Summer 2019

Essays on household income and expenditures

Liqiong Chen
University of Iowa

Follow this and additional works at: https://ir.uiowa.edu/etd
Part of the Economics Commons

ESSAYS ON HOUSEHOLD INCOME AND EXPENDITURES

by

Liqiong Chen

A thesis submitted in partial fulfillment
of the requirements for the
Doctor of Philosophy
degree in Economics in the
Graduate College of
The University of Iowa

August 2019

Thesis Supervisor: Associate Professor Suyong Song

ACKNOWLEDGEMENTS

I would first and foremost like to thank my advisor, Suyong Song for his guidance and support. His passion about research and advice made the completion of my dissertation possible. I would also like to thank my committee members: Linnea Polgreen, Julia Garlick, Anne Villamil and David Frisvold.

I am incredibly grateful to my dad, who is the most important person in my life, for his great love and encouraging me to explore myself. I am also indebted to my mom, my uncle Hongyuan Gu, and my grandparents for constant support throughout my graduate school. I am also thankful for my love Jonathan for his unfailing support and encouragement.

What's more, I feel great gratitude for my college Nankai University. That's the greatest experience I have ever had in my life.

ABSTRACT

This dissertation studies household income and consumption. In the first chapter, I identify the causal effect of retirement on health service utilization in China. In the second chapter, I investigates the impact that retirement has on the family support network of "sandwich" generations in China. In the third chapter, I propose a new estimator for linear quantile regression models with generated regressors, and apply it to study Engel curves for various commodity consumption for families in the UK.

In the first chapter, I apply a regression discontinuity design by exploiting the exogenous mandatory retirement age rules in China in order to identify the causal effect of retirement on health service utilization. In China, the social insurance Urban Employee Basic Medical Insurance (UEBMI) provision continues after individuals retire. Employees, however, stop paying the premium and enjoy reduced cost sharing after they retire. Individual medical expenses, insurance costs, and benefits are recorded in the China Household Finance Survey 2013 (CHFS). Significantly, males and females respond differently to this decrease in the relative price of health insurance at the time of retirement. Females are generally more willing to increase their out-of-pocket expenditures in order to take advantage of better health insurance benefits and utilize more medical care. Males, by contrast, do not respond to this change in relative price in the same manner.

In the second chapter, I investigates the impact that retirement has on the family support networks of "sandwich" generations in China. These middle-aged households have an inter-generational support network that includes both upward transfers (their parents or parents-in-law), as well as downward transfers (their children). I use micro data from CHARLS (China Health and Retirement Longitudinal Study) concerning middle-aged and elderly households in order to evaluate the changes that retirement can have on this family support network, primarily by exploiting the exogenous mandatory retirement age rules in China. I make the identifying assumption that inter-generational transfers would evolve more smoothly if households would not retire and apply a regression discontinuity approach. I find that retirement induces "sandwich" generations to switch roles in the private network as well as in the public transfer channel; indeed, is 55 percent-

age point more likely that households will switch from resource providers to resource recipients in the channel of private transfers. In addition, these “sandwich” generations are about 47 percentage point more likely to receive money from their non-coresident children when they retire.

In the third chapter, we studies estimation and inference for linear quantile regression models with generated regressors. We suggest a practical two-step estimation procedure, where the generated regressors are computed in the first step. The asymptotic properties of the two-step estimator, namely, consistency and asymptotic normality are established. We show that the asymptotic variance-covariance matrix needs to be adjusted to account for the first-step estimation error. We propose a general estimator for the asymptotic variance-covariance, establish its consistency, and develop testing procedures for linear hypotheses in these models. Monte Carlo simulations to evaluate the finite-sample performance of the estimation and inference procedures are provided. Finally, we apply the proposed methods to study Engel curves for various commodities using data from the UK Family Expenditure Survey. We document strong heterogeneity in the estimated Engel curves along the conditional distribution of the budget share of each commodity. The empirical application also emphasizes that correctly estimating confidence intervals for the estimated Engel curves by the proposed estimator is of importance for inference.

PUBLIC ABSTRACT

This dissertation studies household income and consumption. In the first chapter, I identify the causal effect of retirement on health service utilization in China. In the second chapter, I investigates the impact that retirement has on the family support network of "sandwich" generations in China. In the third chapter, I propose a new estimator for linear quantile regression models with generated regressors, and apply it to study Engel curves for various commodity consumption for families in the UK.

In the first chapter, I apply a regression discontinuity design by exploiting the exogenous mandatory retirement age rules in China in order to identify the causal effect of retirement on health service utilization. In China, the social insurance Urban Employee Basic Medical Insurance (UEBMI) provision continues after individuals retire. Employees, however, stop paying the premium and enjoy reduced cost sharing after they retire. Individual medical expenses, insurance costs, and benefits are recorded in the China Household Finance Survey 2013 (CHFS). Significantly, males and females respond differently to this decrease in the relative price of health insurance at the time of retirement. Females are generally more willing to increase their out-of-pocket expenditures in order to take advantage of better health insurance benefits and utilize more medical care. Males, by contrast, do not respond to this change in relative price in the same manner.

In the second chapter, I investigates the impact that retirement has on the family support networks of "sandwich" generations in China. These middle-aged households have an inter-generational support network that includes both upward transfers (their parents or parents-in-law), as well as downward transfers (their children). I use micro data from CHARLS (China Health and Retirement Longitudinal Study) concerning middle-aged and elderly households in order to evaluate the changes that retirement can have on this family support network, primarily by exploiting the exogenous mandatory retirement age rules in China. I make the identifying assumption that inter-generational transfers would evolve more smoothly if households would not retire and apply a regression discontinuity approach. I find that retirement induces "sandwich" generations to switch roles in the private network as well as in the public transfer channel; indeed, is 55 percent-

age point more likely that households will switch from resource providers to resource recipients in the channel of private transfers. In addition, these "sandwich" generations are about 47 percentage point more likely to receive money from their non-coresident children when they retire.

In the third chapter, we studies estimation and inference for linear quantile regression models with generated regressors. We suggest a practical two-step estimation procedure, where the generated regressors are computed in the first step. The asymptotic properties of the two-step estimator, namely, consistency and asymptotic normality are established. We show that the asymptotic variance-covariance matrix needs to be adjusted to account for the first-step estimation error. We propose a general estimator for the asymptotic variance-covariance, establish its consistency, and develop testing procedures for linear hypotheses in these models. Monte Carlo simulations to evaluate the finite-sample performance of the estimation and inference procedures are provided. Finally, we apply the proposed methods to study Engel curves for various commodities using data from the UK Family Expenditure Survey. We document strong heterogeneity in the estimated Engel curves along the conditional distribution of the budget share of each commodity. The empirical application also emphasizes that correctly estimating confidence intervals for the estimated Engel curves by the proposed estimator is of importance for inference.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: GENDER DIFFERENCES IN HEALTH-CARE UTILIZATION: EVIDENCE FROM URBAN EMPLOYEE BASIC MEDICAL INSURANCE (UEBMI) IN CHINA

Introduction

Due to the growing elderly population, policymakers and researchers have given more attention to the effects of retirement on medical expenditures. As the risk of catastrophic medical expenditures and the decision to retire are themselves intertwined, (as pointed out in Blau and Gilleskie (2006)) it is difficult to estimate the causal impacts of each.

In the US, medical expenditures increase dramatically once the insured are covered under Medicare. The fact that underlying health issues may drive both retirement and health insurance decisions is a concern for both researchers and policy makers. The elderly may choose to retire to ensure that they are eligible for Medicare, as it has better health insurance benefits. Growing concerns about health service over-utilization behavior under more generous health insurance have sparked a great deal of literature about the utilization of health services related to health insurance policies Card, Dobkin, and Maestas (2008). Failing to disentangle the relationship between retirement decisions and health insurance decisions makes it difficult for researchers to determine if people are over utilizing their health services.

This paper uses a credible identification method for the purpose of identifying the causal effect of retirement on health-care utilization in China. I apply a regression discontinuity design by exploiting China's exogenous mandatory retirement age rules in order to identify the causal effect of retirement on health service utilization. I use micro data from the China Household Finance Survey 2013 (CHFS), which includes information about individuals' health-insurance premia, overall medical expenses, and those medical expenses covered by health insurance.

There are several compelling aspects to this particular study. First, the "retirement-consumption puzzle" literature challenges the prediction of those life-cycle models where consumers smooth their consumption over their life cycle. Several empirical studies have found that non-durable expenditures decrease after people retire. However, medical expenses, as one of the major expenses in people's lives, are rarely studied due to its endogenous impact on retirement decisions. Medical expenses, indeed, greatly affect how individuals make decisions about their consumption and savings (De Nardi, French, and Jones (2010)). For example, people may save their finances in preparation for catastrophic medical expenses in the future. With this relationship between retirement and health expenditures in mind, this paper will attempt to estimate how this relationship impacts the context of China specifically. Though it is unclear how retirement might change health-service utilization, the overall loss of income and potential future income growth that are associated with retirement may indeed lead to lower health service utilization. However, this effect may be offset by other aspects of retirement, including more available free time and better health insurance benefits, leading to overall higher medical utilization.

Second, the exogenous change in the costs and benefits of the Urban Employee Basic Medical Insurance (UEBMI) at retirement as a natural experiment allows me to both identify the change in the relative price of health insurance and investigate their effects on health-seeking behavior. Employees are covered by the same health insurance before and after retirement, however the insured cease paying health insurance premia, enjoying reduced deductibles and lower coinsurance rates after they retire. The discrete change in health insurance costs and benefits incentivizes individuals to utilize health-care differently. As such, this paper contributes to the literature centered on the effect of changes in cost sharing of health insurance (intensive margin) on the utilization of health services; by contrast, the majority of existing studies focus on the effect of health-insurance provision (extensive margin) on the demand for health care. There is a great deal of literature already available (for instance, Card, Dobkin, and Maestas (2008)) that studies the impact of provisions of Medicare.

In addition, it should be noted that an MSA approach (Medical savings accounts) under the UEBMI elevates the role of the individual choice. An MSA account has disposable income for medical expenses, giving individuals full control over how resources are allocated to various health-care services. Also, individuals have the opportunity to determine when to utilize health-care. Therefore, MSAs allow employees to allocate medical expenses before and after retirement though their accounts.

The main contributions of this paper are as follows: First, the paper estimates the causal effects of retirement on medical expenses; Second, studying the gender differences in health-seeking behavior from an economic perspective is both enlightening and essential. Rather than referring to the strand of literature that studies gender differences in various medical settings, I specifically study such gender differences in response to exogenous change in health insurance benefits at retirement. Examining the patterns of total medical expenditures and medical expenditures covered by the health insurance for males and females respectively, I find that females, upon retirement, allow a discrete increase in both total medical expenditures and those expenditures covered by the health insurance. By contrast, the medical consumption pattern of males do not change upon retirement. Therefore, females delay certain medical expenses and spend more on health-care after retirement. In addition, females are generally more willing to increase their out-of-pocket expenditures in order to both take advantage of better health insurance benefits and utilize more medical care.

Literature Review

The key prediction of the life-cycle models is that consumers make intertemporal plans to smooth their standard of living over the life cycle. However, recent empirical evidence shows that drops in consumption at retirement contradict the consumption-smoothing theory; this phenomenon is known as the “retirement consumption puzzle”(Battistin, Brugiavini, Rettore, and Weber (2009)). Possible reasons for this decrease in consumption include unexpectedly low pensions, increased non-market time, change in preferences, and others.

A majority of the research on the retirement-consumption puzzle focuses on non-durable expenditures, which fails to include expenditures on medical care(Li, Shi, and Wu (2015)). The impact that retirement can have on medical expenses is rarely studied because retirement is an endogenous decision variable. Additionally, it should be noted that concerns about medical expenses could affect an individual’s retirement decision.

For example, French and Jones (2011) provides an empirical analysis of the effects of employer-provided health insurance, Medicare, and Social Security on retirement behavior. Medicare, they found, has a profound impact on retirement behavior. For instance, prior to receiving Medicare at age 65, many individuals receive health insurance only if they continue to work. This incentive to work disappears at age 65, when Medicare provides health insurance to almost everyone. However, retiring before 65 exposes workers who lack retiree health insurance coverage to possible catastrophic medical expenditures. The Social Security system and pensions also provide retirement incentives at age 65. With these factors in mind, it is difficult to determine whether the high job exit rates that are observed at age 65 are due to Medicare, Social Security, pensions or a combination of the three .

Significantly, China has a mandatory retirement age. The exogenous rule forces individuals to exit the labor market immediately upon retirement. Also, health insurance provisions are the same for workers before and after retirement; indeed, there is no switch in health insurance at the extensive margin. This paper, therefore, provides insight into the impact of medical expenditures on relative costs and benefits in China.

The majority of the literature focuses on the effects of changes in health-insurance provisions on health-care utilization at the extensive margin. This paper, however, contributes to the limited literature that focuses on the changes in cost sharing on health insurance at the intensive margin. It should be noted that Baicker and Goldman (2011) offers a review of the difficult task of identifying these effects, due to unobserved characteristics in the presence of self-selection.

For example, Aron-Dine, Einav, and Finkelstein (2013) offer evidence on the effects of cost sharing on health utilization, based on the RAND Health Insurance Experiment (Manning, Newhouse, Duan, Keeler, and Leibowitz (1987)). Chandra, Gruber, and McKnight (2010) also studies a policy change that raised co-payments on the supplemental insurance for retired public employees in California, though the increased cost sharing was restricted to office visits and prescription drugs. Finally, Shigeoka (2014) analyzes a policy in Japan that reduced the cost sharing of patients older than 70 years, and also studies how such cost sharing affects health outcomes and health utilization.

These studies offer some evidence to the effects of cost sharing on health utilization. This paper, however, suggests that a quasi-experiment of change in cost sharing occurs for over the course of an individual's life cycle. Determining how these retiring individuals respond to such changes in cost sharing will provide information that the literature currently needs.

It is worth noting that females and males not only differ in mandatory retirement ages, but they also exhibit different health-seeking behaviors. Mustard, Kaufert, Kozyrskyj, and Mayer (1998) provides evidence that the gender difference in the use of health-care can be substantial at various stages of one's life cycle. For example, females tend to spend more when they are pregnant or giving birth, while males consume more during the period right before they die. Hawkes and Buse (2013)also studies gender as an important influence on health behaviors. This paper provides new insight into the gender differences that exist in health utilization, specifically from an economic perspective.

Health Insurance System In Urban China

Urban residents in China are typically covered under social insurance. The working population, however, is usually covered by the Urban Employee Basic Medical Insurance (UEBMI), which was introduced in order to provide health-care access to urban, working, and retired employees in both public and private sectors. Funds for the UEBMI occupy 8% of the employee's wage (6% are paid by employers and 2% by employee contribution), though these rates vary according to time and municipalities. Unlike other types of insurance schemes, UEBMI is *mandatory*.

The Urban Residents Basic Medical Insurance (URBMI) provides health-care access to urban residents who are not covered by the UEBMI; this includes children, students in school, and other non-working urban residents. These individuals must continue to pay the premium in order to be covered by the health insurance.

Retired employees are still covered by the UEBMI, but the relative costs and benefits change after they retire. Those who have contributed to the UEBMI for a certain number of years beyond the stipulated threshold can stop paying the health insurance premium, enjoying a lower coinsurance rate after they retire.

By contrast, the American health insurance system

- stipulates that non-elderly individuals be offered private health insurance with premia co-payments and deductibles.
- Children of the poor are covered by Medicaid.
- Elderly are covered by Medicare.

Medicare consists of four parts A, B, C and D. Individuals are eligible for premium-free Part A (Hospital Insurance) if they are 65 or older, and if they or their spouses worked and paid Medicare taxes for at least 10 years. These individuals are eligible for Part A at age 65 without having to pay premiums. While most people do not have to pay a premium to be eligible for Part A, all individuals are expected to pay for other parts; for example, it is necessary that an individual pay to be eligible for Part B (Medicare Insurance) should they want it.

Unlike from the American health insurance system, the Chinese health insurance coverage rate does not shift at the point of retirement, though the relative costs and benefits for those insured employees do.

Urban Employee Basic Medical Insurance (UEBMI)

The Urban Employee Basic Medical Insurance (UEBMI) premiums stem from both employers and employees. The contributions of employers are divided into two accounts: 70% goes into a social pooling account (SPA), and 30% is deposited into individual medical savings accounts (MSAs). The funds paid by employees are deposited into their MSAs; each year, a small percentage (2%, for instance) of MSA deposits go to current-year funds in the individual's MSA. The current-year funds are considered disposable income for medical expenses.

Medical expenses are financed according to three tiers: MSAs, out-of-pocket spending in the form of deductibles, and social-risk pooling. MSAs incentivize consumers to be more cost-aware in their demands for health services, deductibles increase cost sharing among patients, and social-risk pooling aims to protect employees from catastrophic expenses.

Employees pay for all of their health-care expenses until the current-year funds in their individual accounts (first tier) have been spent. Whatever funds remain unspent are carried to the next year, going into the past-years funds in MSA; funds that are unspent at the end of a person's life are inheritable. When current-year funds in the individual accounts are exhausted, employees must pay a deductible (second tier) out-of-pocket up to a flat amount. The health care expenses that exceed the deductible are then paid from the social-risk-pool fund (third tier), with patients paying a rate of coinsurance. Such a risk-pool fund limits workers' financial loss.

Medical Savings Accounts (MSA)

Each city in China establishes its Bureau of Social Insurance, which can serve as the group purchaser of services, in order to contract prices and quality of services with health-care providers. MSA funds are deposited into an interest-earning account in the Industrial and Commercial Bank of China.

An MSA approach emphasizes individual Choice, with MSAs themselves providing incentives for consumers to be more cost-aware in their demands for health services. MSAs are personal funds established by individuals or their employers to pay current out-of-pocket medical costs (as current-year funds) and to accumulate funds for future expenses (as past-years funds). MSAs give the individual full control over how their resources are allocated for health-care services. Since most individuals would view health care expenses with respect to other desired consumption goods, it stands to reason that such individuals would be more careful when deciding how and when to use such services.

Past-years funds in an MSA account are funds that the individuals can draw from to pay any deductibles and medical bills that are not covered by health insurance. In addition, unspent current year funds in an employee's MSA are deposited into past-years funds, earning the same interest rate as a 1 year certificate of deposit. In these situations, individuals are responsible for their own health service utilization, and as such there is clear financial incentive to use less care.

Changes In Health Insurance At Retirement

The mandatory health insurance UEBMI does not shift when workers enter retirement; however, it should be noted that the cost of health insurance drops distinctly after they retire. Workers who have paid the premia of their health insurance for at least 15 years cease having to do so after they retire. As shown in Table 1, retirees in Shanghai, for example, also pay a lower deductible.

	before retirement	after retirement
premium	2% of wage	0 RMB
outpatient deductible	1500 RMB	700 RMB
inpatient deductible	1500 RMB	1200 RMB

Table 1: The Premium and Deductibles for UEBMI Health Insurance Before and After Retirement

Insurance policies vary in coinsurance rate depending on the specific tiers that the hospital where the service was provided utilizes. Coinsurance incentivizes patients to seek health care at a lower tier hospital, unless it is necessary to move up to a higher tier. Table 2 shows the comparison of coinsurance rates pre- and post- retirement. Retirees pay about a 5% lower coinsurance rate for outpatient service, and a 7% lower rate for inpatient service.

	before retirement	after retirement
outpatient coinsurance rate		
tier I hospitals	40%	30%
tier II hospitals	30%	25%
tier III hospitals	25%	20%
inpatient coinsurance rate	15%	8%

Table 2: Coinsurance Rate for Health Insurance Before and After Retirement

Identification

According to the notation of the potential outcome framework, (Y_0, Y_1) are the two potential outcomes a household may experience. Y_0 is the measure of interest when an individual is not retired, and Y_1 is the measure of interest when one is retired. In this case, $D = 1$ represents an individual who is retired, and $D = 0$ indicates that an individual is not retired. Therefore, an actual measure of medical expenditure Y can be represented by

$$Y = Y_0(1-D) + Y_1 D = Y_0 + (Y_1 - Y_0)D$$

In this paper, there are three outcome variables of interest. For i-th individual in the sample, Y_{1i} is defined as total medical expenditures. Y_{2i} is defined as medical expenditures that are covered by health insurance. Y_{3i} is defined as out-of-pocket medical expenditures. D_i is a binary variable indicating if an individual is retired or not. A discontinuity design (Thistlethwaite and Campbell (1960)) arises when D_i depends on an observable variable S_i, where S_i measures the age of an individual, and, in support of S_i, the probability of being retired changes discontinuously at the threshold $\bar{s}$. In other words,

$$Pr\{D_i = 1|\bar{s}^+\} \neq Pr\{D_i = 1|\bar{s}^-\}$$

where $\bar{s}^+$ and $\bar{s}^-$ refer to individuals who are marginally above and below $\bar{s}$, respectively.

According to Trochim and Thochim (1984), the distinction between sharp and fuzzy regression discontinuity designs depends on the size of the probability jump at the threshold. A sharp regression discontinuity design occurs if the retirement status deterministically depends on whether an individual's age is above $\bar{s}$; in other words, $D_i = 1(S_i \geqslant \bar{s})$. A fuzzy design is applicable if the size of discontinuity at $\bar{s}$ is smaller than one.

Urban residents in China follow mandatory retirement policies, which are themselves primarily based on age requirements. The normal retirement age for females are either 50 or 55 as a cadre, while for males the normal retirement ages are 55 or 60. In addition, a policy that was established in 1994 during the SOE (state-owned enterprises) reform in the 1990s states that a female can retire at age 45, 50 or 55 as a cadre, while a male can retire at age 50, 55 or 60.

In this paper, I focus on the peak age of retirement for females (age 50) and males (age 60); other retirement-age thresholds are not considered. A fuzzy design is used in this paper because the size of the discontinuity at age thresholds (again, age 50 for female, age 60 for male) is smaller than one, as reaching the age threshold does not necessarily imply that individuals are actually retired. Though the mandatory rule induces a higher probability of being retired, some individuals may choose not to comply with the retirement rule.

The local average treatment effects of retirement on medical expenses can be recovered for retired individuals around $\bar{s}$.

Therefore, three respective treatment effects δ_1, δ_2 and δ_3 can be defined as follows:

$$\delta_1 = \frac{\lim_{S\to\bar{s}^+} E[Y_{1i}|S] - \lim_{S\to\bar{s}^-} E[Y_{1i}|S]}{\lim_{S\to\bar{s}^+} E[D_i|S] - \lim_{S\to\bar{s}^-} E[D_i|S]}$$

$$\delta_2 = \frac{\lim_{S\to\bar{s}^+} E[Y_{2i}|S] - \lim_{S\to\bar{s}^-} E[Y_{2i}|S]}{\lim_{S\to\bar{s}^+} E[D_i|S] - \lim_{S\to\bar{s}^-} E[D_i|S]}$$

$$\delta_3 = \frac{\lim_{S\to\bar{s}^+} E[Y_{3i}|S] - \lim_{S\to\bar{s}^-} E[Y_{3i}|S]}{\lim_{S\to\bar{s}^+} E[D_i|S] - \lim_{S\to\bar{s}^-} E[D_i|S]}$$

Under the following conditions:

Condition 1: $E[Y_{ki}(0)|S=s]$ and $E[Y_{ki}(1)|S=s]$ are continuous functions of S for $k=1,2,$ 3.

Condition 2: $D_i(S)$ is increasing in S at $S=\bar{s}$.

Data And Descriptive Statistics

The 2013 China Household Finance Survey (CHFS) survey covers 262 counties in 29 provinces across China, including roughly 28,000 households and 98,000 individuals. The survey collected detailed information, including demographics and family background, household income and wealth, assets and liabilities, expenditures, social and commercial insurance, and many other types of firsthand raw data that are not represented on a national scale. All data were collected using face-to-face, computer-assisted telephone interviews (CATI).

Availability of insurance enrollment, premium payment, and medical expenses provides information about the relative price of health insurance, as well as the use of medical expenses that are covered by the health insurance. Table 3 reports the medical expenses of females age 40 to 60, while Table 4 shows the medical expenses of males age 50 to 70. About 34% of females and 43%, of males were retired.

Demographic information includes marital status and education level. Marital status is divided into two distinct groups: Married and Single (the latter of which includes divorced, widowed, separated, or never married).

Education attainment is classified into either middle-school degree and above, or lower education. On average, over 90% of females and males were married, and slightly less than 90% of females and males had achieved middle-school degrees.

	count	mean	std	min	25%	50%	75%	max
age	5370	49.923	6.112	40	45	50.0	55.00	60
total medical expenditure	5370	2537.624	4946.138	0	100	800.0	2500.00	41600
expenditure covered by insurance	3652	1454.390	3105.988	0	0	272.5	1300.00	25800
out-of-pocket expenditure	5370	1450.889	3030.011	0	0	300.0	1506.25	30000
retired	5370	0.340	0.474	0	0	0.0	1.00	1
married	5370	0.921	0.270	0	1	1.0	1.00	1
middle-school graduate	5370	0.890	0.313	0	1	1.0	1.00	1

Table 3: Descriptive Statistics for Females

	count	mean	std	min	25%	50%	75%	max
age	4498	58.721	5.758	50	54	59	63	70
total medical expenditure	4498	3413.251	7182.686	0	0	1000	3000	65000
expenditure covered by insurance	3015	2339.419	5293.213	0	0	500	2000	45000
out-of-pocket expenditure	4498	1774.533	3982.420	0	0	300	2000	47635
retired	4498	0.434	0.496	0	0	0	1	1
married	4498	0.954	0.209	0	1	1	1	1
middle-school graduate	4498	0.869	0.337	0	1	1	1	1

Table 4: Descriptive Statistics for Male

Estimation

My goal is to estimate the impact of retirement on health-seeking behavior. The questions of interest for this paper include: Do individuals utilize more, less, or the same degree of medical care after they retire? Also, are individuals willing to pay more out of pocket after they retire?

The main dependent variables in this paper include total medical expenditures, medical expenditures that are covered by health insurance, and out-of-pocket medical expenditures. Y_{1i} is defined as total medical expenditures, Y_{2i} is defined as medical expenditures that are covered by health insurance, and Y_{3i} is defined as out-of-pocket medical expenditures. D_i is a binary variable indicating whether an individual is retired or not. S_i measures the age of the individual.

The following equations are estimated:

$$Y_{1i} = \alpha_1 + \delta_1 D_i + \beta_1 (S_i - \bar{s}) + \beta_2 (S_i - \bar{s}) 1(S_i - \bar{s} \geqslant 0) + \varepsilon_i$$
$$(\bar{s} - h < S_i < \bar{s} + h) \tag{1.1}$$

$$Y_{2i} = \alpha_2 + \delta_2 D_i + \zeta_1 (S_i - \bar{s}) + \zeta_2 (S_i - \bar{s}) 1(S_i - \bar{s} \geqslant 0) + \varepsilon_i$$
$$(\bar{s} - h < S_i < \bar{s} + h) \tag{1.2}$$

$$Y_{3i} = \alpha_3 + \delta_3 D_i + \gamma_1 (S_i - \bar{s}) + \gamma_3 (S_i - \bar{s}) 1(S_i - \bar{s} \geqslant 0) + \varepsilon_i$$
$$(\bar{s} - h < S_i < \bar{s} + h) \tag{1.3}$$

A fundamental problem for the estimation of these equations (1.1)-(1.3) is that the retirement decision itself D_i is endogenous. Some underlying unobservables might drive either the retirement decision or the medical expenditure variables, or both. However, as the Chinese mandatory retirement rule is in place, the age threshold for retirement (50 for females and 60 for male) provides a credible source of exogenous variations in retirement status. Fig 1 and Fig 2 show the proportion of retirees over the life cycle. The overall retirement rate rises about 20% at age 50 for females, and rises about 28% at age 60 for males.

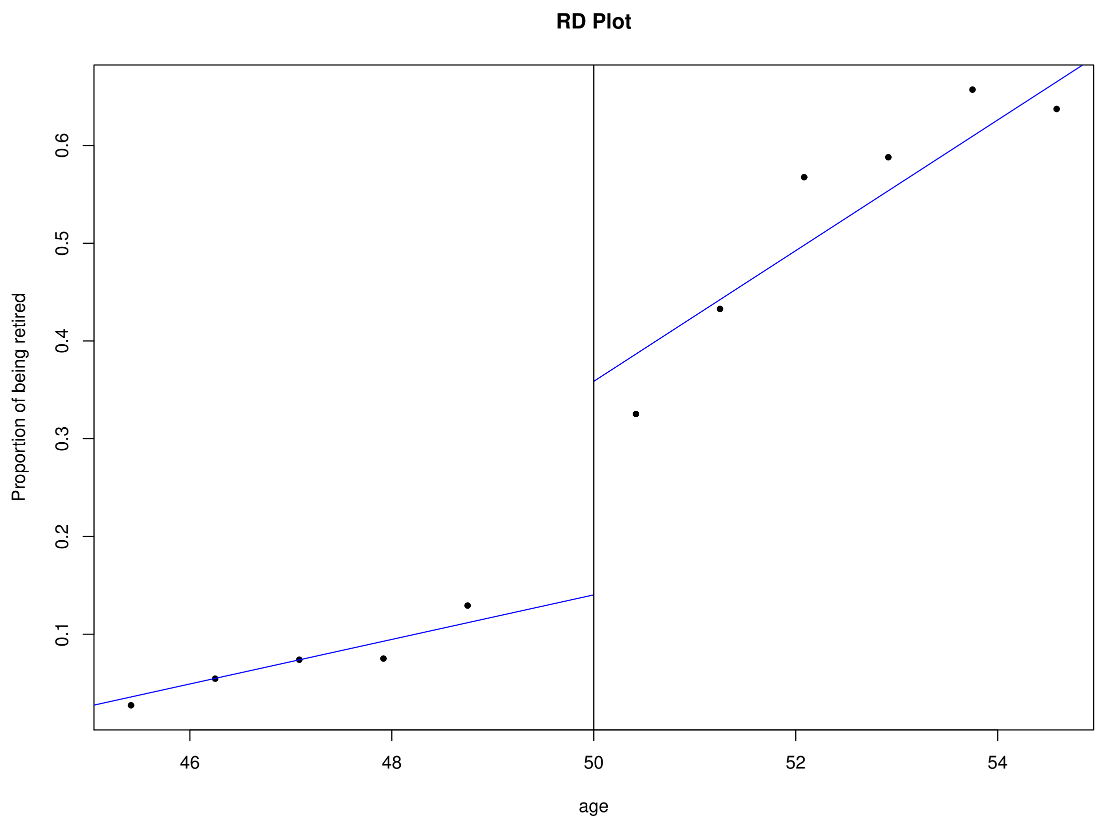


Figure 1: Proportion of Retired Females over Different Ages

The mandatory age rules differ for males and females, and the medical expenditures are measured at the individual level. Therefore, this paper has chosen to apply the estimations to male and female samples separately.

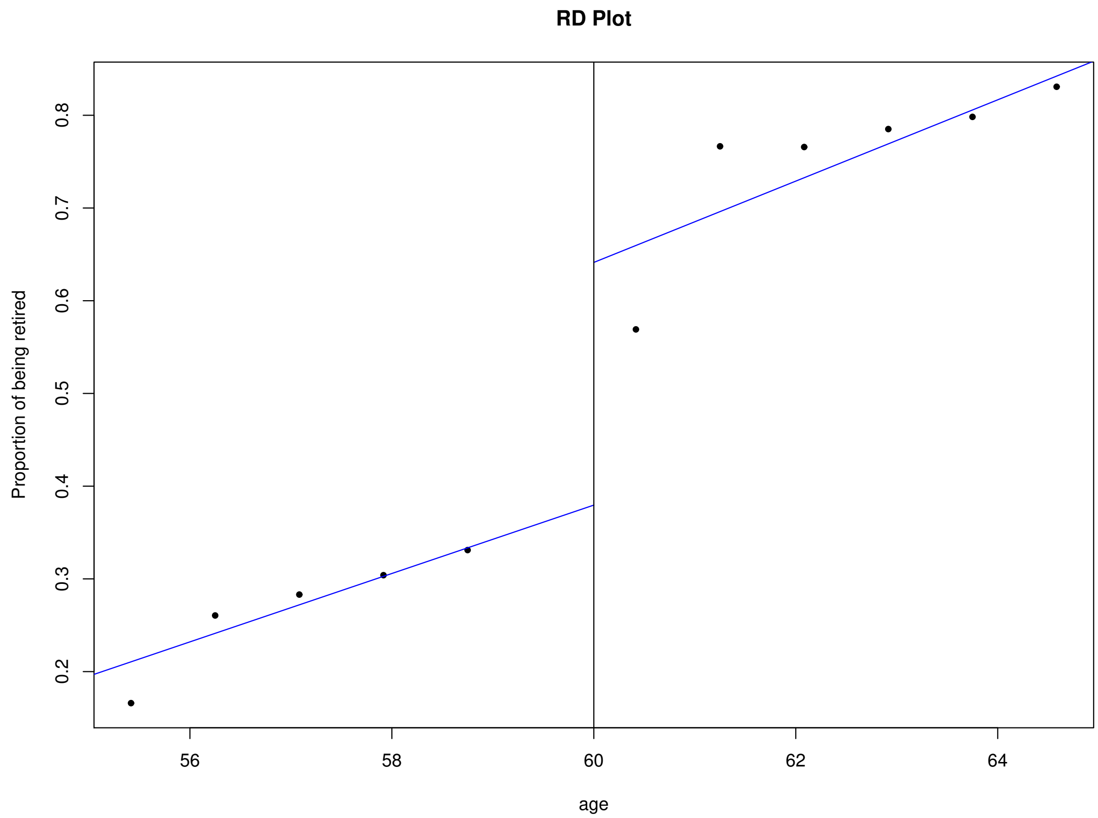


Figure 2: Proportion of Retired Males over Different Ages

Since the OLS estimate using D_i could be subject to selection bias, I have chosen to introduce a second variable $1((S_i \geqslant \bar{s})$ where $\bar{s} = 50$ for women, $\bar{s} = 60$ for men) as an instrumental variable for D_i in order to identify the parameter of interest δ_1,δ_2 and δ_3, the effects of retirement on total medical expenditures, medical expenditures that are covered by health insurance, and out-of-pocket medical expenditures, respectively. A bandwidth of 5 years is applied to both sides of the age threshold for females and males respectively, and uniform kernels are applied in the estimations.

