Essays on Health, Education and Behavioral Choices

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Meng Konishi(Zhao)

IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

PAUL GLEWWE, ADVISER

April, 2010

Acknowledgement

This short passage is dedicated to many people, to whom I would like to express my sincerest appreciation for their kind support throughout my doctoral program.

I am deeply indebted to my adviser Paul Glewwe for his advice, mentorship, and encouragement. He was not only a valuable thesis adviser, but also a supervisor for my research assistantship. Paul spent a large amount of time discussing this thesis as well as other research projects. Without his valuable advice and guidance, this thesis would not be accomplished.

I would like to thank Roger Feldman, Terry Roe, and Baolin Wu for serving in my committee and their helpful comments on this paper.

Numerous professors provided helpful comments and suggestions at various stages of my dissertation work. I truly appreciate Elizabeth Davis, Qiuqiong Huang, Jean Kinsey, Donald Liu, Glenn Pederson, Ford Runge, Ben Senauer, Judy Temple, and Chengyan Yue for their suggestions.

I have also had wonderful friends and colleagues here in Minnesota: Kenji Adachi, Swati Agiwal, Amy Damon, Jennifer Drew, Qihui Chen, Tetsuya Horie, Qinlei Huang, Yang Daisy Liu, Shefali Mehta, Rocky Oishi, Kyong Park, Uttam Sharma, Shinya Takamatsu, Minh Wendt, Haochi Zheng. I sincerely appreciate their friendship and the time we spent together. Special thanks are due to Kyong and Qinlei. Kyong and I studied together for statistics during the first year of my doctoral study, and she has been a valuable friend since then. Qinlei has always been a good friend, and a good listener when I need to talk myself out of daily stress.

I also want to take this opportunity to thank the Hueg-Harrison Fellowship for providing me financial support in 2006. I owe special thank to Dr. William Hueg for his generous contribution to this fellowship.

I have spent the last two years of my doctoral study off campus, residing in western Massachusetts. I am sincerely grateful to Williams College for providing me office space, access to library and many other research resources during these two years.

Lastly, I would like to thank my host family in Minnesota, the Knudson's, my husband Yoshi, my parents and sister for their encouragement and support. I would not have performed well without their constant support.

Thesis Abstract

My dissertation is composed of two essays that investigate the interrelationship between consumers' health, education, behavioral choices, and perceptions. The first essay evaluates the impact of teenage smoking on schooling and estimates the lifetime income loss due to lower educational achievement and attainment caused by youth smoking. Using unusually rich data from China, the study shows that youth smoking can biologically reduce learning productivity and discourage motivation to go to school (where smoking is forbidden), resulting in lower educational outcomes and, consequently, reduced lifetime income. The second essay empirically analyzes the effect of a doctor diagnosis of hypertension (high blood pressure) on food demand and nutrient intake. The study shows that three quarters of the hypertensive population in China are unaware of their condition. Adoctor's diagnosis can lead consumers to update their perceptions about their health and, therefore, make better decisions for their food choices. The study finds that, after a diagnosis of hypertension, consumers significantly reduce their daily fat intake, especially the consumption of animal oil and pork. The effect is stronger for 2004 data, compared to the 1997 and 2000 data. This suggests that consumers have become more health conscious in recent years.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

> "Some activities primarily affect future well-being; the main impact of others is in the present. Some affect money income and others psychic income, that is, consumption. Sailing primarily affects consumption, on-the-job training primarily affects money income, and a college education could affect both. These effects may operate either through physical resources or through human resources." — *Gary Becker*, 1964.

This dissertation is concerned with investments that can affect future incomes through the increase of human capital. Human capital is a term used to describe the stock of knowledge, skills and competence embodied in humans' ability to perform labor. Analogous to physical capital, investment in human capital will be rewarded by monetary or psychic income in the future. There are many forms of investments in human capital, including schooling, on-the-job training, medical care, eating healthy foods, and migration. They differ in their impacts on income and consumption, in the perceived relationship between an investment and its return, and in the amount that is typically invested. But all of these investments can improve knowledge, skills and health, and, therefore, raise future incomes. This dissertation focuses on two most important forms of human capital: education and health.

Education is probably the most important form of human capital. It has been identified by a substantial body of literature in the past three decades as one of the most important sources of economic growth (Lucas, 1988; Barro, 1991; Mankiw, Romer, and Weil, 1992; World Bank, 2001; Krueger and Lindahl, 2001; UNDP, 2003). From a microeconomic perspective, it plays an important role in raising income and, more generally, well-being. Investment in health started to gain the attention of economists in the 1970's, after Grossman's seminal work on the demand for health (1972). Better health can increase labor productivity in many ways. Healthier laborers are more productive in agricultural activities. Eating healthy food can increase an individual's strength and stamina. A lower death rate can extend a person's working life and, therefore, increase lifetime earnings. Different forms of human capital can also interact with each other. For example, health can influence learning ability and educational outcomes, while education may enhance people's knowledge about their health and their efficiency in collecting health information, and so improve their health.

This dissertation is mainly concerned with consumers' investments in education and health in a developing country. Approximately 80% of the world's population live in developing countries. According to the World Bank, countries with a GNI per capita of less than $975 are classified as low-income economies and those with a GNI per capita between $975 and $11,905 are middle-income economies. Both groups are referred to as developing economies. In the September of 2000, eight Millennium Development Goals (MDGs) have been adopted by the United Nation Millennium Summit, two of which focus on education and three of which focus on health. The first is to complete primary education for every child by 2015. The second is to eliminate gender disparity at all education levels by 2015. The third MDG is to reduce the under-five mortality rate by two-thirds between 1990 and 2015. The fourth is to reduce, by 2015, maternal mortality ratio by three quarters and achieve universal access to reproductive health. The fifth is to have halted and begun to reverse the incidence of HIV/AIDS, malaria and other major diseases.

Education in developing countries has improved significantly since the 1960's because of the awareness of its importance. The primary school gross enrollment rate has increased from 65 in 1960 to 102 in 2000 for low-income countries and from 83 to 110 for middle-income countries (Glewwe and Kremer, 2006). Secondary education has experienced an even greater improvement: the secondary gross enrollment rate has increased from 14 to 55 for low-income countries and from 21 to 77 for middle-income countries. However, despite this significant progress, educational attainment and achievement in developing countries are still much lower, compared to those in developed countries. Approximately 15% of the children of the ages associated with primary education are not enrolled in school in low and middle income countries. The rate for secondary education is as high as 55%. On the contrary, almost every child obtains primary education in developed countries and only 10% of the population of the ages associated with secondary education are not enrolled in school. On average, adults aged 15 or older obtain only 5 years of schooling in developing countries, about half of what their counterparts in developed countries usually achieve. (Glewwe and Kremer, 2006)

The quality of schools and teachers is also a problem in developing countries, which usually results in a low quality of education. That is, students learn less per year of schooling than they are expected to because of poor quality of schools and teachers (Lockheed and Verspoor, 1991; Harbison and Hanushek, 1992; Hanushek, 1995; Glewwe, 1999). Some schools in developing countries, especially in rural areas, do not have adequate teachers, basic equipment or school supplies, such as chairs, desks, blackboard, etc. According to the findings of several studies that compare educational achievement across countries, students from developing countries generally score lower in standardized tests on Mathematics, science and reading, compared to their counterparts in developed countries (Glewwe and Kremer, 2006).

Health in developing countries has also been improved significantly. Over the past several decades, many governments of developing countries and international organizations have engaged in providing people in developing countries access to better health care, improved water and sanitation, insecticide-treated mosquito nets for malaria, etc. Consequently, child mortality has been reduced by 27%, from 12.5 million in 1990 to 9 million in 2007 (WHO, 2009b). Life expectancy in both low-income and middle-income countries has also increased by 2-5 years from 1990 to 2007. The prevalence of HIV/AIDS, tuberculosis, and malaria has also decreased slowly over years due to better prevention and treatment.

Infectious diseases are usually the focus of health problem in developing countries.

However, concerns about chronic diseases in developing countries have grown in recent years. According to the World Health Organization, the number of people dying from all infectious diseases (including HIV/AIDS, tuberculosis and malaria) is only half of the number dying from chronic diseases-cardiovascular disease, cancer, diabetes, chronic respiratory disease in recent years. Developing countries are now undergoing a rapid epidemiological transition from infectious diseases to chronic diseases (WHO, 2003 and 2005). It is projected that, without any action, the total deaths due to chronic diseases will increase by 17% from 2005 to 2015 (WHO, 2005). A common misunderstanding is that chronic diseases are more prevalent in developed countries than in developing countries. In fact, 80% of global deaths due to chronic diseases occur in developing countries. The percentage of deaths due to chronic diseases in low-income countries is almost twice as high as that in high-income countries (WHO, 2005). Because of poor health care services and welfare systems, chronic diseases are imposing a much heavier health and economic burden on private families and national economies in developing countries.

This dissertation makes use of rich longitudinal data from China, a developing country that has experienced rapid economic growth during the past three decades. More than a quarter of the total population from developing countries live in China. Starting from the 1980's, the Chinese government has required nine years of compulsory education for students. Usually, a child attends primary school for six years, starting from the age of 6 or 7. After graduating from primary school, a student is expected to complete three years' lower secondary education. There is no restrictive entrance examination that students must pass to enter lower secondary schools. However, at the end of the three years' study, each student needs to take a standard entrance examination in order to continue his or her education in upper secondary school. The entrance examinations are usually conducted by the municipal ministry of education. The scores determine the eligibility of attending upper secondary schools of different quality levels: better schools usually require higher scores. During the three years' of upper secondary education, students study intensively to be prepared for the national college and university entrance examination. The admission to university or college is almost completely dependent on the score on this examination.

China has made significant strides in promoting basic education. China's primary gross enrollment rate was 115% in 2002, compared to an average level of 105% in the other East Asian developing countries. Despite this progress in primary education, the secondary gross enrolment rate in China was only 68% in 2002, much lower than the East Asian regional average of 91% (World Bank, 2004). A major reason for the lower enrolment in secondary education is that many students drop out of school at the end of primary education or during lower secondary school. As a consequence, only about 15% of the population of the age for tertiary education are enrolled in 2002 (MOE, 2003).

Chronic diseases now are responsible for 80% of all deaths and for 70% of disability-adjusted life-year lost in China (Wang et al., 2005; Strong et al., 2005). There are 300 million people who smoke and 160 million people who are hypertensive in China. Cardiovascular diseases and cancer are leading causes of death in China. Death rates from chronic diseases are even higher in China than in some of developed countries, such as Canada and the United Kingdom (Strong, 2005). The major driving forces of the epidemic of chronic diseases include population aging, rapid economic growth and urbanization. Because of population control and the desire for smaller families, the Chinese population is aging fast. The percentage of population aged 65 years or older is expected to increase from 7% in 2000 to 20% in 2040, leading to a 200% increase in the number of death due to chronic diseases (Wang et al., 2005). Rapid economic growth in China in the past three decades has also changed people's lifestyle towards an unhealthy direction. Fast urbanization and a deteriorating environment, accompanied the rapid economic growth, have been found to have negative impacts on chronic health in China (MOHST, 2004).

How much one should invest to improve knowledge, skills and health is a choice made by individuals or households. A rational person makes decisions by weighing the benefits and costs associated with such investments. However, without perfect information, optimal choices may not be made. For example, if the benefit of a healthy diet is underestimated, then the demand for it may be too low. On the other hand, if the cost of smoking is not fully considered, consumption of cigarettes will be too high. The two essays in this dissertation investigate two important pieces that are often ignored by consumers when making decisions regarding investments in education and health.

The first essay estimates the real cost of youth smoking by evaluating the impact of youth smoking on schooling. A substantial literature has estimated the cost of smoking in terms of medical costs of future diseases due to smoking or second-hand smoking, financial costs of mortality and morbidity, reduced labor productivity, and property loss in residential and non-residential cigarette-induced fires. Using unusually rich panel data from China, the first essay of this dissertation shows that youth smoking can also reduce learning productivity and discourage the motivation to go to school. Since education is an important factor that determines earnings, youth smoking can result in a lifetime income loss through its negative impact on schooling. The study finds that smoking one cigarette in adolescence can lower scores on Chinese and Mathematics tests by 0.1 standard deviations (of the distribution of test scores) and reduce the total length of schooling by about 5 days. Smoking one cigarette per day during adolescence is predicted to reduce lifetime income by 0.2%. To the knowledge of the author, this is the first study on the impact of youth smoking on schooling.

The second essay is concerned with consumers' investment in health. Chronic diseases, such as cardiovascular diseases, cancers, and diabetes, are becoming the major causes of global deaths in the past few decades. Unlike communicable diseases and accidental injuries, chronic diseases are usually caused by risk factors that can be modified by changing one's lifestyle. One of the major risk factors that are responsible for chronic diseases is and unhealthy diet. In order to examine the effect of information on moving consumers' diets in a healthier direction, the previous literature focuses on the impact on food choices of providing publicly available information on what constitutes a healthy diet. However, the findings from these studies vary significantly, suggesting that the effect of information on food choices needs to be studied more carefully. The second essay of this dissertation evaluates the effect of health information on consumers' food choices from a new perspective. Since consumers usually respond to publicly available information based on their perceptions about their own health, this study examines the responsiveness of food choices to a doctor's diagnosis of hypertension.

Hypertension is a leading risk factor for stroke, heart disease, heart failure, chronic renal failure, and arterial aneurysm. Without appropriate treatment, serious hypertension can reduce life expectancy significantly. However, because its asymptotic nature, many people are not aware that they have hypertension. For example, 30-50% of people with hypertension in five European countries do not know that they have hypertension. The rate is even higher in developing countries. This study finds that 75% of hypertensive population in China is unaware of this condition. A doctor's diagnosis of hypertension bridges consumers' imperfect perceptions and their true health status. The study analyzes the effect of the diagnosis of hypertension on consumers' food choices, using a regression-discontinuity approach. This method has gained popularity among economists in recent years because it mimics an experimental design. Using nationally representative data from China, this study shows that the diagnosis of hypertension has significant impacts on consumers' food choices: it reduces the consumption of animal oil and pork and increases that of wheat products. Consequently, after a diagnosis of hypertension, daily fat intake decreases significantly.

Chapter 2. Youth Smoking, Schooling and Life Cycle Income in Developing Countries: Evidence from Chinese Panel Data

2.1. Introduction

The detrimental effects of smoking on health have been both well documented and well publicized during the past several decades. Smoking is identified as a leading risk factor that is responsible for 5.4 million global deaths annually (WHO, 2008). Over 80% of these deaths occur in developing countries. There are about one billion smokers in the world, of whom more than 80% live in developing countries and about 35% live in China. While adult smoking rates have slowly decreased in developed countries since the early 1990's, the rate of cigarette smoking has steadily increased in the developing world, especially among teenagers (WHO, 1997; Chaloupka et al., 2000). According to the Global Youth Smoking Survey conducted in 43 developing countries in 2000 by the World Health Organization, about one fourth of currently enrolled high school students in developing countries have smoked cigarettes.

Youth smoking is likely to lead to lifelong smoking (Gruber, 2001; Sloan et al. 2003; USDHHS, 1994). In developing countries, the rate of smoking cessation is still very low. For example, ex-smokers account for 5-10% of the total population in developing countries, versus 30-40% in many developed countries (Gilbert et al., 2004). In these countries, youth smoking is almost equivalent to the start of a lifetime habit. About three quarters of smokers start smoking before they are 19 years old, and the earlier they start, the heavier they smoke in their later lives and the more difficult it is for them to quit.

In addition to the well documented long term adverse health effects, youth smoking can also cause serious damage to learning abilities and decrease intrinsic motivation to go to school. Although at the first glance it seems counterintuitive, given the well known "enhancing effect" of nicotine, there is a sizable body of research that finds a negative effect of smoking on working memory, attention focus, and cognitive abilities. The negative effect is more serious during the time periods when a smoker is not allowed to indulge in the desire for tobacco and if the onset of smoking is early. It is well documented that the pernicious effects of smoking are mainly due to a large potential for nicotine-dependence. The effects of nicotine on human performances are complex. But clinical studies generally conclude that nicotine cannot enhance cognition performance. Although it can reverse abstinence-induced declines in performance for nicotine-dependent individuals, it can only restore functioning to the levels before abstinence, which are not statistically higher than those of either non-abstinent smokers or nonsmokers. Moreover, the "enhancing effect" usually happens within a short period immediate after smoking and mainly affects only motor response. This means that the temporary benefit of the "enhancing effect" of smoking comes at a large cost — teenage smokers may experience learning capacity loss most of the day at school or at home because they are prevented from smoking. (Heishman et al., 1994; Foulds et al., 1996; Pineda et al., 1998; Ernst et al., 2001; Jacobsen et al., 2005)

Smokers usually extract 1-4 mg out of the 7 mg nicotine in a single cigarette. Nicotine delivered from cigarette smoking can rapidly increase the nicotine plasma concentration in smokers' bodies rapidly. The initial distribution half-life of nicotine lasts about 10-20 minutes after inhaling, followed by elimination within 2-3 hours. For a daily smoker, regardless of the time of day cigarette is smoked, the nicotine plasma concentration usually peaks at 10-40 ng/ml in afternoon and declines to 5-10 ng/ml after overnight abstinence. Therefore, a daily smoker is exposed to nicotine constantly. Teenagers are particularly vulnerable to nicotine. Clinical studies have shown that adolescent smokers who have smoked an average of 3-4 years perform poorer in working memory performance accuracy, attention focus and verbal memory than their non-smoking counterparts. Abstinence of only 24 hours can have a much greater adverse impact on teenagers than on adult. Moreover, adolescent smoking has last-lasting chronic adverse effects on cognitive performances caused by the neurotoxic effects of nicotine (Ernst et al., 2001; Jacobsen et al., 2005; Counotte et al., 2009).

Smoking is also linked to child health and nutrition (Grunberg et al., 1984; Bowen et al., 1986; Jorenby et al., 1996). According to the American Cancer Society, cigarette smoking causes serious health problems among children and teens, including coughing, shortness of breath, production of phlegm (mucus), respiratory illnesses, reduced physical fitness, poor lung growth and function, worse overall health and addiction to nicotine. And because smoking could interfere with the absorption of such vital nutrients as folate and vitamin C, it can increase the risk of nutrition deficiency and anemia, which are known to be associated with reduced learning (Alderman et al., 2001; Glewwe, Jacoby and King, 2001; Miguel, Bobonis and Sharma, 2006).

Motivated by these clinical findings, this study uses a rich longitudinal dataset from northwestern China, the Gansu Survey of Children and Families (GSCF), to estimate the impact of youth smoking on teenagers' scores on both Chinese and Mathematics academic tests. The study shows that smoking one additional cigarette per day can reduce test scores by approximately 0.1 standard deviations. Moreover, analysis of the determinants of youth smoking participation indicates that, besides the traditional price effect, parental smoking has a strong impact on both smoking choices and teenage behaviors.

This study also estimates the effect of youth smoking on educational attainment using the rich longitudinal data collected by the China Health and Nutrition Survey (CHNS). The study finds that youth smoking has a negative impact on educational attainment: smoking one cigarette per day at age 12-17 can shorten years of schooling by 0.015. Since both smoking and schooling are personal choices and can be affected by same unobservable factors, an instrumental variable approach is used to minimize omitted variable bias.[1] Cigarette prices, price indices, and the number of registered vendors of alcohol are used as instrumental variables. The estimates indicate that smoking one additional cigarette per day reduces the duration of schooling by about 5.4 days.

[1] In 1982, Farrell and Fuchs (1982) suggested that careful thought be given to the relationship between schooling and smoking, raising the "third variable" hypothesis.

The study predicts that smoking one cigarette per day during adolescence can reduce lifetime income by 0.2%. This finding updates previous estimates of the cost of smoking which mainly consider only medical costs of smoking-caused diseases, financial costs of smoking-caused morbidity and mortality, and property loss in smoking-caused fires. Understanding the real costs of youth smoking is critical for health policy formulation. It could be used to set cigarette taxes, and, more generally, to set tobacco prices at socially optimal levels. It could also be used to assess the benefits of various anti-smoking campaigns that are intended to reduce the epidemic of smoking among teenagers. If young people are provided with an accurate information on the real costs of smoking, so that they can fully consider the negative consequences of smoking, perhaps much less tobacco would be consumed and fewer teenagers would take up this lifelong harmful habit.

The rest of the chapter is organized as follows. Section 2.2 provides a brief review of the existing literature on the economic costs of smoking. Section 2.3 presents the model that describes the decision process by which smoking and schooling choices are made. In Section 2.4, the econometric specifications and estimation strategies are discussed. Section 2.5 discusses the data and provides background information on smoking in China. Section 2.6 presents and discusses the results. Section 2.7 uses estimates obtained in Section 2.6 to simulate the lifetime income loss due to youth smoking, and Section 2.8 concludes.

2.2. Literature review

There is a huge body of literature that has attempted to estimate the economic cost of smoking. These studies vary in their methodological approaches and in the types of costs considered. Yet the number of studies conducted in developing countries is very limited, and only a few focus on China. Cost estimates that consider the impact of smoking on teenagers in developing countries are even rarer: such studies usually focus on passive smoking.

The methodological approaches used in most previous studies can be classified into one of two types: cross-sectional or lifecycle. The cross-sectional approach measures the extra health expenditures and the additional use of health facilities, or the extra economic burden, by comparing smokers and nonsmokers at a specific time point. This approach is by far the most common, especially in studies of developing countries. More recently, some studies of developed countries have used the lifecycle approach, which estimates the present discounted value of all extra medical and nonmedical costs a smoker incurs over his or her lifetime. This approach is preferable because (a) it links the cause (smoking) and the consequences (economic and health burdens) more directly; (b) it is able to account for a wider variety of choices, including starting, quitting or resuming smoking (Miller et al., 1997; Sloan et al., 2004). This study follows the lifecycle approach in order to capture these advantages.

Another major dimension along which studies vary is in the scope of costs they consider. Most studies include both medical costs and nonmedical costs. Some of these costs are private costs and others are external public costs. Medical costs attributable to smoking are by far the most well studied, yet the attention given to nonmedical costs has grown in recent years. Nonmedical costs that have been addressed include the financial costs of mortality and morbidity, reduced labor productivity, property loss in residential and nonresidential cigarette-caused fires (Mudarri, 1994; Leistikow et al., 2000a, 2000b; Sloan et al., 2004), long-term special education care for low-birth-weight babies of smoking mothers (Marks et al., 1990), and expenditures on tobacco prevention and control. Part of these costs are borne by smokers (private costs) while the remaining costs are borne by the others (external costs). There is increasing concern over the spillover costs of smoking. Examples include the economic and health burdens borne by those exposed to secondhand smoke.

The few studies that have attempted to estimate the costs of smoking in China have used cross-sectional data to estimate the medical costs and/or the disease burden attributable to smoking and to passive smoking. Sung et al. (2006) calculated a cost estimate of $5 billion in 2000 in China, which is equivalent to $25.43 per smoker (age over 34). Their cost estimates consisted of direct medical costs (34%), indirect morbidity costs (8%) and indirect mortality costs (58%). Several studies have focused on the disease burden. For example, Gan et al. (2007) found that, among adults, passive smoking alone led to more than 22,000 lung cancer deaths and 33,800 ischaemic heart disease deaths in China in 2002.

To the knowledge of the author, no study to date has addressed the impact of youth smoking on schooling, even though such impact may lead to a different trajectory of lifetime income. In developing countries, child health and nutrition have been identified as important determinants of educational achievement and attainment (Alderman et al., 2001; Glewwe, Jacoby, and King, 2001; Miguel and Kremer, 2004; Miguel, Bobnis, Sharma, 2006, Glewwe and Miguel, 2008). Youth smoking could worsen child health and nutritional status and, thus, may also be an important factor determining children's educational outcomes.

In general, most previous studies of the costs of smoking fail to account for the endogeneity of smoking. Smoking is an individual choice that reflects many unobservable personal, family and community characteristics. Although some studies try to control for some important confounding factors (e.g. Sloan et al. (2004) considered education, obesity, alcohol consumption and time preference), unobservable factors could still lead to severe bias in estimates of the costs of smoking. For example, parental attitudes, mental health, and the capacity for self-control could influence youths' choices with respect to both smoking and schooling. These factors are very hard to measure and, inevitably, will become part of the disturbance term of the equation that describes the impact of smoking on socioeconomic outcomes. Failure to consider the endogeneity issue casts serious doubt on the cost estimates of previous studies. In order to obtain unbiased estimates of the impact of youth smoking on educational outcomes, this study uses unique longitudinal data and an instrumental variable approach to better account for endogeneity problems.

2.3. Theoretical model

Consumers' intertemporal smoking and educational decisions are modeled in the spirit of the rational addiction model of Becker and Murphy (1988). A consumer's preferences in each period are defined over a numeraire consumption good (x) and smoking (s). Following Becker and Murphy, it is assumed that the addictive good s contributes to an addictive stock (A) that also enters the consumer's utility. The one-period utility is thus given by $u(x_t, s_t, A_t)$.

Past consumption of cigarettes can influence current and future consumption decisions, (a) through its indirect effect on the marginal utility of consuming s and (b) through its direct effect on current and future utility due to adverse health consequences or discomfort associated with addiction. More specifically, assume $u_{sA} > 0$, i.e. the marginal utility of smoking is higher if A is high, and $u_A < 0$, i.e. the marginal utility of addiction is negative. The addictive stock in period $t + 1$ depends on the amount of smoking and the addictive stock in period t:

$$A_{t+1} = f(s_t, A_t) \tag{1}$$

The more one smokes during this period or the more one has smoked in the past, the more addicted one is to tobacco in the next period. Moreover, the addiction stock "depreciates" over time — the longer one abstains, the less addicted one is. In addition, the model incorporates the consumer's educational input decisions into the standard addiction model. The educational outcome (in terms of knowledge and skills attained) at the beginning of period $t + 1$ depends on educational inputs e_t in period t, and educational achievement at the beginning of period t:

$$E_{t+1} = \psi h(e_t, E_t) \tag{2}$$

where $\psi > 0$ is a parameter that describes productivity of educational inputs and h a production function of education. Note that this model assumes that the consumer is unaware of the negative impact of smoking on ψ. Although clinical studies have shown that

$d\psi/dA < 0$, consumers, especially teenagers in developing countries, are likely to have incomplete information because the detrimental effects of smoking on learning abilities are not well publicized.

The educational inputs e_t include time and labor devoted to studying, as well as material inputs. It is assumed that the consumer is endowed with constant total labor and time, which are allocated between going to school and working in each period. That is, if e_t increases, the labor and time allocated on working will decrease and, therefore, income in each time period falls. Thus, income is a function of educational inputs and educational achievement in each period: $I(e_t, E_t)$ where $dI/de_t < 0$ and $dI/dE_t > 0$. As in Becker and Murphy (1988) and Becker, Grossman, and Murphy (1994), the consumer lives infinitely and any effect of s or A on the consumers' length of life or other sorts of uncertainty is ignored.

Given this setup, the consumer chooses an optimal consumption path $\{x_t, s_t, e_t\}_{t=0}^{\infty}$, maximizing the discounted sum of utilities:

$$\sum_{t=0}^{\infty} \delta^t u(x_t, s_t, A_t) \tag{3a}$$

subject to (1) and (2), and the intertemporal budget constraint:

$$\sum_{t=0}^{\infty} \left(\frac{1}{1+r}\right)^t [x_t + p_t s_t + w_t e_t] \leq \sum_{t=0}^{\infty} \left(\frac{1}{1+r}\right)^t [I(e_t, E_t) + W_t] \tag{3b}$$

where the consumer's time preference $\delta = 1/(1+r)$, p_t is the price of cigarettes, w_t is the price of education, and r is the interest rate. Note that the price of the numeraire good x is assumed to be constant over time. In earlier periods (e.g. teenage), the consumer may obtain positive non-labor income $W_t > 0$ which is assumed to be exogenous. The

Lagrangian for this problem is:

$$\begin{aligned}\mathcal{L} \;=\; & \sum_{t=0}^{\infty} \delta^t u(x_t, s_t, A_t) + \lambda \sum_{t=0}^{\infty} \left(\frac{1}{1+r}\right)^t [I(e_t, E_t) + W_t - x_t - p_t s_t - w_t e_t] \\ & + \sum_{t=0}^{\infty} \mu_t \,[\psi h(e_t, E_t) - E_{t+1}] + \sum_{t=0}^{\infty} \xi_t \,[f(s_t, A_t) - A_{t+1}] \qquad (4)\end{aligned}$$

Assuming interior solution, the first-order conditions:

$$u_x(t) = \lambda \qquad (5a)$$

$$u_s(t) = \lambda p_t - (\xi_t/\delta^t) f_s(t) \qquad (5b)$$

$$(\mu_t/\lambda\delta^t)\psi h_e(t) = w_t - I_e(t) \qquad (5c)$$

and the transversality conditions:

$$\lim_{t\to\infty} \delta^t A_t = 0 \quad \text{and} \quad \lim_{t\to\infty} \delta^t E_t = 0 \qquad (6)$$

characterize the optimal consumption path.

Equation (5a) is the standard condition that the marginal utility of other consumption in each period equals the marginal utility (or shadow value) of wealth. Equation (5b) implies that the optimal cigarette consumption equates the marginal utility of cigarette consumption with the current price of cigarettes (multiplied by the shadow value of wealth) plus the discounted marginal effect on future utility from increased addiction. Similarly, equation (5c) implies that the optimal educational input in period *t* equates the discounted marginal gain in future income streams from education with the price of education plus the opportunity costs of education.

With some regularity conditions, the program (3) can be reformulated as recursive dynamic programming (Stokey, Lucas, and Prescott, 1989). These conditions include (a) u is concave in x and s for every feasible A, (b) f and h are bounded, real-valued functions of s and e, respectively, for every feasible A and E, (c) $\lim_{t\to\infty} \sum_{t=0}^{\infty} \delta^t u(x_t, s_t, A_t)$ exists for every feasible sequence of $\{x_t, s_t, e_t\}_{t=0}^{\infty}$, and (d) the budget balances each period: i.e. $x_t + p_t s_t + w_t e_t + B_t \leq I(e_t, E_t) + W_t - \frac{1}{1+r} B_{t-1}$ where B_t is intertemporal borrowing. Condition (c) holds if u, f, h, and I are bounded and non-empty valued. The last budget balance condition is consistent with the idea that some families pay W_t to cover educational costs, living expenses, and basic leisure expenditures until children get mature and attain enough skills to earn adequate incomes while poor families do not pay for these costs and therefore children may start working in early ages before they acquire a high level of education.

The solutions to (3) are thus functions of state variables, A_t and E_t, prices, endowed incomes, time preferences (or the interest rate), parameters on utility, addiction, and educational achievement functions:

$$s_t^* = \phi_s(A_t, E_t, p_t, w_t, W_t; \psi, \delta, u, f, h, I) \tag{7a}$$

$$e_t^* = \phi_e(A_t, E_t, p_t, w_t, W_t; \psi, \delta, u, f, h, I) \tag{7b}$$

Clinical studies have found the effect of smoking on human capital productivity is negative: $d\psi/dA < 0$. The objective of this study lies in identifying this effect in the empirical data. If $d\psi/dA < 0$, a negative effect of decreased learning ability ψ on educational outcomes, e_t^* or E_{t+1}, is expected, *conditional on* educational achievement E_t up to period t. More formally, we have:

Effects of Smoking on Education: *Suppose that the value function v of the recursive dynamic programming version of the model (1)-(3) exists, is twice-differentiable, and is concave in endogenous arguments. Then conditional on educational achievement up to period t, E_t, both the demand for education inputs e_t^* and educational achievement E_{t+1} decrease with a decrease in ψ. Because smoking decreases ψ, it has negative effects on both education inputs and educational achievement.*

Proof: Consider the dynamic programming version of the model:

$$v(A,E,B) = \max_{A',E',B'} \left\{ u(x,s,A) + \delta v(A',E',B') \right\},$$

subject to (1), (2), and the intertemporal budget constraint. Primes indicate the next time period. A consumer chooses the optimal levels of A', E', and B' to maximize the total utility. Substituting the constraints, it follows that:

$$\begin{aligned} v(A,E,B) \quad = \quad & \max_{x,s,e} u(x,s,A) + \delta v(f(s,A), \psi h(e,E), I(e,E) + W \\ & + (1+r)B - x - ps - we). \end{aligned} \tag{8}$$

The first-order conditions are:

$$\begin{aligned} \varphi_x \quad &\equiv \quad u_x - \delta v_B = 0, \\ \varphi_s \quad &\equiv \quad u_s - p u_x + \delta v_A f_s = 0, \\ \varphi_e \quad &\equiv \quad \frac{\delta v_E}{u_x} \psi h_e - w + I_e = 0. \end{aligned}$$

These are the recursive analogues of the first-order conditions in (5). To see the impact of ψ on e^*, implicitly differentiate this system with respect to ψ and e^*. By the implicit function theorem, we have:

$$\frac{de^*}{d\psi} = -\frac{1}{\triangle} \begin{vmatrix} \varphi_{x,x} & \varphi_{x,s} & \varphi_{x,\psi} \\ \varphi_{s,x} & \varphi_{s,s} & \varphi_{s,\psi} \\ \varphi_{e,x} & \varphi_{e,s} & \varphi_{e,\psi} \end{vmatrix}$$

where $\triangle$ is the determinant of the Hessian of the objective function (8) and is ≤ 0 since the objective function is concave in endogenous arguments.

$$\begin{vmatrix} \varphi_{x,x} & \varphi_{x,s} & \varphi_{x,\psi} \\ \varphi_{s,x} & \varphi_{s,s} & \varphi_{s,\psi} \\ \varphi_{e,x} & \varphi_{e,s} & \varphi_{e,\psi} \end{vmatrix} = -\frac{\delta v_E}{u_x} h_e \left[(u_{xs})^2 - u_{xx} u_{ss} - \delta v_A u_{xx} f_{ss} \right] \geq 0.$$

By concavity of the utility function, $u_{xx}u_{ss} - (u_{xs})^2 \geq 0$. For the production function of addictive stock, $f_{ss} \geq 0$ as a person gets more addicted to smoking when the consumption of cigarettes is higher. Because $(\delta v_E/u_x)h_e$ is the marginal benefit of educational input which is positive, the term in the brackets is non-positive. Thus we have $de^*/d\psi \geq 0$. Educational achievement E_{t+1} also increases with ψ:

$$\left.\frac{dE_{t+1}}{d\psi}\right|_{E_t} = h(e_t^*, E_t) + h_e \frac{de_t^*}{d\psi} \geq 0$$

Q.E.D.

Given that $d\psi/dA < 0$, it follows that $dE_{t+1}/dA_t < 0$ and $de_t^*/dA_t < 0$ conditional on educational achievement E_t. Note that reduced educational achievement, $dE_{t+1}/dA_t < 0$, might stem either directly from reduced learning ability or indirectly from reduced demand for education or both. This model implicitly assumes that the individual makes decisions without information on $d\psi/dA < 0$. That is, the individual observes ψ, but is not aware of the effect of smoking on ψ. Once fully informed of this negative effect, the individual's demand for cigarettes would decrease because it would add to the (marginal) costs of smoking in Eq. (5b). In the empirical specification, E_{t+1} is approximated by test scores in year t and $\sum_{\tau=0}^{t} e_\tau$ by years of schooling by year t. The obvious endogeneity arises because common factors affect both smoking s_t^* (and A_t^*) and educational input e_t^* (and outcome E_t^*). The next section will discuss the empirical identification strategies to address this problem.

2.4. Empirical strategies

2.4.1. Educational achievement (test scores)

The first empirical strategy is to use variation in test scores on Chinese and math exams conducted in 2004 in Gansu, China. Test scores are a good proxy variable for students' educational achievement in a given year. The effect of smoking status on test scores is estimated by linearly approximating equation (2):

$$T_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + \varepsilon_i \quad (9)$$

where T_i denotes the test scores, $\mathbf{x}$ a vector of covariates, and $\boldsymbol{\beta}$ a vector of parameters. In order to estimate the parameters without bias, the disturbance term ε need to be uncorrelated with $\mathbf{x}$, that is, $E[\varepsilon\mathbf{x}] = 0$. To identify the effect of smoking, smoking status is included as one of the regressors. Because educational inputs are not directly observable (e.g. effort levels and resources used) and because both smoking and educational input decisions are personal choices, some unobserved characteristics that influence educational inputs may also affect smoking status. Thus, OLS estimates are likely to be biased.

Moreover, because smoking status is self-reported, OLS estimates may suffer from measurement error bias. In China, smoking under the age of 18 is forbidden, so teenagers may have an incentive to under-report their smoking habits and this incentive is likely to be greater when parents are present during the survey administration. In the GSCF survey, groups of teenagers completed the questionnaire anonymously in a closed room without school officials or family members present, while the CHNS study uses a regular household survey instrument, in which anonymity is not guaranteed. These two datasets show considerably different rates of smoking among teenagers: about 13% of those aged 14-17 reported smoking at least once in their lives in the GSCF sample, whereas, only 4% reported doing so in the CHNS sample. One of the reasons for this difference could be that smoking behavior information was collected in a more confidential manner in the GSCF. Variables that are measured with random error usually lead to downward bias in the estimates of their associated parameters.

This study uses an instrumental variable (IV) approach to control for both endogeneity and measurement error bias. The instrumental variables for smoking decisions include the count of registered alcohol venders, and price indices of several major categories of consumer goods. Alcohol consumption is very likely to be related to that of cigarette. They may be complements or substitutes. Thus, the number of alcohol vendors is expected to be correlated with cigarette consumption, negatively if they are substitutes or positively if they are complements. Unfortunately, reliable data are not available on registered cigarette vendors in the study area. Price indices may also affect the demand for cigarettes. And they are not likely to be correlated with ε_i in (9) once the demand for educational inputs is controlled for.

A number of covariates are included to account for learning abilities, educational inputs, and educational achievement in the regression equation. For example, parental education is included to reflect the innate ability of children and parental preferences for children's education. Parents with higher education are most likely to help their children with schoolwork. Household income is an indicator of resources allocated to child education (e.g. richer families can spend more on children's schooling).The total years of schooling by the previous time period is also controlled for. However, since years of schooling may be correlated with some unobserved variables, age variable is used instead to approximate years of schooling.[2]

School and teacher characteristics also measure the resources invested in education. Better school infrastructure and skilled teachers may increase student learning productivity and thus increase the return to investment in children's education. Note that children in China are required to attend the school closest to their homes and the school location is determined by local governments instead of parents. This implies that the quality of schools is less likely to reflect parents' attitude and students' innate ability. Information on availability of schools is also used to capture any impacts of community characteristics on schooling, such as the value the community places on education and the opportunity costs of attending school. Another possible way to control for school characteristics is to use a school fixed effect model. However, because the instrumental variables are available at township or county level, a school fixed effect model will wipe out the IVs.

[2] More than 99% of the sample used for the test score regressions were currently enrolled in school. Therefore, their ages can be used to approximate their years of schooling.

2.4.2 Years of schooling

If youth smoking decreases the return to education, it should also reduce the demand for education. The second contribution of this study is to investigate the determinants of the demand for education, as measured by years of schooling. Following Maddala (1983), Glewwe and Jacoby (1994) and Zhao and Glewwe (2010), assume that the latent demand for years of schooling is given by:

$$Y_i^* = \mathbf{z}_i^{'}\boldsymbol{\gamma} + \eta_i \tag{10}$$

where $\mathbf{z}_i$ denotes a vector of covariates, $\boldsymbol{\gamma}$ is a vector of corresponding parameters, and η_i i.i.d. errors. This latent (continuous) demand variable is not observable. Instead, years of schooling, Y_i, a categorical variable, is observed. But the two variables are related to each other: Y_i equals m if and only if $\alpha_m < Y_i^* \leq \alpha_{m+1}$ for some cutoff parameters α's.

If one assumes that the disturbance terms in (10) follow a normal distribution, then the $\boldsymbol{\gamma}$ parameters can be estimated using a censored ordered probit model. A cencored ordered specification is used because the study focuses on teenagers, and therefore, the observed years of schooling variable is "right-censored" for children who are currently enrolled in school, which implies that one has only a lower bound for their final years of schooling. Failure to account for this censoring would yield biased parameter estimates. In the CHNS sample, about 30% of teenagers aged 14-17 are not right censored. The associated likelihood function is derived in Appendix A.

Empirically, years of schooling is regressed on the smoking status, controlling for covariates that affect the benefits and costs of education. However, to the extent that there are unobserved factors in the error term that affect both schooling and smoking status, the estimates will be biased. A two-step estimator is used to address this issue — the first step regresses smoking status on all the exogenous variables in the second stage and on the instrumental variables, cigarette prices when the onset of smoking occurs, and predicts smoking status, and then, in the second step, a censored ordered probit is estimated using the predicted smoking status. Cigarette prices are likely to have little impact on educational decisions, but are likely to be significant determinants of smoking decisions (Gruber, 2001). Community-level prices of cigarettes in both previous and current periods are available in the CHNS data. Using previous prices can minimize possible correlation, if there is any, between the instrumental variables and the unobserved errors in (10). In addition to individual and household characteristics used in (9), average wages for different types of low-skill jobs in the community are also included in the estimation of (10) in order to capture the opportunity cost of education and also community economic characteristics.

2.5. Data and background

2.5.1 China Health and Nutrition Survey (CHNS)

The CHNS is a comprehensive household survey that collected detailed information at individual, household and community levels in China from 1989 to 2004. Actually, the CHNS is still being conducted until now. Thus, new data can be used to update the findings of this study in the future. The two main advantages of the CHNS are that (a) its panel data were collected over a long period of time; and (b) it is approximately nationally representative. Moreover, detailed price data, including cigarette prices, are available for each community, which allows one to use cigarette prices as instrumental variables to predict youth smoking.

