Title
Three Essays on Labor Economics

Author
Li, Fanghua

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Three Essays on Labor Economics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Economics

by

Fanghua Li

2018

© Copyright by

Fanghua Li

2018

ABSTRACT OF THE DISSERTATION

Three Essays on Labor Economics

by

Fanghua Li

Doctor of Philosophy in Economics

University of California, Los Angeles, 2018

Professor Moshe Buchinsky, Chair

This thesis contribute towards the understanding of labor economics and applied econometrics; the thesis is made up of three chapters.

The first chapter explores the causal effect of parents' social capital on the intergenerational occupational inertia in addition to individuals' labor market outcomes. A new data extract was constructed by re-weighting and combining the Panel Study of Income Dynamics (PSID) and the Survey of Income and Program Participation (SIPP) to correct the selection biases induced by children's endogenous moving behaviors post-graduation. By exploiting the recent technological revolution and the resulting changes in occupational skill compositions measured by Dictionary of Titles (DOT) and its successor O*NET, it became possible to isolate the effect of inherited social capital from inherited human capital through a regression discontinuity design. Besides, a correction of the selection bias induced by the social capital advantage through children's occupational switching patterns after the first jobs was made. The results indicate that around 30% of individuals choose the same occupation as their parents for their first job; such people rely more on their parents' social connections in job hunting. Also, they enjoy a positive wage premium of about 5% of the percentile ranks of annual labor income for entry-level jobs but this positive effect fades away in the long-run.

The second chapter studies the estimation and inference of nonlinear econometric model when the economic variables are contained in different datasets. We show that the unknown structural parameters of interest can be possibly uniquely identified if there are some common conditioning variables in different datasets. The identification result is constructive,

which enables us to estimate the unknown parameters based on a simple minimum distance (MD) estimator. We study the asymptotic properties of the MD estimator and provide inference procedure. A simple model specification test on the key identification conditions is also provided.

The third chapter provides an application example of the method developed in the second chapter. It is a long-standing problem in the empirical research that the economic variables are contained in different datasets. One well-accepted solution to this problem is the imputation method, which serves as a crucial step in the seminal work, Blundell, Pistaferri, and Preston (2008) studied the dynamic relationship between consumption and income, with consumption data from CEX and income data from PSID. In this chapter, we first prove that the imputation method is biased because they are significantly different from those based on true data, which is the newly available PSID from 1999 which includes both consumption and income data. Furthermore, we investigate the finite sample performance of our new method with this new PSID data and show that our method delivers comparable results with those based on the true data. We conclude that the imputation gives largely biased estimation compared to the real data results and the new estimator developed in Chapter 2 performs better.

The three chapters share the same interests in the long-lasting question that how we can deal with the situation in which the economic variables or the study population is contained in different datasets. The first chapter starts off from the simplest scenario that the data set is complete in terms of variables but biased in terms of representativeness. The other two chapters deal with the other more difficult and more usual case that the data set is incomplete in terms of economic variables. We not only contribute methodologically by providing a new estimator but also implement the method in an important application case and discuss the implications.

The dissertation of Fanghua Li is approved.

George M De Shazo

Shuyang Sheng

Dora Luisa Costa

Moshe Buchinsky, Committee Chair

University of California, Los Angeles

2018

To my parents . . .

for—among so many other things—

always believing I could do this

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Firstly, I'm indebted to my advisor Prof. Moshe Buchinksy for the continuous support during my Ph.D. studies, I am particularly thankful for your patience, motivation, immense knowledge and guidance which went a long way in helping me carry out this research work and concluding it; especially the help I received last summer. I cannot imagine concluding this work on my own without your guidance, I am forever grateful.

Also, special acknowledgment and enormous respects are reserved for the rest of my thesis committee: Prof. Dora Luisa Costa, Prof. Shuyang Sheng, and Prof. George M De Shazo, for their insightful comments and encouragement. I'm also highly grateful to Adriana Lleras-Muney, Katheleen McGarry, Maurizio Mazzocco, Till von Wachter, Michela Giorcelli and Rodrigo Pinto for their insightful discussion as well as all the participants of Applied Microeconomics Seminar at UCLA for listening and their helpful comments.

The last two chapters of the dissertation are based on a joint research project with Prof. Zhipeng Liao and my advisor Prof. Moshe Buchinsky. I'd like to thank them for their guidance and contributions towards my work, in particular, Prof. Liao for helping me resolve various econometrics modeling and programming issues.

I also want to thank my fellow classmates for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last six years.

Lastly, my deepest appreciation goes to my parents for their unconditional love and support during my studies and at every stage of my life. I'm also grateful to my undergraduate advisor Prof. Xinye Zheng, who has always been there in difficult moments.

VITA

2010	B.S. (Applied Mathematics), B.A. (Economics) Renmin University of China
2012	M.A. (Economics) Renmin University of China
2013	M.A. (Economics) University of California, Los Angeles
2013	C.Phil. (Economics) University of California, Los Angeles
2013–2017	Teaching Assistant, Research Assistant Department of Economics, University of California, Los Angeles
2017–2018	Dissertation Year Fellowship Department of Economics, University of California, Los Angeles

CHAPTER 1

The Role of Social Capital and Intergenerational Occupational Mobility on Labor Market Outcomes

1.1 Introduction

Economists and social scientists have long been interested in the persistence between parents and children's socio-economic outcomes, or the intergenerational immobility.[1] To explain this phenomenon, the current literature focuses on the channels which contribute to children's human capital increase through the automatic genetic transmission, education or/and other hereditary endowments associated with the early years, like family values, preferences, and other inherent characteristics like motivation. This paper proposes another possible explanation that it could reflect the inherited social capital instead of the human capital advantage, particularly the role of networks or direct control in the hiring process to influence employment opportunities in the labor market. More specifically, one individual could benefit by choosing the same or a related job to that of their parents, especially for the individual's first job, which facilitates the transition from schooling to work. By doing this, they rely upon the contacts and information their parents may share with them or on the direct or indirect control parents may have in the hiring process at the chosen place of employment. This benefit, sometimes referred to as nepotism or patronage in the literature, is studied only in few research with small scale data in specific occupations (e.g., Xu, 2017; Bertrand, 2009; Bennedsen et al., 2007) without taking the endogenous occupational choices into consideration or generalizing to the entire labor market.[2] To evaluate this causal effect in the whole labor market is challenging due to the selection biases in the current available data and the endogeneity of children's occupational choices relative to their parents, since intergenera-

[1]See Solon's 1999 and Black and Devereux's 2011 handbook chapters for a review

[2]"Patronage" refers to the discretionary appointment of individuals to governmental or political positions (WebsterâĂŹs II New College Dictionary 1995).

tional occupational correlation may simply result from the intergenerational correlation in innate ability.

The objective of this paper is to inform a discussion surrounding the above-described issue. More specifically, in order to correct for selection biases induced by children's endogenous moving behaviors post-graduation of the currently available data sets, I construct a new data extract for U.S.'s labor market by reweighting and combining *Panel Study of Income Dynamics (PSID)* and *Survey of and Survey of Income and Program Participation (SIPP)*. With this new data extract, I'm the first in my knowledge to descriptively document the stylized facts about the intergenerational occupational transmission and the linkage of which to the inherited social network and lifetime wage inequality.

My main identification strategy relies on the recent technological revolution starting from the late 1980s to the beginning of the 20th century. To be specific, the innate ability of people can be described as a continuously distributed multidimensional random variable, and occupations can be viewed as a categorical variable; hence, individuals' occupational choices can be seen as the assignment rules from the continuous variable to the discrete variable, which is decided by the technology, occupational wage level, and other market factors. I show in the paper that the technological revolution changes the assignment rules significantly enough to give extra variation to disentangle the intergenerational correlation to innate ability and occupations. Also, around the old innate ability cutoffs, there is a discontinuity of whether they will stay in the same occupations as their parents, which enables a Regression Discontinuity design.

Another identification strategy lies in the fact that if childrenâĂŹs true productivity is private information when they enter the labor market, and they stay in the same occupations as their parents only for the social network benefit instead of their true innate ability, as their true abilities are revealed, those people have a higher probability of switching occupations later on. I show this relationship in the paper and these occupational transitions give another variation for identification.

In this line of research, how finely occupation is defined is always believed to affect the results significantly. For robustness, I use both the the broad category of *1990 Census Occupational Classification System (COC)* with 66 occupations and also the detailed category (3-digit) of *1980 Standard Occupational Classification (SOC)* with 242 occupations. To construct comparable measurement for the occupational specific skills in different time periods, I exploit the detailed occupational descriptions in *Dictionary of Occupations (DOT)* and its successor *Occupational Information Network (O*NET)* and create a crosswalks between them two.

My empirical analysis yields the following results. First, for the broad category of COC, around 30% of children choose the same occupations with their parents as their first jobs, the this ratio is relatively higher in years of depression. Besides, the relationship of the occupational inheritance rate and the occupational income standing is a "U" shape, suggesting higher inheritance rate for jobs with high prestige. Secondly, by OLS estimation, staying in the same occupations in first jobs give children around 5% wage premium on the log of yearly earing, while the RD and IV strategies which exploits the variation in occupational specific skills and individuals' occupational transitions gives us around 10% wage premium, which is comparable to return of education. Third, in the long run, the wage premium on the first jobs fade away with time, and the people who choose the same occupations as their first jobs have higher income risks, risks of getting fired and instability in their lifetime. Those results suggest that the social network benefit is not persistent in the long run and leads to a misallocation of talents through distorts individuals' optimal occupational choices when they just enter the labor market.

The contribution of the paper is three-fold. First, since Gary Solon's 1999 Chapter in the Handbook of Labor Economics, the literature of intergenerational mobility has placed increased emphasis on the causal mechanisms that underlie this relationship. The current literature focuses more on education, which is emphasized by almost all the literatures on intergenerational income and social class or status persistence (e.g. Solon, 2004), or/and the automatic genetic transmission or other hereditary endowments associated with the early years, like family values, preferences, and other inherent characteristics like motivation. Those papers emphasize on the human capital increase due to family background, and suggest that children from better-off families and parents with better abilities and attributes being better equipped with intellectual power or earning power to eventual success. As stated in the most recent handbook chapter Black and Devereux (2011), "these phenomena can hardly be regarded as inequality and the policy implications are unclear." This paper explores another channel through the inherited social network instead of human capital increase, and is, to the best of my knowledge, the first study that examines both the short term and the long term casual effect of intergenerational occupational correlation on children's labor income.

Second, this paper sheds new light on the estimation of social network effect. Ever since the seminal work of Granovetter (1973, 1995), there is growing interests on the social networks in labor markets. Exploration in a large number of studies document positive effects for a variety of occupations, skill levels, and socioeconomic backgrounds. But the simultaneous nature of social network and individual characteristics and working experiences make it hard to peel out the impact of social network on income. In this paper, inspired by the Ioannides (2015)'s distinguishment between connections that are the outcome of deliberate decisions by individuals and connections being given exogenously and beyond an individual's control, I try to isolate this simultaneous relationship by discussing later on, which is the initial social network one inherited from their parents. My work is the first to use large scale individual data to show the effect of inherited social network generally exists in the labor market.

Finally, this paper is also related to another growing body of literature that emphasizes the importance of occupational matching for worker outcomes, e.g. Kambourov and Manovskii (2009a, 2009b), Gathmann and Schonberg (2010), Groes et al. (2010) and Antonovics and Golan (2012). If one chooses his/her occupation for the advantages he/she can get from their parents instead of his/her own talent, this talent-occupation mismatch has implications for economic efficiency if the talents of those are under-developed or not fully utilized, as those people will not live up to their productive potential.

The rest of the paper is structured as follows. Section 1 describes the data limitation in the currently available data sets and the construction of the new data extract. Section 2 describes the measurement of parents' occupations and children's occupations and the sample selection. Section 3 presents the stylized facts of intergenerational occupational inheritance and lifetime wage inequality. Section 4 analyzes the mechanism through which the inherited social network affects children's occupational choice and the wage premium. Section 5 provides a conceptional theoretical model and examines the propositions with the real data. Section 6 presents the empirical framework and discusses the identification strategy. Section 7 presents robustness checks. Section 8 concludes.

1.2 Data

In this part, I'll first show that the commonly used data sets in the literature of intergenerational mobility is problematic due to either severe measurement error or selection bias resulting from designs of the surveys. And then I describe the new data extract I create for the following study.

In surveys with both children's and parents' working information available, there are two types in terms of how parents' working information are obtained. The first type is retrospective data, in which only children are the targets of the survey and questions about their parent, especially the parents' main occupation and industry, are parts of the survey, represented by *International Social Survey Programme (ISSP)*. [3] And the other type is self-reported data in which both parents and children are the interviewees and they are asked to report their own job information during the survey, among them, the most commonly used is the longitudinal house survey *Panel Study of Income Dynamics (PSID)*, which began in 1968 and continuously follows and collects the information of the individuals and their descendants from the original sample families.[4] I will discuss the data limitation of these two kinds of data separately.

1.2.1 Problem of the Retrospective Survey Data

The first commonly used data set is the survey data with retrospective data on parents' working information. In those retrospective surveys, people would be asked about their parents' main occupation and industry, and in most cases only one occupation/industry would be listed. Hence the accuracy of those children-reported information is crucial with this kind of data.

In PSID 1997 to 2015 surveys, retrospective questions about their parents were added into the survey for all individuals in the sample, which gives us an opportunity to test the validity of these data since for the people who are the second or third generation of PSID survey, their parents' lifetime working information is also available.[5] I find that only 37.8% people reported at least one of the occupations their parents' ever took, and if we consider that fact

[3] See detailed introduction of ISSP in `http://www.issp.org/menu-top/home/`

[4] See detailed introduction of PSID in `https://psidonline.isr.umich.edu/`

[5] The 1-digit occupation information of head's father's is also available from 1970 to 1996. The corresponding question is " What was your father/mother's usual occupation when you were growing up?"

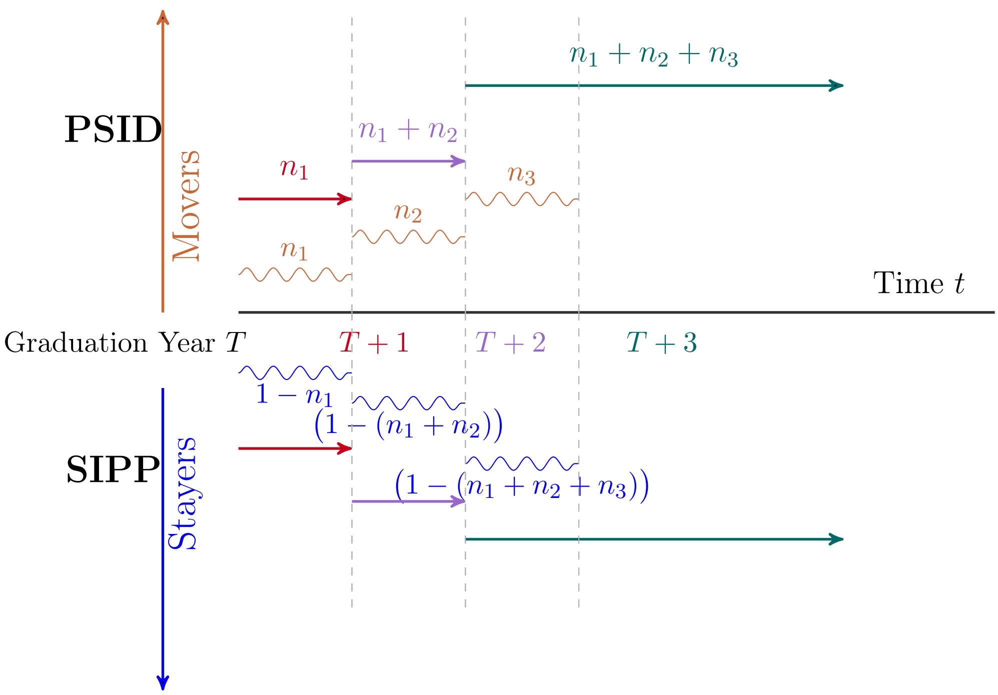
that parents' may have more than one main occupation, the accuracy rate is even lower. [6] This low rate of accuracy would cause significant measurement errors and induce bias.

1.2.2 Problem of PSID and SIPP

For the other type of data, even though we don't have the accuracy problem as we mentioned above, we will have the selection bias problem due to the survey design.

First, for the longitudinal survey data like PSID, even though the design of PSID is to track everybody in the household, but the detailed employment information (occupation/industry etc.) is only available for the heads and wives (in certain years). In other words, when one child graduates from school, we will only observe his/her working information if he/she moves out his/her parents' home and become a head/wife.

Figure 1.1: Moving Behaviors after Graduation



Notes: assume that n_1 percent of individuals move out within one year of graduation, n_2 percent of individuals move out in the 2_{nd} year, and n_3 percent of individuals move out in the 3_{rd} year and so on.

In Figure 1, I depict the individuals' moving behaviors after graduation. We can see that, due to PSID's survey design, we can only observe n_1 percent of individuals' working information

[6]See Appendix Table B.1 for the sample construction and detailed statistics.

from the 1_{st} year of graduation, and (n_1+n_2) percent of individuals' working information from the 2_{nd} year and so on. In other words, if our interest is to study children's labor outcomes in the first years after entering the labor market, the sample of PSID is constrained to those who move out their parents' home relatively earlier which is endogenously selected and highly biased to highly-educated people, since only 10.62% of individuals who graduate at their 15-year-old would move out within the 1_{st} year while that ratio for those who graduate at their 26-year-old is 100% (Appendix Table B.2). [7]

SIPP is another longitudinal data set which will follow people only about 2.5 to 5 years, and all household members age 15 years and older are interviewed by self-response. Differently from PSID, we will only observe both parents' information and children's information if they stay in the same household. [8] As in Figure 1, from SIPP, we can obtain a sample of those $(1-n_1)$ percent of individuals who don't move out. [9] This is also true for census data. Hence either data set separately would induce bias due to the selection sample.

1.2.3 New Data Extract by Data Combination

To solve the selection bias caused by children's moving behavior, I construct a new data extract by reweighting and combining PSID 1986 - 2013 rounds and SIPP 1986-1987, 1990-1993, 1996, 2001, 2004 and 2008 panels. [10]. In this part, I'm going to show that both data sets are nationally representative of the whole U.S. population.

1.2.3.1 Representative of PSID and SIPP

Although both data sets take representative U.S. families as interview units and are conducted by well respected organizations with delicate survey designs and implementations,

[7]Detailed age profile of moving out schedule is provided in Appendix Table B.2

[8]See detailed introduction of SIPP in https://www.census.gov/programs-surveys/sipp/about/sipp-introduction-history.html

[9]SIPP also follows individuals if they move out, but the initial sample would be constrained to those who live with their parents, and in this study, I will concentrate with those who keep staying in the same house with with parents.

[10]The span of SIPP before 1993 is all 2.5 years, while around 4.5 years afterwards. The reason I discard 1988 is that the existing data dictionary is problematic, and for 1989 is that there is no available fully labeled panel data. Before 1986 panels are discarded because they don't have duration information of individuals occupations.

there are still some differences among them that should be taken care of when combining them together.

Unit definition

The definition of the head of the household in the SIPP is the person or one of the persons who owns or rents the unit; this definition is slightly different from the one adopted in the PSID, where the head is always the husband in a couple. Following most previous research, our analysis makes the two definitions compatible. In our study, I'll use both the heads and wives' data, hence the definition of who is the head is not crucial in this study.

Sampling issue of PSID

The genealogical and longitudinal designs of PSID make it a unique resource for addressing particular questions, they nevertheless draw concerns about the sample representativeness since the PSID sample keeps being replenished through births and marriage. Fitzgerald et al.(1998) carefully studied this issue and found no strong evidence of distortion of the representativeness through 1989, and showed with considerable evidence that PSID's cross-sectional representativeness has remained roughly intact from the sample replenishment. Besides, a thorough subsampling procedure was conducted in 1997 in react to the growing size resulted from family splits. Through setting aside entire linkages to a 1968 PSID sample family and adding nationally representative sample of immigrant households and individuals that would not be eligible for PSID under the original 1968 sample recruitment and sample family following rules, this subsampling procedure managed to exemplify the PSID sample while keeping the intergenerational ties in the core panel. Hence, in this study, we undertake previous research results and assume that the PSID data is nationally representative with family weights.

1.2.3.2 Comparison of Descriptive Statistics

Table 1 compares the individuals who don't move out after graduation in these two datasets in terms of average demographic and socioeconomic characteristics for selected years: 1986,

Table 1.1: Summary statistics of PSID and SIPP in selected years

	1986		1993		2008	
	SIPP	PSID	SIPP	PSID	SIPP	PSID
Male	0.612	0.601	0.618	0.616	0.590	0.587
Married	0.093	0.101	0.103	0.096	0.097	0.090
White	0.865	0.881	0.840	0.856	0.804	0.791
Age	22.616	23.289	22.898	22.507	22.540	22.149
HS dropout	0.143	0.149	0.145	0.139	0.189	0.183
HS graduate	0.410	0.393	0.351	0.331	0.330	0.310
College dropout	0.363	0.381	0.412	0.444	0.401	0.442
College graduate	0.080	0.077	0.092	0.086	0.081	0.065
Northeast	0.220	0.205	0.206	0.201	0.186	0.200
Midwest	0.257	0.263	0.294	0.308	0.313	0.331
South	0.325	0.332	0.343	0.318	0.413	0.388
West	0.197	0.201	0.157	0.173	0.089	0.085

1993 and 2008[11]. By pooling these two samples, we can test the difference of each variable separately and all of them are insignificantly different.

Figure 2 depicts the age profile of rate of living with parents, from which we can see that the moving behavior captured in both data sets are also comparable.

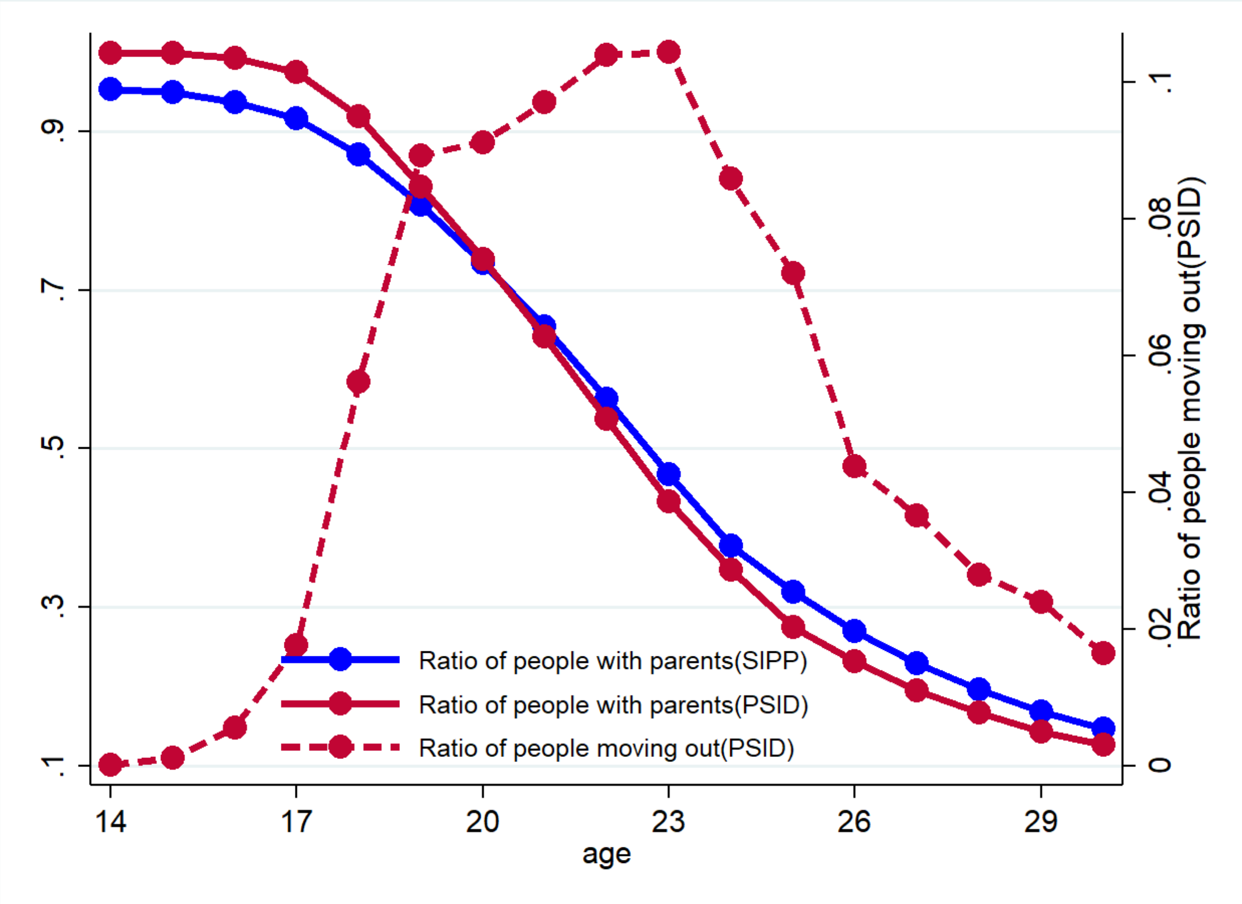
1.2.3.3 New Data Extract

By the above analysis, we can now combine these two data sets by taking PSID as the main data set and pooling together with SIPP after reweighting. The adjusted SIPP weight is the original weight adjusted by the summation of weights in the corresponding synthetic groups defined by education level and years after graduation. [12]

[11]Those three years are the beginning year of SIPP, the last year before SIPP reconstruction, and 2008 is the last panel available. Those variables are the only demographic variables consistently available in PSID individual data index. In PSID, there are only detailed information about the heads and wives, but for the rest of the household, only those basic variables are covered.

[12]Theoretically, finer synthetic groups including age, gender, region and marital status can be constructed, but the sample size of each group would be too small and by testing, these demographic variables are comparable.

Figure 1.2: Age profile of moving out rates



And the summary statistics in the pooled data set is in Table 2.

$$W^{SIPP}_{adjust} = W^{SIPP} * \frac{\sum_G W^{PSID}}{\sum_G W^{SIPP}} \tag{1.1}$$

Table 2 reports the descriptive statistics for the new data extract and two data sets separately. We can see that PSID and SIPP cover quite different subsample in population. The individuals in PSID are significantly higher educated, hence have higher annual labor income and also higher rate of being married.

Table 1.2: Summary Statistics for the New Data Extract

	Full Sample in the New Data Extract				PSID only				SIPP only			
Variable	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
(log) Annual Labor Income	7.746	1.602	-2.926	12.725	8.989	1.086	-2.926	10.295	6.652	1.109	-1.708	12.725
HS Dropout	0.159	0.365	0	1	0.164	0.370	0	1	0.153	0.360	0	1
HS Graduate	0.405	0.491	0	1	0.383	0.486	0	1	0.429	0.495	0	1
College Dropout	0.327	0.469	0	1	0.249	0.433	0	1	0.415	0.493	0	1
College Graduate or Higher	0.109	0.312	0	1	0.204	0.403	0	1	0.002	0.049	0	1
Years of Experience	6.126	4.065	1	61	15.326	9.417	1	59	3.977	2.598	1	61
Female	0.442	0.497	0	1	0.477	0.499	0	1	0.408	0.492	0	1
White	0.715	0.451	0	1	0.614	0.487	0	1	0.811	0.391	0	1
Northeast	0.175	0.380	0	1	0.138	0.345	0	1	0.214	0.410	0	1
Midwest	0.269	0.444	0	1	0.248	0.432	0	1	0.292	0.455	0	1
South	0.418	0.493	0	1	0.454	0.498	0	1	0.379	0.485	0	1
West	0.138	0.345	0	1	0.160	0.366	0	1	0.116	0.320	0	1
Married	0.535	0.499	0	1	0.609	0.488	0	1	0.465	0.499	0	1
	61,840				11,290				50,550			

Notes: Columns 1, 2, 3, and 4 present, respectively, mean, standard deviation, minimum and maximum of characteristics and outcomes of all the 61,840 individuals in the new data extract. Columns 5$\tilde{8}$ and 9$\tilde{1}$2 report those statistics of the same variables, separately, for the PSID sample in the data extract and the SIPP sample in the data extract. For the total 50,550 in the SIPP data, there are 1,816, 1,852, 3,775, 2,156, 3,153, 2,765, 7,773, 5,394, 10,466, 11,400 from the 1986, 1987, 1990, 1991, 1992, 1993, 1996, 2001, 2004 and 2008 SIPP panels.

1.2.4 Description and Classification of Occupations

1.2.4.1 Occupational Classification

As stated in the handbook chapter Black and Devereux (2011), the intergenerational occupational correlation would be quite different if different level of classification system is used. In our study, we don't want the classification too broad hence cannot capture the occupational specific characteristics and in the meantime not too detailed to put for example orthodontist and periodontist in two different occupations. With this aim, the main body of this paper uses the broad category (altogether 66) of the *1990 Census Occupational Classification System* (the list of which is listed in Appendix Table B.3, and the crosswalks of different data sets to this classification is in Appendix Table C.4). I also use the detailed 3-digit *1980 Standard Occupational Classification* for robustness check.

1.2.4.2 Measuring the Occupational Specific Skills

The U.S. Department of Labor released the first edition of the DOT in 1939 to "furnish public employment offices with information and techniques to facilitate proper classification and placement of work seekers."[13] Although the DOT was updated four times in the ensuing five decades (1949, 1965, 1977 and 1991[14]), its structure has been little altered. Based upon first-hand observations of workplaces, DOT examiners using guidelines supplied by the Handbook For Analyzing Jobs rate occupations along 44 objective and subjective dimensions including training times, physical demands, and required worker aptitudes, temperaments, and interests.

In 1998, the National Center for O*NET Development's O*NET database[15] replaced the US Department of Labor's Dictionary of Occupational Titles (DOT) as the primary source of information about US job characteristics. Since then, O*NET has gathered information on hundreds of variables for more than 800 SOC-defined occupations. Prior to 2003, O*NET acquired its data from surveys administered to job analysts and experts. Beginning in 2003,

[13] U.S. Department of Labor (1939:xi) as quoted in Miller et al (1980).

[14] In 1991, based on the time of update, among the 12,741 detailed occupations, 80.76% were not updated, and only 6% were updated after 1986. Hence this revise cannot fully qualify as the up-to-date description in the 1990s.

[15] data are available on website: $https://www.onetcenter.org/db_releases.html$

however, information has come from job incumbent surveys. Compared with the predecessor, the O*NET, offers potentially more up to date information on occupational characteristics by updating twice a year. O*NET categorizes its variables into six distinct surveys, and we here are interested in Abilities, Skills, Knowledge, Working content and Activities surveys. These surveys ask respondents to evaluate the importance of particular abilities (skills, disciplines, activities) required by his/her current job on a scale of 1 to 7.

In this paper, I match the two job description data sets and create longitudinal descriptions of the occupations. From the view of the demand side of the labor market, each job can be describes with a bundle of specific *Activities and Working Contents*, while from the view of the supply side of the labor market, each worker offers a bundle of *Abilities* (Cognitive skills, physical skills), *Skills* (language, communication, analytical skills etc.) and *Knowledge* (physical, history etc.) to match with the demand of jobs. Hence each job should be characterized by pairs of activity-ability. DOT provides only *Aptitude*, *Physical Demands*, *Temperaments* and *General educational development (GED)*, while O*NET provides much more detailed characteristic descriptions. In order to construct time-series descriptions of occupations, I was motivated by past literature and common practice, and aggregate the abilities and activities to several categories, including interactive (or communication) skills, analytical (or reasoning) skills, quantitative (or analytical) skills, routine cognitive skills, routine and non-routine manual tasks and physical abilities. (The detailed matching table is in Appendix Table B.5)[16] Eventually, we can aggregate those categories one step more into broader groups, for example, we can group into cognitive skills, routine skills, manual skills and physical skills as Autor and Dorn (2013). [17] The detailed and major skill aggregation of DOT and O*NET is listed in Table 3.

[16]Some parts of this aggregations are from Autor, Levy and Murnane (2008) and Peri and Sparber (2010)

[17]I'd like to clarify first that I use the term "routine" following Autor, Levy, and Murnane (2003), but it meant the skills required for the "Routine-cognitive" tasks. For example, the description of "Clerk Perception" is "The ability to see detail in manuscript or tabular material. The ability to observe differences in copy, to proofread words and numbers, and to avoid misreading numbers in arithmetic computation.", which requires scrupulousness and patience.

Table 1.3: Aggregation of skills Definition

Detailed	**Major**
Non-routine Cognitive	Communication Interactive Analytical Quantitative
Routine	Routine Cognitive Routine Manual
Manual	Non-routine Manual
Physical	Strength Body Flexibility Visual Auditory

1.2.5 Sample Selection Problem

In the study about intergenerational occupational relationship, it's crucial to measure parents' main occupations correctly and completely. Hence, ideally we need the whole working history of parents. But even in the PSID with 60 years of history, the sample size of children with the whole working history of parents is extremely small, let alone the SIPP only reports parents' current job and the last job.[18] [19] Hence in this part, I study people's work transition behavior, especially the occupational transition and show that individuals' occupations are relatively stable after 28-year-old, and conclude that even if only the later half of the working history is available, we can still obtain parents' main occupations in life without too much information loss.

1.2.5.1 Occupational Transitions

First, we define the annual occupational transition rate as the ratio of individuals who work in a new occupation in the given year.

[18] Appendix Figure B.1 depicts the sample sizes of individuals with different length of coverage in PSID

[19] In SIPP, the core files only report the current jobs and the last jobs are in the *Employment History Module*

$$\text{occupational change rate}_t = \frac{N_t^{\Delta O}}{N_{t-1}^E} \tag{1.2}$$

But the above definition didn't take into consideration of people who change into an old occupation. This is important in the sense that, if people's occupational change behavior is like shopping, maybe he would have change back to the original occupation after trying out other options. To control for this scenario, we need the employment history records for adjustments. Figure 3 depicts the age profile of occupation changing rates on 1-digit and 2-digit levels for people who were recorded through 2008 to 2013.[20] And we also use the 5 years history record of employment to adjust for those who got a new job in 2012 which was different from the previous jobs but the same with an old job he/she took within 5 years. We can see that the occupational transitional rate is especially high for people who just started their jobs (over 20%) and drop to 5% in the mid-30s and 2% after 50. For those who start working before 20 (20%) their job changing rate is as high as 50% on yearly base and 80% of those job changes come with occupation changes.[21] 50% people working in their early 20s when they just graduated from college, their job changes rates and occupation changes rates are also pretty high.

With the large sample size of SIPP, we can statistically describe the duration of people staying in the same line of jobs, which resembles our exercise before of 1-digit/2-digit occupation groups.[22] Figure 4 depicts the average length of time in the same line of work for different age bracket. Considering the individual heterogeneity, I also depict the confidence interval of the average length. We can see that, the confidence interval becomes wider after 55, revealing the situation that people pick up different kinds of jobs after retirements. But before retirement, we can see that the duration of occupation grows steadily with age, which is consistent with the occupation changing rate we calculate before. We also generate the average occupation starting time for different ages. we can see that occupation people started in their 30's last till 55, which is the common retirement time. In other words, the jobs we see people doing in their 30s, are the ones that have very high changes of lasting the

[20] The reason I use the 2008 Panel of SIPP is because it follows the people the longest compared all other SIPPs, which is 5 years and give us enough historical information to control for the old jobs. The reason I choose SIPP over PSID is for the large data sample and the same length of months covered in SIPP.

[21] Here we only consider the people who graduated and take full time jobs

[22] In SIPP, there is a question phrased as " Considering ...'s entire working life, how many years has ... been in this occupation or line of work? "

major working time, or the jobs that we see people doing in their 40s and 50s, have high probability of lasting 10-20 years. Hence we can define the jobs in the range 28 to 64 as the main jobs. In other words, even if we only have a relatively small chunk of available working history, it's enough to define the main jobs. Hence I choose the sample to be those children with their parents' before-retirement working information available.

Figure 1.3: The Age Profile of Occupational Transition

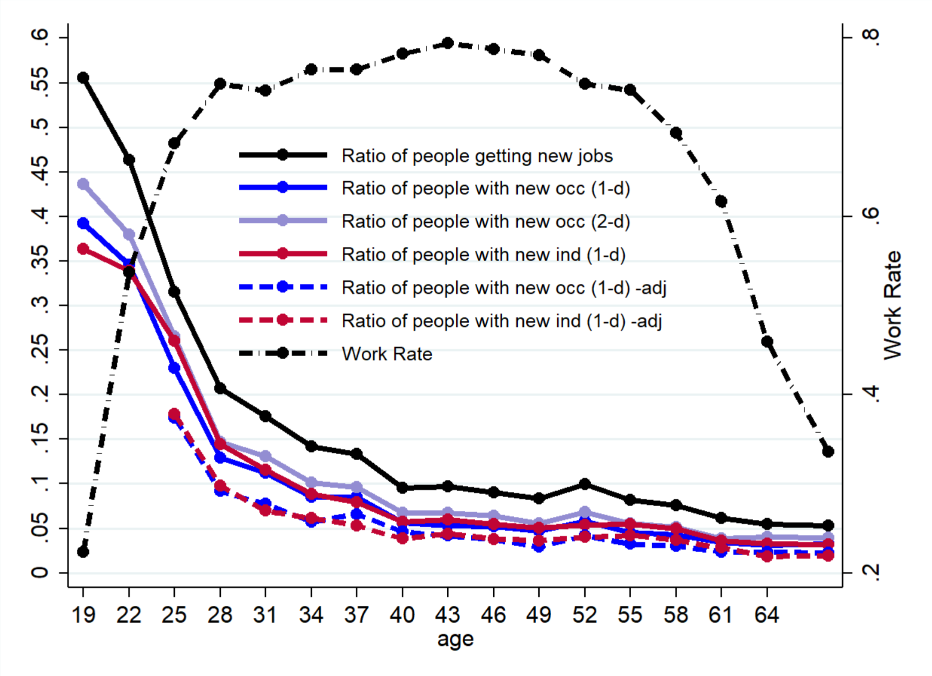
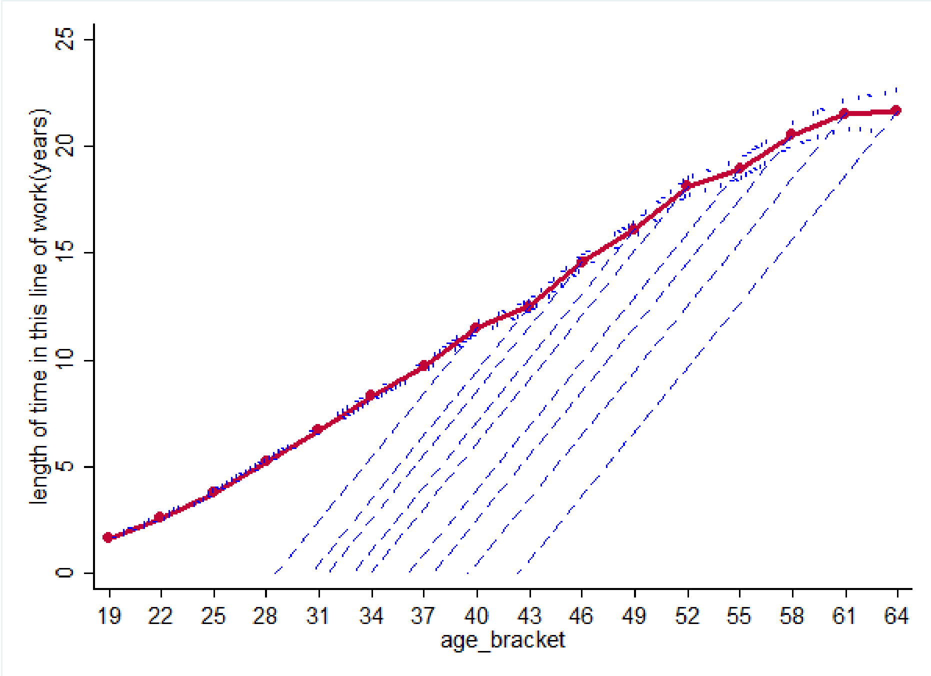


Figure 1.4: The Average Duration of Occupations



Notes 1: The Figure 3 is based on data in SIPP 2008 Panel and Figure 4 is based on all panels in SIPP. In both of those two figures, the "age" on the x-axis means 3-year age group, and each number means the starting age of this group, for example, "19" means 19-year-old to 21-year-old.

Notes 2: In Figure 4, the read line the average length of years that people spend on the current occupation for different ages. The two dotted blue lines around the red line show the confident intervals of the estimation. And the dash lines shows the average starting age of occupations for different ages, calculated by the age minus the average length of duration. For example, for age 55, the average duration is about 19 years, so the average starting age is around 36-year-old.

Notes 3: In Appendix Figure 2 and 3 I show that the job and occupational transitional rate follows the same pattern for different cohorts and along the time, except the 2008 recession, the occupational transitional rates is comparable over the years.

1.3 Stylized Facts

1.3.1 Model Specification

Following the pioneering work by Solon (1992) and Zimmerman (1992), the existing literature estimate the intergenerational income persistence by the Galton-Becker-Solon regression:

$$y_c = \alpha + \beta y_p + u_s$$

in which y is the log income or percentile rank of income.

The primary objective of this paper is to empirically estimate the causal effect of one's career choice in response to his/her parents on labor market outcomes, controlling for the family income and other family background factors. Hence, the main regression equation in this paper is:

$$y_c = \alpha + \tau \mathbf{1}\{O_C == O_P\} + X_c \tau + \beta y_p + u_s \tag{1.3}$$

in which $\mathbf{1}\{O_C == O_P\}$ is a dummy variable which equals 1 if children's occupation is the same with their parents', and y_c can be $log(income)$ or $rank(income)$ as in the Galton-Becker-Solon regression, or some other labor market outcomes, like unemployment rate or variance. And X_c are other variables that already proved to influence one's labor market outcomes, like education level, and other family background variables.

The main threat to identification is the endogeneity of the occupational choice variable $\mathbf{1}\{O_C == O_P\}$. To be specific, there are two kinds of bias, one is the genetic correlation with their parents or other hereditary/occupational training from parents, that the children choose the same occupation with their parents because they are more talented or suitable for the occupations. And the second one is the presence of individual heterogeneity and self-selection giving rise to a sorting gain, which is another form of selection bias in the standard estimator for the causal effect. In this part, I'll show the OLS regression results and some other stylized facts about the intergenerational occupational correlation. And I'll talk about identification in the following sections.

1.3.2 The effect on the entry-level labor markets

First, I concentrate only for the new graduates and their first jobs. Since in this newly constructed data set, we can fully identify the starting and ending time (or still working for) of each job, we can then sample out those people whose first job information is available, and construct the indicator of whether one chooses the same occupation with their parents' as their first job. Also based on the parents' starting and ending time of each job, we can identify whether parents' jobs are their major jobs. In the end, we limit our sample to those who has major' job informations of parents, and construct the following indicator:

$$\mathbf{1}\{O^C_{first} == O^P_{major}\}$$

Estimation of Intergenerational Occupational Correlation

In this part, I estimate the intergenerational occupational correlation for U.S. along with time and also for occupation with different income standings.

Figure 5 depicts the rate of new graduates who choose the same job as their parents (currently have or ever worked in as the main occupation). We can see that, on average, around 30% new graduates choose their parents occupations as their first jobs. In the literature of intergenerational occupational correlation, either the very broad categories of occupations are used or the firm level information is used, and children's occupations are not constrained only in first jobs (e.g. Hellerstein and Morrill, 2008 and Ferrie, 2005)[23]. So this paper cannot be compared with the current literature due to different occupational classifications. Also the ratio peeks when the unemployment level is high, for example year 1992 and 2008, even though it is not perfectly correlated with the unemployment level in U.S. in time, considering the choice of occupation is usually predetermined several years before graduation especially for people with college or even higher education level.

The second figure in Figure 5 depicts people's occupational inheritance behavior by parents' occupational percentile rankings. We can see that, for the very low income group, children are stuck at the low pay jobs. And the whole pattern fits out theory of the U shape inheritance rate (The probit regression results in Appendix Table C.6).

Besides, in order to compare U.S. with other countries, I employ International Social Survey Programme or ISSP, which is an annual program of cross-country surveys on various topics

[23]For example in Hellerstein and Morrill (2008), all jobs are grouped into only 6 categories, including "Managerial and Professional Specialty", "Technical, Sales, and Admin. Support", "Service", "Farming, Forestry, and Fishing", "Precision Production, Craft, and Repair" and "Operators, Fabricators, and Laborers". And in Ferrie (2005), the grouping is even broader, including two "white collar" and "blue collar". Hellerstein and Morrill (2008) show that, in recent cohorts, about 30% of sons and 20% of daughters work in the same occupation as their father.

of social science research, and the topics of which in the year 1999 and 2009 were social inequality, covering 25 and 41 countries and territories respectively. In these two separate years, ISSP provides both respondent's occupation and respondent's parents' occupation, with which we can calculate the rate of occupation inheritance for each region in that specific year, defined as the rate of respondents sharing the same occupation with either hisher mother or father[24] [25]. The occupation classification both follows ISCO88, with 10 major groups (1-digit level), 28 sub-major groups (2-digit level), 116 minor groups (3-digit level), and 390 unit groups (4-digit level). Each survey was carried out around the same time within one year in all the surveyed regions, making the results comparable across countries. In Figure 6(a), I depict the occupation inheritance rate with 1-digit, 2-digit and 4-digit levels of occupation with respect the per capita GDP level in 2009. As we can see, no matter which level we use, the more developed an economy was, the lower the occupation inheritance rate. For China, the 1-digit inheritance rate was as high as 0.56 while for Norway, the rate was close to 0.2. [26]. Since we are interested in the occupational inheritance only here, we want to exclude the effect of the intergenerational education level correlation. Hence, I depict in Figure 6(b) the normalized inheritance rate for 2-digit level by subtracting the inheritance rate by a theoretical inheritance rate if everybody chooses their occupation randomly [27]

[24]The corresponding questions are: *What is your first occupation?* and *What is your current occupation?*

[25]The corresponding question is: *When you were 14-15-16 years old, what kind of work did your father mother do?*

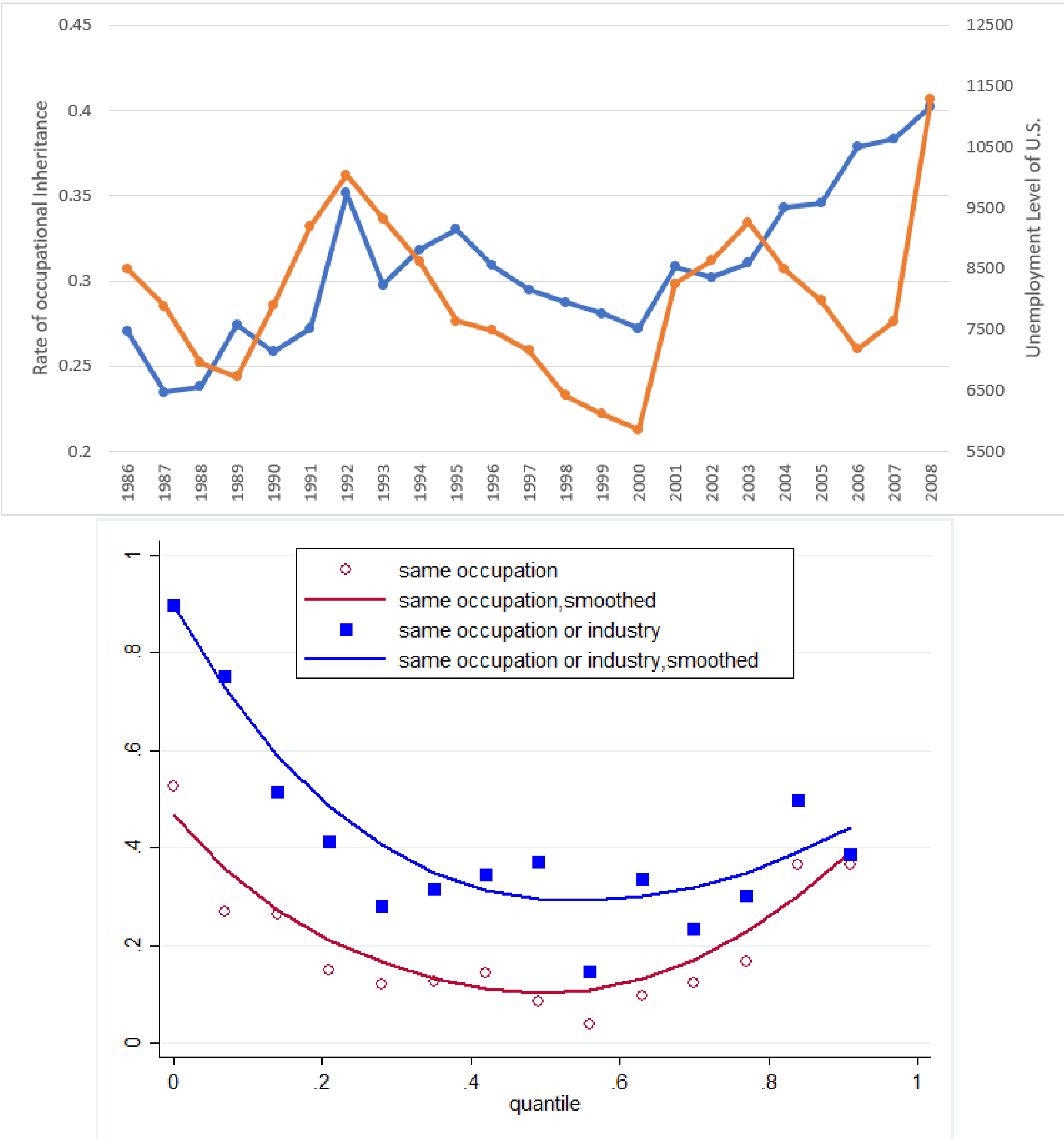
[26]As a robust check, I exclude "farmers" from the sample, and have similar results

[27]The definition is

$$IR_{random} = \pi_{edu} \oslash N_{edu}$$

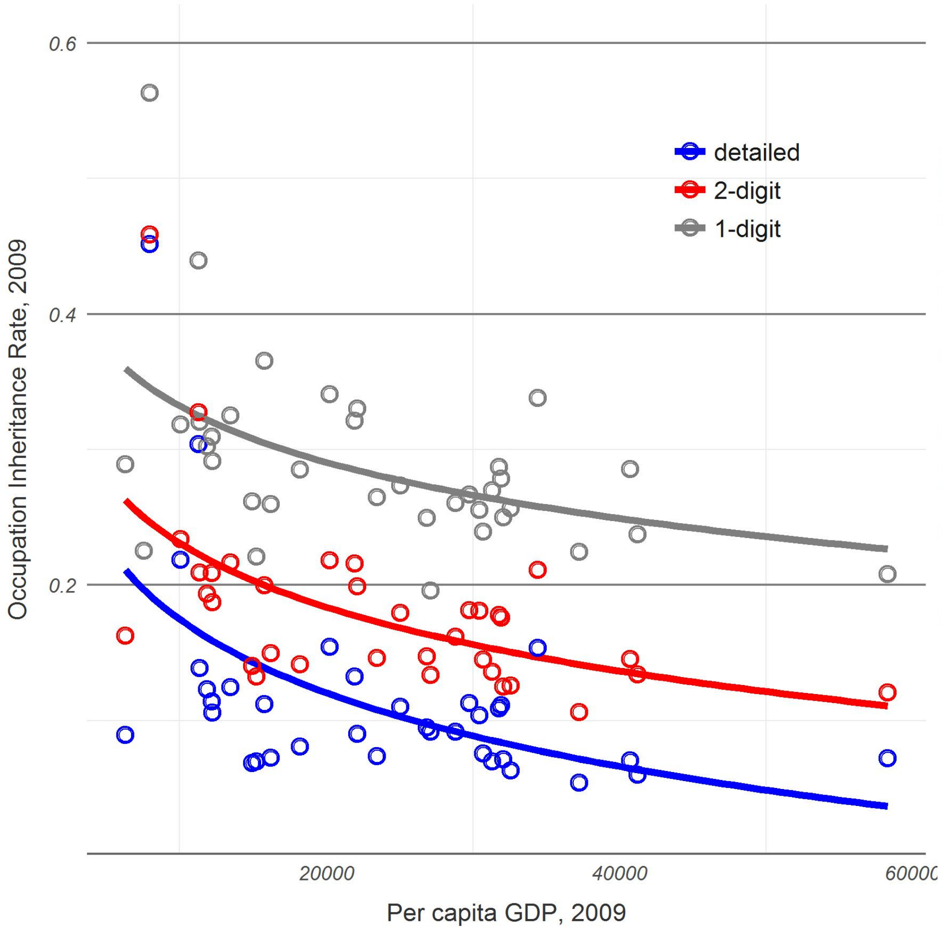
in which $\pi_{edu} = [\pi_{edu=lowest}, \ldots, \pi_{edu=highest}]$ represents the distribution of labor by education level, and $N_{edu} = [N_{edu=lowest}, \ldots, N_{edu=highest}]$ represents the number of available occupations for each education level.

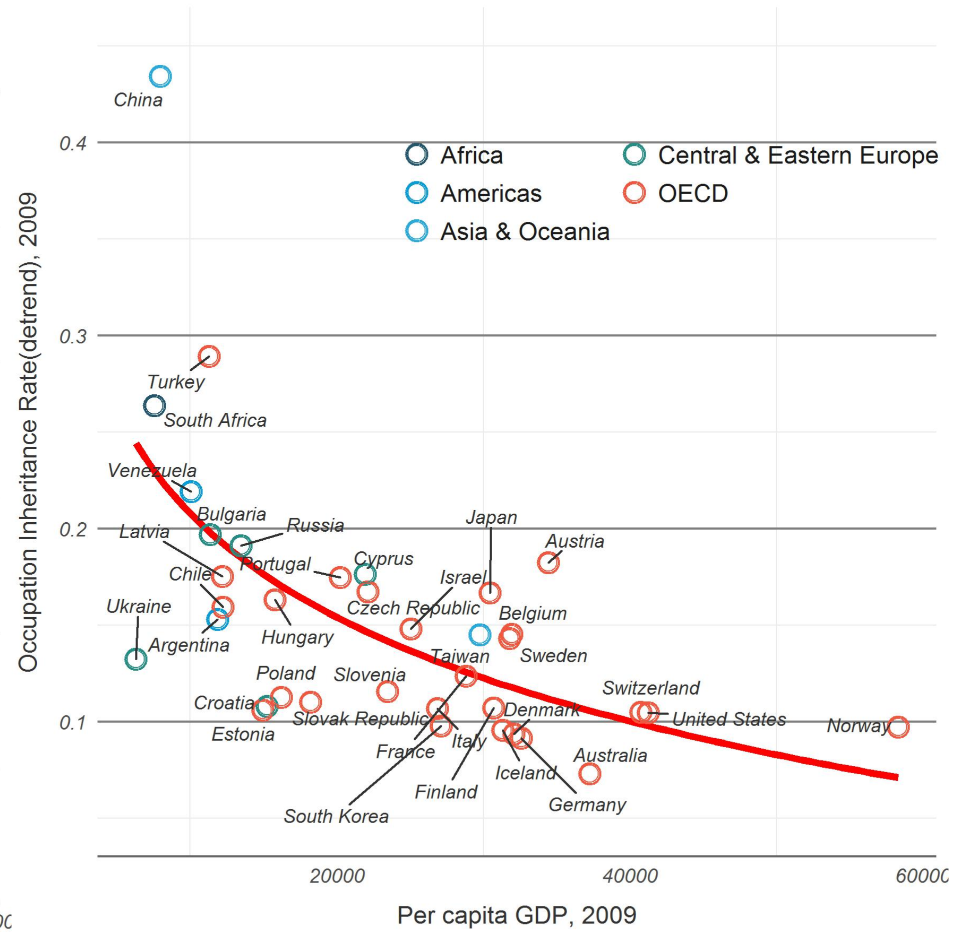
Figure 1.5: Occupational inheritance rate for first jobs (by year, by wage)



Notes: In the first figure, the blue line depicts the time series of the first job inheritance rate and the orange line depicts the time series of the unemployment level in US in the same time period. The source of the latter is Bureau of Labor Statistics Data.

Figure 1.6: Economic Development and Occupational Inheritance Rate





(a) 2-digit, normalized

Notes: For normalization π_{edu} is calibrated using ISSP and N_{edu} is calibrated with ACS 2009. This calibration is upward biased for less developed countries, because for the same occupation, the education requirement should be lower in those countries due to lower technology level.

Income premium[28]

In this part, I'll show the OLS regression results of Eq. (3) in Table 4 with the income (measured by the log and percentile rank of annual labor income) as the dependent variable, to be specific, log of annual labor income and percentile ranks of annual labor incomes are used as y_c, and X includes the demographic variables including the gender, race, region dummy, cohort dummy, year dummy, education level, marital status, children number, and this person's occupation category and the average parents' income ranking y_p is also controlled[29]. Besides, Figure 2 depicts the quantile regression results of Eq. (3) with and without all the control variables.[30]

Table 1.4: Regression results for the entry-level jobs

	log(wage), full	Rank(wage),full	Rank(wage), HS	Rank(wage), College
$\mathbf{1}\{O^C_{first} == O^P_{major}\}$	0.358***	2.538***	2.278**	2.841***
	(0.0311)	(0.520)	(0.720)	(0.756)
experience	0.0854***	0.965***	1.033***	1.080***
	(0.00217)	(0.0397)	(0.0542)	(0.0558)
$\mathbf{1} * experience$	-0.0130**	0.0128	-0.00200	0.0323
	(0.00447)	(0.0705)	(0.102)	(0.0995)
Control	Yes	Yes	Yes	Yes
$UniqueInd.$	30,381	30,381	14,564	15,168
adj. R^2	0.335	0.138	0.083	0.078

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

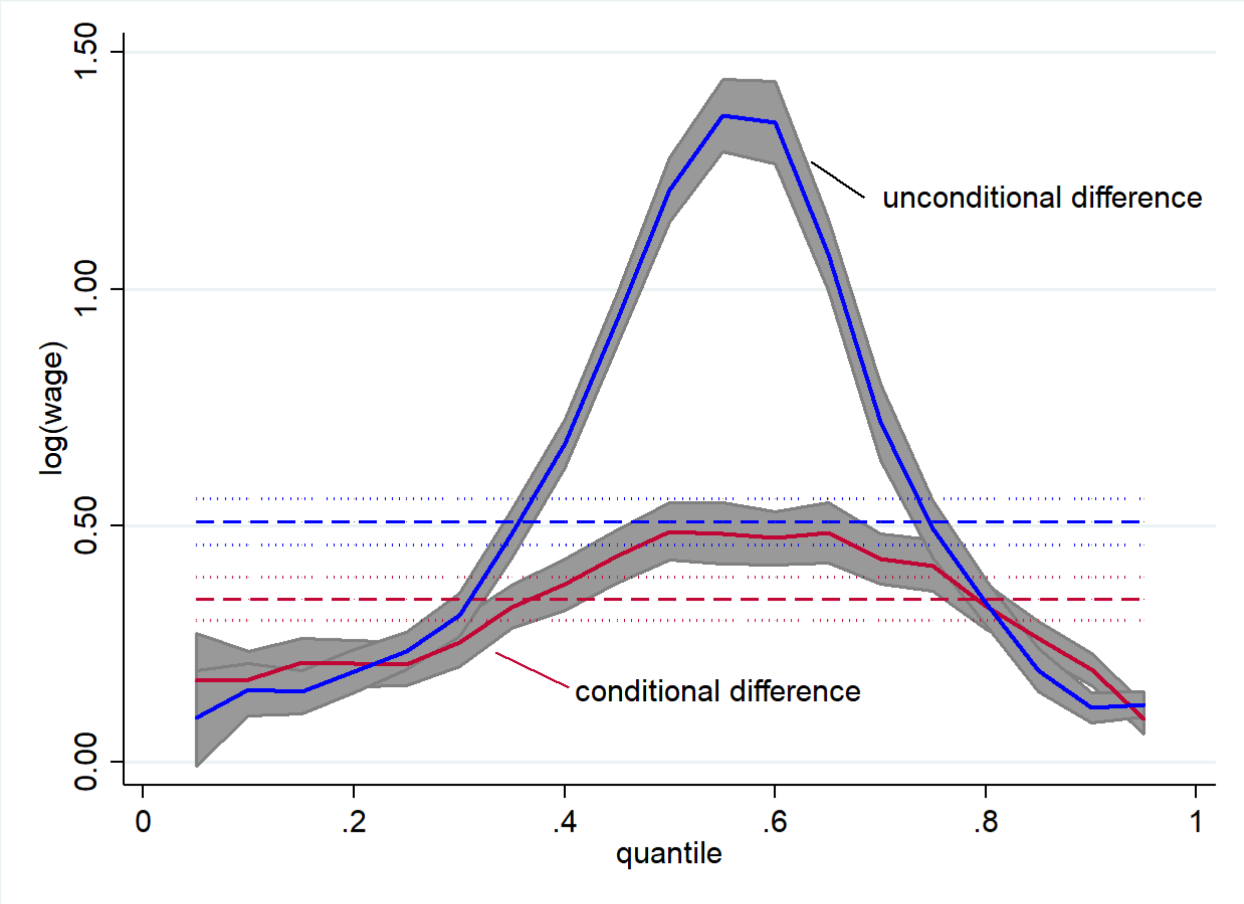
Notes: The control variables are gender, race, region dummy, cohort dummy, year dummy, education level, marital status, children number, and this person's occupation category and the parents' income ranking y_p.

[28]**In Appendix C.2, I also consider the industry choices as well.**

[29]The percentile ranking of wage is calculated within each sample PSID and SIPP and within 3-year age groups. In other words, this variable can be interpreted as the percentile ranking of annual labor income within each age group.

[30]For robustness, I also show the regression results for the detailed groups of SOC (242 groups) in Appendix Table C.6 and the quantile regression results for the percentile rank of annual labor income in Appendix Figure C.4.

Figure 1.7: Unconditional and conditional quantile regression results



We can see that the conditional difference of these two groups are significant and as high as 0.5 on log(wage) or 3.69% on rank(wage) in the median, and 0.357 on log(wage) or 2.54% on rank(wage), suggests that children who choose their parents' occupation as their first jobs have an advantage in earning[31]. And from either scenario, we can see that the difference is biggest around the median wage level. For subgroups of education level, we can see that, this positive wage premium is significant for different education level and people who have some college enjoy higher wage premium than the HS dropouts and HS graduates. The average experience with the first job for this sum-sample is 4.81 years, and we can see that on average, the premium will be persistent during this first job, decreasing slowly along the

[31]It worths mentioning that the log(wage) regression suggests that people who inherit their parents' occupation enjoy a positive wage premium as large as 30%, which is almost equivalent the 4-year college education in the literature of return of education. This results is partly induced by the different income measurements used in the two data sets. In PSID, the information is collected on annual or biannual base, and the labor income definition I used here follows the literature (e.g Blundell, Pistaferri and Preston, 2008) that combines all income from labor (salaries and wages, the labor part of business and farming, roomers and boarders and market gardening. While SIPP collects data on quarter basis, and the composition of income is less clear than PSID. But on average, the SIPP income is lower than PSID

time though.

Lower Worker's quality in terms of education level

Besides the wage level, I also compare these two groups on the education level when they take the first jobs. Since the education level are ordered categorical variables, I change the linear regression equation in Eq.(3) into Probit Model as in Eq.(4)

$$Pr(y_i > s) = F(\beta \mathbf{1}\{O^C_{first} == O^P_{major}\}_{it} + X_{it}\tau - \varepsilon_s) \ , \ s = 1, \ldots, S-1 \tag{1.4}$$

and in this regression y is the education year, whether has some years of college, whether graduated from high school, and the firm size (less than 25, 25-100, over 100).

Table 1.5: Education level, firm size and occupational inheritance

	Education	Whether college Graduate	Whether HS graduate	Firm size
$I(O_c, O_p) = 1$	-0.139***	-0.207***	-0.232***	-0.206***
	(-6.98)	(-8.71)	(-7.46)	(-7.47)
Control	Yes	Yes	Yes	Yes
N	32,746	16,210	15,777	24,385

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The regression result is reported in Table 5. We can see that people who choose the same occupation with their parents has 13.9% probability on average of having one year less of education years. And this effect is even more bigger at the cutoff of education, among all people with 12 years or more education, inheritors have 20.7% less probability to graduate from college or obtain higher education, and among those who don't go to college, inheritors are 23.2% less likely to fully finish high school education. Also, those people are 20.6% more likely to go to relatively smaller firm than those people who choose different occupations with their parents.[32] From those results, we can see that at least measured by the education

[32] For the firm size regression, because it's only available in PSID in limited number of years, so the sample size is smaller than the other regressions.

level, the people who choose their parents' occupations as their first jobs have lower quality.

1.3.3 The persistence of the entry premium

Income Premium

In this part, I expand the sample to the whole available working experience, aiming to study the lifetime effect of occupational inheritance.

Table 6 reports the regression results of the Eq.(3), also for different subsample. Because of the data problem, I run the regression with two different samples. The first sample is all people with full working experience, which is the same with the sample in the first part. This sample is the best to compare with last regression results, but due to the short time span in SIPP, this sample will be oversampled by people with less than 2 years of experience and exclude those in PSID but don't have the first job information because of late move out[33]. Hence I construct the second sample, and redefine the "first job" as the jobs before 25 year old. In this way, I would be able to include more of PSID interviewers and longer working experience. I depict the lifetime premium for high school graduates and people with some college in Figure 7. From the results, we can see that the premium we observed in last part, decreases to zero around 10 years of experience.

As we can notice that because of the smaller sample size of people with higher working experience, the confidence interval is getting bigger. The persistent and magnitude of the occupational inheritance premium is heterogeneous for different education level, which is consistent with our finding in last part. We can see that high school graduates enjoy the biggest premium, which is about 5% higher in percentile ranking within the same region and age group. Compare the two samples, we can see that the second sample gives slightly higher premium at the beginning and even higher decreasing speed because of including more people with longer working experiences.

Besides, I also find that the persistence is different for people coming from different wealthy level. In the last column of Table 4, I construct a cross variable of one's parents' average income level and the inherit dummy. By calculation, we can see that the for the most wealthy family (top 10%), the premium would disappear after over 30 years. I also run the

[33]In my new constructed data set, the number of people with one year's experience is 10,617, that of 10 years' is only 1,668, that of 20 years is 461, and that of 30 years is 62.

regression with log(wage), and the cutoff point is around 40 years. Given the fact that the longest experience in this data set is around 30 years, it's safe to say that those people would have life-long premium. And for the median family, the premium would disappear around 16 years, for the poorest family (lowest 10%), the premium would disappear after 3 years. With all these heterogeneity, there is one fact that is consistent with all the groups. That is on average, the people who started off with their parents' job in their early ages would have a premium at the beginning but this premium would decrease along with the time.

Income Risk and Stability

Besides the income level, we can also measure the income risk along with those people's life time.

There are different measurements we can use. The first one is the percentage of time that one person is unemployed, or layoff, or absent from work without pay among all time in labor force . We can see that people who choose their parents' career path have significantly 1.28% higher unemployment time every year. Also, the probability of changing jobs in any specific years is 40.8% higher for those people, and among those who change their jobs, the probability of getting fired is 79.2% higher.

Another set of measurements is the variance of income. First, I run a regression of the coefficient of variance on the inheritance indicator and other control variables as in Eq.(3). As reported in Table 7 column 2, we can see that with a median of 1.07, the people who inherit their parents' job have 19.5 percent points higher in CV. Next, I follow the literature of income risk and estimate the permanent and transitory risks of the two groups, with the detailed model set up in Appendix C.2 and the results depicted in Figure 8-10 for overall, transitory and permanent variance separately.

We can see that the overall variance is larger for people who choose their parents' occupations as their first jobs, which is mainly explained by higher transitory risk from estimation. All of these suggest that the people in this group suffers higher income risk after their first jobs in their overall life time.

In the last two columns of Table 7, I report the probability of people who choose the same occupations with their parents switch away from their original choice of occupations. We can see that they have much higher chances of switching away then others who don't inherit their parents' job, showing lower level of stability.

1.3.4 Summary

Now we can summarize the stylized facts we get from this section:

- There is a significant premium for people who choose the same job as their parents, about 0.3 on average for log(wage) and 2.54% on percentile ranking of wage.
- This premium diminishes along with the lifetime, and the difference of the premium at the entry-level and 20 years after working is about 6-8%.
- People who choose to stay in the same job have higher income risks, risks of getting fired and lower level of stability in sense of employers, occupations and industries.