Main Results

Fig 3, which displays the average total medical expenses for females by age, offers two interesting findings. Firstly, there is a discrete jump in average total medical expenses around age 50; indeed, the total medical expenses increase significantly by roughly 900 RMB after age 50.

Secondly, there is a significant difference in average medical expenses between females aged 50 years or older, and younger females. Also, an increase in total medical expenses after age 50 is observed.

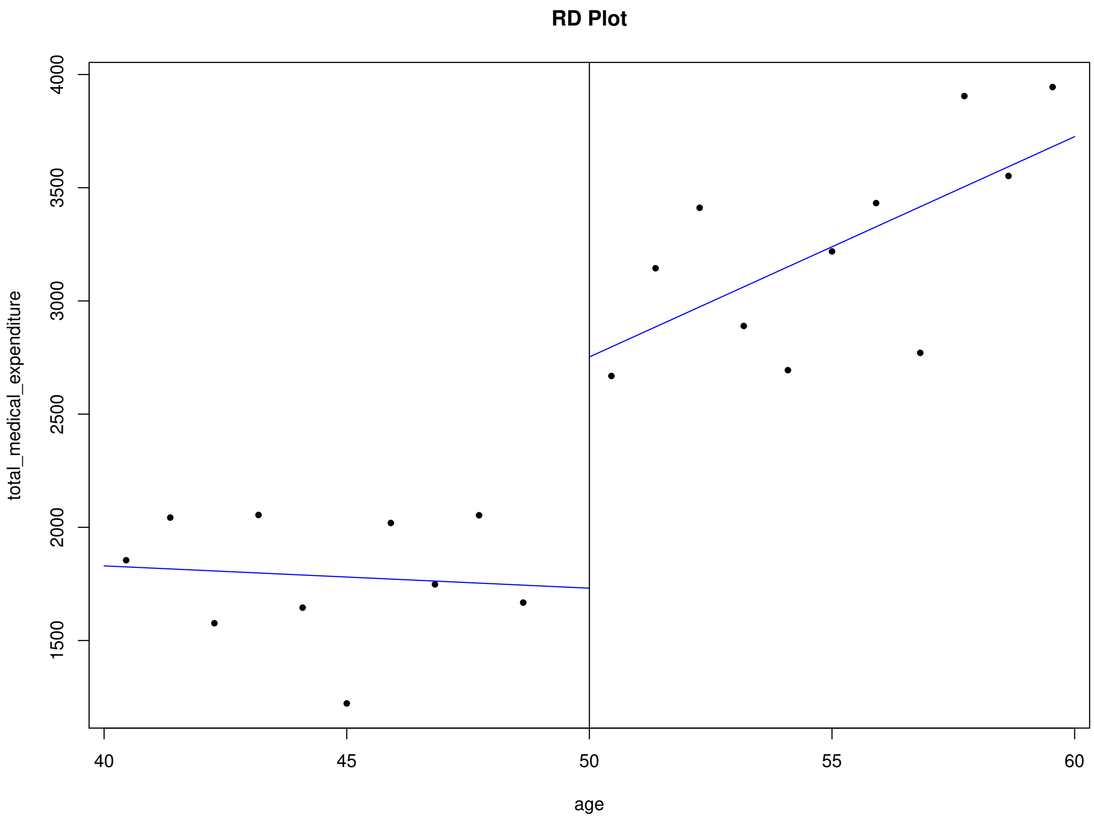


Figure 3: Females' Total Medical Expenditures over Different Ages

Fig 4 shows the average medical expenses that are covered by health insurance for females by age. A similar salient increase in health insurance payment appears around age 50. Is the increase in total medical expenses fully covered by health insurance? Or do females pay more out-of-pocket to utilize their health-care? Fig 5 displays the pattern of average out-of-pocket expenditures by age. A slight decrease in these expenditures appears among females younger than 50. However, a noticeable increase in out-of-pocket expenses occurs in females around age 50. In addition, out-of-pocket expenses begin to increase for females older than 50. These figures superimpose a linear regression, allowing for both a discontinuity at the threshold and linear trends in the running variable (age) on both sides of the threshold.

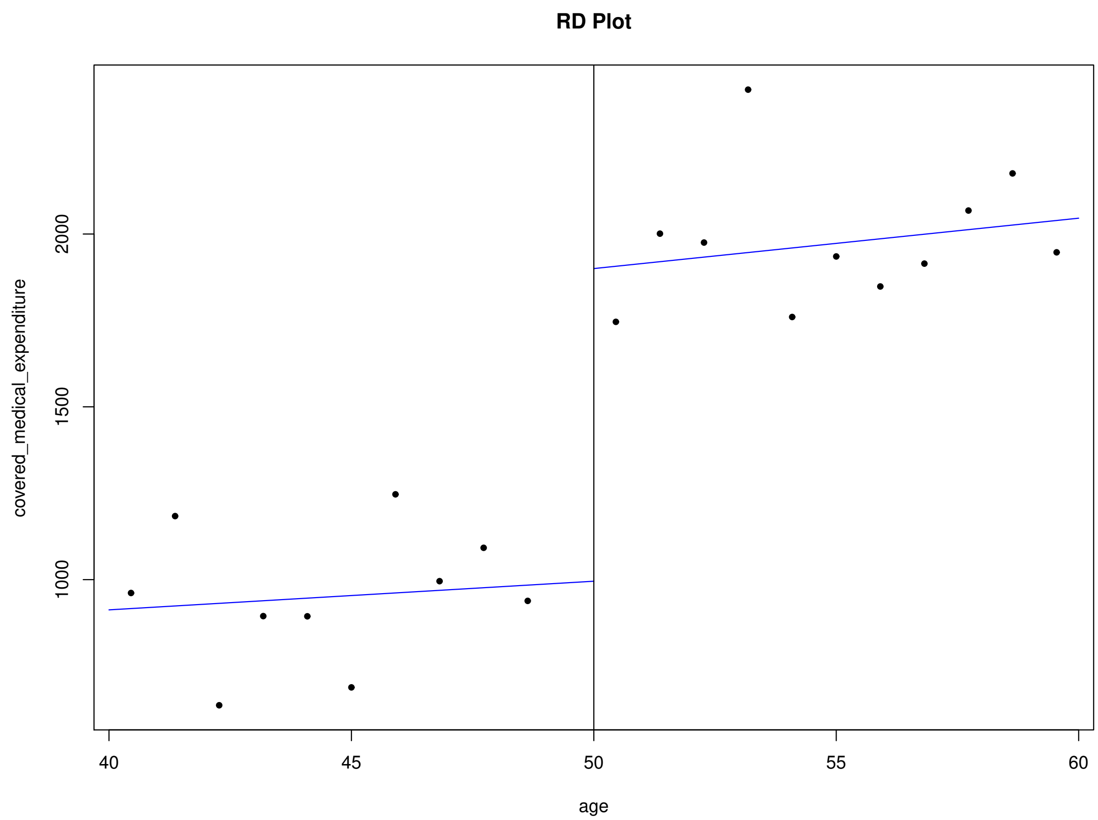


Figure 4: Females' Medical Expenditures Covered by Health Insurance over Different Ages

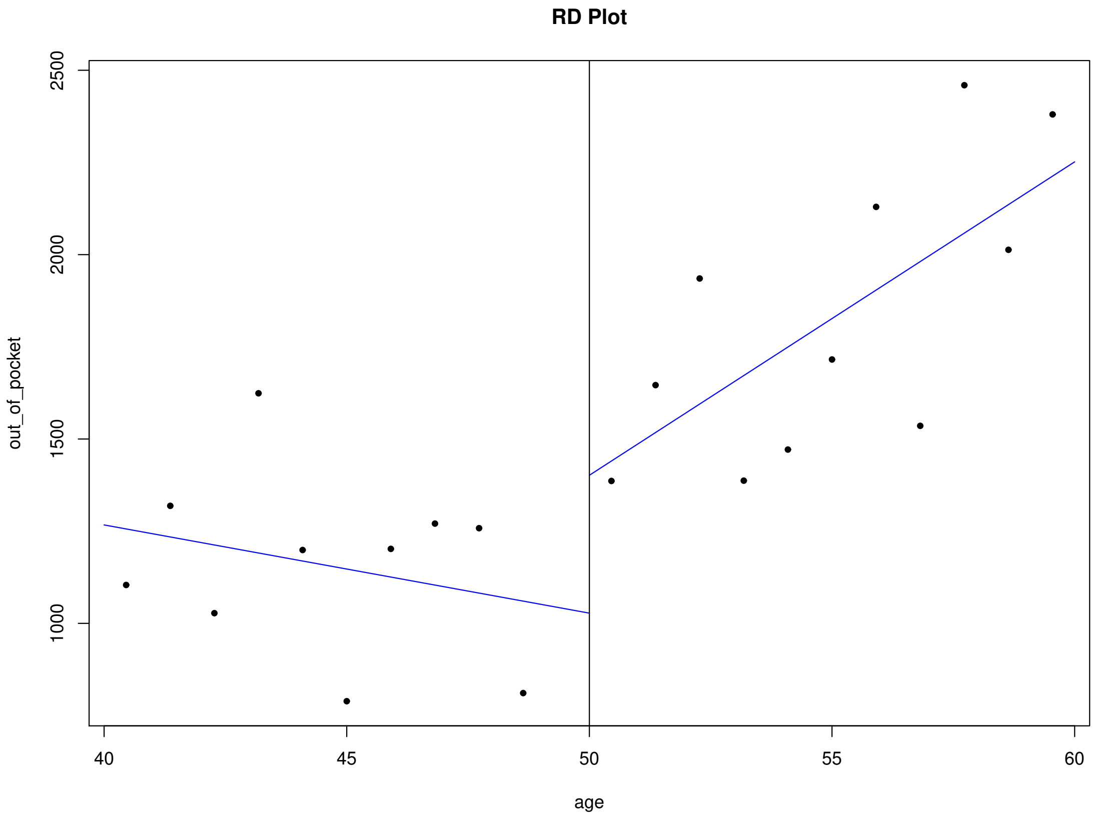


Figure 5: Females' Out-of-pocket Medical Expenditures over Different Ages

Fig 6 through Fig 8 shows the average medical expenses and out-of-pocket expenditures that are covered by health insurance for males by age. In contrast to females, the figures reveal a smooth, consistent pattern as age increases. Though the three medical expenses of interest each increase as the males age, no distinct changes are observed around age 60.

To show the econometric estimates of the effects that appear in these figures, Table 5 (for female) and Table 6 (for male) present the results of the three key measures of medical expenditures, including: total medical expenditures, medical expenditures that are covered by health insurance, and out-of-pocket expenditures. A deeper inspection of the numerical estimates of medical expenses between males and females reveal a significant difference in health service utilization before and after retirement.

The RD estimates of the first row in column 1 and column 2 of Table 5 suggest that the age 50 threshold is associated with a significant increase in total medical expenditures; indeed, about

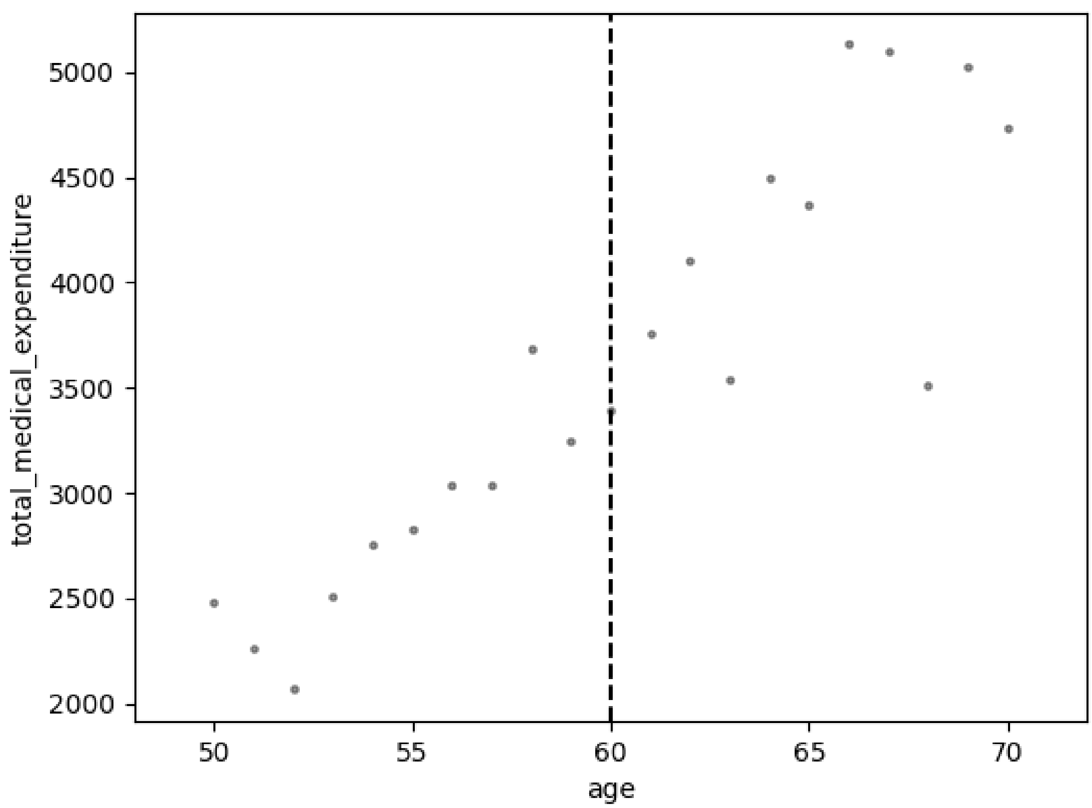


Figure 6: Males' Total Medical Expenditures over Different Ages

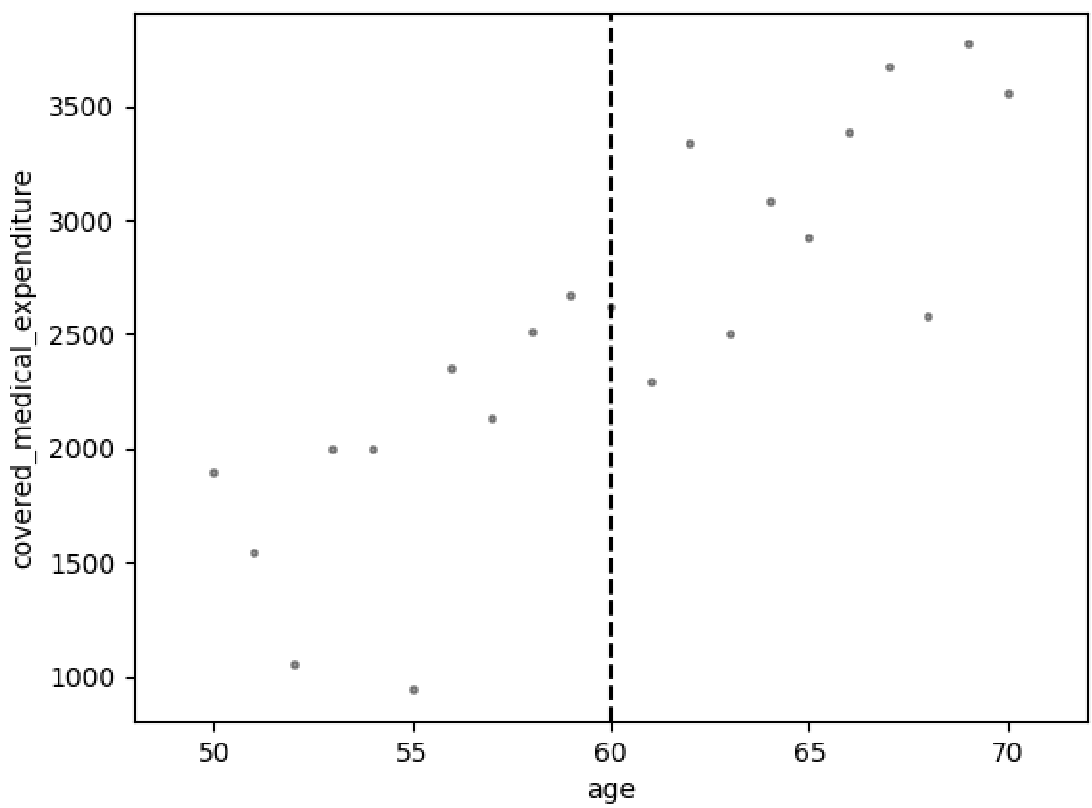


Figure 7: Males' Medical Expenditures Covered by Health Insurance Over Different Ages

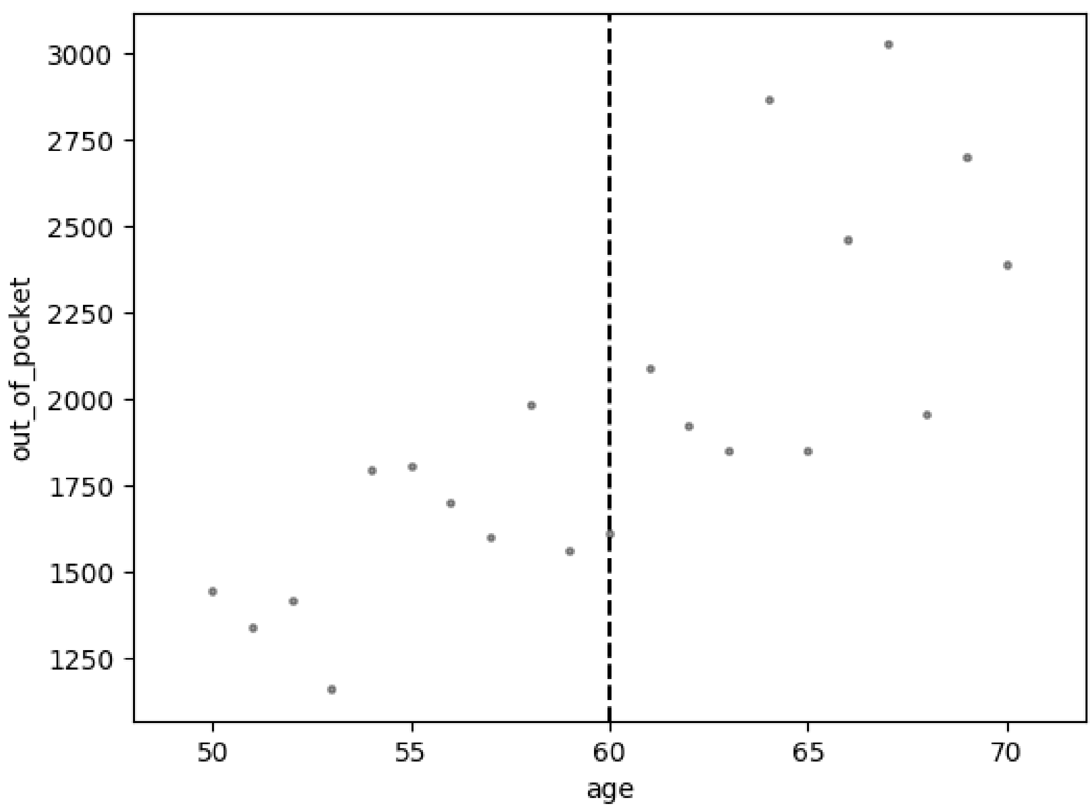


Figure 8: Males' Out-of-pocket Medical Expenditures Over Different Ages

836 RMB (column 1) and 784 RMB (column 2) in expenditures are covered by health insurance for women. Medical expenditures increase at retirement according to the second row of column 1 and 2 of Table 5. To estimate the impact of retirement on medical expenses, the indicator for being older than age 50 is chosen as an IV for the indicator for retirement, the robust standard errors for which are displayed in parentheses. The estimate of total medical expenses is 3872 RMB and the estimate of the amount covered by health insurance is 3782 RMB. As evidenced at the 1% level, a statistical and economical increase occurs at retirement for females.

Furthermore, the willingness of females to pay out-of-pocket in order to utilize increased health care is indicated in column 3 of Table 5. A modest increase in out-of-pocket expenditures (about 1852 RMB) occurs when they retire; the results of this phenomenon are particularly significant at the 5% level.

	total medical expenditures	covered medical expenditures	out-of-pocket expenditures
impact of age $\geqslant$ 50 on y	836.073***	784.989***	403.794***
	(339.993)	(254.345)	(176.484)
impact of retirement on y	3871.885***	3782.301***	1852.399**
(age older than 50 as IV)	(1633.299)	(1332.616)	(837.097)
observations	2712	1819	2712

Table 5: Female: Impact of Retirement on Medical Expenditures

Unlike the estimates for females, the RD estimates in columns 1 and 2 of Table 6 indicate that, for males, the age 60 threshold is unassociated with any significant change in either the total medical expenditures or expenditures covered by health insurance. Additionally, males do not increase their out-of-pocket expenditures as they retire in order to utilize increased health care, as shown in column 3 of Table 6.

	total medical expenditures	covered medical expenditures	out-of-pocket expenditures
impact of age $\geqslant 60$ on y	-112.670 (561.127)	-646.620 (508.270)	93.859 (327.090)
impact of retirement on y (age older than 60 as IV)	-398.916 (1989.260)	-2168.114 (1754.117)	331.425 (1155.979)
observations	2536	1758	2536

Table 6: Male: Impact of Retirement on Medical Expenditures

Interpretation

The magnitude of just how retirement affects the use of health care services is has remained obscure in the literature, involving as it does various incentives for individuals to change their health-seeking behavior. A possible explanation for the ”retirement-consumption drop”, for instance, may involve time cost. As the opportunity cost of time decreases at retirement, and households cut ties to the labor force, expenditures on such ”work-related” categories should decrease even if there are no changes made to lifetime resources or preferences. Similarly, a decrease in the opportunity cost of time after retirement could encourage retirees to utilize more health care.

Another component of health care that should be more heavily evaluated involves financial cost. Indeed, individuals have the opportunity to reduce their health care expenses after retirement because their income is comparatively lower.

Also, medical expense decisions are often made based on health insurance costs and benefits. In the United States, the bulk of medical costs post-retirement are paid by the state *only*; this policy, by extension, creates a strong incentive for individuals to retire. The total medical expenses rise substantially for these retirees, with their out-of-pocket expenses dropping substantially.

By contrast, state-provided health care is available before and after retirement in China, where workers are *mandatorily* enrolled in the social-health-insurance system. An MSA approach gives individuals full control over how their resources are allocated to provide for health-care services over time. Retirees are still covered under the same health insurance, only with lower coinsurance rates and decreased cost sharing. As such, individuals have the opportuny to postpone medical expenses until after retirement, carrying their MSAs over accordingly.

The results of this paper provide some insight into health-seeking behaviors, specifically from an economic perspective. The data display patterns that are in sharp contrast to that of the US. Instead of keeping a smooth, consistent pattern of out-of-pocket expenditures, females in China are willing to increase their out-of- pocket expenditures in order to take advantage of better health insurance benefits and increased medical care. As such, a distinct jump in health care resource utilization occurs when females retire. By contrast, males show smooth, consistent patterns in total medical expenditures as well as out-of-pocket expenditures. These consistent medical expenses continue after males retire.

Effects of Retirement On Pre-assumption Tests

I conduct validity tests for the RD design by examining two features of the underlying assumptions for the RD design. First, individuals do not have precise control over the forcing variable in the neighborhood of the cutoff point. According to this context, the running variable age is unlikely to be manipulated.

Second, I have chosen to test the validity of the RD design by identifying whether other variables correlate with the jump in the probability of retirement for females at age 50. The variables being tested include education level and marriage status. The continuity assumption of the RD design considers the influence of any other factors that might affect the outcome of interest trends at age 50.

Fig 9 and Fig 10 indicate that these variables do not significantly change around age 50. These findings are confirmed by the regressions that are reported in columns (2)-(3) in Table 7 and Table 8, which indicate that the coefficients of the education and marital statuses are insignificant.

Taken as a whole, both the education levels and marital statuses trend relatively smoothly around the age thresholds. As such, they are unlikely to confound my analysis of the impact of retirement on medical expenditures.

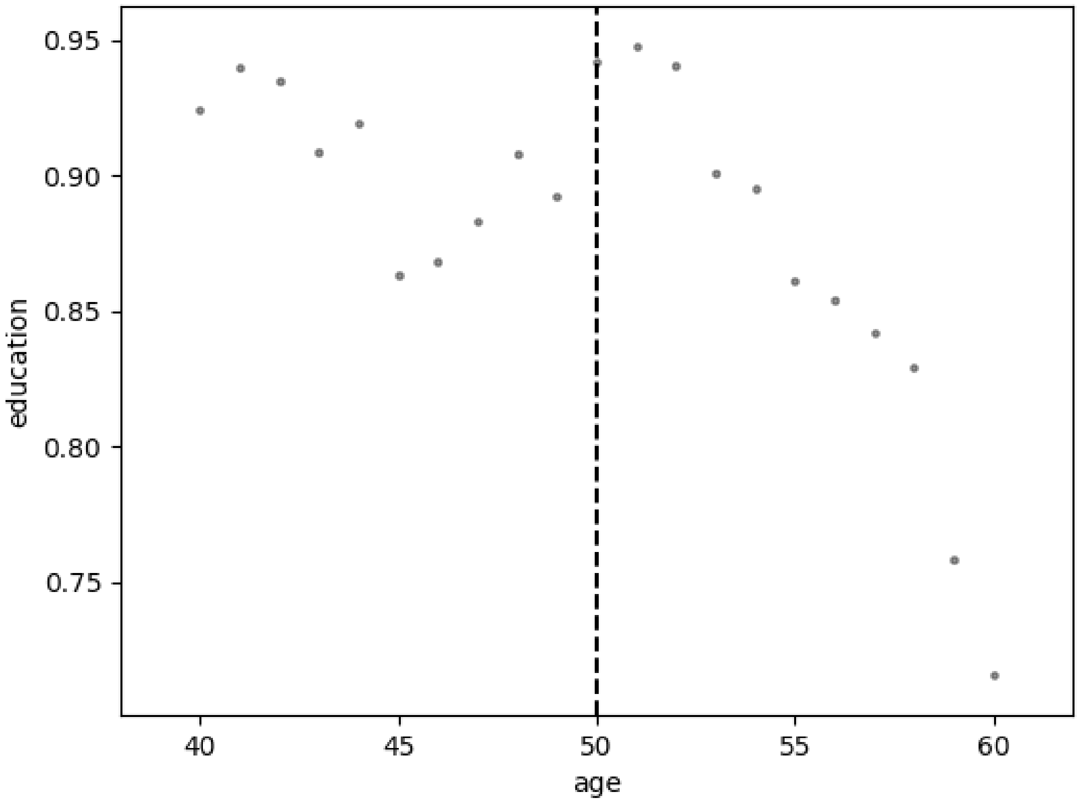


Figure 9: Females' Proportion of Having a Middle School Degree over Different Ages

Conclusion And Policy Implication

Of the existing literature, this is the first paper that specifically studies the causal impact of retirement on medical expenses in the context of China by using a regression discontinuity design. It provides new and essential findings about medical expenses upon retirement, and contributes extensively to the "retirement-consumption puzzle" literature. In addition, this paper highlights the fact that females and males show different reactions to the financial opportunities that occur with regard to medical care upon retirement. As discussed in the previous section, females show a distinct change in how they utilize health care upon retirement, based on the estimates of total medical expenditures and expenditures that are covered by insurance being significant at the 1% level, and out-of-pocket expenses at the 5% level. Males, however, do not exhibit such a distinct

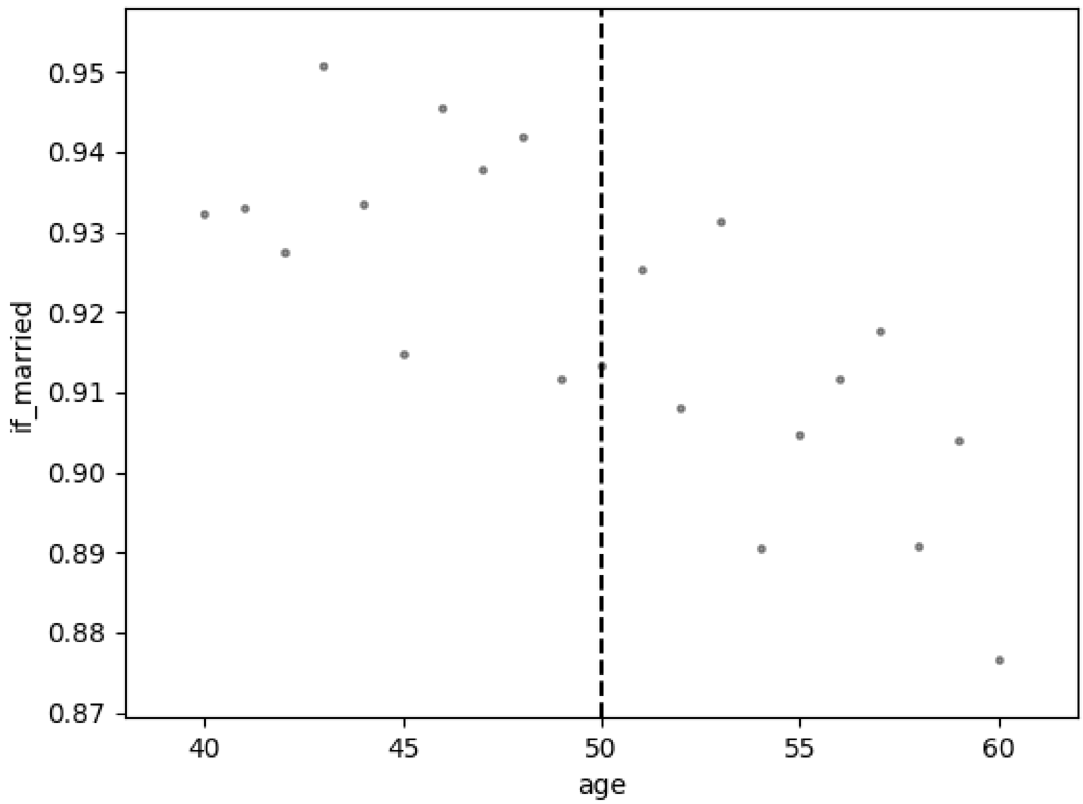


Figure 10: Females' Proportion of Being Married over Different Ages

change in their health-seeking behaviors upon retirement, as indicated by the fact that all of the estimates that appear in Table 6 are insignificant.

This paper also finds that the exogenous increase in the health insurance benefits typically encourages patients to utilize more health care upon retirement. This factor is significant due to the growing concern that overly comprehensive insurance may cause cost growth problems, most notably with regard to the expenditure growth of Medicare. As the government finances both the pension and health insurance system, the basic health care of retirees is big consumption to the government.

The exogenous increase in health insurance benefits upon retirement gives individuals the financial incentive to delay the utilization of their health care until retirement. However, this increase in health insurance benefits does not take into account the type or degree of illness that the

	retired	education	married
impact of age $\geqslant$ 50 on y	0.216***	0.039	-0.007
	(0.027)	(0.024)	(0.020)
impact of retirement on y		0.194	-0.033
(age older than 50 as IV)		(0.123)	(0.090)
observations	2712	2712	2712

Table 7: Female: First Stage Estimation and Pre-assumption Test

	retired	education	married
impact of age $\geqslant$ 60 on y	0.283***	-0.011	-0.017
	(0.036)	(0.025)	(0.016)
impact of retirement on y		-0.042	-0.065
(age older than 60 as IV)		(0.098)	(0.064)
observations	2536	2536	2536

Table 8: Male: First Stage Estimation and Pre-assumption Test

individual may be incurring. Patients may be able to delay their health care until they retire in order to save on medical expenses, but the potential decrease in coinsurance rates could encourage patients to postpone receiving essential medical care. Putting off such medical attention may have dangerous consequences for the patients' health; indeed, delaying certain preventative treatment may put patients at unnecessary risk. A more reliable design might thus be more beneficial for determining the appropriate types of treatment for individuals entering retirement.

CHAPTER 2: DOES RETIREMENT AFFECT "SANDWICH" GENERATION'S FAMILY SUPPORT NETWORK?

Introduction

Previous researchers have attempted to study the conflict and cooperation between the State and the family. It is clear that inter-generational cooperation improves the wellbeing of all generations, as it transfers consumption and education to the younger generation and consumption and care to the older generation. However, to a limited degree these transfers can emerge from an inter-temporal market. The missing degree relies largely on a family support network. It is thus vital that both economists and policy makers study the patterns and determinants for private transfer and their interactive roles with public transfers, as they have important economic implications, including the design and implications of public policies on income distributive issues, etc.

More recent research is needed in order to understand whether the gradual shift from family-based old-age support in developed countries would also occur in developing countries. As in Lee (2016), the main question of interest is: Given the elements of social modernization in developing countries, is the support of older persons within the family sustainable? China, one of the fastest developing countries, is home to ever increasing old-aged populations; as the country's urban pension provisions are linked to the labor market of the current working population, retirees also share the benefits of the growing economy. Indeed, the pension per capita has enjoyed about a 10% growth rate for the past 10 years, inadvertently challenging the traditional family support culture, Urban China therefore provides an effective context from which to study the topic, as its strong filial societal traditions continue to conflict with rising modernization.

Over the years, society has become more familiar with those citizens of the "sandwich" generations (those middle aged households who not only take care of their children but also support their elderly parents). Although the literature has studied many of the relationships that exist between adult children and their parents, they neglect to offer the same degree of attention to the roles that these "sandwich" generations play in the family support network. It is important to notice that these middle-aged generations are "sandwiched" between the competing needs of their children and those of their aging parents/parents-in-law. Such individuals struggle with bilateral altruism (Cigno (2016)) as they embrace their role in the income redistribution among three generations.

A lot of recent research finds that inter vivos transfers play different roles over an individual's life cycle. Switching from working age to retirement is one of the major transitioning points of the life-cycle. Previous studies provide an incomplete picture of the private transfer to retirement due to data limitation and credible identification strategy. In this paper, I study the private transfer patterns against the fast-growing public support system in urban China for these "sandwich" generations. Do their children start to take on the responsibility of supporting them as they retire? What is their overall role that they play in the family support network? I investigate the overall pictures of the financial resource flow that these "sandwich" generations are facing, which includes upward transfers with their parents/ parents-in-law and their downward transfers with their children. Their stepping into retirement has significant impacts on their role in the family support network.