In 1989, all the major provinces in mainland China were contacted to see if they were willing to participate in the CHNS. Among those who showed interest in the survey, eight were selected because of the substantial variation in geographic location, economic development and local culture. The eight selected provinces are Guangxi, Guizhou, Henan, Hubei, Hunan, Jiangsu, Liaoning, and Shandong. A total of 3,793 households and 15,917 individuals were interviewed in 1989. Due to practical reasons, such as migrating away from the sampled neighborhoods, a small percentage of these households and individuals were not interviewed in 1991. More specifically, 3,616 households (14,778 individuals) were re-interviewed in 1991. Two years later in 1993, the sample size fell to 3,441 households (13,893 individuals) in the third round of CHNS. The 1997 CHNS had a slightly larger sample, 14,426 individuals who belonged to 3,875 households, by including newly formed families residing in the sample neighborhoods and replacing households and neighborhoods that left the survey with new households and new neighborhoods. In addition, a new province, Heilongjiang, replaced one of the original sample provinces, Liaoning, during this round. In 2000, Liaoning province returned to the survey and consequently 9 provinces were surveyed in that year. Appendix B and Figure 2.1 provide a detailed description on the sampling method and geographical locations of the sample provinces.

2.5.2 Gansu Survey of Children and Families (GSCF)

The other main data source used in this study is the Gansu Survey of Children and Families, which was conducted in 2000 and 2004 for a random sample of two thousand children in rural areas of Gansu province who were aged 9-13 years old in the year 2000. As shown in Figure 2.2, Gansu is located in northwestern China. The sample was drawn from 20 counties that were randomly selected from all the major regions with different levels of GDP per capita in the province. Within each sample county, 100 children in rural areas were randomly selected from 2-4 townships, yielding a sample of 1,078 boys and 922 girls. Comprehensive data were collected through household and school surveys that conducted interviews of the sample children, their parents, teachers and school principals.

In 2004, the same children were interviewed again, as well as their oldest younger siblings (if aged 7 or older). Of the original 2000 children, the families of 40 boys and 42 girls were not re-interviewed in 2004 because: 71 households moved out of the township, 8 sample children died, 2 children's parents were divorced, and 1 household refused to be interviewed. No information is available for these children, but the low attrition rate of 4.1% suggests that there is little reason to worry about sample selection bias.

2.5.3 Youth smoking and education

Table 2.1 provides a comparison of the prevalence of smoking in China across five different age groups, calculated from the CHNS data. The prevalence of ever smoking presented in the first column and the second column for each age group are different in the sense that the latter are cross checked by the responses of the same individuals from later rounds of the survey. Due to mistaken memory and possible intentional under-reporting by teenagers, the prevalence of ever smoking generally gets higher after cross checking, more obviously for earlier rounds

In general, 0.4-2.7% of youth aged 12-15 and 4.4-12.3% of those aged 15-17 reported that they have ever smoked cigarettes. The prevalence of current smoking among teenagers is much smaller, 0.4-0.8% and 3.7-7.0% for the respective groups. Note that the current smoking rates may be underestimated because they cannot be cross-checked. The smoking rates for youth aged 12-15 are much lower than those reported in the Global Youth Tobacco Smoking Survey, 22% for ever smoking and 11% for current smoking[3]. Although the prevalence of currently smoking for the CHNS sample is relatively low, on average, the daily amount of cigarettes smoked by those who reported currently smoking is as high as 9.5. As for adults, approximately 40-45% have ever smoked and 30-35% are currently smoking in China.

To see when the onset of smoking happened, Table 2.2 reports, for all the ever-smokers, the percentages of those who started to smoke before 18 and those who started smoking between 18 and 24. Approximately 30% of those who have ever smoked lit up their first cigarettes before they turned 18, that is before they became fully legal adults in China. About 55% started to smoke between 18 to 24. In total, 85% of ever-smokers began to smoke before turning 25 years old.

The prevalence of smoking for teenagers reported above does not differentiate between those who were currently enrolled in schools and those who were not. Table 2.3 gives an

[3] Global Youth Tobacco Smoking Surveys were conducted at 73 sites in 43 countries in 1999-2001 for population aged 13-15 (The Global Youth Tobacco Survey Collaborative Group).

overview of the prevalence of smoking for those enrolled in different levels of schools: about 0.3-1.1% for primary school students; 1.5-7.9% for junior high school students; and 3.6-11.1% for senior high school students. Because of the availability of cross checking, the prevalence tends to be higher in earlier rounds. According to the CHNS data, approximately 27% of teenagers aged 12-17 were out of school in 1989. Are teenagers more likely to smoke if they are not enrolled in school? Table 2.4 shows that, among ever-smokers, about half started to smoke after leaving school. The percentage of smoking before leaving school started to increase until 65.6% in 1997 and then dropped to 30% in 2004. Of course, starting to smoke after leaving school is unlikely to have an impact on educational outcomes, which have already been completed.

Table 2.5 summarizes the years of formal education achieved by the start of the surveys for teenagers aged 12 to 17. Note that this is the total completed years of education for those who were already out of school. Although the final levels of education are not observed for those who were still enrolled in school, they are at least as high as the levels achieved by the start of the surveys - the total years of schooling will be the same if one drops out of school right after the survey. According to the completed years of schooling for individuals out of school, educational attainment has clearly increased over time: the percentage of individuals with no education or only 1-6 years of primary education decreases from 13.7% in 1989 to 1.8% in 2004, while the percentage of individuals who dropped out school with some years of secondary education, including training from technical schools, has decreased from 11.2% to 10%. In general, the proportion of teenagers aged 12-17 who were currently enrolled in all levels of school increased steadily from 73% in 1989 to 88% in 2004, mainly due to the expansion of secondary education over this time period. Note that the decrease in the percentage of enrolled primary students, from 32.5% in 1989 to 14.4% in 2004, reflects the fact that children are more likely to be sent to school at an earlier age and less likely to repeat in recent years.

Based on the data from Gansu Survey of Children and Families (GSCF), the prevalence of smoking for teenagers aged 14-17 is around 12%. This rate is much higher than what is found in the CHNS data, and probably reflects that the GSCF data were collected in a more confidential manner. Table 2.6 shows that, on average, these teenagers started to smoke at the age of 11, and 23% reported that they are smoking currently, defined as having smoked in the previous month. The average amount of cigarettes smoked per day, conditional on smoking in previous month, was 3.5 cigarettes. Note that 40% of smoking teenagers reported that they smoked in their friends' houses, 30% smoked in school, 30% smoked at home, with about 20% smoking in public or at social occasions. Although smoking is forbidden in school, many students actually secretly smoke in restrooms when their cigarette cravings are too strong to resist. Such behavior is usually at the risk of being caught and penalized by school authorities.

Of the 2000 sample children in the GSCF data, only 9 never enrolled in school; and of the 1991 who enrolled before 2000, only 19 left school before 2000. In contrast, 225 left school between 2000 and 2004. Thus, 88% of the sampled children, who were aged 14-17 in 2004, were still enrolled in school in that year. As shown in Table 2.7, most students who dropped out did so during or immediately after the third grade (the last grade) of lower secondary school (27.3%), followed by the fifth grade (the second to last grade) of primary school (21.3%). Large gender differences occur at grades four and five of the primary level, with the frequency of female dropping out double or triple that of their male counterparts.

According to the GSCF data, the prevalence of smoking is smaller among those who dropped out of school, compared to their counterparts who are remained in school. About 9.2% of dropouts reported that they had ever smoked, while the rate was 12.4% for currently enrolled students. However, students usually smoke much less, 60% of whom smoked less than 5 cigarettes their entire lives, and only an average of 0.65 cigarettes per day. In contrast, only 23% of dropouts reported having smoked less than 5 cigarettes in their entire life and, on average, smoked more than 4 cigarettes per day.

There is no law in China that specifies a legal smoking age, although several laws mandate that parents, schools and adults should prevent children from smoking before reaching 18 years old . There are also laws that require primary and secondary school students not to smoke, and any smoking on the campus of primary and secondary schools is illegal.[4] Students cannot smoke off campus, either. However, it is not illegal for teenagers who are not students to smoke off campus. Therefore, although primary and secondary schools strictly forbid their students to smoke by imposing school rules and penalties, teenagers have much more freedom to smoke after dropping out of school.

2.5.4 Cigarette prices

Table 2.8 presents the means of cigarette prices in different provinces by area of residence (rural or urban) for different years, based on the village-level price data collected by the CHNS. Cigarette prices can vary significantly in different provinces and between rural and urban areas. In generally, local brand cigarettes, which are the most common, are more expensive in urban areas than in rural areas. However, prices of luxury imported cigarettes, such as Marlboro cigarettes, do not vary very much and tend to be a little bit higher in rural areas: a pack of Marlboro cigarettes costs about 8-12 yuan ($1-$1.5). Although China has made significant progress in developing a market economy, the tobacco industry is still a state monopoly. The China National Tobacco Company (CNTC), under the jurisdiction of the State Tobacco Monopoly Administration, controls cultivation of tobacco, manufacturing tobacco products in contracted tobacco factories, and supplying and distributing tobacco products to retailers in China. Unlike states in the United States, each of which imposes a different tax rate on the purchase of cigarettes, the Chinese government levies several types of nation-wide taxes on tobacco products: a value added tax of 17%, and two types of consumption taxes paid by tobacco factories and the CNTC (a lump-sum tax per pack of cigarettes and an ad valorem tax that is 35-45% of the contract prices agreed to by tobacco factories and the CNTC).

[4] Refer to article 5 in the Law of the People's Republic of China on Tobacco Monopoly, article 10 in the Law of the Peoples Republic of China on the Protection of Minors, and article 15 in the Law of the People's Republic of China on the Prevention of Juvenile Delinquency.

Variation in cigarette prices in China is usually due to several factors. The first is brand. There are hundreds of cigarette brands, including more than one hundred major ones in China. Different brands vary in their quality, flavors, processing technologies and so forth. There are also many variants of each brand and these variants are usually priced at different levels. Because the taxes on the tobacco industry are important sources of revenues for local governments, there usually exists serious protection of local markets against cigarettes made outside of the province. Therefore, it is common to find different brands sold in different places. The second reason why prices vary is that the contract price of a domestic tobacco product is usually set differently inside and outside the manufacturing province. To promote sales, the price of the same product is usually lower outside the manufacturing province. However, the prices of cigarettes made by foreign companies outside China do not vary according to these factors. The prices of foreign brand cigarettes are usually set by their makers, subject to an import tax that is approximately 40% of the original prices. The price may also vary due to transportation costs and between urban and rural areas. Lastly, the retail prices are allowed to vary within a range of 10-15% of the guiding prices set by the CNTC. Retailers who do not follow the guiding prices will be subject to fines or license suspension[5]. Therefore, retail prices may reflect both demand and supply in the local market.

2.5.5 Other descriptive statistics

Table 2.9 presents summary statistics of some of the most important explanatory variables used in this study. The average levels of parental education are similar in both of the datasets, approximately 7 years for fathers and 5 years for mothers. As for parental smoking status, the GSCF data give a slightly higher proportion of fathers who smoke, 77%, and a much lower percentage for mothers, 0.004%, compared with those provided by the CHNS data, 67% and 0.02%, respectively. However, the difference in mothers' smoking rate is not statistically significant (t-statistics=1.3). The rates of mothers' smoking are

[5] A retailer needs to apply for a license with the State Tobacco Monopoly Administration to be eligible to sell any tobacco products in China.

generally very small compared to those of fathers' smoking.

In contrast to the CHNS, the GSCF collected not only household and community information, but also detailed information about sample children's school performance, such as scores on Chinese and Mathematics academic tests. Table 2.8 shows that the most recent average final test scores are 72.9 and 70.2 out of 100 for Chinese and math, respectively. Test scores are averaged for the most recent fall and spring semesters.

2.6 Results

2.6.1 Determinants of youth smoking

Because smoking is an endogenous choice, an instrumental variable approach is adopted to estimate the impact of youth smoking on schooling. The set of instrumental variables used is slightly different for regressions based on the two datasets, due to differences in the availability of instrumental variables (see Table 2.11). Prices of both local brand cigarettes and Marlboro cigarettes from the previous period (the interval is usually 2-4 years) are available in the CHNS dataset, and are used as instrumental variables in regressions using that dataset. Current prices were also tried, but they had little explanatory power.

The instrumental variables used in the regressions based on the GSCF data include the price indices for several major categories of consumer goods and the counts of registered vendors of alcohol. The county-level price indices are for the year of 2007, collected from the Gansu Statistical Yearbook. These price indices track changes in prices. The correlation between youth smoking and these price indices is unlikely to reflect the effects of smoking on prices because teenage purchase of cigarettes accounts for a very small portion of cigarette sales. Because drinking and smoking are related, either as substitutes or as complements, the supply of alcohol may be negatively correlated with the demand for cigarettes negatively (as a substitute) or positively (as a complement). The counts of registered alcohol vendors are based on the records from China's Department of Commerce, which are available for about 40% of the GSCF sample. If the sample attrition is not random, the missing data may cause bias in the estimates. This problem will be addressed in Subsection 2.6.2.

The estimates of the determinants of smoking are reported in Table 2.12 for teenagers aged 12-17 from the CHNS sample and in Table 2.13 for teenagers aged 14-17 from the GSCF sample, respectively. In both tables, the first column reports the results from a probit regression, where the dependent variable equals one if ever smoked and zero if not. The second column presents the regression results from another probit regression, where the dependent variable equals one if the teenager is currently smoking and zero otherwise. The third column reports the tobit estimates for the determinants of the number of cigarettes smoked per day. Lastly, the fourth regression in Table 2.13 is an ordered probit regression for the total number of cigarettes smoked in entire life, available only in the GSCF dataset[6].

As shown in the tables, age and sex are important factors of youth smoking. Males are significantly more likely to smoke. Actually, the sex variable is dropped because there is only 1 female observation who smoked in the CHNS sample used in the regression analysis. The estimates for sex variable are still reported for the regressions based on GSCF data, reflecting the variation in the outcomes caused by 4 female observations who report having smoked.

Although parental education does not appear to have much impact on youth smoking, parental smoking status has a significantly negative effect on children's smoking behavior. Results in Table 2.12 suggest that mothers' smoking has a strong negative impact on participation of youth smoking. Since there are only 3 mothers who smoked in the GSCF sample, the mothers' smoking variable is automatically dropped by the statistical software. Therefore, only the estimates of fathers' smoking are reported, which indicate that

[6] The total number of cigarettes smoked in entire life is measured by a categorical variable which equals 0 if never smoked, 1 if smoked 1-5 cigarettes, 2 if smoked more than 5 cigarettes.

teenagers whose fathers smoke are more likely to smoke and smoke more if they smoke. A possible explanation of the parental smoking effects is that living in a household where parents smoke makes it much easier for a teenager to obtain access to cigarettes. Moreover, seeing one's parents smoke may cause teenagers to underestimate the adverse health consequences of smoking and to imitate their parents. Although not robust to different specifications, household income is found to increase the probability of current smoking.

As for the instrumental variables in Table 2.12, higher Marlboro cigarette prices seem to discourage both the participation and the amount of smoking, significant at the 5% level. On the other hand, the price of local brand cigarettes reported as most commonly consumed, appears insignificant. At a first glance, this may appear counterintuitive because teenagers probably smoke cheap cigarettes, and so they should be less responsive to the price of expensive Marlboro cigarettes than to that of local brand cigarettes. However, these findings persist in different specifications. A possible explanation for this may be that Marlboro cigarette price is a better measure of general market price of cigarettes than the price of popular local brand cigarette, which may be more likely to suffer from measurement error because the definition of "commonly consumed" is vague and, in some sense, arbitrary. Another explanation is that it is probably more "cool" to smoke Marlboro cigarettes for teenagers, whose main motivation for smoking is to be "cool". Note that neither price affects the probability of ever smoking. Indeed, this is reasonable because few teenagers purchase their first cigarettes. The first cigarettes are usually from friends or family members (Forster, et al., 1997; Jones, et al., 2002).

The estimates of the instrumental variables for the GSCF sample are reported in the bottom panel in Table 2.13. In general, the number of alcohol vendors and the price index for clothing are negatively correlated with smoking, and are significant at the 5% level. This is probably because alcohol is a substitute good for cigarettes. However, one needs to be cautious when interpreting the impact of the price index for clothing. It is hard to tell if clothing is a complement or a substitute for cigarettes because the price index reflects changes in prices, rather than the levels of prices. Note that the price index of drink, cigarettes and alcohol is not significant in any of the four specifications. Because this variable is a very noisy indicator of cigarette prices, it does not have much explanatory power.

In order to examine whether the instrumental variables are weak, the F-statistic from regressing the amount of smoking on the excluded instrumental variables only is also checked. A higher F-statistic means that the estimates for instrumental variables are jointly and significantly different from zero and can predict enough exogenous variation in the endogenous variable. The F-statistic for the instrumental variables from the CHNS data, the prices of cigarettes, is 2.92 (p-value=0.0542) and that for the instrumental variables from the GSCF data is 7.92 (p-value=0.0001). These results suggest that the prices of cigarettes from the CHNS data may be weak instrumental variables and the IV estimates may suffer from large standard errors. However, if the instrumental variables and error term in equation (10) are uncorrelated, IV estimators are still consistent (Wooldrige, 2009).

2.6.2 Test scores

The comprehensive data on scores on Chinese and Mathematics academic tests in the GSCF allow us to examine the impact of youth smoking on academic performance, or learning per year of schooling. Table 2.14 presents estimates for test scores, six from OLS regressions (three for math scores and three for Chinese scores) and four from instrumental variable (IV) regressions (two for each subject). The dependent variables are test scores standardized by the means and standard deviations of each grade level. Because the information on alcohol vendors, one of the important instrumental variables, is not available for the whole sample, results from OLS regressions based on both the full sample and the reduced sample are provided in the first two columns for comparison. The second and the third regressions are essentially the same except that the smoking variables in the former measure the amount of cigarettes smoked in the students' lifetime, while that the latter measures the amount smoked per day. The fourth, fifth and sixth regressions are specified in the same way for Chinese test scores. The last four columns present estimates from IV regressions. The seventh and eighth regressions are for math scores, using different smoking variables. The last two regressions are the same as the seventh and eighth regressions, but for Chinese scores.

According to a comparison between the regression results based on the full sample and the reduced sample, the estimates of some variables are sensitive to different sample sizes. For example, the signs and/or significance levels of the estimates of household land assets and some school and teacher variables change dramatically after the sample size is reduced. Therefore, the estimates for these variables are not very reliable. But because these variables are not important for this study, there is no need to worry about them and the discussion of these variables will be very brief.

In the OLS regressions, smoking 5 or more cigarettes over a student's entire life appears to have negative effects on both math and Chinese test scores, although it is only marginal significant for math in the full sample regression. However, the estimates from IV regressions show that smoking more than 5 cigarettes in entire life can reduce the math test score by 0.02 standard deviations, significant at the 5% level. This effect is also observed for the Chinese test score, but it is not statistically significant. Note that the IV estimates are smaller than the OLS estimates in absolute value, suggesting that the OLS regressions tend to overestimate the negative effect of smoking due to the endogeneity issues discussed in Subsection 2.4.1. There are several possible explanations for the negative effects of youth smoking on academic performance. First, smoking can biologically reduce learning abilities, especially during the periods when one can't smoke freely. Smoking can also have a negative impact on health and the absorption of some key nutrients, which can also effect learning negatively. In addition to the health consequences of smoking, teenagers may spend less time on study. For example, teenager smokers may spend much of their spare time on smoking at friends' places.

Unlike the total amount of cigarettes smoked over one's entire life, which may reflect teenagers' experimental smoking experience, daily consumption of cigarettes measures teenage smokers' persistent habit, which may have a much larger impact on biologic learning productivity. Indeed, both the OLS estimates and IV estimates indicate that the daily consumption of cigarettes has a much more significant and greater negative impact on test scores for both subjects. The IV estimates suggest that smoking one more cigarette per day can reduce test scores by 0.14 standard deviations for math and 0.10 standard deviations for Chinese. The estimates in OLS regressions are smaller and less significant, indicating that OLS will underestimate the effect of persistent smoking behavior.

Interestingly, whether one has ever smoked or not does not appear to have any significant impact on test scores. Neither does the years since the first time smoking. This suggests that the negative effect of smoking on educational achievement is due to the habit of smoking instead of experimenting with smoking. For experimental smokers, who are not smoking on a regular basis and not addicted to cigarettes, smoking may change neither their amount of effort devoted to study much, nor is it likely to produce a significant harmful impact on learning. This finding is aligned with medical findings on the effects of nicotine on human performance.

Briefly consider the other variables in the regression, boys generally perform significantly better than girls in both subjects, and the difference in math scores is larger than that in Chinese scores. It is interesting to note that the advantage of boys is observed only after the endogeneity of smoking is corrected. Actually, girls are found to score higher in Chinese in the OLS regressions. For both subjects, parental education is an important factor. A possible explanation is that better educated parents can better assist their children with study and usually have a preference for more educated children. Parental education may also be an important indicator of their children's genetic ability. Children from richer families score higher than those from poorer families. This probably reflects the fact that richer families can allocate more resources to their children to improve both their learning (e.g. study materials) and their nutrition (e.g. better food).

The tuition at the junior high school, and the distance from households is the closest senior high school appear to have strong positive effects on test scores. Because some school and teacher characteristics, especially those of senior high schools, are not controlled for in the regressions, the estimates of school tuition and distance variables may be picking up the effects of these unobserved factors. The regression analysis also identifies a strong effect of having a science lab on test scores. This is consistent with the findings in Zhao and Glewwe (2010) which also reports a positive effect of a science lab on years of schooling. In addition, some teachers' characteristics also affect students' test performance. For example, teachers' experience and monetary incentives for teachers increase test scores in both subjects significantly. However, one needs to be cautious about the estimates of teachers' experience, because they appear to be very sensitive to sample sizes. The estimates of the teachers' bonus variable are more robust and consistent across different estimation specifications. Besides their monthly salary, teachers in China are paid annual bonuses based on the school principal's assessment of their teaching performance. Higher bonuses may provide incentives for teachers to work hard and improve students' academic performance. It is interesting that such stimulating effect is not found for teachers' salary, which may be due to the fact that teachers' monthly salaries are usually fixed based on one's degree level, working experiences, employment status, etc. and, therefore, less relevant to their teaching performance.

Since there are more instrumental variables than the endogenous variables, an overidentification test is conducted for each of the IV regressions. The chi-square statistic (p-value=0.9) fails to reject the null hypothesis that the instrumental variables are uncorrelated with the error term specified in equation (9). This also implies that the IV estimates are less likely to suffer from bias caused by weak instrumental variables. Lastly, since the IV regressions are based on a smaller sample with instrumental variables, the model is also estimated for the whole sample by replacing all the missing instrumental variables with their sample means. The estimates are very similar to those reported in Table 14, indicating that the IV regression results are robust to different sample sizes and may be generalized.

2.6.3 Years of schooling

The impact of youth smoking on educational attainment is estimated using a censored ordered probit (COP), as discussed in Subsection 2.4.1. Table 2.15 shows the results from six regressions based on the CHNS data. The first and the second are essentially the same except that they use different smoking variables: current smoking status versus number of cigarettes smoked per day. The third regression includes both variables. The fourth, fifth, and sixth regressions are instrumental variable (IV) regressions based on the Rivers and Vuong 2-step estimation procedure (1988), where the lags of prices of local brand cigarettes and Marlboro cigarettes are the instrumental variables for youth smoking. Again, these three IV regressions are essentially the same except that one includes only current smoking status, one includes only the amount of cigarettes smoked per day and the last includes both.

Based on the maximum likelihood estimates from the COP regressions, current smoking status appears to have significantly negative impact on years of schooling, regardless whether the daily amount of smoking is included or not. On the other hand, a negative effect of daily consumption of cigarettes on years of schooling is statistically significant only when current smoking status is not included. After using cigarette prices to instrument youth smoking, the negative effect of current smoking status becomes completely insignificant, while the number of cigarettes smoked per day is marginal significant and consistent regardless of the inclusion of current smoking. The comparison indicates that the COP estimates of smoking can be seriously upward biased due to the endogeneity problem. What really matters to educational attainment is not whether a teenager has ever smoked or not, but the persistent habit of smoking as captured by his or her daily smoking behavior.

Based on the IV estimates and the sample means, smoking one more cigarette per day can decrease years of schooling by 0.015, or 5.5 days. At the sample mean of 9.3 cigarettes smoked per day, years of schooling would be shortened by 51 days, due to smoking. Smoking may reduce educational attainment for several reasons. Smoking can biologically reduce learning abilities or reduce learning effort, as discussed in Subsection 2.6.1. Poor academic performance plays an important role in determining years of schooling. It is closely related to the expected return to the investment in education. For example, students, as well as their families, may expect less return to education if they perform poorly in academic tests and, therefore, are more likely to drop out of school. Low test scores can also prevent students from entering high schools or colleges. Another reason why smoking can shorten years of schooling may be related to the motivation to go to school. Smoking is prevented in school, as required by law in China. However, because there is no law that specifies a legal minimum age for smoking, a dropout will have much more freedom to smoke. Therefore, teenage smokers who have become addicted to smoking may have a stronger incentive to drop out of school than their non-smoking counterparts.

Another important finding is that mothers' smoking has a significant effect on children's educational attainment. The impact of mothers's smoking status is very significant after the endogeneity of smoking choices is corrected for. According to many clinical studies (Mark et al., 1990; Pollack et al., 2000; Perreira et al., 2006), maternal smoking can increase the risk of low-birth-weight and premature delivery. Moreover, children of smoking women are more likely to have been exposed to passive smoking, which has also been found to be detrimental to child health and cognitive development (Rechards, 1996; Surgeon General Report, 2006; Gilbert, 2007).

Turning to the other variables in the regression, parents' education and household income have significant positive impacts on children's educational attainment, which is well aligned with the findings in previous literature. Teenagers in rural areas generally attain significantly less years of schooling than their peers in urban areas. In addition, conditional on school and teacher qualities, the distance from home to the closest senior school appears to have a negative impact on years of schooling, suggesting that some students may drop out of school or reduce their optimal amount of education because of the higher opportunity (time) costs of attending a senior high school far away from home.

To test the robustness of the 2-stage (censored) ordered probit, other specifications such as an enrollment probit model were also tried. The results are qualitatively similar except that estimates from other specifications are less robust. One thing to note is that the estimates associated with the amount of cigarettes smoked are sensitive to the use of Tobit specification in the first stage. This is mainly because about 95% of the observations are censored, as they are non-smokers.

2.7. The effect of smoking on future income

In the recent decades, education has been identified as an important determinant of economic growth and of individuals' incomes. Many previous studies have attempted to estimate the economic return to education. Psacharopoulos (1994) presents estimates that the world average return to one year of schooling is about 10%, and that the average rate in Asia is also about 10%. However, the private returns to investment in education in China have been found to be very low during the late 1980's and early 1990's, ranging from 1 to 4 percent. More recent studies report that the rate is higher, and this is attributed to the development of capital and labor markets. Fleisher et al. (2004), Heckman (2003), Zhang et al. (2005) all estimate that the private return to education in China is currently about 10%. Fleisher et al. present evidence that skilled labor is underpaid in China, since the social return to education may be as high as 30-40%.

Drawing on these findings, the simulations done here assume that the return to schooling in China is 10%. To explore the sensitivity of estimates of the income loss due to smoking to this assumption, two other rates of return to education are also tried: 4% and 15%. According to Zhang et al. (2005), as the labor market becomes more developed, the return to schooling could increase to 15% by 2010.[7] Youths today will enter the labor market 5-10 years later. The 15% rate of return is used to see how the predicted increase in the reward to investment in education can affect the simulated income loss.

[7] Zhang et al. estimated the return to schooling in China from 1988 to 2001 year by year. The rate of 15% is projected based on the trend in the series of the estimates they suggested.

The interest rate used also plays an important role in the simulation, because it determines the present discounted value of the cash flows that occur in the future. A 3% rate is commonly used and is recommended by previous studies (see Sloan et al., 2004; Ibbotson et al., 1976; Gold et al., 1996). However, a rate of 5% is also used for comparison purposes.

The simulation proceeds in three steps. First, a series of annual future income, in present value, from age 24 to 60, is projected. Because working experience also explains how annual income changes over time, it is also included to calculate future income. The coefficients for experience and experience-squared used, 0.31 and -0.0004, are taken from Heckman et al. (2004) and Zhang et al. (2005). These numbers are used because both studies found very similar impacts of experience, and they are the most recently published studies that estimate an earnings function for China. Starting from age 24, a person's annual income in the tth working year is calculated as:

$$income_{t+1} = \frac{income_t + 0.031t - 0.0004t^2}{(1+r)^t}$$

where r is the long term interest rate. Secondly, based on the predicted reduction of years of schooling (YOS) attributable to youth smoking (YS) according to the estimates in Table 15, the annual income loss in the tth working year is given by:

$$loss_t = income_t \times \frac{\partial income_t}{\partial YOS_t} \times \frac{\partial YOS_t}{\partial YS}.$$

Lastly, summing up the total of the annual income loss from age 24 to 60 and dividing it by the total of the income will give the percentage of income loss.% $lifetime\ income$

$$loss = \frac{\sum_{t=24}^{60} loss_t}{\sum_{t=24}^{60} income_t} \times 100\%$$

These three steps are repeated for different scenarios that assume different interest rates and rates of return to education.

Table 2.16 presents the simulated percentages of income loss a teenager who smokes 1 cigarette, 3 cigarettes or 5 cigarettes per day would face later in his or her life. The table reports the impact on income of achieving fewer years of schooling due to smoking. The benchmarking percentages for different levels of daily consumption of cigarettes, using 10% as return to schooling and 3% as the interest rate, are 0.2%,0.5% and 0.8%, respectively. The loss grows as the return to schooling increases, so that smoking 5 cigarettes can result in an income loss of as high as 0.9% when education is highly rewarded (15% rate of return). The interest rates also make difference. If the present value of future cash flow is lower, that is, if the interest rate is 5% rather than 3%, then the income loss will be somewhat smaller.

In addition to years of schooling, youth smoking can also affect educational achievement. Thus, ideally, the simulation should also consider the effect on lifetime income of lower test scores per year of schooling caused by smoking. However, reliable and consistent estimates of the effect of educational achievement on earnings are not available. Therefore, the predicted percentages of income loss discussed above under-estimate the real effect of youth smoking on future income.

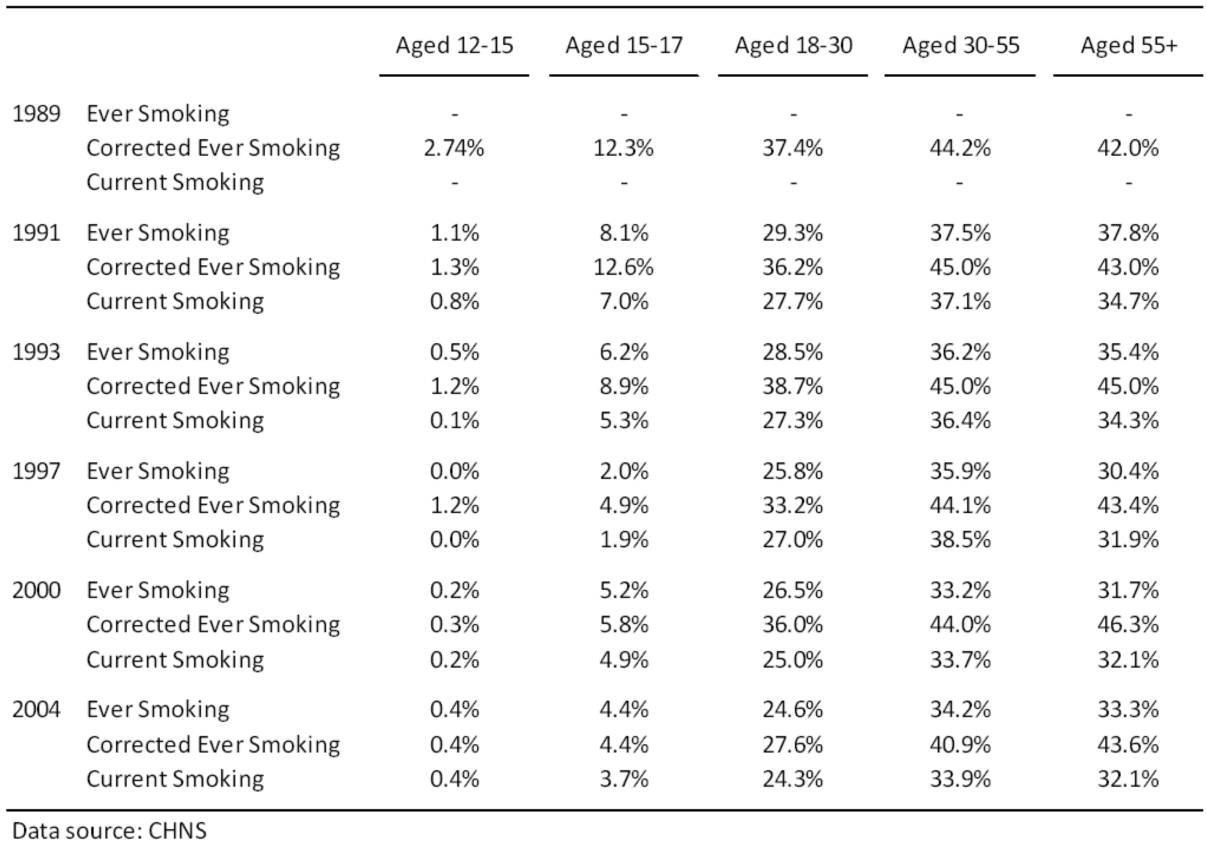
2.8. Conclusions

Using two rich datasets from China, this study has shown that youth smoking has a significant and economically important impact on educational outcomes. Smoking one cigarette per day at ages 14-17 is estimated to reduce test scores in Chinese and math exams by about 0.1 standard deviations. Similarly, smoking one cigarette per day at ages 12-17 is estimated to reduce the years of schooling by 0.015 years, or 5.5 days. The smoking-induced loss in educational attainment can translate into a lifetime income loss of 0.2%. Note that the income loss is actually under-estimated as the effect of smoking-induced reduction in learning per school year is not taken into account.

There are two caveats to the results of this study. First, the lifetime income loss could be underestimated since smoking may plausibly have adverse impacts on the quality of education at the college level. In other words, smoking may not have large impacts on a decision to go to a college, but may affect the quality of colleges to which they are admitted. Second, although the instrumental variables used in this study are plausibly uncorrelated with unobserved errors and have passed specification tests, past cigarette prices could still be correlated with *past* demand for education in the years of schooling equation, as individuals may choose educational inputs and smoking jointly. To address these concerns, future research should investigate the effect of smoking on high school graduates' college admissions.

Tables and figures

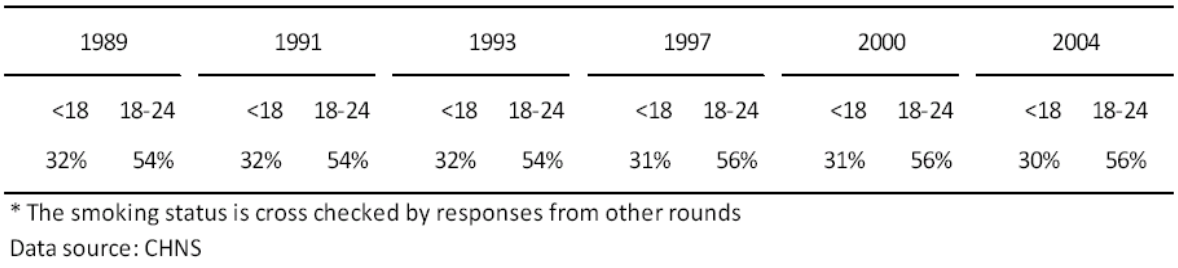
Table 2.1: Prevalence of smoking (CHNS data)



		Aged 12-15	Aged 15-17	Aged 18-30	Aged 30-55	Aged 55+
1989	Ever Smoking	-	-	-	-	-
	Corrected Ever Smoking	2.74%	12.3%	37.4%	44.2%	42.0%
	Current Smoking	-	-	-	-	-
1991	Ever Smoking	1.1%	8.1%	29.3%	37.5%	37.8%
	Corrected Ever Smoking	1.3%	12.6%	36.2%	45.0%	43.0%
	Current Smoking	0.8%	7.0%	27.7%	37.1%	34.7%
1993	Ever Smoking	0.5%	6.2%	28.5%	36.2%	35.4%
	Corrected Ever Smoking	1.2%	8.9%	38.7%	45.0%	45.0%
	Current Smoking	0.1%	5.3%	27.3%	36.4%	34.3%
1997	Ever Smoking	0.0%	2.0%	25.8%	35.9%	30.4%
	Corrected Ever Smoking	1.2%	4.9%	33.2%	44.1%	43.4%
	Current Smoking	0.0%	1.9%	27.0%	38.5%	31.9%
2000	Ever Smoking	0.2%	5.2%	26.5%	33.2%	31.7%
	Corrected Ever Smoking	0.3%	5.8%	36.0%	44.0%	46.3%
	Current Smoking	0.2%	4.9%	25.0%	33.7%	32.1%
2004	Ever Smoking	0.4%	4.4%	24.6%	34.2%	33.3%
	Corrected Ever Smoking	0.4%	4.4%	27.6%	40.9%	43.6%
	Current Smoking	0.4%	3.7%	24.3%	33.9%	32.1%

Data source: CHNS

Table 2.2: Ages when starting to smoke, all ever-smokers (CHNS data)

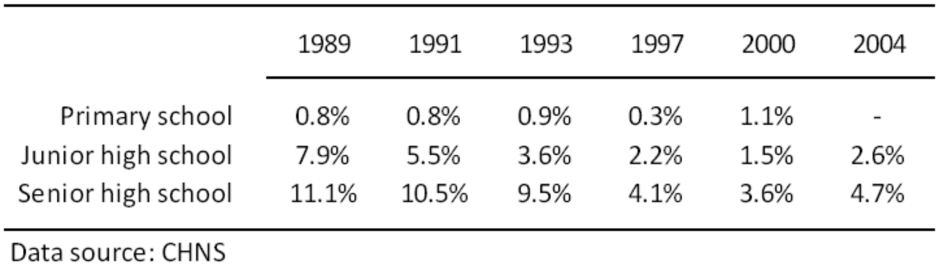


1989		1991		1993		1997		2000		2004	
<18	18-24	<18	18-24	<18	18-24	<18	18-24	<18	18-24	<18	18-24
32%	54%	32%	54%	32%	54%	31%	56%	31%	56%	30%	56%

* The smoking status is cross checked by responses from other rounds

Data source: CHNS

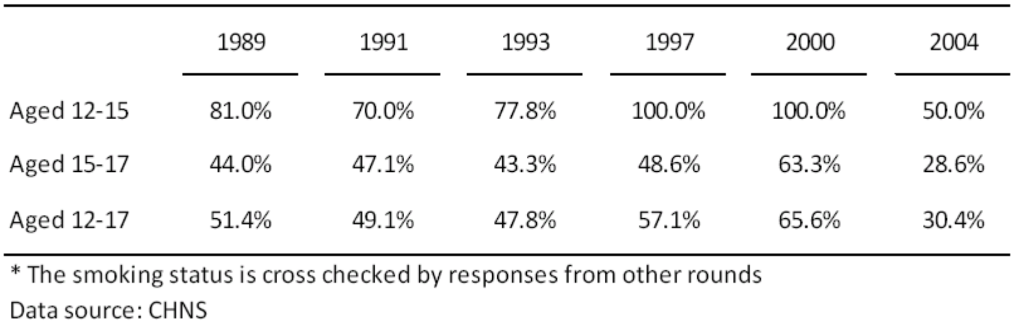
Table 2.3: Prevalence of (corrected) ever smoking in school (CHNS data)



	1989	1991	1993	1997	2000	2004
Primary school	0.8%	0.8%	0.9%	0.3%	1.1%	-
Junior high school	7.9%	5.5%	3.6%	2.2%	1.5%	2.6%
Senior high school	11.1%	10.5%	9.5%	4.1%	3.6%	4.7%

Data source: CHNS

Table 2.4: Smoked before dropping out, ever-smokers aged 15-17 (CHNS data)



	1989	1991	1993	1997	2000	2004
Aged 12-15	81.0%	70.0%	77.8%	100.0%	100.0%	50.0%
Aged 15-17	44.0%	47.1%	43.3%	48.6%	63.3%	28.6%
Aged 12-17	51.4%	49.1%	47.8%	57.1%	65.6%	30.4%

* The smoking status is cross checked by responses from other rounds

Data source: CHNS

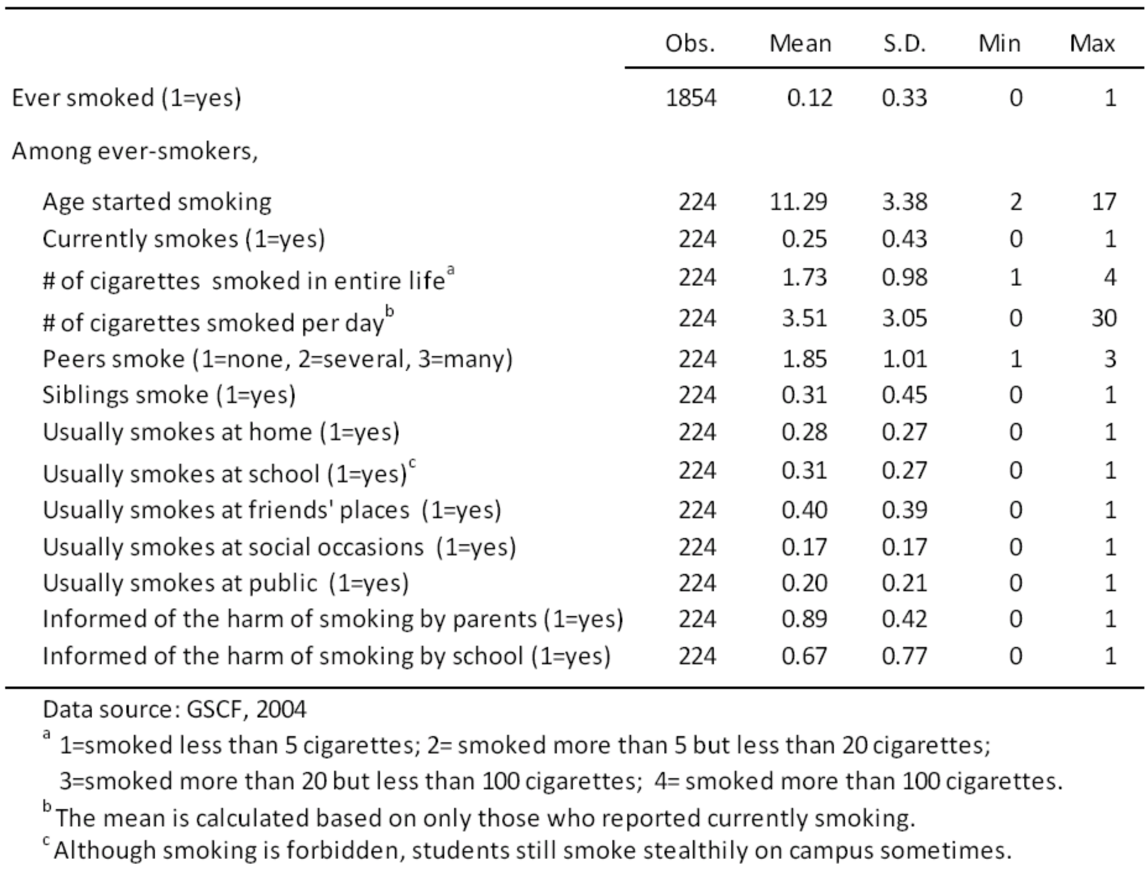
Table 2.5: Completed years of formal education, aged 12-17 (CHNS data)