Figure 1.8: The age profile of the premium

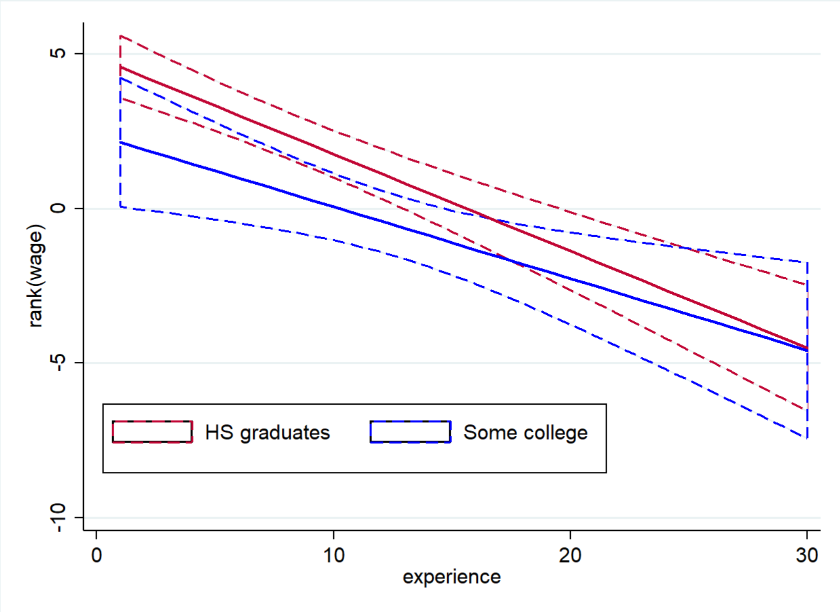


Figure 1.9: Overall Pattern of Total Income Risks

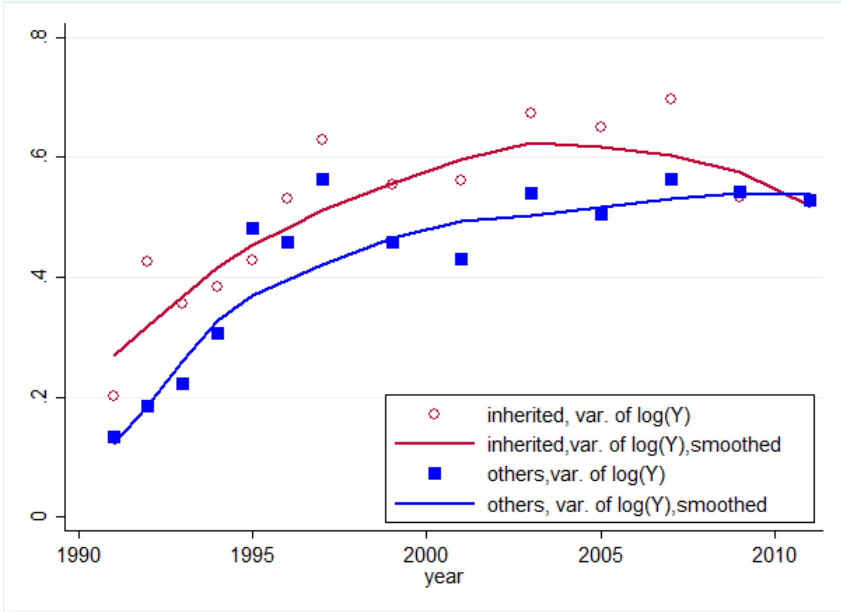


Figure 1.10: variance of transitory risk

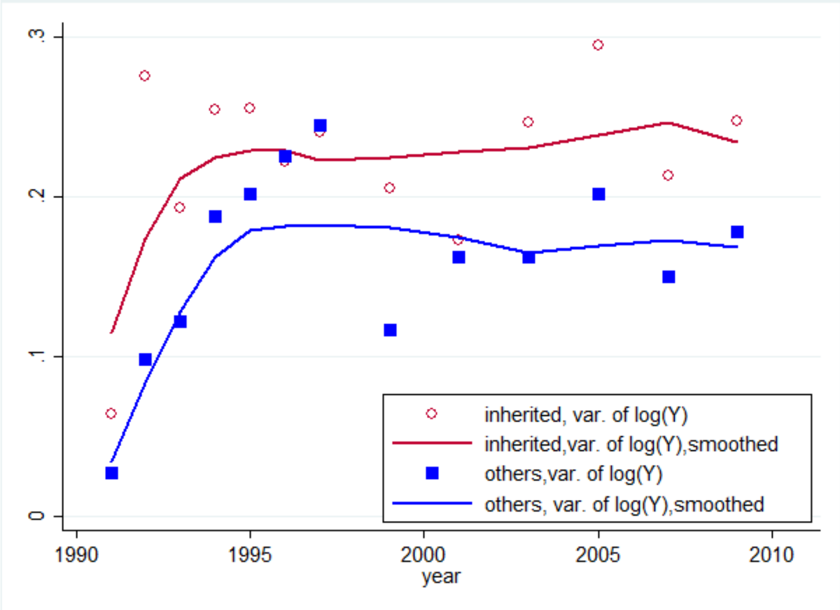


Figure 1.11: variance of permanent risk

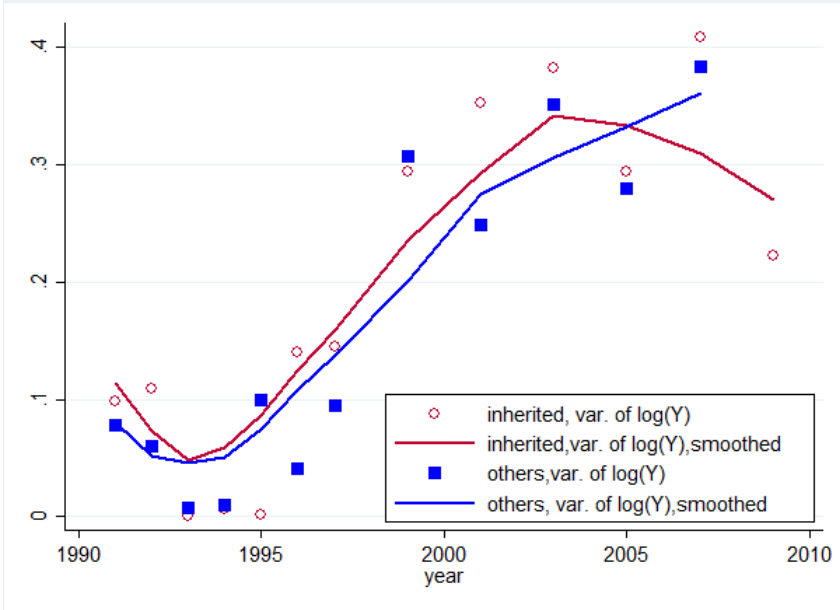


Table 1.6: Persistent effect

	Full Sample	No HS	HS graduate	Some college	Sample 2	log(Wage)	Continuous
$\mathbf{1}\{O^C_{first} = O^P_{major}\} = 1$	1.088**	-0.895	4.892***	2.375*			
	(2.75)	(-0.74)	(8.98)	(2.09)			
I*experience	-0.116**	0.135	-0.314***	-0.232**			-0.176***
	(-3.21)	(1.21)	(-6.93)	(-2.92)			(-6.27)
$\mathbf{1}\{O^C_{before25} = O^P_{major}\} = 1$					1.254**	0.0652*	
					(2.59)	(2.41)	
I*experience					-0.179***	-0.00302	
					(-5.49)	(-1.34)	
I* Wealthy							5.746***
							(7.37)
control	X	X	X	X	X	X	
$Unique Ind.$	20,072	4,394	6,639	9,930	31,572	31,572	31,572
adj. R^2	0.167	0.100	0.071	0.193	0.224	0.435	0.168

t statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.7: Income Risk

	Unemploy Rate	CV of Income	Pr(Fired|job change)	Pr(change jobs)	Pr(change Occ.)	Pr(change Ind.)
$\mathbf{1}\{O^C_{before25} == O^P_{major}\} = 1$	1.277**	0.195**	0.792***	0.408***	2.338***	1.901***
	(3.02)	(2.72)	(5.91)	(6.81)	(36.06)	(29.65)
I*Experience	-0.0257	-0.0157***	-0.0353**	-0.0274***	-0.154***	-0.121***
	(-0.90)	(-3.58)	(-3.27)	(-4.89)	(-24.04)	(-19.69)
Control	Yes	Yes	Yes	Yes	Yes	Yes
UniqueInd.	31,572	31,572	22,228	31,572	31,572	31,572
adj. R^2	0.050	0.011				

t statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.4 Mechanisms of wage premium and social Network

This section begins by documenting the difference of job searching behaviors between people who inherit their parents' job and who don't. The theory developed in the next section is based on the premise that choosing a different career path compared with their parents is accompanied by a loss in these network services, connecting parents' social networks to high occupational inheritance rates, and accompanying labor misallocation, we have documented. This connection will be subjected to greater scrutiny in the empirical analysis that completes the paper.

1.4.1 Job hunting behavior

In the existing network theories, there are two potential channels for social network to influence the labor market opportunities and outcomes. One is the weak tie or network-transmitted information. Besides, favoritism is proved to be fairly common among relatives and close friends who expect preferential treatment in all human societies (Rees, 1966). In essence strong ties are favor-exchange ties (Bian, 2002) or instrumental particular ties (Walder, 1986). These kinds of ties were mobilized to get a favorable job assignment from state authorities before reforms (Bian, 1997), to help people move between employers for better pay or career advancement in China's emerging labor markets (Bian, 2002, 2008; Bian & Huang, 2009), to secure reemployment opportunities after layoffs (Zhao, 2003), and to obtain "soft-skill" jobs whose performance is hard to measure or monitor (Huang, 2008). In PSID 1978 wave, specific questions about interviewers' job hunting histories were added to the survey, including "Do you think there was anyone who may have helped you get the job?" "If yes, what's your relationship with this person?" "What kind of help you got from this person?" etc.

Based on those questions and the existing network researches, I construct the strong-tie and weak-tie variables as below:

$$I_{Strong-tie} = \begin{cases} 1 & \text{if Type = "Direct influence stated"} \\ & \text{or Referee = "Very strong"} \\ 0 & \text{other situation} \end{cases} \tag{1.5}$$

$$I^{Referer}_{Weak-tie} = \begin{cases} 1 & \text{if Type = " Direct inferred, recommendation or mention"} \\ & \text{or Referee = "Modest"} \\ 0 & \text{other situation} \end{cases} \tag{1.6}$$

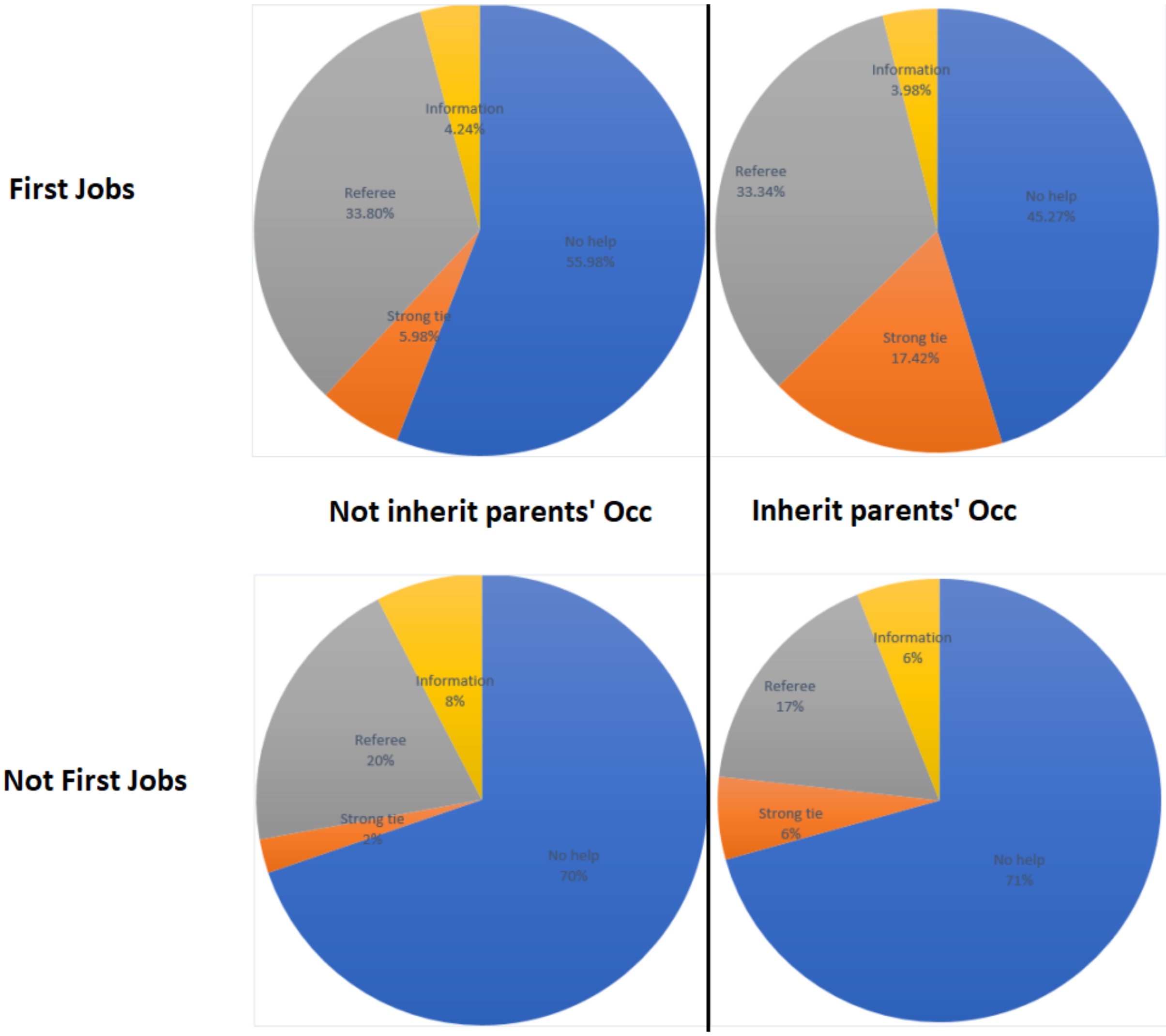
$$I^{Info}_{Weak-tie} = \begin{cases} 1 & \text{if Type = " "Told me to try for/about the job"} \\ & \text{or Referee = "No"} \\ 0 & \text{other situation} \end{cases} \tag{1.7}$$

The variable $I_{Strong-tie}$ is constructed based on the questions "How did they help?", and "Could they have had some say in your getting the job? and How much". If the answer is "Direct influence stated; gave me the job; got me the job" for the first one and "Very much; a lot; gave me the job" for the second one, I define as get the job via strong ties. For the weak ties, I construct two variables based on the mechanism. For the network-transmitted information to the employer, I define as the $I^{Referer}_{Weak-tie}$. It is constructed also based on the above two questions, if the answer is "Direct influence inferred; friend of the foreman" or "Recommended me to employer" or "Told employer about me" for the first one and "Yes and Moderate amount; some" for the second one, I consider the type information transmitted towards the employer. If the answers are "Told me to try for job" or "Told me about job" for the first question and "No" for the second one, I consider it as merely information towards the potential employee.

To prove the relationship of career inheritance and the type of job hunting network usage, I show the distribution of these three channels by the relationship of helpers and whether or not the interviewer inherit his/her parents' career path.

We can see from Figure 11 that for those who inherit their parents' career path, about half

Figure 1.12: Distribution of ties for job hunting



of them got the job through direct influence (strong ties) or direct reference (referrer), and among those people, about 30% of them are through strong ties. Comparatively, strong ties usage for those who chose a different career path with their parents is very limited, only 5%.[34] While for other jobs except the first jobs, the dependence on non labor market channels is decreasing. Though though the difference between the people who inherit their parents' job as their first jobs and who don't is persisting, in the sense that they triple the usage rate of the strong ties even after their first jobs, the magnitude is much less. This is consistent with our finding, that the positive wage premium exists in the entry-level jobs but decreasing along with time.

[34]The distribution of sources of helps are in Appendix Table D.9

1.4.2 The starting time of first jobs

Also, we can see that for people who choose their parents' occupations as their first jobs, the searching time is less compared with the counterparts. For new graduates, June is the most frequent month for people to start their jobs (See details in Appendix Table D.10 the distribution of starting month for the first jobs) and we find that staying in the same occupations give people 7% higher probability of starting their first jobs before or in June compared to the counterparts, which suggests that those people have advantages in the labor market which can make it faster to find a job, or in other words, shorten the waiting time before a suitable job arrives (See the Probit Regression results in Appendix Table D.11).

1.4.3 Firm level evidence

The results we have in this paper also in line with the firm-level intergenerational transition study, which also suggest the usage of parents' social network in the job hunting process, especially for entry-level jobs. For example Corak and Piraino (2011) using firm level data from Canada and Denmark, proving that even in firm level, there is significant level of intergenerational transmission, with 30 to 40% of young adults having at some point been employed with a firm that also employed their fathers (Appendix Figure C.5) depicts the rate of employed by the same employer of fathers from Corak and Piraino, 2011).

1.5 Conceptional Framework

I introduce a model of children's occupational choices for the purpose to guide the interpretation of the empirical analysis. I derive four predictions that guide the empirical analysis in Sections 2 and 3.

This model is based on the occupation choice framework created by Hsieh, Hurst, Jones, and Klenow (Hsieh et al. [2013], HHJK hereafter), which is basically a generalized version of Roy model, initially used in the international trade in Eaton and Kortum(2002).

1.5.1 Preferences and Incentives

There is a continuum of individuals in each generation, and each of them would have one off-spring. There are M occupations in this economy and each individual chooses his occupation $i \in \{1, 2, \ldots, M\}$, with wage w_i as the per unit of efficiency labor. Assume that each occupation requires S_i level of education, without loss of generality, the above occupations are ordered by the education level. Each individual is born with innate talent $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_M)$[35] and the realization of ϵ follows a Fréchet distribution, which is determined by his father's occupation:

$$F_j(\epsilon) = \exp\Big(- \big(\sum_{i=1}^{N} T_{ij} \epsilon_i^{-\frac{\tilde{\theta}}{(1-\rho)}}\big)^{1-\rho}\Big) \tag{1.8}$$

The parameter T_{ij} governs the location of the distribution; a larger T_{ij} implies a larger possibility of high talent in occupation i. The parameter ρ governs the correlation between talents across different occupations for an individual. For the following analysis, I use $\theta = \frac{\tilde{\theta}}{1-\rho}$. If $\rho \to 0$, productivity levels of offspring are uncorrelated across occupations, while in the limit as $\rho \to 1$ they are perfectly correlated, so that productivity is independent of the occupation.

Assume each individual lives 2 periods, human capital can only in accumulated in period 1 (childhood). In this model, innate talent is private information when a person started to work reveal by certain probability in the second period. [36]

Individual's maximization problem is to choose an occupation in order to maximize the expected income net the cost of education[37]:

$$\max_{i \in I, E_i} log(U) = \Big[\beta \log(\sum_{t=1}^{t+2} C_i(t))\Big] + log(1 - S_i) \tag{1.9}$$

in which E_i is the cost of education if occupation i is chosen

Here $C(t)$ is consumption in year t, $1 - s$ is leisure time during the pre-period when human capital investments are made. Note that we assume no discounting of consumption

[35] The reason I choose M occupational specific ability instead of just K abilities as in real life is because with less types of abilities, we cannot get a unique solution for the choice of occupation mathematically. In other words, the M different kinds of abilities can be seen as M different kinds of combination of abilities

[36] We can understand as if there are two sub-periods in the "Young" period, in which people first accumulate human capital and then start to work.

[37] Why don't model the whole life time: because ex-ante, people don't know the whole path of the dynamics of income, the probability of revealing in each year. So basically what they know is the expected advantage, or ex-ante belief of the advantage they are going to get

for simplicity. β parameterizes the tradeoff between lifetime consumption and time spent accumulating human capital. Individuals borrow e in the first period to purchase e units of human capital, a loan they repay over their lifetime, but I assume that there is no financial market, and the amount they can borrow is subject to father's occupation Y_j. I assume the credit constraint that occupation $i <= M_{L1}$ cannot afford the education S_i for $i >= M_{H1}$, and $i <= M_{L2}$ cannot afford the education S_i for $i >= M_{H2}$ with $L_1 < L_2$ and $H_1 < H_2$. This assumption can be understood as the high school education and college education. For people in the lowest income occupation, their parents' don't have enough money to support the children to finish high school while less lower income occupation family can support the high school but not the college. This is the second difference of this model.

The effective labor unit is decided by both the education level and education expenditure through the human capital production function:

$$h_i = S_i^{\phi_i} E_i^{\eta}$$

Hence the income of choosing occupation i is

$$Y_i = w_i \epsilon_i T h_i$$

In the first period, assume that the innate ability is private information, for people who stay in the related occupation with their parents, it will have an inflater $\tau_{ij} > 0$, and for others there will not be any inflater. In the second period, for people who were masked with an inflater in the first period will have p probability of hiding his true productivity, and $(1-p)$ reveals the true productivity.

The individual maximization problem implicitly defines an intergenerational occupation transition matrix , where p_{ij} is the probability of a son choosing occupation i conditional on the fact that his father works in occupation j.

The total amount of efficient labor supply in occupation i thus can be written as:

$$H_i = \sum_{i=1}^{M} \pi_j^L p_{ij} E\left[\epsilon_{ij} | i\right] \tag{1.10}$$

Assume there is a representative firm that hires all occupations of workers and produces final goods according to the CES production function in each period:

$$\max Y - \sum_{i=1}^{M} \exp(1) w_i H_i \tag{1.11}$$

$$Y = \Big(\sum_{i=1}^{M}(A_i H_i)^{\frac{\sigma-1}{\sigma}}\Big)^{\frac{\sigma}{\sigma-1}} \tag{1.12}$$

where A_i is the occupational productivity.

1.5.2 Propositions

From the conceptional framework, we can have four propositions about the inheritance rate, worker's quality, income risk and the total labor productivity.

Proposition 1.5.1 *(Occupational Inheritance Choice): The inheritance rate for the high income and low income occupation is higher than the medium level income occupation. i.e. p_{jj} is decreasing along with w_i and then after certain point, increasing with w_i.*

For the occupations without credit constraint, the higher the return to skill w_i relative to other occupations, the higher the inheritance rate. At the same time, for people from family with credit constraint, the overall choices are limited, ending up with higher possibility of staying at the same jobs. This proposition which predicts the "U" shape relationship of occupational inheritance rate with respect to the occupational income standings fits our empirical results from Section 2.2.

Proposition 1.5.2 *(Average Quality of Workers): The average quality of workers in each occupation, including both human capital and talent, is*

$$E\big[h_{ij}\epsilon_i|Father = j\big] \equiv E_j\big[h_{ij}\epsilon_i\big] = \frac{T_{ij}^{\frac{1}{\theta(1-\eta)}}}{\kappa_{ij}w_i\big[(1-s_i)\big]^{\frac{1}{2\beta}}}\Big(\eta\Phi_j^{\frac{1}{\theta}}\Big)^{\frac{1}{1-\eta}}\Gamma(1-\frac{1}{\theta(1-\eta)}) \tag{1.13}$$

or

$$\Big[s_i^{\phi_i}\Big(\eta\kappa_{ij}w_i\Big)^{\eta}(\frac{1}{p_{ij}})^{\frac{1}{\theta}}\Big]^{\frac{1}{1-\eta}}\Gamma(1-\frac{1}{\theta(1-\eta)})$$

This proposition predicts that the average quality is inversely related to the share of the group working in the occupation p_{ij}, which fits our empirical results in Section 2.2 about worker's quality.

Proposition 1.5.3 *(Occupational Wages): Let $wage_{ij}(t)$ denote the average earnings in* *occupation i at age a of group j. Its value satisfies*

$$\overline{Wage}_{ij}(t) = \begin{cases} \exp(1+\tau_{ij}) w_i T(1) E\big[h_{ij}(t)\epsilon_i\big] = \frac{\exp(1+\tau_{ij})T(1)}{\kappa_{ij}} \frac{\gamma T_{ij}^{\frac{1}{\theta(1-\eta)}}}{\big[(1-s_i)\big]^{\frac{1}{2\beta}}} \Big(\eta \Phi_j^{\frac{1}{\theta}}\Big)^{\frac{1}{1-\eta}} & \textit{if } t = 1 \\ \exp(1+p\tau_{ij}) w_i T(2) E\big[h_{ij}(t)\epsilon_i\big] = \frac{\exp(1+p\tau_{ij})T(2)}{\kappa_{ij}} \frac{\gamma T_{ij}^{\frac{1}{\theta(1-\eta)}}}{\big[(1-s_i)\big]^{\frac{1}{2\beta}}} \Big(\eta \Phi_j^{\frac{1}{\theta}}\Big)^{\frac{1}{1-\eta}} & \textit{if } t = 2 \end{cases} \tag{1.14}$$

in which $\gamma = \Gamma(1 - \frac{1}{\theta(1-\eta)})$

From this proposition, we can see that since

$$\frac{\exp(1+p\tau_{jj})T(2)}{\kappa_{jj}} < 1$$

and

$$\frac{\exp(1+\tau_{jj})T(1)}{\kappa_{jj}} > 1$$

the people who inherit their parents' job would have a wage premium when they are "young", but negative premium when their true ability is revealed, which fits our stylized facts in Section 2.

Based on the above proposition, we can get the following relationship:

$$\exp(\tau_{jj}) = (\frac{T_{jj}}{T_{kj}})^{-\frac{1}{\theta}} (\frac{\overline{Wage}_{ij}}{\overline{Wage}_{kj}})^{1-\eta} (\frac{p_{jj}}{p_{kg}})^{\frac{1}{\theta}} \tag{1.15}$$

or

$$ln(\overline{Wage}_{jj}) - ln(\overline{Wage}_{j,\neq j}) = \frac{1}{1-\eta}\Big[\tau_{jj} + \frac{1}{\theta}(T_{jj} - \bar{T}_{j,\neq j}) - \frac{1}{\theta}(p_{jj} - \bar{p}_{j,\neq j})\Big]$$

This equation reveals the problem of OLS regression, in the sense that, it cannot give a consistent estimation, because of the potential unbalanced distribution of innate ability T, and the consistent estimation of the wage difference in the equilibrium is $\frac{\tau}{1-\eta}$, i.e. the true wage premium adjusted by the return to investment in human capital η. If we average over all the occupation, we can see that

$$ln(\overline{Wage}_{inherit}) - ln(\overline{Wage}_{not}) = \overline{\tau} + \frac{\eta}{1-\eta}\overline{\tau} + \frac{1}{1-\eta}\Big[\frac{1}{\theta}(\overline{T}_{inherit} - \overline{T}_{not}) - \frac{1}{\theta}(\overline{p}_{inherit} - \bar{p}_{not})\Big]$$

The mean has two effects on the average wage. First, $\overline{\tau}$ has a direct effect on the average wage, where the elasticity of the average wage to $\overline{\tau}$ equals one. This effect is given by the first term in the equation for the children who inherit their parents' job. The mean of $\overline{\tau}$ also has an indirect effect on the average wage by changing the return to investment in human capital. The magnitude of this effect depends on η, and is captured by the third term in the equation. The mechanism is that people stay in the same occupation with their parents, which at the same time means that they choose a similar education level with their parents, which may not be the optimal without this labor market friction. In other words, the intergenerational educational correlation that we observe in the data is partially the result of the credit constraint and partially the result of the occupational choices.

Proposition 1.5.4 *(Inequalities): In the "Young" period, the variance of wage for any occupation i from occupation j satisfies*

$$var(wage)_{ij} \propto \Big(\frac{\exp(1+\tau_{jj})T(1)}{\kappa_{jj}}\Big)^2 > 1$$

This proposition suggests that group of those who stay in the same occupations have higher income variance or in other words income risk compared to the counterparts, which fits our prediction in Section 2.3.

Proposition 1.5.5 *(Total Labor Productivity): In general equilibrium, the total labor productivity is negative associated with the social network benefit τ, since the larger the benefit, the larger the misallocation of talent in the economy.*

Our cross-nation empirical result in Section 2.1 suggests that the larger the intergenerational occupational inheritance rate, the lower the per capita GDP level. Though those countries are not exactly the same in technology level, but it still sheds some lights in the relationship of total labor productivity and intergenerational occupational correlation.

1.6 Identification

1.6.1 Intuition of Identification Strategy

This identification strategy aims at disentangling the relationship between the intergenerational correlation of innate abilities and that of occupational choices. In this paper, the main explanatory variable is the dummy variable $\mathbf{1}(O_C = O_P)$, which measures how the children's occupational choice correlated with their parents', hence the major concern is that this variable is decided by the unobserved innate ability of children which is also correlated with their parents, in which case, the parameter we estimates purely comes from the innate ability advantage of those people.

My identification comes from the recent Technology Revolution beginning from the late 1980s to the beginning of 20th century. If we assume that individuals' innate abilities can be described as a multidimensional continuously distributed variable, as depicted in Figure 13. And the occupations follow a categorical variable which can be seen as a partition in the continuously distributed innate ability space, or in other words, individuals' occupational choices follow an assignment rules from the innate ability space to the discrete occupations. The cutoffs of this assignment rules, as illustrated in Figure 13, are determined by market factors like wage schedule and technology. If there is a significant change in market factors leading to a significant change in the assignment rules, for example the Technological Revolution, we will have extra variation to disentangle the co-movement of innate abilities and occupational choices between the two generations.

1.6.2 Regression Discontinuity Design

This section presents the regression discontinuity design following the discussion in Hahn, Todd and van der Klaauw (2001). In this paper, with parents' in the same occupations are the "treatment" and those children with $\mathbf{1}(O_C = O_P) = 1$ are the "treated".

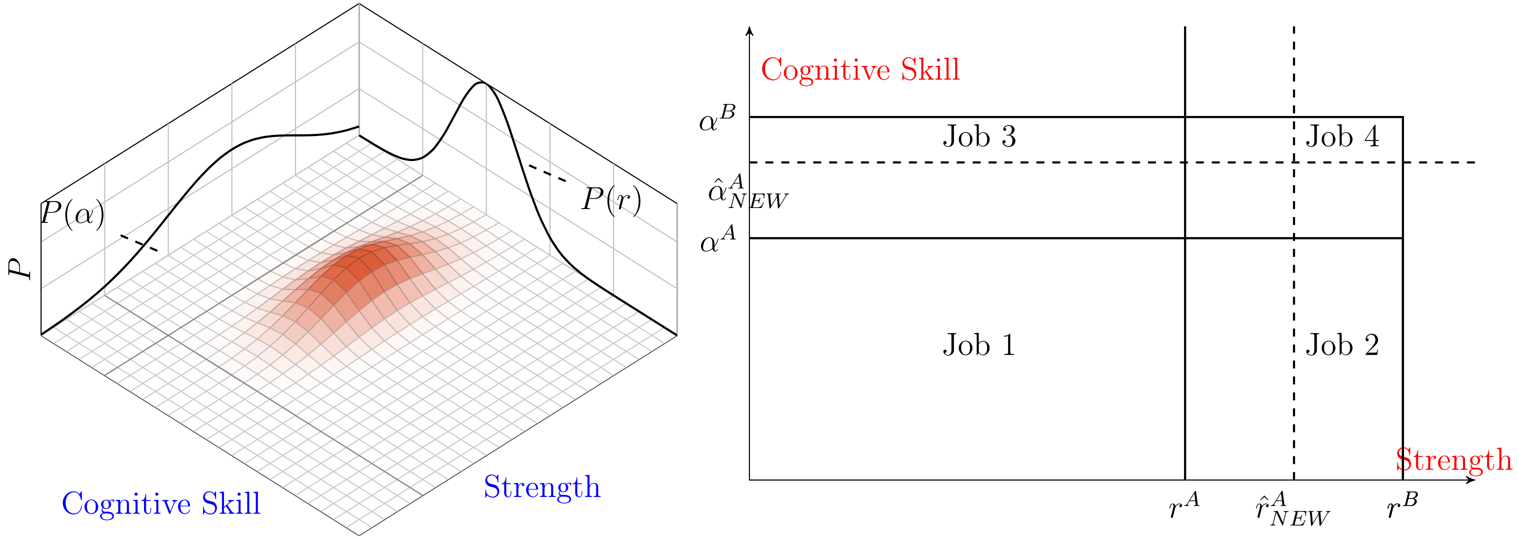
Following the notation of the potential outcome approach to causal inference, let $(Y1, Y0)$ be the two potential outcomes one would experience by choosing the same occupation and not, respectively. In the context of this paper, Y_1 and Y_0 represent the labor market outcome corresponding to the children staying in the same occupation and not, respectively. The causal effect of intergenerational occupational inheritance on labor outcome is then defined as the defference between these outcomes, $\tau = Y_1 - Y_0$, which is not observable. Accordingly, though not observable, τ represents the change in labor outcome corresponding to a difference in the occupational choice, which is our quantity of interest.

Let $\mathbf{1}(O_C = O_P)$ be the binary variable denoting the intergenerational occupational inheritance status, with $\mathbf{1}(O_C = O_P) = 1$ for the treated and $\mathbf{1}(O_C = O_P) = 0$ otherwise. A discontinuity design arises when $\mathbf{1}(O_C = O_P)$ depends on an observable variable S and there exists a known point in the support of S where the probability of being treated changes discontinuously. Formally, if $\bar{s}$ is the discontinuity point, then a regression discontinuity is defined if

$$Pr\{\mathbf{1}(O_C = O_P) = 1|\bar{s}^+\} \neq Pr\{\mathbf{1}(O_C = O_P) = 1|\bar{s}^-\} \tag{1.16}$$

Here and in the following $\bar{s}^+$ and $\bar{s}^-$ refer to those individuals marginally above and below $\bar{s}$, respectively.

Figure 1.13: Illustration of innate ability distribution and occupation partition



Notes: In this illustrative example, we assume there are two abilities, cognitive skills (α) and strength (r). The figure on the left hand depicts the continuous distribution of the two abilities, with the joint p.d.f. on the z-axis. the color shows the density of combinations of those two abilities. The figure on the right hand depicts the partition on the α - r surface, and α^A and r^A represent the cutoffs among the four different occupations. In the extreme case that children's abilities and their parents' abilities are the same, then $\mathbf{1}(O_C = O_P) = 1$ for the whole population. Assume that because of the technology changes, the new cutoffs move to $\hat{\alpha}^A_{NEW}$ and $\hat{r}^A_{NEW}$, then if for the extreme case, children can still end up in the same job even their parents' jobs are different.

In the context of the paper, the expression in Eq.(16) implies that the probability of staying in the same occupation varies discontinuously with an observable variable S. To fix ideas, the assignment variable S is the innate ability. First let's consider the extreme case in which the innate ability of parents and children are exactly the same, then the old cutoffs (α^A and r^A in Figure 13) are the $\bar{s}$, and the children around these cutoffs would choose Job 1, i.e., the dummy variable $\mathbf{1}(O_C = O_P) = 1$ for $\bar{s}^-$ and $\mathbf{1}(O_C = O_P) = 0$ for $\bar{s}^+$, constructing a sharp RD design.

For the real case, if children's innate ability is positive correlated with parents, as assumed in the literature, children's occupational choices neatly fits a fuzzy design conditional on S_P. To be specific, the assignment rule is no longer a deterministic one as in sharp RD design. As a result of the eligibility rule and of self-selection, the probability of being treated for those scoring a value of S above the threshold $\bar{s}$ is zero by definition. The probability of staying in the same occupations for those scoring below $\bar{s}$ is smaller than one because children's innate ability is not the exact as their parents. This implies that the probability of $\mathbf{1}(O_C = O_P)$ is discontinuous at the threshold for eligibility and the size of the discontinuity is less than one[38].

In the rest of this section, I will first show in Section 5.3 that the recent Technological Revolution indeed significantly change the skill composition of occupations in large scale. Then in Section 5.4 I describe how I obtain the measurement of parents' occupational specific skills and cutoffs. In Section 5.5, I consider the measurement error and gives the assumption needed for identification.

1.6.3 The technological Revolution

The major instrument variables in this paper build up on the idea that the rapid technology changes starting in the 1980s, including but not limited to the personal computer (PC), related technologies, and biology etc., reshape the ability requirements in occupations and redefine the industrial structures in U.S.

Following Card and DiNardo's seminal work in 2001 about skill-biased technical change, I

[38] In the whole analysis, I use "below" and "above" with the assumption that the new cutoff point is above the old ones as illustrated in Figure 13. The opposite should apply if the cutoff decreases.

depict the time series of computer usage at work and number of Internet Host in Appendix E.6 as a measure which is not meant to accurately measure the technology change, but only as an index of the major changes, which reshapes the labor market.

As stated in Section 1, I use the DOT and O*NET to measure the multidimensional skills from 1970 to 2010. [39] Figure 14 depicts smoothed changes in the percentile ranking of the detailed skill measures for all occupations between 1980 and 2000, and 1990 to 2010 separately.[40] The 1980s and 1990s witness the spread of Internet, PCs and 2000s more on automation. And in both comparison, we see that the physical demands of some used-to-be high occupations, like visual, auditory, strength are significantly decreased due to the substitution effects of electric devices, and the ability and skills required for routine cognitive tasks also see very big change in distributions.

Among the 321 occupations that are available through the whole time span from 1980 to 2010, only 13 of them only have less 10% changes in all dimension. [41].

To sum up, what we observe in the data is that there indeed exists significant changes in occupational specific skills and the direction of the changes is that the physical, manual and routine skills are largely substituted by machines/technology while relatively little changes happen in the non-routine cognitive skills. Hence we can describe the skill changes in each specific occupation by using the relative changes of physical, manual and routine skills in terms of non-routine cognitive skills. To be specific, I construct the relative changes of non-routine cognitive skills versus other skills for each occupation: $\Delta Cog\_Physical$, $\Delta Cog\_routinecog$, $\Delta Cog\_routineman$ and ΔCog_manual as:

$$\Delta Cog\_Physical = \text{Non-routine Cognitive Skill - Physical Skill} \tag{1.17}$$

$$\Delta Cog\_routinecog = \text{Non-routine Cognitive Skill - Routine cognitive Skill} \tag{1.18}$$

$$\Delta Cog\_routineman = \text{Non-routine Cognitive Skill - Routine manual Skill} \tag{1.19}$$

$$\Delta Cog\_manual = \text{Non-routine Cognitive Skill - Manual Skill} \tag{1.20}$$

[39]To be specific, I use DOT 1977 for 1970s - 1980s, and DOT 1991 for 1990s. I use O*NET version 3.1 for 2000s and 15.1 for 2010s, which are the most up-to-date version of O*NET for the corresponding census year.

[40]Appendix Figure E.7, I also show the real changes vs. the smoothed changes for the four major skill groups to show that the smoothed changes accurately pick up the trend instead of showing some extreme values.

[41]We can also aggregate those multidimensional changes into one dimension, and I show the density of the aggregate changes from 1980 to 2000 and 1990 to 2010 in Appendix Figure E.8.

And the measurement of each skill can be the absolute measure (from scale 1-7) or the percentile ranking of those skills (from scale 0-1). The distribution of these relative changes in absolute values are depicted in Figure 16. [42]. We can see this measurement conveys the same message that the other skills change significant relative to the non-routine cognitive skills. And among these skills, the relative changes of non-routine manual skills are the smallest, the maximum of which is slightly larger than 1 (in scale of 7), while others are around 3 (in scale of 7).

1.6.4 Constructing Individual Level Skill Measurements

The skill measurements we have now are for occupation level (means, std.dev, min, max available), but what we need in the regression discontinuity design is a continuous variable in individual level.

The Committee on Occupational Classification and Analysis of the National Academy of Sciences conducted the surveys of DOT. The Committee also acquired a selection of variables from the April 1971 *Current Population Survey (CPS)* that were gathered from a sample of households which yielded 60,441 workers in the experienced civilian labor force. The CPS survey provided detailed information about the workers and their family backgrounds, education, and employment, with Dictionary of Occupational Titles (DOT) characteristics, e.g., job classification and description, for each worker in the survey. In other words, we get a data sets with demographic information, working information and job descriptions (CPS) and our constructed new data extract with only the the demographic information and the working information.

[42] The distribution of these relative changes in percentile rankings are depicted in Appendix Figure E.10

Figure 1.14: Smoothed changes of percentile ranks for all detailed skill groups between 1980 and 2000

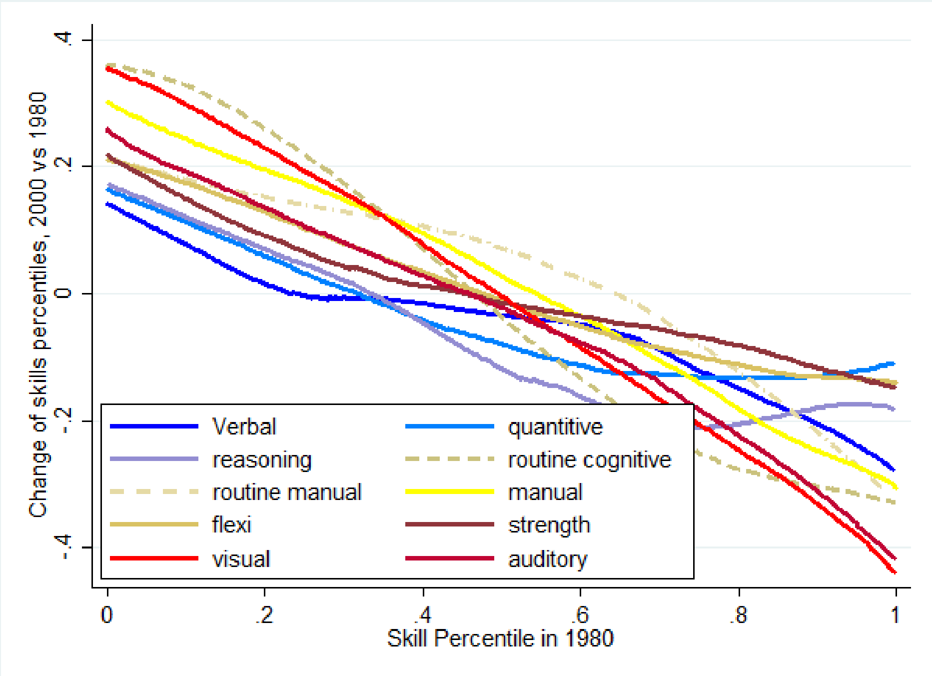
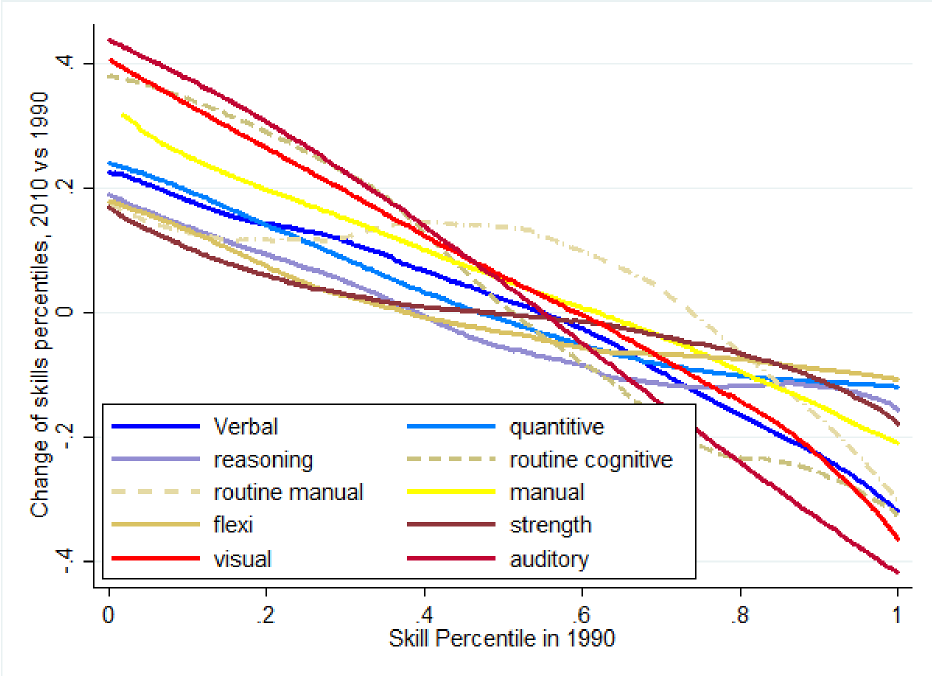
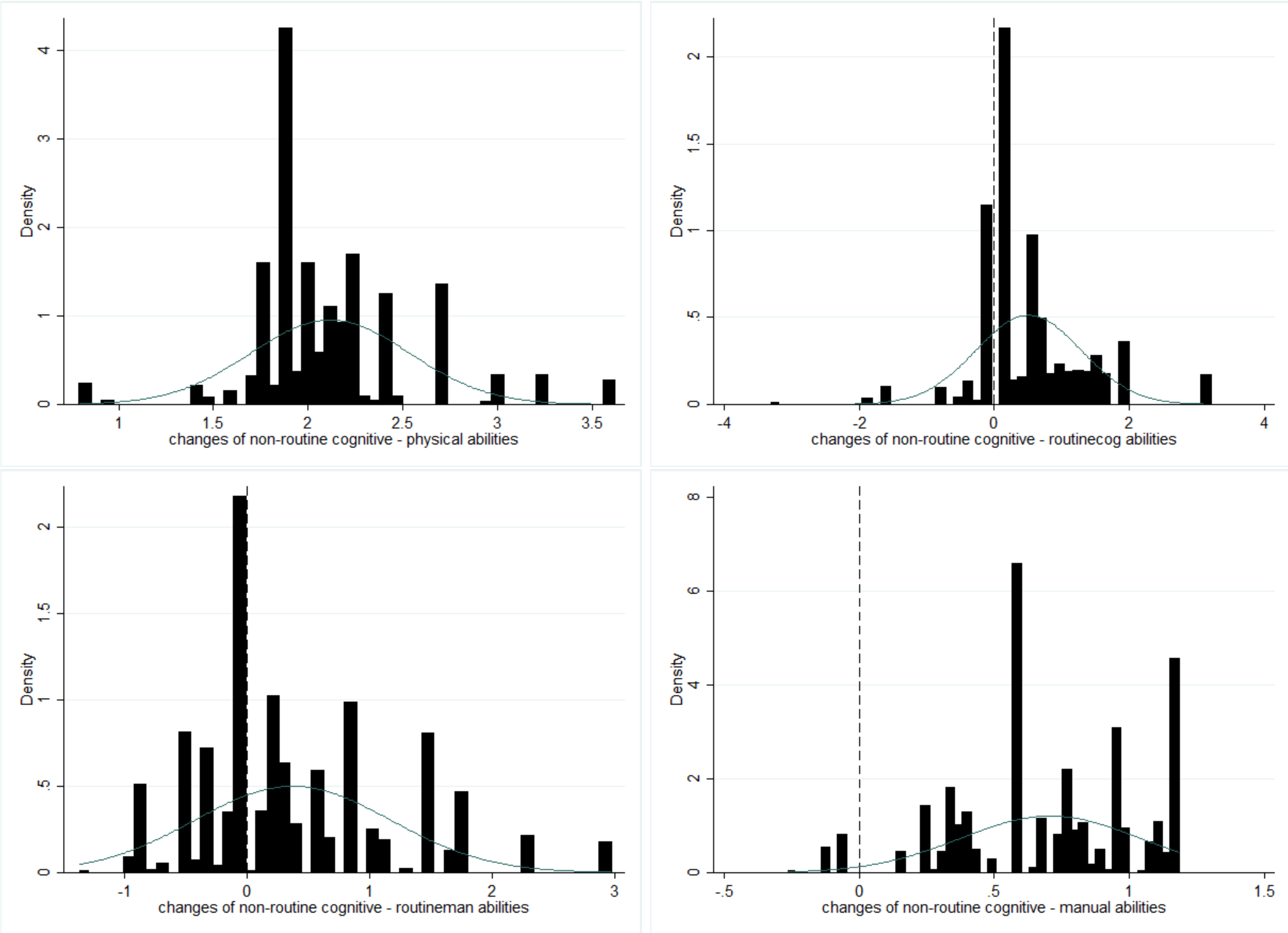


Figure 1.15: Smoothed changes of percentile ranks for all detailed skill groups between 1990 and 2010



Notes: Some of the extremely changes occupations include "Payroll and Timekeeping Clerks", "Billing and Posting Clerks", "Bank Tellers", "Cashiers", "Secretaries and Administrative Assistants", "Agricultural Inspectors", "Bus and Ambulance Drivers and Attendants", "Police Officers and Detectives", "Physician Assistants", "Securities, Commodities, and Financial Services Sales Agents".

Figure 1.16: Relative Changes of Abilities Scores



For this several data sets problem, I use the nearest neighbor matching method based on within-occupation-age income ranking, gender, region, marital status and education level. For the sample selection, I limit the sample to children who take their first jobs after 1997, (which is earliest year of which O*NET 3.0 builds on), given the age differences between parents and children are normally 20-30 years, hence it's safe to say that when their parents chose their jobs in their 30s, the U.S. occupations were still described by the skills distribution before the revolution. The basic assumption based on the matching is that people's relative income is decided by his/her occupational specific skills, along with other demographic variables, which is plausible especially for the later part of their working life. But this method will introduce measurement errors into our assignment variables, which I will discuss in the next part.

I then use the $\mu_{ability} + / - 1.96\sigma_{ability}$ of each occupation as the cutoffs of the ability for each occupation. The reason that I don't use the maximum and minimum is that I want to exclude those outliers. The choice of the cutoffs introduces another measurement errors of the discontinuity point ($\bar{s}$), which I will discuss as well in the next part.

1.6.5 Measurement Errors and Estimation Methods

The above illustrative example is just a simplification of our problem, in which I assume there is only a few occupations and two skills. In real life, there are over 60 of occupations with more than two dimensions of skills. For skills that partition those distribution spaces into occupations, we have four categories, including non-routine cognitive skills, routine, manual and physical abilities.

If we assume that individuals' occupational choices are based on all the categories, it would be a MRD problem with Multiple assignment variables (or one assignment variable depends on our choice of skill category)/dichotomous treatment. But as I describe in last part, the trend of this technology is basically substitution of human being's some physical and manual abilities, for example strength, visual and auditory abilities, and the routine cognitive abilities while the non-routine cognitive stay relatively stable or higher requirement for some jobs. In other words, if we want to describe this technological revolution, the main characteristics is the relative importance of non-routine cognitive skills versus others, which would theoretically give us the biggest change on the occupational choice cutoffs.

For the unidimensional composite distance measurement scenario, the RD problem in our case is the conventional RD with one assignment variable/dichotomous treatment, with various cutoffs for each occupation in the sample instead of being equal for all units. In this part, I'll follow the commonly method, which normalizes the score variable and use the zero cutoff on the normalized score for all observations to estimate a pooled RD treatment effect.

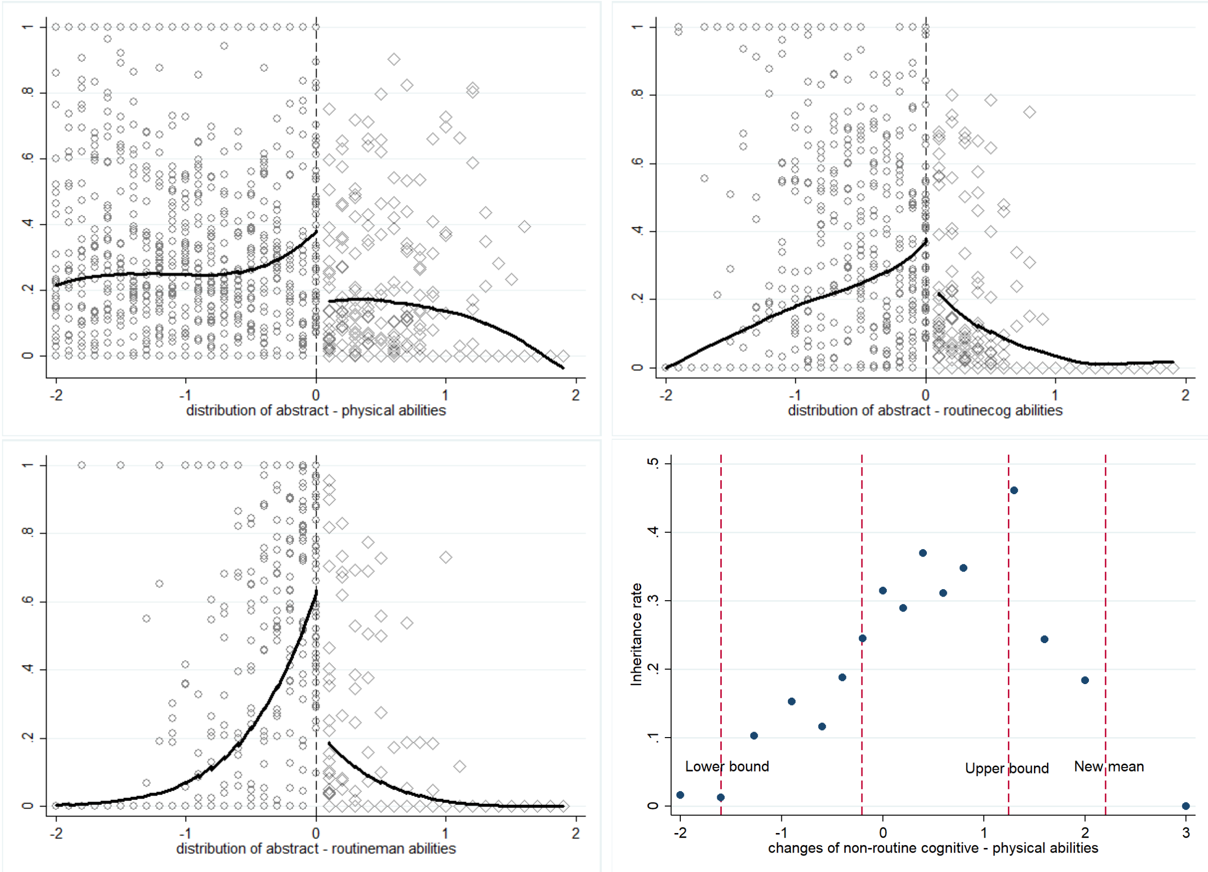
1.6.6 cutoffs of abilities

Based on this histograms of relative changes of occupations skill composition we discussed in last part, I choose three indexes for the continuous assignment variables $\Delta Cog\_Physical$, $\Delta Cog\_routinecog$, $\Delta Cog\_routineman$, since the change in ΔCog_manual is relatively small.

Figure 17 depicts the inheritance rate of each occupation around the original cutoff points.

We can see from Figure 17 that there is a discontinuity around our pooled cutoff points. Since most of the occupations have limited number of individuals to calculate the inheritance rate around the cutoffs, hence I choose the one specific occupation with enough observations in the data set (2881 observations) and shows for one occupation, there is also a significant change in the relative non-routine cognitive and physical abilities. We can see that this occupation change from physical ability as advantage to non-routine cognitive abilities as advantage, and there is a clear discontinuity in inheritance rate around the old cutoff point. As we discussed before, there are measurement errors of the assignment variables and the cutoff points. In the Appendix F.1, I derive the conditions on the measurement error that allow to retrieve the causal parameter $E[\tau|\mathbf{1}(O_C = O_P) = 1, \bar{s}^-]$ from raw data. The main results can be summarized as follows. First, I show that mathematically, the measurement error on the cutoff is the same with the measurement error on the assignment variable. And second, I show that the evidence provided in this part is not consistent with the hypothesis of having classical measurement error in S. A more general model for measurement error is therefore needed. I do that by following Horowitz and Manski (1995) and Battistin et al.

Figure 1.17: Discontinuity on the old cutoffs



(2009) and assuming that individuals whose observed value of the assignment variable is a mixture of the true value S^* and reported value S which is affected by measurement error. Formmaly, the observed value S_{obs} is

$$S_{obs} = S^* Z + S(1 - Z) \tag{1.21}$$

where Z is a binary variable equal to one for the exact reporters and equal to zero otherwise and S is the value contaminated by a measurement error. This is known as the contaminated sampling model discussed, amongst others, by Horowitz and Manski (1995).

Finally, I show that even if in the presence of the measurement error the sample analogue of (21) is inconsistent for the parameter of interest, the latter is nonetheless identifiable provided that conditional on S^* the process generating measurement errors is orthogonal to the process of interest. In particular, if the latter condition is satisfied (see the Appendix F.1 for further details) it is immediate to see that the following ratio

$$\tau = \frac{E\big(Y|S_{obs} = \bar{s}^-\big) - E\big(Y|S_{obs} = \bar{s}^+\big)}{E\big(\mathbf{1}(O_C = O_P)|S_{obs} = \bar{s}^-\big) - E\big(\mathbf{1}(O_C = O_P)|S_{obs} = \bar{s}^+\big)} \tag{1.22}$$

identifies the causal effect of $\mathbf{1}(O_C = O_P)$ on children's labor outcome at the cutoffs.

1.7 Empirical Results

1.7.1 Conventional RD with multiple cutoffs

First, we take the manipulations tests and the Table 8 reports the manipulations tests results for these three indexes. The key idea behind manipulation testing in this context is that, in the absence of systematic manipulation of the unit's index around the cutoff, the density of units should be continuous near this cutoff value. Thus, a manipulation test seeks to formally determine whether there is evidence of a discontinuity in the density of units at the known cutoff. Presence of such evidence is usually interpreted as empirical evidence of self-selection or non-random sorting of units into control and treatment status. All p-values are in parentheses, the effective number of observations are different because of the bandwidth choice, which is data-driven according to Cattaneo and Escanciano (2017), also because I drop all the occupation with percentile changing within 20% or the absolute ability score changing are the lowest 10% to exclude those occupations that don't have large ability composition changes. From the tests, we can see that all of them cannot reject the non hypothesis, suggesting there is no manipulation around the cutoff points.

Table 1.8: Manipulation Tests

Index	**Test Result**	**Sample size**	**bandwidth**
$\Delta Cog\_Physical$	1.244	Left: N = 9,133	h = 0.455
	(0.213)	Right: N = 7,100	h = 0.536
$\Delta Cog\_routinecog$	0.985	Left: N = 2,194	h = 0.397
	(0.324)	Right: N = 2,774	h = 0.521
$\Delta Cog\_routineman$	-1.296	Left: N = 1,072	h = 0.260
	(0.195)	Right: N = 1,677	h = 0.510

Second, Table 9 reports the probit regression results for the three indexes. From the results, we can see that all these three coefficients affect the inheritance rates around the cutoffs and the $\Delta Cog\_Physical$ affects the inheritance rate the most.

Table 1.9: Probit Model Results

Coef.	$\Delta Cog\_Physical$	$\Delta Cog\_routinecog$	$\Delta Cog\_routineman$
β	-0.404^{***}	$-.347^{***}$	$-.124^{***}$
	(-9.470)	(-8.045)	(-6.533)

t statistics is reported in parentheses

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

Table 1.10: $\Delta Cog\_Routineman$

	All, p(1)	All, p(2)	$\Delta PR > .2$,p(1)	$\Delta PR > .2$,p(2)	High Income Jobs	High Edu. Jobs
Conventional RD	4.974**	5.425***	3.369**	4.732***	6.99**	6.006***
	(2.027)	(5.263)	(2.457)	(5.263)	(2.678)	(2.345)
Bias-corrected	5.527**	7.029***	4.371**	6.254***	7.48***	6.357***
	(2.310)	(5.403)	(2.310)	(5.403)	(2.981)	(2.435)
Robust	5.527**	7.029***	4.371**	6.254***	7.48***	6.357**
	(2.161)	(5.266)	(2.161)	(5.266)	(2.630)	(2.279)
$UniqueInd.$	9,122	9,122	9,122	9,122	2,294	3,500
Bandwidth	0.439	0.381	0.439	0.381	0.667	0.566

z statistics in parentheses

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

Table 1.11: $\Delta Cog\_Routinecog$

	All, p(2)	All, p(1)	$\Delta PR > .2$,p(1)	$\Delta PR > .2$,p(2)	High Income Jobs	High Edu. Jobs
Conventional RD	3.671***	4.922**	5.721***	3.351**	6.624**	6.478***
	(2.931)	(2.224)	(2.931)	(2.224)	(2.038)	(3.466)
Bias-corrected	4.685**	5.694**	6.683**	3.955**	7.902*	7.433***
	(2.000)	(2.077)	(2.000)	(2.077)	(1.680)	(2.748)
Robust	4.685*	5.694*	6.683*	3.955*	7.902*	7.433**
	(1.747)	(1.893)	(1.747)	(1.893)	(1.5408)	(2.401)
$UniqueInd.$	9,122	9,122	9,122	9,122	2,294	3,500
Bandwidth	0.469	0.589	0.469	0.589	0.885	0.595

z statistics in parentheses

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

Table 1.12: $\Delta Cog\_Physical$

	All, p(1)	All, p(2)	$\Delta PR > .2$,p(1)	$\Delta PR > .2$,p(2)	High Income Jobs	High Edu. Jobs
Conventional RD	3.319**	4.519***	4.624***	4.967*	5.862**	5.693***
	(2.331)	(2.624)	(2.457)	(1.836)	(2.068)	(3.162)
Bias-corrected	3.699***	4.776***	4.690***	4.068**	5.994**	5.764***
	(3.002)	(3.069)	(2.718)	(2.028)	(2.386)	(3.295)
Robust	3.699***	4.776***	4.690**	4.068**	5.994**	5.764***
	(2.697)	(3.069)	(2.466)	(1.934)	(2.170)	(3.034)
$UniqueInd.$	9,122	9,122	5,864	5,864	2,294	3,500
Bandwidth	0.566	0.589	0.612	0.778	1.014	0.828

z statistics in parentheses

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

Table 10-12 report the results for the three indexes correspondingly. The optimal bandwidths are chosen and used according to Calonico, Cattaneo, and Titiunik (2014). I use the total sample, and the sample with at least 20% percentile changes, occupation income standing larger than 50%,.e. high income jobs, and the occupation education standing larger than 50%, i.e. jobs required high level of education. And for the full sample and large change sample, I report the results with order 1 and 2 for the local polynomial which is used to construct the point estimator. Conventional RD estimator, bias-corrected estimator are reported with also the robust confidence interval.

As we discussed before, this regression discontinuity design is only valid when there exist significantly large economic condition changes which ended up with large changes in terms of the ability compositions of occupations, in other words, the validity is questionable for those with little changes. So, the results from large relative changes samples would be the preferred results.

From these results, we can have two conclusions:

(1) On average, the marginal changes on parents' ability profile would increase their children's entry level rank(wage) on average 4.5%.

(2) For the high income occupations and high educational requirement occupations, the increase would be even larger, up to 7.24%.

As the same with all the regression discontinuity design, this estimation results are only for people with parents whose ability profiles were on the boundary of the old cutoffs. For people far away from the boundary, the effects should be smaller.

1.7.2 Frontier RD

The above summary just take the average over the three indexes we constructed, with the assumption that each of them affects the assignment independently by oneself. There is another potential possibility which is that we take into consideration the interactions effects. In other words, we take the subsample of parents who were just lower than the old cutoffs of one index, for example $\Delta Cov\_physical$, and compare purely the effect of $\Delta Cov\_routinecog$ being just lower than its cutoff.

We need to check the joint distribution of those three indexes, found out whether are moving in the same direction. As it turns out, for $\Delta Cov\_routinecog < 0$, 93.75% of individuals would have also $\Delta Cov\_physical < 0$ as well, which suggests the comovement of the relative changes between non-routine cognitive abilities and routine cognitive abilities and physical abilities. Hence, we end up with $\Delta Cov\_routinecog$ and $\Delta Cov\_routineman$. The Frontier RD can be illustrated as in Figure 13. Before we calculated the average effects around the old cutoffs α_s^A and r_s^A, separately. And for Frontier RD, we will calculate the the effect of merely lower than α_s^A within the range of $r < r_s^A$ and the effect of merely lower than r_s^A within the range of $\alpha < \alpha_s^A$, in other words, we are calculating now if theses two index work together on the occupation distribution, the effect of inheritance for occupation A compared with parents from occupation B, the α advantaged occupation, and C the r advantaged occupation.

Table 1.13: Frontier RD results

	Conventional	Bias-corrected	bandwidth	N
Average effect of merely lower than $\Delta Cov\_routinecog$, for people with parents' $\Delta Cov\_routineman$ within the old range	6.311***	7.231***	0.188	1,821
	(3.906)	(3.933)		
$(robust)$		(3.510)		
Average effect of merely lower than $\Delta Cov\_routineman$, for people with parents' $\Delta Cov\_routinecog$ within the old range	6.058**	6.997***	0.635	1,800
	(2.367)	(2.578)		
$(robust)$		(2.195)		

In Table 14, I report this frontier RD results for these two groups. As expected, the first-stage probit model gets more explanatory power to around $-.509$ and $-.512$ now, and the results are comparable than the original results.

1.7.3 Discussions

In this part, I use the RD design to estimate the causal relationship of the endogenous occupational choice $\mathbf{1}(O_C = O_P)$ on the labor outcomes, and obtain the results that the staying in the same occupation will increase children's income ranking around 5%. The technological revolution we exploit here not only gives extra variation of the innate ability but also the occupational related experience and within family training etc. since the change of the skill composition of occupations will at the same time invalidate the occupational related experience.

1.8 Conclusions

The intergenerational mobility of income has always be an active research area, though the channels behind this phenomenon is still open to discussion. Most research emphasize the genetic difference or/and other hereditary endowment difference or the education investment induced by the income difference. All those factors contribute to the increase of children's human capital.

My paper contributes to answering the question, and exploring the channel which doesn't increase children's human capital but through parents' accumulated social capital in their professional life. Instead of focusing on specific occupations without taking the occupational choice into consideration, I construct a new national representative data extract with the aim to fixing the selection bias in the commonly used survey data caused by children's endogenous moving behaviors after graduation. And I utilize the variation of occupational specific skills induced by the Technological Revolution. This variation is crucial to my study because it helps to disentangle the relationship of the intergenerational correlation in innate abilities and that in occupational choice. At the same time, the significant technological changes we observe in the data set remodeled the working content and activities of occupations, hence the potential explanation that children get benefit from their parents' occupational related experiences can also be excluded through this variation.

Three key findings emerge from my analysis. First, with the new data extract, I find that from 1986 to nowadays, around 30% of individuals would choose their parents' main occupations as their first jobs, and this ratio peaks at the recession. The children of parents with better-off occupations are more likely to stay in the same occupations while the children of poor families also have a high chance of taking the same occupations with their parents. Second, by the regression discontinuity which is induced by the technological revolution, I find that at the entry-level jobs, the individuals who stay in the same occupations with their parents rely significantly more on the strong social connections to find their first jobs, and at the same time, the rank of their annual income is around 5% higher than the counterparts. Third, I find that this wage premium at the entry level jobs fade away slowing along with time, and those people in the long run have higher income risk and instability.

1.9 Appendix

1.9.1 Acronyms List

In this part, I list all the acronyms I use in this paper, in the order of appearance in the paper.

- **PSID** – Panel Study of Income Dynamics
- **SIPP** – Survey of and Survey of Income and Program Participation
- **COC** – 1990 Census Occupational Classification System
- **SOC** – 1980 Standard Occupational Classification
- **DOT** – Dictionary of Occupations
- **O*NET** – Occupational Information Network
- **ISSP** – International Social Survey Programme
- **CPS** – Current Population Survey

1.9.2 Data Appendix

Table 1.14: Accuracy Rate of Retrospective Data of Parents' Occupations

	Individuals	**Jobs**
Number of Matched	1,215	1,517
Number of Total	3,218	6,353
Match Rate	37.76%	23.87%

Notes: I constrain the sample to people whose parents' working information is available since no later than their 40s and before retirement. And the occupational category is COC broad category.

Table 1.15: Age Profile of Moving Out Schedule

Age of Graduation	Move out within 1 year	Move out after 2 years	Move out after 4 years
15	10.62%	79.53%	52.02%
16	11.66%	75.16%	51.36%
17	16.56%	68.35%	45.40%
18	19.79%	66.75%	44.94%
19	26.16%	58.41%	34.95%
20	40.59%	43.04%	22.59%
21	50.90%	33.09%	14.88%
22	48.29%	32.52%	13.87%
23	54.59%	27.81%	13.80%
24	59.92%	20.94%	7.18%
25	87.33%	0.00%	0.00%
26	100.00%	0.00%	0.00%

Notes: this table is based on author's calculation using PSID Individual and PSID Family Data Index from 1986 to 2013 rounds.

Table 1.16: Occupation Categories (COC)

1. Executives, Administrative, & Managerial (1)	34. Police (12)
2. Management Related (2)	35. Guards (12)
3. Architects (3)	36. Food Preparation and Service (13)
4. Engineers (3)	37. Health Service (6)
5. Math and Computer Science (3)	38. Cleaning and Building Service (13)
6. Natural Science (4)	39. Personal Service (13)
7. Health Diagnosing (5)	40. Farm Managers (14)
8. Health Assessment (6)	41. Farm Non-Managers (14)
9. Therapists (6)	42. Related Agriculture (14)
10. Teachers, Postsecondary (7)	43. Forest, Logging, Fishers, and Hunters (14)
11. Teachers, Non-Postsecondary (8)	44. Vehicle Mechanic (15)
12. Librarians and Curators (8)	45. Electronic Repairer (15)
13. Social Scientists and Urban Planners (4)	46. Misc. Repairer (15)
14. Social, Recreation, Religious Workers (4)	47. Construction Trade (15)
15. Lawyers and Judges (5)	48. Extractive Operation (14)
16. Arts and Athletes (4)	49. Precision Production, Supervisor (16)
17. Health Technicians (9)	50. Precision Metal (16)
18. Engineering Technicians (9)	51. Precision Wood (16)
19. Science Technicians (9)	52. Precision Textile (16)
20. Technicians, Other (9)	53. Precision Other (16)
21. Sales, All (10)	54. Precision Food (16)
22. Secretaries (11)	55. Plant and System Operator (17)
23. Information Clerks (11)	56. Metal and Plastic Machine Operator (17)
24. Records Processing, Non-Financial (11)	57. Metal & Plastic Processing Operator (17)

25. Records Processing, Financial (11)
26. Office Machine Operator (11)
27. Computer & Communication Equip. Operator (11)
28. Mail Distribution (11)
29. Scheduling and Distributing Clerks (11)
30. Adjusters and Investigators (11)
31. Misc. Administrative Support (11)
32. Private Household Occupations (13)
33. Firefighting (12)
58. Woodworking Machine Operator (17)
59. Textile Machine Operator (17)
60. Printing Machine Operator (17)
61. Machine Operator, Other (19)
62. Fabricators (18)
63. Production Inspectors (18)
64. Motor Vehicle Operator (19)
65. Non Motor Vehicle Operator (19)
66. Freight, Stock, and Material Handlers (18)
67. Military (20)

Notes: Our 66 market occupations (except military occupations) are based on the 1990 Census Occupational Classification System. We use the 66 sub-headings (shown in the table) to form our occupational classification. Seehttp://www.bls.gov/nls/quex/r1/y97r1cbka1.pdf for the sub-heading as well as detailed occupations that correspond to each sub-heading. The more broader category including only twenty occupations are also provided for robust check. The number in parentheses) refers to how we group these 67 occupations into the twenty broader occupations. For example, all occupations with a 11 in parentheses refers to the fact that these occupations were combined to make the 11th occupation in our broader occupation classification.

