DEPARTMENT OF ECONOMICS
UNIVERSITY OF CYPRUS

The absolute health income hypothesis revisited: A Semiparametric Quantile Regression Approach

Thanasis Stengos and Yiguo Sun

Discussion Paper 2005-08

P.O. Box 20537, 1678 Nicosia, CYPRUS Tel.: ++357-2-892430, Fax: ++357-2-892432
Web site: http://www.econ.ucy.ac.cy

The absolute health income hypothesis revisited: A Semiparametric Quantile Regression Approach*

Thanasis Stengos and Yiguo Sun

Department of Economics, University of Guelph, Guelph, ON, Canada N1G 2W1

October 6, 2005

Abstract

This paper uses the 1998-99 Canadian National Population Health Survey (NPHS) data to examine the health-income relationship that underlies the absolute income hypothesis. To allow for nonlinearity and data heterogeneity, we use a partially linear semiparametric quantile regression model. Among more than dozen of socioeconomic variables, we find that family income, age and the food security status are the most important factors in explaining an individual's overall functional health. The "absolute income hypothesis" is partially true; the negative aging effects appear more pronounced for the ill-healthy population than for the healthy population and when annual income is below 40,000 Canadian dollars.

JEL Classification: C14, C51, I12

---

*Email:yisun@uoguelph.ca and tstengos@uoguelph.ca. The views expressed in this article are those of the authors and do not necessarily reflect the views of Statistics Canada. Both authors acknowledge financial support from the SSHRC of Canada.

1 Introduction

The recent vigorous debate on the role of public health policies and their funding have motivated a good deal of research on the impact that these policies and programs may have on the equal provision of health care to all population groups. Following the seminal paper by Grossman (1972), health is viewed as a durable good that depreciates with age and produces as an output health time. In this case the "shadow price" of health depends on many other factors besides the price of medical care. It is expected that the shadow price rises with age as the stock of health depreciates over the life cycle and decreases with education as more educated people are expected to be more efficient producers of health. Note that the above approach makes an explicit distinction between "medical services" and health. What consumers are after is good health and that frequently is confused as "medical care". Yet, demand for the latter can only be studied properly if there is a model that describes the demand of the former. Given that traditional demand theory takes for granted that goods and services purchased in the market enter the consumers' utility functions, the demand for medical care (which is a directly observable market activity) has been analyzed a lot more extensively at the expense of health as a durable good that produces health time, that is the activity that consumers are really after.

Following the prediction from Grossman's (1972) model that wage income affects health, there have been some important hypotheses that have emerged in this context. Since the health production function has been specified as a function of own variables alone, relative position does not matter. This has given rise to the so called "absolute income hypothesis" that emphasizes that it is income level that matters for health, not income relative to other people's income, nor income inequality. A name that would be at least as good is the "poverty hypothesis", that ill-health is a consequence of low income, in

the sense that more income improves health by more among those with low incomes than among those with high incomes. However, there is also the argument that individuals care about their relative position and status and as such relative position variables need to be included in models of health production. This would give rise to the so called "relative income hypothesis" which implies that health depends not on absolute income, but relative income, that is income inequality affects health. If the "absolute income hypothesis" is correct then policies of income growth will be sufficient to reduce health inequality assuming that the relationship between income and health is concave. In that case, increasing in income would increase health at a decreasing rate, see Contoyannis and Forster (1999). Alternatively, if the "relative income hypothesis" is the empirically relevant hypothesis, then tax policies aiming to reduce income inequalities are of more relevant in reducing health inequality. In this paper, we will concentrate on the validity of the former hypothesis as we explore the relationship between income and health. An important issue in this literature is the shape of the income health relationship, as many studies seem to assume it to be linear even though the evidence for linearity is not strong, see Wagstaff and van Doorslaer (2000).

This paper provides information on how income affects population health status. The empirical results indicate that the "absolute income hypothesis" holds true up to a point before one takes into account age effects, and that the estimated quantile health production function is not globally concave in family income. Two policy-related proposals are given: one is that more effort should be spent to improve the "minimum living conditions" of the whole population; and the other is that more effort should be spent to improve the health status of the young, something that will positively affect their health later on in life and slow down the speed of negative age effects. Consequently, the health status of the whole population will improve over time. In other words, the income effect seems to be stronger for the less healthy segment of

the population than for the more healthy one.

The empirical analysis is based on the 1998-99 Canadian National Population Health Survey (NPHS) data surveyed by Statistic Canada and hosted by the Research Data Centre in the University of Waterloo. Instead of studying the self-assessed health status, we use a continuous health utility index to measure an individual's overall functional health which has an upper bound of one.[1] Looking at the data, we find that nearly 30% of the respondents' health utility index attains this upper bound. Therefore, the dependent variable is censored and a simple linear regression specification will not be applicable. In addition, for the individual microdata, homogeneity may not be a reasonable assumption. The main contribution of our paper is to use a conditional semiparametric quantile health regression model to incorporate data heterogeneity and also handle censoring. In the literature of health economics, both parametric and nonparametric mean regression models of health have been commonly used to analyze the relationship between socio-economic variables including income and health, see Wildman and Jones (2003) for a recent nonparametric mean regression application. Such mean regression model can be used to predict the average response of health to the changes of relevant explanatory variables. However in the presence of heterogeneity, the mean regression model does not provide enough information in predicting how the health status of the whole population will change if systematic changes in socio-economic variables (family income, for instance) occur with a newly proposed health policy. Hence, after detecting the existence of heterogeneity, this paper studies a partially linear quantile health model allowing for both family income and respondents' age to be treated nonparametrically. As a

[1]The overall functional health is not the same as the health we refer to in our everyday life, although the two are closely related, see the World Health Organization (1958) for the importance of functional ability. In this paper, health is always referred to as the overall functional health.

result, by estimating quantile regressions based on a partially linear quantile regression approach at different probability masses, we are able to derive information other than the average predictions resulting from the estimation of the conditional mean regression model.

The rest of the paper is organized as follows: Section 2 describes the data, defines the health production models, and gives the test results on linearity assumption. Section 3 presents the proposed partially linear quantile regression model and explains how to calculate its semiparametric efficient estimator. Section 4 gives our empirical results and the discussion on the "absolute income hypothesis". Section 5 concludes. In the appendix we present Zheng's (1998) test for linearity and we provide a detailed description of the estimation procedure that we use.