	1989		1991		1993		1997		2000		2004	
	Out[a]	In[b]	Out	In	Out	In	Out	In	Out	In	Out	In
None	1.9%	0.2%	1.0%	0.3%	0.4%	0.0%	0.4%	0.0%	0.3%	0.0%	0.0%	0.0%
1-6 years, primary	13.7%	32.5%	11.2%	28.1%	9.6%	24.1%	4.9%	23.5%	4.3%	16.8%	1.8%	14.4%
1-3 years, lower secondary	11.1%	37.1%	9.6%	44.0%	12.8%	47.1%	8.8%	52.3%	10.7%	58.0%	9.5%	61.0%
1-3 years, upper secondary	0.1%	3.1%	0.2%	5.2%	0.1%	4.7%	0.2%	7.7%	0.3%	7.6%	0.1%	11.3%
1-3 years, technical school	0.0%	0.3%	0.2%	0.2%	0.0%	1.2%	0.1%	2.2%	0.0%	1.6%	0.3%	1.4%
1-6 years, college/university	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.3%	0.0%	0.2%
Total	376	1025	268	935	274	923	162	967	183	984	103	768
	27%	73%	22%	78%	23%	77%	14%	86%	16%	84%	12%	88%

[a] Out of school; [b] In school

Data source: CHNS

Table 2.6: Prevalence of smoking and smoking behaviors, aged 14-17 (GSCF data)



	Obs.	Mean	S.D.	Min	Max
Ever smoked (1=yes)	1854	0.12	0.33	0	1
Among ever-smokers,					
Age started smoking	224	11.29	3.38	2	17
Currently smokes (1=yes)	224	0.25	0.43	0	1
# of cigarettes smoked in entire life[a]	224	1.73	0.98	1	4
# of cigarettes smoked per day[b]	224	3.51	3.05	0	30
Peers smoke (1=none, 2=several, 3=many)	224	1.85	1.01	1	3
Siblings smoke (1=yes)	224	0.31	0.45	0	1
Usually smokes at home (1=yes)	224	0.28	0.27	0	1
Usually smokes at school (1=yes)[c]	224	0.31	0.27	0	1
Usually smokes at friends' places (1=yes)	224	0.40	0.39	0	1
Usually smokes at social occasions (1=yes)	224	0.17	0.17	0	1
Usually smokes at public (1=yes)	224	0.20	0.21	0	1
Informed of the harm of smoking by parents (1=yes)	224	0.89	0.42	0	1
Informed of the harm of smoking by school (1=yes)	224	0.67	0.77	0	1

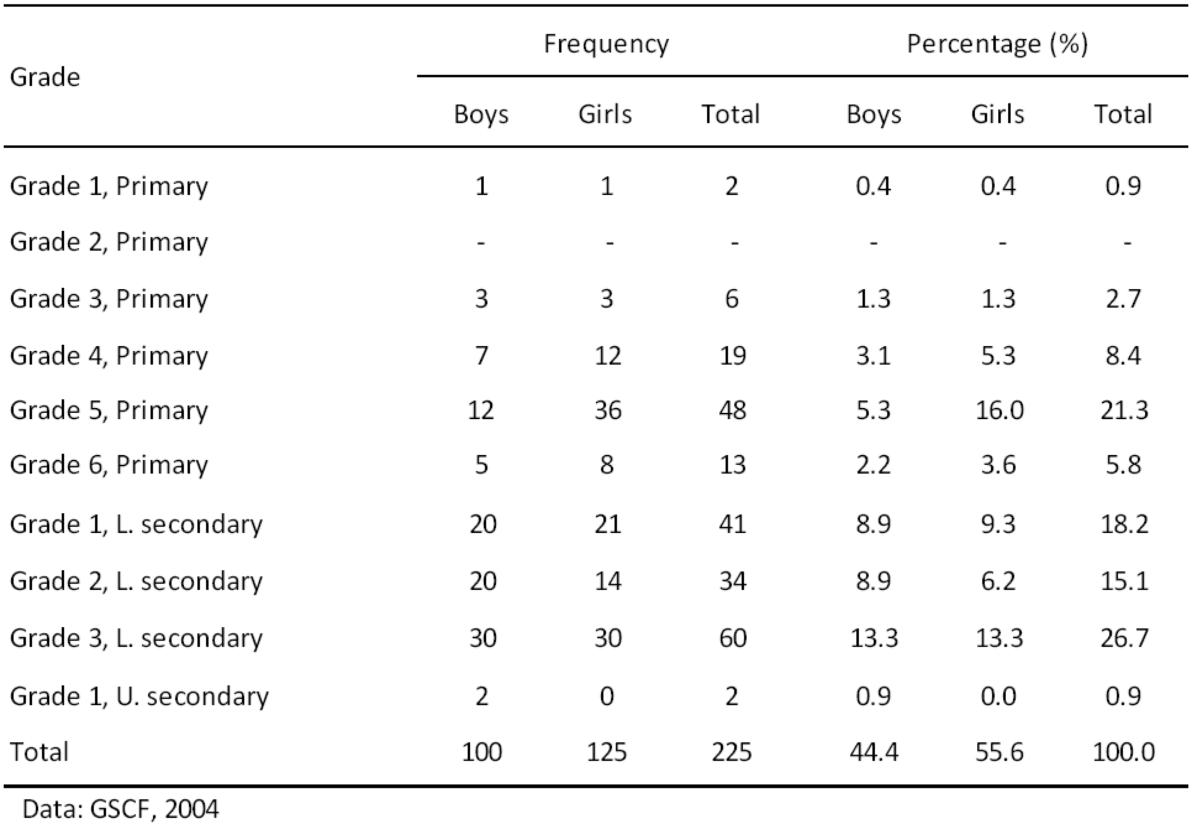
Data source: GSCF, 2004

[a] 1=smoked less than 5 cigarettes; 2= smoked more than 5 but less than 20 cigarettes; 3=smoked more than 20 but less than 100 cigarettes; 4= smoked more than 100 cigarettes.

[b] The mean is calculated based on only those who reported currently smoking.

[c] Although smoking is forbidden, students still smoke stealthily on campus sometimes.

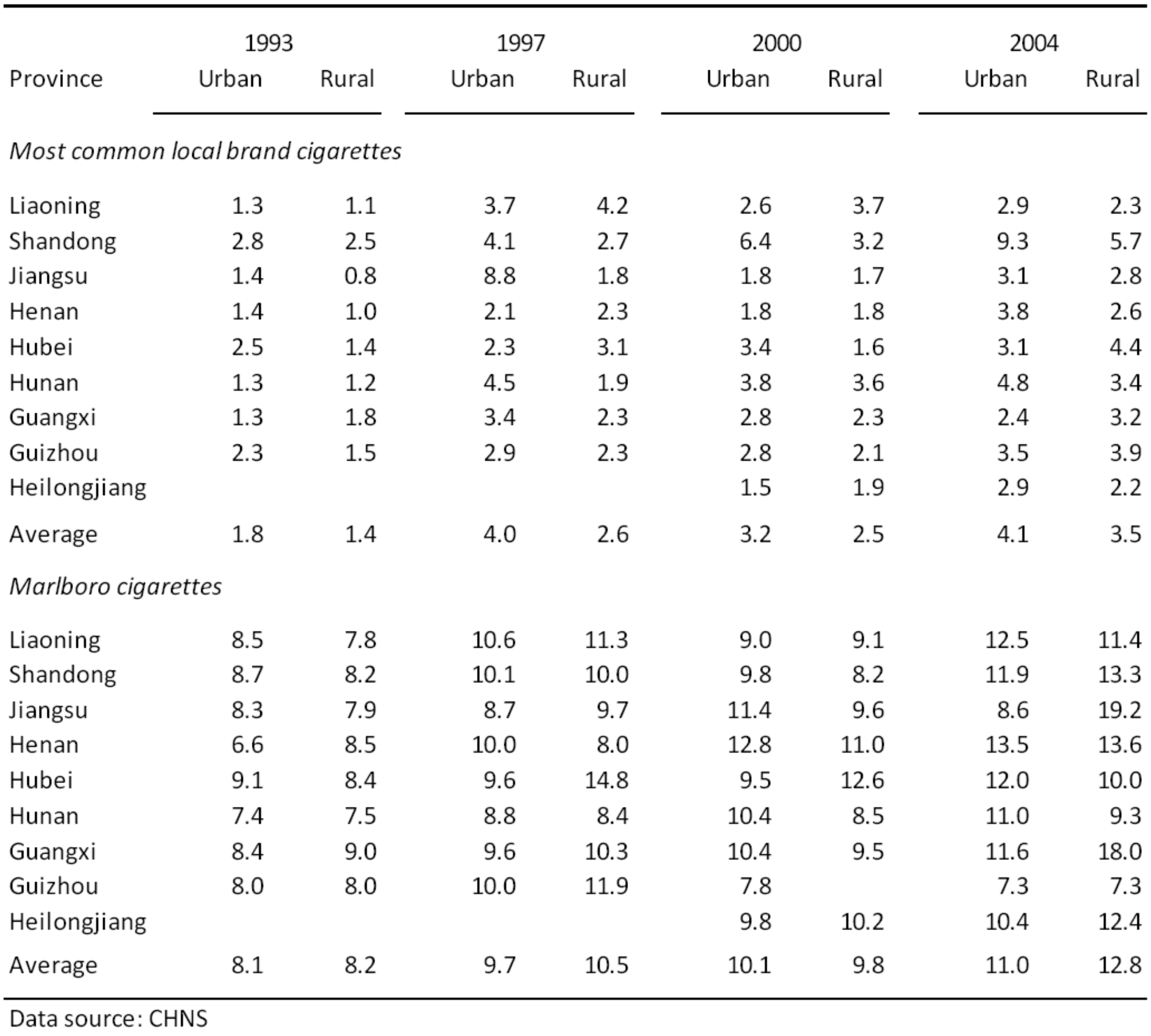
Table 2.7: Highest grades attained by dropouts, aged 14-17 (GSCF data)



Grade	Frequency			Percentage (%)		
	Boys	Girls	Total	Boys	Girls	Total
Grade 1, Primary	1	1	2	0.4	0.4	0.9
Grade 2, Primary	-	-	-	-	-	-
Grade 3, Primary	3	3	6	1.3	1.3	2.7
Grade 4, Primary	7	12	19	3.1	5.3	8.4
Grade 5, Primary	12	36	48	5.3	16.0	21.3
Grade 6, Primary	5	8	13	2.2	3.6	5.8
Grade 1, L. secondary	20	21	41	8.9	9.3	18.2
Grade 2, L. secondary	20	14	34	8.9	6.2	15.1
Grade 3, L. secondary	30	30	60	13.3	13.3	26.7
Grade 1, U. secondary	2	0	2	0.9	0.0	0.9
Total	100	125	225	44.4	55.6	100.0

Data: GSCF, 2004

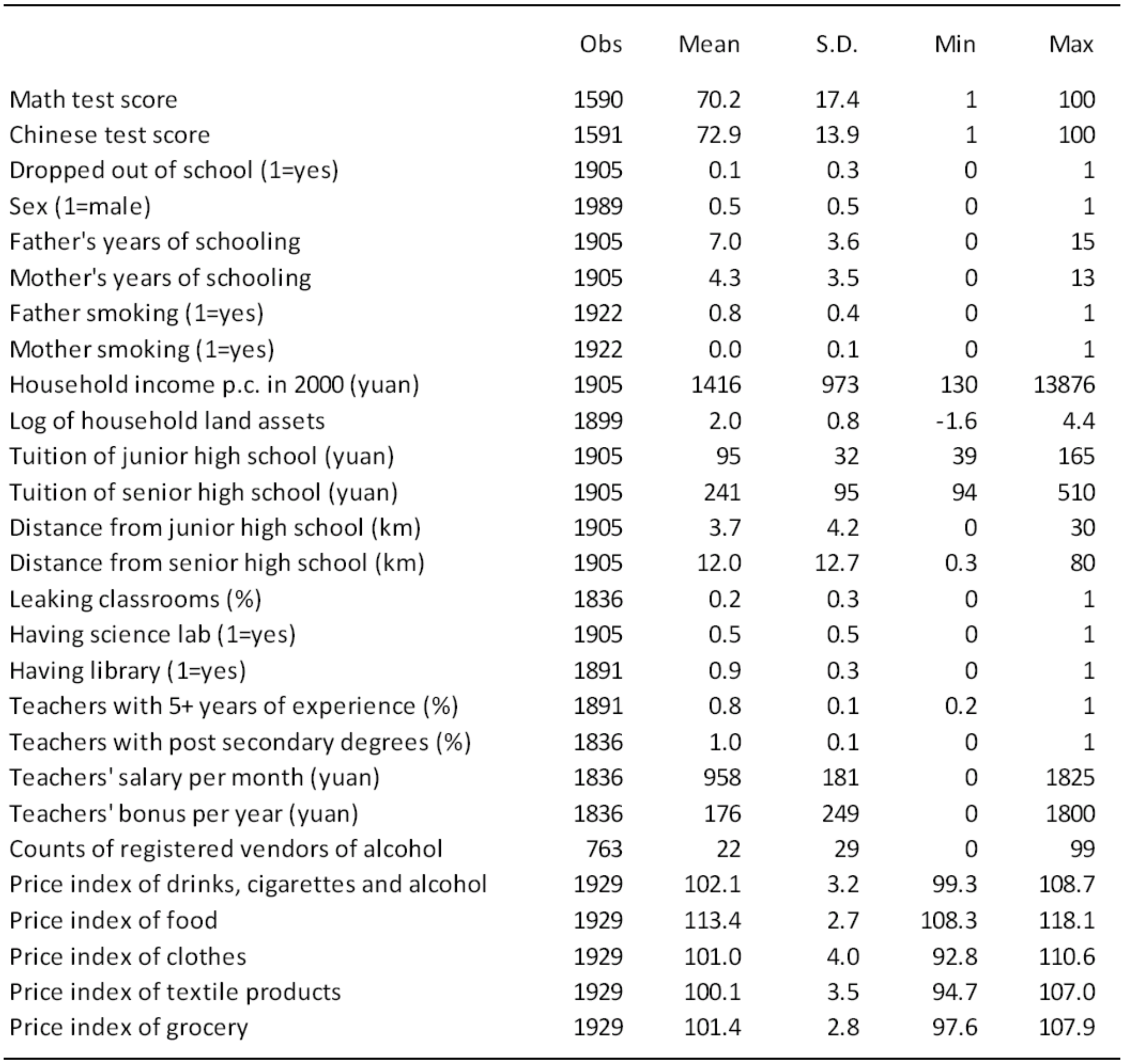
Table 2.8: Cigarette prices (yuan), 1993-2004 (CHNS data)



	1993		1997		2000		2004	
Province	Urban	Rural	Urban	Rural	Urban	Rural	Urban	Rural
Most common local brand cigarettes								
Liaoning	1.3	1.1	3.7	4.2	2.6	3.7	2.9	2.3
Shandong	2.8	2.5	4.1	2.7	6.4	3.2	9.3	5.7
Jiangsu	1.4	0.8	8.8	1.8	1.8	1.7	3.1	2.8
Henan	1.4	1.0	2.1	2.3	1.8	1.8	3.8	2.6
Hubei	2.5	1.4	2.3	3.1	3.4	1.6	3.1	4.4
Hunan	1.3	1.2	4.5	1.9	3.8	3.6	4.8	3.4
Guangxi	1.3	1.8	3.4	2.3	2.8	2.3	2.4	3.2
Guizhou	2.3	1.5	2.9	2.3	2.8	2.1	3.5	3.9
Heilongjiang					1.5	1.9	2.9	2.2
Average	1.8	1.4	4.0	2.6	3.2	2.5	4.1	3.5
Marlboro cigarettes								
Liaoning	8.5	7.8	10.6	11.3	9.0	9.1	12.5	11.4
Shandong	8.7	8.2	10.1	10.0	9.8	8.2	11.9	13.3
Jiangsu	8.3	7.9	8.7	9.7	11.4	9.6	8.6	19.2
Henan	6.6	8.5	10.0	8.0	12.8	11.0	13.5	13.6
Hubei	9.1	8.4	9.6	14.8	9.5	12.6	12.0	10.0
Hunan	7.4	7.5	8.8	8.4	10.4	8.5	11.0	9.3
Guangxi	8.4	9.0	9.6	10.3	10.4	9.5	11.6	18.0
Guizhou	8.0	8.0	10.0	11.9	7.8		7.3	7.3
Heilongjiang					9.8	10.2	10.4	12.4
Average	8.1	8.2	9.7	10.5	10.1	9.8	11.0	12.8

Data source: CHNS

Table 2.9: Descriptive statistics, aged 14-17 (GSCF data)



	Obs	Mean	S.D.	Min	Max
Math test score	1590	70.2	17.4	1	100
Chinese test score	1591	72.9	13.9	1	100
Dropped out of school (1=yes)	1905	0.1	0.3	0	1
Sex (1=male)	1989	0.5	0.5	0	1
Father's years of schooling	1905	7.0	3.6	0	15
Mother's years of schooling	1905	4.3	3.5	0	13
Father smoking (1=yes)	1922	0.8	0.4	0	1
Mother smoking (1=yes)	1922	0.0	0.1	0	1
Household income p.c. in 2000 (yuan)	1905	1416	973	130	13876
Log of household land assets	1899	2.0	0.8	-1.6	4.4
Tuition of junior high school (yuan)	1905	95	32	39	165
Tuition of senior high school (yuan)	1905	241	95	94	510
Distance from junior high school (km)	1905	3.7	4.2	0	30
Distance from senior high school (km)	1905	12.0	12.7	0.3	80
Leaking classrooms (%)	1836	0.2	0.3	0	1
Having science lab (1=yes)	1905	0.5	0.5	0	1
Having library (1=yes)	1891	0.9	0.3	0	1
Teachers with 5+ years of experience (%)	1891	0.8	0.1	0.2	1
Teachers with post secondary degrees (%)	1836	1.0	0.1	0	1
Teachers' salary per month (yuan)	1836	958	181	0	1825
Teachers' bonus per year (yuan)	1836	176	249	0	1800
Counts of registered vendors of alcohol	763	22	29	0	99
Price index of drinks, cigarettes and alcohol	1929	102.1	3.2	99.3	108.7
Price index of food	1929	113.4	2.7	108.3	118.1
Price index of clothes	1929	101.0	4.0	92.8	110.6
Price index of textile products	1929	100.1	3.5	94.7	107.0
Price index of grocery	1929	101.4	2.8	97.6	107.9

Data source: GSCF, 2004

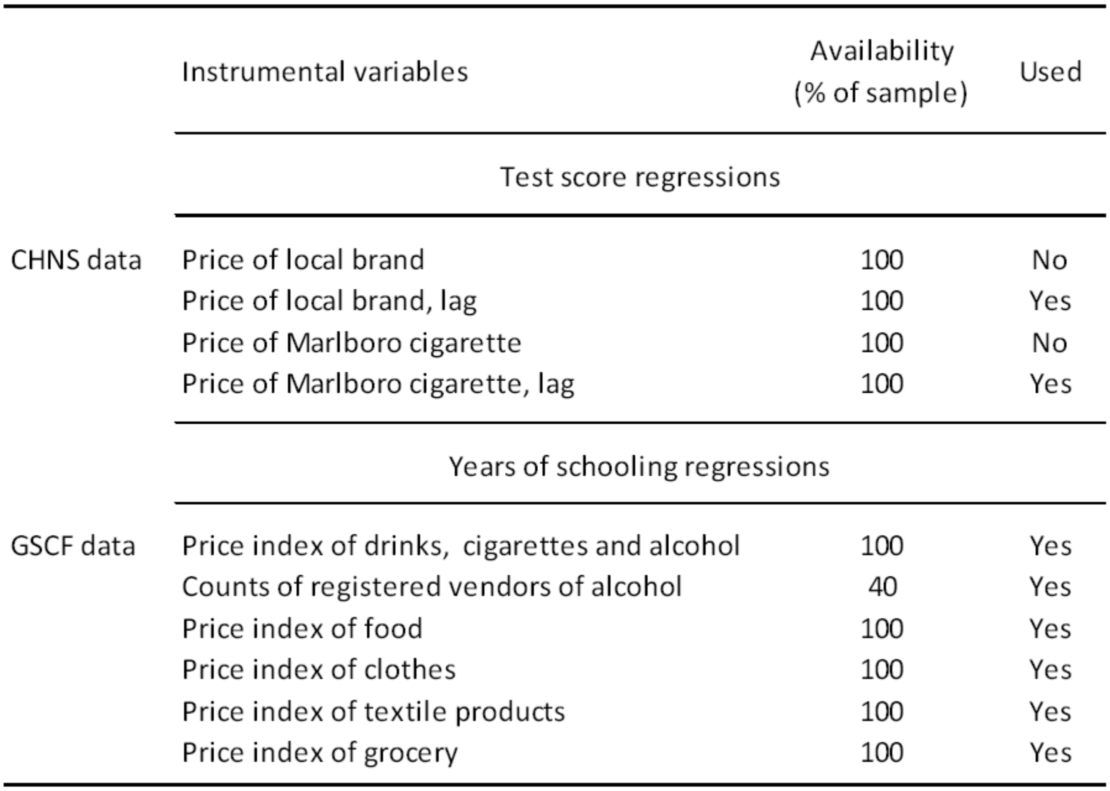
Table 2.10: Descriptive statistics, aged 12-17 (CHNS data)

	Obs	Mean	S.D.	Min	Max
Number of cigarettes smoked per day	85	9.4	7.5	1	30
Age	7907	14.5	1.5	12	17
Father's years of schooling	7121	7.3	3.5	0	18
Mother's years of schooling	7321	5.1	4.1	0	17
Father smoking (1=yes)	6091	0.7	0.5	0	1
Mother smoking (1=yes)	6986	0.0	0.2	0	1
Household income p.c. (in 1988 yuan)[a]	7774	1377	1243	-818	27463
Average wage for babysitting per day (yuan)	5697	13	24	0	200
Average wage for a construction worker per day (yuan)	7177	17	15	0	200
Average wage for a driver per month (yuan)	6668	535	545	0	9000
Distance from primary school (km)	6112	0.2	0.7	0	10
Distance from junior high school (km)	6062	1.6	3.3	0	101
Distance from senior high school (km)	5992	7.4	16.5	0	230
Urban residence (1=yes)	7907	0.3	0.4	0	1

Data source: CHNS

[a] Negative household incomes reflect a loss after taking into account inflation.

Table 2.11: Availability of instrumental variables



	Instrumental variables	Availability (% of sample)	Used
	Test score regressions		
CHNS data	Price of local brand	100	No
	Price of local brand, lag	100	Yes
	Price of Marlboro cigarette	100	No
	Price of Marlboro cigarette, lag	100	Yes
	Years of schooling regressions		
GSCF data	Price index of drinks, cigarettes and alcohol	100	Yes
	Counts of registered vendors of alcohol	40	Yes
	Price index of food	100	Yes
	Price index of clothes	100	Yes
	Price index of textile products	100	Yes
	Price index of grocery	100	Yes

Table 2.12: Determinants of youth smoking, aged 12-17 (CHNS data)

	Ever smoked, probit		Currently smoking, probit		Amount smoked per day, tobit	
	Coef.	Rob. S.E.	Coef.	Rob. S.E.	Coef.	Rob. S.E.
Age[a]	0.295 ***	0.052	0.350 ***	0.070	8.900 ***	2.664
Father's education	-0.014	0.010	-0.011	0.010	-0.250	0.321
Mother's education	0.005	0.009	0.001	0.012	0.013	0.395
Father smoking (1=yes)	0.126	0.161	-0.022	0.207	-1.358	6.343
Mother smoking (1=yes)	0.735 **	0.370	0.999 ***	0.379	26.840	49.355
Household income p.c. (100 yuan)	0.004	0.004	0.005 **	0.003	0.126	0.105
Average wage for babysitting (yuan)	-0.011 **	0.005	-0.004	0.004	-0.115	0.156
Average wage for a construction worker (yuan)	0.010	0.007	0.010	0.008	0.240	0.326
Average wage for a driver (100 yuan)	-0.003	0.012	-0.019	0.016	-0.399	0.677
Distance from primary school (km)	0.076	0.074	-0.011	0.123	0.348	5.666
Distance from junior high school (km)	-0.007	0.041	0.046	0.045	1.065	1.831
Distance from senior high school (km)	-0.001	0.007	-0.038	0.024	-0.881	1.688
Rural (1=yes)	0.008	0.009	0.269	0.215	-0.161	0.500
Instrumental variables						
Price of local brand, lag	0.040	0.057	0.031	0.070	0.396	4.299
Price of Marlboro cigarettes, lag	-0.037	0.035	-0.077 **	0.037	-1.936 **	0.953
No. of Obs.		2017		2017		2017
(Pseudo) R2		0.130		0.170		0.110
Log likelihood		-174		-89		-152

* Significant at 10% level, ** significant at 5% level, *** significant at 1% level

Data source: CHNS

[a] Age-squared is not statistically significant.

Table 2.13: Determinants of youth smoking, aged 14-17 (GSCF data)

	Ever smoked, probit		Currently smoking, probit		Amount smoked per day, tobit		Total amount smoked, ordered probit	
	Coef	Rob. S.E.	Coef	Rob. S.E.	Coef	Rob. S.E.	Coef	Rob. S.E.
Age	0.207 ***	0.061	0.192 ***	0.059	1.077 ***	0.333	0.204 ***	0.055
Sex (1=male)	1.675 ***	0.234	1.662 ***	0.235	7.925 ***	1.186	1.638 ***	0.229
Father's education	0.009	0.021	0.006	0.021	0.057	0.107	0.008	0.020
Mother's education	-0.019	0.023	-0.018	0.023	-0.070	0.111	-0.028	0.022
Father smoking (1=yes)	2.065 **	0.837	2.149 ***	0.835	9.828 ***	3.778	2.376 ***	0.780
Log of household income p.c.	0.059	0.169	0.055	0.169	-0.268	0.759	0.189	0.176
Log of household land assets	-0.224	0.174	-0.223	0.172	-0.418	0.810	-0.208 *	0.155
Log of tuition of junior high school	0.816	1.076	0.784	1.079	2.570	5.372	0.928	1.042
Log of tuition of senior high school	-0.362	0.761	-0.348	0.766	-1.814	3.608	-0.212	0.688
Distance from junior high school (km)	-0.015	0.028	-0.017	0.028	0.102	0.130	-0.011	0.028
Distance from senior high school (km)	0.001	0.008	0.001	0.008	0.000	0.041	-0.001	0.007
Leaking classrooms (%)	0.050	0.303	0.044	0.305	0.564	1.484	0.081	0.274
Having science lab (1=yes)	0.405 **	0.167	0.466 ***	0.163	0.838	0.931	0.431 ***	0.160
Having library (1=yes)	1.276	1.319	1.429	1.359	3.840	6.404	1.388	1.309
Teachers with 5+ years of experience (%)	0.625	1.148	0.646	1.148	0.208	5.466	0.480	1.041
Teachers with post secondary degrees (%)	-6.330 *	3.586	-6.304 *	3.710	-20.815	18.433	-5.265	3.317
Teachers' salary per month(100 yuan)	0.023	0.091	0.022	0.090	0.234	0.396	-0.029	0.103
Teachers' bonus per year (100 yuan)	0.032	0.040	0.034	0.040	0.056	0.021	0.023	0.034

	Ever smoked, probit		Currently smoking, probit		Amount smoked per day, tobit		Total amount smoked, ordered probit	
	Coef	Rob. S.E.	Coef	Rob. S.E.	Coef	Rob. S.E.	Coef	Rob. S.E.
Instrumental variables								
Price index of drinks, cigarettes and alcohol	-0.037	0.050	0.118	0.103	0.693	0.521	0.147	0.094
Counts of registered vendors of alcohol	-0.013 **	0.007	-0.013 **	0.007	-0.060 **	0.031	-0.014 **	0.007
Price index of food	0.110	0.101	-0.036	0.050	-0.038	0.245	-0.038	0.052
Price index of clothes	-0.273 **	0.122	-0.279 **	0.123	-0.949	0.622	-0.260 **	0.113
Price index of textile products	-0.168	0.203	-0.163	0.204	-0.365	0.980	-0.118	0.183
Price index of grocery	0.490	0.320	0.491	0.320	1.362	1.536	0.410	0.286
Constant	-15.312	15.453	-15.986	15.637	-85.744	78.345		
No. of Obs.		649		707		699		711
(Pseudo) R2		0.28		0.28		0.15		0.22
Log value		-203.00		-205.45		-408.00		-294.11

* Significant at 10% level, ** significant at 5% level, *** significant at 1% level

Data source: GSCF

Table 2.14: Youth smoking and test scores, aged 14-17 (GSCF data)

	OLS regressions						IV regressions			
	Math^a	Math^b	Math^b	Chinese^a	Chinese^b	Chinese^b	Math	Math	Chinese	Chinese
1-5 cigarettes	0.058	0.010		0.076	0.140		-0.001		-0.004	
	(0.096)	(0.139)		(0.091)	(0.132)		(0.015)		(0.017)	
5+ cigarettes	-0.178 *	-0.256		-0.256	-0.211		-0.021 **		-0.011	
	(0.101)	(0.170)		(0.170)	(0.292)		(0.009)		(0.011)	
Number of cigarettes smoked per day			-0.051 *			-0.071 **		-0.140 ***		-0.095 ***
			(0.031)			0.035		(0.029)		(0.031)
Age	0.008	0.044	0.043	0.049 **	0.091 **	0.095 ***	0.078 *	0.198 ***	0.116 ***	0.198 ***
	(0.024)	(0.036)	(0.036)	(0.023)	(0.036)	(0.037)	(0.041)	(0.051)	(0.043)	(0.054)
Sex (1=male)	-0.020	0.064	0.067	-0.164 ***	-0.114	-0.079	0.251 *	1.119 ***	0.030	0.618 ***
	(0.055)	(0.082)	(0.078)	(0.055)	(0.078)	(0.075)	(0.140)	(0.230)	(0.142)	(0.237)
Father's education	0.011	0.024 **	0.025 **	0.016 **	0.018	0.019	0.025 **	0.034 ***	0.018	0.025 **
	(0.007)	(0.012)	(0.012)	(0.007)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Mother's education	0.021 **	0.027 **	0.027 **	0.020 **	0.030 **	0.031 **	0.026 *	0.023 *	0.030 **	0.028 **
	(0.008)	(0.013)	(0.013)	(0.008)	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)	(0.012)
Father smoking (1=yes)	-0.747 ***	-0.887 **	-0.902 ***	-0.583 **	-0.636 *	-0.612 *	-0.629	0.447	-0.463	0.273
	(0.252)	(0.379)	(0.379)	(0.242)	(0.361)	(0.365)	(0.404)	(0.476)	(0.407)	(0.495)
Log of household income p.c.	0.091	0.231 **	0.193 *	0.153 ***	0.304 ***	0.255 ***	0.230 **	0.128	0.281 ***	0.215 **
	(0.060)	(0.097)	(0.100)	(0.059)	(0.095)	(0.095)	(0.092)	(0.095)	(0.092)	(0.093)
Log of household land assets	-0.029	0.334 ***	0.353 ***	-0.066 *	0.250 ***	0.264 ***	0.268 ***	0.185 **	0.216 **	0.155 *
	(0.037)	(0.086)	(0.085)	(0.037)	(0.091)	(0.089)	(0.087)	(0.087)	(0.090)	(0.090)
Log of tuition of junior high school	0.118	1.343 ***	1.308 ***	-0.056	0.876 ***	0.833 ***	1.439 ***	1.089 ***	0.899 ***	0.685 ***
	(0.087)	(0.222)	(0.221)	(0.082)	(0.226)	(0.224)	(0.228)	(0.225)	(0.229)	(0.232)
Log of tuition of senior high school	-0.052	0.085	0.110	0.111	0.258	0.272	-0.078	-0.317	0.174	-0.006
	(0.100)	(0.212)	(0.211)	(0.093)	(0.214)	(0.215)	(0.212)	(0.209)	(0.219)	(0.227)

	OLS regressions						IV regressions			
	Math[a]	Math[b]	Math[b]	Chinese[a]	Chinese[b]	Chinese[b]	Math	Math	Chinese	Chinese
Distance from junior high school (km)	-0.014	0.014	0.012	-0.029 ***	0.004	0.002	0.015	0.042 **	0.004	0.023
	(0.009)	(0.016)	(0.016)	(0.009)	(0.014)	(0.014)	(0.016)	(0.017)	(0.014)	(0.015)
Distance from senior high school (km)	0.001	0.005 *	0.006 *	0.004 *	0.011 ***	0.011 ***	0.003	0.003	0.009 ***	0.009 ***
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Leaking classrooms (%)	-0.310 ***	0.052	0.052	-0.128 *	0.286 **	0.277 **	-0.050	-0.003	0.224 **	0.244 **
	(0.083)	(0.125)	(0.125)	(0.077)	(0.108)	(0.108)	(0.127)	(0.124)	(0.108)	(0.106)
Having science lab (1=yes)	-0.077	-0.317 ***	-0.301 ***	-0.037	-0.307 ***	-0.299 ***	-0.273 ***	-0.214 **	-0.280 ***	-0.241 ***
	(0.057)	(0.094)	(0.093)	(0.059)	(0.096)	(0.094)	(0.094)	(0.091)	(0.095)	(0.091)
Having library (1=yes)	0.120	0.880 *	0.843 *	-0.050	0.519	0.499	0.863 *	1.574 ***	0.494	0.968 ***
	(0.131)	(0.485)	(0.486)	(0.132)	(0.359)	(0.357)	(0.475)	(0.405)	(0.368)	(0.340)
Teachers with 5+ years of experience (%)	-0.474 **	1.639 ***	1.671 ***	-0.739 ***	1.441 ***	1.455 ***	1.524 ***	1.291 ***	1.365 **	1.202 **
	(0.204)	(0.493)	(0.497)	(0.209)	(0.508)	(0.516)	(0.491)	(0.488)	(0.527)	(0.525)
Teachers with post secondary degrees (%)	0.973 *	-1.256	-1.291	0.387	-1.335	-1.444	-2.105	-6.573 ***	-1.823	-4.862 ***
	(0.594)	(1.341)	(1.351)	(0.614)	(1.180)	(1.183)	(1.418)	(1.556)	(1.313)	(1.569)
Teachers' salary per month (100 yuan)	-0.028	-0.111 ***	-0.105 ***	-0.079 ***	-0.041	-0.031	-0.092 ***	-0.040	-0.024	0.012
	(0.018)	(0.030)	(0.029)	(0.018)	(0.032)	(0.030)	(0.030)	(0.033)	(0.032)	(0.035)
Teachers' bonus per year (100 yuan)	0.025 **	0.070 ***	0.069 ***	0.078 ***	0.048 **	0.047 **	0.068 ***	0.060 ***	0.046 **	0.040 **
	(0.011)	(0.021)	(0.021)	(0.012)	(0.021)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Constant	-1.323	-9.718 ***	-9.524 ***	-1.707 **	-9.622 ***	-9.297 ***	-8.847 ***	-5.599 ***	-8.982 ***	-6.789 ***
	(0.837)	(1.658)	(1.664)	(0.823)	(1.597)	(1.586)	(1.669)	(1.656)	(1.638)	(1.728)
No. of Obs.	1521	646	646	1522	646	646	649	649	650	650
(Pseudo) R2	0.05	0.17	0.19	0.06	0.18	0.18	0.17	0.20	0.18	0.18

[a] Regressions based on full sample; [b] regressions based on sample for which the information about instrumental variables is available

* Significant at 10% level, ** significant at 5% level, *** significant at 1% level

Robust standard errors are reported in parentheses.

Data source: GSCF

Table 2.15: Youth smoking and years of schooling, aged 12-17 (CHNS data)

	Censored Ordered Probit			2-stage Censored Ordered Probit		
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Currently smoking (1=yes)	-1.152 **		-0.952 **	0.457		1.300
	(0.239)		(0.417)	(2.567)		(2.606)
Number of cigarettes smoked per day		-0.077 ***	-0.015		-0.006 *	-0.006 *
		(0.018)	(0.032)		(0.003)	-(0.003)
Age	0.059 **	0.054 *	0.061 **	0.031	0.085 **	0.078 *
	(0.029)	(0.029)	(0.029)	(0.040)	(0.043)	(0.048)
Sex (1=female)	-0.174 **	-0.152 **	-0.174 **	-0.134	-0.138	-0.141 *
	(0.077)	(0.076)	(0.077)	(0.086)	0.086	(0.086)
Father's education	0.028 ***	0.029 ***	0.029 ***	0.037 ***	0.035 ***	0.035 ***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Mother's education	0.020 ***	0.020 ***	0.020 ***	0.020 **	0.023 ***	0.023 ***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Father smoking (1=yes)	0.094	0.088	0.094	0.048	0.027	0.033
	(0.083)	(0.083)	(0.083)	(0.095)	(0.096)	(0.139)
Mother smoking (1=yes)	-0.288	-0.266	-0.279	-0.771 ***	-0.732 **	-0.798 ***
	(0.205)	(0.205)	(0.205)	(0.290)	(0.237)	(0.291)
Household income p.c. in 2000 (100 yuan)	0.013 ***	0.013 ***	0.013 ***	0.009 **	0.010 **	0.009 **
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)

	Censored Ordered Probit			2-stage Censored Ordered Probit		
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Average wage for babysitting (yuan)	0.015	0.012 **	0.012 **	0.008	0.008	0.008
	(0.047)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
Average wage for a construction worker (yuan)	-0.016	-0.006	-0.005	-0.006	-0.006	-0.005
	(0.019)	(0.006)	(0.007)	(0.007)	(0.008)	(0.009)
Average wage for a driver (100 yuan)	-0.003	0.022 *	0.022 *	0.024	0.019	0.019
	(0.003)	(0.013)	(0.013)	(0.015)	(0.015)	(0.699)
Distance from primary school (km)	0.037 **	0.023	0.016	0.039	0.047	0.043
	(0.016)	(0.047)	(0.047)	(0.054)	(0.054)	(0.054)
Distance from junior high school (km)	-0.007	-0.018	-0.016	-0.016	-0.007	-0.009
	(0.005)	(0.019)	(0.019)	(0.021)	(0.021)	(0.022)
Distance from senior high school (km)	-0.232 **	-0.003	-0.003	-0.002	-0.006 *	-0.006 *
	(0.100)	(0.003)	(0.003)	(0.002)	(0.004)	(0.003)
Rural (1=yes)	-0.233 **	-0.243 **	-0.235 **	-0.288	-0.273 **	-0.287 **
	(0.101)	(0.100)	(0.101)	(0.114)	(0.114)	(0.114)
No. of Obs.	2508	2510	2508	2014	2014	2014
Log likelihood	-940	-940	-939	-703	-702	-701

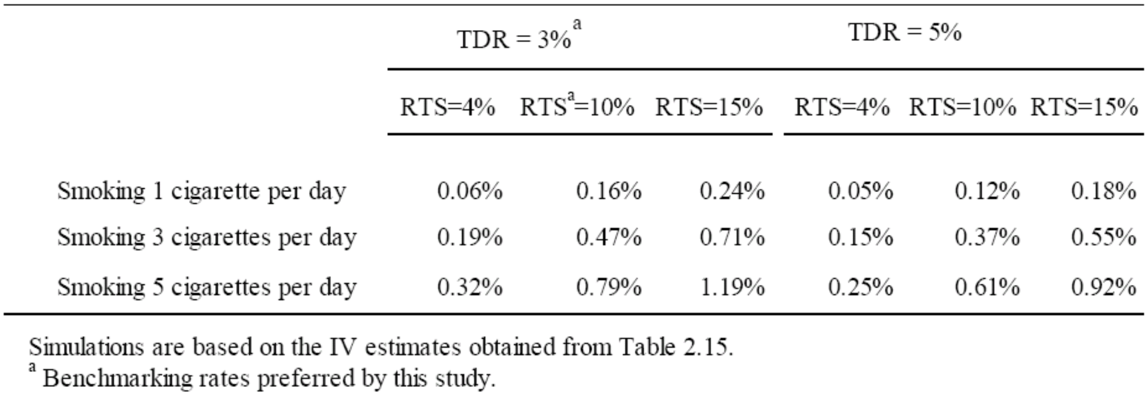
* Significant at 10% level, ** significant at 5% level, *** significant at 1% level

All the regressions control for the province and time fixed effects.

Robust standard errors are reported in parentheses.

Data source: CHNS

Table 2.16: Simulated reduction of lifetime income



	TDR = 3%[a]			TDR = 5%		
	RTS=4%	RTS[a]=10%	RTS=15%	RTS=4%	RTS=10%	RTS=15%
Smoking 1 cigarette per day	0.06%	0.16%	0.24%	0.05%	0.12%	0.18%
Smoking 3 cigarettes per day	0.19%	0.47%	0.71%	0.15%	0.37%	0.55%
Smoking 5 cigarettes per day	0.32%	0.79%	1.19%	0.25%	0.61%	0.92%

Simulations are based on the IV estimates obtained from Table 2.15.

