University of Sussex

Intergenerational Transmission and the Effects of Health on Migration

Mimi Xiao

Submitted for the degree of Doctor of Philosophy

Department of Economics

University of Sussex

January 2015

Declaration

This thesis, whether in the same or different form, has not been previously submitted to this or any other University for the award of a degree. Chapter 2 and 3 are based on joint work with my supervisor Peter Dolton.

Signature:

Mimi Xiao

Abstract

This thesis conducts empirical analysis on the intergenerational transmission of adiposity, using various types of data from various countries; the same intergenerational transmission in China and how it varies with the family socioeconomic factors and age levels; the way in which health impinges on the decision to migrate in China. In the first empirical chapter we find that the intergenerational elasticity of adiposity is relatively constant – at 0.2 per parent, and this elasticity is comparable across time and countries. Quantile estimates suggest that this intergenerational transmission mechanism is more than double for the fattest children as it is for the thinnest children. The second empirical chapter examines the intergenerational transmission of adiposity in China: we use BMI z-score as another measure of adiposity, the longitudinal structure of CHNS data (1993-2009) allows us to control for individual fixed effects or family fixed effects and focus on changes in BMI z-score over the life cycle. We report patterns of the intergenerational relationship of BMI z-score varying by family socio-economic factors and the age of the child, the magnitude of this relationship reaches the peak over the stage between childhood and later adolescence. In the third empirical chapter, which also uses the CHNS data, we examine whether migrants are healthier than those who do not migrate in the places of origin in the context of internal migration in China. Based on the relative wage rates, costs of migration and the assumption of optimization, we set up a theoretical model and estimate the effects of health on the migration probability, we find that people self-evaluating as having “good” or “excellent” health are more likely to migrate, this health effects vary with the type of occupation, we also find evidence on the indirect health effects which operates through the education attainment.

Acknowledgements

I thank my supervisors L. Alan Winters and Peter Dolton for much more than words can express.

I thank Andy Mckay, Andrew Newell, Barry Reilly, Pedro Rosa Dias, Richard Dickens, Richard Tol, Richard Disney and Vikram Pathania for their encouragements and suggestions.

I also thank Antonia Schwarz, Cecillia Poggi, Eva-Maria Egger, Cesar Gustavo Iriarte Rivas, Chowdhury Rashaad Shabab, Edgar Salgado Chavez, Egidio Farina, Farai Jena, Francisco-Javier Cabrera, Hector Gutierrez Rufrancos, Marco Carreras, Pedro Paulo Orraca Romano and Ani Rudra Silwal for being a source of friendships and intellectual stimulations.

I acknowledge financial support from the Sussex-China Scholarship Council joint program.

I dedicate this thesis to my parents, Jian’an Xiao and Lingying Wang.

Table of Contents

List of Tables

List of Figures

List of Appendix Tables

List of Appendix Figures

Acronyms and Glossary

BMI	Body mass index
IBE	Intergenerational elasticity of BMI
IIE	Intergenerational elasticity of income or earnings
IEE	Intergenerational elasticity of education
WHO	World Health Organization
CHNS	China Health and Nutrition Survey
IFLS	Indonesian Family Life Survey
BCS1970	British 1970 Cohort Studies
HSE	Health Survey for England
NHNAES	National Health and Nutrition Examination Survey
ENS-2006	Spanish National Health Survey
ENCELURB	The Survey for the Evaluation of Urban Households in Mexico
OLS	Ordinary Least Squares
FE	Fixed Effects

Introduction

This thesis conducts empirical analysis on the intergenerational transmission of adiposity using data from the UK, USA, China, Indonesia, Spain and Mexico, how this transmission evolves with age and relate to family socioeconomic factors in China. The Health selectivity of migrants in China is also investigated.

With the dramatic socio-cultural and environmental changes which are reshaping human behaviour and their bodies, the rising obesity has become a social phenomenon and a public health problem. According to the media centre of WHO, worldwide obesity has nearly doubled since 1980, particularly among the children. In developed countries, 23.8% of boys and 22.6% of girls were overweight or obese in 2013; in developing countries, this number is around 8.1% to 12.9% in 2013 for boys and from 8.4% to 13.4% for girls (Ng et al. 2014). Studies show that the parental obesity, particularly maternal obesity, has a direct impact on the developmental programming of obesity and metabolic disorders in their children, children from obese mothers are more likely to develop obesity in their lifetime (Battista et al. 2011). Therefore, it is important to study obesity from the intergenerational perspective.

This transmission of adiposity relates to the intergenerational aspect of social mobility more generally. Social mobility includes intergenerational and intragenerational aspects. The intergenerational mobility is more common, it indicates the relationship between children's socioeconomic position relative in children's generation and parents' position relative in parents' generation, and is usually measured by "intergenerational income elasticity". The lower the elasticity, the more mobile is the society. Early child health and education play a key role in this intergenerational mobility. Parental socioeconomic factors (childhood environment) affect child health and then their cognitive or non-cognitive skills to move up on the socioeconomic ladder, this upward mobility also contributes to better adult health (Nyström Peck 1992). In addition, parental health is transmitted to child health, child health affects child's cognitive or education outcomes (von Hinke Kessler Scholder et al. 2012), and then affects adult earnings or other labour market outcomes (such as occupational attainments (Morris 2006)). Therefore, there is also an intergenerational mechanism through which health affects social mobility, by affecting income or education (Manor, Matthews, and Power 2003).

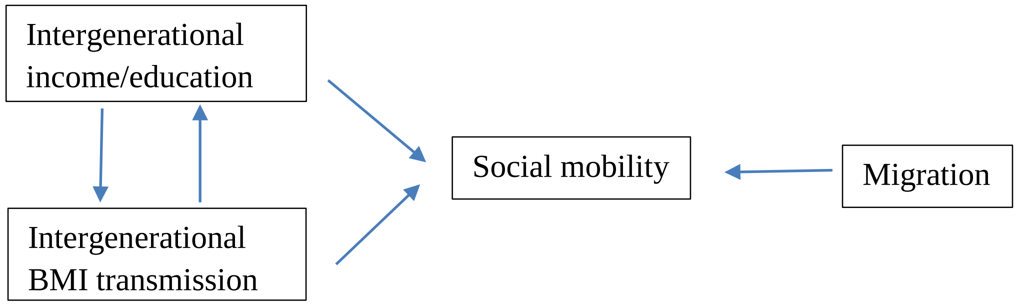
To date intergenerational studies mostly focus on income or education, a comparison of these studies reveals a substantial variation in this intergenerational relationship across countries (Pekkarinen et al. 2009, Björklund et al. 2012). In terms of health, the literature is relatively small though growing. Predominantly intergenerational health study papers are largely published in medical or epidemiological journals. They use a wide variety of different health measures. Most of these studies are conducted by incorporating parental health outcomes into the estimation where the dependent variable is child's health (Ahlburg 1998); few papers have claimed a causal link, due to the difficulty to account for unobserved "environmental" factors, which might influence health outcomes of both parents and children. Using data from different countries and of different types, we estimate the intergenerational elasticity of adiposity measured by BMI. Compared with income or education, adiposity is the product of a process of which a larger fraction might be driven biologically; therefore, the study of the intergenerational transmission of adiposity helps one to understand the underlying process of intergenerational mechanism. In addition, since health affects cognitive or education outcomes (Cesur and Kelly 2010) or labour market outcomes (such as occupation attainments (Morris 2006), the intergenerational transmission of health might interact with the intergenerational transmission of education or income (Case, Fertig, and Paxson 2005), and then influence the social mobility.

China, with the transition in economy, has been undergoing a transition in lifestyle which can be characterised as a falling physical activity and a shift towards energy-dense dietary. As a result, China has changed from one of the leanest populations to a nation with over one fifth of all one billion people in the world (Wu 2006). Studies show that among the adults aged from 18 to 75 years, the proportion of overweight has surged from 14.6% in 1992 to 45.38% in 2011, and the obesity has nearly tripled from 5.2% in 1992 to 15.06% in 2011 (Huynh, Kreinovich, and Sriboonchitta 2014). This trend is rising among children, 23% of Chinese boys under age 20 are overweight or obese, while the comparable figure for girls is 14%. A substantial literature shows parental education (particularly mother's education) (Breierova and Duflo 2004) and parental income are closely correlated with the child health, child health (especially poor health in childhood) might be an important mechanism for intergenerational transmission of economic status

(Grossman 2000), since parental income affects child health, and child health affects future income (Currie and Moretti 2005). Therefore, to obtain a better understanding of the intergenerational mechanism underlying social mobility, it is helpful to investigate the intergenerational correlations in health, and how they vary with family socioeconomic factors (family “environmental” factors), such as father’s occupation and mother’s education.

Migration, as a form of geographic mobility, interacts with the social mobility. People usually migrate for higher income in the destination areas, for higher-income occupations or education opportunities which might be associated with higher incomes, though this might not be achieved until the next generation, therefore, migration is often associated with upward social mobility; on the other hand, studies show occupations at the higher social class are often associated with higher rate of geographic mobility. (Fielding 2007). Internal migration in China might be one of the most extensive in the world, there are over 170 million rural-urban migration from 1979 to 2009 (Chan 2013). One of the significant issues within the migration literature is the selectivity of migrants; it is suggested that migrants are often drawn from the intermediate and higher levels of the skill distribution in the sending communities, people of higher education, better economic status and better health are generally more likely to move. Since self-selective movement contributes to the upward social mobility, examining how the health selectivity of migrants actually applies in the internal migration of China has important implications for an understanding of social mobility in China.

Figure 1.1: Social mobility as the link between intergenerational BMI transmission and migration


Intergenerational income/education
Intergenerational BMI transmission
Social mobility
Migration

This thesis is organized into five chapters---introduction, three empirical analysis and conclusions. The next chapter investigates how adiposity (measured by BMI) is transmitted across generations, based on data from six countries, covering both developed (US, UK, Spain) and developing countries (China, Indonesia and Mexico). Motivated by the prior studies on intergenerational persistence in obesity, we also investigate how the intergenerational elasticity of BMI varies across the distribution of children's BMI. Using the Body Mass Index (BMI) as a measure of adiposity, we find that the elasticity of intergenerational transmission is relatively constant – at 0.2 per parent, and is comparable across time and countries - even if these countries are at different stages of economic development. Additionally, we find this intergenerational transmission mechanism is substantively different across the distribution of children's BMI, it is more than double for the fattest children what it is for the thinnest children.

Chapter 2 provides a broad picture on this intergenerational BMI transmission in different countries. In Chapter 3, using BMI z-score from the WHO software as another measure of adiposity, we estimate the intergenerational transmission of adiposity in China. In addition to the analysis in Chapter 3, the rich information on socioeconomic indicators in China Health and Nutrition Survey (CHNS) allows us to include a variety of covariates in the estimation, and to explore the variability of this transmission with respect to different socioeconomic indicators. Moreover, the longitudinal structure of CHNS allows us to identify the short term environmental effects of BMI by applying individual and household fixed effects. We find that this intergenerational transmission of BMI z-score to be around 0.20 per parent, one standard deviation increase in one parent's BMI z-score is associated with an increase of 0.20 in child's BMI z-score. This falls to around 0.14 when we control for individual fixed effects or family fixed effects and focus on changes in BMI z-score over the life cycle. Additionally, we find that this intergenerational correlation of BMI z-score does not vary substantially with family SES indicators; it tends to be higher among children of higher BMI levels. With respect to age of child, the magnitude of this correlation reaches the maximum over the stage between the childhood and the later adolescence.

Chapter 4 uses the CHNS data (1993-2009) to examine the "healthy migrant hypothesis" in the context of internal migration in China. The "healthy migrant hypothesis" posits that migrants tend to have better health than those who do not migrate in the places of origin. Based loosely on Jasso et al.(2004)'s model of health selectivity, we set up a model in the same way as Borjas (1987)'s model of self-selection; the health effects derived from this selectivity model suggest that the health effects vary with occupation or education, therefore this model allows us to derive an interaction between health and proxies for occupation or education. Based on a sample of people aged 16 to 35 years, we apply a probit model and find that the evidence on positive health selection exists, but is not strong. This might be due to the substantial heterogeneity across households and circumstances and the rather small sample we have to deal with, or the weakness of the measures we have to use. We test the hypothesis on interactions derived from our model, and find that the health effects tend to be larger for the lower skilled workers, which is consistent with what the model predicts. We also test the hypothesis on the indirect effects by which we mean the effects of earlier health on education attainment, we find self-evaluating as having "fair", "good" or "excellent" health between age 13 and 16 years has a positive effect on the highest education degree they obtained after they were 16 years old. In addition, exploiting the longitudinal structure of the CHNS data, we do not find evidence on the effects of lagged health and the effects of improvement in health on migration. Furthermore, we also estimate the main equation using a health index which is created by collapsing various variables into a simple measure, we find the estimates for health effects are sensitive to the type of variables and the weights assigned to variables in the index, and that the estimates appear more significant when the index is based on more health variables and gives more weights to the self-rated, as opposed to "objective" measures of health. To sum up, these results provide positive but relatively weak evidence on health selectivity of migrants, although one needs to remember that there is a large heterogeneity in this rather small sample.

Chapter 2: The Intergenerational Transmission of BMI across Countries

2.1 Introduction

There is a worldwide epidemic of obesity. We are just beginning to understand its consequences for child obesity - which has become one of the foremost public health problems in most countries. This chapter addresses one important component of the crisis – namely the extent to which obesity – or more generally – adiposity - is passed down from one generation to the next.

We examine the extent to which the BMI of the children is inherited from the BMI of their parents. We use data on the heights and weights of approximately 100,000 children and their parents, measured by health care professionals from across six countries[1]: the UK, USA, China, Indonesia, Spain and Mexico. Our analysis applies to all ages of children up to 18 years and in all countries, from the most to the least developed, and with the most (USA) to least (Indonesia) obese population. Using the BMI as a measure of adiposity, we find that the elasticity of intergenerational transmission of BMI is constant – at 0.2 per parent and the effect is additive separable per parent.

In 2013, the US spent 190 billion dollars on obesity-related health expenses. The US is not alone in experiencing this epidemic. Countries like Mexico, the UK and other European countries are all alarmed by the rising obesity prevalence (Popkin and Penny, 2004). It is also the case that many developing countries are seeing huge rises in the fraction of children who are becoming obese-in literally one generation. Countries like China and Indonesia are our relevant comparators. We are beginning to understand the causes and consequences of childhood obesity. This paper addresses the intergenerational dimension of this crisis by examining how adiposity is passed down from one generation to the next and compare it to other intergenerational processes.

[1] We thank Oscar Marcenaro-Gutierrez, Alma Sobrevilla and Qisha Quarina for their research assistance on the Spanish (ENS-2006), British (BCS1970 cohorts), Mexican (ENCELURB) and Indonesian (IFLS) data respectively.

Hence, the underlying question is: what is the driving force behind rising childhood obesity? Adiposity - or fatness - is a result of both genetic inheritance and decisions made in families – loosely termed the 'family environment'. Most clearly, the family decisions relating to what to eat, how much to eat, how much exercise to take, how to spend family time, and other key lifestyle choices will all have a bearing on the outcomes of individuals in the family. However, to what extent is the individual's adiposity as reflected in their BMI 'not directly their responsibility' in the sense that their body shape, weight and height – and hence their BMI - is passed down to them through their parents? This is our central concern in this chapter.

The second focus is to pose the question of whether the process of intergenerational transmission of adiposity is the same across countries – irrespective of their stage of development, degree of industrialisation, or type of economy. The motivation here is to understand the extent to which the process driving intergenerational transmission is related to the type of economy and society under consideration. To this end we sought to examine data from literally all the countries we could retrieve a reasonable sample with the appropriate information. This is a considerable undertaking as there are not many datasets in the world where we have - both children's and parents heights and weights, preferably on more than one occasion, which are mostly medically measured rather than self-reported[2]. We were able to obtain data from diverse countries – from those with the most obese population – USA – to some of the least obese countries in the world – China and Indonesia.

The third line of investigation is to explore the extent to which the intergenerational relationship of BMI is potentially different at different points in the distribution of child's BMI. In other words, to what extent is the intergenerational mechanism the same for fat children and thin children? One could easily hypothesise that the relationship could be different at different points in the distribution. Specifically, for the fatter children, the fact they are being fatter is more to do with that their parents are fat or the decisions they

[2] The height and weight in British Cohort Study 1970 are self-reported when the respondents are aged 26 years old; there are both self-reported and medically-measured height and weight in Health Survey for England; In the Spanish National Health Survey (ENS-2006), adults answer the adult health questionnaire, and members under 16 answer the child health questionnaire.

make by their own as they grow up. Our research findings show this intergenerational correlation varies by child's BMI. Consistently, across all populations studied, we find it to be lowest for the thinnest children and highest for the fattest. The IBE for the former is 0.1 per parent and for the latter, 0.3 per parent.

To understand the process of obesity, it is crucial to understand the intergenerational transmission mechanism behind it. Evidence (Maes, Neale, and Eaves 1997) suggest that adiposity is affected by both environmental and genetic factors. Clearly, the intergenerational transmission mechanism here operates through both these two channels. So it is transmitted through family environmental factors, which directly relates to the intra-household mechanism (how the resources are allocated within the family), and it is also affected by genetic factors through a direct channel. By applying fixed effects model, we attempt to provide some evidence on the effects of short term environmental factors, assuming genetic factors and long term environmental factors are constant over time.

To provide some basic perspective of the underlying relationship between parents and child's BMI – we first present some non-parametric graphs of the aggregate data, using a kernel plot based on the raw data. Figure 2.1 below is the local weighted scatter smoothing of the log of father's BMI variable against the log of their child's BMI variable; similarly, Figure 2.2 presents the local weighted scatter smoothing of the log of mother's BMI variable against the log of their child's BMI variable. In the Mexico data, only pairs of mother-child are available, therefore, the lowess plot for father-child is not presented for Mexico in Figure 2.1. The slopes capture the magnitude of the intergenerational elasticity. They suggest that the slopes have a fairly constant gradient and are nearly parallel across countries. This finding shows that the underlying gradient of the relationship between adiposity across generations is fundamentally constant and that the stage of development of the country only shifts up the intercept with the least developed country having the lowest intercept and the most developed country the highest intercept.

Figure 2.1: Lowess Plot of Log (father's BMI) and Log (child's BMI)

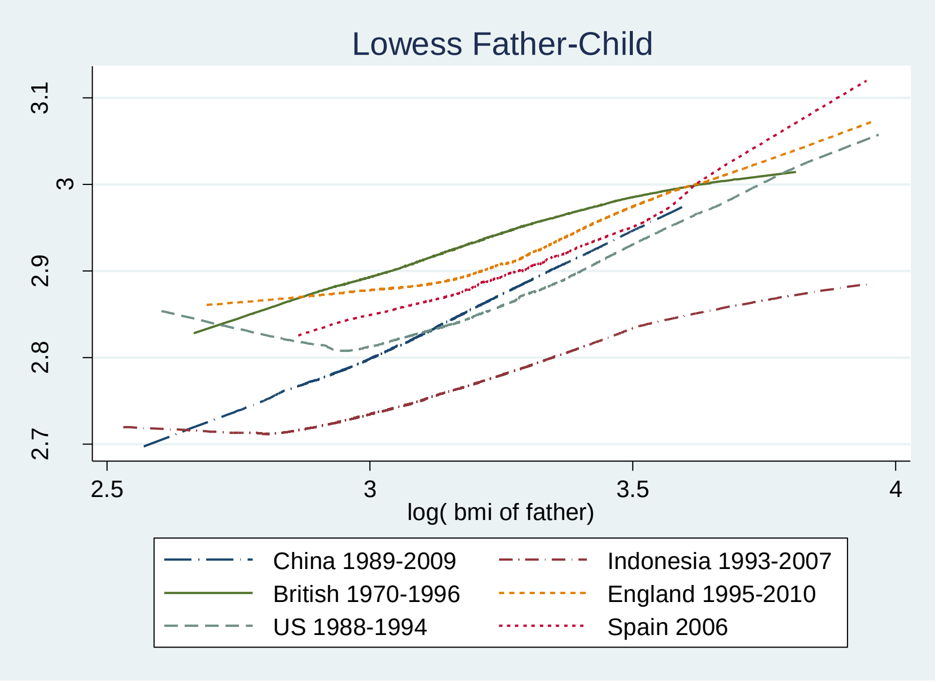
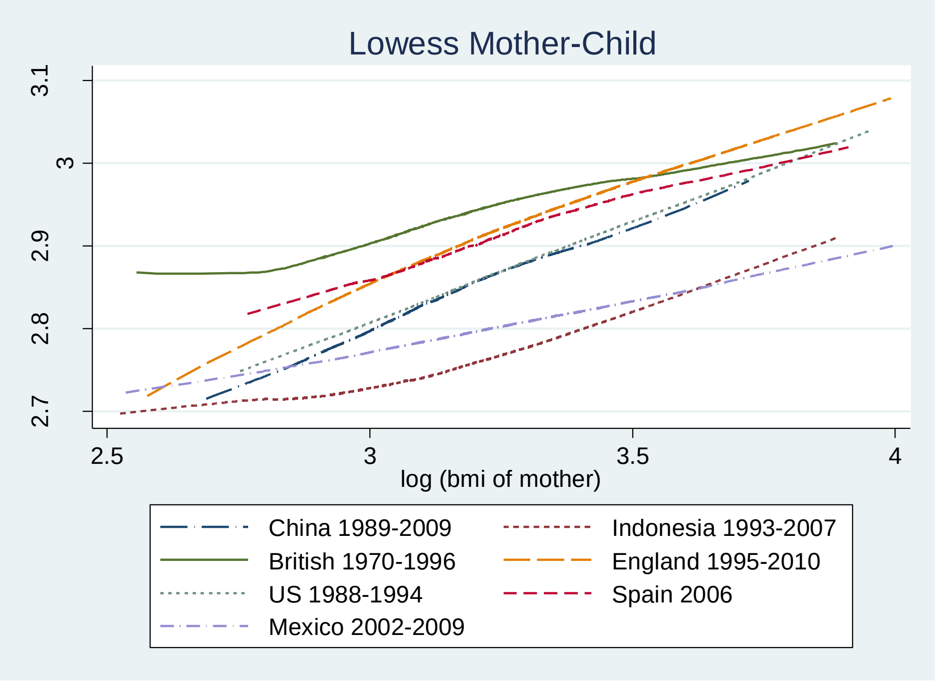


Figure 2.2: Lowess Plot of Log (mother's BMI) and Log (child's BMI)



There are several features in these two figures. Naturally, the western countries, whose populations typically have fatter body types are above the less developed countries whose populations have thinner frames. The other thing we would expect is that some of the country profiles start much further along the x-axis than others – for example, Indonesia and China – simply because there are relatively few fat children in these countries. But the most important thing to notice is our central finding in this research – namely that the lines for each country are, for the most part, parallel. This suggests that the elasticity – here the slope of the line in log-log space - is essentially a similar number in each country. In simple terms, this research presents, the substantive – hitherto unreported finding - that the proportionate increase in a child's BMI which is associated with their parent's BMI, is approximately constant – at around 0.2 across countries and populations which are substantively different in epidemiological terms. This suggests that a unit increase in an adult's BMI will have 20% effect on their child at the mean, and this impact is nearly doubled when we consider the effect of both parents assuming they are additive (we will discuss this later).

2.2 Literature Review

Intergenerational studies originate with Francis Galton (1869). By running a regression of the offspring's height on their parents' height, Francis Galton (1869) argued that an individual's characteristics are correlated with those of their parents and at the same time "regress to mediocrity". More specifically, the individual characteristics (such as height and weight) are closer to the population mean than those of their parents. This finding was the basis of Becker-Tomas model (1986) of intergenerational human capital transmission (Goldberger 1989, Han and Mulligan 2001, Mulligan 1999).

Most of the intergenerational studies concern the transmission of income or education outcomes. The focus in this strand of research relates to the equality of individual opportunity over time, which exerts profound influences on the social mobility. The strength of the income transmission is usually measured by the elasticity of children's income with respect to their parents' income (i.e. the intergenerational elasticity of income, hereafter called IIE). The larger is the IIE the more it means that the children's relative position on the "income ladder" is determined by their parents' income position.

Naturally, we would be concerned if this elasticity of the transmission mechanism was (too) large, it would imply that both equity and efficiency of the society would be undermined.

In terms of the discrepancy in IIE across countries, partly due to the restriction of data which covers multiple generations, most of these studies are conducted in the US or European countries. In the US, the consensus on the estimated IIE is " 0.4 or a bit higher" (for instance, 0.473 using PSID by Grawe (2004), 0.542 using NLSY sample born in 1957-64 by Bratsberg et al. (2007), this is higher than Canada (0.2 using register data (Corak et al. 1999), 0.152 using IID Canadian Intergenerational Income Data and 0.381 using PSID Panel Study of Income Dynamics data (Grawe 2004)) and most of the European countries except for Britain (0.45 using NCDS 1958 cohort (Bratsberg et al. 2007)) and Italy (0.48 using Italian data from the Survey on Household Income and Wealth (SHIW) (Piraino 2007)). The IIE estimates in Nordic countries and Scandinavian societies are often the lowest, ranging from 0.2 to 0.3 (Pekkarinen et al. 2009, Björklund et al. 2012). In contrast, the IIE in China is perhaps at the top of the list with 0.63: i.e. a Chinese father's income 10 percent above the paternal cohort mean will be associated with his son having an income 6.3 percent above the filial cohort mean (Gong 2012). Using the Urban Household Education and Employment Survey (UHEES) and the Urban Household Income and Expenditure Survey in 1987-2004 (UHIFS), Gong et al. (2012) show the IIE in China is 0.63 for father-son, 0.97 for father-daughter, 0.36 for mother-son, and 0.64 for mother-daughter. Education is one of the most crucial channels through which earnings ability is transmitted across generations, however, other factors such as genes and health are also potentially important pathways of intergenerational income transmission.

The intergenerational transmission of education achievement can be thought of in the following, where the child's education achievement is measured by their human capital $H_c = F\left(\mathrm{Y}_{\mathrm{c},}\mathrm{H}_{\mathrm{p},}\mathrm{A}_c\right)$, the mechanism operates through three main channels: First, parental income, higher educated parents tend to have more income, so they have more resources to invest in child's education ($\mathrm{Y}_{\mathrm{c},}$); second, parental education (H_p), since higher-educated parents may invest in child's education in a more efficient way; third, in addition to the two indirect channels above, parental education may affect child's education through a direct channel, which is usually proxied by the genetic inheritance of ability (A_c).

Empirically, the first channel can be decomposed into the effects of current parental income and the effects of permanent parental income, of which the latter normally plays the dominant role and might be measured by family fixed effects (Heckman and Carneiro 2003). The third channel is normally identified by comparing children of twin pairs (Behrman and Rosenzweig 2002) or between biological and adopted children with variation in education (Björklund et al. 2006), the general conclusion is that the intergenerational correlation in education cannot be fully attributed to the genetic factors. The intergenerational education elasticity (hereafter called IEE) varies from 0.14~0.45 in the USA (Mulligan 1999) to 0.25~0.4 in the UK (Dearden et al. 1997). Some studies examine the intergenerational elasticity of IQ, which is considered as a measure of the intergenerational relationship in the third channel, the estimates range from 0.3 to 0.5 (Solon, 2004, Anger and Heineck 2010, Van Leeuwen et al. 2008).

There is also a growing literature on the intergenerational correlation in various health outcomes, such as birth weight (Currie and Moretti 2005), self-rated health (Coneus and Spiess 2012, Thompson 2012), longevity (Trannoy et al. 2010) and smoking behaviour (Loureiro et al. 2006). These studies mostly find strong positive correlations across generations. In terms of adiposity and the related measures, a large proportion of the studies are published on the medical, biological or epidemiological journals, they mostly show parental health outcomes are strongly correlated with children’s. For instance, using data in the US, Canada (national sample), Quebec and Norway, Bouchard (1994) reports the parental-child correlations of BMI are 0.23, 0.20, 0.23 and 0.20, respectively. Using data from the National Longitudinal Survey of Youth 1979 (NLSY 1979) and the Young Adults of the NLSY79, Classen (2010) estimates the intergenerational transmission of BMI between children and their mother when both generations are between the age of 16 and 24, he runs the regression which includes only mother and finds the intergenerational correlation is significant and around 0.35. Applying a similar strategy where parents and children are matched at a similar life stage, Brown and Roberts (2013) use data on mothers and their adolescent children aged 11 to 15 years from the British Household Panel Survey (2004 and 2006), they find the overall intergenerational correlation of BMI is 0.25. In the context of developing countries, using the China Health and Nutrition longitudinal Survey (CHNS) (1989-2009), Eriksson, Pan, and Qin (2014) estimate the intergenerational transmission of health status, using height z-score and weight z-score as the health measure. They find a strong correlation between parents’ health and their children’s health after accounting for various parental socioeconomic factors (education and type of occupation), household characteristics (whether the household has a flush toilet) and the health-care factors (the distance to the nearest health centre in the community). To correct for the unobserved heterogeneity, they use the age and gender adjusted average parents’ BMI in parents’ province as the instrument for parental BMI variable. Additionally, using the decomposition analysis, they find the urban-rural differential in parental health explains 15-27% of urban-rural disparity in child’s health, in addition to the urban-rural differential in parental education and income, which plays a major role. This relates to the transmission mechanism, which is usually considered as operating through two main channels: genetic and environmental. The environmental channels are exploited more often than the genetic channel. Studies usually include a range of parental socioeconomic factors in the estimation, arguably this controls for part of the family “environmental” factors. Based on data from the German Socio-Economic Panel (SOEP), Coneus and Spiess (2008) estimate the intergenerational relationship of both father and mother and children. In addition to the pooled OLS estimation, they apply fixed effects estimation and find that father’s BMI has a significantly positive effect on child’s BMI (with a coefficient of 0.57, the estimates of mother’s BMI effects are not significant), while mother’s obesity is strongly associated with child’s obesity with a coefficient of 0.26. They claim their fixed effects estimates provide a more causal estimate for the intergenerational “transmission” rather than a “relationship”, since fixed effects estimation allows them to differentiate out the time-invariant unobserved heterogeneity. However, we argue the fixed effects estimates mainly capture the effects of rather short term environmental factors, and therefore shield some lights on the underlying transmission mechanism. In addition, in the German Socio-Economic Panel (SOEP), child’s health outcomes are provided by mother rather than medical professionals, and father and mother’s health are self-reported, this might lead to a bias in the estimates due to the measurement error. As Black and Devereux (2003) review, among the studies on intergenerational transmission of health, few have claimed a causal transmission, partly due to the unobserved behaviour or environmental factors, which affect the health outcome of both parents and children.

In addition to “regression to the mean” in the inheritability of BMI, the degree of this inheritability (intergenerational elasticity of BMI, hereafter called IBE) may vary across child’s BMI distribution and this variation usually relates to the family’s socioeconomic status in the society. In the study mentioned earlier, Classen (2010) also estimates the intergenerational BMI relationship across the distribution of child's BMI by applying quantile estimation, the results indicate that the intergenerational BMI relationship tends to be stronger among children with higher levels of BMI. Based on the general population-based Northern Finland Birth Cohort 1986, Jääskeläinen et al. (2011) find that children whose both parents were overweight or obese before pregnancy and after a 16-year follow-up had a high risk of overweight. This relative stronger intergenerational transmission of high levels of BMI is often found in developed countries and among families of lower social class or lower socioeconomic levels (Laitinen et al. 2001). One potential explanation is that in these countries where fast food industry is more developed and "unhealthy" food are generally cheaper than "healthy" food, lower income families might consume more "unhealthy" food which is viewed as one important contributory cause of obesity.

2.3 An Empirical Model of Intergenerational BMI Transmission

In this section, we outline an empirical model on intergenerational transmission of BMI. This model is directly analogous to Becker's model on intergenerational transmission of income. In Becker's model, parents allocate their income between the child's health, and their own consumption, to optimize their utility. In our model, the outcome of interest is a child's health (measured by BMI) which can be invested by parents sacrificing their own consumption. Hence here, Y denotes the child's health as the intergenerational outcome we are interested in. The child's health is a function of parent's income and resources, X, and a genetic endowment, E, which is determined exogenously at birth by the passing on of parental DNA. Since children cannot choose their parents and the genetic traits they inherit from them, then this endowment factor is reasonably taken as exogenous. Assume Y is determined by:

$$Y = \Phi X^{\gamma} E^{\delta} \tag{2.1}$$

where E is decomposed into genetic factors, e, and environmental factors, u .

$$E = e + u \tag{2.2}$$

From this point on, we will use lower case letters to denote observable variables which we obtain data on or can proxy for. Let subscript p index the parent and i index the child, substituting (2.2) into equation (2.1) and taking logs, we obtain

$$logy_i = \log \Phi + \gamma logx_p + \delta \log[e_i + u_i] \tag{2.3}$$

In the empirical work, we assume that mothers and fathers' BMI measures (respectively y_{mi} and y_{fi}) are sufficient statistics for their health and the environmental factors are individual specific and captured by the term f_i, so we estimate the following equation (2.4) in a cross-section framework.

$$\log(y_i) = \delta + \alpha \log(y_{fi}) + \beta \log(y_{mi}) + \gamma logx_p + f_i + \varepsilon_i \tag{2.4}$$

where i indexes individual child observations, ε_i captures the transformed stochastic error term. Equation (2.4) shows that child's health outcome y_i is a function of child i's father's health outcome, y_{fi}, and mother's health outcome, $y_{mi,}$ x_p denotes the age variables of father and mother, and f_i captures child i's age, gender and the interaction between them. Equation (2.4) is the classic equation in intergenerational studies, which is derived from the model of "regression to the mean" (due to Becker but strictly speaking it dates back to Galton). It is noteworthy the intergenerational elasticity here estimates the correlation between parents and child's BMI, rather than a causal relationship. We recognize the possibility that as children grow up, they could influence parents' BMI[3]. However, we cannot control for the reverse causality in this study.

The empirical estimation will be conducted in several stages. First, we estimate the IBE at the aggregate and cross country level. The single parent version (father-child and mother-child) and the both parents version (father-mother-child) of equation (2.4) are then estimated using all the individual-wave observations. Second, applying both parents

[3] For instance, if children are predisposed to do more exercise, this might increase the amount of excises parents take.

version (father-mother-child) version of equation (2.4), we estimate the IBE across different quantiles of child's BMI.

2.4 Data and Measurement Issues

We use data from six countries: China Health and Nutrition Survey (CHNS) data, Indonesian Family Life Survey (IFLS) data, British 1970 Cohort Studies (BCS1970), Health Survey for England (HSE) data, National Health and Nutrition Examination Survey (NHNAES) data, the Spanish National Health Survey (ENS-2006) and the Survey for the Evaluation of Urban Households (ENCELURB) data in Mexico[4]. The heights and weights are mostly medically measured in these data[5]. Compared to self-reported measures, which are widely used in the literature, these medically measured data may help to reduce the bias of our estimates due to measurement error. Our sample includes children aged under five years old[6]. For children aged under five years old, their BMI is likely to be related to their birth weight. Therefore, we restricted the sample to those aged above five and estimate the both-parents version of equation (2.4), the results are presented in Table A 2.2, they suggest that the estimates for intergenerational correlation appear larger than those based on the full sample (Table 2.8). This might be due to a larger fraction of "environmental factors" shared between parents and children when children are aged above five than for those aged under five, since children aged under five might have a different dietary pattern from their parents[7]. In addition, children aged 16 and above might have already left the household and the decision to leave may be related to health/BMI. Therefore, we restrict the sample to those aged between 5 and 16, and estimate the both-parents version of equation (2.4), the estimates are presented in Table A2.3, they are close to those based on children aged above five (Table A2.2), this is reassuring since it suggests that our estimates are not biased significantly by the factor that older children might have left the family.

[4] See Appendix for a detailed description of these data.

[5] Except for the BCS 1970 Cohort Studies and the Spanish National Health Survey (ENS-2006).

[6] The descriptive statistics of children's age are reported in Table A 2.1.

[7] This can be clearly seen in Table 3.9 and Figure 3.7 of Chapter 3, where we analyse the intergenerational BMI correlation by age group.

The most widely used measure of body fat, or adiposity, is the Body Mass Index (BMI) which is calculated using the following formula $\mathrm{BMI} = \left[\frac{\mathrm{weight(kg)}}{\mathrm{height^2(cm)}}\right] * 10{,}000$. As mentioned in the literature review, the majority of intergenerational studies use elasticity (eg. IIE and IEE) as a measure of the intergenerational relationship. To facilitate the comparison of our results on anthropometric data with other intergenerational results, we also use elasticity as the measure of the intergenerational relationship.

A problem we face is exactly how we correlate a child's BMI with their parent's BMI. A child's BMI is a function of their age and gender – so a simple correlation of child's BMI against parents BMI would not allow for this factor. One way to examine the intergenerational transmission is to wait until the child is an adult and then correlate the two BMIs. This is what Classen (2010) did. There are two problems with this – firstly there is very little data relating to when the child's height and weight are observed when they are an adults – as well as having their parents height and weight at the same time. Based on the children aged between 16 and 18 years old, we estimate the intergenerational BMI correlation and report the results in Table A2.8, they suggest the estimates for this correlation appear slightly larger than the estimates based on the full sample[8].The other problem with this is that we are mainly concerned with childhood obesity and so waiting until they are adults does not help us.

To address the potential age bias due to that child's BMI significantly varies with their age, we include child's age, age square and the interaction term of child's age with their gender as controlling regressors in our estimation on the assumption that in doing so we would have conditioned out for the non-linear effect of age on gender[9]. We also take a more flexible approach by including child's age dummies and their interactions with child gender, the results are reported in Table A 2.6, they suggest that the estimates are similar as those from the specification we adopt in this study. We use this method as a robustness check on our findings, but it does not differ much in the findings, we will therefore use the first method in each of our country datasets. We report the second method in an Appendix available on request for those interested.

[8] Another approach to obtain this correlation of "long-term" BMI might be to use the average of the observations in the data as the "long-term" BMI, but in that case we will lose a large number of observations.

[9] The weakness of this method is that we have to assume that we can net out for the whole non-linear process of the child's BMI rising as they age.

In the course of doing this research we had considered if there was an alternative way of retrieving the IBE. We contemplated using the WHO to generate z scores or percentiles and using these logged metrics. Naturally, the estimation of the BMI elasticity is sensitive to any possible transformation of its scale. – i.e., to z scores or percentiles. So keeping the analysis simple has many virtues. It turns out that estimating the model in the log of BMI or the BMI itself does not make much difference – the elasticity is slightly smaller when estimated without logging. But since taking logs allows for general non-linearity in the data and has the nice property that it preserves the constant elasticity across the range of BMI values then we adopt it here[10].

Before estimation, we plot the kernel density of child's BMI, father's BMI, mother's BMI across countries in Figure 2.3, Figure 2.4 and Figure 2.5, respectively. They show that in both generations, the distribution of BMI tend to shift rightwards as the development level of these countries increase, with Indonesian cohorts being the leanest and the UK cohorts (children in British 1970 cohorts and father in the Health Survey for England) being the most obese[11]. This is as expected as the nutrition status of population varies with the development of the nation (Floud et al 2011). In addition, we see the distribution of child's BMI is more concentrated than the distribution of father and mother's BMI, this is consistent with the rise of obesity prevalence in Mexico during the survey period (S. Leeder, et al. 2006).

[10] We naturally relax this assumption in Section 2.6.3 when we consider the quantile regression allowing the elasticity to vary across the range of the child's BMI.

[11] Figure 2.5 suggests that Mexico has the largest fraction of obese mother, this is consistent with the rise of obesity prevalence in Mexico during the survey period (S. Leeder, et al. 2006).

Figure 2.3: The kernel density of child's BMI

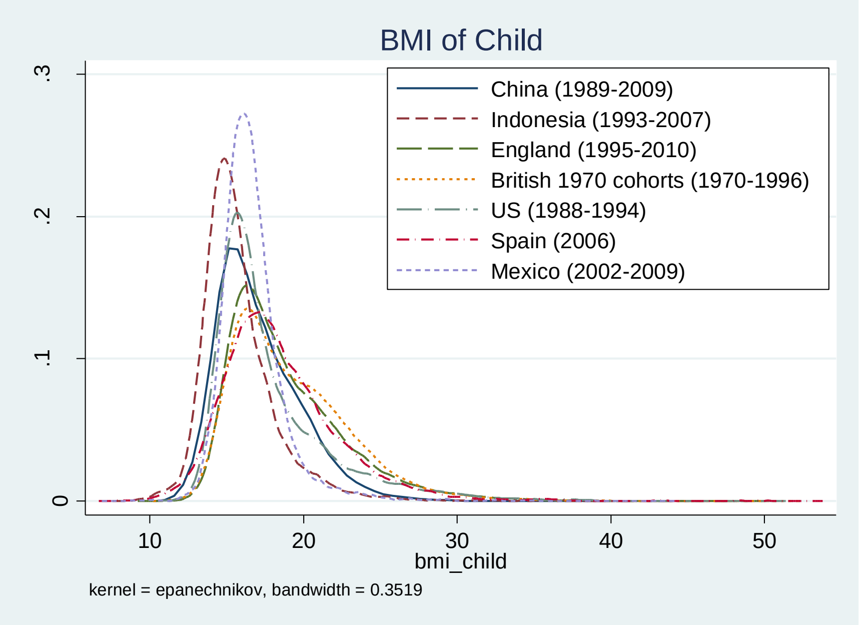


Figure 2.4: The kernel density of father's BMI

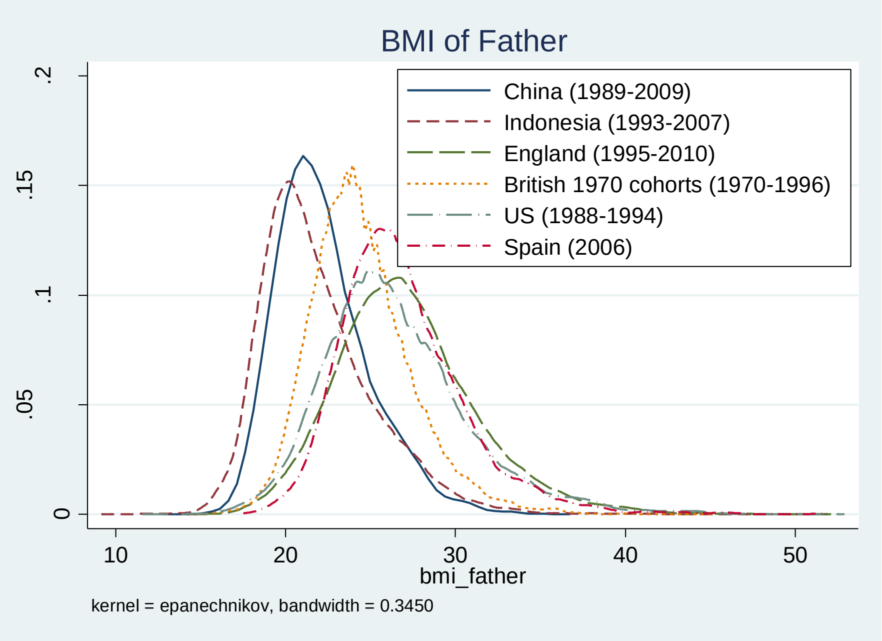
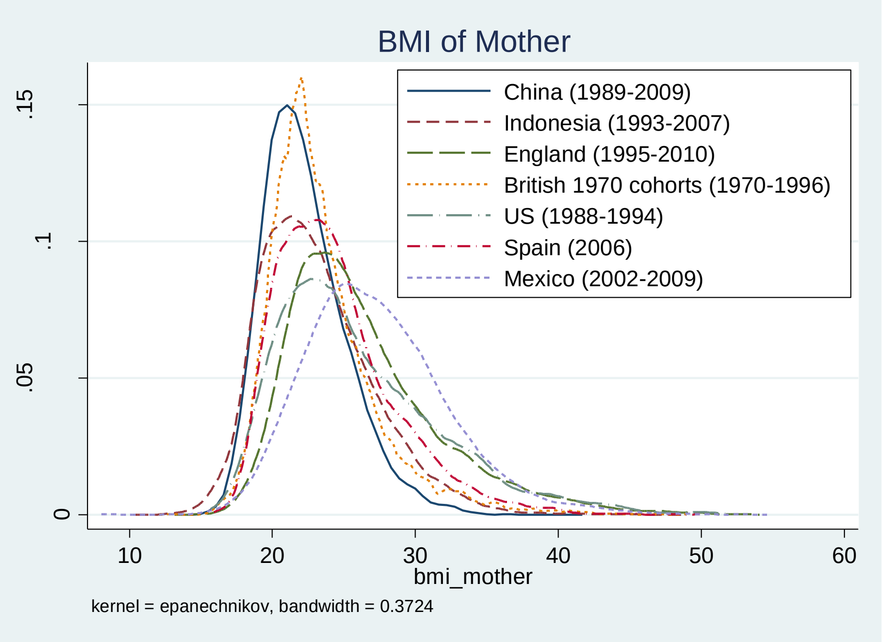


Figure 2.5: The kernel density of mother's BMI



2.5 Transition Matrices

Before estimation, we calculate the conditional transition probabilities to describe the rates of movement across specific categories of the BMI distribution across generations (Bhattacharya and Mazumder 2011). We adopt different BMI measures when we classify the BMI category of mothers and children. We classify mothers' BMI status based on their raw BMI: average BMI under 18.5 are classified as underweight, 18.5~24.9 as normal weight, 25~29.9 as overweight, and above 30 as obese. Whereas the classification of children's BMI status is based on their BMI z-score: underweight if BMI z-score <-1.04; normal if -1.04<=BMI z-score<1.04; overweight if 1.04<=BMI z-score<1.64; obese if BMI z-score>=1.64. This BMI z-score is calculated with respect to the WHO reference population which varies by age and gender rather than with respect to the sample used here. We do not use raw BMI when we classify the BMI status of children because raw BMI levels are interpreted differently for adults and children. For adults, BMI classifications are independent of age or gender, whereas for children aged between 2 and 20 years old, BMI needs to be interpreted relative to a child's age and gender, since the amount of body fat varies by age and gender (CDC, 2011).

Based on this classification, Table 2.1 to Table 2.7 present the transition probabilities of BMI status across generations in the CHNS (1989-2009), IFLS (1993-2007), British 1970 cohorts, HSE (1995-2010), NHANES (1988-1994), ENS-2006 (Spain) and ENCELURB (2002-2009) (Mexico), respectively. These transition probabilities describe the distribution of child's BMI status conditional on mother's BMI status, they are similar to transition matrices across the discretized bivariate distribution. The interaction terms between mother and child of different BMI status provides the matrix of intergenerational transition probabilities. For instance, in Table 2.1, the numbers in the first row of matrix indicate of the total number of children whose mothers were "underweight", 20.56 % were "underweight", 70.33 % were "normal", 4% were "overweight, and 5.11% were "obese". For mothers in the "underweight" category, 20.56 % of their children appear in the same category "underweight", and 70.33 % were in the "normal" category. Compared with other categories, there seems a stronger transmission of the same BMI status in the "underweight" category. In the case of Indonesia, Table 2.2 suggests there is a larger proportion of children in the "underweight" category, and a larger proportion of mothers in the "obesity" category. This distribution seems in line with the recent studies, which suggest a coexistence of "under nutrition" and "obesity" clustering within a single household ("dual burden households") in some developing countries, such as Indonesia (Doak et al. 2004). Moreover, we see there is a stronger intergenerational transmission of "underweight" (26.06%) in the IFLS sample compared to the CHNS sample.

In terms of the UK, as shown in Table 2.3 and 2.4, there is a significant greater fraction of mothers and children in the category of "overweight" and "obesity". Moreover, comparing Table 2.3 (based on BCS 1970 cohorts) and Table 2.4 (HSE sample), the fraction of "overweight" is larger for both mothers and children in the HSE (1995-2010) sample than in the BCS 1970 cohorts survey which follows the cohorts from the time when they were born (1970) up until they were 26 years old (1996). Considering the timing, Table 2.3 and 2.4 indicate an increasing proportion of "overweight" among adults and children over time from the period 1970-1996 to 1995-2010. In the case of the US, based on the NHANES3 sample (1988-1994), Table 2.5 suggests there is a large fraction of "overweight" and "obese" for both mothers and children. Similarly, Table 2.6 suggests a strong transmission of "obese" status (47.34%) from mothers to children. In the case of Mexico, Table 2.7 shows a strong transmission of "obese" status, given a large fraction of "obese" mothers (27.03%) in the mothers' BMI distribution. The relatively larger prevalence of “overweight” and “obese” compared to other developing countries is consistent with the fact there is a substantially rising trend of obesity in Mexico during the survey period (S. Leeder, et al. 2006).