2 Data and the health-income relationship

2.1 Data

This paper analyzes the 1998-99 National Population Health Survey (NPHS) data undertaken by Statistic Canada. Statistic Canada launched the first National Population Health survey in 1994. At the start of this project, only three cycles had been completed: the NPHS Cycle 1 (1994-95), NPHS Cycle 2 (1996-97) and NPHS Cycle 3 (1998-99). Each cycle data includes two files: the general file and the health file. The general file collects household information. Randomly choosing one individual from each interviewed household in the general file, Statistics Canada interviewed this individual on his/her health related information in detail. These records are contained in the health file. Detailed information about the NPHS content has been published elsewhere, see Tambay and Catlin (1995).

The superiority of the 1998-99 cycle data over the two previous surveys

lies in its records on family income—it contains the best estimated family income before tax, which can be considered to be continuously distributed; while the 1994-95 and 1996-97 cycles of the NPHS only contain the categorized family income, which is a discretely distributed variable. In the 1998-99 survey, there are approximately 49,046 households who answered the general portion of the questionnaire, while 17,244 respondents answered the more detailed health portion. Among dozens of health-related factors, the variables of interest in this paper are the health utility index, family wage income,[2] food security status, and the personal information including age, gender, highest education level, living arrangement, and insurance policies in prescription medication and hospital charges, where the food security status and insurance policies are treated as the complements of family income in explaining the respondent's health status and the respondent's highest education level is a proxy of his/her knowledge in health.[3]

The health utility index, h, is an index used to measure the health status of the respondents aged 4 and over. This health utility index, developed at McMaster University's Centre for Health Economics and Policy Analysis, is based on the Comprehensive Health Status Measurement System (CHSMS). It describes an individual's overall functional health, based on eight attributes: vision, hearing, speech, mobility, dexterity (use of hands and fingers), cognition (memory and thinking), emotion, and pain and discomfort. For a detailed explanation of the calculation of the HUI3, see Furlong, Feeny, Torrance (1999) and references therein. In the 1998-99 survey, the index is valued between -0.34 and 1.0, where negative scores reflect

[2]We prefer to use family wage income instead of the equivalent income adjusted with respect to the number of adults and kids, since the validity of equivalence scale and base independence of health expenditure has not been explored.

[3]Since the study on the interaction effect of income adequacy and health behavior on the health of Canadians is inconclusive, see Williamson (2000) and references therein, we did not include variables measuring health behavior in this paper.

health status considered to be worse than death, while nearly 30% of the data has the health utility index being one, which refers to perfect functional health. This of course introduces censoring that is handled through the quantile regression approach that we use. Except for the health utility index, family wage income, and age, the rest of variables are 0-1 dummies in nature. Table 1 presents a detailed description of all the variables used in the empirical analysis.

2.2 The health-income relationship

After removing all the incomplete records and subgroups with less than 40 observations, we end up with 10,018 data observations, which contains 71 different subgroups defined by the thirteen-dummy variables. Following the definitions in Table 1, throughout this paper, we denote

$$X = (INC, AGE, SEX, FFLAG, LVG, ED, ISC), \tag{1}$$

where $LVG = (LVG_1, \cdots, LVG_6)$, $ED = (ED_1, ED_2, ED_3)$ and $ISC = (ISC_1, ISC_2)$.

We treat all the explanatory variables in X as inputs and the health utility index h as output.[4] The function linking health to various socio-economic variables including income is defined to be $h = \min(H(x), 1)$ and the health

[4]The reason that we treat wage income as exogenous is as follows. The health utility index describes an individual's overall functional health and as such it is a proxy for health time. In the case of the stock of health capital, the causation is from wage income to health and not vice-versa, see Grossman (1972). In that case even individuals who are not in the labor force have an incentive to invest in their health and therefore health is not a determinant of the wage rate as it is the case for other forms of human capital where investments in education and on-the-job training raise wage rates.

production regression model becomes

$$h_i^* = H(X_i) + \varepsilon_i, \tag{2}$$

$$h_i = \min(h_i^*, 1) \tag{3}$$

where $E(\varepsilon_i|X_i = x_i) = 0$, $E(\varepsilon_i^2|X_i = x_i) = \sigma_0^2(x_i) < M < \infty$, $i = 1, \cdots, n$.

Such mean regression relationship has been analyzed in the literature using a variety of parametric and nonparametric approaches, see Wildman and Jones (2003). However, as mentioned earlier, nearly 30% of the respondents have their health utility index attain the upper bound of the index; the dependent variable is a censored variable in nature. In such a situation, there are two general ways of analyzing the data: (a) estimating a probit or logistic model under a parametric setup or estimating a semiparametric censored regression model, see Newey, Powell and Walker (1990) and (b) estimating a censored regression model using conditional quantile regression techniques. In this paper, we use the second method since the conditional quantile regression approach provides a natural way of handling the presence of heterogeneity commonly found in microdata, whereas the censored regression approaches do not.

Denote $E_\alpha(h|X = x)$ to be the α-conditional quantile of h given $X = x$, that is

$$\Pr\{h \leq E_\alpha(h|X = x) | X = x\} = \alpha, \alpha \in (0, 1) \tag{4}$$

holds for all x on its domain $\mathcal{X}$. Suppose $\varepsilon_i = \sigma_0(x_i) u_i$, then rewriting (2) yields

$$h_i^* = H(X_i) + \sigma_0(X_i) E_\alpha(u_i) + \sigma_0(X_i)(u_i - E_\alpha(u_i)), i = 1, \cdots, n. \tag{5}$$

Denote $V_i = \sigma_0(X_i)(u_i - E_\alpha(u_i))$, then $E_\alpha(V_i|X_i = x_i) = 0$ and

$$E_\alpha(h_i^*|X_i = x_i) = H(x_i) + \sigma_0(x_i) E_\alpha(u_i), \alpha \in (0, 1), \tag{6}$$

for all i. Therefore,

$$E_\alpha\left(h_i|X_i=x_i\right)=\min\left\{E_\alpha\left(h_i^*|X_i=x_i\right),1\right\} \tag{7}$$

Hence, we aim to estimate the above unknown conditional quantile curves at different probability masses α. For such cases, a two-step semiparametric estimator is developed by Khan and Powell (2001). However, we notice that there is a $\alpha^*\in(0,1)$ such that $E_\alpha\left(h_i^*|X_i=x_i\right)<1$ and $E_\alpha\left(h_i|X_i=x_i\right)=E_\alpha\left(h_i^*|X_i=x_i\right)$—the upper bound is not restrictive and the censor is not involved; in our case, α^* is around 0.71.[5] Therefore, we are able to reduce the above estimation to the case of the usual quantile regression models when the probability mass of interest is less than 0.70. It will simplify the estimation procedure.