Table 1.17: Definition of skill type in DOT and O*NET

Skill Type	**Survey**	**Category**		**Skill Description**
Communication Interactive [43]	O*NET	Abilities	A.1.a	Verbal Abilities
		Activities	A.4	Interacting with others
	DOT	Aptitude	V	Ability to understand and use words effectively
		Temperaments	I	Influencing people in their opinions, attitudes and judgments
		Temperaments	P	Dealing with people
Analytical	O*NET	Abilities	A.1.b	Idea Generation & Reasoning Abilities
		Activities	A.2.b	Reasoning and decision making
	DOT	Aptitude	G	General ability to learn, reason, make judgments
		Temperaments	D	directing, controlling or planning activities of others
		Temperaments	J	Making judgment and decisions
Quantitative	O*NET	Abilities	A.1.c	Quantitative abilities
		Activities	A.2.a	Information/Data Processing
	DOT	Aptitude	N	Ability to understand and perform mathematical functions
Routine Cognitive	O*NET	Abilities	A.2.d	Memorization
			A.2.e	Perceptual Abilities
		Contents	C.3.b.7	Importance of repeating the same tasks
			C.3.b.9	Structured v. Unstructured work (reverse)
	DOT	Aptitude	Q	Clerical Perception

Continued on next page

[43]Notes: Interactive skills include the ability to comprehend and express both oral and written material. They also include the importance of communicating with coworkers and people outside a person's workplace. Strictly speaking, quantitative and analytical skills are not synonymous. Lawyers, for example, require very little mathematical acumen but a high degree of inductive reasoning ability. Nonetheless, we treat quantitative and analytical skills as synonyms so that the terms represent the importance of performing mathematical functions, analysis of data and information, and deductive and inductive reasoning tasks.

Table 1.17 – *Continued from previous page*

Skill Type	Survey	Category		Skill Description
		Temperaments	T	Adaptability to situations requiring the precise attainment of set limits, tolerances, or standards
			U	Working under specific instructions
Routine Manual	O*NET	Abilities	A.2.a	Fine manipulative abilities
		Activities	A.3.a	Performing physical & manual work activ.
	DOT	Aptitude	F	Finger dexterity, ability to finger and manipulate small objects
			M	Ability to handle placing and turning motions
		Temperaments	R	Performing repetitively or short cycle work
Non-routine Manual	O*NET	Abilities	A.2.b	Control movement abilities and reaction time
			I.A.1.f	Spatial Abilities
		Activities	A.3.b	Performing complex/technical activities
	DOT	Aptitude	E	eye/hand/foot coordination, motor responsiveness to visual stimuli
			K	Motor coordination, ability to coordinate eyes, hands, fingers
			S	Ability to visualize three dimensional objects from two
Strength	O*NET	Abilities	A.3.a	Physical Strength Abilities
	DOT	Physical Demands	1	Strength
Body Flexibility	O*NET	Abilities	A.3.c	Flexibility, Balance and Coordination
	DOT	Physical Demands	2-10	Flexibility, Balance and Coordination
Visual	O*NET	Abilities	A.4.a	Visual Abilities

Continued on next page

Table 1.17 – *Continued from previous page*

Skill Type	Survey	Category	Skill Description	
	DOT	Physical Demands	15-20	Visual Abilities
		Aptitude	C	Color Discrimination
		Aptitude	P	Form Perception
Auditory	O*NET	Abilities	A.4.b	Auditory and Speech Abilities
	DOT	Physical Demands	13	Hearing

Figure 1.18: The Sample Size of PSID

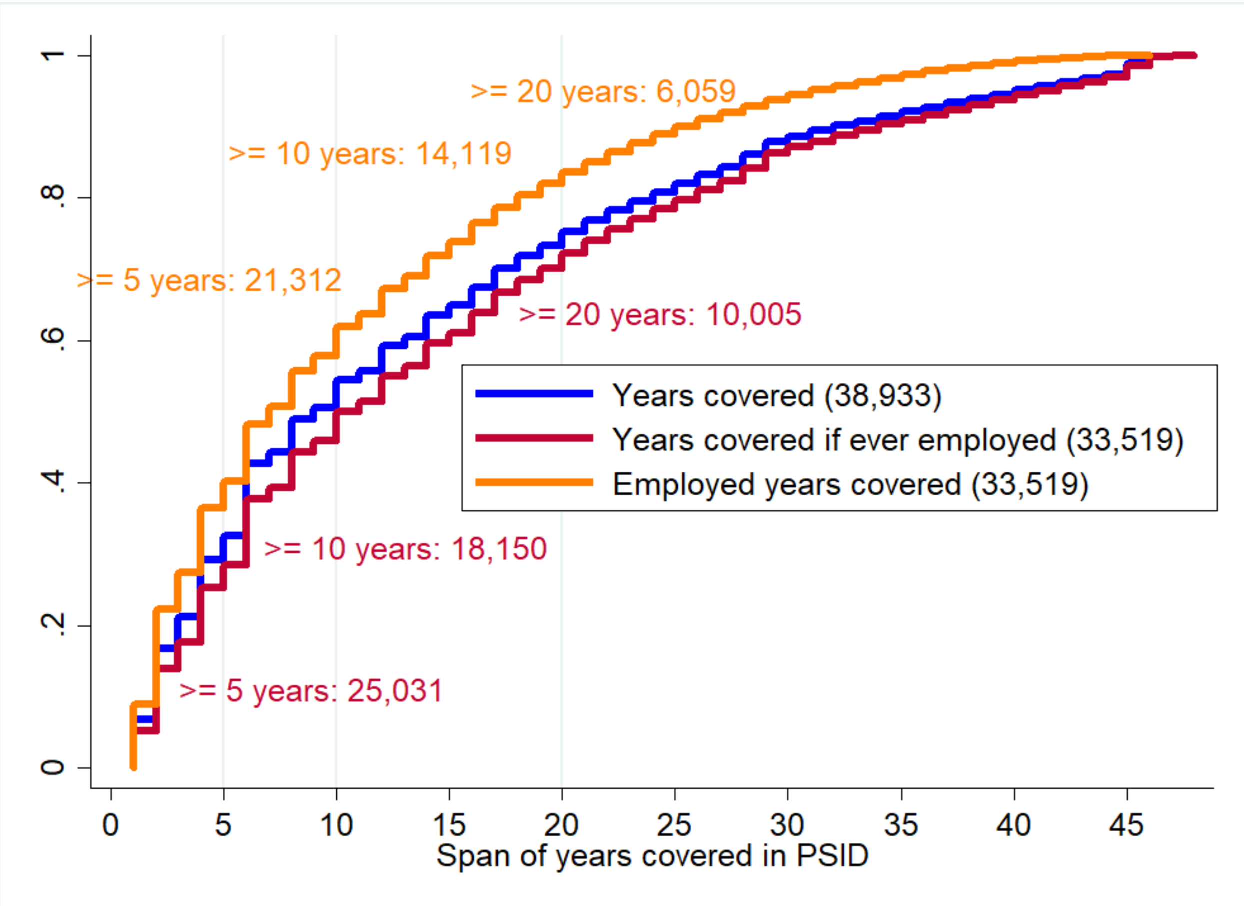
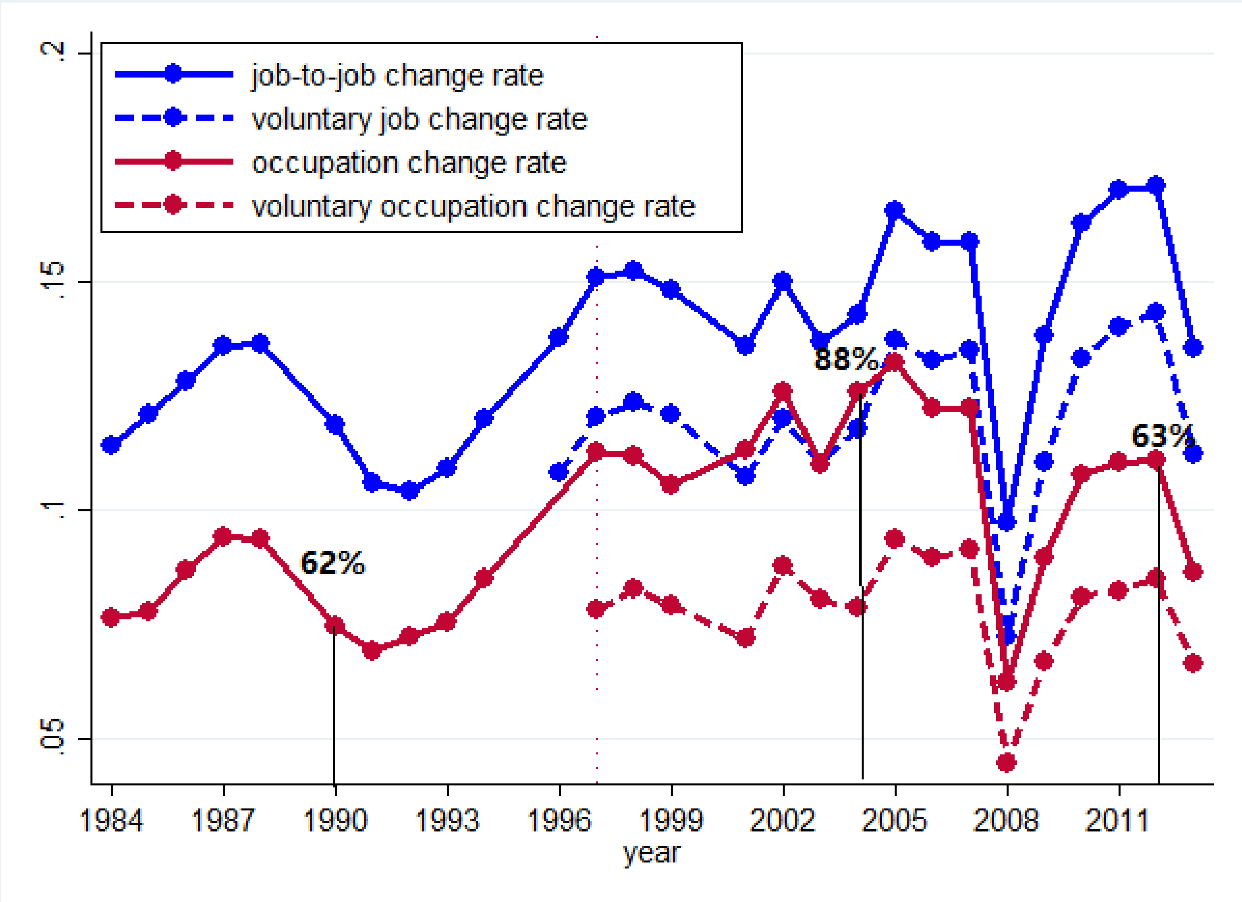


Figure 1.19: Occupational change rate



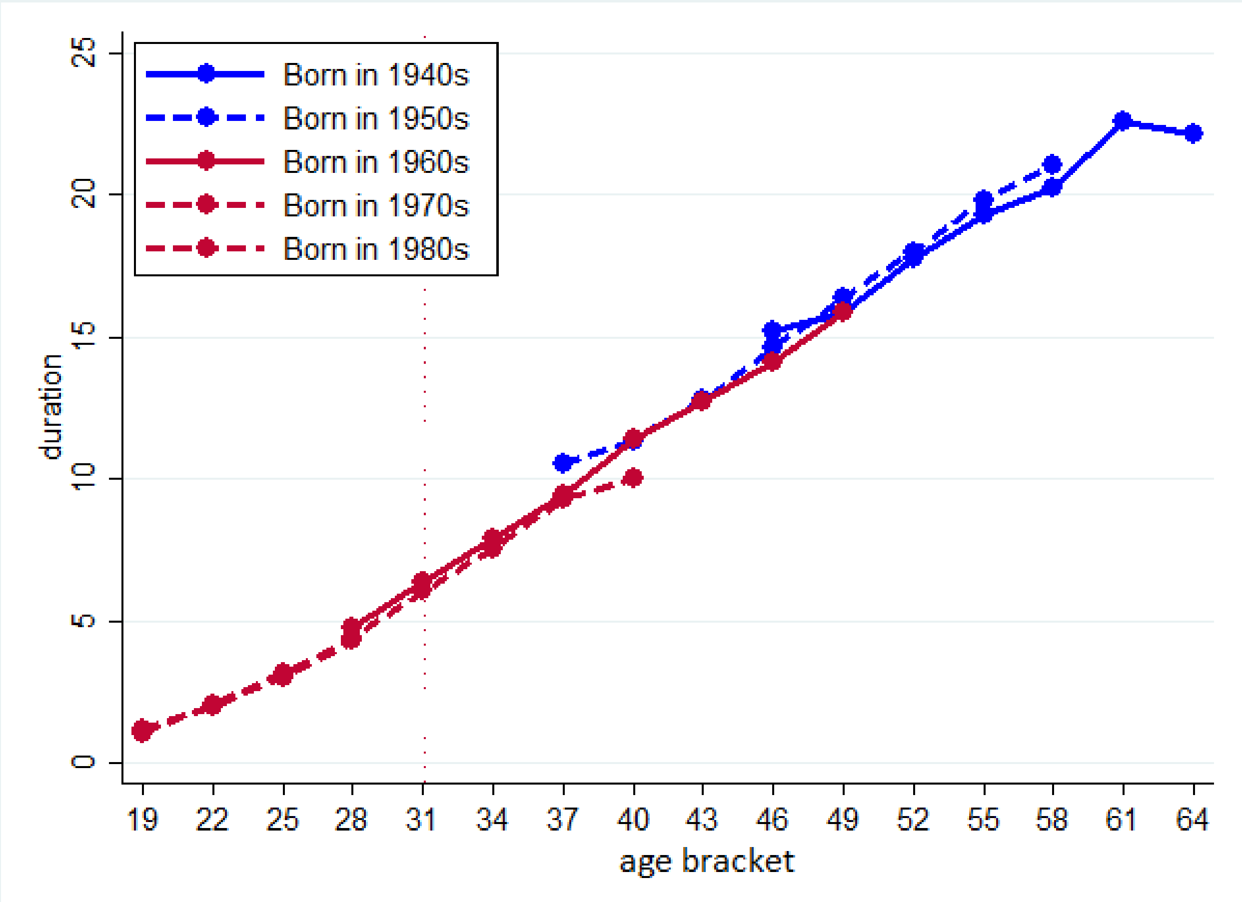
This figure depicts the annual average quarterly occupational change rate (3-digit) from 1984 to 2012. We can see that the occupational change follows almost the same trend with job transition, and around 70% of people who change their jobs would change their occupations at the same time. As to the time trend, the occupational change rate increases higher than the job change rate and reached the peak around 88% in 2004. But after the 2008 Great Recession,the occupational change rate drops to the 1980s level. The above pattern applies to the voluntary job changers as well.

1.9.3 Regression Appendix

1.9.3.1 Intergenerational Occupational Inheritance

1.9.3.2 Robustness check for Entry-level Regression

Figure 1.20: The duration of occupations



Notes: This figure depicts the average length of time in the same line of work for different age groups and different cohorts and shows that this occupation changing behaviors are consistent over different cohorts.

1.9.4 The effect of staying in the same industry

In the main body of the paper, I only consider the occupational choice, but a job consists both the occupational choice and the industrial choice.

I observed in the data set that during the period 1986 to 2013, for people with full experience, for those who choose their parents' occupation as the first occupation, about half of them (13.93% in the total sample) would at the same time choose a different industry, while the other half(11.76% in the total sample) would stay in the same industry cluster. And 13.69% of the total population would choose stay in the same industry cluster but choose a different line of work. Hence, we can

Table 1.18: Children's occupational choices (Probit Mode)

	Occ. Only	Occ. and Ind.
Income standing	-4.004***	-6.388***
	(-19.74)	(-15.82)
Income standing^2	3.678***	5.086***
	(15.55)	(12.11)
change of income standing	2.774***	2.339***
	(10.02)	(6.20)
Parents' performance within occupation	-0.0170	-0.223*
	(-0.25)	(-2.06)
Control	Yes	Yes
Unique Ind.	30,381	30,381

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This is the regression results corresponding to the Figure 5 in Section 2 which depicts the relationship of occupational inheritance with respect to the income standing. In the first regression, I consider only the occupational choice, and in the second equation, the dependent variable equals to 1 if children choose the same occupation or industry, and the income standing is the average of occupational income standing and industry income standing.

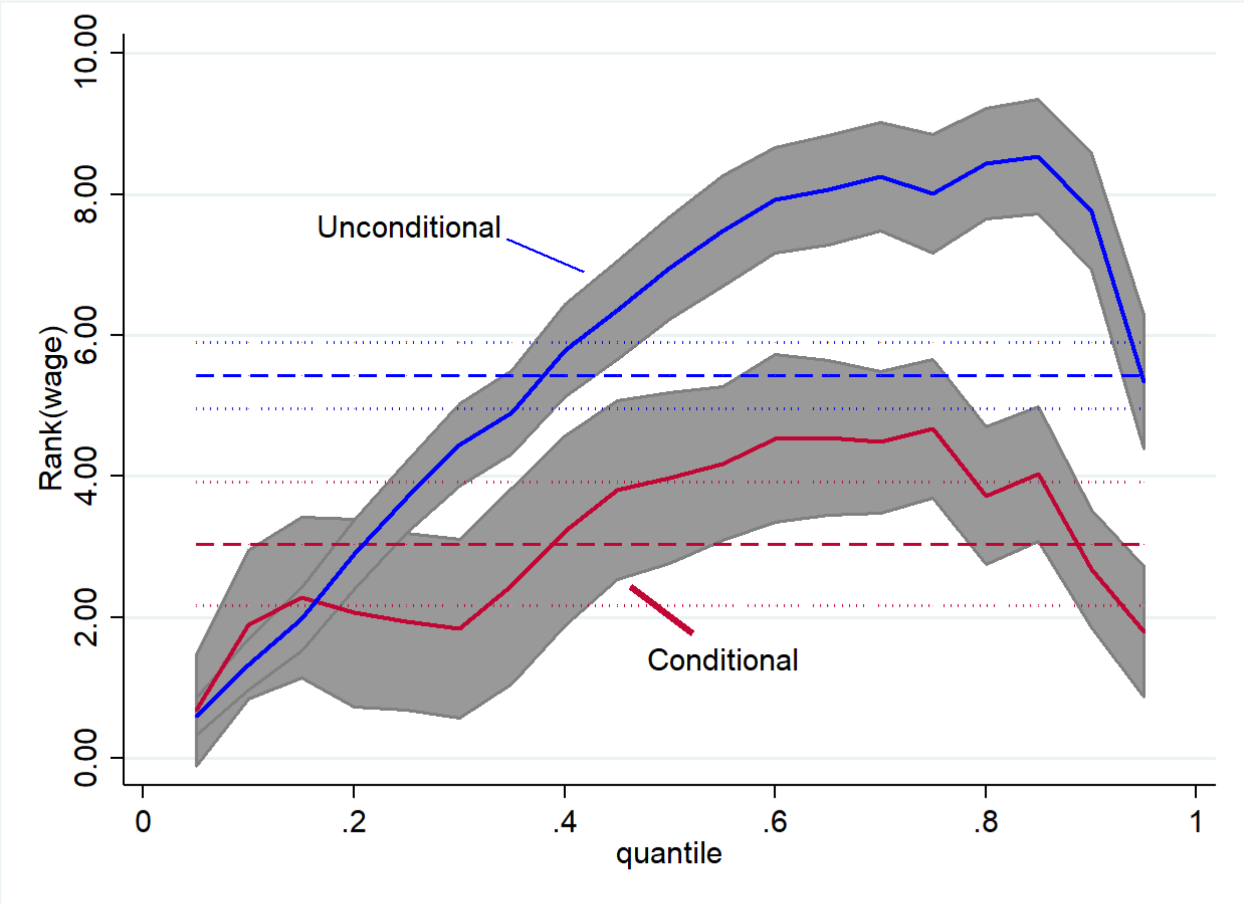
categorize all people's behavior using the following category variable:

$$C_{job} = \begin{cases} 1 & \text{if } \mathbf{1}\{O_C == O_P\} = 1 \& \mathbf{1}\{I_C == I_P\} = 1 : \textbf{both same} \\ 2 & \text{if } \mathbf{1}\{O_C == O_P\} = 1 \& \mathbf{1}\{I_C == I_P\} = 0 : \textbf{same occupation only} \\ 3 & \text{if } \mathbf{1}\{O_C == O_P\} = 0 \& \mathbf{1}\{I_C == I_P\} = 1 : \textbf{same industry only} \\ 4 & \text{if } \mathbf{1}\{O_C == O_P\} = 0 \& \mathbf{1}\{I_C == I_P\} = 0 : \textbf{totally different job} \end{cases} \tag{1.23}$$

In this part, I'll document the stylized facts for those people with the same industry.

I run the Eq.(3) within the sample that $C_{job} = 2$ or $C_{job} = 4$, in other words, I compare the people with only the same industry and different occupation with the people with completely different

Figure 1.21: Unconditional and conditional difference between the two groups



occupation and industry.

The regression results are listed in Appendix Table C.7, with the same control variables in the main body of the paper. We can see that, these results are similar in magnitude with the occupational inheritors.

And I then test the difference of log(wage) and rank(wage) between the group $C_{job} == 1$ and $C_{job} == 3$, i.e., I compare the people who choose the same industry and same occupation with people who choose the same industry but different occupation. Theoretically, if the genetic advance in occupational choice is the main drive of the entry-level premium, then this premium would not exist for people who choose a different occupation but just stays in the same industry. Appendix Table C.8 reports the F test, From which we can see that, the two groups are comparable at least for the entry-level jobs.

Table 1.19: Regression results for the entry-level jobs (SOC 3-digit)

	log(wage), full	log(wage), HS	log(wage), College	Rank(wage),full
$\mathbf{1}\{O^C_{first} == O^P_{major}\}$	0.321***	3.551***	2.472**	4.061***
	(0.0382)	(0.580)	(0.838)	(0.826)
experience	0.0859***	1.009***	1.136***	1.038***
	(0.00197)	(0.0384)	(0.0572)	(0.0513)
$I * experience$	-0.0165**	-0.158*	-0.184	-0.0631
	(0.00566)	(0.0778)	(0.117)	(0.108)
Control	Yes	Yes	Yes	Yes
$UniqueInd.$	30,381	30,381	14,564	15,168
adj. R^2	0.361	0.138	0.081	0.082

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This is the regression results for Eq. (3), also a robustness check for Table 4, with the detailed SOC categories to define the dummy variable $\mathbf{1}\{O^C_{first} == O^P_{major}\}$.

1.9.5 Model Specification for estimating income risks

Assume the real (log) income $logY$ can be decomposed into a permanent component $\mathbf{P}$ and a mean-reverting transitory component v. The income process for each household i is

$$log\mathbf{Y}_{i,t} = \mathbf{Z}'_{i,t}\varphi_t + \mathbf{P}_{i,t} + v_{i,t}$$

where t indexes time and $\mathbf{Z}$ is a set of income characteristics observables and known by consumers at time t. We assume that the permanent component $\mathbf{P}_{i,t}$ follows a martingale process of the form

$$\mathbf{P}_{i,t} = \mathbf{P}_{i,t-1} + \zeta_{i,t}$$

where $\zeta_{i,t}$ is serially uncorrelated and the transitory component $v_{i,t}$ follows an MA(1) process:

$$v_{i,t} = \varepsilon_{i,t} + \theta\varepsilon_{i,t-1}$$

Because of the PSID data problem, starting from 1997, the income data is available every two years. Hence we can get the first difference for income before 1997 and second difference for income afterwards. If we define $y_{i,t} = log\mathbf{Y}_{i,t} - \mathbf{Z}'_{i,t}\varphi_t$ as the log of real income net of predictable individual

Table 1.20: Regression results for the entry-level job: Industry only

	log(wage), Full sample	log(wage), HS	log(wage), college	Rank(wage)
$\mathbf{1}\{I^C_{first} = I^P_{major}\} = 1$	0.324***	0.280***	0.519***	2.530***
	(14.17)	(7.46)	(15.90)	(6.45)
Experience	0.145***	0.170***	0.176***	1.589***
	(68.47)	(55.40)	(51.82)	(41.02)
I*Experience	0.00445	0.0194**	-0.0182**	0.00260
	(0.96)	(2.59)	(-2.85)	(0.04)
Control	X	X	X	X
$UniqueInd.$	12,753	5,755	6,506	12,753
adj. R^2	0.502	0.386	0.393	0.289

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.21: F test $Wage_{C_{job}=1} - Wage_{C_{job}=3}$

Test	**F-stat**	**p_value**
$log(wage)_{C_{job}=1} - log(wage)_{C_{job}=3} \vert X = 0$	0.66	0.4166
$rank(wage)_{C_{job}=1} - rank(wage)_{C_{job}=3} \vert X = 0$	2.60	2.60

components, then the available unexplained income growth is

$$\Delta y_{i,t} = \zeta_{i,t} + \varepsilon_{i,t} + (\theta - 1)\varepsilon_{i,t-1} - \theta\varepsilon_{i,t-2}, t <= 1997$$

$$\Delta_2 y_{i,t} = (\zeta_{i,t} + \zeta_{i,t-1}) + (v_{i,t} - v_{i,t-2}) = (\zeta_{i,t} + \zeta_{i,t-1}) + \Delta_2 v_{i,t}, t >= 1999$$

Hence we can identify the MA coefficient θ, the variance of permanent and transitory income risk before 1997, sums of two year of permanent income afterwards, and the average of two years of transitory income risk afterwards.

Table 1.22: Distribution of ties for job hunting by source of helps

Ties	Friends	Family	Labor Market
Strong tie	5.97	22.41	0
Referee	46.77	49.79	8.97
Information	47.26	27.8	91.03
Total Percentage	35.96	21.57	42.49

Table 1.23: Month distributions of first jobs

Month	Percentage	Cumu. percentage
1	7.7	7.7
2	5.18	12.87
3	10.17	23.04
4	12.48	35.52
5	10.85	46.38
6	17.84	64.22
7	6.53	70.75
8	6.68	77.42
9	6.05	83.48
10	6.06	89.54
11	5.23	94.77
12	5.23	100

1.9.6 Social Network in Job hunting behavior

1.9.7 Derivations and Proofs

The propositions in the paper summarize the key results from the model. This appendix shows how to derive the results. [44]

[44]Here we assume there is only two periods, but it will give the same conclusion if we model the dynamics of income more complicated, for example, with infinite periods, with probability of dying v at each period, and the income level depends on the probability of true ability revealed. Since in this model, the probability of death and timing of revealing is not the main part, so without loss of generality, I simply the whole life of dynamics into two periods.

Table 1.24: Probit Regression Results for starting month

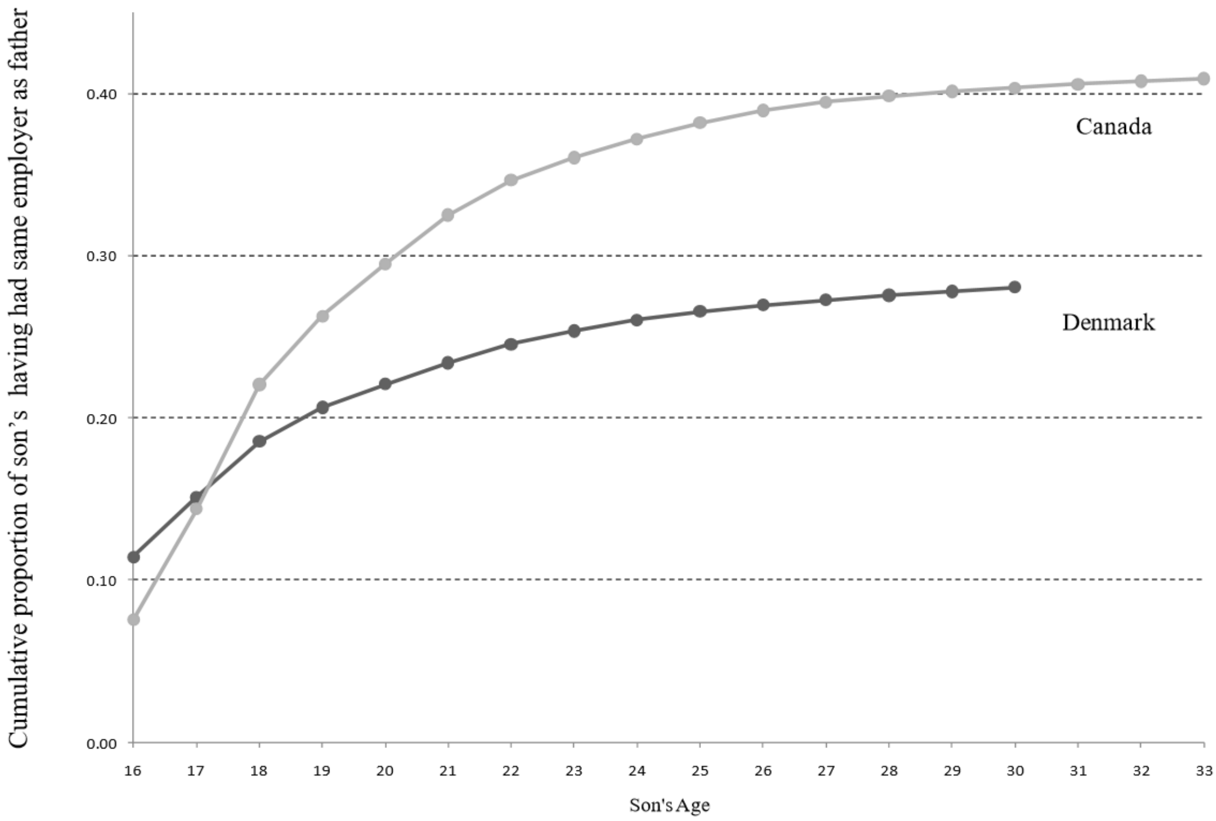
	June	March	April	May
$\mathbf{1}\{O^C_{first} == O^P_{major}\} = 1$	0.0709**	-0.0148	0.00203	-0.0114
	(2.62)	(-0.52)	(0.08)	(-0.48)
Control	X	X	X	X
N	22,228	22,228	22,228	22,228

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This is the results of the probit model in Section 3.2

Figure 1.22: Proportion of sons currently employed or employed at some point with an employer their fathers had worked for in the past: Corak and Piraino(2011)



Proposition 1.9.1 *(Occupational Inheritance Choice)*

Proof. First, we can solve the individual's maximization problem.

Each individual chooses occupation i and education spending E_i, the education level s_i is determined automatically. The efficient labor unit is decided by the education level and the education cost, this is to model the fact that even with same years of education, the type of school, for example private or public high school, or medical major or mathematics major, the cost would be different, and depends on the income level.

S_i and E_i is solved by the FOC:

$$S_i = \frac{1}{1 + \frac{1-\eta}{2\beta\phi_i}}$$

$$E_i^* = \eta(Y_{1,i} + Y_{2,i})$$

Hence

$$E_i^* = \left(\eta\kappa_{ij} w_i \epsilon_i s_i^{\phi_i}\right)^{\frac{1}{1-\eta}}$$

in which

$$\kappa_{ij} = \frac{1}{2}\big[\exp(1+\tau_{ij})T(1)+\exp(1+p\tau_{ij})T(2)\big] = \begin{cases} \frac{1}{2}\big[\exp(1+\tau_{jj})T(1)+\exp(1+p\tau_{ijj})T(2)\big] & \text{if } i=j \\ \frac{1}{2}\exp(1)\big[T(1)+T(2)\big] & \text{if } i \neq j \end{cases}$$

is the premium weighted experience in occupation i from group j, hence if the offspring choose a different occupation, κ_{ij} is the average experience in two periods.

If fathers' occupation $j >= L_2$, then the credit constraint doesn't bind, (s_i, E_i^*) would be optimized choice of education level and education cost.

On the other hand, if $L_1 < j <= L_2$, i.e. the initial education fund is not enough to support the education higher than H_2, then this person should choose another i^{**} within $\{1, \ldots, M_2\}$. For fathers' whose occupations are lower than L_1, it's likewise.

Then this person would choose the next optimal occupation with the s_i lower than the optimal choice. After substituting the expression for human capital into the utility function, indirect utility for an individual from group j working in occupation:

$$U_{i|j} = (\tilde{w}_{ij}\epsilon_i)^{\frac{2\beta}{1-\eta}}$$

in which

$$\tilde{w}_{ij} \equiv \kappa_{ij} w_i s_i^{\phi_i}[(1-s_i)]^{\frac{1-\eta}{2\beta}}$$

Without the labor market distortion, the above return of ability is

$$\tilde{w}_{ij} = \bar{T} w_i s_i^{\phi_i}[(1-s_i)]^{\frac{1-\eta}{2\beta}}$$

So to maximize the utility is equivalent to maximize the expression $W_i = \tilde{w}_{ij}\epsilon_i$ and we can see that it also follows the Fréchet distribution of as

$$G_{i|j}(W_i) = F(W_i <= u) = F(\epsilon_i <= \frac{u}{\tilde{w}_{ij}}) = \exp\Big(- T_{ij}\tilde{w}_{ij}^{\theta}u^{-\theta}\Big)$$

For people who choose the same job as their parents, I use the notation $G_{j|j}$ for the distribution if there were no distortion and $G_{\tau,j|j}$ as the one for the distortion case. And I use the notation $\tau_j = \frac{\kappa_{jj}}{T}$ as the composite premium for taking the same occupation as their parents. [45]

Then we can calculate each element in the transition matrix for different occupations, take $i^* = j$ as an example

$$p_{jj} = Pr(\text{Always takers}) + Pr(\text{Compliers})$$

in which

$$Pr(\text{Alway takers}) = Pr(\frac{1}{\tau_j}W_{jj} >= \{\max W_{kj}, k \in I_j\}|j)$$

and

$$Pr(\text{Compliers}) = Pr(\exists k, \text{ s.t. } \frac{1}{\tau_j}W_{jj} < W_{kj} < W_{jj} \text{ and } W_{kj} >= \{\max W_{gj}, g \neq k, j\}|j)$$

Hence to summarize, we get:

$$p_{ij} = Pr(W_{ij} >= \{\max W_{kj}, k \in I_j\}|j) = \int_0^{\infty} \prod_{k \in I_j} G_{kj}(u)dG_{ij}(u) = \frac{T_{ij}\tilde{w_{ij}}^{\theta}}{\sum_{I_j} T_{kj}\tilde{w_{kj}}^{\theta}} \tag{1.24}$$

in which $\sum_{I_j} T_{kj}\tilde{w_{kj}}^{\theta} \equiv \Phi_j$ is the total efficient labor from family with father's occupation in j. And of course we can have the inheritance rate for each occupation p_{jj} which is higher for high income jobs and low income jobs.

■

Proposition 1.9.2 *(Average Quality of Workers)*

Proof. The ϵ^* follows also the Fréchet distribution that:

$$G_{|j}(x) = Pr(\epsilon^* < x|j) = \exp[-\sum_{I_j}(\frac{\tilde{w}_s}{\tilde{w}^*})x^{-\theta}]$$

And combine with

$$h_{ij}\epsilon_i|j = \left(s_i^{\phi_i}\right)^{\frac{1}{1-\eta}}\left(\eta\kappa_{ij}w_i\right)^{\frac{\eta}{1-\eta}}\epsilon^{\frac{1}{1-\eta}}$$

[45] We assume $\tau_j > 1$ for all occupations, i.e. there is some positive advantage if taking the parents' occupations

And the property of Fréchet distribution

$$E\left[\epsilon^{\frac{1}{1-\eta}} | \text{ choose occupation i and father's occupation is j}\right] = (\frac{1}{p_{ij}})^{\frac{1}{\theta}\frac{1}{1-\eta}}\Gamma(1-\frac{1}{\theta}\frac{1}{1-\eta})$$

■

The proofs the last three are simple algebra changes from Proposition 2

Solve the equilibrium

Given productivity A_i, human capital accumulation coefficient h_{ij} and labor market friction τ_{ij}, the competitive equilibrium is a set C_{ij},s_i,p_{ij},w_i,H_i such that:

- Given the innate talent draw and father's occupation j, each individual chooses the optimal occupation i , and makes optimal consumption decision C_{ij} and study-leisure decision
- The labor market clears

$$H_i^{demand}(t) = (\frac{A_i(t)^{\frac{\sigma-1}{\sigma}}}{w_i(t)})^{\sigma} Y(t)$$

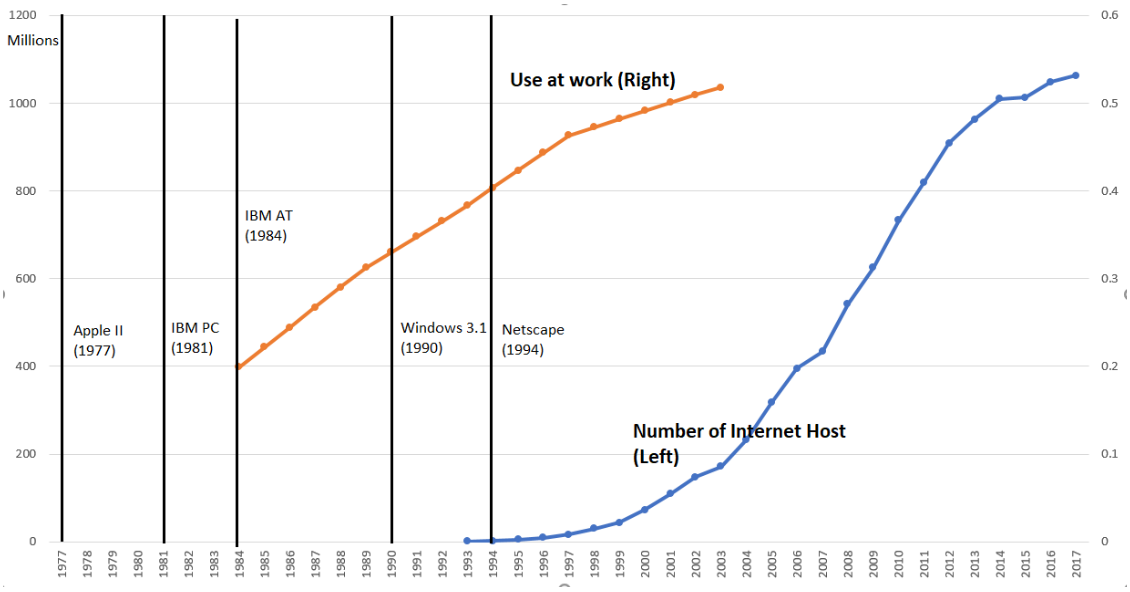
$$\sum_{j=1}^{M} H_{ij}^{supply}(t) = \sum_{j=1}^{M} \pi_j^L p_{ij} T(t) E\left(h_{ij}\epsilon_i\right)$$

- Goods market clears

1.9.8 The Technology Revolution

1.9.8.1 Changes in Occupational Specific Skills

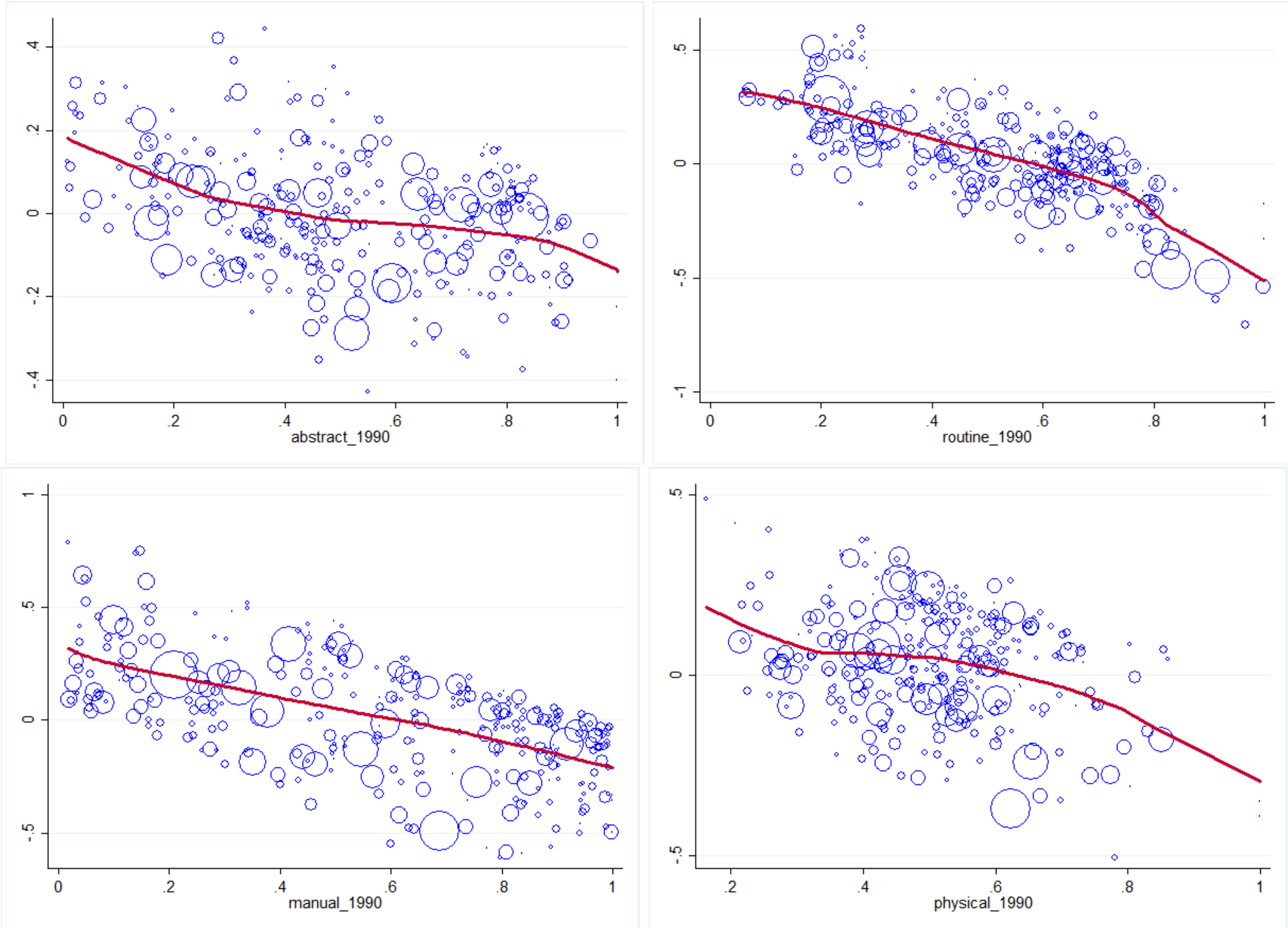
Figure 1.23: Measure of technological change



Notes: this figure presents a time line of key events associated with the development of personal computers, plotted along with two simple measures of the extent of computer-related technological change. Although electronic computing devices were developed during World War II and the Apple II was released in 1977, many observers date the beginning of the "computer revolution" to the introduction of the IBM-PC in 1981. This was followed by the IBM-XT (the first PC with built-in disk storage) in 1982 and the IBM-AT in 1984. As late as 1989, most personal computers used Microsoft's DOS operating system. More advanced graphical interface operating systems only gained widespread use with the introduction of Microsoft's Windows 3.1 in 1990.

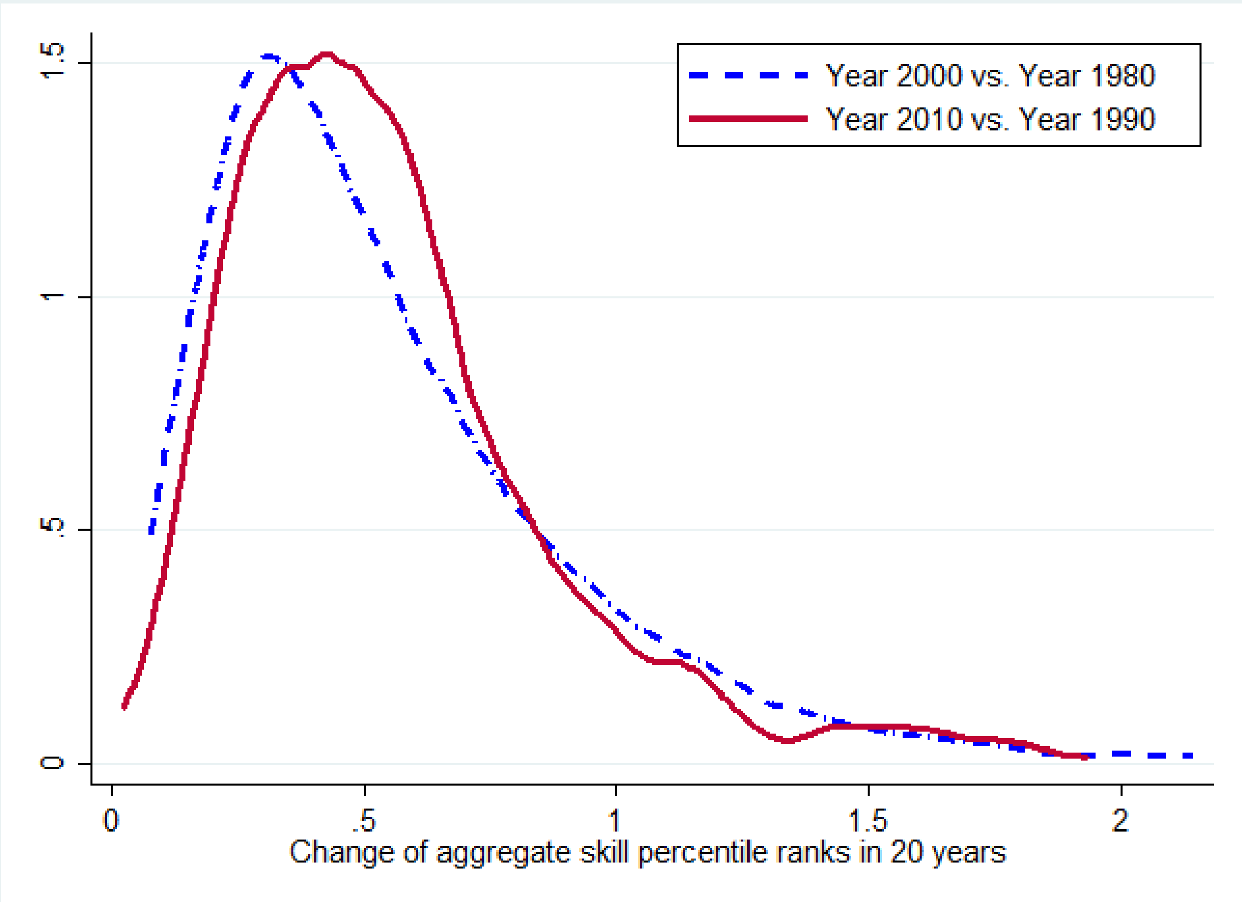
In characterizing the workplace changes associated with the computer revolution, some analysts have drawn a sharp distinction between standalone computing tasks (such as word processing or database analysis)and organization-related tasks (such as inventory control and supply-chain integration and Internet commerce), and have argued that innovations in the latter domain are the major source of SBTC. This reasoning suggests that the evolution of network technologies is at least as important as the development of PC technology. The first network of mainframe computers (the Advances Research Projects Agency Network [ARPANET]) was organized in 1970 and had expanded to about 1,000 host machines by 1984. In the mid-1980s, the National Science Foundation laid the backbone for the modern Internet by establishing the National Science Foundation Network (NSFNET). Commercial restrictions on the use of the Internet were lifted in 1991, and the first U.S. site on the World Wide Web was launched in December 1991. The use of the Internet grew very rapidly after the introduction of Netscape's Navigator program in 1994: the number of Internet hosts rose from about 1 million in 1992 to 20 million in 1997, and to 100 million in 2000.

Figure 1.24: True value and smoothed changes of percentile ranks between 1990 and 2010



Notes: These four figures depict the true values of changes in percentile ranks for the four major skill categories to show that the smoothed changes accurately pick up the trend instead of showing some extreme values.

Figure 1.25: Aggregate changes of percentile ranks in 20 years



Notes: This figure shows the distribution of the aggregate change in percentile rankings of skills from 1980 to 2000 and 1990 to 2010. In Section 5.3, I show the changes in different skills separately, and the aggregation is calculated by the commonly used distance measurement $\Delta PR_A = [\sum_{j=1}^{J} \Delta PR_{A,j}^2]^{1/2}$ for J different skill groups.

1.9.8.2 Technology Revolution, Skill Changes and Wages

In this part, I'd like to use one indicator of technological changes. the computer usage rate change from 1984 to 1993, to directly test the relationship of technological changes to the skill percentile ranks changes. Again, this indicator cannot present all the technological changes happening the labor market, for example, the automation, electric device, biological breakthroughs etc. And it also doesn't show the causal relationship, since it's likely that some occupations due to varies reasons

Table 1.25: Change of computer usage rate and skill distribution

Major Group	**Detailed Group**	$\Delta\text{Computer}_{1984-1993}$	$\Delta\text{Computer}_{1984-2001}$
Abstract	Verbal	0.0644	-0.0081
	Analytical	-0.0464	0.0562
	Quantitative	0.0445	-0.1323
Routine	Routine cognitive	-0.1034	-0.0321
	Routine manual	0.0322	0.0474
Manual	Manual	-0.0661	-0.1268
Physical	Strength	0.0933	0.0686
	Flexibility	0.2900	0.0397
	Auditory	-0.089	-0.2061
	Visual	0.0269	-0.1093

adapt computers earlier than other jobs. The computer usage rate comes from CPS supplements [46] the measurement is "Do you directly use computer at work?". The detailed usage includes emails, word precessing etc. Hence here, in Table 3, I only show the correlation of the change of computer usage rate with the change of each skill percentile ranks in the corresponding year. The change of computer usage rate from 1984 to 1993 corresponds to change from 1980 to 1990, and that of 1984 to 2001 corresponds to change of skills from 1980 to 2000.

Again, I'd like to emphasize that the minor changes in DOT 1991 compared to its previous version, and also some abilities like body flexibility and strength and manual have little to do with computer usage. So it's reasonable that the little correlation with those abilities and computer usage. But we can still see that the quantitative ability decreases with the increase of computer usage, and also the routine cognitive related skills, the visual and auditory abilities.

Now I'd like to look into the relationship of each skill level with the wage, i.e. the return rate of each skill. In Figure E.9, I depict the wage (in log) of each occupation with respect to the percentile of main ability demands. To to specific, I separate occupations by its main ability demands, which is defined as the biggest ability demand among the four categories, and regress the log wage in 1970

[46] Questions regarding Computer and Internet Use have been sporadically included in Current Population Survey supplements since 1984. The questions have primarily appeared as a portion of the October Education supplement, but were also included as a separate supplement in the December 1998, August 2000, and September 2001 samples.

Table 1.26: Marginal value (β) of each skill in 1970 and 2000

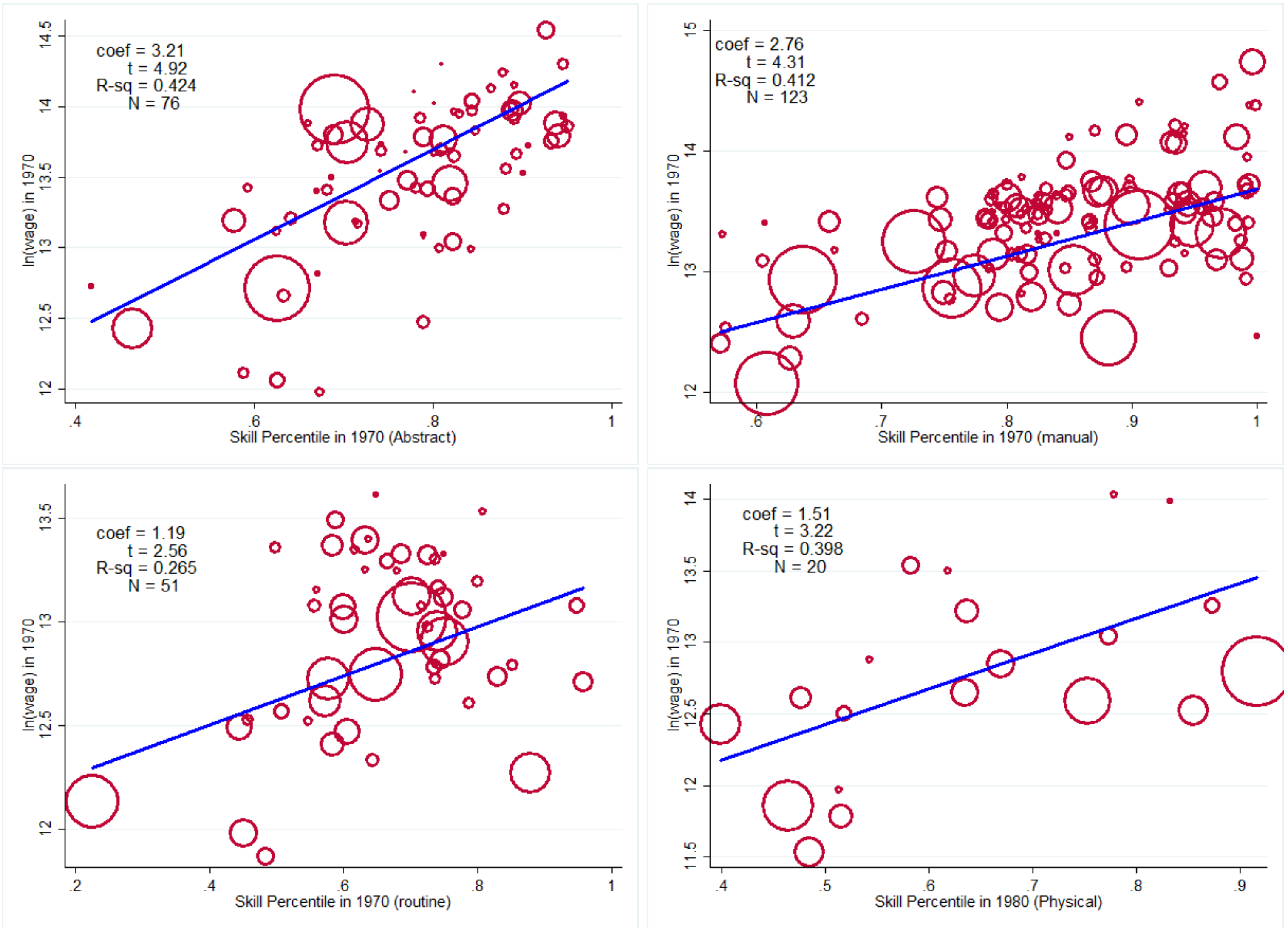
Year	Abstract	Manual	Routine	Physical
1970	3.21***	2.76***	1.19**	1.51***
	(4.92)	(4.31)	(2.56)	(3.22)
R^2	0.424	0.412	0.265	0.398
#	76	123	51	20
2000	2.42***	1.24***	-0.52	1.79***
	(7.40)	3.33	(-0.51)	(3.55)
R^2	0.518	0.254	0.021	0.224
#	153	150	48	68
	(+101.31%)	(+21.95%)	(-4%)	(+240%)
Example	Chief manager, scientists	Chef, technician	Clerk, operator	Dancer, police officer

against the percentile of that main ability demand. For example, if for occupation A, the demand for the abstract ability is bigger than the other three categories, this occupation will be classified as "Abstract job", and qualifies as an observation in the regression of log wage with respect to abstract percentile of occupations. The regression equation is:

$$log(Wage)_{i,j}^{Year} = \alpha_j^{Year} + \beta_j^{Year} PR_{i,j}^{Year} + \varepsilon_{i,j}$$

in which i represents occupation, and j represent the skill type among *Abstract*, *Routine*, *Manual*, *Physics*. And β_j can be interpreted as the marginal value of the comparative advantage of each skill. The regression results for 1970 and 2000 are in Table E.19.

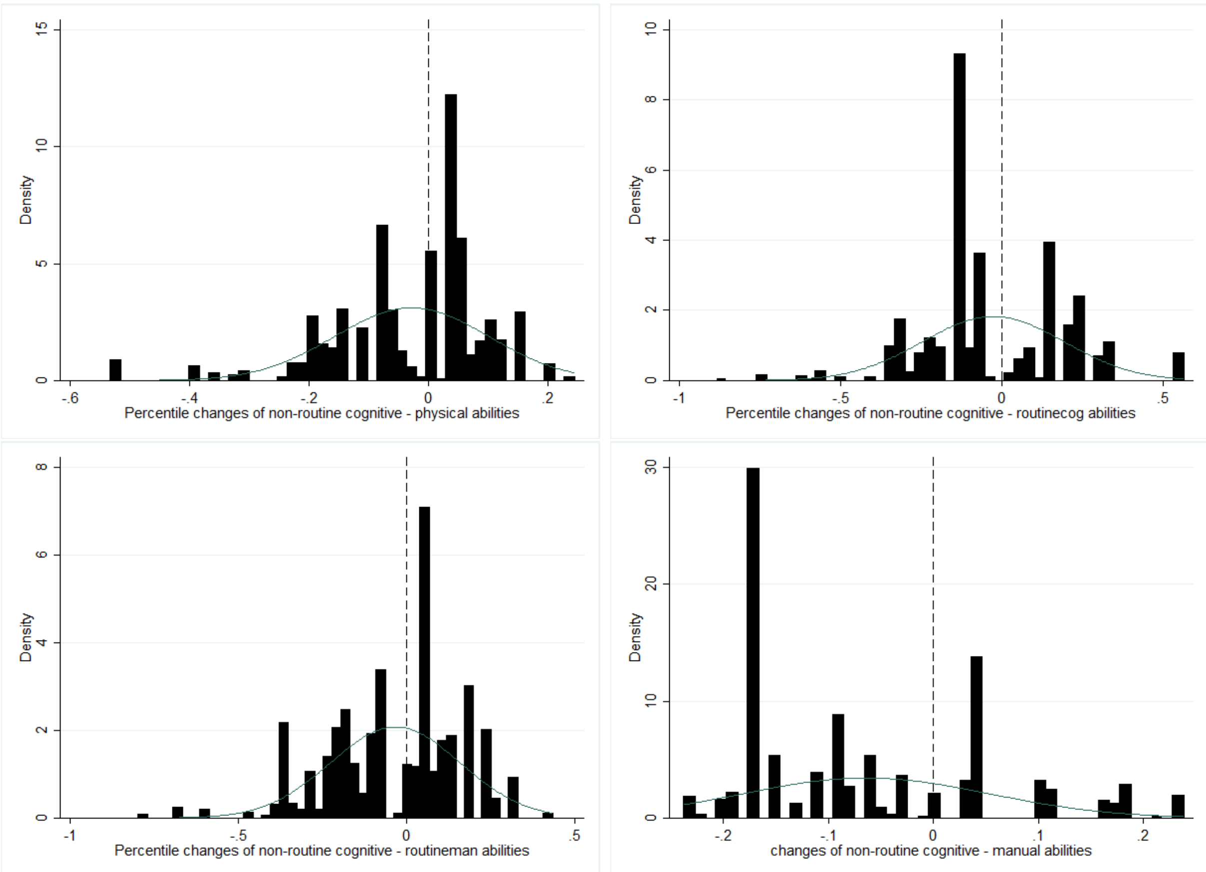
Figure 1.26: Skill percentiles and log(wage) in 1970



From the results, we can see that no matter in which time period, the "Abstract" ability is the most valued, and compared with 1970, the number of jobs that advantaged in the abstract ability. "Manual" ability still significantly affects the wage, but the marginal value decreased more compared with 1970 than abstract ability. "Physical" ability is the only one increases in the marginal value, and the increase of number of jobs is the most. Unsurprisingly, the "Routine" ability is no longer significant.

1.9.8.3 Measurement of Skill changes

Figure 1.27: Relative Changes of percentile ranking of Abilities Scores



1.9.9 Allowing for measurement error in RD

Measurement error in the threshold

In a conventional sharp RD design, the econometricians observe the assignment variable X^*, eligibility/treatment $D^* = 1[X^* < 0]$ and outcome Y. Mathematically, the measurement error on the assignment variable and the threshold is equivalent. To see this, suppose the true eligibility assignment mechanism is $D^* = 1[W^* < c^*]$, where W^* is the actual family income and c^* the eligibility threshold. Suppose the econometricians only see proxies $W = W^* + u$ and $c = c^* + v$. In this case, we can rewrite the eligibility assignment mechanism as $D^* = 1[X - \tilde{u} < 0]$ where $\tilde{u} \equiv u - v$, hence we have a model isomorphic to the original one.

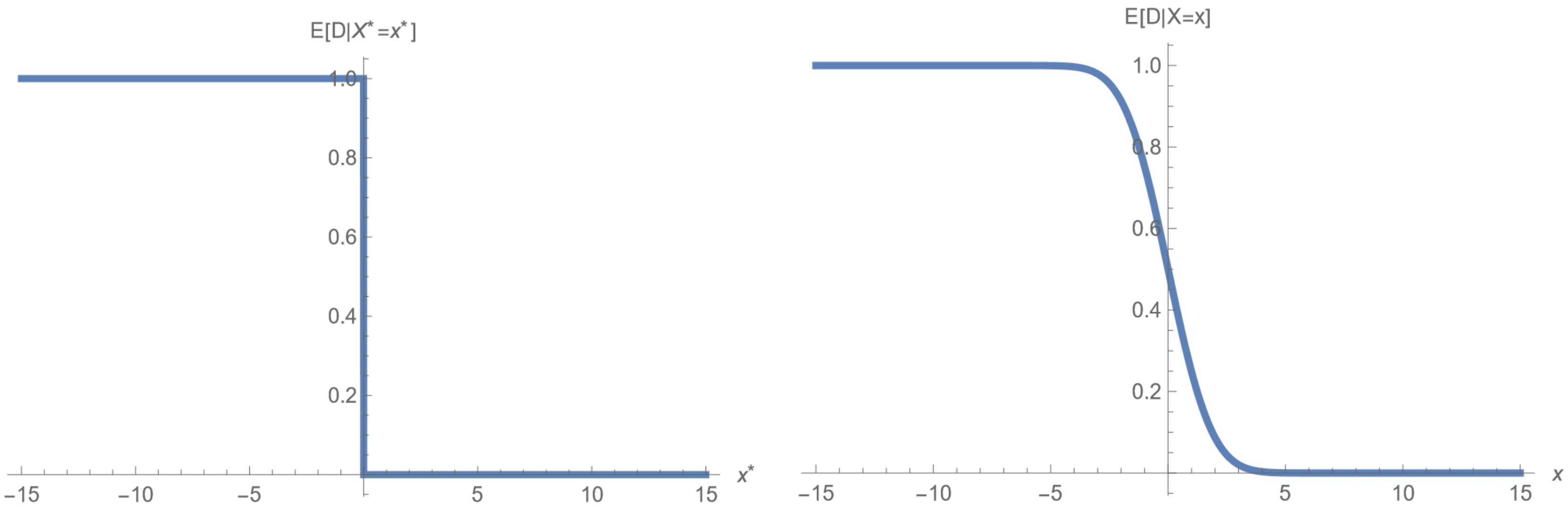
Classical measurement error in Sharp RD

As seen in Appendix Figure F.11, the so-called first stage relationship with the noisy X, $E[D^*|X]$, is smooth at $X = 0$, and we can no longer rely on the discontinuity in that relationship to identify a treatment effect. In a way, this is an extreme form of the attenuation bias, which is typically associated with measurement error. Because of this lack of first stage discontinuity, we cannot treat the problem as a fuzzy design either. In fact, the same force that smooths out the first stage relationship also smooths out the outcome function $E[Y|X]$, and the fuzzy RD simply becomes undefined.

As we can see, this is in contrast to the case where a first stage discontinuity exists despite the presence of measurement error in the assignment variable as in our case. [47]

[47] See Card et al. (2015) for an analogous result in the regression kink design.

Figure 1.28: Theoretical Effect of Classic Measurement Error in the Assignment Variable



Notes: The left panel plots the true first stage relationship $E[D|X^* = x^*]$ in a sharp RD. The right panel plots the observed first stage relationship $E[D|X = x]$. The right panel is generated by assuming that X^* and u are both normally distributed.

[AB17] N. Ashraf and O. Bandiera. *Social Incentives in Organizations.* In preparation for the Annual Review of Economics, 2017.

[ABL16] N. Ashraf, O. Bandiera, and S. Lee. “Do-gooders and Go-getters: Career incentives, Selection, and Performance in Public Service Delivery.” Working paper, STICERD, 2016.

[AGR15] D. Acemoglu, C. Garcia-Jimeno, and J. A. Robinson. “State Capacity and Economic Development: A Network Approach.” *American Economic Review*, **105**:2364–2409, 2015.

[AJR01] D. Acemoglu, S. Johnson, and J. A. Robinson. “The Colonial Origins of Comparative Development: An Empirical Investigation.” *American Economic Review*, **91**:1369–1401, 2001.

[AT97] P. Aghion and J. Tirole. “Formal and Real Authority in Organizations.” *Journal of Political Economy*, **105**:1–29, 1997.

[Ban08] M. Banton. “Administering the Empire.” pp. 1801–1968, 2008.

[BBR09] O. Bandiera, I. Barankay, and I. Rasul. “Social Connections and Incentives in the Workplace: Evidence From Personnel Data.” *Econometrica*, **77**(1047-1094):417–458, 2009.

[BCD12] A. Banerjee, R. Chattopadhyay, E. Duflo, D. Keniston, and N. Singh. “Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training.” Technical report, NBER Working Papers 17912, NBER, 2012.

[BEF17] Dal Bo, F. Finan E., O. Folke, T. Persson, and J. Rickne. “Who becomes a politician?” *e Quarterly Journal of Economics*, 2017. Forthcoming.

[Ber09] M. Bertrand. “CEOs.” *Annual Review of Economics*, **1**:121–150, 2009.

[BER13] Dal Bo, F. Finan E., and M. A. Rossi. “Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service.” *The Quarterly Journal of Economics*, **128**:1169–1218, 2013.

[BP09] T. Besley and T. Persson. “The Origins of State Capacity: Property Rights, Taxation, and Politics.” *American Economic Review*, **99**(1218-44):1–34, 2009.

[Bre16] E. Breza. *Field Experiments, Social Networks, and Development.* Handbook on the Economics of Networks, Oxford, 2016.

[BS03] M. Bertrand and A. Schoar. “Managing with Style: The Effect of Managers on Firm Policies.” *The Quarterly Journal of Economics*, **118**:1169–1208, 2003.

[CH08] V. Chernozhukov and C. Hansen. “The Reduced Form: A Simple Approach to Inference with Weak Instruments.” *Economics Letters*, **100**:68–71, 2008.

[Des17] E. Deserranno. “Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda.” mimeo, 2017.

[DH86] L. Davis and R. Huttenback. *Mammon and the Pursuit of Empire: The Political Economy of British Imperialism.* 1860-1912, Cambridge University Press, 1986.

[DJT99] M. Dewatripont, I. Jewitt, and J. Tirole. “The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies.” *Review of Economic Studies*, **66**:199–217, 1999.

[DP12] M. Dincecco and M. Prado. “Warfare, Fiscal capacity, and Performance.” *Journal of Economic Growth*, **17**:171–203, 2012.

[Fis01] R. Fisman. “Estimating the Value of Political Connections.” *American Economic Review*, **91**:1095–1102, 2001.

[FOP15] F. Finan, B. A. Olken, and R. Pande. “The Personnel Economics of the State.” Working Paper 21825, NBER, 2015.

[Fra92] M. Francis. “Governors and Settlers: Images of Authority in the British Colonies.” pp. 1820–60, 1992.

[FW04] R. Fisman and S. j. Wei. “Tax Rates and Tax Evasion: Evidence from "Missing Imports" in China.” *Journal of Political Economy*, **112**:471–500, 2004.

[Gar12] L. Gardner. *Taxing Colonial Africa: The Political Economy of British Imperialism.* Oxford University Press, 2012.

[GR12] R. Gibbons and J. Roberts. “The Handbook of Organizational Economics.” In *Princeton University Press.* 1 ed, 2012.

[HHM12] B. Henn, L. Hon, M. Macpherson, N. Eriksson, S. Saxonov, I. Peter, and J. Mountain. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples.” *PLoS One*, **7**, 2012.

[IM12] L. Iyer and A. Mani. “Traveling Agents: Political Change and Bureaucratic Turnover in India.” *The Review of Economics and Statistics*, **94**:723–739, 2012.

[Jef38] C. Jeffries. *The Colonial Empire and its Civil Service.* Cambridge University Press, 1938.

[Jia17] R. Jia. “Pollution for Promotion.” mimeo, 2017.

[JKS15] R. Jia, M. Kudamatsu, and D. Seim. “Political Selection in China: The Complementary Roles of Connections and Performance.” *Journal of the European Economic Association*, **13**:631–668, 2015.

[JO05] B. F. Jones and B. A. Olken. “Do Leaders Matter? National Leadership and Growth Since World War II.” *The Quarterly Journal of Economics*, **120**:835–864, 2005.

[Kir00] A. Kirk-Greene. “Britain’s Imperial Administrators.” pp. 1858–1966, 2000.

[KKO15] A. Q. Khan, A. I. Khwaja, and B. A. Olken. *Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors.* The Quarterly Journal of Economics, 2015.

[KKO16] A. Khan, A. Khwaja, and B. Olken. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." mimeo, 2016.

[KT13] F. Kramarz and D. Thesmar. "Social Networks in the Boardroom." *Journal of the European Economic Association*, **11**:780–807, 2013.

[Lai05] Z. Laidlaw. *Colonial Connections 1815-1845: Patronage.* the Information Revolution and Colonial Government, Studies in Imperialism, Manchester University Press, 2005.

[LO12] E. P. Lazear and P. Oyer. "Personnel Economics." In *The Handbook of Organizational Economics.* Princeton University Press, 2012.

[MC99] S. Morris and S. Coate. "Policy Persistence." *American Economic Review*, **89**:1327–1336, 1999.

[NT54] S. Northcote and C. Trevelyan. *Report on the Organization of the Permanent Civil Service.* HMS, 1854.

[NW89] D. North and B. Weingast. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice." *The Journal of Economic History*, **49**:803–832, 1989.

[PT96] C. Prendergast and R. H. Topel. "Favoritism in Organizations." *Journal of Political Economy*, **104**:958–78, 1996.

[PZ16] P. Persson and E. Zhuravskaya. "The Limits of Career Concerns in Federalism: Evidence from China." *Journal of the European Economic Association*, **14**:338–374, 2016.

[RAJ05] J. A. Robinson, D. Acemoglu, and S. Johnson. "Institutions as a Fundamental Cause of Long-Run Growth." *Handbook of Economic Growth*, **1**:386–472, 2005.

[RR17] I. Rasul and D. Rogger. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal*, pp. 1–34, 2017.

[SW16] E. Sanderson and F. Windmeijer. "A weak instrument F-test in linear IV models with multiple endogenous variables." *Journal of Econometrics*, **190**:212–221, 2016.

CHAPTER 2

Estimation and Inference of Semiparametric Models using Data from Several Sources

2.1 Introduction

There are many cases in empirical micro studies where data needed to analyze a particular phenomenon is not always available in one data set. Typically, this hampers the possibility of meaningful empirical research. In fact, a common phenomenon is to make some simplifying assumptions, which would then permit the researcher to use information from more than one data source. For example, as Blundell et al. (2006) note, this is a crucial difficulty when faced by those studying households' consumption and saving behavior because of the lack of panel data on both household expenditures, income, and saving.