Table 2.1: The transition probabilities of mother and child's BMI z-score in CHNS 1989-2009 (China)

Full sample (N= 14011)	BMI z-score		Child’s BMI status by BMI z-score (%)				Mother’s distribution	Observations
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother’s BMI status by BMI(%)	< 18.5	Underweight	20.56	70.33	4	5.11	6.47	900
	18.5-24.9	Normal	10.86	75.25	6.64	7.26	76.53	10,649
	25-29.9	Overweight	6.25	74.71	9.59	9.45	15.29	2,127
	>30	Obese	7.98	65.55	13.45	13.03	1.71	238
	Child’s distribution		10.73	74.68	7.04	7.55		
	Observations		1,493	10,391	979	1,051		13,914

Table 2.2: The transition probabilities of mother and child's BMI z-score in IFLS 1993-2007 (Indonesia)

Full sample (N= 18755)	BMI z-score		Child’s BMI status by BMI z-score (%)				Mother’s distribution	Observations
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother’s BMI status by BMI(%)	< 18.5	Underweight	26.06	64.51	2.91	6.52	9.32	1,719
	18.5-24.9	Normal	17.33	72.62	4.1	5.96	63.06	11,635
	25-29.9	Overweight	12.1	74.62	5.8	7.49	21.86	4,034
	>30	Obese	7.71	72.15	7.9	12.23	5.76	1,063
	Child’s distribution		16.44	72.27	4.58	6.7		
	Observations		3,034	13,335	845	1,237		18,451

Table 2.3: The transition probabilities of mother and child's BMI z-score in British Cohort Studies 1970 (UK)

Full sample			Child's BMI status by BMI z-score (%)				Mother's distribution	Observations
	BMI z-score							
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother's	< 18.5	Underweight	7.72	81.41	7.46	3.4	3.42	764
BMI status	18.5-24.9	Normal	3.85	80.31	9.54	6.31	72.36	16,142
by BMI(%)	25-29.9	Overweight	2.34	70.95	14.72	11.99	18.55	4,137
	>30	Obese	2.37	64.24	16.46	16.93	5.67	1,264
	Child's distribution		3.62	77.7	10.82	7.86		
	Observations		807	17,332	2,414	1,754		22307

Table 2.4: The transition probabilities of mother and child's BMI z-score in HSE 1995-2010 (UK)

Full sample			Child's BMI status by BMI z-score (%)				Mother's distribution	Observations
	BMI z-score							
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother's	< 18.5	Underweig	9.72	63.33	9.17	17.78	1.38	360
BMI status	18.5-24.9	Normal	2.02	59.29	14.21	24.49	46.21	12,092
by BMI(%)	25-29.9	Overweight	1.37	54.62	15.31	28.7	31.54	8,254
	>30	Obese	0.92	44.04	16.24	38.8	20.87	5,461
	Child's distribution		1.69	54.69	14.91	28.71		
	Observations		442	14,310	3,902	7,513		26,167

Table 2.5: The transition probabilities of mother and child's BMI z-score in NHANES 3 1988-1994 (US)

Full sample	BMI z-score		Child's BMI status by BMI z-score (%)				Mother's distribution	Observations
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother's	< 18.5	Underweight	1.44	47.12	24.04	27.4	3.19	208
BMI status	18.5-24.9	Normal	1.55	43.79	18.28	36.38	48.46	3,156
by BMI(%)	25-29.9	Overweight	0.89	40.65	17.99	40.47	25.95	1,690
	>30	Obese	0.55	33.1	16.24	50.1	22.4	1,459
	Child's distribution		1.15	40.69	17.93	40.23		
Observations			75	2,650	1,168	2,620		6,513

Table 2.6: The transition probabilities of mother and child's BMI z-score in ENS-2006 (Spain)

Full sample	BMI z-score		Child's BMI status by BMI z-score (%)				Mother's distribution	Observations
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother's	< 18.5	Underweight	10.53	43.42	14.47	31.58	2.26	76
BMI status	18.5-24.9	Normal	4.44	46.61	13.76	35.19	61.02	2,049
by BMI(%)	25-29.9	Overweight	2.85	42.92	16.21	38.01	26.09	876
	>30	Obese	2.24	36.41	14.01	47.34	10.63	357
	Child's distribution		3.93	44.49	14.44	37.14		
Observations			132	1,494	485	1,247		3,358

Table 2.7: The transition probabilities of mother and child's BMI z-score in ENCELURB (2002-2009) (Mexico)

Full sample	BMI z-score		Child's BMI status by BMI z-score (%)				Mother's distribution	Observations
			<-1.64	-1.64-1.04	1.04-1.64	>1.64		
		Category	Underweight	Normal	Overweight	Obese		
Mother's	< 18.5	Underweight	8.13	75.61	4.88	11.38	1.69	123
BMI status	18.5-24.9	Normal	2.82	75.63	10.86	10.69	33.17	2,413
by BMI(%)	25-29.9	Overweight	1.95	68.65	14.57	14.83	38.11	2,772
	>30	Obese	1.73	62.26	15.67	20.35	27.03	1,966
	Child's distribution		2.28	69.36	13.47	14.89		
Observations			166	5,045	980	1,083		7,274

In summary, these transition probabilities reveal a wide disparity in the joint distribution of mother and child's BMI status across countries. The CHNS sample suggests a stronger persistence of "underweight" between mothers and children in China; the IFLS sample shows a coexistence of "underweight" children and "obese" mothers ("nutrition transition paradox")[12] in Indonesia; there is a significantly larger fraction of mothers and children in the category of "overweight" and "obesity" in the UK, similar in the US and Spain; there is a relatively larger prevalence of "obesity" in Mexico compared to other developing countries. These transition probabilities show a global mobility of BMI status across the entire distribution, in particular reveal the prevalence of large movements in the BMI distribution from one generation to the next.

2.6 Empirical Results

2.6.1 Ordinary Least Squares Estimation

Applying equation (2.4), we estimate the IBE on data from the UK, USA, China, Indonesia, Spain and Mexico. As explained in section 2.4 we regress the log of child's BMI on the log of parents' BMI controlling for child's age, child's age squared, child's gender and the interactions between child's age and their gender. In each of these datasets we are able to control for many different family and parental covariates. We did estimate these models – but here we wanted to focus on a directly comparable equation specification which had the same form in each country. This meant that we had to drop various variables which were not in each dataset as we estimated the 'lowest common denominator' model. Our results – in terms of the sign and size of our main estimated parameter – the IBE – did not change appreciably – no matter what specification we adopted in each country separately when additional regressors were available. So here we focus only on the estimation results we can get for every country – in order that we can directly compare them.

It is clear from all our tables – that as we would expect the additional control variables are all significant with the logical and consistent relative size and signs of the coefficients. This is reassuring and means we can focus our attention on the parameter of interest – the

[12] Relatively, compared with developing countries.

IBE – with some confidence that the underlying relationship we have specified is the reasonable way to approach this estimation problem. Prior to considering the regression results from each country separately we would like to draw attention to our overall benchmark estimates reported in Table A 2.4 in the Appendix. These estimates, of an IBE of 0.2 for father-child and 0.189 for mother-child are the overall estimates derived from all of our combined cross country data. Since the dummy variables for each country are statistically significant then clearly we need to estimate our model separately, by country. In doing so we should be mindful of this benchmark estimate.

Table 2.6 reports the results on IBE when equation (2.4) controls for father's BMI variable alone. It suggests that the father-child IBE estimates range from 0.164 in Indonesian sample to 0.247 in Chinese sample, and they do not vary substantially across countries. For the UK, The IBE estimate on BCS sample (0.211) is close to that from HSE sample (0.198). These results suggest that the responsiveness of child's BMI variable to parents' BMI variable is around 0.20 and the extent of this "inheritability" is relatively constant across countries. In other words, if the father's BMI variable is 50% above the mean of their generation, on average his child's BMI variable would be around 20% above the mean of the children's generation, and this seems to be regardless of the general state of economic development in the country. In a similar way, Table 2.7 presents the mother-child IBE estimates from these samples, and we see a similar pattern as in Table 2.6 which was reported for the father-child IBE estimates. In addition, comparing Table 2.6 and 2.7, we can see that in general, the father-child IBE estimates are larger than mother-child IBE estimates.

Next, we incorporate both father and mother's BMI variables ($\log(BMI_{fi})$ and $log(BMI_{mi})$) into equation (2.4), and the results are reported in Table 2.8. As we expect, once we control for both father and mother's BMI variables, the sizes of paternal and maternal BMI effects shrink significantly compared with Table 2.6 and Table 2.7, with the dominance of father's BMI effects. One also needs to bear in mind that the R^2 of these regressions ranges from 0.232 to 0.553, which suggests that a substantial part of child's BMI is due to factors other than parental BMI. In the literature, studies often estimate the correlation between father or mother and child's BMI (the single parent version of equation (2.4)). When we include both father and mother into the estimation, one might need to be mindful of the assortative mating where obese men tend to form partnerships with obese women, we test this in section 2.6.2, the results suggest that the effects of mothers and fathers' BMI are mainly additively separable.

One important caveat that must be explained in the data we have available is that it all comes from different time periods in the different countries. Some of the data is fairly recent – so for example from China our last wave of data is from 2009. In contrast our data from the US – from NHANES is fairly old – it is from 1988. This means that in many respects true cross country comparisons should be tempered by this limitation. This aspect of our results should be factored into any relevant assessments. At the same time this feature of our results is also an advantage in demonstrating that our relative constant estimate of the IBE is applicable not only across countries but also over time.

Table 2.6: Intergenerational BMI elasticity between father and child across countries

	China	Indonesia	UK		US	Spain
	CHNS	IFLS	BCS	HSE	NHANES 3	ENS-2006
	(1989-2009)	(1993-2007)	(1970-1996)	(1995-2010)	(1988-1994)	(2006)
Dependent variable: Log (BMI of child)						
Log (BMI of father)	0.247***	0.164***	0.211***	0.198***	0.185***	0.212***
	(0.0116)	(0.0076)	(0.0093)	(0.0070)	(0.0124)	(0.0313)
Age of Child	-0.034***	-0.033***	-0.024***	-0.002**	-0.0027	-0.014***
	(0.0009)	(0.0009)	(0.0006)	(0.0010)	(0.002)	(0.0038)
(Age of Child)^2	0.003***	0.004***	0.002***	0.001***	0.002***	0.002***
	(4.97e-05)	(5.86e-05)	(2.34e-05)	(5.19e-05)	(0.0001)	(0.0002)
Male of Child	0.026***	0.040***	-0.031***	0.0178***	0.023***	-0.013
	(0.0048)	(0.0038)	(0.0040)	(0.0036)	(0.0055)	(0.0175)
Male*Age of Child	-0.002***	-0.005***	0.002***	-0.004***	-0.005***	0.0023
	(0.0005)	(0.0005)	(0.0003)	(0.0004)	(0.0009)	(0.0017)
Constant	2.084***	2.262***	2.185***	2.122***	2.148***	2.170***
	(0.0359)	(0.0234)	(0.0298)	(0.0232)	(0.0408)	(0.103)
Obs	14,061	18,570	21,505	26,316	6,515	2,139
R-squared	0.339	0.213	0.537	0.430	0.439	0.141

Robust standard errors in parentheses,*** p<0.01, ** p<0.05, * p<0.1

Table 2.7: Intergenerational BMI elasticity between mother and child across countries

	China	Indonesia	UK		US	Spain	Mexico
	CHNS	IFLS	BCS	HSE	NHANES 3	ENS-2006	ENCELURB
	(1989-2009)	(1993-2007)	(1970-1996)	(1995-2010)	(1988-1994)	(2006)	(2002-2009)
Dependent variable: Log (BMI of child)							
Log (BMI of mother)	0.213***	0.152***	0.184***	0.197***	0.169***	0.166***	0.112***
	(0.0109)	(0.0062)	(0.0075)	(0.0058)	(0.0093)	(0.0189)	(0.0091)
Age of Child	-0.0327***	-0.0334***	-0.0244***	-0.0026***	-0.0033*	-0.0060**	-0.0332***
	(0.0009)	(0.0009)	(0.0005)	(0.0009)	(0.0020)	(0.0027)	(0.0022)
(Age of Child)^2	0.0026***	0.0031***	0.0021***	0.0015***	0.0019***	0.0013***	0.0038***
	(4.99e-05)	(5.86e-05)	(2.31e-05)	(5.12e-05)	(0.0001)	(0.0002)	(0.0002)
Male Child	0.0289***	0.0418***	-0.0308***	0.0184***	0.0222***	0.0277**	0.0145***
	(0.0047)	(0.0037)	(0.0039)	(0.0036)	(0.0054)	(0.0126)	(0.0041)
Male*Age of Child	-0.0021***	-0.0055***	0.0024***	-0.0037***	-0.0046***	-0.0017	0.0003
	(0.0004)	(0.0004)	(0.0003)	(0.0004)	(0.0009)	(0.0013)	(0.0013)
Constant	2.182***	2.294***	2.284***	2.136***	2.209***	2.295***	2.471***
	(0.0336)	(0.0195)	(0.0237)	(0.0193)	(0.0306)	(0.0613)	(0.0297)
Observations	14,061	18,570	22,650	26,316	6,515	3,420	7,413
R-squared	0.333	0.216	0.542	0.445	0.449	0.163	0.094

Note: Spain uses the following, since only pairs of “father-child” or “mother-child” are available.

Table 2.8: Intergenerational BMI elasticity between father, mother and child across countries

	China	Indonesia	UK		US
	CHNS	IFLS	BCS	HSE	NHANES 3
	(1989-2009)	(1993-2007)	(1970-1996)	(1995-2010)	(1988-1994)
Dependent variable: Log (BMI of child)					
Log (BMI of father)	0.211***	0.128***	0.179***	0.161***	0.145***
	(0.0115)	(0.0076)	(0.009)	(0.0068)	(0.0124)
Log(BMI of mother)	0.174***	0.123***	0.162***	0.176***	0.146***
	(0.0106)	(0.0063)	(0.0076)	(0.0057)	(0.0095)
Age of Child	-0.034***	-0.0335***	-0.0241***	-0.0030***	-0.0038*
	(0.0009)	(0.0009)	(0.0006)	(0.0009)	(0.0020)
(Age of Child)^2	0.0027***	0.0031***	0.0021***	0.0015***	0.0019***
	(4.91e-05)	(5.82e-05)	(2.35e-05)	(5.04e-05)	(0.0001)
Male Child	0.0276***	0.0410***	-0.0324***	0.018***	0.0229***
	(0.0047)	(0.0038)	(0.0041)	(0.0035)	(0.0054)
Male*Age of Child	-0.002***	-0.0054***	0.0025***	-0.0037***	-0.005***
	(0.0005)	(0.0005)	(0.0003)	(0.0004)	(0.0008)
Constant	1.658***	1.990***	1.782***	1.680***	1.814***
	(0.0448)	(0.0269)	(0.0352)	(0.0276)	(0.0451)
Observations	14,061	18,570	21,246	26,316	6,515
R-squared	0.359	0.232	0.553	0.462	0.463

Robust standard errors in parentheses,*** p<0.01, ** p<0.05, * p<0.1

2.6.2 Robustness, Fixed Effects and Assortative Mating

It is clear that the data used for the estimation in this paper preclude the use of robust identification strategies like differences-in-differences, regression discontinuity design or other preferred, modern methods of identification. In this paper we mainly rely predominantly on cross-country, cross-section regressions. This means that the most natural question is – to what extent might the results be biased by measurement error, and endogeneity bias. These are difficult questions to answer at the best of times with even the most comprehensive data. It is even more challenging in the context of answering world-wide empirical questions which have not been attempted before. Hence the value added of the present paper is to report on these basic (conditional) correlations – which have such a policy importance – that we need to establish the benchmark of such a fundamental parameter.

Nonetheless – we can report some limited robustness checks in our data. For Indonesia we have good panel data and can use individual fixed effect estimation of our intergenerational transmission elasticity. These results are reported in appendix Table A 2.1. The results – not surprisingly – show an attenuation of our basic IBE – to around 0.11 – rather than 0.2. This is not surprising for two basic reasons. Firstly, the results condition basically for the unobserved heterogeneity at the level of the individual and hence – under the assumption of fixed unobservable heterogeneity across time allow us to estimate the enduring nature of this elasticity. The second factor is that the FE results report – to a large extent (when we have the majority of children observed only twice in the data) - the relationship between the *difference* in the adult parent BMI in consequetive time periods on the *difference* in the child's BMI across time. Such a relationship is, understandably not as strong as the raw correlation we report in our main tables. Notwithstanding these caveats – these results do present robust evidence of the presence of a strong correlation in the intergenerational process which is very comparable across countries.

A secondary concern is the extent to which our assumption that mothers and fathers' BMI each has an additively separable effect on child's BMI. One reason why they may not is that when men and women form partnerships there may be assortative mating with– for example – taller women being attracted to taller men and vice versa. To test this assortative mating, we examine the association of the BMIs of mothers and fathers by running the regression of log father's BMI on log mother's BMI, the results are presented in appendix Table 2.2, it suggests that there is a strong relationship between father and mother's BMI, this might imply some effects of "assortative mating" or effects of cohabitation which captures the correlation of environmental factors shared between spouses. To test whether there is a reinforcing force between father and mother's BMI in this intergenerational transmission, we first introduce simple interactions of father and mother's log BMI, the nature of the nonlinearity of the multiplication of two log values gave understandably strange results. Running the regression without taking logs destroys the elasticity interpretation we seek to use. Hence our solution is to use two dummy variables which relate to having both an underweight father and mother or both an overweight father and mother. We report these results in appendix Table A 2.3. For the most part we do not find large assortative mating effects – although there is a small positive effect of having both an overweight mother and father on child's BMI in the UK and the US and a small positive effect on child's BMI of having an underweight mother and father in Indonesia. The former finding is consistent with the overweight families in the western countries having an overweight child. Based on 7,834 obese probands and from 829 subjects randomly ascertained from the general Swedish population, Jacobson et al. (2007) find assortative mating for obesity is associated a higher risk of obesity in the next generation. The latter finding is consistent with regression to the mean in Indonesia. Including an assortative mating term does not detract from the size or significance of the IBE.

2.6.3 Quantile Estimation

Thus far, the estimates for IBE we have reported are at the conditional mean of child's BMI variable. To explore the variation of IBE across different quantiles of child's BMI distribution, we estimate the quantile elasticities of BMI between father and child at different points in the distribution of child’s BMI, using both parents version of equation (2.4)[13].

The results on father-child IBE [14] are displayed by country in Figure 2.6. We do not present the mother-child IBEs, which show a similar pattern. Figure 2.6 suggests that the degree of BMI transmission is an increasing function throughout child’s BMI distribution in all the samples. This means that the father-child IBE tends to be larger at higher levels of child‘s BMI. In other words, the effects of shared environmental and genetic factors between father and child tend to be larger for the fatter children. However, one needs to be mindful that the results here measure a correlation rather than a causality, since the estimates might be biased by factors such as omitted variables which affect the correlation differently for fatter and thinner children. Nonetheless, to keep the sample size, we do not include more controls.

One possible interpretation of our results is that there is a lower bound to this elasticity of about 0.1 which is more or less a constant at the lower end of the distribution for the thinnest children. This suggests that an IBE of 0.1 could be the lowest feasible value and hence a potential lower bound to what could be measured with a biological transmission mechanism. Any value above 0.1 of this mechanism could be caused by environmental or genetic factors. It is difficult to know what the actual causal underlying mechanism is here but it is difficult to conceive of a biological mechanism which would be higher for fat children than thin children. So – to the extent that a genetically inheritable trait is measured – then potentially the excess of the IBE over 0.1 for the fattest children could be informative.

One way of interpreting these results is to consider what they mean at different points in the child’s adiposity distribution. Take the case of China, at the 95th percentile of child’s adiposity the IBE estimates at the median is 0.30. The 95th percentile bounds of this estimate are 0.25-0.35. The corresponding estimate at the 5th percentile of children’s adiposity at the median is 0.125 and its 95th percentile confidence interval is 0.10-0.15.

[13] We estimate only the mother-child version of equation (2.4) in the Mexican data, as only pairs of mother and child are identifiable in this data; similarly, in the Spanish data, only pairs of father-child or mother-child are available, rather than the whole set of father, mother and child.

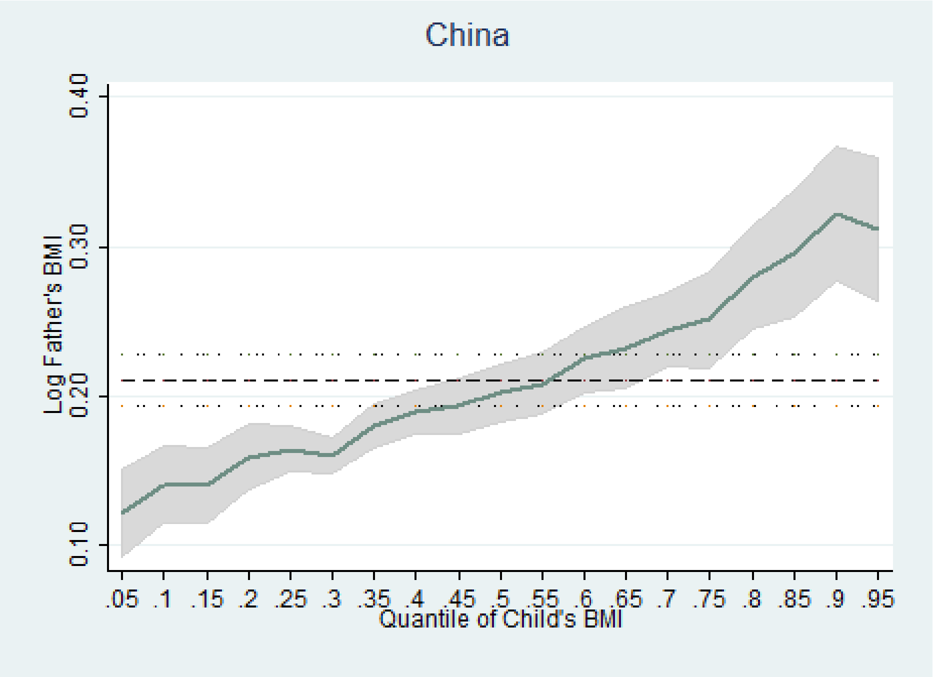
[14] Except for the Mexican data, since only pairs of mother and child are identifiable in this data. We use father-child version of equation (2.4) for the Spanish data, since in the Spanish data, only pairs of father-child or mother-child are available, rather than the whole set of father, mother and child.

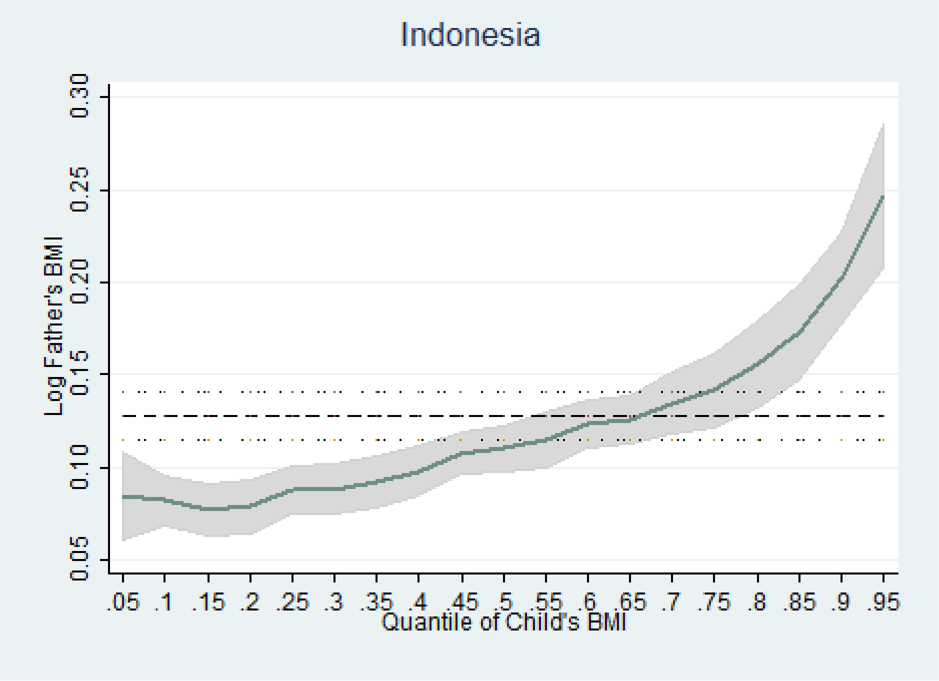
This suggests that the strength of the inheritability process is at least double for the fattest children as it is for the thinnest children.

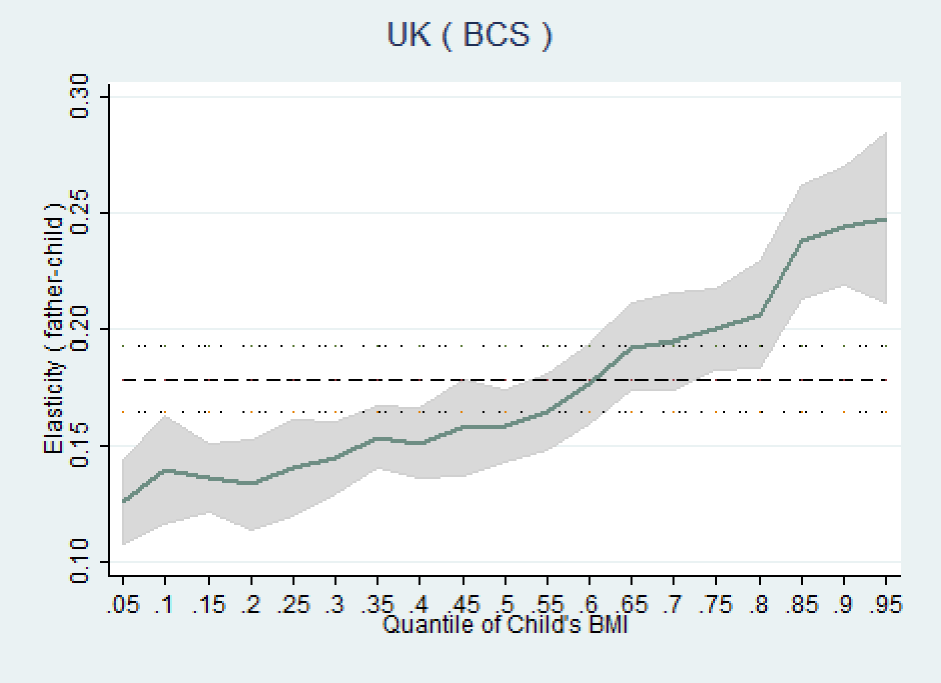
One may wish to hypothesize what the mechanisms might be for this underlying relationship – but a formal proof of any of these possible explanations is not possible with this data. Hence – what we wish to do here is just document and describe this relationship. For China there is limited evidence that the graph turns down slightly for the fattest children – but interestingly for the US the quantile plot turns down quite sharply after the 80th percentile. This indicates that the elasticity is actually falling for the fattest children. This suggests that maybe – in the US – children who are the fattest become that way more of their own accord. The most unusual country is Spain which seems to have a constant IBE across the whole range of children's BMIs.

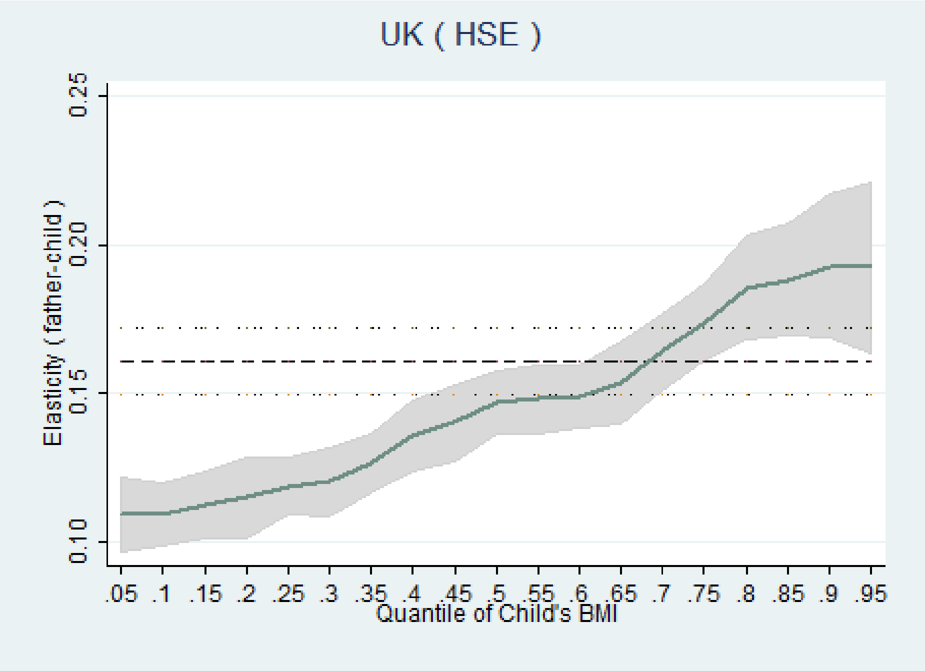
Looking more closely at each of the individual country figures in Figure 2.6 we see that the shape of the graph is quite different. For Indonesia the quantile plots rises at an increasing rate as we move from left to right to consider the fattest children. In contrast the graph for the UK and Mexico is rising monotonically. These figures, taken together, suggest that there is some cross country heterogeneity in the IBE quantile estimates across the distribution of children's adiposity. This may be related to the inherent heterogeneity across countries, or, to some extent, due to the era when the data was collected. Specifically, we should remember that US data is the oldest in that it relates to 1988-1994 and the position may have changed somewhat since then.

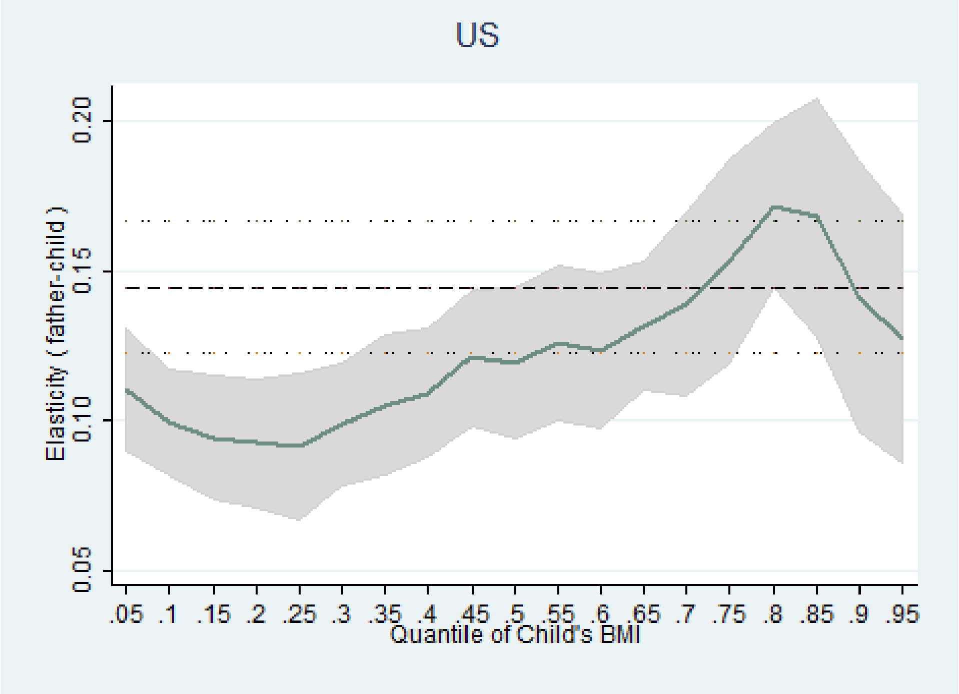
Figure 2.6: Quantile estimates of intergenerational BMI elasticity relative to OLS estimates by country

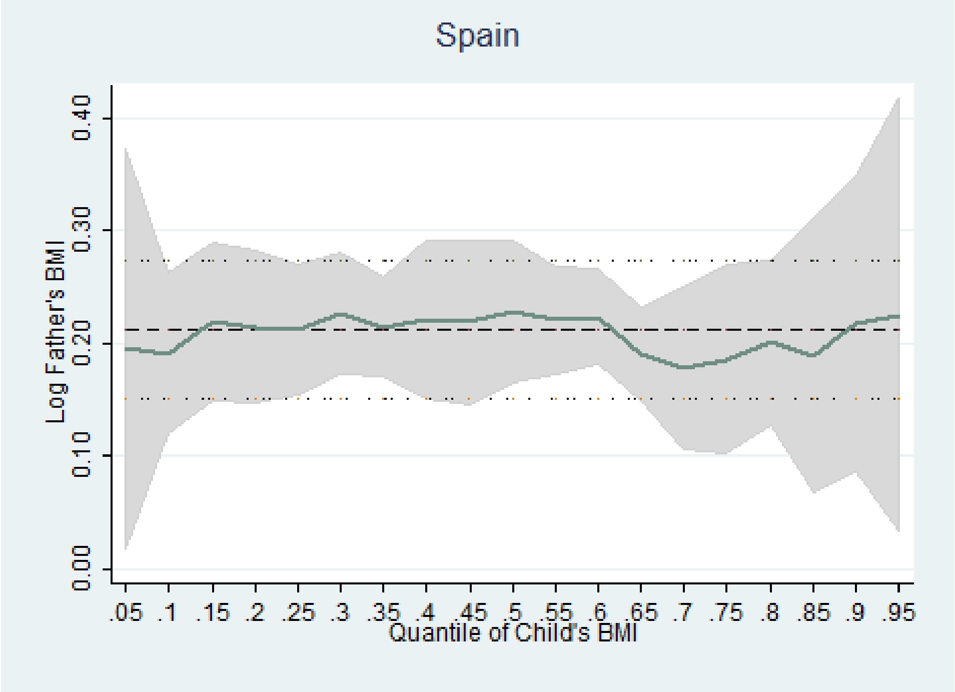


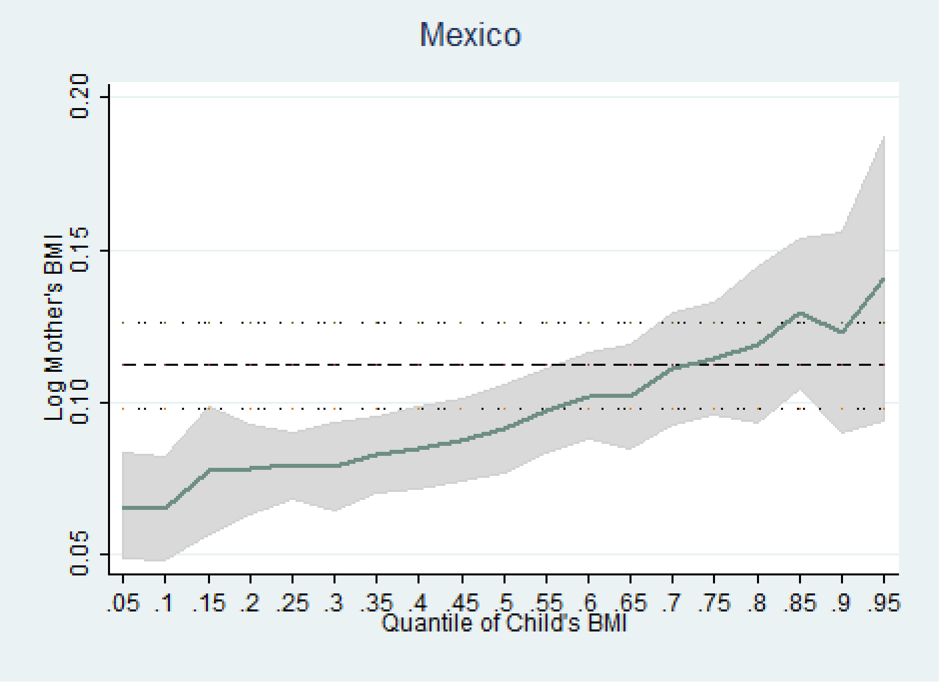



UK (BCS)
Elasticity (father-child)
0.10
0.15
0.20
0.25
0.30
.05
.1
.15
.2
.25
.3
.35
.4
.45
.5
.55
.6
.65
.7
.75
.8
.85
.9
.95
Quantile of Child's BMI


UK (HSE)
Elasticity (father-child)
0.10
0.15
0.20
0.25
.05
.1
.15
.2
.25
.3
.35
.4
.45
.5
.55
.6
.65
.7
.75
.8
.85
.9
.95
Quantile of Child's BMI


US
Elasticity (father-child)
0.05
0.10
0.15
0.20
.05 .1 .15 .2 .25 .3 .35 .4 .45 .5 .55 .6 .65 .7 .75 .8 .85 .9 .95
Quantile of Child's BMI


Spain
Log Father's BMI
0.00
0.10
0.20
0.30
0.40
.05 .1 .15 .2 .25 .3 .35 .4 .45 .5 .55 .6 .65 .7 .75 .8 .85 .9 .95
Quantile of Child's BMI



[15]

[15] Note: shaded area are 95% confidence intervals on estimates.

2.7 Conclusions and Policy Implications

This chapter has examined the intergenerational transmission of BMI or adiposity across generations in six countries across the world. Using the BMI, we find that the intergenerational transmission of adiposity is constant and comparable across time and countries – even if these countries are at different stages in their economic development. These intergenerational correlations determine a significant fraction of the child’s likely BMI as an adult. Since this effect is linear and additively separable for these two parents then we have found that the joint effect of the parents and its associated genetic makeup accounts for around 35% of the child’s likely BMI.

It must be emphasized that the results in this chapter mainly provides a descriptive picture of the intergenerational correlations in BMI. In terms of the channels through which this transmission operates, the fixed effects estimates might provide some evidence for rather short term environmental factors playing an important role, since the fixed-effects models yield significant elasticities of substantial magnitude.

Our second finding is that this intergenerational transmission mechanism is different across the distribution of children’s BMI. It is up to double for the fattest children as it is for the thinnest children. Specifically we find that over 30% of the fattest child’s BMI is determined by the mother and 25% by the father. Hence, jointly they account for over 50% of the fattest child’s likely BMI. In contrast, the corresponding fraction is around 30% for the thinnest child.

To sum up, our evidence from different countries’ data suggests that there is a strong consistency in the IBE estimates across countries. This consistency is different from what the previous studies find with respect to the intergenerational transmission of education or earnings, they found that there is a substantial disparity in the IIE and IEE estimates across different countries and different datasets. The difference of IBE from IIE and IEE hinges on the relative role and the interaction of environmental and genetic forces in the intergenerational transmission, our assumption is that in the transmission of BMI variable, a smaller fraction of the operation forces are open to manipulation (such as the diet change with the household---environmental change), and a larger fraction of the forces are driven by the “natural process”. If this hypothesis is true, our estimation for IBE may provide a lower bound of the intergenerational correlation in any characteristics including the income and education. In other words, assuming the intergenerational transmission of anthropometric outcome is entirely determined by the genetic traits, if our IBE is closer to the IIE in Scandinavian societies where the IIE is the lowest---0.2, it may imply that the relationship between parents and child cannot be lower than this threshold, in spite of the change in either family environment (such as the shift of nutrition pattern) or socioeconomic environment (such as the innovation or marketing campaign in food industry). The IIE in Scandinavian societies is the lowest partly due to the less nepotistic labour market institutional process compared to other societies, so the differences in IIE between these countries might be related to other things going on in the economy, such as the differences in the education and labour market systems. Another way to consider the IBE is to consider it as a process before the society added on this transmission process, therefore it is more reflective of the underlying nature of intergenerational transmission process.

This implication of our research is that it puts the emphasis firmly on the family in terms of understand the large fraction of adiposity determination. Specifically, we need to pay more attention to the inheritance from parents to child and what happens to the child when they are young, to explain a considerable fraction of what they become – as fat or thin adults. We have no way (with the data available to us) of splitting up the IBE into that which is due to genetic inheritance and that which is due to the family environment – but what we do know is that jointly these two influences determine a sizeable faction of what can happen to children. One way of thinking about this process is to suggest that – in the extreme – the thinnest child in the data –inherits 10% of their BMI from their parents – so that this is the lowest bound on how much may be due to the process of inheritance in anthropometric characteristics. Some fraction of the difference between their inheritance, and that of the fat child with a (combined) 0.55 elasticity, may still be due to biology, but it seems likely that this could be more to do with what goes on inside the family – namely how much exercise is taken, what the family diet is like and generally how active they are.

Chapter 3: The Intergenerational Transmission of BMI z-score in China

3.1 Introduction

China is in the throes of a nutritional and epidemiological transformation. Since 1978, the significant improvement of Chinese living standards has been sustained for decades (Klein and Ozmucur 2002). In rural China, the average nominal income in 2011 (6,977 yuan) is over ten times of that in 1990 (686 yuan); in urban China, the average nominal income increased more dramatically, from 1,510 yuan in 1990 to 19,118 yuan in 2011 (The National Bureau of Statistics of China 2012). With the transition of Chinese economy, more and higher quality food has become available, leading to a shift in the dietary habits, an improvement in the nutriental intake and energy composition (Du et al. 2002). The proportion of rice and wheat in the diet has decreased, whilst the share of pork has increased (Guo et al. 2000). More energy was consumed from fat, and an increasing proportion of energy-dense foods are consumed (Popkin 2001). Over the same period, the level of physical activity (such as cycling and walking) has been falling due to the rapid motorization (Bell et al. 2012). As a result, the health outcomes of Chinese population have undergone a dramatic transformation in the past 25 years. Evidence from the China Health and Nutrition Survey (CHNS) data (which is used in this study) shows that there is an increase in the average Body Mass Index (BMI). Particularly, there is a rightward shift in the whole distribution of BMI over the period from 1989 to 1997, the proportion of underweight population is declining, and the proportion of obesity is increasing. This observation is particularly pronounced amongst children and adolescents. For adults aged between 39 and 59, the fraction of underweight has decreased from 14.5% to 13.1%, and the overweight has increased from 6.4% to 7.7% (Popkin 2008). China, which used to have one of the leanest populations in the world, has over one fifth of all one billion obese people in the world nowadays (Wu 2006). In addition, with the rising income inequality (where the Gini coefficient is 0.61) in the economic transition process, there is also a substantial disparity across regions in the process of health transition (Morgan 2000). Given the intergenerational income correlation estimate of 0.6 in China (Gong 2012), the estimation of the intergenerational BMI correlation might help us to gain some insights into the investigation of intergenerational income mobility and the “natural level” of social mobility in China.

As we show in the first chapter, there is a relatively small variability in the magnitude of intergenerational BMI relationship across countries, and this relationship tends to be stronger at the fatter end of child's BMI distribution. In this chapter, we will take China as an example and use another measure-BMI z-score to conduct a more comprehensive analysis. In addition to estimates at the mean and across the distribution (which we already showed in the first chapter), this study makes three distinct contributions to the literature. First, since the CHNS data is a longitudinal dataset which records data on the same individuals repeatedly over time, then this panel structure allows us to explore the dynamic pattern of change in children's adiposity outcomes. We start by conditioning on their own BMI z-score in a previous time period. In addition, we differentiate out for the unobserved intra-household heterogeneity by controlling for household fixed effects, so that we can estimate the mechanism where the change in children's adiposity outcome is a function of the change in parents' adiposity outcome. This is different from most of the previous studies which mainly use levels of BMI in a cross section, taking no account of the potential endogeneity of child and parental BMI through the correlation of these measures with unobserved heterogeneity.

Second, the height and weight in our data are measured medically rather than self-assessed, which is rarely available in other longitudinal data. Self-assessed anthropometric measures tend to be biased, among which weight and BMI tend to be under-reported whereas height tends to be over-reported (Gorber et al. 2009), this may lead to an underestimation of BMI when it is based on the self-assessed data rather than the medically measured data. Therefore, that the height and weight in the CHNS data were recorded by the trained medical staff might help to improve the accuracy of our estimates of IBE (Spencer et al. 2002).

Third, in this chapter we also explore the heterogeneity of this intergenerational relationship of BMI z-score across different age groups of child and different levels of parental socioeconomic status. We find that this relationship tends to grow during the first half part of child's youth stage[16], and then declines until adulthood; we do not find a substantial variability in the magnitude of IBE with respect to different levels of family socioeconomic factors. In addition, though we control for the variable of child's age, the estimates might still be biased due to that the anthropometric system (eg. metabolic) of children is different from that of adults, therefor, in the section of quantile estimation, an attempt is made by using the subsample which

[16] The father-child IBE increases when the child's aged from 0 to 10 years old, whereas the mother-child IBE increases until the child's ages reached around 12 years old.

consists of children aged between 16 and 18 years old, based on the argument that the anthropometric development of children within this age range is similar to that of adults.

3.2 Literature Review

In the previous chapter, the literature review provides a general review on the intergenerational relationship of education, income and health, along with a brief review on the variability of this intergenerational BMI transmission across the BMI distribution. Since this chapter pays more attention to the variability of this intergenerational relationship with respect to different family socioeconomic factors ("environmental factors") and the evolution of this relationship with age. In this section, we will review the literature with an emphasis on these aspects.

One of the critical issues in estimating the intergenerational relationship is how to overcome the life cycle bias. In the circumstances of earnings, this issue relates to the fact that individual earnings vary over the lifecycle. A standard way to address this issue is to link the parents' earnings and the child's earnings when they are at a comparative age, the literature often uses their earnings when they both are at the middle age, though there are limited data which covers the adulthood of the child. Some studies also use the average earnings over the period covered in the data. In terms of the health outcomes such as BMI, some studies adopt a similar approach to overcome this life cycle bias. For instance, Classen (2010) estimates the intergenerational correlations of BMI between women and their children when both are aged between 16 and 24 years. By estimating the equation which simply includes mother and child's BMI, he finds that the measured intergenerational correlation of BMI between mother and child is around 0.35. However, studies also suggest that the correlation between child's BMI and parents' BMI in childhood does not differ substantially from the correlation between child's BMI and parents' BMI in adulthood. Based on two generations in the 1958 British birth cohort (parents' generation with BMI at 7,11,16,23, and 33 years old and a one-third sample of their children selected from 1991 aged 4-18 years old, Li et al. (2009) find that there is not a significant difference in the strength of the parental BMI in childhood and adulthood in the intergenerational BMI correlation.

In terms of the variability of the intergenerational transmission of BMI with respect to different family socioeconomic factors, this relates to the transmission mechanism which involves the interaction of genetic factors with environmental factors. The environmental channels mainly include the dependence of children's health outcome on their families' credit constraint or their parents' socioeconomic status, and the effects of parental health on child's health through family labour supply and therefore family income (Heckman and Carneiro 2003). The estimates for intergenerational correlation of income, education and BMI is a combination of the genetic and environmental effects parents and child share together. The "environmental effects" of parents' BMI on child's BMI may operate through the intra-household resource allocation. On one hand, equality in the allocation may be subject to cultural preference or economic incentive. For instance, in Asia, parents tend to prefer sons over daughters, low-birth order children over higher-birth-order children (Sen 1990, Dasgupta 1993), but this might vary with the bargaining power between fathers and mothers within the family, studies suggest an increase in the female income or income under the control of mother may increase survival rates or nutritional status of daughters, whereas increasing male income worsens the survival rates and the education attainment for daughters (Thomas 1990, Qian 2008). On the other hand, evidence also shows that disparities in intra-household food consumption are, in large part, explained by the disparities in health status, productivity and the energy consumption of the activities across household members (Pitt et al. 1990). To separate environmental effects from genetic effects, some studies compare "sibling mothers" or looking within "twin pairs of mothers", assuming twins share part of the same genetic make-up, therefore the correlation between parental health outcome and child's health outcome attributable to genetic factors can be purged of (Currie and Moretti 2005, Black et al. 2005, Royer 2009). Some studies use the data of twins or siblings, by incorporating siblings' genetic relationships into the intergenerational equation, Martin (2008) investigates the family effects of social characteristics on adolescent weight. Based on the National Health Interview Survey, the National Longitudinal Survey of Youth (NLSY) and the National Longitudinal Survey of Adolescent Health (Add Health), Thompson (2013) finds that 20-30% of the intergenerational correlations in health outcomes can be attributed to genetic mechanisms, by comparing the strength of transmission among biological children and adopted children; in addition, he finds that the intergenerational transmission of health does not differ substantially after including the proxies for environmental factors, such as the SES measures, health care access, health behaviour variables, cognitive test scores and other controls.

Studies suggest the intergenerational correlation of health outcomes tends to be stronger at lower levels of SES in both developed countries and developed countries. Based on a dataset on California births from 1960s to around 2005, Currie and Moretti (2005) find that children of low birth weight mothers are around 50% more likely to have low birth weight, and maternal poverty is strongly correlated with low birth weight of child, they argue that this intergenerational transmission of low birth weight is associated with the intergenerational transmission of low income (i.e. poverty cycle), since parent’s income affects child’s health, and child’s health affects their future education and earnings. In the study mentioned above, using the US data, Thompson (2013) also finds that the intergenerational transmission of health is stronger among families of low SES. In the setting of developing countries, based on individual survey data on 2.24 million children born to 600,000 mothers over the period from 1970 to 2000 in 38 developing countries, Bhalotra and Rawlings (2013) find children of shorter mothers or mothers with weaker health at birth are more sensitive to changes in the socioeconomic environment, and their survival rate is lower.

3.3 Data and Method

3.3.1 Data

The longitudinal data used in this study comes from eight waves (1989, 1991, 1993, 1997, 2000, 2004, 2006, and 2009) of the China Health and Nutrition Survey (CHNS)[17]. CHNS is conducted as a joint project of the Carolina Population Center at the University of North Carolina at Chapel Hill and China’s National Institute for Nutrition and Food Safety and the China Center for Disease Control and Prevention. It covers urban and rural areas of nine provinces that vary substantially in geology, economic development and public resources. This data covers health outcomes, demographic and anthropometric measures of all members of the sampled households, including medically measured heights and weights. It also includes information on social and economic indicators such as education, household income and labor market outcomes such as occupations. The CHNS sample was not representative of China but

[17] The CHNS is publicly available at http://www.cpc.unc.edu/projects/china

is designed to be randomly selected from households in eight provinces[18] --- Liaoning, Shandong, Henan, Jiangsu, Hubei, Hunan, Guizhou and Guangxi (from north to south). They employ a multistage, random cluster process to draw the sample in each of the provinces[19]. Our sample is restricted to children under 18 years old with information (especially anthropometric information) on both the biological father and mother. We choose age 18 as the threshold since this is the age used to distinguish between adult and child in the CHNS physical examination dataset where the anthropometric information is included. Additionally, children within this age range normally live with their parents and rely on their parents for nutritional intake and health care. As a result, this sample includes 14,077 person-wave observations made up by 6,044 children with both their fathers and mothers. In the data, there are 6,274 pairs of father-child and 6,747 pairs of mother-child. Therefore, given the large fraction of children living with both father and mother, the selection of children with both parents might not significantly affect our estimates. As in the previous chapter, we restrict the sample to those aged five years and older, applying equation (3.1), the results are presented in Table A 3.4. We see as in Chapter 2, the estimates for intergenerational correlation are larger than those based on the full sample (Table 3.1). This is consistent with the estimates we will show later when we analyze the correlation by age group, where the intergenerational correlation turns significantly larger on sub-samples aged above five. In terms of the attrition and response rate, based on the definition of response rates that those who participated in previous survey rounds remaining in the current survey, the response rates of this data were 88% at individual level and 90 % at household level. If the response rate is defined based on those who participated in 1989 and remained in the round 2006, then the rates were 63% and 69%, respectively (Popkin, 2010).

Different from the previous chapter, in this chapter we use BMI z-score rather than raw BMI, since BMI z-score reflects the relative position to the WHO reference population (the sample from WHO macro software) which adjusts for age and gender[20]. We also conduct the analysis using the raw BMI (as in the previous) chapter, the corresponding results are presented in Appendix 3 and they are consistent with the results using BMI z-score here. The conversion of z-score can be made using the 2006 WHO Growth Standards for preschool children and the

[18] In 1997 Liaoning was not able to participate and a new province-Hei Longjiang was added as a replacement, then Liaoning returned to the survey in 2000.

[19] See Popkin et al. (2010) for a detailed introduction on the CHNS survey.

[20] See Appendix for a description of BMI z-score and a discussion on BMI z-score and BMI.

2007 WHO Growth Reference for school age children and adolescents. In Stata, this conversion is implemented using a program from the WHO website[21].

3.3.2 Descriptive Statistics

This section presents some descriptive statistics of our sample. Since BMI z-score is adjusted for age and gender, this feature facilitates the comparison of adiposity distribution across groups of different age levels, in our case this is particularly true when we plot the distribution of adult parents' and child's adiposity distribution together, as we show in Figure 3.2 and Figure 3.3. In terms of the classification, we follow the WHO growth reference/standard (Wang and Chen, 2012) and use the following classification: underweight if BMI z-score <-1.04; normal if -1.04<=BMI z-score<1.04; overweight if 1.04<=BMI z-score<1.64; obese if BMI z-score>=1.64. This classification applies to both adults and children since BMI z-score accounts for age and gender, as mentioned earlier.