It is not easy to interpret a pure nonparametric curve defined in many dimensions; therefore, a well-specified parametric model would be preferable if it would pass standard model specification tests. Otherwise, a semiparametric model may be more useful than a pure nonparametric model. Hence, our first attempt will be to use linear quantile health regression models which hold if $H\left(\cdot\right)$ and $\sigma_0\left(\cdot\right)$ are both expressed in linear forms. The consistent, nonparametric test statistic of Zheng (1998) is used to test for the validity of linearity; this is a residual-based statistic using kernel nonparametric estimation, see the appendix for details.

Under the null hypothesis, $E_\alpha\left(h_i|X_i\right)=X_i\beta_\alpha$ is a linear parametric quantile regression model, see Koenker and Bassett (1978). Zheng's test is applied for a range of quantile models at the probability masses $\alpha_j=0.05+0.01\left(j-1\right),\ j=1,\cdots,66$, i.e. $\alpha_j\in[0.05,0.70]$. The empirical results indicate that linear quantile health regression models are rejected for

[5]We apply Hall, Wolff and Yao's (1990) method to estimate the conditional probability $F\left(h_i|x_i\right)$ for each $i=1,2,\cdots,n$. Since the conditional probabilities for individuals with $h_i=1$ are one and those with $h_i<1$ give values of less than one, to be safe, we choose α^* being the fifth largest conditional probability mass.

$\alpha_j \geqslant 0.15$ at the significance level of 5% and they are invariant to the choice of the parameter c of the smoothing matrix.

The above test results indicate that the linear assumption imposed on $H(\cdot)$ and $\sigma_0(\cdot)$ is a good approximation in predicting the health status of individuals which lies at the left tail of the conditional distribution of health status when $\alpha < 0.15$. Then the nonlinearity of quantile regression models at higher probability mass may stem from the nonlinearity of $\sigma_0(x)$. Therefore, we propose a partially linear quantile regression model in the next section and explain how to estimate the model efficiently when $\alpha \geq 0.15$.

2.3 The partially linear quantile specification

The partially linear quantile regression model is based on the assumption that log-income and age enter nonparametrically, but the other discrete variables enter linearly. As a result, it is given by

$$h_i = W_i\beta_{\alpha,0} + g_\alpha(Z_i) + V_i,\ E_\alpha(V_i|W_i, Z_i) = 0, \tag{8}$$

where $Z_i = (X_{1i}, X_{2i}) = (INC_i, AGE_i)$, and $W_i = (X_{3i}, \cdots, X_{15,i})$. For the sake of model identification, the constant term will be absorbed into the unknown function $g_\alpha(\cdot)$.

If $H(\cdot)$ is linear, i.e.

$$H(x) = \beta_0 + \sum_{k=1}^{15} x_{ki}\beta_k, \tag{9}$$

then the above model assumes that $\sigma_0(X) = W'\gamma_0 + \theta_0(Z)$ is nonlinear in both log-income and age. Comparing (9) with (8), we have $\beta_{\alpha,0} = (\beta_3, \cdots, \beta_{15})' + \gamma_0 E_\alpha(u)$, and

$$g_\alpha(Z) = \beta_0 + Z_1\beta_1 + Z_2\beta_2 + \theta_0(Z) E_\alpha(u). \tag{10}$$

Suppose that the conditional probability density function of V is $f(v|w, z)$. If $f(0|w, z) > 0$, and $F(0|w, z) = \alpha$ for all $(w, z) \in \mathcal{X}$, then under certain

regular conditions, the semiparametric efficient,$\sqrt{n}$-asymptotically normally distributed estimator of $\boldsymbol{\beta}_{\alpha,0}$ is given by Lee (2003) and Sun (2005). In the appendix we present the estimation procedure of $\boldsymbol{\beta}_{\alpha,0}$ for a given $\alpha \in (0,1)$.

3 Empirical results

The health utility index as the dependent variable measures a respondent's overall functional health. Among the eight attributes of the index, four of them, such as vision, hearing, mobility and cognition, are unavoidably deteriorating with the increase of his/her age, the so-called negative aging effects. However, in reality we do observe that the aging process is much slower for some people than others, which may result from genetic differences and different life experiences. The latter may relate to the wealth of a family, the family type or living arrangements, and education level. For example, the higher the family income, the better life an individual can enjoy materially, which implies a positive income effect. On the other hand, the higher income may be associated with a job with more responsibility and more pressure, which implies a negative income effect. In this sense, we will say the income effect may be ambiguous and a strictly positive income effect will not be expected. Living arrangements may also have ambiguous effects on an individual's health, since family can bring about not only emotional comfort but also additional pressures.

To offer empirical answers to the questions above we estimated the conditional quantile health income relationship curves at probability masses between 0.05 and 0.70. Based on our test results, we fit our data with the linear quantile regression models if the probability mass α is 0.05 or 0.10; otherwise, with the partially linear quantile regression models. In what follows, we will illustrate our estimation results in Figures 1-3.

Figure 1 plots the estimated parameters of the thirteen dummy variables

and their lower and upper 95% confidence intervals. For the sake of comparison, we also include the estimates from the OLS method (dotted lines) and from the linear quantile regression models (asterisks for the estimations and cross signs stand for the respective 95% confidence intervals). The estimated coefficient in front of SEX is not shown here since it is not statistically different from zero at the significance level of 5% for all cases, in contrast to what is often asserted that women live longer than men but suffer more illness through their lives.[6]

For the other dummy variables, we have the following observations.