An important data for studying consumption is, for example, the Panel Study of Income Dynamics (PSID), a survey that provides longitudinal annual data for household that have been followed since 1968. The PSID collects data on a subset of consumption items, namely food at home and food away from home (with few gaps in some of the survey years) and income. However, the PSID is that it does not provide data on wealth. In contrast, there are few data set that provide detailed data on income and wealth (e.g. Health and Retirement Survey (HRS), or the National Longitudinal Study (NLS)), these data sets provide no information on consumption.

Problems of similar nature exist in many other countries. For example, in the UK, the Family Expenditure Survey (FES) provides comprehensive data on household expenditures, but this is across-sectional data and thus the researcher does not get to observe households over time. In contrast, the British Household Panel Survey (BHPS) is a Panel data set that collects data on income or wealth, but it provides no information on consumption.[1] This is quite puzzling, given the vital need to study the consumption decisions jointly with the income and wealth processes.

[1]Similar problems exist for many other countries, in particular countries in Europe (e.g. France and Spain) that collect detailed data on both consumption, income, and wealth, but the information never exists in a single data set.

Consequently, as is clearly and comprehensively explain in Blundell et al. (2006), studies the aimed at understanding consumption behavior, and testing alternative theories, have resorted to the limited data on food expenditure provided in the PSID. This includes, among others, Hall and Mishkin (1982), Zeldes (1989), Runkle (1991), and Shea (1995), Cochrane (1991), Hayashi, Altonji and Kotlikoff (1996), Cox, Ng and Waldkirch (2004), Martin (2003) and Hurst and Stafford (2004) for tests of many alternative theories. The main problem with all these studies is that they use consumption on limited number of goods (largely necessity goods), and thus putting into question the external validity of the results.

One way that has been used in the literature is to form synthetic panel data sets from repeated cross-section data sets in which consumption is reported (e.g. the CEX or the FES). This is done in, for example, Browning, Deaton and Irish (1985) and Attanasio and Weber (1993).

An alternative empirical approach that have been used occasionally in the literature involve imputation of consumption to the PSID households using information on consumption from the CEX. Specifically, Skinner (1987) proposes to impute total consumption in the PSID using the estimated coefficients of a regression of total consumption on a number of consumption items that are reported in both the PSID and the CEX. While this method seems appealing at first sight, it reduces any variation in total consumption, since it does not take into account the fact the there is considerable idiosyncratic elements that goes into the individual decision making. Ziliak (1998) and Browning and Leth-Petersen (2003) provided alternative method that are variants of that proposed by Skinner (1987).

However, this method has a major weakness, in that it ignored, by construction, the dynamics of the individuals' consumption. Avoiding direct control of Individual's heterogeneity has been shown to provide major obstacle when modeling individual's behavior in general.

The one paper in the literature that provides a method that is related in spirit to the method proposed here is the paper by Blundell et al. (2006).[2] The method is similar in nature to that of Skinner (1987), in that the authors impute consumption data for the households in the PSID using regression parameters estimated from the CEX data. The key difference, is that the authors in Blundell et al. (2006) is that they use "structural" regression of a standard demand function for food that depends not only on other consumption items, but also depend on prices and a set of demographic and socio-economic variable of the household. Assuming monotonicity of the demand for makes it possible to invert these function in order to obtain a structurally based formula

[2]This method is used extensively in Blundell, Pistaferri and Preston (2008).

non-durable consumption, which exists in the CEX, but is missing in the PSID. Nevertheless, the general problem with such an imputation method described above for Skinner's method still apply. If nothing else, the consumption data imputed in this fashion is likely to suffer from the well-known error-in-variable problem. Most importantly, it ignores the inherent individual heterogeneity of consumption.

In two recent papers Fan, Sherman and Shum (2014a, 2014b) address a special case of the problem addressed in our paper, namely the case of treatment effect. Under this scenario, the outcome variables and conditioning variable are observed in two separate data sets, so that the treatment effect parameters are not point-identidfied. The authors provide sharp bounds on the counterfactual distributions and parameter of interest (see Fan, Sherman and Shum (2014b)), and the corresponding inference (see Fan, Sherman and Shum (2014a).

Our case is more general and encompass a more general situation in which some of the variables are available only in one data set, while others are available in a separate data sets. The key insight is that there are some variables that appear in both. Under relatively mild regularity conditions we provide a method that allow one to point-identified the structural parameters of interest in the main data of interest using the information provided in the other data set. The parameters of interest and the "imputation equation" are estimated simultaneously. We also provide the necessary theory for inference including cases in which the number of observation in "imputation" data set do not diverge to infinity at the same rate as that for the main data of interest.

Notation of this paper is standard. For any square matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A respectively. Throughout this paper, we use C to denote a generic finite positive constant which is larger than 1. For any set of real vectors $\{a_l\}_{l\in I}$ where $I = \{l_1, \ldots, l_{d_I}\}$ is a index set with d_I distinct natural numbers, we define $(a_l)_{l\in I} = (a_{l_1}, \ldots, a_{l_{d_I}})$ and $(a_l)'_{l\in I} = (a_{l_1}, \ldots, a_{l_{d_I}})'$. The notation $\|\cdot\|$ denotes the Euclidean norm. A' refers to the transpose of any matrix A. I_k and $\mathbf{0}_l$ are used to denote $k \times k$ identity matrix and $l \times l$ zero matrices respectively. The symbolism $A \equiv B$ means that A is defined as B. the expression $a_n = o_p(b_n)$ signifies that $\Pr(|a_n/b_n| \geq \epsilon) \to 0$ for all $\epsilon > 0$ as n go to infinity; and $a_n = O_p(b_n)$ when $\Pr(|a_n/b_n| \geq M) \to 0$ as n and M go to infinity. As usual, "$\to_p$" and "$\to_d$" imply convergence in probability and convergence in distribution, respectively.

2.2 The Model and the Estimators

We are interested in estimating the following model

$$Y = g(X_1, X_2, \theta_0) + v \tag{2.1}$$

where X_1 and X_2 are sets of regressors, $g(\cdot,\cdot,\cdot) : R^{d_{x_1}} \times R^{d_{x_2}} \times R^{d_\theta} \to R$ is a known function, θ_0 is the unknown parameter of interest and v is a unobservable residual term. Two data sets are available to estimate the unknown parameter: $\{(Y_i, X'_{2,i}, X'_{3,i})\}_{i\in I_1}$ and $\{(X_{1,i}, X'_{2,i}, X'_{3,i})\}_{i\in I_2}$, where I_1 and I_2 are two index sets with cardinalities n_1 and n_2 respectively.

The unknown parameter θ_0 could be conveniently estimated under the conditional moment restriction $E[v|X_1, X_2] = 0$, if we had the joint observations on (Y, X'_1, X'_2). However, such straightforward method is not applicable here because Y and X_1 are contained in different data sets. On the other hand, the common variables X_2 and X_3 contained in both data sets can be useful for identifying and estimating the unknown parameter θ_0. For this purpose, we assume that

$$E[v|X_2, X_3] = 0. \tag{2.2}$$

Using the expression in (2.1) and the conditional mean restriction in (2.2), we get

$$E[Y|\, X_2, X_3] = E[g(X_1, X_2, \theta_0)|\, X_2, X_3], \tag{2.3}$$

which is the key equation for the identification and estimation of θ_0.

For ease of notations, we write $X = (X'_1, X'_2)'$ and $Z = (X'_2, X'_3)'$. Then the model can be written as

$$Y = g(X, \theta_0) + v \text{ with } E[v|Z] = 0. \tag{2.4}$$

For any θ, we define the conditional expectation of $g(X, \theta)$ given Z as

$$\phi(Z, \theta) = E[g(X, \theta)|\, Z].$$

Then we can write

$$Y = \phi(Z, \theta_0) + \varepsilon + v = h_0(Z) + u,$$

where $\varepsilon = g(X, \theta_0) - \phi(Z, \theta_0)$, $u \equiv \varepsilon + v$ and $h_0(Z) = \phi(Z, \theta_0)$.

As the conditioning variable Z is available in both data sets, we have n ($n = n_1 + n_2$) observations: $\{Z_i\}_{i\in I} = \{(X'_{2,i}, X'_{3,i})\}_{i\in I}$ where $I = I_1 \cup I_2$. Let $P_k(z) = [p_1(z), \ldots, p_k(z)]'$ be a k-dimensional vector of basis functions for any positive integer k. For any k and any n, we define $P_{n,k} = (P_k(Z_i))_{i\in I}$ which is an $n \times k$ matrix. Accordingly, we define $P_{n_1,k_1} = (P_{k_1}(Z_i))_{i\in I_1}$ and $P_{n_2,k_2} = (P_{k_2}(Z_i))_{i\in I_2}$ which are $n_1 \times k_1$ and $n_2 \times k_2$ matrices respectively.

The conditional mean function $h_0(Z) = E[Y|Z]$ can be estimated using the first data set by

$$\widehat{h}_{n_1}(z) = P_{k_1}(z)'(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n_1,k_1}Y_{n_1} \tag{2.5}$$

where $Y_{n_1} = (Y_i)'_{i\in I_1}$. Using the second data set, we get the following estimator of the conditional mean function $\phi(Z,\theta)$ for any θ:

$$\widehat{\phi}_{n_2}(Z,\theta) = P_{k_2}(Z)'(P'_{n_2,k_2}P_{n_2,k_2})^{-1}P'_{n_2,k_2}g_{n_2}(\theta) \tag{2.6}$$

where $g_{n_2}(\theta) = (g(X_i,\theta))'_{i\in I_2}$. Using the estimators of $h_0(Z)$ and $\phi(Z,\theta)$, we can construct the estimator of θ_0 via the minimum distance (MD) estimation:

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} n^{-1}\sum_{i\in I}\left[\widehat{w}_n(Z_i)\left|\widehat{h}_{n_1}(Z_i) - \widehat{\phi}_{n_2}(Z_i,\theta)\right|^2\right] \tag{2.7}$$

where Θ denotes the parameter space containing θ_0, and $\widehat{w}_n(\cdot)$ is a non-negative weight function. One simple and straightforward choice of the weight function $\widehat{w}_n(\cdot)$ is the identity function, i.e., $\widehat{w}_n(z) = 1$ for any z. However, as we will show later in this paper, the identity weighted MD estimator may not have the smallest possible variance.

2.3 Asymptotic Properties of the MD Estimator

In this section, we establish the asymptotic properties of the MD estimator. For any positive integer k, Let $\xi_k = \sup_{z\in\mathcal{Z}}\|P_k(z)\|$ and $Q_k = E[P_k(Z)P'_k(Z)]$, where $\mathcal{Z}$ denotes the support of Z. We first state the sufficient conditions for consistency.

Assumption 2.3.1 *(i)* $\{(Y_i,Z_i)\}_{i\in I_1}$ *and* $\{(X_i,Z_i)\}_{i\in I_2}$ *are independent with i.i.d. observations; (ii)* $E[Y|Z] < C$*; (iii)* $C^{-1} \le \lambda_{\min}(Q_k) \le \lambda_{\max}(Q_k) \le C$ *for all* k*; (iv) there exist* $\beta_{h,k} \in R^k$ *and* $r_h > 0$ *such that*

$$\sup_{z\in\mathcal{Z}}\left|h_0(z) - P_k(z)'\beta_{h,k}\right| = \sup_{z\in\mathcal{Z}}|h_0(z) - h_{0,k}(z)| = O(k^{-r_h}); \tag{2.8}$$

(v) $\max_{j=1,2}\xi^2_{k_j}\log(k_j)n_j^{-1} = o(1)$ *and* $k_1n_1^{-1} + k_1^{-1} = o(1)$.

Assumption 2.3.1 includes mild and standard conditions on nonparametric series estimation of conditional mean function (see, e.g. Andrews (1991) and Newey (1997)).

Define

$$L_n(\theta) = n^{-1}\sum_{i\in I}w_n(Z_i)\left|h_0(Z_i) - \phi(Z_i,\theta)\right|^2 \text{ and } L_n^*(\theta) = E\left[w_n(Z)\left|h_0(Z) - \phi(Z,\theta)\right|^2\right]$$

for any $\theta\in\Theta$, where $w_n(\cdot)$ is defined in Assumption 2.3.2(v) below.

Assumption 2.3.2 *(i)* $\sup_{\theta\in\Theta} E\left[\phi^2(Z,\theta)\right] < C$; *(ii)* $n^{-1}\sum_{i\in I}|\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta)|^2 = o_p(1)$ *uniformly over* θ; *(iii) for any* $\varepsilon > 0$, *there is* $\eta_\varepsilon > 0$ *such that*

$$E\left[|h_0(Z) - \phi(Z,\theta)|^2\right] > \eta_\varepsilon \text{ for any } \theta \in \Theta \text{ with } ||\theta - \theta_0|| \geq \varepsilon;$$

(iv) $\sup_{\theta\in\Theta}|L_n(\theta) - L_n^*(\theta)| = o_p(1)$; *(v)* $\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z) - w_n(z)| = O_p(\delta_{w,n})$ *where* $\delta_{w,n} = O(n_1^{-1/4} + n_2^{-1/4})$ *and* $w_n(\cdot)$ *is a sequence of non-random functions with* $C^{-1} \leq w_n(z) \leq C$ *for any* n *and any* $z \in \mathcal{Z}$.

Assumption 2.3.2(i) imposes uniform finite second moment condition on the function $\phi(Z,\theta)$. Assumption 2.3.2(ii) requires that the nonparametric estimator $\widehat{\phi}_{n_2}(Z_i,\theta)$ of $\phi(Z_i,\theta)$ is consistent under the empirical L_2-norm uniformly over $\theta \in \Theta$. Assumption 2.3.2(iii) is the identification condition of θ_0. Assumption 2.3.2(iv) is a uniform law of large numbers of the function $w(Z_i)\,|h_0(Z_i) - \phi(Z_i,\theta)|^2$ indexed by θ. Assumption 2.3.2(v) requires that $\widehat{w}_n(\cdot)$ is approximated by a sequence of nonrandom function $w_n(\cdot)$ uniformly over z. For the consistency of the MD estimator, it is sufficient to have $\delta_{w,n} = o(1)$ in Assumption 2.3.2(v). The rate condition $\delta_{w,n} = O(n_1^{-1/4} + n_2^{-1/4})$ is needed for deriving the asymptotic normality of the MD estimator. It is clear that Assumption 2.3.2(v) holds trivially if $\widehat{w}_n(\cdot)$ is the identity function.

Theorem 2.3.1 *Under Assumptions 2.3.1 and 2.3.2, we have* $\widehat{\theta}_n = \theta_0 + o_p(1)$.

For ease of notations, we define

$$\begin{aligned} g_\theta(X,\theta) &= \tfrac{\partial g(X,\theta)}{\partial\theta}, & g_{\theta\theta}(X,\theta) &= \tfrac{\partial^2 g(X,\theta)}{\partial\theta\partial\theta'}, \\ \phi_\theta(Z,\theta) &= E\left[g_\theta(X,\theta)|\,Z\right], & \phi_{\theta\theta}(Z,\theta) &= E\left[g_{\theta\theta}(X,\theta)|\,Z\right], \\ \widehat{\phi}_{\theta,n_2}(Z,\theta) &= \tfrac{\partial\widehat{\phi}_{n_2}(Z,\theta)}{\partial\theta}, & \widehat{\phi}_{\theta\theta,n_2}(Z,\theta) &= \tfrac{\partial^2\widehat{\phi}_{n_2}(Z,\theta)}{\partial\theta\partial\theta'}. \end{aligned}$$

By the consistency of $\widehat{\theta}_n$, there exists a positive sequence $\delta_n = o(1)$ such that $\widehat{\theta}_n \in \mathcal{N}_{\delta_n}$ with probability approaching 1, where $\mathcal{N}_{\delta_n} = \{\theta\in\Theta : ||\theta - \theta_0|| \leq \delta_n\}$. Define $H_{0,n} = E\left[w_n(Z)\phi_\theta(Z,\theta_0)\phi_\theta'(Z,\theta_0)\right]$. Let $\phi_{\theta_j}(z,\theta)$ denote the j-th component of $\phi_\theta(Z,\theta)$.

We next state the sufficient conditions for asymptotic normality of $\widehat{\theta}_n$.

Assumption 2.3.3 *The following conditions hold:*

(i) $\sup_{\theta\in\mathcal{N}_n} n_2^{-1}\sum_{i\in I_2}||g_{\theta\theta}(X_i,\theta)||^2 = O_p(1)$;

(ii) $\lambda_{\min}(H_{0,n}) > C^{-1}$;

(iii) $n^{-1}\sum_{i\in I}||\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0)||^2 = o_p(n_2^{-1/2})$;

(iv) $E\left[\|\phi_\theta(Z,\theta_0)\|^4\right] < \infty$;

(v) $E\left[u^2 \middle| Z\right] > C^{-1}$, $E\left[\varepsilon^2 \middle| Z\right] > C^{-1}$ *and* $E\left[u^4 + \varepsilon^4 \middle| Z\right] < C$;

(vi) $\sup_{z\in\mathcal{Z}} |w_n(z)\phi_{\theta_j}(z,\theta_0) - P_k'(z)\beta_{w\phi_j,n,k}| = o(1)$ *where* $\beta_{w\phi_j,n,k} \in R^k$ $(j = 1,\ldots,d_\theta)$;

(vii) $\max_{j=1,2}(k_j n_j^{-1/2} + k_j^{-r_h} n_j^{1/2}) = o(1)$.

Assumptions 2.3.3(i) holds when $\|g_{\theta\theta}(x,\theta)\|^2 < C$ for any x and any θ in the local neighborhood of θ_0. The lower bound of the eigenvalue of $H_{0,n}$ in Assumptions 2.3.3(ii) ensures the local identification of θ_0. Assumptions 2.3.3(iii) requires that the convergence rate of $\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)$ under the empirical L_2-norm is faster than $n_2^{-1/4}$. Assumptions 2.3.3(iv) imposes finite second moment on the derivative function $\phi_\theta(Z,\theta_0)$. Assumption 2.3.3(v) imposes moment conditions on the projection errors u and ε which are useful for deriving the asymptotic normality of the MD estimator. Assumption 2.3.3(vi) requires that the function $w_n(z)\phi_{\theta_j}(z,\theta_0)$ can be approximated by the basis functions. Assumption 2.3.3(vii) imposes restrictions on the number of basis functions and the smoothness of the unknown function h_0.

Let $\sigma_u^2(Z) = E\left[u^2 \middle| Z\right]$, $\sigma_\varepsilon^2(Z) = E\left[\varepsilon^2 \middle| Z\right]$ and $\phi_{w\theta,n} = (w_n(Z_i)\phi_\theta(Z_i,\theta))_{i\in I}$. Define

$$\Sigma_{n_1} \equiv \frac{\phi_{w\theta,n} P_{n,k_1} Q_{n_1,k_1}^{-1} Q_{n_1,u} Q_{n_1,k_1}^{-1} P_{n,k_1}' \phi_{w\theta,n}'}{n^2 n_1}$$

where $Q_{n_1,u} = n_1^{-1}\sum_{i\in I_1}\sigma_u^2(Z_i)P_{k_1}(Z_i)P_{k_1}'(Z_i)$, and

$$\Sigma_{n_2} \equiv \frac{\phi_{w\theta,n} P_{n,k_2} Q_{n_2,k_2}^{-1} Q_{n_2,\varepsilon} Q_{n_2,k_2}^{-1} P_{n,k_2}' \phi_{w\theta,n}'}{n^2 n_2}$$

where $Q_{n_2,\varepsilon} = n_2^{-1}\sum_{i\in I_2}\sigma_\varepsilon^2(Z_i)P_{k_2}(Z_i)P_{k_2}'(Z_i)$.

Theorem 2.3.2 *Under Assumptions 2.3.1, 2.3.2 and 2.3.3, we have*

$$\widehat{\theta}_n - \theta_0 = O_p(n_1^{-1/2} + n_2^{-1/2}) \tag{2.9}$$

and moreover

$$\gamma_n'(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})^{1/2}(\widehat{\theta}_n - \theta_0) \to_d N(0,1) \tag{2.10}$$

for any non-random sequence $\gamma_n \in R^{d_\theta}$ *with* $\gamma_n'\gamma_n = 1$.

Remark 3.1. The first result of Theorem 2.3.2, i.e., (2.9), implies that the convergence rate of the MD estimator is of the order $\max\{n_1^{-1/2}, n_2^{-1/2}\}$.

Remark 3.2. By the Cramer-Wold device and Theorem 2.3.2, we know that

$$(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})^{1/2}(\widehat{\theta}_n - \theta_0) \to_d N(0_{d_\theta}, I_{d_\theta}), \tag{2.11}$$

which together with the continuous mapping theorem (CMT) implies that,

$$(\widehat{\theta}_n - \theta_0)'(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})(\widehat{\theta}_n - \theta_0) \to_d \chi^2(d_\theta). \tag{2.12}$$

Moreover, let ι_j^* be the $d_\theta \times 1$ selection vector whose j-th ($j = 1, \ldots, d_\theta$) component is 1 and rest components are 0. Define

$$\gamma_{j,n} = \frac{(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})^{-1/2}}{(\iota_j^{*\prime}(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})^{-1}\iota_j^*)^{1/2}}\iota_j^*, \text{ for } j = 1, \ldots, d_\theta.$$

It is clear that $\gamma_{j,n}'\gamma_{j,n} = 1$, and by Theorem 2.3.2, we have

$$\begin{aligned}&\gamma_{j,n}'(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})^{1/2}(\widehat{\theta}_n - \theta_0)\\ &= \frac{\widehat{\theta}_{j,n} - \theta_{j,0}}{(\iota_j^{*\prime}(H_{0,n}(\Sigma_{n_1} + \Sigma_{n_2})^{-1}H_{0,n})^{-1}\iota_j^*)^{1/2}} \to_d N(0,1)\end{aligned} \tag{2.13}$$

where $\widehat{\theta}_{j,n} = \iota_j^{*\prime}\widehat{\theta}_n$ and $\theta_{j,0} = \iota_j^{*\prime}\theta_0$. Results in (2.12) and (2.13) can be used to conduct inference on $\theta_{j,0}$ and θ_0 if the consistent estimators of $H_{0,n}$, Σ_{n_1} and Σ_{n_2} are available.

2.4 Semiparametric Nonlinear GLS Estimator

In this section, we study the properties of the semiparametric nonlinear GLS (SNGLS) estimator:

$$\widehat{\theta}_{ls} = \arg\min_{\theta\in\Theta} n_1^{-1}\sum_{i\in I_1}\left[\widehat{w}_n(Z_i)\left|Y_i - \widehat{\phi}_{n_2}(Z_i,\theta)\right|^2\right] \tag{2.14}$$

where Θ denotes the parameter space containing θ_0, and $\widehat{w}_n(\cdot)$ is a non-negative weight function.

Assumption 2.4.1 *(i)* $\{(Y_i, Z_i)\}_{i\in I_1}$ *and* $\{(X_i, Z_i)\}_{i\in I_2}$ *are independent with i.i.d. observations; (ii)* $E\left[u^4\middle|\, Z\right] < C$*; (iii)* $C^{-1} \le \lambda_{\min}(Q_k) \le \lambda_{\max}(Q_k) \le C$ *for all* k*; (iv)* $\sup_{\theta\in\Theta} E\left[\phi^4(Z,\theta)\right] < C$*; (v)* $n_1^{-1}\sum_{i\in I_1}|\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta)|^2 = o_p(1)$ *uniformly over* θ*; (vi) for any* $\varepsilon > 0$*, there is* $\eta_\varepsilon > 0$ *such that*

$$E\left[|\phi(Z,\theta_0) - \phi(Z,\theta)|^2\right] > \eta_\varepsilon \text{ for any } \theta \in \Theta \text{ with } ||\theta - \theta_0|| \ge \varepsilon;$$

(vii) $\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z) - w_n(z)| = o_p(1)$ *where* $w_n(\cdot)$ *is a sequence of non-random functions with* $C^{-1} \le w_n(z) \le C$ *for any* n *and any* $z \in \mathcal{Z}$*.*

Assumption 2.4.2 *The following conditions hold:*

(i) $n_1^{-1}\sum_{i\in I_1} u_i(\widehat{w}_n(Z_i) - w_n(Z_i))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) = o_p(n_1^{-1/2})$;

(ii) $n_1^{-1}\sum_{i\in I_1}(\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0))^2 = o_p(1)$;

(iii)

$$\begin{aligned} & n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(\phi_n(Z_i,\theta_0)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) \\ & = n_1^{-1}\phi_{w\theta,n_1}P_{n_2,k}(P'_{n_2,k}P_{n_2,k})^{-1}\sum_{i\in I_2}P_k(Z_i)\varepsilon_i + o_p(n_1^{-1/2}+n_2^{-1/2}); \end{aligned}$$

(iv) $\sup_{\theta\in\mathcal{N}_n} n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i\widehat{\phi}_{\theta\theta,n_2}(Z_i,\theta)=o_p(1)$.

Theorem 2.4.1 *Under Assumption 2.4.1, we have* $\widehat{\theta}_{ls}=\theta_0+o_p(1)$.

Lemma 2.4.1 *Under Assumptions 2.3.1 and 2.3.2, we have*

$$n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) = n_1^{-1}\sum_{i\in I_1}w_n(Z_i)u_i\phi_\theta(Z_i,\theta_0)+o_p(n_1^{-1/2}). \tag{2.15}$$

Lemma 2.4.2 *Under Assumptions 2.3.1 and 2.3.2, we have*

$$\begin{aligned} & n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(\phi(Z_i,\theta_0)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) \\ & = n_1^{-1}\phi_{w\theta,n_1}P_{n_2,k}(P'_{n_2,k}P_{n_2,k})^{-1}\sum_{i\in I_2}P_k(Z_i)\varepsilon_i + o_p(n_1^{-1/2}). \end{aligned} \tag{2.16}$$

Proof of Lemma 2.4.2. By definition, ■

Let $\sigma_u^2(Z)=E\left[u^2\middle| Z\right]$, $\sigma_\varepsilon^2(Z)=E\left[\varepsilon^2\middle| Z\right]$ and $\phi_{w\theta,n}=(w_n(Z_i)\phi_\theta(Z_i,\theta))_{i\in I}$. Define

$$\Sigma_{n_1}\equiv n_1^{-1}\sum_{i\in I_1}w_n^2(Z_i)\sigma_u^2(Z_i)\phi_\theta(Z_i,\theta_0)\phi_\theta(Z_i,\theta_0)'$$

and

$$\Sigma_{n_2}\equiv\frac{\phi_{w\theta,n_1}P_{n_1,k}Q_{n_2,k}^{-1}Q_{n_2,\varepsilon}Q_{n_2,k}^{-1}P'_{n_1,k}\phi'_{w\theta,n_1}}{n_1^2n_2}$$

where $Q_{n_2,\varepsilon}=n_2^{-1}\sum_{i\in I_2}\sigma_\varepsilon^2(Z_i)P_k(Z_i)P'_k(Z_i)$.

By the definition of $\widehat{\theta}_n$, we have the following first order condition

$$n_1^{-1}\sum_{i\in I_1}w_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_n))\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)=0. \tag{2.17}$$

Applying the first order expansion to (2.17), we get

$$
\begin{aligned}
0 &= n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&- n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)'(\widehat{\theta}_n-\theta_0)\\
&+ n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_n))\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_n)(\widehat{\theta}_n-\theta_0),
\end{aligned}
\tag{2.18}
$$

where $\widetilde{\theta}_n$ is between $\widehat{\theta}_n$ and θ_0 and it may differ across rows. We next show that

$$
\begin{aligned}
& n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&= n_1^{-1}\sum_{i\in I_1}w_n(Z_i)u_i\phi_\theta(Z_i,\theta_0)\\
&+ n_1^{-1}\phi_{w\theta,n_1}P_{n_2,k}(P_{n_2,k}'P_{n_2,k})^{-1}\sum_{i\in I_2}P_k(Z_i)\varepsilon_i\\
&+ o_p(n_1^{-1/2}+n_2^{-1/2}),
\end{aligned}
\tag{2.19}
$$

and

$$
n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)'\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)=H_{0,n}+o_p(1),
\tag{2.20}
$$

and

$$
n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_n))\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_n)=o_p(1).
\tag{2.21}
$$

By definition

$$
\begin{aligned}
& n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&= n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&+ n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(\phi(Z_i,\theta_0)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0),
\end{aligned}
\tag{2.22}
$$

which together with Assumption 2.4.2(iii) and Lemma 2.4.1 proves (2.19). (2.20) and (2.21) has been proved in the proof of Theorem 2.3.2.

Therefore from (2.18), (2.19), (2.20) and (2.21), we have

$$
\begin{aligned}
(\widehat{\theta}_n-\theta_0) &= n_1^{-1}H_{0,n}^{-1}\sum_{i\in I_1}w_n(Z_i)u_i\phi_\theta(Z_i,\theta_0)\\
&+ n_1^{-1}H_{0,n}^{-1}\phi_{w\theta,n_1}P_{n_1,k}(P_{n_2,k}'P_{n_2,k})^{-1}\sum_{i\in I_2}P_k(Z_i)\varepsilon_i\\
&+ o_p(n_1^{-1/2}+n_2^{-1})
\end{aligned}
$$

We next show that

$$n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)=n_1^{-1}\sum_{i\in I_1}w_n(Z_i)u_i\phi_\theta(Z_i,\theta_0)+o_p(n_1^{-1/2}) \tag{2.23}$$

and

$$\begin{aligned}&n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(\phi(Z_i,\theta_0)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\&=\phi_{w\theta,n_1}P_{n_2,k}(P'_{n_2,k}P_{n_2,k})^{-1}\sum_{i\in I_2}P_k(Z_i)\varepsilon_i+o_p(n_1^{-1/2}+n_2^{-1/2}).\end{aligned} \tag{2.24}$$

$$\begin{aligned}&n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\&=n_1^{-1}\sum_{i\in I_1}w_n(Z_i)u_i\phi_\theta(Z_i,\theta_0)\\&+n_1^{-1}\sum_{i\in I_1}(\widehat{w}_n(Z_i)-w_n(Z_i))u_i\phi_\theta(Z_i,\theta_0)\\&+n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(\phi(Z_i,\theta_0)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\phi_\theta(Z_i,\theta_0)\\&+n_1^{-1}\sum_{i\in I_1}\widehat{w}_n\widehat{\phi}_{n_2}(Z_i,\theta_0)\end{aligned}$$

$$\begin{aligned}&n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)(Y_i-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\&=n_1^{-1}\sum_{i\in I_1}u_iw_n(Z_i)\phi_\theta(Z_i,\theta_0)\\&+n_1^{-1}\sum_{i\in I_1}w_n(Z_i)(\phi(Z_i,\theta_0)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\phi_\theta(Z_i,\theta_0)\end{aligned}$$

Proof of Theorem 2.4.1. Define the empirical criterion function of the SNGLS estimation problem as

$$\widehat{L}_{n_1}(\theta)=n_1^{-1}\sum_{i\in I_1}\left[\widehat{w}_n(Z_i)\left|Y_i-\widehat{\phi}_{n_2}(Z_i,\theta)\right|^2\right]\text{ for any }\theta\in\Theta. \tag{2.25}$$

Define $L_n^*(\theta) = E\left[w_n(Z)\,|Y - \phi(Z,\theta)|^2\right]$. The conditional mean restriction $E[u|Z] = 0$ implies that

$$L_n^*(\theta) = E\left[w_n(Z)\,|\phi(Z,\theta) - \phi(Z,\theta_0)|^2\right] + E\left[w_n(Z)u^2\right]. \tag{2.26}$$

By Assumptions 2.4.1(ii) and 2.4.1(vii), $E\left[w_n(Z)u^2\right] \le C$ for any n which together with Assumption 2.4.1(vi) implies that θ_0 is uniquely identified as the minimizer of $L_n^*(\theta)$. Hence, to prove the consistency of $\widehat{\theta}_n$, it is sufficient to show that

$$\sup_{\theta\in\Theta}\left|\widehat{L}_{n_1}(\theta) - L_n^*(\theta)\right| = o_p(1). \tag{2.27}$$

Note that we can decompose $L_n(\theta)$ as

$$\begin{aligned}\widehat{L}_{n_1}(\theta) &= n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i^2 \\ &\quad + n_1^{-1}\sum_{i\in I_1}\left[\widehat{w}_n(Z_i)|\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta_0)|^2\right] \\ &\quad + 2n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i(\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta_0)).\end{aligned} \tag{2.28}$$

In the following, we show that

$$n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i^2 - E\left[w_n(Z)u^2\right] = o_p(1), \tag{2.29}$$

$$n_1^{-1}\sum_{i\in I_1}\widehat{w}_n(Z_i)u_i(\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta_0)) = o_p(1), \tag{2.30}$$

and

$$\begin{aligned}&n_1^{-1}\sum_{i\in I_1}\left[\widehat{w}_n(Z_i)|\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta_0)|^2\right] \\ &- -E\left[w_n(Z)\,|\phi(Z,\theta) - \phi(Z,\theta_0)|^2\right] = o_p(1)\end{aligned} \tag{2.31}$$

which together with (2.26) proves (2.27) and hence the claim of the theorem.

By the triangle inequality and Assumption 2.4.1(vii),

$$\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)| \le \sup_{z\in\mathcal{Z}}|\widehat{w}_n(z) - w_n(z)| + \sup_{z\in\mathcal{Z}}|w_n(z)| < 2C \tag{2.32}$$

with probability approaching 1. By Assumptions 2.4.1(ii) and 2.4.1(v),

$$\begin{aligned}&\left|n_1^{-1}\sum_{i\in I_1}\left[(\widehat{w}_n(Z_i) - w_n(Z_i))u_i^2\right]\right| \\ &\le \sup_{z\in\mathcal{Z}}|\widehat{w}_n(z) - w_n(z)|\, n_1^{-1}\sum_{i\in I_1}u_i^2 = o_p(1).\end{aligned} \tag{2.33}$$

By Assumptions 2.4.1(i), 2.4.1(ii) and 2.4.1(v),

$$n_1^{-1} \sum_{i \in I_1} \left[w_n(Z_i) u_i^2 - E\left[w_n(Z) u^2 \right] \right] = o_p(1) \tag{2.34}$$

which combined with (2.33) proves (2.29).

By Assumption 2.4.1(v) and (2.32)

$$n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta)|^2 \right] = o_p(1). \tag{2.35}$$

Using similar arguments in proving (2.29) but replacing Assumption 2.4.1(ii) with Assumption 2.4.1(iv), we can show that

$$\begin{aligned} & n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\phi(Z_i, \theta) - \phi(Z_i, \theta_0)|^2 \right] \\ & - E\left[w_n(Z) \left| \phi(Z, \theta) - \phi(Z, \theta_0) \right|^2 \right] = o_p(1), \end{aligned} \tag{2.36}$$

which together with Assumptions 2.4.1(iv) and 2.4.1(vii) implies that

$$n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\phi(Z_i, \theta) - \phi(Z_i, \theta_0)|^2 \right] = O_p(1). \tag{2.37}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} & \left| n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta))(\phi(Z_i, \theta) - \phi(Z_i, \theta_0)) \right] \right|^2 \\ & \leq n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta)|^2 \right] \\ & \times n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\phi(Z_i, \theta) - \phi(Z_i, \theta_0)|^2 \right] = o_p(1) \end{aligned} \tag{2.38}$$

where the equality is by (2.35) and (2.37). Combining the results in (2.35), (2.36) and (2.38), we deduce that

$$\begin{aligned} & n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0)|^2 \right] - E\left[w_n(Z) \left| \phi(Z, \theta) - \phi(Z, \theta_0) \right|^2 \right] \\ & = n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta)|^2 \right] - E\left[w_n(Z) \left| \phi(Z, \theta) - \phi(Z, \theta_0) \right|^2 \right] \\ & + n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) |\phi(Z_i, \theta) - \phi(Z_i, \theta_0)|^2 \right] \\ & + 2 n_1^{-1} \sum_{i \in I_1} \left[\widehat{w}_n(Z_i) (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta))(\phi(Z_i, \theta) - \phi(Z_i, \theta_0)) \right] = o_p(1) \end{aligned} \tag{2.39}$$

which proves (2.31).

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
& n_1^{-1} \sum_{i \in I_1} (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0))^2 \\
& \leq 2 n_1^{-1} \sum_{i \in I_1} (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta))^2 \\
& + 2 n_1^{-1} \sum_{i \in I_1} (\phi(Z_i, \theta) - \phi(Z_i, \theta_0))^2 = O_p(1) \qquad (2.40)
\end{aligned}
$$

where the equality is by Assumptions 2.4.1(i), 2.4.1(iv) and 2.4.1(v). By the triangle inequality and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
& \left| n_1^{-1} \sum_{i \in I_1} (\widehat{w}_n(Z_i) - w_n(Z_i)) u_i (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0)) \right| \\
& \leq \sup_{z \in \mathcal{Z}} |\widehat{w}_n(z) - w_n(z)| \left(n_1^{-1} \sum_{i \in I_1} u_i^2 \right)^{1/2} \\
& \times \left(n_1^{-1} \sum_{i \in I_1} (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0))^2 \right)^{1/2} = o_p(1) \qquad (2.41)
\end{aligned}
$$

where the equality is by Assumptions 2.4.1(i), 2.4.1(ii), 2.4.1(vii) and (2.40). By Assumption 2.4.1(i) and the conditional restriction $E[u|Z] = 0$,

$$
\begin{aligned}
& E\left[\left| n_1^{-1} \sum_{i \in I_1} u_i w_n(Z_i) (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0)) \right|^2 \middle| I_2, \{Z_i\}_{i \in I_1} \right] \\
& = n_1^{-2} \sum_{i \in I_1} E[u_i^2 | Z_i] (w_n(Z_i))^2 (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0))^2 \\
& \leq C n_1^{-2} \sum_{i \in I_1} (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0))^2 = O_p(n_1^{-1}) \qquad (2.42)
\end{aligned}
$$

which together with the Markov inequality implies that

$$
n_1^{-1} \sum_{i \in I_1} u_i w_n(Z_i) (\widehat{\phi}_{n_2}(Z_i, \theta) - \phi(Z_i, \theta_0)) = o_p(1). \qquad (2.43)
$$

Combining the results in (2.41) and (2.43) we proves (2.30). This finishes the proof of the theorem. ■

Proof of Lemma 2.4.1. By Assumption 2.3.2.(i)

$$
\begin{aligned}
& n_1^{-1} \sum_{i \in I_1} \widehat{w}_n(Z_i) u_i \widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - n_1^{-1} \sum_{i \in I_1} w_n(Z_i) u_i \phi_\theta(Z_i,\theta_0) \\
&= n_1^{-1} \sum_{i \in I_1} u_i (\widehat{w}_n(Z_i) - w_n(Z_i)) \widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) \\
&+ n_1^{-1} \sum_{i \in I_1} u_i w_n(Z_i) (\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0)) \\
&= n_1^{-1} \sum_{i \in I_1} u_i w_n(Z_i) (\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0)) + o_p(n_1^{-1/2}).
\end{aligned} \tag{2.44}
$$

Therefore, it is sufficient to show that

$$
n_1^{-1} \sum_{i \in I_1} u_i w_n(Z_i) (\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0)) = o_p(n_1^{-1/2}). \tag{2.45}
$$

By Assumption 2.4.1(i) and the conditional restriction $E[u|Z] = 0$,

$$
\begin{aligned}
& E\left[\left| n_1^{-1} \sum_{i \in I_1} u_i w_n(Z_i) (\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0)) \right|^2 \middle| I_2, \{Z_i\}_{i \in I_1} \right] \\
&= n_1^{-2} \sum_{i \in I_1} E[u_i^2 | Z_i] (w_n(Z_i))^2 (\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0))^2 \\
&\leq C n_1^{-2} \sum_{i \in I_1} (\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) - \phi_\theta(Z_i,\theta_0))^2 = o_p(n_1^{-1})
\end{aligned} \tag{2.46}
$$

which together with the Markov inequality implies (2.45). This finishes the proof. ■

2.5 Optimal Weighting

In this section, we compare the MD estimators through their finite sample variances. The comparison leads to an optimal weight matrix which gives MD estimator with smallest finite sample variance, as well as asymptotic variance, among all MD estimators. The following lemma simplifies the finite sample variance-covariance matrix which facilitates the comparison of the MD estimators.

Lemma 2.5.1 *Under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.3.2(v) and 2.3.3(iv)-2.3.3(vi),*

$$
H_{0,n}^{-1} (\Sigma_{n_1} + \Sigma_{n_2}) H_{0,n}^{-1} = V_{n,\theta}(1 + o_p(1)).
$$

where $V_{n,\theta} = H_{0,n}^{-1} E\left[w_n^2(Z) \left(n_1^{-1} \sigma_u^2(Z) + n_2^{-1} \sigma_\varepsilon^2(Z) \right) \phi_\theta(Z,\theta_0) \phi_\theta'(Z,\theta_0) \right] H_{0,n}^{-1}$.

If the sequence of the weight function is set to be

$$
w_n^*(Z) = (n_1^{-1} + n_2^{-1})(n_1^{-1} \sigma_u^2(Z) + n_2^{-1} \sigma_\varepsilon^2(Z))^{-1}, \tag{2.47}
$$

then the finite sample variance of the MD estimator becomes

$$V_{n,\theta}^{*}=\left(E\left[\left(\frac{\sigma_u^2(Z)}{n_1}+\frac{\sigma_\varepsilon^2(Z)}{n_2}\right)^{-1}\phi_\theta(Z,\theta_0)\phi_\theta'(Z,\theta_0)\right]\right)^{-1}. \tag{2.48}$$

The next lemma shows that V_θ^* is the smallest asymptotic variance-covariance of the MD estimator.

Theorem 2.5.1 *For any sequence of weight functions $w_n(Z)$, we have $V_{n,\theta}\geq V_{n,\theta}^*$ for any n_1 and any n_2.*

We call the MD estimator whose finite sample variance-covariance matrix equals V_θ^* optimal MD estimator. To ensure the optimal MD estimator is feasible, we have to: (i) show that $C^{-1}<w_n^*(z)<C$ for any $z\in\mathcal{Z}$ and any n_1, n_2; and (ii) construct an empirical weight function $\widehat{w}_n^*(z)$ such that $\sup_{z\in\mathcal{Z}}|\widehat{w}_n^*(z)-w_n^*(z)|=O_p(\delta_{w,n})$, where $\delta_{w,n}=O(n_1^{-1/4}+n_2^{-1/4})$. In the rest of this section, we show that $w_n^*(z)$ is bounded from above and from below. Construction of the empirical weight function $\widehat{w}_n^*(\cdot)$ is studied in the next section.

Lemma 2.5.2 *Under Assumption 2.3.3(v), $C^{-1}<w_n^*(z)<C$ for any $z\in\mathcal{Z}$ and any n_1, n_2.*

2.6 Estimation of the Variance and Optimal Weighting

The estimator of the variance-covariance matrix is constructed by its sample analog. Let $\widehat{u}_i=Y_i-\widehat{h}_{n_1}(Z_i)$ for any $i\in I_1$, and $\widehat{\varepsilon}_i=g(Z_i,\widehat{\theta}_n)-\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_n)$ for any $i\in I_2$. Define

$$\begin{aligned}
\widehat{H}_n&=n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)',\\
\widehat{\Sigma}_{n_1}&=\frac{\widehat{\phi}_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}\widehat{Q}_{n_1,u}Q_{n_1,k_1}^{-1}P_{n,k_1}'\widehat{\phi}_{w\theta,n}'}{n^2n_1},\\
\widehat{\Sigma}_{n_2}&=\frac{\widehat{\phi}_{w\theta,n}P_{n,k_2}Q_{n_2,k_2}^{-1}\widehat{Q}_{n_2,\varepsilon}Q_{n_2,k_2}^{-1}P_{n,k_2}'\widehat{\phi}_{w\theta,n}'}{n^2n_2},
\end{aligned}$$

where $\widehat{\phi}_{w\theta,n}=(\widehat{w}_n(Z_i)\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n))_{i\in I}$, $\widehat{Q}_{n_1,u}=n_1^{-1}\sum_{i\in I_1}\widehat{u}_i^2P_{k_1}(Z_i)P_{k_1}'(Z_i)$ and $\widehat{Q}_{n_2,\varepsilon}=n_2^{-1}\sum_{i\in I_2}\widehat{\varepsilon}_i^2P_{k_2}(Z_i)P
The variance estimator is defined as

$$\widehat{V}_n=\widehat{H}_n^{-1}(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})\widehat{H}_n^{-1}. \tag{2.49}$$

The following conditions are needed to show the consistency of $\widehat{V}_n$ and the empirical optimal weight function constructed later in this section.

Assumption 2.6.1 *(i)* $\sup_{\theta\in\mathcal{N}_n} n_2^{-1}\sum_{i\in I_2}\|g_\theta(X_i,\theta)\|^2 = O_p(1)$*; (ii) there exist* $\beta_{u,k}\in R^k$ *and* $r_u>0$ *such that*

$$\sup_{z\in\mathcal{Z}}\left|\sigma_u^2(z)-P_k(z)'\beta_{u,k}\right| = O(k^{-r_u}); \tag{2.50}$$

(iii) there exist $\beta_{\varepsilon,k}\in R^k$ *and* $r_\varepsilon>0$ *such that*

$$\sup_{z\in\mathcal{Z}}\left|\sigma_\varepsilon^2(z)-P_k(z)'\beta_{\varepsilon,k}\right| = O(k^{-r_\varepsilon}); \tag{2.51}$$

(iv) $\max_{j=1,2}(\xi_{k_j}k_j^{1/2}n_j^{-1/2}+\xi_{k_j}k_j^{-r_h}) = o(1)$*; (v)* $E[\|g_\theta(X,\theta_0)\|^4]\le C$.

Assumption 2.6.1(i) requires that the sample average of $\|g_\theta(X_i,\theta)\|$ is stochastically bounded uniformly over the local neighborhood of θ_0. Assumptions 2.6.1(ii) and 2.6.1(iii) implies that the conditional variances $\sigma_u^2(z)$ and $\sigma_\varepsilon^2(z)$ can be approximated by the basis functions $P_k(z)$. Assumption 2.6.1(iv) imposes restrictions on the numbers of basis functions and the smoothness of the conditional variance functions. Assumption 2.6.1(v) imposes finite fourth moment on $g_\theta(X,\theta_0)$.

Theorem 2.6.1 *Suppose Assumptions 2.3.1, 2.3.2, 2.3.3, 2.6.1(i) and 2.6.1(iv) hold. If* $(k_1+k_2)\delta_{w,n}^2=o(1)$*, then we have*

$$\widehat{H}_n^{-1}(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})\widehat{H}_n^{-1} = H_{0,n}^{-1}(\Sigma_{n_1}+\Sigma_{n_2})H_{0,n}^{-1}(1+o_p(1) \tag{2.52}$$

and moreover,

$$\gamma_n'(\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H}_n)^{\frac{1}{2}}(\widehat{\theta}_n-\theta_0)\to_d N(0,1), \tag{2.53}$$

for any non-random sequence $\gamma_n\in R^{d_\theta}$ *with* $\gamma_n'\gamma_n=1$.

Remark 5.1. By the consistency of the $\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H}_n$ and CMT,

$$(\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H})^{1/2}(\widehat{\theta}_n-\theta_0)\to_d N(0,I_{d_\theta}),$$

which together with the CMT implies that

$$W_n(\theta_0)=(\widehat{\theta}_n-\theta_0)'(\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H})(\widehat{\theta}_n-\theta_0)\to_d\chi^2(d_\theta). \tag{2.54}$$

Recall that ι_j^* is the $d_\theta\times 1$ selection vector whose j-th ($j=1,\ldots,d_\theta$) component is 1 and rest components are 0. By the consistency of the $\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H}_n$, we have

$$\iota_j^{*\prime}\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H}_n\iota_j^* = \iota_j^{*\prime}H_0^{-1}(\Sigma_{n_1}+\Sigma_{n_2})H_0^{-1}\iota_j^*(1+o_p(1)$$

which together with (2.13) and the CMT implies that

$$t_{j,n}(\theta_{j,0})=\frac{\widehat{\theta}_{j,n}-\theta_{j,0}}{\sqrt{\iota_j^{*\prime}\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_1})^{-1}\widehat{H}\iota_j^*}}\to_d N(0,1). \tag{2.55}$$

The Student-t statistic in (2.55) and the Wald-statistic in (2.54) can be applied to conduct inference on $\theta_{j,0}$ for $j = 1, \dots, d_\theta$ and joint inference on θ_0 respectively.

Remark 5.2. Theorem 2.6.1 can be applied to conduct inference on θ_0 using the identity weighted MD estimator $\widehat{\theta}_{1,n}$ defined as

$$\widehat{\theta}_{1,n} = \arg\min_{\theta\in\Theta} n^{-1} \sum_{i\in I} (\widehat{h}_{n_1}(Z_i) - \widehat{\phi}_{n_2}(Z_i, \theta))^2. \tag{2.56}$$

As the identity weight function satisfies Assumption 2.3.2(v) and the condition $(k_1 + k_2)\delta^2_{w,n} = o(1)$ holds trivially, under Assumptions 2.3.1, 2.3.2(i)-(iv) and 2.3.3, Theorem 2.3.2 implies that

$$\widehat{\theta}_{1,n} - \theta_0 = O_p(n_1^{-1/2} + n_2^{-1/2}). \tag{2.57}$$

The identity weighted MD estimator can be used to construct the empirical weight function which enables us to construct the optimal MD estimator.

Let $\widehat{u}_i = Y_i - \widehat{h}_{n_1}(Z_i)$ for any $i \in I_1$, and $\widetilde{\varepsilon}_i = g(Z_i, \widehat{\theta}_{1,n}) - \widehat{\phi}_{n_2}(Z_i, \widehat{\theta}_{1,n})$ for any $i \in I_2$. Define

$$\widehat{w}^*_n(z) = (n_1^{-1} + n_2^{-1})(n_1^{-1}\widehat{\sigma}^2_{n,u}(z) + n_2^{-1}\widehat{\sigma}^2_{n,\varepsilon}(z))^{-1}, \tag{2.58}$$

where $\widehat{\sigma}^2_{n,u}(z)$ and $\widehat{\sigma}^2_{n,\varepsilon}(z)$ are the estimators of the conditional variances $\sigma^2_u(z)$ and $\sigma^2_\varepsilon(z)$:

$$\widehat{\sigma}^2_{n,u}(z) = n_1^{-1} P'_{k_1}(z) Q^{-1}_{n_1,k_1} P'_{n_1,k_1} \widehat{U}_{2,n_1} \text{ and } \widehat{\sigma}^2_{n,\varepsilon}(z) = n_2^{-1} P'_{k_2}(z) Q^{-1}_{n_2,k_2} P'_{n_2,k_2} \widehat{e}_{2,n_2}, \tag{2.59}$$

where $\widehat{U}_{2,n_1} = (\widehat{u}^2_i)'_{i\in I_1}$ and $\widehat{e}_{2,n_2} = (\widetilde{\varepsilon}^2_i)'_{i\in I_2}$. The optimal MD estimator is defined as

$$\widehat{\theta}^*_n = \arg\min_{\theta\in\Theta} n^{-1} \sum_{i\in I} \widehat{w}^*_n(Z_i)(\widehat{h}_{n_1}(Z_i) - \widehat{\phi}_{n_2}(Z_i, \theta))^2. \tag{2.60}$$

To show the optimality of $\widehat{\theta}^*_n$, it is sufficient to show that $\widehat{w}^*_n(Z_i)$ satisfies the high level conditions in Assumption 2.3.2(v). For this purpose, we first derive the convergence rates of $\widehat{\sigma}^2_{n,u}(z)$ and $\widehat{\sigma}^2_{n,\varepsilon}(z)$.

Lemma 2.6.1 *Under Assumptions 2.3.1, 2.3.2(i)-(iv), 2.3.3 and 2.6.1, we have*

$$\sup_{z\in\mathcal{Z}} \left|\widehat{\sigma}^2_{n,u}(z) - \sigma^2_u(z)\right| = O_p(\xi_{k_1}(k_1^{1/2} n_1^{-1/2} + k_1^{-r_u}) + \xi^2_{k_1} k_1^{-2r_h}),$$

and

$$\sup_{z\in\mathcal{Z}} \left|\widehat{\sigma}^2_{n,\varepsilon}(z) - \sigma^2_\varepsilon(z)\right| = O_p(\xi_{k_2}(k_2^{1/2} n_2^{-1/2} + k_2^{-r_\varepsilon}) + \xi^2_{k_2}(n_1^{-1} + k_2^{-2r_h})).$$

Remark 5.3. Under Assumption 2.6.1(iv) and

$$\xi_{k_1} k_1^{-r_u} + \xi_{k_2} k_2^{-r_\varepsilon} + \xi^2_{k_2} n_1^{-1} = o(1), \tag{2.61}$$

Lemma 2.6.1 implies that

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}^2_{n,u}(z)-\sigma^2_u(z)\right|=o_p(1) \text{ and } \sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}^2_{n,\varepsilon}(z)-\sigma^2_\varepsilon(z)\right|=o_p(1), \tag{2.62}$$

which means that $\widehat{\sigma}^2_{n,u}(z)$ and $\widehat{\sigma}^2_{n,\varepsilon}(z)$ are consistent estimators of $\sigma^2_u(z)$ and $\sigma^2_\varepsilon(z)$ under the uniform metric.

Theorem 2.6.2 *Under (2.61), Assumptions 2.3.1, 2.3.2(i)-(iv), 2.3.3 and 2.6.1, we have*

$$\sup_{z\in\mathcal{Z}}|\widehat{w}^*_n(z)-w^*(z)|=O_p(\delta_{w,n})$$

where $\delta_{w,n}=\max_{j=1,2}(\xi_{k_j}k_j^{1/2}n_j^{-1/2}+\xi_{k_j}^2k_j^{-2r_h})+\xi_{k_1}k_1^{-r_u}+\xi_{k_2}k_2^{-r_\varepsilon}+\xi_{k_2}^2n_1^{-1}$.

Remark 5.4. When the power series are used as the basis functions $P_k(z)$, we have $\xi_{k_j}\le Ck_j$. Then the convergence rate of $\delta_{w,n}$ is simplified as

$$\delta_{w,n}=\max_{j=1,2}(k_j^{3/2}n_j^{-1/2}+k_j^{2-2r_h})+k_1^{1-r_u}+k_2^{1-r_\varepsilon}+k_2^2n_1^{-1}.$$

Hence in this case $\delta_{w,n}=o(1)$, if $\max_{j=1,2}k_j^3n_j^{-1}+k_2^2n_1^{-1}=o(1)$, $r_h>1$, $r_u>1$ and $r_\varepsilon>1$. The condition $\delta_{w,n}=O(n_1^{-1/4}+n_2^{-1/4})$ hold when

$$\max_{j=1,2}k_j^6n_j^{-1}+k_2^{8/3}n_1^{-1}=O(1) \text{ and } \max_{j=1,2}n_j^{1/4}k_j^{2-2r_h}+n_1^{1/4}k_1^{1-r_u}+n_2^{1/4}k_2^{1-r_\varepsilon}=O(1). \tag{2.63}$$

Moreover, $(k_1+k_2)\delta^2_{w,n}=o(1)$ holds under (2.63) and $k_2^2n_1^{-1}=o(1)$.

Remark 5.5. When the splines or trigonometric functions are used as the basis functions $P_k(z)$, we have $\xi_{k_j}\le Ck_j^{1/2}$. Then the convergence rate of $\delta_{w,n}$ is simplified as

$$\delta_{w,n}=\max_{j=1,2}(k_jn_j^{-1/2}+k_j^{1-2r_h})+k_1^{1/2-r_u}+k_2^{1/2-r_\varepsilon}+k_2n_1^{-1}.$$

Hence in this case $\delta_{w,n}=o(1)$, if $\max_{j=1,2}k_j^2n_j^{-1}+k_2^2n_1^{-1}=o(1)$, $r_h>1/2$, $r_u>1/2$ and $r_\varepsilon>1/2$. The condition $\delta_{w,n}=O(n_1^{-1/4}+n_2^{-1/4})$ hold when

$$\max_{j=1,2}k_j^4n_j^{-1}+k_2^{8/3}n_1^{-1}=O(1) \text{ and } \max_{j=1,2}n_j^{1/4}k_j^{1-2r_h}+n_1^{1/4}k_1^{1-r_u}+n_2^{1/4}k_2^{1-r_\varepsilon}=O(1). \tag{2.64}$$

Moreover, $(k_1+k_2)\delta^2_{w,n}=o(1)$ holds under (2.64) and $k_2^2n_1^{-1}=o(1)$.

2.7 Monte Carlo Simulation

In this section, we study the finite sample performances of the MD estimator and the proposed inference method. The simulated data is from the following model

$$Y_i = g(X_i, \theta_0) + v_i, \tag{2.65}$$

where Y_i, X_i and v_i are scale random variables, $g(X_i, \theta_0)$ is a function specified in the following

$$g(X_i, \theta_0) = \begin{cases} X_i\theta_0 & \text{in Model 1} \\ \log(1 + X_i^2\theta_0), & \text{in Model 2} \end{cases}, \tag{2.66}$$

where $\theta_0 = 1$ is the unknown parameter.

To generate the simulated data, we first generate $(X_{1,i}^*, X_{2,i}^*, v_i)'$ from joint normal distribution with mean zero and identity variance-covariance matrix. Let

$$Z_i = X_{2,i}^*(1 + X_{2,i}^{*2})^{-1/2} \text{ and } X_i = Z_i + X_{1,i}^* \log(Z_i^2). \tag{2.67}$$

We assume that (Y_i, Z_i) are observed together and (X_i, Z_i) are observed together. We generate the first data set $\{(Y_i, Z_i)\}_{i \in I_1}$ with sample size n_1, and then independently generate the second data set $\{(X_i, Z_i)\}_{i \in I_2}$ with sample size n_2. As the both the magnitudes of n_1, n_2 and their relative magnitude are important to the finite sample properties of the MD estimator, we consider two sampling schemes (i.e., equal sampling and unequal sampling) separately. In the first scheme (equal sampling), $n_1 = n_2 = n_0$ where n_0 starts from 50 with increment 50 and ends at 1000. In the second scheme (unequal sampling), $n_1 + n_2 = 1000$ where n_1 starts from 100 with increment 50 and ends at 900. For each combination of n_1 and n_2, we generate 10000 simulated samples to evaluate the performances of the MD estimator and the proposed inference procedure.

In addition to the MD estimator, we also study two alternative estimators based on data imputation. The first estimator (which is called X-imputed estimator in this section) is defined as

$$\widehat{\theta}_{X,n} = \arg\min_{\theta \in \Theta} n_1^{-1} \sum_{i \in I_1} (Y_i - g(\widehat{X}_i, \theta))^2 \tag{2.68}$$

where $\widehat{X}_i = n_2^{-1} P_{k_2}'(Z_i) Q_{n_2,k_2}^{-1} \sum_{i \in I_2} X_i P_{k_2}(Z_i)$ for any $i \in I_1$ is the predicted value of X_i in the first data set based on nonparametric regression. The second estimator (which is called the Y-imputed estimator in this section) is defined as

$$\widehat{\theta}_{Y,n} = \arg\min_{\theta \in \Theta} n_2^{-1} \sum_{i \in I_2} (\widehat{Y}_i - g(X_i, \theta))^2 \tag{2.69}$$

where $\widehat{Y}_i = n_1^{-1} P_{k_1}'(Z_i) Q_{n_1,k_1}^{-1} \sum_{i \in I_1} Y_i P_{k_1}(Z_i)$ for any $i \in I_2$ is the predicted value of Y_i in the second data set based on nonparametric regression. In the simulation studies, we set $k_1 = k_2 = 5$

and $P_{k_1}(Z) = P_{k_2}(Z) = (1, Z, Z^2, Z^3, Z^4)$. The minimization problem in the MD estimation and the nonlinear regressions (in (2.68) and (2.69)) are solved by grid search with $\Theta = [0, 2]$ and equally spaced grid points with grid length 0.001).

The finite sample properties of the identity weighted MD estimator (the green dashed line), the optimal weighted MD estimator (the black solid line), the X-imputed estimator (the blue dotted line) and the Y-imputed estimator (the red dash-dotted line) are provided in Figures 6.1 and 6.2. In Figure 6.1, we see that the bias and variance of the two MD estimators converge to zero with the growth of two sample sizes n_1 and n_2. The optimal weighted MD estimator has smaller bias and smaller variance, and hence smaller RMSE than the identity weighted MD estimator. The improvement of the optimal MD estimator over the identity weighted MD estimator is particularly clear in model 1. The X-imputed estimator has almost the same finite sample bias and variance as the identity weighted MD estimator in the linear model (i.e., model 1). But it has large and non-convergent finite sample bias in model 2 which indicates that it is inconsistent in the general nonlinear model. The Y-imputed estimator has large and non-convergent finite sample bias in both model 1 and model 2 which shows that it is an inconsistent estimator in general. The finite sample performances of the MD estimators under unequal sampling scheme are presented in Figure 6.2. In this figure, we see that when n_1 (or n_2) is small, the finite sample bias and variance of the MD estimators are large regardless how big n_2 (or n_1) is. This means that the main part in the estimation error of MD estimator is from the component estimated by the small sample, which is implied by Theorem 2.3.2.

The finite sample properties of the inference procedure based on the identity weighted MD estimator and the optimal weighted MD estimator are provided in Figures 6.3 and 6.4. In Figure 6.3, we see that the finite coverage probabilities of the confidence intervals based on the MD estimators converge to the nominal level 0.9 with both n_1 and n_2 increase to 1000. In model 1, the coverage probability of the confidence interval based on the optimal MD estimator is closer to the nominal level than that based on the identity weighted MD estimator in all sample sizes we considered. In model 2, the confidence interval based on the optimal MD estimator is slightly worse than that based on the identity weighted MD estimator when the sample sizes n_1 and n_2 are small, and the coverage probabilities of the two confidence intervals are identical and close to the nominal level when n_1 and n_2 are larger than 250. In both model 1 and model 2, the average length of the confidence interval of the optimal MD estimator is much smaller than that of the confidence interval of the identity weighted MD estimator, which is because the optimal MD estimator has smaller variance. The finite sample performances of the confidence intervals based on the MD estimators under unequal

Figure 2.1: Figure 6.1. Properties of the MD and the Imputation Estimators ($n_1 = n_2$)

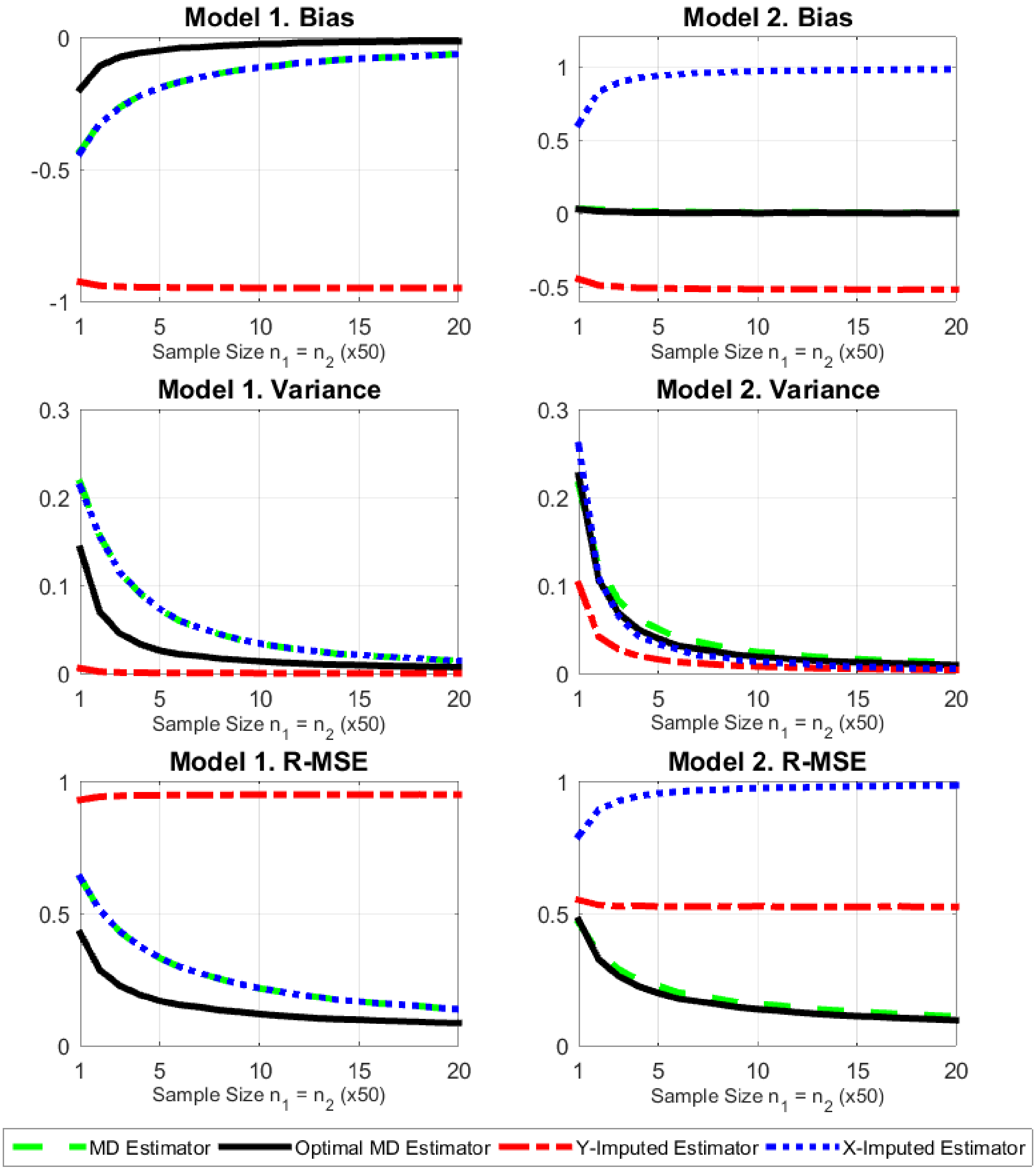


Figure 2.2: Figure 6.2. Properties of the MD and the Imputation Estimators ($n_1 + n_2 = 1000$)

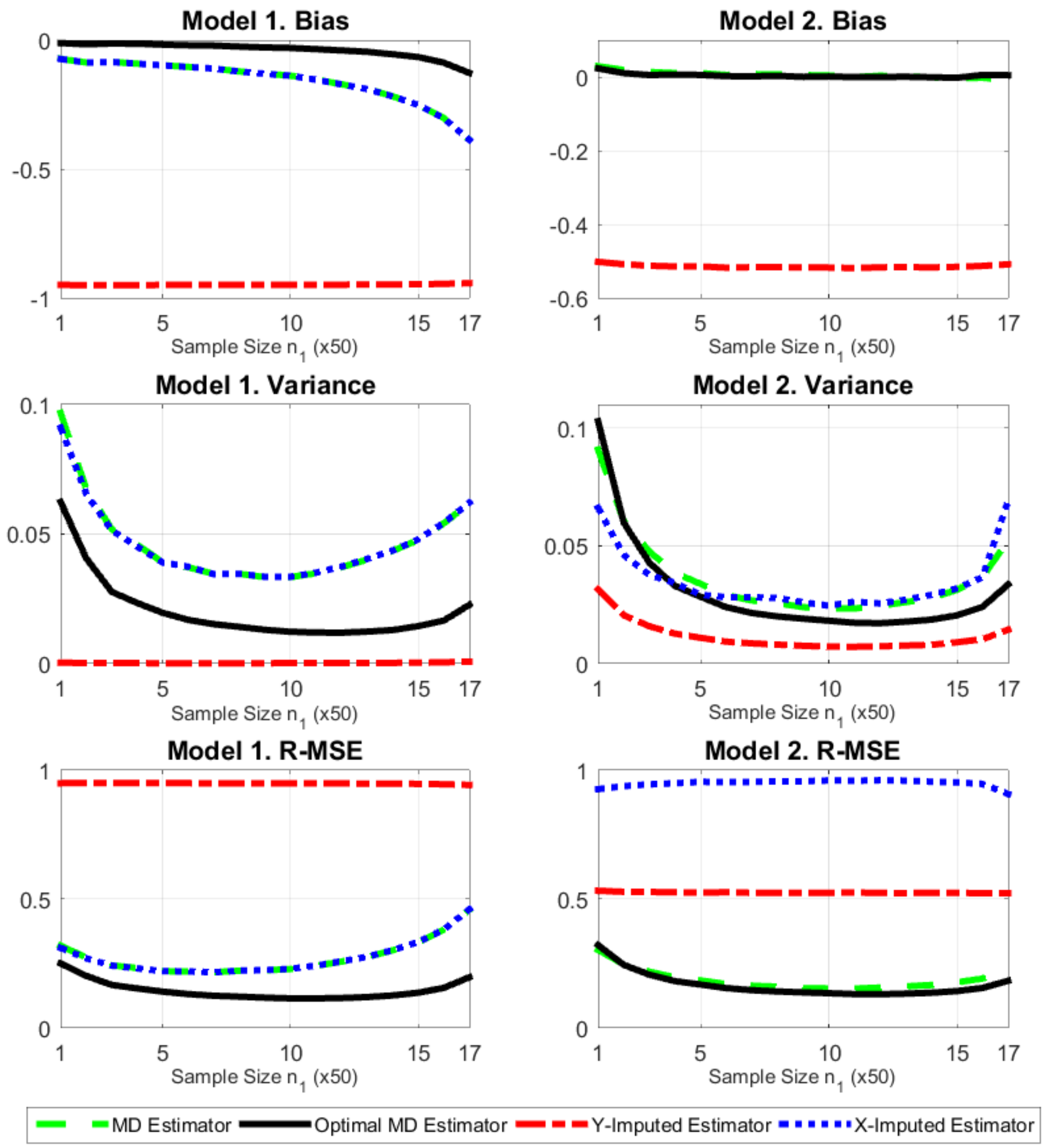


Figure 2.3: Figure 6.3. Properties of the Confidence Intervals ($n_1 = n_2$)

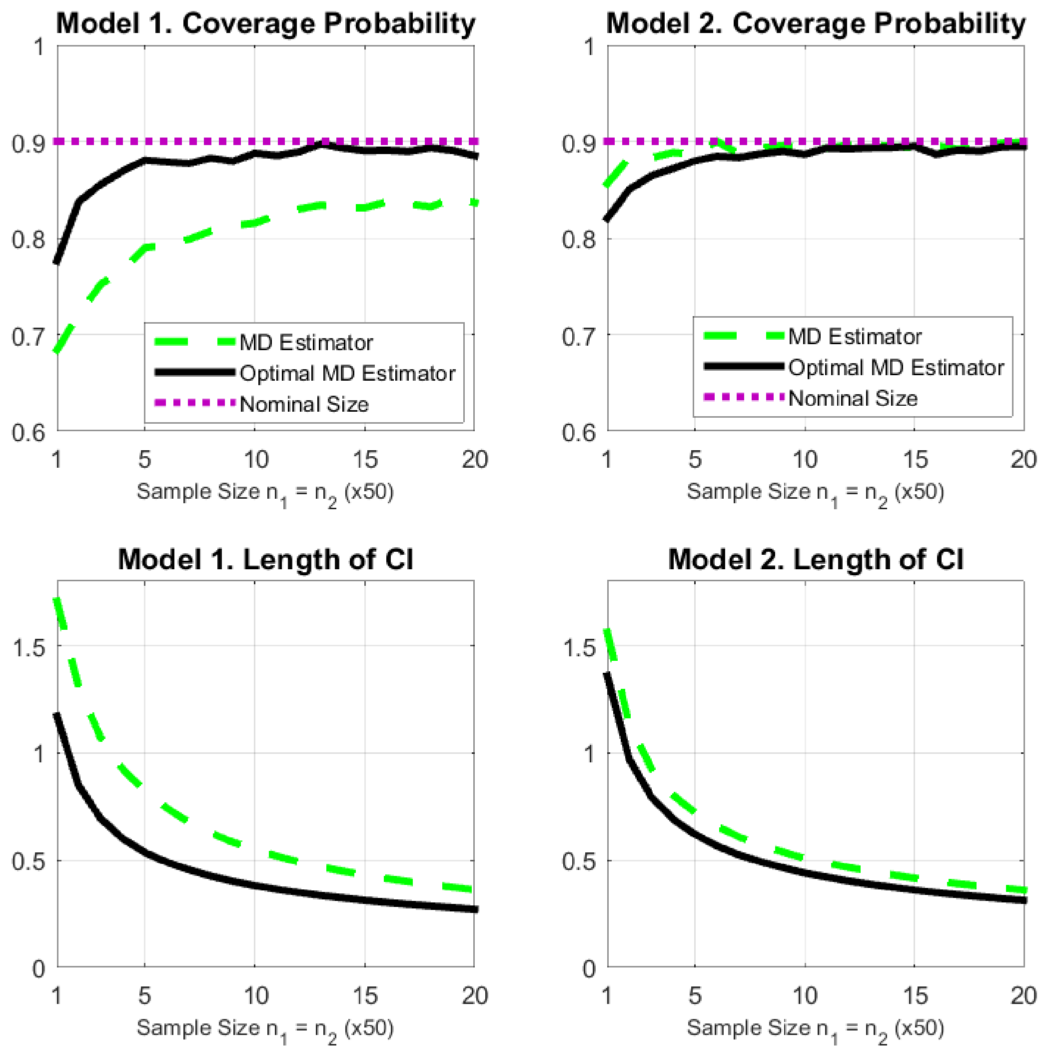
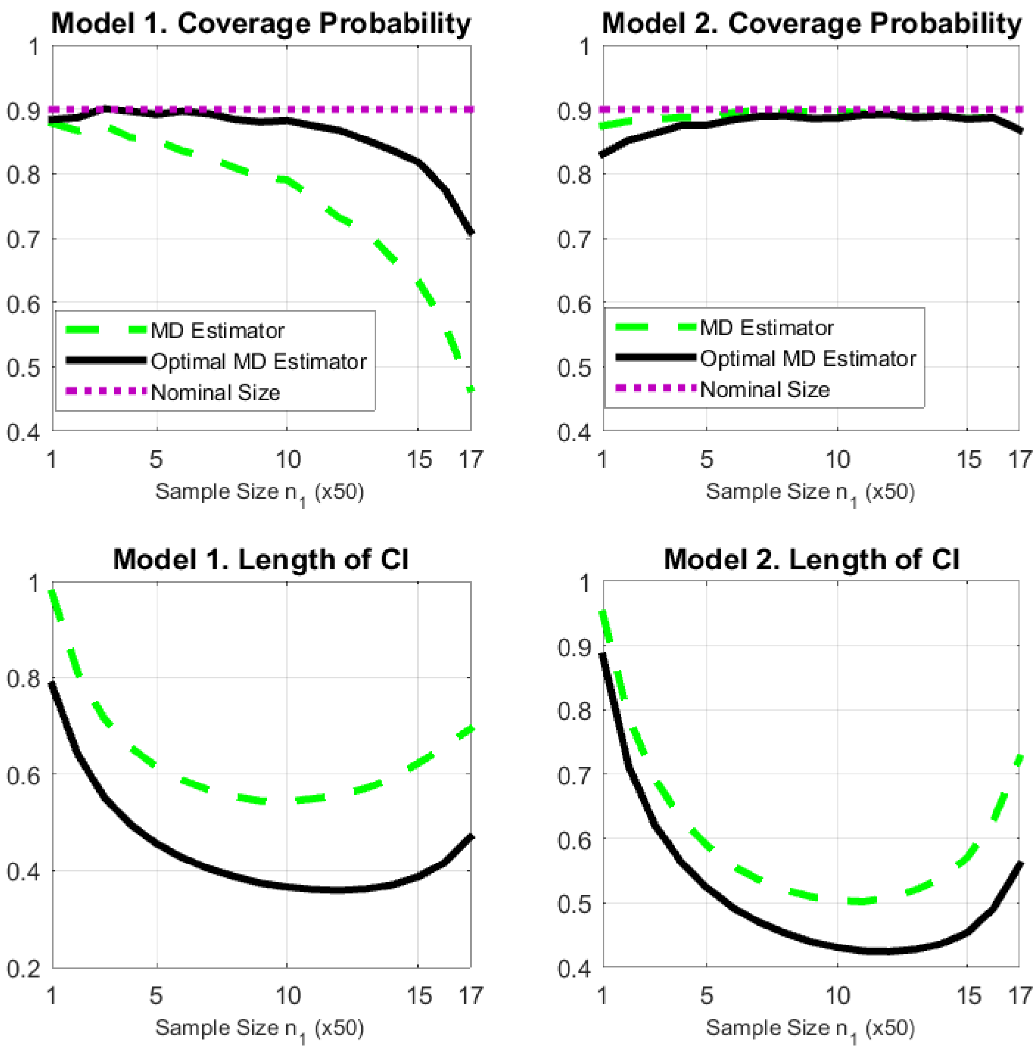


Figure 2.4: Figure 6.4. Properties of the Confidence Intervals ($n_1 + n_2 = 1000$)



sampling scheme are presented in Figure 6.4. In this figure, we see that when n_1 (or n_2) is small, the coverage probabilities of the confidence intervals of the two MD estimators are away from the nominal level. The performance of the inference based on the identity weighted MD estimator is poor in model 1 when the sample size n_2 is small regardless the size of the other sample n_1 is big (close to 1000). From Figure 6.4, we also see that the average length of the confidence intervals of the MD estimators is large when n_1 or n_2 is small.