[a] Benchmarking rates preferred by this study.

Figure 2.1: Map of China, sample provinces marked (CHNS data)

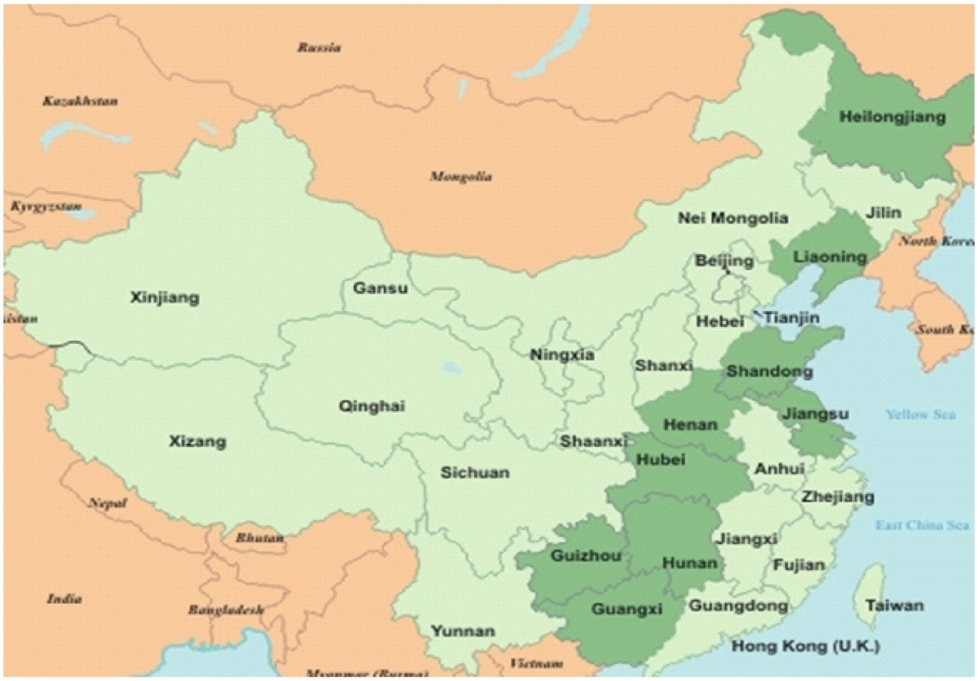
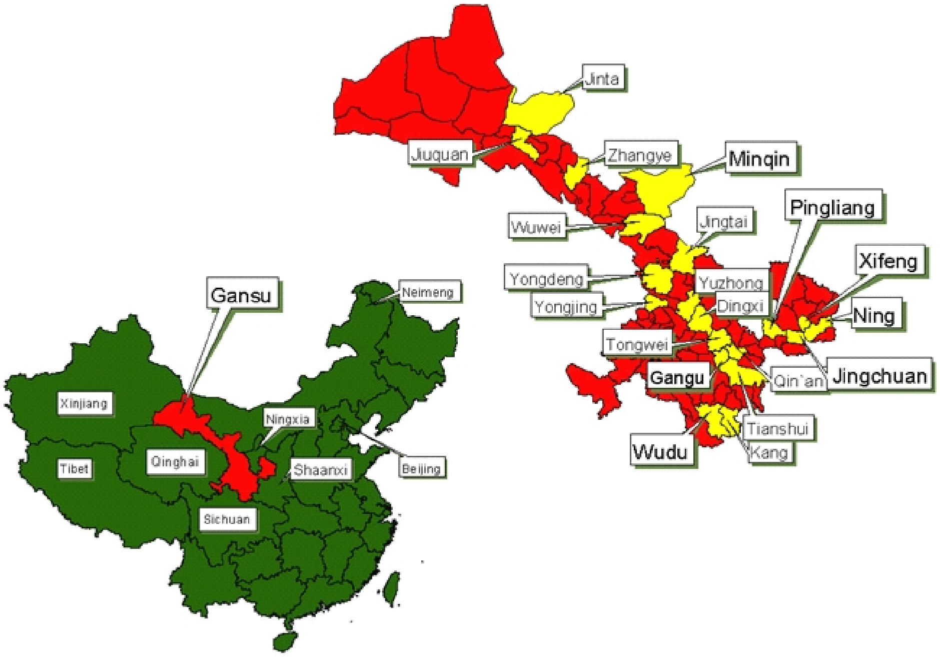


Figure 2.2: Map of Gansu province, GSCF sample counties marked



Chapter 3. Does Doctor Diagnosis of Hypertension Affect Food Choices? A Regression-Discontinuity Approach

3.1. Introduction

The World Health Organization (WHO) reported that 35 million people died of chronic diseases (e.g. cardiovascular diseases, cancers, and diabetes) in 2005, and that chronic diseases are the leading cause of death, accounting for 60% of global deaths. One of the major risk factors of chronic diseases is an unhealthy diet (WHO, 2005). Healthy eating can reduce existing chronic health problems and prevent future health risks, while eating unhealthy food can increase the chance of developing chronic diseases. Increased availability of public information on diet and health may persuade consumers to alter food choices in a direction that improves their health:

> "There is a widely held belief that concerns about diet-related consequences for health have altered the landscape of food-consumption patterns in many industrialized countries. Analysts have frequently attempted to validate this belief and to provide empirical evidence on the subject. However, the economics literature has shown that it is very difficult to quantify precisely the impacts."
> — *Chern and Rickertsen*, 2003.

To estimate the impacts of health concerns on food demand, previous studies focused on how consumers' food demand responds to the provision of public information on what constitutes a healthy diet. A commonly used measure of information is an index created based on the counts of medical reports or media articles on the health consequences of different diets (e.g. Brown and Schrader, 1990; Chern, Loehman and Yen, 1995; Kim and Chern, 1999; Kabiaa et al., 2001; Roosen et al., 2009). Another strand of the literature examines the effects of nutrition labels, food advertisements and social marketing on consumers' food choices (e.g. Martin et al, 1994; Alston et al., 1998 and 1999; Duffy, 1999;

Crutchfield et al., 2001). However, the wide variation of findings regarding the impacts of healthy-diet information on food demand suggests a need for more careful measurement of consumers' concerns about health and more careful analysis of their underlying decision process when choosing foods.

This study proposes an alternative approach to quantify the impacts of health concerns on food choices by investigating the effects on food demand of providing information on consumers' true heath status, as measured by hypertension. Although consumers might accumulate general health-risk information slowly over time, they are unlikely to make a long term and significant effort to improve diet habits unless they are faced with life-threatening health shocks[8]. Consumers' responses to public information on a healthy diet may depend on their perceptions of their own health status. Furthermore, information on health status could be of particular importance when analyzing the impacts of health concerns on food items that may have both positive and negative health consequences, depending on the amount consumed. For example, red meats are important providers of protein but could be detrimental to health if over-consumed.

Hypertension is one of the major risk factors for cardiovascular disease. Because hypertension is usually asymptomatic at moderate or even serious levels, many individuals with hypertension are unaware that they have this condition. According to a cross country study based on national survey data collected in the 1990's, the prevalence of hypertension for persons 35 to 64 years was 28% in the United States, 27% in Canada, and 40-55% in five European countries (Germany, Italy, Spain, Sweden, and the United Kingdom). About 20% of the hypertensive population in United States and Canada were unaware of their condition, compared to higher rates , from 30% to 50%, in the five European countries. Lack of awareness of the condition is even more serious in developing countries. According to national representative survey data from the China Health and Nutrition Surveys,

[8] For example, as found in a field experiment conducted in France, warning of poison in fish only slightly modified household fish consumption and the impact became insignificant after only three months. Roosen et al. (2009) attributed such lack of impact to consumers' weak memory of information provided.

three-quarters of the hypertensive population in China were unaware of their illness. If information were provided to consumers regarding their hypertension status, would they become more conscious about the health consequences of their diet and, therefore, change their food choices to healthier ones? To answer this question, this study investigates the gap between consumers' perceived hypertension and clinical hypertension, and the impact of bridging this gap on their food consumption.

This study uses unique panel data from China to analyze the impact of a doctor's diagnosis of hypertension on consumers' nutrient intake and food consumption. China's economy has experienced rapid growth over the past three decades, with GDP growing at an average annual rate of 8.4%. The accompanying increases in personal income and improvement in living conditions have pushed food demand in a direction that adversely affects health status (see Figure 3.1). During the same period, the prevalence of chronic diseases in China has rapidly increased. Chronic diseases now account for about 80% of total deaths in China (Wang et al., 2005). Indeed, the WHO projects that 560 billion U.S. dollars will be foregone from 2000-2015 due to chronic diseases in China (WHO, 2005), by far the highest loss among all of the countries examined in that study. On the other hand, growing health concerns may, to some extent, reverse the rising demand for red meat and the falling demand for grains and vegetables.

The rest of the chapter is organized as follows. Section 3.2 provides background information on food demand in China and on the data. Section 3.3 provides a brief literature review. Section 3.4 presents a simple model that describes how health information can interact with consumers' decisions regarding food choices. Section 3.5 proposes the method to identify the effects of diagnosis of hypertension on food choices. In Section 3.6, graphical evidence and results from local linear regressions are discussed. Lastly, Section 3.7 concludes.

3.2. Background and data

Both the quantity and structure of private food consumption in China have undergone considerable changes during the economic prosperity of the past two decades. Figure 3.1 shows trends in average per capita demand for seven staple foods in urban China from 1981 to 2004, as reported in China's Statistical Yearbooks. Except for fruit (due to the lack of data from 1980 to 1990), the demand for all the other foods is standardized by the levels of 1981. Consumption of poultry increased the fastest, followed by that of beef, which started to drop slightly after 1997. The demand for seafood and fruit also rose steadily, with a sharp jump in seafood consumption after 2000. Pork consumption in urban China was constant until 2000, at which time it experienced a sudden increase. A slight decreasing trend can be observed for vegetable consumption from 1981 to 1993. After that, it stayed constant and then started to rise slightly after 2001. Grain consumption dropped more than any other food group over the last two decades. In 2004, the average urban Chinese resident consumed only half of the amount consumed in 1981.

This study is based on approximately national representative data that are collected from the China Health and Nutrition Survey (CHNS), a comprehensive household panel survey that collected six rounds of data in China from 1989 to 2004. Actually, the survey continues to collect data since 2004. These new data can be used to complement this study when they become available in the future. Unusually detailed information was collected on household characteristics, individual health status and food consumption. Following a standard procedure, households' food intake was measured over three consecutive days by trained investigators. The three consecutive interview days were randomly selected from Monday to Sunday and were spread throughout a whole week. The investigators weighed food inventory in the morning and after the last meal for each day. In addition, household food purchases and transfers throughout the three days are also recorded. In total, there are more than 1500 types of food items recorded by the survey. At the end of each round of the survey, physical examinations were also conducted. However, the eligibility for physical examinations was defined slightly differently over the different years. Considering the inconsistency and availability, this study will mainly use the data from the three most recent rounds: 1997, 2000 and 2004.

Table 3.1 presents descriptive statistics for the 3-day per capita consumption of several major food categories from 1997 to 2004. Comparing the average levels across years, consumption of rice, wheat products and vegetables first decreased and then increased. By contrast, average consumption of animal oil first increased and then decreased. Lastly the consumption of pork, poultry and plant oil has been increasing. These trends suggest two possible effects on food consumption: an income effect and the effect of health concerns. As income increased over years, the demand for meat and oil generally increases, while that for staple foods and vegetables decreases. However, as consumers' concerns regarding health and health consequences of their diets grow, the demand for healthy foods, such as vegetables and wheat products, may display an increasing trend. Animal foods such as animal oil are well-known for their long-run detrimental effects on health and, therefore, may over time become less attractive to consumers. The trend in the demand for animal oil seems to be consistent with this hypothesis. Unfortunately, the trend in the consumption of beef and mutton is not observable due to the lack of data in 1997 and 2000.

Based on the 3-day food intake data collected by the CHNS and a Chinese food nutrition table compiled by Yang, et al.(2002), the Carolina Population Center calculated the daily intakes of four nutrients: energy, fat, carbohydrates and protein. Table 3.2 presents descriptive statistics for these nutrients. From 1997 to 2004, averages of energy and fat intake first increased and then decreased. These trends match those in the consumption of animal oil. During the same time period, the average carbohydrate intake has continuously dropped. This is mainly due to a decrease in the consumption of staple foods which are the major sources of carbohydrates. Daily intake of protein has been steadily increasing. Protein comes from both animal products (e.g. meat, fish, poultry, and eggs) and vegetables (e.g. beans, nuts, and whole grains). Animal protein and vegetable protein have the same effect on health. But different types of foods provide protein in combination with different levels of other nutrients.

Chronic diseases are usually referred to as “diseases of the wealthy”, which reflects the increasing incidence of these diseases as incomes grow and diet patterns change. One of the leading risk factors of chronic diseases is hypertension. The CHNS data collected two measures of hypertension: (a) diagnosis information of hypertension; (b) systolic and diastolic blood pressure levels from individual physical examination conducted by professionally trained investigators. The diagnosis information was not collected directly from doctors. Instead, it was from a household health survey. In the individual questionnaire, each adult aged 18 years old or plus was asked: "Has a doctor ever told you that you suffer from high blood pressure?" The self-reported diagnosis information is clearly different from self-reported health status commonly collected by health surveys. The former is based on a matter of fact, while the latter is very subjective and primarily depends on self-assessment. However, there may be some noise in the self-reported diagnosis data due to heterogeneity in memory, access to medical care, and so forth.

A unique feature of the CHNS is that it also conducted physical examinations for each individual after the health survey questions were asked. Professionally trained interviewers measured systolic and diastolic blood pressure at three separate times for each individual. Therefore, hypertension status is measured both by objective data and by self-reported recall of a clinical diagnosis. The survey personnel did inform people of the results of the physical examinations. However, the hypertensive people might not be informed of their illness explicitly because many people who have high blood pressure did not report so in the later rounds of the survey. As shown in Figure 3.2, the prevalence of hypertension, calculated from objective measures of blood pressure, among population aged 18-70 in China has been increasing steadily, from 14% in 1991 to 22% in 2004[9]. Since China is now experiencing population aging, the increase in the prevalence of hypertension may be due, or

[9] According to the National Institute of Health, hypertension is a high blood pressure with systolic pressure reading of 140 mmHg or higher and/or diastolic pressure reading of 90 mmHg or higher.

at least in part, to the change in population's age structure. Indeed, after adjusting for age, the increase in the prevalence of hypertension falls by one-third.

Table 3.3 provides a comparison between the prevalence of hypertension based on the self-reported diagnosis results and the objective measures of blood pressure for the three most recent rounds of the CHNS. As suggested by table 3.3, on average, approximately 20% of the sample were hypertensive from 1997 to 2004, more than three quarters of whom were unaware of their illness. However, people are becoming better informed over time, as indicated by the increase in the percentage of those who reported having been diagnosed with hypertension: 9% of total population were diagnosed with hypertension in 2004, compared to 4.4% in 1997. This change mainly reflects the fact the late middle age and elderly populations were better informed in more recent years (e.g. from 1991 to 2004, the proportion of the population aged over 65 with a diagnosis of hypertension who were aware of their status increased from about one-third to about one half).

The raw correlation between the self-reported diagnosis of hypertension and an objective measure of hypertension increased slightly from 0.31 in 1997 to 0.35 in 2004. Interestingly, the rate is not the same for different socioeconomic groups. For people in the lowest income quintile, the correlation is 0.25, whereas, it is 0.41 for people in the highest income quintile.

3.3. Literature review

Most of the literature that examines the impact of health information on food choices focuses on changes in consumers' food demand in response to the provision of public information on what constitutes a healthy diet. For example, Brown and Schrader (1990) created a health information index based on the counts of journal articles that found links between cholesterol and heart disease. Their findings suggest that health information, as measured by the health information index, reduced the per capita demand for eggs by 16% to 25% in United States from 1955 to 1987. Kim and Chern (1999) created a cholesterol information index using a modified weighting method, assuming that articles published in specific time periods can have carry over and decay effects. The study found evidence that health information on fat and cholesterol increased the consumption of fish oil and reduced the use of lard, tallow and palm oil in Japan. Roosen et al. (2009) conducted a field experiment in France to investigate the impact of providing households with a warning on the risks of methylmercury contamination in fish. They found only a weak decrease in the consumption of the contaminated fish. The explanation for the ineffectiveness of the warning is that consumers could not remember the fish types quoted in the warning. Some studies have also analyzed the effects of nutrition labels, food advertisements and social marketing on consumers' food choices (e.g. Alston et al., 1998 and 1999b; Martin et al., 1994; Crutchfield et al., 2001). Crutchfield et al. (2001) analyzed the impact of nutrition labels to estimate the economic benefit of new rules that require the provision of nutrition information for all the raw meat and poultry products. They show that providing these nutrition labels decreases the intake of fat and cholesterol and, therefore, reduces the risks of developing future cases of stroke, cancer and heart disease.

The pioneer of China's demand analysis is Houthakker (1957), who studied the demand for four categories of purchases, namely food, clothing, housing and miscellaneous, using 1927 data for Beijing and 1929 data for Shanghai. The first recent study was conducted by Chow (1984) and followed by a World Bank report on demand patterns in rural China. Since the late 1980's, many more demand studies have been carried out. These studies vary by both the type of data and the methodologies used.

Aggregate time series data (e.g. Kueh, 1988; Lewis and Andrews, 1989; Peterson et al., 1991; Fan et al., 1994) were commonly used before provincial panel data (e.g. Wang and Chern, 1992; Fan et al., 1995; Gao et al., 1996b) and household data (e.g. Halbrendt et al., 1994; Gao et al., 1996a; Huang and Rozelle, 1998; Fang and Beghin, 2002; Yen et al., 2004) became available and so gained more popularity for demand analysis. Estimates obtained using either aggregate time series data or provincial cross-sectional data could suffer from lack of precision due to small sample. For example, Lewis and Andrews (1989) analyzed data only for 1981-1985 and Fan et al. (1994) had only 9 years in their sample. Another problem with many of these studies is that they do not account for heterogeneity due to household characteristics. Halbrendt et al. (1994) and Gao et al. (1996a) started to use household survey data collected in 1994 and 1990 to estimate rural demand in two Chinese provinces, Guangdong and Jiangsu, respectively. Their studies included household characteristics such as family size, household education, and household type.

Huang and Rozelle (1998) took into account the heterogeneity on the supply side, as measured by market development, in their food demand analysis, using 1993 household survey data in Hebei province. More recently, Fang and Beghin (2002) estimated a small 3-equation system for fats and oil demand using national representative household survey data. Yen et al. (2004) also used a nationally representative dataset to estimate a translog demand system that account for zero consumption. Moreover, except for Gao et al. (1996a), which followed the Cox and Wohlgenant approach, all of these studies ignored the possibility of endogeneity bias due to the use of "unit values", the ratio of household food expenditure over units purchased, as individual prices. As pointed out by Deaton (1988), "unit value" expenditure could be correlated with the quality of food purchased by the household. As the quality choices of the household may be correlated with unobserved household characteristics, estimates of price elasticities will be biased.

This study contributes to the literature in three ways. First, studies on the impact of health information on food choices in China are extremely rare. Most of studies on this topic are conducted in developed countries. However, 80% of deaths due to chronic diseases occur in developing countries. In order to address the global epidemic of chronic diseases, there is an urgent need to understand how to persuade consumers in developing countries to alter their food choices in a direction that improves their health. Second, this study offers a new perspective on how to identify the effect of health information. Previous studies focus on the effect of public information on healthy diet. However, the findings from these studies can be misleading if consumers' perceptions about their own health are not taken into account. In many cases, public information takes effect mainly through interaction with individuals' beliefs about their health conditions. Third, this study uses a non-parametric method to identify the effects of information about health status on food consumption, making fewer assumptions about the functional form of the demand for food. Therefore, the applicability of the results of this study is wider.

3.4. Theoretical model

Based on Gary Becker's household production model (Becker, 1965), Michael Grossman's (1972) seminal paper suggests that rational consumers make investments in health by allocating time to exercise and to purchase medical services. In return, better health generates more healthy time and enhances the consumption of other commodities. Therefore, consumers make optimal health investment choices, which in turn determine health in the next period. The model assumes that consumers are perfectly informed. Food consumption is often included as an input into health production, so Grossman's model has implications for food demand analysis (see a review of applications of Grossman's model in food demand analysis in Chern and Rickertsen, 2003).

Given a fixed budget, there is always a trade-off between current well being (e.g. current consumption) and investment (e.g. better health or higher wealth in the future). If the money allocated to a given time period is spent on consumption as one's taste prefers, without considering its potential impacts on future health, even though current utility will be maximized, the person's lifetime utility may not be maximized. Consider a consumer who, in each time period t, chooses optimal levels of non-food goods $\mathbf{c}_t$ (bold letters always denote vectors in this chapter) and food $\mathbf{x}_t$ to maximize the sum of current utility and expected future utilities discounted by a time constant rate of β. Assuming no credit constraints, this utility maximization is subject to a lifetime budget constraint. Following standard assumptions, assume that the utility function is concave, twice differentiable, time separable and time independent. A higher stock of health in period t, h_t, indicates a "better" health and $h_{\min}$ is the minimum health required for survival.

The consumer also faces a constraint on her health stock in each time period. Analogous to capital stock, health stock is determined by investment in, and depreciation of, health. Any increase in a person's health stock is due to investment in health, through a health production function, $I(\mathbf{x}_t)$ which characterizes the relationship between current food consumption and health in the next time period. That is, $h_{t+1} = I(\mathbf{x}_t) - (1-\delta)h_t$, where δ is the depreciation rate for the stock of health. The health production function is assumed to be concave and differentiable. A person chooses the time path of investments in health to maximize lifetime utility[10].

More formally, following Grossman (1972), and assuming all future prices and incomes are known, a consumer solves the following optimization problem:

$$\begin{aligned} & Max \sum_{t=0}^{\infty} \beta^t U(\mathbf{c}_t, \mathbf{x}_t, h_t) \\ s.t.\ h_{t+1} &= I(\mathbf{x}_t) - (1-\delta)h_t \quad t = 0, 1, ... \\ \sum_{t=0}^{\infty} \frac{\mathbf{c}_t + \mathbf{p}_t \mathbf{x}_t}{(1+r)^t} &= \sum_{t=0}^{\infty} \frac{m_t}{(1+r)^t} + A_0 \\ h_0, A_0 \text{ given; } \mathbf{c}_t, \mathbf{x}_t &> 0;\ h_t \geq h_{\min} > 0;\ t = 0, 1, ... \end{aligned}$$

where δ is the depreciation rate of the health stock, r is the interest rate, m is income at time t, A_0 is the initial wealth endowment, and $\mathbf{p}_t$ is the vector of prices for the goods in $\mathbf{x}_t$. Assuming non-food good as a numeraire, the price of $\mathbf{c}_t$ in each time period is 1. The first order conditions that solve this optimization problem are:

$$\beta^t U'_{ct} = \frac{\lambda}{(1+r)^t} \tag{1}$$

$$\beta^t U'_{xt} + \tilde{\mu}_t I'_{xt} = \frac{\mathbf{p}_t \lambda}{(1+r)^t} \tag{2}$$

[10] Actually, Ried (1998) pointed out that models assuming fixed and free terminal time give the same set of results.

where λ and μ_t denote the Lagrangian multipliers for the budget constraint and the health stock constraint, respectively. As long as better health increases utility, μ_t is positive. Note that $\tilde{\mu}_t = \mu_t + (1-\delta)\mu_{t+1} + (1-\delta)^2\mu_{t+2} + ...$, which is the lifetime marginal utility of health at time period t. Equation (2) equates the marginal cost of $\mathbf{x}_t$ and its marginal benefits, through the marginal utility of $\mathbf{x}_t$ and the lifetime marginal utility of better health generated by the health production function I'_{xt}. Dividing (1) by (2) yields:

$$\frac{U'_{ct}}{U'_{xt} + \beta_t^{-t}\tilde{\mu}_t I'_{xt}} = \frac{1}{\mathbf{p}_t} \tag{3}$$

which implies the condition that decides the optimal demand for $\mathbf{c}_t^*$ and $\mathbf{x}_t^*$. Since β^{-t} and $\tilde{\mu}_t$ are positive, I'_{xt}, the health consequences of $\mathbf{x}_t$ play an important role in determining the marginal rate of substitution between $\mathbf{c}_t$ and $\mathbf{x}_t$. If $I'_{xt} > 0$, that is, $\mathbf{x}$ is nutritious food and increases health stock, $\frac{U'_{ct}}{U'_{xt}+\beta^{-t}\tilde{\mu}_t I'_{xt}} < \frac{U'_{ct}}{U'_{xt}}$, the demand for $\mathbf{x}$ will be larger than it could have been when health factor is not considered. On the other hand, for health detrimental food whose $I'_{xt} < 0$, there should be less demand for it.

3.5. Identification and estimation

This study analyzes the effect of informing consumers of their true health status, more specifically, informing them that they have hypertension, on food demand. Diagnoses of hypertension are made if one's blood pressure is above a certain cutoff. For example, a person will be diagnosed as hypertensive if either her systolic blood pressure (SBP) is above 140 mmHg or her diastolic blood pressure (DBP) is above 90 mmHg. As a result, the diagnostic rule allows one to analyze the impact of a diagnosis of hypertension using a regression discontinuity (RD) design, which is a very powerful quasi-experimental design. The RD design was first introduced by Thistlethwaite and Campbell (1960) and gained popularity in empirical research in economics in the late 1990s. Some early applications of the RD approach in economics include Angrist and Lavy (1999), who estimated the effect of class size on student test scores, Van der Klaauw (2002), who investigated the effect of financial aid offers on students' decisions on accepting offers of admission to colleges, and Black (1999), who analyzed parents' willingness to pay for school quality. See Lee and Lemieux (2009) for a summary of papers that have used RD approach.

Consider a random sample of individuals that includes data on the outcome measure, Y_i, and the treatment indicator T_i. The subscript i indicates the ith individual. T_i equals one if an individual receives the treatment and zero otherwise. The common econometric specification to evaluate the treatment effect is

$$Y_i = \alpha + \beta T_i + u_i \tag{4}$$

whereβ measures the treatment effect and u_i is the variation in Y_i that cannot be explained by T_i. If the assignment of the treatment is random, then β can be estimated by OLS. However, if the treatment is not randomly assigned, then in general $E[u|T] \neq 0$ and the OLS estimate of β will be biased. In this study, the treatment of interest is diagnosis of hypertension. Whether an individual is diagnosed as having hypertension is affected by many unobservable factors that may also affect the outcome measure Y_i. For example, a person who is more concerned about his or her health may have a healthier diet and also may use preventive care more often and, therefore, is less likely to be diagnosed as having hypertension. In this case, the OLS estimate of β is biased and so it does not measure the causal relationship between the treatment (diagnosis of hypertension) and the outcome measures of interest (food consumption).

The RD design makes use of additional information to help to identify the treatment effect. In an RD design, the treatment is assigned based (at least partly) on the value of an underlying continuous and observable variable, z_i, relative to a cutoff, z_0, such that the probability of receiving treatment is discontinuous in z at z_0, i.e.

Assumption (RD): $T^+ \neq T^-$ where $T^+ = \lim_{z \to z_0^+} E[T_i|z_i = z]$, $T^- = \lim_{z \to z_0^-} E[T_i|z_i = z]$.

There are two kinds of RD design, referred to as "sharp" RD design and "fuzzy" RD design, respectively[11]. With a sharp design, the assignment of treatment depends on z deterministically, i.e. $T_i = T(z_i) = 1\{z_i \geq z_0\}$. With a fuzzy RD design, the probability of receiving treatment variable is randomly given $z_{i,}$ i.e. $\Pr(T_i = 1|z_i = z] = E[T_i|z_i = z]$, but it is discontinuous at z_0. Generally, the sharp design is regarded as a special case of the fuzzy design. Note that the probability of receiving treatment in the fuzzy design can also be determined by other factors that may have an impact on the outcome Y. This is not the case for the design of standard randomized experiment.

Diagnosis of hypertension is usually based on the value of one's blood pressure measure, the assignment variable in this case, relative to the standard cutoff. That is, a person will be diagnosed as hypertensive if her SBP is above the cutoff of 140 mmHg or her DBP is above 90 mmHg. If the probability of being diagnosed is discontinuous at the cutoff, a regression discontinuity approach can still be used to identify the effect of receiving a diagnosis of hypertension even if the probability of being diagnosed is affected by other factors. This study will use a fuzzy RD design, and the discussion below focuses on the identification using a fuzzy design.

To see how a fuzzy RD design can identify and estimate the treatment effect, consider the sample of individuals within a small interval close to the cutoff point. Since these individuals have essentially the same value of z, their characteristics are likely to be the same. The only difference among them is that some fall slightly to the left, and some fall slightly to the right, of the cutoff point due to very small differences in observed and unobserved variables that determine z. The average outcomes in the absence of treatment for this sample should be similar. Or to put it a different way, the outcomes of all of these individuals, if they receive the treatment, are expected to be similar, and all of these individuals, if they do not receive the treatment, are also expected to be similar. Thus, at the cutoff point, this is very similar to a randomized experiment. More formally, to guarantee the treatment effect is identifiable, the following assumption needs to hold:

[11] See Trochim(1984)

Assumption (A1): $E[u_i|z_i = z]$ is continuous in z at z_0.

Following Hahn et al.(2001), if both assumption (RD) and assumption (A1) are satisfied, the treatment effect using fuzzy design is:

$$\beta = \frac{Y^+ - Y^-}{T^+ - T^-} \tag{5}$$

where $Y^+ = \lim_{z \to z_0^+} E[Y_i|z_i = z]$, $Y^- = \lim_{z \to z_0^-} E[T_i|z_i = z]$. For the case when the treatment effect is variable, that is, the treatment effect is heterogeneous, one more assumption is needed,

Assumption (A2): the average treatment effect $E[\beta_i|z_i = z]$ is continuous in z at z_0.

Suppose assumptions (RD), (A1) and (A2) hold, then the treatment effect at the margin can be estimated as follows:

$$E[\beta_i|z_i = z_0] = \frac{Y^+ - Y^-}{T^+ - T^-}. \tag{6}$$

Note that the subscript "i" of β implies that the treatment effect is variable. If the treatment effect varies with z in a deterministic way, that is, $Y_i = \alpha + \beta(z_i)T_i + u_i$, Eq. (6) will identify the local treatment effect $\beta(z_0)$ at z_0. Note that the denominator will be less than one, because the assignment of treatment in a fuzzy design can also determined by other factors, yet greater than zero as long as the discontinuity in the propensity score, $Pr[T_i = 1|z = z_i]$, exists.

Therefore, with consistent estimates of one-sided limits $\hat{Y}^+$, $\hat{Y}^-$, $\hat{T}^+$, $\hat{T}^-$, the treatment effect can be identified by equations (5) or (6). There are many ways to estimate these one-sided limits in the literature, including both nonparametric methods and (semi)parametric ones. First consider a one-sided kernel estimation. In a special case where the kernel regression is based on a uniform kernel, the estimates of the limits are equivalent to:

$$\begin{aligned}\hat{Y}^{+} &= \frac{\sum_{i\in\Omega} Y_i w_i}{\sum_{i\in\Omega} w_i}, \hat{Y}^{-} = \frac{\sum_{i\in\Omega} Y_i(1-w_i)}{\sum_{i\in\Omega}(1-w_i)}, \\ \hat{T}^{+} &= \frac{\sum_{i\in\Omega} T_i w_i}{\sum_{i\in\Omega} w_i}, \hat{T}^{-} = \frac{\sum_{i\in\Omega} T_i(1-w_i)}{\sum_{i\in\Omega}(1-w_i)}.\end{aligned}$$

where Ω stands for the subsample where $z_0 - h < z_i < z_0 + h$, w_i equals one if $z_0 < z_i < z_0 + h$ and h denotes the bandwidth. However, since kernel estimates have poor properties at boundary points where the treatment effect is evaluated in a RD design, Hahn et al.(2001) proposed to estimate the one sided limits by a local linear regression (LLR). The LLR estimator for Y^+, Y^-, T^+, T^- are given by $\hat{a}_{Yr}$, $\hat{a}_{Yl}$, $\hat{a}_{Tr}$, $\hat{a}_{Tr}$ in the following equations:

$$(\hat{a}_{Yr}, \hat{b}_{Yr}) \equiv \arg\min_{a,b} \sum_{i:z_i \geq z_0} [Y_i - a - b(z_i - z_0)^2]\lambda_i \tag{7a}$$

$$(\hat{a}_{Yl}, \hat{b}_{Yl}) \equiv \arg\min_{a,b} \sum_{i:z_i < z_0} [Y_i - a - b(z_i - z_0)^2]\lambda_i \tag{7b}$$

$$(\hat{a}_{Tr}, \hat{b}_{Tr}) \equiv \arg\min_{a,b} \sum_{i:z_i \geq z_0} [T_i - a - b(z_i - z_0)^2]\lambda_i \tag{7c}$$

$$(\hat{a}_{Tl}, \hat{b}_{Tl}) \equiv \arg\min_{a,b} \sum_{i:z_i < z_0} [T_i - a - b(z_i - z_0)^2]\lambda_i \tag{7d}$$

where $\lambda_i = K(\frac{z_i - z_0}{h})$ is a kernel function. A triangular kernel is commonly used, although Lee and Lemieux (2009) pointed out that the choice of kernel function "typically has little impact in practice". Following the literature, a triangular kernel is used in this study.

Since the RD design is mainly based on comparing the average outcomes for the sample that are within a small interval close to the cutoff point, the choice of bandwidth of the interval can make a difference in the estimates. If the window of the observations is too wide, the estimates may be biased and fail to account for the treatment effect. That is, the comparison on both sides of the cutoff points is not reliable because it is not comparing those just to the left and the right of the cutoff points. However, if it is too narrow, the estimates may not be precise because less data are used. Imbens and Kalyanaraman (2009) recently proposed a method to choose the optimal bandwidth for the regression discontinuity estimator. The optimal bandwidth minimizes an approximation of mean squared error of the RD estimates, $E[(\hat{\beta}-\beta)^2]$. Empirically, it can be obtained by estimating:

$$\hat{h}_{opt} = C_k \left(\frac{2 \cdot \hat{\sigma}^2(z_0)/\hat{f}(z_0)}{(\hat{m}_+^{(2)}(z_0) - \hat{m}_-^{(2)}(z_0))^2 + (\hat{r}_+ + \hat{r}_-)} \right)^{1/5} \cdot N^{-1/5}.$$

where $\hat{f}(z_0)$ is the estimator of the density function of the assignment variable at the cutoff point, $\hat{\sigma}^2$ is an estimator for the conditional variance of Y given the forcing, evaluated at the threshold, $\hat{m}_+^{(2)}(z_0)$ and $\hat{m}_+^{(2)}(z_0)$ are estimators for the second derivatives of the regression function from the left and the right, as a function of the forcing variable, evaluated at the threshold. The remaining components, $\hat{r}_+$ and $\hat{r}_-$ are regularization parameters that are used to avoid instabilities associated with low values of the difference in the second derivatives from the left and the right. The multiplicative constant C_k is a function of the kernel. Refer to Appendix C for a detailed discussion on how to compute the standard errors of the treatment effect β.

3.6. Results

3.6.1 Graphical evidence

In order to make sure that the impact of a diagnosis of hypertension can be estimated by a regression discontinuity approach, the discontinuity in the treatment, in this context the self-reported diagnosis of hypertension, is first examined at the cutoff values of blood pressure. The thresholds commonly used for defining hypertension are 140 for SBP and 90 for DBP. Although not everyone with blood pressure above the cutoffs is diagnosed with hypertension, the probability of being diagnosed, referred to as the propensity score, is likely to be much higher among those with blood pressure above the thresholds. As long as a discontinuity in the propensity score exists, a fuzzy regression discontinuity approach can be used to identify the effect of a doctor's diagnosis of hypertension.

Figure 3.4 shows the percentages of individuals diagnosed as having hypertension against objective blood pressure for 25,777 adults aged 18 years or older, from 1997 to 2004. The graph on the left shows the rates for population groups within intervals of every 5 units of SBP, and the one on the right is plotted against DBP. The graphs suggest an obvious jump in the rates of the diagnosis of hypertension at the cutoff points. Each graph in Figure 3.5 presents nonparametric prediction from a local polynomial smoother with different degrees and bandwidths. The first two are obtained using bandwidth of 1 and the last two are based on bandwidth of 2.5. These graphs also show a jump in the probability of receiving treatment at the cutoff points.

Probit regressions are also conducted to explore the discontinuity at the cutoff, using the whole sample. The dependent variable is equal to one if diagnosed with hypertension and zero if otherwise. Table 3.4 reports 4 specifications. The first regression includes a dummy variable indicating whether one's blood pressure is above the cutoff or not, controlling for continuous blood pressure measure, SBP and DBP, time and province fixed effects, as well as important variables such as age, gender, education and household income. The second regression adds the squared terms of SBP and DBP. The third regression drops demographic variables and controls for household fixed effect. The last regression is essentially the same as the second one, except that it controls for household fixed effect. In all specifications, whether above the cutoff or not is a significant indicator of the probability of receiving treatment, significant at the 1% level.

To compare the probability of being diagnosed between population within a small interval right below and above the cutoff, a probit regression is also conducted for the sample right and left to the cutoff separately. Figure 3.6 shows the predicted propensity from these regressions. A jump in the predicted probability is clearly visible at the cutoff. In order to test how robust the discontinuity is, a few different bandwidths are tried: 30, 20, 10, and 5. The discontinuity persists when different bandwidths are used.

The visible discontinuity in the propensity scores suggests that a fuzzy design should be feasible. The next step is to see if there is any discontinuity in the average outcomes at the cutoff points. According to RD design, since the sample within the small interval around cutoffs is essentially the same, the difference in average outcomes between the sample immediately below and above the cutoffs can be attributed to the treatment effect. In a fuzzy design, even if the diagnosis may be endogenous, that is, the treatment is correlated with some other characteristics that can also affect the outcomes, the treatment effect can still be identified as long as the propensity scores are significantly different at the cutoffs. Figures 3.7-3.17 present the scatter plots of the intakes of four major nutrients and the consumption of several food categories from 1997 to 2004. The vertical lines indicate the cutoffs, 140 mmHg for SBP and 90 mmHg for DBP, respectively.

Figure 3.7 presents the scatter plots only for observations within a narrow "window" close to the cutoff points. The bandwidth used is 20 for SBP and 10 for DBP, which are close to the optimal bandwidths estimated by the method discussed in Section 3.5. The horizontal lines denote the average energy intake for the samples left and right to the cutoff points. The energy intake of individuals with blood pressure right below the cutoffs, on average, is higher than that of those right above the cutoffs. The gap is smaller in the plot for DBP. Note that, although the discontinuity may appear visibly small, the treatment effect can be much larger. Recall that the treatment effect in a fuzzy RD design is obtained by dividing the difference in the outcomes by the difference in the probability of receiving treatment at the cutoff. Unlike a sharp design, the denominator in a fuzzy design is always smaller than one, which means that the treatment effect will be larger than what appears in a graphic presentation.

Similar patterns are observed for the intakes of fat, carbohydrate and protein. Graphically speaking, the average intakes of these nutrients are generally slightly higher for those whose blood pressure levels are below the cutoff points. As for the consumption of foods, the figures show that, on average, less rice, pork and animal oil is consumed by the sample to the right of the cutoffs. But the consumption of wheat products and plant oil is slightly higher for this sample. The difference in the consumption of vegetables is not clear. Again, the differences are generally smaller for the DBP cutoff point.

3.6.2 Estimation results

As suggested by the graphs discussed above, the probability of being diagnosed as having hypertension is significantly higher when one's blood pressure is above the cutoff. At the same time, a lower average level of nutrient intake is observed for the sample whose blood pressure is right above the cutoffs. Therefore, a fuzzy RD design can be used to identify the causal impact of a diagnosis of hypertension on food consumption and nutrient intake. The local linear regression (LLR) method discussed in Section 3.5 is used to estimate the effect of being diagnosed with hypertension on nutrient intake and food consumption. In brief, the method first estimates the optimal bandwidth used to determine the interval. Then the sample within this interval will be used in the LLR in the next step to estimate the right and left limits of outcomes and the propensity scores at the cutoff points. Lastly, the standard errors are calculated for inference of statistical significance.

Tables 3.5-3.8 summarize the results from LLR for the intakes of four nutrients. The intake data are all standardized by their sample means and standard deviations. Although DBP is an important indicator of hypertension, physicians usually pay more attention to SBP (Rutan, McDonald, and Kuller, 1989; Kannel, 2000). In addition, according to the probit estimates of the effects of SBP and DBP on the probability of being diagnosed with hypertension, SBP is a stronger predictor of the diagnosis of hypertension than DBP. Thus, for simplicity, this study focuses on SBP, using it as the forcing variable in the LLR estimation. But this will cause a problem; that is, some individuals in the sample with SBP lower than 140 mmHg may be diagnosed with hypertension if their DBP is above 90 mmHg. The treatment effect will be underestimated when comparing these observations with those with SBP above the cutoff. To correct this problem, the treatment effect is estimated by comparing only the population just below on both to the population just above on both. In this way, individuals with SBP above the cutoff point

In each table, the top panel reports the estimates of optimal bandwidths, treatment effects, standard errors, and corresponding z-statistics. All dependent variables are standardized by their sample means and sample standard deviations. SBP is also standardized by its cutoff and sample standard deviation. Thus, the estimates of optimal bandwidths are expressed in terms of the standard deviations from the cutoff point. The middle panel presents the intermediate estimates of the four limits used to calculate the treatment effect. The bottom panel provides some descriptive statistics, including the numbers of observations and the means of outcomes, separately for the samples below and above the cutoff point.

According to the LLR estimates in Table 3.5, an individual will reduce her daily fat intake by approximately 50 grams if she is diagnosed with hypertension. The reduction accounts for almost half of the average daily intake of fat among hypertensive people. The effect is very strong, and significant at the 5% level. Considering a possible change in the effect of diagnosis over time, the estimation is also done for different years separately. The negative effect on fat intake is found for the data from all three rounds, yet it is strongest for the 2004 data. The magnitude of the estimate of the effect found for 2004 data is also greater than that estimated based on three years of data. This suggests that consumers in China have become more health conscious and responsive to the diagnosis of hypertension in recent years.