(a) The estimated coefficients of the food security status are relatively stable and take values around 0.20 except for $\alpha = 0.15$ and 0.70. The contribution from this variable is bigger than what the OLS estimator predicts from the linear mean regression model. The linear quantile regression models do not provide a plausible explanation especially for higher conditional quantiles when α exceeds 0.42, where log-income, age and food security status all lose their significance in explaining the respondents' overall health status. Therefore, both the OLS method and the linear quantile regression models will understate the role played by food security status, while the proposed quantile models correctly recognize that having enough food is a necessary condition in maintaining the respondents' good health status. This is consistent with results of Vozoris and Tarasuk (2003), who analyze the 1996-97 NPHS data and found that individuals from households with insufficient food intake had significantly higher odds of reporting poor/fair health.

(b) For the effects of living arrangement, except for $\alpha = 0.15$, we find that the coefficient of LVG_5 (the children living with a single parent) is not

[6]For example, examing the 1996-97 NPHS data within a framework of a parametric logistic model, Rosengerg and Wilson (2000) found that men are less likely to report a chronic condition. The different conclusion reached here may be attribuatble to the different definitions of health status.

statistically different from zero at the significance level of 5%, and that the coefficients of LVG_1-LVG_3 (single adults and adults living with a spouse or partner with or without children) are statistically positive for $\alpha < 0.45$ although the values are very small—below 0.1 in general. Roughly speaking, among ill-healthy respondents, single parents and children living with single parents find themselves at slightly weak positions.

(c) For the effects of education to the respondents' overall functional health, all three approaches show that those with BA degree and higher perform slightly better than those with a lower degree. However, we notice that the OLS estimator underestimates the positive effects of higher education among the worst ill-healthy population ($\alpha \leq 0.2$).

(d) For the effects of insurance policies, the OLS estimates understate both the negative effect of insurance for prescription drugs and the positive effect of insurance for hospital charges when $\alpha = 0.05$ and 0.10. Holding these two types of insurance policies has no impact in explaining the respondents' overall functional health at the significance level of 5% for all other segments. That is, having insurance policies matters most to that segment of the population with the worst health, and almost has no effect on the relatively healthier population. We notice that the linear quantile model overestimates the effects of these insurance policies relative to the proposed partially additive model.

Next, we are going to explain the effects of family income and age on health. Firstly, we present the joint roles played by income and age. In Figure 2 we plot the three-dimensional surface curve $\widehat{g}_\alpha(inc, age)$ with respect to (inc, age) for $\alpha = 0.25$ and 0.56. The graphs may be under-smoothed, and they appear not to be globally concave in log-income and age. Four features are observed. (a) It seems that the linear plane is not a bad approximation except for observations with relatively low and high income values and those with age beyond fifty five. (b) Strong positive income effects are identified

among respondents with family income greater than $100,000 per year across all ages. (c) The negative age effects are more pronounced for the respondents with low family income and age around sixty and beyond eighty. (d) $\widehat{g}_{0.25}(inc, age)$ is more variant than $\widehat{g}_{0.56}(inc, age)$ where the standard deviation of the former is 0.142 and 0.076 for the latter; in other words, age and income are more important factors affecting the respondent's health among the ill-healthy segment of the population than the heathier population.

Secondly, in order to separate income effects from age effects, in Figure 3 we plot the predicted health index against income (or age) at the probability masses 0.2, 0.25, 0.30, 0.56, and 0.65, where only income (or age) is allowed to vary across observations, while the rest of explanatory variables are fixed. In particular, we choose the observations from the male adult respondents living with a spouse or partner without children who do not worry about obtaining nutritious food, and have post secondary diploma and both types of insurance policies.[7]

Consider the income effect first. Six graphs are plotted with respect to six different age groups: 27, 33, 42, 56, 65, and 69. The quantile health production functions are concave in income except for the group aged 56. Roughly speaking, among the healthy population, the higher the family income, the higher the predicted health utility index; among the ill-healthy population, the income effect is ambiguous as argued earlier. The group of respondents aged 33 is the most well-off among the others and the negative age effect begins with age beyond 60 (see the ranges of the estimated health utility index for different age groups at the same probability mass).

We next consider the age effect. The predicted quantile health production functions are downward sloping and concave in the direction of age given family income for almost all cases. The sharp reduction in health status is more prominent among respondents with family income below 40,000 Cana-

[7]More such graphs can be obtained from the authors on request.

dian dollars per year. If age plays a bigger role at the 0.65-quantile curve, then the fact that the quantile curves at lower probability mass have steeper downward trend than that at the probability 0.65 implies even faster aging effects among the unhealthy segment of the population. When the family income is above 100,000 Canadian dollars a year, the speed of the age effect is much slower than in the other cases. In addition, the speed of the age effect is also significantly slower among the healthy population ($\alpha = 0.56$ and 0.65) than the unhealthy population ($\alpha = 0.2$).

To sum up, low family income, older age, and ill health together are associated with the most disadvantaged health status. The fact that the negative aging effect is more significant among the unhealthy segment of the population suggests that efforts should be placed to improve an individual's health when he/she is young. This will slow the speed of negative aging effects in the future and as such the health status of the whole population will improve with time.

3.1 Implications for the absolute income hypothesis

The absolute income hypothesis assumes a concave relationship between individual income and health, see Rogers (1979). That will be the result of individuals assigning declining marginal utility to additional units of health or the presence of diminishing returns in the production of health with respect to income (or health inputs purchased by income). The concavity of the health production function means that a dollar transferred from a rich person to a poorer person raises average health, holding average income constant. In that case policies of income growth will be sufficient to reduce health inequality.

The "absolute income hypothesis" implies that the rich have better health than the poor; this emphasizes the important positive contribution of income on individuals' health. However, as we know, an individual's health is

a result of many respects of his/her life other than income. Besides family income, the commonly considered socioeconomic variables include education level, health behavior (exercise, smoking, alcohol consumption, sleep, etc., see Williamson 2000), and gender, see Rosenberg and Wilson (2000). Cairney and Arnold (1996) look at socio-economic determinants of self-assessed health and morbidity in elderly non-institutionalized Canadians. They demonstrate a strong inverse relationship between income and self-assessed health. Using the 1994 Canadian National Population Health Survey data, Humphries and van Doorslaer (2000) consider the income related inequality in self-assessed health by means of concentration indices. They find that significant inequalities in self-reported ill-health exist and favor the higher income groups– the higher the level of income, the better the level of self-assessed health.