In order to identify the causal effect of retirement on "sandwich" generations' financial transfers between generations, I apply a regression discontinuity framework (see Hahn, Todd, and Van der Klaauw (2001)) and exploit the exogenous mandatory retirement age rules. These strategies benefited substantially from CHARLS (China Health and Retirement Longitudinal Study) dataset, which itself includes information concerning middle-aged households on detailed transfers with both their children and their parents or parents-in-law. Detailed knowledge of extended family transfer networks allows me to shed new light on the inter-generational transfers (both elaborate and realistic) that a real household might face. To correct for the endogenous nature of the retirement decisions, I have chosen to both explore the exogenous variability in age eligibility, and engage an identification strategy that assumes that all private transfers would be the same around the threshold for age eligibility if households do not retire.

The moment that the "sandwich" generation steps beyond the accumulation phase of their life cycle into retirement, their family support network is substantially impacted. As such, the first goal of this paper is to depict a more realistic picture of the inter-generational transfers that middle-aged Chinese households are currently facing – households that are not only dealing with downward transfers with their children but also upward transfers with their parents/parents-in-law. Secondly, this paper will explore how retirement induces these "sandwich" generations to switch from resource providers to resource takers in the private transfer channel, similarly to how their roles switch in the public transfer channel. By retiring, these individuals move from supporting the elderly by paying taxes to being supported by the younger generation through pension; the children of these individuals in turn take on the responsibilities of supporting their parents when they transfer financially to middle aged households as they retire.

Background

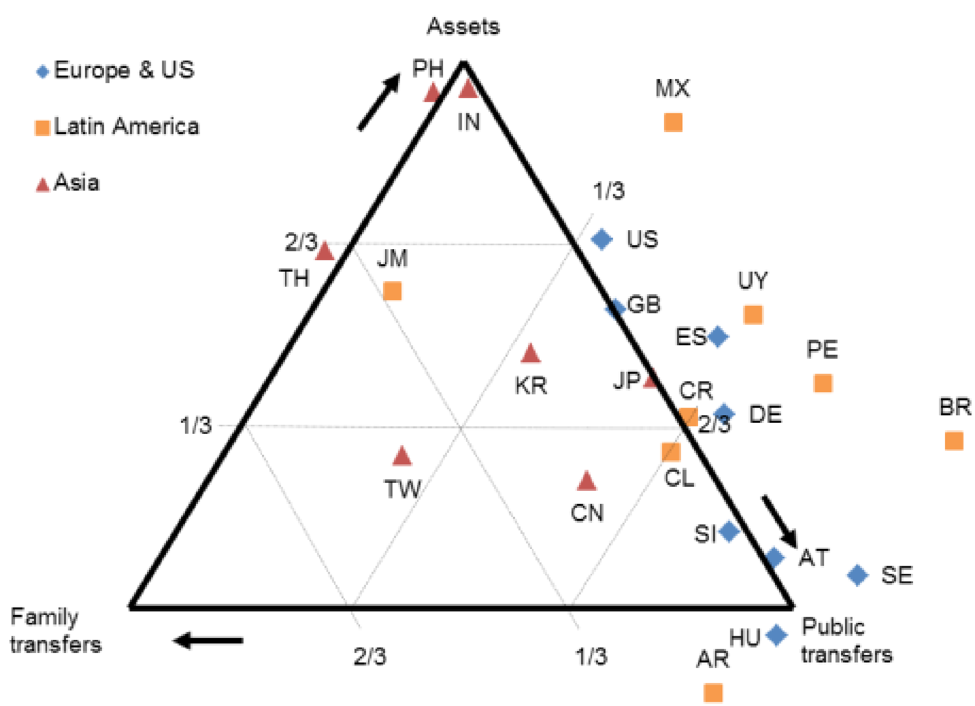
Old-age Support System

Countries tend to vary greatly in how consumption is funded for elderly people. Though the elderly continue to work in certain countries, their labor income is usually too low to cover their consumption. How, then, do the elderly fund their consumption if they consume much more than the labor income they earn? Other sources of income, including assets, public transfer, and private transfer, help to fill in these gaps.

As pointed out in the survey paper (Lee (2016)), the elderly in developed countries rely largely on asset income; this is due, in part, to the more elaborate financial and insurance markets, as well as the more advanced social protection, that are available in developed countries. Using these findings to make inferences about old-age support systems in developing countries would likely be misleading, however, since developing countries vary fundamentally in terms of private transfer traditions and the sizes of their public transfer systems.

Altruism between generations is more pervasive in developing countries because they have less sufficient public transfer, as well as a capital market that is less poorly developed when compared to those of other countries. Some researches document altruism motives by studying the predictions of the altruism models (Becker (1974)), which indicate that an increase in public transfers would be neutralized by private transfers. For example, Fan (2010) finds that private transfers displace the benefits of public transfers in South Africa. Jensen (2004), as well, finds that a reduction of private transfers occurs in response to the Farmer's Pension Program of Taiwan, etc. Finally, Chuanchuan and Binkai (2014) studies what happens to the private transfers of rural residents' when introduced to the rural pension systems.

Many of these researches focus on "the consequences of transitioning from an unfunded to a funded public pension program." However, it should be noted that different levels of the public transfer systems in developed and developing countries have varying impacts on private transfers.



Source: Lee (2016)
Figure 11: Gap Between Consumption and Labor Income in Old Age is Funded

As in Lee (2016), Fig 11 shows how the gap for the elderly is funded. Each vertex refers to a single source of funding. The funding from the various sources is equivalent to 1. Each country, coded according to a two-letter abbreviation, is presented as a point in the figure. For example, in the United States, more than half the funding for the elderly relies on asset income; its representative point, which lies outside the triangle on the right, indicates that it has negative net private transfer. In other words, the elderly in United States make net transfers to the younger generation. However, the point representing China lies inside the triangle, which means it has a mixture of all three funding sources; such funding relies mostly on public transfers, while roughly 30 percent stems from private transfers.

Urban Pension System In China And Family Support Tradition

The urban residents in China provide a good opportunity to study whether a consistent private transfer pattern for these "sandwich" generations is capable of existing over the course of their life cycle regardless of the impact of modernization. These urban residents participate in a pension system that reflects the growth benefits of modernization and development since pensions are linked to the average local wage. As exhibited in Fig 12, the pension per capita in China has enjoyed a 10 percent growth rate over the past 10 years.

According to the policies in place, both employees and employers are required to make contributions to the pension system. Workers contribute based upon their individual wage (a rate of up to 8 percent) while employers contribute a percentage of the total wages paid to their workforce (usually around 20 percent). The exact contribution rates vary from region to region. Workers' contributions are paid into a personal account; upon retirement, the balance of the account, including interest, is divided into 120 installments that are to be paid out monthly over a ten-year period. In addition, the worker also receives general pension payments that are payable until death. The general pension payments are determined by the number of years of employment, the average local wage, and the worker's life expectancy. These general pension payments are funded by the employer's contributions, though the government is legally obligated to cover any shortfalls.

Conversely, the emerging literature has identified the degree to which private transfers are constrained by cultural, familial or religious norms. Indeed, researchers (Lindbeck and Nyberg (2006), Cigno, Giannelli, Rosati, and Vuri (2006), Cigno, Komura, and Luporini (2017)) have at-

Source: China Economic Weekly

Figure 12: Growth Rate of Pension Per Capita over Years

tempted to explain the emergence and persistence of social and familial norms between generation. In China, for instance, the expectation that children will care for the elderly is a deep-rooted tradition. According to Confucian ethics, filial piety is a sign of respect for one's parents, elders, and ancestors. In more general terms, filial piety means to be good to one's parents by taking care of them. Such filial piety is considered a key virtue in Chinese and other East Asian Cultures; indeed, while China has always had a diversity of religious beliefs, filial piety is valued amongst almost all of them. In fact, the expectation that adult children will support their elderly parents is included in both the Constitution of the People's Republic of China as well as the Marriage Law.

Given these societal expectations, it is vital that researchers consider how the "sandwich" generations' decision to retire impacts the family support network. China's strong family support tradition and growing pension support system establishes it as an ideal setting to assess these important topics, predominantly how both private transfer and public pension scheme support elderly population.

Two-sided Altruism

The standard models of transfers are based on simple altruism (Becker (1974)), exchange (Cox (1987)), and accidental bequests (Barro (1974)), as well as Altonji, Hayashi, and Kotlikoff (1997), etc. Regarding the patterns and determinants of inter-generational transfers, several empirical studies focus on studying the responsiveness of private transfers to the recipients' income. The evidence of these studies is mixed. Much of the empirical research in the US has suggested a negative correlation (McGarry and Schoeni (1995)), though a positive correlation is found in Peru (Cox, Eser, and Jimenez (1998)). Failures of standard models that explain the emerging various empirical findings call for both more elaborate models and a more holistic picture of the circulation of private transfers in both developed and less developed countries.

Unlike the one-sided altruism that is predominant in the US (Laitner (2001)), inter vivos financial transfers from adult children to parents are limited; such inter vivos financial transfers from parents to adult children far exceed the transfers from adult children to their elderly parents.

According to (Genicot (2016), Cremer, Pestieau, and Roeder (2016), Aoki and Nishimura (2017), Ottoni-Wilhelm, Vesterlund, and Xie (2017)), inter vivos transfers stemming from two-sided altruism occur during the life cycles of Chinese "sandwich" generations.

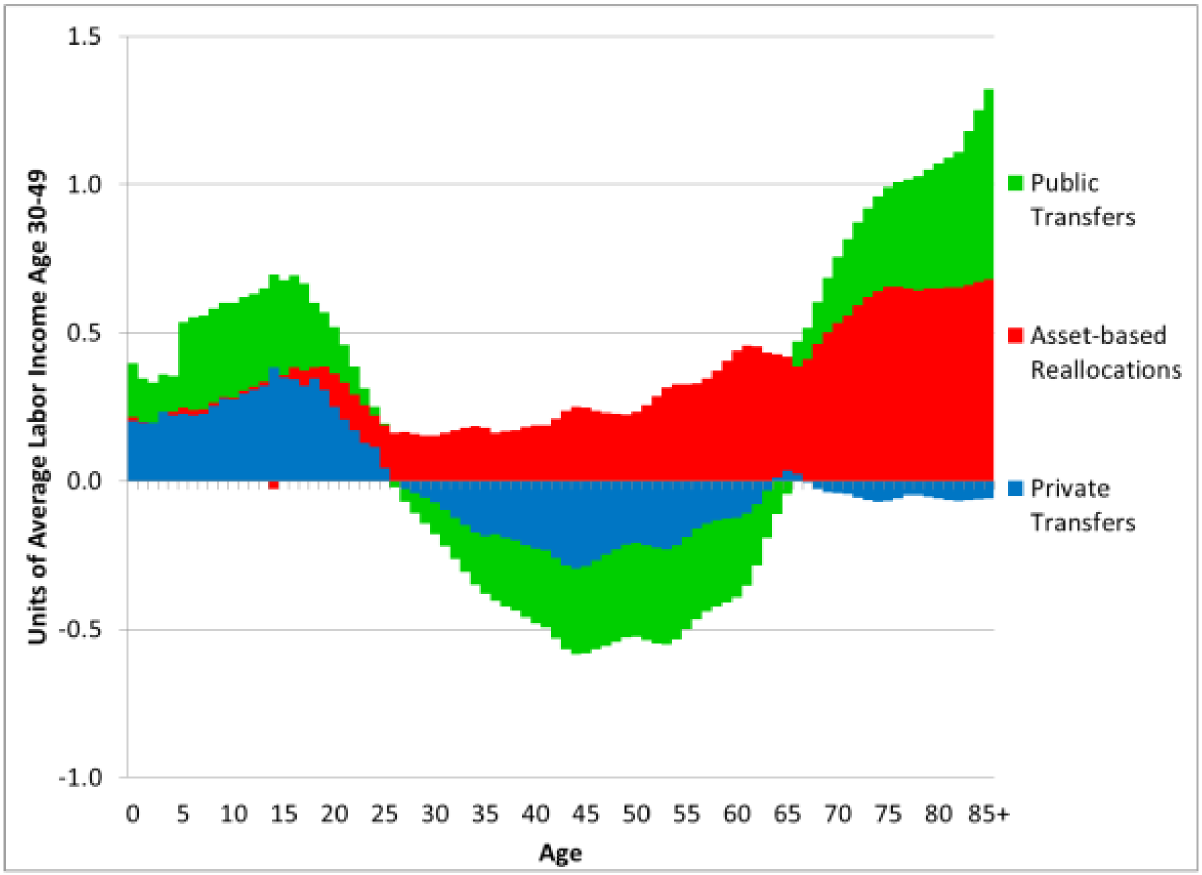
Generally, households derive utility not only from their own lifetime consumption, but also from the well-being of their parents/parents-in-law and children. As a result, such households transfer certain resources to their children and in support of their parents/parents-in-law via inter vivos transfers.

Those households that abide by two-sided altruism adjust their consumption and saving level according to the flow of financial resources to parents/parents-in-law and downward children. This invariably leads to different consumptions and saving profiles over the course of one's life cycle. For example, Lee, Lee, and Mason (2008) finds that while the consumption patterns of young individuals are similar in the US and Taiwan, the consumption profile of those 40 years or older are very different. Indeed, the difference sources of support for multi-generational relationships contribute to the differences in these consumption profiles (Mason, Lee, Tung, Lai, and Miller (2009)).

Heterogeneity of Private Transfer Over The Life-cycle

Due to the strong heterogeneity of inter vivos transfers, modelling and predicting appears quite unrealistic. One of the most neglected questions in the literature is whether these private transfers occur over the course of a life-cycle, as pointed out in Arrondel and Masson (2006).

Financial transfers lead to different implications according to what stage of one's life cycle they are received. For example, children's consumption and human capital investments are largely funded by their parents. Such transfers add presumably to the children's present and future incomes. Liquidity-constrained adult children can extend their access to credit with their parents' financial assistance; these transfers add primarily to the children's consumption. Private transfers received later in life are presumably added to their wealth.



Source: Lee (2016)

Figure 13: How the Gap Between Consumption and Labor Income at Each is Made Up in the US

According to Lee (2016) in Fig 13, the United States' main source of funding for elderly consumption is assets. Despite this percentage, however, American net private transfers between parents and children continue to be negative, including those in working age to older ages. In other words, downward transfer from parents to children is still dominant even after retirement. There is no change in the private transfer behaviors at retirement.

China, unlike the American public transfer institutions and family support environment, has different private transfer patterns. Due to the limitation of data availability, researchers have begun to study the private transfer patterns in China, including Secondi (1997), Cai, Giles, and Meng (2006), Goh (2009). The findings of these studies provide some idea about the magnitude and direction of private transfers in China, however they fail to study the life cycle patterns of these private transfers.

China's Mandatory Retirement Policy

One of the reasons that it is difficult to observe the life cycle patterns of private transfers in the United States, especially changes at retirement, is due to voluntary retirement decisions, as most professions do not have a mandatory retirement age. There are ages, however, when people can claim their full Social Security, Medicare, and other benefits. More importantly, the retirement policy in the United States allows retirees to receive pensions or benefits from the government without necessarily having to leave the company or their position; it follows, then, that retirees can choose to continue to work at the same position even after they process their retirement.

In China, however, processing for retirement requires individuals to leave the company and their position. The competitive and relatively more tough labor market for retirees typically forces them to leave the labor market permanently, which has a huge impact on retired households.

In addition, urban Chinese residents abide by the Chinese mandatory retirement policies According to a series of government documents, these polices are mainly based on age requirement; a policy issued in 1994 states that, due to the process of SOE (state-owned enterprises) reform in the 1990s, a woman can retire at age 45, 50 or 55 as a cadre, while a man can retire at age 50, 55 or 60. ZHAO (2010) identifies discrepancies in the probability of retirement at those retirement ages, using the 2005 1% population sample survey.

There is emerging literature studying the change of consumption patterns at retirement according to the regression discontinuity design, such as that of Battistin, Brugiavini, Rettore, and Weber (2009), Cho (2012), Battistin, Brugiavini, Rettore, and Weber (2007), Li, Shi, and Wu (2015), etc. However, there are a small number of studies that use the regression discontinuity design to identify the causal impact of retirement on private transfers. As such, this paper will apply the RDD framework to study the private transfers in China.

Data And Descriptive Evidence

Data

China Health and Retirement Longitudinal Study (CHARLS) 2011 baseline survey is a nationally representative survey of the middle-aged and elderly population, as well as their spouses, across 150 counties in 28 provinces across China. Within each household in China, one person was randomly chosen to be the main respondent, after which their spouse was automatically included. Based on this sampling procedure, 1 or 2 individuals in each household were interviewed depending on the marital status of the main respondent. The total sample size was 10,257 households and 17,708 individuals.

The CHARLS main questionnaire in the 2011 survey consists of 7 modules; those modules covered demographics, family background, socioeconomic status, health status (including physical and psychological health), and environment (community questionnaire and county-level questionnaire). The data for this survey was collected using face-to-face, computer-aided personal interviews (CAPI).

In the family module of this 2011 survey, private transfer is defined as the incidence and magnitude of transfer between the household and non-coresident children, as well as between the household and their non-coresident parents and parents-in-law. In terms of the upward transfer to parents or parents-in-law, CHARLS respondents were asked: ”In the past year, did you or your spouse receive any economic supports from your non-coresident parents/ parents-in-law?” and ”How much did you receive from your non-coresident parents / parents-in-law in the past year?” They were instructed to include regular and non-regular monetary support and in-kind support. They were also asked about any economic supports given to their non-coresident parents/ parents-in-law. The same set of questions was asked with regard to any downward transfer to their non-coresident children.

Due to the social norm of “sandwich” generations supporting both parents and parents-in-law, upward transfers with older generations include all the financial transfers with both parents and parents-in-law. In addition, downward transfers include all financial transfers with any non-coresident children. In this paper, all regular and non-regular support - including both money and in-kind support - are calculated into the gross measures.

Variables Definition And Descriptive Evidence

In this paper, private transfer is defined as both the incidence and magnitude of transfer behavior between “sandwich” generations and non-coresident children as well as non-coresident parents/parents-in-law. The key outcome variables of interest are defined as: (1) a binary variable indicating the incidence of the household receiving economic transfer, in addition to any cash and in-kind transfers from any of their non-coresident family members, including their children and parents/parents-in-law, or zero otherwise; (2) a continuous variable measuring the net amount of the overall transfers, which may include the summation of net transfers from households to their non-coresident children, as well as net transfers from households to their non-coresident parents and parents-in-law; and (3) a dichotomous variable representing whether the household is overall a net recipient, taking the value of 1 in the case of positive net transfer, and zero otherwise.

	mean	std	count
If_receiving_money_from_kids	0.333	0.471	1720
if_giving_money_to_kids	0.151	0.358	1720
if_receiving_money_from_either_parents	0.048	0.214	1290
if_giving_money_to_either_parents	0.479	0.500	1290
if_receiving_any_money	0.275	0.447	2292
if_giving_any_money	0.347	0.476	2292
if_positive_net_overall	0.457	0.498	1211
if_positive_net_kids	0.711	0.454	736
age	58.583	10.523	2487
if_retired	0.568	0.495	2487
if_adls	0.149	0.356	2487
if_middle_school	0.690	0.463	2487
if_married	0.818	0.386	2487
ln(income)	10.140	0.699	1878

Table 9: Descriptive Statistics

Table 9 shows a descriptive summary of the entire sample of urban households. It depicts

	count	mean	std	min	25%	50%	75%	max
net_kids	736	366.940	15248.024	-85000	-500	1000	3625	80000
net_parents	616	-1868.604	3151.367	-16000	-2500	-1000	-400	10000
net_overall	1211	-685.175	10417.887	-61000	-2000	-200	1800	62400

Table 10: Descriptive Statistics for Net Amount of Transfers

a list of transfer patterns between "sandwich" generations and other non-coresident generations, including their non-coresident children and parents/parents-in-law. As shown in this table, intergenerational transfers are prevalent.

From a certain perspective, upward transfers toward non-coresident parents/parents-in-law are widespread; indeed, 48 percent of the "sandwich" generations are themselves donors of these transfers. Only 5 percent of households receive economic resources from their elder parents/parents-in-law, which implies that these "sandwich" generations are substantially supporting their elderly parents/parents-in-law. This is a strong indication of the filial piety that is so predominant in China.

Conversely, these "sandwich" generations have financial relationships with their non-coresident children. About 33 percent of the households receive private transfers from their children, while 15 percent of the households give cash or in-kind transfers to their adult children.

In table 10, the average net amount of transfers reveals a great deal about the overall flow of economic resources between generations. The upward net amount toward parents/parents-in-law show distinctive transfer patterns from a downward net amount toward their children. On average, "sandwich" generations give a higher amount of transfers toward their elderly parents/parents-in-law. However, such generations mainly receive a positive net amount from their children. This is distinct from patterns in the US, where households dominantly transfer resources to the adult children. Indeed, combining the two-sided transfer income flows reveals a more complete picture of the reality that Chinese households are facing. Overall, they are donors in the family support network, as their overall net transfers are negative on average.

In line with previous studies on private transfer, several variables reflecting households' demographic and socio-economic characteristics are specified. For instance, marital status is divided into two distinct groups: Married and Single (the latter of which includes divorced, widowed, separated, or never married). Education attainment is classified into either Middle-school degree and above, or lower education. Household income includes the pooled income of the main respondent and his/her spouse's income. Adl, which indicates if any of the main respondents and his/her spouses has functional difficulties when undergoing everyday activities, is used as a proxy for health measure.

Identification

In this paper, I estimate the causal effects of retirement on the private transfer network; however, the key coefficients of interest are problematic due to the potential endogeneity issue. For example, households that expect to receive economic support from their children may have more incentive to retire, which could be a reverse causality. Secondly, other unobservable characteristics (preferences, ability, etc...) may simultaneously determine one's retirement decision and private transfer behavior. The usual ordinary least square (OLS) estimates would thus be biased since the endogeneity issue could contaminate the estimates. In my context, however, the existence of a mandatory retirement rule establishes an effective identification strategy for exogenously identifying the causal link.

Following the notation of the potential outcome framework, (Y_0, Y_1) are the two potential outcomes a household would experience. Y_0 is the measure of interest when one household is not retired, and Y_1 is the measure of interest when one is retired. In this case, $D = 1$ represents a household that is retired and $D = 0$ indicates a household that is not retired. Therefore, an actual measure of private transfer Y can be represented by

$$Y = Y_0(1 - D) + Y_1 D = Y_0 + (Y_1 - Y_0)D$$

In this paper, three outcome variables are of interest. For i-th household in the sample, Y_{1i} is defined as a binary variable, indicating if a household receives transfer from non-coresident children or parents/parents-in-law. Y_{2i} is a continuous variable measuring the overall net amount from non-coresident children and parents/parents-in-law. Y_{3i} is defined as a binary variable indicating if a household gains from the overall private transfer – in other words, if the value of transfers received exceeds the value that is transferred out.

D_i is the binary variable indicating the retirement status of a household, while $D_i = 1$ indicates if a household has at least a retiree and $D_i = 0$ otherwise. A discontinuity design (Thistlethwaite and Campbell (1960)) arises when D_i depends on an observable variable S_i, where S_i measures the age of a household, and in the support of S_i the probability of being retired changes discontinuously at the threshold $\bar{s}$. In other words,

$$Pr\{D_i = 1|\bar{s}^+\} \neq Pr\{D_i = 1|\bar{s}^-\}$$

where $\bar{s}^+$ and $\bar{s}^-$ refer to households that are marginally above and below $\bar{s}$, respectively.

According to Trochim and Thochim (1984), the distinction between sharp and fuzzy regression discontinuity designs depends on the size of the probability jump at the threshold. A sharp regression discontinuity design occurs if the retirement status is deterministically dependent on whether the household's age is above $\bar{s}$; in other words, if $D_i = 1(S_i \geqslant \bar{s})$. A fuzzy design is applicable if the size of discontinuity at $\bar{s}$ is smaller than one.

These private transfers are recorded at the household level in the survey. In addition, any private transfers in a Chinese family are mutually made or received by both husband and wife. In this study, a household's age is represented by the mean age of both the husband and wife residing in the household. A household's status becomes retired once any one of the main respondent and spouse retires. Hence, the probability of a household being retired is discontinuous at age 50. A fuzzy design is used in this paper since the discontinuity at age 50 is smaller than one, and because reaching age 50 does not necessarily imply that the household is actually retired. Though the mandatory rule induces a higher probability of being retired, some households may not comply to the retirement rule.

It can be shown that the local average treatment effects of retirement on private transfer can be recovered for retired households around $\bar{s}$.

Therefore, three respective treatment effect δ_1, δ_2 and δ_3 can be defined as follows:

$$\delta_1 = \frac{\lim_{S\to\bar{s}^+} E[Y_{1i}|S] - \lim_{S\to\bar{s}^-} E[Y_{1i}|S]}{\lim_{S\to\bar{s}^+} E[D_i|S] - \lim_{S\to\bar{s}^-} E[D_i|S]}$$

$$\delta_2 = \frac{\lim_{S\to\bar{s}^+} E[Y_{2i}|S] - \lim_{S\to\bar{s}^-} E[Y_{2i}|S]}{\lim_{S\to\bar{s}^+} E[D_i|S] - \lim_{S\to\bar{s}^-} E[D_i|S]}$$

$$\delta_3 = \frac{\lim_{S\to\bar{s}^+} E[Y_{3i}|S] - \lim_{S\to\bar{s}^-} E[Y_{3i}|S]}{\lim_{S\to\bar{s}^+} E[D_i|S] - \lim_{S\to\bar{s}^-} E[D_i|S]}$$

Under the following conditions:

Condition 1: $E[Y_{ki}(0)|S = s]$ and $E[Y_{ki}(1)|S = s]$ are continuous functions of S for $k = 1,2,$ 3.

Condition 2: $D_i(S)$ is increasing in S at $S = \bar{s}$.

Empirical Analysis

Estimation

The primary task of this paper is to evaluate the impact of retirement on the family support network; three outcome variables are of interest. With this in mind, the following equations will be considered:

$$Y_{1i} = \alpha_1 + \delta_1 D_i + \beta_1(S_i - \bar{s}) + \beta_2(S_i - \bar{s})1(S_i - \bar{s} \geqslant 0) + \varepsilon_i$$

$$(\bar{s} - h < S_i < \bar{s} + h)$$

where Y_{1i} is a binary variable indicating if a household receives transfer from non-coresident children or parents/parents-in-law. D_i indicates whether the household is treated (equal to 1 if a household has at least a retiree) or not. The forcing variable S_i measures the age of a household, and the threshold $\bar{s}$ equals to 50.

Similar estimations are carried out by replacing the dependent variable with Y_{2i}, a continuous variable measuring the overall net amount from non-coresident children or parents/parents-in-law. The following equations are estimated:

$$Y_{2i} = \alpha_2 + \delta_2 D_i + \zeta_1 (S_i - \bar{s}) + \zeta_2 (S_i - \bar{s}) 1 (S_i - \bar{s} \geqslant 0) + \varepsilon_i$$

$$(\bar{s} - h < S_i < \bar{s} + h)$$

Another outcome variable of interest Y_{3i} is defined as a binary variable indicating if a household gains from the overall private transfer – in other words, if the value of transfers received exceeds the value that is transferred out.

$$Y_{3i} = \begin{cases} 1 & if\, overall\, net\, transfer > 0 \\ 0 & if\, overall\, net\, transfer \leqslant 0 \end{cases}$$

$$Y_{3i} = \alpha_3 + \delta_3 D_i + \gamma_1 (S_i - \bar{s}) + \gamma_2 (S_i - \bar{s}) 1 (S_i - \bar{s} \geqslant 0) + \varepsilon_i$$

$$(\bar{s} - h < S_i < \bar{s} + h)$$

Since the OLS estimate using D_i could be subject to selection bias, I have chosen to introduce a second variable $1(S_i \geqslant \bar{s})$ where $\bar{s} = 50$ as an instrumental variable for D_i to identify the parameter of interest δ_1, δ_2 and δ_3, the effects of retirement on the probability of obtaining transfer from children, on the overall net amount and probability of having positive overall net transfer at $\bar{s}$, respectively.

Main Results

By plotting the relationship between the running variable (age) and the private transfer variables of interest, the visual representation allows this study to determine whether the private transfer patterns have a discrete response to retirement.

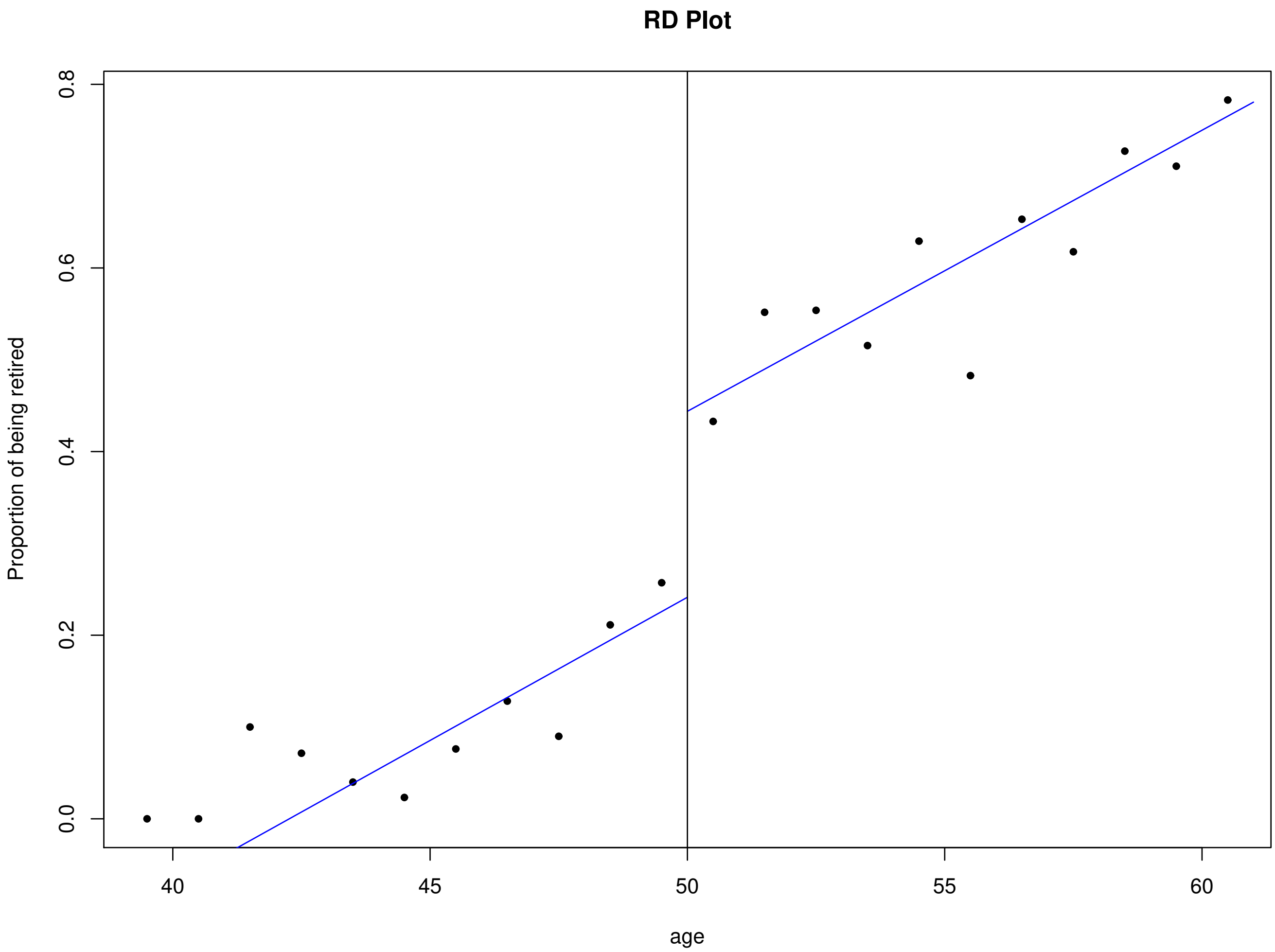


Figure 14: Proportion of Being Retired over Age

	retired	ln(income)	education	being married	health
impact of age $\geqslant 50$ on y	0.188***	-0.196*	0.025	0.032	-0.002
	(0.045)	(0.102)	(0.038)	(0.032)	(0.060)
impact of retirement on y		-0.820*	0.122	0.156	-0.010
(age older than 50 as IV)		(0.488)	(0.179)	(0.154)	(0.288)
observations	1309	1086	1309	1309	1309

Table 11: First Stage Results, Impacts of Retirement on Household Income and Pre-assumption Tests.

	Probability of receiving transfer	Net amount of total transfer	Probability of net transfer being positive
impact of age $\geqslant 50$ on y	0.094*** (0.039)	1704.658 (1302.906)	0.112** (0.056)
impact of retirement on y (age older than 50 as IV)	0.501** (0.248)	9664.546 (8593.335)	0.553* (0.332)
observations	1309	722	722

Table 12: Impacts of Retirement on Overall Transfers

In the first stage, being 50 years of age or older can strongly predict the probability of retirement. Fig 14 shows a sudden jump in the probability of retirement at the age 50.

Regression results reported in the first column of table 11 confirm the graphical findings. I regress the dummy variable for retirement on the dummy variable of being older than 50. The coefficient on the dummy variable for "older than 50" is 0.19, which is significant at the 1% level, suggesting that the probability of retirement jumps by 19 percents at age 50.

I indicate the two patterns of private transfer in response to retirement in Fig 15 and Fig 16, respectively. Table 12 also shows the RD results from estimating the equations in 5.1. Rectangular kernel is used, and a bandwidth of 10 years is applied on both sides of the cutoff. The numbers reported in the parentheses are robust standard errors.

Fig 15 shows the impact of age on the likelihood of receiving transfer from non-coresident children or parents/parents-in-law. A discrete upward probability jump of receiving any transfers occurs around age 50. Estimation results are reported in the first column of Table 12, suggesting that the likelihood of receiving transfers from children or parents/parents-in-law increases by 50 percentage point, which is significantly different from zero at the 5% level.