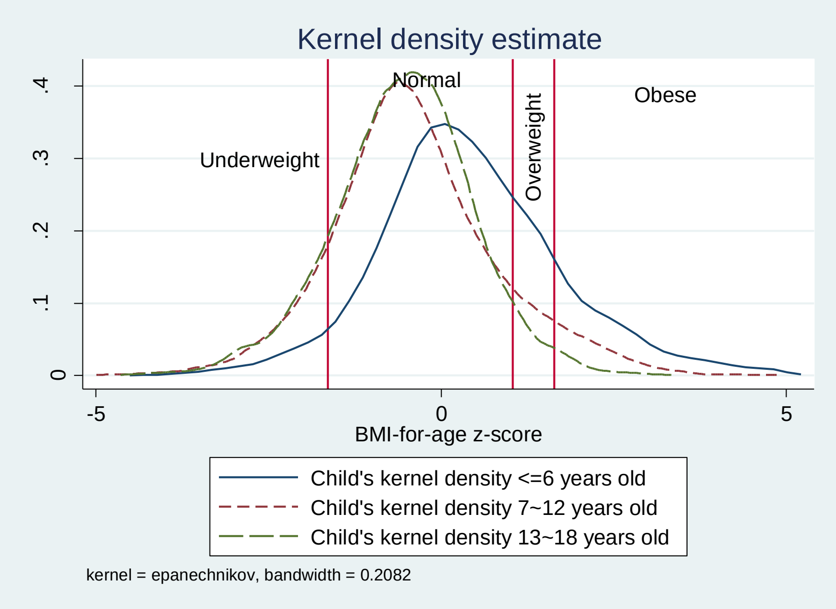
Figure 3.1 suggests the density of child's BMI z-score for children in the cohorts aged less than six, seven to twelve and thirteen to eighteen. The BMI z-score is normalised for age, these are the children who are less than six years old, seven to twelve and thirteen to eighteen, so Figure 3.1 shows a cohort effect. The distribution of child's BMI z-score shifts left-wards as their age increases because the child's BMI values in our sample tend to be lower than the reference population in the Anthro software, the average BMI of Asian population tends to be lower than that of non-Asian population (WHO 2004). This decline is relative to the external (world) reference population rather than the internal (our CHNS sample) reference population, and the left-wards shift of child's BMI density does not imply a decrease in the child's true BMI values as their age increase. When it comes to the intergenerational relationship of BMI z-score, Figure 3.2 and Figure 3.3 suggest that the distribution of child's BMI z-score shifts towards the left relative to their father and mother, this trend indicates a shift towards lower BMI z-score among children relative to their parents. The reverse case is found by Classen (2010) using NLSY79 data in the United States, he shows a shift towards higher BMI levels among children relative to their mother.

[21] They can be downloaded from http://www.who.int/childgrowth/software/en/ for Child growth standards (0~5 years old) and http://www.who.int/growthref/tools/en/ for Growth reference (5~19 years old).

There are three factors which explain these patterns. First, Figure 3.1 shows that growing children are getting fatter in the sense that the proportion of young children aged below six, who are obese is greater than the share of older children who are obese. Second, although the children's distribution, relative to their mother and father is shifting leftwards (due to the bias in the WHO reference group), it is the case that the fraction of children who are obese is larger than the fraction of mothers (Figure 3.2) or fathers who are obese (Figure 3.3).Third, the fraction of fathers and mothers who are obese is increasing over time. To see this, we cluster them by survey period (1989, 1991 and 1993, 1997 and 2001, 2004, 2006 and 2009) and find that the fractions of fathers and mothers who are obese have shifted to the right over time. Likewise, for children within the same age range (<=6, 7-12, and 13-18), the fraction of obesity is also increasing over time[22].

Figure 3.1: Distribution of child's BMI z-score by age group

(Using WHO Age and Gender Adjustment)

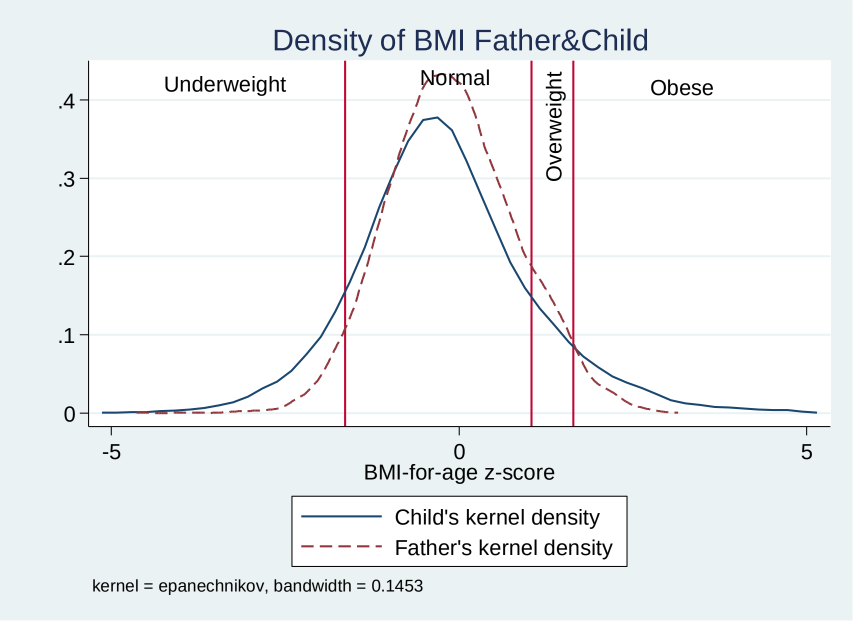


Source: own calculation

[22] These figures are available on request.

Figure 3.2: Distribution of father and child's BMI z-score

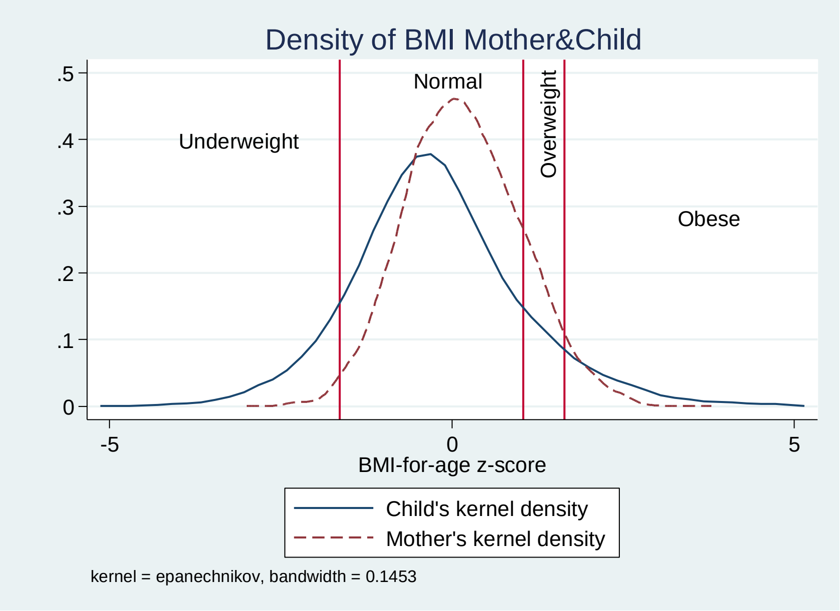
(Using WHO Age and Gender Adjustment)



Source: own calculation

Figure 3.3: Distribution of mother and child's BMI z-score

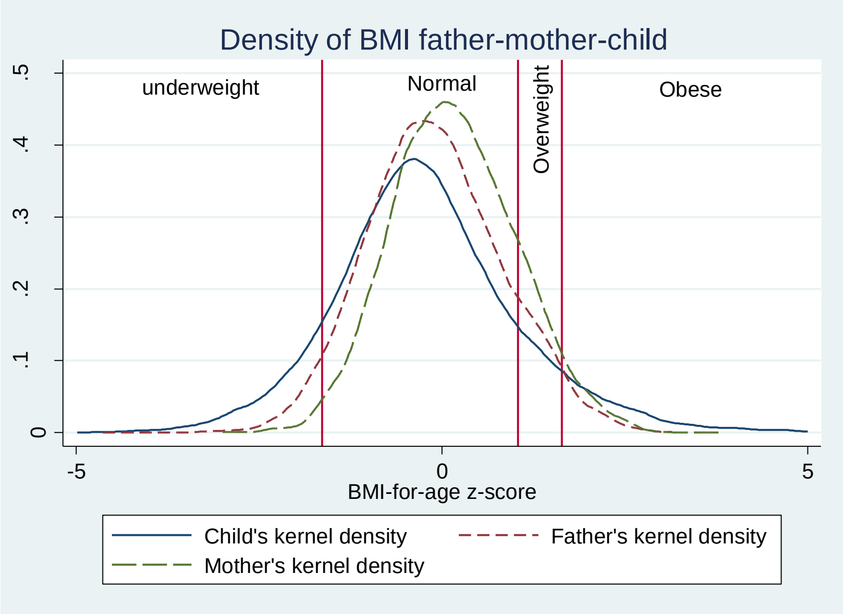
(Using WHO Age and Gender Adjustment)



Source: own calculation

Figure 3.4: Distribution of father, mother and child's BMI z-score

(Using WHO Age and Gender Adjustment)



Source: own calculation

3.4 Empirical Model

As in the previous chapter, the equation we mainly employ here is as follows:

$$y_i = \delta + \alpha \mathrm{y}_{fi} + \beta \mathrm{y}_{mi} + \gamma x_p + f_i + \varepsilon_i \tag{3.1}$$

where i indexes individual child observations and ε_i captures the transmitted stochastic error term. Therefore, child's health outcome y_i is a function of child i's father's health outcome, y_{fi}, and mother's health outcome, $y_{mi,}$ x_p denotes the age variables of father and mother, and f_i captures child i's age annual dummies, gender and the interaction term between them. Notice here we use age annual dummies rather than the continuous variable of BMI to account for the potential non-linear relationship between age and BMI z-score. Equation (3.1) is the baseline equation we use in this study, it derives the intergenerational correlation of parents' BMI z-score with child's BMI z-score. We recognize the possibility of a reverse causality from children's BMI to parents' BMI, however, we cannot identify this reverse causality.

In addition to the pooled OLS estimation, based on the longitudinal structure, we also investigate the simplistic dynamic pattern of child's BMI measure, equation (3.1) is estimated with the incorporation of lagged child's BMI z-score, $\mathrm{y}_{\mathrm{i,t-1}}$, in doing so we wish to net out for the individual unobserved heterogeneity .

$$y_{it} = \delta + \alpha \mathrm{y}_{ift} + \beta \mathrm{y}_{imt} + \gamma y_{i,t-1} + \varepsilon_{it} \qquad (3.2)$$

Where y_{it} denotes father's BMI z-score, y_{mt} denotes mother's BMI z-score, and $y_{i,t-1}$ represents the lagged value of child's BMI z-score.

As said in the literature, the IBE measures to what extent (i) biological/genetic factors and (ii) a shared environment learning contribute to intergenerational correlation of body weight from parents to their offspring. It does not tell us much on the mechanism in terms of how much is specifically attributable to genetic factors and how much of this correlation is specifically due to environmental factors. The longitudinal structure of CHNS data provides a possibility to differentiate out unobserved genetic effects and part of the unobserved environmental effects which are not changing over time, thus the fixed effects estimates mainly capture the effects of change in shared environmental factors (i.e. short term environmental factors).

First, as both child's health and parents' health are affected by time-invariant unobserved individual heterogeneity, f_i, such as genetic components and part of the environmental factors which are fixed over time. The panel structure of the data allows us to estimate equation (3.1) in an individual fixed effect framework.

$$y_{it} = \delta + \alpha y_{fit} + \beta \mathrm{y}_{mit} + \gamma x_{pt} + f_i + \varepsilon_{it} \qquad (3.3)$$

Where t denote observations referenced to a specific time period (or wave of the data). This equation takes into account an individual fixed effect f_i. It should be acknowledged that the fixed effects estimates do not condition out for factors which are variant over time, such as the eating habits which change over time, especially as children age. In addition, the pattern of food allocation among household members also varies with time. These patterns, together with, who is in control of the family income (Thomas 1990), who takes a larger share of energy-intensive activities (Pitt et al. 1990), whether the parents have a preference for sons (Qian 2008) or lower-birth-order children (Dasgupta 1993), can all affect both parental and children's BMI outcomes. Thus, household fixed effects are applied to estimate equation (3.1), i.e., fixed effects model is estimated using the following equation (3.4).

$$\mathrm{y}_{ijt} = \delta + \alpha \mathrm{y}_{fijt} + \beta \mathrm{y}_{mijt} + \gamma x_{pjt} + h_j + \varepsilon_{it} \qquad (3.4)$$

Where j indexes household observations. In household j, child i's health is a function of father's health, y_{fjt}, and mother's health, y_{mjt}, h_j denotes the household fixed effects. This equation can only be identified when we have data on siblings for which the f_i effects are distinct and the h_j are the same. We can estimate this model on the data as in a subset of this data there are more than one child in each household. In all of the estimation which follows our interest is on the α and β coefficients which respectively measure the relationship of BMI z-score for father-child and mother-child. Biological laws of nature define the both coefficients will be positive. What is at issue here is how large are, compared with the OLS estimates from equation (3.1).

In addition, we estimate equation (3.1) with respect to different levels of family social economic status, measured by family income, mother's education, father's occupation and the time duration when the family was in poverty, respectively; we estimate the IBE across different quantiles of child's BMI; we estimate the intergenerational correlation of BMI z-score by age group, and conduct quantile estimation on samples aged 16~18 years old to examine the probability that this relationship is structurally different when the children have become adults.

3.5 Empirical Results

3.5.1 Ordinary Least Squares (OLS) and Fixed Effects (FE) Estimation

3.5.1.1 OLS Estimation

Before presenting the results, it is necessary to clarify that the intergenerational relationship of BMI z-score cannot be used to isolate the genetic effects from the environmental effects[23]. Instead, it explores the role of common genes and environment in the intergenerational transmission of anthropometric tendency (Gruber 2009).

Table 3.1 presents the baseline pooled OLS estimates (with around 14,000 observations) in single-parent and both-parents version of equation (3.1). Column (1) shows the correlation between father's BMI z-score and child's BMI z-score when the regression only controls for father's BMI z-score, the coefficient of 0.223 suggests that one standard deviation increase in father's BMI z-score is associated with an increase of 0.223 in child's BMI z-score. Similarly, Column (2) suggests that the association between mother and child's BMI z-score in the sample is 0.208. The coefficients for child's age dummies are mostly negative, column (4) suggests that this intergenerational correlation appears stronger after we control for child's age and the interaction of child's age with their gender [24], this indicates child's BMI z-scores declines with age, which corresponds to Figure 3.1.

The results imply a marginally greater role for the environment and genes that a father and child share together than the mother and child share together, in the intergenerational transmission of BMI z-score. Column (3) shows that this result is robust when we control for both father and mother's BMI z-score, the magnitude of the coefficients on both the mother and father fall slightly. This result is counter to some studies in other countries, which find a stronger influence of maternal health status (eg, obesity) than paternal health on child's BMI (Anderson 2012), Anderson (2012) attributes this relative importance of mother's health to the fact that mother is usually the primary caregiver in the family responsible for the diet and health care of the child. However, some studies also find that this intergenerational correlation does not differ substantially with father or mother, using 4,654 complete parent–offspring trios in Avon Longitudinal Study of Parents and Children (ALSPAC), Smith et al.(2007) find that the correlation between parental BMI and children's BMI at age 7.5 was similar for both parents. As we discussed in the previous chapter, we do not find strong evidence on the interaction between father and mother's BMI variable. Therefore, we mainly use the both-parents version of intergenerational equation.

[23] Note that the intergenerational correlation of BMI z-score shares this property with the intergenerational correlation of income and education which cannot distinguish between 'inherited' factors from family and shared environment influences.

[24] In results not reported we included an interaction term in mother and father BMI which was always statistically insignificant.

Table 3.1: OLS estimates of the intergenerational correlation of BMI z-score

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: BMI z-score of child					
BMI z-score of father	0.223***		0.191***	0.246***	0.244***
	(0.0155)		(0.0152)	(0.0141)	(0.0141)
BMI z-score of mother		0.208***	0.166***	0.234***	0.235***
		(0.0174)	(0.0168)	(0.0166)	(0.0166)
Male of child				0.0224	0.0194
				(0.171)	(0.171)
Age dummies of child				Y	Y
Age dummies of child* Male of child				Y	Y
Age of father					-0.00654
					(0.00448)
Age of mother					0.00173
					(0.00503)
Constant	-0.190***	-0.248***	-0.222***	0.778***	0.917***
	(0.0147)	(0.0152)	(0.0147)	(0.122)	(0.143)
Observations	13,943	13,943	13,943	13,943	13,943
R-squared	0.027	0.021	0.039	0.188	0.189

Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Next we include household characteristics, regional dummies, year dummies and the interaction between them in the equation, the results are presented in Table 3.2. Column (1) suggest after we control for household socioeconomic factors, they include the type of father's occupation, the highest education degree of mother, the number of people in the household and the category of household income per capita, the estimates for father and mother's BMI effects decrease slightly compared with the baseline estimates (column (5) of Table 3.1). Column (3) suggests there is an appreciable decrease in the estimates for father and mother's BMI effects after we control for the fixed effects for child's province of residence, this is consistent with the prior studies that there is a substantial disparity in child's health status across regions in China. In column (4) we control for the time trend by including the survey year dummies in the estimation; in column (5) we control for provincial-varying time trends that may have occurred during the survey period used in this study. The results suggest that the inclusion of potential environmental factors does not significantly reduce the estimates of intergenerational BMI correlation, which is consistent with the prior studies (Thompson 2013). We also estimate this correlation between father and son, father and daughter, mother and son and mother and daughter, separately. The estimates are presented in Table A 3.5, they suggest the estimates are slightly larger than those based on the full sample (column (5) of Table 3.2), this is understandable due to that we include one parent only in the estimation.

Table 3.2: OLS estimates of the intergenerational correlation of BMI z-score with more controls

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: BMI z-score of child					
BMI z-score of father	0.237***	0.235***	0.202***	0.200***	0.198***
	(0.0171)	(0.0173)	(0.0173)	(0.0175)	(0.0175)
BMI z-score of mother	0.221***	0.221***	0.194***	0.193***	0.191***
	(0.0200)	(0.0201)	(0.0192)	(0.0193)	(0.0195)
Household characteristics	Y	Y	Y	Y	Y
Year fixed effects		Y		Y	Y
Province fixed effects			Y	Y	Y
Province*Year					Y
N	9,536	9,536	9,536	9,536	9,536
R-squared	0.193	0.194	0.213	0.214	0.224

Note: the regression also includes: age dummies of child, gender of child and the interactions between them, father and mother's age. Household characteristics include the category of household income per capita, household size , the type of father's occupation , the highest education degree mother attained. "Province*Year" are interactions of child's residential province dummies with the survey year. Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

In Table 3.3 we turn to a basic (flexible) first difference model in which we control for the child's BMI z-score at a previous time period [25]. This reduces the sample to 7,918 observations. Here, the magnitude of these correlations understandably shrink slightly (to around 0.17) when equation (3.1) includes the child's lagged BMI z-score, where the coefficient of this term is around 0.34 (column (2)). This result shows that the change of child's BMI status is strongly correlated with his/her BMI status in the previous year, i.e. there is a strong persistence in child's BMI over time. Column (3) and (4) suggest this result is robust to the inclusion of household characteristics, regional and time effects.

[25] Nickell (1981) suggests that a 'quasi-fixed effects estimate with a lagged dependent variable may downwardly bias the coefficient of the lagged dependent variable. But this coefficient is not our central concern in this study.

Table 3.3: OLS estimates of the intergenerational correlation of BMI z-score, controlling for the lagged value of Child's BMI z-score

Dependent Variable: BMI z-score of child	(1)	(2)	(3)	(4)
Lagged BMI z-score of child	0.387***	0.340***	0.335***	0.328***
	(0.0114)	(0.0125)	(0.0147)	(0.0151)
BMI z-score of father	0.171***	0.198***	0.201***	0.181***
	(0.0135)	(0.0139)	(0.0174)	(0.0177)
BMI z-score of mother	0.146***	0.176***	0.172***	0.160***
	(0.0160)	(0.0171)	(0.0215)	(0.0216)
Age dummies of child, gender of child and the interactions between them, father and mother's age.		Y	Y	Y
Household characteristics			Y	Y
Year fixed effects, Province fixed effects Province*Year effects				Y
Constant	-0.334***	1.452***	1.043***	0.958***
	(0.0131)	(0.361)	(0.146)	(0.262)
Observations	7,918	7,918	5,484	5,484
R-squared	0.243	0.278	0.276	0.298

Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

3.5.1.2 Panel Fixed Effects (FE)

Next we use the longitudinal element of the data and model a child's BMI z-score over the life course of their childhood (or whatever part of it we observe) between 1989 and 2009. In doing so we are able to - by turns - control for individual unobserved heterogeneity and family specific unobserved heterogeneity. Our results are presented in the Table 3.4 and Table 3.5, respectively. Care needs to be taken in interpreting these results in comparison to Table 3.1. In Table 3.1 we report simple correlations taking no account of the panel element of the data - treating all observations occurring at any point in time as independent. In contrast, the panel estimates specify the dynamic underlying intergenerational correlation of BMI z-score over the childhood life course after having netting out for family and individual unobserved heterogeneity.

The individual fixed effects estimates from equation (3) are provided in Table 3.4. It suggests that the intergenerational relationship of BMI z-score remains significant after the regression controls for time-invariant unobserved individual characteristics. The individual fixed effect estimates (0.151 and 0.160) are lower than the previous pooled OLS estimates (0.244 and 0.235), this indicates a potential upward bias in the pooled OLS estimates due to the omission of unobserved individual heterogeneity.

Table 3.4: Individual Fixed Effects Estimates of the intergenerational correlation of BMI z-score

	(1)	(2)	(3)
Dependent Variable: BMI z-score of child			
BMI z-score of father	0.151*** (0.0301)	0.151*** (0.0377)	0.138*** (0.0374)
BMI z-score of mother	0.160*** (0.0350)	0.135*** (0.0383)	0.118*** (0.0403)
Age dummies of child, gender of child and the interactions between them, father and mother's age.	Y	Y	Y
Household characteristics		Y	Y
Year fixed effects, Province*Year effects			Y
Constant	3.533** (1.482)	2.549 (1.741)	-5.511** (2.752)
Observations	13,943	9,536	9,536
R-squared	0.159	0.154	0.186
Number of individuals	6,027	4,341	4,341

Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Applying equation (3.4), the household fixed effects estimates in Table 3.5 suggest that the household fixed effect estimates are close to the individual fixed effect estimates. As discussed earlier in the previous chapter, compared with the OLS estimates, these fixed effects estimates provide some information on the short-term environmental effects through differentiating out the unobserved genetic effects and the long-term environmental effects which are assumed invariant with time. In other words, OLS estimates provide a descriptive correlation in BMI, if we want to say something on the mechanism- separating different channels of intergenerational transmission, the fixed effects provide some evidence for the effects of change in environmental factors (i.e. short term environmental effects), such as change in the dietary and the type of transport.

More specifically, the individual effect results indicate the effects of individual specific parents on the individual specific child, and the household fixed effect results are more likely to be associated with the difference between child i and his/her sibling j in terms of the way they are treated, the longer the age gap between child i and j, the greater differences in terms of the way they are treated, and the more likely that what the household fixed effect results capture is accounted for by these differences. In other words, the family fixed effect results indicate the effects of the difference between child i and j on child i's BMI. Therefore, the subsample of children with siblings may be different from the full sample[26].

Table 3.5: Household Fixed Effects Estimates of the intergenerational correlation of BMI z-score

	(1)	(2)	(3)
Dependent Variable: BMI z-score of child			
BMI z-score of father	0.152***	0.157***	0.138***
	(0.0273)	(0.0352)	(0.0357)
BMI z-score of mother	0.152***	0.130***	0.117***
	(0.0291)	(0.0349)	(0.0349)
Age dummies of child, gender of child and the interactions between them, father and mother's age.	Y	Y	Y
Household characteristics		Y	Y
Year fixed effects, Province*Year effects			Y
Constant	0.568***	0.509	0.792**
	(0.194)	(0.310)	(0.330)
Observations	13,943	9,536	9,536
R-squared	0.166	0.166	0.187
Number of households	3,708	2,917	2,917

Notes: standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

[26] The identification of household fixed effects is coming off the 47.18% of sample which have more than one child, of which 39.18% have two children, 11.63% have three children and 1.53% have four children. It is possible there might be some sample selection problem in the household fixed effects estimation, due to the one-child policy in China. We examined this and found no evidence that the decision of having the second or more children might be associated with health status of the first child. We also found evidence that the one child policy was not rigorously enforced in rural areas during the period when the CHNS was collected.

To summarize, the fixed effects estimates suggest when we account for unobserved time-invariant factors, the magnitude of this intergenerational correlation estimates drop by a significant amount. This is consistent with the prior studies applying fixed effects model to the estimation of intergenerational health transmission. Coneus and Spiess (2012) use fixed effects as a robustness check for their cross-section estimates, and they find most of their cross-section estimates are robust when they control for the fixed effects.

3.5.2 Estimation by Family Socioeconomic Group

So far in our analysis, we have focused on the estimation of the intergenerational BMI correlation at the conditional mean. In this section, we explain the heterogeneity of this effect by family income, mother's education, father's occupation, and the poverty status of the family. We then investigate the intergenerational correlation of BMI across the distribution of children's BMI by using quantile estimation.

To test whether the intergenerational correlation of BMI z-score varies with the parental socioeconomic factors, we include the interactions of father and mother's BMI variable with different measures of socioeconomic factors. We use three main variables as the measure of family socioeconomic factors: the quartiles of household income per capita relatively within the CHNS sample; the proportion of time that the family was in poverty during the survey years, the family was classified as being in poverty if the income per household member inflated to 2009 was below the world poverty line in 2009 (3,100 yuan per capita per year[27]); the highest education degree of mother, lower education levels (primary school and below) middle education levels (lower middle school and upper middle school) and higher education levels (technical school , college and beyond); the type of father's occupation, farmers, workers (skilled/non-skilled/service worker and other) and professionals (Professional/technical/ administrator/executive/manager/office). The results are presented in Table 3.6. Column (1) displays the estimates of the correlation between mother and child's BMI z-score when the equation includes the interactions between mother's BMI z-score and different socioeconomic factors. Model 1 includes the interactions between mother's BMI z-score and the quartiles of household income per capita; model 2 includes the interactions between mother's BMI z-score

[27] The world poverty line here is calculated according to the poverty line 1.25 dollars/ day from the world bank in 2008.

and mother's education level, and model 3 includes the interactions between mother's BMI z-score and the indicators of family poverty duration. The results suggest that the relationship between mother and child's BMI z-score does not vary substantially with the socioeconomic indicators. In terms of the variability in the correlation between father and child's BMI z-score, column (2) shows similar results as column (1), the correlation between father and child's BMI z-score barely varies with the socioeconomic indicators. In summary, the intergenerational correlation of BMI z-score does not vary substantially by socioeconomic factors used here.

Table 3.6: OLS estimates of the intergenerational correlation of BMI z-score, including the interactions between parental BMI z-score and family socioeconomic factors

	(1) Mother-child		(2) Father-child
Dependent variable : BMI z-score of child			
Model1		**Model1**	
BMI z-score of mother	0.200*** (0.0389)	BMI z-score of father	0.205*** (0.0389)
Income quarter: (Ref.: 0-25th percentile of Income)		Income quarter: (Ref.: 0-25th percentile of Income)	
25-50th percentile of	-0.0337 (0.0397)	25-50th percentile of	-0.0415 (0.0427)
50-75th percentile of	-0.0385 (0.0418)	50-75th percentile of	-0.0490 (0.0435)
>75th percentile of Income	-0.0294 (0.0469)	>75th percentile of	-0.0332 (0.0473)
25-50th* BMI z-score of mother	-0.00408 (0.0472)	25-50th* BMI z-score of father	-0.0325 (0.0468)
50-75th* BMI z-score of	-0.0425 (0.0460)	50-75th* BMI z-score of	-0.0346 (0.0457)
>75th* BMI z-score of	0.0124 (0.0528)	>75th* BMI z-score of	0.0317 (0.0465)
Observations	9,420	Observations	9,420
R-squared	0.225	R-squared	0.225

	(1) Mother-child		(2) Father-child
Dependent variable : BMI z-score of child			
Model 2		**Model 2**	
BMI z-score of mother	0.201***	BMI z-score of father	0.172***
	(0.0297)		(0.0266)
Highest education degree obtained: (Ref.: Primary and below)		Occupation: (Ref.: famer)	
High school	0.0481	Skilled/non- skilled/	0.0210
	(0.0357)	service worker and other	(0.0350)
Technical and Tertiary	0.161**	Professional/technical/	0.0662
	(0.0713)	administrator/ executive/manager/ office	(0.0488)
High school* BMI z-score	-0.0150	Skilled/non-	0.0496
of mother	(0.0380)	skilled/service worker and other* BMI z-score of father	(0.0344)
Technical and Tertiary*	-0.0524	Professional/technical/	0.0146
BMI z-score of mother	(0.0700)	administrator/executive/ manager/office* BMI z-score of father	(0.0442)
Observations	9,536	Observations	9,439
R-squared	0.224	R-squared	0.224

	(1) Mother-child		(2) Father-child
Dependent variable : BMI z-score of child			
Model 3		**Model 3**	
BMI z-score of mother	0.212***	BMI z-score of father	0.155***
	(0.0327)		(0.0330)
The proportion of time in		The proportion of time in	
50-75% of time in poverty	-0.0381	50-75% of time in poverty	-0.0246
	(0.0518)		(0.0537)
1-50% of time in poverty	-0.00639	1-50% of time in poverty	-0.0117
	(0.0492)		(0.0498)
Never in poverty	0.0671	Never in poverty	0.0698
	(0.0640)		(0.0637)
50-75% of time in poverty	-0.0390	50-75% of time in poverty	0.0706
*BMI z-score of mother	(0.0632)	*BMI z-score of father	(0.0583)
1-50% of time in poverty	-0.0412	1-50% of time in poverty	-0.0186
*BMI z-score of mother	(0.0539)	*BMI z-score of father	(0.0503)
Never in poverty* BMI	-0.0251	Never in poverty* BMI	0.0682*
z-score of mother	(0.0416)	z-score of father	(0.0383)
Observations	9,536	Observations	9,536
R-squared	0.225	R-squared	0.225

Note: the regression also includes: age dummies of child, gender of child and the interactions between them, father and mother's age, the category of household income per capita, household size, the type of father's occupation, the highest education degree of mother, provincial fixed effects, year fixed effects and the interactions between them. Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

As another test for the variability of intergenerational BMI z-score correlation with respect to family socioeconomic factors, we estimate this correlation in sub-samples divided with respect to these indicators. The results are presented in Table 3.7. Using the quartiles of household income per capita, the correlation between mother and child's BMI z-score ranges from 0.158 in the third quartile to 0.240 in the second quartile; using sub samples divided by mother's education levels, the results suggest there is a stronger correlation between mother and child's BMI z-score at the lower levels (primary and below) and higher levels (technical and tertiary) of mother's education; the results from sub-samples divided by poverty duration provides a similar pattern: the correlation between mother and child's BMI z-score tends to be higher for families that were observed in poverty for the longest time (15-100%) and never in poverty than those that were in poverty for 1-50% and 50-75% of the time. Therefore, based on sub-samples divided with respect to three socioeconomic indicators, the correlation between mother and child's BMI z-score seems stronger for the poorer and richer families, though this pattern is not particularly strong.

Table 3.7: Intergenerational correlation of BMI z-score between mother and child by SES measures: by income level, mother's education and world poverty line

	Sample	Coefficient	Std. Error	R-squared
<25th percentile of Income	1,777	0.186***	0.0433	0.289
25-50th percentile of Income	2,162	0.240***	0.0353	0.280
50-75th percentile of Income	2,621	0.158***	0.0319	0.264
>75th percentile of Income	2,860	0.217***	0.0398	0.225
Primary school and below	3,066	0.211***	0.0301	0.276
lower and upper middle school	5,870	0.190***	0.0253	0.222
Technical and Tertiary	600	0.231***	0.0702	0.299
75-100% of time in poverty	2,577	0.226***	0.0344	0.271
50-75% of time in poverty	1,009	0.161***	0.0580	0.360
1-50% of time in poverty	1,091	0.171***	0.0486	0.285
Never in poverty	4,859	0.188***	0.0280	0.220

Note: the regression also includes: age dummies of child, gender of child and the interactions between them, father and mother's age, the category of household income per capita, household size, the type of father's occupation, the highest education degree of mother, provincial fixed effects, year fixed effects and the interactions between them. Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Similarly, we estimate the correlation between father and child's BMI z-score in sub samples divided by different socioeconomic indicators. The results are given in Table 3.8. They suggest that measured by the quartile of household income, the correlation between father and child's BMI seems slightly stronger for families at the lower income level (<25th percentile of income) and higher income level (>75th percentile of income) than for those at the middle levels, but we do not see a pattern in this correlation when the sample is divided with respect to the type of father's occupation and poverty durations. To summarize, the intergenerational correlation of BMI does vary significantly with the socioeconomic indicators used in this study.

Table 3.8: Intergenerational correlation of BMI z-score between father and child by SES measures: by income level, father's occupation and world poverty line

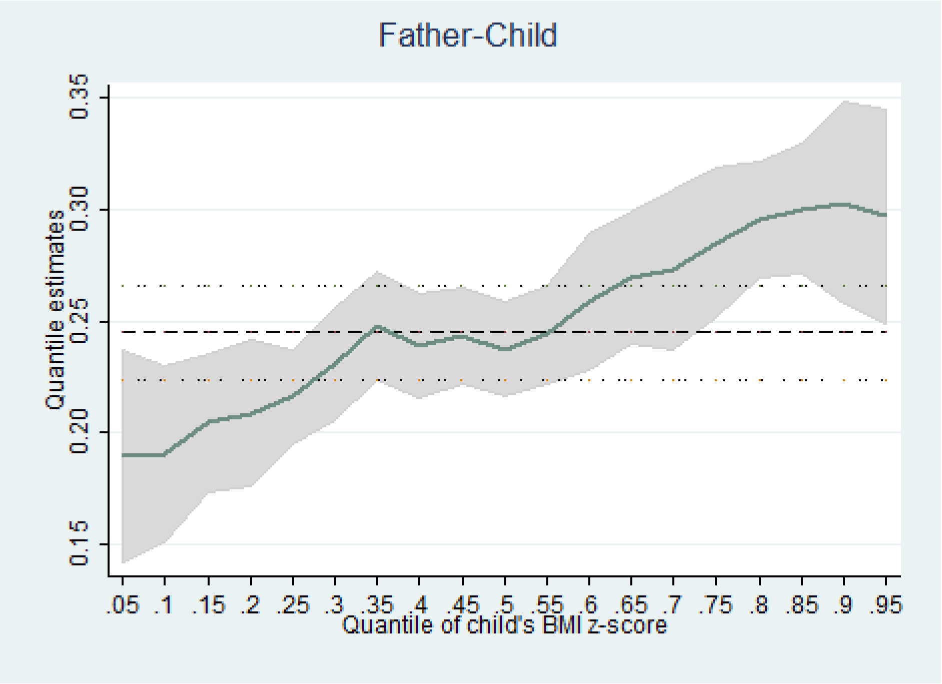
	Sample	Coefficient	Std. Error	R-squared
<25th percentile of Income	1,777	0.197***	0.0428	0.289
25-50th percentile of Income	2,162	0.178***	0.0351	0.280
50-75th percentile of Income	2,621	0.180***	0.0286	0.264
>75th percentile of Income	2,860	0.228***	0.0285	0.225
Farmer	4,126	0.172***	0.0271	0.272
Skilled/non-skilled/service worker and other	3,670	0.215***	0.0261	0.233
Professional/technical/ administrator/executive/manager/office	1,643	0.158***	0.0389	0.224
75-100% of time in poverty	2,577	0.165***	0.0341	0.271
50-75% of time in poverty	1,009	0.204***	0.0559	0.360
1-50% of time in poverty	1,091	0.126***	0.0444	0.285
Never in poverty	4,859	0.219***	0.0221	0.220

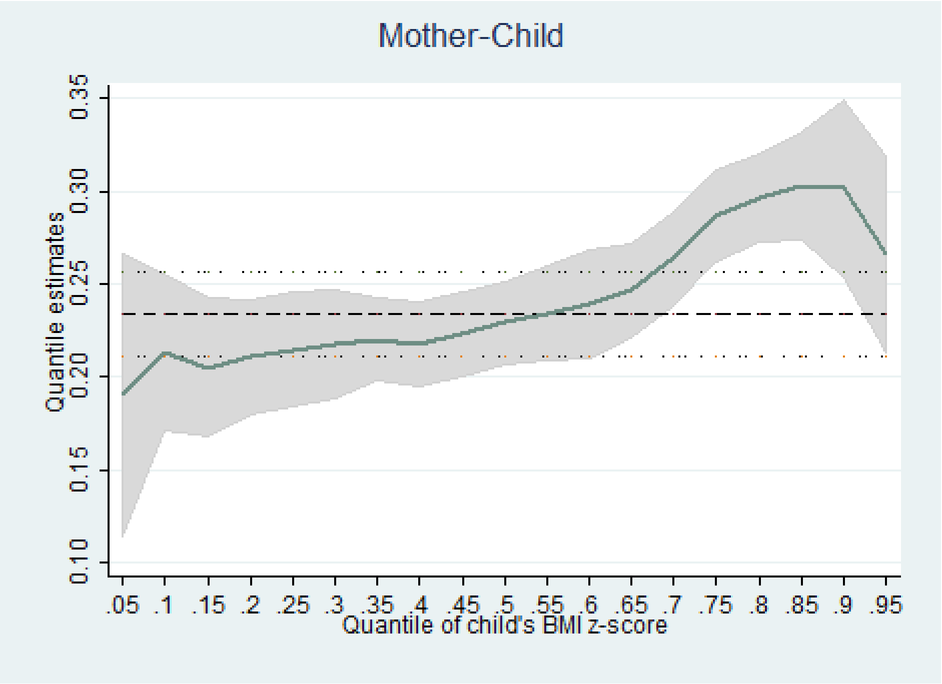
Note: the regression also includes: age dummies of child, gender of child and the interactions between them, father and mother's age, the category of household income per capita, household size, the type of father's occupation, the highest education degree of mother, provincial fixed effects, year fixed effects and the interactions between them. Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

3.5.3 Quantile Estimation

Thus far, the intergenerational correlation of BMI z-score are estimated at the conditional mean of child's BMI. It is likely this relationship varies significantly for children at the thinner end and fatter end of the distribution, therefore, next we apply quantile estimation to explore the variation of this intergenerational relationship across the distribution of child's BMI z-score.

Figure 3.5: Quantile estimates of the intergenerational correlation of BMI z-score relative to OLS estimates[28]





[28] Shaded area is 95% confidential intervals.

Based on equation (3.1), Figure 3.5 shows quantile estimates of the relationship between father and child's BMI z-score across the distribution of child's BMI z-score in the sample. We see this relationship increases throughout the quantiles of child's BMI, with coefficients at the median (around 0.23) close to coefficients at the mean (the OLS estimates, 0.24 in column (5) of Table 3.1), the correlation between father's and child's BMI z-score is around 0.31 at the fattest end of the distribution (the 90th), and around 0.18 at the thinnest end of the distribution (the 5th). A similar pattern emerges when we apply quantile estimation to estimate the relationship between mother and child's BMI z-score, with a slightly lower magnitude than in the case of father and child's BMI correlation. Therefore, the estimates of intergenerational dependence in BMI tends to be stronger among children of higher BMI levels, these results imply that the common environmental and genetic factors shared between parents and child tend to have a larger effect on children of higher BMI.

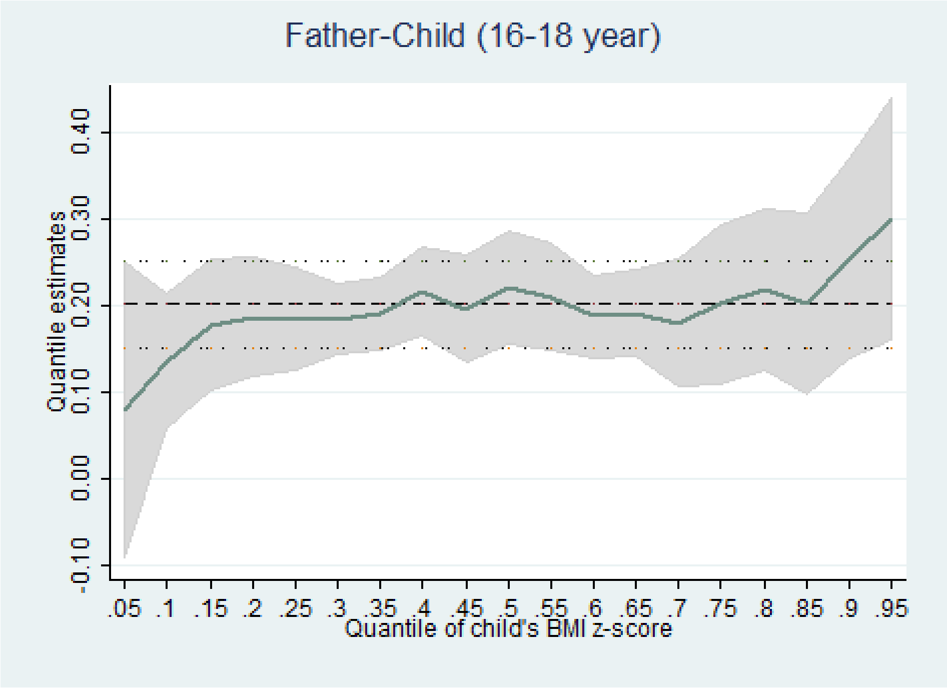
These results are of general interest in that they suggest that the transmission of "obesity" is a trait that is much more strongly transmitted across generations for families with fatter children. Those children with the highest adiposity, who are fat, are much more likely to have inherited this from their parents.

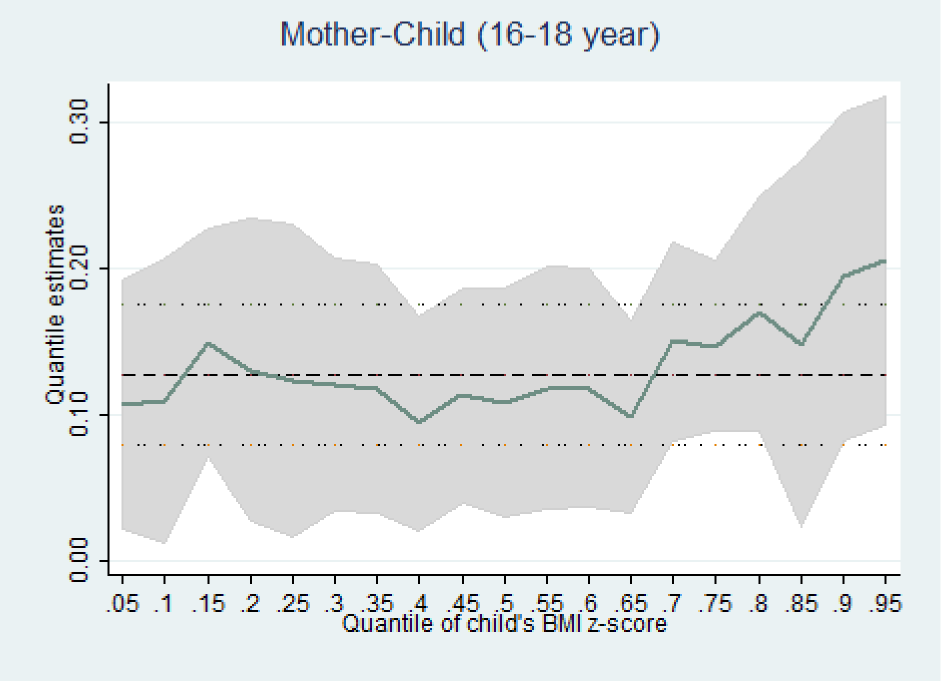
3.5.4 Quantile Estimation for Children Aged 16-18 years Old

As discussed in the literature review, the studies on intergenerational income or education often concern the potential life cycle bias. In the case of health, this bias might affect both biological and environmental channels in the transmission mechanism: biological, the metabolism of body varies with age, studies show there is a decrease in resting metabolic rate (the number of calories burned when the body is at rest) with the increase of age (Fukagawa et al. 1990); environmental, the time and the way parents and children share the dietary and lifestyle varies with time, children might have more decisions over their dietary as they go to school and become more independent of their parents. To address this life cycle bias, some studies follow a similar approach as the studies on intergenerational earnings transmission (Classen 2010). However, this same life stage match approach requires the data to cover the adulthood of the child and hence this approach is not implementable with the CHNS data we have. Nonetheless, as a response to the potential bias due to the unobserved heterogeneity associated with age, we restrict our sample to children aged between 16 and 18 years (approaching adults) and apply quantile estimation on this sample.

In this sample, there are 1,360 observations of children with father and mother. For this restricted sample, using equation (3.1), the quantile estimates are displayed in Figure 3.6. They suggest that for these children aged between 16 and 18 years old, the intergenerational persistence in BMI z-score are highest at the fattest end of BMI distribution (with an estimate of around 0.30 for father-child and 0.21 for mother-child). Compared with the pattern from the full sample (Figure 3.5), we can see that the degree of intergenerational transmission in BMI z-score varies with the stage of lifecycle, this motivates our analysis by age group in the next section.

Figure 3.6: Quantile estimates of intergenerational correlation of BMI z-score on Children aged 16-18 years old





3.5.5 Estimation by Age Group

We see the intergenerational correlation of BMI z-score varies with cohorts of different life stages. In this respect our results may be sensitive to the age range over which children are observed. To explore the change of this correlation with age, we estimate the intergenerational correlation of BMI z-score separately by age group at two year intervals. The results are provided in Table 3.9 and Figure 3.7[29]. They suggest that this correlation increases until the children are aged between eight and ten years old, when the estimates of the intergenerational BMI correlation reach around 0.29 for father-child and 0.23 for mother-child. This increase is followed by a decline over the rest of the childhood before the children enter adulthood for mother-child, whereas a fluctuation for father-child. Therefore, the common environment and genes that parents and child share together, play a greater role in the intergenerational transmission of BMI when children are aged between 8-12 years old, than other childhood stages prior to adulthood. One potential explanation is that before this pre-puberty stage, the effects of inherited genes from their parents exert the maximum influence. Whereas, after this stage, children spend less time with their parents, and exercise more control over their own dietary and exercise choices and hence the effects of a common family environment decline.

[29] It should be remembered that this is the pooled sample, so each individual children can appear more than once in the data as they age.

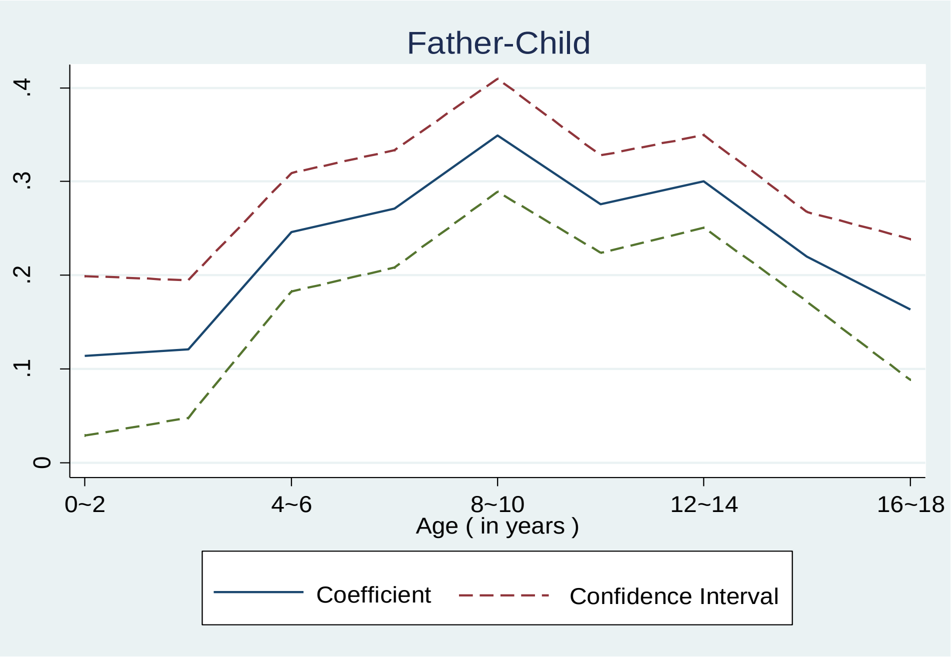
Based on the sample aged 16-18 years old, the estimates for the BMI correlation between parents and children within this age range might be close to the estimates for the correlation between parents and adult children's BMI (the correlation of long-term BMI). However, this is not our central interest in this study, we are mainly interested in the correlation between parents and children's BMI.