Figure 2 shows that a positive income effect holds for the ill-healthy segment of the population. It indicates that low family income affects negatively the health of the seriously unhealthy population especially the elderly. Money is more valuable for the unhealthy population before age becomes important. In addition, among the unhealthy population, two significant leaps in income effects are observed: one occurs when family income per year increases to \$20,000 - \$30,000 from below \$10,000; and the other occurs when family income per year is more than \$100,000. The general impression of Figure 2 is that an individual's health does not increase monotonically with family income. In this sense, the absolute income hypothesis is only partially true — family income and age have to be taken into account simultaneously.

4 Conclusion

In this paper we consider the health-income relationship that underlies the absolute income hypothesis using the 1998-99 Canadian National Population Health Survey data. Having tested a linear specification using a test pro-

posed by Zheng (1998) we fit our data with a partially linear semiparametric quantile regression model at probability masses above 0.15. Our empirical results indicate that (a) income and age do not enter into the quantile health production models with global concavity; (b) the negative aging effects are stronger among the relatively unhealthy population aged beyond sixty with family income below $40,000 per year; (c) the positive income effects are observed among the healthy population, and the mixed income effects are identified among the ill-healthy population; (d) gender itself may not be indicative of an individual's overall functional health status; (e) having enough variety of food is a very important factor in maintaining an individual's good health; (f) the effects of living arrangements are identified only among the unhealthy population; (g) higher education and holding insurance policies on prescription drugs and hospital charges have limited effects among the unhealthy population.

Using the fact that family income, age, and with or without enough nutritious food are the most important factors affecting population's health status, we suggest that more effort shall be spent educating people to have a healthy diet and improving an individual's health when he/she is young. This will slow the speed of negative aging effects in the future and as a result the health status of the whole population will improve over time, since the health of the healthy segment of the population declines with age at a slower pace than that of the unhealthy one.

The empirical results indicate that the "absolute income hypothesis" holds true only partially, since family income and age would have to be taken into account simultaneously. Two policy-related proposals can be drawn from this analysis: one is that more effort should be spent to improve the "minimum living conditions" of the whole population; and the other is that more effort should be spent to improve the health status of the young, something that will positively affect their health later on in life and slow down the speed

of negative aging effects.

5 Appendix

5.1 A test for functional form

As mentioned in Section 2.2, we need to test for nonlinearities in $H(\cdot)$ by using Zheng's (1998) statistic. Suppose $\{(h_i, X_i)\}_{i=1}^n$ are a sequence of i.i.d. samples from a common distribution $F(h, x)$, where $(h_i, X_i) \in R \times \mathcal{X}$, and $\mathcal{X}$ is the domain of X. The null and alternative hypotheses to be tested are

$$H_0 \ : \ \Pr\left\{g_\alpha(x) = \alpha_{\alpha,0} + x'\beta_{\alpha,0}\right\} = 1 \text{ for some } \left(\alpha_{\alpha,0}, \beta_{\alpha,0}\right) \in R^{16} \text{ and all } x \in \mathcal{X},$$

$$H_1 \ : \ \Pr\left\{g_\alpha(x) \neq \alpha_{\alpha,0} + x'\beta_{\alpha,0}\right\} < 1 \text{ for any } \left(\alpha_{\alpha,0}, \beta_{\alpha,0}\right) \in R^{16} \text{ and any } x \in \mathcal{X},$$

where $g_\alpha(x) = E_\alpha(h|X = x)$ to be the $\alpha-$conditional quantile of h given $X = x$. Zheng's test is defined as $J_n = \sqrt{\frac{n-1}{n}} n\sqrt{|B|} V_n \Big/ \sqrt{\widehat{\Sigma}}$ with

$$\begin{aligned} V_n &= \frac{1}{n(n-1)|B|} \sum_{i=1}^n \sum_{j=1, j\neq i}^n K\left(B^{-1}(X_j - X_i)\right) \widehat{u}_i \widehat{u}_j, \\ \widehat{\Sigma} &= \frac{2\alpha^2(1-\alpha)^2}{n(n-1)|B|} \sum_{i=1}^n \sum_{j=1, j\neq i}^n K^2\left(B^{-1}(X_j - X_i)\right), \end{aligned} \tag{11}$$

where $\{\widehat{u}_i\}_{i=1}^n$ are the estimated residuals under the null hypothesis $\widehat{u}_i = I\left(h_i \leq X_i\widehat{\beta}_\alpha\right) - \alpha$, $B = diag(b_1, b_2)$ is a 2×2 diagonal matrix containing the bandwidths, and $|B| = b_1 b_2$. Among the fifteen-explanatory variables, only income is a continuous variable by nature. Age is treated as a continuous variable since the range of age is so wide that its distribution is quite close to the normal distribution. The rest of the variables are discrete. Therefore, the product kernel function $K(\cdot)$ at point $u = (u_1, u_2, \cdots, u_{15})' \in R^{15}$ is defined as

$$K(u) = k_1(u_1) k_1(u_2) \prod_{k=3}^{15} k_2(u_k), \tag{12}$$

where the Epanechnikov kernel, $k_1(t) = \frac{3}{4}(1-t^2) I(|t| \leq 1)$, is used for log-income and age, and $k_2(t) = I(t=0)$ is defined for the discretely distributed variables. Since the data ranges of INC and AGE are so different, we decide to use two different bandwidths here, i.e. $B = diag(b_1, b_2)$ is a 2×2 diagonal matrix containing the bandwidths with $|B| = b_1 b_2$; where $b_1 = c\widehat{\sigma}_{inc} n^{-1/6}$ and $b_2 = c\widehat{\sigma}_{age} n^{-1/6}$ are the respective bandwidths corresponding to log-income and age, and $\widehat{\sigma}_{inc}$ and $\widehat{\sigma}_{age}$ are their respective standard errors. $c = 0.8, 1, 1.1$ are used to measure how sensitive the test statistic is to the choice of the smoothing matrix B. Under the null hypothesis, $\widehat{\beta}_{\alpha}$ is estimated as suggested by Koenker and Bassett (1978) and the algorithm belongs to Koenker and d'Orey (1987).