Considering the fact that these "sandwich" generations are charged with the responsibility of supporting their parents/ parents-in-law, it is essential that the occurrence of the positive net amount of total transfers (including transfers to their children and transfers to their parents/ parents-in-law) be investigated; such an investigation will offer vital information about their private transfer flow status, and will determine if these "sandwich" generations are resource providers

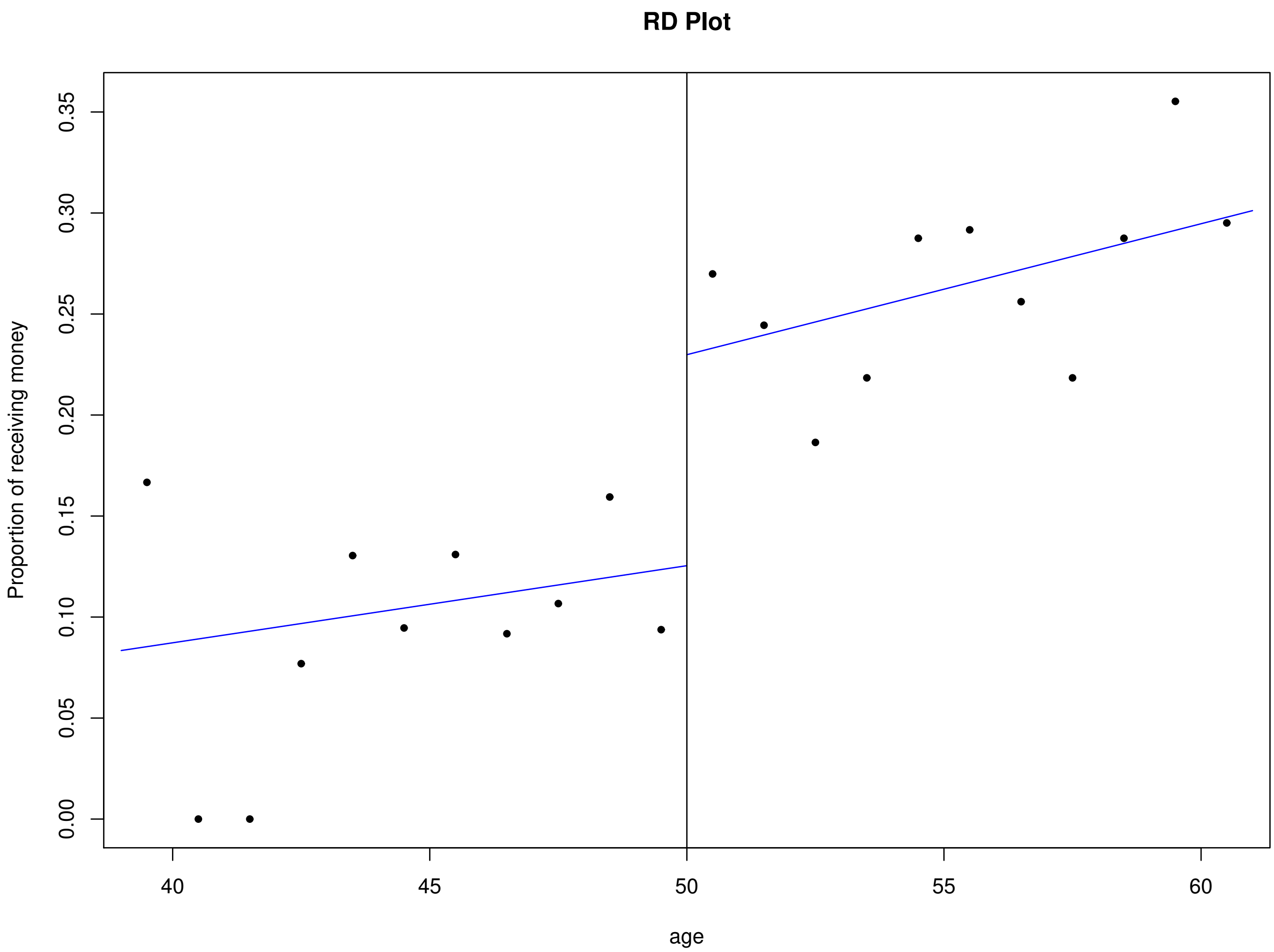


Figure 15: Probability of Receiving Any Transfer over Age

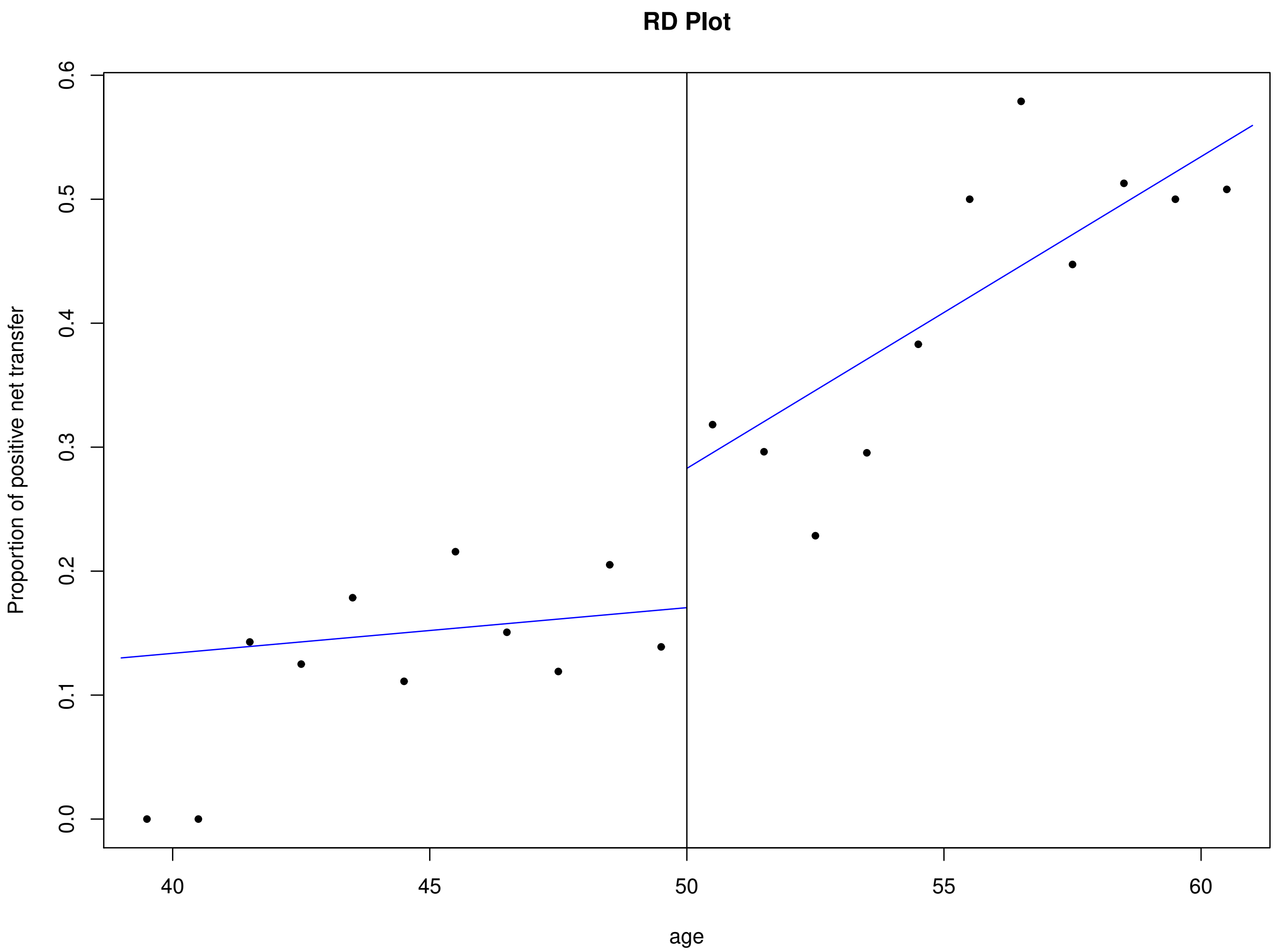


Figure 16: Probability of Net Transfer Being Positive over Age

or takers. Fig 16 shows the effect of retirement on the probability of gaining from the overall intergenerational transfer. Estimation results are reported in the third column of Table 12, suggesting that households switch roles from resource givers to resource takers by about 55 percentage point, which is significantly distinct from zero at the 10% level.

Transfers With Children Only

Indoctrinated as they are according to the old-age support tradition, how children of "sandwich" generations respond to their parents' retirement warrant further study. The estimated effects of receiving private transfers, net amount of transfers, and the occurrence of positive net transfers are presented in Table 13.

	Probability of receiving transfer	Net amount of total transfer	Probability of net transfer being positive
impact of age $\geqslant 50$ on y	0.168***	3470.410	0.124
	(0.070)	(3487.053)	(0.150)
impact of retirement on y	0.473**	9170.820	0.280
(age older than 50 as IV)	(0.227)	(10688.102)	(0.366)
observations	644	238	238

Table 13: Impacts of Retirement on Transfers with Children

The estimations are implemented by replacing the variables of equations in the Estimation section with corresponding variables of non-coresident children only. A bandwidth of 8 years is applied to the estimations due to the non-existence of non-coresident children at younger ages.

Consistent with conventional expectations, children do indeed respond to retirement of "sandwich" generations at an extensive margin, as shown in Fig 17; this shows the role of private transfers from children as an informal form of old age support. Despite the fact that the mandatory retirement age is known in advance, retirement is nonetheless a strong indication that it is time for the children to begin supporting their parents. This suggests how deeply rooted the societal expectation of children supporting their elders is in Chinese culture.

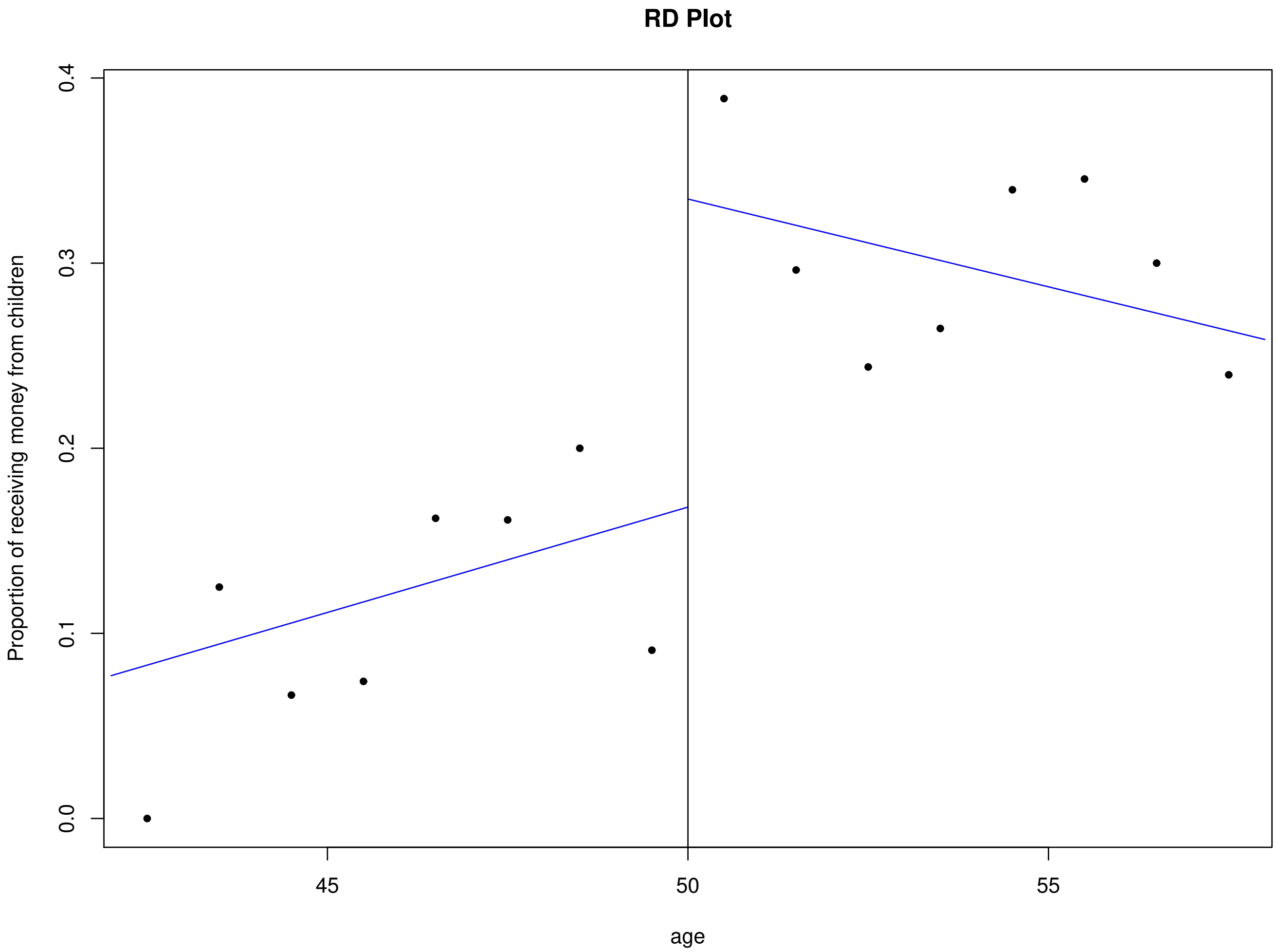


Figure 17: Probability of Receiving Transfer from Children

A possible explanation for these results, which indicate that "sandwich" generations over time switch from resource providers to resource takers, may stem from the fact that children begin offering financial resources to these "sandwich" generations upon retirement. However, in terms of the net transfer amount of children only (without taking into account the net transfer amount with their parents/ parents-in-law) conditional on those households who have financial transfer behaviors with their children, there is no discrete jump in the net amount, or the occurrence of being resource taker.

When combined, the findings from Table 12 and Table 13 suggest that households make financial transfer decisions about both flows inclusively. Overall, retirement leads to a reduction of transferring out resources, which by extension encourages "sandwich" generations to transition from providing resources to taking resources.

Robustness Checks

I conduct validity tests for the RD design by examining two features of the underlying assumptions for the RD design. First, individuals do not have precise control over the forcing variable in the neighborhood of the cutoff point. According to this context, the running variable age is unlikely to be manipulated.

Second, the mean values of the observable characteristics in relation to the outcome variable should evolve smoothly around the cutoff. In order to determine whether other variables are correlated with the jump in the probability of retirement at age 50. I estimate regressions in the Estimation section by using 3 different observable characteristics as dependent variables in order to test whether they exhibit discontinuity at the cutoff. These variables include marriage status, education level, and health status. Estimation results in columns 3 - 5 in Table 11 indicate that none of these variables jump around age 50, which is also shown in the Fig 18 - Fig 20.

Moreover, I investigate whether household income discretely decreases at retirement. I report regression with household income in log term as the dependent variable in the second column of table 11. The coefficient on the dummy variable for retirement is significant at 10% level, suggesting that the household income drops at retirement.

Policy Implication And Discussion

In this paper, I assess the life cycle nature of private transfer patterns at retirement in urban China. China itself has both a mandatory public old-age support system as well as traditional family support. Within the analytical framework of the regression discontinuity design, detailed economic transfers between generations as identified by CHARLS data offers some compelling

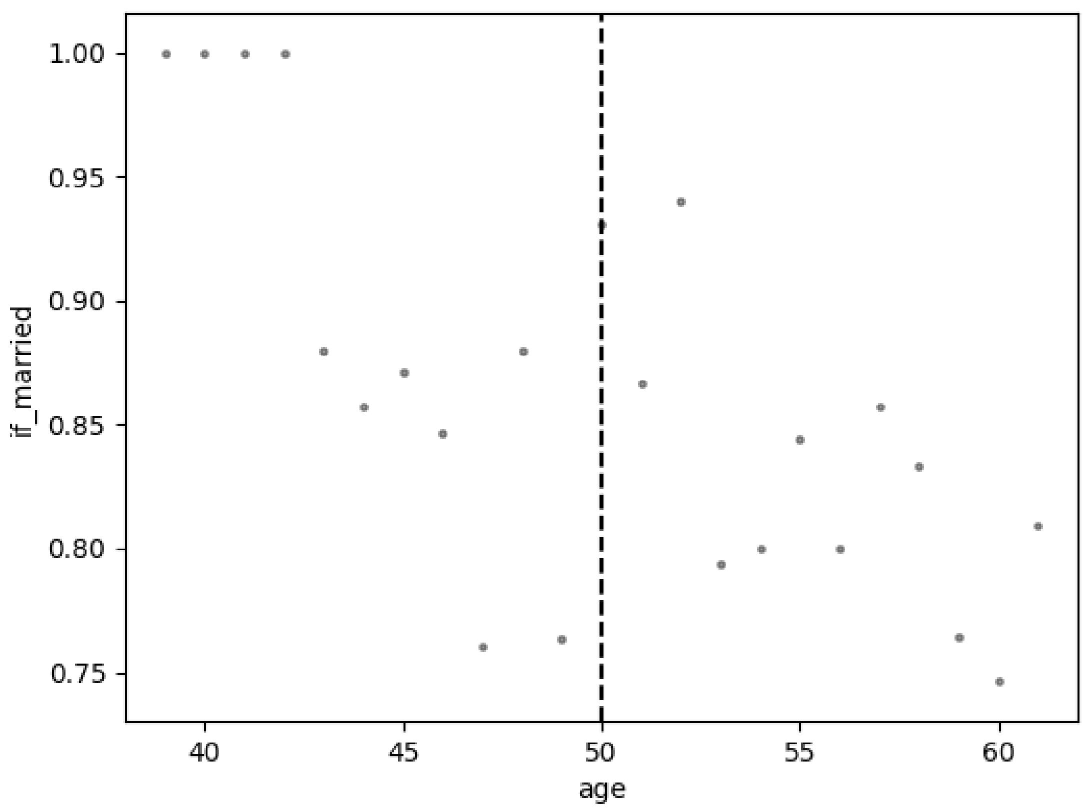


Figure 18: Proportion of Being Married over Age

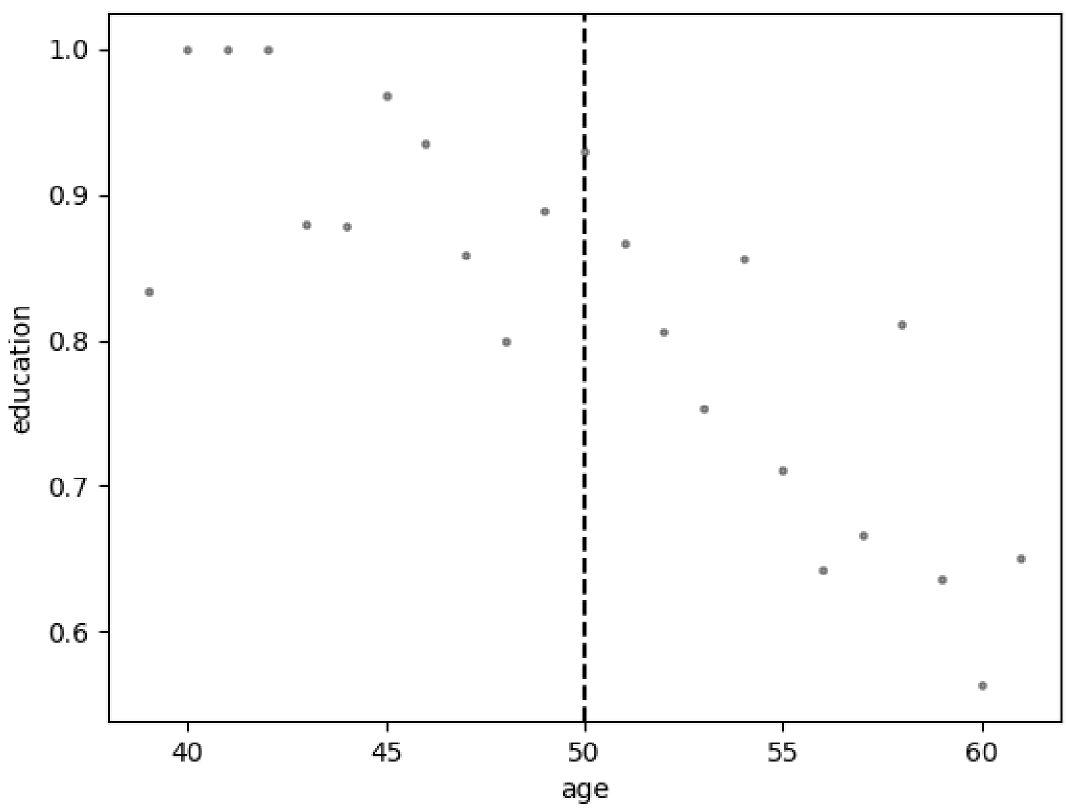


Figure 19: Proportion of the Sample Having Middle School Degree over Age

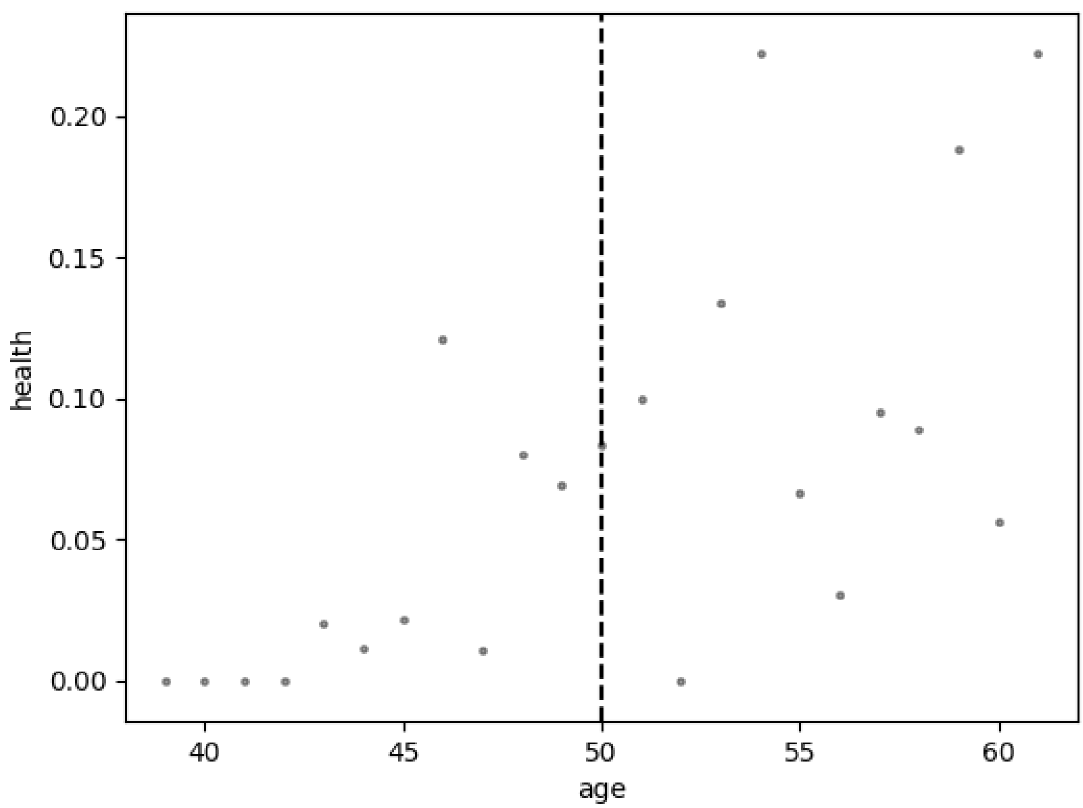


Figure 20: Health Measure over Age

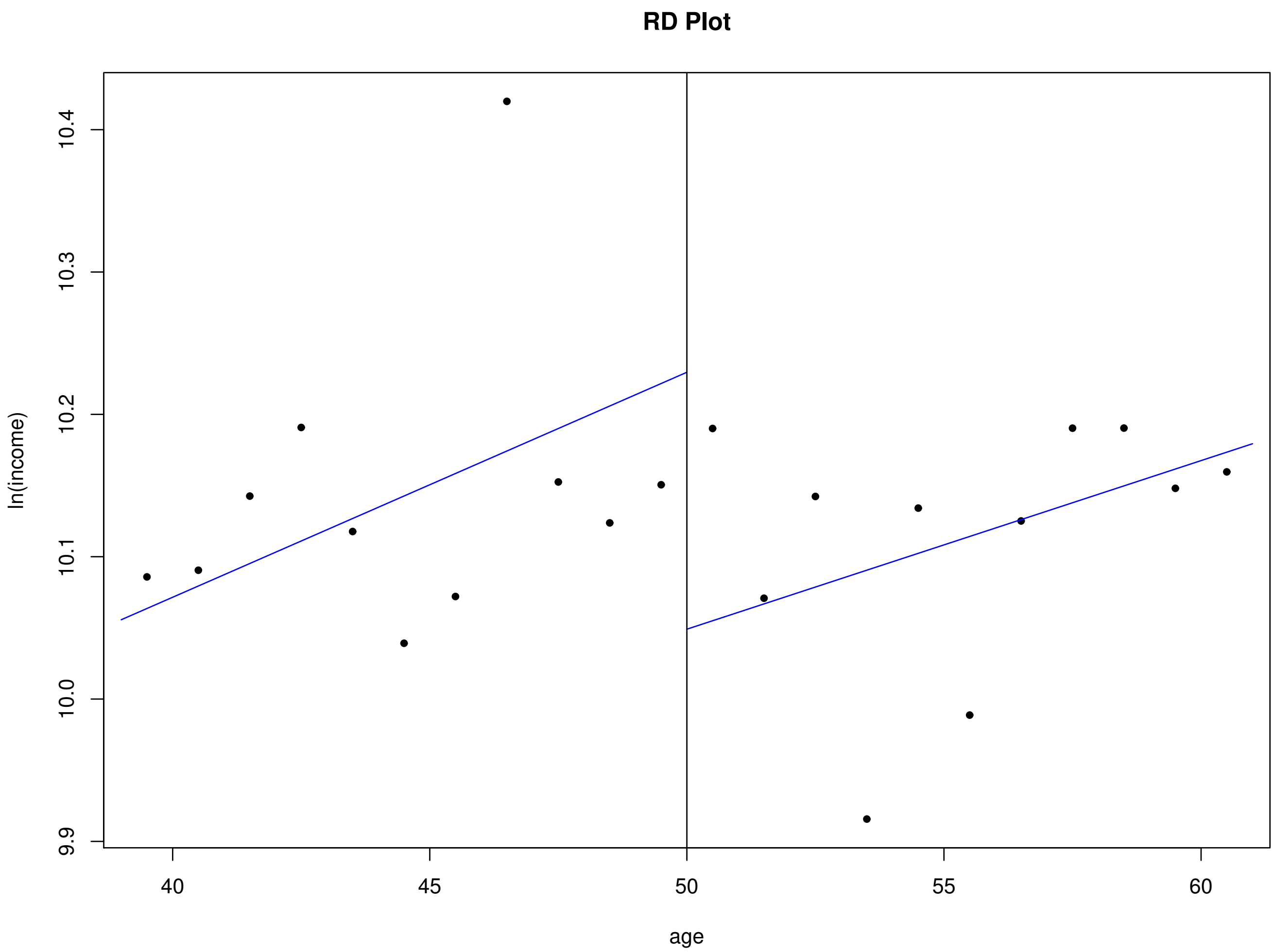


Figure 21: Log Income over Age

findings, including: Firstly, switching from working status to retirement changes the “sandwich” generation’s overall financial transfers between generations. Retirement not only changes their status in the labor market, but also encourages them to switch from a resource donor to a resource taker within the family support channel. Secondly, when the “sandwich” generation enters retirement, the likelihood that their children will be expected to start supporting their parents increases distinctly. This finding demonstrates that, when their parents retire, children in China assume the responsibility of supporting them. Given the steadily growing pension support in the public transfer system, it is clear that the concepts of traditional filial piety still remain in the younger generation.

An effective example of social modernization, the urban pension support system shares the benefits of the local market’s economic growth. Higher average wages of the current population also raise the pension benefits for the retirees. Despite this, however, filial piety is still considered a vital source pf old age support. A significant concern, then, is whether these children will continue to provide old age support despite modernization and development, as both are inextricably related to fertility decisions. This is especially concerning for developing countries who, like China, are experiencing both a decline in fertility rates and an increasing older population. Strategies like the universal two-child policy are implemented in response to declining fertility rate.

In addition, culturally driven private transfers compel younger generations to transfer money not only through the public system but also through private channels. It is worth considering how much of a role the government should play in such matters, given the respective advantages of private family support and public support. These coexistent support systems give policymakers more room to design and adjust the costs and benefits of public transfer. For example, new policies about Individual Income Tax Law were implemented in 2019, allowing adult children who support their parents to enjoy some tax waive benefits.

CHAPTER 3: QUANTILE REGRESSION WITH GENERATED REGRESSORS

Introduction

[1]Since the seminal work of Koenker and Bassett (1978), quantile regression (QR) models have provided a valuable tool in economics, finance, and statistics as a way of capturing heterogeneous effects of covariates on the outcome of interest, exposing a wide variety of forms of conditional heterogeneity under weak distributional assumptions. Also importantly, QR provides a framework for robust inference.

Applied researchers are commonly confronted with the absence of observable regressors in practice. In some cases, proxies for the unobservable variables can be found in data, while in other cases, these regressors need to be estimated. Thus, a very common strategy to deal with unobservable variables is to replace them with estimated values, that is, generated regressors (GR).[2] These GR have important implications for the reliability of general standard estimation and inference procedures. Pagan (1984) and Murphy and Topel (2002) point out that even though consistent estimates of parameters of interest are produced when the unobserved regressors are replaced with their estimated values, the conventional ways to estimate standard errors are incorrect. Recently, Mammen, Rothe, and Schienle (2012) provide a general theory for the impact of GR on the final estimator's asymptotic properties in nonparametric mean regression. Hahn and Ridder (2013) derive the asymptotic distribution of three-step estimators of a finite-dimensional parameter in a semiparametric mean regression models where GR is estimated parametrically or nonparametrically in the first-step.

Though estimation and inference for conditional mean models with GR have been widely studied and used in practice, the literature on QR with GR is more sparse. Nevertheless, a number of important statistical applications requires estimation of a conditional quantile function when some of the covariates are not directly observed, but have themselves only been estimated in a

[1]This chapter is joint work with Antonio F. Galvao, Suyong Song.

[2]Examples of generated regressors include models of interest involving expectations of future variables, such as expected prices or sales or inflation that have been generated as the predictions of some dynamic model (Engle (1982)). "Unanticipated" components of aggregate money growth in macroeconomic models (Barro (1977), Barro (1978)).

preliminary step. Examples include, among others, stochastic volatility time-series models, triangular simultaneous equation models, censoring, sample selection models, and treatment effects models. In particular, Xiao and Koenker (2009) develop QR with GR in the context of GARCH models. Chen, Dolado, and Gonzalo (2015) propose a quantile factor model. Regarding triangular simultaneous equation models with endogeneity, Lee (2007) applied a control function approach to generate instruments and resolve the endogeneity, and Ma and Koenker (2006) develop QR for recursive structural equation models. Chernozhukov, Fernandez-Val, and Kowalski (2015) develop QR with censoring and endogeneity. Arellano and Bonhomme (2016) discuss the correction of the QR estimates for nonrandom sample selection. Chernozhukov and Hansen (2005, 2006) develop a model of quantile treatment effects. Therefore, the presence of estimated regressors in different QR frameworks suggests that such strategies need to be closely examined, systematized, and formalized for a general but yet simpler and important case of linear QR.

This paper contributes to the literature by systematically studying and formalizing estimation and inference for linear QR models with a general semiparametric GR. The main contributions are as following. First, we suggest a practical two-step estimation procedure to estimate the parameters of interest in the QR-GR model. The first-step applies a general (semi)-parametric estimator to compute the GR. The second step uses the GR as regressor variables in a QR estimation. We establish the asymptotic properties of the two-step QR-GR estimator, namely, consistency and asymptotic normality. We show that the asymptotic variance-covariance matrix needs to be adjusted to account for the first-step estimation error. The basic idea is as follows. Since the first-step estimation of the unobservable regressors produces consistent estimates of the corresponding true parameters, the GR are consistently estimated. However, the GR are included in the QR model of interest with sampling error, which introduces additional noise into the asymptotic variance-covariance matrix of the coefficients of interest. In other words, the sampling error from the first stage contaminates the second stage estimation. Therefore, the usual way of calculating the QR variance-covariance matrix fails to account for the additional source of error. Under some general conditions, the estimated limiting distribution of the first-step is used to consistently estimate the variance-covariance matrix of the parameters of interest.

The second contribution of the paper is to develop inference procedures for the QR model with GR. We develop testing procedures for general linear hypotheses in these models based on Wald-type tests, and derive their associated limiting distributions. To implement the tests, we propose an estimator for the asymptotic variance-covariance of the QR-GR coefficients, and formally establish its consistency. An important advantage of the proposed tests is that they are simple to compute and implement in practice.

Compared to the existing procedures, the QR-GR methods proposed in this paper present several distinctive advantages for applied researchers. First, instead of imposing a linear structure in the first step, as in Xiao and Koenker (2009), or imposing triangular structure between two steps, as in Ma and Koenker (2006) and Lee (2007), we work with the general case of QR-GR where no structural restrictions between the two steps or any specific functional forms and estimation strategies in the first step are imposed. That is useful in applied work since practitioners have a large range of alternatives to construct the unobservable variables by using different estimation strategies or even different data sets. Second, we establish the asymptotic properties of the QR-GR estimator for non-iid data under weak conditions. This is an important generalization for practitioners since it allows for inference in a more general class of models. Third, we develop practical inference procedures. Finally, the weak conditions we imposed for QR-GR allow for simple computational implementation.[3] Linear QR models have been the workhorse of the applied research and the methods lead to a simple algorithm that can be conveniently implemented in empirical applications. Researchers can simply use existing software packages for the first-step estimation and to construct the regressors needed, for example, MLE, OLS, QR or GMM, and then apply the QR procedure with our described variance-covariance matrix adjustment.

Monte Carlo simulations assess the finite-sample properties of the proposed methods. We evaluate the QR-GR estimator in terms of empirical bias and root mean squared error, and compare its performance with methods that are not designed for dealing with GR issues. In addition, we

[3] R codes are provided for all methods, simulations, and applications.

compute the corresponding standard errors of the QR-GR estimator and evaluate its bias. The experiments suggest that the proposed approach performs very well in finite samples and effectively removes the bias of the standard errors induced by the GR. Thus, the proposed variance estimator is approximately unbiased and approximates well the true variance.