The point estimates of the propensity scores show an increasing trend over time. The estimate of the left limit has increased from 0.05 in 1997 to 0.13 in 2004, and that of the right limit has increased from 0.13 to 0.26 during the same time period. The gap between the left and right limits has also grown a little wider, from 0.08 in 1997 to 0.13 in 2004. These trends suggest that, in general, hypertensive individuals are more likely to be informed of their conditions by doctors in recent years. On average, the probability of being diagnosed with hypertension for those whose SBP is above the cutoff is approximately 10% higher than that for those with an SBP lower than the cutoff.

The LLR estimates in Tables 3.6 and 3.7 show that the diagnosis of hypertension may reduce the intakes of energy and protein. However, although their signs and magnitudes are relatively consistent in different specifications, these estimates are not statistically significant. The estimates for carbohydrate intake in Table 3.8 are generally positive, but again statistically insignificant.

The lack of effect on these nutrients may be reasonable. In China, doctors' advice on diet for a patient with hypertension usually includes: (a) consuming less fats and red meats; (b) reducing the consumption of salt; (c) cutting back calorie intake if overweight or obese; (d) maintaining a moderate amount of protein intake; (e) replacing the consumption of simple sugars by complex sugars (major sources of complex sugars are grains, vegetables, and fruits). The advice on the consumption of fats, red meats and salt is pretty clear, while the advice on intakes of energy and protein is relatively vague. According to the World Health Organization (2009), the prevalence of overweight or obesity is still very low in China, below 5% and 15% respectively, so consumers may be less responsive to the advice on energy intake. The advice that recommends maintaining a moderate amount of protein intake is also vague and hard for patients to follow. This may explain why consumers are less responsive to such advice. Carbohydrate intake should increase if one follows doctors' advice. This is because grain products are rich sources of both complex sugars and carbohydrates. Since complex sugars are recommended to replace simple sugars (e.g. candy and soft drink), increasing consumption of grains will result in an increase in carbohydrate intake. In fact, the signs of the LLR estimates of the effects of diagnosis of hypertension on the intakes of energy, protein and carbohydrate are quite consistent with Chinese doctors' advice on diet for hypertensive individuals.

In addition to nutrient intake, the effect of a diagnosis of hypertension on the consumption of several major food categories is further examined. Tables 3.9-3.16 report the LLR estimates for the consumption of pork, animal oil, wheat products, plant oil, beef, mutton, poultry, and vegetables. In summary, a significant effect of being better informed of hypertension status is found for the consumption of pork, animal oil, and wheat products.

The results suggest that the diagnosis of hypertension decreases the consumption of pork by approximately 100 grams per day. When using data from different years separately, the effect is found to be greatest for the 2004 data.

As shown in Table 3.10, the consumption of animal oil also decreases significantly after a diagnosis of hypertension. The detrimental health consequences of animal oil consumption are pretty well known, and the increased availability of plant oil can serve as a substitute. Consumers respond to the diagnosis of hypertension by reducing their consumption of animal oil by approximately 30 grams. On the other hand, the consumption of wheat products is found to have increased after a diagnosis of hypertension, but the effect is significant only at the 10% level, and only for the 2004 data. The LLR estimates show little impact of a diagnosis of hypertension on the demand for beef, mutton, poultry and vegetables. Actually, the consumption levels of beef and mutton are extremely low in China. According to the CHNS data, only 5% of the sample reported consuming beef or mutton within the three interview days. The diagnosis of hypertension has a marginally significant positive effect on the consumption of plant oil, but only for the 1997 data. This is consistent with a decline in the consumption of animal oil.

As discussed in Lee and Lemieux (2009), a good way to test the robustness of the LLR estimates is to try different bandwidths, in comparison with the optimal bandwidth. The signs of the LLR estimates are generally very robust to the choices of bandwidth. However, the magnitude of the estimates of treatment effect vary slightly at different bandwidths. According to the results using different bandwidths, increasing the bandwidth by 0.1 standard deviation from the optimal bandwidth results in a 5% change in the magnitude of the estimated treatment effect.

Lastly, when comparing the difference in probability of receiving treatment and in the outcomes between the sample right below and above the cutoff, one concern is that some patients may be having their blood pressure under control by taking anti-hypertension drugs. In that sense, the treatment will affect the forcing variable and cause a downward bias in the estimates of the treatment effects. However, little difference is found in blood pressure between the hypertensive people who were taking anti-hypertension drugs and those who were not. The estimate of the effect of anti-hypertension drugs on SBP is generally statistically insignificant, even controlling for basic personal and household background, such as age, gender, education and household income. Another way to check is to take advantage of the longitudinal nature of the dataset. The SBP of people who were taking anti-hypertension drugs is replaced by their SBP measures from previous rounds before they reported to have been diagnosed of hypertension. The results actually do not differ much.

3.7. Conclusions

Chronic diseases account for the majority of global deaths and are becoming a major threat to people in developing countries who have less access to information, health facilities and education. Increased information on healthy diets from the media and from health organizations may make consumers more aware of the risks associated with an unhealthy diet. However, if consumers have imperfect information on their current health status, their dietary choices may not be optimal and they may not give sufficient attention to information on healthy diets. Indeed, it is well documented that a large proportion of the hypertensive population are unaware of their condition.

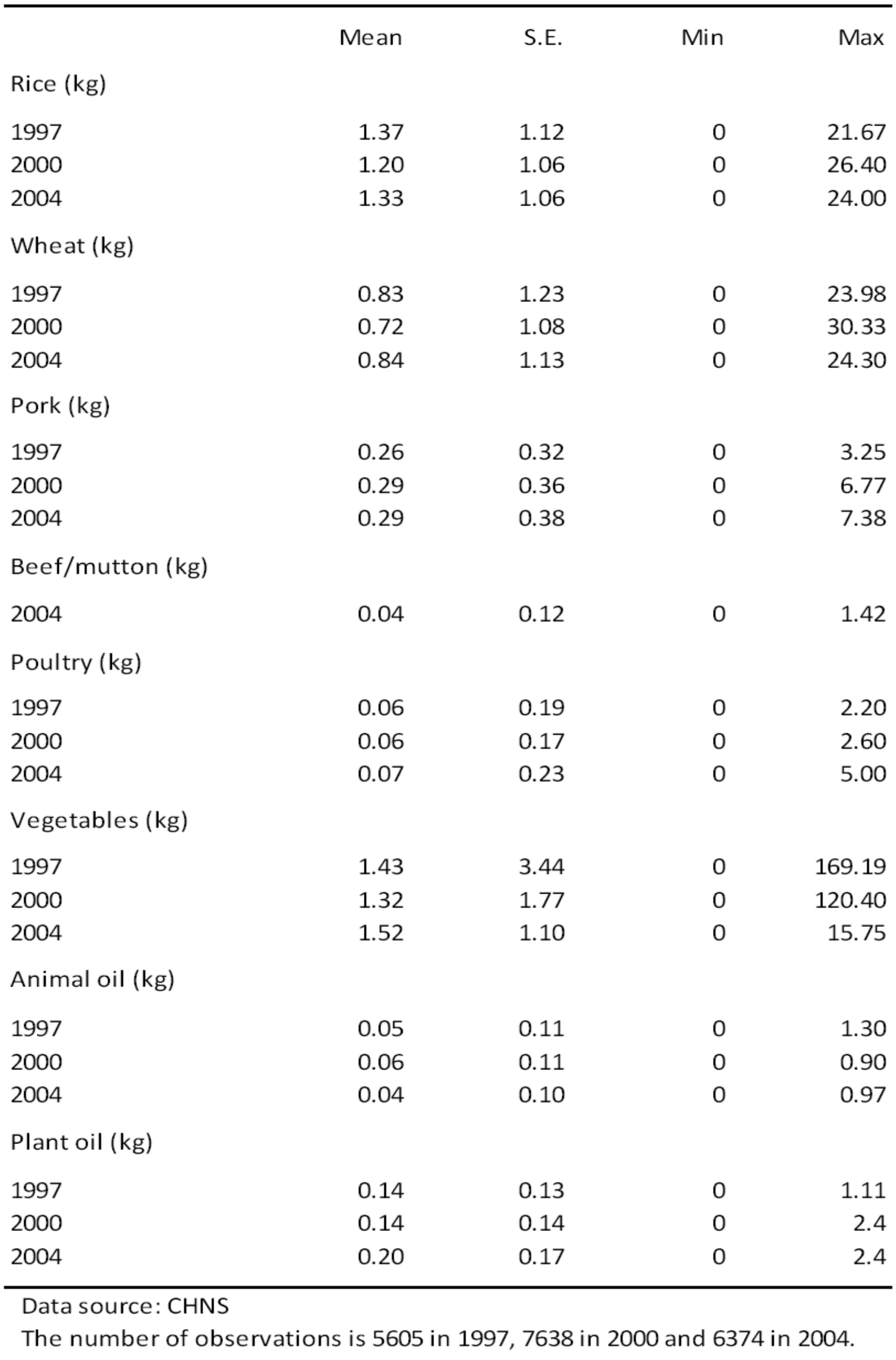
This study uses data from China to evaluate the responsiveness of individuals' dietary choices to information concerning their true health status, as measured by information on hypertension. A fuzzy regression discontinuity approach is used to estimate the causal impact of knowledge of hypertension status on food choices. This method has gained popularity among empirical economists because it mimics an experimental design. The diagnosis rules for hypertension, either SBP above 140 mmHg or DBP above 90, allows for the use of regression discontinuity design. A discontinuity in the probability of being diagnosed is observed at these cutoffs. Because the average outcomes for the sample within a small interval right below the cutoffs are likely to be the same as the average outcomes for the sample right above the cutoffs, the difference in the outcomes between these two may be used to estimate the treatment effect, that is, the effect of being diagnosed as having hypertension.

The findings show that providing information on hypertension status has significant impacts on food choices. A person tends to reduce his or her daily fat intake by 50 grams after being diagnosed with hypertension. The effect on the intake of fat can be partially explained by the impact on several major food categories. Once diagnosed as having hypertension, the consumption of pork and of animal oil decreases significantly, while that of wheat products tends to increase. These changes in food choices are consistent with doctors' advice and are likely to help reduce blood pressure. These effects are stronger for the 2004 data, suggesting that consumers in China are becoming more concerned about their chronic health status in recent years. A healthier diet is chosen after a diagnosis of hypertension.

The findings of this study have very important policy implications. It suggests that consumers are responsive in changing their diet in a healthier direction if they are better informed of their true chronic health status. Health policies or interventions that promote better monitoring of chronic health status may have important impacts on reducing the risk of developing chronic diseases, such as cardiovascular diseases, cancers, and diabetes. For example, some regulations on the health insurance of preventive care may be necessary to help consumers to become better informed, treated and, most importantly, directed to a healthier lifestyle.

Tables and figures

Table 3.1: Descriptive statistics of 3-day food consumption per capita

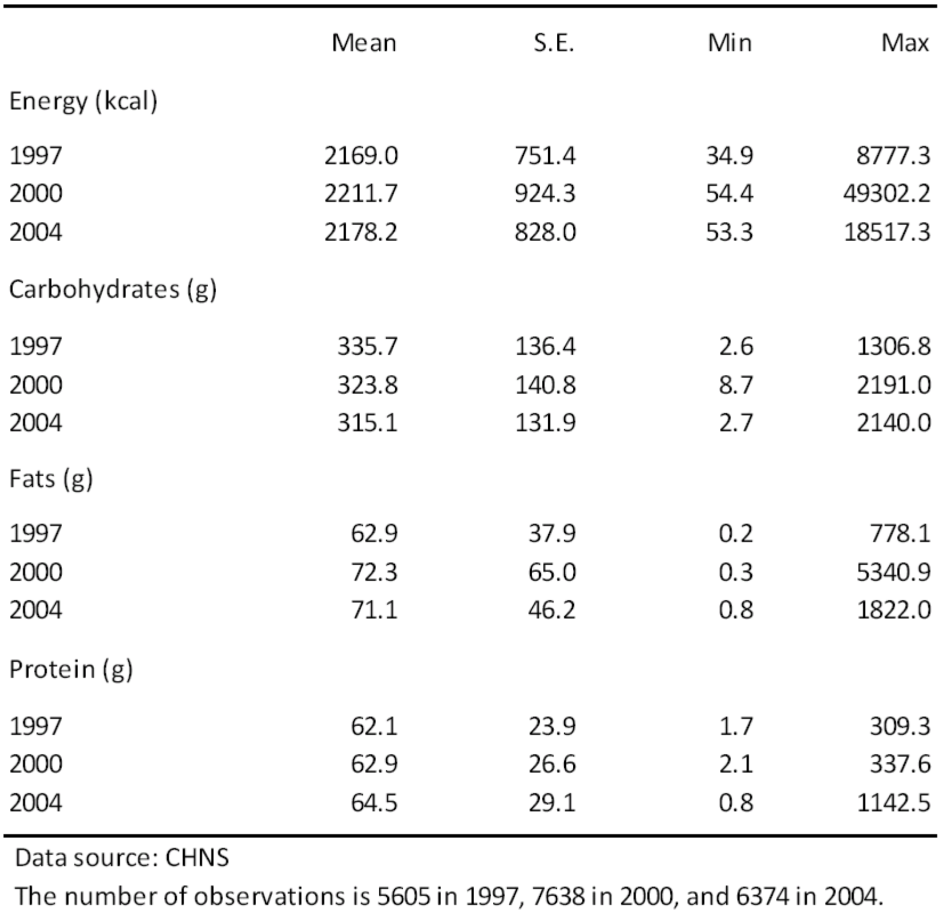


	Mean	S.E.	Min	Max
Rice (kg)				
1997	1.37	1.12	0	21.67
2000	1.20	1.06	0	26.40
2004	1.33	1.06	0	24.00
Wheat (kg)				
1997	0.83	1.23	0	23.98
2000	0.72	1.08	0	30.33
2004	0.84	1.13	0	24.30
Pork (kg)				
1997	0.26	0.32	0	3.25
2000	0.29	0.36	0	6.77
2004	0.29	0.38	0	7.38
Beef/mutton (kg)				
2004	0.04	0.12	0	1.42
Poultry (kg)				
1997	0.06	0.19	0	2.20
2000	0.06	0.17	0	2.60
2004	0.07	0.23	0	5.00
Vegetables (kg)				
1997	1.43	3.44	0	169.19
2000	1.32	1.77	0	120.40
2004	1.52	1.10	0	15.75
Animal oil (kg)				
1997	0.05	0.11	0	1.30
2000	0.06	0.11	0	0.90
2004	0.04	0.10	0	0.97
Plant oil (kg)				
1997	0.14	0.13	0	1.11
2000	0.14	0.14	0	2.4
2004	0.20	0.17	0	2.4

Data source: CHNS

The number of observations is 5605 in 1997, 7638 in 2000 and 6374 in 2004.

Table 3.2: Descriptive statistics of daily nutrient consumption



	Mean	S.E.	Min	Max
Energy (kcal)				
1997	2169.0	751.4	34.9	8777.3
2000	2211.7	924.3	54.4	49302.2
2004	2178.2	828.0	53.3	18517.3
Carbohydrates (g)				
1997	335.7	136.4	2.6	1306.8
2000	323.8	140.8	8.7	2191.0
2004	315.1	131.9	2.7	2140.0
Fats (g)				
1997	62.9	37.9	0.2	778.1
2000	72.3	65.0	0.3	5340.9
2004	71.1	46.2	0.8	1822.0
Protein (g)				
1997	62.1	23.9	1.7	309.3
2000	62.9	26.6	2.1	337.6
2004	64.5	29.1	0.8	1142.5

Data source: CHNS

The number of observations is 5605 in 1997, 7638 in 2000, and 6374 in 2004.

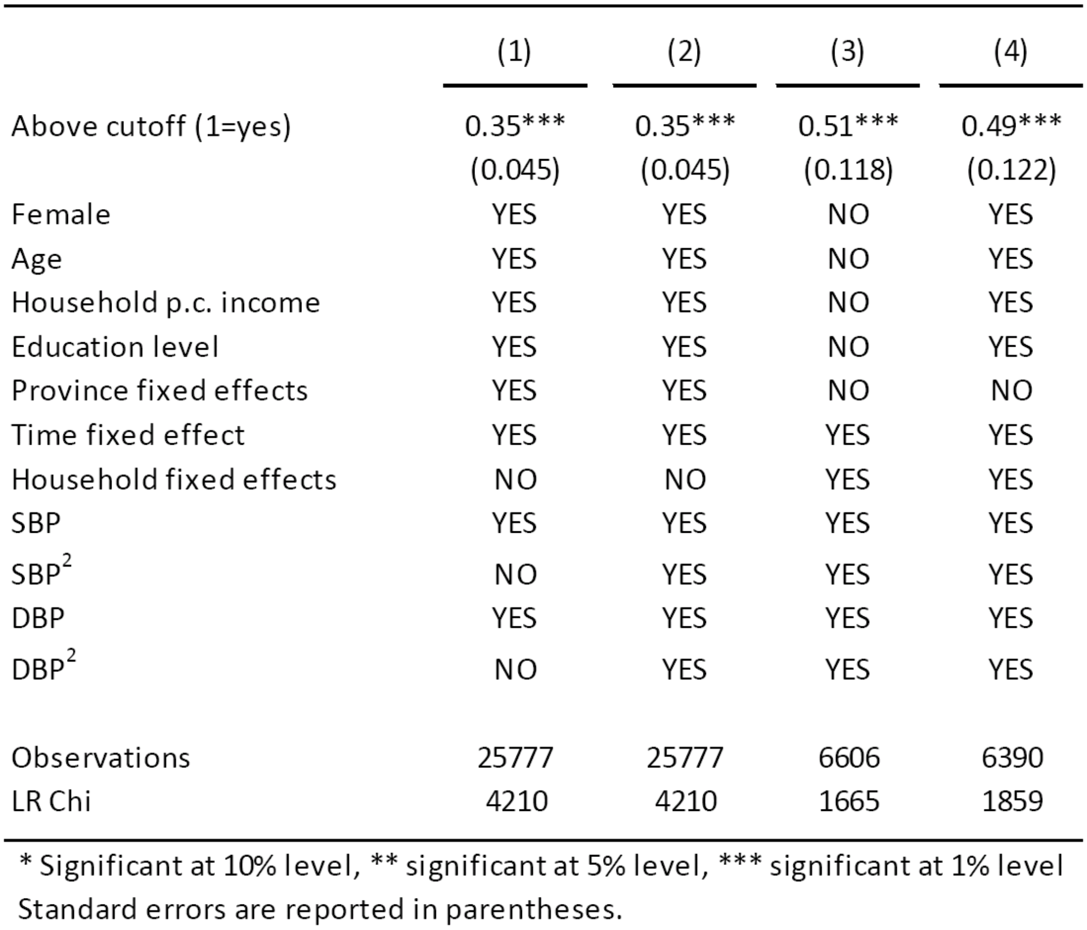
Table 3.3: Prevalence of hypertension

	Total			Pop aged 18-45			Pop aged 45-65			Pop aged 65+		
	Diag.	O.M.[a]	Diff.	Diag.	O.M.	Diff.	Diag.	O.M.	Diff.	Diag.	O.M.	Diff.
1997 (Obs= 8127)												
Hypertensive (%)	4.4	19.0	-14.6	0.7	8.5	-7.8	6.2	26.6	-20.4	16.5	46.7	-30.2
Among hypertensive, % of taking anti-hypertension drugs	65.4	11.6		51.6	2.9		64.9	12.1		68.5	18.1	
2000 (Obs=9142)												
Hypertensive (%)	7.1	19.5	-12.4	1.4	8.8	-7.3	10.1	25.6	-15.5	20.7	43.9	-23.2
Among hypertensive, % of taking anti-hypertension drugs	68.7	18.3		50.0	6.4		66.3	19.2		78.3	25.8	
2004 (Obs=9537)												
Hypertensive (%)	9.0	21.6	-12.7	1.6	10.0	-8.4	11.1	25.7	-14.7	23.9	42.6	-18.7
Among hypertensive, % of taking anti-hypertension drugs	72.8	20.8		60.7	7.4		71.2	21.2		77.1	28.8	

Data source: CHNS

[a] O.M. stands for objectively measured.

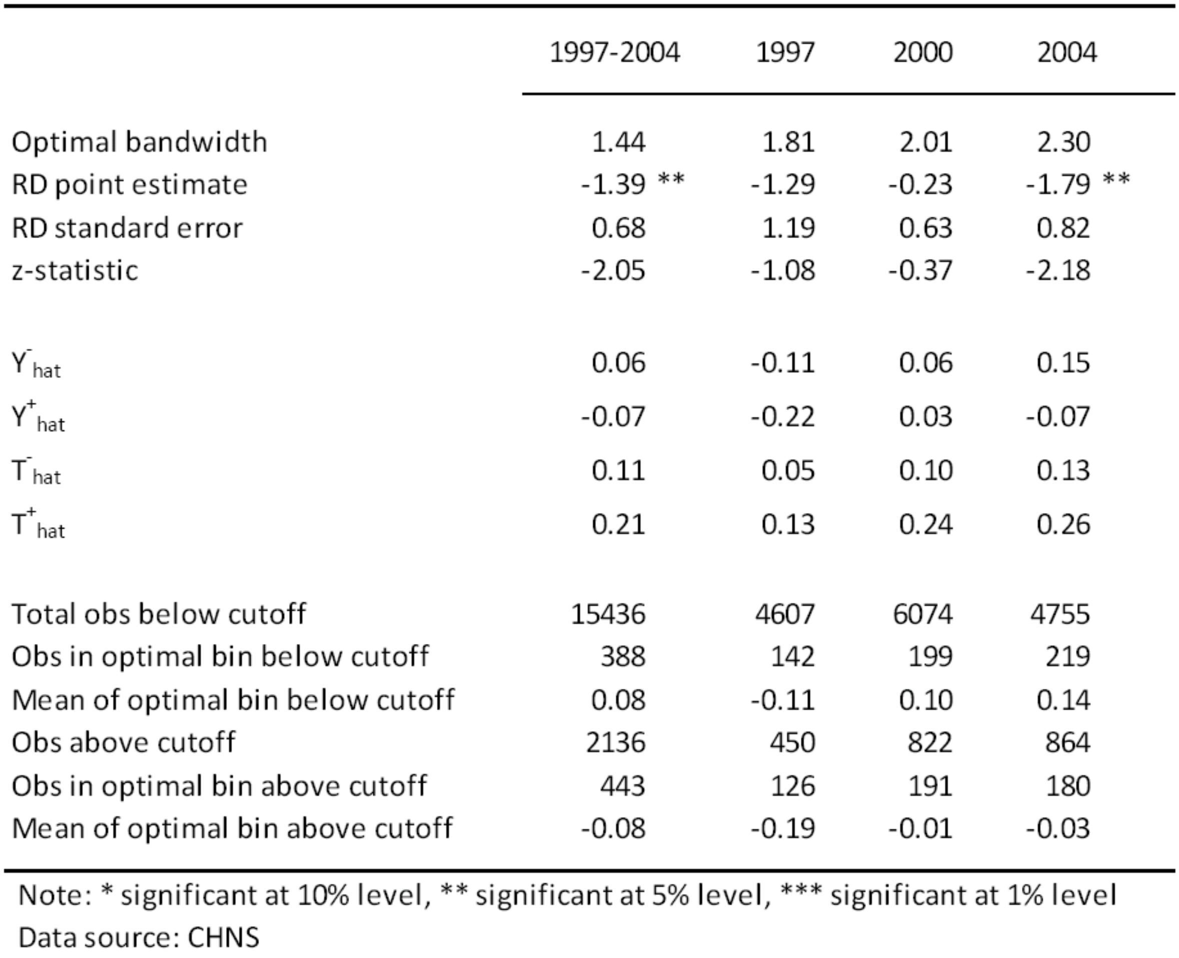
Table 3.4: Probability of the diagnosis of hypertension at the cutoff



	(1)	(2)	(3)	(4)
Above cutoff (1=yes)	0.35***	0.35***	0.51***	0.49***
	(0.045)	(0.045)	(0.118)	(0.122)
Female	YES	YES	NO	YES
Age	YES	YES	NO	YES
Household p.c. income	YES	YES	NO	YES
Education level	YES	YES	NO	YES
Province fixed effects	YES	YES	NO	NO
Time fixed effect	YES	YES	YES	YES
Household fixed effects	NO	NO	YES	YES
SBP	YES	YES	YES	YES
SBP^2	NO	YES	YES	YES
DBP	YES	YES	YES	YES
DBP^2	NO	YES	YES	YES
Observations	25777	25777	6606	6390
LR Chi	4210	4210	1665	1859

* Significant at 10% level, ** significant at 5% level, *** significant at 1% level
Standard errors are reported in parentheses.

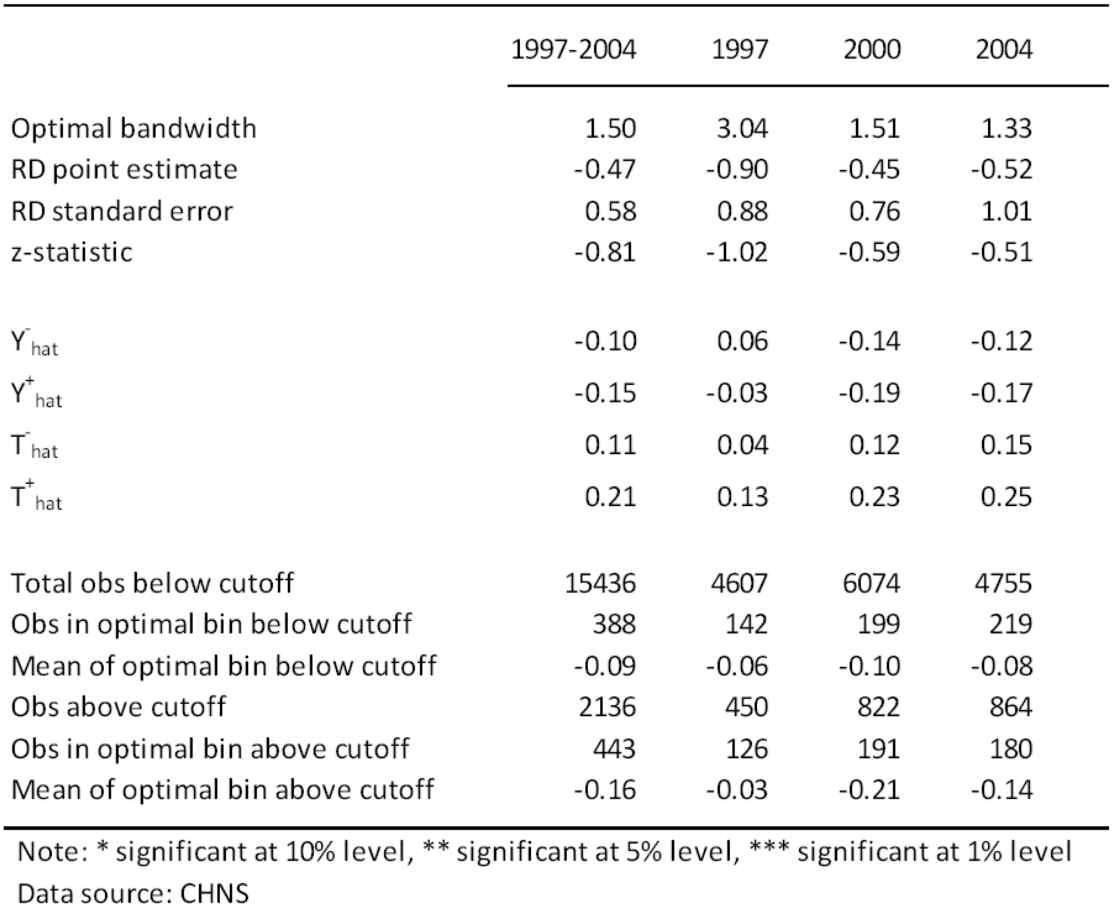
Table 3.5: Local linear regression estimates, daily fat intake



	1997-2004	1997	2000	2004
Optimal bandwidth	1.44	1.81	2.01	2.30
RD point estimate	-1.39 **	-1.29	-0.23	-1.79 **
RD standard error	0.68	1.19	0.63	0.82
z-statistic	-2.05	-1.08	-0.37	-2.18
Y^{-}_{hat}	0.06	-0.11	0.06	0.15
Y^{+}_{hat}	-0.07	-0.22	0.03	-0.07
T^{-}_{hat}	0.11	0.05	0.10	0.13
T^{+}_{hat}	0.21	0.13	0.24	0.26
Total obs below cutoff	15436	4607	6074	4755
Obs in optimal bin below cutoff	388	142	199	219
Mean of optimal bin below cutoff	0.08	-0.11	0.10	0.14
Obs above cutoff	2136	450	822	864
Obs in optimal bin above cutoff	443	126	191	180
Mean of optimal bin above cutoff	-0.08	-0.19	-0.01	-0.03

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

Table 3.6: Local linear regression estimates, daily energy intake

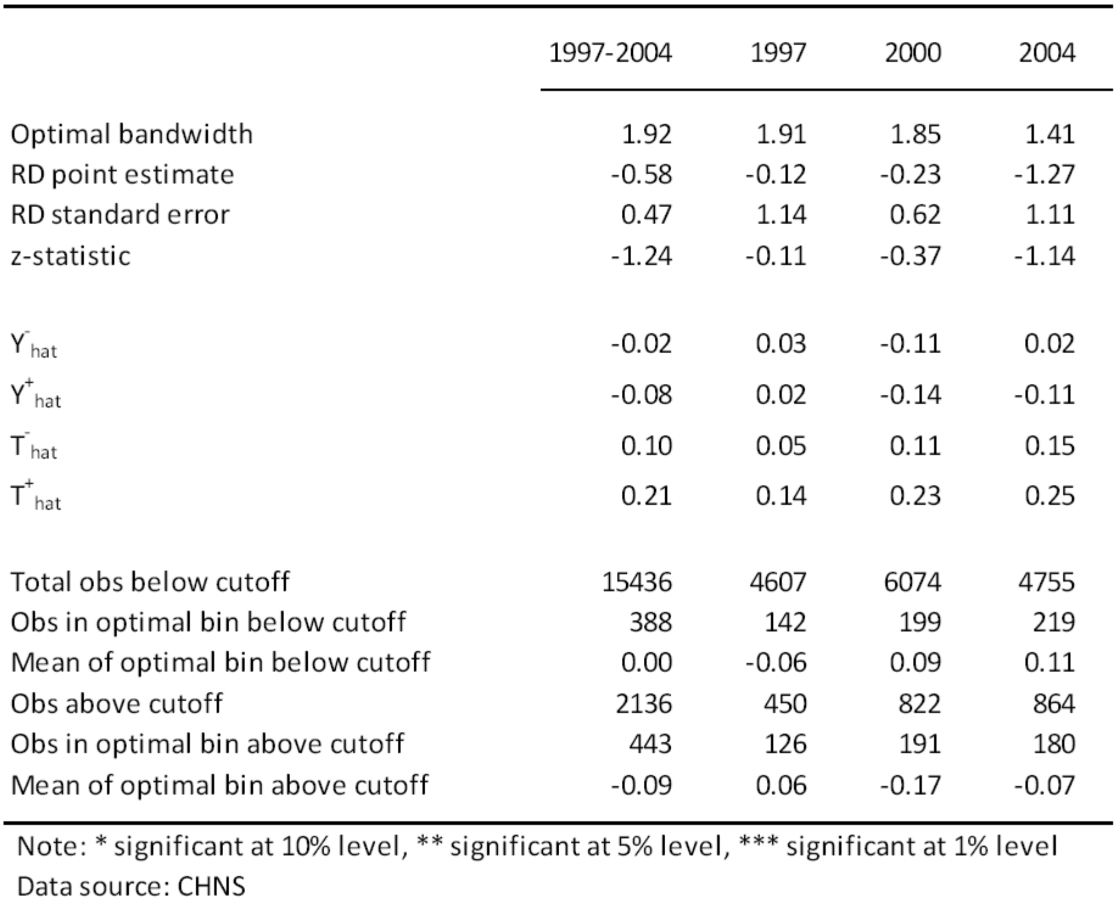


	1997-2004	1997	2000	2004
Optimal bandwidth	1.50	3.04	1.51	1.33
RD point estimate	-0.47	-0.90	-0.45	-0.52
RD standard error	0.58	0.88	0.76	1.01
z-statistic	-0.81	-1.02	-0.59	-0.51
Y^{-}_{hat}	-0.10	0.06	-0.14	-0.12
Y^{+}_{hat}	-0.15	-0.03	-0.19	-0.17
T^{-}_{hat}	0.11	0.04	0.12	0.15
T^{+}_{hat}	0.21	0.13	0.23	0.25
Total obs below cutoff	15436	4607	6074	4755
Obs in optimal bin below cutoff	388	142	199	219
Mean of optimal bin below cutoff	-0.09	-0.06	-0.10	-0.08
Obs above cutoff	2136	450	822	864
Obs in optimal bin above cutoff	443	126	191	180
Mean of optimal bin above cutoff	-0.16	-0.03	-0.21	-0.14

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level

Data source: CHNS

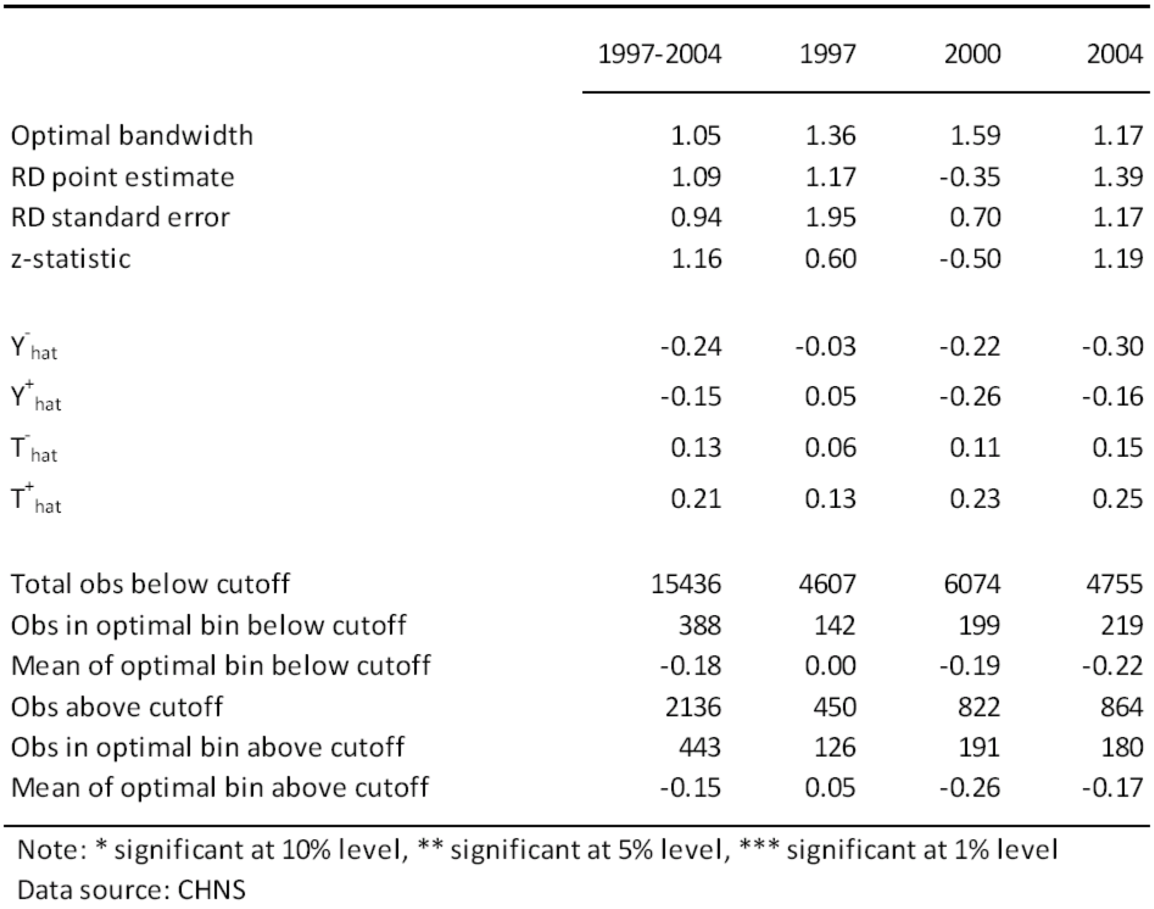
Table 3.7: Local linear regression estimates, daily protein intake



	1997-2004	1997	2000	2004
Optimal bandwidth	1.92	1.91	1.85	1.41
RD point estimate	-0.58	-0.12	-0.23	-1.27
RD standard error	0.47	1.14	0.62	1.11
z-statistic	-1.24	-0.11	-0.37	-1.14
Y^{-}_{hat}	-0.02	0.03	-0.11	0.02
Y^{+}_{hat}	-0.08	0.02	-0.14	-0.11
T^{-}_{hat}	0.10	0.05	0.11	0.15
T^{+}_{hat}	0.21	0.14	0.23	0.25
Total obs below cutoff	15436	4607	6074	4755
Obs in optimal bin below cutoff	388	142	199	219
Mean of optimal bin below cutoff	0.00	-0.06	0.09	0.11
Obs above cutoff	2136	450	822	864
Obs in optimal bin above cutoff	443	126	191	180
Mean of optimal bin above cutoff	-0.09	0.06	-0.17	-0.07

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

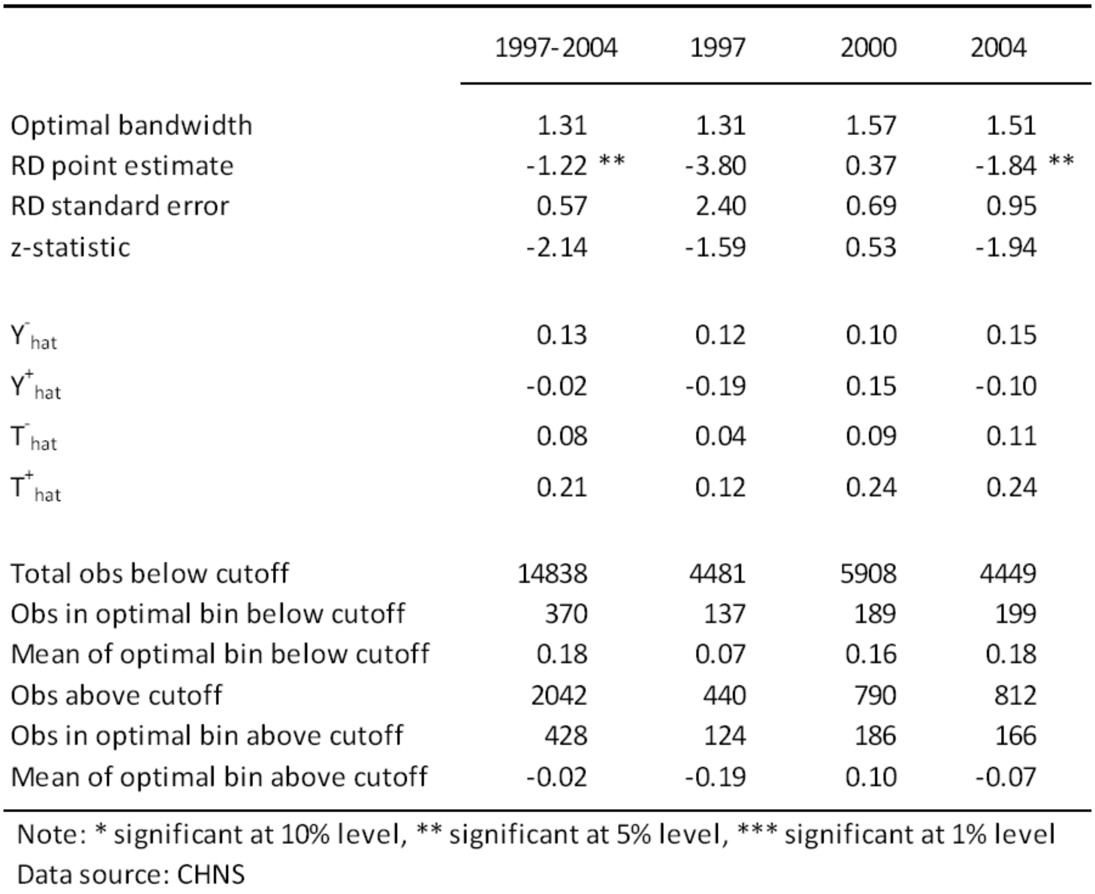
Table 3.8: Local linear regression estimates, daily carbohydrate intake



	1997-2004	1997	2000	2004
Optimal bandwidth	1.05	1.36	1.59	1.17
RD point estimate	1.09	1.17	-0.35	1.39
RD standard error	0.94	1.95	0.70	1.17
z-statistic	1.16	0.60	-0.50	1.19
Y^{-}_{hat}	-0.24	-0.03	-0.22	-0.30
Y^{+}_{hat}	-0.15	0.05	-0.26	-0.16
T^{-}_{hat}	0.13	0.06	0.11	0.15
T^{+}_{hat}	0.21	0.13	0.23	0.25
Total obs below cutoff	15436	4607	6074	4755
Obs in optimal bin below cutoff	388	142	199	219
Mean of optimal bin below cutoff	-0.18	0.00	-0.19	-0.22
Obs above cutoff	2136	450	822	864
Obs in optimal bin above cutoff	443	126	191	180
Mean of optimal bin above cutoff	-0.15	0.05	-0.26	-0.17