5.2 Estimating the parameters of the Quantile Partially Linear Regression Model

Section 2.3. defines the following partially linear quantile regression model

$$h_i = W_i \beta_{\alpha,0} + g_\alpha(Z_i) + V_i, \; E_\alpha(V_i | W_i, Z_i) = 0, \tag{13}$$

where $Z_i = (X_{1i}, X_{2i}) = (INC_i, AGE_i)$, and $W_i = (X_{3i}, \cdots, X_{15,i})$. For the sake of model identification, the constant term will be absorbed into the unknown function $g_\alpha(\cdot)$. Under certain regular conditions, Lee (2003) and Sun (2005) develope the semiparametric efficient,$\sqrt{n}$-asymptotically normally distributed estimator for $\boldsymbol{\beta}_{\alpha,0}$. Below we present the estimation procedure of $\boldsymbol{\beta}_{\alpha,0}$ for a given $\alpha \in (0,1)$.

Step 1. At each point i, solve for $\widehat{\beta}_{\alpha,i}$ and $\widehat{g}_\alpha(z_i)$ by minimizing the following objective function over γ and β

$$\widehat{Q}_n(\beta, \gamma) = \sum_{j=1}^{n} \rho_\alpha(h_j - W_j \beta - \gamma_0 - (Z_j - z_i)\gamma_1) \widetilde{K}^{(-i)}\left(B^{-1}(Z_j - z_i)\right), \tag{14}$$

where $\widetilde{K}^{(-i)}(\cdot) = 0$ if $Z_j = z_i$ (the leave-one-out technique), and the check function $\rho_\alpha(u) = u(\alpha - I(u < 0))$. Then the solution of the above minimization problem, $\widehat{\boldsymbol{\beta}}_{\alpha,i}$, converges to $\boldsymbol{\beta}_{\alpha,0}$ at the usual nonparametric convergence rate under certain conditions, see Sun (2005). The kernel $\widetilde{K}(u_1, u_2) = k_1(u_1) k_1(u_2)$ is the product of two Epanechnikov kernels.

Step 2. Calculate the quantile-based quantile (QQR) estimator

$$\widehat{\beta}_{\alpha,QQR} = \arg\min_{\beta} \widehat{Q}_n(\beta, a) = \sum_{j=1}^{n} \rho_\alpha(h_j - \widehat{g}_\alpha(Z_i) - W_j\beta), \tag{15}$$

where $\widehat{g}_\alpha(Z_i)$ is the leave-one-out estimation of $g_\alpha(Z_i)$ obtained from Step 1.

Step 3. Given a$\sqrt{n}$-consistent estimator $\widehat{\beta}_\alpha$ of β_α, Lee (2003) shows that

$$\widehat{\beta}_\alpha^* = \widehat{\beta}_\alpha + \left[\sum_{i=1}^{n} \partial \widehat{S}_\alpha\left(\widehat{\beta}_\alpha\right) \Big/ \partial \beta_\alpha\right]^{-1} \sum_{i=1}^{n} \widehat{S}_\alpha\left(\widehat{\beta}_\alpha\right) \tag{16}$$

is the semiparametric efficient estimator of β_α such that

$$\sqrt{n}\left(\widehat{\beta}_\alpha^* - \beta_\alpha\right) \overset{n\to\infty}{\to} N(0, V), \tag{17}$$

where

$$\widehat{S}_\alpha(\beta_\alpha) = \frac{\widehat{f}_v(0|W,Z)}{\alpha(1-\alpha)}\left[\alpha - 1 + J\left(\frac{Y - \widehat{g}_a(Z) - W\widehat{\beta}_\alpha}{j_n}\right)\right]\left[W - \widehat{T}(Z)\right], \tag{18}$$

$\widehat{f}_v(0|W,Z)$ and $\widehat{T}(Z)$ are the kernel estimates of $f_v(0|W,Z)$ and

$$T(Z) = E\left[f_v^2(0|W,Z) W|Z\right] / E\left[f_v^2(0|W,Z)|Z\right], \tag{19}$$

respectively. The asymptotic variance is estimated by

$$\widehat{V} = \frac{\alpha(1-\alpha)}{n}\left[\sum_{i=1}^{n} \partial \widehat{S}_\alpha\left(\widehat{\boldsymbol{\beta}}_\alpha\right) \Big/ \partial \boldsymbol{\beta}_\alpha\right]^{-1}. \tag{20}$$

The trimming function $J(x)$ is defined as

$$J(x) = \begin{cases} 0 & if \quad v < -1 \\ 0.5 + \frac{15}{16}\left(x - \frac{2}{3}x^3 + \frac{1}{5}x^5\right) & if \quad |v| \leq 1 \\ 1 & if \quad v > 1 \end{cases}$$

and the trimming parameter $j_n \to 0$ as $n \to \infty$. Therefore, given $\widehat{\beta}_{\alpha,QQR}$, we can calculate the efficient estimators $\widehat{\beta}^*_{\alpha,QQR}$.

The first step is a pure nonparametric estimation procedure.[8] Related work in this area includes the kernel and the k-nearest neighbor estimator of Bhattacharya and Gangopadtyay (1990), spline smoothing estimator of Koenker, Ng, and Portnoy (1994), the local linear regression approach of Fan, Hu, and Truong (1994), and the double kernel method of Yu and Jones (1998). We choose to use the local linear regression approach here because the bias of this estimator is adaptive to the underlying data generating mechanism and it has better properties at the boundaries.[9]

In Step 3, the QQR estimator of Sun (2005) is used to calculate the semiparametric efficient estimator $\widehat{\beta}^*_{\alpha}$, instead of the average quantile estimator of Lee (2003), since the former is more robust to extreme observations if the effective sample size is not sufficiently large.

According to Lee (2003) and Sun (2005), the smoothing parameter matrix $B\tilde{}O\left(n^{-1/7}\right)$, since we have a two-dimension smoothing case. Since there are no results available in choosing the optimal smoothing parameter for the nonparametric conditional quantile curve estimation, we choose B according to the rule-of-thumb method of Yu and Jones (1998): define $b_0 = (\widehat{\sigma}_{inc}, \widehat{\sigma}_{age})\, n^{-\frac{1}{7}}$, then use $b_\alpha = b_0 \left\{\alpha(1-\alpha)/\phi\left(\Phi^{-1}(\alpha)\right)^2\right\}^{\frac{1}{7}}$ to esti-

[8] Lee (2003) uses Chaudhuri's (1991) method in Step 1. The essential difference between these two nonparametric smoothing techniques lies in the kernel function used in Step 1: the latter uses a uniform kernel, which may generate non-smooth curves.