Finally, to motivate and illustrate the applicability of the methods, we consider an empirical application to study demand models, using data from the UK Family Expenditure Survey. Demand models (also known as Engel curves) represent the relationship between total expenditure and the share of various commodities to total expenditure. The QR approach is a useful tool in this example because it allows us to capture the heterogeneity in the expenditures of the different commodities along the conditional distribution of each commodity share. The first step in this exercise is to estimate the unobserved factors for budget shares of different commodities using the factor analysis proposed by Barigozzi and Moneta (2016). The unobserved factors consist of the motives of consumption as necessities, luxuries, and unitary elasticity goods. In the second step, we apply the proposed QR-GR estimator to obtain Engel curves for the commodities such as food, housing and leisure, by regressing each commodity share on the estimated factors from the first step. We found that the motives of consumption play different roles for various commodities and contributions of factors to each budget share vary over the total expenditure. Furthermore, the empirical results document important heterogeneity in the Engel curves. The estimated curves present strong heterogenous effect of the consumption motives on the budget share along the conditional distribution of the budget share in most commodities. Importantly, the empirical study underscores the importance of obtaining correct confidence intervals for the estimated Engel curves by taking into account the GR issue.

Quantile Regression With Generated Regressors

This section describes the quantile regression (QR) model with generated regressors (GR) we consider in this paper and the two-step estimation procedure.

Model

For each fixed $\tau \in (0,1)$, we consider the following model

$$y_i = \boldsymbol{x}_i^\top \beta_0(\tau) + u_i, \quad i = 1, ..., n, \tag{3.1}$$

where y_i is a response variable, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ is a k–dimensional vector of explanatory variables, $\beta_0(\tau)$ is a $k \times 1$ vector of parameters, and the innovation term u_i has conditional τ-quantile zero, that is $F_\tau^{-1}(u_i|\boldsymbol{x}_i) = 0$.

When all the regressors $\boldsymbol{x}_i$ in model (3.1) are observable, the model can be written as the following standard QR model

$$Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \beta_0(\tau), \tag{3.2}$$

where $Q_\tau(y_i|\boldsymbol{x}_i)$ is the conditional τ-quantile of y_i given $\boldsymbol{x}_i$. In general, β_0 can depend on τ. The model is semiparametric in the sense that the functional form of the conditional distribution of y_i given $\boldsymbol{x}_i$ is left unspecified.

The parameter of interest for the researcher is $\beta_0(\tau)$ in model (3.2). However, in many applications one or more elements of the vector of regressors $\boldsymbol{x}_i$ may not be directly observable, but instead estimated from a model with other given variables, that is, the GR. In this paper, we assume that some of the regressors in the vector $\boldsymbol{x}_i$ are not observable to the researcher, i.e., $(x_{i1}, \ldots, x_{iq})$ are not directly observable, but $(x_{iq+1}, \ldots, x_{ik})$ are observed, where $q \leqslant k$. In particular, the GR are assumed to have the following form

$$x_{ij} = g_j(\boldsymbol{w}_i, \boldsymbol{\theta}_j),$$

where the function $g_j(\cdot, \cdot)$ is differentiable and known up to the unknown $p_j \times 1$ parameter vector $\boldsymbol{\theta}$ for $j = 1, \cdots, q$, and the variable $\boldsymbol{w}_i = (w_{i1}, ..., w_{iq})$ is a q-dimensional vector of observables. Thus, one can still estimate the QR model in (3.2) by replacing x_{ij} with the GR $\hat{x}_{ij}$ for $j = 1, \cdots, q$.

The GR $\widehat{x}_{ij}$ is obtained from the following first-step estimation

$$\widehat{x}_{ij} = g_j(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}_j), \tag{3.3}$$

where $\widehat{\boldsymbol{\theta}}_j$ satisfies very general weak conditions as: $\sqrt{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) = n^{-1/2}\sum_{i=1}^{n} \boldsymbol{r}_i(\boldsymbol{\theta}_j) + o_p(1)$, and $\boldsymbol{r}(\cdot)$ is a generic function satisfying $E[\boldsymbol{r}_i(\boldsymbol{\theta}_j)] = 0$. The model in equation (3.3) defines the GR. Notice that most estimators used in empirical applications satisfy this first-order representation. We will discuss these conditions more formally below. To complete the model, together with (3.2) and (3.3), we assume that $F_{\tau}^{-1}(u_i|\boldsymbol{w}_{i1}, \cdots, \boldsymbol{w}_{iq}, x_{iq+1}, \cdots, x_{ik}) = 0$.

Remark 3.1. One or more elements of the regressors $\boldsymbol{x}_i$ may be estimated in the first step, but we impose no additional restrictions on how they are related to each other. Each of the GR could be related to different observable variables in different functional forms. For simplicity, we assume that each of the GR equations are estimated separately.

To illustrate how this QR-GR framework widely appears in practice, we include the following examples:

Example 3.1(a) (Two-stage regression with proxy variables)**.** . *A very important example of GR occurs when the variables $\boldsymbol{x}_i$ are not directly observable. For instance, assume that the variables $(x_{i1}, \ldots, x_{iq})$ are related to additional observable variables, $\boldsymbol{w}_{ij}$, as following*

$$\begin{aligned}
\widetilde{x}_{i1} &= x_{i1} + v_{i1} = g_1(\boldsymbol{w}_{i1}, \boldsymbol{\theta_1}) + v_{i1}, \\
&\vdots \\
\widetilde{x}_{iq} &= x_{iq} + v_{iq} = g_q(\boldsymbol{w}_{iq}, \boldsymbol{\theta_q}) + v_{iq}
\end{aligned}$$

where $(\widetilde{x}_{i1}, \ldots, \widetilde{x}_{iq})$ are proxy variables, or endogenous observables, $\boldsymbol{w}_{ij}$ is a vector of exogenous observable variables, $\boldsymbol{\theta}_j$ is a $p_j \times 1$ vector, the functions $g_j(\cdot)$ are unknown up to the vector $\boldsymbol{\theta}_j$, and v_{ji} are mutually independent innovation terms, for $j = 1, \cdots, q$.

In this case, to complete the definition of the GR one needs to impose more structure on the innovation term v_{ji} to estimate the parameters $\boldsymbol{\theta}_j$, for $j = 1, \cdots, q$. As special cases of a more general procedure which generates the regressors, consider a simple but commonly used linear model to generate one regressor, x_{i1}, as a function of several variables $\boldsymbol{w}_i$, that is, $\tilde{x}_1 = g(\boldsymbol{w}, \boldsymbol{\theta}) + v = \boldsymbol{w}^\top \boldsymbol{\theta} + v$. The following are two standard examples.

Example 3.1(b) (Conditional average)**.** *A simple example for the GR model is a linear conditional expectation. In this case, the GR model is defined as*

$$E[\tilde{x}_{i1}|\boldsymbol{w}_i] = \boldsymbol{w}_i^\top \boldsymbol{\theta}.$$

Example 3.1(c) (Conditional quantile)**.** *Another simple model for the GR is a linear conditional quantile. In this case, for a given quantile τ', the GR model is defined as*

$$Q_{\tau'}(\tilde{x}_{i1}|\boldsymbol{w}_i) = \boldsymbol{w}_i^\top \boldsymbol{\theta}(\tau'),$$

In practice one needs to estimate the parameters $\boldsymbol{\theta}$ in both examples and compute the GR.

Example 3.1(d) (Quantile regression with endogeneity: control function approach)**.** . *Consider the following model*

$$y_i = x_{i1}\beta_1 + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2 + u_i \tag{3.4}$$

where x_{i1} is the endogenous explanatory variable, β_1 is a scalar parameter for simplicity, and $\boldsymbol{z}_i$ is the $(k-1) \times 1$ vector of exogenous variables, $\boldsymbol{\beta}_2$ is a $(k-1) \times 1$ vector of parameters. The endogenous explanatory variable is modeled as

$$x_{i1} = g_1(\boldsymbol{w}_{i1}, \boldsymbol{\theta}_1) + v_{1i} \tag{3.5}$$

where x_{i1} is related to some exogenous variable $\boldsymbol{w}_{i1}$ which contains at least one element not in $\boldsymbol{z}_i$. The endogeneity of x_{i1} arises because v_{1i} is correlated with u_i in (3.4). Write the linear projection *of* u_i *on* v_{1i} *as*

$$u_i = \gamma v_{1i} + \varepsilon_i. \tag{3.6}$$

Plugging (3.6) into equation (3.4) gives

$$y_i = x_{i1}\beta_1 + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2 + \gamma v_{1i} + \varepsilon_i,$$

where v_{1i} *is treated as an explanatory variable. Notice* ε_i *is uncorrelated with* v_{i1} *and* $\boldsymbol{z}_i$*. Since* x_{i1} *is a function of* $\boldsymbol{w}_{i1}$ *and* v_{1i}*,* ε_i *is also uncorrelated with* x_{i1}*. Since we do not observe* v_{1i}*, we can replace* v_{1i} *with* $\widehat{v}_{1i} \equiv x_{i1} - g_1(\boldsymbol{w}_{i1}, \widehat{\boldsymbol{\theta}}_1)$*, which gives*

$$y_i = x_{i1}\beta_1 + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2 + \gamma \widehat{v}_{1i} + \eta_i$$

where $\eta_i = \varepsilon_i + \gamma[g_1(\boldsymbol{w}_{i1}, \boldsymbol{\theta}_1) - g_1(\boldsymbol{w}_{i1}, \widehat{\boldsymbol{\theta}}_1)]$ *which depends on the sampling error in* $\widehat{\boldsymbol{\theta}}_1$*. Use the consistency result for QR-GR in the Asymptotic Properties Section, we can consistently estimate* $\beta_1, \boldsymbol{\beta}_2$ *and* γ*.*

Example 3.1(e) (Quantile regression with endogeneity: instrumental variable approach)**.** . *Consider model (3.4) and (3.5), apply the 'fitted value' approach to (3.4) and replace* x_{i1} *with the fitted value of* $g_1(\boldsymbol{w}_{i1}, \boldsymbol{\theta}_1)$ *in (3.5). As shown in*

$$y_i = [g_1(\boldsymbol{w}_{i1}, \boldsymbol{\theta}_1)]\beta_1 + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2 + \varepsilon_i,$$

where $\varepsilon_i = u_i + \beta_1 v_{1i}$*, estimating* β_1 *and* $\boldsymbol{\beta}_2$ *consistently only requires that* $Q_\tau(\varepsilon_i | \boldsymbol{w}_{i1}, \boldsymbol{z}_i)$ *is independent of* $(\boldsymbol{w}_{i1}, \boldsymbol{z}_i)$*, which is easy to satisfy since both* u_i *and* v_{1i} *are independent of* $\boldsymbol{w}_{i1}$ *and* $\boldsymbol{z}_i$*. In practice,* $g_1(\boldsymbol{w}_{i1}, \boldsymbol{\theta}_1)$ *needs to be replaced with* $g_1(\boldsymbol{w}_{i1}, \widehat{\boldsymbol{\theta}}_1)$*.*

Estimation

The estimation of the parameters of interest in model (3.2) involves a two-step estimation. In the first step one estimates and computes the GR from model (3.3). In the second step one uses the GR and other regressors, and computes the QR of interest from (3.2).

Estimation of the unobservables usually comes from the same sample of data – may come from a different dataset or even from the parameter estimates by another researcher. Models used to estimate the unknown parameter $\boldsymbol{\theta}_j$ may generally include linear or nonlinear models. Also, the parameters can be estimated by various strategies. The QR-GR two-step estimation procedure is as following:

Step 1 Estimate $\boldsymbol{\theta}_j$ from (3.3) and compute the fitted values $\widehat{x}_{ij} = g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)$ for $j = 1, \cdots, q$, and then obtain the generated regressors $\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, \cdots, \widehat{x}_{iq}, x_{iq+1}, \cdots, x_{ik})^\top$ for $i = 1, \cdots, n$.

Step 2 Compute $\boldsymbol{\beta}$ from the following QR

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}), \tag{3.7}$$

where $\rho_\tau(u) := \{\tau - I(u \leqslant 0)\}u$ is the check function as in Koenker and Bassett (1978).

Thus, one uses the estimates of $\boldsymbol{\theta}_j$, denoted by $\widehat{\boldsymbol{\theta}}_j$, in the first step, to obtain the GR $\widehat{\boldsymbol{x}}_i$. In the examples for the average and quantile models discussed previously we have the following for the first step.

Example 3.1(f) (Average continued)**.** . *In this example, one employs the standard OLS estimator and obtains*

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{w}^\top \boldsymbol{w})^{-1} \boldsymbol{w}^\top \widetilde{x}_1$$

and computes $\widehat{x}_{i1} = \boldsymbol{w}^\top \widehat{\boldsymbol{\theta}}$*, and also* $\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, x_{i2}, ..., x_{ik})^\top$*. Then, the* τ*-th QR estimator* $\widehat{\boldsymbol{\beta}}(\tau)$ *can be obtained by* (3.7)*.*

Example 3.1(g) (Quantile continued)**.** . *In this case, for a given quantile* τ' *one applies the usual QR procedure to estimate* $\boldsymbol{\theta}(\tau')$ *and obtain* $\widehat{x}_{i1} = \boldsymbol{w}_i^\top \widehat{\boldsymbol{\theta}}(\tau')$*. Thus, the first-step estimation is given* *by the following QR*

$$\widehat{\boldsymbol{\theta}}(\tau') = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(\widetilde{x}_{i1} - \boldsymbol{w}_i^\top \boldsymbol{\theta}).$$

Then, in the second step, for a given τ-quantile, the QR estimator $\widehat{\boldsymbol{\beta}}(\tau)$ can be obtained by (3.7).

These two practically-common cases illustrate the simple implementation of QR model with GR. We have made `R` routines for the QR-GR estimator and inference in the QR-GR framework available for the practitioners.

Asymptotic Properties

We now establish consistency and asymptotic normality of the QR-GR two-step estimator, $\widehat{\boldsymbol{\beta}}(\tau)$, defined in the previous section. Proofs are collected in the Appendix. We consider the following regularity conditions:

A1. $\{(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)\}_{i=1}^n$ is independent across i. The conditional distribution functions of the error term $\{F_i(u|\boldsymbol{x}_i)\}$ have continuous densities $f_i(u|\boldsymbol{x}_i)$ with a unique conditional τ-th quantile equal to 0, and $f_i(0|\boldsymbol{x}_i)$ are uniformly bounded away from 0 and ∞.

A2. There exist positive definite matrices D_0 and D_1 such that

(i) $D_0 = \text{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top,$

(ii) $D_1 = \text{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} f_i(0|\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top,$

(iii) $\max_{i=1,\ldots,n} \|\boldsymbol{x}_i\|/\sqrt{n} \xrightarrow{p} 0.$

A3. $\text{plim}_{n\to\infty} \widehat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j$, where each $\widehat{\boldsymbol{\theta}}_j$ is estimated individually for $j = 1, \cdots, q$.

A4. $\sqrt{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) = n^{-1/2} \sum_{i=1}^n \boldsymbol{r}_i(\boldsymbol{\theta}_j) + o_p(1)$ where $\boldsymbol{r}_i(\cdot)$ is a continuous function which satisfies that $E[\boldsymbol{r}_i(\boldsymbol{\theta}_j)] = 0$ and $Var[\boldsymbol{r}_i(\boldsymbol{\theta}_j)] = V_j$, for $j = 1, \cdots, q$.

A5. There exist nonsingular and positive definite matrices D_{12}^j and M such that

(i) $D_{12}^j = \text{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f_i(0|\boldsymbol{x}_i) \beta_j(\tau) \boldsymbol{x}_i \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top$, where $\nabla_{\theta_j} g_j(\boldsymbol{w}_j, \boldsymbol{\theta}_j)^\top$ is $1 \times p_j$ Jacobian of $g_j(\boldsymbol{w}_j, \boldsymbol{\theta}_j)$ for $j = 1, \cdots, q$, (*ii*) $M = \operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\{\psi_\tau(y_i - \boldsymbol{x}_i^\top \beta_0(\tau)) f_i(0|\boldsymbol{x}_i) \sum_{j=1}^{q} \beta_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\}$, where $\psi_\tau(u) := \tau - I(u \leqslant 0)$.

Conditions **A1** and **A2** are the usual conditions in the QR literature. Assumptions **A1** allows for non-iid sampling, and **A2** requires limiting matrices to be well defined. Assumptions **A3–A5** refer to the GR estimation in the first step and need to be verified for each empirical application. These conditions are very mild. Assumptions **A3** and **A4** impose consistency and asymptotic normality, respectively, for the first-step estimator of the GR. These conditions hold for most estimators employed in empirical work in the family of M- and Z-estimators. We note that no restrictions are imposed on the functional form of $g_j(\cdot)$ except condition **A5** which is a weak smoothness condition needed only for nonlinear models.

The following result states the consistency of the QR-GR estimator.

Theorem 3.1 (Consistency)**.** *Consider the model in* (3.2) *and* (3.3)*. Under the conditions* ***A1–A3*** *and* ***A5****, as* $n \to \infty$

$$\widehat{\beta}(\tau) \xrightarrow{p} \beta_0(\tau).$$

Remark 3.2. A consistent estimate of the unknown parameter $\boldsymbol{\theta}_j$ for $j = 1,\cdots,q$ in the first step suffices for the consistency of QR with GR. In other words, replacing $\boldsymbol{x}$ by $\widehat{\boldsymbol{x}}$ in a quantile regression still gives us a consistent QR estimator.

The next result establishes the asymptotic normality of the QR-GR two-step estimator.

Theorem 3.2 (Asymptotic normality)**.** *Consider the model in* (3.2) *and* (3.3)*. Under conditions* ***A1–A5****, as* $n \to \infty$*, we have that,*

$$\sqrt{n}(\widehat{\beta}(\tau) - \beta_0(\tau)) \xrightarrow{d} N(0, \Omega(\tau)),$$

where $\Omega(\tau) := \tau(1-\tau)D_1^{-1}D_0D_1^{-1} + \sum_{j=1}^{q}\left[D_1^{-1}D_{12}^{j}V_jD_{12}^{j\top}D_1^{-1}\right] - 2D_1^{-1}MD_1^{-1}$.

The result in Theorem 3.2 shows that the limiting distribution of $\sqrt{n}(\widehat{\beta}(\tau) - \beta_0(\tau))$ generally depends on the statistical properties of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)$ since the sampling error of generating regressor in model (3.3) contaminates the variance-covariance estimation in the second stage. Hence, the asymptotic variance-covariance of the two-step estimator needs to be adjusted when the regressors are generated. The adjusted variance-covariance matrix is larger, since the sampling error V_j from the first-step estimation enters variance-covariance matrix. Nevertheless, the sampling variation of $\hat{\boldsymbol{\theta}}_j$ can be ignored, at least asymptotically, if the coefficients β_j of the GR are zero, i.e. $D_{12}^j = 0$ and the corresponding part in M becomes zero. In that case, replacing the true regressors with GR will not impact the asymptotic distribution of the estimator.

Remark 3.3. In general, the consistency of QR estimator with GR holds when the true regressor is replaced by GR. However, the asymptotic variance-covariance matrix of the estimator $\hat{\beta}(\tau)$ needs to be adjusted because of the sampling variation introduced by estimation of $\hat{\boldsymbol{\theta}}_j$ in the first step.

Remark 3.4. Comparing the QR-GR estimator $\hat{\beta}(\tau)$ with regular QR estimator with the true unobserved regressors, we see that the additional second term appears in the variance-covariance matrix coming from the first-step estimation. The first-step estimation contaminates the variance-covariance matrix of $\hat{\beta}(\tau)$ in two ways: one is the sampling error V_j of coefficient estimates $\boldsymbol{\theta}_j$ in the first stage, the other is the gradient of model specification with respect to the parameter in the first stage. For the simple linear model in the first stage, the gradient is simply the regressors in the first stage. But for nonlinear model in the first stage, the coefficient estimates show up in D_{12}^j which makes the variance-covariance matrix of $\hat{\beta}(\tau)$ larger than when the true regressors are used for estimation.

The two-step estimation procedure in this paper is easy to implement in practice. First, since weak conditions are needed for QR-GR procedure, different estimation strategies may be used in the first step to construct the GR. Most common estimators in practice satisfy the weak conditions: for example, the simple OLS, QR, MLE methods, etc. Second, both linear and nonlinear model specifications in the first step are allowed. Finally, it is important to notice that weak conditions in the QR step include non-iid models which allow practitioners to proceed inference in a general class of models.

To further illustrate the estimation of QR models with GR, we further discuss the two common models, OLS and QR with GR. We discuss the case where only one regressor is generated in the first step. For both models, after estimating $\widehat{\boldsymbol{\theta}}$ which satisfies the condition $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n} r_i(\boldsymbol{\theta}) + o_p(1)$ where $E[r_i(\boldsymbol{\theta})] = 0$ and $Var[r_i(\boldsymbol{\theta})] = V$ in the first step, one obtains the GR: $\widehat{\boldsymbol{x}} = (\widehat{x}_1, x_2, \cdots, x_k)^\top$ where $\widehat{x}_1 = g(\boldsymbol{w}, \widehat{\boldsymbol{\theta}}) = \boldsymbol{w}^\top \widehat{\boldsymbol{\theta}}$. The following examples derive the asymptotic variance-covariance matrix for both OLS and QR with GR: **Example 3.1** (OLS with GR). *In this example, $\widehat{\boldsymbol{\beta}}$ is estimated from the model*

$$E(y_i|\widehat{\boldsymbol{x}}_i) = \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}.$$

*Since the OLS estimator has a closed-form solution, one obtains $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i y_i\right)$. Finally, the variance-covariance matrix is given by the following result which is proved in the Appendix, for completeness. Note that according to **Proposition 2** in the Appendix, under conditions **C1–C5**,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N(0, H^{-1} Var[\boldsymbol{x}_i u_i - G r_i(\boldsymbol{\theta})] H^{-1}),$$

where $H = \text{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i \boldsymbol{x}_i^\top)$, and $G = \text{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta} \otimes \boldsymbol{x}_i^\top)^\top([1\ 0\ \ldots\ 0]^\top \otimes w_i^\top)$. When there is no GR or the coefficients of the GR are zero, the asymptotic variance displayed above simplifies since $G = 0$. In the case of $Var(u|\boldsymbol{X}) = \sigma_u^2 I_n$, where $\sigma_u^2 > 0$ and I_n is an $n \times n$ identity matrix, we obtain that $\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = \sigma_u^2 H^{-1}$, since $E(\boldsymbol{x}_i u_i^2 \boldsymbol{x}_i^\top) = \sigma_u^2 H$. In addition, in the case of $Var(u|\boldsymbol{X}) = \Gamma_0$ and $\text{plim}_{n\to\infty} n^{-1}(X^\top \Gamma_0 X) = \Gamma_X$, where Γ_X is positive definite matrix, since $E(\boldsymbol{x}_i u_i^2 \boldsymbol{x}_i^\top) = \Gamma_X$, we have that $\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = H^{-1}\Gamma_X H^{-1}$.

Example 3.1(h) (QR with GR). . *In this example, $\widehat{\boldsymbol{\beta}}(\tau)$ is estimated from the model*

$$Q_\tau(y_i|\widehat{\boldsymbol{x}}_i) = \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}(\tau).$$

Unlike the simple closed-form solution in the above OLS estimate, one applies the usual QR pro-

cedure to estimate

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}).$$

Theorem 3.2 shows that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)) \xrightarrow{d} N(0, D_1^{-1}\{Var[\boldsymbol{x}_i \psi_\tau(y_i - \xi_i(\tau))] + Var[D_{12}\boldsymbol{r}_i(\boldsymbol{\theta})] - 2M\}D_1^{-1}),$$

where D_0, D_1, and D_{12} are defined in conditions ***A2*** *and* ***A5****, and M is simplied to $M = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\{\psi_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)) f_i(0|\boldsymbol{x}_i)\beta_1(\tau)\boldsymbol{w}_i^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\}$. In addition, one can notice that the above asymptotic variance-covariance matrix can be simplified to*

$$D_1^{-1}\left(\tau(1-\tau)D_0 + D_{12}VD_{12}^\top - 2M\right)D_1^{-1}.$$

To conclude this section we have two remarks. First, from the second example, one can notice that the variance for the QR with GR has the additional terms $D_1^{-1}D_{12}VD_{12}^\top D_1^{-1}$ and $-2D_1^{-1}MD_1^{-1}$ relative to the standard QR model without GR, which has the variance $\tau(1-\tau)D_1^{-1}D_0D_1^{-1}$. Second, intuitively, as shown in Theorem 3.1 and Proposition 1, in both OLS and QR frameworks, estimation with GR still produces consistent estimates. However, as shown in Theorem 3.2 and Proposition 2, the corresponding variance-covariance matrices need to account for two sources of error – the usual estimation error in the OLS or QR method, and the sampling error in generating the regressors. As a special case, the variance-covariance matrix with GR reduces to the usual variance-covariance matrix in both OLS and QR when the coefficient of the GR, θ_j for $j = 1, \cdots, q$, are zero, i.e. both G in the OLS case and D_{12}^j and M in the QR case vanish.

Inference

In this section, we turn our attention to inference in the QR-GR model. First, we suggest an estimator for the asymptotic variance-covariance matrix of the QR-GR estimator. Second, we propose a Wald-type test for general linear hypotheses.

Variance-covaraince Matrix Estimation

In applications, the variance-covariance matrices are unknown and need to be estimated. Now we suggest an estimator for the corresponding variance of the QR-GR estimator. The estimator is closely related to those suggested by Hendricks and Koenker (1991) and Powell (1991) and given as the following form

$$\widehat{\Omega}(\tau) = \tau(1-\tau)\widehat{D}_1^{-1}\widehat{D}_0\widehat{D}_1^{-1} + \sum_{j=1}^{q}[\widehat{D}_1^{-1}\widehat{D}_{12}^{j}\widehat{V}_j\widehat{D}_{12}^{j\top}\widehat{D}_1^{-1}] - 2\widehat{D}_1^{-1}\widehat{M}\widehat{D}_1^{-1} \tag{3.8}$$

where

$$\widehat{D}_0 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top} \tag{3.9}$$

$$\widehat{D}_1 = \frac{1}{2nc_n}\sum_{i=1}^{n}I(|y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\boldsymbol{\beta}}(\tau)| < c_n)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top} \tag{3.10}$$

$$\widehat{D}_{12}^{j} = \frac{1}{2nc_n}\sum_{i=1}^{n}I(|y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\boldsymbol{\beta}}(\tau)| < c_n)\widehat{\beta}_j(\tau)\widehat{\boldsymbol{x}}_i\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^{\top} \tag{3.11}$$

$$\widehat{M} = \frac{1}{2nc_n}\sum_{i=1}^{n}\{\psi_\tau(y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\beta}(\tau))I(|y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\beta}(\tau)| < c_n)\cdot\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji}\widehat{\boldsymbol{\theta}}_j)^{\top}\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top}, \tag{3.12}$$

To establish the consistency of $\widehat{\Omega}(\tau)$ we impose the following assumptions:

B1. There exists a positive sequence of bandwidths $\{c_n\}$ such that $c_n \to 0$ and $\sqrt{n}c_n \to \infty$.

B2. $E(\|\boldsymbol{x}_i\|^4) \leqslant H_1 < \infty$ for all i and for some constant H_1.

B3. $||\nabla_{\theta_j}g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)|| \leqslant D_j(\boldsymbol{w}_j)$ for all $\boldsymbol{\theta}_j \in \Theta_j$ where $\max_j E[D_j(\boldsymbol{w}_j)^2] \leqslant D_0 < \infty$ and $||\nabla_{\theta_j}g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j) - \nabla_{\gamma_j}g_j(\boldsymbol{w}_j,\boldsymbol{\gamma}_j)|| \leqslant J_j||\boldsymbol{\theta}_j - \boldsymbol{\gamma}_j||$ for some constant $J_j < \infty$.

B4. There exists a function $A_f(\boldsymbol{x}_i)$ such that $f_i(\lambda|\boldsymbol{x}_i) \leqslant A_f(\boldsymbol{x}_i)$ for all i and for some λ near zero a.s.[4]

[4]We note that $\boldsymbol{x}_i = (x_{i1},\cdots,x_{iq},x_{iq+1},\cdots,x_{ik})$ where $x_{ij} = g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)$ for $j = 1,\cdots,q$. And we also note that for any $\boldsymbol{\theta}_j \in \Theta_j$ and for all i, $E[\{\|\boldsymbol{x}_i\|^3 +$

B5. $f_i(\lambda|\boldsymbol{x}_i)$ satisfies the Lipschitz condition that $|f_i(\lambda_1|\boldsymbol{x}_i) - f_i(\lambda_2|\boldsymbol{x}_i)| \leqslant L_i|\lambda_1 - \lambda_2|$ for some constant $L_i < \infty$.

Assumption **B1** is a restriction on the bandwidth c_n, which is commonly used to estimate unknown conditional densities. **B2** imposes some moment condition, which is typical in the QR literature. Assumption **B3** imposes smoothness and dominance conditions that are typical for nonlinear models (see, e.g., Powell (1991)). **B4** imposes some local restriction on the conditional density function near zero and satisfies some moment bounds, which can be weakened at the cost of moment bounds for various cross products of the bounding functions. (see, e.g., Assumption ER of Buchinsky and Hahn (1998)). **B5** imposes smoothness condition on the conditional densities and is standard in the QR literature.

The next result states the consistency of the variance-covariance QR-GR estimator.

Theorem 3.3 (Variance-covariance matrix estimation)**.** *Under the assumptions **B1–B5** and conditions of Theorem 3.2, as $n \to \infty$*

$$\widehat{\Omega}(\tau) \xrightarrow{p} \Omega(\tau).$$

*In other words, $\widehat{D}_0 \xrightarrow{p} D_0$, $\widehat{D}_1 \xrightarrow{p} D_1$, $\widehat{D}_{12}^j \xrightarrow{p} D_{12}^j$ and $\widehat{M} \xrightarrow{p} M$, for $j = 1, \cdots, q$ where D_0, D_1 and D_{12}^j are defined in conditions **A2** and **A5**, and M is defined in **Theorem 3.2**.*

As a special case of more general procedure which generates the regressors, consider a linear model $\tilde{x}_1 = g(\boldsymbol{w}, \boldsymbol{\theta}) + v = \boldsymbol{w}^\top \boldsymbol{\theta} + v$. The gradient of $g(\boldsymbol{w}, \boldsymbol{\theta})$, denoted by $\nabla_{\theta} g(\boldsymbol{w}, \boldsymbol{\theta})$, is reduced to $\boldsymbol{w}$, in the following examples for the average and quantile models:

Example 3.1(i) (Average continued)**.** . *In this example, one employs the standard OLS estimator in the first step to obtain*

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{w}^\top \boldsymbol{w})^{-1} \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_1,$$

and the standard error $\widehat{V}$. Then, one is able to compute $\widehat{x}_{1i} = \boldsymbol{w}_i^\top \widehat{\boldsymbol{\theta}}$, as well as $\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, x_{i2}, \ldots, x_{ik})^\top$.

$\max_j D_j(\boldsymbol{w}_j)\|\boldsymbol{x}_i\|^2\} A_f(\boldsymbol{x}_i)] < A < \infty$ and $E[D_j(\boldsymbol{w}_j)\{\|\boldsymbol{x}_i\|^2 + \max_{j'} D_{j'}(\boldsymbol{w}_{j'})\|\boldsymbol{x}_i\|\} A_f(\boldsymbol{x}_i)] < B_j < \infty$. Also, $E\left[\sum_{j=1}^{q} \|\widehat{\beta}_j(\tau) \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j) \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top\| \cdot [\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)] A_f(\boldsymbol{x}_i)\right]$ is bounded by applying **A2**, **A4** and **A5**.

Finally, with the second step estimation $\widehat{\boldsymbol{\beta}}$, the estimator of the corresponding variance-covariance matrix is simply $\widehat{\Omega}(\tau) := \tau(1-\tau)\widehat{D}_1^{-1}\widehat{D}_0\widehat{D}_1^{-1} + \widehat{D}_1^{-1}\widehat{D}_{12}\widehat{V}\widehat{D}_{12}^{\top}\widehat{D}_1^{-1} - 2\widehat{D}_1^{-1}\widehat{M}\widehat{D}_1^{-1}$.

$$\widehat{D}_{12} = (2nc_n)^{-1}\sum_{i=1}^{n} I(|y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\boldsymbol{\beta}}(\tau)| < c_n)\widehat{\beta}_1(\tau)\widehat{\boldsymbol{x}}_i\boldsymbol{w}_i^{\top}$$

$$\widehat{D}_0 = n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top}$$

$$\widehat{D}_1 = (2nc_n)^{-1}\sum_{i=1}^{n} I(|y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\boldsymbol{\beta}}(\tau)| < c_n)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top}$$

$$\widehat{M} = (2nc_n)^{-1}\sum_{i=1}^{n}\Big\{[\tau - I(y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\boldsymbol{\beta}}(\tau) < 0)]I(|y_i - \widehat{\boldsymbol{x}}_i^{\top}\widehat{\boldsymbol{\beta}}(\tau)| < c_n)\widehat{\beta}_1(\tau)\boldsymbol{w}_i^{\top}(\boldsymbol{w}^{\top}\boldsymbol{w})^{-1}\boldsymbol{w}^{\top}\widehat{\boldsymbol{v}}\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top}\Big\}.$$

with $\widehat{\boldsymbol{v}}$ being the residual from the first-step estimation.

Example 3.1(j) (Quantile continued)**.** . *In this example, for a given quantile τ' one applies the usual QR procedure to estimate $\boldsymbol{\theta}(\tau')$ and obtains $\widehat{x}_1$. Thus, the first-step estimation is given by the following QR*

$$\widehat{\boldsymbol{\theta}}(\tau') = \underset{\theta}{\text{argmin}}\sum_{i=1}^{n}\rho_\tau(\widetilde{x}_{i1} - \boldsymbol{w}_i^{\top}\boldsymbol{\theta}).$$

So one obtains the standard error $\widehat{V}$ and computes $\widehat{x}_{1i} = \boldsymbol{w}_i^{\top}\widehat{\boldsymbol{\theta}}(\tau')$ to obtain $\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, x_{i2}, \ldots, x_{ik})^{\top}$. The final formula for the estimator of the variance covariance matrix is analogous to the previous example.

Testing

In the independent and identically distributed setting, the conditional quantile functions of the response variable, given the covariates, are all parallel, implying that covariate effects shift the location of the response distribution but do not change the scale or shape. However, slope estimates often vary across quantiles, implying that it is important to test for equality of slopes across quantiles. Wald tests designed for this purpose were suggested by Koenker and Bassett (1982a), Koenker and Bassett (1982b), and Koenker and Machado (1999). It is possible to formulate a wide variety of tests using variants of the proposed Wald test, from simple tests on a single quantile regression coefficient to joint tests involving many covariates and distinct quantiles at the same time.