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

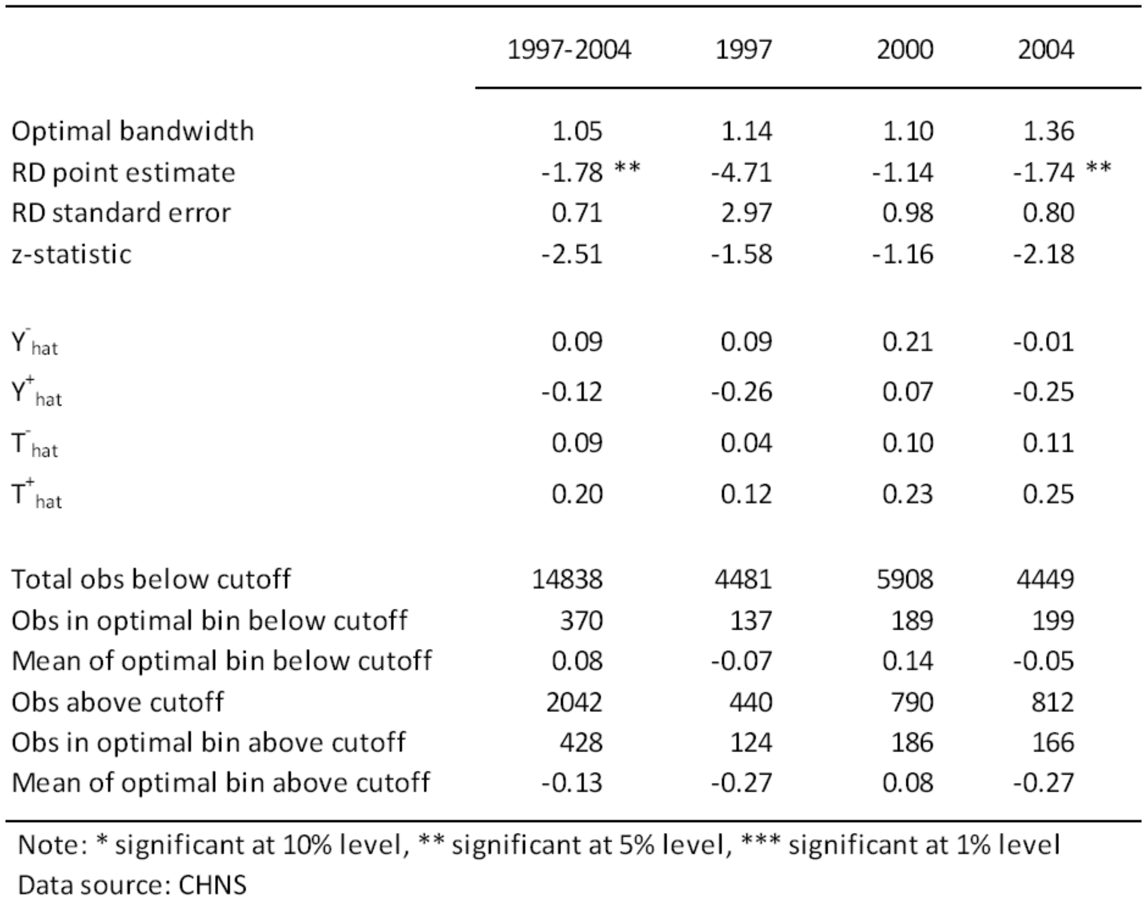
Table 3.9: Local linear regression estimates, 3-day pork consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	1.31	1.31	1.57	1.51
RD point estimate	-1.22 **	-3.80	0.37	-1.84 **
RD standard error	0.57	2.40	0.69	0.95
z-statistic	-2.14	-1.59	0.53	-1.94
Y^{-}_{hat}	0.13	0.12	0.10	0.15
Y^{+}_{hat}	-0.02	-0.19	0.15	-0.10
T^{-}_{hat}	0.08	0.04	0.09	0.11
T^{+}_{hat}	0.21	0.12	0.24	0.24
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	0.18	0.07	0.16	0.18
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	-0.02	-0.19	0.10	-0.07

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

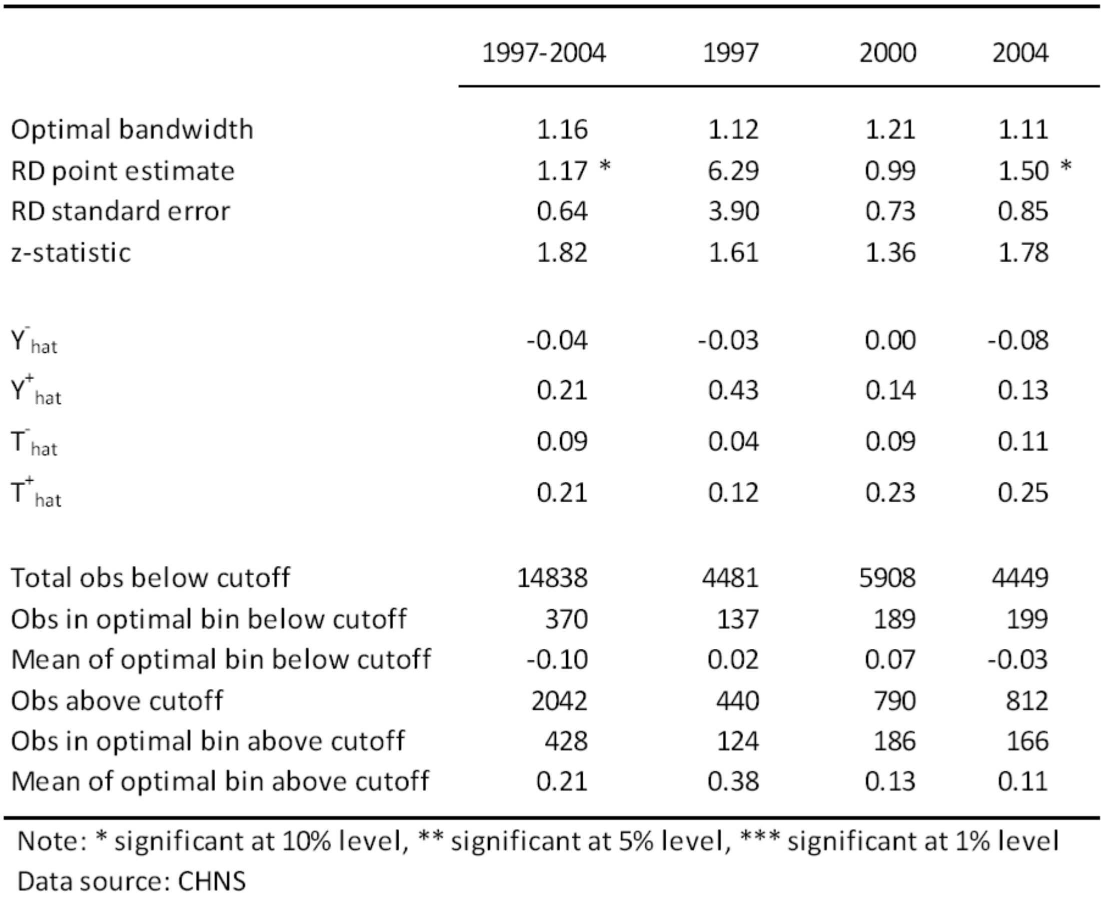
Table 3.10: Local linear regression estimates, 3-day animal oil consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	1.05	1.14	1.10	1.36
RD point estimate	-1.78 **	-4.71	-1.14	-1.74 **
RD standard error	0.71	2.97	0.98	0.80
z-statistic	-2.51	-1.58	-1.16	-2.18
Y^{-}_{hat}	0.09	0.09	0.21	-0.01
Y^{+}_{hat}	-0.12	-0.26	0.07	-0.25
T^{-}_{hat}	0.09	0.04	0.10	0.11
T^{+}_{hat}	0.20	0.12	0.23	0.25
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	0.08	-0.07	0.14	-0.05
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	-0.13	-0.27	0.08	-0.27

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

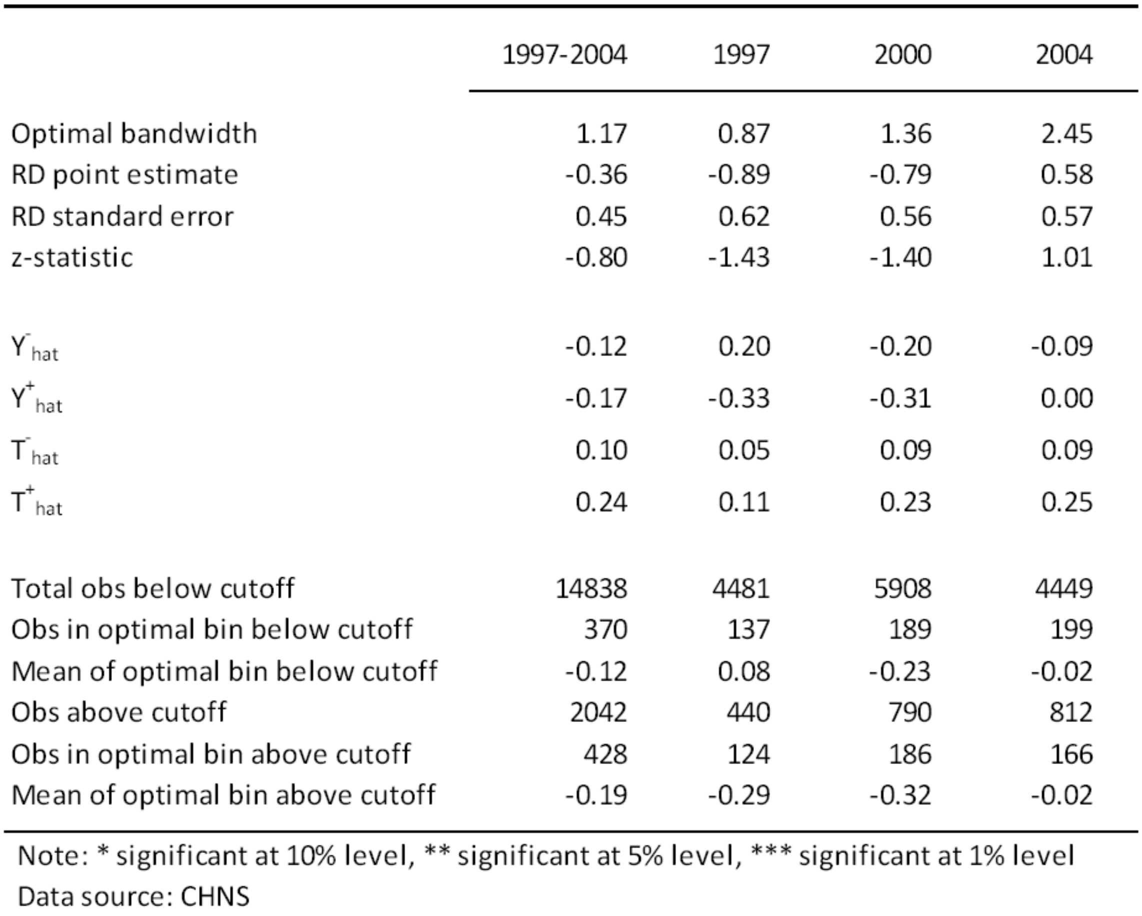
Table 3.11: Local linear regression estimates, 3-day wheat consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	1.16	1.12	1.21	1.11
RD point estimate	1.17 *	6.29	0.99	1.50 *
RD standard error	0.64	3.90	0.73	0.85
z-statistic	1.82	1.61	1.36	1.78
Y^{-}_{hat}	-0.04	-0.03	0.00	-0.08
Y^{+}_{hat}	0.21	0.43	0.14	0.13
T^{-}_{hat}	0.09	0.04	0.09	0.11
T^{+}_{hat}	0.21	0.12	0.23	0.25
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	-0.10	0.02	0.07	-0.03
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	0.21	0.38	0.13	0.11

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

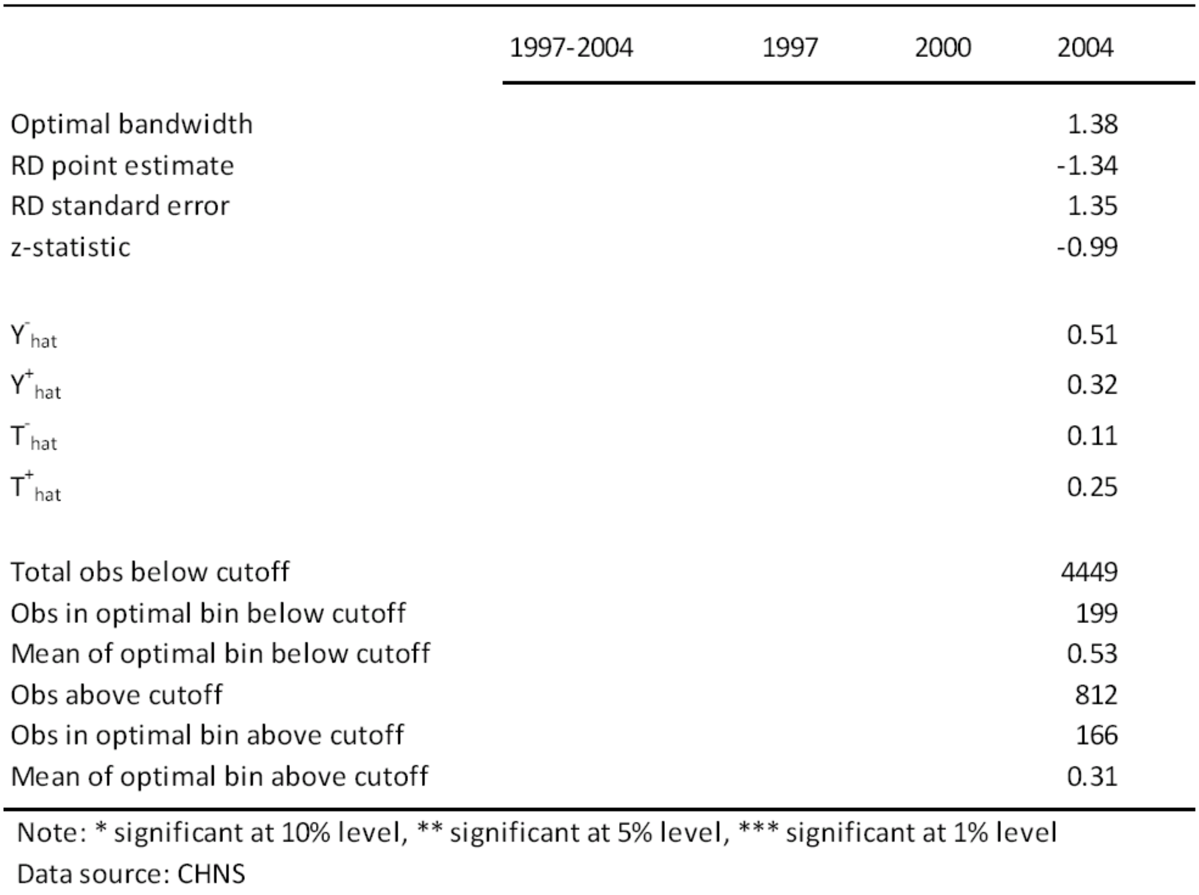
Table 3.12: Local linear regression estimates, 3-day rice consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	1.17	0.87	1.36	2.45
RD point estimate	-0.36	-0.89	-0.79	0.58
RD standard error	0.45	0.62	0.56	0.57
z-statistic	-0.80	-1.43	-1.40	1.01
Y^{-}_{hat}	-0.12	0.20	-0.20	-0.09
Y^{+}_{hat}	-0.17	-0.33	-0.31	0.00
T^{-}_{hat}	0.10	0.05	0.09	0.09
T^{+}_{hat}	0.24	0.11	0.23	0.25
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	-0.12	0.08	-0.23	-0.02
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	-0.19	-0.29	-0.32	-0.02

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

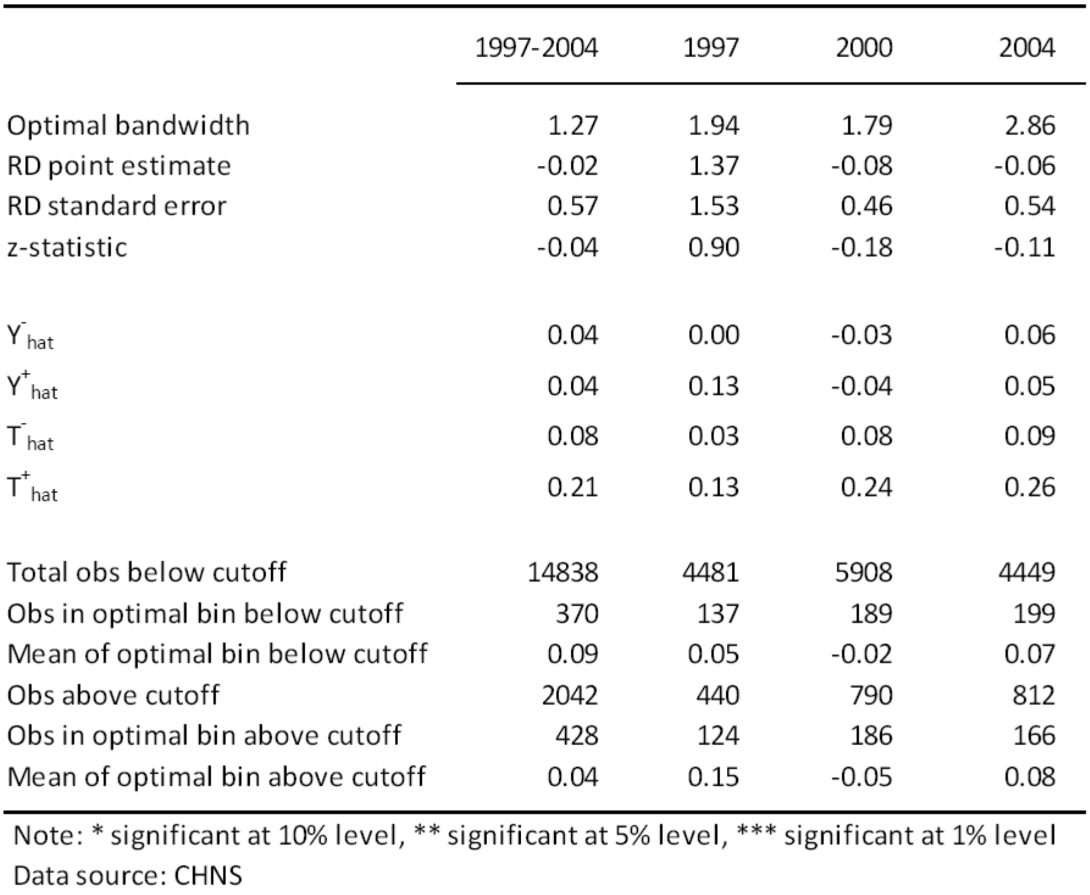
Table 3.13: Local linear regression estimates, 3-day beef and mutton consumption



	1997-2004	1997	2000	2004
Optimal bandwidth				1.38
RD point estimate				-1.34
RD standard error				1.35
z-statistic				-0.99
Y^{-}_{hat}				0.51
Y^{+}_{hat}				0.32
T^{-}_{hat}				0.11
T^{+}_{hat}				0.25
Total obs below cutoff				4449
Obs in optimal bin below cutoff				199
Mean of optimal bin below cutoff				0.53
Obs above cutoff				812
Obs in optimal bin above cutoff				166
Mean of optimal bin above cutoff				0.31

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

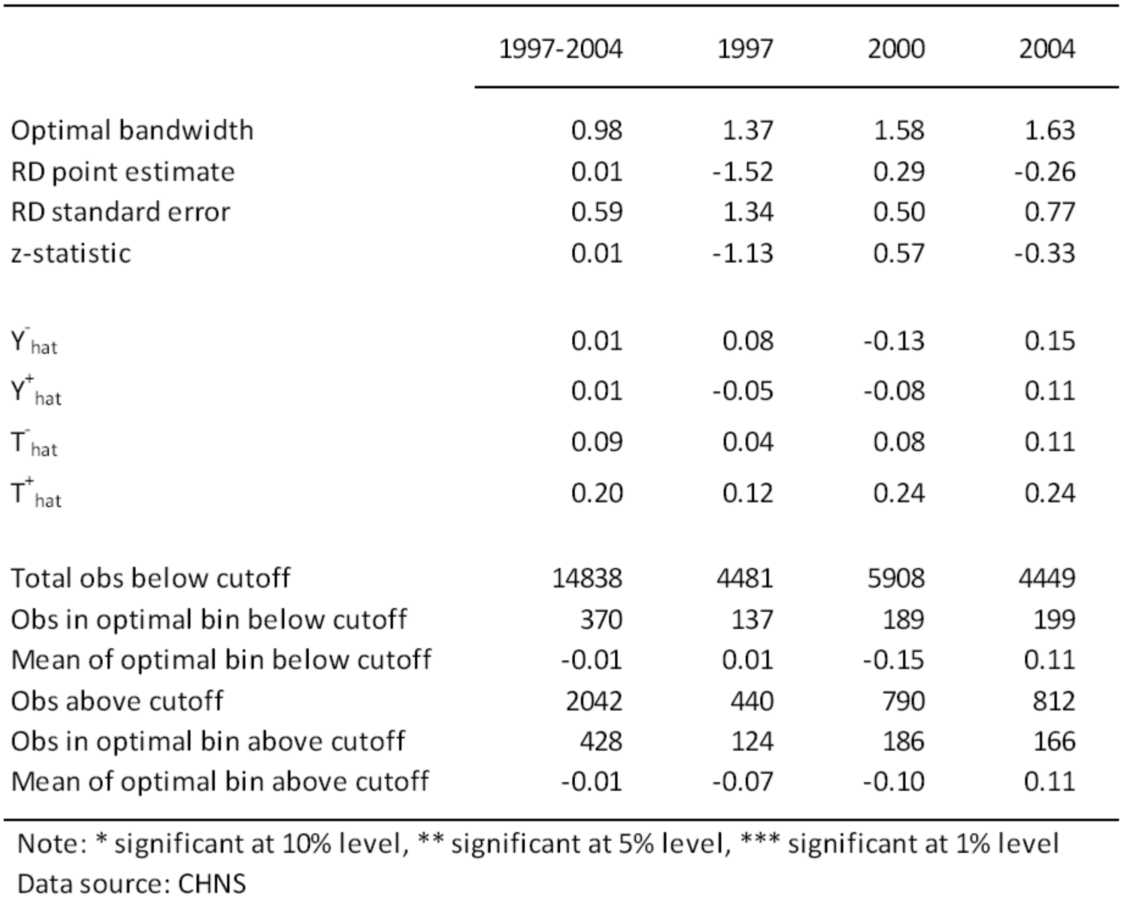
Table 3.14: Local linear regression estimates, 3-day poultry consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	1.27	1.94	1.79	2.86
RD point estimate	-0.02	1.37	-0.08	-0.06
RD standard error	0.57	1.53	0.46	0.54
z-statistic	-0.04	0.90	-0.18	-0.11
Y^{-}_{hat}	0.04	0.00	-0.03	0.06
Y^{+}_{hat}	0.04	0.13	-0.04	0.05
T^{-}_{hat}	0.08	0.03	0.08	0.09
T^{+}_{hat}	0.21	0.13	0.24	0.26
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	0.09	0.05	-0.02	0.07
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	0.04	0.15	-0.05	0.08

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

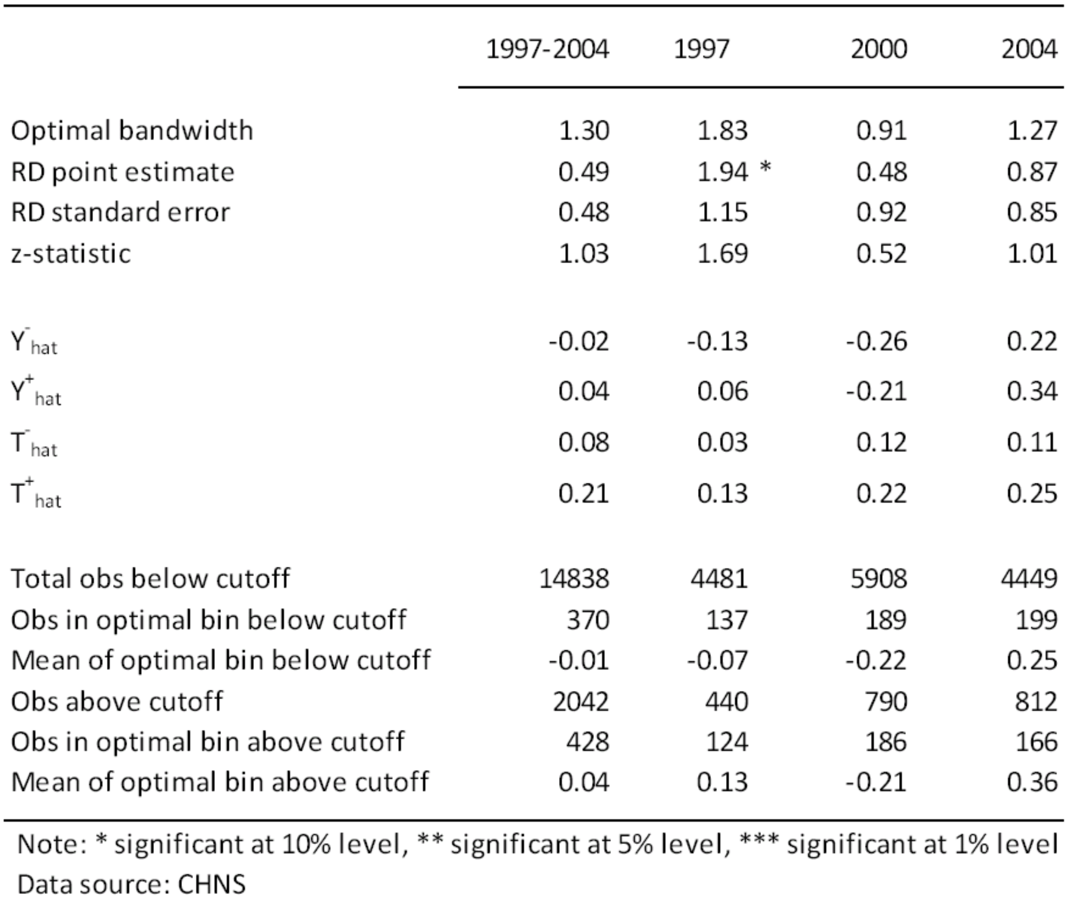
Table 3.15: Local linear regression estimates, 3-day vegetable consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	0.98	1.37	1.58	1.63
RD point estimate	0.01	-1.52	0.29	-0.26
RD standard error	0.59	1.34	0.50	0.77
z-statistic	0.01	-1.13	0.57	-0.33
Y^{-}_{hat}	0.01	0.08	-0.13	0.15
Y^{+}_{hat}	0.01	-0.05	-0.08	0.11
T^{-}_{hat}	0.09	0.04	0.08	0.11
T^{+}_{hat}	0.20	0.12	0.24	0.24
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	-0.01	0.01	-0.15	0.11
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	-0.01	-0.07	-0.10	0.11

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

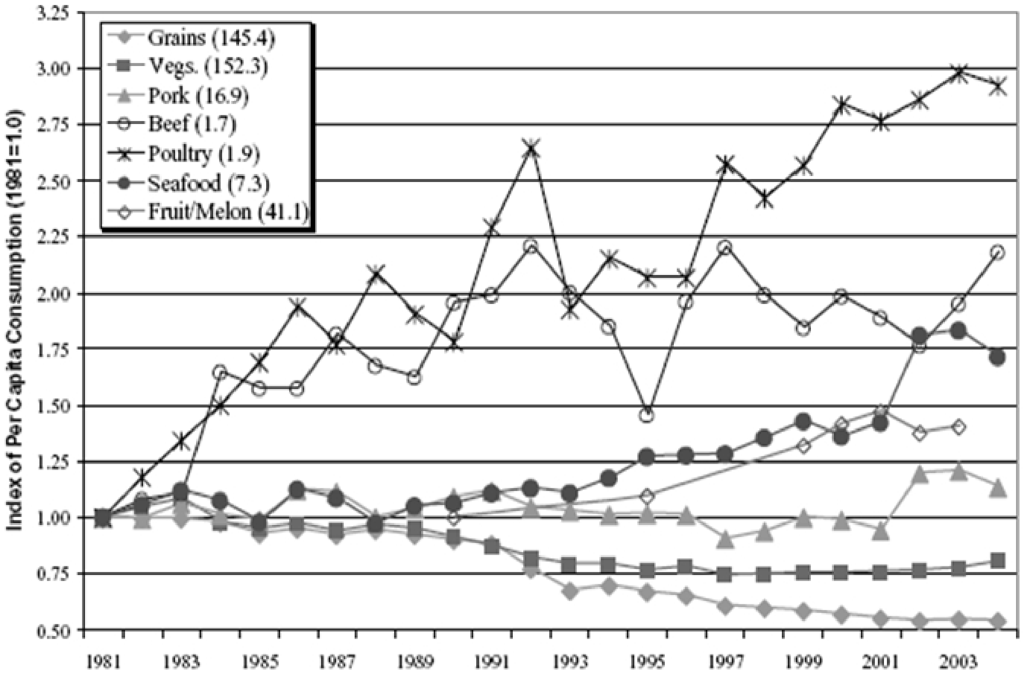
Table 3.16: Local linear regression estimates, 3-day plant oil consumption



	1997-2004	1997	2000	2004
Optimal bandwidth	1.30	1.83	0.91	1.27
RD point estimate	0.49	1.94 *	0.48	0.87
RD standard error	0.48	1.15	0.92	0.85
z-statistic	1.03	1.69	0.52	1.01
Y^{-}_{hat}	-0.02	-0.13	-0.26	0.22
Y^{+}_{hat}	0.04	0.06	-0.21	0.34
T^{-}_{hat}	0.08	0.03	0.12	0.11
T^{+}_{hat}	0.21	0.13	0.22	0.25
Total obs below cutoff	14838	4481	5908	4449
Obs in optimal bin below cutoff	370	137	189	199
Mean of optimal bin below cutoff	-0.01	-0.07	-0.22	0.25
Obs above cutoff	2042	440	790	812
Obs in optimal bin above cutoff	428	124	186	166
Mean of optimal bin above cutoff	0.04	0.13	-0.21	0.36

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level
Data source: CHNS

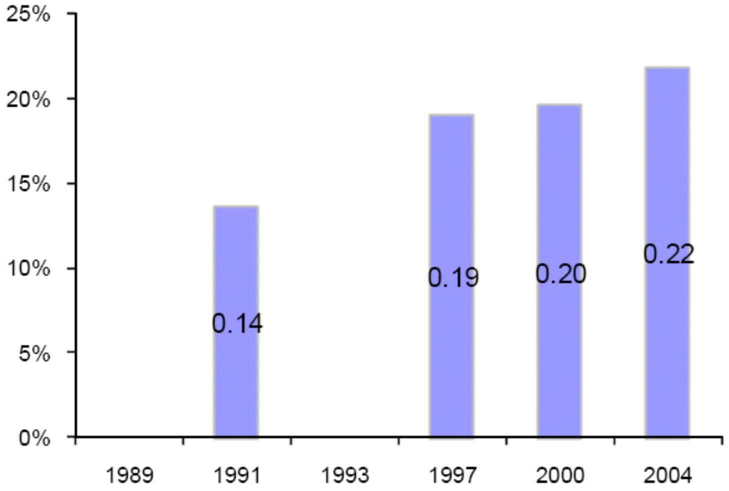
Figure 3.1: Food demand trends in urban China



Note: 1981 Kg. values in parentheses (except for fruit)

Source: Gould and Villarreal (2006), Cited China Statistical Yearbook, various years

Figure 3.2: Prevalence of hypertension, aged 18-70



Source: China Health and Nutrition Survey, 1989-2004

Figure 3.3: Distribution of hypertension prevalence in sample provinces

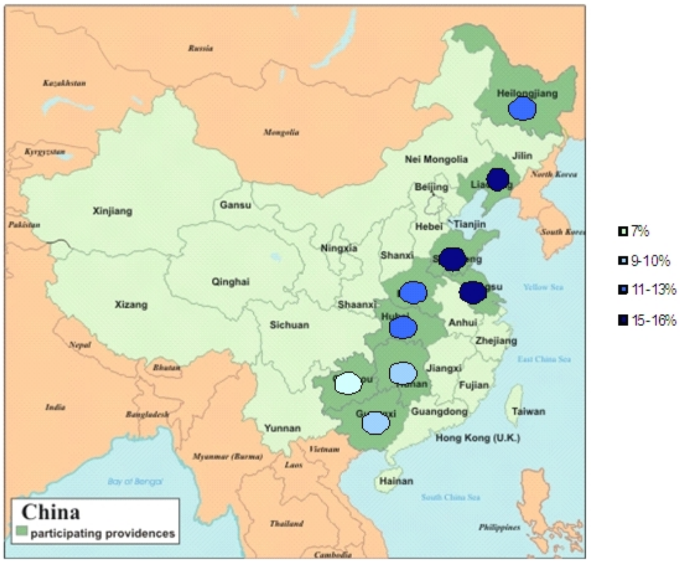


Figure 3.4: Percentages of individuals diagnosed with hypertension

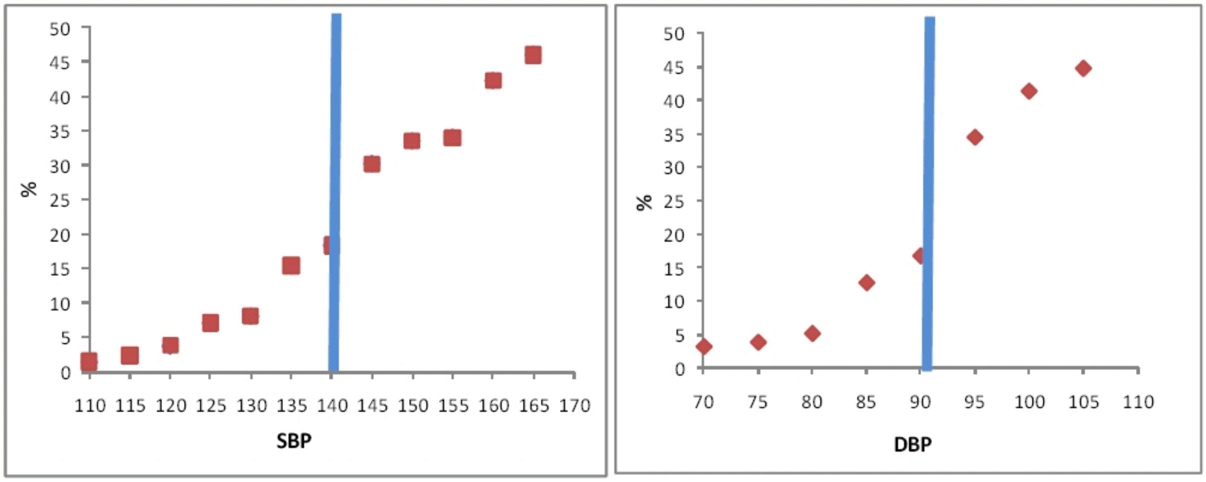
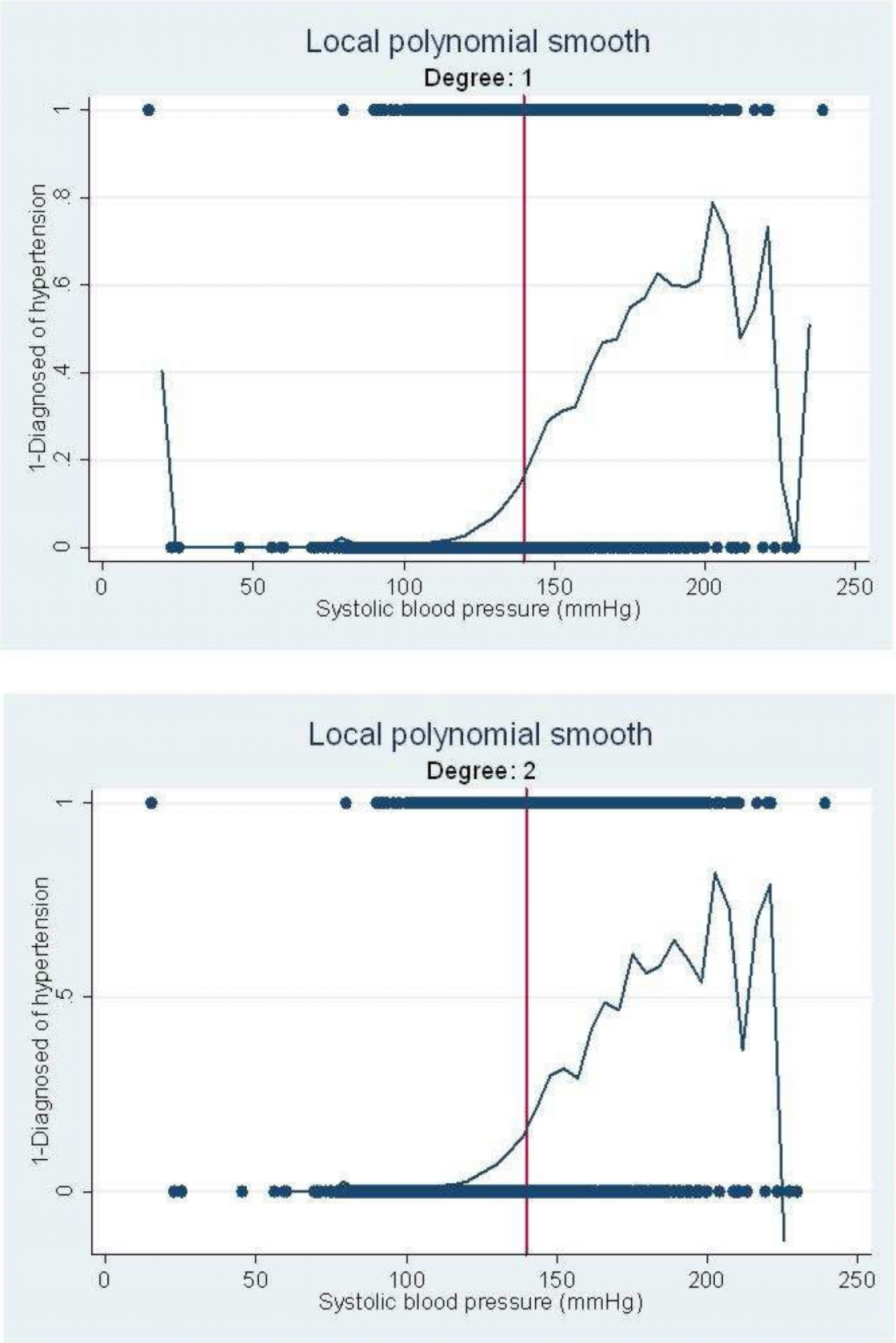
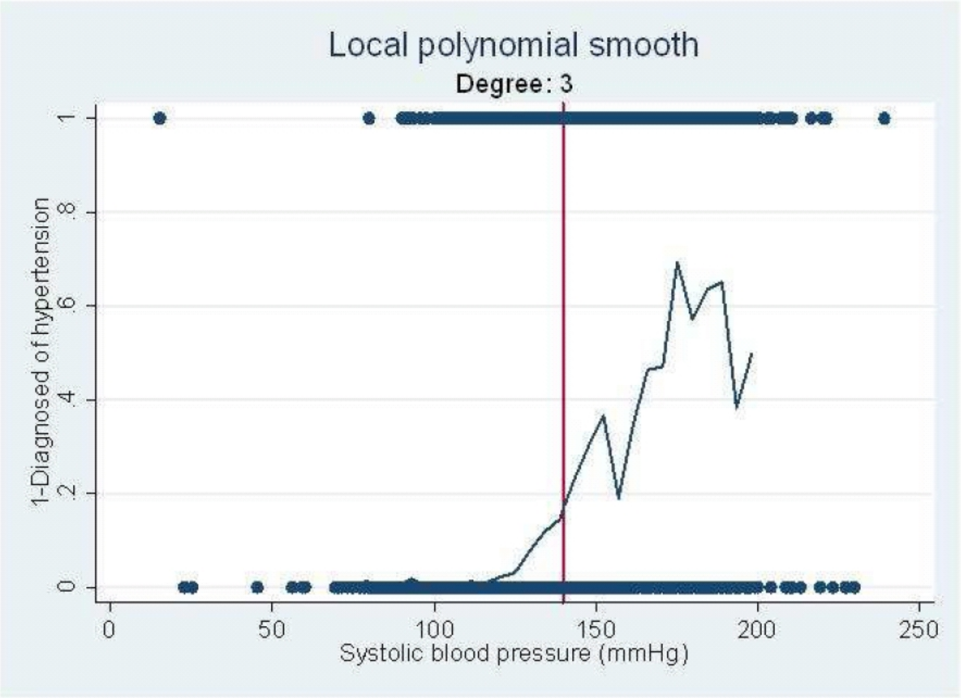


Figure 3.5: The diagnosis of hypertension




Local polynomial smooth
Degree: 3
1-Diagnosed of hypertension
0
.2
.4
.6
.8
1
0
50
100
150
200
250
Systolic blood pressure (mmHg)