[9] The interior algorithm of Portnoy and Koenker (1997) is used to solve the optimization problem in (14).

mate $\beta_{\alpha,i}$ and $g_\alpha(z_i)$ in Step 1, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution. Finally, the trimming parameter j_n is set to 0.05; the trimming function $\tau_z(z) = I(INC \in [a_1, a_2], AGE \in [6, 78])$ is used in estimating the unknown conditional pdf $f_v(0|w, z)$, where a_1 and a_2 are the empirical quantiles of income at the lower and upper probabilities 1%, respectively.

Moreover, we also calculate the unknown function $g_\alpha(z)$ by replacing $\beta_{\alpha,0}$ with $\widehat{\beta}^*_{\alpha,QQR}$ and the optimization problem solved is

$$\min_{a,\gamma} \sum_{j=1}^{n} \rho_\alpha \left(h_j - W_j \widehat{\beta}^*_\alpha - a - (Z_j - z_i)\gamma \right) \widetilde{K}\left(S^{-1}(Z_j - z_i) \right), \tag{21}$$

where $S = diag(s_1, s_2) = O\left(n^{-1/6}\right)$, the optimal rate of smoothing parameters in estimating $g_\alpha(z_i)$. $\widehat{g}_\alpha(z_i) = \widehat{a}$ will the estimate of $g_\alpha(z_i)$ for $i = 1, 2, \cdots, n$. Specifically, the optimal smoothing parameters are defined as $s_k = s_0 \left[\alpha(1-\alpha) / \phi\left(\Phi^{-1}(\alpha)\right)^2 \right]^{\frac{1}{6}}$, $k = 1, 2$ with $s_0 = (\widehat{\sigma}_{inc}, \widehat{\sigma}_{age}) n^{-\frac{1}{6}}$.

References

[1] Bhattacharya, P.K., and A.K. Gangopadtyay (1990) 'Kernel and Nearest-Neighbor Estimation of a Conditional Quantile,' *The Annals of Statistics 18, 1400-1415.*

[2] Cairney, J., and Arnold, R. (1996), "Social class, health, and aging: socioeconomic determinants of self-reported morbidity among the non-institutionalized elderly in Canada," *Canadian Journal of Public Health, 87, 199-203.*

[3] Chaudhuri, P. (1991) 'Nonparametric estimates of regression quantiles and their local Bahadur representation,' *The Annals of Statistics 19, 760-777.*

[4] Contoyannis, P. and M. Forster (1999) 'The Distribution of Health and Income: A Theoretical Framework,' *Journal of Health Economics, 18, 605-622.*

[5] Fan, J., T.C. Hu, and Y.K. Truong (1994) 'Robust Non-parametric function estimation,' *Scandinavian Journal of Statistics 21, 433-446.*

[6] Furlong, W.J., D.H. Feeny, and G.W. Torrance (1999) 'Health Utility Index (HUI): Algorithm for determining HUI Mark2 (HUI2)/ Mark 3 (HUI3) health status classification levels, health states, health-related quality of life utility scores and single-attribute utility score from 40-item interviewer-administered health status questionnaires,' *Dundas, Canada: Health Utilities Inc.*

[7] Grossman, M. (1972) 'On the concept of health capital and the demand for health' *Journal of Political Economy, 80, 223-255.*

[8] Hall, P., R.C. Wolff, and Q. Yao (1999) 'Method for Estimating Conditional Distribution Function,' *Journal of the American Statistical Association 94, 154-163.*

[9] Humphries, K.H., and van Doorslaer, E. (2000), "Income-related inequalities in health in Canada," *Social Science and Medicine, 50, 673-748.*

[10] Khan, S. and J.L. Powell, 2001. Two-step estimation of semiparametric censored regression models. *Journal of Econometrics 103, 73-110.*

[11] Koenker, R., and G. Bassett (1978) 'Regression quantiles,' *Econometrica 46, 33-50.*

[12] Koenker, R., and V. d'Orey (1987) 'Computing regression quantiles,' *Applied Statistics 36, 383-393.*

[13] Koenker, R., P. Ng, and S. Portnoy (1994) 'Quantile Smoothing Spline,' *Biometrika 81, 673-680.*

[14] Lee, S. (2003) 'Efficient Semiparametric estimator of a partially linear quantile regression model,' *Econometric Theory 19, 1-31.*

[15] Portnoy, S. and R. Koenker (1997) 'The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators,' *Statistical Science 12, 279-300.*

[16] Newey, W.K., J.L. Powell and J.M. Walker (1990) "Semiparametric Estimation of Selection Models: Some Empirical Results", *American Economic Review Papers and Proceedings, 80, 324-328.*

[17] Rogers, G.B. (1979) 'Income Inequality as Determinants of Mortality: An International Cross-Section Analysis' *Population Studies, 33, 343-351.*

[18] Rosenberg, M.W. and K. Wilson (2000) 'Gender, poverty and location: how much difference do they make in the geography of health inequalities,' *Social Science & Medicine 51, 275-287.*

[19] Sun, Y. (2005) 'Semiparametric efficient estimation of partially linear quantile regression models,' *The Annals of Economics and Finance 6, 105-127.*

[20] Tambay, J.L., and G. Catlin (1995) 'Sample design of the National Population Health Survey,' *Health Reports 7, 29-38.*

[21] Vozoris, N.T. and V.S. Tarasuk (2003) 'Household food insufficiency is associated with poorer health,' *Journal of Nutrition 133, 120-126.*

[22] Wagstaff, A and E. van Doorslaer (2000) 'Income Inequality and Health: What does the Literature Tell Us?' *Annual review of Public Health, 21, 543-567.*

[23] Wildman, J. and A.M. Jones (2003) 'Is it Absolute Income or Relative Deprivation that Leads to Poor Psychological Well being? A Test based on Individual-level Longitudinal Data,' manuscript, Department of Economics, University of Newcastle.