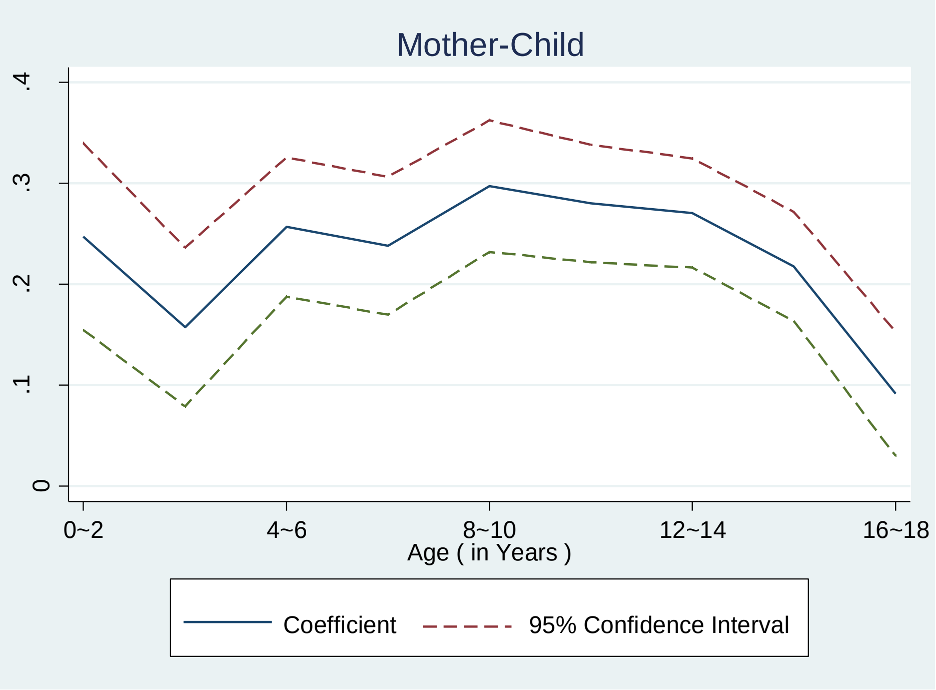
Table 3.9 : OLS estimates of the intergenerational correlation of BMI by age group

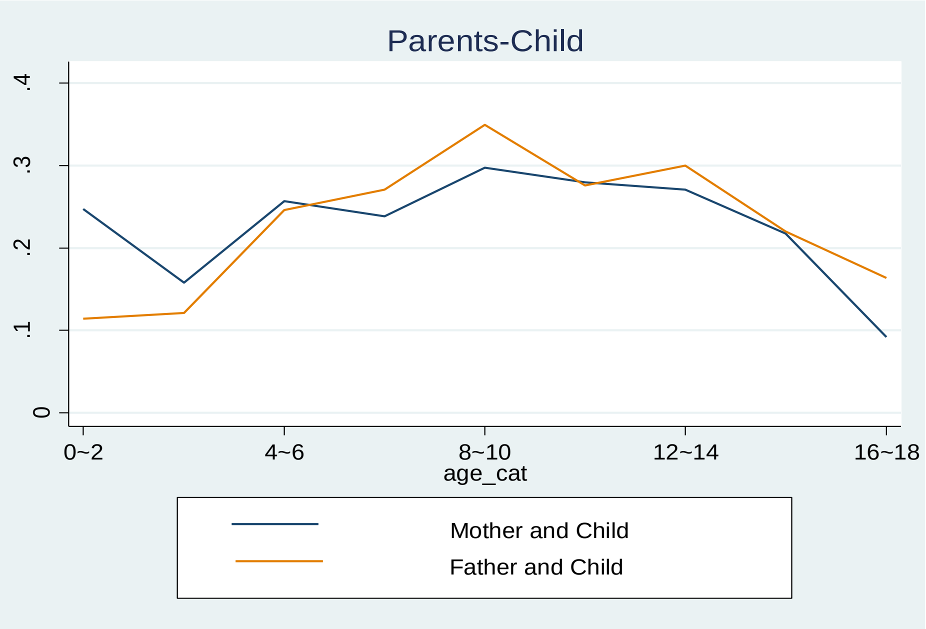
		Father and Child		Mother and Child	
Age Group (years)	Obs	Coefficient	Std. Error	Coefficient	Std. Error
0-2	1,536	0.121***	0.0452	0.243***	0.0514
2-4	1,438	0.122***	0.0386	0.153***	0.0433
4-6	1,554	0.249***	0.0356	0.257***	0.0388
6-8	1,620	0.274***	0.0333	0.238***	0.0369
8-10	1,775	0.349***	0.0314	0.302***	0.0344
10-12	1,966	0.276***	0.0291	0.280***	0.0292
12-14	1,876	0.300***	0.0262	0.268***	0.0295
14-16	1,572	0.221***	0.0248	0.217***	0.0294
16-18	606	0.163***	0.0449	0.0925**	0.0398

Note: the regression also includes: age dummies of child, gender of child and the interactions between them, father and mother's age. Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Figure 3. 7 : Estimates of the intergenerational relatsionship of BMI z-score by age group







3.6 Conclusions and Implications

Based on the CHNS longitudinal data from 1989 to 2009 and using BMI z-score as the measure of adiposity, we estimate the intergenerational correlation of BMI z-score in China. We use the OLS estimates as the main estimates of the intergenerational BMI correlation, since it indicates the basic underlying intergenerational correlation, and it is comparable with the intergenerational correlation of income or education, which are mostly based on cross-section data. The OLS estimates suggest one standard deviation increase in father's BMI z-score is associated with an increase of 0.20 in child's BMI z-score, and this figure is around 0.22 for the correlation between mother and child's BMI z-score. These estimates decreases to around 0.14 for father-child and 0.12 for mother-child when we control for the household fixed effects, similarly when we control for the individual fixed effects. The fixed effects estimates provide some evidence for the rather strong effects of short term environmental factors in the intergenerational transmission of body weight.

In terms of the heterogeneity of this correlation, this intergenerational correlation of BMI z-score does not vary substantially with family SES indicators; it tends to be higher among children of higher BMI levels, though this tendency becomes weaker when we use the sub sample of approaching adult children (children aged 16-18 years old).

Additionally, the change of this intergenerational relationship during child’s growth indicates that it tends to be higher over the stage between the childhood and the later adolescence.

As in the previous chapter, our findings indicate the importance of family and parents in determining the health status of children, they enhance the increasingly prevalent view that government policies to promote child health should be directed towards the parents which are the health care providers (Graham and Chris 2004). In particular the strong short term environmental effects suggested by the fixed effects estimates imply a potential role for the family-based interventions to play by promoting a healthy lifestyle for the parents (Marion 2006, Moria 2006).

Chapter 4: Health Selectivity of Migrants: The Case of Internal Migration in China

4.1 Introduction

Health, as an important component of human capital, is connected with migration in various ways. Studies on migration and health mostly concern the trajectory of migrant health associated with migration, which includes what happens before migration in terms of health (called health selectivity where migrants are selected on health traits) and what happens after people migrate (called acculturation and partly concerns the impact of migration on migrant health) (Jasso et al. 2004). The latter strand of literature largely compares the health of migrants with that of the population in the destination, which is comprised of one of the most significant propositions in the related studies: the "Epidemiological paradox" (or "health immigrant effects" or "healthy migrant phenomena"). It states that immigrants appear healthier when compared to native-born populations, in spite of the socioeconomic disadvantages and limited access to health care, with this health outcome often indicated by mortality rates, chronic conditions or disabilities, mental health and self-reported health (Chen, Wilkins, and Ng 1996, Marmot et al. 1984, Frisbie, Cho, and Hummer 2001, Hummer et al. 2007). There are three main explanations for this phenomenon: "healthy migrant theory" (migrants are healthier because they only represent a selectively healthy group rather than the whole population at the origin), cultural factors (migrants are healthier because of their better health habits, behaviours from their origins), and "salmon bias hypothesis"[30] (migrants are healthier because less healthy migrants return to their origins). Some studies also argue that the better health of migrants might be attributed to other unobservable factors, such as certain activities or cultural factors shared by the same community (Abraido-Lanza et al. 1999, Kennedy, McDonald, and Biddle 2006).

[30] "Salmon bias hypothesis" postulates that Hispanic people return to Mexico after temporary employment, retirement or severe illness, meaning that their deaths occur in Mexico and are not taken into account by mortality reports in the United States (Abraido-Lanza et al. 1999).

Among these explanations, “healthy migrant theory” posits that migrants tend to be positively selected on health traits and are in better health than those who do not migrate[31] (Findley 1988, Palloni and Morenoff 2001). There has been little research into the theoretical investigation of this relationship and empirical evidence on this “healthy migrant theory” remains scarce. This is largely due to the lack of data, which requires information on migrants and those who do not migrate in the places of origin prior to migration. Based on the limited data, existing studies usually compare migrants and those who do not migrate when they are observed just before migration and when they are observed just after migration. The relationship obtained from this short time-period “difference in difference” does not allow for the long term effects of health (proxied by the lagged health) on migration behaviour. Additionally, health effects might operate through education and/or occupation. These distant effects of health (lagged health effects and health effects via other factors, such as occupation) are important but have received little attention in previous studies. This current study will investigate both the indirect and direct effects.

This chapter is organised as follows. Firstly, we review relevant literature focusing on the context of international migration and internal migration in China. Then, we establish our theoretical model to ascertain the selectivity of health. Thirdly, we discuss the data and provide summary statistics for variables used in the empirical analysis. Fourthly, we describe the empirical model and present and discuss the empirical results. Finally, we summarise the main findings and present concluding remarks.

4.2 Literature Review

4.2.1 International Evidence

Current studies are mainly conducted in the context of US-Mexican migration. They are often flawed by making a comparison with an inappropriate reference group. For instance, using New Immigrant Survey 2003 cohort data, Akresh and Frank (2008) compare the self-assessed health of migrants in the US with that of residents in the origin communities, finding that the extent of positive health selection varies significantly across

[31] There is another version of “healthy migrant theory” stating that migrants tend to be healthier than the residents when they arrive at the destination.

immigrant groups and is related to compositional differences in migrants' socioeconomic profiles. However, this comparison is based on health outcomes after migration, so the health of migrants in the US and that of residents in the origin communities have been affected by different factors. Therefore, this "post-migration sample based" comparison is not a test of the "healthy migrant hypothesis", where the comparison is supposed to be made between migrants and those who do not migrate in the sending communities prior to migration. In the latter category, Rubalcava et al. (2008) use nationally representative longitudinal data from a Mexican Family Life Survey to examine whether recent migrants from Mexico to the United States are healthier than other Mexicans. By applying a logistic model, they investigate the effects of health and education on migration decision, where the migration occurs between surveys in 2002 and 2005, and the health and education indicators were measured in 2002. Their results suggest weak "positive selectivity" (the association of migrant health with their subsequent migration) among females and rural males. However, few health indicators were found to be statistically significant. Largely owing to the longitudinal structure of MxFLS data, which allows one to observe migrants and non-migrants in their origin communities at the initial time of migration, these results might provide some valid evidence on how health differs between migrants and non-migrants before migration, thus shedding some light on the verification of the "healthy migrant hypothesis".

Based on 1997 and 2000 waves in the Indonesian Family Life Survey (IFLS), Lu (2008) applies the same strategy to a sample comprised of individuals aged from 18-75 years old. Using a logistic model, she estimates the effects of health on migration, where health was measured by "problem with ADLs"[32] and other health variables in 1997, and migration occurs between 1997 and 2000. To estimate how the selection varies according to the reasons for migration, she also conducts multinomial logistic regressions to disaggregate migration by purpose, and applies household fixed effects to adjust for household unobserved heterogeneity. She finds that migrants tend to be selected on health traits, with the direction and size varying with the type of migration. Younger migrants are positively selected with respect to health, whereas older migrants are negatively selected. She argues that this might be because older people often migrate to seek health care, whereas younger migrants migrate mainly for labour market outcomes, so for them,

[32] The question is asked as "Having difficulties to carry out daily activities during the last three months" in the 2004 question, and as "Trouble working due to illness for the last 3 months" in the 2009 questionnaire.

especially for the labour migrants, they appear to be negatively selected for chronic health conditions and disabilities, as reflected in the inability to perform “Activities of Daily Living”.

In summary, in international literature, current studies on “health selectivity” remain scarce, largely due to the lack of data on the health of movers and stayers in the sending communities at the time of migration. The existing studies are mostly based on two-wave longitudinal surveys and mainly predict migration behaviours in between the two waves on health outcomes in the previous wave. These results mainly suggest a weak and partial “positive selectivity” among migrants. However, due to the short term nature of this longitudinal data, what these results provide is mainly a short-term correlation between health and migration.

4.2.2 Evidence for China

In the context of China, studies on the health selectivity of migrants are scarce. Some studies provide indirect evidence for positive health selectivity among migrants. Using a rural household survey conducted by a research institute in China’s Ministry of Agriculture and covering two provinces from 2003 to 2006, Wu (2010) applies a two-step selection bias correction model in the estimation of earnings. In this two-step setting, the 1st step, the employment choice model (actually an occupation model), is conducted to generate bias correction terms for the 2nd step, earnings, so as to purge the selection bias due to the unobserved characteristics associated with migration. Since this 1st step is a model for self-selection in migration, it generates predictions about how migrants compare with their home population. Therefore, it provides insight into the determinants of individuals’ migration decisions. Wu (2010)’s results suggest that youths, men, better educated and healthy individuals are more likely to participate in migration.

A recent study is based on a longer panel survey that covers four waves (1997-2009) of the China Health and Nutrition Survey (CHNS). Using a sample composed of individuals aged 16-35 years old, Tong and Piotrowski (2012) apply binary probit regressions of current migration status on the health variables in the previous wave, finding that migrants are positively selected on the basis of health, with the relationship between health and migration becoming less marked in later years. Though Tong and

Piotrowski (2012) use a relatively long span of longitudinal data, they basically pool the data in waves and only estimate the association of health with migration between one wave and the next for around three years. In this study, we attempt to provide evidence on the effects of earlier health on migration by exploiting the repeated observations in this longitudinal data. In addition, health selectivity might vary with the type of occupation migrants expect to get at the destination, since different occupational types require different health levels. Given that the occupations at the place of origin are often closely correlated to the prospective occupations migrants take in the destination, we will explore the variation of health effects by occupation. In addition, some exercises will be conducted to ascertain how these effects vary by education and age groups, and how we measure the health (using different health indices).

4.3 Theoretical Model

This section develops a migration model that describes health effects on migration. Firstly, we discuss Jasso et al. (2004)'s model, which is based on a benefit-cost framework and in which health effects mainly operate through skill and labour supply. However, the relationship between health and other factors in Jasso et al. (2004)'s model is too complicated for practical use (predicting health effects is not straightforward) and was not used by its authors in any formal way. Instead, we modify Borjas (1987)'s model on the self-selection of US immigrants, although in his case, the selection is based on unobserved individual characteristics. We follow Borjas (1987)'s structure and develop a probit model based on selectivity by health (illustrating the marginal effects of health).

In the model, migration is considered as an investment in a benefit-cost framework (Sjaastad 1962). Migration costs include monetary costs (such as the increase in expenditure for lodgings and transportation) C_0, and non-money costs, such as "psychic" costs $C,$ which continues over time (since people are usually reluctant to leave familiar surroundings). The expected benefits if people remain in their original communities are denoted by W_s, and the expected benefits if people migrate to the receiving communities denoted by W_r. Since this study is dominated by rural-urban migration, we define the rural area as the sending area and the urban area as the receiving area. For convenience of exposition, the costs and expected benefits are assumed constant through time.

Under the Chinese household registration system, the medical care systems are directly shaped by the rural-urban dualist structure. In the rural areas, a rural Cooperative Medical System was started in the end of the 1960s, it was dropped by counties and the coverage rate was only around 5% in 1985 (Liu and Cao, 1992). The rural population were mostly uninsured during the period between 1985 and 2003. To solve this lack of health insurance among the rural residents, the Chinese government launched the New Cooperative Medical Insurance in 2003, this program has expanded rapidly, the number of counties covered rose from 310 in 2004 to 2451 in 2007, and the number of participants reached 0.73 billion (Lei and Lin, 2009). In urban areas, the medical system was different, this system requires all the employees of urban enterprises to join the system, and this medical care scheme does not cover migrant workers[33]. Migrants do not have adequate access to health care, a survey in 2000 found that less than 3% were covered by health insurance schemes (Tang et al., 2008). This lack of access to urban health care system for rural migrants might affect the self-selection of migrants and also the health effects on migration: First, young and healthy people are more likely to migrate than elderly and unhealthy people; second, elderly and sick migrants tend to return to avoid the high medical costs in cities (Hu, Cook, & Salazar, 2008)

4.3.1 Model

We start with what is essentially Jasso et al. (2004)'s model. To simplify, we do not discuss how the length of time T migrants expect to settle at the receiving communities is determined; rather, T is assumed infinite and the same for everyone. People foresee and discount the future, with the discount rate assumed to be constant and denoted by i ($|i| < 1$). As a result, the present value of the expected migration benefits are denoted by

$$\sum_{t=0}^{\infty} W(1-i)^t = \frac{W}{i} \qquad (4.1)$$

where the discounted benefits are summarised over the migration period T (from period 0 to infinity). Applying this to both the expected benefits and costs, the migration decision

[33] See Biao (2003) for a detailed description of the urban medical care system.

will be made if the present value (discounted stream) of the net benefits of migration exceed the cost.

$$\frac{W_r}{i} - \frac{W_s}{i} - \frac{C}{i} - C_0 \geq 0 \qquad (4.2)$$

Multiplying the equation (4.2) by discount rate i, we obtain

$$W_r - W_s - iC_0 - C > 0 \qquad (4.3)$$

where iC_0 denotes the annualised amount of fixed costs. Following Jasso et al. (2004)'s model on migrant selectivity, the expected benefits are determined by the skills k and labour supply l of the migrants, and wage w in the receiving community r and the sending community s:

$$W_s = w_s\, k_s l_s \qquad\qquad W_r = w_r\, k_r l_r \qquad (4.4)$$

The wage w here is the "basic" wage, and it is augmented by skill k and labour supply l. Since these factors (w, k, l) might not be perfectly transferable across areas, according to Jasso et al. (2004)'s model, the relationship of these factors between the sending communities s and receiving communities r might be as follows :

$$k_r = \theta\, k_s \qquad\qquad w_s = \beta_0 + \beta w_r \qquad\qquad l_r = \gamma\, l_s \qquad (4.5)$$

where θ, β and γ represent the degree of transferability in factor k, w and l, respectively. θ, β and γ might be indexed to reflect different levels of k, w and l. For instance, θ might be larger for low skills than for high skills, since low skills might be more homogeneous across areas; on the other hand, there might also be reasons to presume that θ is larger for high skills since the recognition of high skills might be more general across the regions. Substituting equation (4.4) and (4.5) into equation (4.3), migration occurs if:

$$w_r\, k_s l_s \left(\theta\gamma - \frac{\beta_0}{w_r} - \beta\right) - iC_0 - C > 0 \qquad (4.6)$$

Based on Jasso et al. (2004)'s model, health enters the migration decision mainly through skills k and labour supply l. Let the base skill level be denoted by k_0, skill in the sending communities k_s is a function of h_s, and the same applies to labour supply l.

$$k_s = k_0 + \delta h_s \qquad\qquad l_s = l_0 + \varepsilon h_s \qquad (4.7)$$

Substituting equation (4.7) into (4.6), we obtain a migration model that incorporates the health factors.

$$w_r\,(k_0 + \delta h_s)(l_0 + \varepsilon h_s)\left(\theta\gamma - \frac{\beta_0}{w_r} - \beta\right) - iC_0 - C > 0 \qquad (4.8)$$

Thus far, this model essentially follows Jasso et al. (2004)'s migration model on initial health selectivity, which might be the only formal statement of a model on the health selection of migrants. However, this model is rather arbitrary and complicated. As equation (4.8) shows, there are many parameters and interactions; it does not really define selectivity and it is not clear how they derive the relationship of the degree of selectivity with other factors based on the model. Additionally, Jasso et al. (2004) do not actually use the model in their empirical work; their theoretical model is based on wages in sending areas w_r , whereas in their empirical work, they use real GDP per worker in the home country.

Jasso et al. (2004)'s empirical work mainly tests the relationship of health and skill selectivity with skill prices in the home country. Using the log of real GDP per worker as the country-specific skill price determinant and a self-reported health index (scaled from 1 (=excellent) to 5 (=poor)) as the measure of health, Jasso et al. (2004) estimate the determinants of ln (home country earnings) in a GLS model include the log of real GDP per worker and the average worker skill in the home country. Similarly, they estimate an ordered logit model for self-reported health. The results suggest that the log of real GDP per worker positively correlates with home country earnings and negatively correlates with the health index; the average worker skill negatively correlates with home country earnings and positively correlates with the health index. Jasso et al. (2004) argue that these results together suggest immigrants from countries with high skill prices might be positively selected according to their skill and health.

To make this model more formal and more empirically applicable, we turn to Borjas (1988)'s approach (Borjas selection model), which is a simple formulation of the Roy model. Roy (1951) associates the distribution of earnings with the distributions of various kinds of human capital and techniques in different occupations. More specifically, it states that there are three factors that affect the optimising choices of workers' selected occupations: the distribution of skills and abilities; the correlations among these skills in the population; and the technologies for applying these skills. Borjas' (1987) paper on

“Self-selection and the earnings of immigrants” is the first paper presenting a simple, parametric 2-sector Roy model (Autor 2003). In this model, Borjas (1987) assumes that the log of wages in the sending countries is normally distributed,

$$\ln w_0 = \mu_0 + \varepsilon_0 \qquad \text{where } \varepsilon_0 \sim N(0, \sigma_0^2)$$

And the same with the log of income in the United States (the receiving country),

$$\ln w_1 = \mu_1 + \varepsilon_1 \qquad \text{where } \varepsilon_1 \sim N(0, \sigma_1^2)$$

μ_0 and μ_1 are the observable socioeconomic variables, ε_0 and ε_1 are the unobserved characteristics. The model focuses on the impact of selection bias on ε_0 and ε_1. If π denotes a “time-equivalent” measure of migration costs, the probability of migration from the sending countries can be written as a probit model:

$$P = \Pr[v > -(\mu_1 - \mu_0 - \pi)] = 1 - \Phi\,(Z) \tag{4.9}$$

where $v = \varepsilon_1 - \varepsilon_0$, $Z = -(\mu_1 - \mu_0 - \pi)/\sigma_v$, and Φ is the standard normal distribution function.

Borjas' (1987) model is driven by the unobserved heterogeneity $v = \varepsilon_1 - \varepsilon_0$; however, our model is driven by the psychic costs C, which is assumed to be normally distributed to capture the heterogeneity across individuals. We adopt a more normal notation $\tilde{v}^j$ for this random element C, $\tilde{v}^j = v^j + \bar{v}$, where $\bar{v}$ denotes the average psychic costs of being away, which is absorbed into the fixed costs iC_0; v^j captures the part that varies across individuals. In other words, $\tilde{v}^j \sim N(\bar{v}, \sigma^2)$ and $v^j = \left(\tilde{v}^j - \bar{v}\right) \sim N(0, \sigma^2)$. We apply Borjas' (1987) selection model to model the selection of initial health. Putting equation (4.3) in the probit model, the probability of migration can be written as:

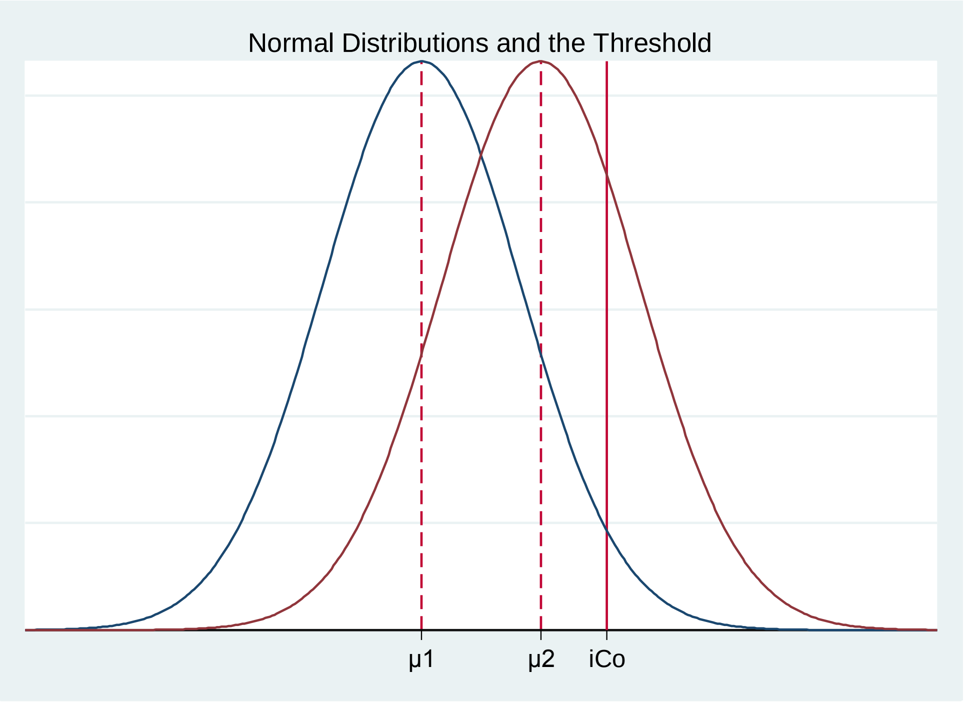
$$Prob\,\left(m_j\right) = Prob(v^j \leq W_r - W_s - iC_0) \tag{4.10}$$

where W_r and W_s are exogenous, $W_r - W_s - iC_0$ can be seen as the net benefits, they are the deterministic factors that comprise:

$$Prob\left(v^j \leq Z\right) = \Phi\,(Z) \tag{4.11}$$

The probability of random elements being less than the deterministic factors is the cumulative distribution function (CDF) of the standard normal random variable Z, with $\Phi(Z)$ being the univariate normal distribution.

Figure 4.1: The normal distribution and the threshold



One way of thinking about this model is in the following way: the wage differential $W_r - W_s$ is exogenous, the psychic costs C is normally distributed and the fixed costs iC_0 is the threshold. As Figure 4.1 suggests, there are two normal distributions of $W_r - W_s - C$, with means of $\mu 1$ and $\mu 2$respectively. Variability is captured by C and is normally distributed. Assuming less than half of the population migrate and thus the mean (μ) of the distribution is lower than the threshold, the threshold stands at the right tail of the distribution. The probability of migration P depends on how close the mean of the distribution is to the threshold. For instance, for the distribution with the mean $\mu 2$, the probability of migration is higher than a situation where the mean is $\mu 1$, since $\mu 2$ is closer to the threshold iC_0 than $\mu 1$. Similarly, $\frac{dP}{dZ}$ and $\frac{dP}{dC}$ would be higher when the mean is $\mu 2$ than a situation when the mean is $\mu 1$.

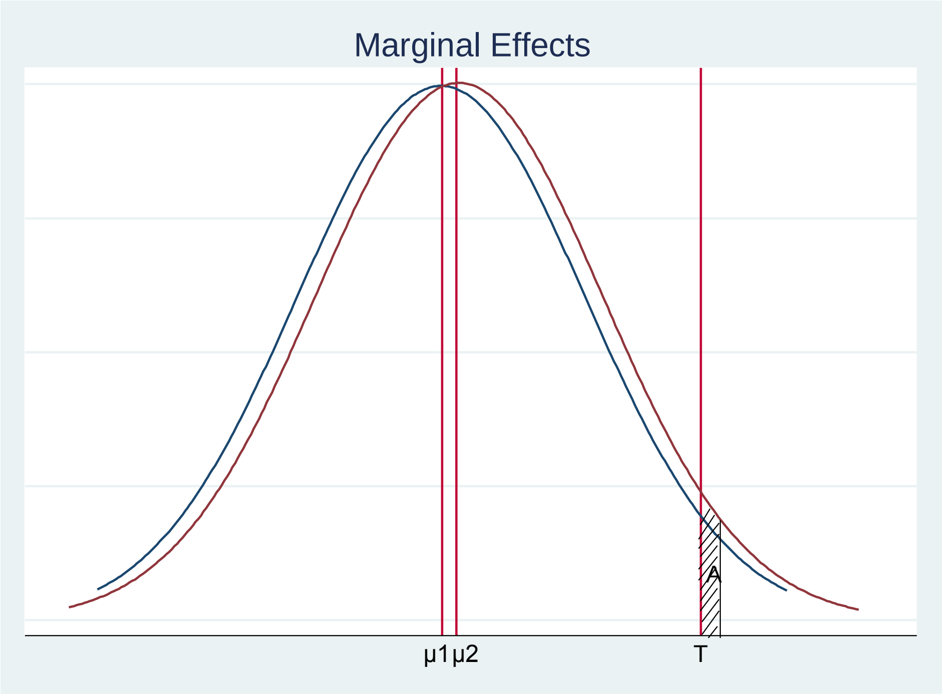
Selectivity for health concerns whether the probability of migration is positively or negatively related to health. In the context of the migration model established earlier (see equation (4.10)), the health effects relate to the change in the net benefits that are associated with the change in health. This marginal effect of health is obtained by differentiating the probability of migration with respect to health h:

$$\frac{\partial Prob(m=1)}{\partial h} = \frac{\partial P}{\partial Z}\frac{\partial Z}{\partial h} \tag{4.12}$$

where P denotes the probability of migration. Equation (4.12) suggests that the marginal effects of health depend on the values $\frac{\partial P}{\partial Z}$ and $\frac{\partial Z}{\partial h}$. In other words, the effects of health on migration probability depend on how much the move in the mean of Z affects the migration probability and how much health h affects Z.

As mentioned earlier, Z subsumes the deterministic factors W_r, W_s and iC_0, based on Jasso et al. (2004)'s model mentioned earlier (equation (4.8)), which, in turn, depend on the factors w, k, l, h, C and iC_0.

Figure 4.2: Marginal effects



As Figure 4.2 suggests, Z comprises the deterministic factors, such as health plus C, and so is normally distributed, the threshold T exceeds which migration might occur is fixed. As in Figure 4.1, the threshold always stands at the right tail of the distribution. Any increase in Z increases the probability of migration, since any increase in Z increases the number of people above the threshold. Put in Figure 4.2, when the mean of the distribution shifts slightly from $\mu 1$ to $\mu 2$, the marginal shift of the distribution creates an additional amount of migration by exceeding the threshold by accordingly more; the amount of this extra increased migration depends on the height of the normal curve $\frac{dP}{dZ}$ at T. All who would migrate with μ migrate with μ_2, and in addition, people falling in shaded area A now also migrate. For very small changes in μ, area A essentially corresponds to the height of the normal curve at T.

Turning now to $\frac{\partial Z}{\partial h}$, based on Jasso et al. (2004)'s framework mentioned earlier, substituting equations (4.4), (4.5) and (4.7) into equation (4.3), we obtain:

$$Z = w_r\,(k_0 + \delta h_s)(l_0 + \varepsilon h_s)\left(\theta\gamma - \frac{\beta_0}{w_r} - \beta\right) - iC_0 \tag{4.13}$$

Unfolding it, equation (4.13) can be written as:

$$Z = w_r\,[k_0 l_0 + (k_0\varepsilon + l_0\delta)h_s + \ \delta\varepsilon h_s^2]\left(\theta\gamma - \frac{\beta_0}{w_r} - \beta\right) - iC_0 \tag{4.14}$$

The quadratic term in equation (4.14) might imply a quadratic effect of health if the health variable h_s is continuous[34]. Differentiating equation (4.14) with respect to h, we obtain $\frac{\partial Z}{\partial h}$, which indicates how Z function moves from change h.

$$\frac{\partial Z}{\partial h} = w_r\left(\theta\gamma - \frac{\beta_0}{w_r} - \beta\right)[(k_0\varepsilon + l_0\delta) + \ 2\delta\varepsilon h_s] \tag{4.15}$$

Equation (4.15) suggests that $\frac{\partial Z}{\partial h}$ depends on the initial level of health and on k_0 and l_0, which vary by individual, and it also depends on w_r.

Equation (4.15) is overly complicated. To simplify it, we start by only considering wage w; assuming w depends on health h^j which is assumed fixed now, the migration decision is made if the net wage gains exceed the costs. Let superscript j index the individual, although for now the moving costs iC_0^j are assumed equal across individuals.

$$W_r^j - W_s^j - iC_0^j - C^j > 0 \tag{4.16}$$

Assuming the relationship between wages in the receiving area W_r^j and wages in the sending area W_s^j is written as follows:

[34] We tested this in the empirical model but it was not significant, so we dropped it.

$$W_r^j = \alpha\, W_s^j \tag{4.17}$$

where $\alpha > 1.$ Substituting equation (4.17) into (4.16), the equation can be expressed in terms of wages in the sending area W_s^j , which is what we have information on:

$$(\alpha - 1)W_s^j - iC_0^j \geq v^j \tag{4.18}$$

As mentioned earlier, suppose W_s^j is a function of health, $W_s^j = W\left(h^j\right) = W_0(1 + \lambda h^j)$, where W_0 denotes the wage of this individual at the base level of health, λ denoting the marginal (average) effect of health on the wage, $\lambda > 0$, so W_s^j increases as the level of health h^j increases.

Therefore, we have

$$Z^j = (\alpha - 1)W_s^j \ - iC_0^j = (\alpha - 1)W_0(1 + \lambda h^j) \ - iC_0^j \geq v^j \tag{4.19}$$

$(\alpha - 1)W_s^j \ - iC_0^j$ are deterministic factors and they are denoted by Z^j, with v^j denoting the random elements coming from the psychic costs C_0^j. Differentiating this with respect to h^j, we have:

$$(\frac{\partial Z}{\partial h})^j = (\alpha - 1)\frac{\partial W_s^j}{\partial h^j} = (\alpha - 1)W_0\lambda \tag{4.20}$$

4.3.2 Health Interacting with Wages

Equation (4.20) suggests that the health effects $\frac{dP^j}{dZ^j}\frac{\partial Z^j}{\partial h^j}$ vary with the level of W_s^j. Since the wage is not measured sufficiently well in the data (the best we can do is to measure it by dividing household income by the number of adults), in the empirical work, we use occupation and education as the proxies for wage W_s^j. Taking occupation as a proxy for W_0 suggests that a “better” occupation will show a greater degree of health selectivity-i.e. $\frac{\partial Z}{\partial h}$ will be higher for better paid occupations. Unfortunately, however, both λ and α may also vary by occupation, possibly in off-setting ways. For example, the sensitivity of wage with respect to health, λ, might be smaller for service work than construction work because it is more demanding in physical health, since usually work requiring a lower education or skill level[35] involves a higher standard (or level) of physical labour (Gagnon, Xenogiani, and Xing 2011); Similarly, $(\alpha - 1)$ may also vary by occupation, with the rural-urban wage ratio α potentially being larger for skilled more than unskilled work. In addition, $\frac{dP^j}{dZ^j}$ also depends on the type of occupation. As discussed earlier in Figure 4.2, since Z^j comprises W_s^j and C_0^j, Z^j increases as the level of occupation increases, and this increases $\frac{dP^j}{dZ^j}$ by moving the mean to the right towards the threshold, thus increasing the probability of migration. Therefore, the marginal effects of health on migration probability $\frac{dP^j}{dZ^j}\frac{\partial Z^j}{\partial h^j}$ varies with occupation (though we are unable to disentangle through which channel), which provides justification for the interaction of health h^j with occupation.

4.3.3 Direct Effects

The health effects discussed thus far are mainly indirect effects that operate through wage W_s^j. In addition, health might also affect the migration decision in a direct way. For instance, unhealthy people might be less capable of handling hardship on the journey, especially for long distance migration. In that case, health might directly interact with the moving costs C_0^j. Suppose (4.21)

C_0^j might be higher for unhealthy people, thus $\frac{\partial C_0^j}{\partial h^j} = \tau < 0$. Substituting equation (4.21) into equation (4.19), we obtain

$$Z^j = (\alpha - 1)W_s^j(h^j) - iC_0^j = (\alpha - 1)W_0(1 + \lambda h^j) - i(\tilde{C} + \tau h^j) \tag{4.22}$$

$$(\frac{\partial Z}{\partial h})^j = (\alpha - 1)W_0\lambda \ - i\tau \tag{4.23}$$

where $(\alpha - 1)W_0\lambda$ varies with occupation, and $i\tau$ captures the direct effects.

[35] In this study, work at the lower education or skill levels can refer to a farmer or non-skilled worker, which includes: senior professional/technical worker; junior professional/technical worker; administrator/executive/manager and office staff.

4.3.4 Indirect Effects

In addition to these effects of health via occupation or education (the interaction), there is another “indirect” channel-health might operate through “skill selectivity”. These skills are often measured by educational attainment. Specifically, if there is education selectivity, it might pick up some of the health effects because health, especially early health, might affect migration through education (attainment). Using data from a birth cohort that has been followed from birth into middle age, Case, Fertig, and Paxson (2005) present that children who experience poor health from the age of 7 to 16 years have significantly lower educational attainment, with childhood health conditions having a lasting impact on health and socioeconomic status in middle adulthood. Based on panel data from the US (the NLSY79 survey), Gan and Gong (2007) apply a structural four-stage model to clarify the mechanisms by which health and education interact with each other, finding that, on average, experiencing sickness before the age of 21 decreases education by 1.4 years. To account for the fact that k^j might interact with h^j, let k^j be a function of h^j, $k^j = k^j(h^j)$, hence the wage W_s^j is a function of skill k^j and health h^j:

$$W_s^j\left(k^j, h^j\right) = W_{00}k^j(h^j)(1 + \lambda h^j) \qquad (4.24)$$

where W_{00} is W_0 purged of the effects of k^j, i.e. W_0 without skill elements. The relationship between W_0 and W_{00} can be written as:

$$W_0 = W_{00}k^j \qquad (4.25)$$

Thus, W_{00} in a sense captures the mean of the wages across skill levels, and is the same for all levels of skills within the community. Substituting equation (4.25) into equation (4.19) and (4.20) accordingly, we have

$$(\frac{dZ}{dh})^j = (\alpha - 1)W_{00}[\frac{\partial k^j}{\partial h^j} + \lambda k^j + \lambda h^j \frac{\partial k^j}{\partial h^j}]$$

$$=(\alpha - 1)W_{00}\,\lambda k^j + (\alpha - 1)W_{00}(1 + \lambda h^j)\frac{\partial k^j}{\partial h^j} \qquad (4.26)$$

Equation (4.26) suggests $(\frac{dZ}{dh})^j$ depends on the levels of both k^j and h^j. This implies a quadratic term of h^j in Z^j after we incorporate $k^j(h^j)$.

In fact, it is likely that skill k^j is a function of lagged health (h_{-1}^j), rather than current health, and therefore can be treated as pre-determined and exogenous. In this case, the lagged health (h_{-1}^j) might have two effects, one via k^j and another correlating with current health (h^j); in the empirical work we will explore the relationship between lagged health and current health.

4.3.5 Empirical Implementation

Health selectivity is the derivative of migration probability with respect to health $\frac{dP^j}{dh^j}$, it is the selectivity of individual effect and is positive if $\frac{dP^j}{dh^j} > 0$. This is the definition we adopt in this study. There is an alternative definition of selectivity that is measured by how the average health of migrants differs from the average health of non-migrants, allowing for other characteristics and as one might see from tables of descriptive statistics. However, in that case, there is not necessarily a monotonic relationship between health and the probability of migration, as illustrated in Appendix 4.

In our model, for any given value of v^j, migration occurs if

$$(\alpha - 1)W_s^j(h^j) \geq iC_0^j \tag{4.27}$$

Equation (4.27) suggests that if the costs iC_0^j get higher, it requires a higher $(\alpha - 1)W_s^j(h^j)$ to overcome the threshold iC_0^j , implying a higher α (the rural-urban wage difference in large cities) or a higher wage $W_s^j\left(h^j\right)$, and then a higher level of health. In other words, having high health h^j will help to overcome the higher threshold iC_0^j. Thinking about internal migration in China in our model, iC_0 might be relatively high due to the household registration system, with the selectivity in h^jpotentially only being there for people with better health.

In the context of China, over half (around 65%) of the migrants are educated at the lower middle school level (Shi 2008), with a large proportion of the migrants working in manufacturing and construction[36] (Meng and Zhang 2001). In the meantime, an increasing fraction of younger generation migrants are employed in the manufacturing industry[37] and tertiary sector[38], while a declining proportion go into the construction sector[39]. Therefore, average health selectivity might change over time, even though average selectivity by occupation might remain the same.

4.4 Data and Empirical Model

4.4.1 Data

This study uses the China Health and Nutrition Longitudinal Survey, ranging from 1989 to 2011[40]. This survey contains detailed information on health outcomes, demographics and the anthropometric measures of all members of the sampled households, including height and weight. In addition, it includes information on economic and non-economic indicators, such as education, household income and labour market outcomes.

The sample used in this study comprises of individuals aged between 16 and 35 years old, by survey wave (i.e., aged 16-35 in 1997, 16-35 in 2000 and 16-35 in 2006; N=8,528 cases pooled from the 1997-2009 waves) because this study mainly concerns work migration. Table 4.1 presents the number of times that individuals aged 16-35 years in the CHNS raw data (1989-2009) are repeatedly observed (i.e. the number of individuals observed for different period lengths in the longitudinal data). Column 2 (observations 3,323 with frequency 6,646) shows that 3,323 individuals were observed for two waves, with column 7 suggesting that 11 individuals were observed for seven waves. Table 4.1 presents the number of times that individuals aged 16-35 years old in the CHNS raw data

[36] According to the National Bureau of Statistics, in 2009, nearly 39.1% of the migrants worked in manufacturing, about 17.3% in construction and more than 7.8% in wholesale and retail. Based on data from Beijing, Tianjing, Shanghai and Guangzhou in 2008, Cheng et al. (2013) present that around 76.9% of rural migrants work as competitive general workers, with “general” employees generally working as frontline commercial and service workers, manual workers and factory workers, undertaking repetitive tasks on assembly lines, low-skilled machine work and equipment operators.

[37] 44.4% compared to 31.5 percent of the previous generation.

[38] From http://www.mckinsey.com/insights/urbanization/preparing_for urban billion in_china

[39] 9.8% Compared to 27.8 percent of the previous generation.

[40] See appendix for a detailed introduction of the CHNS data.

(1989-2009) are repeatedly observed (i.e. the number of individuals which are observed for different lengths of period in the longitudinal data). In total there are 24,915 observations. Column 2 (observations 3,323 with frequency 6,646) shows that 3,323 individuals are observed for two waves,…, and column 7 suggests that 11 individuals are observed for seven waves. In our sample, the 8,528 observations are those who were observed at least once with all the variables used in the replication estimates (Table A4.5) realized. As we see, the attrition rate of the survey is relatively high[41], so this might underestimate the amount of migration. However, in this study our main interest is not the propensity to migrate, rather, we are interested in the effects of health on migration. For people who were observed only once, we observed their health the time we observed them, then they were missing, which we treated as migration. It is not that we do not treat them as migrants when they are missing. Therefore, the fact that almost 50% of the respondents are observed only once might not significantly affect our estimates of the health effects on migration. There might be a problem when the whole households were missing from the sample, since the migrant statuses were reported by household members, the missing of the entire households would not be treated as migrants. Therefore, the high attrition rate and the fact that a large number of respondents were observed only once might not significantly affect our estimates for the health effects on migration, though it might cause an underestimation of the migration propensity when the whole households migrate.

Table 4.1: The number of times individuals aged 16-35 years old were observed in CHNS (1989-2009)

Waves	1	2	3	4	5	6	7	Total
Obs	5,328	3,323	2,078	1,059	384	79	11	12,262
Frequency	5,328	6,646	6,234	4,236	1,920	474	77	24,915

Note: 5,328 individuals are observed for one wave, 3,323 individuals are observed for two waves, the sum of observations made on 12,262 individuals is 24,915.

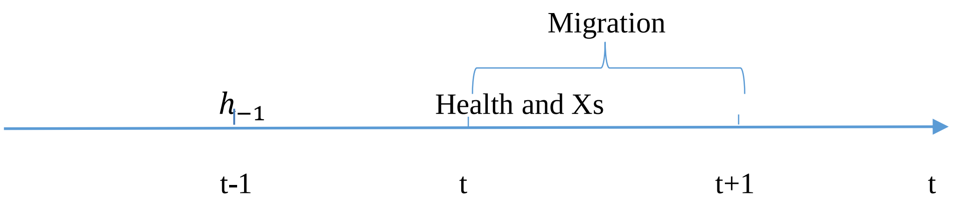
[41] See Popkin (2010) for a detailed description of the attrition rate in the CHNS data.

In terms of the age range of the sample, we adopt 16 as the bottom age range based on their argument that 16 years old is the starting point of the legal working age in China. Concerning the upper age limit, we use 35 because those older than 35 years might return due to deterioration in health[42]. It is worth noting that since we use a certain age level as the cut-off point, the sample size varies with the way this cut-off point is treated. Specifically, the number of individuals aged between 16 and 35 years old depends on whether the age is rounded into integers or not. The sample size presented in the baseline estimates (8,528) is the one when age is rounded into integers, as adopted from Tong and Piotrowski (2012)'s study. Here we thank Yuying Tong and Martin Piotrowski for their correspondence; we follow some codes in their stata program file. However, fewer observations would be left in the sample (8,062) if we used the two-decimal age points in the original data. This is because by taking the integers, some individuals aged between 15.5 and 16 years might be subsumed into the sample, thus those who actually did not meet the working age (16 years old) criteria would be included in the sample. Similarly, those aged between 35 and 35.5 years would be included in the sample because their age is rounded up as 35 years old. Therefore, more people would be included in the sample when age is rounded up into integers, rather than the two-decimal age points in the original data. Nonetheless, for comparability with Tong and Piotrowski (2012)'s study, we still round the age up into integers in this study.

The definition of the outcome variable "migrant status" is also based on Tong and Piotrowski (2012)'s paper and the programme file sent by one of the authors. Those who changed their hukou status (notice this requires this "hukou" variable not to be missing in the adjacent waves), with those who are absent for military, employment or other reasons in the next wave defined as migrants; those who remain at home, or are not living at home, but are in the same village/neighbourhood or the same county, or those who have gone to school in the next wave are defined as non-migrants; those who are dead in the next wave are missing. As Figure 4.3 suggests, the migration variable is measured as a change in residence across waves, and in the estimation, migration is a flow over period between t and $t+1$ and is explained by health and other characteristics at t.

[42] This is called the "salmon bias" hypothesis, which posits that people might return after temporary employment, retirement or severe illness (Abraido-Lanza et al. 1999).

Figure 4.3: The timing of the measure of migrant status



The health indicators adopted here include both objective health, such as acute and chronic conditions, and subjective health measures, such as a self-evaluation of overall health. Self-rated health is obtained by asking the respondents to rate their status relative to other people of a similar age and measured as a series of dummy variables that fall into the following four categories: “poor”, “fair”, “good” and “excellent”. Other indicators include dichotomous measures concerning whether the respondent had difficulty carrying out daily activities during the previous three months (henceforth referred as “ADLs”)[43], had a history of bone fractures or had ever smoked. “ADLs”, as an indicator of physical functioning, is a measure of long-term health condition and is particularly associated with limitations, such as severe chronic disease and disability (Johnson and Wolinsky 1993). It has been often used to study the health of prime-age adults in previous studies (Frankenberg and Jones 2004).

To facilitate the comparison with Tong and Piotrowski (2012)’s estimates, we first include both self-rated health and objective health. As self-rated health and objective health include almost identical information, in the bulk of the following analysis, we only use self-rated health because it is a more comprehensive health indicator. In addition, as a subjective indicator, self-rated health might have stronger predictive power of individual behaviour and thus might be a more significant determinant of the propensity to migrate.

In terms of other variables, for the “occupation” variable in the raw data, there are sixteen occupation types. Table 4.2 presents our classification of these occupations, classified into six main categories that are mutually exclusive. Though the distinction of “non-farm worker” from other types of worker is unclear, it is more like the category

[43] It is referred as “having trouble working due to illness last 3 months" in 2009 longitudinal data.

"professional and administrative worker". We adopt this classification from Tong and Piotrowski (2012)'s study.

Table 4.2: The categories of occupations

The categories of occupations in this study	The categories of occupations in the raw data	Sample size
The unemployed or student	The unemployed or student	1,975
Farmer	Farmer, fisherman, hunter	3,313
Non-farm worker	senior professional/technical worker, junior professional/ technical worker, administrator/executive/manager and office staff	1,024
Service worker	army officer, police officer, ordinary soldier, policeman, driver and service worker	590
Skilled worker	skilled worker	847
Non-skilled worker	non-skilled worker	1,041

The variables associated with family members, such as the residence of spouse and parents, are mainly obtained based on the information from the "roster" file, one of the 40 data files from the 1989-2011 longitudinal data. The variable "spouse's presence" is constructed by combining the variables "does spouse live at home" and "spouse's line number" because there is a relatively high proportion of missing values (90.54%) for the variable "does spouse live at home"[44]. The constructed variable "spouse's presence" is a dichotomous variable that is equal to one when the respondent has a spouse present at home (which is for the married respondents); while it is equal to zero when the respondent does not have a spouse or has a spouse but the spouse is not living at home. In other words, the respondents with "spouse=0" includes both non-married people (never-married, widowed, divorced, separated) and people who are married but without the spouse's presence at home. Therefore, this is not a variable that is only observed for the married people[45]. Rather, this "spouse's presence" variable is defined based on the whole sample which includes both married and unmarried people. In terms of the variable

[44] The proportion of missing values for the "spouse's line number" is 42.61%.

[45] In this case (if the variable "spouse's presence" is only for married people), the proportion of "spouse's presence" will be around 90%.

“parents’ presence and age”, parents’ “presence” is a dichotomised variable, which is defined based on the question “Does your father/mother live in the home?”, their ages are merged from the “physical examination” file through the parents’ identification number (“father/mother’s line number”). Based on the definitions above, the descriptive statistics for these variables are presented in Table 4.3.

As mentioned earlier, Tong and Piotrowski (2012)’s study might be one of the closest studies on the “healthy migrant hypothesis”. We wish to extend and refine their analysis for the following reasons. Firstly, in Tong and Piotrowski (2012)’s study, there are still a variety of results they have not explored. For instance, they do not include interactions in their estimation nor explore the effects of lagged health. Secondly, their study has little relationship to economic theory. In our study, we derive a subtle model from Jasso et al. (2004) and Borjas (1988)’s migration model, in which we show that the health effects might vary with wage. Nonetheless, a sensible starting point seems to be to try to replicate their estimates. We downloaded the CHNS 2011 longitudinal survey and use the same waves (1997-2009) as their study; our sample size is larger than theirs and the descriptive statistics appear different to theirs (discussions on these differences are presented in Appendix 4). To test whether this difference comes from differences in the data versions, we use the 2009 longitudinal survey, even though the sample size and descriptive statistics remain the same as those from the 2011 longitudinal survey. We will conduct various tests to investigate the differences between Tong and Piotrowski (2012)’s sample and our sample. For instance, when checking the parental residence variables, we use 1990 Chinese census data, and similar periods (the waves 1991 and 1993) in the 2011 longitudinal survey, constructing the parental variables and finding that the descriptive statistics based on these data are closer to our sample than to Tong and Piotrowski (2012)’s sample. To check the spouse’s presence variable, we contacted Ahn et al. (2013) and Chen (2012) who created the same variable using this data (we thank them for their information and follow their approach when constructing this variable). Based on this replication, we attempt to re-estimate the “healthy migrant hypothesis” in China and conduct several extension analyses.