2.8 Conclusion

This paper studies estimation and inference of nonlinear econometric models when the economic variables of the models are contained in different data sets in practice. We provide a semiparametric MD estimator based on conditional moment restriction with common conditioning variables which are contained in different data sets. The MD estimator is show to be consistent and has asymptotic normal distribution. We provide the specific form of optimal weight for the MD estimation, and show that the optimal weighted MD estimator has the smallest asymptotic variance among all MD estimators. Consistent estimator of the variance-covariance matrix of the MD estimator, and hence inference procedure of the unknown parameter is also provided. The finite sample performances of the MD estimator and the inference procedure are investigated in simulation studies.

2.9 Appendix

2.9.1 Proof of the Main Results in Section 2.3

Proof of Theorem 2.3.1. Define the empirical criterion function of the MD estimation problem as

$$\widehat{L}_n(\theta) = n^{-1}\sum_{i\in I}\left[\widehat{w}_n(Z_i)\left|\widehat{h}_{n_1}(Z_i) - \widehat{\phi}_{n_2}(Z_i,\theta)\right|^2\right] \text{ for any } \theta \in \Theta. \tag{2.70}$$

By Assumptions 2.3.2(iii) and 2.3.2(v),

$$\inf_{\{\theta\in\Theta:\ ||\theta-\theta_0||\geq\varepsilon\}} L_n^*(\theta) \geq \eta_{C,\varepsilon} \tag{2.71}$$

where $\eta_{C,\varepsilon} = C\eta_\varepsilon > 0$ is a fixed constant which only depends on ε. (2.71) implies that θ_0 is uniquely identified as the minimizer of $L_n^*(\theta)$. Hence, to prove the consistency of $\widehat{\theta}_n$, it is sufficient to show that

$$\sup_{\theta\in\Theta}\left|\widehat{L}_n(\theta) - L_n^*(\theta)\right| = o_p(1). \tag{2.72}$$

Note that we can decompose $L_n(\theta)$ as

$$\begin{aligned}
\widehat{L}_n(\theta) = {} & n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(|\widehat{h}_{n_1}(Z_i) - h_0(Z_i)|^2 + |\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta)|^2 + |h_0(Z_i) - \phi(Z_i,\theta)|^2) \\
& - 2n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i) - h_0(Z_i))(\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta)) \\
& - 2n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta))(h_0(Z_i) - \phi(Z_i,\theta)) \\
& + 2n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i) - h_0(Z_i))(h_0(Z_i) - \phi(Z_i,\theta)).
\end{aligned} \tag{2.73}$$

Using Assumption 2.3.1(i), one can use Rudelson's law of large numbers for matrices (see, e.g., Lemma 6.2 in Belloni, et. al. (2015)) to get

$$Q_{n,k_j} - Q_{k_j} = O_p(n^{-1/2}\xi_{k_j}(\log(k_j))^{1/2}) \text{ and } Q_{n_j,k_j} - Q_{k_j} = O_p(n_j^{-1/2}\xi_{k_j}(\log(k_j))^{1/2}) \tag{2.74}$$

where $Q_{n,k_j} = n^{-1}\sum_{i\in I} P_{k_j}(Z_i)P'_{k_j}(Z_i)$, $Q_{n_j,k_j} = n_1^{-1}\sum_{i\in I_j} P_{k_j}(Z_i)P'_{k_j}(Z_i)$ and the convergence is under the operator norm of matrix. By (2.74), Assumptions 2.3.1(iii) and 2.3.1(v),

$$C^{-1} \leq \lambda_{\min}(Q_{n,k_j}) \leq \lambda_{\max}(Q_{n,k_j}) \leq C \text{ and } C^{-1} \leq \lambda_{\min}(Q_{n_j,k_j}) \leq \lambda_{\max}(Q_{n_j,k_j}) \leq C, \tag{2.75}$$

with probability approaching 1. Under Assumption 2.3.1 and (2.75), (A.2) in the proof of Theorem 1 in Newey (1997) implies that

$$\left\|\widehat{\beta}_{k_1,n_1} - \beta_{h,k_1}\right\|^2 = O_p(k_1 n_1^{-1} + k_1^{-2r_h}). \tag{2.76}$$

By the triangle inequality,

$$
\begin{aligned}
& n^{-1}\sum_{i\in I}\left|\widehat{h}_{n_1}(Z_i)-h_0(Z_i)\right|^2 \\
& \le 2n^{-1}\sum_{i\in I}\left|\widehat{h}_{n_1}(Z_i)-h_{0,k_1}(Z_i)\right|^2+2n^{-1}\sum_{i\in I}|h_{0,k_1}(Z_i)-h_0(Z_i)|^2 \\
& \le 2(\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1})'Q_{k_1,n}(\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1})+2\sup_{z\in\mathcal{Z}}|h_{0,k_1}(z)-h_0(z)|^2 \\
& \le 2\lambda_{\max}(Q_{k_1,n})\left\|\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1}\right\|^2+2\sup_{z\in\mathcal{Z}}|h_{0,k_1}(z)-h_0(z)|^2 \\
& = O_p(k_1n_1^{-1}+k_1^{-2r_h})=o_p(1)
\end{aligned}
\tag{2.77}
$$

where the first equality is by Assumption 2.3.1(iv), (2.75) and (2.76), the second equality is by Assumption 2.3.1(v).

By the triangle inequality and Assumption 2.3.2(v),

$$
\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)|\le\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)-w_n(z)|+\sup_{z\in\mathcal{Z}}|w_n(z)|<2C \tag{2.78}
$$

with probability approaching 1. By (2.77) and (2.78),

$$
n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\left|\widehat{h}_{n_1}(Z_i)-h_0(Z_i)\right|^2\le\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)|\sum_{i\in I}\left|\widehat{h}_{n_1}(Z_i)-h_0(Z_i)\right|^2=o_p(1). \tag{2.79}
$$

By (2.78) and Assumption 2.3.2(ii),

$$
\begin{aligned}
& \sup_{\theta\in\Theta}n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\left|\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta)\right|^2 \\
& \le\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)|\sup_{\theta\in\Theta}n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta)\right|^2=o_p(1).
\end{aligned}
\tag{2.80}
$$

Using (2.77), (2.78), Assumption 2.3.2(ii) and the Cauchy-Schwarz inequality, we get

$$
\begin{aligned}
& \sup_{\theta\in\Theta}\left|n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))(\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta))\right| \\
& \le\sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)|\sqrt{n^{-1}\sum_{i\in I}\left|\widehat{h}_{n_1}(Z_i)-h_0(Z_i)\right|^2}\sqrt{\sup_{\theta\in\Theta}n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta)\right|^2}=o_p(1).
\end{aligned}
\tag{2.81}
$$

By Assumption 2.3.1(ii), $E\left[h_0^2(Z)\right]<C$, which together with Assumption 2.3.2(i) implies that

$$
\sup_{\theta\in\Theta}E\left[|h_0(Z)-\phi(Z,\theta)|^2\right]\le 2E\left[h_0^2(Z)\right]+2\sup_{\theta\in\Theta}E\left[\phi^2(Z,\theta)\right]<C. \tag{2.82}
$$

By (2.82), Assumptions 2.3.2.(iv) and 2.3.2.(v),

$$\begin{aligned}
&\sup_{\theta\in\Theta} n^{-1}\sum_{i\in I}|h_0(Z_i)-\phi(Z_i,\theta)|^2\\
&\leq C\sup_{\theta\in\Theta} n^{-1}\sum_{i\in I} w_n(Z_i)\,|h_0(Z_i)-\phi(Z_i,\theta)|^2\\
&\leq (C+o_p(1))\sup_{\theta\in\Theta} E\left[w_n(Z_i)\,|h_0(Z_i)-\phi(Z_i,\theta)|^2\right]\\
&\leq (C+o_p(1))\sup_{\theta\in\Theta} E\left[|h_0(Z_i)-\phi(Z_i,\theta)|^2\right]=O_p(1). \qquad (2.83)
\end{aligned}$$

Using (2.78), (2.83), Assumptions 2.3.2(ii) and (iv), and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta))(h_0(Z_i)-\phi(Z_i,\theta))\right|\\
&\leq \sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)|\sup_{\theta\in\Theta}\sqrt{n^{-1}\sum_{i\in I}|h_0(Z_i)-\phi(Z_i,\theta)|^2}\sqrt{n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta)\right|^2}=o_p(1). \qquad (2.84)
\end{aligned}$$

Similarly, using (2.78), (2.77), (2.83), Assumptions 2.3.2(iv) and the Cauchy-Schwarz inequality, we get

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))(h_0(Z_i)-\phi(Z_i,\theta))\right|=o_p(1). \qquad (2.85)$$

Collecting the results in (2.73), (2.79), (2.80), (2.81), (2.84) and (2.85), we get

$$\sup_{\theta\in\Theta} L_n(\theta)=\sup_{\theta\in\Theta} n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\,|h_0(Z_i)-\phi(Z_i,\theta)|^2+o_p(1). \qquad (2.86)$$

By (2.83) and Assumption 2.3.2(v),

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i\in I}(\widehat{w}_n(Z_i)-w_n(Z_i))\,|h_0(Z_i)-\phi(Z_i,\theta)|^2\right|\\
&\leq \sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)-w_n(z)|\sup_{\theta\in\Theta} n^{-1}\sum_{i\in I}|h_0(Z_i)-\phi(Z_i,\theta)|^2=o_p(1) \qquad (2.87)
\end{aligned}$$

which together with (2.86) and Assumption 2.3.2(iv),

$$\sup_{\theta\in\Theta}\left|\widehat{L}_n(\theta)-L_n^*(\theta)\right|=\sup_{\theta\in\Theta}|L_n(\theta)-L_n^*(\theta)|+o_p(1)=o_p(1). \qquad (2.88)$$

This proves (2.72) and hence the claim of the theorem. ■

Lemma 2.9.1 *By Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v) and 2.3.3(i), we have*

$$\sup_{\theta\in\mathcal{N}_{\delta_n}} n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta\theta,n_2}(Z_i,\theta)\right\|^2=O_p(1).$$

Proof of Lemma 2.9.1. By definition,

$$\widehat{\phi}_{\theta\theta,n_2}(z,\theta) = n_2^{-1}P_{k_2}(z)'Q_{n_2,k_2}^{-1}\sum_{i\in I}P_{k_2}(Z_i)g_{\theta\theta}(X_i,\theta).$$

Let $g_{\theta_{j_1}\theta_{j_2}}(X_i,\theta)$ denote the (j_1,j_2)-th component of $g_{\theta\theta}(X_i,\theta)$, for any $j_1 = 1,\ldots,d_\theta$ and any $j_2 = 1,\ldots,d_\theta$. Let

$$\widehat{\phi}_{\theta_{j_1}\theta_{j_2},n_2}(z,\theta) = n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}P_{n_2,k_2}'g_{\theta_{j_1}\theta_{j_2},n_2}(\theta),$$

where $g_{\theta_{j_1}\theta_{j_2},n_2}(\theta) = (g_{\theta_{j_1}\theta_{j_2}}(X_i,\theta))_{i\in I_2}'$. Then by definition,

$$\begin{aligned}
n^{-1}\sum_{i\in I}\widehat{\phi}^2_{\theta_{j_1}\theta_{j_2},n_2}(Z_i,\theta) &= g_{\theta_{j_1}\theta_{j_2},n_2}(\theta)'P_{n_2,k_2}Q_{n_2,k_2}^{-1}Q_{n,k_2}Q_{n_2,k_2}^{-1}P_{n_2,k_2}'g_{\theta_{j_1}\theta_{j_2},n_2}(\theta)\\
&\leq \frac{\lambda_{\max}(Q_{n,k_2})}{\lambda_{\min}(Q_{n_1,k_2})}\frac{g_{\theta_{j_1}\theta_{j_2},n_2}(\theta)'P_{n_2,k_2}(P_{n_2,k_2}'P_{n_2,k_2})^{-1}P_{n_2,k_2}'g_{\theta_{j_1}\theta_{j_2},n_2}(\theta)}{n_2}\\
&\leq \frac{\lambda_{\max}(Q_{n,k_2})}{\lambda_{\min}(Q_{n_1,k_2})}n_2^{-1}\sum_{i\in I_2}(g_{\theta_{j_1}\theta_{j_2}}(X_i,\theta))^2
\end{aligned}$$

which together with (2.75) (which holds under Assumptions 2.3.1(i), 2.3.1(iii) and 2.3.1(v)), and Assumption 2.3.3(i) implies that

$$\sup_{\theta\in\mathcal{N}_{\delta_n}} n^{-1}\sum_{i\in I}\widehat{\phi}^2_{\theta_{j_1}\theta_{j_2},n_2}(Z_i,\theta) \leq \frac{\lambda_{\max}(Q_{n,k_2})}{\lambda_{\min}(Q_{n_1,k_2})}\sup_{\theta\in\mathcal{N}_{\delta_n}} n_2^{-1}\sum_{i\in I_2}(g_{\theta_{j_1}\theta_{j_2}}(X_i,\theta))^2 = O_p(1)$$

for any $j_1 = 1,\ldots,d_\theta$ and any $j_2 = 1,\ldots,d_\theta$. This finishes the proof. ■

Lemma 2.9.2 *By Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(iv), 2.3.1(v), 2.3.3(v) and 2.3.3(vii), we have*

$$n^{-1}\sum_{i\in I}(\widehat{h}_{n_1}(Z_i) - \widehat{\phi}_{n_2}(Z_i,\theta_0))^2 = o_p(n_1^{-1/2}+n_2^{-1/2}).$$

Proof of Lemma 2.9.2. By (2.77) (which holds under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(iv) and 2.3.1(v)),

$$\begin{aligned}
&n^{-1}\sum_{i\in I}(\widehat{h}_{n_1}(Z_i) - \widehat{\phi}_{n_2}(Z_i,\theta_0))^2\\
&\leq 2n^{-1}\sum_{i\in I}(\widehat{h}_{n_1}(Z_i) - h_0(Z_i))^2 + 2n^{-1}\sum_{i\in I}(\widehat{\phi}_{n_2}(Z_i,\theta_0) - h_0(Z_i))^2\\
&= 2n^{-1}\sum_{i\in I}(\widehat{\phi}_{n_2}(Z_i,\theta_0) - h_0(Z_i))^2 + O_p(k_1n_1^{-1}+k_1^{-2r_h}).
\end{aligned} \tag{2.89}$$

Let $\widehat{\beta}_{\phi,n_2} = (P_{n_2,k_2}'P_{n_2,k_2})^{-1}P_{n_2,k_2}'g_{n_2}(\theta_0)$, where $g_{n_2}(\theta_0) = (g(X_i,\theta_0))_{i\in I_2}'$. Then

$$\begin{aligned}
\left\|\widehat{\beta}_{\phi,n_2} - \beta_{h,k_2}\right\|^2 &= (g_{n_2}(\theta_0) - H_{n_2,k_2})'P_{n_2,k_2}(P_{n_2,k_2}'P_{n_2,k_2})^{-2}P_{n_2,k_2}'(g_{n_2}(\theta_0) - H_{n_2,k_2})\\
&\leq \frac{(g_{n_2}(\theta_0) - H_{n_2,k_2})'P_{n_2,k_2}(P_{n_2,k_2}'P_{n_2,k_2})^{-1}P_{n_2,k_2}'(g_{n_2}(\theta_0) - H_{n_2,k_2})}{n_2\lambda_{\min}(Q_{n_2,k_2})},
\end{aligned} \tag{2.90}$$

where $H_{n_2,k_2} = (h_{0,k_2}(Z_i))'_{i\in I_2}$. By Assumptions 2.3.1(iv),

$$\begin{aligned} & n_2^{-1}(H_{n_2} - H_{n_2,k_2})'P_{n_2,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}P'_{n_2,k_2}(H_{n_2} - H_{n_2,k_2}) \\ & \leq n_2^{-1}(H_{n_2} - H_{n_2,k_2})'(H_{n_2} - H_{n_2,k_2}) = O(k_2^{-2r_h}), \end{aligned} \tag{2.91}$$

where $H_{n_2} = (h_0(Z_i))'_{i\in I_2}$. By Assumptions 2.3.1(i), 2.3.1(iii) and 2.3.3(v),

$$\begin{aligned} & E\left[n_2^{-1}(g_{n_2}(\theta_0) - H_{n_2})'P_{n_2,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}P'_{n_2,k_2}(g_{n_2}(\theta_0) - H_{n_2}) \middle| \{Z_i\}_{i\in I_2}\right] \\ & = n_2^{-1} tr\left((P'_{n_2,k_2}P_{n_2,k_2})^{-1}P'_{n_2,k_2}E\left[(g_{n_2}(\theta_0) - H_{n_2})(g_{n_2}(\theta_0) - H_{n_2})' \middle| \{Z_i\}_{i\in I_2}\right]P_{n_2,k_2}\right) \\ & \leq \sup_{z\in\mathcal{Z}} \sigma_\varepsilon^2(z)k_2 n_2^{-1} = O(k_2 n_2^{-1}) \end{aligned}$$

which together with the Markov inequality implies that

$$n_2^{-1}(g_{n_2}(\theta_0) - H_{n_2})'P_{n_2,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}P'_{n_2,k_2}(g_{n_2}(\theta_0) - H_{n_2}) = O_p(k_2 n_2^{-1}). \tag{2.92}$$

Combining the results in (2.90), (2.91) and (2.92), and then applying (2.75), we get

$$\left\|\widehat{\beta}_{\phi,n_2} - \beta_{h,k_2}\right\|^2 = O_p(k_2 n_2^{-1} + k_2^{-2r_h}). \tag{2.93}$$

By (2.75) and (2.93)

$$\begin{aligned} & n^{-1}\sum_{i\in I}(\widehat{\phi}_{n_2}(Z_i,\theta_0) - h_{0,k_2}(Z_i))^2 \\ & = (\widehat{\beta}_{\phi,n_2} - \beta_{h,k_2})'Q_{n,k_2}(\widehat{\beta}_{\phi,n_2} - \beta_{h,k_2}) \\ & \leq \lambda_{\max}(Q_{n,k_2})\left\|\widehat{\beta}_{\phi,n_2} - \beta_{h,k_2}\right\|^2 = O_p(k_2 n_2^{-1} + k_2^{-2r_h}) \end{aligned} \tag{2.94}$$

which together with (2.89) and Assumption 2.3.3(vii) proves the claim of the lemma. ■

Lemma 2.9.3 *Under Assumptions 2.3.1(i), 2.3.2(v), 2.3.3(iv) and 2.3.3(vi), we have*

$$n^{-1}\phi_{w\theta,n}P_{n,k_1}(P'_{n,k_1}P_{n,k_1})^{-1}P'_{n,k_1}\phi'_{w\theta,n} = E\left[w_n^2(Z)\phi_\theta(Z,\theta_0)\phi'_\theta(Z,\theta_0)\right] + o_p(1).$$

Proof of Lemma 2.9.3. For $j = 1,\ldots,d_\theta$, let $\phi_{w\theta_j,k_1,n}(z,\theta_0) = P'_k(z)\beta_{w\phi_j,k_1}$, $\phi_{w\theta_j,k_1,n} = (\phi_{w\theta_j,k_1,n}(Z_i,\theta_0))_{i\in I}$ and $\phi_{w\theta,k_1,n} = (\phi'_{w\theta_j,k_1,n})'_{j=1,\ldots,d_\theta}$. For ease of notations, we define $M_{k_1,n} = P_{n,k_1}(P'_{n,k_1}P_{n,k_1})^{-1}P'_{n,k_1}$.

By definition,

$$\begin{aligned} \phi_{w\theta,n}M_{k_1,n}\phi'_{w\theta,n} &= \phi_{w\theta,k_1,n}M_{k_1,n}\phi'_{w\theta,k_1,n} + (\phi_{w\theta,n} - \phi_{w\theta,k_1,n})\, M_{k_1,n}\,(\phi_{w\theta,n} - \phi_{w\theta,k_1,n})' \\ &\quad + (\phi_{w\theta,n} - \phi_{w\theta,k_1,n})\, M_{k_1,n}\phi'_{w\theta,k_1,n} + \phi_{w\theta,k_1,n}M_{k_1,n}\,(\phi_{w\theta,n} - \phi_{w\theta,k_1,n})'. \end{aligned} \tag{2.95}$$

For any $j = 1, \ldots, d_\theta$, let $\phi_{w\theta_j,n}$ denote the j-th row of $\phi_{w\theta,n}$. By the Cauchy-Schwarz inequality, for any $j_1 = 1, \ldots, d_\theta$ and any $j_2 = 1, \ldots, d_\theta$,

$$\begin{aligned}
&\left| n^{-1}(\phi_{w\theta_{j_1},n} - \phi_{w\theta_j,k_1,n}) M_{k_1,n} (\phi_{w\theta_{j_2},n} - \phi_{w\theta_{j_2},k_1,n})' \right|^2 \\
&\leq n^{-1}(\phi_{w\theta_{j_1},n} - \phi_{w\theta_j,k_1,n}) M_{k_1,n} (\phi_{w\theta_{j_1},n} - \phi_{w\theta_j,k_1,n})' \\
&\times n^{-1}(\phi_{w\theta_{j_2},n} - \phi_{w\theta_{j_2},k_1,n}) M_{k_1,n} (\phi_{w\theta_{j_2},n} - \phi_{w\theta_{j_2},k_1,n})' \\
&\leq n^{-1} \sum_{i \in I} \left| w_n(Z_i)\phi_{\theta_{j_1}}(Z_i, \theta_0) - \phi_{w\theta_{j_1},k_1}(Z_i, \theta_0) \right|^2 \\
&\times n^{-1} \sum_{i \in I} \left| w_n(Z_i)\phi_{\theta_{j_2}}(Z_i, \theta_0) - \phi_{w\theta_{j_2},k_1}(Z_i, \theta_0) \right|^2 = o(1)
\end{aligned} \tag{2.96}$$

where the last equality is by Assumption 2.3.3(vi), and the fact that $M_{k_1,n}$ is an idempotent matrix. (2.96) then implies that

$$n^{-1}\left(\phi_{w\theta,n} - \phi_{w\theta,k_1,n}\right) M_{k_1,n} \left(\phi_{w\theta,n} - \phi_{w\theta,k_1,n}\right) = o(1). \tag{2.97}$$

For any $j_1 = 1, \ldots, d_\theta$ and any $j_2 = 1, \ldots, d_\theta$, by definition we can write

$$\begin{aligned}
&n^{-1}\phi_{w\theta_{j_1},k_1,n} M_{k_1,n} \phi'_{w\theta_{j_2},k_1,n} \\
&= n^{-1}\beta'_{w\phi_{j_1},k} P'_{n,k_1} P_{n,k_1} (P'_{n,k_1} P_{n,k_1})^{-1} P'_{n,k_1} \phi'_{w\theta_{j_2},k_1,n} \\
&= n^{-1} \sum_{i \in I} \phi_{w\theta_{j_1},k_1,n}(Z_i, \theta_0) \phi_{w\theta_{j_2},k_1,n}(Z_i, \theta_0) \\
&= n^{-1} \sum_{i \in I} w_n^2(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0) \phi_{\theta_{j_2}}(Z_i, \theta_0) \\
&+ n^{-1} \sum_{i \in I} (\phi_{w\theta_{j_1},k_1,n}(Z_i, \theta_0) - w_n(Z_i)\phi_{\theta_{j_1}}(Z_i, \theta_0)) \phi_{w\theta_{j_2},k_1,n}(Z_i, \theta_0) \\
&+ n^{-1} \sum_{i \in I} w_n(Z_i)\phi_{\theta_{j_1}}(Z_i, \theta_0)(\phi_{w\theta_{j_2},k_1}(Z_i, \theta_0) - w_n(Z_i)\phi_{\theta_{j_2}}(Z_i, \theta_0)).
\end{aligned} \tag{2.98}$$

By Assumptions 2.3.1(i), 2.3.2(v), 2.3.3(iv) and the Markov inequality, we have

$$n^{-1} \sum_{i \in I} w_n^2(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0) \phi_{\theta_{j_2}}(Z_i, \theta_0) - E\left[w_n^2(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0) \phi_{\theta_{j_2}}(Z_i, \theta_0) \right] = O_p(n^{-1/2}), \tag{2.99}$$

where under Assumptions 2.3.2(v) and 2.3.3(iv)

$$\left| E\left[w_n^2(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0) \phi_{\theta_{j_2}}(Z_i, \theta_0) \right] \right| < C. \tag{2.100}$$

By Assumption 2.3.3(vi),

$$\begin{aligned}
&\left| n^{-1} \sum_{i \in I} (\phi_{w\theta_{j_1},k_1,n}(Z_i, \theta_0) - w_n(Z_i)\phi_{\theta_{j_1}}(Z_i, \theta_0))(\phi_{w\theta_{j_2},k_1,n}(Z_i, \theta_0) - w_n(Z_i)\phi_{\theta_{j_2}}(Z_i, \theta_0)) \right| \\
&\leq \left(\max_{j=1,\ldots,d_\theta} \sup_{z \in \mathcal{Z}} |\phi_{w\theta_j,k_1,n}(z, \theta_0) - w_n(z)\phi_{\theta_j}(z, \theta_0)| \right)^2 = o(1),
\end{aligned} \tag{2.101}$$

which implies that

$$
\begin{aligned}
&n^{-1}\sum_{i\in I}(\phi_{w\theta_j,k_1,n}(Z_i,\theta_0)-w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0))\phi_{w\theta_{j_2},k_1,n}(Z_i,\theta_0)\\
&=n^{-1}\sum_{i\in I}(\phi_{w\theta_j,k_1,n}(Z_i,\theta_0)-w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0))w_n(Z_i)\phi_{\theta_{j_2}}(Z_i,\theta_0)+o_p(1).
\end{aligned}
\tag{2.102}
$$

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\left|n^{-1}\sum_{i\in I}(\phi_{w\theta_j,k_1,n}(Z_i,\theta_0)-w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0))w_n(Z_i)\phi_{\theta_{j_2}}(Z_i,\theta_0)\right|^2\\
&\leq n^{-1}\sum_{i\in I}\left|\phi_{w\theta_j,k_1,n}(Z_i,\theta_0)-w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\right|^2 n^{-1}\sum_{i\in I}w_n^2(Z_i)\phi^2_{\theta_{j_2}}(Z_i,\theta_0)=o_p(1)
\end{aligned}
\tag{2.103}
$$

where the equality is by Assumption 2.3.3(vi), (2.99) and (2.100). Combining the results in (2.102) and (2.103), we get

$$
n^{-1}\sum_{i\in I}(\phi_{w\theta_j,k_1,n}(Z_i,\theta_0)-w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0))\phi_{w\theta_{j_2},k_1,n}(Z_i,\theta_0)=o_p(1). \tag{2.104}
$$

Similarly, we can show that

$$
n^{-1}\sum_{i\in I}w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)(\phi_{w\theta_{j_2},k_1,n}(Z_i,\theta_0)-w_n(Z_i)\phi_{\theta_{j_2}}(Z_i,\theta_0))=o_p(1). \tag{2.105}
$$

Collecting the results in (2.98), (2.99), (2.104) and (2.105), we have

$$
n^{-1}\phi_{w\theta_{j_1},k_1,n}M_{k_1,n}\phi'_{w\theta_{j_2},k_1,n}=E\left[w_n^2(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)\right]+o_p(1) \tag{2.106}
$$

for any $j_1=1,\ldots,d_\theta$ and any $j_2=1,\ldots,d_\theta$, which implies that

$$
n^{-1}\phi_{w\theta,k_1,n}M_{k_1,n}\phi'_{w\theta,k_1,n}=E\left[w_n^2(Z_i)\phi_\theta(Z_i,\theta_0)\phi'_\theta(Z_i,\theta_0)\right]+o_p(1). \tag{2.107}
$$

By the Cauchy-Schwarz inequality, for any $j_1=1,\ldots,d_\theta$ and any $j_2=1,\ldots,d_\theta$,

$$
\begin{aligned}
&\left|\frac{(\phi_{w\theta_{j_1},n}-\phi_{w\theta_{j_1},k_1,n})M_{k_1,n}\phi'_{w\theta_{j_2},k_1,n}}{n}\right|^2\\
&\leq\frac{(\phi_{w\theta_{j_1},n}-\phi_{w\theta_{j_1},k_1,n})M_{k_1,n}(\phi_{w\theta_{j_1},n}-\phi_{w\theta_{j_1},k_1,n})'}{n}\frac{\phi_{w\theta_{j_2},k_1,n}M_{k_1,n}\phi'_{w\theta_{j_2},k_1,n}}{n}=o_p(1)
\end{aligned}
\tag{2.108}
$$

where the equality is by (2.96), (2.100) and (2.106). (2.108) then implies that

$$
n^{-1}\left(\phi_{w\theta,n}-\phi_{w\theta,k_1,n}\right)M_{k_1,n}\phi'_{w\theta,k_1,n}=o_p(1) \tag{2.109}
$$

and similarly

$$
n^{-1}\phi_{w\theta,k_1,n}M_{k_1,n}\left(\phi_{w\theta,n}-\phi_{w\theta,k_1,n}\right)'=o_p(1). \tag{2.110}
$$

Combining the results in (2.95), (2.97), (2.107), (2.109) and (2.110), we immediately get the claimed result. ∎

Proof of Theorem 2.3.2. By the definition of $\widehat{\theta}_n$, we have the following first order condition

$$n^{-1}\sum_{i\in I} w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_n))\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)=0. \tag{2.111}$$

Applying the first order expansion to (2.111), we get

$$\begin{aligned}
0 &= n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&\quad - n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)'(\widehat{\theta}_n-\theta_0)\\
&\quad + n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_n))\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_n)(\widehat{\theta}_n-\theta_0),
\end{aligned} \tag{2.112}$$

where $\widetilde{\theta}_n$ is between $\widehat{\theta}_n$ and θ_0 and it may differ across rows.

For any $j=1,\ldots,d_\theta$, by the mean value expansion and the Cauchy-Schwarz inequality,

$$\left|\widehat{\phi}_{\theta_j,n_2}(Z_i,\widetilde{\theta}_{j,n})-\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0)\right| \le \sup_{\theta\in\mathcal{N}_{\delta_n}}\left\|\widehat{\phi}_{\theta_j\theta,n_2}(Z_i,\theta)\right\|\left\|\widetilde{\theta}_{j,n}-\theta_0\right\|$$

which together with the triangle inequality and Lemma 2.9.1 implies that

$$n^{-1}\sum_{i\in I}(\widehat{\phi}_{\theta_j,n_2}(Z_i,\widetilde{\theta}_{j,n})-\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0))^2 \le \sup_{\theta\in\mathcal{N}_{\delta_n}} n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta_j\theta,n_2}(Z_i,\theta)\right\|^2\left\|\widetilde{\theta}_{j,n}-\theta_0\right\|^2 = o_p(1). \tag{2.113}$$

By Assumption 2.3.3(iii) and (2.113),

$$\begin{aligned}
&n^{-1}\sum_{i\in I}(\widehat{\phi}_{\theta_j,n_2}(Z_i,\widetilde{\theta}_{j,n})-\phi_{\theta_j}(Z_i,\theta_0))^2\\
&\le 2n^{-1}\sum_{i\in I}(\widehat{\phi}_{\theta_j,n_2}(Z_i\widetilde{\theta}_{j,n})-\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0))^2+2n^{-1}\sum_{i\in I}(\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0)-\phi_{\theta_j}(Z_i,\theta_0))^2=o_p(1).
\end{aligned} \tag{2.114}$$

By Assumption 2.3.3(iv) and the Markov inequality,

$$n^{-1}\sum_{i\in I}(\phi_{\theta_j}(Z_i,\theta_0))^2=O_p(1), \tag{2.115}$$

which together with (2.114) implies that

$$n^{-1}\sum_{i\in I}(\widehat{\phi}_{\theta_j,n_2}(Z_i,\widetilde{\theta}_n))^2=O_p(1) \tag{2.116}$$

for any $j=1,\ldots,d_\theta$. For any $j_1=1,\ldots,d_\theta$ and any $j_2=1,\ldots,d_\theta$, we can use the triangle inequality and the Cauchy-Schwarz inequality, Assumptions 2.3.2(v), (2.114), (2.115) and (2.116) to deduce that

$$\begin{aligned}
&\left|n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\widehat{\phi}_{\theta_{j_1},n_2}(Z_i,\widetilde{\theta}_{j_1,n})\widehat{\phi}_{\theta_{j_2},n_2}(Z_i,\widetilde{\theta}_{j_2,n})-n^{-1}\sum_{i\in I}w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)\right|\\
&\leq\left|n^{-1}\sum_{i\in I}(\widehat{w}_n(Z_i)-w_n(Z_i))\widehat{\phi}_{\theta_{j_1},n_2}(Z_i,\widetilde{\theta}_{j_1,n})\widehat{\phi}_{\theta_{j_2},n_2}(Z_i,\widetilde{\theta}_{j_2,n})\right|\\
&+\left|n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{\phi}_{\theta_{j_1},n_2}(Z_i,\widetilde{\theta}_{j_1,n})-\phi_{\theta_{j_1}}(Z_i,\theta_0))\widehat{\phi}_{\theta_{j_2},n_2}(Z_i,\theta_0)\right|\\
&+\left|n^{-1}\sum_{i\in I}w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)(\widehat{\phi}_{\theta_{j_2},n_2}(Z_i,\widetilde{\theta}_{j_2,n})-\phi_{\theta_{j_2}}(Z_i,\theta_0))\right|=o_p(1).
\end{aligned}\tag{2.117}$$

Under Assumptions 2.3.1(i), 2.3.2(v) and 2.3.3(iv), we can the Markov inequality to deduce that

$$n^{-1}\sum_{i\in I}w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)=E\left[w_n(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)\right]+o_p(1)\tag{2.118}$$

for any $j_1=1,\ldots,d_\theta$ and any $j_2=1,\ldots,d_\theta$. Collecting the results in (2.117) and (2.118), we get

$$n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)'=H_0+o_p(1).\tag{2.119}$$

By the second order Taylor expansion and the triangle inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&n^{-1}\sum_{i\in I}\left|\phi(Z_i,\widetilde{\theta}_{j,n})-\phi(Z_i,\theta_0)\right|^2\\
&\leq 2n^{-1}\sum_{i\in I}\|\phi_\theta(Z_i,\theta_0)\|^2\,||\widetilde{\theta}_{j,n}-\theta_0||^2\\
&+2^{-1}\sup_{\theta\in\mathcal{N}_{\delta_n}}n^{-1}\sum_{i\in I}\|\phi_{\theta\theta}(Z_i,\theta)\|^2\,||\widetilde{\theta}_{j,n}-\theta_0||^4=o_p(1)
\end{aligned}\tag{2.120}$$

where the equality is by (2.115), Lemma 2.9.1 and $||\widetilde{\theta}_{j,n}-\theta_0||=o_p(1)$ for any $j=1,\ldots,d_\theta$. (2.120) together with Assumptions 2.3.2(ii) then implies that

$$\begin{aligned}
&n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_{j,n})-\phi(Z_i,\theta_0)\right|^2\\
&\leq 2n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_{j,n})-\phi(Z_i,\widetilde{\theta}_n)\right|^2+2n^{-1}\sum_{i\in I}\left|\phi(Z_i,\widetilde{\theta}_{j,n})-\phi(Z_i,\theta_0)\right|^2=o_p(1)
\end{aligned}\tag{2.121}$$

for any $j=1,\ldots,d_\theta$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left\|n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_n))\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_n)\right\|\\
&\leq\sup_z|\widehat{w}_n(z)|\sqrt{\max_{j=1,\ldots,d_\theta}n^{-1}\sum_{i\in I}(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_{j,n}))^2}\sqrt{\max_{j=1,\ldots,d_\theta}n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_{j,n})\right\|^2}\\
&\leq 2\sup_z|\widehat{w}_n(z)|\sqrt{\max_{j=1,\ldots,d_\theta}n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_n)\right\|^2}\\
&\times\sqrt{n^{-1}\sum_{i\in I}\left|\widehat{h}_{n_1}(Z_i)-h_0(Z_i)\right|^2+\max_{j=1,\ldots,d_\theta}n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_{j,n})-\phi(Z_i,\theta_0)\right|^2}=o_p(1)
\end{aligned}\tag{2.122}$$

where the last equality is by (2.77), (2.78), Lemma 2.9.1 and (2.121).

By definition,

$$
\begin{aligned}
& n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&= n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\phi_\theta(Z_i,\theta_0)\\
&+ n^{-1}\sum_{i\in I}(\widehat{w}_n(Z_i)-w_n(Z_i))(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&+ n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))(\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)-\phi_\theta(Z_i,\theta_0)). \qquad (2.123)
\end{aligned}
$$

By Assumptions 2.3.3(iv) and 2.3.3(v), and the Markov inequality

$$
n^{-1}\sum_{i\in I}||\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)||^2 \leq 2n^{-1}\sum_{i\in I}||\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)-\phi_\theta(Z_i,\theta_0)||^2+2n^{-1}\sum_{i\in I}\|\phi_\theta(Z_i,\theta_0)\|^2 = O_p(1). \qquad (2.124)
$$

By the triangle inequality and the Cauchy-Schwarz inequality, (2.124), Lemma 2.9.2, Assumptions 2.3.2(v) and 2.3.3(vii),

$$
\begin{aligned}
&\left\|n^{-1}\sum_{i\in I}(\widehat{w}_n(Z_i)-w_n(Z_i))(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\right\|\\
&\leq \sup_{z\in\mathcal{Z}}|\widehat{w}_n(z)-w_n(z)|\sqrt{n^{-1}\sum_{i\in I}(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))^2}\sqrt{n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\right\|^2}\\
&= o_p(n_1^{-1/2}+n_2^{-1/2})\sqrt{n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\right\|^2} = o_p(n_1^{-1/2}+n_2^{-1/2}). \qquad (2.125)
\end{aligned}
$$

By the triangle inequality and the Cauchy-Schwarz inequality, Lemma 2.9.2, Assumptions 2.3.2(v), 2.3.3(iii) and 2.3.3(vii),

$$
\begin{aligned}
&\left\|n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))(\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)-\phi_\theta(Z_i,\theta_0))\right\|\\
&\leq \sup_{z\in\mathcal{Z}}|w_n(z)|\sqrt{n^{-1}\sum_{i\in I}(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))^2}\sqrt{n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)-\phi_\theta(Z_i,\theta_0)\right\|^2}\\
&= o_p(n_1^{-1/2}+n_2^{-1/2}). \qquad (2.126)
\end{aligned}
$$

Combining the results in (2.123), (2.125) and (2.126), we get

$$\begin{aligned}
& n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0) \\
&= n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\phi_\theta(Z_i,\theta_0)+o_p(n_1^{-1/2}+n_2^{-1/2}) \\
&= n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))\phi_\theta(Z_i,\theta_0) \\
&\quad - n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{\phi}_{n_2}(Z_i,\theta_0)-h_0(Z_i))\phi_\theta(Z_i,\theta_0)+o_p(n_1^{-1/2}+n_2^{-1/2}). \qquad (2.127)
\end{aligned}$$

By the definition of $\widehat{h}_{n_1}(Z_i)$, we can write

$$\begin{aligned}
& n^{-1}\sum_{i\in I}w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))\phi_\theta(Z_i,\theta_0) \\
&= \frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n_1,k_1}U_{n_1}}{n} \\
&+ \frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n_1,k_1}(H_{n_1}-H_{n_1,k_1})}{n}+\frac{\phi_{w\theta,n}(H_n-H_{n,k_1})}{n}. \qquad (2.128)
\end{aligned}$$

where $H_n=(h_0(Z_i))'_{i\in I}$, $H_{n_1}=(h_0(Z_i))'_{i\in I_1}$, $U_{n_1}=(u_i)'_{i\in I_1}$, $H_{n,k_1}=(h_{0,k_1}(Z_i))'_{i\in I}$, $H_{n_1,k_1}=(h_{0,k_1}(Z_i))'_{i\in I_1}$ and $\phi_{w\theta,n}=(w_n(Z_i)\phi_\theta(Z_i,\theta_0))_{i\in I}$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
& \frac{\left|\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n_1,k_1}(H_{n_1}-H_{n_1,k_1})\right|^2}{n^2} \\
&\leq \frac{\phi_{w\theta,n}P_{n,k_1}P'_{n,k_1}\phi'_{w\theta,n}}{n^2}(H_{n_1}-H_{n_1,k_1})'P_{n_1,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-2}P'_{n_1,k_1}(H_{n_1}-H_{n_1,k_1}) \\
&\leq \frac{\lambda_{\max}(Q_{n,k_1})}{\lambda_{\min}(Q_{n_1,k_1})}\frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n,k_1}P_{n,k_1})^{-1}P'_{n,k_1}\phi'_{w\theta,n}}{n} \\
&\times \frac{(H_{n_1}-H_{n_1,k_1})'P_{n_1,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n_1,k_1}(H_{n_1}-H_{n_1,k_1})}{n_1} \\
&\leq \frac{\sup_{z\in\mathcal{Z}}\left|w_n^2(z)\right|\lambda_{\max}(Q_{k_1,n})}{\lambda_{\min}(Q_{k_1,n_1})}n^{-1}\sum_{i\in I}\|\phi_\theta(Z_i,\theta_0)\|^2\times n_1^{-1}\sum_{i\in I_1}|h_{0,k_1}(Z_i)-h_0(Z_i)|^2=O_p(k_1^{-2r_h}),
\end{aligned}$$

(2.129)

where the last equality is by (2.75), (2.115), Assumptions 2.3.1(iv) and 2.3.2(v). By the triangle inequality,

$$\begin{aligned}
& \left\|n^{-1}\sum_{i\in I}w_n(Z_i)(h_{0,k_1}(Z_i)-h_0(Z_i))\phi_\theta(Z_i,\theta_0)\right\| \\
&\leq \sup_{z\in\mathcal{Z}}|w_n(z)|\,n^{-1}\sum_{i\in I}\|(h_{0,k_1}(Z_i)-h_0(Z_i))\phi_\theta(Z_i,\theta_0)\| \\
&\leq C\frac{\sup_z|h_0(z)-h_{k_1}(z)|}{n}\sum_{i\in I}\|\phi_\theta(Z_i,\theta_0)\|=O_p(k_1^{-r_h}), \qquad (2.130)
\end{aligned}$$

where the last equality is by (2.115), Assumptions 2.3.1(iv) and 2.3.2(v). Combining the results in (2.128), (2.129) and (2.130), we get

$$\begin{aligned} & n^{-1}\sum_{i\in I} w_n(Z_i)(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))\phi_\theta(Z_i,\theta_0) \\ & = \frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}}{n}\sum_{i\in I_1} u_i P_{k_1}(Z_i) + O_p(k_1^{-r_h}). \end{aligned} \tag{2.131}$$

By the definition of $\widehat{\phi}_{n_2}(Z_i,\theta_0)$, we can write

$$\begin{aligned} & n^{-1}\sum_{i\in I} w_n(Z_i)(\widehat{\phi}_{n_2}(Z_i,\theta_0)-h_0(Z_i))\phi_\theta(Z_i,\theta_0) \\ & = \frac{\phi_{w\theta,n}P_{n,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}}{n}\sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) \\ & + \frac{\phi_{w\theta,n}P_{n,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}(H_{n_2}-H_{n_2,k_2})}{n} + \frac{\phi_{w\theta,n}(H_n-H_{n,k_2})}{n} \end{aligned} \tag{2.132}$$

where $H_{n_2,k_2} = (h_{0,k_2}(Z_i))'_{i\in I_2}$ and $H_{n,k_2} = (h_{0,k_2}(Z_i))_{i\in I}$. Using similar arguments in showing (2.129) and (2.130), we get

$$\frac{\phi_{w\theta,n}P_{n,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}(H_{n_2}-H_{n_2,k_2})}{n} = O_p(k_2^{-r_h}) \text{ and } \frac{\phi_{w\theta,n}(H_n-H_{n,k_2})}{n} = O_p(k_2^{-r_h}),$$

which together with (2.132) implies that

$$\begin{aligned} & n^{-1}\sum_{i\in I} w_n(Z_i)(\widehat{\phi}_{n_2}(Z_i,\theta_0)-h_0(Z_i))\phi_\theta(Z_i,\theta_0) \\ & = \frac{\phi_{w\theta,n}P_{n,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}}{n}\sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) + O_p(k_2^{-r_h}). \end{aligned} \tag{2.133}$$

By Assumption 2.3.3(v), $C^{-1}Q_{n_1,k_1} \leq n_1\sum_{i\in I_1}\sigma_u^2(Z_i)P_{k_1}(Z_i)P_{k_1}(Z_i)' \leq CQ_{n_1,k_1}$, which together with (2.75) implies that

$$C^{-1} < \lambda_{\min}(Q_{n_1,u}) \leq \lambda_{\max}(Q_{n_1,u}) < C, \tag{2.134}$$

with probability approaching 1. Similarly, we can show that

$$C^{-1} \leq \lambda_{\min}(Q_{n_2,\varepsilon}) \leq \lambda_{\max}(Q_{n_2,\varepsilon}) \leq C \tag{2.135}$$

with probability approaching 1.

Under the i.i.d. assumption,

$$\begin{aligned}
&E\left[\left|\left|\frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}}{n}\sum_{i\in I_1}u_iP_{k_1}(Z_i)\right|\right|^2\Big|\{Z_i\}_{i\in I}\right]\\
&=\frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}Q_{n_1,u}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n,k_1}\phi'_{w\theta,n}}{n^2n_1^{-1}}\\
&\leq\frac{C\lambda_{\max}(Q_{n_1,u})}{\lambda^2_{\min}(Q_{k_1,n_1})}\frac{\phi_{w\theta,n}P_{n,k_1}P'_{n,k_1}\phi'_{w\theta,n}}{n^2n_1}\\
&\leq\frac{\lambda_{\max}(Q_{n_1,u})\lambda_{\max}(Q_{n,k_1})}{n_1\lambda^2_{\min}(Q_{n_1,k_1})}\frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n,k_1}P_{n,k_1})^{-1}P'_{n,k_1}\phi'_{w\theta,n}}{n}\\
&\leq\sup_{z\in\mathcal{Z}}\left|w_n^2(z)\right|\frac{\lambda_{\max}(Q_{n_1,u})\lambda_{\max}(Q_{n,k_1})}{n_1\lambda^2_{\min}(Q_{n_1,k_1})}n^{-1}\sum_{i\in I}\|\phi_\theta(Z_i,\theta_0)\|^2=O_p(n_1^{-1}),
\end{aligned}\tag{2.136}$$

where the last equality is by (2.134), (2.75), (2.115), Assumptions 2.3.2(v) and 2.3.3(iv). Combined with the Markov inequality, (2.136) implies that

$$\frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}}{n}\sum_{i\in I_1}u_iP_{k_1}(Z_i)=O_p(n_1^{-1/2}).\tag{2.137}$$

Similarly, we can show that

$$\frac{\phi_{w\theta,n}P_{n,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}}{n}\sum_{i\in I_2}\varepsilon_iP_{k_2}(Z_i)=O_p(n_2^{-1/2}).\tag{2.138}$$

By (2.119) and (2.122),

$$n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{\phi}_{\theta,n_2}(Z_i,\widetilde{\theta}_n)\widehat{\phi}'_{\theta,n_2}(Z_i,\widetilde{\theta}_n)+(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\widetilde{\theta}_n))\widehat{\phi}_{\theta\theta,n_2}(Z_i,\widetilde{\theta}_n))=H_{0,n}+o_p(1),\tag{2.139}$$

which together with (2.112) implies that

$$[H_{0,n}+o_p(1)]\,(\widehat{\theta}_n-\theta_0)=-n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0).\tag{2.140}$$

By (2.127), (2.131) and (2.133), and Assumption 2.3.3(vii),

$$\begin{aligned}
&n^{-1}\sum_{i\in I}\widehat{w}_n(Z_i)(\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0))\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)\\
&=\frac{\phi_{w\theta,n}P_{n,k_1}(P'_{n_1,k_1}P_{n_1,k_1})^{-1}}{n}\sum_{i\in I_1}u_iP_{k_1}(Z_i)\\
&\quad-\frac{\phi_{w\theta,n}P_{n,k_2}(P'_{n_2,k_2}P_{n_2,k_2})^{-1}}{n}\sum_{i\in I_2}\varepsilon_iP_{k_2}(Z_i)+o_p(n_1^{-1/2}+n_2^{-1/2})
\end{aligned}\tag{2.141}$$

which together with (2.137) and (2.138) implies that

$$\frac{1}{n}\sum_{i\in I}\left[\widehat{h}_{n_1}(Z_i)-\widehat{\phi}_{n_2}(Z_i,\theta_0)\right]\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)=O_p(n_1^{-1/2}+n_2^{-1/2}).\tag{2.142}$$

Using (2.140), (2.141) and (2.142), and then applying Assumption 2.3.3(ii), we get

$$(\widehat{\theta}_n-\theta_0)=\frac{-H_0^{-1}\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}}{nn_1}\sum_{i\in I_1}u_iP_{k_1}(Z_i)$$
$$+\frac{H_0^{-1}\phi_{w\theta,n}P_{n,k_2}Q_{n_2,k_2}^{-1}}{nn_2}\sum_{i\in I_2}\varepsilon_iP_{k_2}(Z_i)+o_p(n_1^{-1/2}+n_2^{-1/2}), \tag{2.143}$$

which together with Assumption 2.3.3(ii), (2.137) and (2.138) implies that $\widehat{\theta}_n-\theta_0=O_p(n_1^{-1/2}+n_2^{-1/2})$.

By Assumption 2.3.3(v), $Q_{n_1,u}\geq C^{-1}Q_{n_1,k_1}$ which implies that

$$\begin{aligned}
n_1\Sigma_{n_1}&=\frac{\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}Q_{n_1,u}Q_{n_1,k_1}^{-1}P_{n,k_1}'\phi_{w\theta,n}'}{n^2}\\
&\geq\frac{\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}P_{n,k_1}'\phi_{w\theta,n}}{Cn^2n_1^{-1}}\\
&\geq\frac{\lambda_{\min}(Q_{n,k_1})}{C\lambda_{\max}(Q_{n_1,k_1})}\frac{\phi_{w\theta,n}P_{n,k_1}(P_{n,k_1}'P_{n,k_1})^{-1}P_{n,k_1}'\phi_{w\theta,n}}{n}\\
&\geq\frac{\lambda_{\min}(Q_{n,k_1})}{C\lambda_{\max}(Q_{n_1,k_1})}H_{0,n}+o_p(1)
\end{aligned} \tag{2.144}$$

where the last equality is by (2.75), Assumption 2.3.2(v) and Lemma 2.9.3. Using (2.75), (2.144) and Assumption 2.3.3(ii), we have

$$\lambda_{\min}(n_1\Sigma_{n_1})\geq C^{-1} \tag{2.145}$$

with probability approaching 1. Similarly, we can show that

$$\lambda_{\min}(n_2\Sigma_{n_2})\geq C^{-1} \tag{2.146}$$

with probability approaching 1. By (2.145),

$$\lambda_{\max}\left((n_1^{-1}+n_2^{-1})(\Sigma_{n_1}+\Sigma_{n_2})^{-1}\right)\leq\lambda_{\min}^{-1}(n_1\Sigma_{n_1})+\lambda_{\min}^{-1}(n_2\Sigma_{n_2})\leq 2C \tag{2.147}$$

with probability approaching 1. Combining the results in (2.143), (2.147) and $\lambda_{\min}(H_0)>0$ in Assumption 2.3.3(ii), we get

$$\begin{aligned}
&(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}(\widehat{\theta}_n-\theta_0)\\
&=\frac{-(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}H_{0.n}^{-1}\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}}{nn_1}\sum_{i\in I_1}u_iP_{k_1}(Z_i)\\
&+\frac{(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}H_{0.n}^{-1}\phi_{w\theta,n}P_{n,k_2}Q_{n_2,k_2}^{-1}}{nn_2}\sum_{i\in I_2}\varepsilon_iP_{k_2}(Z_i)+o_p(n_1^{-1/2}+n_2^{-1/2}).
\end{aligned} \tag{2.148}$$

Define

$$\omega_{i,n}=\left\{\begin{array}{ll}\frac{-(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}H_{0.n}^{-1}\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}P_{k_1}(Z_i)u_i}{nn_1}, & 1\leq i\leq n_1\\ \frac{(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}H_{0.n}^{-1}\phi_{w\theta,n}P_{n,k_2}Q_{n_2,k_2}^{-1}P_{k_2}(Z_i)\varepsilon_i}{nn_2}, & n_1<i\leq n\end{array}\right. .$$

Then by (2.143), we can write

$$(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}(\widehat{\theta}_n-\theta_0)=\sum_{i=1}^n \omega_{i,n}+o_p(n_1^{-1/2}+n_2^{-1/2}). \tag{2.149}$$

Let $\mathcal{F}_{i,n}$ be the sigma field generated by $\{\omega_{1,n},\ldots,\omega_{i,n},\{Z_i\}_{i\in I}\}$ for $i=1,\ldots,n$. Then under Assumption 2.3.1(i), $E\left[\gamma_n'\omega_{i,n}\middle|\mathcal{F}_{i-1,n}\right]=0$ which means that $\{\gamma_n'\omega_{i,n}\}_{i=1}^n$ is a martingale difference array. We next use the Martingale CLT to show the claim. There are two sufficient conditions to verify:

$$\sum_{i=1}^n E\left[(\gamma_n'\omega_{i,n})^2\middle|\mathcal{F}_{i,n}\right]\to_p 1;\text{ and} \tag{2.150}$$

$$\sum_{i=1}^n E\left[(\gamma_n'\omega_{i,n})^2 I\left\{\left|\gamma_n'\omega_{i,n}\right|>\varepsilon\right\}\middle|\mathcal{F}_{i,n}\right]\to_p 0 \text{ for } \forall\varepsilon>0. \tag{2.151}$$

For ease of notations, we define $D_n=(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}H_{0,n}^{-1}$. By definition, we have

$$\begin{aligned}\sum_{i=1}^n E\left[(\gamma_n'\omega_{i,n})^2\middle|\mathcal{F}_{i,n}\right]&=\sum_{i=1}^n\gamma_n' E\left[\omega_{i,n}\omega_{i,n}'\middle|\mathcal{F}_{i,n}\right]\gamma_n\\&=\gamma_n'\frac{D_n\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}Q_{n_1,u}Q_{n_1,k_1}^{-1}P_{n,k_1}'\phi_{w\theta,n}'D_n'}{n^2n_1}\gamma_n\\&+\gamma_n'\frac{D_n\phi_{w\theta,n}P_{n,k_2}Q_{n_2,k_2}^{-1}Q_{n_2,\varepsilon}Q_{n_2,k_2}^{-1}P_{n,k_2}'\phi_{w\theta,n}'D_n'}{n^2n_2}\gamma_n\\&=\gamma_n'D_n(\Sigma_{n_1}+\Sigma_{n_2})D_n'\gamma_n=\gamma_n'\gamma_n=1\end{aligned} \tag{2.152}$$

which proves (2.150). By the monotonicity of expectation,

$$\begin{aligned}&\sum_{i=1}^n E\left[(\gamma_n'\omega_{i,n})^2 I\left\{\left|\gamma_n'\omega_{i,n}\right|>\varepsilon\right\}\middle|\mathcal{F}_{i,n}\right]\\&\leq\frac{1}{\varepsilon^2}\sum_{i=1}^n E\left[(\gamma_n'\omega_{i,n})^4\middle|\mathcal{F}_{i,n}\right]\\&=\frac{1}{\varepsilon^2}\sum_{i\in I_1}E\left[\frac{\left|\gamma_n'D_n\phi_{w\theta,n}P_{n,k_1}Q_{n_1,k_1}^{-1}P_{k_1}(Z_i)u_i\right|^4}{n^4n_1^4}\middle|\mathcal{F}_{i,n}\right]\\&+\frac{1}{\varepsilon^2}\sum_{i\in I_2}E\left[\frac{\left|\gamma_n'D_n\phi_{w\theta,n}P_{n,k_2}Q_{n_2,k_2}^{-1}P_{k_2}(Z_i)\varepsilon_i\right|^4}{n^4n_2^4}\middle|\mathcal{F}_{i,n}\right].\end{aligned} \tag{2.153}$$

By Assumptions 2.3.2(v) and 2.3.3(iv),

$$H_{0,n}=E\left[w_n(Z_i)\phi_\theta(Z_i,\theta_0)\phi_\theta'(Z_i,\theta_0)\right]\leq C. \tag{2.154}$$

By (2.99) in the proof of Lemma 2.9.3,

$$n^{-1}\phi_{w\theta,n}\phi_{w\theta,n}'=E\left[w_n^2(Z_i)\phi_\theta(Z_i,\theta_0)\phi_\theta'(Z_i,\theta_0)\right]+o_p(1), \tag{2.155}$$

which together with (2.154), Assumptions 2.3.2(v) and 2.3.3(ii) implies that

$$C^{-1} < \lambda_{\max}(n^{-1}\phi_{w\theta,n}\phi'_{w\theta,n}) < C \tag{2.156}$$

with probability approaching 1. By (2.145),

$$\lambda_{\min}(n_1(\Sigma_{n_1}+\Sigma_{n_2})) \geq \lambda_{\min}(n_1\Sigma_{n_1}) > C^{-1}. \tag{2.157}$$

For any $\gamma \in R^{d_\theta}$, we have

$$\begin{aligned}
\frac{\gamma' D_n D_n' \gamma}{n_1} &= \frac{\gamma'(H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n})^{1/2}H_{0,n}^{-2}(H_0(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_0)^{1/2}\gamma}{n_1} \\
&\leq \frac{1}{\lambda^2_{\min}(H_{0,n})}\frac{\gamma' H_{0,n}(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_{0,n}\gamma}{n_1} \\
&\leq \frac{\gamma' H_{0,n}^2\gamma}{\lambda^2_{\min}(H_{0,n})\lambda_{\min}(n_1(\Sigma_{n_1}+\Sigma_{n_2}))} \\
&\leq \frac{\lambda^2_{\max}(H_{0,n})}{\lambda^2_{\min}(H_{0,n})\lambda_{\min}(n_1(\Sigma_{n_1}+\Sigma_{n_2}))},
\end{aligned} \tag{2.158}$$

which combined with Assumption 2.3.3(ii), (2.154) and (2.157) implies that

$$\lambda_{\max}(n_1^{-1}D_n D_n') \leq C \tag{2.159}$$

with probability approaching 1. By Assumption 2.6.1(v) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\frac{1}{\varepsilon^2}\sum_{i\in I_1} E\left[\left|\frac{\left|\gamma_n' D_n\phi_{w\theta,n}P_{n,k_1}Q^{-1}_{n_1,k_1}P_{k_1}(Z_i)u_i\right|^4}{n^4 n_1^4}\right| \mathcal{F}_{i,n}\right] \\
&\leq \frac{C}{\varepsilon^2}\sum_{i\in I_1}\frac{\left|\gamma_n' D_n\phi_{w\theta,n}P_{n,k_1}Q^{-1}_{n_1,k_1}P_{k_1}(Z_i)\right|^4}{n^4 n_1^4} \\
&\leq \frac{C\xi^2_{k_1}\gamma_n' D_n\phi_{w\theta,n}P_{n,k_1}Q^{-2}_{n_1,k_1}P'_{n,k_1}\phi'_{w\theta,n}D_n'\gamma_n}{\varepsilon^2}\sum_{i\in I_1}\frac{\left|\gamma_n' D_n\phi_{w\theta,n}P_{n,k_1}Q^{-1}_{n_1,k_1}P_{k_1}(Z_i)\right|^2}{n^4 n_1^4} \\
&\leq \frac{C\lambda_{\max}(Q_{n,k_1})}{\varepsilon^2\lambda^2_{\min}(Q_{n_1,k_1})}\frac{\xi^2_{k_1}}{n^2 n_1^4}\frac{\gamma_n' D_n\phi_{w\theta,n}\phi'_{w\theta,n}D_n'\gamma_n}{n}\sum_{i\in I_1}\left|\gamma_n' D_n\phi_{w\theta,n}P_{n,k_1}Q^{-1}_{n_1,k_1}P_{k_1}(Z_i)\right|^2 \\
&= \frac{C\lambda_{\max}(Q_{n,k_1})}{\varepsilon^2\lambda^2_{\min}(Q_{n_1,k_1})}\frac{\xi^2_{k_1}}{n_1}\frac{\gamma_n' D_n\phi_{w\theta,n}\phi'_{w\theta,n}D_n'\gamma_n}{nn_1}\frac{\gamma_n' D_n\phi_{w\theta,n}P_{n,k_1}Q^{-1}_{n_1,k_1}P'_{n,k_1}\phi'_{w\theta,n}D_n'\gamma_n}{n^2 n_1} \\
&\leq \frac{C\lambda^2_{\max}(Q_{n,k_1})}{\varepsilon^2\lambda^3_{\min}(Q_{n_1,k_1})}\frac{\xi^2_{k_1}}{n_1}\left|\frac{\gamma_n' D_n\phi_{w\theta,n}\phi'_{w\theta,n}D_n'\gamma_n}{nn_1}\right|^2 \\
&\leq \frac{C\lambda^2_{\max}(Q_{n,k_1})\lambda^2_{\max}(n^{-1}\phi_{w\theta,n}\phi'_{w\theta,n})}{\varepsilon^2\lambda^2_{\min}(Q_{n_1,k_1})}\frac{\xi^2_{k_1}}{n_1}\left|\frac{\gamma_n' D_n D_n'\gamma_n}{n_1}\right|^2 = o_p(1),
\end{aligned} \tag{2.160}$$

where the last equality is by (2.75), (2.156), (2.159) and Assumptions 2.3.1(v). Similarly, we can show that

$$\frac{1}{\varepsilon^2}\sum_{i\in I_2} E\left[\left|\frac{\left|\gamma_n' D_n\phi_{w\theta,n}P_{n,k_2}Q^{-1}_{n_2,k_2}P_{k_2}(Z_i)\varepsilon_i\right|^4}{n^4 n_2^4}\right| \mathcal{F}_{i,n}\right] = o_p(1), \tag{2.161}$$

which together with (2.153) and (2.160) proves (2.151). As a result, the asymptotic normality of $\widehat{\theta}_n$ follows by the martingale CLT. ■

2.9.2 Proof of the Main Results in Section 2.5

Lemma 2.9.4 *Under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.3.2(v) and 2.3.3(iv)-2.3.3(vi), we have*

$$
\begin{aligned}
n_1\Sigma_{n_1} &= E\left[u^2 w_n^2(Z)\phi_\theta(Z,\theta_0)\phi_\theta(Z,\theta_0)'\right] + o_p(1),\\
\text{and } n_2\Sigma_{n_2} &= E\left[\varepsilon^2 w_n^2(Z)\phi_\theta(Z,\theta_0)\phi_\theta(Z,\theta_0)'\right] + o_p(1).
\end{aligned}
$$

Proof of Lemma 2.9.4. For $j = 1,\ldots,d_\theta$, let $\widetilde{\phi}_{w\theta_j,k_1,n}(z) = P_{k_1}(z)'Q^{-1}_{n_1,k_1}P'_{n,k_1}\phi'_{w\theta_j,n}$ where $\phi_{w\theta_j,n}$ denotes the j-th row of $\phi_{w\theta,n}$. We first show that

$$
n^{-1}\sum_{i\in I}|\widetilde{\phi}_{w\theta_j,k_1,n}(Z_i) - w_n(Z_i)\phi_\theta(Z_i,\theta_0)|^2 = o_p(1) \tag{2.162}
$$

for any $j = 1,\ldots,d_\theta$. Let $\widetilde{\beta}_{w\phi_j,k_1} = Q^{-1}_{n_1,k_1}P'_{n,k_1}\phi'_{w\theta_j,n}$. Then

$$
\begin{aligned}
& n^{-1}\sum_{i\in I}|\widetilde{\phi}_{w\theta_j,k_1,n}(Z_i) - w_n(Z_i)\phi_\theta(Z_i,\theta_0)|^2\\
&\leq 2n^{-1}\sum_{i\in I}|P'_{k_1}(Z_i)\widetilde{\beta}_{w\phi_j,k_1} - P'_{k_1}(Z_i)\beta_{w\phi_j,k_1,n}|^2\\
&+ 2n^{-1}\sum_{i\in I}|P'_{k_1}(Z_i)\beta_{w\phi_j,k_1,n} - w_n(Z_i)\phi_\theta(Z_i,\theta_0)|^2\\
&= (\widetilde{\beta}_{w\phi_j,k_1} - \beta_{w\phi_j,k_1,n})'Q_{n,k_1}(\widetilde{\beta}_{w\phi_j,k_1} - \beta_{w\phi_j,k_1,n}) + o(1)
\end{aligned} \tag{2.163}
$$

where the equality is by Assumption 2.3.3(vi). Moreover

$$
\widetilde{\beta}_{w\phi_j,k_1} - \beta_{w\phi_j,k_1,n} = Q^{-1}_{n_1,k_1}P'_{n,k_1}(\phi_{w\theta_j,n} - \phi_{w\theta_j,n,k_1})' \tag{2.164}
$$

where $\phi_{w\theta_j,n,k_1} = (P'_{k_1}(Z_i)\beta_{w\phi_j,k_1,n})_{i\in I}$, which implies that

$$
\begin{aligned}
&(\widetilde{\beta}_{w\phi_j,k_1} - \beta_{w\phi_j,k_1,n})'Q_{n,k_1}(\widetilde{\beta}_{w\phi_j,k_1} - \beta_{w\phi_j,k_1,n})\\
&\leq \frac{\lambda^2_{\max}(Q_{n,k_1})}{n\lambda^2_{\min}(Q_{n_1,k_1})}(\phi_{w\theta_j,n} - \phi_{w\theta_j,n,k_1})P_{n,k_1}(P'_{n,k_1}P_{n,k_1})^{-1}P'_{n,k_1}(\phi_{w\theta_j,n} - \phi_{w\theta_j,n,k_1})'\\
&\leq \frac{\lambda^2_{\max}(Q_{n,k_1})}{\lambda^2_{\min}(Q_{n_1,k_1})}n^{-1}\sum_{i\in I}|P'_{k_1}(Z_i)\beta_{w\phi_j,k_1,n} - w_n(Z_i)\phi_\theta(Z_i,\theta_0)|^2 = o_p(1)
\end{aligned} \tag{2.165}
$$

where the equality is by Assumption 2.3.3(vi) and (2.75). Combining the results in (2.163) and (2.165), we immediately get (2.162).