General hypotheses on the vector $\boldsymbol{\beta}(\tau)$ can be accommodated by Wald-type tests. The Wald process and associated limiting theory provide a natural foundation for the hypothesis

$$H_0 : R\boldsymbol{\beta}(\tau) = r,$$

where R is a full-rank matrix imposing s number of restrictions on the parameters and r is a column vector of s elements. We consider a Wald-type test where we test the coefficients for selected quantiles of interest. For simplicity, we use the model stated in equation (3.2) with a single variable in the $\boldsymbol{x}$ matrix. The following example is a hypotheses that may be considered in the former framework.

Example 3.2 (Location shifts)**.** *The hypotheses of location shifts for $\boldsymbol{\beta}(\tau)$ can be accommodated in the model. For instance, $H_0 : \beta_1(\tau) = \beta_1$, so $R = [1\ 0\ \ldots\ 0]$ and $r = \beta_1$.*

In general, for given τ, the regression Wald process can be constructed as

$$\mathscr{W}_n = n(R\widehat{\boldsymbol{\beta}}(\tau) - r)^\top [R\widehat{\Omega}(\tau)R^\top]^{-1}(R\widehat{\boldsymbol{\beta}}(\tau) - r), \tag{3.13}$$

where $\widehat{\Omega}(\tau) = \tau(1-\tau)\widehat{D}_1^{-1}\widehat{D}_0\widehat{D}_1^{-1} + \sum_{j=1}^{q}[\widehat{D}_1^{-1}\widehat{D}_{12}^{j}\widehat{V}_j\widehat{D}_{12}^{j\top}\widehat{D}_1^{-1}] - 2\widehat{D}_1^{-1}\widehat{M}\widehat{D}_1^{-1}$.

In order to implement the test it is necessary to estimate $\Omega(\tau)$ consistently. It is possible to obtain such an estimator as suggested in Theorem 3.3 in the previous section, and the main components of $\widehat{\Omega}(\tau)$ can be obtained as in equations (3.9)–(3.11).

Given the results on consistency of $\widehat{\Omega}(\tau)$, if we are interested in testing $R\boldsymbol{\beta}(\tau) = r$ at a particular quantile $\tau = \tau_0$, a Chi-square test can be conducted based on the statistic $\mathscr{W}_n(\tau_0)$. Under H_0, the statistic $\mathscr{W}_n$ is asymptotically χ_s^2 with s-degrees of freedom, where s is the rank of the matrix R. The limiting distribution of the test is summarized in the following theorem.

Theorem 3.4 (Wald Test Inference)**.** *Under $H_0 : R\boldsymbol{\beta}(\tau) = r$, and conditions **A1-A5** and **B1–B5**,* *for fixed* τ,

$$\mathscr{W}_n(\tau) \overset{a}{\sim} \chi_s^2.$$

Monte Carlo Simulations

In this section, we evaluate the performance of two-step QR-GR estimator and compare its performance with the usual QR estimator which does not account for the first-step estimation. Also, the performance of the proposed variance-covariance estimator is evaluated. The computational results are obtained in the `R` language.

Monte Carlo Design

We consider the following model as a data generating process:

$$y_i = \beta_0 + \beta_1 x_i^* + (1 + \gamma x_i^*)\varepsilon_i,$$

where $\varepsilon \sim N(0,1)$, and β_0 and β_1 are the parameters of interest. We set $(\beta_0, \beta_1) = (4,3)$. The parameter γ captures the heterogeneity, hence we let $\gamma = \{0,1\}$. When $\gamma = 0$ we have a location-shift model and for $\gamma = 1$ the location-scale-shift. Thus, for the later model, we have that $\beta_1(\tau) = \beta_1 + \gamma F_\varepsilon^{-1}(\tau)$.

The regressor x^* is unobserved but its observable counterpart x is related to observed variables w and z as follows:

$$x_i = x_i^* + \nu_i = \theta_0 + \theta_1 w_i + \theta_2 z_i + \nu_i,$$

where w, z, ν are generated as the following: $z \sim t_5$, $w \sim N(10,5)$, $\nu \sim N(0,25)$. The parameter vectors are specified as following: $(\theta_0, \theta_1, \theta_2) = (1,3,2)$. In the simulations, we consider cases where sample size $n = \{100, 1000\}$, quantiles $\tau = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and we set the number of replications to be 1,000.

For comparison, we consider two estimators of β_1: (i) the standard (infeasible) QR using the unobserved regressors x^*, which we label QR; (ii) the QR estimates with the GR as described above, which is defined as QR-GR. For the two-step QR-GR estimator, the estimation process is as following. In step 1, using the OLS estimation, we obtain the generated regressor (GR) from the model: $\widehat{x} = \widehat{\theta}_0 + \widehat{\theta}_1 w + \widehat{\theta}_2 z$, where (x, w, z) are observables. In step 2, for each $\tau = (0.1, 0.3, 0.5, 0.7, 0.9)$, we estimate β_1 using the QR-GR estimator of y on $\widehat{x}$. We also present results for the corresponding standard errors (SE) of the estimators. For the QR-GR estimator we use the estimator in equation $c_n = k \cdot (\Phi^{-1}(\tau + h_n) - \Phi^{-1}(\tau - h_n))$ where $k = \min\{SE(\widehat{u}), (quantile(\widehat{u}, 0.75) - quantile(\widehat{u}, 0.25))/1.34\}$, $\widehat{u}$ is the residual of the QR estimation, and h_n is the default bandwidth in the R package. k is a robust estimate of the scale and the bandwidth c_n is also commonly chosen in the R package.

Simulation Results

Location Shift Model

The results for the location-shift model are provided in Tables 14–15. The bias, SE, and root mean squared error (RMSE) of both QR-GR and QR estimators are presented in Table 14. Different sample sizes $n = 100$ and $n = 1000$ in the experiments are reported in the table. According to the theorems in the Asymptotic Properties Section, both estimators QR-GR and QR should be consistent. As shown in Table 14, both estimators have empirical bias very close to zero, even for small samples.

Table 14 also shows that the standard error decreases as the sample size increases. However, the QR-GR have substantially larger standard error and RMSE than the regular QR with the true regressor. This result is expected from the theoretical results. These observations reflect the fact that having a GR estimation in the first step does not affect the bias performance but induces a substantially larger variance. This confirms that the sampling error of obtaining by the GR contaminates the standard error in the second-stage QR estimation.

In Table 15 we evaluate the performance of proposed variance-covariance estimator discussed in the Inference Section for $n = 1000$. We report three statistics in Table 15. First, in column 1, we report the sample standard deviation of the QR-GR estimates based on the Monte

Carlo repetitions, which approximates the true standard error of the parameter β_1 of interest. Second, the average of the proposed standard error of the QR-GR estimator is reported in the column 2. Finally, for comparison, the standard error of the usual (infeasible) QR is given in the column 3.

By comparing columns 1 and 2 of Table 15, we can see that the estimated standard errors (SE) of the proposed QR-GR is very closely to the true value given in column 1. However, the average of the estimated standard error calculated in the usual way is severely biased downwards. This result confirms the theoretical predictions and reflects that the sampling error from the first-step estimation induces a larger variance-covariance matrix in the second step. Thus, the estimated standard errors from the conventional formula without considering the GR problem underestimates the population counterpart, and in turn severely affects the inference procedure.

		$n = 100$			$n = 1000$		
		Bias	SE	RMSE	Bias	SE	RMSE
$\tau = 0.1$	QR	0.000	0.012	0.012	0.000	0.004	0.004
	QR-GR	-0.002	0.107	0.107	-0.002	0.032	0.032
$\tau = 0.3$	QR	0.000	0.009	0.009	0.000	0.003	0.003
	QR-GR	0.003	0.101	0.101	0.000	0.032	0.032
$\tau = 0.5$	QR	0.000	0.008	0.008	0.000	0.003	0.003
	QR-GR	-0.002	0.102	0.102	0.001	0.031	0.031
$\tau = 0.7$	QR	0.000	0.009	0.009	0.000	0.003	0.003
	QR-GR	0.003	0.099	0.099	-0.002	0.032	0.032
$\tau = 0.9$	QR	0.000	0.011	0.011	0.000	0.004	0.004
	QR-GR	0.005	0.102	0.102	0.000	0.032	0.032

Table 14: Simulation Results: Bias, SE, RMSE for $n = 100, 1000$, Population Value of $\beta_1 = 3$.

	Empirical SE	Proposed QR-GR SE	Naive QR SE
$\tau = 0.1$	0.032	0.032	0.004
$\tau = 0.3$	0.032	0.032	0.003
$\tau = 0.5$	0.031	0.032	0.003
$\tau = 0.7$	0.032	0.032	0.003
$\tau = 0.9$	0.032	0.032	0.004

Table 15: Simulation Results for Variance and Covariance Matrix. $n = 1000$.

Location-scale Shift Model

The results for the bias and RMSE for sample sizes $n = 100$ and $n = 1000$ are reported in Table 16. For the location-scale shift model, both QR and QR-GR estimators have small bias so that they are close to the population value of $3 + F_{\varepsilon}^{-1}(\tau)$. However, as expected, the QR-GR have larger standard error and root mean square error (RMSE) than the regular QR with the true regressor for both sample sizes. Thus, the results corroborates the theoretical findings that the GR variable has no asymptotic effects in the bias performance but induces a larger variance.

As in the previous case, in Table 17 we assess the performance of the proposed variance-covariance estimator discussed in the Inference Section for the location-scale model. The results are analogous. We see that the proposed QR-GR standard errors in column 2 closely approximates the true standard error in column 1. However, the estimated standard error calculated in the usual way in column 3 is severely smaller and does not approximate well the true standard error.

		$n = 100$			$n = 1000$		
		Bias	SE	RMSE	Bias	SE	RMSE
$\tau = 0.1$	QR	0.053	0.286	0.291	0.138	0.092	0.166
	QR-GR	0.087	0.447	0.455	0.140	0.140	0.197
$\tau = 0.3$	QR	0.014	0.207	0.207	0.028	0.061	0.067
	QR-GR	0.048	0.510	0.512	0.037	0.148	0.152
$\tau = 0.5$	QR	0.001	0.183	0.183	-0.002	0.055	0.055
	QR-GR	0.005	0.573	0.573	0.006	0.165	0.165
$\tau = 0.7$	QR	-0.015	0.205	0.205	-0.028	0.064	0.070
	QR-GR	-0.009	0.635	0.635	-0.047	0.200	0.206
$\tau = 0.9$	QR	-0.048	0.270	0.274	-0.144	0.094	0.172
	QR-GR	-0.109	0.803	0.811	-0.158	0.237	0.285

Table 16: Simulation Results: Bias, SE, RMSE for $n = 100, 1000$, Population Value of $\beta_1 = 3 + F_\varepsilon^{-1}(\tau)$

	Empirical SE	Proposed QR-GR SE	Naive QR SE
$\tau = 0.1$	0.140	0.136	0.092
$\tau = 0.3$	0.148	0.162	0.088
$\tau = 0.5$	0.165	0.185	0.089
$\tau = 0.7$	0.200	0.207	0.088
$\tau = 0.9$	0.237	0.241	0.093

Table 17: Simulation Results for Variance and Covariance Matrix

Application

A Brief Literature Review On Engel Curves

Engel curves describe how household expenditures on particular goods and services depend on household income. The analysis of Engel curves has a long history of estimating the expenditure-income relationship (see Engel (1857) and Working (1943)). They are regression functions where the dependent variable is the level or the budget share of total expenses used to purchase a commodity of goods or services, and the explanatory variable, total expenditure, is usually used as a proxy for income. A very robust empirical result referred to as 'Engel's law' states that the poorer a family is, the larger the budget share it spends on food. Other categories of expenditure present a less robust pattern.

Many researchers explored different functional forms for Engel curves which better fit the data. For example, Lewbel (1997) proposed a functional form for Engel curves that contains a linear function of logarithm of total expenditure and some nonlinear function of total expenditure. Nonparametric estimations also have been incorporated in the estimation of Engel curves like Blundell, Chen, and Kristensen (2007). Methods for comparing different regression functions are discussed in Lewbel (2008). These various shapes of Engel curves suggest a deeper understanding of underlying motives which drive household expenditure decisions. However, it is problematic to assume that only one motive drives the consumption of one particular category of goods or services (Chai and Moneta (2010)). For example, luxury, as a relative concept, is possible in all sorts of consumption. Barigozzi and Moneta (2016) studied a system of budget share and extracted multiple factors that span the same space of basic Engel curves. To understand how the patterns of consumption may be driven by a mixture of different motives, we use the quantile regression with generated regressor (QR-GR) framework laid out in the beginning, in order to explore the heterogeneous effects of motives over different commodity.

To be specific, we estimate unobserved factors which represent underlying motives of expenditure in the first step using the factor model in Barigozzi and Moneta (2016). We then apply quantile regression (QR) methods to the Engel curve model in the second step where budget share is regressed on the estimated factors. The proposed QR-GR model is used to study how each type of household expenditure is driven by different underlying motives. The QR-GR model has two advantages. First, an important difficulty with the model is that the estimator of the variance-covariance matrix needs to take into account the fact that unobserved factors are estimated. And the proposed method is able to provide statistically reliable inference via correct estimation of the variance-covariance matrix. Second, we accommodate possible heterogeneity in the effects of total expenditure over the conditional distribution of budget shares. We find that the estimated Engel curves of budget shares are driven by a mixture of three underlying forces which are motives for a household to consume necessities, luxuries, and goods or services on which is spent the same percentage of the total budget. Also, we find heterogeneity in each motive along the conditional distribution of budget shares.

Data Description

The data we use in this paper is the household data from the UK Family Expenditure Survey 1997-2001 and the Expenditure and Food Survey 2002-2006.[5] The data contains the information about household expenditures on different goods and services. About 7000 households were randomly selected and each household's expenditures were recorded for 2 weeks, which enables researchers to explore various household consumption patterns. We use information about the number of family members, total expenditures, and expenditures on 13 aggregated categories: (1) housing (net); (2) fuel, light, and power; (3) food; (4) alcoholic drinks; (5) tobacco; (6) clothing and footwear; (7) household goods; (8) household services; (9) personal goods and services; (10) motoring; (11) fares and other travel; (12) leisure goods; (13) leisure services.[6]

In this paper, we study a sample of about 4000 households which have 2 to 4 family members, and the budget shares of these categories are pooled over ten years. Table 18 reports some descriptive statistics for total expenditure and 13 categories of expenditures. As shown in the table,

[5]This data has been previously used by Barigozzi and Moneta (2016).
[6]The way to aggregate consumption follows Barigozzi and Moneta (2016).

on average, about 20% of the budget is spent on food and housing, followed by leisure (about 16%) which includes leisure goods and leisure services in our analysis. In order to analyze the deflated data, we pick 2005 as the base year, and use the aggregate price index and the price indices for different categories of expenditure from price indices data (RPI).[7]

sample size $n = 43702$	Min	Max	Mean	Std. Dev.
nper	2.00	4.00	2.67	0.81
total_expenditure	9.97	1587.95	442.92	253.96
housing_net	-0.18	0.97	0.18	0.12
fuel_light_power	-0.15	0.79	0.04	0.04
food	0.00	0.88	0.19	0.09
alcoholic_drink	0.00	0.53	0.04	0.05
tobacco	0.00	0.81	0.02	0.05
clothing_and_footwear	0.00	0.63	0.05	0.06
household_goods	0.00	0.84	0.07	0.08
household_services	0.00	0.89	0.06	0.05
personal_goods_and_services	0.00	0.82	0.04	0.04
motoring	-1.90	0.88	0.13	0.12
fares_and_other_travel	0.00	0.76	0.02	0.05
leisure_goods	0.00	0.85	0.04	0.05
leisure_services	0.00	1.15	0.12	0.12

Table 18: Descriptive Statistics

Empirical Analysis

Barigozzi and Moneta (2016) study a system of budget shares that are driven by latent factors. Using factor analysis in Bai and Ng (2002), they determine the number of basic Engel curves (i.e., the rank of the system) and found that budget shares of each expenditure can be approximated by a three-factor model: (i) a decreasing function (necessities), (ii) an increasing function (luxuries), (iii) a constant function over the total expenditure (unitary elasticity goods).

[7]RPI is obtained from UK Office for National Statistics http://www.ons.gov.uk/. To deflate the data, we divide the nominal total expenditure by the aggregate price index. Also, the nominal budget share, as a ratio of nominal level of expenditure over nominal total budget, is multiplied by a ratio of the total price index over a price index for the particular expenditure.

We estimate the following conditional quantile function:

$$Q_\tau(y_i|f_{1i}, f_{2i}) = \beta_0(\tau) + \beta_1(\tau) f_{1i} + \beta_2(\tau) f_{2i},$$

where the quantity y_i denotes the household budget share of a commodity, and f_{1i} and f_{2i} are the first and second factors (motives) of total expenditure, respectively. Since the third factor is constant over the total expenditure, a constant is also included.

In order to estimate the quantile model in practice, we need to replace the unobserved factors f_1 and f_2 with the observed counterparts $\widehat{f}_1$ and $\widehat{f}_2$ (i.e., generated regressors) by following the functional specifications of each motives established in Barigozzi and Moneta (2016). We thus use the following estimation steps:

Step 1 Estimate the first two motives from the factor analysis and obtain their fitted values by taking the following regression on functions of the total expenditure x:

$$\widehat{f}_1 = \widehat{\theta}_1 + \widehat{\theta}_2 \log(x) \quad \text{where} \quad \widehat{\theta}_2 < 0,$$
$$\widehat{f}_2 = \widehat{\gamma}_1 + \widehat{\gamma}_2 x \log(x) \quad \text{where} \quad \widehat{\gamma}_2 > 0.$$

Step 2 Run a quantile regression of budget shares y on the three factors where the third factor f_3 is associated with a constant:

$$\widehat{\beta}(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \beta_1 \widehat{f}_1 - \beta_2 \widehat{f}_2),$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and $\beta_0 = f_3$.

Remark 3.5. In step 1, in order to generate regressors $\widehat{f}_1$ and $\widehat{f}_2$, we need to estimate the coefficients $(\theta_1, \theta_2, \gamma_1, \gamma_2)$. Using the sample of representative households in Barigozzi and Moneta (2016), we follow their approximate principal component analysis to obtain the three normalized independent factors and then regress each identified factor on their established functional form of the total expenditure to estimate the coefficients for a representative household $(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\gamma}_1, \widehat{\gamma}_2)$.

Remark 3.6. In step 2, each factor represents specific reactions to consumption changes. The first factor f_1 is interpreted as a motive to consume necessities since it is decreasing in the total expenditure, while the second factor f_2 represents the motive to consume luxuries, which is increasing in the total expenditure. The third factor f_3 is related to unitary elasticity goods which represents goods or services allocated the same percentage of total budget by both rich and poor households.

We estimate the coefficients of the factors over different quantiles. For conciseness, we only report the results for food, housing and leisure. The main results in Figure 22 show how the budget shares of each commodity relate to each factors, respectively. We only show the results for f_1 and f_2 since the third factor f_3 does not vary with the family expenditure. The coefficient estimates over different quantiles reflect marginal impact of each motive on the distribution of budget shares at different quantiles. In all cases, their magnitudes varies over the level of the quantile τ. Thus our quantile model well identifies apparent heterogenous effects of consumption motives on the the budget shares.

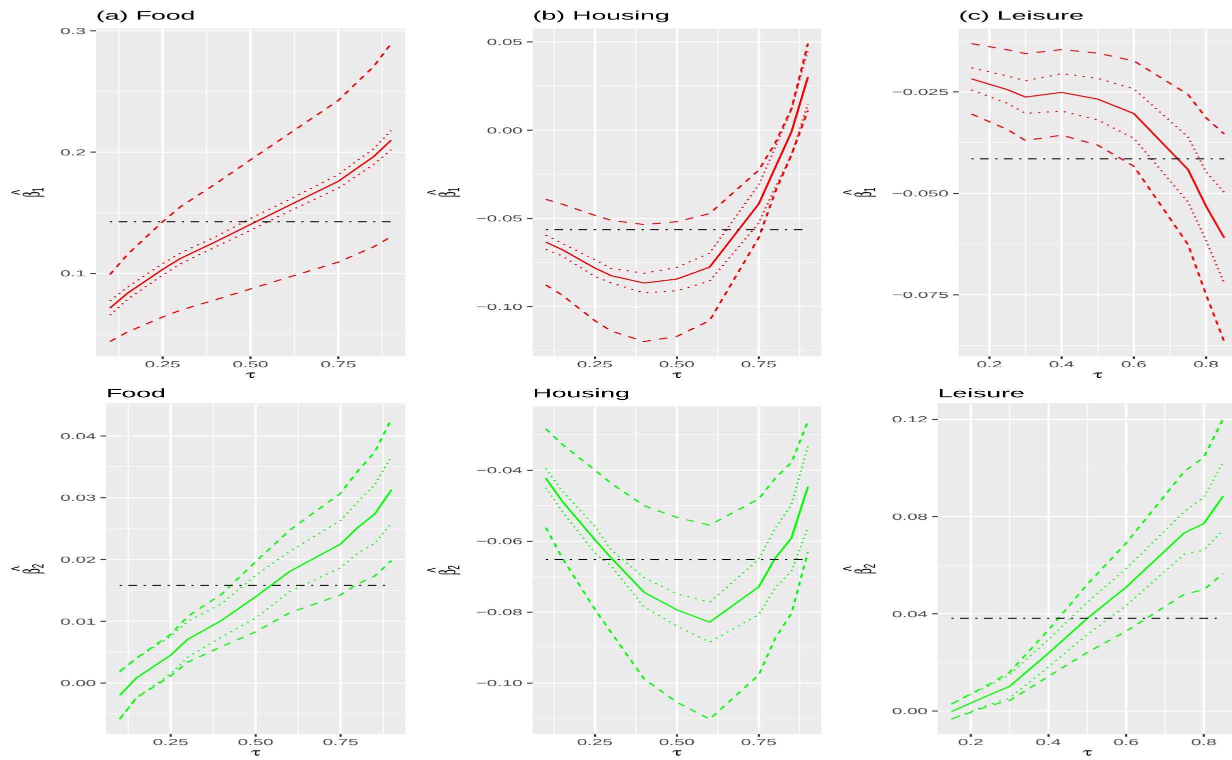


Figure 22: Coefficient Estimates of Factor 1 and 2 and Their 95% Confidence Intervals

Notes: (a) food budget share; (b) housing budget share; (c) leisure budget share. Upper row: coefficient estimates of factor 1; Lower row: coefficient estimates of factor 2. Black line: OLS coefficient estimate; Solid line: QR coefficient estimate; Dotted line: 95% confidence intervals obtained by conventional estimation; Dashed line: 95% confidence intervals obtained by our proposed estimation.

The 95% confidence intervals in dashed lines are based on our proposed estimator of the asymptotic variance-covariance matrix, while the 95% confidence intervals in dotted lines are calculated by conventional QR variance-covariance estimation. Clearly, the adjusted confidence bands for QR-GR are much larger than the conventional bands which do not adjust for the GR issue. This can be observed more clearly in Table 19 which summarizes the estimated standard errors for a subset of quantiles $\tau \in \{0.25, 0.5, 0.75\}$. The table reports that the estimated standard errors from the proposed estimator (QR-GR SE) is always larger than those from the conventional estimator (QR SE). These results underscore the importance of estimating standard errors correctly by taking into account the GR issue in Engel curves.

Since both f_1 and f_2 are normalized measures of different motives, comparing the relative

	$\widehat{\beta}_1$	QR-GR SE	QR SE	$\widehat{\beta}_2$	QR-GR SE	QR SE
τ=0.25	0.103	0.020	0.002	0.005	0.002	0.001
τ= 0.5	0.140	0.027	0.003	0.014	0.003	0.002
τ=0.75	0.176	0.034	0.003	0.023	0.004	0.002

(a) Food

	$\widehat{\beta}_1$	QR-GR SE	QR SE	$\widehat{\beta}_2$	QR-GR SE	QR SE
τ=0.25	-0.078	0.015	0.002	-0.060	0.010	0.002
τ= 0.5	-0.084	0.017	0.003	-0.079	0.013	0.002
τ=0.75	-0.042	0.010	0.006	-0.073	0.013	0.004

(b) Housing

	$\widehat{\beta}_1$	QR-GR SE	QR SE	$\widehat{\beta}_2$	QR-GR SE	QR SE
τ=0.25	-0.024	0.005	0.002	0.007	0.002	0.002
τ= 0.5	-0.027	0.006	0.003	0.038	0.007	0.003
τ=0.75	-0.044	0.009	0.004	0.073	0.013	0.005

(c) Leisure

Table 19: Estimates of Factor 1 and Factor 2 at quantiles $\tau = 0.25, 0.5, 0.75$ for Various Commodities

Notes: Coefficient estimates of factor 1 and factor 2 and their standard error at quantiles $\tau = 0.25, 0.5, 0.75$ for (a) food, (b) housing and (c) leisure. Column 1 and 4: coefficient estimates; column 2 and 5: standard error calculated in adjusted QR estimation; column 3 and 6: standard error calculated in conventional QR estimation.

magnitudes of their coefficient estimates helps us to tell which motive plays a leading role in a particular commodity. For example, Figure 22(a) shows that the motive to consume food as necessity clearly plays a leading role relative to luxuries at all quantiles. In Figure 22(c), the motive to consume leisure as luxuries is dominating a role as necessity. However, consuming housing in Figure 22(b) has a mixed result where both necessity and luxuries play a similar role.

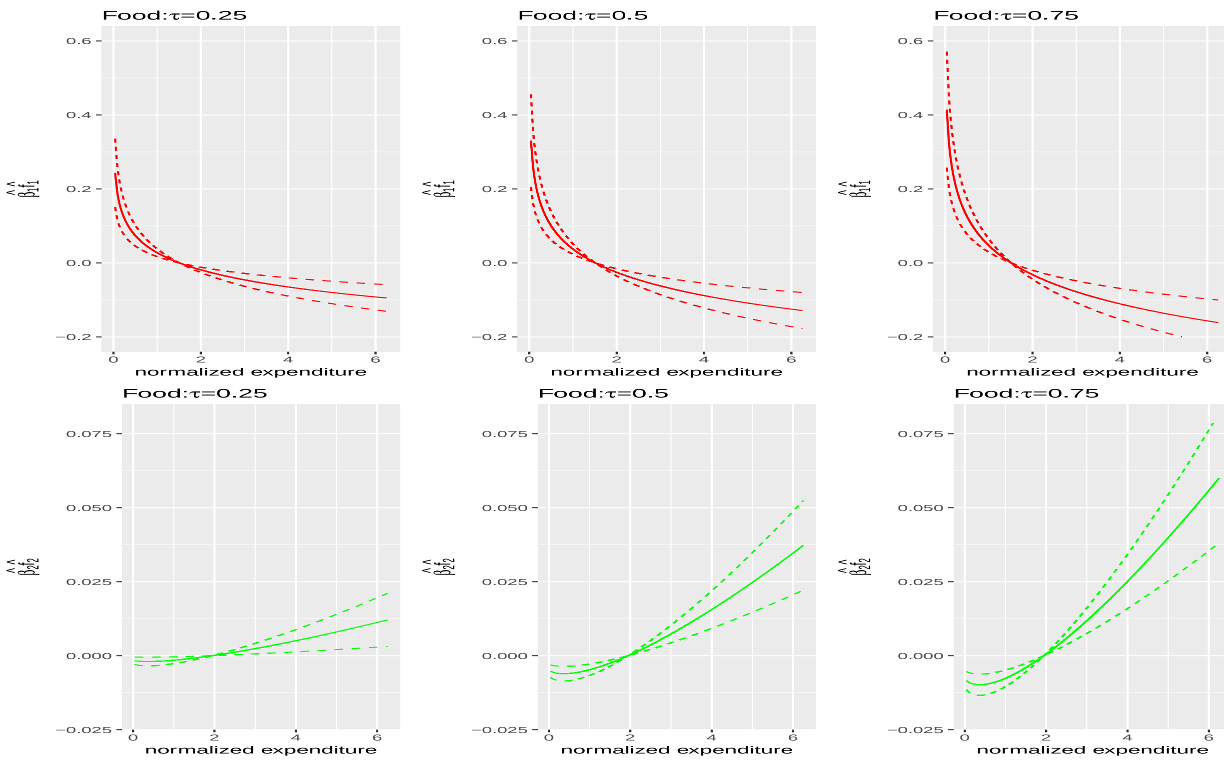


Figure 23: Contributions of Factor 1 and Factor 2 to Food Budget Share

Notes: Factor 1 (upper row) and factor 2 (lower row) to food budget share; dotted lines are the corresponding 95% confidence intervals.

To help further interpret how the patterns of budget shares are related to these unobservable motives, we show how the budget share of a particular commodity varies over the total expenditure in Figure 23 through Figure 25. We estimate the contribution of each motive to the budget share at different quantiles $\tau \in \{0.25, 0.5, 0.75\}$. It is interesting to see how the motive to consume a particular commodity as necessity or luxury contributes to its budget share at different levels of total expenditure.

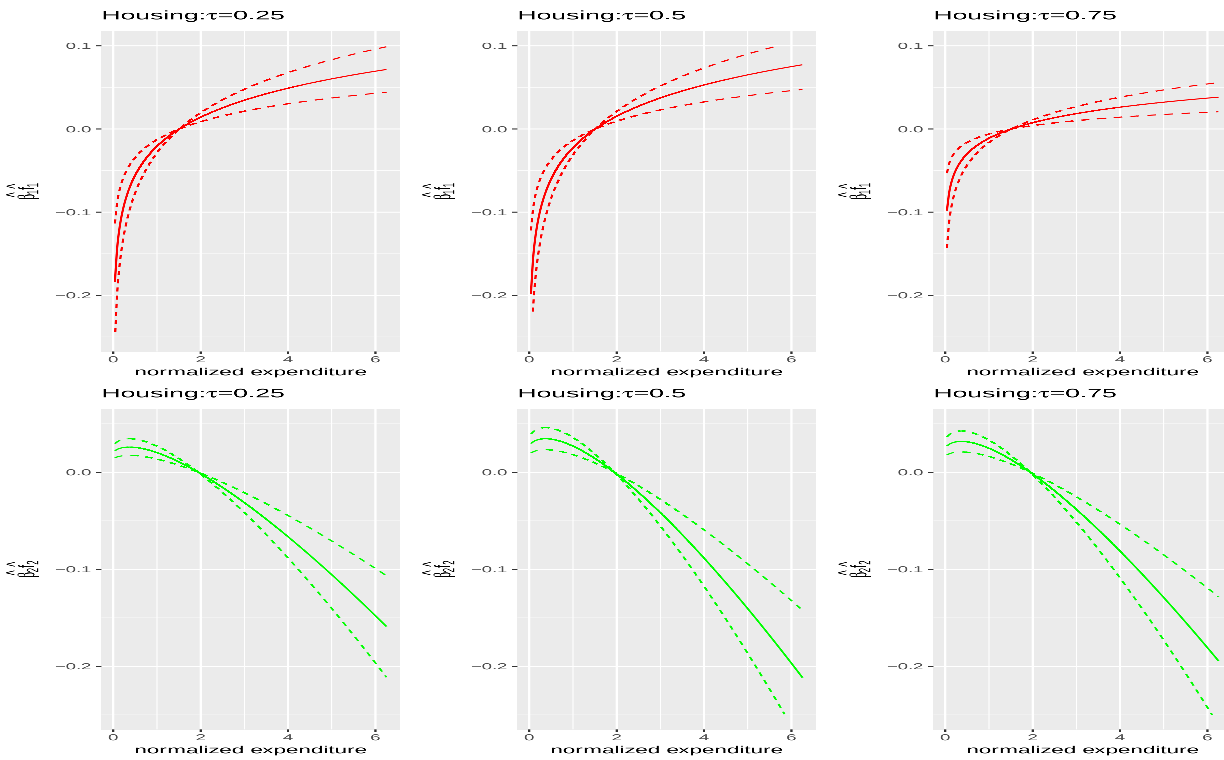


Figure 24: Contributions of Factor 1 and Factor 2 to Housing Budget Share

Notes: factor 1 (upper row) and factor 2 (lower row) to housing budget share; dotted lines are the corresponding 95% confidence intervals.

For instance, Figure 23 shows that, as total expenditure increases, the contribution of the first factor, f_1 (necessity), to food budget share decreases at all τ, while the second factor, f_2 (luxuries), increases, as well documented in the literature. This implies that as total expenditure increases, the motive to consume food as a necessary good decreases, while the motive to consume food as a luxury good increases. The pattern in Figure 24, which is very different from food, reflects that individuals consume housing more as necessity but less as luxury, as their total expenditure increases. The motives to consume food and housing as necessity or luxuries play opposite roles in the budget share as total expenditure changes. However, as shown in Figure 25, both motives to consume leisure increase as total expenditure increases as well. Another interesting point is that the magnitudes of those estimated factors or motives are different across different quantiles of the conditional distribution of budget shares in all three commodities.