Table 4.3: The descriptive statistics for independent variables

Wave	Pooled	1997	2000	2004	2006
	Mean	Mean	Mean	Mean	Mean
Health					
Self-rated health					
Poor	0.02	0.01	0.02	0.02	0.02
Fair	0.17	0.15	0.17	0.22	0.17
Good	0.60	0.66	0.56	0.54	0.58
Excellent	0.21	0.18	0.25	0.22	0.23
Difficulty with ADLs	0.03	0.02	0.04	0.03	0.04
Bone fracture	0.02	0.01	0.03	0.03	0.03
Ever smoked	0.26	0.25	0.25	0.27	0.27
Demographic					
Age	26.97	25.86	26.82	28.11	28.45
Gender (male)	0.49	0.50	0.47	0.48	0.49
Ever married	0.62	0.54	0.62	0.69	0.70
Highest degree earned					
Primary or lower	0.23	0.26	0.24	0.20	0.16
Lower middle	0.48	0.49	0.50	0.47	0.48
Upper middle	0.16	0.17	0.14	0.17	0.15
Technical/vocational	0.08	0.06	0.07	0.11	0.12
College and beyond	0.05	0.02	0.05	0.07	0.09
Occupation					
None/student	0.22	0.18	0.20	0.30	0.28
Farmer	0.38	0.45	0.45	0.26	0.26
Non-farm	0.12	0.10	0.12	0.13	0.13
Skilled	0.07	0.06	0.07	0.07	0.07
Non-skilled	0.09	0.10	0.07	0.10	0.10
Service	0.12	0.10	0.10	0.14	0.16
Ever migrated since 1993	0.09	0.07	0.05	0.10	0.18
Household					
Rural	0.71	0.73	0.72	0.67	0.70
Size	4.35	4.40	4.25	4.27	4.45

Real income in 2006[46] currency[47]	3427.64	2485	2978.5	4443.12	5086.25
Log income	11.98	11.98	11.98	11.99	11.99
Parents					
Both parents <56	0.31	0.36	0.31	0.26	0.24
One parent >55	0.11	0.11	0.10	0.11	0.13
Both parents > 55	0.10	0.09	0.08	0.12	0.11
No parents	0.49	0.45	0.51	0.52	0.52
Spouse	0.61	0.54	0.62	0.68	0.69
Child	0.56	0.55	0.54	0.58	0.59
Region					
Coastal	0.21	0.22	0.20	0.20	0.21
Northeast	0.19	0.14	0.27	0.20	0.19
Inland	0.34	0.38	0.27	0.33	0.33
Southern mountain	0.26	0.27	0.26	0.26	0.26
Wave					
1997	0.40	-	-	-	-
2000	0.23	-	-	-	-
2004	0.20	-	-	-	-
2006	0.17	-	-	-	-
Total number of cases	8,528	3,423	1,956	1,738	1,411

[46] Note: The "income in 2006 currency" is calculated using the price index from the World Bank (2005=100) and converted from the income in 2011 currency . In addition, we follow the stata dofile sent by one of the authors, to avoid losing the negative values of the income, we shift the income distribution to the right by a distance of absolute value of minimum income, through adding this value to the income before taking the logarithm. Also, to avoid losing observations with minimum income, we also add one unity to the income before taking the logarithm. In sum, before taking the logarithm, we add the absolute value of minimum income (scaling to zero) and 1 (one unity) to the income, in order to keep all the observations (rather than losing the observations with negative values) in the sample.

As mentioned in the theoretical model, health might enter into the model as an additive factor, in a similar way as skill. At the same time, it might operate through being a determinant of skill, and therefore multiply with skills or other human capital factors (measured by occupation or education here). Therefore, we estimate the probit model:

$$Prob(migration_{i,t}) = \Phi(health_{i,t}\alpha + occupation_{i,t}\beta + Interactions_{i,t}\gamma + X_{i,t}^{'}\vartheta + \varepsilon_{i,t}) \quad (4.28)$$

where the variable $migration_{i,t}$ equals one if migration occurs over the period from t to $t+1$, zero otherwise. The variables included in this porbit model are as follows: health, occupation, education, the interaction of health with occupation, the interaction of health with education and other characteristics measured at t. Therefore, the probability of migration between t and $t+1$ is a function of health and other characteristics at t.

4.5 Empirical Results

4.5.1 Baseline Estimates

Table 4.4 presents the estimates from equation (4.28) and are used as baseline estimates in this study. The results across the pooled sample suggest those self-evaluating as having “excellent”, “good” or “fair” health to be more likely to migrate than those self-evaluating as having “poor” health, indicating that most of the distinction comes from “poor” and the rest three categories. Concerning the waves, those self-evaluating as having better health are more likely to migrate in earlier waves (“good” or “excellent” in 1997 and “excellent” in 2000). Though these health effects appear insignificant in other waves, their signs are mostly positive for all except the last wave (2006) where “excellent” health is negative. These results support the hypothesis that there might be a positive health selection on migrants, which is consistent with related studies, claiming that there is a weak and partial “positive selectivity” among migrants (Rubalcava et al. 2008). Moreover, these results also accord with studies showing that the health effects vary with the type of migration and the age of migrants (Lu 2008), finding younger migrants to be positively selected on health, whereas older migrants are negatively selected. These effects might offset each other, therefore, together positive health effects might not appear strong.

In terms of other health measures, the estimate for the “having difficulties to carry out daily activities during the last three months” variable is not significant in the pooled data, but we might still be able to draw some inference from the positive sign that those who have “ADLs” are more likely to migrate. The results across the waves suggest that the effect of having “ADLs” is positive in the 1997 and 2000 waves, negative in the 2004 wave and significantly positive in the 2006 wave. Using the 1997 and 2000 waves of the Indonesia Family life Survey (IFLS), Lu (2008) finds that ADLs are negatively associated with the possibility of migration for people aged 18-45 years old. Thus, based on our sample, aged 16-35 years old, we might expect to see a negative correlation between “having ADLs” and the probability of migration. As another indicator of chronic health, the effects of bone fracture appear insignificant, though they are mostly positive across the waves. Table 4.4 also suggests that the effects of “ever smoking” are not significant in the pooled sample and across the waves, except for the 2000 wave, in which those who are habitual smokers seem more likely to migrate. The signs of the effects are mostly positive until the latest 2006 wave, in which the sign is negative. However, smoking might not be an adequate indicator of adverse health, since smoking is more like health behaviour than a health outcome. In addition, there might be potential collinearity between these health measures. As mentioned earlier, the following equations will not include these objective health measures.

Table 4.4: Probit regression of migration status on health

	(1) Pooled b/se	(2) 1997 b/se	(3) 2000 b/se	(4) 2004 b/se	(5) 2006 b/se
Self-rated health: Poor (Ref.)					
Fair health	0.291*	0.603	0.352	0.361	0.175
	(0.16)	(0.38)	(0.32)	(0.32)	(0.29)
Good health	0.361**	0.663*	0.392	0.395	0.257
	(0.16)	(0.38)	(0.31)	(0.32)	(0.28)
Excellent health	0.400**	0.714*	0.566*	0.508	-0.015
	(0.16)	(0.39)	(0.32)	(0.33)	(0.29)
Trouble working due to illness	0.190	0.225	0.050	-0.070	0.518**
in the last three months	(0.12)	(0.27)	(0.20)	(0.26)	(0.22)
History of Bone Fracture	0.094	0.106	0.306	-0.291	0.007
	(0.13)	(0.26)	(0.21)	(0.28)	(0.26)
Ever Smoked	0.057	0.039	0.205**	0.071	-0.066
	(0.05)	(0.08)	(0.10)	(0.11)	(0.12)
Demographic					
Age (in Yrs)	-0.044***	-0.032***	-0.047***	-0.048***	-0.051***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Gender (Male=1)	0.111**	0.078	0.101	0.173	0.266**
	(0.04)	(0.07)	(0.09)	(0.11)	(0.12)
Ever married	0.068	0.104	0.047	0.034	0.050
	(0.14)	(0.24)	(0.23)	(0.37)	(0.40)
Highest degree: Primary or lower (Ref.)					
Lower middle school	-0.010	-0.034	-0.033	-0.091	0.237*
	(0.05)	(0.07)	(0.10)	(0.11)	(0.14)
Upper middle school	-0.066	-0.238**	0.002	-0.108	0.309*
	(0.07)	(0.11)	(0.14)	(0.15)	(0.17)
Technical/Vocational school	-0.138	0.061	-0.199	-0.519***	0.083
	(0.09)	(0.15)	(0.19)	(0.20)	(0.19)
College and beyond	0.033	-0.050	-0.055	-0.034	0.506**
	(0.12)	(0.25)	(0.27)	(0.22)	(0.23)
Occupation: None/student (Ref.)					
Farmer	0.037	-0.046	0.011	0.199*	0.069
	(0.05)	(0.09)	(0.11)	(0.12)	(0.14)
Non-farm	-0.172**	-0.035	-0.671***	-0.007	-0.345*
	(0.09)	(0.14)	(0.19)	(0.17)	(0.18)
Skilled	0.034	-0.107	0.077	-0.236	0.266
	(0.08)	(0.15)	(0.17)	(0.19)	(0.19)
Non-skilled	-0.045	-0.144	-0.161	-0.120	0.084
	(0.08)	(0.13)	(0.18)	(0.17)	(0.17)
Service	-0.000	-0.050	-0.108	0.144	-0.070
	(0.07)	(0.12)	(0.16)	(0.13)	(0.14)
Previous Migration Experience	0.392***	0.783***	0.120	0.388***	0.211**

	(0.05)	(0.10)	(0.14)	(0.10)	(0.09)
Rural/Urban(Rural=1)	0.383***	0.424***	0.431***	0.374***	0.294**
	(0.05)	(0.08)	(0.10)	(0.11)	(0.12)
The number of people in household	0.077***	0.032	0.080**	0.076**	0.130***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)
Household Income per capita (in 2006 currency, logged)	-1.110	1.694	3.243	-2.730	-0.860
	(1.06)	(2.70)	(2.22)	(2.17)	(1.73)
Parents: Both parents <56 (Ref.)					
One parent's age > 55	-0.000	-0.051	-0.136	0.112	0.162
	(0.07)	(0.11)	(0.14)	(0.16)	(0.17)
Both parents' age > 55	-0.006	-0.028	0.025	-0.010	0.013
	(0.08)	(0.12)	(0.16)	(0.17)	(0.18)
No parents	-0.062	-0.264**	0.013	0.007	0.083
	(0.07)	(0.12)	(0.14)	(0.15)	(0.17)
Spouse	-0.187	-0.416*	-0.056	-0.135	0.108
	(0.14)	(0.23)	(0.21)	(0.36)	(0.39)
Child	-0.149***	0.131	-0.321***	-0.252**	-0.340**
	(0.06)	(0.09)	(0.11)	(0.13)	(0.14)
Region: Coastal (Ref.)					
Northeast	-0.292***	-0.241*	-0.485***	-0.408***	0.285
	(0.07)	(0.13)	(0.13)	(0.16)	(0.18)
Inland	0.198***	0.101	0.319***	0.177	0.432***
	(0.06)	(0.09)	(0.11)	(0.12)	(0.15)
Southern mountain	0.213***	0.178*	0.182	0.269*	0.440***
	(0.06)	(0.10)	(0.12)	(0.14)	(0.16)
Wave: 1997 (Ref.)					
2000	0.256***				
	(0.05)				
2004	0.248***				
	(0.06)				
2006	0.147**				
	(0.06)				
Constant	12.323	-21.550	-39.553	32.049	8.987
	(12.70)	(32.33)	(26.59)	(26.02)	(20.79)
Observations	8528	3423	1956	1738	1411

Standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Concerning education, with primary school and below as the base category, education effects are not significant in the pooled sample, but are significant in several waves. Specifically, the estimates suggest that those with middle-higher levels of education are less likely to migrate in earlier waves (1997 and 2004) but more likely to migrate in the recent wave (2006). Previous studies suggest that migrants are mainly drawn from the intermediate level (especially those who have completed junior secondary school and above) of education distribution in the sending communities (Yang and Guo 1999, Li and Zahniser 2002, Wu and Zhou 1996). With respect to occupation, we see those with initial an occupation of "non-farm" worker to be less likely to migrate than those who are students or unemployed in the pooled data and the 2000 and 2006 waves. As mentioned earlier, "non-farm" workers are mainly "professional or administrative workers". This negative effect of the "non-farm" occupation on migration appears surprising because based on the fact that the base wages for these high skilled occupations are higher than less skilled occupations, people who are more highly paid should have more incentive to migrate. However, studies suggest that rural migrants are treated differently to their urban counterparts in terms of occupational attainment and wages (Knight and Song 1999, Meng 2000). Using data from two comparative surveys in Shanghai, Meng and Zhang (2001) present that 6% of rural migrants who would have been suitable for white-collar jobs were forced to take blue-collar jobs in these urban labour markets where rural migrants are discriminated against, with skilled migrants potentially having to accept work in the unskilled occupations, thus having less incentive to move to the city.

Regarding other variables, Table 4.4 suggests that among people aged between 16 and 35 years, age has a negative effect on the probability of migration, with the respondents less likely to migrate when they grow older. Males are more likely to be migrants in the pooled sample and the later 2004 and 2006 waves, indicating that migration might become more male dominated over time. The prior migration experience is significantly positively related to migration in all but the 2000 wave. Those from rural households are more likely to migrate in all the waves. Household size is significantly positively related to migration in all the waves apart from 1997, which is consistent with related studies (Rozelle, Taylor, and DeBrauw 1999, Taylor, Rozelle, and DeBrauw 2003) , as larger households have more labour to allocate across activities. Household income per capita seems not significantly related to migration, though previous studies suggest an inverted-U-shaped relationship between household endowments and the likelihood of migration (Du, Park, and Wang 2005). For the relational variables, "residence with no parents" reduced the potential to migrate in 1997 and "having a child aged less than 12 years old at home" was negatively related to migration in all waves but 1997. In terms of regional variation, compared to coastal regions (the reference group, includes the provinces Shandong, Jiangsu and Heilongjiang), respondents from the less developed Northeast region are less likely to migrate, except in 2006, when the effects were non-significant. However, those from inland regions and southern mountain regions, which are also less developed areas, were more likely to migrate, but it was not seen in all the waves.

As we see in Table 4.4, there are not many significant effects in the results, which might reflect the fact that there is insufficient information in the sample. Our sample consists of only around 8000 people from across China, whereas migration is a complex, patterned, multi-dimensional and dynamic process, and there is a large amount of heterogeneity and noise in this process (Castles 2012). Thus, it is not very surprising that most estimates are not very well-defined or significant from this small amount of information. Additionally, there is potential collinearity between health measures because these health measures might contain similar information, making it more difficult to identify health effects. Similarly, the potential multi-collinearity between education and occupation measures might confound the identification of education and occupation effects. Nonetheless, based on the pattern of these estimates, we can still gain some insights into this "healthy migrant hypothesis", so we will carry forward and conduct some extension analysis.

Our estimates of the health effects are weaker than Tong and Piotrowski (2012)'s estimates of the health effects[48]. Collapsing "poor" and "fair" into one category, Tong and Piotrowski (2012)'s estimates suggest that those self-evaluating as having "excellent" health are significantly more likely to migrate than those self-evaluating as having "poor or fair" health, at a 1% significance level. Using the four-category version of self-rated health, our estimates suggest that most of the distinction comes from those self-evaluating as having "poor" health (accounting for only around 2% of the sample) and that those self-evaluating as having "fair", "good" or "excellent" health are significantly more likely to migrate than those self-evaluating as having "poor" health, at a lower significance level (10% and 5%, respectively). Compared with Tong and Piotrowski (2012)'s estimates, our estimates provide weaker evidence to claim "health selectivity" among the migrants, with our estimates being consistent with relevant studies that suggest weak and partial "positive health selectivity" among migrants. Additionally, we conducted various tests to

[48] The details on the replication and the comparison between our estimates and Tong and Piotrowski (2012)'s estimates are presented in Appendix 4.

try to replicate the data, with the results lending confidence to the validity of our estimated results.

4.5.2 Health Interacts with Occupation

As discussed above,$(\frac{\partial Z}{\partial h})^j = (\alpha - 1)\frac{\partial W_s^j}{\partial h^j}$ - $i\tau$, which might vary by occupation, since $\frac{\partial W_s^j}{\partial h^j}$ (the sensitivity of wage W_s^j with health) varies with occupation, $\frac{\partial W_s^j}{\partial h^j}$ might be larger for lower skilled workers than for higher skilled workers, and α may vary by occupation. The direct costs $i\tau$ might not vary by occupation because the effects of health on the costs of making the trip seem independent of occupation.

To test whether $(\frac{\partial Z}{\partial h})^j$ varies with occupation, we create interaction terms between occupation and self-rated health and include them in the estimation. The key results are presented in Table 4.5 and suggest that these interaction terms are mostly insignificant and the coefficients of other variables do not change significantly. The sample size changes from 8528 in Table 4.4 to 8779 in Table 4.5, since the estimates in Table 4.5 use the specification which does include objective health measures as in Table 4.4. To facilitate the comparison, we report the coefficients for each health/occupation interaction term in Table 4.6. We test the joint significance of interactions of “fair” health with occupations (the p-value of the χ^2 test is 0.393), suggesting that the interactions of “fair” health with occupations are jointly insignificant; similarly, for the joint significance of interactions of “good” health with occupations (the p-value of the χ^2 test is 0.538); the interactions of “excellent” health with occupations (the p-value of the χ^2 test is 0.358); and also tested the joint significance of all these interaction terms (the p-value of the χ^2 test is 0.524). They suggest that these interactions are not jointly significant. Table 4.6 suggests that “excellent” health has a larger positive effect on migration probability for people with an initial occupation as a lower skilled worker (“unemployed or student”, “farmer” and “non-skilled”) than for those who worked as a higher skilled worker (“non-farm”, “skilled” and “service”) at the places of origin. Therefore, these results are consistent with the model above, with the positive health effects tending to be larger for lower skilled workers than for higher skilled workers. Additionally, we see the coefficients of these interactions increase as health gets better in each occupation, except for "non-farm" and "service". Although these coefficients are mostly insignificant, this pattern indicates that the health effects might become larger with improvement in health. In addition, using the binary version of health variable, we estimate a more parsimonious model and the sample size is hence expanded, the estimates are presented in Appendix 4. We re-estimate the baseline equation (Table 4.4) and the equation on the health effects estimates by occupation (Table 4.5), the results are presented in Table A 4.1 and Table A 4.2, respectively.

Table 4.5: The estimates of health effects by occupation

Dependent variable: Probability of migration	Pooled coeff	s.e.
Self-rated health: Poor (Ref.)		
Fair	0.380	(0.29)
Good	0.468	(0.28)
Excellent	0.543*	(0.29)
Occupation: Unemployed/student (Ref.)		
Farmer	0.415	(0.34)
Non-farm	-0.210	(0.14)
Skilled	-0.177	(0.15)
Non-skilled	-0.054	(0.55)
Service	0.594	(0.56)
Fair* Farmer	-0.315	(0.35)
Fair* Non-farm	-0.121	(0.25)
Fair* Skilled	0.408*	(0.24)
Fair* Non-skilled	-0.059	(0.58)
Fair* Service	-0.508	(0.58)
Good* Farmer	-0.379	(0.34)
Good * Non-farm	0.088	(0.17)
Good * Skilled	0.244	(0.18)
Good * Non-skilled	0.041	(0.56)
Good * Service	-0.605	(0.57)
Excellent * Farmer	-0.436	(0.35)
Excellent * Non-skilled	0.230	(0.57)
Excellent * Service	-0.675	(0.58)
Observations	8779	

Note: The equation also includes other controls in the baseline equation (except for the objective health measures); there are only three interactions of "excellent" health with occupations (rather than five) because of collinearity; standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table 4.6: Partial interaction of health with occupation

		Poor		Fair		Good		Excellent	
		Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
(1)	Unemployed /students	0	0	0.38	(0.29)	0.468	(0.29)	0.54*	(0.29)
(2)	Farmer	0.42	(0.34)	0.48*	(0.29)	0.50*	(0.28)	0.52*	(0.29)
(3)	Non-farm	-0.21	(0.14)	0.05	(0.34)	0.35	(0.30)	0.33	(0.31)
(4)	Skilled	-0.18	(0.15)	0.61*	(0.33)	0.54*	(0.29)	0.37	(0.31)
(5)	Non-skilled	-0.05	(0.55)	0.27	(0.32)	0.46	(0.29)	0.72*	(0.31)
(6)	Service	0.59	(0.56)	0.47	(0.31)	0.46	(0.29)	0.46	(0.31)

The estimates in Table 4.6 reflect how $\frac{\partial Z}{\partial h}$ varies by occupation. Based on the equation $(\frac{\partial Z}{\partial h})^j = (\alpha - 1)\frac{\partial W_s^j}{\partial h^j} - i\tau$, these differentials might come from the differential in α (the ratio of average urban wage to average rural wage) or the differential in $\frac{\partial W_s^j}{\partial h^j}$ across occupations. Above, we have proceeded as if α is constant across occupations, but to test whether this is a reasonable assumption to make, we calculate the ratio of average urban wage to average rural wage by occupation in our pooled sample (N=8790) (the results are presented in Table 4.7). The wage here is approximated by the household income divided by the number of adults, an admittedly inadequate measure. Table 4.7 suggests that there is some variation in α across occupations.

Table 4.7: The ratio of average urban wage to average rural wage by occupation ($\boldsymbol{\alpha}$)

Occupation	Mean of urban wage (yuan)	S.d	Mean of rural wage	S.d.	The ratio of urban wage/rural wage
Unemployed or student	5612	5483	4112	4419	1.36
Farmer	3961	3424	3696	5106	1.07
Non-farm	11186	11702	8632	8802	1.30
Skilled	7830	5912	6474	4762	1.21
Non-skilled	7075	6046	5650	4227	1.25
Service	8486	8340	6502	6264	1.31

To test whether α is common across occupations, we estimate the following equation:

$$\ln W_{oa}^{j} = \eta + \sum_{o=2}^{6} \beta_o D_o + \delta\, D_r + \sum_{o=2}^{6} \gamma_o\, D_o * D_r \qquad (4.29)$$

where $\ln W_{oa}^{j}$ denotes the log wages of individual j, dummy D_o denotes the type of occupation, among which the reference group (o=1) is "unemployed or student", it equals one if the occupation is o and zero otherwise; dummy D_r equals one if the respondent is from the rural area and zero otherwise, $D_o * D_r$ equals one if the occupation of individual j is o and they are from a rural area, and η is the constant for urban unemployed/students. For instance, when o equals four (the skilled worker occupation), $D_4 * D_r$=1 captures all the rural skilled workers. Therefore, coefficient β_o captures the effects of being a skilled worker, δ captures the effects of being rural areas and γ_o captures the difference in δ by occupation, testing whether the effects of coming from a rural area is the same across occupations. If it is the same across occupations, it indicates that α is common across occupations.

The estimates are presented in Table 4.8. We can see that interactions for "non-farm", "skilled" and "non-skilled" with the rural dummy are significant, suggesting that the differentials are significantly different from the unemployed or students. Through testing the interactions, the coefficients between occupations and rural dummy D_r do not significantly differ across the five occupations (in the test we ignored the interaction between rural area and farmer because urban farmer is a small special group). We also tested the joint significance of interactions between the "rural" dummy with occupations (the p-value of the χ^2 test is 0.151), suggesting these interactions are not jointly significant. Overall, these tests suggest that α varies by occupation but not significantly and not particularly systematically.

Table 4.8: The estimation of wage equation for testing the urban-rural wage differences by occupation

Dependent variables: Log(wage)	Pooled coeff	s.e.
Occupations: Unemployed or student (Ref.)		
Farmer	-0.306***	(0.05)
Non-farm	0.726***	(0.05)
Skilled	0.447***	(0.06)
Non-skilled	0.323***	(0.06)
Service	0.441***	(0.06)
Rural/Urban(Rural=1)	-0.340***	(0.04)
Farmer* rural	0.189***	(0.06)
Non-farm* rural	0.123*	(0.07)
Skilled* rural	0.163*	(0.08)
Non-skilled* rural	0.147*	(0.08)
Service* rural	0.084	(0.07)
Constant	8.282***	(0.03)
Observations	8677	

As we see, $(\frac{\partial Z}{\partial h})^j = (\alpha - 1)\frac{\partial W_s^j}{\partial h^j}$ - $i\tau = (\alpha - 1)W_0\lambda$ - $i\tau$, α varies by occupation, but not greatly and $i\tau$ is assumed constant over occupations, so the differences in the coefficients $(\frac{\partial Z}{\partial h})^j$ by occupation might reflect the differences in the response of wages to health $\frac{\partial W_s^j}{\partial h^j}$ (or λW_0) by occupation. Since we know the coefficients $(\frac{\partial Z}{\partial h})^j$ (Table 4.6) and α (Table 4.7), we can obtain λW_0 by dividing $(\frac{\partial Z}{\partial h})^j$ by $(\alpha - 1)$ (the results are reported in Table 4.9). $\frac{\partial W_s^j}{\partial h^j}$ is the product of λ and W_0, among which λ (the marginal (average) effects of health on the wage) varies by occupation, and W_0 (the individual wage at the base level of health) also varies by occupation. For instance, for skilled workers, λ might decline whereas W_0 might increase, but it is unknown which force is stronger. Also, it is difficult to test, partly due to the wage here not being an adequate measure. Table 4.9 suggests that in most of the occupations, $\frac{\partial W_s^j}{\partial h^j}$ increases as health improves, a result that accords with the estimates of $(\frac{\partial Z}{\partial h})^j$ in Table 4.6.

Table 4.9: The sensitivity of wage with respect to health by occupation

		Poor	Fair	Good	Excellent
(1)	Unemployed or student	0	1.06	1.3	1.5
(2)	Farmer	6	6.86	7.14	7.43
(3)	Non-farm	-0.7	0.17	1.17	1.1
(4)	Skilled	-0.86	2.90	2.57	1.76
(5)	Non-skilled	-0.2	1.08	1.84	2.88
(6)	Service	1.90	1.52	1.48	1.48

In summary, the estimates for $(\frac{\partial Z}{\partial h})^j$ suggest that the effects of being self-evaluated as having "good" or "excellent" health on the migration probability are larger for people with an initial occupation of a lower skilled worker than for those who worked as a higher skilled worker. Based on $(\frac{\partial Z}{\partial h})^j = (\alpha - 1)\frac{\partial W_s^j}{\partial h^j} - i\tau,$ assuming $i\tau$ is constant over occupations, we find α varies by occupation, though not greatly. The differences in $(\frac{\partial Z}{\partial h})^j$ by occupation might also be driven by the variation in $\frac{\partial W_s^j}{\partial h^j}$ (or $W_0\lambda$), among which the sensitivity of wage to health, λ, which tends to be larger for construction work than higher service work, might be the dominating force. Additionally, sensitivity to monetary returns (higher urban wages, α) might be different across occupations. Overall, we admit that we can not make much order out of these results, partly because the wage here is not a very accurate measure.

4.5.3 Health Interacts with Education

Using education as an alternative proxy for wages, we interact health with education and repeat a similar exercise to the above. The estimates are reported in Table 4.10, with the coefficients for the education variables capturing the increments of having different levels of education relative to primary education or lower, for people in poor or fair health. Using the baseline equation but without the objective health measures, now the sample size now becomes 8769.

Table 4.10: The estimates of health effects by education

	Pooled coeff	s.e.
Self-rated health: Poor (Ref.)		
Fair	0.092	(0.21)
Good	0.194	(0.21)
Excellent	0.226	(0.22)
Highest degree: Primary or lower (Ref.)		
Lower Middle	0.036	(0.29)
Upper Middle	-0.120	(0.13)
Technical/Vocational	-0.198	(0.17)
College and Beyond	-0.028	(0.21)
Interactions		
Fair* Lower Middle	0.017	(0.31)
Fair* Upper Middle	0.300	(0.19)
Fair* Technical/Vocational	-0.235	(0.26)
Fair* College and Beyond	-0.378	(0.36)
Good* Lower Middle	-0.077	(0.30)
Good* Upper Middle	0.021	(0.15)
Good* Technical/Vocational	0.153	(0.19)
Good* College and Beyond	0.201	(0.23)
Excellent* Lower Middle	-0.011	(0.31)
Observations	8769	

Note: The equation also includes other controls in the baseline equation (except for the objective health measures); standard errors are in Parentheses, *** p<0.01, ** p<0.05, * p<0.1

To facilitate the comparison, the direct coefficients are presented in Table 4.11. The interactions are insignificant, suggesting no significant variation in health effects across education levels.

Table 4.11: Partial interaction of health with education

	Poor		Fair		Good		Excellent	
	Coef.	Sd.	Coef.	Sd.	Coef	Sd.	Coef	Sd.
Primary	0	0	0.09	(0.21)	0.19	(0.21)	0.23	(0.22)
Lower middle	0.04	(0.29)	0.15	(0.21)	0.15	(0.20)	0.25	(0.21)
Upper middle	-0.12	(0.13)	0.27	(0.23)	0.09	(0.21)	0.11	(0.22)
Technical	-0.20	(0.17)	-0.34	(0.27)	0.15	(0.22)	0.03	(0.24)
College	-0.03	(0.21)	-0.31	(0.35)	0.37	(0.24)	0.20	(0.27)

4.5.4 Indirect Channel and Lagged Health

As discussed in the theoretical discussion, prior studies suggest that earlier health (especially childhood health) has a lasting impact on later education attainment (Case, Fertig, and Paxson 2005). To account for the fact that skill k might pick up the effects of h_{-1} (the indirect effects of earlier health), we introduce k^j as a function of lagged health (h_{-1}^j), $k^j = k^j(h_{-1}^j)$ into the model, thus we have

$$(\frac{dZ}{dh_{-1}})^j = (\alpha-1)W_{00}[\frac{\partial k^j}{\partial h_{-1}^j} + \lambda k^j \frac{\partial h^j}{\partial h_{-1}^j} + \lambda h^j \frac{\partial k^j}{\partial h_{-1}^j}]$$

$$=(\alpha-1)W_{00}\,\lambda \frac{\partial h^j}{\partial h_{-1}^j} + (\alpha-1)W_{00}(1+\lambda h^j)\frac{\partial k^j}{\partial h_{-1}^j} \qquad (4.30)$$

Equation (4.30) suggests that we estimate an equation that includes the interaction of lagged health with current health. However, since the health variable here is a categorical variable that includes four categories, the interaction of two four-category categorical variables might introduce a complication into the estimation. Therefore, for now, we do not include these interactions in the estimation.

To investigate how h_{-1}^j affects k^j , we estimate the effects of lagged health on education. These health effects might operate through promoting the probability of moving on to a higher degree or improving performance during the same degree. We cannot estimate the latter type of effects here, due to the lack of information on schooling performance. For the first type of effect, substantial evidence suggests that children who are in poor health tend to have lower education attainments, which are often measured by years of schooling (Behrman 1996, Smith 2009).

To examine the effects of earlier health on the highest education degree obtained later in life, we go back to the original CHNS data and used a sample consisting of those who were observed when they were aged between 13 and 16 years. Based on this sample, we estimate the effects of their self-rated health when they were aged between 13 and 16 years on the highest degree they obtained after they were 16 years old. In the literature (Smith 2009), the classical equation for this is:

$$Education\ level\ at\ adult = health\ at\ earlier\ age + family\ characteristics + \varepsilon \qquad (4.31)$$

Using an ordered logit model, we follow the basic shape of equation (4.31) and also include parental socioeconomic factors and regional fixed effects in the estimation. The education degree ranges from the lowest (“primary and below”) to the highest (“college and beyond”), including five categories. The results are reported in Table 4.12 and suggest that self-evaluating as having “fair” “good” or “excellent” health at 13-16 years of age has a significantly positive effect on the probability of obtaining a higher education degree after the age of 16. This result indicates that better earlier health improves the probability of obtaining a higher degree later in life. In addition, the coefficient is larger for “fair”, small for “good” and smaller for “excellent”. This result implies that beyond the small fraction (2%) of children with “poor” health, who barely had the chance of an education, children with “excellent” health might be sent to work rather than go to school, whereas those with “fair” or “good” health received an increased chance of attaining a higher education. The above shows the response of education outcome to earlier health and earlier the estimation of our main equation (Table 4.4) showed the effects of education on the propensity to migrate. One might consider estimating the indirect effects of health on migration by substituting the equation for earlier health on education into the main migration equation. However, the limited sample size (N=1262) does not allow us to create this reduced form equation. Nonetheless, Table 4.12 provides some evidence that children with “fair” or “good” or “excellent” health are more likely to migrate than those with “poor” health.

Table 4.12: Ordered logit estimates of the health effects at age 13-16 years on the highest education degree obtained after age 16

Dependent variable: The probability of obtaining a higher education degree after age 16		
	Coeff.	s.e.
Self-rated health aged 13-16: Poor (Ref.)		
Fair	1.734***	(0.40)
Good	1.454***	(0.35)
Excellent	1.335***	(0.38)
Age	0.004	(0.03)
Gender (Male=1)	-0.032	(0.16)
Father's occupation: Unemployed/student (Ref.)		
Farmer	-0.800***	(0.18)
Non-farm	0.288	(0.29)
Skilled	-0.156	(0.26)
Non-skilled	-0.644***	(0.23)
Service	0.558**	(0.27)
Mother's education: Primary and below (Ref.)		
Lower Middle	7.181***	(1.17)
Upper Middle	12.853***	(1.43)
Technical/Vocational	15.579***	(1.50)
College and Beyond	34.689***	(1.56)
Household size	-0.024**	(0.01)
Household Income per capita (in 2011 currency, logged)	-0.909	(2.53)
Region: Coastal (Ref.)		
Northeast	-0.653***	(0.24)
Inland	-0.190	(0.21)
Southern Mountain	-0.217	(0.23)
Observations	1262	

Standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1

In summary, earlier health might have a positive effect on the later education outcome; however, it is worth noting that education might depend on expected migration. Studies suggest that since the returns to upper middle school or a higher level are not higher than those for lower education levels (Schultz 2004), the opportunity costs of attending upper middle school might be higher than the opportunity costs of attending lower middle school. As a consequence, upon the completion of lower middle school, many youths in rural China often migrate than pursue a higher education degree. Therefore, there is a negative relationship between migrant opportunity and upper middle school enrolment (DeBrauw and Giles 2008). These relationships of health with education and education with expected migration greatly complicate the study of the effects of health on migration. Similarly, early health investment might rely on the expectation of migration. Unfortunately, with our limited information in this data, we cannot deal with these potential reverse causalities in this study. However, we recognise this as a potential complication in our estimates of the relationship running from education to migration and health to migration.

Next, we examine the effects of lagged health on migration. In the literature, the long term effects of heath have not been widely examined due to data limitations, with the examination of long term effects usually requiring a longitudinal survey that follows people for a given period. The CHNS longitudinal survey provides the possibility of investigating this effect, although as Table 4.1 suggests, there are not a large number of people tracked for more than two waves. Nonetheless, we can still try to estimate the effects of lagged health to ascertain some insight on the long term effects.

Before estimating the effects of lagged health on migration, it is useful to get a sense of the correlation between lagged health and current health. Based on our pooled sample aged between 16 and 35 years old (N=8790), Table 4.13 presents the transition matrix for lagged health with current health. Through describing the distribution of current health status conditional to the previous health status, Table 4.13 shows the transition probabilities of health status from the previous period $(t-1)$ to the current period (t), and provides a sense of how health status evolves over time. As Table 4.13 shows, for those with "good" health at $t-1$, 21% saw their health get better (changed to "excellent") in the next period, whereas 22.04% saw their health worsen (changed to "poor" or "fair"); more than half (57%) saw their health status stay the same. Therefore, Table 4.13 reveals a stronger transmission of "good" health status from period t-1 to period t, compared to the health status "excellent" and "poor/fair", with there being a tendency for people across different health statuses converging to "good" health in the next period. The χ^2 test rejects the null hypothesis that health at $(t-1)$ and health at t are independent; health at $t-1$ is correlated with health at t. Therefore, the significant effects of current health in the baseline equation might capture the effects of lagged health.

Table 4.13: The transition of health (t) from health (t-1)

		Health (t)				
		Poor	Fair	Good	Excellent	Total
Health (t-1)	Poor	12.5	29.17	50	8.33	100
	Fair	4.29	25.04	55.23	15.44	100
	Good	2.24	19.8	56.97	21	100
	Excellent	0.81	13.73	52.49	32.97	100
	Total	2.42	19.5	55.59	22.49	100

Pearson chi2(9) = 118.7767 Pr = 0.000

As a result, instead of current health, we now estimate the effects of lagged health alone on migration. The results are reported in column (1) of Table 4.14 and suggest that lagged health effects are insignificant. After, we added current health into the estimation, with neither lagged health or current health being significant (as shown in Table 4.14, column (2)). We tested the joint significance of lagged health and current health (the p-value of the χ^2 test is 0.376) and suggest that lagged health and current health are not jointly significant. Based on the sample equation in column (2), Table 4.14, Column (3) presents the results when the equation includes only current health, with the results suggesting that the effects of current health are insignificant. Table 4.14, together with Table 4.13, imply that lagged health might not have significant effects on migration, as well as lowering the significance of current health, although they closely correlate with each other. However, this might be due to the limited information on lagged health in this small sample.

Table 4.14: Probit regression of migration status on lagged health (t-1)

	(1) Pooled Coeff	s.e.	(2) Pooled Coeff	s.e.	(3) Pooled Coeff	s.e.
Self-rated health: Poor (Ref.)						
Fair			0.145	(0.19)	0.148	(0.19)
Good			0.167	(0.18)	0.171	(0.18)
Excellent			0.244	(0.19)	0.254	(0.19)
Fair t-1	-0.290	(0.22)	-0.297	(0.23)		
Good t-1	-0.200	(0.21)	-0.197	(0.22)		
Excellent t-1	-0.125	(0.22)	-0.127	(0.23)		
Observations	3437		3384		3384	

Note: The equation also includes other controls in the baseline equation (except for the objective health measures); standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1

4.5.5 The Effects of Change in Health Status

As an extension of the analysis of lagged health effects, we will now look at the relationship between the change in health status from $t-1$ to t and migration at t. In doing so, we aim to explore whether the improvement in health raises the possibility of migration; more specifically, whether there is a group of unhealthy people who postponed migration until their health improved.

Based on our pooled sample aged between 16 and 35 years old, Table 4.15 presents this relationship in a transition matrix form. It suggests that the proportion of migrants is larger for those whose health statuses improved (16.36%) than those whose health statuses remained the same (14.2%) and those whose health declined (14.88%). The χ^2 test here tests the independence of the variable for “health improved or not” from the variable for “migration status” (the p-value for this test is 0.394), with the distribution of “health declined”, “health remained the same” and “health improved” not being significantly different for migrants and non-migrants. The improvement in health is not significantly associated with the migration decision.

Table 4.15: Change in health from (t-1) to (t) and migration at (t)

		Migration status at t		
		Non-migrant	Migrant	Total
Health from t-1 to t	Decline	85.12	14.88	100
	Remained the same	85.8	14.2	100
	Improved	83.64	16.36	100
	Total	85.09	14.91	100
		2,649	464	3,113

Pearson chi2(1) = 1.8637 Pr = 0.394

The estimates above might be subject to bias due to the unobserved heterogeneity associated with both health status and the probability of migration, such as previous life exposure and genetics. The observed relationship might be indications of highly selective characteristics of migrants that affect both health status and the decision to migrate. To allow for the unobserved heterogeneity fixed at the household level, we follow Lu (2008)'s study and apply a household fixed effect (FE) model. As mentioned earlier, using the 1997 and 2000 waves from the Indonesian longitudinal survey (IFLS), Lu (2008) tested the health selectivity hypothesis and adopted the household fixed effects model to test the robustness of her results. Our household fixed effects estimates are reported in Table 4.16, column (1) and suggest that the change in health status does not significantly correlate with the change in migration probability, assuming household heterogeneity, such as family background and genetic disposition, are constant over time. Similarly, column (2) reports the individual fixed-effect (FE) estimates and suggests that the health effects are not significant; it is important to note that the sample sizes are small though.

In addition, we also apply the individual random effects model, with the results presented in Table 4.16, column (3). They suggest that "excellent" health has a significant effect on migration probability. Notice the assumption for random effects is strong and the unobserved effect is independent of all explanatory variables across all time periods. Additionally, these random effects estimates are close to the pooled probit estimates shown in Table 4.4, since the individual random effects logit model is very similar to the probit model on the pooled sample (as shown in equation (4.28)). As fixed effects model are estimated for individuals or households that are repeatedly observed, the sample for the fixed effects estimation are substantially smaller than those used in the random effects estimation. Table 4.16, column (4) presents the individual random effects estimates using the fixed effects model sample and shows that the significance of health effects disappear because the sample is too small.

Table 4.16: Logit fixed effects and random effects on pooled sample

	(1)	(2)	(3)	(4)
	Household FE	Individual FE	Individual RE	Individual RE
Fair health	-0.116	-13.167	0.304	-0.091
	(0.40)	(2179.45)	(0.29)	(0.72)
Good health	-0.114	-12.565	0.405	-0.324
	(0.39)	(2179.45)	(0.28)	(0.71)
Excellent health	-0.088	-12.738	0.489*	-0.251
	(0.41)	(2179.45)	(0.29)	(0.72)
Observations	2801	1074	8790	1074
Pseudo R^2	0.069	0.926		

Note: The equation also includes other controls in the baseline equation (except for the objective health measures); standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

In conclusion, the change in health is not significantly associated with the migration decision, we cannot identify the health effects with fixed effects estimation, potentially due to the small sample size.

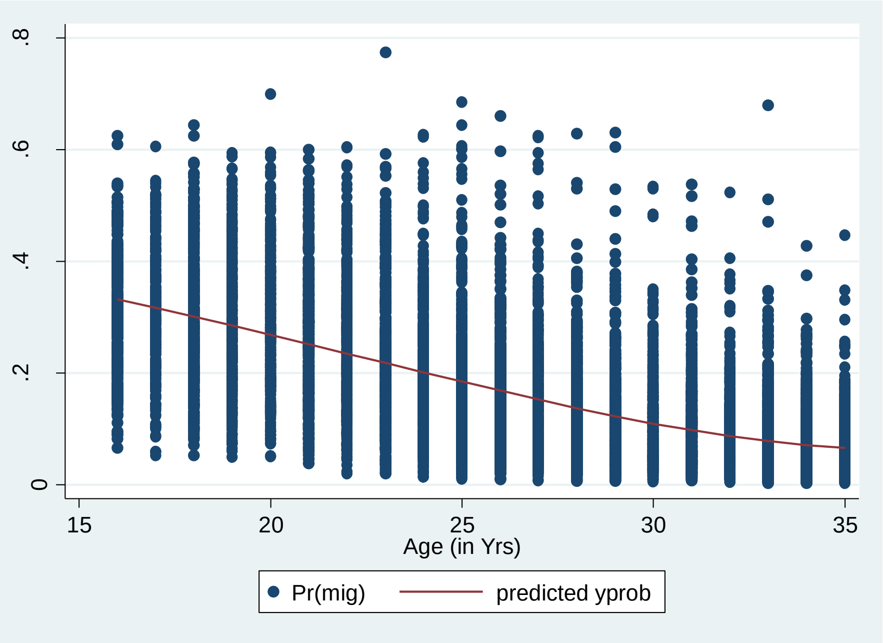
4.5.6 Health Interacts with Age

Recall that in the theoretical model, the time horizon is infinite and the same for everyone, so the migration probability is not expected to be higher for the young than it is for the old. However, standing outside the model, according to the standard human capital framework that views migration as an investment, the time horizon is finite. Therefore, the time for the expected higher income to offset the migration costs (i.e., the payoff period) falls as the worker gets older, with the migration probability expected to be higher for the young than for the old. To illustrate this, using our pooled sample aged 16 to 35 years old, we obtained the predicted migration probability from the baseline equation (without objective health measures)[49], and plotted it against age in Figure 4.4. It suggests that the migration probability declines with age and that this declining slope

[49] The equation here is the one shown in Table 4.4 without the variables “ADLs”, “bone fracture” and “ever smoked”.

reflects the age effects on migration, with people migrating less as they get older in this sample.

Figure 4.4: The migration probability and age



To explore these age effects further, in addition to the age continuous variable, we create annual dummies for each age level and include these 20 age dummies (age 16-35 years) in the baseline equation. Based on the pooled sample in our baseline estimation (N=8790)[50], the estimates are presented in Table 4.17, along with the estimates from the baseline equation. They suggest that compared with those who are aged 16 years old, almost all those who are older than 16 are less likely to migrate, which might be related to the fact that age 16 is the legal working age in China, so many youths aged 16 migrate to work. However, including these annual age dummies does not make a large difference to the estimates for health and other variables. The health effects estimates are barely affected by the inclusion of these age dummies, which might be due to there not being a large variation in health over this age range (16-35 years).

[50] The sample size is different from the one in Table 4.4, since here (Table 4.17) we do not include the objective health measures.

Table 4.17: Migration equation including 20 age dummies

	(1) Pooled b/se	(2) Pooled b/se
Dependent variable: the probability of migration		
Self-rated health: Poor (Ref.)		
Fair health	0.180	0.179
	(0.15)	(0.15)
Good health	0.239	0.238
	(0.15)	(0.15)
Excellent health	0.285*	0.281*
	(0.15)	(0.15)
Age (years)		-0.044***
		(0.01)
Age dummies: 16 years (Ref.)		
19 age dummies from 17-35 years old	Y	
Gender (Male=1)	0.142***	0.142***
	(0.04)	(0.04)
Marital Status	0.080	0.080
	(0.14)	(0.14)
Highest degree: Primary and lower (Ref.)		
Lower middle school	-0.013	-0.010
	(0.05)	(0.05)
Upper middle school	-0.067	-0.069
	(0.07)	(0.07)
Technical/Vocational school	-0.132	-0.138
	(0.09)	(0.08)
College and beyond	0.045	0.042
	(0.12)	(0.12)
Occupation: None/student (Ref.)		
Farmer	0.052	0.040
	(0.06)	(0.05)
Non-farm	-0.157*	-0.170**
	(0.09)	(0.08)
Skilled	0.048	0.035
	(0.08)	(0.08)
Non-skilled	0.032	0.018
	(0.08)	(0.07)
Service	0.008	-0.004
	(0.07)	(0.06)
Previous Migration Experience	0.393***	0.390***
	(0.05)	(0.05)
Rural/Urban(Rural=1)	0.393***	0.395***
	(0.05)	(0.05)
The number of people in household	0.074***	0.073***
	(0.02)	(0.02)

Household Income per capita (in 2006 currency, logged)	-0.936	-0.940
	(1.06)	(1.05)
Parents: Both parents <56 (Ref.)		
One parent's age > 55	-0.008	-0.020
	(0.07)	(0.07)
Both parents' age > 55	-0.004	-0.021
	(0.08)	(0.07)
No parents	-0.073	-0.081
	(0.07)	(0.07)
spouse	-0.188	-0.195
	(0.14)	(0.14)
child	-0.133**	-0.135**
	(0.06)	(0.05)
Region: Coastal (Ref.)		
Northeast	-0.286***	-0.289***
	(0.07)	(0.07)
Inland	0.204***	0.204***
	(0.06)	(0.06)
Southern mountain	0.206***	0.204***
	(0.06)	(0.06)
2000	0.255***	0.248***
	(0.05)	(0.05)
2004	0.251***	0.245***
	(0.06)	(0.06)
2006	0.155**	0.146**
	(0.06)	(0.06)
Constant	9.716	10.415
	(12.66)	(12.56)
Observations	8790	8790

Since including the annual age dummies does not significantly change the coefficients of other variables, next, when we introduced the interactions of health with age, for the sake of brevity, we collapsed these age dummies into four groups and interact health with these four age groups. These four groups are 16-18, 19-24, 25-30 and 31-35 years of age. We choose the ages 18, 24 and 30 as the thresholds for the following reasons: 18 is another education milestone due to the fact that 18 is the typical age for upper middle school completion, also the age dummies are significant until the age 19; age 24 and 30 are the breaks over which there are significant changes in the magnitude of their coefficients[51]. The estimates are presented in Table 4.18, column (1) and suggest that the health effects do not vary much with the age group.

[51] The estimates for these age dummies are available on request.

Table 4.18: Probit regression of migration including age groups and the interactions between health and age group

	(1)		(2)	(3)
Dependent variable: the probability of migration				
Self-rated health: Poor		Self-rated health: Poor		
Fair health	0.388	Fair	0.012	0.176
	(0.44)		(0.08)	(0.24)
Good health	0.349	Good	0.044	0.218
	(0.43)		(0.07)	(0.23)
Excellent health	0.502	Excellent	0.079	0.187
	(0.44)		(0.08)	(0.24)
Age group: 16-18 years (Ref.)		Age group: 16~25 years (Ref.)		
Age 19-24 * Fair	-0.411	26~35 years	-0.204***	-0.229
	(0.54)		(0.06)	(0.30)
Age 19-24 * Good	-0.311	36~45 years	-0.278***	-0.084
	(0.52)		(0.10)	(0.29)
Age 19-24 * Excellent	-0.524	46~55 years	-0.229*	0.016
	(0.53)		(0.14)	(0.29)
Age 25-30 * Fair	-0.263	56~65 years	-0.162	-0.006
	(0.52)		(0.18)	(0.32)
Age 25-30 * Good	-0.137	Fair * 26~35 years		-0.006
	(0.50)			(0.31)
Age 25-30 * Excellent	-0.179	Fair * 36~45 years		-0.272
	(0.51)			(0.28)
Age 31-35 * Fair	-0.060	Fair * 46~55 years		-0.227
	(0.53)			(0.27)
Age 31-35 * Good	0.046	Fair * 56~65 years		-0.094
	(0.51)			(0.28)
Age 31-35 * Excellent	-0.044	Good * 26~35 years		0.009
	(0.52)			(0.30)
		Good * 36~45 years		-0.191
				(0.27)
		Good * 46~55 years		-0.297
				(0.26)
		Good * 56~65 years		-0.279
				(0.27)
		Excellent *26~35		0.098
				(0.30)
		Excellent * 36~45		-0.137
				(0.28)
		Excellent * 46~55		-0.173
				(0.28)
		Excellent * 56~65		0.010
				(0.30)
Observations	8790	Observations	26998	26998

Note: The equation also includes annual age dummies and other controls in the baseline equation (except for the objective health measures); standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1

We then raised the upper age limit from 35 years to 65 years old and estimated the baseline equation (without objective health measures) for this sample. The results are not reported here, with age as a continuous variable, suggesting that on average, the health effects are not significant. One potential explanation is that the positive health effects from people outside of the age range 16~35 years might be smaller, and, as mentioned earlier, might even be negative. This force dilutes or offsets some of the positive health effects from those aged 16-35 years, so overall, the positive health effects disappear.

Next, we created a categorical variable defined by 10-year age groups ranging from 16 to 65 years, and included it in the equation. The results are presented in Table 4.18, column (2) and suggest that people aged 26-35 and 36-45 years are less likely to migrate, compared to those aged between 16 and 25 years, with this negative age effects smaller for those aged 46-55 years and 56-65 years old. In other words, these estimates indicate that the middle aged are least likely to migrate, but the old are relatively more likely to move than the middle aged. This accords with the "salmon bias effects" theory that states people are likely to migrate when they get old.