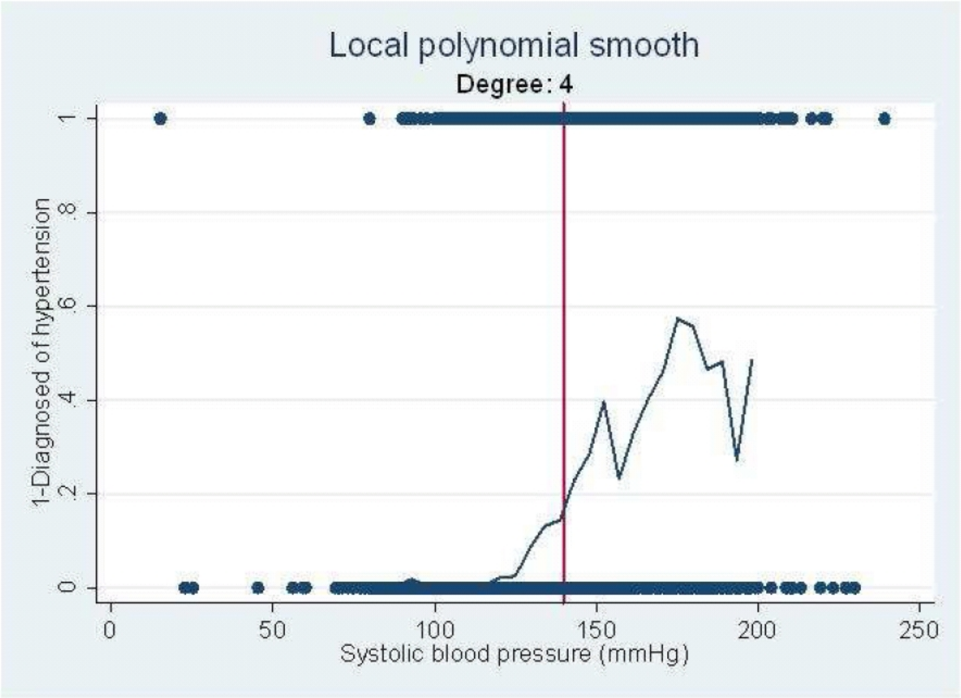
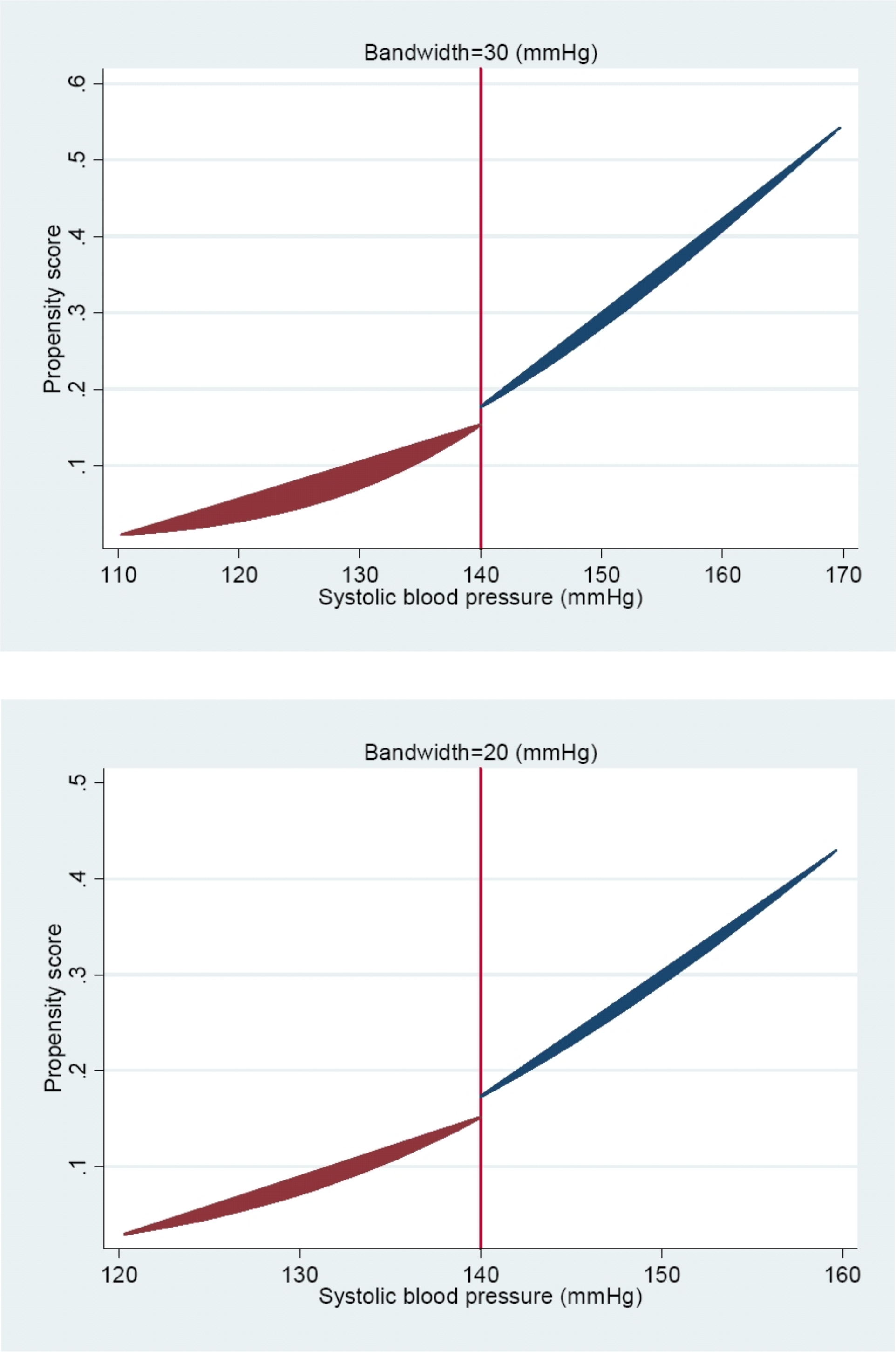
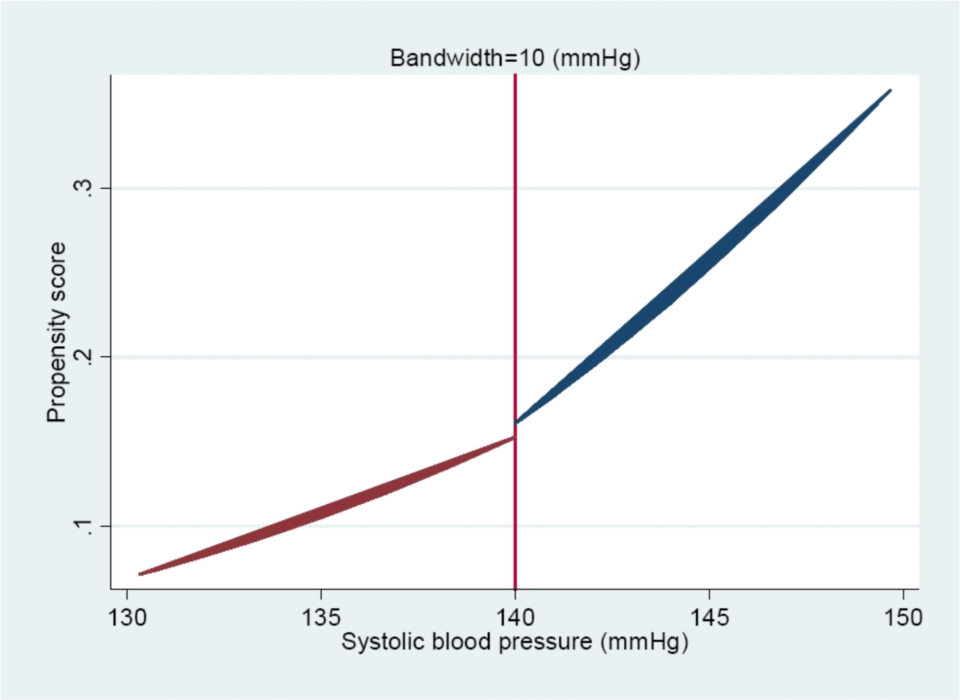

Local polynomial smooth
Degree: 4
1-Diagnosed of hypertension
0
.2
.4
.6
.8
1
0
50
100
150
200
250
Systolic blood pressure (mmHg)

Figure 3.6: The predicted propensity score of a diagnosis of hypertension




Bandwidth=10 (mmHg)
Propensity score
.1
.2
.3
130
135
140
145
150
Systolic blood pressure (mmHg)

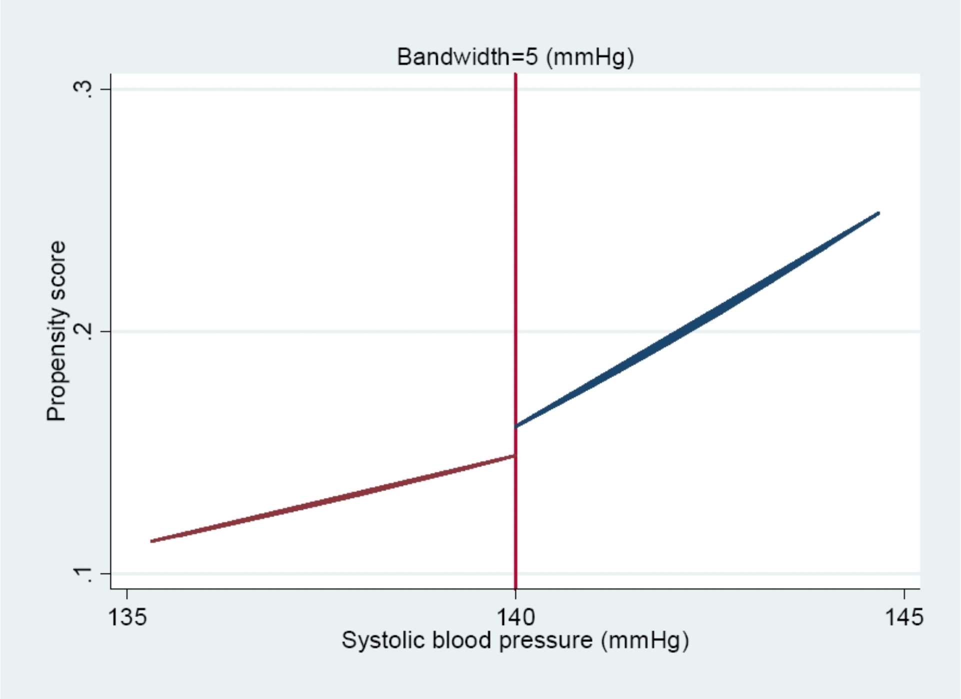

Bandwidth=5 (mmHg)
Propensity score
.1
.2
.3
135
140
145
Systolic blood pressure (mmHg)

Figure 3.7: Comparison of daily energy intake at cutoffs, Obs=7,096

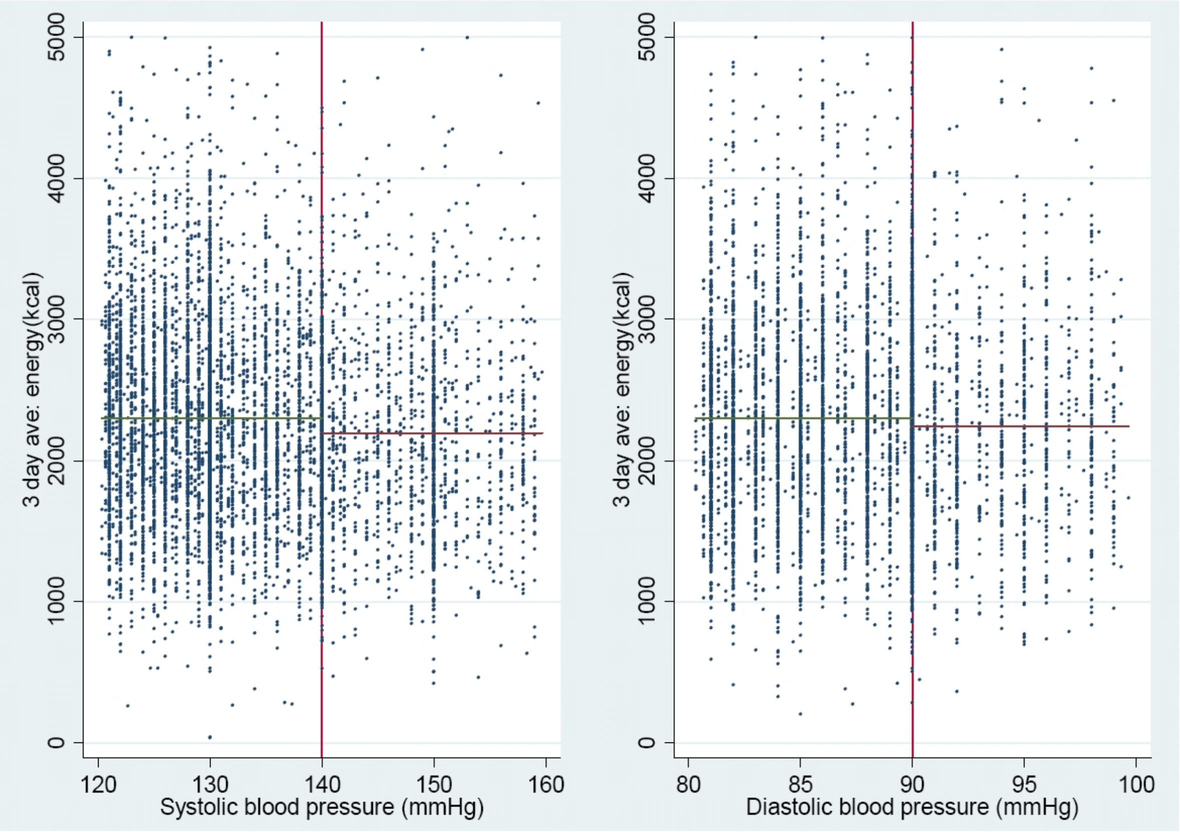


Figure 3.8: Comparison of daily fat intake at cutoffs, Obs=7,096

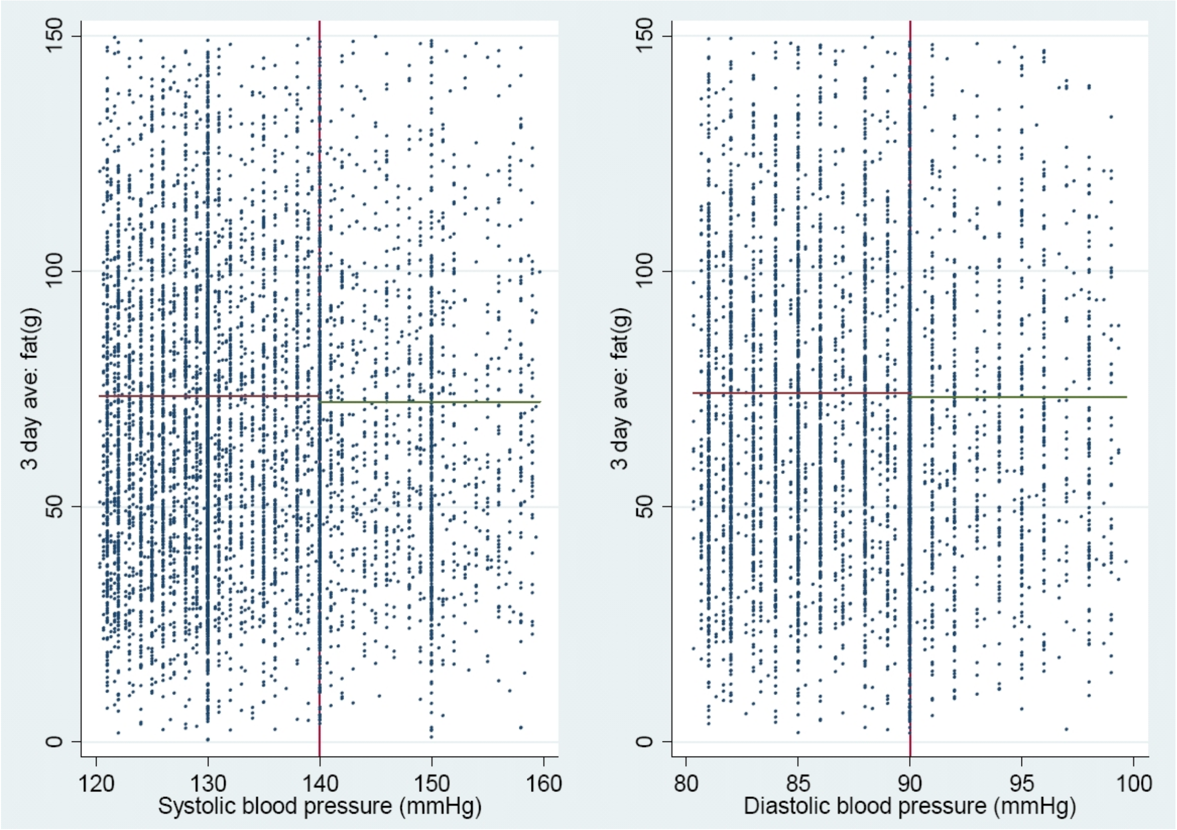


Figure 3.9: Comparison of daily carbohydrate intake at cutoffs, Obs=7,096

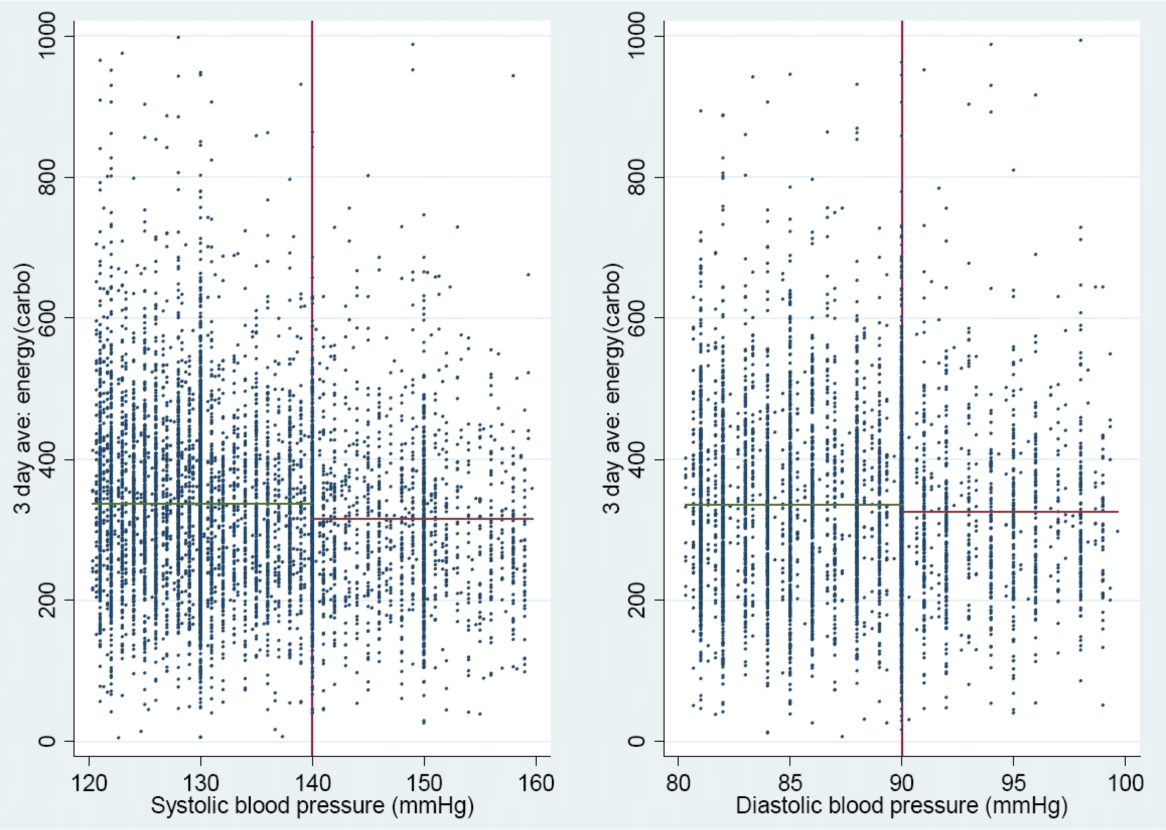


Figure 3.10: Comparison of daily protein intake at cutoffs, Obs=7,096

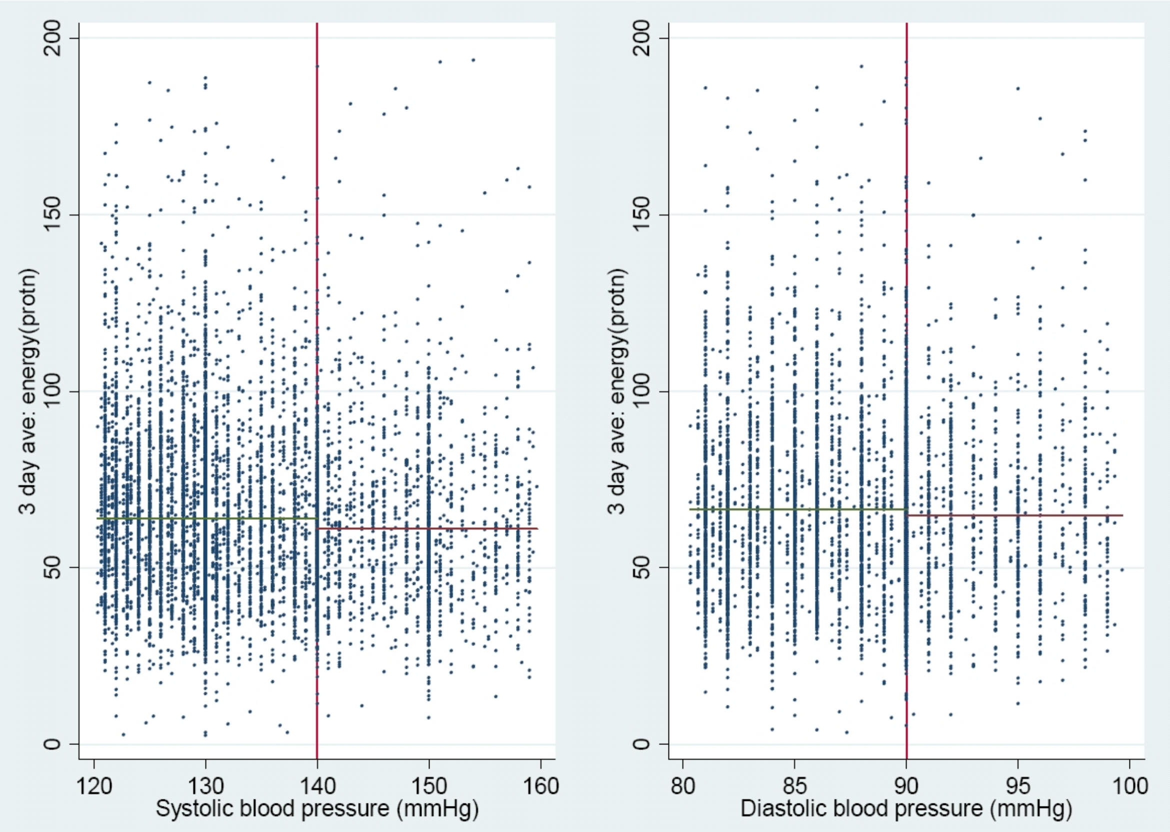


Figure 3.11: Comparison of daily rice consumption at cutoffs, Obs=7,096

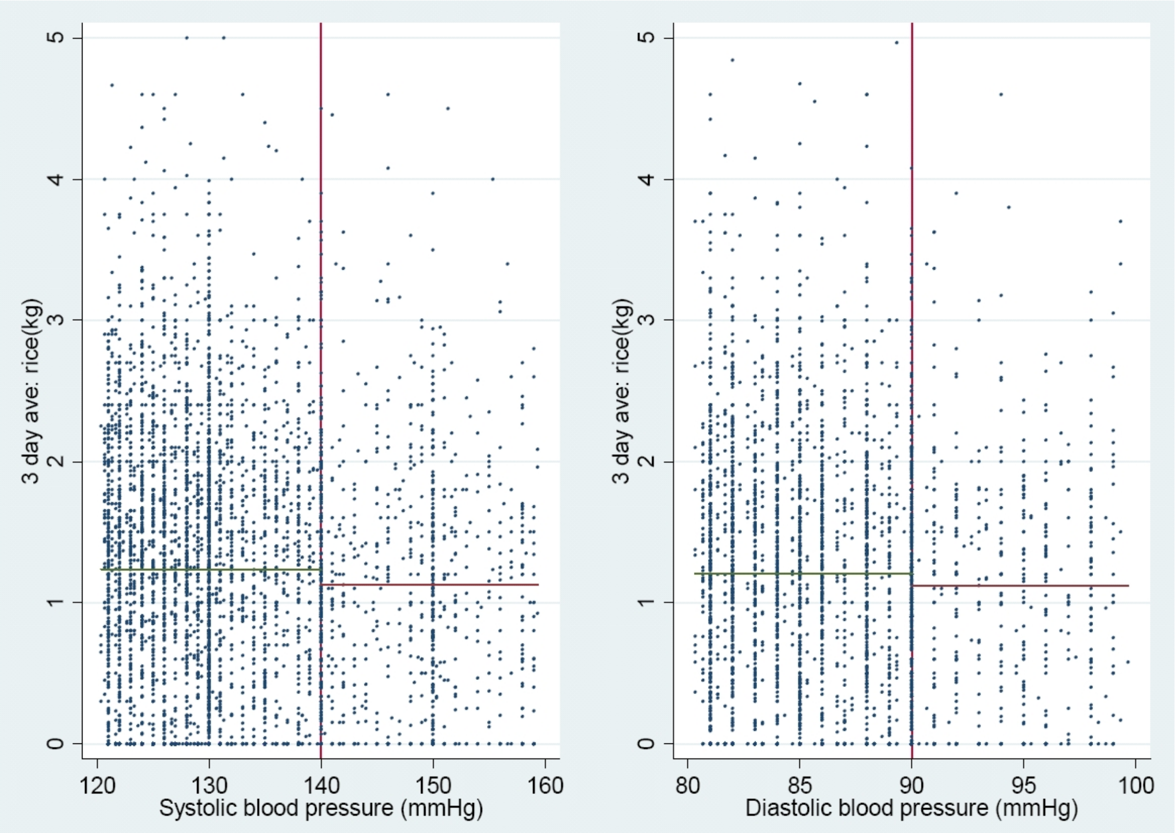


Figure 3.12: Comparison of daily wheat consumption at cutoffs, Obs=7,096

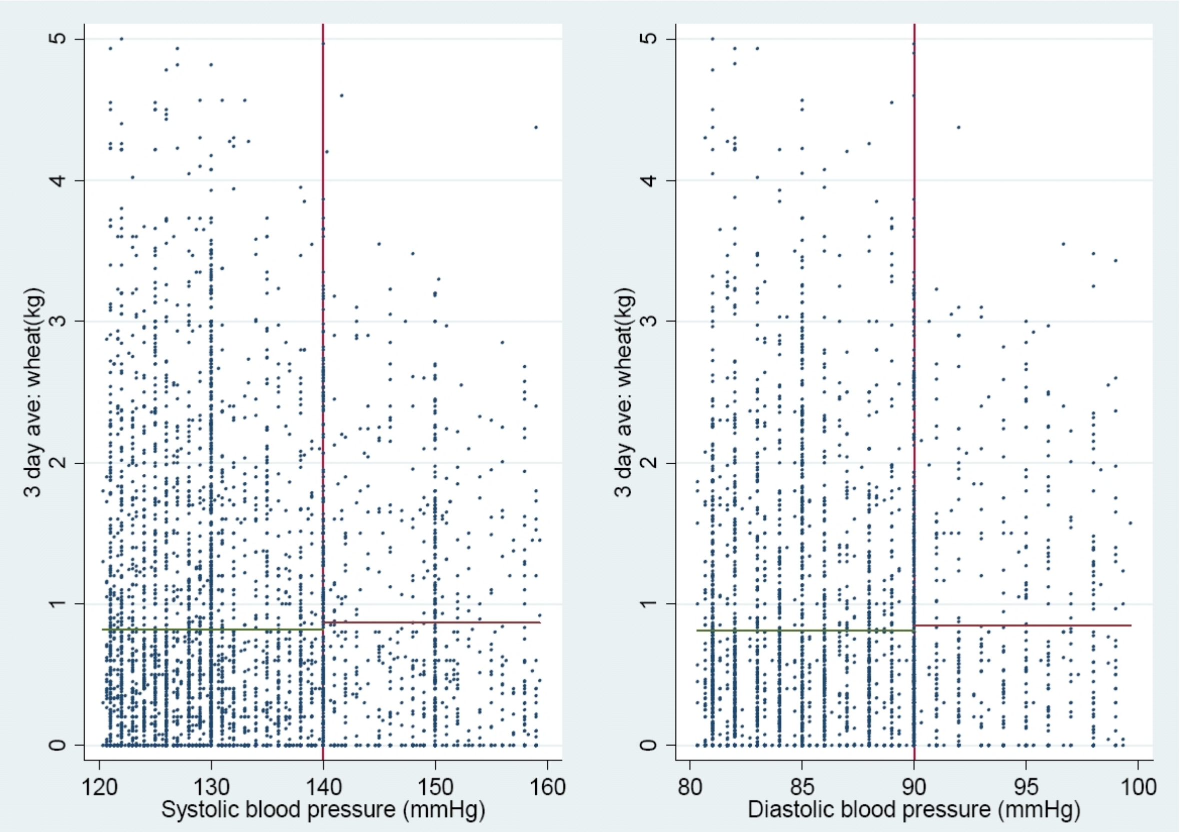


Figure 3.13: Comparison of daily pork consumption at cutoffs, Obs=7,096

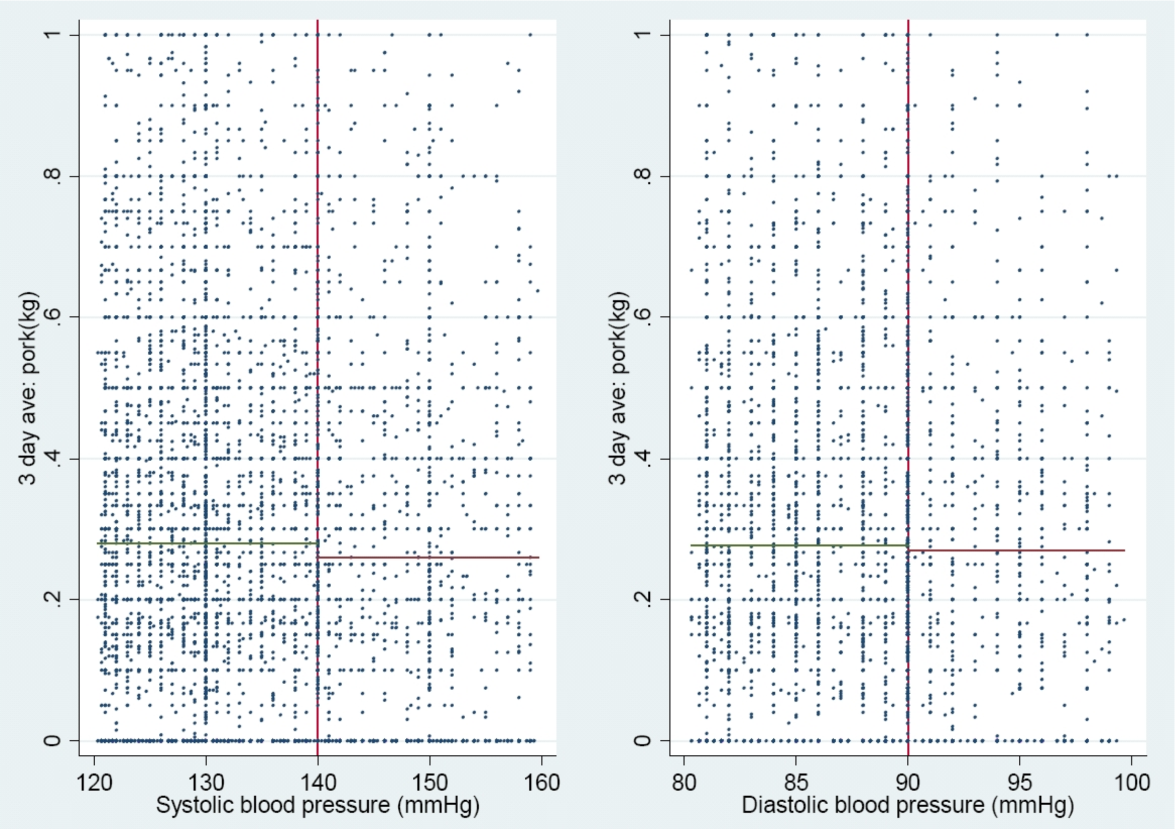


Figure 3.14: Comparison of daily beef/mutton consumption at cutoffs, Obs=7,096

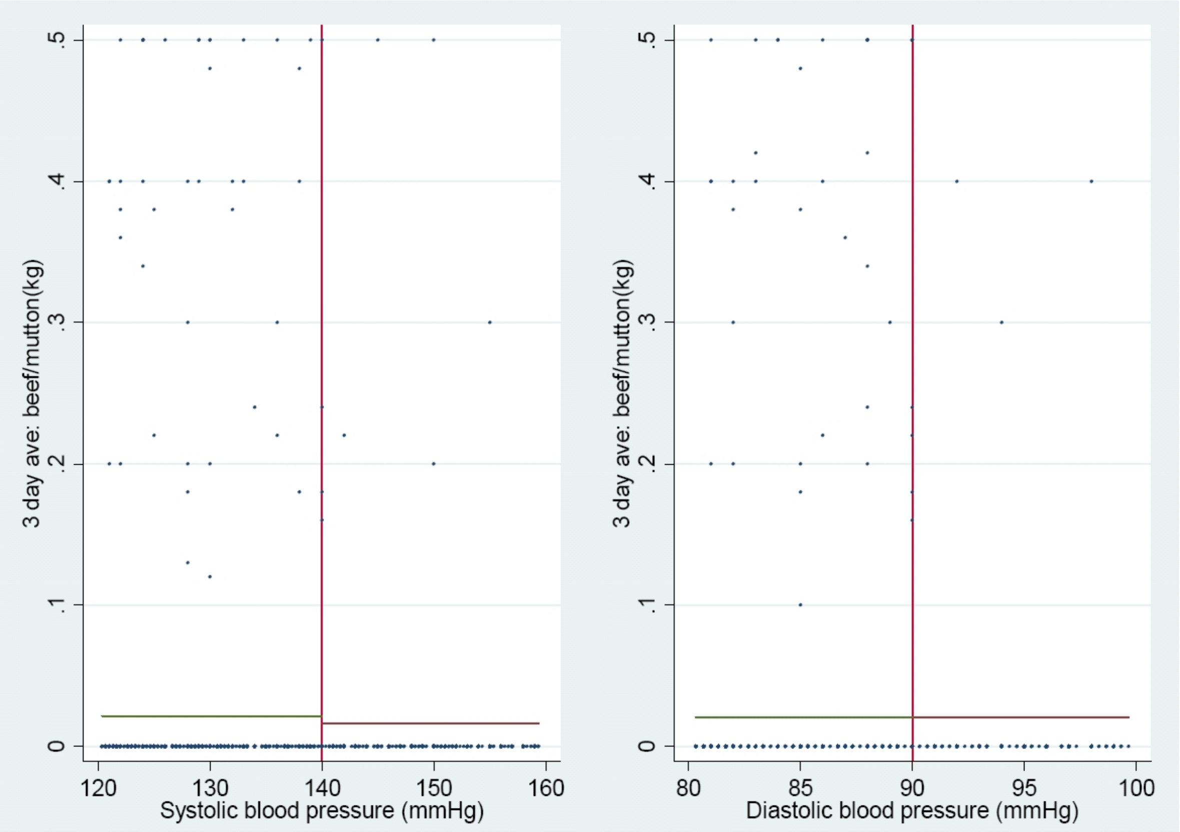


Figure 3.15: Comparison of daily vegetable consumption at cutoffs, Obs=7,096

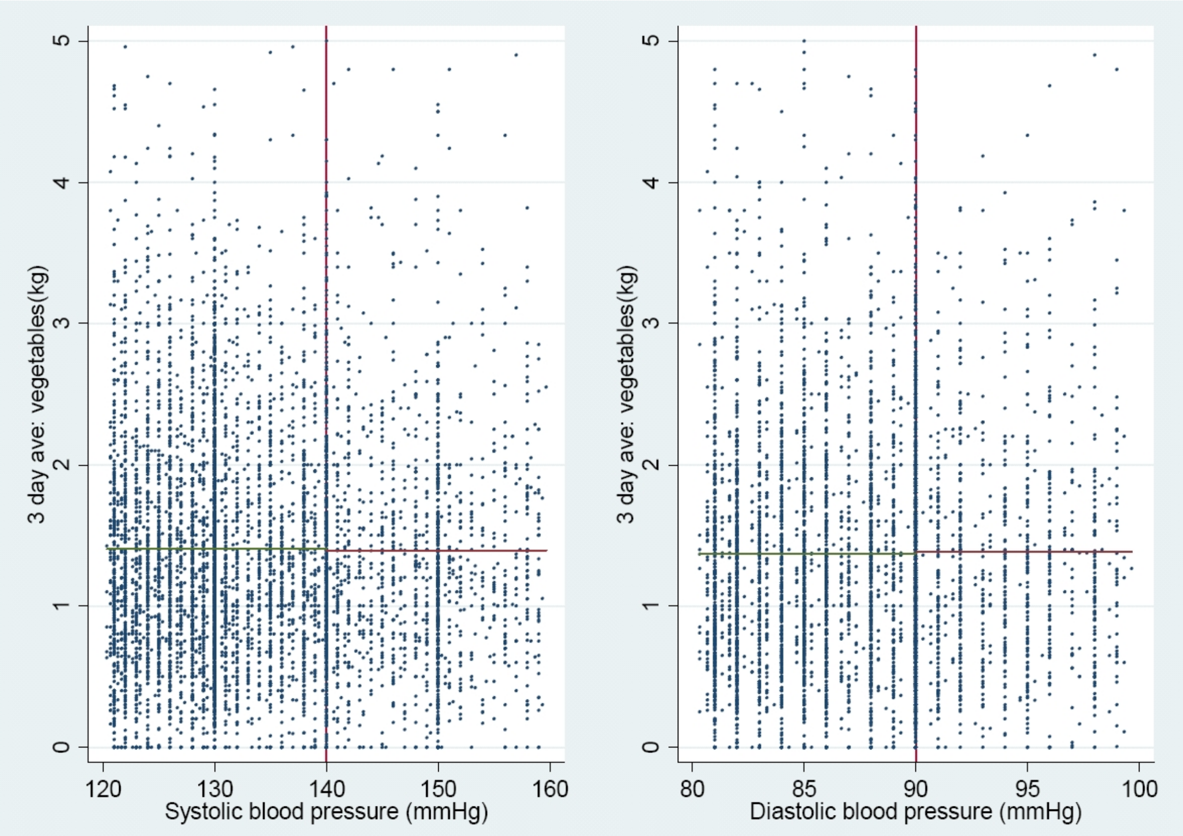


Figure 3.16: Comparison of daily animal oil consumption at cutoffs, Obs=7,096

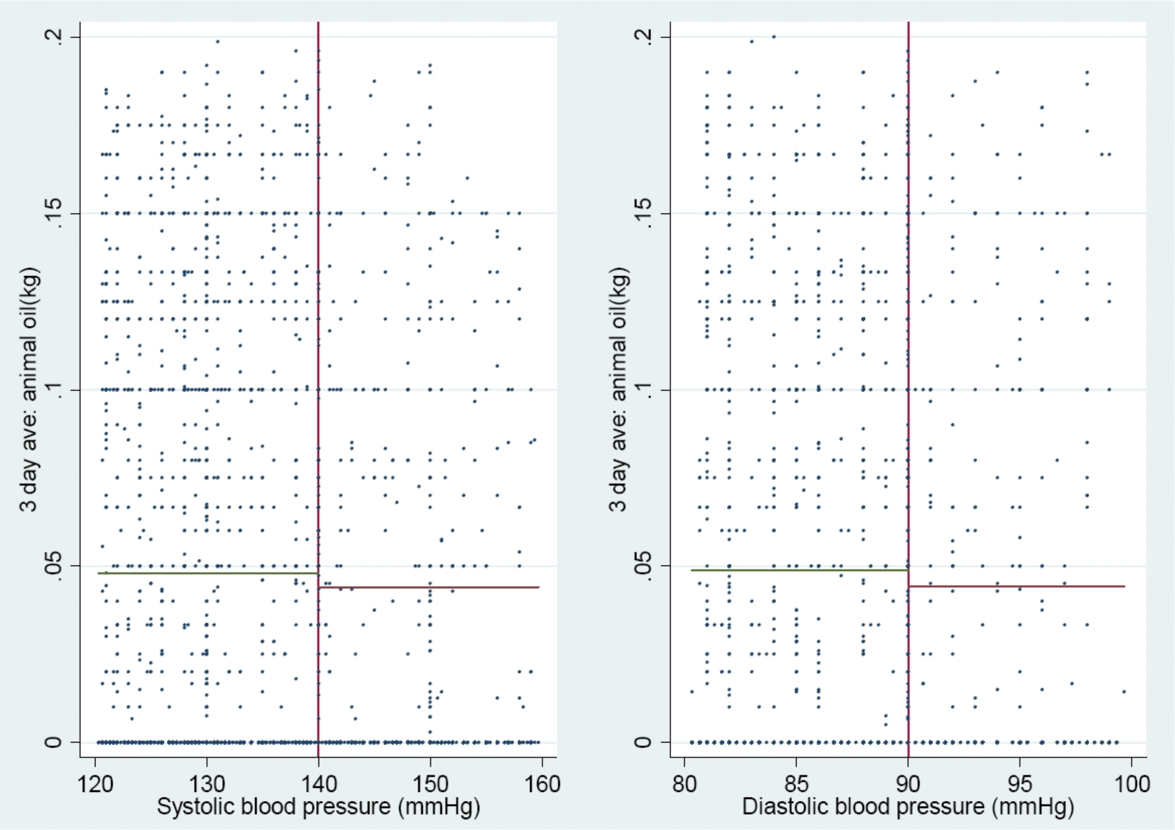
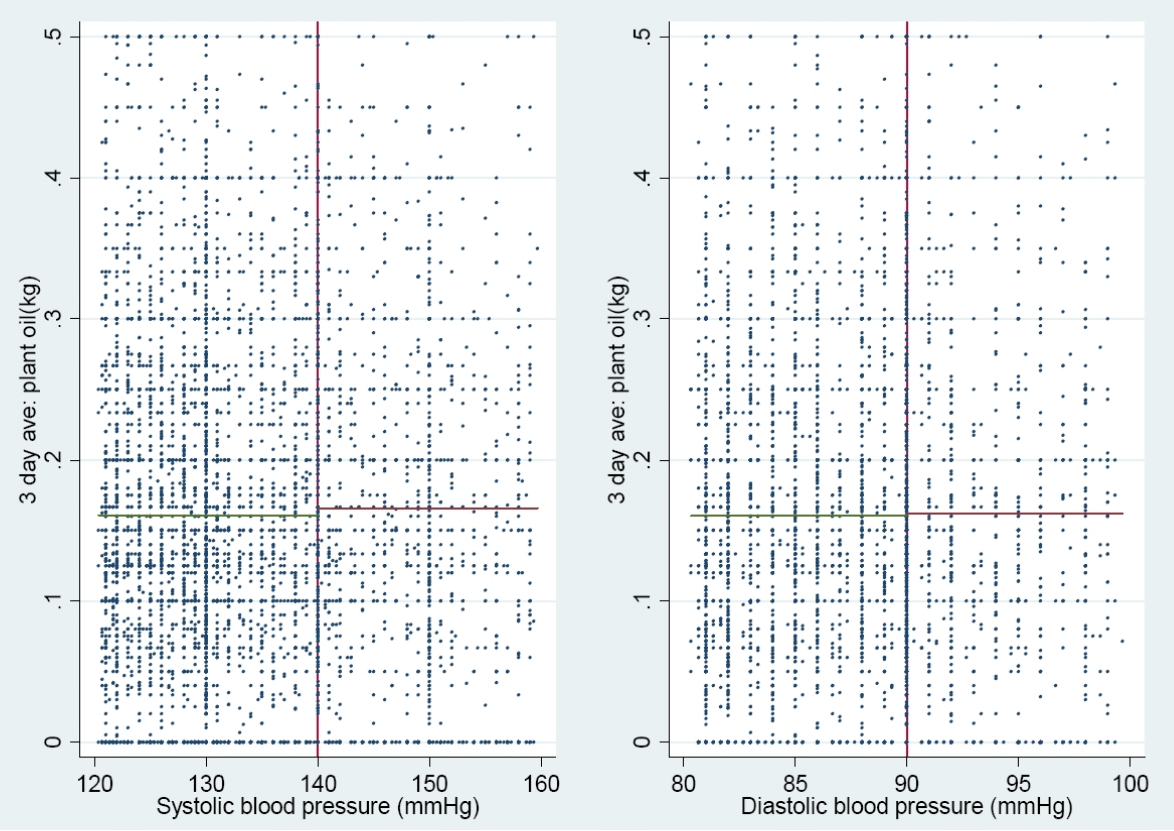


Figure 3.17: Comparison of daily plant oil consumption at cutoffs, Obs=7,096



Chapter 4. Conclusions

This dissertation explores the interrelationships among health, education, health perceptions, and behavioral choices such as smoking and food consumption. Consumers' investments in health and education are important determinants of their personal wellbeing. They are also of critical importance for governments because they can determine a country's overall productivity and its economic burden of chronic diseases. The behavioral choices studied in this dissertation, schooling, smoking and food choices, all have large impacts on one's stock of human capital. Considered as inputs that can enhance human capital, they are chosen based on their benefits, compared to the associated costs. It is of critical importance for policy makers to understand how consumers make decisions regarding these choices in order to design policies that can persuade consumers to make choices in a direction that enhances their human capital.