Recall that $\sigma_u^2(z) = E\left[u^2 \middle| Z = z\right]$. By definition

$$n_1 \Sigma_{n_1} = n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) \widetilde{\phi}_{w\theta,k_1}(Z_i) \widetilde{\phi}_{w\theta,k_1}(Z_i)' \tag{2.166}$$

where $\widetilde{\phi}_{w\theta,k_1,n}(z) = (\widetilde{\phi}_{w\theta_j,k_1,n}(z))'_{j=1,\ldots,d_\theta}$. First note that

$$\begin{aligned}
&n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) \widetilde{\phi}_{w\theta,k_1,n}(Z_i) \widetilde{\phi}_{w\theta,k_1,n}(Z_i)' - n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) w_n^2(Z_i) \phi_\theta(Z_i, \theta_0) \phi_\theta(Z_i, \theta_0)' \\
&= n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0)) w_n(Z_i) \phi_\theta(Z_i, \theta_0)' \\
&+ n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) w_n(Z_i) \phi_\theta(Z_i, \theta_0) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0))' \\
&+ n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0)) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0))'.
\end{aligned} \tag{2.167}$$

For any any $j_1 = 1, \ldots, d_\theta$ and any $j_2 = 1, \ldots, d_\theta$, by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\left| n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta_{j_1},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0)) (\widetilde{\phi}_{w\theta_{j_2},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_2}}(Z_i, \theta_0)) \right|^2 \\
&\leq n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta_{j_1},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0))^2 \\
&\times n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta_{j_2},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_2}}(Z_i, \theta_0))^2 \\
&\leq C n_1^{-1} \sum_{i \in I_1} (\widetilde{\phi}_{w\theta_{j_1},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0))^2 \\
&\times n_1^{-1} \sum_{i \in I_1} (\widetilde{\phi}_{w\theta_{j_2},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_2}}(Z_i, \theta_0))^2 = o_p(1)
\end{aligned} \tag{2.168}$$

where the second inequality is by Assumption 2.3.3(v), the equality is by (2.162). (2.168) then implies that

$$n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0)) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0))' = o_p(1). \tag{2.169}$$

For any any $j_1 = 1, \ldots, d_\theta$ and any $j_2 = 1, \ldots, d_\theta$, by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\left| n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) w_n(Z_i) \phi_{\theta_{j_1}}(Z_i, \theta_0) (\widetilde{\phi}_{w\theta_{j_2},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_2}}(Z_i, \theta_0)) \right|^2 \\
&\leq n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) w_n^2(Z_i) \phi_{\theta_{j_1}}^2(Z_i, \theta_0) \times n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta_{j_2},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_2}}(Z_i, \theta_0))^2 \\
&\leq C n_1^{-1} \sum_{i \in I_1} \phi_{\theta_{j_1}}^2(Z_i, \theta_0) \times n_1^{-1} \sum_{i \in I_1} (\widetilde{\phi}_{w\theta_{j_2},k_1,n}(Z_i) - w_n(Z_i) \phi_{\theta_{j_2}}(Z_i, \theta_0))^2 = o_p(1)
\end{aligned} \tag{2.170}$$

where the second inequality is by Assumptions 2.3.2(v) and 2.3.3(v), the equality is by (2.162) and (2.115). (2.170) implies that

$$n_1^{-1} \sum_{i \in I_1} \sigma_u^2(Z_i) (\widetilde{\phi}_{w\theta,k_1,n}(Z_i) - w_n(Z_i) \phi_\theta(Z_i, \theta_0)) w_n(Z_i) \phi_\theta(Z_i, \theta_0)' = o_p(1) \tag{2.171}$$

and

$$n_1^{-1}\sum_{i\in I_1}\sigma_u^2(Z_i)w_n(Z_i)\phi_\theta(Z_i,\theta_0)(\widetilde{\phi}_{w\theta,k_1,n}(Z_i)-w_n(Z_i)\phi_\theta(Z_i,\theta_0))'=o_p(1). \tag{2.172}$$

Combining the results in (2.167), (2.169), (2.171) and (2.172), we have

$$n_1^{-1}\sum_{i\in I_1}\sigma_u^2(Z_i)\widetilde{\phi}_{w\theta,k_1,n}(Z_i)\widetilde{\phi}_{w\theta,k_1,n}(Z_i)'-n_1^{-1}\sum_{i\in I_1}\sigma_u^2(Z_i)w_n^2(Z_i)\phi_\theta(Z_i,\theta_0)\phi_\theta(Z_i,\theta_0)'=o_p(1). \tag{2.173}$$

For any $j_1=1,\ldots,d_\theta$ and any $j_2=1,\ldots,d_\theta$,

$$\begin{aligned}
&E\left[\left|\sigma_u^2(Z_i)w_n^2(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)\right|^2\right]\\
&\le CE\left[\left|\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)\right|^2\right]\\
&\le CE\left[\phi_{\theta_{j_1}}^4(Z_i,\theta_0)\right]E\left[\phi_{\theta_{j_2}}^4(Z_i,\theta_0)\right]<C
\end{aligned} \tag{2.174}$$

where the first inequality is by Assumptions 2.3.2(v) and 2.3.3(v), the second inequality is by the Hölder inequality, and the last inequality is by Assumption 2.3.3(iv). The i.i.d. assumption together with (2.174) and the Markov inequality implies that

$$n_1^{-1}\sum_{i\in I_1}\sigma_u^2(Z_i)w_n^2(Z_i)\phi_{\theta_{j_1}}(Z_i,\theta_0)\phi_{\theta_{j_2}}(Z_i,\theta_0)-E\left[\sigma_u^2(Z)w_n^2(Z)\phi_{\theta_{j_1}}(Z,\theta_0)\phi_{\theta_{j_2}}(Z,\theta_0)\right]=o_p(1). \tag{2.175}$$

Collecting the results in (2.173) and (2.175), we have

$$n_1\Sigma_{n_1}=E\left[u^2w_n^2(Z)\phi_\theta(Z,\theta_0)\phi_\theta(Z,\theta_0)'\right]+o_p(1) \tag{2.176}$$

which proves the first claim of the lemma. The proof of the second claim of the lemma is similar and hence omitted. ■

Proof of Lemma 2.5.1. By Lemma 2.9.4,

$$\Sigma_{n_1}+\Sigma_{n_2}=E\left[w_n^2(Z)\left(\frac{\sigma_u^2(Z)}{n_1}+\frac{\sigma_\varepsilon^2(Z)}{n_2}\right)\phi_\theta(Z,\theta_0)\phi_\theta'(Z,\theta_0)\right]+o_p(n_1^{-1}+n_2^{-1}). \tag{2.177}$$

By Assumptions 2.3.2(v), 2.3.3(ii) and 2.3.3(v),

$$E\left[w_n^2(Z)\left(\frac{\sigma_u^2(Z)}{n_1}+\frac{\sigma_\varepsilon^2(Z)}{n_2}\right)\phi_\theta(Z,\theta_0)\phi_\theta'(Z,\theta_0)\right]>C^{-1}(n_1^{-1}+n_2^{-1}) \tag{2.178}$$

which together with (2.177) implies that

$$\Sigma_{n_1}+\Sigma_{n_2}=E\left[w_n^2(Z)\left(\frac{\sigma_u^2(Z)}{n_1}+\frac{\sigma_\varepsilon^2(Z)}{n_2}\right)\phi_\theta(Z,\theta_0)\phi_\theta'(Z,\theta_0)\right](1+o_p(1)). \tag{2.179}$$

The claim of the lemma then follows by combining (2.179) and Assumption 2.3.3(ii). ■

Proof of Theorem 2.5.1. Let $w_n^*(Z) = (n_1^{-1} + n_2^{-1})(n_1^{-1}\sigma_u^2(Z) + n_2^{-1}\sigma_\varepsilon^2(Z))^{-1}$ and $H_{0,n}^* = E\left[w_n^*(Z)\phi_\theta(Z,\theta_0)\phi_\theta'(Z,\theta_0)\right]$. Then by definition

$$V_{n,\theta}^* = (n_1^{-1} + n_2^{-1})(H_{0,n}^*)^{-1}. \tag{2.180}$$

For any $w_n(\cdot)$, define

$$A_n(Z, w_n) = H_{0,n}^{-1}\phi_\theta(Z,\theta_0)w_n(Z) - H_{0,n}^{*-1}\phi_\theta(Z,\theta_0)w_n^*(Z). \tag{2.181}$$

Then we have

$$V_{n,\theta} - V_{n,\theta}^* = E\left[A_n(Z,w_n)\Omega_n(Z)A_n(Z,w_n)'\right] \geq 0, \tag{2.182}$$

for any n_1 and any n_2, where $\Omega_n(Z) = n_1^{-1}\sigma_u^2(Z) + n_2^{-1}\sigma_\varepsilon^2(Z)$ and the inequality is by the fact that $\Omega_n(Z) \in (0,\infty)$ and $A_n(Z,w_n)\Omega_n(Z)A_n(Z,w_n)'$ is a positive semidefinite matrix almost surely. ■

Proof of Lemma 2.5.2. By definition $w_n^*(z)^{-1} = (n_1^{-1}\sigma_u^2(z) + n_2^{-1}\sigma_\varepsilon^2(z))(n_1^{-1} + n_2^{-1})^{-1}$. Then for any n_1, n_2,

$$\min\left\{\inf_{z\in\mathcal{Z}}\sigma_u^2(z), \inf_{z\in\mathcal{Z}}\sigma_\varepsilon^2(z)\right\} \leq w_n^*(z)^{-1} \leq \max\left\{\sup_{z\in\mathcal{Z}}\sigma_u^2(z), \sup_{z\in\mathcal{Z}}\sigma_\varepsilon^2(z)\right\}$$

which together with Assumption 2.3.3(v) proves the claim of the lemma. ■

2.9.3 Proof of the Main Results in Section 2.6

Lemma 2.9.5 *Under Assumptions 2.3.1, 2.3.2, 2.3.3 and 2.6.1(i), we have*

$$n_2^{-1}\sum_{i\in I_2}\left|\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_n) - \phi(Z_i,\theta_0)\right|^2 = O_p(n_1^{-1} + k_2 n_2^{-1} + k_2^{-2r_h}) \tag{2.183}$$

and moreover,

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\phi}_{n_2}(z,\widehat{\theta}_n) - \phi(z,\theta_0)\right| = O_p(\xi_{k_2}n_1^{-1/2} + \xi_{k_2}k_2^{1/2}n_2^{-1/2} + \xi_{k_2}k_2^{-r_h}). \tag{2.184}$$

Proof of Lemma 2.9.5. By definition,

$$\widehat{\phi}_{n_2}(z,\widehat{\theta}_n) = P_{k_2}'(z)(P_{k_2,n_2}'P_{k_2,n_2})^{-1}P_{k_2,n_2}'g_{n_2}(\widehat{\theta}_n);$$
$$\widehat{\phi}_{n_2}(z,\theta_0) = P_{k_2}'(z)(P_{k_2,n_2}'P_{k_2,n_2})^{-1}P_{k_2,n_2}'g_{n_2}(\theta_0),$$

where $g_{n_2}(\widehat{\theta}_n) = (g(Z_i,\widehat{\theta}_n))_{i\in I_2}'$ and $g_{n_2}(\theta_0) = (g(Z_i,\theta_0))_{i\in I_2}'$. By the triangle inequality, for any z,

$$\begin{aligned}\left|\widehat{\phi}_{n_2}(z,\widehat{\theta}_n) - \phi(z,\theta_0)\right| &\leq \left|\widehat{\phi}_{n_2}(z,\widehat{\theta}_n) - \widehat{\phi}_{n_2}(z,\theta_0)\right| \\ &\quad + \left|\widehat{\phi}_{n_2}(z,\theta_0) - h_{0,k_2}(z)\right| + \left|h_{0,k_2}(z) - h_0(z)\right|.\end{aligned} \tag{2.185}$$

By the mean value expansion and the Cauchy-Schwarz inequality,

$$\left|g(X_i,\widehat{\theta}_n)-g(X_i,\theta_0)\right| \leq \sup_{\theta\in\mathcal{N}_n}\|g_\theta(X_i,\theta)\| \left\|\widehat{\theta}_n-\theta_0\right\| \tag{2.186}$$

which together with Assumption 2.6.1(i) and (2.9) in Theorem 2.3.2 implies that

$$n_2^{-1}\sum_{i\in I_2}\left|g(Z_i,\widehat{\theta}_n)-g(Z_i,\theta_0)\right|^2 \leq \sup_{\theta\in\mathcal{N}_n} n_2^{-1}\sum_{i\in I_2}\|g_\theta(X_i,\theta)\|^2\left\|\widehat{\theta}_n-\theta_0\right\|^2 = O_p(n_1^{-1}+n_2^{-1}). \tag{2.187}$$

Under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(iv) and 2.3.3(v), we can use similar arguments in proving (2.94) to show that

$$n_2^{-1}\sum_{i\in I_2}(\widehat{\phi}_{n_2}(Z_i,\theta_0)-h_{0,k_2}(Z_i))^2 = O_p(k_2n_2^{-1}+k_2^{-2r_h}). \tag{2.188}$$

Using (2.187), we get

$$\begin{aligned}
& n_2^{-1}\sum_{i\in I_2}\left|\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_n)-\widehat{\phi}_{n_2}(Z_i,\theta_0)\right|^2 \\
&= \frac{[g_{n_2}(\widehat{\theta}_n)-g_{n_2}(\theta_0)]'P_{k_2,n_2}(P'_{k_2,n_2}P_{k_2,n_2})^{-1}P'_{k_2,n_2}[g_{n_2}(\widehat{\theta}_n)-g_{n_2}(\theta_0)]}{n_2} \\
&\leq n_2^{-1}\sum_{i\in I_2}\left|g(Z_i,\widehat{\theta}_n)-g(Z_i,\theta_0)\right|^2 = O_p(n_1^{-1}+n_2^{-1}),
\end{aligned} \tag{2.189}$$

which together with (2.185), (2.188) and Assumption 2.3.1(iv) implies that

$$n_2^{-1}\sum_{i\in I_2}\left|\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_n)-\phi(Z_i,\theta_0)\right|^2 = O_p(n_1^{-1}+k_2n_2^{-1}+k_2^{-2r_h}). \tag{2.190}$$

This proves the first claim of the lemma.

By (2.93) and the Cauchy-Schwarz inequality,

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\phi}_{n_2}(z,\theta_0)-h_{0,k_2}(z)\right| = O_p(\xi_{k_2}k_2^{1/2}n_2^{-1/2}+\xi_{k_2}k_2^{-r_h}). \tag{2.191}$$

Using (2.187), we get

$$\begin{aligned}
& \sup_{z\in\mathcal{Z}}\left|\widehat{\phi}_{n_2}(z,\widehat{\theta}_n)-\widehat{\phi}_{n_2}(z,\theta_0)\right| \\
&\leq \xi_{k_2}\left\|[g_{n_2}(\widehat{\theta}_n)-g_{n_2}(\theta_0)]'P_{k_2,n_2}(P'_{k_2,n_2}P_{k_2,n_2})^{-1}\right\| \\
&\leq \xi_{k_2}\left(n_2^{-1}\sum_{i\in I_2}\left|g(Z_i,\widehat{\theta}_n)-g(Z_i,\theta_0)\right|^2\right)^{1/2} \\
&= O_p(\xi_{k_2}n_1^{-1/2}+\xi_{k_2}n_2^{-1/2})
\end{aligned} \tag{2.192}$$

which together with (2.185), (2.191) and Assumption 2.3.1(iv) implies that

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\phi}_{n_2}(z,\widehat{\theta}_n)-\phi(z,\theta_0)\right| = O_p(\xi_{k_2}n_1^{-1/2}+\xi_{k_2}k_2^{1/2}n_2^{-1/2}+\xi_{k_2}k_2^{-r_h}). \tag{2.193}$$

This proves the second claim of the lemma. ■

Lemma 2.9.6 *Under the conditions of Theorem 2.6.1, we have*

(i) $\widehat{H}_n = H_{0,n} + o_p(1)$*;*

(ii) $n^{-1}(\widehat{\phi}_{w\theta,n} - \phi_{w\theta,n})' P_{n,k_1} = o_p(1)$*;*

(iii) $n^{-1}(\widehat{\phi}_{w\theta,n} - \phi_{w\theta,n})' P_{n,k_2} = o_p(1)$*;*

(iv) $\sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \left[\gamma_k'(\widehat{Q}_{n_1,u} - Q_{n_1,u})\gamma_k\right] = o_p(1)$*;*

(v) $\sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \left[\gamma_k'(\widehat{Q}_{n_2,\varepsilon} - Q_{n_2,\varepsilon})\gamma_k\right] = o_p(1)$*.*

Proof of Lemma 2.9.6. (i) The proof of the first claim follows by the consistency of $\widehat{\theta}_n$ and similar arguments in deriving (2.119). Hence it is omitted.

(ii) By definition,

$$\begin{aligned} n^{-1}(\widehat{\phi}_{w\theta,n} - \phi_{w\theta,n})' P_{n,k_1} &= n^{-1} \sum_{i \in I} (\widehat{w}_n(Z_i) - w_n(Z_i)) \widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) P_{k_1}(Z_i) \\ &\quad + n^{-1} \sum_{i \in I} w_n(Z_i)(\widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) - \phi_\theta(Z_i, \theta_0)) P_{k_1}(Z_i). \end{aligned} \tag{2.194}$$

By (2.9) in Theorem 2.3.2, Lemma 2.9.1 and similar arguments in showing (2.113),

$$\begin{aligned} & n^{-1} \sum_{i \in I} \left\| \widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) - \widehat{\phi}_{\theta,n_2}(Z_i, \theta_0) \right\|^2 \\ & \leq \sup_{\theta \in \mathcal{N}_{\delta_n}} n^{-1} \sum_{i \in I} \left\| \widehat{\phi}_{\theta_j\theta,n_2}(Z_i, \theta) \right\|^2 \left\| \widehat{\theta}_n - \theta_0 \right\|^2 = O_p(n_1^{-1} + n_2^{-1}), \end{aligned} \tag{2.195}$$

which together with Assumption 2.3.3(iii) implies that

$$\begin{aligned} & n^{-1} \sum_{i \in I} \left\| \widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) - \phi_\theta(Z_i, \theta_0) \right\|^2 \\ & \leq 2n^{-1} \sum_{i \in I} \left\| \widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) - \widehat{\phi}_{\theta,n_2}(Z_i, \theta_0) \right\|^2 \\ & + 2n^{-1} \sum_{i \in I} ||\widehat{\phi}_{\theta,n_2}(Z_i, \theta_0) - \phi_\theta(Z_i, \theta_0)||^2 = o_p(n_1^{-1/2} + n_2^{-1/2}). \end{aligned} \tag{2.196}$$

By (2.75), (2.115), (2.196), Assumption 2.6.1(iv), and the Cauchy-Schwarz inequality,

$$\begin{aligned} & \left\| n^{-1} \sum_{i \in I} (\widehat{w}_n(Z_i) - w_n(Z_i)) \widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) P_{k_1}(Z_i) \right\|^2 \\ & \leq \sup_{z \in \mathcal{Z}} |\widehat{w}_n(z) - w_n(z)|^2 \left(n^{-1} \sum_{i \in I} \left\| \widehat{\phi}_{\theta,n_2}(Z_i, \widehat{\theta}_n) \right\|^2 \right) n^{-1} \sum_{i \in I} \|P_{k_1}(Z_i)\|^2 \\ & = o_p(k_1 \delta_{w,n}^2) = o_p(1), \end{aligned} \tag{2.197}$$

where the $n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)\right\|^2 = O_p(1)$ is used in the first equality which is by (2.115) and (2.196). By Assumptions 2.3.2(v) and 2.3.3(vii), (2.196) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left\|n^{-1}\sum_{i\in I} w_n(Z_i)(\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)-\phi_\theta(Z_i,\theta_0))P_{k_1}(Z_i)\right\|^2 \\
&\leq \sup_{z\in\mathcal{Z}}|w_n(z)|\, n^{-1}\sum_{i\in I}\left\|\widehat{\phi}_{\theta,n_2}(Z_i,\widehat{\theta}_n)-\phi_\theta(Z_i,\theta_0)\right\|^2 n^{-1}\sum_{i\in I}\|P_{k_1}(Z_i)\|^2 = o_p(1). \qquad (2.198)
\end{aligned}$$

Collecting the results in (2.194), (2.197) and (2.198), we immediately prove the second claim of the lemma.

(iii) The proof of this result is similar to the arguments in the proof in (ii) and hence is omitted.

(iv) By definition,

$$\begin{aligned}
\widehat{Q}_{n_1,u} - Q_{n_1,u} &= n_1^{-1}\sum_{i\in I_1}(Y_i-\widehat{h}_{n_1}(Z_i))^2 P_{k_1}(Z_i)P'_{k_1}(Z_i) - Q_{n_1,u} \\
&= n_1^{-1}\sum_{i\in I_1}\left(u_i^2-\sigma_u^2(Z_i)\right)P_{k_1}(Z_i)P'_{k_1}(Z_i) \\
&+ n_1^{-1}\sum_{i\in I_1}(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))^2 P_{k_1}(Z_i)P'_{k_1}(Z_i) \\
&- 2n_1^{-1}\sum_{i\in I_1}(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))u_i P_{k_1}(Z_i)P'_{k_1}(Z_i). \qquad (2.199)
\end{aligned}$$

By (2.75), Assumptions 2.3.1(i), 2.3.3(v),

$$\begin{aligned}
&E\left[\left\|n_1^{-1}\sum_{i\in I_1}\left(u_i^2-\sigma_u^2(Z_i)\right)P_{k_1}(Z_i)P'_{k_1}(Z_i)\right\|^2\middle|\{Z_i\}_{i\in I_1}\right] \\
&= n_1^{-2}\sum_{i\in I_1}E\left[\left(u_i^2-\sigma_u^2(Z_i)\right)^2\middle| Z_i\right]\left|P'_{k_1}(Z_i)P_{k_1}(Z_i)\right|^2 \\
&\leq C\xi_{k_1}^2 n_1^{-2}\sum_{i\in I_1}P'_{k_1}(Z_i)P_{k_1}(Z_i) \\
&\leq C\lambda_{\max}(Q_{k_1,n_1})\xi_{k_1}^2 k_1 n_1^{-1} = O_p(\xi_{k_1}^2 k_1 n_1^{-1}), \qquad (2.200)
\end{aligned}$$

which together with the Markov inequality and Assumption 2.6.1(iv) implies that

$$n_1^{-1}\sum_{i\in I_1}\left(u_i^2-\sigma_u^2(Z_i)\right)P_{k_1}(Z_i)P'_{k_1}(Z_i) = o_p(1). \qquad (2.201)$$

Using Assumption 2.3.1(iv), (2.76), we get

$$\begin{aligned}
\sup_{z\in\mathcal{Z}}\left|\widehat{h}_{n_1}(z)-h_0(z)\right| &\leq \sup_{z\in\mathcal{Z}}\left|\widehat{h}_{n_1}(z)-h_{k_1}(z)\right| + \sup_{z\in\mathcal{Z}}|h_{0,k_1}(z)-h_0(z)| \\
&\leq \sup_{z\in\mathcal{Z}}\left|(\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1})'P_{k_1}(z)\right| + \sup_{z\in\mathcal{Z}}|h_{0,k_1}(z)-h_0(z)| \\
&\leq \left\|\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1}\right\|\xi_{k_1} + \sup_{z\in\mathcal{Z}}|h_{0,k_1}(z)-h_0(z)| \\
&= O_p(\xi_{k_1}k_1^{1/2}n_1^{-1/2}+\xi_{k_1}k_1^{-r_h}), \qquad (2.202)
\end{aligned}$$

where the first inequality is by the triangle inequality, the second inequality is by the Cauchy-Schwarz inequality. For any $\gamma_k \in R^k$ with $\gamma_k' \gamma_k = 1$, we have

$$
\begin{aligned}
&\sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} n_1^{-1} \sum_{i \in I_1} (\widehat{h}_{n_1}(Z_i) - h_0(Z_i))^2 \gamma_k' P_{k_1}(Z_i) P_{k_1}'(Z_i) \gamma_k \\
&\leq \sup_z \left| \widehat{h}_{n_1}(z) - h_0(z) \right|^2 \sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \frac{1}{n_1} \sum_{i \in I_1} \gamma_k' P_{k_1}(Z_i) P_{k_1}'(Z_i) \gamma_k \\
&\leq \sup_z \left| \widehat{h}_{n_1}(z) - h_0(z) \right|^2 \lambda_{\max}(Q_{k_1,n_1}) \\
&= O_p(\xi_{k_1}^2 k_1 n_1^{-1} + \xi_{k_1}^2 k_1^{-2r_h}) = o_p(1) \qquad (2.203)
\end{aligned}
$$

where the first equality is by (2.75) and (2.202), the last equality is by Assumption 2.6.1(iv). By the triangle inequality,

$$
\begin{aligned}
&\sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \left\| n_1^{-1} \sum_{i \in I_1} (h_{0,k_1}(Z_i) - h_0(Z_i)) u_i \gamma_k' P_{k_1}(Z_i) P_{k_1}'(Z_i) \gamma_k \right\| \\
&\leq \left\| n_1^{-1} \sum_{i \in I_1} (h_{0,k_1}(Z_i) - h_0(Z_i)) u_i P_{k_1}(Z_i) P_{k_1}'(Z_i) \right\|. \qquad (2.204)
\end{aligned}
$$

By Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(iv) and 2.3.3(v),

$$
\begin{aligned}
&E\left[\left\| n_1^{-1} \sum_{i \in I_1} (h_{0,k_1}(Z_i) - h_0(Z_i)) u_i P_{k_1}(Z_i) P_{k_1}'(Z_i) \right\|^2 \right] \\
&= n_1^{-1} E\left[\sum_{i \in I_1} (h_{0,k_1}(Z_i) - h_0(Z_i))^2 u_i^2 \left| P_{k_1}'(Z_i) P_{k_1}(Z_i) \right|^2 \right] \\
&\leq C n_1^{-1} \sup_{z \in \mathcal{Z}} |h_{0,k_1}(z) - h_0(z)|^2 \xi_{k_1}^2 E\left[P_{k_1}'(Z_i) P_{k_1}(Z_i) \right] = O_p(\xi_{k_1}^2 k_1^{1-2r_h} n_1^{-1}) \qquad (2.205)
\end{aligned}
$$

which together with (2.204), Assumption 2.6.1(iv), and the Markov inequality implies that

$$
\sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \left\| n_1^{-1} \sum_{i \in I_1} (h_{0,k_1}(Z_i) - h_0(Z_i)) u_i \gamma_k' P_{k_1}(Z_i) P_{k_1}'(Z_i) \gamma_k \right\| = o_p(1). \qquad (2.206)
$$

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \left\| n_1^{-1} \sum_{i \in I_1} (\widehat{h}_{n_1}(Z_i) - h_{0,k_1}(Z_i)) u_i \gamma_k' P_{k_1}(Z_i) P_{k_1}'(Z_i) \gamma_k \right\|^2 \\
&\leq \left\| \widehat{\beta}_{k_1,n_1} - \beta_{h,k_1} \right\|^2 \sup_{\{\gamma_k \in R^k:\ \gamma_k'\gamma_k = 1\}} \sum_{j=1}^{k_1} \left| n_1^{-1} \sum_{i \in I_1} p_j(Z_i) u_i \gamma_k' P_{k_1}(Z_i) P_{k_1}'(Z_i) \gamma_k \right|^2 \\
&\leq \left\| \widehat{\beta}_{k_1,n_1} - \beta_{h,k_1} \right\|^2 \sum_{j=1}^{k_1} || n_1^{-1} \sum_{i \in I_1} p_j(Z_i) u_i P_{k_1}(Z_i) P_{k_1}'(Z_i) ||^2. \qquad (2.207)
\end{aligned}
$$

By Assumptions 2.3.1(i), 2.3.1(iii) and 2.3.3(v),

$$
\begin{aligned}
& E\left[\sum_{j=1}^{k_1} ||n_1^{-1}\sum_{i\in I_1} p_j(Z_i)u_i P_{k_1}(Z_i)P_{k_1}'(Z_i)||^2\right] \\
& = \sum_{j=1}^{k_1} n_1^{-1} E\left[p_j^2(Z_i)u_i^2\left|P_{k_1}'(Z_i)P_{k_1}(Z_i)\right|^2\right] \\
& \leq Cn_1^{-1}\xi_{k_1}^4 \sum_{j=1}^{k_1} E\left[p_j^2(Z_i)\right] = O(\xi_{k_1}^4 k_1 n_1^{-1}),
\end{aligned} \tag{2.208}
$$

which together with (2.76), (2.207), Assumption 2.6.1(iv), and the Markov inequality implies that

$$
\begin{aligned}
& \sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} \left\| n_1^{-1}\sum_{i\in I_1}(\widehat{h}_{n_1}(Z_i)-h_{0,k_1}(Z_i))u_i\gamma_k' P_{k_1}(Z_i)P_{k_1}'(Z_i)\gamma_k \right\| \\
& = O_p(\xi_{k_1}^2 k_1^{1/2} n_1^{-1/2}(k_1^{1/2}n_1^{-1/2}+k_1^{-r_h})) = o_p(1).
\end{aligned} \tag{2.209}
$$

Combining the results in (2.206), (2.209) and the triangle inequality, we get

$$
\sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} \left\| n_1^{-1}\sum_{i\in I_1}(\widehat{h}_{n_1}(Z_i)-h_0(Z_i))u_i\gamma_k' P_{k_1}(Z_i)P_{k_1}'(Z_i)\gamma_k \right\| = o_p(1). \tag{2.210}
$$

Combining the results in (2.199), (2.201), (2.203) and (2.210), we prove the fourth claim of the lemma.

(v) By definition,

$$
\begin{aligned}
\widehat{Q}_{n_2,\varepsilon} - Q_{n_2,\varepsilon} &= n_2^{-1}\sum_{i\in I_2}\left(\varepsilon_i^2-\sigma_\varepsilon^2(Z_i)\right)P_{k_2}(Z_i)P_{k_2}'(Z_i) \\
&+ 2n_2^{-1}\sum_{i\in I_2}\left[g(X_i,\widehat{\theta}_n)-\widehat{\phi}(Z_i,\widehat{\theta}_n)-g(X_i,\theta_0)+\phi(Z_i,\theta_0)\right]\varepsilon_i P_{k_2}(Z_i)P_{k_2}'(Z_i) \\
&+ n_2^{-1}\sum_{i\in I_2}\left[\left|g(X_i,\widehat{\theta}_n)-\widehat{\phi}(Z_i,\widehat{\theta}_n)-g(X_i,\theta_0)+\phi(Z_i,\theta_0)\right|^2\right]P_{k_2}(Z_i)P_{k_2}'(Z_i).
\end{aligned} \tag{2.211}
$$

Using similar arguments in showing (2.201), we get

$$
n_2^{-1}\sum_{i\in I_2}\left(\varepsilon_i^2 - E[\varepsilon_i^2|Z_i]\right)P_{k_2}(Z_i)P_{k_2}'(Z_i) = O_p(\xi_{k_2}k_2^{1/2}n_2^{-1/2}) = o_p(1). \tag{2.212}
$$

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
& \sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} n_2^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_n)-g(X_i,\theta_0))^2\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k \\
& \leq \xi_{k_2}^2 n_2^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_n)-g(X_i,\theta_0))^2 = O_p(\xi_{k_2}^2 n_1^{-1}+\xi_{k_2}^2 n_2^{-1}) = o_p(1)
\end{aligned} \tag{2.213}
$$

where the first equality is by (2.187), the second equality is by Assumption 2.6.1(iv). Similarly

$$\sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} n_2^{-1}\sum_{i\in I_2}(\widehat{\phi}(Z_i,\widehat{\theta}_n)-\phi(Z_i,\theta_0))^2\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k$$
$$\leq \xi_{k_2}^2 n_2^{-1}\sum_{i\in I_2}(\widehat{\phi}(Z_i,\widehat{\theta}_n)-\phi(Z_i,\theta_0))^2 = O_p(\xi_{k_2}^2 k_2 n_2^{-1}+\xi_{k_2}^2 k_2^{-2r_h}) = o_p(1) \tag{2.214}$$

where the first equality is by (2.183), the second equality is by Assumption 2.6.1(iv). Collecting the results in (2.213) and (2.214), we have

$$\sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} n_2^{-1}\sum_{i\in I_2}\left[\left|g(X_i,\widehat{\theta}_n)-\widehat{\phi}(Z_i,\widehat{\theta}_n)-g(X_i,\theta_0)-\phi(Z_i,\theta_0)\right|^2\right]\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k$$
$$\leq \sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} 2n_2^{-1}\sum_{i\in I_2}\left[\left|g(X_i,\widehat{\theta}_n)-g(X_i,\theta_0)\right|^2\right]\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k$$
$$+\sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} 2n_2^{-1}\sum_{i\in I_2}\left[\left|\widehat{\phi}(Z_i,\widehat{\theta}_n)-\phi(Z_i,\theta_0)\right|^2\right]\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k = o_p(1). \tag{2.215}$$

Next, note that by the Cauchy-Schwarz inequality and the triangle inequality,

$$\left|n_2^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_n)-\widehat{\phi}(Z_i,\widehat{\theta}_n)-g(X_i,\theta_0)-\phi(Z_i,\theta_0))\varepsilon_i\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k\right|$$
$$\leq \left|n_2^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_n)-g(X_i,\theta_0))\varepsilon_i\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k\right|$$
$$+\left|n_2^{-1}\sum_{i\in I_2}(\widehat{\phi}(Z_i,\widehat{\theta}_n)-\phi(Z_i,\theta_0))\varepsilon_i\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k\right|. \tag{2.216}$$

By the definition of γ_k, we can use the Cauchy-Schwarz inequality to show that

$$\sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} n_2^{-1}\sum_{i\in I_2}\varepsilon_i^2\left|\gamma_k' P_{k_2}(Z_i)P_{k_2}'(Z_i)\gamma_k\right|^2$$
$$\leq \sup_{\{\gamma_k\in R^k:\ \gamma_k'\gamma_k=1\}} \xi_{k_2}^2\gamma_k'\left(n_2^{-1}\sum_{i\in I_2}\varepsilon_i^2 P_{k_2}(Z_i)P_{k_2}'(Z_i)\right)\gamma_k$$
$$\leq \xi_{k_2}^2\lambda_{\max}\left(n_2^{-1}\sum_{i\in I_2}\varepsilon_i^2 P_{k_2}(Z_i)P_{k_2}'(Z_i)\right). \tag{2.217}$$

By Assumptions 2.3.1(i), 2.3.1(iii), 2.3.3(v),

$$E\left[\left\|n_1\sum_{i\in I_1}(\varepsilon_i^2-\sigma_\varepsilon^2(Z_i))P_{k_1}(Z_i)P_{k_1}(Z_i)'\right\|^2\right]$$
$$= n_1^{-1}E\left[(\varepsilon_i^2-\sigma_\varepsilon^2(Z_i))^2\left|P_{k_1}(Z_i)'P_{k_1}(Z_i)\right|^2\right]$$
$$\leq Cn_1^{-1}\xi_{k_1}^2 E\left[P_{k_1}(Z_i)'P_{k_1}(Z_i)\right] \leq Ck_1\xi_{k_1}^2 n_1^{-1} \tag{2.218}$$

which together with the Markov inequality and Assumption 2.6.1(iv) implies that

$$n_1 \sum_{i \in I_1} \varepsilon_i^2 P_{k_1}(Z_i) P_{k_1}(Z_i)' - Q_{n_1,\varepsilon} = o_p(1). \tag{2.219}$$

By (2.135) and (2.219), we have

$$\lambda_{\max}\left(n_2^{-1} \sum_{i \in I_2} \varepsilon_i^2 P_{k_2}(Z_i) P_{k_2}'(Z_i)\right) \leq C \tag{2.220}$$

with probability approaching 1, which together with (2.187), (2.217) and Assumption 2.6.1(iv), implies that

$$\begin{aligned}
&\left(n_2^{-1} \sum_{i \in I_2} (g(X_i, \widehat{\theta}_n) - g(X_i, \theta_0))^2 \times n_2^{-1} \sum_{i \in I_2} \varepsilon_i^2 \left|\gamma_k' P_{k_2}(Z_i) P_{k_2}'(Z_i) \gamma_k\right|^2\right)^{1/2} \\
&= O_p(\xi_{k_2} n_1^{-1/2} + \xi_{k_2} n_2^{-1/2}) = o_p(1). \qquad (2.221)
\end{aligned}$$

Similarly, by (2.183), (2.217), (2.220) and Assumption 2.6.1(iv)

$$\begin{aligned}
&\left(n_2^{-1} \sum_{i \in I_2} (\widehat{\phi}(Z_i, \widehat{\theta}_n) - \phi(Z_i, \theta_0))^2 \times n_2^{-1} \sum_{i \in I_2} \varepsilon_i^2 \left|\gamma_k' P_{k_2}(Z_i) P_{k_2}'(Z_i) \gamma_k\right|^2\right)^{1/2} \\
&= O_p(\xi_{k_2} k_2^{1/2} n_2^{-1/2} + \xi_{k_2} k_2^{-r_h}) = o_p(1), \qquad (2.222)
\end{aligned}$$

which together with (2.216), (2.221), (2.222) and the Cauchy-Schwarz inequality implies that

$$n_2^{-1} \sum_{i \in I_2} (g(X_i, \widehat{\theta}_n) - \widehat{\phi}(Z_i, \widehat{\theta}_n) - g(X_i, \theta_0) - \phi(Z_i, \theta_0)) \varepsilon_i \gamma_k' P_{k_2}(Z_i) P_{k_2}'(Z_i) \gamma_k = o_p(1) \tag{2.223}$$

uniformly over γ_k with $\gamma_k' \gamma_k = 1$. Collecting the results in (2.211), (2.212), (2.215) and (2.223), we prove the last claim of the theorem. ■

Proof of Theorem 2.6.1. By definition,

$$\begin{aligned}
\widehat{\Sigma}_{n_1} - \Sigma_{n_1} &= \frac{\widehat{\phi}_{w\theta,n} P_{n,k_1} - \phi_{w\theta,n} P_{n,k_1}}{n} \frac{Q_{n_1,k_1}^{-1} \widehat{Q}_{n_1,u} Q_{n_1,k_1}^{-1}}{n_1} \frac{P_{n,k_1}' \widehat{\phi}_{w\theta,n}'}{n} \\
&+ \frac{\phi_{w\theta,n} P_{n,k_1}}{n} Q_{n_1,k_1}^{-1} \frac{\widehat{Q}_{n_1,u} - Q_{n_1,u}}{n_1} Q_{n_1,k_1}^{-1} \frac{P_{n,k_1}' \widehat{\phi}_{w\theta,n}'}{n} \\
&+ \frac{\phi_{w\theta,n} P_{n,k_1}}{n} Q_{n_1,k_1}^{-1} \frac{Q_{n_1,u}}{n_1} Q_{n_1,k_1}^{-1} \frac{P_{n,k_1}' \widehat{\phi}_{w\theta,n}' - P_{n,k_1}' \phi_{w\theta,n}'}{n}. \qquad (2.224)
\end{aligned}$$

For any $j = 1, \ldots, d_\theta$,

$$\begin{aligned}
\frac{\phi_{w\theta_j,n} P_{n,k_1}}{n} \frac{P_{n,k_1}' \phi_{w\theta_j,n}'}{n} &\leq \lambda_{\max}(Q_{n,k_1}) n^{-1} \phi_{w\theta_j,n}' \phi_{w\theta_j,n} \\
&\leq \lambda_{\max}(Q_{n,k_1}) n^{-1} \sum_{i \in I} w_n^2(Z_i) \phi_{\theta_j}^2(Z_i, \theta_0) \\
&= \lambda_{\max}(Q_{n,k_1}) E[w_n^2(Z_i) \phi_{\theta_j}^2(Z_i, \theta_0)] + o_p(1) = O_p(1) \qquad (2.225)
\end{aligned}$$

where the first equality is by (2.75) and (2.99), the second equality is by (2.75) and (2.100). Moreover,

$$\frac{\widehat{\phi}_{w\theta_j,n}P_{n,k_1}}{n}\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_j,n}}{n} \leq 2\frac{(\widehat{\phi}_{w\theta_j,n}-\phi_{w\theta_j,n})P_{n,k_1}}{n}\frac{P'_{n,k_1}(\widehat{\phi}_{w\theta_j,n}-\phi_{w\theta_j,n})}{n} + 2\frac{\phi_{w\theta_j,n}P_{n,k_1}}{n}\frac{P'_{n,k_1}\phi'_{w\theta_j,n}}{n} = O_p(1) \tag{2.226}$$

where the equality is by Lemma 2.9.6(ii) and (2.225). By Lemma 2.9.6(iv) and (2.134), we know that

$$C^{-1} \leq \lambda_{\min}(\widehat{Q}_{n_1,u}) \leq \lambda_{\max}(\widehat{Q}_{n_1,u}) \leq C \tag{2.227}$$

with probability approaching 1. For any $j_1 = 1,\ldots,d_\theta$ and any $j_2 = 1,\ldots,d_\theta$, by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\left|\frac{\widehat{\phi}_{w\theta_{j_1},n}P_{n,k_1}-\phi_{w\theta_{j_1},n}P_{n,k_1}}{n}\frac{Q^{-1}_{n_1,k_1}\widehat{Q}_{n_1,u}Q^{-1}_{n_1,k_1}}{n_1}\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_{j_2},n}}{n}\right| \\
&\leq n_1^{-1}\left\|n^{-1}(\widehat{\phi}_{w\theta_{j_1},n}-\phi_{w\theta_{j_1},n})P_{n,k_1})\right\|\sqrt{\frac{\widehat{\phi}_{w\theta_{j_2},n}P_{n,k_1}}{n}\left(Q^{-1}_{n_1,k_1}\widehat{Q}_{n_1,u}Q^{-1}_{n_1,k_1}\right)^2\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_{j_2},n}}{n}} \\
&\leq \frac{\lambda_{\max}(\widehat{Q}_{n_1,u})}{n_1\lambda^2_{\min}(Q_{k_1,n_1})}\left\|n^{-1}(\widehat{\phi}_{w\theta_{j_1},n}-\phi_{w\theta_{j_1},n})P_{n,k_1})\right\|\sqrt{\frac{\widehat{\phi}_{w\theta_{j_2},n}P_{n,k_1}}{n}\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_{j_2},n}}{n}} = o_p(n_1^{-1})
\end{aligned} \tag{2.228}$$

where the equality is by (2.75), (2.227), Lemma 2.9.6(ii) and (2.226). Similarly for any $j_1 = 1,\ldots,d_\theta$ and any $j_2 = 1,\ldots,d_\theta$,

$$\begin{aligned}
&\left|\frac{\phi_{w\theta_{j_1},n}P_{n,k_1}}{n}Q^{-1}_{n_1,k_1}\frac{\widehat{Q}_{n_1,u}-Q_{n_1,u}}{n_1}Q^{-1}_{n_1,k_1}\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_{j_2},n}}{n}\right| \\
&\leq \frac{\sup_{\{\gamma_k\in R^k:\ \gamma'_k\gamma_k=1\}}\left[\gamma'_k(\widehat{Q}_{n_1,u}-Q_{n_1,u})\gamma_k\right]}{n_1}\sqrt{\frac{\phi_{w\theta_{j_1},n}P_{n,k_1}}{n}Q^{-2}_{n_1,k_1}\frac{P'_{n,k_1}\phi'_{w\theta_{j_1},n}}{n}} \\
&\times\sqrt{\frac{\widehat{\phi}_{w\theta_{j_2},n}P_{n,k_1}}{n}Q^{-2}_{n_1,k_1}\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_{j_2},n}}{n}} \\
&\leq \frac{\sup_{\{\gamma_k\in R^k:\ \gamma'_k\gamma_k=1\}}\left[\gamma'_k(\widehat{Q}_{n_1,u}-Q_{n_1,u})\gamma_k\right]}{n_1\lambda^2_{\min}(Q_{n_1,k_1})} \\
&\times\sqrt{\frac{\phi_{w\theta_{j_1},n}P_{n,k_1}}{n}\frac{P'_{n,k_1}\phi'_{w\theta_{j_1},n}}{n}}\sqrt{\frac{\widehat{\phi}_{w\theta_{j_2},n}P_{n,k_1}}{n}\frac{P'_{n,k_1}\widehat{\phi}'_{w\theta_{j_2},n}}{n}} = o_p(n_1^{-1})
\end{aligned} \tag{2.229}$$

where the equality is by (2.75), Lemma 2.9.6(iv), (2.225) and (2.226). Similarly for any $j_1 = 1,\ldots,d_\theta$

and any $j_2 = 1, \ldots, d_\theta$,

$$\left| \frac{\phi_{w\theta_{j_1},n} P_{n,k_1}}{n} Q_{n_1,k_1}^{-1} \frac{Q_{n_1,u}}{n_1} Q_{n_1,k_1}^{-1} \frac{P'_{n,k_1} \widehat{\phi}'_{w\theta_{j_2},n} - P'_{n,k_1} \phi_{w\theta_{j_2},n}}{n} \right|$$
$$\leq \sqrt{\frac{\phi_{w\theta_{j_1},n} P_{n,k_1}}{n} Q_{n_1,k_1}^{-1} \frac{Q_{n_1,u}}{n_1} Q_{n_1,k_1}^{-2} \frac{Q_{n_1,u}}{n_1} Q_{n_1,k_1}^{-1} \frac{P'_{n,k_1} \phi'_{w\theta_{j_1},n}}{n}} \left\| n^{-1} P'_{k_1,n} (\widehat{\phi}_{\theta_{j_2},n} - \phi_{\theta_{j_2},n}) \right\|$$
$$\leq \frac{\lambda_{\max}(Q_{n_1,u})}{n_1 \lambda^2_{\min}(Q_{n_1,k_1})} \sqrt{\frac{\phi_{w\theta_{j_1},n} P_{n,k_1}}{n} \frac{P'_{n,k_1} \phi'_{w\theta_{j_1},n}}{n}} \left\| n^{-1} P'_{k_1,n} (\widehat{\phi}_{\theta_{j_2},n} - \phi_{\theta_{j_2},n}) \right\| = o_p(n_1^{-1}) \tag{2.230}$$

where the equality is by (2.75), (2.134), Lemma 2.9.6(ii) and (2.225). Collecting the results in (2.224), (2.228), (2.229) and (2.230), we get

$$\widehat{\Sigma}_{n_1} - \Sigma_{n_1} = o_p(n_1^{-1}). \tag{2.231}$$

Using similar arguments in proving (2.231), we can show that

$$\widehat{\Sigma}_{n_2} - \Sigma_{n_2} = o_p(n_2^{-1}). \tag{2.232}$$

By (2.145), (2.146), (2.231) and (2.232),

$$\widehat{\Sigma}_{n_1} + \widehat{\Sigma}_{n_2} = (\Sigma_{n_1} + \Sigma_{n_2})(1 + o_p(1)). \tag{2.233}$$

By (2.232), we deduce that

$$H_0^{-1}(\Sigma_{n_1} + \Sigma_{n_2}) H_0^{-1} \widehat{H}_n (\widehat{\Sigma}_{n_1} + \widehat{\Sigma}_{n_2})^{-1} \widehat{H}_n$$
$$= H_0^{-1}(\Sigma_{n_1} + \Sigma_{n_2}) H_0^{-1} \widehat{H}_n (\Sigma_{n_1} + \Sigma_{n_2})^{-1} \widehat{H}_n (1 + o_p(1))$$
$$= H_0^{-1}(\Sigma_{n_1} + \Sigma_{n_2}) H_0^{-1} (\widehat{H}_n - H_0)(\Sigma_{n_1} + \Sigma_{n_2})^{-1} \widehat{H}_n (1 + o_p(1))$$
$$+ H_0^{-1}(\widehat{H}_n - H_0)(1 + o_p(1)) + I_{d_\theta}(1 + o_p(1))$$
$$= I_{d_\theta} + H_0^{-1}(\Sigma_{n_1} + \Sigma_{n_2}) H_0^{-1} (\widehat{H}_n - H_0)(\Sigma_{n_1} + \Sigma_{n_2})^{-1} H_0 (1 + o_p(1)) + o_p(1) \tag{2.234}$$

where the last equality is by Lemma 2.9.6(i). Using (2.75), (2.134), (2.135) and Lemma 2.9.3, we have

$$\lambda_{\max}(n_1 \Sigma_{n_1}) \leq C \text{ and } \lambda_{\max}(n_2 \Sigma_{n_2}) \leq C \tag{2.235}$$

with probability approaching 1. For any $\gamma_1, \gamma_2 \in R^{d_\theta}$ with $\gamma_1' \gamma_1 = 1$ and $\gamma_2' \gamma_2 = 1$, by the Cauchy-Schwarz inequality we have

$$\left| \gamma_1' \Sigma_{n_1} H_0^{-1} (\widehat{H}_n - H_0)(\Sigma_{n_1} + \Sigma_{n_2})^{-1} H_0 \gamma_2 \right|$$
$$\leq (\gamma_1'(n_1 \Sigma_{n_1}) H_0^{-1} (\widehat{H}_n - H_0)^2 H_0^{-1} (n_1 \Sigma_{n_1}) \gamma_1)^{1/2} (\gamma_2' H_0 (n_1 \Sigma_{n_1} + n_1 \Sigma_{n_2})^{-2} H_0 \gamma_2)^{1/2}$$
$$\leq \frac{C \lambda_{\max}(n_1 \Sigma_{n_1}) ||\widehat{H}_n - H_0||}{\lambda_{\min}(H_0) \lambda_{\min}(n_1 \Sigma_{n_1} + n_1 \Sigma_{n_2})} = o_p(1) \tag{2.236}$$

where the equality is by Assumptions 2.3.3(ii), Lemma 2.9.6(i), (2.145) and (2.235). Similarly, we can show that

$$\left|\gamma_1'\Sigma_{n_2}H_0^{-1}(\widehat{H}_n - H_0)(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_0\gamma_2\right| = o_p(1). \tag{2.237}$$

Let γ_1' be any row of H_0^{-1} and γ_2 be any column of H_0. Then we can use (2.236) and (2.237) to deduce that

$$H_0^{-1}(\Sigma_{n_1}+\Sigma_{n_2})H_0^{-1}(\widehat{H}_n - H_0)(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_0 = o_p(1) \tag{2.238}$$

which together with (2.234) implies that

$$\widehat{H}_n(\widehat{\Sigma}_{n_1}+\widehat{\Sigma}_{n_2})^{-1}\widehat{H}_n = (H_0(\Sigma_{n_1}+\Sigma_{n_2})^{-1}H_0)(I_{d_\theta}+o_p(1)). \tag{2.239}$$

This shows (2.52). Using (2.52) and Theorem 2.3.2, and then applying CMT, we immediately prove the claim of the theorem. ■

Proof of Lemma 2.6.1. By definition,

$$\widehat{u}_i = Y_i - \widehat{h}_{n_1}(Z_i) = u_i + h_0(Z_i) - \widehat{h}_{n_1}(Z_i) \tag{2.240}$$

which implies that

$$\widehat{u}_i^2 = u_i^2 + (h_0(Z_i) - \widehat{h}_{n_1}(Z_i))^2 + 2u_i(h_0(Z_i) - \widehat{h}_{n_1}(Z_i)). \tag{2.241}$$

Hence,

$$\begin{aligned}\widehat{\sigma}^2_{n,u}(z) &= n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_i^2P_{k_1}(Z_i) + n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}(h_0(Z_i)-\widehat{h}_{n_1}(Z_i))^2P_{k_1}(Z_i)\\ &\quad + 2n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-\widehat{h}_{n_1}(Z_i))P_{k_1}(Z_i). \end{aligned} \tag{2.242}$$

Let $\widehat{\beta}_{u,n_1} = n_1^{-1}Q_{k,n_1}^{-1}\sum_{i\in I_1}u_i^2P_k(Z_i)$. By Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.3.3(v) and 2.6.1(i), we can use similar arguments in showing (2.76) to deduce that

$$\left\|\widehat{\beta}_{u,n_1} - \beta_{u,k}\right\|^2 = O_p(k_1n_1^{-1} + k_1^{-2r_u}). \tag{2.243}$$

By (2.243) and Assumptions 2.6.1(ii),

$$\sup_{z\in\mathcal{Z}}\left|P_{k_1}'(z)\widehat{\beta}_{u,n_1} - \sigma_u^2(z)\right| = O_p(\xi_{k_1}k_1^{1/2}n_1^{-1/2} + \xi_{k_1}k_1^{-r_u}). \tag{2.244}$$

By the triangle inequality, the Cauchy-Schwarz inequality, (2.75) and (2.77),

$$\begin{aligned}&\left|n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}(h_0(Z_i)-\widehat{h}_{n_1}(Z_i))^2P_k(Z_i)\right|\\ &\leq \sup_{z\in\mathcal{Z}}(P_{k_1}'(z)Q_{n_1,k_1}^{-1}P_{k_1}(z))^2n_1^{-1}\sum_{i\in I_1}(h_0(Z_i)-\widehat{h}_{n_1}(Z_i))^2\\ &= O_p(\xi_{k_1}^2k_1n_1^{-1} + \xi_{k_1}^2k_1^{-2r_h}). \end{aligned} \tag{2.245}$$

By definition,

$$\begin{aligned}
&n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-\widehat{h}_{n_1}(Z_i))P_k(Z_i)\\
&=n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-h_{0,k_1}(Z_i))P_k(Z_i)\\
&+n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_iP_k(Z_i)P_k'(Z_i)(\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1}).
\end{aligned}\tag{2.246}$$

By Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(iv) and 2.3.3(v),

$$\begin{aligned}
&E\left[\left\|n_1^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-h_{0,k_1}(Z_i))P_{k_1}(Z_i)\right\|^2\right]\\
&=n_1^{-1}E\left[u^2(h_0(Z)-h_{0,k_1}(Z))^2P_{k_1}(Z)'P_{k_1}(Z)\right]\\
&\leq Cn_1^{-1}\sup_{z\in\mathcal{Z}}|h_0(z)-h_{0,k_1}(z)|^2E\left[P_{k_1}(Z)'P_{k_1}(Z)\right]=O(k_1^{1-2r_h}n_1^{-1})
\end{aligned}\tag{2.247}$$

which together with the Markov inequality implies that

$$n_1^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-h_{0,k_1}(Z_i))P_{k_1}(Z_i)=O_p(k_1^{1/2-r_h}n_1^{-1/2}).\tag{2.248}$$

By (2.75), (2.77), (2.248) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left|n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-h_{0,k_1}(Z_i))P_k(Z_i)\right|\\
&\leq\sup_{z\in\mathcal{Z}}(P_{k_1}'(z)Q_{n_1,k_1}^{-2}P_{k_1}(z))^{1/2}\left\|n_1^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-h_{0,k_1}(Z_i))P_{k_1}(Z_i)\right\|\\
&=O_p(\xi_{k_1}n_1^{-1/2}k_1^{1/2-r_h}).
\end{aligned}\tag{2.249}$$

By Assumptions 2.3.1(i), 2.3.1(iii) and 2.3.3(vi),

$$E\left[\left\|n_1^{-1}\sum_{i\in I_1}u_iP_k(Z_i)P_k'(Z_i)\right\|^2\right]=\frac{E\left[u^2(P_{k_1}(Z)'P_{k_1}(Z))^2\right]}{n_1}\leq\frac{C\xi_{k_1}^2E\left[P_{k_1}(Z)'P_{k_1}(Z)\right]}{n_1}=C\xi_{k_1}^2k_1n_1^{-1}$$

which together with the Markov inequality implies that

$$\left\|n_1^{-1}\sum_{i\in I_1}u_iP_k(Z_i)P_k'(Z_i)\right\|=O_p(\xi_{k_1}k_1^{1/2}n_1^{-1/2}).\tag{2.250}$$

By (2.75), (2.76), (2.250) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_iP_k(Z_i)P_k'(Z_i)(\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1})\right|\\
&\leq\sup_{z\in\mathcal{Z}}(P_{k_1}'(z)Q_{n_1,k_1}^{-2}P_{k_1}(z))^{1/2}\left\|n_1^{-1}\sum_{i\in I_1}u_iP_k(Z_i)P_k'(Z_i)\right\|\left\|\widehat{\beta}_{k_1,n_1}-\beta_{h,k_1}\right\|\\
&=O_p(\xi_{k_1}^2k_1^{1/2}n_1^{-1/2}(k_1^{1/2}n_1^{-1/2}+k_1^{-r_h})).
\end{aligned}\tag{2.251}$$

Combining the results in (2.246), (2.249) and (2.251), we get

$$\sup_{z\in\mathcal{Z}}\left|n_1^{-1}P_{k_1}'(z)Q_{n_1,k_1}^{-1}\sum_{i\in I_1}u_i(h_0(Z_i)-\widehat{h}_{n_1}(Z_i))P_k(Z_i)\right| = O_p(\xi_{k_1}^2k_1n_1^{-1}+\xi_{k_1}^2k_1^{-2r_h}), \tag{2.252}$$

which together with (2.242), (2.244) and (2.245) implies that

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,u}^2(z)-\sigma_u^2(z)\right| = O_p(\xi_{k_1}(k_1^{1/2}n_1^{-1/2}+k_1^{-r_u})+\xi_{k_1}^2(k_1n_1^{-1}+k_1^{-2r_h})). \tag{2.253}$$

This together with Assumptions 2.6.1(iv) proves the first claim of the lemma.

By definition,

$$\widetilde{\varepsilon}_i = g(X_i,\widehat{\theta}_{1,n})-\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n}) = \varepsilon_i+(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))-(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n})-\phi(Z_i,\theta_0)).$$

Hence,

$$\begin{aligned}\widehat{\sigma}_{n,\varepsilon}^2(z) &= n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}\varepsilon_i^2P_{k_2}(Z_i)+n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))^2P_{k_2}(Z_i)\\
&+n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n})-\phi(Z_i,\theta_0))^2P_{k_2}(Z_i)\\
&+2n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}\varepsilon_i(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))P_{k_2}(Z_i)\\
&-2n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}\varepsilon_i(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n})-\phi(Z_i,\theta_0))P_{k_2}(Z_i)\\
&-2n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n})-\phi(Z_i,\theta_0))P_{k_2}(Z_i).\end{aligned} \tag{2.254}$$

Using similar arguments in showing (2.244), we can show that

$$\sup_{z\in\mathcal{Z}}\left|P_{k_2}'(z)\widehat{\beta}_{\varepsilon,n_2}-\sigma_\varepsilon^2(z)\right| = O_p(\xi_{k_2}k_2^{1/2}n_2^{-1/2}+\xi_{k_2}k_2^{-r_\varepsilon}). \tag{2.255}$$

By the Cauchy-Schwarz inequality, (2.75) and (2.187)

$$\begin{aligned}&\left|n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))^2P_{k_2}(Z_i)\right|\\
&\le \sup_{z\in\mathcal{Z}}(P_{k_2}'(z)Q_{n_2,k_2}^{-1}P_{k_2}(z))n_2^{-1}\sum_{i\in I_2}(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))^2\\
&= O_p(\xi_{k_2}^2n_1^{-1}+\xi_{k_2}^2n_2^{-1}).\end{aligned} \tag{2.256}$$

Similarly, by the Cauchy-Schwarz inequality, (2.75) and (2.190)

$$\begin{aligned}&\left|n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n})-\phi(Z_i,\theta_0))^2P_{k_2}(Z_i)\right|\\
&\le \sup_{z\in\mathcal{Z}}(P_{k_2}'(z)Q_{n_2,k_2}^{-1}P_{k_2}(z))n_2^{-1}\sum_{i\in I_2}(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n})-\phi(Z_i,\theta_0))^2\\
&= O_p(\xi_{k_2}^2n_1^{-1}+\xi_{k_2}^2k_2n_2^{-1}+\xi_{k_2}^2k_2^{-2r_h}).\end{aligned} \tag{2.257}$$

By the Cauchy-Schwarz inequality, (2.75), (2.187) and (2.190)

$$\begin{aligned}
&\left| n_2^{-1} P_{k_2}'(z) Q_{n_2,k_2}^{-1} \sum_{i\in I_2} (g(Z_i,\widehat{\theta}_{1,n}) - g(Z_i,\theta_0))(\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n}) - \phi(Z_i,\theta_0)) P_{k_2}(Z_i) \right|^2 \\
&\leq \sup_{z\in\mathcal{Z}} (P_{k_2}'(z) Q_{n_2,k_2}^{-1} P_{k_2}(z))^2 n_2^{-1} \sum_{i\in I_2} (g(Z_i,\widehat{\theta}_{1,n}) - g(Z_i,\theta_0))^2 \\
&\times n_2^{-1} \sum_{i\in I_2} (\widehat{\phi}_{n_2}(Z_i,\widehat{\theta}_{1,n}) - \phi(Z_i,\theta_0))^2 \\
&= O_p(\xi_{k_2}^4 (n_1^{-1} + n_2^{-1})(n_1^{-1} + k_2 n_2^{-1} + k_2^{-2r_h})). \qquad (2.258)
\end{aligned}$$

By the second order expansion,

$$g(X_i,\widehat{\theta}_{1,n}) - g(X_i,\theta_0) = g_\theta(X_i,\theta_0)'(\widehat{\theta}_{1,n} - \theta_0) + 2^{-1}(\widehat{\theta}_{1,n} - \theta_0)' g_{\theta\theta}(X_i,\widetilde{\theta}_{i,n})(\widehat{\theta}_{1,n} - \theta_0) \qquad (2.259)$$

where $\widetilde{\theta}_{i,n}$ is between $\widehat{\theta}_n$ and θ_0. By Assumption 2.3.1(i), 2.3.3(v) and 2.6.1(v)

$$\begin{aligned}
E\left[\left\| n_2^{-1} \sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) g_\theta(X_i,\theta_0)' \right\|^2\right] &= n_2^{-1} E\left[\varepsilon_i^2 \left\| P_{k_2}(Z_i) g_\theta(X_i,\theta_0)' \right\|^2\right] \\
&\leq n_2^{-1} \xi_{k_2}^2 E\left[\varepsilon_i^2 \left\| g_\theta(X_i,\theta_0) \right\|^2\right] \\
&\leq n_2^{-1} \xi_{k_2}^2 \sqrt{E\left[\varepsilon_i^4\right] E\left[\left\| g_\theta(X_i,\theta_0) \right\|^4\right]} = O(\xi_{k_2}^2 n_2^{-1}) \qquad (2.260)
\end{aligned}$$

which together with the Markov inequality implies that

$$n_2^{-1} \sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) g_\theta(X_i,\theta_0)' = O_p(\xi_{k_2} n_2^{-1/2}). \qquad (2.261)$$

By (2.57), (2.75), the triangle inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \left| n_2^{-1} P_{k_2}'(z) Q_{n_2,k_2}^{-1} \sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) g_\theta(X_i,\theta_0)'(\widehat{\theta}_{1,n} - \theta_0) \right| \\
&\leq \sup_{z\in\mathcal{Z}} (P_{k_2}'(z) Q_{n_2,k_2}^{-2} P_{k_2}(z))^{1/2} \left\| n_2^{-1} \sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) g_\theta(X_i,\theta_0)' \right\| \left\| \widehat{\theta}_{1,n} - \theta_0 \right\| \\
&= O_p(\xi_{k_2}^2 n_2^{-1/2} (n_1^{-1/2} + n_2^{-1/2})). \qquad (2.262)
\end{aligned}$$

By Assumptions 2.3.1(i), 2.3.3(i) and 2.3.3(v), and the Cauchy-Schwarz inequality,

$$\max_{i\in I_2} \left\| n_2^{-1} \sum_{i\in I_2} |\varepsilon_i| g_{\theta\theta}(X_i,\widetilde{\theta}_{i,n}) \right\|^2 \leq n_2^{-1} \sum_{i\in I_2} \varepsilon_i^2 \times \sup_{\theta\in\mathcal{N}_n} n_2^{-1} \sum_{i\in I_2} \left\| g_{\theta\theta}(X_i,\theta) \right\|^2 = O_p(1). \qquad (2.263)$$

By (2.57), the triangle inequality and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left|n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}\varepsilon_i P_{k_2}(Z_i)(\widehat{\theta}_{1,n}-\theta_0)'g_{\theta\theta}(X_i,\widetilde{\theta}_{i,n})(\widehat{\theta}_{1,n}-\theta_0)\right| \\
&\le \sup_{z\in\mathcal{Z}}(P_{k_2}'(z)Q_{n_2,k_2}^{-1}P_{k_2}(z))\max_{i\in I_2}\left\|n_2^{-1}\sum_{i\in I_2}|\varepsilon_i|g_{\theta\theta}(X_i,\widetilde{\theta}_{i,n})\right\|\left\|\widehat{\theta}_{1,n}-\theta_0\right\|^2 \\
&= O_p(\xi_{k_2}^2(n_1^{-1}+n_2^{-1})). \qquad (2.264)
\end{aligned}
$$

Combining the results in (2.259), (2.262) and (2.254), we get

$$
\sup_{z\in\mathcal{Z}}\left|n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}\varepsilon_i(g(X_i,\widehat{\theta}_{1,n})-g(X_i,\theta_0))P_{k_2}(Z_i)\right| = O_p(\xi_{k_2}^2(n_1^{-1}+n_2^{-1})). \qquad (2.265)
$$

By (2.75), (2.188), Assumptions 2.3.1(i), 2.3.1(iv) and 2.3.3(v),

$$
\begin{aligned}
&E\left[\left\|n_2^{-1}\sum_{i\in I_2}\varepsilon_i(\widehat{\phi}_{n_2}(Z_i,\theta_0)-\phi(Z_i,\theta_0))P_{k_2}(Z_i)\right\|^2\middle|\{Z_i\}_{i\in I_2}\right] \\
&= n_2^{-2}\sum_{i\in I_2}\sigma_\varepsilon^2(Z_i)(\widehat{\phi}_{n_2}(Z_i,\theta_0)-\phi(Z_i,\theta_0))^2P_{k_2}(Z_i)'P_{k_2}(Z_i) \\
&\le C\xi_{k_2}^2n_2^{-2}\sum_{i\in I_2}(\widehat{\phi}_{n_2}(Z_i,\theta_0)-\phi(Z_i,\theta_0))^2 = O_p(\xi_{k_2}^2n_2^{-1}(k_2n_2^{-1}+k_2^{-2r_h})), \qquad (2.266)
\end{aligned}
$$

which together with (2.75), the Markov inequality, the triangle inequality and the Cauchy-Schwarz inequality implies that

$$
\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left|n_2^{-1}P_{k_2}'(z)Q_{n_2,k_2}^{-1}\sum_{i\in I_2}\varepsilon_i(\widehat{\phi}_{n_2}(Z_i,\theta_0)-\phi(Z_i,\theta_0))P_{k_2}(Z_i)\right| \\
&\le \sup_{z\in\mathcal{Z}}(P_{k_2}'(z)Q_{n_2,k_2}^{-2}P_{k_2}(z))^{1/2}\left\|n_2^{-1}\sum_{i\in I_2}\varepsilon_i(\widehat{\phi}_{n_2}(Z_i,\theta_0)-\phi(Z_i,\theta_0))P_{k_2}(Z_i)\right\| \\
&= O_p(\xi_{k_2}^2n_2^{-1/2}(k_2^{1/2}n_2^{-1/2}+k_2^{-r_h})). \qquad (2.267)
\end{aligned}
$$

By Assumptions 2.3.1(i) and 2.3.1(iii),

$$
E\left[\left\|n_2^{-1}\sum_{i\in I_2}\varepsilon_iP_{k_2}(Z_i)P_{k_2}(Z_i)'\right\|^2\right] = n_2^{-1}E\left[\sigma_\varepsilon^2(Z_i)|P_{k_2}(Z_i)'P_{k_2}(Z_i)|^2\right] = O(\xi_{k_2}^2k_2n_2^{-1}), \qquad (2.268)
$$

which together with the Markov inequality implies that

$$
\left\|n_2^{-1}\sum_{i\in I_2}\varepsilon_iP_{k_2}(Z_i)P_{k_2}(Z_i)'\right\| = O_p(\xi_{k_2}k_2^{1/2}n_2^{-1/2}). \qquad (2.269)
$$

Recall that $g_{n_2}(\widehat{\theta}_{1,n}) = (g(X_i, \widehat{\theta}_{1,n}))'_{i\in I_2}$ and $g_{n_2}(\theta_0) = (g(X_i, \theta_0))'_{i\in I_2}$. By (2.75), (2.187), (2.269), the triangle inequality and the Cauchy-Schwarz inequality

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left| n_2^{-1} P'_{k_2}(z) Q^{-1}_{n_2,k_2} \sum_{i\in I_2} \varepsilon_i (\widehat{\phi}_{n_2}(Z_i, \widehat{\theta}_{1,n}) - \widehat{\phi}_{n_2}(Z_i, \theta_0)) P_{k_2}(Z_i) \right| \\
&= \left| n_2^{-1} P'_{k_2}(z) Q^{-1}_{n_2,k_2} \sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) P_{k_2}(Z_i)' (P'_{n_2,k_2} P_{n_2,k_2})^{-1} P'_{n_2,k_2} (g_{n_2}(\widehat{\theta}_{1,n}) - g_{n_2}(\theta_0)) \right| \\
&\leq \sup_{z\in\mathcal{Z}} (P'_{k_2}(z) Q^{-2}_{n_2,k_2} P_{k_2}(z))^{1/2} \left\| n_2^{-1} \sum_{i\in I_2} \varepsilon_i P_{k_2}(Z_i) P_{k_2}(Z_i)' \right\| \\
&\times \sqrt{(g_{n_2}(\widehat{\theta}_{1,n}) - g_{n_2}(\theta_0))' P_{n_2,k_2} (P'_{n_2,k_2} P_{n_2,k_2})^{-2} P'_{n_2,k_2} (g_{n_2}(\widehat{\theta}_{1,n}) - g_{n_2}(\theta_0))} \\
&= O_p(\xi^2_{k_2} k_2^{1/2} n_2^{-1/2}) \sqrt{n_2^{-1} \sum_{i\in I_2} (g(X_i, \widehat{\theta}_{1,n}) - g(X_i, \theta_0))^2} \qquad (2.270) \\
&= O_p(\xi^2_{k_2} k_2^{1/2} n_2^{-1/2} (n_1^{-1/2} + n_2^{-1/2})).
\end{aligned}$$

Combining the results in (2.267) and (2.270), we get

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left| n_2^{-1} P'_{k_2}(z) Q^{-1}_{n_2,k_2} \sum_{i\in I_2} \varepsilon_i (\widehat{\phi}_{n_2}(Z_i, \widehat{\theta}_{1,n}) - \phi(Z_i, \theta_0)) P_{k_2}(Z_i) \right| \\
&= O_p(\xi^2_{k_2} k_2^{1/2} n_2^{-1} + \xi^2_{k_2} n_2^{-1/2} k_2^{-r_h} + \xi^2_{k_2} k_2^{1/2} n_2^{-1/2} n_1^{-1/2}). \qquad (2.271)
\end{aligned}$$

Combining the results in (2.254), (2.255), (2.256), (2.257), (2.258), (2.265) and (2.271), and then applying Assumption 2.6.1(iv), we get

$$\sup_{z\in\mathcal{Z}} \left| \widehat{\sigma}^2_{n,\varepsilon}(z) - \sigma^2_\varepsilon(z) \right| = O_p(\xi^2_{k_2} n_1^{-1} + \xi^2_{k_2} k_2^{-2r_h} + \xi_{k_2} k_2^{1/2} n_2^{-1/2} + \xi_{k_2} k_2^{-r_\varepsilon}).$$

■

Proof of Theorem 2.6.2. By the triangle inequality,

$$\begin{aligned}
|\widehat{w}^*_n(z) - w^*(z)| &= (n_1^{-1} + n_2^{-1}) \left| (n_1^{-1} \widehat{\sigma}^2_{n,u}(z) + n_2^{-1} \widehat{\sigma}^2_{n,\varepsilon}(z))^{-1} - (n_1^{-1} \sigma^2_u(z) + n_2^{-1} \sigma^2_\varepsilon(z)) \right| \\
&\leq (n_1^{-1} + n_2^{-1}) \frac{n_1^{-1} \left| \widehat{\sigma}^2_{n,u}(z) - \sigma^2_u(z) \right| + n_2^{-1} \left| \widehat{\sigma}^2_{n,\varepsilon}(z) - \sigma^2_\varepsilon(z) \right|}{\left| (n_1^{-1} \widehat{\sigma}^2_{n,u}(z) + n_2^{-1} \widehat{\sigma}^2_{n,\varepsilon}(z))(n_1^{-1} \sigma^2_u(z) + n_2^{-1} \sigma^2_\varepsilon(z)) \right|} \\
&\leq (n_1^{-1} + n_2^{-1})^2 \frac{\left| \widehat{\sigma}^2_{n,u}(z) - \sigma^2_u(z) \right| + \left| \widehat{\sigma}^2_{n,\varepsilon}(z) - \sigma^2_\varepsilon(z) \right|}{\left| (n_1^{-1} \widehat{\sigma}^2_{n,u}(z) + n_2^{-1} \widehat{\sigma}^2_{n,\varepsilon}(z))(n_1^{-1} \sigma^2_u(z) + n_2^{-1} \sigma^2_\varepsilon(z)) \right|}.
\end{aligned}$$

(2.272)

By Lemma 2.6.1 and the triangle inequality

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left|n_1^{-1}\widehat{\sigma}_{n,u}^2(z)+n_2^{-1}\widehat{\sigma}_{n,\varepsilon}^2(z)-(n_1^{-1}\sigma_u^2(z)+n_2^{-1}\sigma_\varepsilon^2(z))\right| \\
&\leq n_1^{-1}\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,u}^2(z)-\sigma_u^2(z)\right|+n_2^{-1}\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,\varepsilon}^2(z)-\sigma_\varepsilon^2(z)\right| \\
&= O_p(\delta_{w,n}(n_1^{-1}+n_2^{-1})). \qquad (2.273)
\end{aligned}$$

By Assumption 2.3.3(v),

$$n_1^{-1}\sigma_u^2(z)+n_2^{-1}\sigma_\varepsilon^2(z)\geq(n_1^{-1}+n_2^{-1})C^{-1} \qquad (2.274)$$

for any $z\in\mathcal{Z}$, which together with (2.273) and $\delta_{w,n}=o(1)$ implies that

$$n_1^{-1}\widehat{\sigma}_{n,u}^2(z)+n_2^{-1}\widehat{\sigma}_{n,\varepsilon}^2(z)=(n_1^{-1}\sigma_u^2(z)+n_2^{-1}\sigma_\varepsilon^2(z))(1+o_p(1)) \qquad (2.275)$$

uniformly over $z\in\mathcal{Z}$. Combining Lemma 2.6.1, the results in (2.272), (2.273) and (2.275), we have

$$\begin{aligned}
\sup_{z\in\mathcal{Z}}\left|\widehat{w}_n^*(z)-w^*(z)\right| &\leq (n_1^{-1}+n_2^{-1})^2\frac{\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,u}^2(z)-\sigma_u^2(z)\right|+\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,\varepsilon}^2(z)-\sigma_\varepsilon^2(z)\right|}{\inf_{z\in\mathcal{Z}}(n_1^{-1}\sigma_u^2(z)+n_2^{-1}\sigma_\varepsilon^2(z))^2(1+o_p(1))} \\
&\leq \sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,u}^2(z)-\sigma_u^2(z)\right|+\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}_{n,\varepsilon}^2(z)-\sigma_\varepsilon^2(z)\right|=O_p(\delta_{w,n}) \qquad (2.276)
\end{aligned}$$

which finishes the proof. ■

2.10 Low-level Sufficient Conditions

In this section, we provide low-level sufficient conditions for Assumptions 2.3.2(i), 2.3.2(ii), 2.3.2(iv), 2.3.3(i), 2.3.3(iii)-(iv), 2.6.1(i) and 2.6.1(v).