To examine the variability of health effects with age level, we also interact the self-rated health with age group and include them into the equation (the results are presented in Table 4.18, column (3)). Those interactions are not significant and we tested the joint significance of these interactions (the p-value of the χ^2 test is 0.546), with the results suggesting that they are not jointly significant. However, the positive signs for the interaction term of the "26-35 years" and "36-45 years" age groups with "good health" and negative signs for the interaction term of the "46-55 years" age group, "56-65 years" group with "good health" indicate a pattern as the theory predicted: younger people with good health are more likely to move than those with poor/fair health, whereas old people with good health are less likely to move than those with poor health.

4.5.7 An Alternative: Health Index

Using self-reported health alone might lose some useful information, but using several health measures might cause a decrease in the sample size. Next, we attempted to obtain a health index that has three main advantages: first, this index concentrates various health information in the data down to one single effect; second, this index allows us to extend the data and make more use of the data by using more health measures in the data; and third, since this index is continuous, it allows us to examine some effects that are difficult to estimate when health is a categorical variable.

To start with, we converted the categorical variable self–rated health to a binary variable that is equal to one if the respondents evaluate their health as being “good” or “excellent”, and zero otherwise. Using the pooled sample, the results are presented in Table 4.19, column (1) suggests that those self-evaluating as having better health are more likely to migrate. Since using self-rated health alone might lose some health information in the data, to achieve a better coverage of health information in the data, we created a health index that absorbs both self-rated health and objective measures. The three objective health measures are mainly the objective measures used in Tong and Piotrowski (2012)’s study (except for “ever smoked”): bone fracture “Do you have a history of bone fracture”, ADLs “did you have trouble working due to illness in the last 3 months”, and high blood pressure “diagnosed with higher blood pressure or not”. They are coded as binary variables, which is equal to one if the answer to those questions is “No”, and zero otherwise. Therefore, for variables used in the index, a higher value indicates better health. We assigned equal weight to the binary self-rated health variable and three objective health measures individually, and take the sum of them as an index[52] (the estimates are reported in Table 4.19, column (2)). After absorbing the e objective health measures, the health effects become insignificant.

We next used the categorical version of self-rated health that takes the value 0 if the respondents evaluate themselves as having “poor” health, 1 if “fair” health, 2 if “good” health and 3 if “excellent” health. The results are presented in Table 4.19, column (3) and are consistent with the earlier results when we used the binary version of self-rated health (column (1)), with those self-evaluating as having better health more likely to migrate. Next we assigned weights to these health measures; first, we assigned equal weights to the self-rated health and objective measures, then gave half (1/2) weight and one and half (3/2) weights to the self-rated health as to objective measures[53] (the results are presented in Table 4.19, columns (4), (5) and (6), respectively). They suggest that the health effects are insignificant, except when the self-rated health is assigned one and half weights in the index. This suggests that the health effects become significant as the weights for self-

[52] Henceforth we will refer to the indices used in column (1) and (2) as Type 1 index.

[53] Henceforth we will refer to the indices used in column (3), (4),(5) and (6) as Type 2 index.

rated health increase.

Table 4.19: Probit regression of migration using different indices

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Type1 index		Type 2 index				Type 3 index	
	Index1	Index2	Index3	Index4	Index5	Index6	Index7	Index8
Health	0.079^{*}	0.038	0.059^{**}	0.040	0.045	0.032^{*}	0.322^{***}	0.390^{**}
index	(0.05)	(0.04)	(0.03)	(0.03)	(0.04)	(0.02)	(0.11)	(0.15)
Obs	8782	8536	8790	8536	8536	8536	8897	8959

Notes: health index includes: in column (1), self-rated health, a binary variable which is valued 1 if "good" or "excellent", 0 otherwise; in column (2), self-rated health, a binary variable which is valued 1 if "good" or "excellent", 0 otherwise, and three objective measures. They are weighted equally in the index; in column (3), self-rated health, a variable which is valued 0 if "poor", 1 if "fair", 2 if "good" and 3 if "excellent"; in column (4), self-rated health, a variable which is valued 0 if "poor", 1 if "fair", 2 if "good" and 3 if "excellent", and three objective measures. They are weighted equally in the index; in column (5), self-rated health, a variable which is valued 0 if "poor", 1 if "fair", 2 if "good" and 3 if "excellent", and three objective measures. The self-rated health is assigned half weight as the objective measures in the index; in column (6), self-rated health, a variable which is valued 0 if "poor", 1 if "fair", 2 if "good" and 3 if "excellent", and three objective measures. The self-rated health is assigned one and half weights as the objective measures in the index; in column (7), self-rated health, a variable which is valued 0 if "poor", 1 if "fair", 2 if "good" and 3 if "excellent", long term and short term health, we assign triple, double and single weights to them, respectively; in column (8), self-rated health, a variable which is valued 0 if "poor", 1 if "fair", 2 if "good" and 3 if "excellent", long term and short term health, we assign triple, double and single weights to them, respectively; the missing values of the objective health are imputed with positive responses (ie. "no, I do not suffer from this problem"). The equation also includes other controls in the baseline equation (except for the objective health measures).

However, there might not be enough information in the self-rated health and three objective measures used here to obtain a measure with a larger coverage of the information, so we need to go back to the original data and absorb a variety of other health measures. The measures we use are listed in Table 4.20. All the binary variables are recoded as those that are equal to one if the respondents did not have those symptoms or diseases, and zero otherwise. Self-rated health is maintained as a variable that is equal to 0 if the respondent evaluated their health as being "poor", 1 if "fair", 2 if "good" and 3 if "excellent". Based on our sample comprised of full sets of observations (N=8897), the summary statistics for those variables and health index are presented in Table 4.21. We assigned different weights to these variables according to their relative importance. Since self-rated health is an indicator that reflects overall health and individual behaviour, bone fracture, high blood pressure, overweight, diabetes, myocardial infarction, apoplexy and ADLs that tend to reflect long-term health, whilst health conditions in the last four weeks concern short-term health relatively, we applied triple weights to the self-rated health, a single weight to those regarding health conditions in the last four weeks and double weights to the long-term health indicators. As Table 4.21 shows, there is a variety of missing rates across these variables. To maximise the information from these variables, for any individual with at least four observations across these variables, we take the mean of their values and use it as a health index. Using this index, the estimation results are presented in Table 4.19, column (7) and suggest that those with a larger health index are more likely to migrate. Together with columns (1), (3) and (6), these results suggest that the health effects turn out more strongly when the index uses self-rated health alone or gives more weight to self-rated health. Table 4.21 reveals that there is a high missing rate among the short term health variables (in the last four weeks). In case those without health problems might be coded as missing, we next impute the missing values with positive responses (i.e. "no, I do not suffer from this problem")[54]. The results are presented in Table 4.19, column (8) and based on a larger sample obtained from the imputation, with the results suggesting that those with larger health index are more likely to migrate.

Table 4.20: The description of variables used in the health index

Variable	Variable description
Self-rated health	health (current health status (self-report))
ADLs	trouble working due to illness in the last three months? =1 if yes; =0 if no
Bone fracture	Have a history of Bone Fracture? =1 if yes; =0 if no
High blood pressure	diagnosed with high blood pressure? ? =1 if yes; =0 if no
Overweight	= 1 if bmi>=30, =0 otherwise
diabetes	diagnosed with diabetes? ? =1 if yes; =0 if no
myocardial infarction	diagnosed with myocardial infarction? ? =1 if yes; =0 if no
apoplexy	diagnosed with apoplexy? =1 if yes; =0 if no
Sick in the last 4week	been sick or injured in last 4 weeks ? =1 if yes; =0 if no
fever in the last 4week	last 4 wks: fever, sore throat, cough? =1 if yes; =0 if no
headache in the last 4week	last 4 wks: headache, dizziness? =1 if yes; =0 if no
Muscle pain in the last 4week	last 4 wks: joint, muscle pain? =1 if yes; =0 if no
Heart chest in the last 4week	last 4 wks: heart disease/chest pain? =1 if yes; =0 if no
Seek health care in the last 4	last 4 wks: preventative hlth service? =1 if yes; =0 if no
Seek formal medical care in the last 4week	last 4 wks: seek formal medical care? ? =1 if yes; =0 if no

[54] Henceforth, we will refer to the indices used in column (7) and (8) as Type 3 index.

Table 4.21: The summary statistics of health variables used in the health index

Variable	Obs	Mean	Std. Dev.	Min	Max	Freq. Missings in our sample (N=8897) (%)
Health index	8897	1.713	0.227	0.727	2.667	0
Self-rated health	8782	2.009	0.672	0	3	1.293
ADLs	8701	0.028	0.165	0	1	2.203
Bone fracture	8817	0.021	0.142	0	1	0.899
High blood pressure	8846	0.004	0.065	0	1	0.573
Overweight	7502	0.016	0.127	0	1	15.68
diabetes	8731	0.002	0.048	0	1	1.866
myocardial infarction	8826	0.001	0.021	0	1	0.798
apoplexy	8757	0.001	0.021	0	1	1.574
Sick in the last 4week	8776	0.050	0.217	0	1	1.36
fever in the last 4week	3392	0.092	0.289	0	1	61.87
headache in the last 4week	3386	0.038	0.190	0	1	61.94
Muscle pain in the last 4week	3382	0.014	0.116	0	1	61.99
Heart chest in the last 4week	3382	0.003	0.054	0	1	61.99
Seek health care in the last 4week	8706	0.018	0.134	0	1	2.147
Seek formal medical care in the last 4week	3002	0.012	0.110	0	1	66.26

As mentioned earlier, since the health index is continuous, we can examine some effects that might be intractable when health is a discrete variable. Therefore, we interact different health indices with occupation (the results are presented in Table 4.22). It suggests that there are interactive effects when we use Type 2 and Type 3 indices (columns (3) to (8)). Using the Type 2 index apart from the one in which self-rated health is assigned half weights when combined with three objective health measures, columns (3), (4) and (6) suggest that for respondents with an initial occupation type as unemployed or student, those who have a larger health index are significantly more likely to migrate. The coefficients for the "skilled workers" are significantly positive, implying that skilled workers are more likely to migrate than those who are unemployed or students. The coefficients for the interaction term of skilled worker with the health index are significantly negative, suggesting that compared to those who are unemployed and a student, health has a less strong positive relationship to migration probability for skilled workers. When the indices also absorbs other health information (Type 3 index), columns

(7) and (8) suggest that the health effects are positive (though not significant in column (8)) for those who are unemployed or students; those who are non-skilled workers are significantly less likely to migrate than those who are unemployed or students; the interaction terms of non-skilled worker with the health index are significantly negative, suggesting that positive health effects are stronger for non-skilled workers than those who are unemployed or students in terms of promoting the propensity to migrate.

Table 4.22: The estimates of health effects by occupation using various health indices

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Type 1 index		Type 2 index				Type 3 index	
	Index1	Index2	Index3	Index4	Index5	Index6	Index7	Index8
Health index	0.142	0.087	0.103^{**}	0.090^{*}	0.120	0.067^{**}	0.285^{*}	0.342
	(0.09)	(0.08)	(0.05)	(0.05)	(0.08)	(0.03)	(0.17)	(0.28)
Occupation: None/student (Ref.)								
Farmer	0.135	0.208	0.205	0.336	0.315	0.315	0.351	0.114
	(0.11)	(0.36)	(0.14)	(0.31)	(0.42)	(0.26)	(0.36)	(0.57)
Non-farm	-0.300	-0.757	-0.183	-0.239	-0.432	-0.184	-0.608	-1.249
	(0.21)	(0.75)	(0.25)	(0.52)	(0.74)	(0.44)	(0.55)	(0.92)
Skilled	0.239	0.821	0.457^{*}	0.992^{**}	1.165^{*}	0.858^{**}	0.253	1.613^{*}
	(0.19)	(0.57)	(0.25)	(0.50)	(0.64)	(0.43)	(0.58)	(0.92)
Non-skilled	-0.089	0.147	-0.260	-0.171	0.032	-0.215	-1.355^{***}	-1.943^{**}
	(0.17)	(0.58)	(0.22)	(0.51)	(0.73)	(0.42)	(0.52)	(0.93)
Service	0.135	0.600	0.234	0.694	0.899	0.580	0.075	0.449
	(0.15)	(0.50)	(0.20)	(0.43)	(0.58)	(0.36)	(0.45)	(0.75)
Farmer *health	-0.112	-0.045	-0.080	-0.060	-0.070	-0.046	-0.172	-0.048
	(0.12)	(0.09)	(0.07)	(0.06)	(0.11)	(0.04)	(0.20)	(0.37)
Non-farm *health	0.159	0.158	0.008	0.017	0.070	0.005	0.259	0.690
	(0.22)	(0.19)	(0.11)	(0.10)	(0.18)	(0.07)	(0.32)	(0.59)
Skilled *health	-0.236	-0.205	-0.201^{*}	-0.188^{*}	-0.281^{*}	-0.134^{*}	-0.121	-1.016^{*}
	(0.20)	(0.15)	(0.11)	(0.10)	(0.16)	(0.07)	(0.33)	(0.59)
Non-skilled *health	0.122	-0.045	0.132	0.029	-0.014	0.031	0.775^{***}	1.251^{**}
	(0.18)	(0.15)	(0.10)	(0.10)	(0.18)	(0.07)	(0.29)	(0.59)
Service *health	-0.161	-0.157	-0.114	-0.138	-0.225	-0.095	-0.039	-0.286
	(0.16)	(0.13)	(0.09)	(0.09)	(0.15)	(0.06)	(0.26)	(0.48)
Obs	8782	8536	8782	8536	8536	8536	8897	8959

Notes: The indices used here are the same as those in Table 4.19; the equation also includes other controls in the baseline equation (except for the objective health measures).

Similarly, we interact current health with lagged health and included it in the estimation because based on equation (29) the response of migration probability to lagged health, which is captured by the coefficient of lagged health, depends on current health. The results are not reported here and suggest that the interactions between current health and lagged health are not significant. This result indicates that the effects of lagged health on migration seem to not significantly depend on current health.

In summary, Table 4.19 presents the results when we used three main types of health indices. Using these health indices as another approach, we found evidence for positive health effects, which indicates that there might be some health effects there but they are sensitive to the measure of health. In addition, we interact this continuous health index with occupation and lagged health, finding that positive health effects are less strong for skilled workers than for those who are unemployed or students when the self-rated health is coded as a variable that takes four values ranging from zero to three and given larger than equal weights when combined with three objective measures (mainly the Type 2 index); when we absorbed other health information in the data (Type 3 index), the positive health effects appear stronger for non-skilled workers than for those who are unemployed or students in terms of promoting migration probability. This result hints that positive health effects might be relatively stronger for non-skilled workers than skilled workers, which is consistent with the results when we used the categorical version of health variable (Table 4.6) and the theoretical model.

4.6 Conclusion

This chapter developed a theoretical model to assess the effects of health on migration. Based loosely on Jasso et al.(2004)’s model of health selectivity, we established a model in the same way as Borjas' (1987) self-selection model; the health effects derived from this selectivity model suggest that health effects vary with occupation or education and allowed us to derive the interaction between health and proxies for occupation and education. Based on this framework, we applied a probit model and found that those self-evaluating as having “fair”, “good” or “excellent” health were more likely to migrate than those self-evaluating as having “poor” health; in other words, the distinction seems to be driven by those self-evaluating as having “poor” health being less likely to migrate.

We tested the hypothesis on the interaction of health with occupation or education derived from our model, finding that the health effects tend to be larger for lower skilled workers, which is consistent with what the model predicts, although not larger for people with lower education levels. We also tested the hypothesis on the indirect effects, by which we mean the effects of earlier health on education attainment, finding that self-evaluating as having “fair”, “good” or “excellent” health between the ages of 13 and 16 has a positive effect on the highest education degree they obtained after they were 16 years old. To gain insight into the long-term effects of health, we estimated the effects of lagged health on migration, finding that the effects of lagged health on migration not to be significant. Next, we examined the effects of changes in health, but did not find evidence that improvements in health led to increased migration probability, with the fixed effects estimate and the random effects estimates also suggesting that the effects of a change in health are not significant. Interestingly, we did find that health effects estimates are sensitive to the measure of health; when we estimated the main equation using a health index created by collapsing various variables into a simple measure, we found the estimates for health effects to be sensitive to the type of variables and the weights assigned to variables in the index, and that the estimates appear more significant when the index is based on more health variables and gives more weight to the self-rated, as opposed to the “objective” measures of health.

To conclude, we found positive but relatively weak evidence on the health selectivity of migrants. We conducted various tests to investigate these health effects, and although we did not find conventionally statistically significant effects, this might be due to the substantial heterogeneity across households and circumstances, as well as the rather small sample we had and the weaknesses associated with the measures we had to use. Additionally, the variation in health might not be substantial due to the age range (16-35 years) of the sample. More importantly, it is noteworthy that when we extracted more information from the data to construct a simple continuous health index, the health effects appeared more significant, especially when the index gave more weight to the self-rated, as opposed to the “objective” measures of health. This result offers some suggestion that there might be a stronger health effect if we use more health information from the data.

Conclusions and Policy Recommendations

This thesis conducted empirical analysis on the intergenerational transmission of adiposity across countries and in China, and the relationship between health and migration decision in China.

We conducted three sets of separate and related analysis in this thesis. In the first empirical chapter, we set up an empirical model on the intergenerational transmission (of income, education or BMI). Using different datasets from around the world: China Health and Nutrition Survey (CHNS), Indonesian Family Life Survey (IFLS), British 1970 Cohort Studies (BCS1970), Health Survey for England (HSE), National Health and Nutrition Examination Survey (NHNAES), the Spanish National Health Survey (ENS-2006) and the Survey for the Evaluation of Urban Households (ENCELURB) data in Mexico, we estimate the intergenerational transmission of adiposity in these countries. We find that the elasticity of intergenerational transmission is relatively constant – at 0.2 per parent, this elasticity is comparable across time and countries, regardless of the economic development degree and the main ethnic composition of the country. To investigate the variation in this intergenerational elasticity across the BMI distribution, we conduct quantile estimation and the results suggest that this intergenerational transmission mechanism is more than double for the fattest children as it is for the thinnest children. The results indicate a large fraction of adiposity determination within the family, particularly for the fatter children. This seems to be a general pattern across different countries. Therefore, one policy implication is to put more attention on family and parents, making parents better informed or educated on healthy lifestyle and healthy dietary could be the options.

In the second chapter, using BMI z-score as another measure of adiposity, we estimate the intergenerational transmission of adiposity in China. Based on the CHNS longitudinal data from 1989 to 2009, the OLS estimates suggest one standard deviation increase in father's BMI z-score is associated with an increase of 0.20 in child's BMI z-score, and this figure is around 0.22 for the correlation between mother and child's BMI z-score. These estimates decreases to around 0.14 for father-child and 0.12 for mother-child when we control for the household fixed effects, similarly when we control for the individual fixed effects. The fixed effects estimates might provide some evidence for the short term environmental effects of parents' BMI on child's BMI. By applying quantile estimation, we find that the correlation between father and child's BMI z-score tends to be higher among fatter children, it is around 0.31 at the fattest end (90th) of child's BMI z-score distribution, and around 0,18 at the thinnest end (5th) of the distribution. To alleviate the lifecycle bias, we estimate the quantile elasticities on the sample of children aged 16~18 years old ("the approaching adults"), the pattern of the estimates are similar to the quantile estimates on the full sample, the correlation tends to be higher at the fatter end of child's BMI z-score distribution. As another dimension of the heterogeneous effects in the elasticity, this correlation is estimated by family socioeconomic level, we find this correlation does not vary substantially with family SES indicators. Additionally, the correlations by age group reveals that this intergenerational relationship increases during the first stage of the childhood and then decreases, it reaches the maximum over the period between childhood and the later adolescence.

In the third empirical chapter, using the CHNS data (1993-2009), we examine the "healthy migrant hypothesis" in the context of internal migration in China. Based on a framework set up in the same way as Borjas (1988)'s model of self-selection, we find those self-evaluating as having "fair", "good" or "excellent" health are more likely to migrate than those self-evaluating as having "poor" health. We find that the health effects tend to be larger for the lower skilled workers, which is consistent with what the model predicts, although not larger for people with lower education levels. We also test the indirect effects by which we mean the effects of earlier health on education attainment, we find self-evaluating as having "fair", "good" or "excellent" health between age 13 and 16 years has a positive effect on the highest education degree they obtained after they were 16 years old. To gain an insight into the long term effects of health, we estimate the effects of lagged health on migration, we find that the effects of lagged health on migration are not significant. In addition, the fixed effects estimate also suggest the effects of change in health are not significant. However, we find the health effects estimates are sensitive to the measure of health; when we estimate the main equation using a health index which is created by collapsing various variables into a simple measure, we find the estimates for health effects are sensitive to the type of variables and the weights assigned to variables in the index, and that the estimates appear more significant when the index is based on more health variables and gives more weights to the self-rated, as opposed to “objective” measures of health. This result offers some hints that there might be a stronger health effect if we use more health information from the data.

Bibliography

Abraido-Lanza, Ana F., Bruce P. Dohrenwend, Daisy S. Ng-Mak, and J. Blake Turner. "The Latino mortality paradox: a test of the" salmon bias" and healthy migrant hypotheses." *Ame'rican Journal of Public Health* 89, no. 10 (1999): 1543-1548.

Ahlburg, Dennis. "Intergenerational transmission of health." *American Economic Review* (1998): 265-270.

Ahn, SangNam, Matthew Lee Smith, Jinmyoung Cho, James E. Bailey, and Marcia G. Ory. "Hypertension awareness and associated factors among older Chinese adults." *Frontiers in Public Health* 1 (2013).

Akresh, Ilana Redstone, and Reanne Frank. "Health selection among new immigrants." *American Journal of Public Health* 98, no. 11 (2008): 2058.

Anderson, Patricia M. "Parental employment, family routines and childhood obesity." *Economics & Human Biology* 10, no. 4 (2012): 340-351.

Anger, Silke, and Guido Heineck. "Do smart parents raise smart children? The intergenerational transmission of cognitive abilities." *Journal of Population Economics* 23, no. 3 (2010): 1105-1132.

Battista, Marie-Claude, Marie-France Hivert, Karine Duval, and Jean-Patrice Baillargeon. "Intergenerational cycle of obesity and diabetes: how can we reduce the burdens of these conditions on the health of future generations?." *Experimental diabetes research* (2011).

Becker, Gary S., and Nigel Tomes. "Human capital and the rise and fall of families." In *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition),* pp. 257-298. The University of Chicago Press, 1994.

Behrman, Jere R. "The impact of health and nutrition on education." *The World Bank Research Observer* 11, no. 1 (1996): 23-37.

Behrman, Jere R., and Mark R. Rosenzweig. "Does increasing women's schooling raise the schooling of the next generation?." *American Economic Review* (2002): 323-334.

Bhalotra, Sonia, and Samantha Rawlings. "Gradients of the Intergenerational Transmission of Health in Developing Countries." *Review of Economics and Statistics* 95, no. 02 (2013): 660-672.

Bhattacharya, Debopam, and Bhashkar Mazumder. "A nonparametric analysis of black–white differences in intergenerational income mobility in the United States." *Quantitative Economics* 2, no. 3 (2011): 335-379.

Bell, A. Colin, Keyou Ge, and Barry M. Popkin. "The road to obesity or the path to prevention: motorized transportation and obesity in China." *Obesity Research* 10, no. 4 (2002): 277-283.

Biao, Xiang. "Migration and health in China: problems, obstacles and solutions." *Singapore: Asian Metacentre for Population and Substainable Development Analysis* (2003): 1-40.

Björklund, Anders, Jesper Roine, and Daniel Waldenström. "Intergenerational top income mobility in Sweden: Capitalist dynasties in the land of equal opportunity?." *Journal of Public Economics* 96, no. 5 (2012): 474-484.

Björklund, Anders, Mikael Lindahl, and Erik Plug. "The origins of intergenerational associations: Lessons from Swedish adoption data." *The Quarterly Journal of Economics* (2006): 999-1028.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. *Why the apple doesn't fall far: Understanding intergenerational transmission of human capital*. No. w10066. National Bureau of Economic Research, 2003.

Blumenthal, David, and William Hsiao. "Privatization and its discontents—the evolving Chinese health care system." *New England Journal of Medicine* 353, no. 11 (2005): 1165-1170.

Borjas, George J. "Self-Selection and the Earnings of Immigrants." *American Economic Review* 77, no. 4 (1987): 531-553

Bouchard Claude, ed. Genetics of obesity. Boca Raton, Fla.: CRC Press, 1994.

Bratsberg, Bernt, Knut Røed, Oddbjørn Raaum, Robin Naylor, and Tor Eriksson. "Nonlinearities in Intergenerational Earnings Mobility: Consequences for Cross-Country Comparisons*." *The Economic Journal* 117, no. 519 (2007): C72-C92.

Breierova, Lucia, and Esther Duflo. *The impact of education on fertility and child mortality: Do fathers really matter less than mothers?*. No. w10513. National Bureau of Economic Research, 2004.

Brown, Heather, and Jennifer Roberts. "Born to be wide? Exploring correlations in mother and adolescent body mass index." *Economics Letters* 120, no. 3 (2013): 413-415.

Case, Anne, Angela Fertig, and Christina Paxson. "The lasting impact of childhood health and circumstance." *Journal of health economics* 24, no. 2 (2005): 365-389.

Castles, Stephen. "Methodology and Methods: Conceptual Issues." In African Migration Research: Innovative Methods and Methodologies, ed. Berriane, Mohamed, and Hein de Haas, Africa World Press, Berriane, 2012.

Centers for Disease Control. "Body mass index: considerations for practitioners." (2011).

Cesur, Resul, and Inas Rashad Kelly. "From cradle to classroom: high birth weight and cognitive outcomes." In *Forum for Health Economics & Policy*, vol. 13, no. 2. 2010.

Chan, Kam Wing, “China, Internal Migration,” in Immanuel Ness and Peter Bellwood, eds. The Encyclopedia of Global Migration, Blackwell Publishing, 2013.

Chen, Feinian. "Family division in China's transitional economy." *Population studies* 63, no. 1 (2009): 53-69.

Chen, Jiajian, Russell Wilkins, and Edward Ng. "Health expectancy by immigrant status, 1986 and 1991." *Health Reports-Statistics Canada* 8 (1996): 29-38.

Cheng, Zhiming, Fei Guo, Graeme Hugo, and Xin Yuan. "Employment and wage discrimination in the Chinese cities: A comparative study of migrants and locals." *Habitat International* 39 (2013): 246-255.

Classen, Timothy J. "Measures of the intergenerational transmission of body mass index between mothers and their children in the United States, 1981–2004." *Economics & Human Biology* 8, no. 1 (2010): 30-43.

Cole, Tim J., Katherine M. Flegal, Dasha Nicholls, and Alan A. Jackson. "Body mass index cut offs to define thinness in children and adolescents: international survey." *Bmj* 335, no. 7612 (2007): 194.

Coneus, Katja, and C. Katharina Spiess. "The intergenerational transmission of health in early childhood—Evidence from the German Socio-Economic Panel Study." *Economics & Human Biology* 10.1 (2012): 89-97.

Corak, Miles, and Andrew Heisz. "The intergenerational earnings and income mobility of Canadian men: Evidence from longitudinal income tax data." *Journal of Human Resources* (1999): 504-533.

Currie, Janet, and Enrico Moretti. "Biology as destiny? Short and long-run determinants of intergenerational transmission of birth weight." No. w11567. National Bureau of Economic Research, 2005.

Dasgupta, Partha. "An inquiry into well-being and destitution." *OUP Catalogue*(1995).

David H. Autor. "Self-Selection - The Roy Model." Lecture, MIT 14.661, November 14, 2003

Dearden, Lorraine, Stephen Machin, and Howard Reed. "Intergenerational mobility in Britain." *The Economic Journal* (1997): 47-66.

De Brauw, Alan, and John Giles. "Migrant opportunity and the educational attainment of youth in rural China." *World Bank Policy Research Working Paper Series, Vol* (2008).

Deolalikar, Anil B. "Nutrition and Labor Producvivity in Agriculture: Estimates for Rural South India." *The review of Economics and Statistics* (1988): 406-413.

Currie, Janet, and Enrico Moretti. *Biology as destiny? Short and long-run determinants of intergenerational transmission of birth weight*. No. w11567. National Bureau of Economic Research, 2005.

Doak, Colleen M., Linda S. Adair, Margaret Bentley, Carlos Monteiro, and Barry M. Popkin. "The dual burden household and the nutrition transition paradox."*International journal of obesity* 29, no. 1 (2004): 129-136.

Du, Shufa, Bing Lu, Fengying Zhai, and Barry M. Popkin. "A new stage of the nutrition transition in China." *Public health nutrition* 5, no. 1a (2002): 169-174.

Du, Yang, Albert Park, and Sangui Wang. "Migration and rural poverty in China." Journal of comparative economics 33, no. 4 (2005): 688-709.

Eriksson, Tor, Jay Pan, and Xuezheng Qin. “The Intergenerational Inequality of Health in China.” *China Economic Review* 31, no.0 (2014): 392–409.

Fielding, A. J. *Migration and social mobility in urban systems: national and international trends*. Edward Elgar: Cheltenham, 2007.

Findley, Sally E. "The directionality and age selectivity of the health-migration relation: Evidence from sequences of disability and mobility in the United States." *International Migration Review* (1988): 4-29.

Floud, Roderick, Robert W. Fogel, Bernard Harris, and Sok Chul Hong. *The changing body: Health, nutrition, and human development in the western world since 1700*. Cambridge University Press, 2011.

Frankenberg, Elizabeth, and Nathan R. Jones. "Self-rated health and mortality: does the relationship extend to a low income setting?." *Journal of Health and Social Behavior* 45, no. 4 (2004): 441-452.

Frisbie, W. Parker, Youngtae Cho, and Robert A. Hummer. "Immigration and the health of Asian and Pacific Islander adults in the United States." *American Journal of Epidemiology* 153, no. 4 (2001): 372-380.

Fukagawa, NAOMI K., LINDA G. Bandini, and JAMES B. Young. "Effect of age on body composition and resting metabolic rate." *Am J Physiol* 259, no. 2 Pt 1 (1990): E233-E238. University Press, 2011.

Gagnon, Jason, Theodora Xenogiani, and Chunbing Xing. "Are all migrants really worse off in urban labour markets: new empirical evidence from China." (2009).

Galton, Francis. *Hereditary genius*. Macmillan and Company, 1869.

Gan, Li, and Guan Gong. *Estimating interdependence between health and education in a dynamic model*. No. w12830. National Bureau of Economic Research, 2007.

Golan, Moria. "Parents as agents of change in childhood obesity-from research to practice." *International Journal of Pediatric Obesity* 1, no. 2 (2006): 66-76.

Goldberger, Arthur S. "Economic and mechanical models of intergenerational transmission." *The American Economic Review* (1989): 504-513.

Gong, Honge, Andrew Leigh, and Xin Meng. "Intergenerational income mobility in urban China." *Review of Income and Wealth* 58, no. 3 (2012): 481-503.

Gorber, S. Connor, M. Tremblay, David Moher, and B. Gorber. "A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review." *Obesity reviews* 8, no. 4 (2007): 307-326.

Graham, Hilary, and Chris Power. "Childhood disadvantage and health inequalities: a framework for policy based on lifecourse research." *Child: care, health and development* 30, no. 6 (2004): 671-678.

Grawe, Nathan D. "Intergenerational mobility for whom? The experience of high-and low-earning sons in international perspective." *Generational income mobility in North America and Europe* (2004): 58-89.

Grossman, Michael. "The human capital model." *Handbook of health economics* 1 (2000): 347-408.

Gruber, Jonathan, ed. *The Problems of Disadvantaged Youth: An Economic Perspective*. University of Chicago Press, 2009.

Guo, Xuguang, Thomas A. Mroz, Barry M. Popkin, and Fengying Zhai. "Structural change in the impact of income on food consumption in China, 1989–1993." *Economic Development and Cultural Change* 48, no. 4 (2000): 737-760.

Han, Song, and Casey B. Mulligan. "Human capital, heterogeneity and estimated degrees of intergenerational mobility." *The Economic Journal* 111, no. 470 (2001): 207-243.

Heckman, James, and Pedro Carneiro. *Human capital policy*. No. w9495. National Bureau of Economic Research, 2003.

Hu, Xiaojiang, Sarah Cook, and Miguel A. Salazar. "Internal migration and health in China." *The Lancet* 372, no. 9651 (2008): 1717-1719.

Hummer, Robert A. "Adult mortality differentials among Hispanic subgroups and non-Hispanic whites." *Social Science Quarterly* 81, no. 1 (1999): 459-476.

Hummer, Robert A., Daniel A. Powers, Starling G. Pullum, Ginger L. Gossman, and W. Parker Frisbie. "Paradox found (again): infant mortality among the Mexican-origin population in the United States." *Demography* 44, no. 3 (2007): 441-457.

Jacobson, Peter, Jarl S. Torgerson, Lars Sjöström, and Claude Bouchard. "Spouse resemblance in body mass index: effects on adult obesity prevalence in the offspring generation." *American journal of epidemiology* 165, no. 1 (2007): 101-108.

Jasso, Guillermina, Douglas S. Massey, Mark R. Rosenzweig, and James P. Smith. "Immigrant health: selectivity and acculturation." *Critical perspectives on racial and ethnic differences in health in late life* (2004): 227-266.

Jääskeläinen, A., J. Pussinen, O. Nuutinen, U. Schwab, J. Pirkola, M. Kolehmainen, M. R. Järvelin, and J. Laitinen. "Intergenerational transmission of overweight among Finnish adolescents and their parents: a 16-year follow-up study." *International Journal of Obesity* 35, no. 10 (2011): 1289-1294.

Johnson, Robert J., and Fredric D. Wolinsky. "The structure of health status among older adults: disease, disability, functional limitation, and perceived health." *Journal of health and social behavior* (1993): 105-121.

Kalmijn, Matthijs. "Assortative mating by cultural and economic occupational status." *American Journal of Sociology* (1994): 422-452.

Klein, Lawrence R., and Süleyman Özmucur. "The estimation of China's economic growth rate." *Journal of Economic and Social Measurement* 28, no. 4 (2003): 187-202.

Knight, John, and Lina Song. "The rural-urban divide: economic disparities and interactions in China." *OUP Catalogue* (1999).

Laitinen, Jaana, Chris Power, and Marjo-Riitta Järvelin. "Family social class, maternal body mass index, childhood body mass index, and age at menarche as predictors of adult obesity." *The American journal of clinical nutrition* 74, no. 3 (2001): 287-294.

Lei, Xiaoyan, and Wanchuan Lin. "The new cooperative medical scheme in rural China: Does more coverage mean more service and better health?."*Health Economics* 18, no. S2 (2009): S25-S46.

Li, Haizheng, Zahniser, Steven. "The determinants of China's temporary rural–urban migration." *Urban Studies* 39, no. 12 (2002): 2219–2235.

Li, Leah, Catherine Law, Rossella Lo Conte, and Chris Power. "Intergenerational influences on childhood body mass index: the effect of parental body mass index trajectories." *The American journal of clinical nutrition* 89, no. 2 (2009): 551-557.

Loureiro, Maria L., Anna Sanz-de-Galdeano, and Daniela Vuri. "Smoking Habits: Like Father, Like Son, Like Mother, Like Daughter?*." *Oxford Bulletin of Economics and Statistics* 72, no. 6 (2010): 717-743.

Lu, Yao. "Test of the 'healthy migrant hypothesis': a longitudinal analysis of health selectivity of internal migration in Indonesia." *Social science & medicine*67, no. 8 (2008): 1331-1339.

Maes, Hermine HM, Michael C. Neale, and Lindon J. Eaves. "Genetic and environmental factors in relative body weight and human adiposity." *Behavior genetics* 27, no. 4 (1997): 325-351.

Mare, Robert D. "Five decades of educational assortative mating." *American Sociological Review* (1991): 15-32.

Manor, Orly, Sharon Matthews, and Chris Power. "Health selection: the role of inter-and intra-generational mobility on social inequalities in health." *Social science & medicine* 57, no. 11 (2003): 2217-2227.

Marmot, Michael G., Abraham M. Adelstein, and Lak Bulusu. "Lessons from the study of immigrant mortality." *The Lancet* 323, no. 8392 (1984): 1455-1457.

Martin, Molly A. "The intergenerational correlation in weight: how genetic resemblance reveals the social role of families." *AJS; American journal of sociology* 114, no. Suppl (2008): S67.

Meng, Xin. *Labour market reform in China*. Cambridge University Press, 2000.

Meng, Xin, and Junsen Zhang. "The two-tier labor market in urban China: occupational segregation and wage differentials between urban residents and rural migrants in Shanghai." *Journal of comparative Economics* 29, no. 3 (2001): 485-504.

Morris, Stephen. "Body mass index and occupational attainment." *Journal of health economics* 25, no. 2 (2006): 347-364.

Morgan, Stephen L. "Richer and taller: stature and living standards in China, 1979-1995." *The China Journal* (2000): 1-39.

Mulligan, Casey B. "Galton versus the human capital approach to inheritance." *Journal of political Economy* 107, no. S6 (1999): S184-S224.

Nestle, Marion. "Food marketing and childhood obesity—a matter of policy."*New England Journal of Medicine* 354, no. 24 (2006): 2527-2529.

Ng, Marie, Tom Fleming, Margaret Robinson, Blake Thomson, Nicholas Graetz, Christopher Margono, Erin C. Mullany et al. "Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013." *The Lancet* (2014).

Onis, Mercedes de, Adelheid W. Onyango, Elaine Borghi, Amani Siyam, Chizuru Nishida, and Jonathan Siekmann. "Development of a WHO growth reference for school-aged children and adolescents." *Bulletin of the World Health Organization* 85, no. 9 (2007): 660-667.

Palloni, Alberto, and Jeffrey D. Morenoff. "Interpreting the paradoxical in the Hispanic paradox." *Annals of the New York Academy of Sciences* 954, no. 1 (2001): 140-174.

Peck, AM Nyström. "Childhood environment, intergenerational mobility, and adult health--evidence from Swedish data." *Journal of Epidemiology and Community Health* 46, no. 1 (1992): 71-74.

Pekkarinen, Tuomas, Roope Uusitalo, and Sari Kerr. "School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform." *Journal of Public Economics* 93, no. 7 (2009): 965-973.

Pollak, Robert A. *Bargaining power in marriage: Earnings, wage rates and household production*. No. w11239. National Bureau of Economic Research, 2005.

Popkin, Barry M. "The nutrition transition and obesity in the developing world."*The Journal of nutrition* 131, no. 3 (2001): 871S-873S.

Popkin, Barry M. "The nutrition transition: an overview of world patterns of change." *Nutrition reviews* 62, no. s2 (2004): S140-S143.

Popkin, Barry M., and Penny Gordon-Larsen. "The nutrition transition: worldwide obesity dynamics and their determinants." *International journal of obesity* 28 (2004): S2-S9.

Popkin, Barry M., Shufa Du, Fengying Zhai, and Bing Zhang. "Cohort Profile: The China Health and Nutrition Survey—monitoring and understanding socio-economic and health change in China, 1989–2011." *International journal of epidemiology* 39, no. 6 (2010): 1435-1440.

Piraino, Patrizio. "Comparable estimates of intergenerational income mobility in Italy." *The BE Journal of Economic Analysis & Policy* 7, no. 2 (2007).

Pitt, Mark M., Mark R. Rosenzweig, and Md Nazmul Hassan. "Productivity, health, and inequality in the intrahousehold distribution of food in low-income countries." *The American Economic Review* (1990): 1139-1156.

Qian, Nancy. "Missing women and the price of tea in China: The effect of sex-specific earnings on sex imbalance." *The Quarterly Journal of Economics* 123, no. 3 (2008): 1251-1285.

Raymond, Susan U., Stephen Leeder, and Henry M. Greenberg. "Obesity and cardiovascular disease in developing countries: a growing problem and an economic threat." *Current Opinion in Clinical Nutrition & Metabolic Care* 9, no. 2 (2006): 111-116.

Rosenzweig, Mark R., and T. Paul Schultz. "The stability of household production technology: A replication." *Journal of Human Resources* (1988): 535-549.

Roy, Andrew Donald. "Some thoughts on the distribution of earnings." *Oxford economic papers* 3, no. 2 (1951): 135-146.

Royer, Heather. "Separated at girth: US twin estimates of the effects of birth weight." *American Economic Journal: Applied Economics* 1, no. 1 (2009): 49-85.

Rozelle, Scott, J. Edward Taylor, and Alan DeBrauw. "Migration, remittances, and agricultural productivity in China." *American Economic Review* (1999): 287-291.

Rubalcava, Luis N., Graciela M. Teruel, Duncan Thomas, and Noreen Goldman. "The healthy migrant effect: new findings from the Mexican Family Life Survey." *Journal Information* 98, no. 1 (2008).

Schultz, T. Paul. "Human resources in China: The birth quota, returns to schooling, and migration." *Pacific Economic Review* 9, no. 3 (2004): 245-267.

Sen, Amartya. "More than 100 million women are missing." *The New York Review of Books* (1990).

Shaheen, Seif O., Jonathan AC Sterne, Scott M. Montgomery, and Hossain Azima. "Birth weight, body mass index and asthma in young adults." *Thorax*54, no. 5 (1999): 396-402.

Shi, Li. *Rural migrant workers in China: scenario, challenges and public policy*. Geneva: ILO, 2008.

Sjaastad, Larry A. "The costs and returns of human migration." *The journal of political economy* (1962): 80-93.

Smith, George Davey, Colin Steer, Sam Leary, and Andy Ness. "Is there an intrauterine influence on obesity? Evidence from parent–child associations in the Avon Longitudinal Study of Parents and Children (ALSPAC)." *Archives of disease in childhood* 92, no. 10 (2007): 876-880.

Smith, James P. "The impact of childhood health on adult labor market outcomes." *The review of economics and statistics* 91, no. 3 (2009): 478-489.

Solon, Gary. "A model of intergenerational mobility variation over time and place." *Generational income mobility in North America and Europe* (2004): 38-47.

Spencer, Elizabeth A., Paul N. Appleby, Gwyneth K. Davey, and Timothy J. Key. "Validity of self-reported height and weight in 4808 EPIC–Oxford participants." *Public health nutrition* 5, no. 04 (2002): 561-565.

Tang, Shenglan, Qingyue Meng, Lincoln Chen, Henk Bekedam, Tim Evans, and Margaret Whitehead. "Tackling the challenges to health equity in China."*The Lancet* 372, no. 9648 (2008): 1493-1501.

Taylor, J. Edward, Scott Rozelle, and Alan De Brauw. "Migration and incomes in source communities: A new economics of migration perspective from China*." *Economic Development and Cultural Change* 52, no. 1 (2003): 75-101.

Thailand Econometric Society. International Conference, Van-Nam Huynh, Vladik Kreinovich, and Songsak Sriboonchitta. *Modeling Dependence in Econometrics*. Springer, 2014.

Thomas, Duncan."Intra-household resource allocation: An inferential approach." *Journal of human resources* (1990): 635-664.

Thompson, O. "The Intergenerational Transmission of Health Status: Estimates and Mechanisms." (Ph.D diss., University of Wisconsin, 2013).

Tong, Yuying, and Martin Piotrowski. "Migration and health selectivity in the context of internal migration in China, 1997–2009." *Population Research and Policy Review* 31, no. 4 (2012): 497-543.

Trannoy, Alain, Sandy Tubeuf, Florence Jusot, and Marion Devaux. "Inequality of opportunities in health in France: a first pass." *Health economics* 19, no. 8 (2010): 921-938.

Van Leeuwen, Marieke, Stéphanie M. Van Den Berg, and Dorret I. Boomsma. "A twin-family study of general IQ." *Learning and Individual Differences* 18, no. 1 (2008): 76-88.

von Hinke Kessler Scholder, Stephanie, George Davey Smith, Debbie A. Lawlor, Carol Propper, and Frank Windmeijer. "The effect of fat mass on educational attainment: Examining the sensitivity to different identification strategies." *Economics & Human Biology* 10, no. 4 (2012): 405-418.

Wang, Youfa and Hsin-Jen Chen. "Use of Percentiles and Z-Scores in Anthropometry." In *Handbook of Anthropometry: Physical Measures of Human Form in Health and Disease*, ed.Victor R. Preedy, pp. 29-48. Springer Science+Business Media, LLC, New York, 2012.

WHO, Expert Consultation. "Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies." *Lancet* 363, no. 9403 (2004): 157.

World Health Organization. "The World health report: 2002: Reducing the risks, promoting healthy life." (2002).

Wu, Harry X., and Li Zhou. "Rural-to-Urban Migration in China*." *Asian‐Pacific Economic Literature* 10, no. 2 (1996): 54-67.

Wu, Yangfeng. "Overweight and obesity in China." *Bmj* 333, no. 7564 (2006): 362-363.

Wu, Zheren. "Self-selection and Earnings of Migrants: Evidence from Rural China." *Asian Economic Journal* 24, no. 1 (2010): 23-44.

Xingzhu, Liu, and Cao Huaijie. "China's cooperative medical system: Its historical transformations and the trend of development." *Journal of public health policy* (1992): 501-511.

Yang, Xiushi, and Fei Guo. "Gender differences in determinants of temporary labor migration in China: A multilevel analysis." *International Migration Review*(1999): 929-953.

Zhang, B., F. Y. Zhai, S. F. Du, and B. M. Popkin. "The China Health and Nutrition Survey, 1989–2011." *Obesity Reviews* 15, no. S1 (2014): 2-7.

Appendices

Appendix 2

Discussion of the interaction between father and mother’s BMI

In our model, the interaction term between father’s BMI and mother’s BMI is also included in the transmission regression due to the potential interactive “reinforcing” effects between parents and their children. The argument here is that there is a potential assortative mating (Mare 1991, Kalmijn 1994) between father and mother, the subsequent sharing of a household environment and the nutrition regime may enforce the effects of father and mother’s similar BMI status. For instance, having an overweight father and an overweight mother may generate an interaction effect which is greater than the sum of these two terms together. On the other hand, if the father is overweight whereas the mother is normally weighted or underweighted, the interaction effects may depend on the role of them in this family (such as who is in charge of the food preparation or allocation) and their bargaining power within the household (Pollak 2005) .

However, in the empirical analysis we find that this interaction term is omitted when it is incorporated into the model, this indicates that there might not an independent role for the interaction effect to play in the intergenerational transmission, in other words, our hypothesis that the potential “ assortative mating” of father and mother has a “ reinforcing” effect on the BMI development of their child is not supported by our data.