The second chapter of this dissertation investigated the decision making process regarding education and smoking. It analyzed the impact of teenage smoking on educational outcomes. It found that regular smoking can significantly reduce educational attainment and achievement. This may be due to the biological damage to cognitive ability caused by constant nicotine exposure. Regular smoking can also affect schooling in many other ways. For example, it can discourage students from attending school, where smoking is strictly forbidden. For regular teenage smokers, the incentive to attend school may be particularly low in a developing country like China, where youth smoking is treated differently inside and outside of school because there is no law that specifies the legal age of smoking. Smoking can also affect educational outcomes simply because smokers may devote less effort and time to studying.

Using two rich datasets from China, this dissertation finds that smoking one cigarette per day during adolescence can reduce test scores on both Chinese and math by about 0.1 standard deviations. It can also reduce the total years of schooling by about 5 days. Because education is an important determinant of personal income, the negative effects of smoking on educational outcomes can reduce one's lifetime income. This study predicts that smoking 1-5 cigarettes per day can reduce lifetime income by 0.2-0.8%. These findings update the estimate of the real cost of smoking and suggest that the current cost of smoking perceived by many people is downward biased.

The updated estimate of the real cost of smoking has important policy implications. If correct information about the cost of smoking is provided, the demand for cigarettes, especially by teenage consumers, may be reduced. It also suggests that the reduction in the aggregate level of human capital due to smoking is larger than policy makers consider, which implies that the benefit of policies or interventions reducing smoking is actually much greater than previously thought.

The third chapter of this dissertation focuses on an important risk factor of chronic diseases - hypertension. Chronic diseases are now the major causes of global deaths, in both developed countries and developing countries. They impose enormous economic and health burdens on private households and governments. To a large extent, chronic diseases are affected by one's lifestyle, e.g. choices of diet, smoking and physical activity. If information can be provided to persuade consumers to choose a healthier lifestyle, 80% of chronic diseases can be prevented. Thus, this dissertation investigates the effect of providing information on consumers' food choices.

Unlike the traditional literature, which studies the effect of publicly available information on what constitutes a healthy diet, this study examines the effect on food choices of providing information on personal chronic health status, as measured by hypertension status. Using a regression-discontinuity approach, the study investigates the effect of receiving a diagnosis of hypertension on consumers' nutrient intakes and food demand. Based on rich longitudinal data from China with a large sample size, the study finds that a large proportion of hypertensive people, approximately 75%, are not aware of their hypertension status. After being diagnosed with hypertension, they cut back their consumption of pork and animal oil and increase that of wheat products. Overall, their daily fat intake decreases significantly.

These findings may have important policy implications. Since chronic health status is usually asymptomatic, many consumers are not aware of their chronic health problem and, therefore, may make mistakes in evaluating the benefits and costs of food choices. Policies or interventions that can help consumers keep better track of their true chronic health status can correct consumers' misperceptions and help them make optimal choices. Moreover, when people are better informed of their health status, they may also absorb publicly available health information more efficiently. Lastly, health insurance policies should also be designed in the way that promotes preventive care and health status monitoring.

References

[1] Abdalla, C.W., B. A. Roach, and D. J. Epp, 1992. "Valuing Environmental Quality Change Using Averting Expenditures: An Application to Groundwater Contamination," *Land Economics* 68: 163-169.

[2] Alderman, H., J. R. Behrman, V. Lavy, and R. Menon , 2001. "Child Health and School Enrollment: a Longitudinal Analysis", *Journal of Human Resources* 36: 185-205.

[3] Alston, J. M., J. A. Chalfant, and N. E. Piggott, 1998. "A globally flexible model of the effects of generic advertising of beef and pork on U.S. meat demand,". *American Journal of Agricultural Economics* 80(5): 1174.

[4] Alston, J. M., J. A. Chalfant, and J.S. James, 1999. "Doing well by doing a body good: An evaluation of the industry-funded nutrition education program conducted by the dairy council of California," *Agribusiness* 15(3): 371-392.

[5] Angrist, J. D., and V. Lavy, 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *Quarterly Journal of Economics* 114(2): 533–575.

[6] Aubert, P. L., J. P. Bovet, A. Gervasoni, B. W. Rwedbogora, and F. Paccaud, 1998. "Knowledge, attitudes, and practices in a country in epidemiological transition, hypertension," *Journal of American Heart Association* 31: 1136-1145.

[7] Auld, Christopher M., 2005. "Smoking, Drinking, and Income", *Journal of Human Resources* 40(2): 505-518.

[8] Barro, R., 1991. "Economic growth in a cross section of countries," *Quarterly Journal of Economics* 106(2): 407-443.

[9] Becker, G. S, 1964. *Human capital*. New York: Columbia University Press.

[10] Becker, G. S., 1965. "A Theory of the Allocation of Time," *Economic Journal* 965(75): 493-517.

[11] Becker, G. S., M. Grossman, and K. M. Murphy, 1994. "An Empirical Analysis of Cigarette Addiction", *The American Economic Review* 84(3): 396-418.

[12] Becker, G. S., and Kevin M. Murphy, 1988. "A theory of rational addiction", *Journal of Political Economy* 96(4): 675-700.

[13] Black, S., 1999. "Do better schools matter? Parental valuation of elementary education", *Quarterly Journal of Economics* 114(2): 577–599.

[14] Blundell, R., and R. J. Smith, 1994. "Coherency and estimation in simultaneous models with censored or qualitative dependent variables," *Journal of Econometrics* 64: 355-373.

[15] Bowen, D. J, S. E. Eury, and N. E. Grunberg, 1986. "Nicotine's Effects on Female Rats' Body Weight: Caloric Intake and Physical Activity", *Pharmacology, Biochemistry and Behavior* 25: 1131–1136.

[16] Brown, J. D., 1969. "Effect of a health hazard “scare” on consumer demand," *American Journal of Agricultural Economics* 51: 676-678.

[17] Brown, J. D., and L. F. Schrader, 1990. "Cholesterol information and shell eggs consumption," *American Journal of Agricultural Economics* 72: 548-555.

[18] Buse, A., 1994. "Evaluating the linearize almost ideal demand system," *American Journal of Agricultural Economics* 76(4): 781-793.

[19] Campbell, D. T., and J. C. Stanley, 1963. “Experimental and quasi-experimental designs for research on teaching,” in N. L. Gage, ed., *Handbook of Research on Teaching*, Chicago: Rand McNally.

[20] Capps, O., J. D. Schmitz, 1991. "A recognition of health and nutrition factors in food demand analysis," *Western Journal of Agricultural Economics* 16(1): 21-35.

[21] Chaloupka, F. J., T. Hu, K. E. Warner, R. Jacobs, and A. Yurekli, 2000. *Tobacco Control in Developing Countries*, Oxford University Press, Oxford.

[22] Chern, W. S., E. T. Loehman, and S. T. Yen, 1995. "Information, health risk beliefs, and the demand for fats and oils," *The Review of Economics and Statistics* 77: 555-564.

[23] Chern, W. S., and K. Rickertsen, 2003. *Health, Nutrition and Food Demand*, MA: CABI Publishing.

[24] Chern, W. S., and J. Zuo, 1995. "Alternative measures of changing consumer information on fat and cholesterol," paper presented at the AAEA annual meeting, Indianapolis.

[25] Chow, G., 1984. *The Chinese Economy*, Harper and Row, New York.

[26] Counotte, D. S., S. Spijker, L. H. Van de Burgwal, et al., 2009. "Long-lasting cognitive deficits resulting from adolescent nicotine exposure in rats", *Neuropsychopharmacology* 34: p299–306.

[27] Crutchfield, S., F. Kuchler and J. N. Variyam, 2001. "The economic benefits of nutrition labeling: a case study for fresh meat and poultry products," *Journal of Consumer Policy* 24(2): 183-207.

[28] Deaton, A., 1988. "Quality, quantity, and spatial variation of price," *The American Economic Review* 78(3): 418-430.

[29] Deaton, A., and J. Muellbauer, 1980. "An Almost Ideal Demand System," *The American Economic Review* 70(3): 312-326.

[30] DeCicca, P, D. Kenkel, and A. Mathios, 2002. "Putting Out the Fires: Will Higher Taxes Reduce the Onset of Youth Smoking?", *Journal of Political Economy* 110(1): 144-169.

[31] Duffy, M, 1999. *Advertising in Consumer Allocation Models: Choice of Functional Form*. Report No. 9909, Manchester, UK: Manchester School of Management, University of Science and Technology.

[32] Duflo, E., 2001. “Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment,” *American Economic Review* 91(4): 795-814.

[33] Ernst,.M., S. J. Heishman, L. Spurgeon, and E. D. London, 2001. "Smoking History and Nicotine Effects on Cognitive Performance", *Neuropsychopharmacology* 25(3): 313-319.

[34] Fan, S., G. L. Cramer, and E. J. Wales, 1994. "Food demand in rural China: evidence from household survey," *Agricultural Economics* 11(1): 61-69.

[35] Fan, S., E. J. Wales, and G. L. Cramer, 1995. "Food demand in rural China: evidence from household survey," *American Journal of Agricultural Economics* 77(1): 54-62.

[36] Fang, C., and J. Beghin, 2002. "Urban Demand for Edible Oils and Fats in China. Evidence from Household Survey Data," *Journal of Comparative Economics* 30(4): 732-753.

[37] Farrell, P., and V. Fuchs, 1982. "Schooling and Health: the Cigarette Connection", *Journal of Health Economics* 1: 217-230.

[38] Fleisher, B., and X. Wang, 2004. "Skill Differentials, Return to Schooling, and Market Segmentation in a Transition Economy: the Case of Mainland China", *Journal of Development Economics* 73: 315-328

[39] Forster, J. L., M. Wolfson, D. M. Murray, et al., 1997. "Perceived and measured availability of tobacco to youths in 14 Minnesota communities: the TPOP Study," *American Journal of Preventive Medicine* 13: 167-74.

[40] Foulds, J., J. Stapleton, J. Swettenham, N. Bell, K. McSorley, and M. Russell, 1996. "Cognitive Performance Effects of Subcutaneous Nicotine in Smokers and Never-smokers", *Psychopharmacology* 127: 31-38.

[41] Gale, F., P. Tang, X. Bai, and H. Xu, 2005. *Commercialization of food consumption in rural China*, USDA ERS Report.

[42] Gan, Q., K. R. Smith, and S. K. Hammond, 2000. "Disease Burden of Adult Lung Cancer and Ischaemic Heart Disease from Passive Tobacco Smoking in China", *Tobacco Control* 16: 417-422.

[43] Gao, X. M., E. J. Wailes, and G. L. Cramer, 1996a. "A two-stage rural household demand analysis: microdata evidence from Jiangsu province, China, American," *American Journal of Agricultural Economics* 78: 604-613.

[44] Gao, X. M., E. J. Wailes, and G. L. Cramer, 1996b. "Partial rationing and Chinese urban household food demand analysis," *Journal of Comparative Economics* 22: 43-62.

[45] Gilbert, A. R., C. Pinget, P. Bovet, J. Cornuz, C. Shamlaye, and F. Paccaud, 2004. "The Cost Effectiveness of Pharmacological Smoking Cessation Therapies in Developing Countries: a Case Study in the Seychelles", *Tobacco Control* 13: 190-195.

[46] Gilbert, S., 2007. *Scientific Consensus Statement on Environmental Agents Associated with Neurodevelopmental Disorders*, the Collaborative on Health and the Environment's Learning and Developmental Disabilities Initiative.

[47] Glewwe, P. 1999. *The Economics of School Quality Investments in Developing Countries*, St. Martin's Press, New York.

[48] Glewwe, P., and H. Jacoby, 1994. "Student Achievement and Schooling Choice in Low Income Countries: Evidence from Ghana", *Journal of Human Resources* 29(3): 843-864.

[49] Glewwe, P., H. Jacoby, and E. King, 2001. "Early Childhood Nutrition and Academic Achievement: a Longitudinal Analysis", *Journal of Public Economics* 81: 245-368.

[50] Glewwe, P., and M. Kremer, 2006. "Schools, teachers and education outcomes in developing countries", IN: *Handbook of the Economics of Education*, edited by E. Hanushek and F. Welch. North Holland.

[51] Glewwe, P., and E. A. Miguel, 2008. "The impact of child health and nutrition on education in less developed countries", IN: *Handbook of Development Economics*, volume 4, edited by T. Paul Schultz and J. Strauss. New York, NY: Elsevier: 3561-3606.

[52] Gold, M. R., J. E. Siegel, L. B. Russell, and M. C. Weinstein (eds.), 1996. *Cost-Effectiveness in Health and Medicine, New York*: Oxford Press.

[53] Gould, B. W., and H. J. Villarreal, 2006. "An assessment of the current structure of food demand in urban China", *Agricultural Economics* 34: 1-16

[54] Green, R., and J. M. Alston, 1990. "Elasticities in AIDS models", *American Journal of Agricultural Economics* 72: 442-445.

[55] Grossman, M., 1972. "On the concept of health capital and the demand for health", *Journal of Political Economy* 80: 223-255.

[56] Gruber, J., 2001. "Youth Smoking in the 1990's: Why Did it Rise and What Are the Long Run Implications?", *The American Economics Review* 91(2): 85-90.

[57] Grunberg, N. E., D. J. Bowen, and D. E. Morse, 1984. "Effects of Nicotine on Body Weight and Food Consumption of Rats", *Psychopharmacology* 83: 93–98.

[58] Hahn, J., P. Todd, and W. V. Klaauw, 2001. “Regression discontinuity", *Econometrica* 69: 201-209.

[59] Halbrendt, C., F. Tuan, C. Gempesaw, and D. Dolk-Etz, 1994. "Rural Chinese food consumption: the case of Guangdong", *American Journal of Agricultural Economics* 76: 794-799.

[60] Hanushek, E., 1995. “Interpreting recent research on schooling in developing countries”, *World Bank Research Observer* 10(2): 227-246.

[61] Harbison, R., and E. Hanushek, 1992. *Educational Performance of the Poor: Lessons from Rural Northeast Brazil*, Oxford University Press.

[62] Heckman, J. J., 2003. "China’s Investment of Human Capital", *Economic Development and Cultural Change* 51(4): 795-804.

[63] Heckman, J. J., and X. Li, 2004. "Selection Bias, Comparative Advantage and Heterogeneous Returns to Education: Evidence from China in 2000", *Pacific Economic Review* 9(3): 155-171.

[64] Heishman, S. J., R. C. Taylor, and J. E. Henningfield, 1994. "Nicotine and smoking: A Review of Effects on Human Performance", *Experimental and Clinical Psychopharmacology* 2(4): 345-95.

[65] Houthakker, H. S., 1957. "An international comparison of household expenditure patterns, commemorating the centenary of Engel’s Law," *Econometrica* 25(4): 532-50.

[66] Huang, J., and H. Bouis, 1996. "Structural changes in the demand for food in Asia", IFPRI 2020 Vision Discussion Series No.11.

[67] Huang, J., and S. Rozelle, 1998. "Market development and food demand in rural China," *China Economic Review* 9(1): 25.

[68] Ibbotson, R., and R. Sinquefeld, 1976. "Stocks, bonds, bills, and inflation: year-by-year historical returns (1926-1974)," *Journal of Business* 49: 11-47.

[69] Imbens, G., and K. Kalyanaraman, 2009. “Optimal bandwidth choice for the regression discontinuity estimator" *NBER Working Paper Series* No.14726.

[70] Imbens, G., and T. Lemieux, 2008. “Regression Discontinuity Designs: A Guide to Practice,” *Journal of Econometrics* 142(2): 615–635.

[71] Jacobsen, L. K., J. H. Krystal, W. E. Mencl, M. Westerveld, S. J. Frost, and K. R. Pugh, 2005. "Effects of Smoking and Smoking Abstinence on Cognition in Adolescent Tobacco Smokers", *Biological Psychiatry* 57(1): 56-66.

[72] Johnston, D. W., C. Propperb, and M. A. Shieldse, 2009. "Comparing subjective and objective measures of health: Evidence from hypertension for the income/health gradient," *Journal of Health Economics*, 28(3): 540-552.

[73] Jones, S. E., D. J. Sharp, C. G. Husten, and L. S. Crossett, 2002. "Cigarette acquisition and proof of age among US high school students who smoke," *Tobacco Control* 11: 20-25.

[74] Jorenby, D. E., D. K. Hatsukami, S. S. Smith, M. C. Fiore, S. Allen, J. Jensen, and T. B. Baker, 1996. "Characterization of Tobacco Withdrawal Symptoms: Transdermal Nicotine Reduces Hunger and Weight Gain", *Psychopharmacology* 128: 130–138.

[75] Kaabia, B. M., A. M. Angulo, and J. M. Gil, 2001. "Health information and the demand for meat in Spain," *European Review Agricultural Economics* 28(4): 499-517.

[76] Kannel, W. B., 2000. "Elevated systolic blood pressure as a cardiovascular risk factor - principal results," *The American Journal of Cardiology* 85(2): 251-255.

[77] Kim, S., W. S. Chern, 1999. "Alternative measures of health information and demand for fats and oils in Japan," *Journal of Consumer Affair* 33: 92-109.

[78] Kinnucan, H. W., H. Xiao, C. J. Hsia, and J. D. Jackson, 1997. "Effect of health information and generic advertising on U.S. meat demand," *American Journal of Agricultural Economics* 79: 13-23.

[79] Krueger, A. B., and M. Lindahl, 2001. "Education for growth: why and for whom?" *Journal of Economic Literature* 39(4): 1101-1136.

[80] Kueh, Y. Y., 1988. "Food consumption and peasant incomes in the post-Mao era," *The China Quarterly* 116: 634–70.

[81] Lee, D., and T. Lemieux, 2009. “Regression discontinuity designs in economics," *NBER Working Paper Series No.*14723.

[82] Leistikow, B., N., D. C. Martin, and C. E. Milano, 2000a. "Estimates of Smoking-attributable Deaths at Ages 15-54: Motherless or Fatherless Youths, and Resulting Social Security Costs in the United States in 1994", *Preventive Medicine* 30(5): 353-360.

[83] Leistikow, B. N., D. C. Martin, and C. E. Milano, 2000b. "Fire Injuries, Disasters, and Costs from Cigarettes and Cigarette Lights: a Global Overview", *Prevention* 31(2): 91-99.

[84] Lewbel, A., 1996. "Demand estimation with expenditure measurement errors on the left and right hand side," T*he Review of Economics and Statistics* 78(4): 718-725.

[85] Lewis, P., and N. Andrews, 1989. "Household demand in China," *Applied Economics* 21(6): 793-807.

[86] Liu, B.Q., R. Peto, and Z. M. Chen., 1998. "Emerging tobacco hazards in China: Retrospective Proportional Mortality Study of One Million Deaths", *British Medical Journal* 317: 1411–22.

[87] Lockheed, M., and A. Verspoor, 1991. *Improving Primary Education in Developing Countries*, Oxford University Press, New York.

[88] Lucas, R., 1988. "On the mechanics of economic development," *Journal of Monetary Economics* 22: 3-42.

[89] Maddala, G. S., 1983. *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

[90] Mankiw, N. G., D. Romer, and D. N. Weil, 1992. "A contribution to the empirics of economic growth," *Quarterly Journal of Economics* 107(2): 407-437.

[91] Marks, J. S., J. P. Koplan, C. H. Hogue, and M. E. Dalmat, 1990. "A cost-benefit /cost-effectiveness analysis of smoking cessation for pregnant women", *American Journal of Preventive Medicine* 6: 282-289.

[92] Martin, H. D., T. L. Mader, and M. Pedersen, 1994. "Influencing diet and health through project LEAN," *Journal of Nutrition Education* 26(4): 191-194.

[93] Mathios, A. D., and P. Ippolito, 1999. "Health claims in food advertising and labeling-disseminating nutrition information to consumers," in *America's Eating Habits: Changes and Consequences*, edited by E. Frazao, chap. 11, p189-212.

[94] Mendez, M. A., and B. M. Popkin, 2004. "Globalization, urbanization and nutritional change in the developing world," electronic *Journal of Agricultural and development Economics* 1(2): 220-241.

[95] Meyerhoefer, C. D., C. K. Ranney, and D. E. Sahn, 2005. "Consistent estimation of censored demand systems using panel data," *American Journal of Agricultural Economics* 87(3): 660-672.

[96] Miguel, E., G. Bobonis, and C. Sharma, 2006. "Iron Deficiency Anemia and School Participation", *Journal of Human Resources* 41(4): 692-721.

[97] Miguel, E., and M. Kremer, 2004, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities", *Econometrica* 72(1): 159-217.

[98] Miller, V. P., C. R. James, C. Ernst, and F. Collin, 1997. *Smoking Attributable Medical Care Costs: Models and Results*. Berkeley, California: Berkeley Economic Research Associates.

[99] Ministry of Education of the People's Republic of China (MOE), 2003. *Report of Education Statistics* 1(26), Department of Planning, Beijing.

[100] Ministries of Health and Science and Technology (MOHST) and the National Bureau of Statistics of the Peoples Republic of China, 2004. *The Nutrition and Health Status of the Chinese People*, State Information Office, Beijing.

[101] Mudarri, D., 1994. *The Costs and Benefits of Smoking Restrictions: An Assessment of the Smoke-free Environment Act of 1993 (H. R. 3434)*, Washington, D.C., Indoor Air Division, U.S. Environmental Protection Agency.

[102] Murphy, K. M., and R. H. Topel, 1985. "Estimation and inference in two-step econometric models," *Journal of Political Economy* 95: 730-739.

[103] Perreira, K. M., K. and E. Cortes, 2006. "Race/ethnicity and nativity differences in alcohol and tobacco use during pregnancy," *American Journal of Public Health* 96(9): 1629-36.

[104] Peterson, E., F. Wesley, L. Jin, and S. Ito, 1991. "An Econometric analysis of rice consumption in the People's Republic of China," *Agricultural Economics* 6(1): 67-78.

[105] Pineda, J. A., C. Herrera, C. Kang, and A. Sandler, 1998. "Effects of Cigarette Smoking and 12-h Abstention on Working Memory during a Serial-probe Recognition Task", *Psychopharmacology* 139: 311-321.

[106] Pollack, H., P. M. Lantz, and J. G. Frohna, 2000. "Maternal smoking and adverse birth outcomes among singletons and twins," *American Journal of Public Health* 90(3): 395-400.

[107] Pollak, R.A., and T. J. Wales, 1981. "Demographic variable in demand analysis," *Econometrica* 49(6): 1533-1551.

[108] Pollak, R.A., and T. J. Wales, 1992. *Demand System Specification and Estimation*. New York: Oxford University Press.

[109] Psacharopoulos, G., 1985. "Returns to Education: A Further International Update and Implications," *Journal of Human Resources* 20(4): 583-604.

[110] Psacharopoulos, G., 1994. "Returns to Investment in Education: A Global Update," *World Development* 22: 1325-1344.

[111] Richards, G. A., A. P. Terblanche, A. J. Theron, et al., 1996. "Health effects of passive smoking in adolescent children," *South African Medical Journal* 86 (2): 143–7.

[112] Ried, W., 1998. "Comparative dynamic analysis of the full Grossman model," *Journal of Health Economics* 17: 383-425.

[113] Rivers, D., and Q. H. Vuong, 1988. "Limited Information Estimators and Exogeneity Tests," *Journal of Econometrics* 39: 347-366.

[114] Roosen, J., S. Marette, S. Blanchemanche, and P. Verger, 2009. "Does health information matter for modifying consumption? A field experiment measuring the impact of risk information on fish consumption," *Review of Agricultural Economics* 31(1): 2-20.

[115] Ruttan G. H., R. H. McDonald, and L. H. Kuller, 1989. "A historical perspective of elevated systolic vs diastolic blood pressure from an epidemiological and clinical trial viewpoint," *Journal of Clinical Epidemiology*, 42 (7): 663-673.

[116] Shonkwiler, J. S., and S. T. Yen, 1999. "Two-step estimation of a censored system of equations," *American Journal of Business and Economic Statistics* 3: 370-9.

[117] Silverman, B. W., 1986. *Density Estimation*, London: Chapman and Hall.

[118] Sloan, F., J. Ostermann, G. Picone, C. Conover, and D. Taylor, 2004. *The Price of Smoking*, the MIT Press, Cambridge, Massachusetts.

[119] Stokey, N., R. Lucas, and E. Prescott, 1989. *Recursive Methods in Economic. Dynamics,* Harvard University Press, Cambridge Massachusetts.

[120] Strauss, J., 1986. "Does better nutrition raise farm productivity?," *The Journal of Political Economy* 94(2): 297-320.

[121] Strong, K., C. Mathers, S. Leeder, R. Beaglehole, 2005. "Preventing chronic diseases: how many lives can we save",.*Lancet* Vol 366(9496), p1578-1582.

[122] Sung, H-Y., L. Wang, S. Jin, T. Hu, and Y. Jiang, 2006. "Economic Burden of Smoking in China, 2000", *Tobacco Control* 15(Supplement I): i5–i11.

[123] The Global Youth Tobacco Survey Collaborative Group, 2002. "Tobacco use among youth: a cross country comparison," *Tobacco Control* 11: 252-270.

[124] Thistlethwaite, D. L., and D. T. Campbell, 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment", *Journal of Educational Psychology* 51: 309–317.

[125] Trochim, W. M. K., 1984. *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, Sage Publications, Beverly Hills, CA.

[126] UNDP, 2003. *Human Development Report*, United Nations Development Program, New York.

[127] U. S. Department of Health and Human Services, 1994. *Preventing Tobacco Use among Young People: A Report of the Surgeon General*, Washington, DC, U.S. Government Printing Office.

[128] U. S. Department of Health and Human Services, 2006. *The Health Consequences of Involuntary Exposure to Tobacco Smoke, A Report of Surgeon General*, Washington, DC, U.S. Government Printing Office.

[129] Van der Klaauw, W., 2002. "Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach", *International Economic Review* 43 (4): 1249–1287.

[130] Viscusi, K., and J. Hersch, 2001. "Cigarette Smokers Are Job Risk Takers," *Review of Economics and Statistics* 83: 269-280.

[131] Wang, L., L. Kong, F. Wu, Y. Bai, and R. Burton, 2005. "Preventing chronic diseases in China", *Lancet* 366(9499): p1821-1824.

[132] Wang, Z., and W. S. Chern, 1992. "Effects of rationing on the consumption behavior of Chinese urban households during 1981-1987," *Journal of Comparative Economics* 16: 1-26.

[133] Strong, K., C. Mathers, S. Leeder, and R. Beaglehole, 2005. "Preventing chronic diseases: how many lives can we save?" Lancet 366: 1821-24.

[134] Wolf-Maier, K., R. S. Cooper, H. Kramer, J. R. Banegas, S. Giampaoli, M. R. Joffres, N. Poulter, P. Primatesta, B. Stegmayr, and M. Thamm, 2004. "Hypertension treatment and control in five European countries, Cananda and the United States," *Hypertension* 43: 10-17.

[135] Wooldridge, J. M., 2001. *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, Massachusetts.

[136] Wooldridge, J. M., 2009. *Introductory Econometrics: A Modern Approach*, South-Western, a part of Cengage Learning.

[137] World Bank, 1985. *China: Long-term Development Issues and Options*, Johns Hopkins University Press, Baltimore.

[138] World Bank, 2001. *World Development Report 2000/2001: Attacking Poverty*, Washington, DC.

[139] World Bank, 2004. *World Development Report 2004: Making Services Work for Poor People*, Washington, DC.

[140] World Health Organization, 1997. *Tobacco or Health: A Global Status Report*, WHO, Geneva.

[141] World Health Organization, 2003. *The World Health Report 2003: Shaping the Future*, WHO, Geneva.

[142] World Health Organization, 2005. *Preventing Chronic Diseases: a Vital Investment: WHO Global Report*.

[143] World Health Organization, 2008. *WHO Report on the Global Tobacco Epidemic 2008: the Mpower Package*, WHO, Geneva.

[144] World Health Organization, 2009. *Global Strategy on Diet, Physical Activity and Health,* WHO, Geneva.

[145] World Health Organization, 2009. *World Health Statistics 2009,* WHO, Geneva.

[146] Wu, Y., E. Li, and S. N. Samuel, 1995. "Food consumption in urban China: an empirical analysis," *Applied Economics* 27: 509-515.

[147] Yang, X., 2002. *Food Nutrient Table in China*, Beijing: Peking University Medical Press.

[148] Yen, S. T., and W.S. Chern, 1992. "Flexible demand systems with serially correlated errors: fat and oil consumption in the United States," *American Journal of Agricultural Economics* 74: 689-694..

[149] Yen, S. T., F. Cheng, and S.J. Su, 2004. "Household food demand in urban China: a censored system approach," *Journal of Comparative Economics* 32: 564-585.

[150] Yen, S. T., K. Kan, and S.J. Su, 2002. "Household demand for fats and oils: two-step estimation of a censored demand system," *Applied Economics* 14: 1799-1806.

[151] Zhang, J., Y. Zhao, A. Park, X. Song, 2005. "Economic Returns to Schooling in Urban China, 1988-2001," *Journal of Comparative Economics* 33: 730–752.

[152] Zhao, M., and P. Glewwe, 2010. "What determines basic school attainment in developing countries? Evidence from rural china," *Economics of Education Review* (in press).

Appendix A: The censored order probit

Following Maddala (1983) and Glewwe and Jacoby (1994), assume that for each person i ($i = 1, \dots, N$), the demand for education is a linear function of $\mathbf{z}_i$, where α is a vector of coefficients associated with all the variables in $\mathbf{z}_i$ The observed years of schooling, Y_i, is assumed to follow:

$$
\begin{aligned}
Y_i &= 0 \quad\; if\; -\infty \leq Y_i^* < \alpha_0 \iff -\infty \leq \mathbf{z}_i^{'}\boldsymbol{\gamma} + \eta_i < \alpha_0 \iff \eta_i < \alpha_0 - \mathbf{z}_i^{'}\boldsymbol{\gamma} \\
Y_i &= 1 \quad\; if\; \alpha_1 \leq Y_i^* < \alpha_2 \iff \alpha_1 \leq \mathbf{z}_i^{'}\boldsymbol{\gamma} + \eta_i < \alpha_2 \iff \alpha_1 - \mathbf{z}_i^{'}\boldsymbol{\gamma} \leq \eta_i < \alpha_2 - \mathbf{z}_i^{'}\boldsymbol{\gamma} \\
& \dots \\
Y_i &= m-1 \; if\; \alpha_{m-1} \leq Y_i^* < \alpha_m \iff \alpha_{m-1} \leq \mathbf{z}_i^{'}\boldsymbol{\gamma} + \eta_i < \alpha_m \iff \alpha_{m-1} - \mathbf{z}_i^{'}\boldsymbol{\gamma} \leq \eta_i < \alpha_m - \mathbf{z}_i^{'}\boldsymbol{\gamma} \\
Y_i &= m \quad\; if\; \alpha_m \leq Y_i^* < \infty \iff \alpha_m \leq \mathbf{z}_i^{'}\boldsymbol{\gamma} + \eta_i \iff \alpha_m - \mathbf{z}_i^{'}\boldsymbol{\gamma} \leq \eta_i
\end{aligned}
$$

with m the highest level of Y_i and α's the underlying cutoffs that determine, jointly with Y_i^*, the observed years of schooling. As η_i is assumed *i.i.d* with a standard normal distribution, the probability of observing Y_i at each level can be expressed as the follows:

$$
\begin{aligned}
\Pr(Y_i &= 0|\mathbf{z}_i) = F(\alpha_0 - \mathbf{z}_i^{'}\boldsymbol{\gamma}) \\
\Pr(Y_i &= 1|\mathbf{z}_i) = F(\alpha_2 - \mathbf{z}_i^{'}\boldsymbol{\gamma}) - F(\alpha_1 - \mathbf{z}_i^{'}\boldsymbol{\gamma}) \\
& \dots \\
\Pr(Y_i &= m-1|\mathbf{z}_i) = F(\alpha_m - \mathbf{z}_i^{'}\boldsymbol{\gamma}) - F(\alpha_{m-1} - \mathbf{z}_i^{'}\boldsymbol{\gamma}) \\
\Pr(Y_i &= m|\mathbf{z}_i) = 1 - F(\alpha_m - \mathbf{z}_i^{'}\boldsymbol{\gamma})
\end{aligned}
$$

where F is the *c.d.f.* of a standard normal distribution. If $Y_i = j$ ($j = 0, \dots, m$) and is censored (e.g. person i is currently enrolled in level j), all we know is that her final years of schooling will be greater than or equal to $(j-1)$. Hence, the probability of observing j years of schooling should be $\Pr(Y_i = j|\mathbf{z}_i) = 1 - F(\alpha_j - \mathbf{z}_i^{'}\boldsymbol{\gamma})$. Letting $I_{ij} = 1$ if $Y_i = j$, $I_{ij} = 0$ otherwise; $d_i = 1$ if Y_i is censored and $d_i = 0$ otherwise, the log likelihood of observing the whole sample of size N can be expressed as:

$$
\begin{aligned}
\ln L(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \textstyle\sum_{i=1}^{N} \ln \Pr(Y_i|\mathbf{z}_i) \\
\ln L(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \textstyle\sum_{i=1}^{N} \sum_{j=1}^{J} I_{ij} \ln[F(\alpha_j - \mathbf{z}_i^{'}\boldsymbol{\gamma})^{1-d_i} - F(\alpha_{j-1} - \mathbf{z}_i^{'}\boldsymbol{\gamma})]
\end{aligned}
$$

Censored ordered probit estimators for the $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s are those that maximize the above log likelihood function.

Appendix B: Sampling design of the CHNS

In 1989, all the major provinces in mainland China were contacted to see if they were willing to participate in the CHNS. Among those who showed interest in the survey, 8 were selected because of the substantial variation in geographic location, economic development, health indicators and local culture (Guangxi, Guizhou, Henan, Hubei, Hunan, Jiangsu, Liaoning, and Shandong). All areas in each province were categorized into urban and rural areas, followed by an income stratification- that is, low, middle and high income groups in urban areas and in rural areas. Sampling with probability proportional to size (PPS), one high income city, usually the capital city which has the largest population, and one low income city were randomly selected from urban areas. Following the same weighted sampling scheme, one high income county, two middle income counties and one low income county were selected in rural areas. The next step was neighborhood selection: two urban neighborhoods and two suburban ones were randomly selected from each sampled city while, in each rural county, one neighborhood from county capital city and three villages were randomly selected. Due to a misunderstanding, one city did not select the required two suburban neighborhoods, which resulted in a total of 32 urban neighborhoods, 30 (should have been 32) suburban neighborhoods, 32 town neighborhoods and 96 villages. Within each selected neighborhood/village, 20 households were randomly selected.

Appendix C: Local linear regression estimates of standard errors of the treatment effect

The estimates of the standard errors are obtained by the following steps. First, define the residuals from the regressions of equations (7a)-(7d) as

$$\hat{e}_{Y,i} = \begin{cases} Y_i - \hat{a}_{Y,l} - \hat{b}_{Y,l}(z_i - z_0) \;\; if \;\; z_i < z_0 \\ Y_i - \hat{a}_{Y,r} - \hat{b}_{Y,r}(z_i - z_0) \;\; if \;\; z_i \geq z_0 \end{cases}$$

and

$$\hat{e}_{T,i} = \begin{cases} Y_i - \hat{a}_{T,l} - \hat{b}_{T,l}(z_i - z_0) \;\; if \;\; z_i < z_0 \\ Y_i - \hat{a}_{T,r} - \hat{b}_{T,r}(z_i - z_0) \;\; if \;\; z_i \geq z_0. \end{cases}$$

Then define

$$\begin{aligned} \hat{\Delta}_{Y,l} &= \sum_{i:z_i<z_0} \lambda_i^2 \cdot \hat{e}_{Y,i} \cdot A_i, \qquad \hat{\Delta}_{T,l} = \sum_{i:z_i<z_0} \lambda_i^2 \cdot \hat{e}_{T,i} \cdot A_i \\ \hat{\Delta}_{Y,r} &= \sum_{i:z_i>z_0} \lambda_i^2 \cdot \hat{e}_{Y,i} \cdot A_i, \qquad \hat{\Delta}_{T,r} = \sum_{i:z_i>z_0} \lambda_i^2 \cdot \hat{e}_{T,i} \cdot A_i \\ \hat{\Delta}_{YT,r} &= \sum_{i:z_i<z_0} \lambda_i^2 \cdot \hat{e}_{Y,i} \cdot \hat{e}_{T,i} \cdot A_i, \qquad \hat{\Delta}_{YT,r} = \sum_{i:z_i>z_0} \lambda_i^2 \cdot \hat{e}_{Y,i} \cdot \hat{e}_{T,i} \cdot A_i. \end{aligned}$$

where

$$A_i = \begin{pmatrix} 1 & (z_i - z_0) \\ (z_i - z_0) & (z_i - z_0)^2 \end{pmatrix}$$

Finally, define

$$\hat{\Delta} = \begin{pmatrix} \hat{\Delta}_{Y,l} & \hat{\Delta}_{YT,l} & 0 & 0 \\ \hat{\Delta}_{YT,l} & \hat{\Delta}_{T,l} & 0 & 0 \\ 0 & 0 & \hat{\Delta}_{Y,r} & \hat{\Delta}_{YT,r} \\ 0 & 0 & \hat{\Delta}_{YT,r} & \hat{\Delta}_{T,r} \end{pmatrix}$$

and

$$\hat{\Gamma}_l = \sum_{i:z_i<z_0} \lambda_i A_i, \; \hat{\Gamma}_r = \sum_{i:z_i>z_0} \lambda_i A_i, \text{ and } \hat{\Gamma} = \begin{pmatrix} \hat{\Gamma}_l & 0 & 0 & 0 \\ 0 & \hat{\Gamma}_l & 0 & 0 \\ 0 & 0 & \hat{\Gamma}_r & 0 \\ 0 & 0 & 0 & \hat{\Gamma}_r \end{pmatrix}.$$

Then the covariance matrix of $\hat{\theta} = (\hat{a}_{Y,l}, \hat{b}_{Y,l}, \hat{a}_{T,l}, \hat{b}_{T,l}, \hat{a}_{Y,r}, \hat{b}_{Y,r}, \hat{a}_{T,r}, \hat{b}_{T,r})'$ can be estimated by $\mathbf{V} = \hat{\Gamma}^{-1}\hat{\Delta}\hat{\Gamma}^{-1}$. Since the estimate of the treatment effect is

$$\hat{\beta} = \frac{\hat{a}_{Y,r} - \hat{a}_{Y,l}}{\hat{a}_{T,r} - \hat{a}_{T,l}} = \frac{\hat{\theta}_5 - \hat{\theta}_1}{\hat{\theta}_7 - \hat{\theta}_3},$$

following delta method, the variance of $\hat{\beta}$ can be estimated as $g'\hat{\mathbf{V}}g$ where g is a 8×1 column vector with all the elements equal to zero except those corresponding to $\hat{a}'$s:

$$\begin{aligned} g_1 &= -1/(\hat{\theta}_7 - \hat{\theta}_3), \;\; g_3 = (\hat{\theta}_5 - \hat{\theta}_1)/(\hat{\theta}_7 - \hat{\theta}_3)^2 \\ g_5 &= 1/(\hat{\theta}_7 - \hat{\theta}_3), \;\; g_7 = -(\hat{\theta}_5 - \hat{\theta}_1)/(\hat{\theta}_7 - \hat{\theta}_3)^2. \end{aligned}$$