. Because $\sup_{\pi\in\Pi} |\hat{\mathscr{T}}_L(\pi)| = \mathrm{O}_P(1)$ (Lemma C.13(i)), and φ is monotonic (Assumption 3.15), we have that $\hat{\varphi}_L = \mathrm{O}_P(1)$. It follows that there exists a real number Δ such that

$$1 - \bar{F}_L(\Delta) < b\varepsilon/2, \quad L \in \mathbb{N}.$$

Note that whenever $1 - \hat{F}_L(\Delta) \leq \varepsilon$ and $\varphi_L \geq \Delta$, it holds that $1 - \hat{F}_L(\varphi_L) \leq \varepsilon$. It follows that

$$P[\hat{p}_L > \varepsilon] = P[1 - \hat{F}_L(\varphi_L) > \varepsilon] \leq P[1 - \hat{F}_L(\Delta) > \varepsilon] + P[\varphi_L < \Delta].$$

It thus suffices to show that (A) $P[1 - \hat{F}_L(\varphi_L) > \varepsilon] < b/2$ for each $L \in \mathbb{N}$ and that (B) $P[\varphi_L < \Delta] < b/2$ for almost all $L \in \mathbb{N}$.

To show (A), we use the fact that for each $L \in \mathbb{N}$, $\mathrm{E}[\hat{F}_L(\Delta)] = \bar{F}_L(\Delta)$. Using this fact, we obtain that

$$\mathrm{E}[1 - \hat{F}_L(\Delta)] = 1 - \bar{F}_L(\Delta) < b\varepsilon/2, \quad L \in \mathbb{N}.$$

Given that $1 - \hat{F}_L(\Delta)$ is a positive random variable, condition (A) follows from this inequality by the Markov inequality (Davidson, 1994, p. 132).

For condition (B), recall that $\{\hat{\alpha}_L(\pi)\}_{L\in\mathbb{N}}$ and $\{\hat{V}_L(\pi)\}_{L\in\mathbb{N}}$ are respectively consistent for $\{\alpha_L^*(\pi)\}_{L\in\mathbb{N}}$ and $\{\bar{V}_L(\pi)\}_{L\in\mathbb{N}}$ uniformly in $\pi \in \Pi$ (Theorem 3.1 and Theorem 3.4(i)). Also, we know that $\{\bar{V}_L(\pi) : \pi \in \Pi, L \in \mathbb{N}\}$ is uniformly bounded and uniformly positive (Assumption 3.13(b) and Theorem 3.4(i)). Thus, we have that

$$\sup_{\pi\in\Pi} \left| L^{-1/2} \mathscr{T}_L(\pi) - \frac{\alpha_L^*(\pi)}{\bar{V}_L} \right| \to 0 \text{ in probability-}P. \tag{C.24}$$

By hypothesis, we also have that $\liminf \inf_{\pi\in\bar{\Pi}} \alpha_L^*(\pi) > 0$, so that

$$c \equiv \liminf \inf_{\pi\in\bar{\Pi}} \frac{\alpha_L^*(\pi)}{\bar{V}_L} > 0.$$

It follows that

$$P\left[\inf_{\pi\in\bar{\Pi}} L^{-1/2}\mathscr{T}_L(\pi) < c/2\right] \to 0.$$

Note that when $\inf_{\pi\in\bar{\Pi}} L^{-1/2}\mathscr{T}_L(\pi) \geq c/2$, it holds that

$$\varphi_L \geq \varphi(1\{\pi \in \bar{\Pi}\} \cdot L^{1/2}c),$$

where $1\{C\}$ is the indicator function for the condition C. Because the right-hand side of the above inequality, which diverges to infinity by Assumption 3.15, is no smaller than Δ for almost all $L \in \mathbb{N}$, we have that for almost all $L \in \mathbb{N}$,

$$P[\varphi_L < \Delta] \leq P\big[\,\varphi_L < \varphi(1\{\pi \in \bar{\Pi}\} \cdot L^{1/2}c)\,\big] \leq P\left[\inf_{\pi\in\bar{\Pi}} L^{-1/2}\mathscr{T}_L(\pi) < c/2\right].$$

The right-hand side of this inequality converging to zero by (C.24). Condition (B) therefore follows, and so does claim (ii). □