Table A 2.1: The age of child by country

	Variable	Obs	Mean	Std. Dev.	Min	Max
British	Age of child	44899	14.91	4.83	1	19
China	Age of child	14081	9.085	4.78	0	17
England	Age of child	26476	9.84	4.56	2	17
Indonesia	Age of child	18650	7.17	4.23	0	14
Spain	Age of child	11114	8.05	4.81	0	16.05
US	Age of child	6581	7.24	4.31	2	16

Table A2.2: Intergenerational BMI elasticity by country on children aged above five

	China	Indonesia	UK	US	China
	CHNS (1989-2009)	IFLS (1993-2007)	BCS (1970-1996)	HSE (1995-2010)	CHNS (1989-2009)
Dependent variable: Log (BMI of child)					
Log (BMI of father)	0.241***	0.161***	0.185***	0.186***	0.195***
	(0.0127)	(0.0094)	(0.0097)	(0.0077)	(0.0175)
Log(BMI of mother)	0.190***	0.147***	0.174***	0.196***	0.202***
	(0.0122)	(0.0078)	(0.0080)	(0.0065)	(0.0138)
Age of Child	-0.006***	-0.0306***	0.0472***	0.0284***	0.0390***
	(0.002)	(0.0029)	(0.0035)	(0.0019)	(0.0045)
(Age of Child)^2	0.0015***	0.0031***	-0.0005***	0.0001*	-0.0001
	(8.87e-05)	(0.0002)	(0.000122)	(8.59e-05)	(0.0002)
Male Child	0.0503***	0.0641***	-0.0741***	0.0093	0.0193
	(0.0076)	(0.007)	(0.0056)	(0.0060)	(0.0125)
Male*Age of Child	-0.0038***	-0.0077***	0.0052***	-0.0031***	-0.0042***
	(0.00064)	(0.0008)	(0.0004)	(0.0005)	(0.0014)
Constant	1.363***	1.792***	1.267***	1.367***	1.270***
	(0.0517)	(0.0359)	(0.0433)	(0.0321)	(0.0665)
Observations	11,082	12,884	19,594	22,103	4,207
R-squared	0.403	0.293	0.568	0.439	0.423

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A2.3: Intergenerational BMI elasticity by country on children aged between 5 and 16 years old

	China	Indonesia	UK		US
	CHNS (1989-2009)	IFLS (1993-2007)	BCS (1970-1996)	HSE (1995-2010)	CHNS (1989-2009)
Dependent variable: Log (BMI of child)					
Log (BMI of father)	0.246***	0.161***	0.173***	0.184***	0.195***
	(0.0130)	(0.00939)	(0.00957)	(0.00793)	(0.0175)
Log(BMI of mother)	0.195***	0.147***	0.161***	0.193***	0.202***
	(0.0126)	(0.00781)	(0.00813)	(0.00666)	(0.0138)
Age of Child	-0.0142***	-0.0306***		0.0221***	0.0390***
	(0.00226)	(0.00292)		(0.00209)	(0.00448)
(Age of Child)^2	0.00190***	0.00306***	0.00149***	0.000517***	-0.000129
	(0.000106)	(0.000154)	(1.78e-05)	(9.99e-05)	(0.000224)
Male Child	0.0520***	0.0641***	0.000229	0.0210***	0.0193
	(0.00785)	(0.00704)	(0.00732)	(0.00625)	(0.0125)
Male*Age of Child	-0.00394***	-0.00768***	-0.00171***	-0.00438***	-0.00418***
	(0.000687)	(0.000757)	(0.000645)	(0.000604)	(0.00137)
Constant	1.368***	1.792***	1.623***	1.407***	1.270***
	(0.0532)	(0.0359)	(0.0369)	(0.0333)	(0.0665)
Observations	10,474	12,884	15,658	20,431	4,207
R-squared	0.378	0.293	0.450	0.421	0.423

Note: the variable for “age of child” is omitted due to collinearity. Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table A 2.4: Individual fixed effects on Indonesian data

	Indonesia IFLS (1993-2007)
Dependent variable: Log (BMI of child)	
Log (BMI of father)	0.107*** (0.0309)
Log(BMI of mother)	0.130*** (0.0266)
Age of Child	-0.0413*** (0.00176)
(Age of Child)^2	0.00335*** (7.64e-05)
Male Child	0.0986*** (0.0321)
Male*Age of Child	-0.00562*** (0.000669)
Age of Father	0.00311*** (0.00105)
Age of Mother	0.00168 (0.00120)
Constant	1.860*** (0.117)
Observations	18,570
Number of pid	14,347
R-squared	0.429

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A 2.5: Assortative mating: the association between father and mother's BMI

	China	Indonesia	UK		US
	CHNS	IFLS	BCS	HSE	NHANES 3
	(1989-2009)	(1993-2007)	(1970-1996)	(1995-2010)	(1988-1994)
Dependent variable: Log (BMI of father)					
Log (BMI of mother)	0.199*** (0.0112)	0.223*** (0.00744)	0.132*** (0.00804)	0.138*** (0.00660)	0.166*** (0.0100)
Constant	2.478*** (0.0346)	2.378*** (0.0231)	2.776*** (0.0252)	2.841*** (0.0214)	2.727*** (0.0322)
Observations	14,081	18,650	37,197	26,476	6,581
R-squared	0.044	0.065	0.027	0.030	0.048

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A 2.6: Intergenerational BMI elasticity by country with interaction terms of obese or underweight parents

	China	Indonesia	UK		US
	CHNS	IFLS	BCS	HSE	NHANES 3
	(1989-2009)	(1993-2007)	(1970-1996)	(1995-2010)	(1988-1994)
Dependent variable: Log (BMI of child)					
Log (BMI of father)	0.212***	0.133***	0.182***	0.157***	0.139***
	(0.0121)	(0.0082)	(0.0094)	(0.0071)	(0.0131)
Log (BMI of mother)	0.177***	0.130***	0.164***	0.173***	0.146***
	(0.0113)	(0.0069)	(0.00778)	(0.0060)	(0.0104)
Obese father* mother	-0.0049	0.0147	-0.0280*	0.0121**	0.0194*
	(0.101)	(0.0193)	(0.0158)	(0.0054)	(0.0103)
Underweight father*mother	0.0017	0.00753**	-0.0018	-0.0133	0.0120
	(0.0049)	(0.0036)	(0.0098)	(0.0199)	(0.018)
Age of Child	-0.0346***	-0.0347***	-0.0242***	-0.0043***	-0.00415**
	(0.0010)	(0.0009)	(0.0006)	(0.0010)	(0.0020)
(Age of Child)^2	0.00267***	0.00318***	0.0021***	0.0015***	0.0019***
	(4.96e-05)	(5.94e-05)	(2.37e-05)	(5.07e-05)	(0.0001)
Male Child	0.0271***	0.0422***	-0.0333***	0.0165***	0.0195***
	(0.0048)	(0.0038)	(0.0040)	(0.0036)	(0.0055)
Male*Age of Child	-0.0019***	-0.0055***	0.0026***	-0.0036***	-0.0041***
	(0.0005)	(0.0005)	(0.0003)	(0.0004)	(0.0008)
Constant	1.651***	1.958***	1.765***	1.710***	1.837***
	(0.0501)	(0.0321)	(0.0369)	(0.0308)	(0.0521)
Observations	14,081	18,650	21,253	26,476	6,581
R-squared	0.355	0.225	0.552	0.452	0.449

Robust standard errors in parentheses,*** p<0.01, ** p<0.05, * p<0.1

Table A 2.7: Intergenerational BMI elasticity for parents and child on pooled data, Indonesian as the reference group

	(1)	(2)	(3)	(4)	(5)
Dependent variable: Log(BMI of child)					
Log (BMI of	0.161***	0.200***		0.166***	0.161***
father)	(0.0040)	(0.0041)		(0.00451)	(0.0042)
Log (BMI of	0.163***		0.189***	0.165***	0.162***
mother)	(0.0034)		(0.0034)	(0.0039)	(0.0036)
Obese					0.0129***
father*mother					(0.0045)
Underweight					0.0125***
father*mother					(0.0027)
Age of Child	-0.0202***	-0.0193***	-0.0199***	-0.020***	-0.020***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
(Age of	0.0021***	0.0021***	0.0021***	0.0021***	0.0021***
Child)2	(1.76e-05)	(1.78e-05)	(1.74e-05)	(1.73e-05)	(1.76e-05)
Male Child	0.0132***	0.0127***	0.0134***	-0.0002	0.0132***
	(0.0019)	(0.0019)	(0.0019)	(0.0021)	(0.0020)
Male*Age of	-0.0018***	-0.0018***	-0.0018***	-0.0004*	-
					0.0018***
Child	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
China	0.0435***	0.0372***	0.0471***	0.0446***	0.0440***
	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
British	0.0440***	0.0403***	0.0618***	0.0477***	0.0452***
	(0.0016)	(0.0016)	(0.0015)	(0.0016)	(0.0016)
England	0.0706***	0.0824***	0.102***	0.0707***	0.0712***
	(0.0017)	(0.0017)	(0.0015)	(0.0017)	(0.0017)
US	0.0668***	0.0778***	0.0947***	0.0657***	0.0674***
	(0.0022)	(0.0023)	(0.0022)	(0.0023)	(0.0022)
Constant	1.741***	2.126***	2.154***	1.726***	1.743***
	(0.0151)	(0.0127)	(0.0108)	(0.0170)	(0.0167)
Observations	87,041	87,300	88,445	87,041	87,041
R-squared	0.499	0.479	0.487	0.525	0.499

Notes: Robust standard errors in parentheses, column (4) results are sample weighted.The dummy Obese_father*mother=1 if the BMI of father and mother are above 30, Under_father*mother=1 if the BMI of father and mother are below 20, *** p<0.01, ** p<0.05, * p<0.1

Table A 2.8: Intergenerational BMI elasticity on sample with approaching adult children (age>16), Indonesian as the reference group

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: BMI of child						
BMI of father	0.189***	0.246***		0.195***	0.234***	
	(0.0102)	(0.0102)		(0.0108)	(0.0110)	
BMI of mother	0.163***		0.206***	0.185***		0.210***
	(0.0088)		(0.0088)	(0.0090)		(0.0090)
Male child	-0.39***	-0.37***	-0.38***	-0.28***	-0.28***	-0.29***
	(0.0649)	(0.0658)	(0.0650)	(0.0641)	(0.0654)	(0.0642)
British				1.866***	1.850***	2.209***
				(0.0760)	(0.0767)	(0.0749)
England				0.876***	1.388***	1.696***
				(0.104)	(0.105)	(0.0983)
US				1.368***	2.030***	2.116***
				(0.274)	(0.283)	(0.285)
Constant	13.58***	16.11***	17.29***	11.46***	14.82***	15.27***
	(0.288)	(0.251)	(0.207)	(0.302)	(0.256)	(0.217)
Observations	13,881	13,967	14,409	13,881	13,967	14,409
R-squared	0.099	0.063	0.064	0.128	0.085	0.094

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A 2.9: OLS estimates of the intergenerational BMI elasticity, controlling for age dummies and the interactions between age and gender

	China	Indonesia	US		CHNS
	CHNS	IFLS	BCS	HSE	
	(1989-2009)	(1993-2007)	(1970-1996)	(1995-2010)	(1989-2009)
Dependent variable: log (BMI of child)					
Log (BMI of father)	0.212***	0.129***	0.177***	0.166***	0.150***
	-0.0115	-0.0077	-0.0092	-0.0067	-0.0123
Log (BMI of mother)	0.173***	0.126***	0.163***	0.178***	0.152***
	-0.0108	-0.0064	-0.0076	-0.0057	-0.0096
Constant	1.689***	1.960***	1.747***	1.699***	1.817***
	-0.046	-0.0281	-0.0351	-0.0274	-0.045
Observations	14,081	18,650	21,253	26,476	6,581
R-squared	0.37	0.231	0.566	0.468	0.469

Notes: The regression also includes the child age dummies and their interactions with gender. Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Figure A 2.1: BMI of child and the elasticity of father’s BMI with child’s BMI across countries

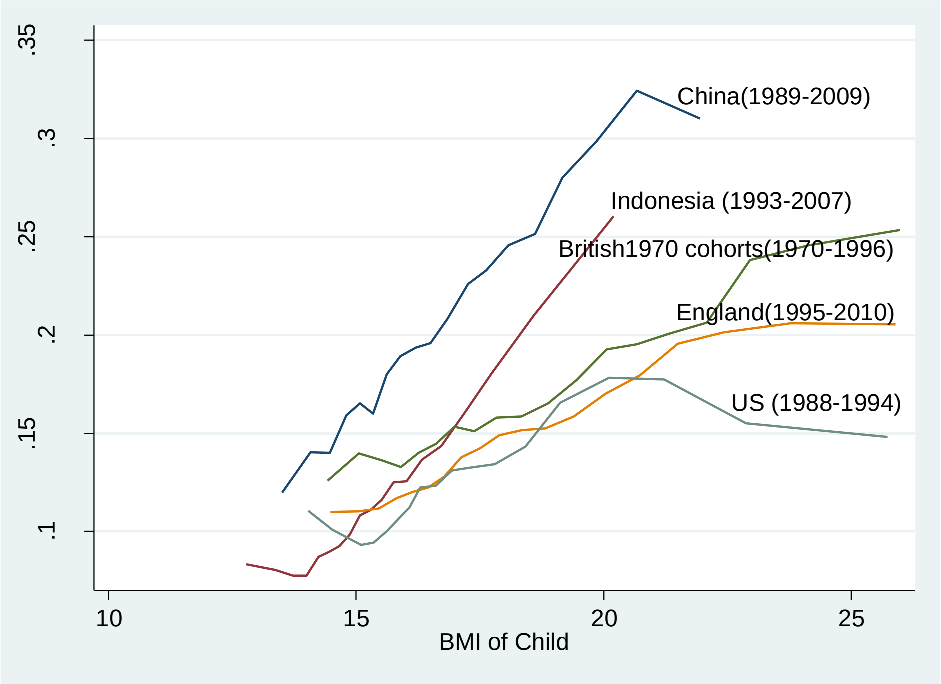


Figure A 2.1 shows how the IBE varies across child’s BMI within each country. However, the relative position of the same child may vary with country, for instance, an obese child in Indonesian data might not be seen as obese in the US data. Therefore, now we pool these data together, calculate the quantiles of child’s BMI distribution in these countries, and then obtain the mean of child’s BMI in each quantile by country. Next we plot the mean of child’s BMI in each quantile (of child’s BMI distribution in these countries) by country against the corresponding elasticity estimates, in doing so we are able to see how this elasticity varies with the BMI levels across countries. The results are presented in Figure A 2.1, it suggests that the elasticity of father’s BMI with child’s BMI in developing countries (China and Indonesia) seems to vary more with BMI levels than that in developed countries (US and UK).

The Description of the Data[55]

Indonesian Family Life Survey (IFLS)

The Indonesian Family Life Survey (IFLS) is an on-going longitudinal survey data which started in 1993. The sample used here is drawn from 1993, 2000 and 2007 waves of the survey, it is representative of 83% of the Indonesian population and contains over 30,000 individuals living in 13 of the 27 provinces in Indonesia. This survey includes a range of health measures for both parents and children. It is noteworthy that as in CHNS data, the anthropometric outcome in IFLS survey was also measured by trained nurses rather than self-reported. Additionally, the IFLS data also includes information on socioeconomic factors such as education and income. Thus, the IFLS data is similar to CHNS data in terms of the survey design and measure methods, this similarity improves the comparability of results based on these two datasets. The sample is restricted to those aged from 0 to 14 years old in each wave and have both parents and household’s information. It is noteworthy that this is different from the CHNS data, where the child sample comprises those aged between 0 and 18 years old.

In addition, in the Indonesian Family Life Survey (IFLS), we also consider step/adopted children as the sample. The adopted or step children account for around 1% of the whole sample in each wave, for these children, the information on their parents use the step parents’ rather than biological parents’.

British Cohort Study 1970 (BCS)

The 1970 British Cohort Study is an ongoing follow up study of 17,200 babies born in England, Scotland, Wales and Northern Ireland between 5 and 11 April 1970 who are still living in Britain (excluding Northern Ireland). The survey was conducted when the cohorts at birth, aged 5 (in 1975),10 (in 1980), 16 (in 1986),26 (in 1996), 30 (in 1999-2000),34 (in 2004-2005) and 38 (in 2008-2009). The samples at the age 5 and 10 were augmented since immigrants born in the same week were added in. In this paper we use the cohorts in the first five waves (sweeps).

[55] The description of the CHNS data is presented in Chapter 3.

At the birth, the questionnaires were completed by midwife and the supplementary information was collected from clinical records. As the cohorts got older, the approach of survey changed, parents were interviewed by the health stuff and questionnaires were completed by teachers. In terms of the anthropometric information, the height and weight were measured at the age of 10 and self-reported at the age of 26 (Shaheen et al. 1999)

Health Survey for England (HSE)

The Health Survey for England is designated to be nationally representative of people of different age, gender, geographic region and socio-demographic circumstances [56]. It was started in 1991 and has been conducted annually since then. The survey combines questionnaire-based answers with physical measurements and the analysis of blood sample. Each year's survey has a particular focus on a disease or condition or population group, but height, weight and general health are covered each year. An interview with household members is followed by a nurse visit. Thus, there are both self-reported and medically-measured height and weight in this data. In the computation of BMI, we use "htval" and "wtval" in the survey which are referred to as the "valid" height and weight.

The National Health and Nutrition Examination Survey III (NHANES) (US)

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. Four surveys of this type have been conducted since 1970:

1. 1971-75—National Health and Nutrition Examination Survey I (NHANES I);
2. 1976-80—National Health and Nutrition Examination Survey II (NHANES II);
3. 1982-84—Hispanic Health and Nutrition Examination Survey (HHANES);

[56] "The 1991 and 1992 surveys had a limited population sample of about 3,000 and 4,000 adults respectively. For 1993 to 1996 adult sample was boosted to about 16,000 to enable analysis by socio-economic characteristics and health regions. In 1995 for the first time a sample of about 4,000 children was also introduced. In the 1997 Health Survey the sample was about 7,000 children and 9,000 adults. In 1998 the sample was again about 16,000 adults and 4,000 children. "

4. 1988-94—National Health and Nutrition Examination Survey (NHANES III);
5. 1999-present--National Health and Nutrition Examination Survey (Continuous NHANES).

Note in NHANES data, there is only a personal identification variable (seqn), there is no household id on the public release file, the relationship of a participant to the household reference person is not publicly released[57]. Thus, we cannot track down the participants' parents via father and mother's id (as in CHNS and IFLS data), or identify the potential parents via the household id (as in English HSE data). In other words, there is no way to identify the parents by ID. However, in one of these surveys---NHANES III, there is a family background section in the youth file, where limited characteristics of the parents were collected, including mother and father's height and weight.

NHANES III, conducted between 1988 and 1994, included about 40,000 people selected from households in 81 counties across the United States. In NHANES III, black Americans and Mexican Americans were selected in large proportions, each of these groups comprised separately 30 percent of the sample. It was the first survey to include infants as young as 2 months of age and to include adults with no upper age limit. Our sample is obtained by merging the youth data which includes child's age and parents' height weight with examination data which includes child's final (medically measured) height weight. Our final sample includes 6,582 pairs of father, mother and child.

The Spanish National Health Survey (ENS-2006)

The Spanish data used here is from the Spanish National Health Survey (ENS-2006), which is the most recent statistical data collection of its type conducted by the Instituto Nacional de Estadistica (INE). This survey is representative at both the national and autonomous regional level. All the members residing at home are requested to provide information on certain demographic variables, adults answer the adult health questionnaire, and members under 16 answer the child health questionnaire. The survey covers the period between June 2006 and June 2007. The final sample used here includes

[57] With the exception of dietary data, the relationship of the sample participant to the proxy is not publicly released, either.

more than 7,000 individuals, which consists of 2,139 pairs of father-child and 3,420 pairs of mother-child. The anthropometric measures such as height and weight are self-reported.

The Survey for the Evaluation of Urban Households (ENCELURB) (Mexico)

This survey is a longitudinal data for three years (2002, 2004 and 2009) from the Survey for the Evaluation of Urban Households (ENCELURB). This survey contains comprehensive anthropometric and general health outcomes (such as weight, height, hemoglobin levels, diabetes status, etc), and all the anthropometric measures such as weight and height have been collected by medical personnel, instead of self-reported.

This survey only includes pairs of mother-child and does not contain information on fathers, as the programme was initially designed to help children and their mothers, therefore the anthropometric information collected for children (under four years old at the beginning of the program in 2002) is more specific.

The sample used in this study considers 7,413 person-wave observations constituted by 2338 pairs of children and mothers for 2002; 3,459 for 2004; and 1,616 for 2009[58]. Since children are not necessarily observed in all waves, Table A 2.7 shows the number of parent-child pairs that were observed more than once. We see that almost 50 percent of the individuals were observed at least twice in the time horizon being considered, this may allow us to apply individual fixed effects.

Table A 2.10: The number of times children were observed in Mexican data

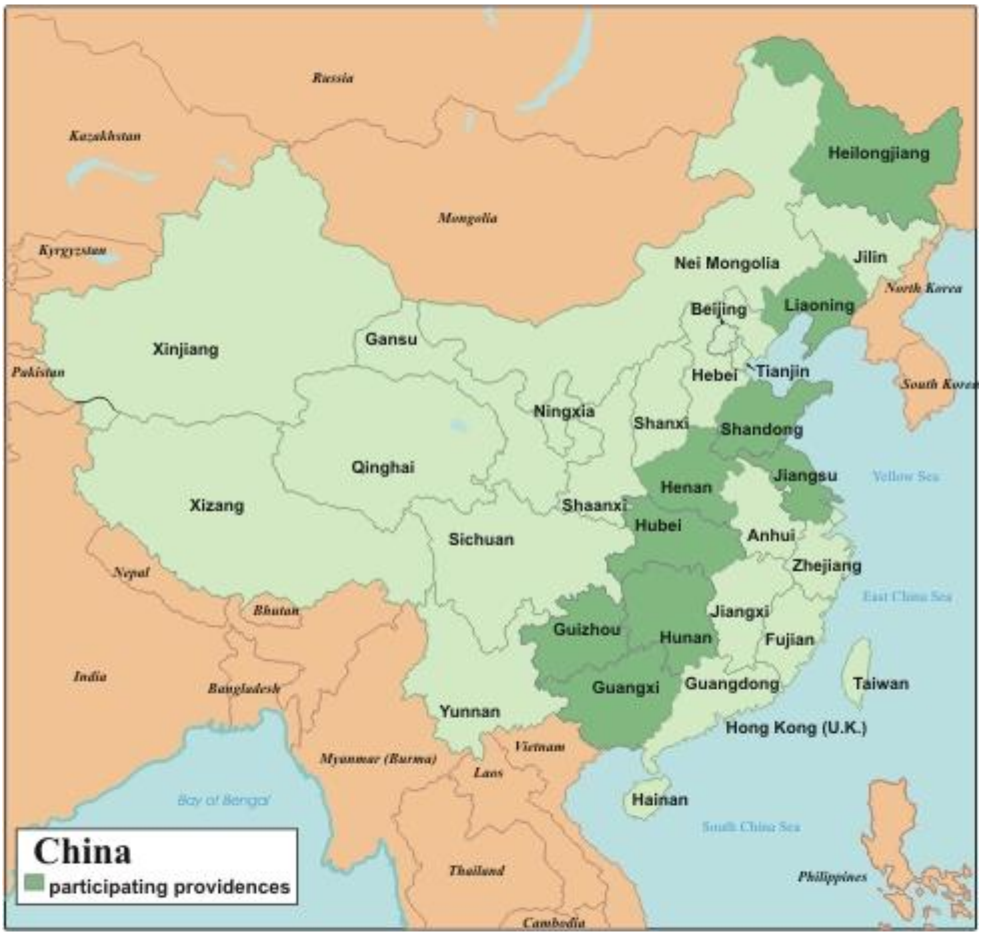
Waves (Years)	1	2	3
Observations	3,709	2,936	768

Source: ENCERLUB 2002, 2004 and 2009.

[58] The data relative to the external evaluation for the *Oportunidades* programme for Urban Households is also available for 2003, we omit this wave since the survey did not collect anthropometric measures this year.

Appendix 3

Figure A 3.1: Map of China Health and Nutrition Survey (CHNS) Regions



Note: The darker shaded regions are the provinces in which the survey has been conducted. They are: Guangxi; Guizhou; Heilongjiang; Henan; Hubei; Hunan; Jiangsu; Liaoning; Shandong. http://www.cpc.unc.edu/projects/china/proj_desc/chinamap; the CHNS data is not nationally representative, rather, this is a purposeful sample of selected provinces and within the provinces, counties and large urban areas. Thus, this result does not provide representativeness at the national, provincial or community levels.

Some descriptive statistics of the CHNS data

Table A 3.1 displays the number of parent-child pairs that were observed for multiple times. These repeated observations facilitate the possibility of netting out for time-invariant unobserved individual heterogeneity through individual fixed effects.

Table A 3.1: The number of times that children were observed in CHNS (1989-2009)

Waves	1	2	3	4	5	6	7	Total
Obs	1,840	1,933	1,164	707	352	45	3	6044
Frequency of Obs	1,840	3,866	3,492	2,828	1,760	270	21	14077

Note: In this longitudinal data, 1,840 individuals are observed for one wave, 1,933 individuals are observed for two waves, the sum of observations is 14,077.

Table A 3.2: Summary of BMI z-score when they were observed for the last time

	Obs	Mean	Std.	Min	Max
Father					
Age	6044	40.28	6.77	18.97	69.66
Height	6044	166.42	6.39	144.8	189
Weight	6044	62.70	10.08	38	115.3
BMI	6044	22.58	2.93	13.06	36.39
BMI z-score	6044	0.019	0.96	-4.64	3.15
Mother					
Age	6044	38.64	6.32	19.57	66.01
Height	6044	155.89	5.78	131	175.5
Weight	6044	55.17	8.71	33.2	98
BMI	6044	22.66	3.06	15.08	41.32
BMI z-score	6044	0.28	0.87	-2.79	3.79
Child					
Age	6044	12.28	4.438	0.02	17.99
Height	6044	142.87	24.50	50	186
Weight	6044	38.66	15.08	3.2	97.7
BMI	6044	18.02	2.89	10.74	39.07
BMI z-score	6028[59]	-0.35	1.19	-4.85	5

[59] Due to the potential error in data recording for height and/or weight, some children's BMI z-scores are considered as biologically implausible and flagged as missing by the Anthro software. In addition, we drop BMI z-scores outside of the commonly applied range (-5, 5).

Table A 3.3: Summary of height, weight and BMI z-score (from Pooled Data)

	Obs	Mean	Std.	Min	Max
Father					
Age	14077	37.51	6.93	18.97	69.66
Height	14077	166.22	6.26	144.8	189
Weight	14077	61.49	9.49	38	115.3
BMI	14077	22.20	2.76	13.06	36.39
BMI z-score	14077	-0.10	0.93	-4.64	3.15
Mother					
Age	14077	35.92	6.50	18.31	66.01
Height	14077	155.70	5.68	131	179
Weight	14077	54.11	8.36	32.5	98
BMI	14077	22.28	2.93	14.72	41.32
BMI z-score	14077	0.169	0.85	-3	3.79
Child					
Age	14077	9.59	4.78	0.02	17.99
Height	14077	128.01	27.84	50	186
Weight	14077	30.02	15.01	3.2	97.7
BMI	14077	17.07	2.72	10.74	39.07
BMI z-score	14006	-0.21	1.25	-4.99	5.38

Table A 3.2 presents the basic summary statistics of the maximum number of complete sets of child, mother and father the last survey time we observe them, Table A 3.3 presents the summary statistics of the pooled observations. They suggest that the Chinese are still, predominantly shorter and lighter than people in developed western countries, but there is still a high variance to be explained.

Discussion of BMI z-score and BMI

BMI z-scores are computed based on the comparison with reference population in terms of mean and standard deviation. In this paper, the reference population comes from the World Health Organization (WHO), there are three main versions of references: the 1978 WHO/NCHS Growth References (for children up to age10), the WHO Growth References (for children and adolescents up to age 19), and the 2006 WHO Growth Standards (for preschool children, under 6 years of age) (Wang and Chen 2012).

Most of the earlier versions are based on growth references developed and used in the US. For 1978 WHO/NCHS (National Center for Health Statistics), the growth reference for infants was developed based on data collected from the Fels Longitudinal study, which followed mainly formula-fed children in Ohio State in the USA (Wang and Chen, 2012). In 2006, WHO generate the 2006 WHO Child Growth Standards for which the data was collected from Brazil, Ghana, India, Norway, Oman, and the USA. In 2007, the 2007 WHO reference was released for children and adolescents aged 5 to 19 years, using the same sample as for the 1978 WHO/NCHS growth references (Onis 2007). Thus, the references used in this paper are the 2006 WHO Child Growth Standards for preschool children under 5 years of age and the 2007 WHO Growth Reference for school-age children and adolescent aged 5 to 19 years of age.

In stata, the BMI zscores are calculated using the igrowup_standard.ado file downloaded from the WHO website[60]. This file calculates z-scores for eight anthropometric indicators based on the WHO Child Growth Standards, body mass index (BMI)-for-age is one of them. The macro produces sex- and age-specific estimates for the prevalence of under/over nutrition and summary statistics (mean and SD) of the z-scores for each indicator. Extreme (i.e. biologically implausible) BMI z-scores (less than -5 and more than 5) are flagged.

The BMI z-score system assumes the BMI values to be normally distributed and computes the z-score based on the formula $z = \frac{x-u}{\sigma}$, where x is the individual's BMI value , μ is the mean BMI value of reference population , and σ is the standard deviation

[60] They can be downloaded from http://www.who.int/childgrowth/software/en/ for Child growth standards (0~5 years old) and http://www.who.int/growthref/tools/en/ for Growth reference (5~19 years old).

of reference population, here the reference population refers to the sample from the WHO macro software[61]. Thus, the anthropometric measure is expressed as a range of standard deviations or z-scores above or below the mean or median value of reference population. The BMI z-score adjusts the BMI value for age and gender, thus it is more comparable across ages and genders compared to BMI values which are age and gender independent (Wang and Chen 2012). In the conversion of BMI z-score, since the maximum age available for 2007 WHO reference is 19 years old, the age of parents over 19 years old will be treated as 19-year when their BMI z-scores are calculated, in other words, we assume that parents' BMI z-score follows the distribution of reference population aged 19 years old if the parents are over 19 years old. In addition, due to the potential error in data recording for height and/or weight, some children's BMI z-scores (below -5 and above 5) are considered as biologically implausible and flagged as missing by the Anthro software.

However, Note most of the papers argue that BMI is better than z-score, and z-score makes things more complicated. Cole (2007) argues that BMI or BMI % are more appropriate scales than BMI z-scores to measure the changes in adiposity, as the within-child variability over time depends on the child's level of adiposity rather than the relative position to the reference population. Though in the case of a single cross section, BMI z-score is a better measure for the level of adiposity.

[61] It is argued that it might not be appropriate to use the WHO criteria to apply to some populations in the Western Pacific Region due to their anthropometric characteristics.

Table A 3.4 : Intergenerational correlation of BMI z-score on children aged above five

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: BMI z-score of child					
BMI z-score of father	0.308***		0.267***	0.275***	0.274***
	(0.0156)		(0.0152)	(0.0151)	(0.0152)
BMI z-score of mother		0.277***	0.219***	0.244***	0.245***
		(0.0194)	(0.0184)	(0.0186)	(0.0186)
Male of child				-0.212***	-0.211***
				(0.0744)	(0.0744)
Age dummies of child				Y	Y
Age dummies of child* Male of child				Y	Y
Age of father					-0.00799*
					(0.00470)
Age of mother					0.00566
					(0.00525)
Constant	-0.402***	-0.478***	-0.453***	-0.0959*	-0.0122
	(0.0150)	(0.0158)	(0.0149)	(0.0534)	(0.114)
Observations	10,969	10,969	10,969	10,969	10,969
R-squared	0.063	0.044	0.090	0.129	0.129

Standard errors are clustered at the household level in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A 3.5: Intergenerational correlation of BMI z-score (father-son, father-daughter, mother-son, mother-daughter)

	(1)	(2)	(3)	(4)
Dependent variable: BMI z-score of child				
	Boys	Girls	Boys	Girls
BMI z-score of father	0.223***	0.212***		
	(0.0250)	(0.0242)		
BMI z-score of mother			0.215***	0.209***
			(0.0298)	(0.0249)
Constant	1.074***	0.594**	1.070***	0.593**
	(0.281)	(0.278)	(0.276)	(0.279)
Observations	5,137	4,399	5,137	4,399
R-squared	0.196	0.236	0.193	0.232

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A 3.6: The number of children in households

The number of children in the household	The number of observations	Percentage (%)
1	6,644	47.18
2	5,518	39.18
3	1,638	11.63
4	216	1.53
5	60	0.43
6	6	0.04
Total	14,082	100

Corresponding results using raw BMI (rather than BMI z-score)

Instead of BMI z-score, we also estimate the intergenerational elasticity of BMI using the log of raw BMI based on formula BMI $= \left[\frac{\text{weight(kg)}}{\text{height}^2\text{(cm)}}\right] * 10{,}000$, controlling for age, age square, gender of child, the interactions between them and the age of father and mother, as we did in Chapter 2. The results are presented below, they suggest the results of intergenerational elasticity of BMI are consistent with those of intergeneration correlation of BMI z-score (as shown in the main text of Chapter 3). This consistency implies that the intergeneration correlation of adiposity is robust to the measure of adiposity.

Table A 3.7: OLS estimates of the intergenerational BMI elasticity

Table A 3.7 is directly comparable to Table 3.1.

	(1)	(2)	(3)	(4)	(5)
Dependent variable: Log(BMI of child)					
Log (BMI of Father)	0.291***		0.238***	0.211***	0.210***
	(0.013)		(0.013)	(0.012)	(0.012)
Log (BMI of Mother)		0.282***	0.234***	0.173***	0.174***
		(0.012)	(0.012)	(0.011)	(0.011)
Age of Child				-0.035***	-0.034***
				(0.001)	(0.001)
(Age of Child)^2/100				0.267***	0.267***
				(0.005)	(0.005)
Male				0.027***	0.027***
				(0.005)	(0.005)
Age*Male				-0.002***	-0.002***
				(0.000)	(0.000)
Father’s age					-0.001
					(0.000)
Mother’s age					0.000
					(0.000)
Constant	1.926***	1.954***	1.363***	1.665***	1.680***
	(0.040)	(0.037)	(0.050)	(0.046)	(0.046)
Observations	14,082	14,082	14,082	14,082	14,082
R-squared	0.053	0.056	0.090	0.355	0.355

Notes: Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

Table A 3.8: OLS estimates of the intergenerational BMI elasticity with more controls

Table A 3.8 is directly comparable to Table 3.2.

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: Log (child's BMI)					
Log(BMI of father)	0.201***	0.198***	0.173***	0.169***	0.168***
	(0.0126)	(0.0126)	(0.0127)	(0.0127)	(0.0127)
Log(BMI of mother)	0.164***	0.163***	0.144***	0.142***	0.141***
	(0.0116)	(0.0117)	(0.0115)	(0.0116)	(0.0116)
Household characteristics	Y	Y	Y	Y	Y
Year fixed effects		Y		Y	Y
Province fixed effects			Y	Y	Y
Province*Year					Y
N	9,588	9,588	9,588	9,588	9,588
R-squared	0.347	0.348	0.362	0.364	0.372

Note: the regression also includes: age of child, $(\text{age of child})^2$, gender of child, gender*age of child, gender*$(\text{age of child})^2$, father and mother's age. Household characteristics include household income per capita, household size , father's occupation, mother's education. "Province*Year" are interactions of dummies for the child's province of residence and the survey year. Standard errors clustered at the village (or town) and year level in parentheses.

Table A 3.9: OLS estimates of the intergenerational BMI elasticity, controlling for the lagged value of Child's BMI

Table A 3.9 is directly comparable to Table 3.3.

	(1)	(2)	(3)
Dependent variable: Log(BMI of child)			
Log (BMI of Father)	0.211***	0.169***	0.169***
	(0.013)	(0.011)	(0.011)
Log (BMI of Mother)	0.195***	0.135***	0.135***
	(0.012)	(0.011)	(0.011)
Age of Child		-0.004**	-0.004**
		(0.002)	(0.002)
(Age of Child)^2/100		0.116***	0.116***
		(0.009)	(0.009)
Male		0.025***	0.025***
		(0.007)	(0.007)
Age*Male		-0.0027***	-0.002***
		(0.001)	(0.001)
Father's age			-0.0001
			(0.000)
Mother's age			-0.0003
			(0.000)
Log (Child's BMI in t-1)	0.471***	0.384***	0.384***
	(0.014)	(0.014)	(0.014)
Constant	0.259***	0.718***	0.731***
	(0.056)	(0.057)	(0.057)
Observations	8,037	8,037	8,037
R-squared	0.274	0.487	0.487

Notes: Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

Table A 3.10: Fixed Effects Estimates of the intergenerational elasticity

Table A 3.10 is directly comparable to Table 3.4 and Table 3.5.

	(1)	(2)	(3)	(4)
	Individual	Individual	Household	Household
Dependent variable: Log(BMI of child)				
Log (BMI of father)	0.130***	0.138***	0.140***	0.148***
	(0.024)	(0.024)	(0.023)	(0.023)
Log(BMI of mother)	0.124***	0.131***	0.110***	0.116***
	(0.023)	(0.023)	(0.021)	(0.021)
Obese father*mother		-0.013		-0.006
		(0.118)		(0.116)
Underweight father*mother		0.012**		0.013**
		(0.006)		(0.006)
Age of Child	-0.0334***	-0.033***	-0.036***	-0.036***
	(0.005)	(0.006)	(0.001)	(0.001)
(Age of Child)^2/100	0.277***	0.276***	0.271***	0.270***
	(0.006)	(0.006)	(0.006)	(0.006)
Male of Child			0.030***	0.030***
			(0.006)	(0.006)
Male*Age of Child	-0.003***	-0.003***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
Age of Father	0.008	0.008	0.005***	0.005***
	(0.005)	(0.005)	(0.002)	(0.002)
Age of Mother	-0.010*	-0.010*	-0.004**	-0.004**
	(0.005)	(0.005)	(0.002)	(0.002)
Constant	2.118***	2.075***	2.052***	2.005***
	(0.170)	(0.171)	(0.083)	(0.086)
Observations	14,082	14,082	14,014	14,014
Number of individuals	6,045	6,045		
Number of households			3,711	3,711
R-squared	0.399	0.399	0.356	0.356

Notes: Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

Table A 3.11: OLS estimates of the intergenerational BMI elasticity with the interactions of parental BMI variable with family socioeconomic factors

Table A 3.11 is directly comparable to Table 3.6.

	Mother-child		Father-child
Dependent variable: Log (BMI of child)			
Model1		**Model1**	
Log(BMI of mother)	0.197***	Log(BMI of father)	0.207***
	(0.0195)		(0.0232)
Income quarter: (Ref.: 0-25th percentile)		Income quarter: (Ref.: 0-25th percentile)	
25-50th percentile of income	0.0886 (0.0825)	25-50th percentile of income	0.0757 (0.0913)
50-75th percentile of income	0.126	50-75th percentile of income	0.0613
	(0.0803)		(0.0893)
>75th percentile of income	0.0278	>75th percentile of income	-0.0898
	(0.0867)		(0.0928)
25-50th* Log(BMI of mother)	-0.0313	25-50th* Log(BMI of father)	-0.0272
	(0.0269)		(0.0299)
50-75th* Log(BMI of mother)	-0.0435*	50-75th* Log(BMI of father)	-0.0225
	(0.0262)		(0.0292)
>75th* Log(BMI of mother)	-0.00861	>75th* Log(BMI of father)	0.0292
	(0.0282)		(0.0302)
Observations	13,707	Observations	13,707
R-squared	0.359	R-squared	0.359

	Mother-child		Father-child
Dependent variable: Log (BMI of child)			
Model 2		**Model 2**	
Log(BMI of mother)	0.215***	Log (BMI of father)	0.174***
	(0.0121)		(0.0158)
Highest degree: (Ref.: Primary and less)		Occupation: (Ref.: famer)	
High school	0.00656	Skilled/non- worker and other	-0.131*
	(0.0720)	skilled/service	(0.0701)
Technical and Tertiary	0.178	Professional/ /manager/office	-0.0879
	(0.139)	technical/executive/	(0.0952)
		administrative	
High school* log (BMI of	0.000494	Skilled/non-skilled	0.0432*
mother)	(0.0233)	*log (BMI of father)	(0.0228)
Technical and Tertiary*	-0.0498	Professional/	0.0328
log (BMI of mother)	(0.0448)	*log (BMI of father)	(0.0307)
Observations	10,346	Observations	13,237
R-squared	0.339	R-squared	0.360

	Mother-child		Father-child
Dependent variable: Log (BMI of child)			
Model 3		**Model 3**	
Log (BMI of mother)	0.206***	Log (BMI of father)	0.182***
	(0.0107)		(0.0183)
The proportion of time in		The proportion of time	
50-75% of time in	0.101	50-75% of time in	-0.0730
poverty	(0.102)	poverty	(0.110)
1-50% of time in poverty	0.163*	1-50% of time in poverty	0.0537
	(0.0878)		(0.103)
Never in poverty	0.0707	Never in poverty	-0.135*
	(0.0662)		(0.0710)
50-75% of time in	-0.0333	50-75% of time in	0.0233
Poverty*log (BMI of	(0.0331)	Poverty*log (BMI of	(0.0358)
mother)		mother)	
1-50% of time in	-0.0534*	1-50% of time in	-0.0178
Poverty*log (BMI of	(0.0285)	Poverty*log (BMI of	(0.0334)
mother)		mother)	
Never in poverty*	-0.0214	Never in poverty*	0.0453*
log (BMI of mother)	(0.0215)	log (BMI of mother)	(0.0231)
Observations	14,082	Observations	14,082
R-squared	0.356	R-squared	0.356

Notes: the regression also includes: age of child, (age of child)^2, gender of child, gender*age of child, gender*(age of child)^2, father and mother's age. Standard errors clustered at the village (or town) and year level in parentheses.

Table A 3.12: OLS estimates of the intergenerational BMI elasticity between mother and child by SES measures: by income level, mother's education and world poverty line

Table A 3.12 is directly comparable to Table 3.7.

	Sample size	Elasticity	Std. Error	R-squared
<25th percentile of Income	3,323	0.194***	0.0188	0.349
25-50th percentile of Income	3,462	0.171***	0.0190	0.372
50-75th percentile of Income	3,461	0.159***	0.0191	0.375
>75th percentile of Income	3,461	0.186***	0.0230	0.327
Primary school	3,250	0.176***	0.0201	0.3784
High school	6,457	0.174***	0.0169	0.3265
Technical and Tertiary	639	0.139***	0.0478	0.2801
75-100% of time in poverty	4,796	0.194***	0.0167	0.394
50-75% of time in poverty	1,522	0.170***	0.0364	0.276
1-50% of time in poverty	1,481	0.144***	0.0258	0.296
Never in poverty	6,283	0.167***	0.0165	0.341

Notes: Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

Table A 3.13: OLS estimates of the intergenerational BMI elasticity between father and child by SES measures: by income level, father's occupation and world poverty line

Table A 3.13 is directly comparable to Table 3.8.

	Sample	Elasticity	Std.	R-
<25th percentile of Income	3,323	0.204***	0.0235	0.349
25-50th percentile of Income	3,462	0.182***	0.0207	0.372
50-75th percentile of Income	3,461	0.190***	0.0193	0.375
>75th percentile of Income	3,461	0.233***	0.0208	0.327
Professional/technical/administrator/executive/manager/office	1,811	0.203***	0.0298	0.311
Skilled/non-skilled/service worker and	4,347	0.223***	0.0188	0.347
Farmer	7,079	0.174***	0.0164	0.378
75-100% of time in poverty	4,796	0.177***	0.0202	0.394
50-75% of time in poverty	1,522	0.217***	0.0349	0.276
1-50% of time in poverty	1,481	0.171***	0.0297	0.296
Never in poverty	6,283	0.229***	0.0160	0.341

Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

Figure A 3.2: Quantile estimates of the intergenerational BMI elasticity

Figure A 3.2 is directly comparable to Figure 3.5.

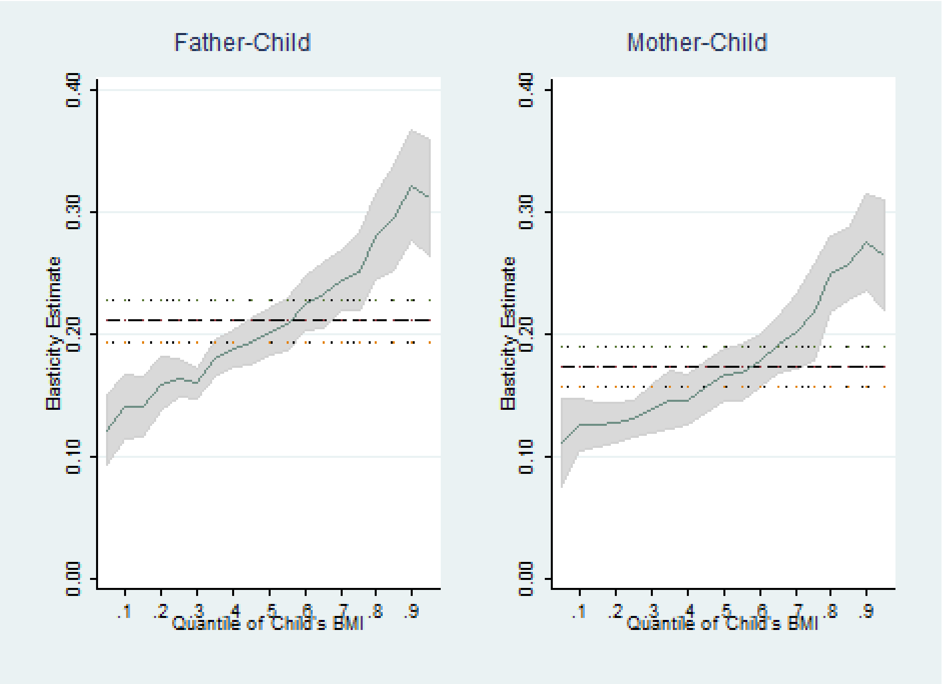


Figure A 3.3: Quantile estimates of the intergenerational BMI elasticity on Children aged 16~18 years old

Figure A 3.3 is directly comparable to Figure 3.6.

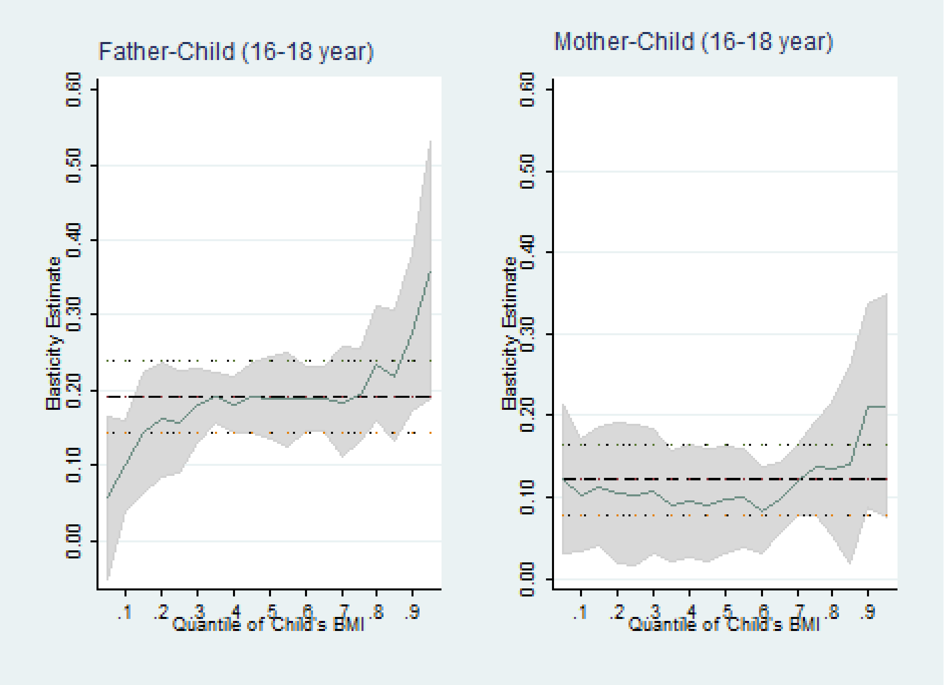


Table A 3.14: Estimates of the intergenerational BMI elasticity by age group

Table A 3.14 is directly comparable to Table 3.9.

Age Stage	Obs	Father and Child (s.e)	Mother and Child (s.e)
0-2	1551	0.087***	0.0137***
2-4	1448	0.081***	0.093***
4-6	1634	0.186***	0.156***
6-8	1623	0.221***	0.163***
8-10	1783	0.306***	0.214***
10-12	1976	0.257***	0.213***
12-14	1882	0.273***	0.227***
14-16	1576	0.208***	0.182***
16-18	609	0.158***	0.095***

Notes: Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

Figure A 3.4 : Estimates of the intergenerational BMI elasticity by age group

Figure A 3.4 is directly comparable to Figure 3.7.

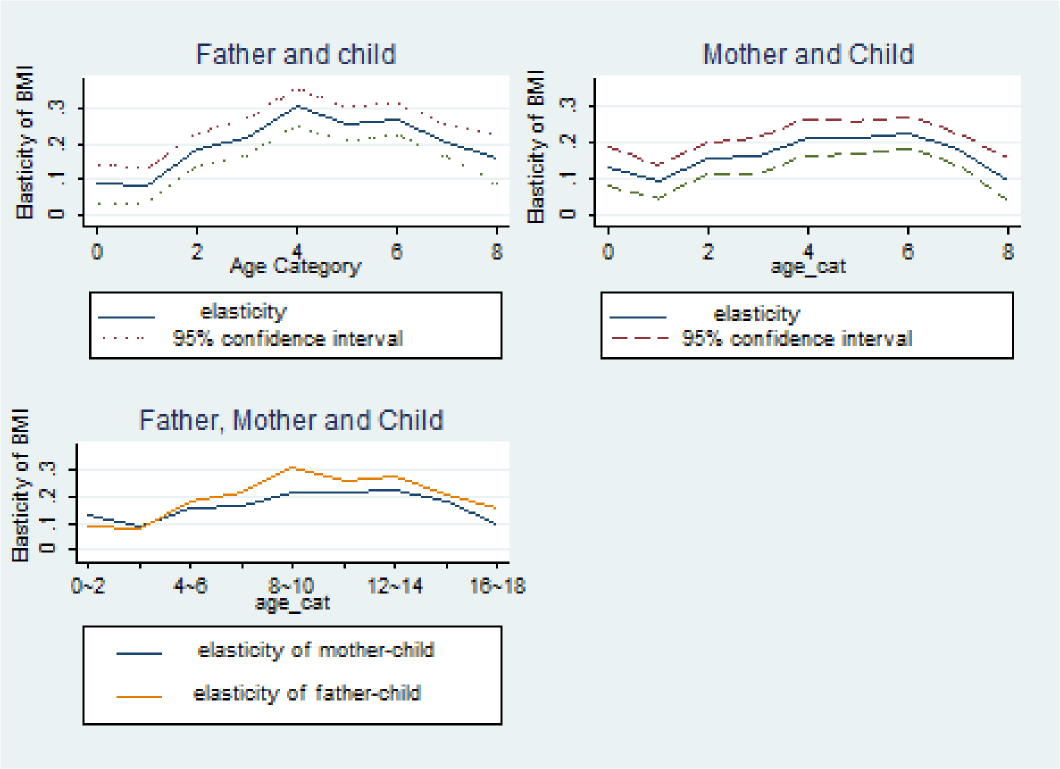


Table A 3.15: OLS estimates of the intergenerational BMI elasticity, controlling for interactions between a dummy indicating whether both parents are obese and a dummy indicating whether both parents are underweight

	(1)	(2)	(3)
Dependent Variable: Log (BMI of Child)			
Log (BMI of Father)	0.219***	0.174***	0.173***
	(0.014)	(0.012)	(0.012)
Log (BMI of Mother)	0.203***	0.139***	0.139***
	(0.013)	(0.011)	(0.011)
Obese father*mother	-0.024	-0.024	-0.026
	(0.063)	(0.069)	(0.069)
Under weight father*mother	0.015**	0.008	0.008
	(0.007)	(0.006)	(0.006)
Age of Child		-0.004**	-0.004**
		(0.002)	(0.002)
(Age of Child)^2/100		0.115***	0.116***
		(0.009)	(0.009)
Male		0.025***	0.025***
		(0.007)	(0.007)
Age*Male		-0.002***	-0.002***
		(0.001)	(0.001)
Father's age			-0.0001
			(0.000)
Mother's age			-0.0004
			(0.000)
Log (Child's BMI in t-1)	0.471***	0.384***	0.384***
	(0.014)	(0.014)	(0.014)
Constant	0.211***	0.691***	0.704***
	(0.061)	(0.061)	(0.062)
Observations	8,037	8,037	8,037
R-squared	0.275	0.487	0.487

Notes: Robust standard errors in parentheses, standard errors clustered at individual level, *** p<0.01, ** p<0.05, * p<0.1

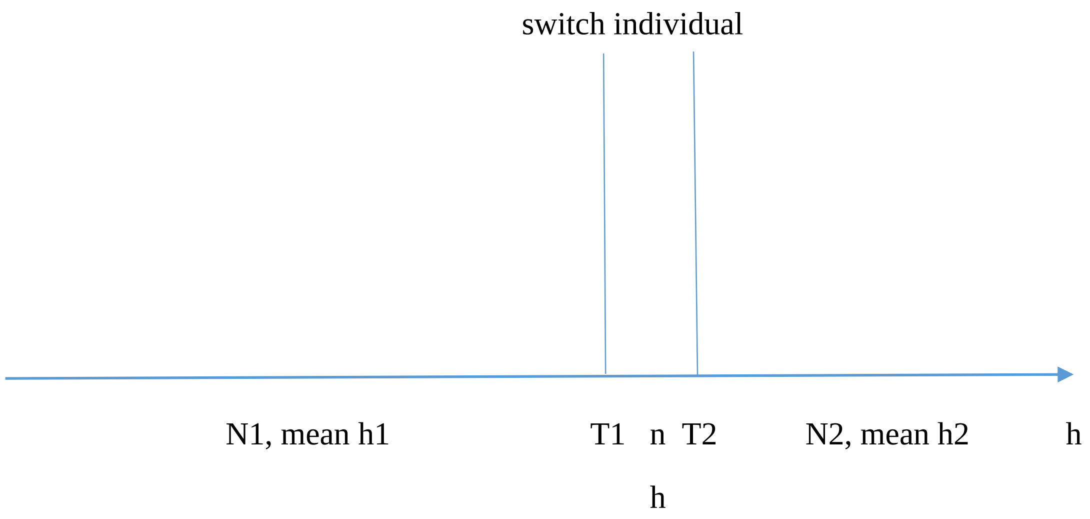
Appendix 4

China Health and Nutrition Survey (CHNS)

In chapter 3 we already have a brief introduction of CHNS data. Here we provide more information, with an emphasis on the sampling and follow-up of this longitudinal survey.