Assumption 2.10.1 *(i) For any θ, there exist $\beta_{\theta,k}\in R^k$ and $r_\varphi>0$, such that*

$$\sup_{z\in\mathcal{Z}}\left|\varphi(z,\theta)-P_k(z)'\beta_{\theta,k}\right|=O(k^{-r_\varphi})$$

uniformly over $\theta\in\Theta$; (ii) $\sup_{x,\theta}\left[\|g(x,\theta)\|+\|g_\theta(x,\theta)\|+\|g_{\theta\theta}(x,\theta)\|\right]\leq C$; (iii) Θ is a compact subspace of R^{d_θ}; (iv) there exist $\beta_{\varphi_\theta.j,k}\in R^k$ and $r_{\varphi_\theta,j}>0$, such that

$$\sup_{z\in\mathcal{Z}}\left|\varphi_{\theta_j}(z,\theta_0)-P_k(z)'\beta_{\varphi_\theta.j,k}\right|=O(k^{-r_{\varphi_\theta,j}})$$

for any $j=1,\ldots,d_\theta$; (v) $\max_{j=1,\ldots,d_\theta}n_2^{-1/4}k_2^{-r_{\varphi_\theta,j}}=o(1)$.

It is clear that Assumption 2.10.1(ii) implies that Assumptions 2.3.2(i), 2.3.3(i), 2.3.3(iv), 2.6.1(i) and 2.6.1(v) hold. In the rest of the section, we verify Assumptions 2.3.2(ii), 2.3.2(iv) and 2.3.3(iii).

Lemma 2.10.1 *Under Assumptions 2.3.1(i), 2.3.3(v), 2.3.3(vii), 2.10.1(ii) and 2.10.1(iii),*

$$\sup_{\gamma_{k_2}\in U_{k_2}}\left|n_2^{-1}\sum_{i\in I_2}\gamma_{k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta)-\varphi(Z_i,\theta)\right)\right| = O_p(\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}),$$

uniformly over $\theta\in\Theta$, where $U_{k_2}=\{\gamma_{k_2}\in R^{k_2}:\gamma_{k_2}'\gamma_{k_2}=1\}$

Proof of Lemma 2.10.1. For any $\gamma_{1,k_2},\gamma_{2,k_2}\in U_{k_2}$ and any $\theta_1,\theta_2\in\Theta$, using the triangle inequality we get

$$\begin{aligned}
&\left|\gamma_{1,k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)\right)-\gamma_{2,k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta_2)-\varphi(Z_i,\theta_2)\right)\right|\\
&\leq\left|(\gamma_{1,k_2}-\gamma_{2,k_2})'P_{k_2}(Z_i)\left(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)\right)\right|\\
&+\left|\gamma_{2,k_2}'P_{k_2}(Z_i)(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)-g(X_i,\theta_2)+\varphi(Z_i,\theta_2))\right|. \qquad (2.277)
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|(\gamma_{1,k_2}-\gamma_{2,k_2})'P_{k_2}(Z_i)\left(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)\right)\right|\\
&\leq\|\gamma_{1,k_2}-\gamma_{2,k_2}\|\sqrt{\left(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)\right)^2P_{k_2}'(Z_i)P_{k_2}(Z_i)}\\
&\leq C\xi_{k_1}\|\gamma_{1,k_1}-\gamma_{2,k_1}\| \qquad (2.278)
\end{aligned}$$

where the second inequality is by Assumptions 2.10.1(ii) and the definition of $\varphi(z,\theta)$. By the mean value theorem,

$$g(X_i,\theta_1)-g(X_i,\theta_2)=g_\theta(X_i,\widetilde{\theta}_i)'(\theta_1-\theta_2), \qquad (2.279)$$

which together with the Cauchy-Schwarz inequality and Assumptions 2.10.1(ii) implies that for any x,

$$|g(x,\theta_1)-g(x,\theta_2)|\leq C\|\theta_1-\theta_2\|. \qquad (2.280)$$

Similarly, we can show that for any z,

$$|\varphi(z,\theta_1)-\varphi(z,\theta_2)|\leq C\|\theta_1-\theta_2\|. \qquad (2.281)$$

By the Cauchy-Schwarz inequality, the triangle inequality, (2.280) and (2.281),

$$\left|\gamma_{2,k_2}'P_{k_2}(Z_i)(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)-g(X_i,\theta_2)+\varphi(Z_i,\theta_2))\right|\leq C\|\theta_1-\theta_2\|. \qquad (2.282)$$

Combining the results in (2.277), (2.278) and (2.282), we get

$$\begin{aligned}
&\left|\gamma_{1,k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta_1)-\varphi(Z_i,\theta_1)\right)-\gamma_{2,k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta_2)-\varphi(Z_i,\theta_2)\right)\right|\\
&\leq C\xi_{k_1}\left[\|\gamma_{1,k_1}-\gamma_{2,k_1}\|+\|\theta_1-\theta_2\|\right]. \qquad (2.283)
\end{aligned}$$

Let γ_{m_0,k_2} $(m_0 = 1,\ldots,M_{\gamma,n})$ be a set of points such that $\min_{m_0\leq M_{\gamma,n}}\|\gamma_{k_2}-\gamma_{m_0,k_2}\| \leq C^{-1}\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}\xi_{k_2}^{-1}$ for any $\gamma_{k_2}\in U_{k_2}$. Similarly, let θ_{m_1} $(m_1 = 1,\ldots,M_{\theta,n})$ be a set of points in $\times$ such that $\min_{m_1\leq M_{\theta,n}}\|\theta-\theta_{m1}\| \leq C^{-1}\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}\xi_{k_2}^{-1}$ for any $\theta\in\times$. As U_{k_2} is compact in R^{k_2}, we know that $M_{\gamma,n}\leq C(n_2^{1/2}\xi_{k_2}\log^{-1/2}(n_2)k_2^{-1/2})^{k_2}$. Similarly, $M_{\theta,n}\leq C(n_2^{1/2}\xi_{k_2}\log^{-1/2}(n_2)k_2^{-1/2})^{d_\theta}$, which implies that

$$M_{\gamma,n}M_{\theta,n}\leq C(n_2^{1/2}\xi_{k_2}\log^{-1/2}(n_2)k_2^{-1/2})^{k_2+d_\theta}. \tag{2.284}$$

Hence, by the triangle inequality,

$$\begin{aligned}
&\sup_{\theta\in\times,\ \gamma_{k_2}\in U_{k_2}}\left|n_2^{-1}\sum_{i\in I_2}\gamma_{k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta)-\varphi(Z_i,\theta)\right)\right| \\
&\leq 2C\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2} \\
&+\max_{m_0\leq M_{\gamma,n},m_1\leq M_{\theta,n}}\left|n_2^{-1}\gamma_{m_0,k_2}'\sum_{i\in I_2}P_{k_2}(Z_i)\left[g(X_i,\theta_{m_1})-\varphi(Z_i,\theta_{m_1})\right]\right|.
\end{aligned} \tag{2.285}$$

For any m_0 and m_1 and for any i, by the Cauchy-Schwarz inequality, and Assumption 2.10.1(i)

$$\left|n_2^{-1}\gamma_{m_0,k_2}'P_{k_2}(Z_i)\left[g(X_i,\theta_{m_1})-\varphi(Z_i,\theta_{m_1})\right]\right|\leq C\xi_{k_2}n_2^{-1}. \tag{2.286}$$

By Assumptions 2.3.1(i) and 2.3.3(v)

$$\begin{aligned}
&E\left[\left|n_2^{-1}\gamma_{m_0,k_2}'P_{k_2}(Z_i)\left[g(X_i,\theta_{m_1})-\varphi(Z_i,\theta_{m_1})\right]\right|^2\right] \\
&\leq Cn_2^{-2}\gamma_{m_0,k_2}'E[P_{k_2}(Z_i)P_{k_2}'(Z_i)]\gamma_{m_0,k_2}\leq C\lambda_{\max}(Q_{k_2})n_2^{-2}\leq Cn_2^{-2}.
\end{aligned} \tag{2.287}$$

By (2.286) and (2.287), we can apply the Bernstein inequality to get

$$\begin{aligned}
&\Pr\left(\left|n_2^{-1}\sum_{i\in I_2}\gamma_{m_0,k_2}'P_{k_2}(Z_i)(g(X_i,\theta_{m_1})-\varphi(Z_i,\theta_{m_1}))\right|>B\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}\right) \\
&\leq 2\exp\left[-\frac{B^2\log(n_2)k_2n_2^{-1}}{2C(n_2^{-1}+B\log^{1/2}(n_2)k_2^{1/2}n_2^{-3/2})}\right] \\
&=2\exp\left[-\frac{B^2\log(n_2)k_2}{2C\left(1+B\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}\right)}\right]\leq 2\exp\left[-\frac{B\log(n_2)k_2}{2C}\right],
\end{aligned} \tag{2.288}$$

where the last inequality is by Assumption 2.3.3(vii). (2.288) together with the Bonferroni inequality implies that

$$
\begin{aligned}
&\Pr\left(\max_{m_0\le M_{\gamma,n},m_1\le M_{\theta,n}}\left|n_2^{-1}\gamma'_{m_0,k_2}P'_{n_2,k_2}(g_{n_2}(\theta_{m_1})-\varphi_{n_2}(\theta_{m_1}))\right|>B\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}\right)\\
&\le 2M_{\gamma,n}M_{\theta,n}\exp\left[-\frac{B\log(n_2)k_2}{2C}\right]\\
&\le 2C(n_2^{1/2}\xi_{k_2}\log^{-1/2}(n_2)k_2^{-1/2})^{k_2+d_\theta}\exp\left[-\frac{B\log(n_2)k_2}{2C}\right]\\
&\le 2C\exp\left[-\frac{B\log(n_2)k_2}{2C}+(k_2+d_\theta)\left(2^{-1}\log(n_2)+\log(\xi_{k_2})-2^{-1}\log(k_2)\right)\right]\\
&\le 2C\exp\left[-\frac{B\log(n_2)k_2}{8C}\right],
\end{aligned}
\tag{2.289}
$$

where the last inequality is by the assumption that $k_2\to\infty$, as $n_2\to\infty$. As C is a fixed constant, from (2.289), we can choose B sufficiently large such that for any (fixed but) small $\varepsilon>0$, there is

$$
\Pr\left(\max_{m_0\le M_{\gamma,n},m_1\le M_{\theta,n}}\left|n_2^{-1}\gamma'_{m_0,k_2}P'_{n_2,k_2}(g_{n_2}(\theta_{m_1})-\varphi_{n_2}(\theta_{m_1}))\right|>B\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}\right)\le\varepsilon
$$

for all large n_2, which implies that

$$
\max_{m_0\le M_{\gamma,n},m_1\le M_{\theta,n}}\left|n_2^{-1}\gamma'_{m_0,k_2}P'_{n_2,k_2}(g_{n_2}(\theta_{m_1})-\varphi_{n_2}(\theta_{m_1}))\right|=O_p(\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}).
\tag{2.290}
$$

Combining the results in (2.285) and (2.290), the claimed result immediately follows. ■

Lemma 2.10.2 *Under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.3.3(v), 2.3.3(vii), 2.10.1(i)-2.10.1(iii),*

$$
\sup_{\theta\in\Theta}n_2^{-1}\sum_{i\in I}|\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta)|^2=o_p(1).
$$

Proof of Lemma 2.10.2. Let $\varphi_{k_2}(z,\theta)=P_{k_2}(z)'\beta_{\theta,k_2}$. By the triangle inequality and Assumption 2.10.1(i),

$$
\begin{aligned}
&n^{-1}\sum_{i\in I}|\widehat{\phi}_{n_2}(Z_i,\theta)-\phi(Z_i,\theta)|^2\\
&\le 2n^{-1}\sum_{i\in I}|\widehat{\phi}_{n_2}(Z_i,\theta)-\varphi_{k_2}(Z_i,\theta)|^2+2n_2^{-1}\sum_{i\in I}|\varphi_{k_2}(Z_i,\theta)-\varphi(Z_i,\theta)|^2\\
&\le(\widehat{\beta}_{\theta,k_2}-\beta_{\theta,k_2})'Q_{n,k_2}(\widehat{\beta}_{\theta,k_2}-\beta_{\theta,k_2})+o(1)\\
&\le\lambda_{\max}(Q_{n,k_2})\left\|\widehat{\beta}_{\theta,k_2}-\beta_{\theta,k_2}\right\|^2+o(1).
\end{aligned}
\tag{2.291}
$$

By definition, $\widehat{\beta}_{\theta,k_2} - \beta_{\theta,k_2} = n^{-1}Q_{n,k_2}^{-1}\sum_{i\in I} P_{k_2}(Z_i)(g(X_i,\theta) - \varphi_{k_2}(Z_i,\theta))$. Hence

$$\begin{aligned}
& n_2^{-1}\sum_{i\in I}|\widehat{\phi}_{n_2}(Z_i,\theta) - \phi(Z_i,\theta)|^2 \\
& \leq 2(g_n(\theta) - \varphi_n(\theta))' P_{n,k_2}(P_{n,k_2}'P_{n,k_2})^{-2}P_{n,k_2}'(g_n(\theta) - \varphi_n(\theta)) \\
& + 2(\varphi_{k_2,n}(\theta) - \varphi_n(\theta))' P_{n,k_2}(P_{n,k_2}'P_{n,k_2})^{-2}P_{n,k_2}'(\varphi_{k_2,n}(\theta) - \varphi_n(\theta)) \\
& \leq 2\lambda_{\min}^{-2}(Q_{n,k_2})n^{-2}(g_n(\theta) - \varphi_n(\theta))' P_{n,k_2}P_{n,k_2}'(g_n(\theta) - \varphi_n(\theta)) \\
& + 2n^{-1}\sum_{i\in I}|\varphi_{k_2}(Z_i,\theta) - \varphi(Z_i,\theta)|^2. \qquad (2.292)
\end{aligned}$$

By Lemma 2.10.1, we have

$$\begin{aligned}
& n^{-2}(g_n(\theta) - \varphi_n(\theta))' P_{n,k_2}P_{n,k_2}'(g_n(\theta) - \varphi_n(\theta)) \\
& \leq \left\|n^{-1}P_{n,k_2}'(g_n(\theta) - \varphi_n(\theta))\right\| \sup_{\gamma_{k_2}\in U_{k_2}}\left|n_2^{-1}\sum_{i\in I_2}\gamma_{k_2}'P_{k_2}(Z_i)\left(g(X_i,\theta) - \varphi(Z_i,\theta)\right)\right| \\
& = \left\|n^{-1}P_{n,k_2}'(g_n(\theta) - \varphi_n(\theta))\right\| O_p(\log^{1/2}(n_2)k_2^{1/2}n_2^{-1/2}) \qquad (2.293)
\end{aligned}$$

uniformly over $\theta \in \Theta$. (2.293) together with Assumption 2.3.3(v) then implies that

$$\sup_{\theta\in\Theta}\left\|n^{-1}P_{n,k_2}'(g_n(\theta) - \varphi_n(\theta))\right\| = o_p(1). \qquad (2.294)$$

Combining the results in (2.291), (2.292) and (2.294), and then applying Assumption 2.10.1(i), we immediately prove the claim of the lemma. ■

Lemma 2.10.3 *Under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.3.2(v), 2.3.3(v), 2.3.3(vii), 2.10.1(i)-2.10.1(iii),*

$$\sup_{\theta\in\Theta}|L_n(\theta) - L_n^*(\theta)| = o_p(1).$$

Proof of Lemma 2.10.3. For any $\theta_1, \theta_2 \in \Theta$, using the triangle inequality and Assumption 2.10.1(ii), we get

$$\begin{aligned}
& \left|w_n(Z_i)\left|h_0(Z_i) - \phi(Z_i,\theta_1)\right|^2 - w_n(Z_i)\left|h_0(Z_i) - \phi(Z_i,\theta_2)\right|^2\right| \\
& \leq w_n(Z_i)\left|(2h_0(Z_i) - \phi(Z_i,\theta_1))(\phi(Z_i,\theta_2) - \phi(Z_i,\theta_1))\right| \\
& + w_n(Z_i)\left|(\phi(Z_i,\theta_2) - \phi(Z_i,\theta_1))^2\right| \\
& \leq C\left|\phi(Z_i,\theta_2) - \phi(Z_i,\theta_1)\right| + C\left|\phi(Z_i,\theta_2) - \phi(Z_i,\theta_1)\right|^2 \leq C\left\|\theta_2 - \theta_1\right\|. \qquad (2.295)
\end{aligned}$$

let θ_{m_1} $(m_1 = 1,\ldots,M_{\theta,n})$ be a set of points in $\times$ such that $\min_{m_1\leq M_{\theta,n}}\|\theta - \theta_{m_1}\| \leq n^{-1/2}\log(n)$ for any $\theta \in \times$. As $\times$ is compact in R^{d_θ}, we know that $M_{\theta,n} \leq C(n^{1/2}\log(n))^{d_\theta}$. Hence, by the triangle inequality,

$$\sup_{\theta\in\times}|L_n(\theta)-L_n^*(\theta)|\leq 2Cn^{-1/2}\log(n)+\max_{m_1\leq M_{\theta,n}}|L_n(\theta_{m_1})-L_n^*(\theta_{m_1})|\,. \tag{2.296}$$

For any m_1 and for any i, by Assumptions 2.3.2(v) and 2.10.1(ii)

$$n^{-1}w_n(Z_i)\,|h_0(Z_i)-\phi(Z_i,\theta_{m_1})|^2\ \leq Cn^{-1}. \tag{2.297}$$

By Assumptions 2.3.1(i), 2.3.2(v) and 2.10.1(ii),

$$\mathrm{Var}\left[n^{-1}w_n(Z_i)\,|h_0(Z_i)-\phi(Z_i,\theta_{m_1})|^2\ \right]\leq n^{-2}E\left[w_n^2(Z_i)\,|h_0(Z_i)-\phi(Z_i,\theta_{m_1})|^4\right]\leq Cn^{-2}. \tag{2.298}$$

By (2.297) and (2.298), we can apply the Bernstein inequality to get

$$\begin{aligned}
&\Pr\left(|L_n(\theta_{m_1})-L_n^*(\theta_{m_1})|>Bn^{-1/2}\log(n)\right)\\
&\leq 2\exp\left[-\frac{B^2\log^2(n)n^{-1}}{2C(n^{-1}+B\log(n)n^{-3/2})}\right]\\
&=2\exp\left[-\frac{B^2\log^2(n)}{2C\left(1+B\log(n)n^{-1/2}\right)}\right]\leq 2\exp\left[-\frac{B\log(n)}{2C}\right],
\end{aligned} \tag{2.299}$$

where the last inequality is by Assumption 2.3.3(vii). (2.299) together with the Bonferroni inequality implies that

$$\begin{aligned}
&\Pr\left(\max_{m_1\leq M_{\theta,n}}|L_n(\theta_{m_1})-L_n^*(\theta_{m_1})|>Bn^{-1/2}\log(n)\right)\\
&\leq 2M_{\theta,n}\exp\left[-\frac{B\log(n)}{2C}\right]\\
&\leq 2C(n^{1/2}\log(n))^{d_\theta}\exp\left[-\frac{B\log(n)}{2C}\right]\\
&\leq 2C\exp\left[-\frac{B\log(n)}{2C}+2d_\theta\log(n)\right]\\
&\leq 2C\exp\left[-\frac{B}{8C}\log(n)\right],
\end{aligned} \tag{2.300}$$

As C is a fixed constant, from (2.300), we can choose B sufficiently large such that for any (fixed but) small $\varepsilon>0$, there is

$$\Pr\left(\max_{m_1\leq M_{\theta,n}}|L_n(\theta_{m_1})-L_n^*(\theta_{m_1})|>Bn^{-1/2}\log(n)\right)\leq\varepsilon$$

for all large n_2, which implies that

$$\max_{m_1\leq M_{\theta,n}}|L_n(\theta_{m_1})-L_n^*(\theta_{m_1})|=O_p(n^{-1/2}\log(n)). \tag{2.301}$$

Combining the results in (2.296) and (2.301), the claimed result immediately follows. ■

Lemma 2.10.4 *Under Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.3.2(v), 2.3.3(v), 2.3.3(vii), 2.10.1(i)-2.10.1(iv),*

$$n^{-1}\sum_{i\in I}||\widehat{\phi}_{\theta,n_2}(Z_i,\theta_0)-\phi_\theta(Z_i,\theta_0)||^2=o_p(n_2^{-1/2}).$$

Proof of Lemma 2.10.4. By Assumption 2.10.1(ii),

$$E\left[\|g_\theta(X,\theta_0)-\phi_\theta(Z,\theta_0)\|^2\Big| Z\right]\leq E\left[\|g_\theta(X,\theta_0)\|^2\Big| Z\right]<C \tag{2.302}$$

for any Z. Using Assumptions 2.3.1(i), 2.3.1(iii), 2.3.1(v), 2.10.1(iv) and (2.302), we can use similar arguments in showing (2.94) to show that

$$n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0)-\phi_{\theta_j,k_2}(Z_i,\theta_0)\right|^2=O_p(k_2n_2^{-1}+k_2^{-2r_{\varphi_\theta,j}}) \tag{2.303}$$

for any $j=1,\ldots,d_\theta$, where $\phi_{\theta_j,k_2}(z,\theta_0)=P_k(z)'\beta_{\varphi_\theta.j,k_2}$. By Assumption 2.10.1(iv) and (2.303),

$$\begin{aligned}&n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0)-\phi_{\theta_j}(Z_i,\theta_0)\right|^2\\&\leq 2n^{-1}\sum_{i\in I}\left|\widehat{\phi}_{\theta_j,n_2}(Z_i,\theta_0)-\phi_{\theta_j,k_2}(Z_i,\theta_0)\right|^2\\&+2n^{-1}\sum_{i\in I}\left|\phi_{\theta_j,k_2}(Z_i,\theta_0)-\phi_{\theta_j}(Z_i,\theta_0)\right|^2=O_p(k_2n_2^{-1}+k_2^{-2r_{\varphi_\theta,j}})\end{aligned} \tag{2.304}$$

for any $j=1,\ldots,d_\theta$. Using the result in (2.304), Assumptions 2.3.3(vii) and 2.10.1(v), we prove the claim of the lemma. ■

[AB17] N. Ashraf and O. Bandiera. *Social Incentives in Organizations.* In preparation for the Annual Review of Economics, 2017.

[ABL16] N. Ashraf, O. Bandiera, and S. Lee. "Do-gooders and Go-getters: Career incentives, Selection, and Performance in Public Service Delivery." Working paper, STICERD, 2016.

[AGR15] D. Acemoglu, C. Garcia-Jimeno, and J. A. Robinson. "State Capacity and Economic Development: A Network Approach." *American Economic Review*, **105**:2364–2409, 2015.

[AJR01] D. Acemoglu, S. Johnson, and J. A. Robinson. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, **91**:1369–1401, 2001.

[AT97] P. Aghion and J. Tirole. "Formal and Real Authority in Organizations." *Journal of Political Economy*, **105**:1–29, 1997.

[Ban08] M. Banton. "Administering the Empire." pp. 1801–1968, 2008.

[BBR09] O. Bandiera, I. Barankay, and I. Rasul. "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica*, **77**(1047-1094):417–458, 2009.

[BCD12] A. Banerjee, R. Chattopadhyay, E. Duflo, D. Keniston, and N. Singh. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training." Technical report, NBER Working Papers 17912, NBER, 2012.

[BEF17] Dal Bo, F. Finan E., O. Folke, T. Persson, and J. Rickne. "Who becomes a politician?" *e Quarterly Journal of Economics*, 2017. Forthcoming.

[Ber09] M. Bertrand. "CEOs." *Annual Review of Economics*, **1**:121–150, 2009.

[BER13] Dal Bo, F. Finan E., and M. A. Rossi. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service." *The Quarterly Journal of Economics*, **128**:1169–1218, 2013.

[BP09] T. Besley and T. Persson. "The Origins of State Capacity: Property Rights, Taxation, and Politics." *American Economic Review*, **99**(1218-44):1–34, 2009.

[Bre16] E. Breza. *Field Experiments, Social Networks, and Development.* Handbook on the Economics of Networks, Oxford, 2016.

[BS03] M. Bertrand and A. Schoar. "Managing with Style: The Effect of Managers on Firm Policies." *The Quarterly Journal of Economics*, **118**:1169–1208, 2003.

[CH08] V. Chernozhukov and C. Hansen. "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters*, **100**:68–71, 2008.

[Des17] E. Deserranno. “Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda.” mimeo, 2017.

[DH86] L. Davis and R. Huttenback. *Mammon and the Pursuit of Empire: The Political Economy of British Imperialism.* 1860-1912, Cambridge University Press, 1986.

[DJT99] M. Dewatripont, I. Jewitt, and J. Tirole. “The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies.” *Review of Economic Studies*, **66**:199–217, 1999.

[DP12] M. Dincecco and M. Prado. “Warfare, Fiscal capacity, and Performance.” *Journal of Economic Growth*, **17**:171–203, 2012.

[Fis01] R. Fisman. “Estimating the Value of Political Connections.” *American Economic Review*, **91**:1095–1102, 2001.

[FOP15] F. Finan, B. A. Olken, and R. Pande. “The Personnel Economics of the State.” Working Paper 21825, NBER, 2015.

[Fra92] M. Francis. “Governors and Settlers: Images of Authority in the British Colonies.” pp. 1820–60, 1992.

[FW04] R. Fisman and S. j. Wei. “Tax Rates and Tax Evasion: Evidence from "Missing Imports" in China.” *Journal of Political Economy*, **112**:471–500, 2004.

[Gar12] L. Gardner. *Taxing Colonial Africa: The Political Economy of British Imperialism.* Oxford University Press, 2012.

[GR12] R. Gibbons and J. Roberts. “The Handbook of Organizational Economics.” In *Princeton University Press.* 1 ed, 2012.

[HHM12] B. Henn, L. Hon, M. Macpherson, N. Eriksson, S. Saxonov, I. Peter, and J. Mountain. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples.” *PLoS One*, **7**, 2012.

[IM12] L. Iyer and A. Mani. “Traveling Agents: Political Change and Bureaucratic Turnover in India.” *The Review of Economics and Statistics*, **94**:723–739, 2012.

[Jef38] C. Jeffries. *The Colonial Empire and its Civil Service.* Cambridge University Press, 1938.

[Jia17] R. Jia. “Pollution for Promotion.” mimeo, 2017.

[JKS15] R. Jia, M. Kudamatsu, and D. Seim. “Political Selection in China: The Complementary Roles of Connections and Performance.” *Journal of the European Economic Association*, **13**:631–668, 2015.

[JO05] B. F. Jones and B. A. Olken. “Do Leaders Matter? National Leadership and Growth Since World War II.” *The Quarterly Journal of Economics*, **120**:835–864, 2005.

[Kir00] A. Kirk-Greene. “Britain’s Imperial Administrators.” pp. 1858–1966, 2000.

[KKO15] A. Q. Khan, A. I. Khwaja, and B. A. Olken. *Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors.* The Quarterly Journal of Economics, 2015.

[KKO16] A. Khan, A. Khwaja, and B. Olken. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." mimeo, 2016.

[KT13] F. Kramarz and D. Thesmar. "Social Networks in the Boardroom." *Journal of the European Economic Association*, **11**:780–807, 2013.

[Lai05] Z. Laidlaw. *Colonial Connections 1815-1845: Patronage.* the Information Revolution and Colonial Government, Studies in Imperialism, Manchester University Press, 2005.

[LO12] E. P. Lazear and P. Oyer. "Personnel Economics." In *The Handbook of Organizational Economics.* Princeton University Press, 2012.

[MC99] S. Morris and S. Coate. "Policy Persistence." *American Economic Review*, **89**:1327–1336, 1999.

[NT54] S. Northcote and C. Trevelyan. *Report on the Organization of the Permanent Civil Service.* HMS, 1854.

[NW89] D. North and B. Weingast. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice." *The Journal of Economic History*, **49**:803–832, 1989.

[PT96] C. Prendergast and R. H. Topel. "Favoritism in Organizations." *Journal of Political Economy*, **104**:958–78, 1996.

[PZ16] P. Persson and E. Zhuravskaya. "The Limits of Career Concerns in Federalism: Evidence from China." *Journal of the European Economic Association*, **14**:338–374, 2016.

[RAJ05] J. A. Robinson, D. Acemoglu, and S. Johnson. "Institutions as a Fundamental Cause of Long-Run Growth." *Handbook of Economic Growth*, **1**:386–472, 2005.

[RR17] I. Rasul and D. Rogger. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal*, pp. 1–34, 2017.

[SW16] E. Sanderson and F. Windmeijer. "A weak instrument F-test in linear IV models with multiple endogenous variables." *Journal of Econometrics*, **190**:212–221, 2016.

CHAPTER 3

Consumption Inequality and Partial Insurance: A Revisit with New Estimation Methods

3.1 Introduction

There has been a long-lasting interest in the dynamic relationship between consumption and income. This line of research started from primary descriptions in Cutler and Katz (1992) and Johnson and Smeeding (1998), then was comprehensively explained by Blundell, Pistaferri and Preston in their seminal work in 2008 (we refer to this as BPP in the following paper) and has been growing rapidly ever since. The intensive academic attention on this topic is drawn from both the theoretical and empirical spheres. Theoretically, the joint evolution of consumption distribution and income distribution serves as a primary key to disentangle competing hypotheses. On one hand, it is assumed that consumption is fully insured against idiosyncratic shocks, both transitory and permanent, to income in the complete markets hypothesis. On the other hand, in permanent income hypothesis, only the former shock, i.e. the transitory shock can be smoothed by self-insurance behaviors (Deaton(1992)). Unfortunately, neither of these two mainstream hypotheses possesses indisputable support from real data. Attanasio and Davis (1996) constructed synthetic panel data of consumption, label supply, and wages in the 1980s for the U.S. but failed to find any between-group consumption insurance. Besides, a series of work also discovered that consumption appears to react little towards permanent income shocks (Campbell and Deaton (1989); Attanasio and Pavoni 2006) but sensitive to transitory shocks (Hall and Mishkin (1982)). Empirically, this relationship builds up the foundations of several policy-oriented questions. From the micro perspective, this question is closely related to the transmission of inequality over the life cycle, hence can be used to answer questions like "how should government insurance programs be optimally designed?" in Low and Pistaferri (2010). Meanwhile, the macro perspective of this question can be rephrased as the relationship in aggregate savings, consumption and economic growth, hence can be used to answer questions like "What are the likely effectiveness of stabilization or stimulus policies?" in Hacker (2006).

Despite the importance and ever-lasting fever surrounding this topic, the data problem hampers the possibility of meaningful empirical research. To be specific, the lack of individual-level longitudinal panel data on both household expenditures, income, and saving gives rise to crucial difficulty in this study; Bundell (2014) stated "...With repeated cross-section measurements of income alone we cannot distinguish permanent from transitory income shocks, let alone identify the evolution of those variances and the insurance parameters...". For instance, in the US, the Panel Study of Income Dynamics (PSID) provides longitudinal annual income data for all households since 1968. Unfortunately, the PSID only collected data on a limited subset of consumption items, to be precise, it only contains food data (with few gaps in some of the survey years) till 1998. In contrast, there are a few datasets that provide detailed data on the spending habits of US households, like the Consumer Expenditure Survey (CEX), which are all repeated cross-sectional data by design. Problems of similar nature exist in many other countries, in the UK, the Family Expenditure Survey (FES) provides comprehensive data on household expenditures, but this is a cross-sectional data and thus the researcher does not get to observe households over time. In contrast, the British Household Panel Survey (BHPS) is a Panel dataset that collects data on income or wealth but provides no information on consumption.[1]

With this data constraint, various empirical methods have been proposed when confronting this difficulty. Hall and Mishkin (1982) were the first to take up the challenge, they utilized the food consumption data as an index of the total consumption. This method of using limited number of goods (largely necessity goods) was also employed by Zeldes (1989), Runkle (1991), Cochrane (1991), Shea (1995), Altonji, Hayashi and Kotlikof (1996), Martin (2003), Cox, Ng and Waldkirch (2004) and Hurst and Stafford (2004) and they all suffered the questioning of the validity of this substitution due to the different dynamics between food (or other necessities) and total consumption. Browning, Deaton and Irish (1985) and Attanasio and Weber (1993) contributed another method which involves forming of synthetic panel data sets from repeated cross-section data sets in which consumption is reported (e.g. the CEX or the FES). But in this case, the common employed mean-based construction method (proposed by Bourguignon, Goh and Kim (2004)) would iron out the individual differences in shocks that are of interest to us. Lastly, Jonathan Skinner was the first to propose imputation in 1987 and this was quickly accepted in the empirical world and hence developed by his followers Ziliak (1998) and Browning, Leth-Petersen (2003) and the seminal work BPP.

[1] Similar problems exist for many other countries, in particular countries in Europe (e.g. France and Spain) that collect detailed data on both consumption, income, and wealth, but the information never exists in a single data set.

Specifically, this method obtains total consumption in the PSID using the estimated relationship between total consumption and some other consumption items that are reported in both the PSID and the CEX. The difference between the original imputation method by Skinner (1987) and the BPP method is that the original method regresses total consumption on other consumption items while the BPP used “structural” regression of a standard demand function for food that depends not only on other consumption items, but also on prices and a set of demographic and socio-economic variable of the household. Assuming monotonicity of the demand for food makes it possible to invert these function in order to obtain a structurally based formula total consumption, which exists in the CEX, but is missing in the PSID. In one sentence, the original method treats the missing variable as the "Y" while the developed BPP method treats it as an "X". Though this imputation method seems appealing, at first sight, it has a major weakness since it reduces any variation in total consumption; it does not take into account the fact the there is considerable idiosyncratic elements that go into the individual decision making. Besides, the BPP method also suffers from the well-known error-in-variable problem and gives biased results. We will discuss those problems in details later. This method is so well accepted that Heathcote, Storesletten, and Violante (2014) directly used the imputed data by BPP in the calibration of their structural model.

To deal with the lack of complete data set in this topic and other similar circumstances in other research fields, we follow the work of Klevmarken (1982), Angrist and Krueger (1992), Arellano and Meghir (1992) to study the estimation and inference of nonlinear econometric model when the economic variables are contained in different data sets in Buchinsky, Li and Liao (2016,a). In that paper, a minimum distance (MD) estimator of the unknown structural parameter of interest was constructed with some common containing variables in different data sets to provide the specific form of optimal weight for the MD estimation. Hence, in utilizing this new method, we would like to join the efforts and re-evaluate the relationship of the dynamic processes of consumption and income. At the same time, the inclusion of more consumption data in PSID starting from 1999 wave gives us a wonderful chance to verify our new method with empirical data. The validity of the new method makes it possible for us to revisit the topic using the historical incomplete data in this topic and also gives confidence for application of the new method to other topics with similar data obstacle.

It’s worth mentioning that the new PSID consumption data is still not sufficient for the estimation of the insurance parameter. The reason lies in the fact that the PSID consumption only contains several but not all categories of the nondurables. A study by Blundell, Pistaferri, and Saporta (2012) shows that the average of PSID nondurables is only about 55% of the total nondurable consumption. Theoretically, if we treat PSID consumption as a constant fraction of nondurable expenditure (which is true by real data), the degree of insurance of this partial consumption with respect to income shocks reflects partly the true degree of insurance of nondurable consumption and partly the relationship between this partial consumption and total nondurable consumption (i.e. the budget elasticity). Hence what we can estimate with this consumption measurement is just a part of the parameter that we are truly interested in. But this estimation can still serve as a verification baseline for BPP and our method. The logic here is that if we construct an identical partial consumption measurement in CEX, then an accurate multi-dataset estimation method should give similar results with that from the true PSID data. After the method verification, we can, therefore, employ the true total CEX consumption to estimate the parameters that we are interested in, employing the validated estimation method.

Hence we will follow the idea in the seminal BPP paper by assuming some, but not necessarily in regards to insurance as well as taking into consideration the distinctions between transitory and permanent shocks. Section II introduces the new available PSID biannual panel data of some categories of consumption from 1999 wave and discusses various data problems. Also, the descriptive empirical results based on the new data will be provided at the end. In section III, we will restate the consumption model formulated in BPP and provide identification strategy for both our new method and the imputation method. Section IV gives the estimation results using both methods without the newly available longitudinal consumption data, and the result with the complete data. If we consider the latter as the "true" result, the comparison of former with the latter can be used as a test of the validity of different empirical strategies. Section VI discusses the theoretical hypotheses with the empirical results and conclusion.

3.2 Introduction of new PSID and CEX

Panel Study of Income Dynamics (PSID)[2]

As already stated in the introduction, the Panel Study of Income Dynamics (PSID) is a survey that provides longitudinal data describing roughly 5,000 households that have been followed since 1968. It has been used as the major data set in researches regarding of income due to its excellency in providing information on income and also various socioeconomic micro-level data. However it's also

[2]**Please see Martha Hill (1992) for a more detailed description of the PSID. We only provide basic facts and related information about the 1999 reform here.**

known as its lack in consumption data. Only food consumption and rental data[3] is consistently available (with the exception of 1973, 1988, and 1989) while other categories of consumptions are all missing.[4]

Starting with the 1999 wave, however, two major reforms were made. First, the original annual survey was switched to a biannual survey. Second, the PSID began collecting more consumption information, including health expenditures, utilities, transportation related expenditure, education and child care, which covers almost 70% of nondurable expenditure from national accounts. Another minor change is implemented in 2005 with a few additional consumption categories added (such as clothing and entertainment).

The Consumer Expenditure Survey (CEX)[5]

The Consumer Expenditure Survey, on the other hand, is intended to investigate the expenditure habits of American consumers, and used primarily for revising the CPI[6]. This data contains two components, the Diary Survey, which is designed to obtain detailed expenditures data on small and frequently purchased items, and the Interview Survey which on the other hand follows survey households for a maximum of five consecutive quarters, and is designed to capture large purchases, such as spending on rent property and vehicles, as well as those expenditures that occur on a regular basis such as rent or utility payments.

Data comparison in PSID and CEX

Although both data sets take representative U.S. families as interview units and are conducted by well-respected organizations with delicate survey designs and implementations, there are still some differences among them that should be taken considered when combining them. And no matter

[3]Only rental consumption data is available but not the rent equivalents for non-homeowners, so it's impossible to construct a consistent rent consumption for all households

[4]Before 1999, the survey occasionally collected other consumption data, such as home insurance etc..

[5]**Please see "Chapter 16: Consumer Expenditures and Income" from BLS *Handbook of Methods* for a detailed description of the survey, including sample design, interview procedures, etc.**

[6]A description of the survey, including more details on sample design, interview procedures, etc., may be found in "Chapter 16: Consumer Expenditures and Income," from the BLS Handbook of Methods.

which method we use, the BPP method or our proposed method, the comparability of the two data sets is crucial. In this section, we explain the adjustments made to the two data sets for later use.

Unit definition

In those two surveys, the basic unit is always household but the unit head is defined differently. In PSID, the head of one unit or household is the husband in a couple while that in CEX is the person or one of the persons who owns or rents the unit. Following most of the previous research, our analysis makes the two definitions compatible.

Sampling issue of PSID

The genealogical and longitudinal designs of the PSID make it a unique resource for addressing particular questions; nevertheless there are concerns about the sample representativeness since the PSID sample is continually replenished through births and marriage. Fitzgerald et al.(1998) carefully studied this issue and found no strong evidence of distortion in the representativeness through 1989, and showed with considerable evidence that the PSID's cross-sectional representativeness has remained roughly intact from the sample replenishment. In addition, a thorough subsampling procedure was conducted in 1997 in reaction to the growth of the sample size resulting from family splits. By setting aside entire linkages to a 1968 PSID sample family and adding a nationally representative sample of immigrant households and individuals that would not be eligible for the PSID under the original 1968 sample recruitment and sample family following rules, this subsampling procedure managed to exemplify the PSID sample while maintaining the intergenerational ties in the core panel. Hence, in this study, we undertake previous research results and assume that the PSID data is nationally representative with family weights.[7]

Construction of PSID equivalent CEX consumption

PSID has two major changes with respect to consumption data after 1999. The first change added most consumption categories including housing, transportation, health, education and child care, while the second change completed the consumption data even more by adding clothing, recreation, etc. According to Blundell, Pistaferri and Saporta (2012), the data collected from the 1999 wave to 2003 wave accounted for about 66% of the NIPA data while the percentage for the 2005 to 2009 wave was about 70%. Since the topic of this research is estimating the relationship between consumption and income, no matter what definition of consumption we use, the consumption data has

[7] In BPP, weighting was not considered for both the CEX data set and PSID data set, and this is one of the minor changes we made when we implemented their method.

to be consistent among different years, in the sense that for those years after 2005, even when new categories were added, we still used the categories existing in all of the years starting from 1999 wave.

Table 3.1: Comparison of PSID data with NIPA

	1998	2000	2002	2004	2006	2008
PSID Total	3,276	3,769	4,285	5,058	5,926	5,736
NIPA Total	5,139	5,915	6,447	7,224	8,190	9,021
Ratio	0.64	0.64	0.66	0.7	0.72	0.64
PSID Nondurables	746	855	867	1,015	1,188	1,146
NIPA Nondurables	1,330	1,543	1,618	1,831	2,089	2,296
Ratio	0.56	0.55	0.55	0.55	0.57	0.5
PSID Services	2,530	2,914	3,398	4,043	4,739	4,590
NIPA Services	3,809	4,371	4,829	5,393	6,101	6,725
Ratio	0.66	0.67	0.7	0.75	0.78	0.68

- This table is from Blundell,Pistaferri and Saporta (2012).
- PSID weights are applied for the non-sampled PSID data.
- Total consumption is defined as Non-durables + Services, PSID consumption categories include food, gasoline, utilities, health, rent (or rent equivalent), transportation, child care, education and other insurance.
- NIPA numbers are from NIPA table 2.3.5.
- All numbers are nominal.

In addition to the time consistency of consumption data, we also needed to ensure that the consumption data was comparable in the two data sets. Here, we borrowed from the wisdom of Charles et.al (2007) and Cooper (2010) who did an excellently job of matching the CEX categories (UCC coded) with their counterparts in the PSID for the 1999-2006 waves and the post 2006 waves separately. (See the mapping table 2 below).

Table 3.2: Mapping of CE UCC codes into PSID categories

PSID Consumption Category	CE UCC Code
Food	
At home	190904, 790220, 790230
Delivered	
Away from home	190902, 190903, 790410, 790420, 800700
Health Care	
Hospital & nursing home	570110, 570210, 570220, 570230
Doctor	560110, 560210, 560310, 560330, 560400
Prescription drugs	340906, 540000, 550110, 550320, 550340, 570901, 570903, 570240
Insurance	580111, 580112, 580113, 580114, 580311, 580312, 580901, 580903, 580904, 580905, 580906
Housing	
Mortgage	220311, 220312, 220321, 220322, 830201, 830202
Rent	210110, 800710
Insurance	220121, 220122
Property Tax	220211, 220212
Utility	250111, 250112, 250113, 250114, 250211, 250212, 250213, 250214, 250221, 250222, 250223, 250224, 250901, 250902, 250903, 250904, 260111, 260112, 260113, 260114, 260211, 260212, 260213, 260214, 270211, 270212, 270213, 270214, 270310, 270411, 270412, 270413, 270414, 270901, 270902, 270903, 270904
Transportation	

Vehicle Loan Payment	870103, 8701014, 8702023, 870204, 850300
Down Payment	870101, 870102, 870201, 870202, 870801
Vehicle Lease Payment	450310, 450313, 450314, 450410, 450413, 450414
Insurance	450311, 450411, 500110
Gasoline	470220, 470211, 470212, 480110, 480213, 480214, 490110, 490211, 490212, 490221, 490231, 490232, 490311, 490312, 490313, 490314, 490318, 490319, 490411, 490412, 490413, 490501, 490502, 490900, 520410
Other Vehicle Payments	
Parking	520531, 520532
Bus	530311, 530312, 530501, 530902, 530210
Taxicab	530411, 530412
Other Transportation	520511, 520512, 520521, 520522, 520542, 520902, 520903, 520904, 520905, 520906, 520907, 530110, 530901
Education	
Schooling	190901, 210310, 370903, 390901, 660110, 660210, 660310, 660900, 670110, 670210, 670901, 670902, 800802, 800804
Other School-related	690111, 690112
Child Care	340211, 340212, 670310
Clothing	360110 - 370902, 370904 - 390322, 390902 - 430120, 640130
Trips & Vacations	470113, 470212, 520212, 520522, 520532, 520542, 520905 - 530210, 530312, 530411, 530510, 530901, 610900, 620122, 620212, 620222, 620903, 620909, 620919, 290116, 810400
Other Recreation	310240, 310340 - 310350, 590111 - 590410, 600210 - 610320, 620111, 620121, 620211, 620221, 620310, 620330, 620904 - 620908, 620912, 620921 - 620930

Household Furnishing & Equipment	220612, 220615, 220616, 230133, 230134, 280110, 280120 - 310230, 310311 - 310334, 320111 - 320522, 320633 -320904, 340902, 340904, 340905, 340907, 990900, 230117, 230118, 790611
Home Repair & Maintenance	230112 - 230115, 230121, 230122, 230123, 230142 - 230150, 240111 - 240323, 270211 - 270214, 270901 - 270904, 320611 - 320633, 330511, 340620, 340630, 340901, 340903, 340914, 790600, 990930, 990940

- Categories from *Food* to *Child Care* are for pre-2005 mapping from Charles et.al (2007). The remaining categories are for the post-2005 mapping from Cooper (2010).
- 200900, 790310, 790320, and 790410 (alcohol) are included in the total food expenditures.
- 270101 and 270102 (telephone) are included for the total housing expenses.

For comparison purpose, we consider several consumption measurements in our paper: non-durable[8]-durable with services from some durable (housing and vehicles), non-durable with health care, education and total consumption. We calculate the ratio of CEX and adjusted CEX consumption with PSID consumption as a measure of the comparability between these two data sets and the result is in Table 3. Each number is defined as follows:

$$Ratio_{Before} = \frac{\text{CEX consumption}}{\text{PSID consumption}}$$

$$Ratio_{After} = \frac{\text{PSID-equivalent CEX consumption}}{\text{PSID consumption}}$$

We can see that the adjustment significantly decreases the differences between these two data sets. For example, before the adjustment, the consumption calculated using CEX category is significantly larger than the PSID counterpart, the difference between these two measurements is as large as 72% and the adjustment shrinks this difference to around 10%. And among those different definitions, the total consumption matched the best for all years (the difference between measures from these two data sets is around 2%), which suggests that we should have the result with the least noise by employing this consumption definition.

In addition, we also provide the matching results by category. And we can see that, even with careful matching, there are still quite a few differences between these two data sets which maybe due to the wording differences in the questionnaires, for example. Also, the differences between these two data sets are not time consistent. Although we could not do better with the matching for each category, fortunately we didn't really need the detailed consumption data but rather the total consumption.

The last minor problem was that the PSID survey has been conducted every other year since 1999 while the CEX has been collected on annual base. Therefore, to make the consumption data in the PSID comparable to the imputed consumption data based on BPP and the CEX data used in BLL, for both the imputation and our proposed method, we only chose those years in which PSID data was available.

[8]defined in Attanasio and Guglielmo Weber (1995), which is the sum of food (defined above), alcohol, tobacco, and expenditure on other nondurable goods, such as services, heating fuel, public and private transport (including gasoline), personal care, and semi-durables, defined as clothing and footwear. Since the semi-durables are only available after 2005, we don't include them in our calculation.

Table 3.3: Comparison of PSID data with CEX, PSID-equivalent CEX consumption

	Non-durables		**Non-durables with edu&health**		**Non-durables with housing**		**Total**	
	Before	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*
1998	120.95%	84.31%	107.87%	95.47%	98.11 %	89.57%	108.47%	97.48%
2000	142.80%	88.20%	119.61%	97.68%	113.48%	93.62%	120.51%	99.91%
2002	129.90%	84.47%	111.69%	94.14%	99.27%	90.39%	109.16%	96.53%
2004	172.26%	94.78%	135.18%	99.65%	127.14%	100.27%	132.16%	102.50%
2006	155.56%	97.63%	125.05%	101.25%	113.95%	103.68%	118.44%	104.73%
2008	117.42%	97.59%	101.30%	99.33%	90.57%	101.16%	97.70%	101.33%
2010	131.07%	105.52%	105.03%	102.54%	97.17%	106.36%	101.05%	103.59%

- "*Before*" here means before the matching process and each number is the ratio of CEX consumption and PSID consumption under that specific definition.
- "*After*" means after the matching process and each number is the ratio of adjusted CEX consumption and PSID consumption under that specific definition.

Table 3.4: Comparison of PSID data with PSID-equivalent CEX consumption, by category

	Food	**Health**	**Housing**	**Transport**	**Education**	**Child care**
1998	91.96%	84.15%	108.42%	82.25%	163.38%	148.02%
2000	92.24%	96.93%	113.02%	86.29%	145.02%	98.95%
2002	91.44%	94.71%	100.84%	88.61%	143.98%	103.86%
2004	94.47%	118.70%	117.83%	84.0%	122.37%	111.95%
2006	106.52%	120.67%	117.85%	81.19%	141.36%	62.87%
2008	133.95%	117.43%	117.88%	62.70%	105.00%	75.82%
2010	132.41%	117.12%	108.25%	78.77%	97.36%	61.22%

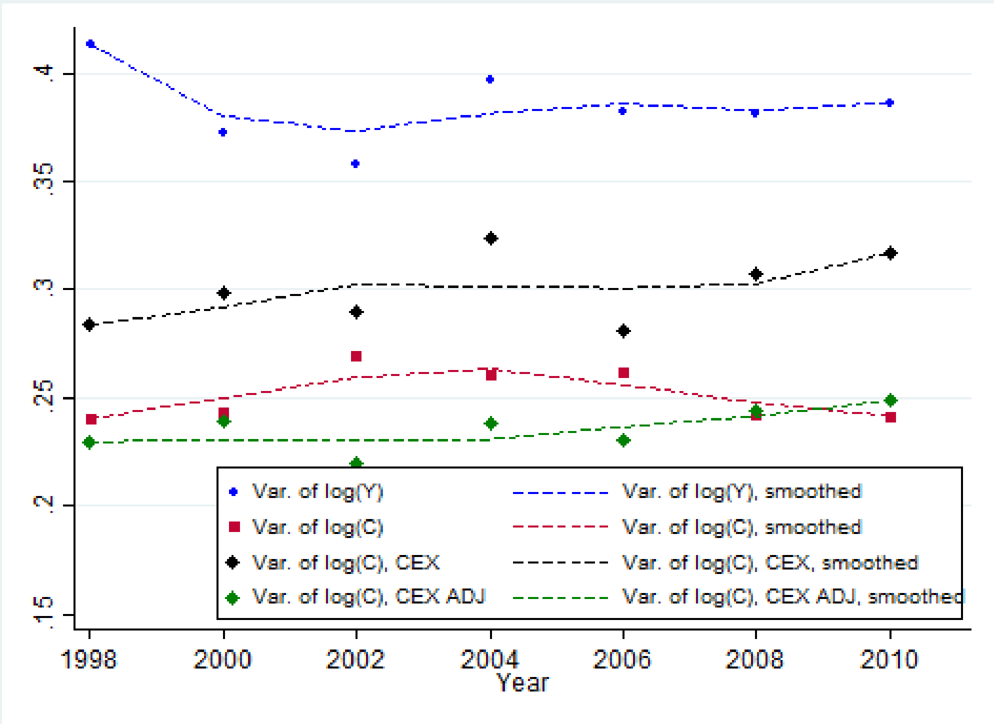
Characteristics of Consumption and Income Inequality

As stated in the introduction, the scarcity of consumption data in any longitudinal dataset used to make it impossible to study the comovement of consumption and income. The new PSID data for the first time gives us the possibility of documenting some basic features of the evolution of consumption and income inequality with more precision.

From Figure 1, which uses PSID data on log income and log consumption for the year 1998 onward, and CEX data on log consumption from 1980 to 2010, we plot both the point estimators of the variances and the polynomial smoothing lines through those. The income we used here is the after tax non-financial income (non-financial income - taxes on non-financial income), and the after tax total income gives the exactly same trend. The consumption is the PSID total consumption, BPP defined CEX total consumption and adjusted CEX consumption as we defined before (We also depict the non durable consumption trends and present it in the Figure 2 of Appendix 1).

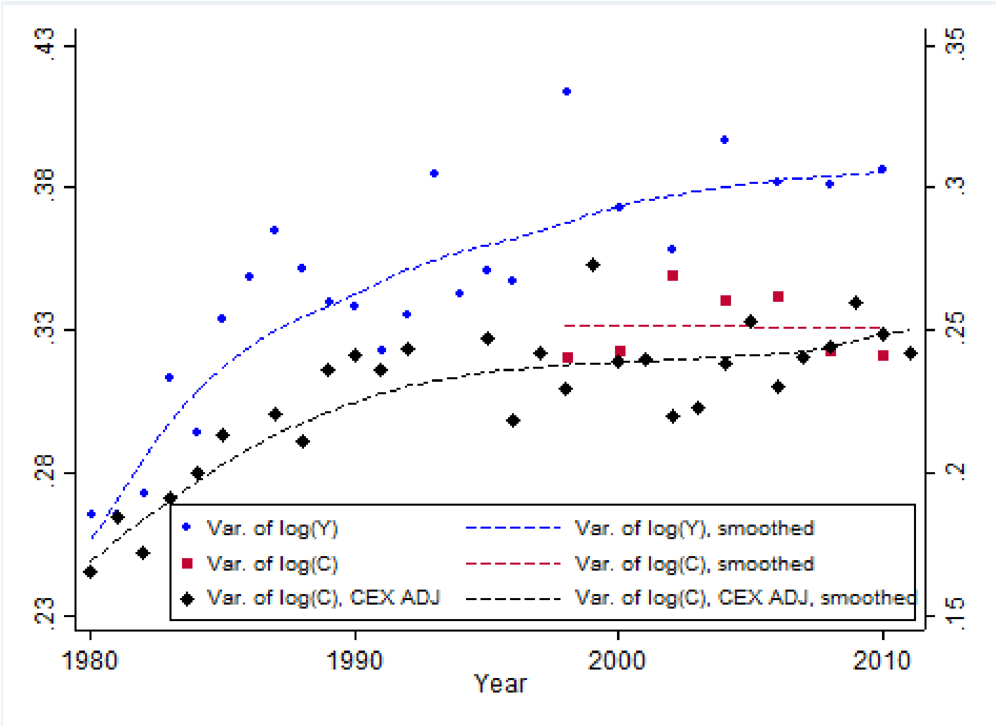
Before our discussion of the evolution of consumption and income equality in the 21st century, it seems necessary to take a closer look at the different definitions of consumption. CEX provides detailed expenditure data for each quarter on the UCC level, and at the same time provides expenditures by category in the family characteristics, income and consumption files (except for 1982 and 1983), in which the consumption is aggregated using the detailed expenditure file by formula. But this formula changes every year due to the adjustments to the UCC system. In other words, if we directly use the total consumption defined by CEX, we actually get time-inconsistent defined consumption data, for which the time trend analysis is meaningless. This is the reason that BPP calculated time-consistent consumption by applying one year's definition of consumption to all of the years that they were interested in. In other words, the BPP defined consumption is also a "partial" consumption, and we sacrifice some expenditure items for the consistency of the consumption panel. We also calculated the mean and variance of the CEX published consumption by various definitions, and put them in Appendix 1 for reference . The CEX adjusted consumption is the trimmed CEX consumption adjusted to match the PSID definition and available consumption categories. We can understand it as another "partial" consumption with even fewer categories (clothing, recreation and furnishing etc.) So understandably, the variance of the adjusted CEX consumption is smaller than the BPP-defined CEX consumption. Fortunately, they two follow very similar patterns and are both quite flat during the period under study. In fact, the former is just a downward-translated version (by about 0.05 units) of the latter. And most importantly, we can see that the PSID consumption and the the CEX adjusted consumption share a similar flat pattern as well, which serves as the base

Figure 3.1: Overall Pattern of Inequality



- The income data comes from PSID 1998 onward, and here we use the after-tax, non-financial income (i.e., non-financial income - taxes on non-financial income). We also checked with the after-tax total income, and put it in Appendix 1.
- The consumption data also comes from PSID 1998 onward.
- There are two CEX consumptions here; the one labeled "log(C), CEX" is the same one employed by BPP, which includes all categories of consumption. The other, which labeled as "log(C), CEX ADJ", is the adjusted CEX consumption we use in this paper, which aims at matching the exact categories of consumption included by the PSID.
- Since CEX data is collected annually, and is collected PSID biannually, the PSID income and consumption trends have only 7 samples while the CEX consumption trends have 14.

Figure 3.2: Overall Pattern of Inequality from 1980 to 2010



- We use Lowess smoothing for the long income series from the PSID and the long consumption series from the CEX
- We use linear fit for the short consumption series from the PSID. The reasoning behind this is that, with a much shorter length of time, each point would have a larger weight when smoothed compared to the long consumption series from the CEX.

Figure 3.3: Overall Pattern of Inequality - Nondurable

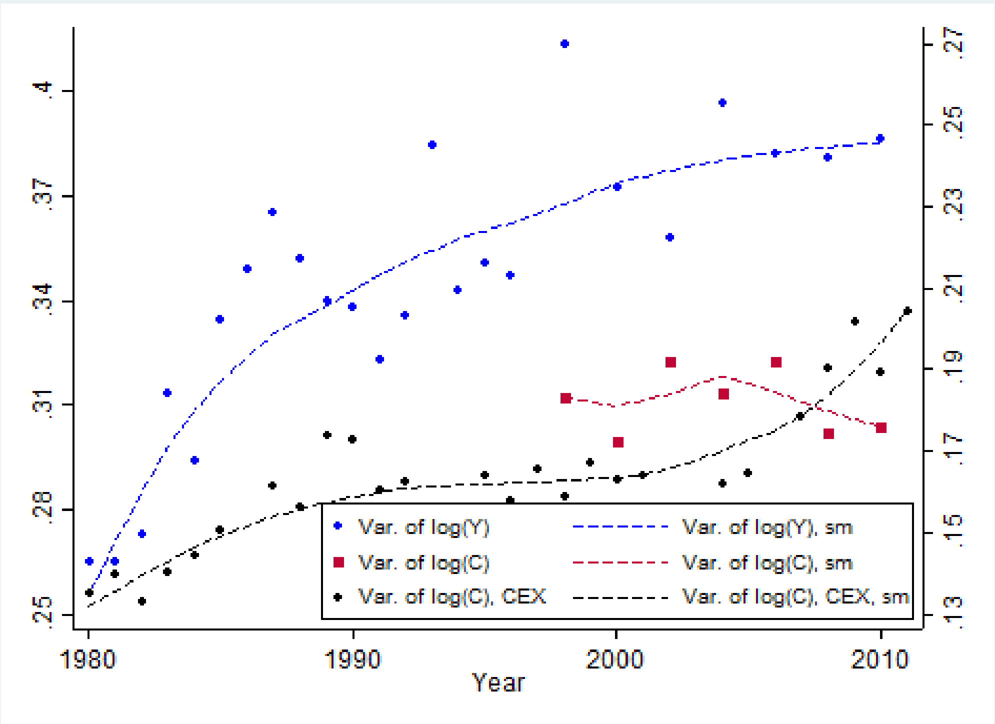
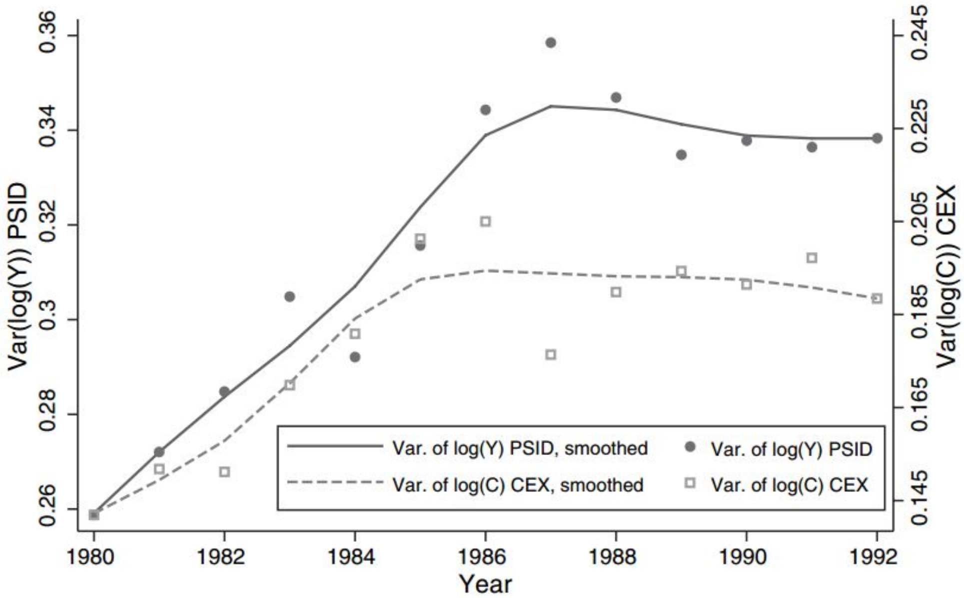


Figure 3.4: Overall Pattern of Inequality from 1980 to 1992, from BPP



- They use Lowess smoothing for the long income series from the PSID and the long consumption series from the CEX
- They didn't use the weight of either survey, which leads to a slightly different result than ours.

of our estimation using the two data sets.

There are two distinct features of the evolution of consumption and income inequality in the 21st century that are worth noting. The first distinct feature is that the scale of the income variance is greater than that of the consumption variance which suggests a certain degree of self-insurance against income shocks. The second is that both income and consumption inequality are quite flat during these periods. There are indeed some small fluctuations, for example, income equality was relatively high in 1998 and 2008, possibly due to the financial crisis. But if we pool data from a longer period, from 1980 to 2010 (as shown in Figure 2), we can obtain a clearer picture of the trend in which the consumption inequality turns to completely flat in the late 1990s, whereas income inequality slows down the rising trend in the 1980s.