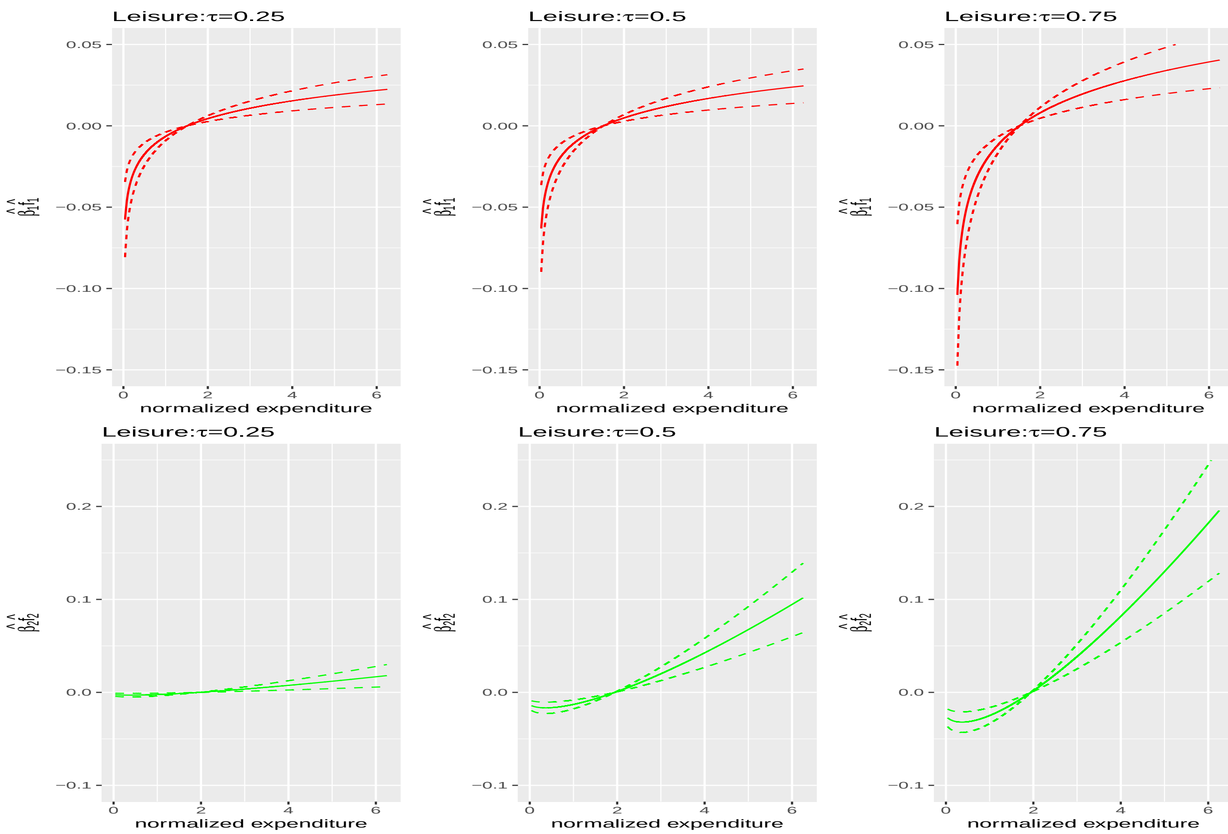


Figure 25: Contributions of Factor 1 and Factor 2 to Leisure Budget Share

Notes: factor 1 (upper row) and factor 2 (lower row) to leisure budget share; dotted lines are the corresponding 95% confidence intervals.

Figures 23 through 25 suggest that the motive of consumption as necessity or luxury plays different roles over the level of total expenditure. The overall effects of total expenditure on budget shares are presented in Figure 26. The estimated budget shares are calculated by adding all three factors together. Figure 26(a) shows that the budget share of food decreases as total expenditure increases, which is classically referred to as Engel's law. In Figure 26(b) we see a concave shape curve, that is, the budget share of housing increases for low expenditures, but decreases as total expenditure increases. In addition, individuals always raise the budget share of leisure as they consume more, as shown in Figure 26(c). Finally, and importantly, in all three commodities, the shapes of the curves vary over different quantiles τ. In all, the results show strong evidence that individual heterogeneity in the shape of Engel curves is uncovered by the QR.

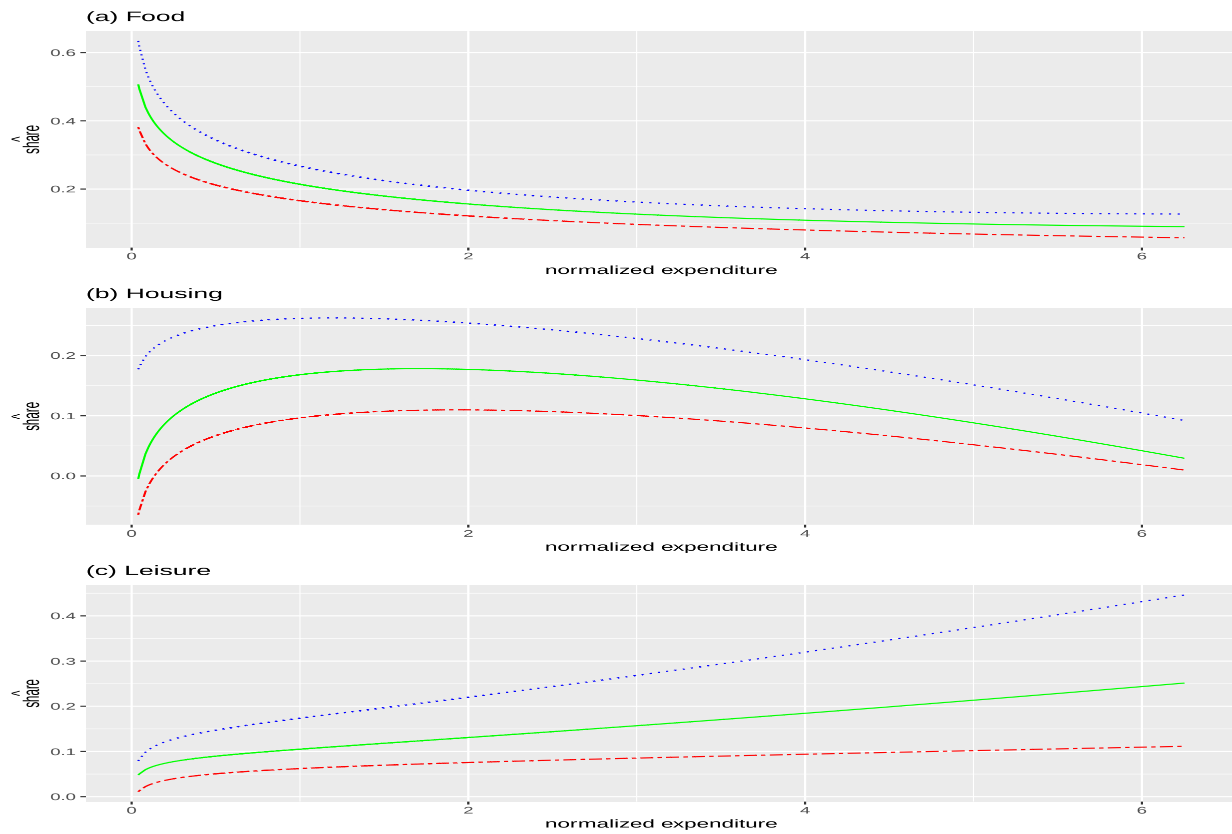


Figure 26: Fitted Values of Budget Share where $\widehat{share} = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{f}_1 + \widehat{\beta}_2 \widehat{f}_2$.

Notes: f(a) food; (b) housing; (c) leisure. Red two dashed line: corresponding coefficients estimated at $\tau = 0.25$; green solid line: corresponding coefficients estimated at $\tau = 0.5$; blue dotted line: corresponding coefficients estimated at $\tau = 0.75$.

Conclusion

We study estimation and inference for linear quantile regression (QR) models with generated regressors (GR). We propose a QR-GR two-step estimator for the parameters of interest and an estimator for the corresponding asymptotic variance-covariance matrix. We establish the asymptotic properties of the estimators. Monte Carlo simulation and estimation of the Engel curves using data from the UK Family Expenditure Survey confirm that taking into account the GR problem in the QR framework is essential for correct inference. Furthermore, the empirical application shows strong heterogeneity of the Engel curves over different quantiles of the conditional distribution of budget shares in most commodities.

REFERENCES

ALEMAYEHU, B., AND K. E. WARNER (2004): “The lifetime distribution of health care costs,” *Health services research*, 39(3), 627–642.

ALTONJI, J. G., F. HAYASHI, AND L. J. KOTLIKOFF (1997): “Parental altruism and inter vivos transfers: Theory and evidence,” *Journal of Political Economy*, 105(6), 1121–1166.

AOKI, T., AND K. NISHIMURA (2017): “Global convergence in an overlapping generations model with two-sided altruism,” *Journal of Evolutionary Economics*, 27(5), 1205–1220.

ARELLANO, M., AND S. BONHOMME (2016): “Sample selection in quantile regression: A survey,” in *Handbook of Quantile Regression*, ed. by R. Koenker, V. Chernozhukov, X. He, and L. Peng. CRC/Chapman-Hall, forthcoming.

ARON-DINE, A., L. EINAV, AND A. FINKELSTEIN (2013): “The RAND health insurance experiment, three decades later,” *Journal of Economic Perspectives*, 27(1), 197–222.

ARRONDEL, L., AND A. MASSON (2006): “Altruism, exchange or indirect reciprocity: what do the data on family transfers show?,” *Handbook of the economics of giving, altruism and reciprocity*, 2, 971–1053.

BAI, J., AND S. NG (2002): “Determining the number of factors in approximate factor models,” *Econometrica*, 70, 191–221.

BAICKER, K., AND D. GOLDMAN (2011): “Patient cost-sharing and healthcare spending growth,” *Journal of Economic Perspectives*, 25(2), 47–68.

BARIGOZZI, M., AND A. MONETA (2016): “Identifying the independent sources of consumption variation,” *Journal of Applied Econometrics*, 31(2), 420–449.

BARRO, R. J. (1974): “Are government bonds net wealth?,” *Journal of political economy*, 82(6), 1095–1117.

——— (1977): "Unanticipated money growth and unemployment in the United States," *The American Economic Review*, 67(2), 101–115.

——— (1978): "Unanticipated money, output, and the price level in the United States," *Journal of Political Economy*, 86(4), 549–580.

BATTISTIN, E., A. BRUGIAVINI, E. RETTORE, AND G. WEBER (2007): "How Large is the Retirement Consumption Drop in Italy?," Discussion paper, Working Paper.

——— (2009): "The retirement consumption puzzle: evidence from a regression discontinuity approach," *American Economic Review*, 99(5), 2209–26.

BECKER, G. S. (1974): "A theory of social interactions," *Journal of political economy*, 82(6), 1063–1093.

BERTAKIS, K. D., R. AZARI, L. J. HELMS, E. J. CALLAHAN, AND J. A. ROBBINS (2000): "Gender differences in the utilization of health care services," *Journal of family practice*, 49(2), 147–147.

BLAU, D. M., AND D. B. GILLESKIE (2006): "Health insurance and retirement of married couples," *Journal of Applied Econometrics*, 21(7), 935–953.

BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75(6), 1613–1669.

BUCHINSKY, M., AND J. HAHN (1998): "An alternative estimator for the censored quantile regression model," *Econometrica*, pp. 653–671.

CAI, F., J. GILES, AND X. MENG (2006): "How well do children insure parents against low retirement income? An analysis using survey data from urban China," *Journal of public Economics*, 90(12), 2229–2255.

CARD, D., C. DOBKIN, AND N. MAESTAS (2008): “The impact of nearly universal insurance coverage on health care utilization: evidence from Medicare,” *American Economic Review*, 98(5), 2242–58.

CHAI, A., AND A. MONETA (2010): “Retrospectives: Engel curves,” *Journal of Economic Perspectives*, 24(1), 225–40.

CHANDRA, A., J. GRUBER, AND R. MCKNIGHT (2010): “Patient cost-sharing and hospitalization offsets in the elderly,” *American Economic Review*, 100(1), 193–213.

CHEN, L., J. J. DOLADO, AND J. GONZALO (2015): “Quantile factor models,” Mimeo.

CHERNOZHUKOV, V., I. FERNANDEZ-VAL, AND A. KOWALSKI (2015): “Quantile regression with censoring and endogeneity,” *Journal of Econometrics*, 186, 201–221.

CHERNOZHUKOV, V., AND C. HANSEN (2005): “An IV model of quantile treatment effects,” *Econometrica*, 73, 245–261.

——— (2006): “Instrumental quantile regression inference for structural and treatment effects models,” *Journal of Econometrics*, 132, 491–525.

CHO, I. (2012): “The retirement consumption in Korea: evidence from the Korean Labor and Income Panel Study,” *Global Economic Review*, 41(2), 163–187.

CHUANCHUAN, Z., AND C. BINKAI (2014): “Can” Public Pension System” Substitutes” Family Mutual Insurance”?,” *Economic Research Journal*, (11), 9.

CIGNO, A. (2016): “Chapter 10 - Conflict and Cooperation Within the Family, and Between the State and the Family, in the Provision of Old-Age Security,” vol. 1 of *Handbook of the Economics of Population Aging*, pp. 609 – 660. North-Holland.

CIGNO, A., G. C. GIANNELLI, F. C. ROSATI, AND D. VURI (2006): “Is there such a thing as a family constitution? A test based on credit rationing,” *Review of Economics of the Household*, 4(3), 183–204.

CIGNO, A., M. KOMURA, AND A. LUPORINI (2017): "Self-enforcing family rules, marriage and the (non) neutrality of public intervention," *Journal of Population Economics*, 30(3), 805–834.

COX, D. (1987): "Motives for private income transfers," *Journal of political economy*, 95(3), 508–546.

COX, D., Z. ESER, AND E. JIMENEZ (1998): "Motives for private transfers over the life cycle: An analytical framework and evidence for Peru," *Journal of Development Economics*, 55(1), 57–80.

CREMER, H., P. PESTIEAU, AND K. ROEDER (2016): "Social long-term care insurance with two-sided altruism," *Research in Economics*, 70(1), 101–109.

DE NARDI, M., E. FRENCH, AND J. B. JONES (2010): "Why do the elderly save? The role of medical expenses," *Journal of Political Economy*, 118(1), 39–75.

ENGEL, E. (1857): "Die productions-und consumtionsverhältnisse des königreichs sachsen," *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern*, 8, 1–54.

ENGLE, R. F. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, pp. 987–1007.

FAN, E. (2010): "Who benefits from public old age pensions? Evidence from a targeted program," *Economic Development and Cultural Change*, 58(2), 297–322.

FINKELSTEIN, A. (2007): "The aggregate effects of health insurance: Evidence from the introduction of Medicare," *The quarterly journal of economics*, 122(1), 1–37.

FINKELSTEIN, A., S. TAUBMAN, B. WRIGHT, M. BERNSTEIN, J. GRUBER, J. P. NEWHOUSE, H. ALLEN, K. BAICKER, AND O. H. S. GROUP (2012): "The Oregon health insurance experiment: evidence from the first year," *The Quarterly journal of economics*, 127(3), 1057–1106.

FRENCH, E., AND J. B. JONES (2011): "The effects of health insurance and self-insurance on retirement behavior," *Econometrica*, 79(3), 693–732.

GAN, L., Z. YIN, N. JIA, S. XU, S. MA, AND L. ZHENG (2014): *Data you need to know about China*. Springer.

GENICOT, G. (2016): "Two-sided altruism and signaling," *Economics Letters*, 145, 92–97.

GOH, E. C. (2009): "Grandparents as childcare providers: An in-depth analysis of the case of Xiamen, China," *Journal of Aging Studies*, 23(1), 60–68.

HAHN, J., AND G. RIDDER (2013): "Asymptotic variance of semiparametric estimators with generated regressors," *Econometrica*, 81, 315–340.

HAHN, J., P. TODD, AND W. VAN DER KLAAUW (2001): "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 69(1), 201–209.

HALL, R. E., AND C. I. JONES (2007): "The value of life and the rise in health spending," *The Quarterly Journal of Economics*, 122(1), 39–72.

HAWKES, S., AND K. BUSE (2013): "Gender and global health: evidence, policy, and inconvenient truths," *The Lancet*, 381(9879), 1783–1787.

HENDRICKS, W., AND R. KOENKER (1991): "Hierarchical spline models for conditional quantiles and the demand for electricity," *Journal of the American Statistical Association*, 87, 58–68.

HUANG, F., AND L. GAN (2017): "The impacts of China's urban employee basic medical insurance on healthcare expenditures and health outcomes," *Health economics*, 26(2), 149–163.

JENSEN, R. T. (2004): "Do private transfers 'displace'the benefits of public transfers? Evidence from South Africa," *Journal of Public Economics*, 88(1-2), 89–112.

KOENKER, R., AND G. BASSETT (1978): "Regression quantiles," *Econometrica*, 46, 33–50.

——— (1982a): "Robust tests for heteroscedasticity based on regression quantiles," *Econometrica*, 50(1), 43–61.

——— (1982b): "Tests of linear hypotheses and l_1 estimation," *Econometrica*, 50, 1577–1584.

KOENKER, R., AND J. A. F. MACHADO (1999): "Goodness of fit and related inference processes for quantile regression," *Journal of the American Statistical Association*, 94, 1296–1310.

LADWIG, K.-H., B. MARTEN-MITTAG, B. FORMANEK, AND G. DAMMANN (2000): "Gender differences of symptom reporting and medical health care utilization in the German population," *European journal of epidemiology*, 16(6), 511–518.

LAITNER, J. (2001): "Wealth accumulation in the US: do inheritances and bequests play a significant role?," *Michigan Retirement Research Center Research Paper No. WP*, 19.

LEE, R. (2016): "Macroeconomics, aging, and growth," in *Handbook of the Economics of Population Aging*, vol. 1, pp. 59–118. Elsevier.

LEE, R., S.-H. LEE, AND A. MASON (2008): "Charting the Economic Life Cycle," *Population and Development Review*, 34, 208–237.

LEE, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141(2), 1131–1158.

LEWBEL, A. (1997): "Consumer demand systems and household equivalence scales," *Handbook of Applied Econometrics*, 2, 167–201.

——— (2008): "Engel curves," *The New Palgrave Dictionary of Economics*, 2.

LI, H., X. SHI, AND B. WU (2015): "The retirement consumption puzzle in China," *American Economic Review*, 105(5), 437–41.

LINDBECK, A., AND S. NYBERG (2006): "Raising children to work hard: altruism, work norms, and social insurance," *The Quarterly Journal of Economics*, 121(4), 1473–1503.

MA, L., AND R. KOENKER (2006): "Quantile regression methods for recursive structural equation models," *Journal of Econometrics*, 134(2), 471–506.

MAMMEN, E., C. ROTHE, AND M. SCHIENLE (2012): "Nonparametric regression with nonparametrically generated covariates," *Annals of Statistics*, 40, 1132–1170.

MANNING, W. G., J. P. NEWHOUSE, N. DUAN, E. B. KEELER, AND A. LEIBOWITZ (1987): "Health insurance and the demand for medical care: evidence from a randomized experiment," *The American economic review*, pp. 251–277.

MASON, A., R. LEE, A.-C. TUNG, M.-S. LAI, AND T. MILLER (2009): "Population aging and intergenerational transfers: Introducing age into national accounts," in *Developments in the Economics of Aging*, pp. 89–122. University of Chicago Press.

MCGARRY, K., AND R. F. SCHOENI (1995): "Transfer behavior in the health and retirement study: Measurement and the redistribution of resources within the family," *Journal of Human resources*, pp. S184–S226.

MURPHY, K. M., AND R. H. TOPEL (2002): "Estimation and inference in two-step econometric models," *Journal of Business & Economic Statistics*, 20(1), 88–97.

MUSTARD, C. A., P. KAUFERT, A. KOZYRSKYJ, AND T. MAYER (1998): "Sex differences in the use of health care services," *New England Journal of Medicine*, 338(23), 1678–1683.

OTTONI-WILHELM, M., L. VESTERLUND, AND H. XIE (2017): "Why do people give? Testing pure and impure altruism," *American Economic Review*, 107(11), 3617–33.

OWENS, G. (2008): "Gender differences in health care expenditures, resource utilization, and quality of care," *Journal of Managed Care Pharmacy*, 14(3), 2–6.

PAGAN, A. (1984): "Econometric issues in the analysis of regressions with generated regressors," *International Economic Review*, 25, 221–247.

POWELL, J. L. (1991): "Estimation of monotonic regression models under quantile regressions," in *Nonparametric and semiparametric models in econometrics*, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge University Press, Cambridge.

REDONDO-SENDINO, Á., P. GUALLAR-CASTILLÓN, J. R. BANEGAS, AND F. RODRÍGUEZ-ARTALEJO (2006): "Gender differences in the utilization of health-care services among the older adult population of Spain," *BMC public health*, 6(1), 155.

SECONDI, G. (1997): "Private monetary transfers in rural China: are families altruistic?," *The Journal of Development Studies*, 33(4), 487–511.

SHAIKH, B. T., AND J. HATCHER (2004): "Health seeking behaviour and health service utilization in Pakistan: challenging the policy makers," *Journal of public health*, 27(1), 49–54.

SHIGEOKA, H. (2014): "The effect of patient cost sharing on utilization, health, and risk protection," *American Economic Review*, 104(7), 2152–84.

THISTLETHWAITE, D. L., AND D. T. CAMPBELL (1960): "Regression-discontinuity analysis: An alternative to the ex post facto experiment.," *Journal of Educational psychology*, 51(6), 309.

TROCHIM, W. M., AND W. THOCHIM (1984): *Research design for program evaluation: The regression-discontinuity approach*. Sage Beverly Hills, CA.

VAIDYA, V., G. PARTHA, AND M. KARMAKAR (2012): "Gender differences in utilization of preventive care services in the United States," *Journal of women's health*, 21(2), 140–145.

WORKING, H. (1943): "Statistical laws of family expenditure," *Journal of the American Statistical Association*, 38, 43–56.

XIAO, Z., AND R. KOENKER (2009): "Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models," *Journal of the American Statistical Association*, 104, 1696–1712.

ZHAO, X. (2010): "Does Retirement Affect Health?," *China Economic Quarterly*, 4, 019.

Proof of The Main Results

Proof of Theorem of Consistency. Let $\widehat{\beta}(\tau)$ be the two-step quantile regression (QR) with generated regressors (GR) estimator, and $\breve{\beta}(\tau)$ be the usual (infeasible) QR estimator with true unobservable variables. Consider

$$\widehat{\beta}(\tau)-\beta_0(\tau)=(\widehat{\beta}(\tau)-\breve{\beta}(\tau))+(\breve{\beta}(\tau)-\beta_0(\tau)). \tag{A.1}$$

Under the standard QR conditions **A1** and **A2**, we have that the second term in equation (A.1) satisfies that

$$\breve{\beta}(\tau)-\beta_0(\tau)\xrightarrow{p}0, \tag{A.2}$$

and the following linear representation

$$\breve{\beta}(\tau)-\beta_0(\tau)=D_1^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\psi_\tau(y_i-\boldsymbol{x}_i^\top\beta_0(\tau))+o_p(1) \tag{A.3}$$

where $\psi_\tau(u):=\tau-I(u\leqslant 0)$.

Now, by noticing that $\widehat{\beta}=\widehat{\beta}(\widehat{x}_1,\cdots,\widehat{x}_q,x_{q+1},\cdots,x_k)$ and $\breve{\beta}=\breve{\beta}(x_1,\cdots,x_q,x_{q+1},\cdots,x_k)$, and by expanding the first term in equation (A.1) $\widehat{\beta}(\tau)-\breve{\beta}(\tau)$, we have

$$\begin{aligned}\widehat{\beta}(\tau)-\breve{\beta}(\tau) &= \left(\frac{\partial\breve{\beta}(\tau)}{\partial x_1}^\top,\cdots,\frac{\partial\breve{\beta}(\tau)}{\partial x_q}^\top\right)\begin{bmatrix}\widehat{x}_1-x_1\\ \vdots\\ \widehat{x}_q-x_q\end{bmatrix}+o_p(1)\\ &= \left(\frac{\partial\breve{\beta}(\tau)}{\partial x_1}^\top,\cdots,\frac{\partial\breve{\beta}(\tau)}{\partial x_q}^\top\right)\begin{bmatrix}g_1(\boldsymbol{w}_1,\widehat{\boldsymbol{\theta}}_1)-g_1(\boldsymbol{w}_1,\boldsymbol{\theta}_1)\\ \vdots\\ g_q(\boldsymbol{w}_q,\widehat{\boldsymbol{\theta}}_q)-g_q(\boldsymbol{w}_q,\boldsymbol{\theta}_q)\end{bmatrix}+o_p(1)\\ &= \left(\frac{\partial\breve{\beta}(\tau)}{\partial x_1}^\top,\cdots,\frac{\partial\breve{\beta}(\tau)}{\partial x_q}^\top\right)\begin{bmatrix}\nabla_{\theta_1}g_1(\boldsymbol{w}_1,\boldsymbol{\theta_1})^\top(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1)\\ \vdots\\ \nabla_{\theta_q}g_q(\boldsymbol{w}_q,\boldsymbol{\theta_q})^\top(\widehat{\boldsymbol{\theta}}_q-\boldsymbol{\theta}_q)\end{bmatrix}+o_p(1),\end{aligned}$$

where the last equality follows from an expansion for $g_j(\cdot)$ for $j=1,\cdots,q$.

As seen in (A.3), $\breve{\beta}(\tau)$ can be rewritten as

$$\breve{\beta}(\tau)=\left(\frac{1}{n}\sum_{i=1}^n f_i(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\left[f_i(0|\boldsymbol{x}_i)\boldsymbol{x}_i^\top\beta_0(\tau)+\psi_\tau(y_i-\boldsymbol{x}_i^\top\beta_0(\tau))\right]+o_p(1).$$

By taking the derivative of $\breve{\beta}(\tau)$ with respect to x_j, one can notice that the contribution of the $\psi_\tau(y_i-\boldsymbol{x}_i^\top\beta_0(\tau))$ term is $o_p(1)$.

Now we apply Lemma 2 about G-inverse from Ma and Koenker (2006), and obtain that for $j=1,\cdots,q$

$$\begin{aligned}\left(\frac{\partial\breve{\beta}(\tau)}{\partial x_j}\right)^\top\nabla_{\theta_j}g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)^\top &= -D_1^{-1}\left(\frac{1}{n}\sum_{i=1}^n f_i(0|\boldsymbol{x}_i)\beta_j(\tau)\boldsymbol{x}_i\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top\right)+o_p(1)\\ &= -D_1^{-1}D_{12}^j+o_p(1),\end{aligned}$$

where

$$D_{12}^j=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n f_i(0|\boldsymbol{x}_i)\beta_j(\tau)\boldsymbol{x}_i\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top.$$

It follows that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(\tau)-\breve{\boldsymbol{\beta}}(\tau) &= \sum_{j=1}^{q}\left(\frac{\partial\breve{\boldsymbol{\beta}}(\tau)}{\partial x_j}\right)^{\top}\nabla_{\theta_j}g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)^{\top}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)+o_p(1)\\ &= \sum_{j=1}^{q}(-D_1^{-1}D_{12}^{j}+o_p(1))(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)+o_p(1)\\ &= \sum_{j=1}^{q}[-D_1^{-1}D_{12}^{j}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)]+o_p(1).\end{aligned}$$

For any $j=1,\cdots,q$: $\text{plim}_{n\to\infty}\widehat{\boldsymbol{\theta}}_j=\boldsymbol{\theta}_j$ by assumption **A3** and $D_1^{-1}D_{12}^{j}$ is bounded by assumptions **A2** and **A5**, we have that $-D_1^{-1}D_{12}^{j}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)=o_p(1)$. Therefore, as $n\to\infty$, we have

$$\widehat{\boldsymbol{\beta}}(\tau)-\breve{\boldsymbol{\beta}}(\tau)\xrightarrow{p}0. \tag{A.4}$$

Combining (A.2) and (A.4), we have

$$\widehat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}_{\mathbf{0}}(\tau)=(\widehat{\boldsymbol{\beta}}(\tau)-\breve{\boldsymbol{\beta}}(\tau))+(\breve{\boldsymbol{\beta}}(\tau)-\widehat{\boldsymbol{\beta}}_0(\tau))\xrightarrow{p}0.$$

□

Proof of Theorem of Asymptotic Normality. Recall that $\widehat{\boldsymbol{\beta}}(\tau)$ is the two-step QR-GR estimator and $\breve{\boldsymbol{\beta}}(\tau)$ is the infeasible QR estimator. Consider the following

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}_{\mathbf{0}}(\tau))=\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau)-\breve{\boldsymbol{\beta}}(\tau))+\sqrt{n}(\breve{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}_0(\tau)). \tag{A.5}$$

Under the usual QR conditions **A1** and **A2**, we have that the second term in equation (A.5) has the standard QR expansion as

$$\sqrt{n}(\breve{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}_0(\tau))=D_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i\psi_\tau(y_i-\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_0(\tau))+o_p(1), \tag{A.6}$$

and satisfies a Central Limit Theorem such that

$$\sqrt{n}(\breve{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)) \xrightarrow{d} N(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}), \tag{A.7}$$

where $\psi_\tau(u) := \tau - I(u \leqslant 0)$.

Now, by noticing that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{x}_1, \cdots, \widehat{x}_q, x_{q+1}, \cdots, x_k)$ and $\breve{\boldsymbol{\beta}} = \breve{\boldsymbol{\beta}}(x_1, \cdots, x_q, x_{q+1}, \cdots, x_k)$, and by expanding the first term in equation (A.5) $\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \breve{\boldsymbol{\beta}}(\tau))$, we have

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \breve{\boldsymbol{\beta}}(\tau)) &= \sqrt{n}\left(\frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_1}^\top, \cdots, \frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_q}^\top\right)\begin{bmatrix} \widehat{x}_1 - x_1 \\ \vdots \\ \widehat{x}_q - x_q \end{bmatrix} + o_p(1) \\
&= \sqrt{n}\left(\frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_1}^\top, \cdots, \frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_q}^\top\right)\begin{bmatrix} g_1(\boldsymbol{w}_1, \widehat{\boldsymbol{\theta}}_1) - g_1(\boldsymbol{w}_1, \boldsymbol{\theta}_1) \\ \vdots \\ g_q(\boldsymbol{w}_q, \widehat{\boldsymbol{\theta}}_q) - g_q(\boldsymbol{w}_q, \boldsymbol{\theta}_q) \end{bmatrix} + o_p(1) \\
&= \sqrt{n}\left(\frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_1}^\top, \cdots, \frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_q}^\top\right)\begin{bmatrix} \nabla_{\theta_1} g_1(\boldsymbol{w}_1, \boldsymbol{\theta}_1)^\top (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \\ \vdots \\ \nabla_{\theta_q} g_q(\boldsymbol{w}_q, \boldsymbol{\theta}_q)^\top (\widehat{\boldsymbol{\theta}}_q - \boldsymbol{\theta}_q) \end{bmatrix} + o_p(1),
\end{aligned}$$

where the last equality follows from an expansion for $g_j(\cdot)$ for $j = 1, \cdots, q$.

As seen in (A.6), the Bahadur representation for $\sqrt{n}\breve{\boldsymbol{\beta}}(\tau)$ can be rewritten as

$$\sqrt{n}\breve{\boldsymbol{\beta}}(\tau) = \left(\frac{1}{n}\sum_{i=1}^{n} f_i(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i \left[f_i(0|\boldsymbol{x}_i)\boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau) + \psi_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau))\right] + o_p(1).$$

Take the derivative of $\sqrt{n}\breve{\boldsymbol{\beta}}(\tau)$ with respect to x_j by applying Lemma 2 about G-inverse from Ma and Koenker (2006) and notice that the contribution of the $\psi_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau))$ term is $o_p(1)$. For

$j = 1, \cdots, q$,

$$\left(\frac{\partial \breve{\boldsymbol{\beta}}(\tau)}{\partial x_j}\right)^\top \nabla_{\theta_j} g_j(\boldsymbol{w}_j, \boldsymbol{\theta}_j)^\top = -D_1^{-1}\left(\frac{1}{n}\sum_{i=1}^n f_i(0|\boldsymbol{x}_i)\beta_j(\tau)\boldsymbol{x}_i \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top\right) + o_p(1)$$
$$= -D_1^{-1} D_{12}^j + o_p(1),$$

where

$$D_{12}^j = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n f_i(0|\boldsymbol{x}_i)\beta_j(\tau)\boldsymbol{x}_i \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top. \tag{A.8}$$

Thus, we have the following representation

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \breve{\boldsymbol{\beta}}(\tau)) = \sum_{j=1}^q \left[-D_1^{-1} D_{12}^j \sqrt{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\right] + o_p(1). \tag{A.9}$$

For any $j = 1, \cdots, q$: $\sqrt{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) \xrightarrow{d} N(0, V_j)$ by assumption **A4** and $D_1^{-1} D_{12}^j$ is bounded by assumptions **A2** and **A5**, we have that $-D_1^{-1} D_{12}^j \sqrt{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) \xrightarrow{d} N(0, D_1^{-1} D_{12}^j V_j D_{12}^{j\top} D_1^{-1})$. Therefore, we have by assumption **A3** that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \breve{\boldsymbol{\beta}}(\tau)) \xrightarrow{d} N\left(0, \sum_{j=1}^q \left[D_1^{-1} D_{12}^j V_j D_{12}^{j\top} D_1^{-1}\right]\right). \tag{A.10}$$

Recall from (A.6) that $\sqrt{n}(\breve{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)) = D_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{x}_i \psi_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)) + o_p(1)$. In addition, (A.9) together with (A.8) implies that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\tau) - \breve{\boldsymbol{\beta}}(\tau)) = -D_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n f_i(0|\boldsymbol{x}_i) \sum_{j=1}^q \beta_j(\tau)\boldsymbol{x}_i \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top (\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) + o_p(1).$$

Substituting **A4** where $\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{r}_i(\boldsymbol{\theta}_j)$ in the above equation, we obtain that

$$
\begin{aligned}
&Cov(\sqrt{n}(\breve{\beta}(\tau) - \beta_0(\tau)), \sqrt{n}(\widehat{\beta}(\tau) - \breve{\beta}(\tau))) \\
&\quad = -D_1^{-1} E\{\frac{1}{n}\sum_{i=1}^{n} \psi_\tau(y_i - \boldsymbol{x}_i^\top \beta_0(\tau)) f_i(0|\boldsymbol{x}_i) \sum_{j=1}^{q} \beta_j(\tau) \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j) \boldsymbol{x}_i \boldsymbol{x}_i^\top\} D_1^{-1} \\
&\quad = -D_1^{-1} M D_1^{-1} + o_p(1)
\end{aligned}
$$

where $M = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\{\psi_\tau(y_i - \boldsymbol{x}_i^\top \beta_0(\tau)) f_i(0|\boldsymbol{x}_i) \sum_{j=1}^{q} \beta_j(\tau) \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j) \boldsymbol{x}_i \boldsymbol{x}_i^\top\}$. Finally, combining the above covariance term with (A.7) and (A.10) with (A.5), we obtain

$$
\sqrt{n}(\widehat{\beta}(\tau) - \beta_0(\tau)) \xrightarrow{d} N\left(0, \tau(1-\tau) D_1^{-1} D_0 D_1^{-1} + \sum_{j=1}^{q} \left[D_1^{-1} D_{12}^{j} V_j D_{12}^{j\top} D_1^{-1}\right] - 2D_1^{-1} M D_1^{-1}\right).
$$

□

Proof of Theorem of Consistency of Variance-Covariance Matrix. (i) Claim that $\widehat{D}_0 \xrightarrow{p} D_0$, in other words,

$$
n^{-1}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top - n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top = o_p(1).
$$

By definition:

$$
\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, \cdots, \widehat{x}_{iq}, x_{iq+1}, \cdots, x_{ik})^\top, \boldsymbol{x}_i = (x_{i1}, \cdots, x_{iq}, x_{iq+1}, \cdots, x_{ik})^\top.
$$

We show that $n^{-1}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top$ and $n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ are close to each other element by element:

(a) For $j = 1, \cdots, q$, $j' = 1, \cdots, q$, we have

$$
\widehat{x}_{ij} = g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j),
$$

$$
\widehat{\boldsymbol{\theta}}_j \xrightarrow{p} \boldsymbol{\theta}_j.
$$

Assumption **B3** implies the following uniform convergence by applying the uniform weak law of large number (UWLLN): ($E(\cdot)$ means expectation in terms of joint distribution)

$$n^{-1}\sum_{i=1}^{n} x_{ij}x_{ij'} = n^{-1}\sum_{i=1}^{n} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)g_{j'}(\boldsymbol{w}_{j'i},\boldsymbol{\theta}_{j'}) \xrightarrow{p} E[g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)g_{j'}(\boldsymbol{w}_{j'},\boldsymbol{\theta}_{j'})],$$

and assumption **A3** that $\widehat{\boldsymbol{\theta}}_j \to \boldsymbol{\theta}_j$ for $j = 1,\cdots,q$ implies that

$$n^{-1}\sum_{i=1}^{n} \widehat{x}_{ij}\widehat{x}_{ij'} = n^{-1}\sum_{i=1}^{n} g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)g_{j'}(\boldsymbol{w}_{j'i},\widehat{\boldsymbol{\theta}}_{j'}) \xrightarrow{p} E[g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)g_{j'}(\boldsymbol{w}_{j'},\boldsymbol{\theta}_{j'})].$$

It follows that

$$n^{-1}\sum_{i=1}^{n} \widehat{x}_{ij}\widehat{x}_{ij'} - n^{-1}\sum_{i=1}^{n} x_{ij}x_{ij'} = o_p(1).$$

(b) For $j = 1,\cdots,q$, $s = q+1,\cdots,k$, similarly by applying UWLLN, we have

$$n^{-1}\sum_{i=1}^{n} x_{ij}x_{is} = n^{-1}\sum_{i=1}^{n} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)x_{is} \xrightarrow{p} E[g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)x_{is}],$$

and the consistency of $\boldsymbol{\theta}_j$ implies that

$$n^{-1}\sum_{i=1}^{n} \widehat{x}_{ij}x_{is} = n^{-1}\sum_{i=1}^{n} g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)x_{is} \xrightarrow{p} E[g_j(\boldsymbol{w}_j,\boldsymbol{\theta}_j)x_{is}].$$

It follows that

$$n^{-1}\sum_{i=1}^{n} \widehat{x}_{ij}x_{is} - n^{-1}\sum_{i=1}^{n} x_{ij}x_{is} = o_p(1).$$

After claiming (a) and (b), we have

$$n^{-1}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^{\top} - n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\top} = o_p(1).$$

(ii) Claim that $\widehat{D}_1 \xrightarrow{p} D_1$.