Popkin (2014) provides a detailed description of the CHNS data. The CHNS is a collaborative project between the Carolina Population Center (CPC), University of North Carolina at Chapel Hill and the National Institute of Nutrition and Food Safety, CCDC. The CHNS was designed as a household-based study which covers nine provinces and eight rounds of surveys between 1989 and 2009. A multistage, random cluster design was used in eight provinces (Liaoning, Jiangsu, Shangdong, Henan, Hubei, Hunan, Guangxi and Guizhou) to select a stratified probability sample. Under this sampling scheme, two cities (one large and one small, usually a low income city and the provincial capital) and four counties (stratified by income, one high-, one low- and two middle income) per province were selected. Within cities, two urban and two suburban communities were randomly selected. Within counties, one community in the capital city and three rural villages were randomly chosen, twenty households were then randomly selected from each community. In 1997, Heilongjiang province joined the survey since Liaoning province dropped out of the survey, in 2000 Liaoning was added back into the survey.

Figure A 4.1: Alternative Definition of health selectivity



In addition to the definition we use in Chapter 4, there is an alternative definition of health selectivity: Since $W_s^j\left(h^j\right) = W_0(1+\lambda h^j)$, for the sake of exposition, if we hold α and W_0 fixed, $(\alpha-1)W_s^j(h^j)$ depends only on h^j. Suppose there is a health axis h, we rate everyone along the axis, there is a threshold T above which people migrate, this threshold corresponds to the threshold T (iC_0) in Figure 1 and 2 , as α and W_0 are fixed now. As Figure A 2.1 suggests, suppose there are two thresholds T_1 and T_2, those who are rated between them are those who switch the migration status when the threshold shifts from T_1 to T_2, the number of them is n and the mean of their health is h. Initially the threshold is located at T_1, those who are rated on the left side of the threshold are non-migrants, the number of them is N_1, and the mean of their health is h_1. By contrast, there are $n+N_2$ people on the other side of the threshold, the average health of these migrants is $(N_2h_2+nh)/(N_2+n)$. When the threshold shifts to T_2, the number of non-migrants increases to N_1+n, and the number of migrants declines to N_2. As a consequence, the average health of non-migrants is $(N_1h_1+nh)/(N_1+n)$, and the average health of migrants now is h_2. The change in these mean health captures the selectivity, if the change in the mean health of migrants is larger than that of non-migrants when the threshold shifts up, it implies a positive health selectivity on migrants. As we see, for non-migrants, the change in their mean health is $\frac{N_1h_1+nh}{N_1+n} - \frac{(N_1+n)h_1}{N_1+n} = \frac{n(h-h_1)}{N_1+n}$, whereas this change for migrants is $\frac{N_2+n}{N_2+n}h_2 - \frac{N_2h_2+nh}{N_2+n} = \frac{n(h_2-h)}{N_2+n}$. Which is larger depends on $N_1, N_2, (h - h_1)$ and $(h_2 - h)$. In other words, what happens to the mean depends partly on the number of people moving (n) when the threshold shifts, relative to the initial number of people in each group, the same number of people might account for a larger proportion for one group (normally migrants) than for the other.

A more parsimonious model

In addition to the specification we apply in Chapter 4, we estimate a more parsimonious model which allows us to have a larger sample size. As we showed earlier in Table 4.4, the results across the pooled sample suggest that the likelihood of migration is higher for those self-evaluating as having "excellent", "good" or "fair" health than for those self- evaluating as having "poor" health, this indicates most of the distinction comes from "poor" and the rest three categories. Therefore, we combine "fair", "good" and "excellence" together and convert the variable of self-rated health into a binary variable which equals one if the respondents rate themselves as having "fair", "good" or "excellent" health, zero if they evaluate themselves as having "poor" health. Using this binary version of health variable, we estimate the baseline model (Table 4.4) and present the results in Table A 4.1, they suggest that those who self-evaluate themselves as having fair or good or excellent health are more likely to migrate than those who evaluate themselves as having poor health, this result is consistent with the results when we use the four-category version of health variable (Table 4.4).

Table A 4.1: Probit regression of migration status on health using a more parsimonious model (corresponds to Table 4.4)

	(1) Pooled Coeff.	Se.
Fair/Good/Excellent	0.344**	(0.16)
Trouble working due to illness in the last three months	0.171	(0.11)
history of Bone Fracture	0.082	(0.13)
Ever Smoked	0.060	(0.05)
Observations	8528	

Note: The equation also includes other controls in the baseline equation (except for the objective health measures); standard errors are in Parentheses, *** p<0.01, ** p<0.05, * p<0.1

Next we interact this binary version of health variable with occupation and include them in the estimation, the results are presented in Table A 4.2. They suggest that those who have fair or good or excellent health are more likely to migrate than those who evaluate their health as being poor. In addition, the negative coefficients of the variables "non-farm" and "skilled" suggest that for those who evaluate their health as being "poor", those whose occupations are “non-farm” and “skilled” are less likely to migrate than those who are unemployed or students, this might be due to the fact that those who are unemployed or students with poor health. The coefficients of the interaction terms “Fair/Good/Excellent *Non-farm” and “Fair/Good/Excellent * Skilled” are positive, they suggest that compared with those who are unemployed or students, health has a stronger positive relationship to the migration probability for non-farm and skilled workers.

Table A 4.2: The estimates of health effects by occupation using a more parsimonious model (corresponds to Table 4.5)

Dependent variable: Probability of migration	Pooled coeff	s.e.
Self-rated health: Poor (Ref.)		
Fair/Good/Excellent	0.469*	(0.28)
Occupation: Unemployed/student (Ref.)		
Farmer	0.415	(0.34)
Non-farm	-3.021***	(0.32)
Skilled	-3.521***	(0.36)
Non-skilled	-0.051	(0.55)
Service	0.592	(0.56)
Fair/Good/Excellent * Farmer	-0.383	(0.34)
Fair/Good/Excellent * Non-farm	2.852***	(0.32)
Fair/Good/Excellent * Skilled	3.555***	(0.37)
Fair/Good/Excellent * Non-skilled	0.069	(0.56)
Fair/Good/Excellent * Service	-0.603	(0.56)
Observations	8790	

Note: The equation also includes other controls in the baseline equation (except for the objective health measures); standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

The replication of Tong and Piotrowski (2012)'s work and several issues in the replication:

Motivated by the "healthy migrant hypothesis" which posits that migrants tend to be positively selected on health, we started to investigate the effects of health on migration using data from China. In the literature, we found a study by Tong and Piotrowski (2012). Based on a pooled sample composed of individuals aged 16-35 years old from five waves (1997-2009) of the China Health and Nutrition Survey (CHNS), they employ binary probit regressions to estimate the health effects on migration decision, where health and other control variables are measured in the wave preceding the measure of migration status. For instance, those who are defined as migrants in 2004 had their health measured in 2000. They find that migrants are positively selected on the basis of health, but more strongly at the beginning of their sample than at the end, their results suggest a strong positive selection on health among migrants. However, their study includes only a limited menu of tests, they do not explore the interactions between health and other factors; they do not explore the effects of lagged health; their study does not have a theoretical model to support their empirical analysis. Therefore, we attempt to make some extension analysis in this study, and a natural place to start was replication[62]. However, eventually we find we could not replicate their results. In the following we will go through the process how we conducted this replication, including alternative definitions for some variables, data versions and what differences these alternatives make to the estimate results. Also we will discuss some concerns about the way they choose the sample and how the corresponding estimates match up with the literature.

[62] we are grateful to Yuying Tong and Martin Piotrowski (2012) for some of their stata dofiles and their guidance (from the conversation with them through email) in helping me define several variables at the earlier stage of this replication. In this appendix we will quote some of their emails, which provide hints on how they define some variables and the version of data they used. (They were trying to be helpful, though they felt constrained from sharing their data by confidentiality arguments, either able to fully describe the process or share the dofile on how they process the data.)

Difference in data

We start by trying to replicate Tong and Piotrowski (2012)'s sample, and find we could not replicate their data, the descriptive statistics based on our sample are presented in Table A 4.3, in comparison with the corresponding statistics from Tong and Piotrowski (2012)'s sample, we see there are substantial differences between our sample and their sample. Tong and Piotrowski (2012) are not able to supply their data, we tried various ways to replicate the data. Next we discuss potential explanations for the differences between our sample and their sample.

First, the data we use in the replication might be different from the data used in Tong and Piotrowski (2012)'study. We use the 1997-2009 waves of the 2011 longitudinal data[63]; the descriptive statistics based on this data are different from those from Tong and Piotrowski (2012)'s study. We also collected data using 2009 longitudinal data (downloaded in October 2012), but find the descriptive statistics based on 2011 longitudinal data are almost the same as those based on 2009 longitudinal data (downloaded in October 2012). Therefore, if the differences in the statistics between Tong and Piotrowski (2012)'s sample and our sample are due to the differences in the version of data, this data difference appears to come from the difference between 2009, 2011 longitudinal data and the 1989-2009 data used in Tong and Piotrowski (2012)'s study, rather than the difference between 2009 longitudinal data and 2011 longitudinal data. According to one of the authors---Yuying Tong, this data difference might be due to "their research design or because of different data version." Since they started the study in 2010, "at that time, CHNS data they released are different from now" and requires "more data management than now" [64]. Tong and Piotrowski (2012) "used SAS program for most of the data management" and "switched back and forth between SAS and stata"[65]. This transition between the programs might be why they could not provide a perfectly clean account about their data and definitions. We speculate the data Tong and Piotrowski (2012) use might be pooled from the earlier version of cross section datasets, which might

[63] Downloaded from the website of China Health and Nutrition Survey in February 2014 http://www.cpc.unc.edu/projects/china, overall it includes data in the wave 1989, 2000,2004, 2006, 2009 and 2011.

[64] From Tong's email on 7 March 2014.

[65] From Tong's email on 7 Febuary 2014.

be downloaded before the release of the longitudinal CHNS data. Further evidence of the differences in data appears when we generated the variable for “real income in 2006 currency”. We find in the stata codes sent by one of the authors[66], this variable is directly available in their data, but not in either the 2011 longitudinal data or the 2009 longitudinal data[67]; rather, this variable “real income” is inflated to “2011 currency” (in the 2011 longitudinal data) or “2009 currency” (in the 2009 longitudinal data). Therefore, the data used in Tong and Piotrowski (2012)’s study might be different from the 2009 or 2011 longitudinal data and it might not be available anymore.

As Table A4.1 suggests, the sample size of our sample is larger than Tong and Piotrowski (2012)’s sample. One potential explanation for this larger size of our sample is the following: since “the study was initiated with a different topic of health and skill selection”[68], this data might be set up and configured for a particularly different research question, as a consequence, Tong and Piotrowski (2012) might start with a larger number of variables. In the process of data construction, since they use only observations with all the variables realized, their sample size might shrink due to the absence of some observations in some of the variables that they expected to use but eventually did not. In other words, they may have retained other variables where there is a large number of observations in the original dataset, but they did not use them in the paper. However, we started the replication directly with their paper, and constructed our sample directly based on the realization of the variables used in the paper. As a consequence, the sample size is larger than theirs. Secondly, in Tong and Piotrowski (2012)’s study, they “remove cases from the Northeast region (N=337) in 2004 because no one in the sample from that region migrated in that year”. However, in our sample, we find there are 25 migrants of the 329 individuals from the northeast area between the wave 2004 and 2006, so we keep the northeast area in the wave 2004.

[66] Here we are grateful to Yuying Tong and Martin Piotrowski (2012) for their helpful stata codes on the regression process, they do not supply stata codes on how they construct the variables.

[67] Online there is a pdf file “Household Income Variable Construction”, it describes how income variables such as “the income inflated to 2006 Yuan currency values” are constructed in the CHNS data, but the dataset is not now available online.
http://www.cpc.unc.edu/projects/china/data/datasets/Household%20Income%20Variable%20Construction.pdf

[68] According to Yuying Tong’s email on 9Feb 2014, she also mentioned they “have quite a large number of SAS program for this study and many of them do not really exactly apply this study”.

Difference in descriptive statistics

Based on the 2011 longitudinal data, Table A 4.3 presents the descriptive statistics of our sample in comparison against Tong and Piotrowski (2012)'s sample. As in Tong and Piotrowski (2012)'s study, we apply the complete case analysis on cross section data by wave; the observations with missing data on any of the independent variables are removed from the sample.

Table A 4.3: The descriptive statistics for independent variables (in comparison with those in Tong and Piotrowski (2012))

wave	Pooled		1997		2000		2004		2006	
variable	Mean		Mean		Mean		Mean		Mean	
	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao
Self-rated health										
Poor/fair	0.19	0.19	0.16	0.16	0.21	0.19	0.25	0.24	0.21	0.19
Good	0.60	0.60	0.66	0.66	0.56	0.56	0.54	0.54	0.56	0.58
Excellent	0.21	0.21	0.19	0.18	0.22	0.25	0.21	0.22	0.23	0.23
Difficulty with ADLs	0.03	0.03	0.03	0.02	0.03	0.04	0.03	0.03	0.04	0.04
Bone fracture	0.02	0.02	0.01	0.01	0.03	0.03	0.02	0.03	0.03	0.03
Ever smoked	0.27	0.26	0.24	0.25	0.28	0.25	0.29	0.27	0.29	0.27
Demographic										
Age	26.99	26.97	25.86	25.86	27.13	26.82	28.10	28.11	28.31	28.45
Gender (male)	0.53	0.49	0.52	0.50	0.53	0.47	0.55	0.48	0.53	0.49
Ever married	0.56	0.62	0.49	0.54	0.57	0.62	0.62	0.69	0.66	0.70
Highest degree earned										
Primary or lower	0.22	0.23	0.25	0.26	0.23	0.24	0.17	0.20	0.16	0.16
Lower middle	0.49	0.48	0.5	0.49	0.50	0.50	0.46	0.47	0.47	0.48
Upper middle	0.15	0.16	0.16	0.17	0.14	0.14	0.17	0.17	0.13	0.15
Technical/vocational	0.08	0.08	0.06	0.06	0.07	0.07	0.11	0.11	0.12	0.12
College and beyond	0.06	0.05	0.03	0.02	0.06	0.05	0.09	0.07	0.11	0.09
Occupation										
None/student	0.25	0.22	0.22	0.18	0.26	0.20	0.27	0.30	0.26	0.28
Farmer	0.34	0.38	0.42	0.45	0.34	0.45	0.25	0.26	0.25	0.26
Non-farm	0.14	0.12	0.11	0.10	0.17	0.12	0.15	0.13	0.16	0.13
Skilled	0.06	0.07	0.06	0.06	0.06	0.07	0.06	0.07	0.06	0.07
Non-skilled	0.10	0.09	0.10	0.10	0.07	0.07	0.11	0.10	0.10	0.10

	Pooled		1997		2000		2004		2006	
	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao
Service	0.12	0.12	0.09	0.10	0.10	0.10	0.16	0.14	0.16	0.16
Ever migrated since 1993	0.08	0.09	0.05	0.07	0.05	0.05	0.12	0.10	0.13	0.18
Household characteristics										
Rural	0.77	0.71	0.77	0.73	0.78	0.72	0.75	0.67	0.78	0.70
Household size	4.43	4.35	4.44	4.40	4.34	4.25	4.53	4.27	4.42	4.45
Real income in 2006 currency	4359.28	3427.64	4118.24	2485.00	4305.25	2978.50	5473.47	4443.12	6130.00	5086.25
Log (income per capita)	9.58	11.98	9.51	11.98	9.58	11.98	9.63	11.99	9.67	11.99
Parents										
Both parents <56	0.09	0.31	0.08	0.36	0.07	0.31	0.12	0.26	0.12	0.24
One parent >55	0.10	0.11	0.10	0.11	0.10	0.10	0.11	0.11	0.12	0.13
Both parents > 55	0.30	0.10	0.36	0.09	0.30	0.08	0.24	0.12	0.24	0.11
No parents	0.50	0.49	0.46	0.45	0.54	0.51	0.52	0.52	0.51	0.52
Spouse	0.43	0.61	0.41	0.54	0.38	0.62	0.42	0.68	0.56	0.69
Child	0.43	0.56	0.39	0.55	0.36	0.54	0.46	0.58	0.55	0.59
Region										
Coastal	0.22	0.21	0.22	0.22	0.22	0.20	0.26	0.20	0.20	0.21
Northeast[69]	0.13	0.19	0.13	0.14	0.17	0.27	-	0.20	0.20	0.19
Inland	0.36	0.34	0.38	0.38	0.31	0.27	0.41	0.33	0.33	0.33
Southern mountain	0.29	0.26	0.28	0.27	0.29	0.26	0.33	0.26	0.27	0.26
Wave										
1997	0.41	0.40	-	-	-	-	-	-	-	-
2000	0.23	0.23	-	-	-	-	-	-	-	-
2004	0.18	0.20	-	-	-	-	-	-	-	-
2006	0.18	0.17	-	-	-	-	-	-	-	-
Total number of cases	7,986	8,528	3,313	3,423	1,818	1,956	1,419	1,738	1,436	1,411

[69] As said above, Tong and Piotrowski (2012) “remove cases from the Northeast region (N=337) in 2004”.

As Table A 4.3 suggests, there are appreciable differences in the descriptive statistics for some variables, these variables are mostly dummy variables. The variables where the difference between our descriptive statistics and Tong and Piotrowski (2012)'s are greater than 10% are shaded grey. For instance, the proportion of males is around 53% in Tong and Piotrowski (2012)'s statistics, while around 49% in our statistics, and this difference is prevalent in the samples for separate waves[70].Conversely, for the variable "ever married", 56% of the respondents in Tong and Piotrowski (2012)'s pooled sample were reported as "ever married" (divorce, separation and widow-"previously married" and "currently married"), while this proportion is around 62% in our sample[71]. For the variable of "occupation", sixteen types of occupations in the raw data are classified into six main categories (Tong and Piotrowski 2012)[72]. Table A 4.3 suggests that 25% of the pooled sample in Tong and Piotrowski (2012)'s study are students or unemployed, while this fraction is around 22% in our pooled sample[73]. Similarly, there are more non-farm workers in Tong and Piotrowski (2012)'s pooled sample (14%) than in our pooled sample (12%), and fewer farmers in Tong and Piotrowski (2012)'s pooled sample (34%) than those in our sample (38%). However, the fraction of households living in rural areas is greater in Tong and Piotrowski (2012)'s sample (77%) than that in our sample (71%)[74]. In addition, the statistics of "real income in 2006 currency" in our sample is significantly different from those in Tong and Piotrowski (2012)'s, this is not surprising since the "real income in 2006 currency" is not available in our data, we generate this variable by converting "the income in 2011 currency" using the consumer price index (2005=100) from World Bank.

[70] The gender is coded based on the codebook of CHNS 1989-2011 longitudinal data.

[71] Notice there are 31 observations valued as "6", where the corresponding meaning is not found in the codebook. They are coded as missing in our sample.

[72] "Farmer, fisherman, hunter" in the data are coded as farmer, "senior professional/technical worker", "junior professional/technical worker", "administrator/executive/manager" and "office staff " are coded as non-farm worker", "skilled worker" as "skilled worker", "non-skilled worker" as "non-skilled worker", "army officer, police officer", "ordinary soldier, policeman", "driver" and "service worker" as "service worker".

[73] The variable "student/unemployed" is defined based on the question "Are you presently working?" and "currently in school?", together with one of the occupation categories as a student.

[74] The rural/urban status is defined using the variable "rural/urban site", which is the same as in the stata dofile sent by one of the authors. The proportion of rural residents based on this binary variable is different (71% in our sample compared to 77% in Tong and Piotrowski (2012)'s sample), this might suggest significant differences between our sample and Tong and Piotrowski (2012)'s sample.I also try to define this variable using question on "the type of household registration", and geographic code in the household questionnaire, but in the end we adopt the variable "rural/urban site", since this one appears closer to the definition "from rural or urban area".

There are appreciable differences in the statistics of the variable for migration experience---“ever migrated since 1993”. This variable is generated based on the definition of “migration status”. In Tong and Piotrowski (2012)’s study, migration status is defined as a change in residence across panels. It equals to “1” if the individual moved out of his/her county, city or province, and changed her household registration status (hukou) in the next wave; or was absent due to work, military service or other reasons in the next wave; otherwise, it equal to “0”. Derived from this variable, the variable of “ever migrated since 1993” is measured as a dummy variable which equals to “1” if the individual “migrates” in any of the previous waves after 1993 (i.e. equals to “1” if the variable of “migration status” equals to “1” in any of the previous waves since 1993). For instance, for migrants in 2006, the variable “ever migrated since 1993” equals to “1” if they migrated in 1993 or 1997 or 2000 or 2004.

Perhaps the most significant differences between Tong and Piotrowski (2012)’s descriptive statistics and ours come from the variable “the residence and age of the respondent’s parents”. Parents’ “residence” is a dichotomised variable, which is defined based on the question “Does your father/mother live in the home?”, their ages are merged from the file of “physical examination” through the parents’ identification number---“father/mother’s line number”. As Table A 4.3 suggests, for around 30% of respondents in our pooled sample, both their father and mother reside at home and are aged below 56 years old, compared with around 9% in the pooled sample used by Tong and Piotrowski (2012). By contrast, only around 10% of the respondents in our pooled sample reported having both parents at home and parents’ age over 55 years, in comparison with around 30% in Tong and Piotrowski (2012)’s statistics.

This difference is significant, so in order to find out which one is more plausible, we turn to other sources of information on household structure in China. Based on the sample of individuals aged between 16 and 35 years old, Table A 4.4 presents the summary statistics of “parents’ residence and age” variable in China’s 1990 census data[75], and the wave 1991 and 1993[76] of the 1989-2011 CHNS longitudinal data[77]. As we see, the

[75] Downloaded from: https://international.ipums.org/international/index.shtml

[76] We choose these two waves since their timing is close to the China 1990 census data, we do not use data in the wave of 1989, since the information on father, mother's presence and line number are not collected in the 1989 wave of CHNS survey.

[77] The samples in comparison here consist of individuals aged from 16 to 35 years old, in order to consist with the range of age chosen by Tong and Piotrowski (2012). In Table 2, Tong and Piotrowski (2012)’s

descriptive statistics on these samples are close to each other, and they are closer to those from our sample than to Tong and Piotrowski (2012)'s sample. Specifically, column (1) suggests that in the 1990 China census data, around one quarter of the respondents live with both parents who are aged 56 years and below, this figure is closer to the corresponding fraction 31% from our sample, than the 9% from Tong and Piotrowski (2012)'s sample. Similarly, for the third category---respondents with "no parents", the fraction from Census 1990 data (7%) (or 1991, 1993 of the CHNS data (10%)) is significantly closer to the corresponding fraction (10%) in our sample, compared to 30% from Tong and Piotrowski (2012)'s sample. Similarly, these fractions in the wave 1991 and 1993 of the 1989-2011 CHNS longitudinal data are also closer to those in our sample, compared to those in Tong and Piotrowski (2012)'s sample.

Table A 4.4: The descriptive statistics of the variable of parental residence based on different samples (age 16~35 years old)

	(1) Census 1990	(2) CHNS1991 longitudinal data	(3) CHNS1993 longitudinal data	(4) Our pooled sample	(5) Tong (2012) Pooled sample
Parents					
Both parents < 56	0.25	0.26	0.29	0.31	0.09
One parent > 55	0.10	0.12	0.12	0.11	0.10
Both parents > 55	0.07	0.10	0.10	0.10	0.30
No parents	0.58	0.52	0.49	0.49	0.50
The number of observations	4,317,690	3,453	3,011	8,528	7,986
Spouse[78]	0.59	0.55	0.53	0.61	0.43
Child[79]	0.61	0.56	0.56	0.56	0.43
Observations	4,534,873	3,988	3,463	8,528	7,986

sample and our sample consist of "complete cases", which comprises observations without missing data on any of the independent variables; while the sample based on 1990 census and 1991, 1993 wave of the CHNS data only use the variable for "parents' residence and age".

[78] Here the respondents with "spouse=1" includes those who are married and the spouse is present, those with "spouse=0" includes both non-married people (never-married, widowed, divorced, separated) and people who are married but spouse is absent. In the census1990 "spouse=0" includes "Single/never married, Separated/divorced/spouse absent, Widowed', "spouse=1" includes "Married/in union".

[79] Children are defined as those younger than 12 years old.

In addition, in order to check whether “the age of parents” in our sample is identified in a plausible way, we also follow up those households which were observed for multiple waves and explore whether these parents and their children age through time in a right way (i.e. whether their age increase with the interval between waves)[80]. Table A 4.5 presents the number of times that individuals aged 16-35 years old in the CHNS unbalanced data (1989-2009) are repeatedly observed (i.e. the number of individuals which are observed for different lengths of period in the longitudinal data). Column 2 (observations 3,323 with frequency 6,646) suggests that 3,323 individuals are observed for two waves,…, and column 7 suggests that 11 individuals are observed for seven waves. For those who were observed for seven waves (between 1991 and 2009), Table A 4.6 presents the summary statistics of their age and their parents’ age (if the parents’ ages are available), it suggests that the age of parents and children mostly increase by the correct interval between waves. Figure A 4.1 provides a visualization of this pattern. For instance, in Figure A 4.1 (a), the lines connect the age of father observed in separate waves, each line traces the age of father over time for a given father, the increasing trend indicates that the age of father increases with waves.

Table A 4.5: The number of times that individuals (aged 16-35 years old) were observed in the CHNS unbalanced data (1989-2009)

Waves	1	2	3	4	5	6	7	Total
Obs	5,328	3,323	2,078	1,059	384	79	11	12,262
Frequency	5,328	6,646	6,234	4,236	1,920	474	77	24,915

Note: 5,328 individuals are observed for one wave, 3,323 individuals are observed for two waves, the sum of observations made on 12,262 individuals is 24,915.

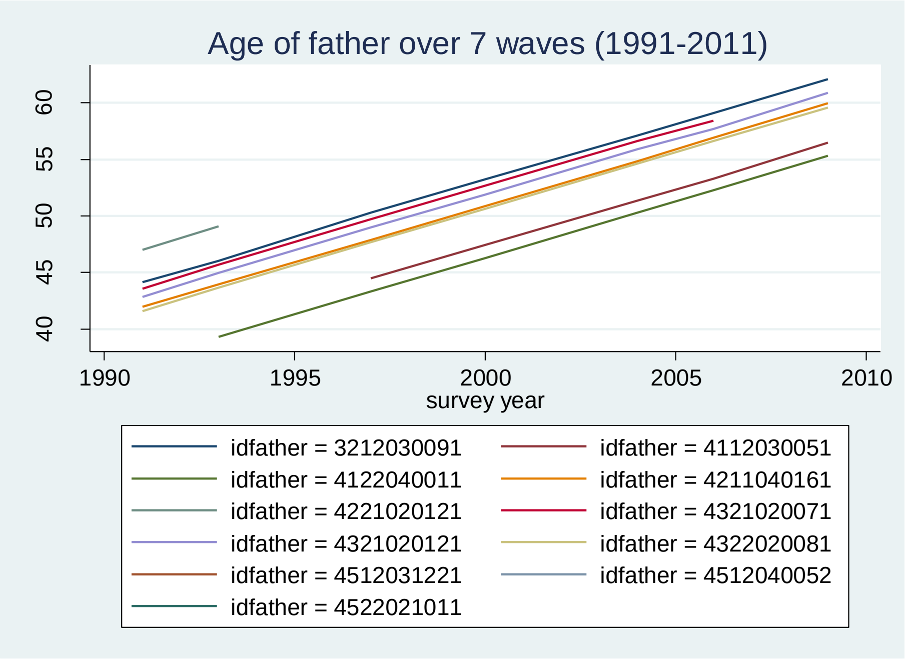
[80] We also browse these 11 households which stayed in the sample for seven waves (from 1991 to 2009), and find their age mostly increase by the interval between waves, the reason their age do not increase precisely by the interval between waves is: first, the intervals between survey waves might not be the integral number of years (i.e, might be 1.7 years---one year and eight months rather than 2 years); second, the age of father and mother are missing for several waves.

Table A 4.6: The summary statistics of the age of parents and children which were followed through over seven waves (from 1991 to 2009)

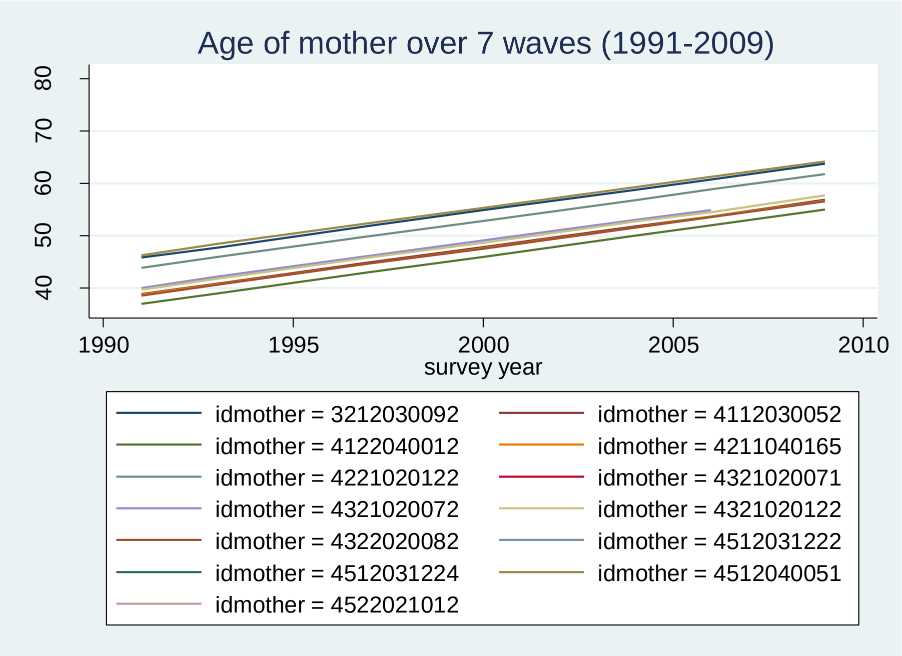
	Obs	Mean	Std. Dev.	Min	Max
1991					
Age of father	6	43.52	1.95	41.61	46.99
Age of mother	9	40.93	3.41	36.94	46.17
Age of child	11	16.49	0.28	16.04	17.03
1993					
Age of father	7	44.67	2.96	39.31	49.07
Age of mother	9	42.97	3.43	38.94	48.36
Age of child	11	18.55	0.29	18.09	19.04
1997					
Age of father	7	47.48	2.64	43.31	50.29
Age of mother	9	47.00	3.45	42.94	52.32
Age of child	11	22.59	0.30	22.14	23.15
2000					
Age of father	6	50.05	2.66	46.3	53.24
Age of mother	8	50.05	3.67	45.93	55.28
Age of child	11	25.52	0.30	25.04	26.11
2004					
Age of father	6	54.35	2.85	50.31	57.09
Age of mother	8	54.20	3.53	49.94	59.23
Age of child	11	29.49	0.27	29.14	29.96
2006					
Age of father	7	56.33	2.58	52.29	59.08
Age of mother	9	55.88	3.46	51.92	61.28
Age of child	11	31.47	0.29	31.07	31.96
2009					
Age of father	6	59.06	2.61	55.33	62.09
Age of mother	9	60.63	5.86	54.95	73.44
Age of child	11	34.49	0.26	34.03	34.95

Figure A 4.2: The age of parents and children who were observed for seven waves

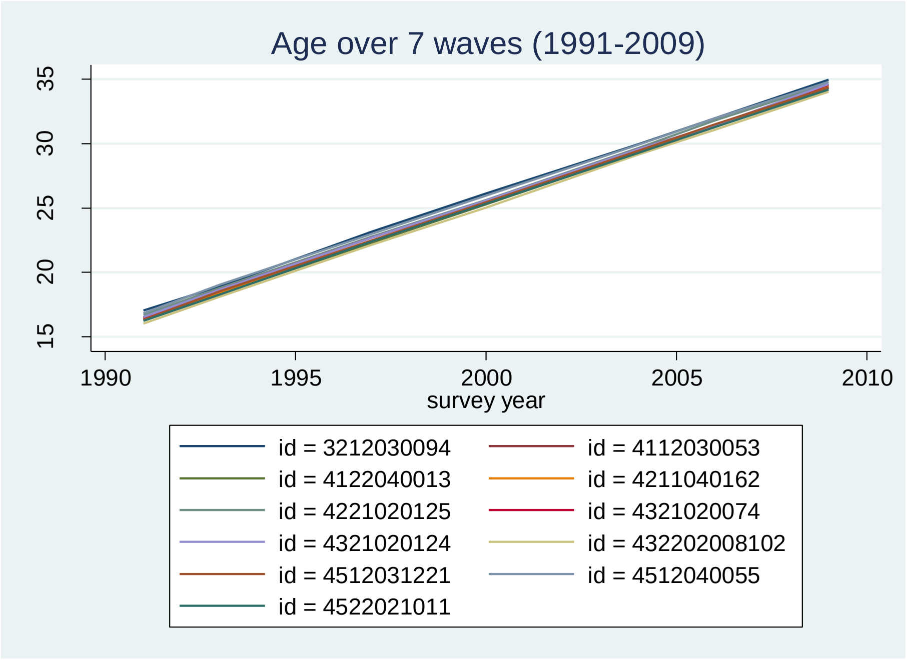
(a)



(b)



(c)



Another variable where the significant difference between Tong and Piotrowski (2012)'s sample and our sample is found is the variable for spouse's presence. Table A 4.3 suggests that there are substantial differences between Tong and Piotrowski (2012)'s sample (0.43) and our sample (0.60). Here the "spouse's presence" is a dichotomous variable which equals to 1 when the respondent has a spouse present at home (which is for the married respondents); while equals to 0 when the respondent does not have a spouse or he has a spouse but the spouse is not at home. Notice the respondents with "spouse=0" includes both non-married people (never-married, widowed, divorced, separated) and people who are married but without spouse's presence at home. Therefore, this is not a variable which is observed only on the married people [81]. Rather, this variable "spouse's presence" is defined relative to the whole sample[82]. This is the same way the variable "whether live with a spouse" is defined in Ahn et al. (2013)'s study. Using a

[81] In that case (If the variable "spouse's presence" is only for the married people), the proportion of "spouse's presence" would be around 90%.

[82] we realized this when we was looking into the "selection issue", which concerns that the households without a family member aged 16-35 years old at home might disappear from the sample. we find that the migration status of individuals aged 16-35 years old would be reported if any family member was present at home (regardless of the age of this family member).

sample of individuals aged 60 years and older in the 1997-2006 wave of the CHNS data, Ahn et al. (2013) find that over 60% of the sample “live with a spouse”, where the variable of “whether live with a spouse” is defined based on the dichotomization of the variable “marital status” (i.e. the variable of “whether live with a spouse” equals to 1 if the variable of “marital status” equals to 1; while equals to 0 if the variable of “marital status” equals to 0)[83]. In addition, informed by Chen (2012), who examined the effects of spousal absence/presence on the health trajectory, we also use the variable “husband at home or not”, which is designed for the ever-married women in the data file of “marriage history”, as a source of supplementary information to help construct the variable of “spouse’s presence”. However, we find that for some observations, the answers to the question “does spouse live at home” contradicts with those to the question “does husband live at home”. Since from the stata dofile sent by Tong and Piotrowski (2012), it seems that Tong and Piotrowski (2012) construct the variable based on the question “does spouse live at home” from the data file of “roster”, we follow them and also adopt “does spouse live at home” as the source of information to construct our variable for “spouse’s presence”. One of the problems with this source is that there is a relatively high proportion of missing values (90.54%)[84] for this variable. Therefore, we combine this variable with another variable “spouse’s line number” where the proportion of missing values is 42.61%, to construct the variable for “spouse’ presence”. As a test on the plausibility, Table 4.A.2 also presents the descriptive statistics of this variable in 1990 census, and the wave 1991, 1993 of the 2011 longitudinal data. As the left two columns of Table A 4.4 suggest, the statistics of these variables appear closer to our sample compared to Tong and Piotrowski (2012)’s sample. And the same holds for the variable “child’s residence”, where child is defined as those younger than 12 years old. Therefore, the comparison of the statistics of “spouse” variable between Tong and Piotrowski (2012)’s sample and our sample, again, lends some confidence to the credibility of our sample.

To summarize, as Table A 4.3 suggests, based on the samples composed of different number of observations, there are significant differences in the magnitude of descriptive statistics for several variables between our sample and Tong and Piotrowski (2012)’s.

[83] In Ahn et al. (2013)’s answer to our email on how they obtain the variable of “Living with a spouse (or not)”, they told me they “simply dichotomized” the variable of “marital status”, though it is not clear to me why they used the variable of “marital status” as equal to “Living with a spouse (or not)”.

[84] Given by the codebook of the China Health and Nutrition Study (CHNS) Roster File 1989-2011.

Migrant status

Thus far, we have discussed the definition of explanatory variables. In terms of the definition of outcome variable “migrant status”, we follow Tong and Piotrowski (2012)’s paper and the do files sent by the authors. we define those who change hukou status (notice this requires this “hukou” variable is not missing in the adjacent waves) and those who are absent for military, employment or other reason in the next wave as migrants; those who remain at home, or are not living in home, but in the same village/neighbourhood or the same county[85], or those who have gone to school in the next wave are defined as non-migrants, those who are dead in the next wave as missing. It is noteworthy since this variable is measured as a change in residence across waves. In other words, it depends on the next wave, therefore the migrant variable for observations which are observed only once and at the last time in the panel (the last occurrence) are coded as missing.

The comparison of estimate results

Nonetheless, based on this sample, Table A 4.7 presents the baseline estimates of the probit regression, in comparison with those from Tong and Piotrowski (2012)’s study. It suggests the health effects estimates based on our sample are close to those from Tong and Piotrowski (2012)’s sample, those self-reporting as having “excellent” and “good” health are significantly more likely to migrate than those self-evaluating as having “poor or fair” health. This result also holds in the wave 2000. Other health variables are mostly not significant, the estimates for “ADLs” in the wave 2006 in our sample suggest those with “ADLs” are more likely to migrate. In addition, those who are smokers are more likely to move in the wave 2000 in either our sample or Tong and Piotrowski (2012)’s sample. In terms of the estimates for other variables, those from our sample are also similar to Tong and Piotrowski (2012)’s estimates. For instance, age has a negative effect

[85] The definition of migrant mainly uses variable “a5e” (the reasons for migration), supplemented by “aa13” (where lives now). For those who stayed in “the same village”, “same county”, “same city”, “same province”, they are coded as migrants if “a5e”equals 3 (for military service), or 4 (sought employment elsewhere), or 6 (other); coded as non-migrants if “a5e” is missing.

on the probability of migration, indicating the respondents are less likely to migrate when they grow older. Males are more likely to be migrants in the pooled sample and in the wave 2004, 2006. The prior migration experience is significantly positively related to migration in the pooled data and most of the waves. Those who are from rural areas are more likely to migrate. Compared to the coastal region (the reference group, which includes provinces Shandong, Jiangsu and Heilongjiang), respondents from the less developed Northeast region are less likely to migrate, but those from inland region and southern mountain regions, which are also the less developed areas, are more likely to migrate, though not seen in all the waves. These are variables where there are no major differences between our sample and Tong and Piotrowski (2012)'s.

There are, however, significant differences in the estimates for some variables. For instance, the coefficient for the variable "spouse's presence" in our estimates is not statistically significant, whereas it appears significant in Tong and Piotrowski (2012)'s estimates. Our estimate indicates that respondents with spouse are no more likely to migrate in the next wave than those who do not have a spouse (including those who are never-married and separated or divorced or widowed). This difference is not surprising given the aforementioned significant differences in the proportion of individuals with "spouse's presence" between Tong and Piotrowski (2012)'s sample and our sample. Similarly, the difference in the estimates for "log (income per capita in 2006 currency" between our sample and Tong and Piotrowski (2012)'s sample might also be expected given that this variable is calculated in our sample whereas might directly existed in Tong and Piotrowski (2012)'s sample. In addition, there are some differences in the estimates for "non-farm occupation", "no parents", "child" and "the wave 2004". Our estimates suggest those with initial occupation as "non-farm" are less likely to migrate than those who are students or unemployed. The occupations of "non-farm" consist of "senior professional/technical worker", "junior professional/technical worker", "administrator/executive/manager" and "office staff". In addition, in our estimates, "with no parents' residence" appears not to have an effect on the probability of migration, whereas "having a child aged less than 12 years old at home" might reduce the propensity to migrate. Regarding the wave of the survey, our estimates show those in 2004 are more likely to migrate than those in the wave 1997, similar to 2000 and 2006.

Table A 4.7: The probit regression results (in comparison with Tong and Piotrowski (2012))

	Pooled		1997		2000		2004		2006	
	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao
Health										
Self-rated health: Poor/fair (Ref.)										
Good	0.097*	0.088*	0.14	0.075	0.13	0.072	-0.037	0.055	0.11	0.099
	(0.052)	(0.05)	(0.092)	(0.08)	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.12)
Excellent	0.20***	0.127**	0.30**	0.126	0.37***	0.244**	0.12	0.168	0.011	-0.172
	(0.066)	(0.06)	(0.12)	(0.11)	(0.13)	(0.12)	(0.14)	(0.12)	(0.15)	(0.16)
Difficulty with	0.029	0.128	-0.077	0.077	-0.079	-0.010	0.012	-0.143	0.28	0.479**
ADLs	(0.11)	(0.11)	(0.24)	(0.25)	(0.23)	(0.20)	(0.26)	(0.25)	(0.21)	(0.22)
Bone fracture	-0.16	0.100	-0.11	0.118	-0.07	0.295	-0.66	-0.282	-0.052	0.020
	(0.13)	(0.13)	(0.29)	(0.26)	(0.2)	(0.21)	(0.48)	(0.28)	(0.23)	(0.26)
Ever smoker	0.055	0.059	-0.057	0.041	0.25**	0.203**	0.14	0.076	-0.01	-0.064
	(0.05)	(0.05)	(0.083)	(0.08)	(0.11)	(0.10)	(0.11)	(0.11)	(0.11)	(0.12)
Demographic										
Age	-0.055***	-0.044***	-0.047***	-0.033***	-0.073***	-0.047***	-0.040***	-0.048***	-0.051***	-0.051***
	(0.0058)	(0.01)	(0.0099)	(0.01)	(0.013)	(0.01)	(0.014)	(0.01)	(0.013)	(0.01)
Male	0.10**	0.111**	0.11	0.081	-0.0046	0.103	0.25**	0.172	0.24**	0.261**
	(0.044)	(0.04)	(0.068)	(0.07)	(0.099)	(0.09)	(0.12)	(0.11)	(0.11)	(0.12)
Ever married	-0.14*	0.073	0.00077	0.112	-0.12	0.057	-0.37**	0.035	0.13	0.058
	(0.078)	(0.14)	(0.14)	(0.24)	(0.15)	(0.23)	(0.18)	(0.37)	(0.19)	(0.40)
Highest degree earned: Primary and lower (Ref.)										
Lower middle	0.023	-0.008	-0.00023	-0.033	-0.11	-0.034	-0.0027	-0.080	0.23*	0.242*
	(0.049)	(0.05)	(0.076)	(0.07)	(0.11)	(0.10)	(0.13)	(0.11)	(0.13)	(0.14)
Upper middle	-0.034	-0.062	-0.14	-0.236**	-0.27	0.004	0.053	-0.097	0.27*	0.317*
	(0.07)	(0.07)	(0.11)	(0.11)	(0.15)	(0.14)	(0.17)	(0.15)	(0.16)	(0.17)
Technical	-0.027	-0.135	-0.032	0.065	-0.056	-0.199	-0.019	-0.507**	0.15	0.088

	Pooled		1997		2000		2004		2006	
	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao
/vocational	(0.084)	(0.09)	(0.14)	(0.15)	(0.19)	(0.19)	(0.2)	(0.20)	(0.18)	(0.19)
College and	-0.027	0.037	0.12	-0.054	-0.22	-0.054	-0.035	-0.022	0.28	0.513**
beyond	(0.11)	(0.12)	(0.2)	(0.25)	(0.26)	(0.27)	(0.26)	(0.22)	(0.21)	(0.23)
Occupation: Unemployed or student (Ref.)										
Farmer	0.088	0.039	0.026	-0.044	0.052	0.011	0.43***	0.205*	-0.029	0.072
	(0.056)	(0.05)	(0.09)	(0.09)	(0.12)	(0.11)	(0.14)	(0.12)	(0.13)	(0.14)
Non-farm	-0.0034	-0.169**	0.04	-0.033	-0.14	-0.667***	0.0019	0.001	-0.057	-0.344*
	(0.074)	(0.09)	(0.13)	(0.14)	(0.16)	(0.19)	(0.18)	(0.17)	(0.15)	(0.18)
Skilled	0.12	0.038	0.13	-0.104	-0.12	0.082	0.063	-0.226	0.13	0.273
	(0.088)	(0.08)	(0.15)	(0.15)	(0.18)	(0.17)	(0.21)	(0.19)	(0.2)	(0.19)
Non-skilled	0.075	-0.043	0.019	-0.146	-0.3	-0.155	0.28*	-0.110	0.069	0.086
	(0.076)	(0.08)	(0.12)	(0.13)	(0.19)	(0.18)	(0.16)	(0.17)	(0.16)	(0.17)
Service	0.091	0.002	0.073	-0.051	0.056	-0.102	0.065	0.149	0.13	-0.068
	(0.069)	(0.07)	(0.12)	(0.12)	(0.16)	(0.16)	(0.15)	(0.13)	(0.13)	(0.14)
Ever migrated	0.45***	0.393***	0.74***	0.783***	0.52***	0.120	0.46***	0.390***	0.14	0.213**
	(0.059)	(0.05)	(0.12)	(0.10)	(0.14)	(0.14)	(0.11)	(0.10)	(0.12)	(0.09)
Household characteristics										
Rural	0.25***	0.382***	0.23***	0.423***	0.37**	0.433***	0.19	0.372***	0.15	0.290**
	(0.058)	(0.05)	(0.089)	(0.08)	(0.13)	(0.10)	(0.14)	(0.11)	(0.12)	(0.12)
Household size	0.025*	0.077***	0.021	0.033	0.004	0.079**	-0.043	0.077**	0.12***	0.130***
	(0.014)	(0.02)	(0.024)	(0.03)	(0.034)	(0.03)	(0.035)	(0.03)	(0.032)	(0.04)
Log (income pc)	-0.50***	-1.059	-0.26	1.831	-0.19	3.246	-1.06***	-2.681	-0.34*	-0.824
	(0.11)	(1.06)	(0.22)	(2.69)	(0.24)	(2.22)	(0.23)	(2.17)	(0.18)	(1.73)
One parent>55	0.046	0.001	-0.021	-0.049	-0.16	-0.136	0.1	0.112	0.32*	0.163
	(0.084)	(0.07)	(0.13)	(0.11)	(0.2)	(0.14)	(0.18)	(0.16)	(0.17)	(0.17)
Both parents>55	-0.11	-0.007	-0.2*	-0.030	-0.25	0.012	0.034	-0.006	0.16	0.013
	(0.077)	(0.08)	(0.12)	(0.12)	(0.18)	(0.16)	(0.17)	(0.16)	(0.17)	(0.18)

	Pooled		1997		2000		2004		2006	
	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao	Tong	Xiao
No parents	-0.20***	-0.062	-0.18	-0.259**	-0.25	0.009	-0.52***	0.013	0.26	0.079
	(0.075)	(0.07)	(0.13)	(0.12)	(0.18)	(0.14)	(0.15)	(0.15)	(0.16)	(0.17)
Spouse	-0.18**	-0.189	-0.64***	-0.422*	-0.031	-0.061	-0.01	-0.134	-0.11	0.102
	(0.072)	(0.14)	(0.14)	(0.23)	(0.14)	(0.22)	(0.16)	(0.36)	(0.15)	(0.39)
Child	0.0071	-0.149***	0.091	0.130	-0.079	-0.319***	0.095	-0.252**	-0.27*	-0.338**
	(0.071)	(0.06)	(0.14)	(0.09)	(0.14)	(0.11)	(0.16)	(0.13)	(0.14)	(0.14)
Region:										
Northeast	-0.47***	-0.291***	-1.12***	-0.244*	-1.25***	-0.485***	-	-0.404***	0.32**	0.286
	(0.09)	(0.07)	(0.17)	(0.13)	(0.26)	(0.13)	-	(0.16)	(0.16)	(0.18)
Inland	0.23***	0.199***	0.13	0.100	0.40***	0.319***	0.31**	0.177	0.30**	0.434***
	(0.056)	(0.06)	(0.088)	(0.09)	(0.11)	(0.11)	(0.13)	(0.12)	(0.14)	(0.15)
Southern	0.22***	0.213***	0.24**	0.176*	0.24*	0.182	0.15	0.271**	0.33**	0.442***
	(0.061)	(0.06)	(0.095)	(0.10)	(0.12)	(0.12)	(0.14)	(0.14)	(0.14)	(0.16)
Wave:										
2000	0.22***	0.255***	-		-		-		-	
	(0.05)	(0.05)								
2004	0.11	0.247***	-		-		-		-	
	(0.06)	(0.06)								
2006	0.39***	0.144**	-		-		-		-	
	(0.059)	(0.06)								
_cons	4.68***	11.990	2.33	-22.592	2.5	-39.258	9.97***	31.788	2.3	8.715
	(1.03)	(12.67)	(2.14)	(32.25)	(2.28)	(26.67)	(2.36)	(25.99)	(1.84)	(20.74)
Observations	7986	8528	3313	3423	1818	1956	1419	1738	1436	1411

In conclusion, starting from the appreciable differences in the descriptive statistics between our sample and Tong and Piotrowski (2012)'s, this appendix speculates on the difference in the data, discusses the way we construct the variables (including the outcome variable), and compares the estimate results. Basically we failed in the replication, our estimates for the health effects appear less significant than Tong and Piotrowski (2012)'s estimates, since we conducted various tests to testify our sample, the results of these tests lend some credibility to our estimates. Our estimates suggest that there is a positive but relatively weak evidence on health selectivity of migrants, though this effect is not very significant in a statistical sense, this might be due to the substantial heterogeneity across households and circumstances and the rather small sample we have to deal with, or the weakness of the measures we have to use. Additionally, since our sample aged between 16 and 35 years old, there might not be substantial variation in health within this age range.