The nondurable consumption gives a slightly different pattern, which we can see clearly in Figure 3. The non-durables start to flatten out in the late 1980s and started to increase again in the 2000s. This is slightly different from BPP's observations due to the availability of longer time series (we attach the BPP result here for comparison). But one problem with non-durables is that the CEX and the PSID give two different patterns after 2000. The PSID is almost flat, very similar to the total consumption, while the CEX consumption inequality increases. This is one reason we use total consumption as our main result. But in either case, we can see that the relationship of income and consumption is quite stable after 2000, no matter which definition of consumption we use. This is the first time that we can use the real data to do the above exercise, though there has been a lot of attempts for the pre-1999 period with the defective data. We can see that the original BPP's work gives a similar trend for the pre-1999 period, but they concludes that both consumption in late 1980s while in the larger picture, we can see that actually consumption inequality was still rising in that period but flattens out in the late 1990s instead.

3.3 The income and consumption process

Here we completely follows the setting in BPP. Suppose real (log) income, $logY$ can be decomposed into a permanent component P and a mean-reverting transitory component v. The income process for each household i is

$$logY_{i,t} = Z'_{i,t}\varphi_t + P_{i,t} + v_{i,t}$$

where t indexes time and Z is a set of income characteristics observable and known by consumers at time t. These will include demographic, education, ethnic, and other variables.
Assume the permanent component $P_{i,t}$ follows a martingale process of the form

$$P_{i,t} = P_{i,t-1} + \zeta_{i,t}$$

where $\zeta_{i,t}$ is serially uncorrelated, and the transitory component $v_{i,t}$ follows an MA(q) process, where the order of q is set to 1 as it is in Blundell et al. (2008)

$$v_{i,t} = \varepsilon_{i,t} + \theta\varepsilon_{i,t-1}$$

It follows that the unexplained income growth is

$$\Delta y_{i,t} = \zeta_{i,t} + \varepsilon_{i,t} + (\theta - 1)\varepsilon_{i,t-1} - \theta\varepsilon_{i,t-2} \tag{3.1}$$

Suppose the real log consumption $logC$ follows the process

$$logC_{i,t} = Z'_{i,t}\varphi^c_t + P^c_{i,t}$$

and the error component follows

$$P^c_{i,t} = P^c_{i,t-1} + \phi\zeta_{i,t} + \psi\varepsilon_{i,t} + \xi_{i,t}$$

in other words, we can write (unexplained) change in log consumption as follows:

$$\Delta c_{i,t} = \phi\zeta_{i,t} + \psi\varepsilon_{i,t} + \xi_{i,t}. \tag{3.2}$$

We allow permanent income shocks $\zeta_{i,t}$ to have an impact on consumption with a loading factor of $\phi_{i,t}$, which may potentially vary across individuals and time; the impact of transitory income shocks $\varepsilon_{i,t}$ is measured by the loading factor $\psi_{i,t}$. The random term $\xi_{i,t}$ represents innovations in consumption that are independent of those in income. This may capture the measurement error in consumption, preference shocks, innovation to higher moments of the income process, etc. We call $\phi_{i,t}$ and $\psi_{i,t}$ partial insurance parameters. Equation (2) nests the two extreme cases of full insurance of income shocks ($\phi_{i,t} = \psi_{i,t} = 0$) as contemplated by the complete markets hypothesis, and no insurance ($\phi_{i,t} = \psi_{i,t} = 1$) as in autarky, as well as intermediate cases in which $0 < \phi_{i,t} < 1$ and $0 < \psi_{i,t} < 1$. The closer the coefficient is to zero, the higher the degree of insurance.

In Appendix we discuss identification details of the model in more in detail; it is similar to BPP in most respects, but slightly different because we have biannual data instead of the annual data they used.

3.3.1 Skinner method

In his 1987 paper, Jonathan Skinner was the first to propose the imputation method. The procedure he proposed is that firstly run a regression of total consumption on all commonly presented individual consumption categories with CEX data set, then with the PSID data, consumption is constructed with the estimated parameter and the corresponding PSID consumption categories. Consider the following total consumption equation in the CEX:

$$c_{i,x} = D'_{i,x}\beta + \gamma f_{i,x} + e_{i,x}$$

in which c is the total consumption (or non-durable consumption), f is the food consumption, and D contains prices and a set of conditioning variables (also available in both data sets). For simplicity, we follow Skinner's linear assumption and notification that we have one input data set (in this case is CEX) and one target data set (in this case is PSID), and we use the subscript x to indicate an input data set variable and the subscript p to indicate a target data set variable. We assume that x and p are two random samples drawn from the same underlying population. Then Skinner's method will generate the following imputed consumption in PSID:

$$c_{i,p}^{IM} = D'_{i,p}\hat{\beta} + \hat{\gamma} f_{i,p} = c_{i,p} - \hat{e}_{i,p}$$

In other words, the imputed consumption can be viewed as error-ridden measurement of the true consumption $c_{i,p}$, but the measurement error $\hat{e}_{i,p}$ is uncorrelated with $c_{i,p}^{IM}$.[9] Then if the imputed consumption were used as a variable in a linear regression model, the OLS regression would still give us an unbiased and consistent estimation for the parameters, but would fail if the linear assumption doesn't hold.

Alternatively, the BPP method proposed a structural model for the food consumption, in other words, they assume the following:

$$f_{i,x} = D'_{i,x}\beta + \gamma c_{i,x} + e_{i,x}$$

[9] In the original regression, we have that the fitted values and residuals are uncorrelated, i.e., $\hat{c}'_x \hat{e}_x = 0$. Since the CEX and the PSID are random samples from the same population, $\hat{c}'_p \hat{e}_p = 0$.

Again, we assume linearity for simplicity. Then the imputed consumption by BPP would be

$$c_{i,p}^{IM} = \frac{1}{\hat{\gamma}}(f_{i,p} - D'_{i,p}\hat{\beta}) = c_{i,p} + \frac{\hat{e}_{i,p}}{\hat{\gamma}}$$

where $\hat{e}_{i,p}$ is the residual from the original food consumption structural regression equation. This imputation would also produce a measurement error but the measurement error $\frac{\hat{e}_{i,p}}{\hat{\gamma}}$ would be correlated with $c_{i,p}^{IM}$. Hence, even with linear assumption, any regression employing the imputed consumption as a variable would fail to generate an unbiased and consistent estimation.

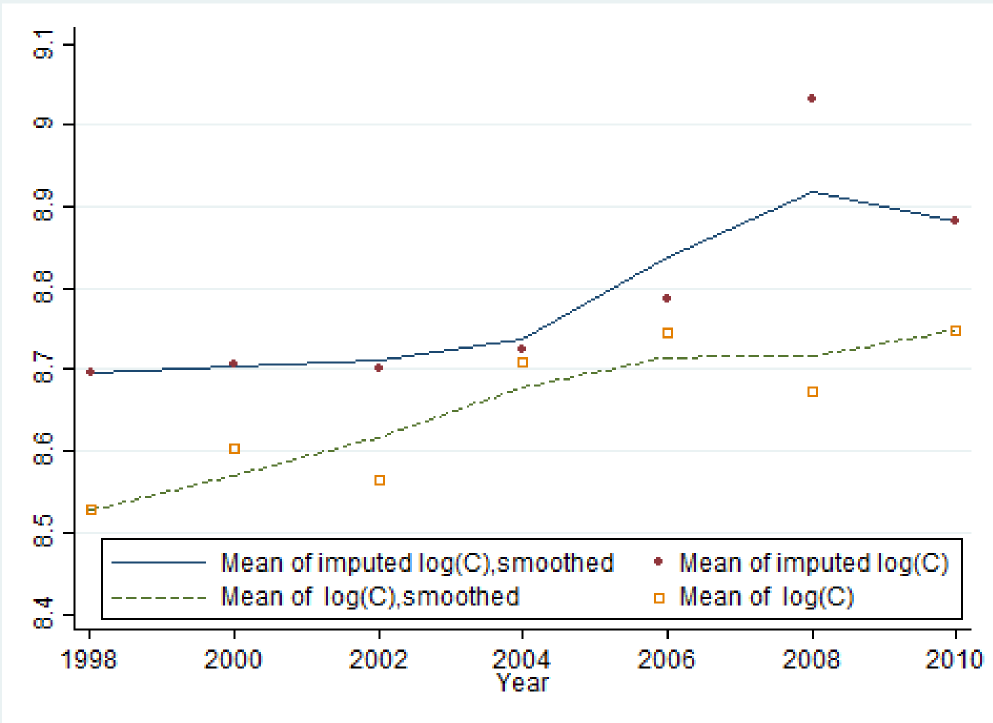
3.3.2 BPP method

Panel consumption Data

We follows the BPP imputation methodology and compare the average of variance of the imputed consumption with the true PSID consumption. The result is presented in Figure 3, in which we can see the following two distinct characteristics. The first is that the averages of these two measurements are similar, which is also proven in Cooper (2010). This is easy to understand: after averaging out the individual idiosyncratic errors, the imputation can provide a good semblance of the population. The second is that the variance of the imputation consumption is significantly smaller than the true ones and gives totally different trend. This is because the imputation rules out the individual differences, and hence decreases the variation of the sample.

We can observe the above conclusion more clearly by comparing the kernel density of the true consumption from the PSID and the imputed consumption. Or we can also turn to the kernel density of the individual difference between the true consumption and the imputed ones. We can draw very similar conclusions from Table 5 in the sense that the mean of the difference is very small compared to the mean of $log(C)$, for example; the smallest difference is in 2006, which is only 0.3% of the mean of $log(C)$. This suggests that the imputation can do a very good job in predicting the sample mean. But the variance of the difference is comparable to $log(C)$, implying the imputation cannot capture the individual idiosyncrasy.
In the figure below, we also provide the kernel density of the difference by year.

Figure 3.5: Mean of the true and imputed consumption



3.4 BLL and BPP

BPP

In this part, we first estimate the model with the real PSID consumption data and the BPP imputed consumption data and then compare the two results. In Table 6, we show the estimation results for three different definitions of non-durables that were defined in last section. We choose to follow BPP's main sample selection, which is a sample of continuously married couples headed by a male (with or without children) aged 30 to 65. Hence we eliminate households in which the head or the head's spouse changes; therefore we can focus on income risk without modeling divorce, widowhood or other household breaking-up factors.

Table 3.6: Optimal Minumum Distance Estimates

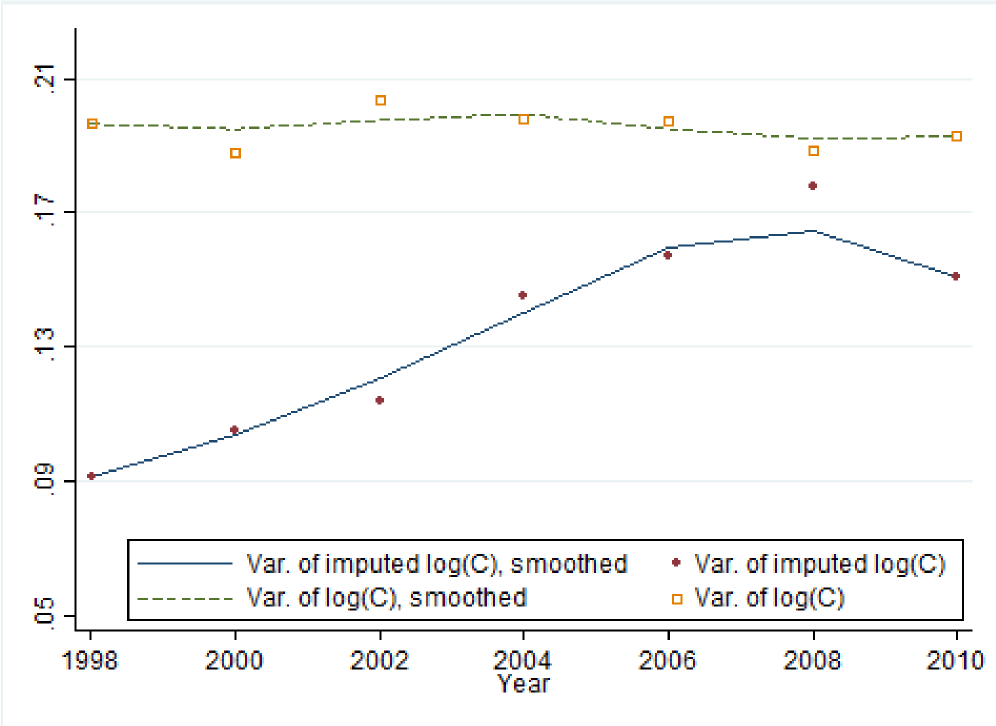
	Non-durables		Non-durables + Hous.		Non-durables + Edu.	
	Real	BPP	Real	BPP	Real	BPP
Variance of Permanent Shocks						
Year = 2000	0.081	0.043	0.062	0.115	0.047	0.109

	(0.017)	(0.014)	(0.014)	(0.017)	(0.012)	(0.017)
Year = 2002	0.059	0.027	0.056	0.087	0.048	0.09
	(0.013)	(0.01)	(0.013)	(0.012)	(0.013)	(0.012)
Year = 2004	0.037	0.039	0.039	0.057	0.039	0.055
	(0.009)	(0.009)	(0.009)	(0.006)	(0.009)	(0.006)
Year = 2006	0.031	0.029	0.033	0.041	0.032	0.039
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)
Year = 2008	0.043	0.042	0.044	0.042	0.044	0.042
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Year = 2010	0.092	0.06	0.017	0.015	0.077	0.042
	(2.252)	(0.331)	(0.279)	(0.018)	(0.409)	(0.012)
Variance of Transitory Shocks						
Year = 1999	0.271	0.493	0.277	0.089	0.151	0.044
	(2.48)	(0.369)	(0.424)	(0.183)	(0.463)	(0.078)
Year = 2000	0.028	0.042	0.032	0.017	0.034	0.015
	(0.032)	(0.006)	(0.008)	(0.007)	(0.018)	(0.007)
Year = 2001	0.345	1.01	0.387	0.229	0.218	0.106
	(2.966)	(1.002)	(0.614)	(0.504)	(0.733)	(0.18)
Year = 2002	0.042	0.05	0.042	0.024	0.04	0.022
	(0.045)	(0.009)	(0.01)	(0.005)	(0.021)	(0.006)
Year = 2003	0.314	1.802	0.26	0.293	0.185	0.144
	(2.335)	(2.572)	(0.298)	(0.658)	(0.527)	(0.247)
Year = 2004	0.055	0.058	0.055	0.031	0.053	0.031
	(0.051)	(0.015)	(0.018)	(0.007)	(0.028)	(0.008)
Year = 2005	0.016	1.926	0.147	0.234	0.06	0.125
	(0.94)	(2.898)	(0.136)	(0.518)	(0.186)	(0.212)
Year = 2006	0.041	0.037	0.038	0.026	0.038	0.027
	(0.028)	(0.005)	(0.007)	(0.006)	(0.016)	(0.007)
Year = 2007	0.043	1.23	0.045	0.441	0.007	0.171
	(0.66)	(1.498)	(0.289)	(1.002)	(0.382)	(0.291)
Year = 2008	0.04	0.045	0.041	0.051	0.041	0.044
	(0.028)	(0.006)	(0.009)	(0.012)	(0.018)	(0.013)
Year = 2009	0.033	0.712	0.005	0.591	0.051	0.181

	(1.23)	(0.576)	(0.455)	(1.359)	(0.349)	(0.307)
Year = 2010	0.003	0.026	0.077	0.065	0.014	0.043
	(2.284)	(0.328)	(0.283)	(0.018)	(0.425)	(0.012)
Partial insurance permanent shock						
ϕ	0.128	0.198	0.2	0.067	0.244	0.054
	(0.041)	(0.054)	(0.051)	(0.023)	(0.057)	(0.025)
Partial insurance transitory shock						
ψ	0.116	0.219	0.251	0.67	0.202	1.019
	(0.572)	(0.219)	(0.354)	(0.729)	(0.504)	(0.744)
Serial correlation of transitory shock						
θ	-0.157	-0.036	-0.136	-0.109	-0.206	-0.218
	(1.074)	(0.041)	(0.167)	(0.227)	(0.567)	(0.325)
Variance of unobservables (slope heterogeneity)						
σ^2_ξ	0.071	0.019	0.059	0.013	0.061	0.012
	(0.011)	(0.029)	(0.019)	(0.021)	(0.015)	(0.021)

From the results based on the real consumption data, we can arrive at the following conclusions. First, the estimate of ϕ, the partial insurance coefficient for the permanent shock, which is significant in all the three measurements, provides evidence in favor of some partial insurance. In particular, a 10 percent permanent income shock induces a 1.28 percent permanent change in consumption. The estimates on ψ are all insignificant, which accords with a simple PIH model. Here we have to keep in mind that these parameters are not comparable with the results from BPP or any other research performed on this topic since the "consumption" we defined using the PSID data is only partial consumption; hence this parameter only reflects the relationship based on this partial consumption and income instead of the total non-durable consumption. We defer the discussion of the economic implication to the next part. Second, with more "durable" components in the consumption definition, the estimate of ϕ is higher than the baseline estimation, indicating that the health, education and durable related service consumption are more difficult to insure compared to common non-durables. Third, we can see that the estimation of both permanent shock variances and transitory shock variances are not sensitive to the use of different consumption definitions, and are relatively stable within these periods. The MA parameter for the transitory shock is stably

Figure 3.6: Variance of the true and imputed consumption



small and insignificant. The variance of unobserved heterogeneity in the consumption is significant and comparable in amount with the variances of shocks.

We can then verify the validity of the BPP method by comparing the results from the imputed consumption with those from the real consumption. First, the comparison based on non-durables plus are more reliable given the comparability of the PSID and CEX data sets we discussed in last section. And the results differ most with the non-durable plus education and health definition, which is the most comparable definition; for example, the estimate of ϕ, which we are most interested in, is about 1/5 of the real estimate. But notably that the estimates of the variance of permanent shocks, except for the beginning and ending periods, are fairly close when the BPP method is compared with the real data. This is due to the fact that the identification of the permanent shocks âĂŞ aside from the beginning and ending periods âĂŞ comes from the income data, which are identical in the two methods.

The BPP paper claims that the imputation error can be included in the model setup as a measurement error in the consumption, and that the variances of those errors can be precisely measured. Hence, we can compare the estimation based on the real data with that based on the imputation data with measurement errors. For a more reliable comparison, we provide the results based on

Figure 3.7: Kernel Density of the true and imputed consumption

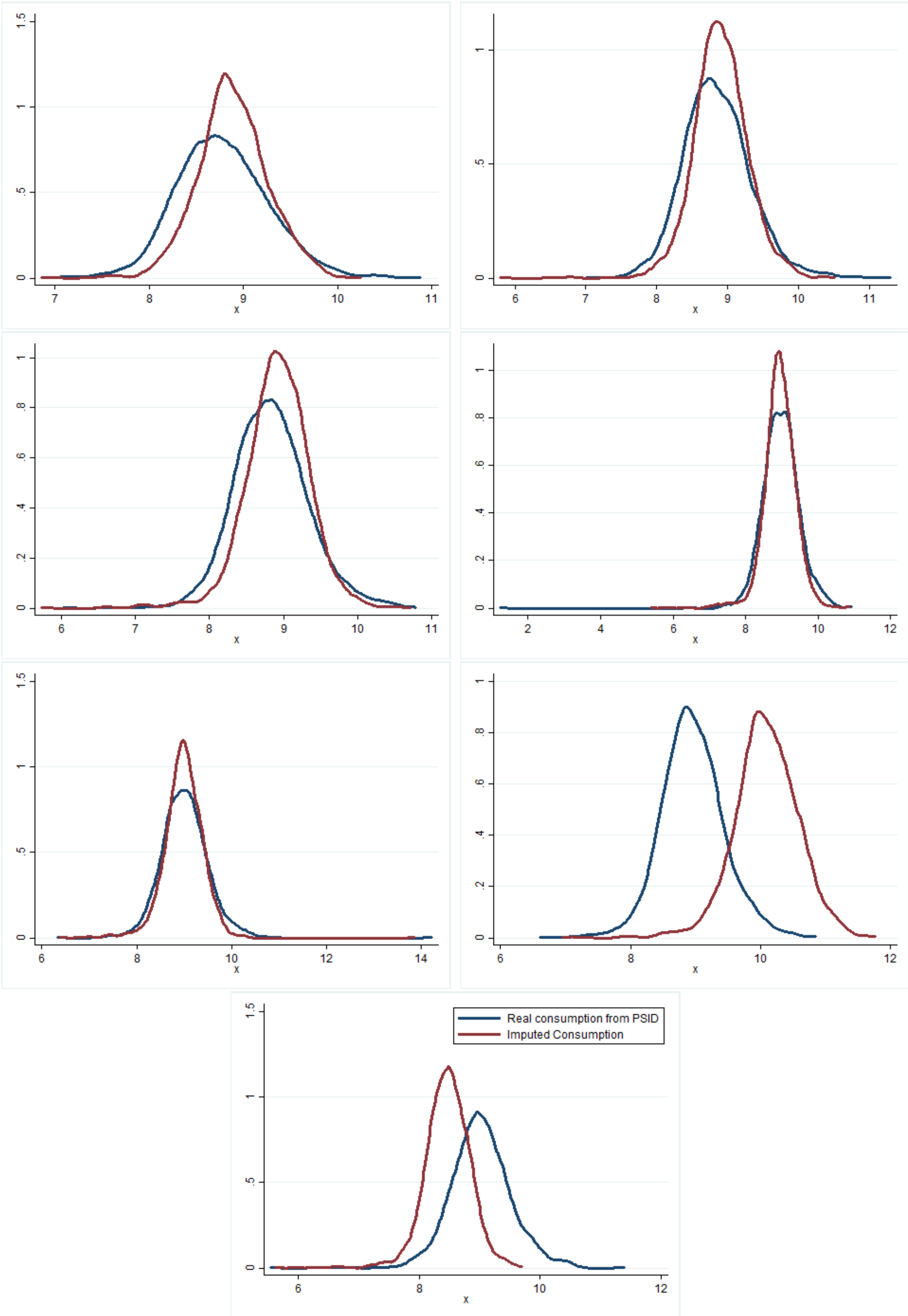
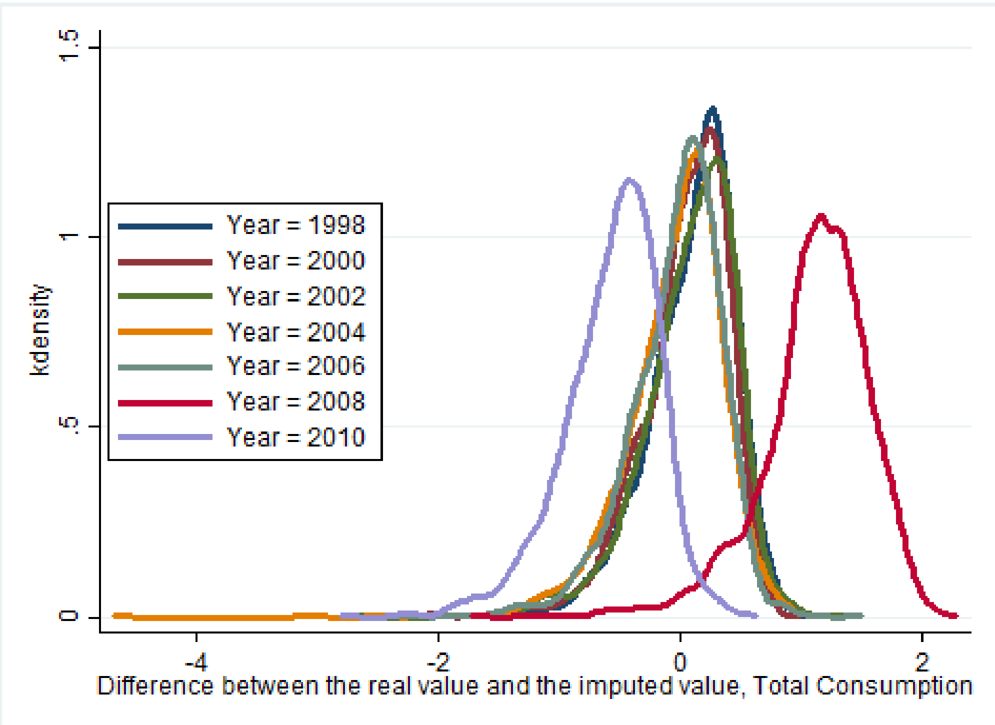


Table 3.5: Summary statistics of the difference

Year	1998	2000	2002	2004	2006	2008	2010
Obs	1,629	1,792	1,832	1,827	1,841	1,825	1,661
Mean of $\Delta log(C)$	0.105	0.044	0.082	-0.038	-0.029	1.124	-0.563
Mean of $log(C)$	8.777	8.856	8.838	8.975	9.010	8.955	9.024
Var of $\Delta log(C)$	0.360	0.367	0.393	0.416	0.385	0.445	0.396
Var of $log(C)$	0.489	0.482	0.504	0.524	0.498	0.491	0.500
Min of $\Delta log(C)$	-2.038	-2.255	-2.750	-4.667	-2.327	-1.711	-2.788
Max of $\Delta log(C)$	1.122	0.995	1.361	0.933	1.500	2.283	0.632

Figure 3.8: Kernel Density of difference between the true and imputed consumption



the non-durable plus here. We can see that with the extra assumption that the two ending periods share the same permanent variances with the third and fourth last periods, and that the the last period shares the same transitory variances with the second last periods decreases the estimates of the partial insurances. But the inclusion of the measurement errors does not necessarily improve the estimations; for example, for the first definition, including measurement errors makes the BPP estimation differ more from the real results, though not significantly.

Table 3.7: Estimates with Measurement Errors

	Nondurables + Hous.			Nondurables + Edu.		
	Real	BPP	BPP with m.e.	Real	BPP	BPP with m.e.
Variance of Permanent Shocks						
Year = 2000	0.067	0.084	0.074	0.056	0.096	0.074
	(0.015)	(0.017)	(0.017)	(0.013)	(0.016)	(0.017)
Year = 2002	0.053	0.042	0.048	0.048	0.052	0.05
	(0.013)	(0.012)	(0.013)	(0.013)	(0.012)	(0.013)
Year = 2004	0.038	0.038	0.04	0.038	0.046	0.038
	(0.009)	(0.008)	(0.009)	(0.009)	(0.006)	(0.009)
Year = 2006	0.03	0.035	0.033	0.028	0.033	0.033
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)
Year = 2008	0.047	0.044	0.043	0.048	0.042	0.044
	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
Variance of Transitory Shocks						
Year = 2000	1.65	0.423	0.003	0.336	0.085	0.206
	(6.702)	(0.626)	(0.935)	(1.827)	(0.09)	(0.466)
Year = 2001	0.037	0.031	0.035	0.038	0.022	0.034
	(0.007)	(0.006)	(0.077)	(0.02)	(0.006)	(0.013)
Year = 2002	1.872	1.432	0.13	0.466	0.147	0.831
	(7.366)	(1.658)	(1.817)	(2.888)	(0.16)	(4.789)
Year = 2003	0.046	0.05	0.039	0.043	0.033	0.044
	(0.009)	(0.008)	(0.101)	(0.022)	(0.007)	(0.019)

Year = 2004	0.368	2.463	0.12	0.313	0.18	0.593
	(5.143)	(3.556)	(1.395)	(1.53)	(0.202)	(2.665)
Year = 2005	0.057	0.059	0.047	0.052	0.04	0.057
	(0.016)	(0.014)	(0.117)	(0.03)	(0.01)	(0.025)
Year = 2006	1.027	1.937	0.047	0.032	0.164	0.097
	(2.576)	(2.697)	(0.215)	(2.277)	(0.179)	(0.698)
Year = 2007	0.037	0.033	0.033	0.037	0.032	0.037
	(0.006)	(0.005)	(0.079)	(0.022)	(0.006)	(0.012)
Year = 2008	0.5	0.512	0.008	0.235	0.123	0.005
	(2.324)	(0.689)	(1.076)	(1.149)	(0.129)	(1.565)
Year = 2009	0.037	0.043	0.044	0.034	0.038	0.045
	(0.007)	(0.005)	(0.097)	(0.018)	(0.008)	(0.014)
Partial insurance of permanent shock						
ϕ	0.155	0.089	0.087	0.192	0.105	0.13
	(0.043)	(0.033)	(0.036)	(0.047)	(0.032)	(0.041)
Partial insurance of transitory shock						
ψ	0.091	0.183	0.109	0.105	0.874	0.103
	(0.368)	(0.167)	(1.096)	(0.735)	(0.471)	(0.365)
Serial correlation of transitory shock						
θ	-0.029	-0.026	-0.277	-0.092	-0.223	-0.09
	(0.092)	(0.033)	(3.342)	(0.534)	(0.205)	(0.371)
Variance unobservable (slope heterogeneity)						
σ^2_ξ	0.059	0.026	0.005	0.061	0	0.003
	(0.031)	(0.021)	(0.011)	(0.025)	(0.028)	(0.009)

We can also test the fit of the model to determine how far the BPP estimation is from the true data. In Figure 9, we plot the actual variances of income growth, consumption growth and the covariance of income growth and consumption growth together with their predicted values using the real data and imputed consumption data respectively. We repeat this exercise for all three definitions of consumption and we can observe a quite obvious divergence from the BPP estimation and the real data. It's worth mentioning that this doesn't mean the BPP method doesn't do a good job in terms of fitting the model, only that the data it is fitted to is inaccurate. In Figure 10, we provide the comparison of the variance of the imputed consumption and the covariance of imputed consumption with income to their estimated counterparts respectively for all three consumption definitions. In this figure, we can see that the BPP estimation did a very good job fitting the imputed data.

Figure 3.9: Goodness of fit of BPP

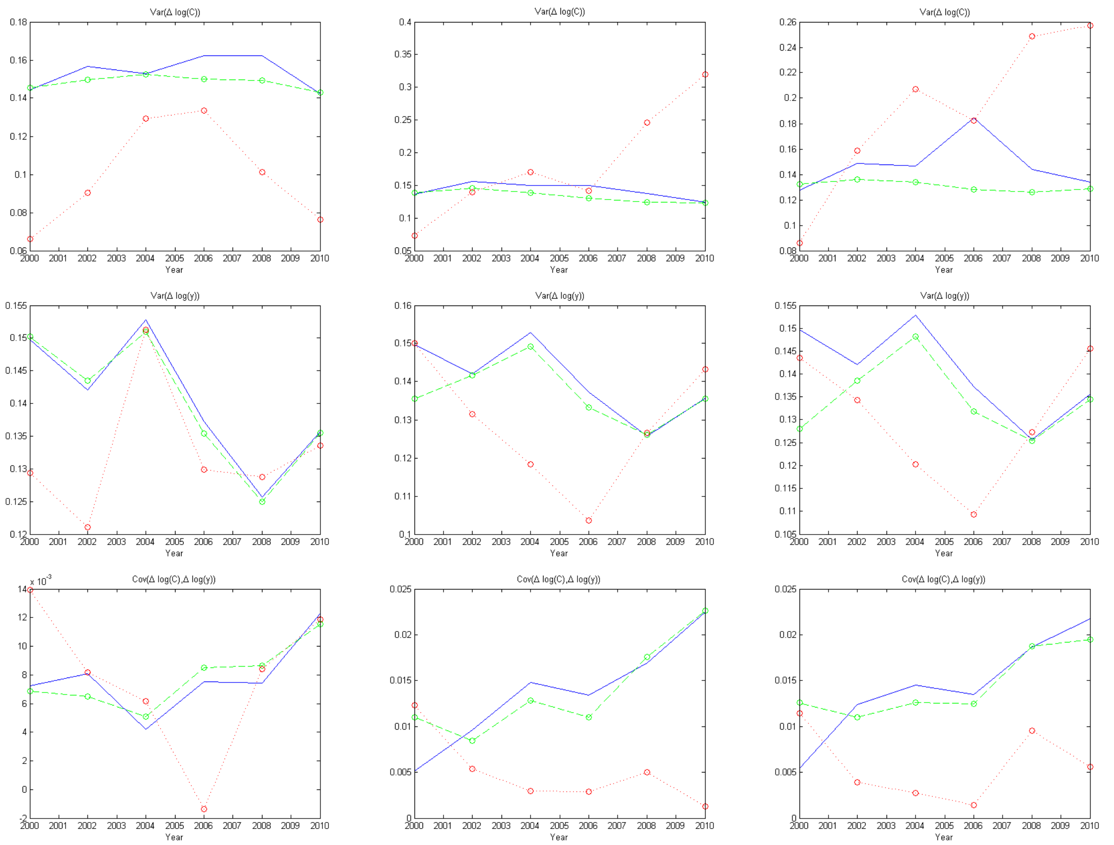
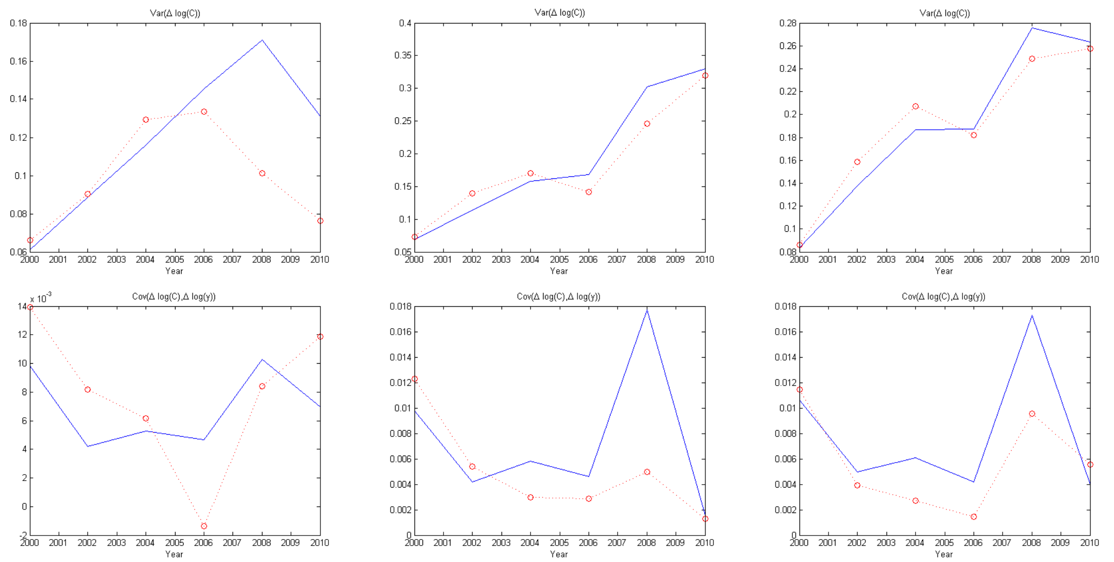


Figure 3.10: Goodness of fit of BPP



BLL

Now, we turn to our proposed method, which we refer to as BLL. Due to the fact that the consumption data and income data are not in the same data sets âĂŞ and what's worse is that the consumption data in CEX is repeated cross-sectional data âĂŞ the consumption data for the different years are also not in the same data set. This means that there is a situation in which there is no way to calculate $cov(\Delta log(C), \Delta log(y))$. With the lack of the series of $cov(\Delta log(C), \Delta log(y))$, we can no longer identify all the variance in the model (see Appendix 2 for details), but we can still identify the parameters that we are interested in, which are the two insurance parameters.

Partial Insurance for the Permanent Shock

From Eq.(3.1), we can write

$$\zeta_{i,t} + \zeta_{i,t-1} = \Delta_2 y_{i,t} - \Delta_2 v_{i,t}$$

which together with (3.2) implies that

$$\Delta_2 c_{i,t} = \phi \Delta_2 y_{i,t} + \eta_{i,t} \tag{3.3}$$

where we define

$$\begin{aligned}\eta_{i,t} &= (\xi_{i,t} + \xi_{i,t-1}) + \psi(\varepsilon_{i,t} + \varepsilon_{i,t-1}) - \phi \Delta_2 v_{i,t} \\ &= (\xi_{i,t} + \xi_{i,t-1}) + (\psi - \phi)\varepsilon_{i,t} + (\psi - \phi\theta)\varepsilon_{i,t-1} - \phi(\varepsilon_{i,t-2} + \theta\varepsilon_{i,t-3})\end{aligned}$$

Because in our case $\Delta_2 c_{i,t}$ is unavailable since $P^c_{i,t}$ and $P^c_{i,t-2}$ are in two different data sets, we rewrite the above equation as

$$\mathbf{P}^c_{i,t} - \mathbf{P}^c_{i,t-2} = \phi \Delta_2 y_{i,t} + \eta_{i,t} \tag{3.4}$$

The permanent income component ζ is in both $\Delta_2 c_{i,t}$ and $\Delta_2 y_{i,t}$, while it is not in $\eta_{i,t}$. Hence, valid IVs $z_{y;i;t}$ for estimating equation (8) could be the variables that are correlated with $\zeta_{i,t}$ but uncorrelated with $\varepsilon_{i,t}$ and $\xi_{i,t}$, and at the same time, contained in any two successive CEX datasets and also in the PSID dataset. Here we choose wives' characteristics (wives' ages). And we can prove that ϕ can be identified as follows:

$$\phi = \frac{E\left[\mathbf{P}^c_{i,t} z^{CEX_t}_{y,i,t}\right] - E\left[\mathbf{P}^c_{i,t-2} z^{CEX_{t-1}}_{y,i,t}\right]}{E\left[\Delta_2 y_{i,t} z^{PSID}_{y,i,t}\right]} \tag{3.5}$$

As the instrument variable z_y is available in all three data sets, we have n observations: $\{z_{y,i,t}\}_{i\in I}$ where $I = I_1 \cup I_2 \cup I_3$. Let $P_k(z) = [p_1(z), \ldots, p_k(z)]'$ be a k-dimensional vector of basis functions for any positive integer k. For any k and any n, we define

$$P_{n,k} = [P_k(Z_1), \ldots, P_k(Z_n)]' .$$

It is clear that $P_{n,k}$ is an $n \times k$ matrix. Accordingly, we define P_{n_1,k_1} , $P_{n_2,k_2}$and P_{n_3,k_3} which are $n_1 \times k_1$, $n_2 \times k_2$ and $n_3 \times k_3$ matrices respectively.

The conditional mean $h_0(Z) = E\left[Y|\, Z\right]$ can be estimated by

$$\widehat{h}_{n_1,n_2}(Z) = P_{k_1}(Z)'(P'_{n_1,k_1}P_{n_1,k_1})^{-1}P'_{n_1,k_1}\mathbf{P}^c_{n_1,t} - P_{k_2}(Z)'(P'_{n_2,k_2}P_{n_2,k_2})^{-1}P'_{n_2,k_2}\mathbf{P}^c_{n_2,t-2} \tag{3.6}$$

Using the third data set, we get the following estimator of the conditional mean function $\phi_0(Z, \theta) = E\left[g(X, \theta)|\, Z\right]$ for any θ:

$$\widehat{\phi}_{n_3}(Z, \phi) = P_{k_3}(Z)'(P'_{n_3,k_3}P_{n_3,k_3})^{-1}P'_{n_3,k_3}\phi\Delta_2 y_{n3,t} \tag{3.7}$$

Using the estimators of $h_0(Z)$ and $\phi_0(Z, \phi)$, we can construct the estimator of ϕ_0 via the minimum distance (MD) estimation:

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} n^{-1}\sum_{i\in I}\left[\left|\widehat{h}_{n_1,n_2}(Z_i) - \widehat{\phi}_{n_3}(Z_i, \phi)\right|^2\right] \tag{3.8}$$

where Θ denote some subspace of R^{d_θ}.

Results Comparison and Discussion

In Table 8, we report the results from the BPP and BLL methods, and compare them with the results from the true data.

Due to the identification problem we mentioned earlier, we can only compare the parameters of the two kinds of partial insurance shocks, which are the ones of greatest interest in this study. We can see that the BLL method provides results that are very similar to the results from the true data, implying the reliance of this new method.

We can see from the results that, the estimate of ϕ, which is the partial insurance coefficient for the permanent shock, is significantly positive but significantly smaller than 1, indicates that if consumption is only partially insured from the permanent shock. In particular, we find that a 10 percent permanent income shock induces a 1.3 percent permanent change in consumption. This finding is consistent with the excessively smooth consumption discovered in both aggregate and micro data, i.e., it reacts insufficiently to permanent income shocks to be consistent with the theory (Campbell and Deaton (1989); Attanasio and Pavoni (2006)).

In the original BPP work, the parameter is as high as 0.64, which means that a 10 percent permanent income shock induces a 6.4 percent permanent change in consumption. Hence in their work, they conclude that their results can be simply explained by PIH but not support the excessively smooth consumption phenomenon. Through this comparison, we can see that the imputation method delivers a significantly higher (almost double the size) estimation results compared to those based on true data. Although, it's hard to identify the direction of the estimation bias of their method theoretically from the estimation process, this comparison gives us a rough estimation of the size of the overestimation in their original work. An overestimation at this degree would completely change the supporting theories and lead to totally different results. As to the source of the bias, it may included several aspects, but one most important of these is misspecification of the consumption variance. In contrast, our estimation differs slightly with the real data result, but gives a much closer estimation with less than 10% upscaling.

For the partial insurance for transitory shock (i.e., ψ), all three columns show that the parameter is insignificant. But again, the point estimator of BPP varies largely with the read data results. This evidence accords with a simple PIH model and we cannot reject the null that there is full insuring behavior with respect to transitory shocks.

3.5 Conclusion

In this paper, we re-evaluate the relationship between the dynamic processes of consumption and income with a new estimation method described in Buchinsky, Li and Liao (2016,a) which deals with the data problem of having common containing variables in different data sets. The inclusion of additional consumption data in the PSID starting with the 1999 wave grants us with an precious opportunity to verify our new method with empirical data. The validity of the new method makes it possible for us to revisit the topic using the historical incomplete data in this topic and also gives confidence for application of the new method to other topics with similar data obstacle.

We proceeded in the same spirit as the seminal BPP paper by assuming some, but not necessarily full, insurance and taking into consideration the distinctions between transitory and permanent shocks. We show in our paper that the BPP method gives biased estimation results compared with those from the true data, while the BLL succeeds in giving similar results. If we consider the results based on true data as the "true" result, the comparison of estimators with those based on true data can be used as a test of the validity for different empirical strategies. After theoretically proving the commonly used BPP method is biased, We support our theoretical finding by making comparisons among the estimation results of BPP, BLL and the true data and concluding that BLL gives much closer results compared to the true data results while BPP differs to that significantly. This results suggest the validity of the newly introduced BLL result. In the future, we will employ this method to revisit the dynamic relationship of consumption and income between 1980 to the most current in the hope of shedding more lights in this line of reseach.

3.6 Appendix

3.6.1 The income and consumption process

In line with BPP(2008), we assume the real (log) income $logY$ can be decomposed into a permanent component $\mathbf{P}$ and a mean-reverting transitory component v. The income process for each household i is

$$log\mathbf{Y}_{i,t} = \mathbf{Z}'_{i,t}\varphi_t + \mathbf{P}_{i,t} + v_{i,t}$$

where t indexes time and $\mathbf{Z}$ is a set of income characteristics observables and known by consumers at time t. We assume that the permanent component $\mathbf{P}_{i,t}$ follows a martingale process of the form

$$\mathbf{P}_{i,t} = \mathbf{P}_{i,t-1} + \zeta_{i,t}$$

where $\zeta_{i,t}$ is serially uncorrelated and the transitory component $v_{i,t}$ follows an MA(1) process:

$$v_{i,t} = \varepsilon_{i,t} + \theta\varepsilon_{i,t-1}$$

If we define $y_{i,t} = log\mathbf{Y}_{i,t} - \mathbf{Z}'_{i,t}\varphi_t$ as the log of real income net of predictable individual components, then the unexplained two-period income growth is

$$\Delta_2 y_{i,t} = (\zeta_{i,t} + \zeta_{i,t-1}) + (v_{i,t} - v_{i,t-2}) = (\zeta_{i,t} + \zeta_{i,t-1}) + \Delta_2 v_{i,t}$$

Similarly, we write the unexplained two-period consumption growth as

$$\Delta_2 c_{i,t} = \phi(\zeta_{i,t} + \zeta_{i,t-1}) + \psi(\varepsilon_{i,t} + \varepsilon_{i,t-1}) + (\xi_{i,t} + \xi_{i,t-1})$$

where the random term $\xi_{i,t}$ presents innovations in consumption that are independent of those in income, this may capture measurement error in consumption, preference shocks, innovation to higher moments of the income process, etc.

In other words, we assume the consumption follows the process:

$$log\mathbf{C}_{i,t} = \mathbf{Z}'_{i,t}\varphi^{\mathbf{c}}{}_t + \ \mathbf{P}^c_{i,t}$$

where

$$\mathbf{P}^c_{i,t} = \mathbf{P}^c_{i,t-1} + \phi\zeta_{i,t} + \psi\varepsilon_{i,t} + \xi_{i,t}$$

We as well assume that $\zeta_{i,t}$,$\varepsilon_{i,t}$ and $\xi_{i,t}$ are mutually uncorrelated processes. then we can impose covariance restrictions on the bivariate process to identify the parameters of interest. In particular, equation (1) can be used to derive the following covariance restrictions in panel data:

$$cov(\Delta_2 y_t, \Delta_2 y_{t+s}) = \begin{cases} \sigma^2_{\zeta,t} + \sigma^2_{\zeta,t-1} + \sigma^2_{v,t} + \sigma^2_{v,t-2} & \text{if } s = 0 \\ \sigma^2_{v,t} & \text{if } s = 2 \\ 0 & \text{if } s > 2 \end{cases} \tag{3.9}$$

in which

$$\sigma^2_{v,t} = \sigma^2_{\varepsilon,t} + \theta^2\sigma^2_{\varepsilon,t-1}$$

The panel data restrictions on consumption growth from equation (2) are as follows:

$$cov(\Delta_2 c_t, \Delta_2 c_{t+s}) = \begin{cases} \phi^2(\sigma^2_{\zeta,t} + \sigma^2_{\zeta,t-1}) + \psi^2(\sigma^2_{\varepsilon,t} + \sigma^2_{\varepsilon,t-1}) + (\sigma^2_{\xi,t} + \sigma^2_{\xi,t-1}) & \text{if } s = 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

Finally, the covariance between income growth and consumption growth at various lags is

$$cov(\Delta_2 c_t, \Delta_2 y_{t+s}) = \begin{cases} \phi(\sigma^2_{\zeta,t} + \sigma^2_{\zeta,t-1}) + \psi(\sigma^2_{\varepsilon,t} + \theta\sigma^2_{\varepsilon,t-1}) & \text{if } s = 0 \\ -\psi(\sigma^2_{\varepsilon,t} + \theta\sigma^2_{\varepsilon,t-1}) & \text{if } s = 2 \\ 0 & \text{if } s > 2 \text{ or } s < 0 \end{cases} \tag{3.11}$$

3.6.2 Identification

From the Eq.(1) and Eq.(2), we can see that the parameters to identify are:θ,σ^2_ζ,σ^2_ε,σ^2_ξ,ϕ and ψ. Here I show how the model can be identified with N years of biannual panel data (label as $l = 2t - 1$ for $t = 1, \ldots, N$), or $N - 1$ years of biannual differences, $\{\Delta_2 y_{i,t}, \Delta_2 c_{i,t}\}$ for and discuss the extension with measure errors on consumption.

I start with the simplest model with MA(1) process of transitory component and no measurement error.

1. **Identification of** $(\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}})$

It's clear that $\sigma^2_{\zeta_t}$ and $\sigma^2_{\zeta_{t-1}}$ cannot be identified separately, but can be identified as a unit. From Eq. (3)

$$\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}} = E[\Delta_2 y_t] + cov(\Delta_2 y_{t-2}, \Delta_2 y_t) + cov(\Delta_2 y_t, \Delta_2 y_{t+2}) \tag{3.12}$$

for $\quad t \in \{5, \dots, 2N-3\}$

2. **Identification of ϕ**

   From Eq.(3) and (5), we can have

$$\phi = \frac{cov(\Delta_2 c_t, \Delta_2 y_t) + cov(\Delta_2 c_t, \Delta_2 y_{t+2})}{\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}}} \tag{3.13}$$

   which in turn gives us

$$\sigma^2_{\zeta_{t-2}} + \sigma^2_{\zeta_{t-3}} = \frac{cov(\Delta_2 c_{t-2}, \Delta_2 y_{t-2}) + cov(\Delta_2 c_{t-2}, \Delta_2 y_t)}{\phi}$$

   so we can identify $\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}}$ for $t = 3$

3. **Identification of ψ, θ and $\sigma^2_{\varepsilon t}$**

   After identifying $(\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}})$ for the first two periods and ϕ, we have the following $2 * (N-2) + (N-1)$ equations with $3 + 2 * (N-2)$ unknowns $\{\psi, \theta, \sigma^2_{\xi}, \sigma^2_{\varepsilon_l}\}$ for $l \in [1, 2N-3]$

$$\begin{aligned}
-\psi(\sigma^2_{\varepsilon_l} + \theta\sigma^2_{\varepsilon_{l-1}}) &= cov(\Delta_2 c_l, \Delta_2 y_{l+2}) && l = 3, \dots, 2N-3 \\
-(\sigma^2_{\varepsilon_l} + \theta^2\sigma^2_{\varepsilon_{l-1}}) &= cov(\Delta_2 y_l, \Delta_2 y_{l+2}) && l = 3, \dots, 2N-3 \\
\psi^2(\sigma^2_{\varepsilon_l} + \sigma^2_{\varepsilon_{l-1}}) + 2\sigma^2_{\xi} &= E(\Delta_2 c^2_l) - \phi^2(\sigma^2_{\zeta_l} + \sigma^2_{\zeta_{l-1}}) && l = 3, \dots, 2N-3
\end{aligned} \tag{3.14}$$

   It's an overidentification problem if $N > 4$, an exact-identification problem if $N = 4$, and under-identified if $N < 4$

4. **Identification of the variances of the last period**

   For the last period, the only available covariances are $E(\Delta_2 y_t)$, $E(\Delta_2 c^2_t)$ and $cov(\Delta_2 y_t, \Delta_2 c_t)$, but since we've already identified all the insurance coefficients ψ, ϕ and also the MA(1) coefficient θ and the idiosyncratic variance σ^2_{ξ}, the only unknown left are three: $(\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}})$, $\sigma^2_{\varepsilon_t}$ and $\sigma^2_{\varepsilon_{t-1}}$. So these three can be exact identified with the last three covariances.

5. **Identification with time-varying measurement errors**

   If instead of the above consumption process, we assume there is a time-varying measurement error for consumption, i.e.

$$logC^*_{i,t} = logC_{i,t} + u^c_{i,t}$$

where C^* denote measured consumption, C is true consumption, and u^c the measurement error, which is a draw from a distribution with variance $\sigma^2_{u^c_t}$.

Then the Eq. (2) becomes

$$\Delta_2 c^*_{i,t} = \phi(\zeta_{i,t} + \zeta_{i,t-2}) + \psi(\varepsilon_{i,t} + \varepsilon_{i,t-1}) + (\xi_{i,t} + \xi_{i,t-1}) + (u^c_t - u^c_{t-2}) \qquad (2^*)$$

and the covariance of consumption growth, i.e. Eq.(4), becomes

$$cov(\Delta_2 c^*_t, \Delta_2 c^*_{t+s}) = \begin{cases} E(\Delta_2 c^2_t) + \sigma^2_{u^c,t} + \sigma^2_{u^c,t-2} & \text{if } s = 0 \\ -\sigma^2_{u^c,t} & \text{if } s = 2 \\ 0 & \text{otherwise} \end{cases}$$

Then at Step 3, first of all, with $N-2$ equations of $cov(\Delta_2 c^*_t, \Delta_2 c^*_{t+2})$, we can identify $\{\sigma^2_{u^c_t}\}$, for $t = 3, \ldots, 2N-3$, then with the other $2*(N-2)+(N-1)$ equations, we have the old $2*(N-2)+3$ unknowns, and 1 extra unknowns to identify $\{\sigma^2_{u^c,1}\}$, which means the system is still overidentified if $N > 5$, exact-identified if $N = 5$ and under-identified if $N < 5$.

But for the last period, with still three equations for covariances $E(\Delta_2 y_t)$, $E(\Delta_2 {c^*}^2_t)$ and $cov(\Delta_2 y_t, \Delta_2 c^*_t)$, we now have four unknowns: $(\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}})$, $\sigma^2_{\varepsilon_t}$, $\sigma^2_{\zeta_{t-1}}$ and $\sigma^2_{u^c_t}$. So we cannot identify all the parameters of the last period.

3.6.3 Identification with consumption and income growths are in separate data sets

Now with the income growth $\{\Delta_2 y_t\}$ and the consumption growth $\{\Delta_2 c_t\}$, we can still identify $(\sigma^2_{\zeta_t} + \sigma^2_{\zeta_{t-1}})$ for $t = 5, \ldots, 2N-3$ as we did in Step 1 before, the major differences are for insurance parameters ϕ and ψ.

1. **Identification of ϕ**

From Eq.(3.1), we can write

$$\zeta_{i,t} + \zeta_{i,t-1} = \Delta_2 y_{i,t} - \Delta_2 v_{i,t}$$

which together with (3.2) implies that

$$\Delta_2 c_{i,t} = \phi \Delta_2 y_{i,t} + \eta_{i,t} \qquad (3.15)$$

where we define

$$\eta_{i,t} = (\xi_{i,t} + \xi_{i,t-1}) + \psi(\varepsilon_{i,t} + \varepsilon_{i,t-1}) - \phi\Delta_2 v_{i,t}$$
$$= (\xi_{i,t} + \xi_{i,t-1}) + (\psi - \phi)\varepsilon_{i,t} + (\psi - \phi\theta)\varepsilon_{i,t-1} - \phi(\varepsilon_{i,t-2} + \theta\varepsilon_{i,t-3})$$

Because in our case, $\Delta_2 c_{i,t}$ is unavailable since $P^c_{i,t}$ and $P^c_{i,t-2}$ are in two different data sets, we rewrite the above equation into

$$\mathbf{P}^c_{i,t} - \mathbf{P}^c_{i,t-2} = \phi\Delta_2 y_{i,t} + \eta_{i,t} \tag{3.16}$$

Since the permanent income component ζ is in both $\Delta_2 c_{i,t}$ and $\Delta_2 y_{i,t}$, while it is not in $\eta_{i,t}$. Hence, valid IVs $z_{y;i;t}$ for estimating equation (8) could be the variables which are correlated with $\zeta_{i,t}$ but uncorrelated with $\varepsilon_{i,t}$ and $\xi_{i,t}$, and in the same time, contained in any two successive CEX datasets and also in PSID dataset. Here we choose wives' characteristics (wives' ages).

Then one can prove:

$$\phi = \frac{E\left[\mathbf{P}^c_{i,t} z^{CEX_t}_{y,i,t}\right] - E\left[\mathbf{P}^c_{i,t-2} z^{CEX_{t-1}}_{y,i,t}\right]}{E\left[\Delta_2 y_{i,t} z^{PSID}_{y,i,t}\right]} \tag{3.17}$$

2. **Identification of ψ and θ**

Similarly, from the equation (3.2),

$$\varepsilon_{i,t} + \varepsilon_{i,t-1} = \frac{\Delta_2 c_{i,t} - \phi(\zeta_{i,t} + \zeta_{i,t-1}) - (\xi_{i,t} + \xi_{i,t-1})}{\psi}$$

which together with (3.1) implies that

$$\begin{aligned} \Delta_2 y_{i,t} &= \frac{\Delta_2 c_{i,t}}{\psi} - \frac{\theta\Delta_2 c_{i,t-2}}{\psi} + u_{i,t} \\ &= \frac{\mathbf{P}^c_{i,t} - \mathbf{P}^c_{i,t-2}}{\psi} - \frac{\theta(\mathbf{P}^c_{i,t-2} - \mathbf{P}^c_{i,t-4})}{\psi} + u_{i,t} \end{aligned} \tag{3.18}$$

where we define

$$\begin{aligned} u_{i,t} &= \frac{(\psi - \phi)(\zeta_{i,t} + \zeta_{i,t-1})}{\psi} + \frac{\theta\phi(\zeta_{i,t-2} + \zeta_{i,t-3})}{\psi} - \frac{\xi_{i,t} + \xi_{i,t-1}}{\psi} \\ &\quad + \frac{\theta(\xi_{i,t-2} + \xi_{i,t-3})}{\psi} + (\theta - 1)(\varepsilon_{i,t-1} + \varepsilon_{i,t-2}). \end{aligned} \tag{3.19}$$

It is clear that the transitory income components $\varepsilon_{i,t}$ and $\varepsilon_{i,t-3}$ are in $\Delta_2 y_{i;t}$, $\Delta_2 c_{i;t}$ and $\Delta_2 c_{i;t-2}$, while they are not in $u_{i;t}$. Hence, valid IVs $z_{c;i;t}$ for estimating equation (13) could be the variables which are correlated with $\varepsilon_{i,t}$ and $\varepsilon_{i,t-3}$ but uncorrelated with $\zeta_{i,t}$, $\xi_{i,t}$ or

$\varepsilon_{i,t-1}$, $\varepsilon_{i,t-2}$. But it's potentially hard to find any variables correlated with $\varepsilon_{i,t}$ but not with the lagged ones. For example, the change of per capita GDP growth rate of the residential states would be a potential choice, but it's hard to argue that is not correlated with last two periods' transitory income shocks . But supposedly we are lucky enough to find at least two valid instruments $z_{c,i,t}^{PSID}$, we can identify ψ and θ at the same time through the following moment conditions.

$$\psi\mathbf{E}[\Delta_2 y_{i,t} z_{c;i;t}^{PSID}] = \mathbf{E}[\mathbf{P}_{i,t}^c z_{c;i;t}^{CEX}] - (\theta+1)[\mathbf{P}_{i,t-2}^c z_{c;i;t}^{CEX}] + \theta[\mathbf{P}_{i,t-4}^c z_{c;i;t}^{CEX}]$$

3. **Identification of σ_ε^2 and σ_ξ^2**

After identifying $(\sigma_{\zeta_t}^2 + \sigma_{\zeta_{t-1}}^2)$ for the first two periods and ϕ, we have the following $2*(N-4)$ equations with $2*(N-4)+1$ unknowns $\{\sigma_\xi^2, \sigma_{\varepsilon_l}^2\}$ for $l \in [5, 2N-3]$

$$\begin{aligned} -(\sigma_{\varepsilon_l}^2 + \theta^2\sigma_{\varepsilon_{l-1}}^2) &= cov(\Delta_2 y_l, \Delta_2 y_{l+2}) & l = 5,\dots,2N-3 \\ \psi^2(\sigma_{\varepsilon_l}^2 + \sigma_{\varepsilon_{l-1}}^2) + 2\sigma_\xi^2 + \phi^2(\sigma_{\zeta_l}^2 + \sigma_{\zeta_{l-1}}^2) &= E(\Delta_2 c_l^2) & l = 5,\dots,2N-3 \end{aligned} \tag{3.20}$$

in which

$$E(\Delta_2 c_l^2) = \sigma_{\mathbf{P}_t^c}^2 - \sigma_{\mathbf{P}_{t-2}^c}^2$$

The system is under-identified.This failure of

3.6.4 Identification of the starting and ending periods

Both in BPP and in our paper, it's been shown that with four biannual data, the parameters $\{\theta, \sigma_{\zeta_t}^2, \sigma_{\varepsilon_t}^2, \sigma_\xi^2, \phi, \psi\}$ can be identified. And in BPP, and I quote, "if one has T years of data, only $T-3$ variances of the permanent shock can be identified and only $T-2$ variances of the i.i.d. transitory shock can be identified". But actually with more than four periods' data, it's an overidentified problem for $\{\theta, \phi, \psi, \sigma_\xi^2\}$. Here we can show that that the variances at the starting and ending periods can also be identified.

Let's take BPP data as an example, i.e. we have data from 1978 to 1992 or we can say that we have first difference data from 1979 to 1992. Then with data from 1979 to 1982, the parameters $\{\theta, \phi, \psi, \sigma_\xi^2\}$ are already identified, together with $\sigma_{\zeta_{1981}}^2, \sigma_{\varepsilon_{1981}}^2$. Then with θ known and the following two equations:

$$\sigma_\varepsilon^2 = \frac{E\left[\Delta y_{i,t}\Delta y_{i,t+1}\right]}{\theta - 1} - E\left[\Delta y_{i,t+1}\Delta y_{i,t-1}\right] \tag{3.21}$$

$$\sigma_\zeta^2 = E\left[\Delta y_{i,t}(\Delta y_{i,t} + \frac{\Delta y_{i,t+1} - \theta\Delta y_{i,t-1}}{\theta - 1}) - (\theta + \frac{1}{\theta})\Delta y_{i,t+1}\Delta y_{i,t-1})\right] \tag{3.22}$$

$\sigma^2_{\zeta_{1980}}, \sigma^2_{\varepsilon_{1980}}$ are also identified. With the second moment of Δy:

$$E\left[\Delta y_{i,t}\Delta y_{i,t-2}\right] = -\theta\sigma^2_{\varepsilon,t-2}$$

$\sigma^2_{\varepsilon_{1979}}$ is also identified ($t-2 = 1979$). And together with another second moment of Δy as followed

$$E\left[\Delta y_{i,t}\Delta y_{i,t-1}\right] = (\theta-1)\sigma^2_{\varepsilon,t-1} - \theta(\theta-1)\sigma^2_{\varepsilon,t-2}$$

$\sigma^2_{\varepsilon_{1978}}$ is also identified ($t-1 = 1979$). If together with the second moment of Δc

$$E\left[\Delta c^2_{i,t}\right] = \phi^2\sigma^2_{\zeta_t} + \psi^2\sigma^2_{\varepsilon_t} + \sigma^2_\xi$$

$\sigma^2_{\zeta_{1979}}$ can be identified ($t = 1979$). The last step is to go back to the second moment of Δy:

$$E\left[\Delta y^2_{i,t}\right] = \sigma^2_{\zeta,t} + \sigma^2_{\varepsilon,t} + (\theta-1)^2\sigma^2_{\varepsilon,t-1} + \theta^2\sigma^2_{\varepsilon,t-2}$$

$\sigma^2_{\varepsilon_{1977}}$ is also identified ($t = 1979$). For the ending periods, there are two unknowns left $\sigma^2_{\varepsilon_{1992}}$ and $\sigma^2_{\zeta_{1992}}$ with two equations of $E\left[\Delta y^2_{i,t}\right]$ and $E\left[\Delta c^2_{i,t}\right]$ ($t = 1992$), so clearly those two are also identified.

To sum up, with T periods of first difference data, T variances of permanent shocks can be identified, and $T+2$ variances of transitory shocks can be identified.

If there is measurement errors, which can be identified by

$$E(\Delta c^*_t \Delta c^*_{t+1}) = -\sigma^2_{u^c_t}$$

for all the periods but the first and last ones.

For the last period, we have not two but three unknowns $\sigma^2_{\varepsilon_{1992}}$, $\sigma^2_{\zeta_{1992}}$ and $\sigma^2_{u^c_{1992}}$ with still two equations. Then the last three are underidentified.

The same applies to the first period, we have two equations, but now we have an extra $\sigma^2_{u^c_{1978}}$.

So to make the results comparable, let's assume the last periods variances equal to penultimate ones, in other words, we assume there are $T-1$ permanent shocks, $T+1$ transitory shocks and $T-1$ measurement errors to identify.

Table 3.8: Comparisons of BLL and BPP

	Real	BLL	BPP
ϕ	0.1331	0.1407	0.2404
(Partial insurance perm. shock)	(0.0432)	(0.0453)	(0.0592)
ψ	0.1584	0.1922	-0.0981
(Partial insurance trans. shock)	(0.9839)	(1.0290)	(0.2523)

[AB17] N. Ashraf and O. Bandiera. *Social Incentives in Organizations.* In preparation for the Annual Review of Economics, 2017.

[ABL16] N. Ashraf, O. Bandiera, and S. Lee. "Do-gooders and Go-getters: Career incentives, Selection, and Performance in Public Service Delivery." Working paper, STICERD, 2016.

[AGR15] D. Acemoglu, C. Garcia-Jimeno, and J. A. Robinson. "State Capacity and Economic Development: A Network Approach." *American Economic Review*, **105**:2364–2409, 2015.

[AJR01] D. Acemoglu, S. Johnson, and J. A. Robinson. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, **91**:1369–1401, 2001.

[AT97] P. Aghion and J. Tirole. "Formal and Real Authority in Organizations." *Journal of Political Economy*, **105**:1–29, 1997.

[Ban08] M. Banton. "Administering the Empire." pp. 1801–1968, 2008.

[BBR09] O. Bandiera, I. Barankay, and I. Rasul. "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica*, **77**(1047-1094):417–458, 2009.

[BCD12] A. Banerjee, R. Chattopadhyay, E. Duflo, D. Keniston, and N. Singh. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training." Technical report, NBER Working Papers 17912, NBER, 2012.

[BEF17] Dal Bo, F. Finan E., O. Folke, T. Persson, and J. Rickne. "Who becomes a politician?" *e Quarterly Journal of Economics*, 2017. Forthcoming.

[Ber09] M. Bertrand. "CEOs." *Annual Review of Economics*, **1**:121–150, 2009.

[BER13] Dal Bo, F. Finan E., and M. A. Rossi. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service." *The Quarterly Journal of Economics*, **128**:1169–1218, 2013.

[BP09] T. Besley and T. Persson. "The Origins of State Capacity: Property Rights, Taxation, and Politics." *American Economic Review*, **99**(1218-44):1–34, 2009.

[Bre16] E. Breza. *Field Experiments, Social Networks, and Development.* Handbook on the Economics of Networks, Oxford, 2016.

[BS03] M. Bertrand and A. Schoar. "Managing with Style: The Effect of Managers on Firm Policies." *The Quarterly Journal of Economics*, **118**:1169–1208, 2003.

[CH08] V. Chernozhukov and C. Hansen. "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters*, **100**:68–71, 2008.

[Des17] E. Deserranno. “Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda.” mimeo, 2017.

[DH86] L. Davis and R. Huttenback. *Mammon and the Pursuit of Empire: The Political Economy of British Imperialism.* 1860-1912, Cambridge University Press, 1986.

[DJT99] M. Dewatripont, I. Jewitt, and J. Tirole. “The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies.” *Review of Economic Studies*, **66**:199–217, 1999.

[DP12] M. Dincecco and M. Prado. “Warfare, Fiscal capacity, and Performance.” *Journal of Economic Growth*, **17**:171–203, 2012.

[Fis01] R. Fisman. “Estimating the Value of Political Connections.” *American Economic Review*, **91**:1095–1102, 2001.

[FOP15] F. Finan, B. A. Olken, and R. Pande. “The Personnel Economics of the State.” Working Paper 21825, NBER, 2015.

[Fra92] M. Francis. “Governors and Settlers: Images of Authority in the British Colonies.” pp. 1820–60, 1992.

[FW04] R. Fisman and S. j. Wei. “Tax Rates and Tax Evasion: Evidence from "Missing Imports" in China.” *Journal of Political Economy*, **112**:471–500, 2004.

[Gar12] L. Gardner. *Taxing Colonial Africa: The Political Economy of British Imperialism.* Oxford University Press, 2012.

[GR12] R. Gibbons and J. Roberts. “The Handbook of Organizational Economics.” In *Princeton University Press.* 1 ed, 2012.

[HHM12] B. Henn, L. Hon, M. Macpherson, N. Eriksson, S. Saxonov, I. Peter, and J. Mountain. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples.” *PLoS One*, **7**, 2012.

[IM12] L. Iyer and A. Mani. “Traveling Agents: Political Change and Bureaucratic Turnover in India.” *The Review of Economics and Statistics*, **94**:723–739, 2012.

[Jef38] C. Jeffries. *The Colonial Empire and its Civil Service.* Cambridge University Press, 1938.

[Jia17] R. Jia. “Pollution for Promotion.” mimeo, 2017.

[JKS15] R. Jia, M. Kudamatsu, and D. Seim. “Political Selection in China: The Complementary Roles of Connections and Performance.” *Journal of the European Economic Association*, **13**:631–668, 2015.

[JO05] B. F. Jones and B. A. Olken. “Do Leaders Matter? National Leadership and Growth Since World War II.” *The Quarterly Journal of Economics*, **120**:835–864, 2005.

[Kir00] A. Kirk-Greene. “Britain’s Imperial Administrators.” pp. 1858–1966, 2000.

[KKO15] A. Q. Khan, A. I. Khwaja, and B. A. Olken. *Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors.* The Quarterly Journal of Economics, 2015.

[KKO16] A. Khan, A. Khwaja, and B. Olken. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." mimeo, 2016.

[KT13] F. Kramarz and D. Thesmar. "Social Networks in the Boardroom." *Journal of the European Economic Association*, **11**:780–807, 2013.

[Lai05] Z. Laidlaw. *Colonial Connections 1815-1845: Patronage.* the Information Revolution and Colonial Government, Studies in Imperialism, Manchester University Press, 2005.

[LO12] E. P. Lazear and P. Oyer. "Personnel Economics." In *The Handbook of Organizational Economics.* Princeton University Press, 2012.

[MC99] S. Morris and S. Coate. "Policy Persistence." *American Economic Review*, **89**:1327–1336, 1999.

[NT54] S. Northcote and C. Trevelyan. *Report on the Organization of the Permanent Civil Service.* HMS, 1854.

[NW89] D. North and B. Weingast. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice." *The Journal of Economic History*, **49**:803–832, 1989.

[PT96] C. Prendergast and R. H. Topel. "Favoritism in Organizations." *Journal of Political Economy*, **104**:958–78, 1996.

[PZ16] P. Persson and E. Zhuravskaya. "The Limits of Career Concerns in Federalism: Evidence from China." *Journal of the European Economic Association*, **14**:338–374, 2016.

[RAJ05] J. A. Robinson, D. Acemoglu, and S. Johnson. "Institutions as a Fundamental Cause of Long-Run Growth." *Handbook of Economic Growth*, **1**:386–472, 2005.

[RR17] I. Rasul and D. Rogger. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal*, pp. 1–34, 2017.

[SW16] E. Sanderson and F. Windmeijer. "A weak instrument F-test in linear IV models with multiple endogenous variables." *Journal of Econometrics*, **190**:212–221, 2016.