To obtain a consistent estimator of D_1 which contains unknown conditional density function, the estimator $\widehat{D}_1$ replaces the conditional density functions by uniform kernel weights.

Define

$$\widetilde{D}_1 \equiv (2nc_n)^{-1} \sum_{i=1}^{n} I(|\varepsilon_i| < c_n) \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

Denote $\widehat{\varepsilon}_i \equiv y_i - \widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau)$ and $\varepsilon_i \equiv y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)$.

(a) Consider

$$\begin{aligned}\|\widehat{D}_1 - \widetilde{D}_1\| = \Bigg\| (2nc_n)^{-1} \times \sum_{i=1}^{n} \Bigg\{ & [I(|\widehat{\varepsilon}_i| < c_n) - I(|\varepsilon_i| < c_n)] \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top \\ & + I(|\varepsilon_i| < c_n)[\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i] \widehat{\boldsymbol{x}}_i^\top + I(|\varepsilon_i| < c_n) \boldsymbol{x}_i [\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i]^\top \Bigg\} \Bigg\|.\end{aligned}$$

Notice that

$$\begin{aligned}|I(|\widehat{\varepsilon}_i| < c_n) - I(|\varepsilon_i| < c_n)| \leqslant\ & I(|\varepsilon_i - c_n| < |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)|) \\ & + I(|\varepsilon_i + c_n| < |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)|),\end{aligned} \tag{A.11}$$

and

$$\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i\| \leqslant \max_j \{\|\nabla g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j^*)\| \cdot \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|\}. \tag{A.12}$$

Combining (A.11) and (A.12), we have

$$\begin{aligned}
\|\widehat{D}_1 - \widetilde{D}_1\| \leqslant (2nc_n)^{-1} \times \sum_{i=1}^{n} \Bigg\{ \Bigg[& I(|\varepsilon_i - c_n| < |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)|) \\
& + I(|\varepsilon_i + c_n| < |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)|) \Bigg] ||\widehat{\boldsymbol{x}}_i||^2 \\
& + I(|\varepsilon_i| < c_n) \max_j \{\|\nabla g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j^*)\| \cdot \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|\} \|\widehat{\boldsymbol{x}}_i\| \\
& + I(|\varepsilon_i| < c_n) ||\widehat{\boldsymbol{x}}_i|| \max_j \{\|\nabla g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j^*)\| \cdot \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|\} \Bigg\} \\
& \equiv T_1 + 2T_2.
\end{aligned}$$

For any $\delta > 0$ and any $j \in \{1, \cdots, q\}$, n is large enough that

$$\begin{aligned}
c_n &< 1, \\
c_n^{-1} \cdot \|\boldsymbol{\beta}_0(\tau)\| \cdot \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\| &< \delta, \\
c_n^{-1} \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\| &< \delta.
\end{aligned} \tag{A.13}$$

We can find a bound for $\|\widehat{D}_1 - \widetilde{D}_1\|$ such that it can be arbitrarily small in probability if δ is chosen sufficiently small. Since the bound for T_2 is easy to derive, we only show that $E(T_1) = O(\delta)$.

To show this, applying assumption **B3**, (A.12) and (A.13), we have

$$\begin{aligned}
|\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)| &= |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(\tau) + \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)| \\
&\leqslant |\widehat{\boldsymbol{x}}_i^\top (\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau))| + |(\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i)^\top \boldsymbol{\beta}_0(\tau)| \\
&\leqslant \|\widehat{\boldsymbol{x}}_i\| \cdot \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\| + \max_j \{\|\nabla g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j^*)\| \cdot \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|\} \cdot \|\boldsymbol{\beta}_0(\tau)\| \\
&\leqslant \delta c_n \|\widehat{\boldsymbol{x}}_i\| + \delta c_n \max_j D_j(\boldsymbol{w}_j) \\
&= \delta c_n [\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)].
\end{aligned}$$

Inequality (A.11) can be rewritten as

$$|I(|\widehat{\varepsilon}_i| < c_n) - I(|\varepsilon_i| < c_n)| \leqslant I(|\varepsilon_i - c_n| < \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]) \qquad \text{(A.14)}$$
$$+ I(|\varepsilon_i + c_n| < \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]).$$

Then we have

$$\begin{aligned}
E(T_1) &\leqslant E\Bigg\{(2nc_n)^{-1}\sum_{i=1}^{n}\Bigg[I(|\varepsilon_i - c_n| < \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]) \\
&\quad + I(|\varepsilon_i + c_n| < \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)])\Bigg]\|\widehat{\boldsymbol{x}}_i\|^2\Bigg\} \\
&\leqslant E\Bigg\{\|\widehat{\boldsymbol{x}}_i\|^2(2nc_n)^{-1}\sum_{i=1}^{n}2\int_{c_n-\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_j D_j(\boldsymbol{w}_j)]}^{c_n+\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_j D_j(\boldsymbol{w}_j)]} f_i(\lambda|\boldsymbol{x}_i)d\lambda\Bigg\} \\
&\leqslant E\Bigg\{\|\widehat{\boldsymbol{x}}_i\|^2(2nc_n)^{-1}4\delta c_n\sum_{i=1}^{n}[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]A_f(\boldsymbol{x}_i)\Bigg\} \\
&\leqslant n^{-1}\sum_{i=1}^{n}2\cdot\delta\cdot A = 2A\cdot\delta = O(\delta),
\end{aligned}$$

where A is some bound for $E[(\|\widehat{\boldsymbol{x}}_i\|^3 + \max_j D_j(\boldsymbol{w}_j)\|\widehat{\boldsymbol{x}}_i\|^2)A_f(\boldsymbol{x}_i)]$ by assumption **B4**.

(b) We now show $\widetilde{D}_1 - D_1 = o_p(1)$. Note that

$$\begin{aligned}
\widetilde{D}_1 - D_1 &= (2nc_n)^{-1}\sum_{i=1}^{n}\Big\{I(|\varepsilon_i| < c_n)\boldsymbol{x}_i\boldsymbol{x}_i^\top - E[I(|\varepsilon_i| < c_n)|\boldsymbol{x}_i]\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big\} \\
&\quad + n^{-1}\sum_{i=1}^{n}\Big\{(2c_n)^{-1}E[I(|\varepsilon_i| < c_n)|\boldsymbol{x}_i]\boldsymbol{x}_i\boldsymbol{x}_i^\top - [f_i(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top]\Big\}.
\end{aligned}$$

For the first term, it has zero expectation and if $\boldsymbol{a}$ and $\boldsymbol{b}$ are k-dimensional vectors with

$\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = 1$ and consider the variance of the first term,

$$
\begin{aligned}
&Var[\boldsymbol{a}^\top\{(2nc_n)^{-1}\sum_{i=1}^{n} I(|\varepsilon_i| < c_n)\boldsymbol{x}_i\boldsymbol{x}_i^\top - E[I(\varepsilon_i| < c_n)|\boldsymbol{x}_i]\boldsymbol{x}_i\boldsymbol{x}_i^\top\}\boldsymbol{b}] \\
&= E\left\{(2nc_n)^{-1}\sum_{i=1}^{n} I(|\varepsilon_i| < c_n)\boldsymbol{a}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{b} - E[I(\varepsilon_i| < c_n)|\boldsymbol{x}_i]\boldsymbol{a}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{b}\right\}^2 \\
&= (2nc_n)^{-2}E\left\{\sum_{i=1}^{n}\{I(|\varepsilon_i < c_n) - E[I(\varepsilon_i| < c_n)|\boldsymbol{x}_i]\}^2 \times [\boldsymbol{a}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{b}]^2\right\} \\
&\leqslant (2nc_n)^{-2}\sum_{i=1}^{n} E[\|\boldsymbol{x}_i\|\|\boldsymbol{x}_i\|]^2 \\
&= (4nc_n^2)^{-1}H_1 = o(1),
\end{aligned}
$$

where the second equality holds because all the cross-product terms are zero by the law of iterated expectations, and H_1 is some bound for $E(\|\boldsymbol{x}_i\|^4)$ by assumption **B2**. Therefore, the first term converges in quadratic mean to zero, which implies that it converges to zero in probability.

For the second term,

$$
\begin{aligned}
|(2c_n)^{-1}E[I(|\varepsilon_i| < c_n)|\boldsymbol{x}_i] - f_{\varepsilon_i}(0|\boldsymbol{x}_i)| &= \left|(2c_n)^{-1}\int_{-c_n}^{c_n} f_i(\lambda|\boldsymbol{x}_i)d\lambda - f_i(0|\boldsymbol{x}_i)\right| \\
&\leqslant |(2c_n)^{-1}2c_n f_i(\lambda^*|\boldsymbol{x}_i) - f_i(0|\boldsymbol{x}_i)| \\
&\leqslant L_i|c_n| = o_p(1),
\end{aligned}
$$

where $f_i(\lambda^*|\boldsymbol{x}_i) \equiv \max_{\lambda\in[-c_n,c_n]} f_i(\lambda|\boldsymbol{x}_i)$ and the last inequality uses assumption **B5**.

Combining claim (a) and (b), we have $\widehat{D}_1 - D_1 = o_p(1)$.

(iii) Claim that $\widehat{D}_{12}^j - D_{12}^j = o_p(1)$ for $j = 1, \cdots, q$:

$$
\begin{aligned}
\widehat{D}_{12}^j &= (2nc_n)^{-1}\sum_{i=1}^{n} I(|\widehat{\varepsilon}_i| < c_n)\widehat{\beta}_j(\tau)\widehat{\boldsymbol{x}}_i\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top, \\
D_{12}^j &= n^{-1}\sum_{i=1}^{n} f_i(0|\boldsymbol{x}_i)\beta_j(\tau)\boldsymbol{x}_i\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top.
\end{aligned}
$$

Define

$$\widetilde{D}_{12}^{j} \equiv (2nc_n)^{-1}\sum_{i=1}^{n} I(|\varepsilon_i| < c_n)\beta_j(\tau)\boldsymbol{x}_i\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top,$$

where $\widehat{\varepsilon}_i \equiv y_i - \widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau)$ and $\varepsilon_i \equiv y_i - \boldsymbol{x}_i^\top \beta_0(\tau)$.

Similar to Part (ii), we need to show that $\|\widehat{D}_{12}^{j} - \widetilde{D}_{12}^{j}\| = o_p(1)$ and $\|\widetilde{D}_{12}^{j} - D_{12}^{j}\| = o_p(1)$. $\|\widetilde{D}_{12}^{j} - D_{12}^{j}\| = o_p(1)$ is easy to derive by following the idea in Part (ii)(b), we only show that $\|\widehat{D}_{12}^{j} - \widetilde{D}_{12}^{j}\| = o_p(1)$.

$$\begin{aligned}\|\widehat{D}_{12}^{j} - \widetilde{D}_{12}^{j}\| = \Bigg\|(2nc_n)^{-1} \times \sum_{i=1}^{n}\Big\{&[I(|\widehat{\varepsilon}_i| < c_n) - I(|\varepsilon_i| < c_n)]\widehat{\beta}_j(\tau)\widehat{\boldsymbol{x}}_i\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \\ &+ I(|\varepsilon_i| < c_n)(\widehat{\beta}_j(\tau) - \beta_j(\tau))\widehat{\boldsymbol{x}}_i\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \\ &+ I(|\varepsilon_i| < c_n)\beta_j(\tau)[\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i]\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \\ &+ I(|\varepsilon_i| < c_n)\beta_j(\tau)\boldsymbol{x}_i[\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top - \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top]\Big\}\Bigg\|.\end{aligned}$$

Applying assumption **B3**, (A.12) and (A.13), we have

$$\begin{aligned}||\beta_j(\tau)[\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i]\nabla_{\boldsymbol{\theta}_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)|| &= \beta_j(\tau)\max_{j'}\{||\widehat{\boldsymbol{\theta}}_{j'} - \boldsymbol{\theta}_{j'}|| \cdot ||\nabla_{\boldsymbol{\theta}_{j'}} g_{j'}(\boldsymbol{w}_{j'i}, \boldsymbol{\theta}_{j'}^*)||\} \cdot ||\nabla_{\boldsymbol{\theta}_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)|| \\ &\leqslant \delta c_n \max_{j'} D_{j'}(\boldsymbol{w}_{j'}) \cdot D_j(\boldsymbol{w}_j),\end{aligned}$$

and using assumption **B3** and (A.13), we have

$$\begin{aligned}||I(|\varepsilon_i| < c_n)\beta_j(\tau)\boldsymbol{x}_i[\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top - \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \boldsymbol{\theta}_j)^\top]|| &\leqslant I(|\varepsilon_i| < c_n)\beta_j(\tau)||\boldsymbol{x}_i|| \cdot J_j||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j|| \\ &\leqslant I(|\varepsilon_i| < c_n)\delta c_n J_j\|\boldsymbol{x}_i\|.\end{aligned}$$

Combining the above two inequalities, assumption **B3** and (A.14), we have

$$\|\widehat{D}_{12}^{j}-\widetilde{D}_{12}^{j}\| \leqslant (2nc_n)^{-1}\sum_{i=1}^{n}\Bigg\{\Bigg[I(|\varepsilon_i-c_n|<\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})])$$
$$+I(|\varepsilon_i+c_n|<\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})])\Bigg]|\widehat{\beta}_j(\tau)|\cdot\|\widehat{\boldsymbol{x}}_i\|D_j(\boldsymbol{w}_j)$$
$$+I(|\varepsilon_i|<c_n)\delta c_n\|\widehat{\boldsymbol{x}}_i\|D_j(\boldsymbol{w}_j)$$
$$+I(|\varepsilon_i|<c_n)\delta c_n\max_{j'}D_{j'}(\boldsymbol{w}_{j'})D_j(\boldsymbol{w}_j)$$
$$+I(|\varepsilon_i|<c_n)\delta c_n J_j\|\boldsymbol{x}_i\|\Bigg\}$$
$$\equiv N_1+N_2+N_3+N_4.$$

Similar to Part (ii)(a), we can find a bound for $\|\widehat{D}_{12}-\widetilde{D}_{12}\|$, such that it can be arbitrarily small in probability if δ is chosen sufficiently small. Since the bounds for N_2, N_3 and N_4 are easy to derive, we only show that $E(N_1)=O(\delta)$ as follows:

$$E(N_1)=E\Bigg\{(2nc_n)^{-1}\sum_{i=1}^{n}\Bigg[I(|\varepsilon_i-c_n|<\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})])$$
$$+I(|\varepsilon_i+c_n|<\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})])\Bigg]|\widehat{\beta}_j(\tau)|\cdot\|\widehat{\boldsymbol{x}}_i\|D_j(\boldsymbol{w}_j)\Bigg\}$$
$$\leqslant E\Bigg\{|\widehat{\beta}_j(\tau)|\cdot\|\widehat{\boldsymbol{x}}_i\|D_j(\boldsymbol{w}_j)(2nc_n)^{-1}\sum_{i=1}^{n}2\int_{c_n-\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})]}^{c_n+\delta c_n[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})]}f_i(\lambda|\boldsymbol{x}_i)d\lambda\Bigg\}$$
$$\leqslant E\Bigg\{|\widehat{\beta}_j(\tau)|\cdot\|\widehat{\boldsymbol{x}}_i\|D_j(\boldsymbol{w}_j)(2nc_n)^{-1}4\delta c_n\sum_{i=1}^{n}[\|\widehat{\boldsymbol{x}}_i\|+\max_{j'}D_{j'}(\boldsymbol{w}_{j'})]A_f(\boldsymbol{x}_i)\Bigg\}$$
$$\leqslant n^{-1}\sum_{i=1}^{n}2|\widehat{\beta}_j(\tau)|\cdot\delta\cdot B_j=O(\delta),$$

where the last inequality holds by assumption **B4**.

(iv) Claim that $\widehat{M} \xrightarrow{p} M$.

$$
\begin{aligned}
M &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_\tau\left(y_i - \boldsymbol{x}_i^\top \beta_0(\tau)\right) f_i(0|\boldsymbol{x}_i) \sum_{j=1}^{q} \beta_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\} \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left\{[\tau - I(\varepsilon_i<0)] f_i(0|\boldsymbol{x}_i) \sum_{j=1}^{q} \beta_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\} \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left\{[\tau f_i(0|\boldsymbol{x}_i) - I(\varepsilon_i<0) f_i(0|\boldsymbol{x}_i)] \sum_{j=1}^{q} \beta_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\}.
\end{aligned}
$$

To obtain a consistent estimator of M which contains unknown conditional density function, the estimator $\widehat{M}$ replaces the conditional density functions by uniform kernel weights:

$$
\begin{aligned}
\widehat{M} &= \frac{1}{2nc_n}\sum_{i=1}^{n}\left\{[\tau - I(\widehat{\varepsilon}_i<0)] I(|\widehat{\varepsilon}_i|<c_n) \sum_{j=1}^{q} \widehat{\beta}_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top\right\} \\
&= \frac{1}{2nc_n}\sum_{i=1}^{n}\left\{[\tau I(|\widehat{\varepsilon}_i|<c_n) - I(\widehat{\varepsilon}_i<0) I(|\widehat{\varepsilon}_i|<c_n)] \sum_{j=1}^{q} \widehat{\beta}_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top\right\}.
\end{aligned}
$$

Define

$$
\begin{aligned}
\widetilde{M} &= \frac{1}{2nc_n}\sum_{i=1}^{n}\left\{[\tau - I(\varepsilon_i<0)] I(|\varepsilon_i|<c_n) \sum_{j=1}^{q} \beta_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\} \\
&= \frac{1}{2nc_n}\sum_{i=1}^{n}\left\{[\tau I(|\varepsilon_i|<c_n) - I(\varepsilon_i<0) I(|\varepsilon_i|<c_n)] \sum_{j=1}^{q} \beta_j(\tau)\nabla_{\theta_j} g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top \boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\}.
\end{aligned}
$$

Denote $\widehat{\varepsilon}_i \equiv y_i - \widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau)$ and $\varepsilon_i \equiv y_i - \boldsymbol{x}_i^\top \beta_0(\tau)$.

Similar to Part (ii), we need to show that $\|\widehat{M} - \widetilde{M}\| = o_p(1)$ and $\|\widetilde{M} - M\| = o_p(1)$. $\|\widetilde{M} - M\| = o_p(1)$ is easy to derive by following the idea in Part (ii)(b), we only show that $\|\widehat{M} - \widetilde{M}\| = o_p(1)$.

Consider

$$\|\widehat{M}-\widetilde{M}\| = \|(2nc_n)^{-1}\times\sum_{i=1}^{n}\{[\tau I(|\widehat{\varepsilon}_i|<c_n)-I(\widehat{\varepsilon}_i<0)I(|\widehat{\varepsilon}_i|<c_n)]\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top$$

$$-[\tau I(|\varepsilon_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]\sum_{j=1}^{q}\beta_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top\boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\|\}$$

$$=||(2nc_n)^{-1}\times\sum_{i=1}^{n}\Big\{\Big([\tau I(|\widehat{\varepsilon}_i|<c_n)-\tau I(|\varepsilon_i|<c_n)]$$

$$-[I(\widehat{\varepsilon}_i<0)I(|\widehat{\varepsilon}_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]\Big)\times\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top$$

$$+[\tau I(|\varepsilon_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]$$

$$\times\left[\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top-\sum_{j=1}^{q}\beta_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top\boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right]\Big\}\|$$

$$\leqslant\|(2nc_n)^{-1}\times\sum_{i=1}^{n}\{[\tau I(|\widehat{\varepsilon}_i|<c_n)-\tau I(|\varepsilon_i|<c_n)]$$

$$-[I(\widehat{\varepsilon}_i<0)I(|\widehat{\varepsilon}_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]\}$$

$$\times\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top\|$$

$$+\|(2nc_n)^{-1}\times\sum_{i=1}^{n}[\tau I(|\varepsilon_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]\times[\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i\top$$

$$-\sum_{j=1}^{q}\beta_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top\boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top]\|$$

$$\leqslant\|(2nc_n)^{-1}\times\sum_{i=1}^{n}\{[\tau I(|\widehat{\varepsilon}_i|<c_n)-\tau I(|\varepsilon_i|<c_n)]\times\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top\}\|$$

$$+\|[(2nc_n)^{-1}\times\sum_{i=1}^{n}I(\widehat{\varepsilon}_i<0)I(|\widehat{\varepsilon}_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]$$

$$\times\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top\}\|$$

$$+\|(2nc_n)^{-1}\times\sum_{i=1}^{n}[\tau I(|\varepsilon_i|<c_n)-I(\varepsilon_i<0)I(|\varepsilon_i|<c_n)]\times[\sum_{j=1}^{q}\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top$$

$$-\sum_{j=1}^{q}\beta_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\boldsymbol{\theta}_j)^\top\boldsymbol{r}_i(\boldsymbol{\theta}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top]\|$$

$$=Z_1+Z_2+Z_3.$$

Notice that

$$I(\widehat{\varepsilon}_i < 0)I(|\widehat{\varepsilon}_i| < c_n) - I(\varepsilon_i < 0)I(|\varepsilon_i| < c_n) = I(-c_n < \widehat{\varepsilon}_i < 0) - I(-c_n < \varepsilon_i < 0)$$

and

$$|I(-c_n < \widehat{\varepsilon}_i < 0) - I(-c_n < \varepsilon_i < 0)| \leqslant I(|\varepsilon_i + c_n| < |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)|)$$
$$+ I(|\varepsilon_i| < |\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)|).$$

From (8.12), we already have

$$|\widehat{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{x}_i^\top \boldsymbol{\beta}_0(\tau)| \leqslant \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)].$$

We can find a bound for $\|\widehat{M} - \widetilde{M}\|$ such that it can be arbitrarily small in probability if δ is chosen sufficiently small. Since the bounds for Z_1 and Z_3 are easy to derive using the similar argument we used in part (iii), we only show that $E(Z_2) = O(\delta)$ as following:

$$E(Z_2) \leqslant E\{(2nc_n)^{-1} \times \sum_{i=1}^{n} \big[I(|\varepsilon_i + c_n| < \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)])$$
$$+ I(|\varepsilon_i| < \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)])\big] \|\sum_{j=1}^{q} \widehat{\beta}_j(\tau) \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j) \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top \|\}$$
$$\leqslant E\{\|\sum_{j=1}^{q} \widehat{\beta}_j(\tau) \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j) \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top \|(2nc_n)^{-1}$$
$$\times \sum_{i=1}^{n} [\int_{c_n - \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]}^{c_n + \delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]} f_i(\lambda|\boldsymbol{x}_i) d\lambda + \int_{-\delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]}^{\delta c_n[\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)]} f_i(\lambda|\boldsymbol{x}_i) d\lambda\}$$
$$\leqslant E\{\|\sum_{j=1}^{q} \widehat{\beta}_j(\tau) \nabla_{\theta_j} g_j(\boldsymbol{w}_{ji}, \widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j) \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top \|(2nc_n)^{-1}$$
$$\times 4\delta c_n \sum_{i=1}^{n} [\|\widehat{\boldsymbol{x}}_i\| + \max_j D_j(\boldsymbol{w}_j)] A_f(\boldsymbol{x}_i)\}$$
$$\leqslant n^{-1} \sum_{i=1}^{n} 2 \cdot \delta \cdot J = 2J \cdot \delta = O(\delta)$$

where J is some bound for $E\left[\sum_{j=1}^{q}\|\widehat{\beta}_j(\tau)\nabla_{\theta_j}g_j(\boldsymbol{w}_{ji},\widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^\top\|\cdot[\|\widehat{\boldsymbol{x}}_i\|+\max_j D_j(\boldsymbol{w}_j)]A_f(\boldsymbol{x}_i)\right]$ by assumptions **A2**, **A4** and **A5**.

□

Proof of Theorem of Wald Test. The proof of Theorem of Wald Test is simple; it follows from observing that for any fixed τ, by Theorem of Normality

$$\sqrt{n}(\widehat{\beta}(\tau)-\beta_0(\tau)) \xrightarrow{d} N(0,\Omega(\tau)),$$

and under the null hypothesis,

$$\sqrt{n}(R\widehat{\beta}(\tau)-r) \xrightarrow{d} N(0,R\Omega(\tau)R^\top).$$

Since by Theorem of Asymptotic Variance, $\widehat{\Omega}(\tau)$ is a consistent estimator of $\Omega(\tau)$, by the Slutsky's theorem,

$$\mathscr{W}_T = T(R\widehat{\beta}(\tau)-r)^\top[R\widehat{\Omega}(\tau)R^\top]^{-1}(R\widehat{\beta}(\tau)-r) \overset{a}{\sim} \chi^2_s.$$

□

Ols With Generated Regressors

In order to compare our quantile regression with generated regressor framework with the OLS with generated regressor (OLS-GR), we include the detailed assumptions and propositions for the OLS case in this section.

$$y = \boldsymbol{x}^\top\beta_0 + u \tag{A.15}$$

where $E(u|\boldsymbol{x}) = 0$, β_0 is a $k\times 1$ vector, $\boldsymbol{x} = h(\boldsymbol{w},\boldsymbol{\delta})$, $\boldsymbol{\delta}$ is a $p\times 1$ vector and $h(\cdot)$ is known but $\boldsymbol{\delta}$ is unknown and $E(u|\boldsymbol{w}) = 0$.

Let $\widehat{\boldsymbol{\delta}}$ be a $\sqrt{n}$-consistent estimator of $\boldsymbol{\delta}$, and we obtain the generated regressors as $\widehat{\boldsymbol{x}}_i =$

$h(\boldsymbol{w}_i, \widehat{\boldsymbol{\delta}})$. Let $\widehat{\boldsymbol{\beta}}$ be the OLS-GR estimator from the equation

$$y_i = \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + e_i,$$

where $\widehat{\boldsymbol{x}}_i = h(\boldsymbol{w}_i, \widehat{\boldsymbol{\delta}})$. We impose the following regularity conditions.

Conditions:

C1. $n \geqslant k$, $rank(\boldsymbol{X}) = k$.

C2. $E(u|\boldsymbol{x}) = 0$.

C3. The observations $\{(y_i, \boldsymbol{x}_i) : i = 1, 2, \cdots, n\}$ are i.i.d. across i with $E(y_i^{4+\varepsilon}) < \infty$ and $E(||\boldsymbol{x}_i||^{4+\varepsilon}) < \infty$ for some $\varepsilon > 0$, $\text{plim}_{n\to\infty} n^{-1} \sum_{i=1}^n (\boldsymbol{x}_i \boldsymbol{x}_i^\top) = H$ is a nonsingular matrix, and $E(u|\boldsymbol{w}) = 0$.

C4. $\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = n^{-1/2} \sum_{i=1}^n \boldsymbol{r}_i(\boldsymbol{\delta}) + o_p(1)$, where $E[r_i(\boldsymbol{\delta})] = 0$.

C5. $\text{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^n (\boldsymbol{\beta} \otimes \boldsymbol{x}_i^\top)^\top \nabla_{\boldsymbol{\delta}} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top = G$, which is bounded.

We first state the consistency of the OLS-GR estimator.

Proposition 1. In model (A.15), under the assumptions **C1–C3**, the OLS-GR estimator is consistent, i.e., $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

Proof of Proposition 1. Recall that

$$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{X}}^\top \widehat{\boldsymbol{X}})^{-1} \widehat{\boldsymbol{X}}^\top y = (\sum_{i=1}^n \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top)^{-1} [\sum_{i=1}^n \widehat{\boldsymbol{x}}_i y_i].$$

Thus,

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\frac{1}{n} \sum_{i=1}^n \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top)^{-1} \{\frac{1}{n} \sum_{i=1}^n \widehat{\boldsymbol{x}}_i [(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_0 + u_i]\} = \widehat{H}^{-1} \{\frac{1}{n} \sum_{i=1}^n \widehat{\boldsymbol{x}}_i [(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_0 + u_i]\},$$

where $\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top$. Using the mean value expansion, we have

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i u_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i u_i + [\frac{1}{n}\sum_{i=1}^{n} \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top u_i](\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + o_p(1).$$

Since $E(u|\boldsymbol{w}) = 0$, $E[\nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top u_i] = 0$. It implies that $\frac{1}{n}\sum_{i=1}^{n} \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top u_i = o_p(1)$. And since $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} = o_p(1)$ and $E(\boldsymbol{x}_i u_i) = 0$, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i u_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i u_i + o_p(1) = o_p(1).$$

Or,

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i (\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \beta_0 = -[\frac{1}{n}\sum_{i=1}^{n} (\beta_0 \otimes \boldsymbol{x}_i^\top)^\top \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top](\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + o_p(1) = -G(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + o_p(1),$$

where $G = E[(\beta_0 \otimes \boldsymbol{x}_i^\top)^\top \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top]$, using $\widehat{\boldsymbol{x}}_i (\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \beta_0 = (\beta_0 \otimes \widehat{\boldsymbol{x}}_i^\top)^\top (\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)$. Since $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} = o_p(1)$ and G is bounded, then

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i (\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \beta_0 = o_p(1).$$

Therefore, $\widehat{\boldsymbol{\beta}} - \beta_0 = o_p(1)$. □

Next, we state the asymptotic normality of the OLS-GR estimator.

Proposition 2. In model (A.15), under the assumptions **C4–C5** and the conditions in Proposition 1, the OLS-GR is asymptotically normal, i.e.,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, H^{-1} M H^{-1}),$$

where $M = Var[\boldsymbol{x}_i u_i - G r_i(\boldsymbol{\delta})]$.

Proof of Proposition 2. Note that from

$$\widehat{\boldsymbol{\beta}} = (\widehat{X}^\top \widehat{X})^{-1}\widehat{X}^\top y = (\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top)^{-1}[\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i y_i],$$

by rearranging we obtain

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= (\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top)^{-1}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i[(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_0 + u_i]\} \\ &= \widehat{H}^{-1}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i[(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_0 + u_i]\},\end{aligned}$$

where $\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top$. Using the mean value expansion, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i u_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i u_i + [\frac{1}{n}\sum_{i=1}^{n} \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top u_i]\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + o_p(1).$$

Since $E(u|\boldsymbol{w}) = 0$, $E[\nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top u_i] = 0$. It implies that $\frac{1}{n}\sum_{i=1}^{n} \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top u_i = o_p(1)$. And since $\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = O_p(1)$, it follows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i u_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i u_i + o_p(1).$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{\boldsymbol{x}}_i(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_0 = -[\frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{\beta}_0 \otimes \boldsymbol{x}_i^\top)^\top \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top]\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + o_p(1) = -G\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + o_p(1),$$

where $G = E[(\boldsymbol{\beta}_0 \otimes \boldsymbol{x}_i^\top)^\top \nabla_{\delta} h(\boldsymbol{w}_i, \boldsymbol{\delta})^\top]$, using $\widehat{\boldsymbol{x}}_i(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^\top \boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0 \otimes \widehat{\boldsymbol{x}}_i^\top)^\top (\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)$. By Assumption **C4**,

$$\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = n^{-1/2}\sum_{i=1}^{n} \boldsymbol{r}_i(\boldsymbol{\delta}) + o_p(1).$$

Hence, $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\boldsymbol{x}_i u_i - G\boldsymbol{r}_i(\boldsymbol{\delta})] + o_p(1)$.

By the Central Limit Theorem , we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, H^{-1}MH^{-1}),$$

where $M = Var[\boldsymbol{x}_i u_i - G\boldsymbol{r}_i(\boldsymbol{\delta})]$. □