[24] Williamson, D.L. (2000) 'Health behaviors and health: evidence that the relationship is not conditional on income adequacy,' *Social Science & Medicine 51, 1741-1754.*

[25] World Health Organization (1958) The first ten years of World Health Organization. *Geneva: World Health Organization.*

[26] Yu, K., and M.C. Jones (1998) 'Local linear quantile regression,' *Journal of the American Statistical Association 93, 228-237.*

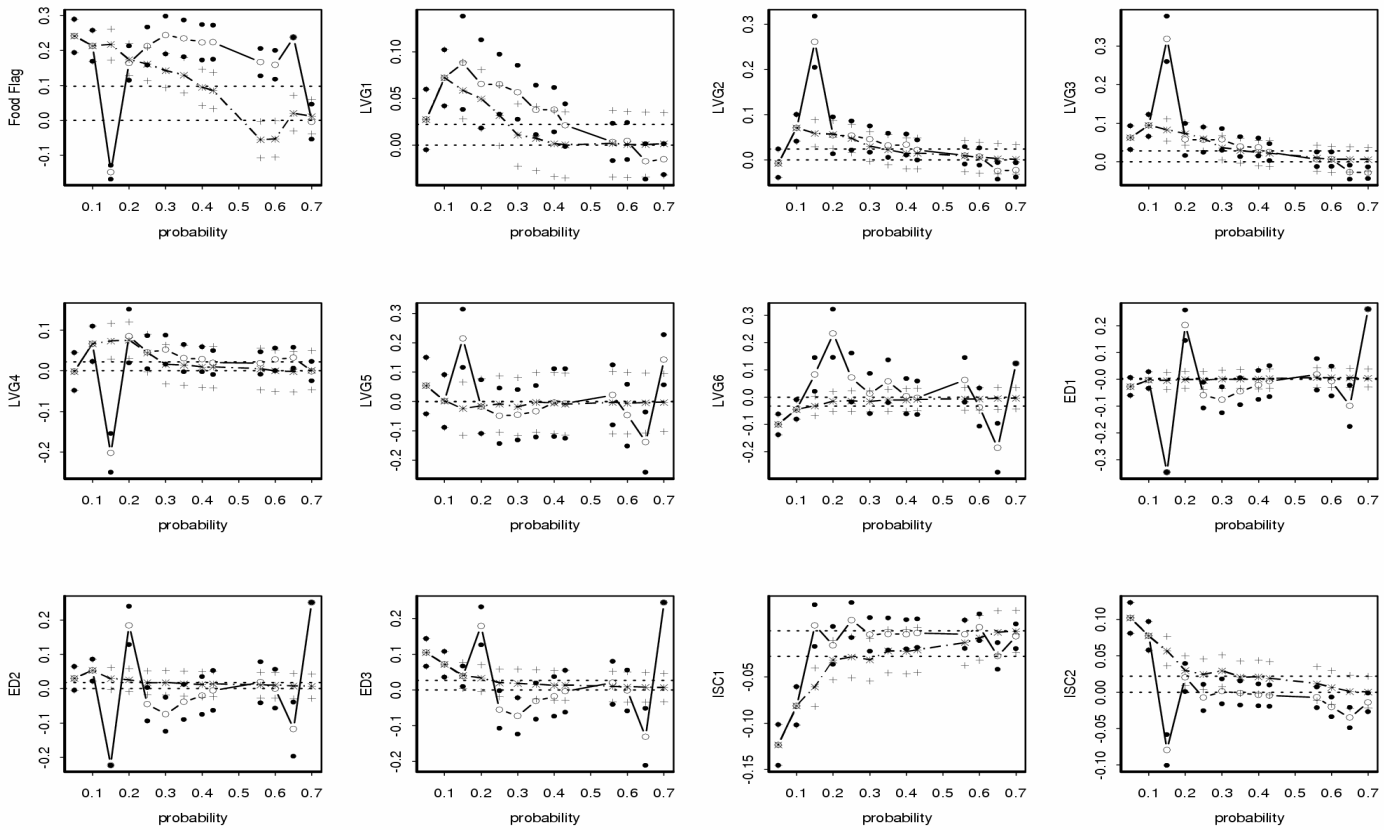
[27] Zheng, J.X. (1998) 'A consistent nonparametric test of parametric regression models under conditional quantile restrictions,' *Econometric Theory 14, 123-138.*

Table 1. Data description

Variable	Definition
HUI	health utility index takes value between -0.34 and 1 in 1998-99 NHPS
INC	logarithm of the respondent's family income
AGE	the respondent's age
FFLAG	takes a value of one if the respondent does not worry about food shortage or lack of satisfaction of the food he/she ate in the past 12 months; zero otherwise
SEX	takes a value of one if the respondent is female; zero otherwise
LVG	The living arrangement describes how the respondent relates to others. Of seven types considered, we define six dummy variables
LVG1	takes a value of one if the respondent is a single adult living alone; zero otherwise
LVG2	takes a value of one if the respondent is an adult living with a spouse or partner without child; zero otherwise
LVG3	takes a value of one if the respondent is an adult living with a spouse or partner with children; zero otherwise
LVG4	takes a value of one if the respondent is a single parent living with children; zero otherwise
LVG5	takes a value of one if the respondent is a child living with single parent; zero otherwise
LVG6	takes a value of one if the respondent is a child living with two parents; zero otherwise
ED	the highest education levels of the respondent; of four groups, we define three dummies, and no dummy for non-schooling for children.
ED1	takes a value of one if the respondent has secondary school or lower; zero otherwise
ED2	takes a value of one if the respondent has post secondary diploma; zero otherwise
ED3	takes a value of one if the respondent has BA degree and above; zero otherwise
ISC	captures insurance coverage
ISC1	takes a value of one if the respondent is covered by prescription medication; zero otherwise
ISC2	takes a value of one if the respondent is covered by hospital charges; zero otherwise

Except for the health utility index, family wage income, and age, the rest of variables are 0-1 dummies in nature.

Figure 1. Estimates of Partially Linear Quantile Regression Models



The empty circles are the estimates from the partially linear quantile regressions and the 95% confidence intervals are marked by the solid points. The estimates from the linear quantile regressions are marked by asterisks and the cross signs give the corresponding 95% confidence intervals. The OLS estimates and zero are in dotted lines.

Figure 2. Data-grid surfaces of $\left(inc, age, \hat{g}_{\alpha}\left(inc, age\right)\right)$

The vertical axis is scaled up by 10; and $\alpha = \left(0.25, 0.56\right)$

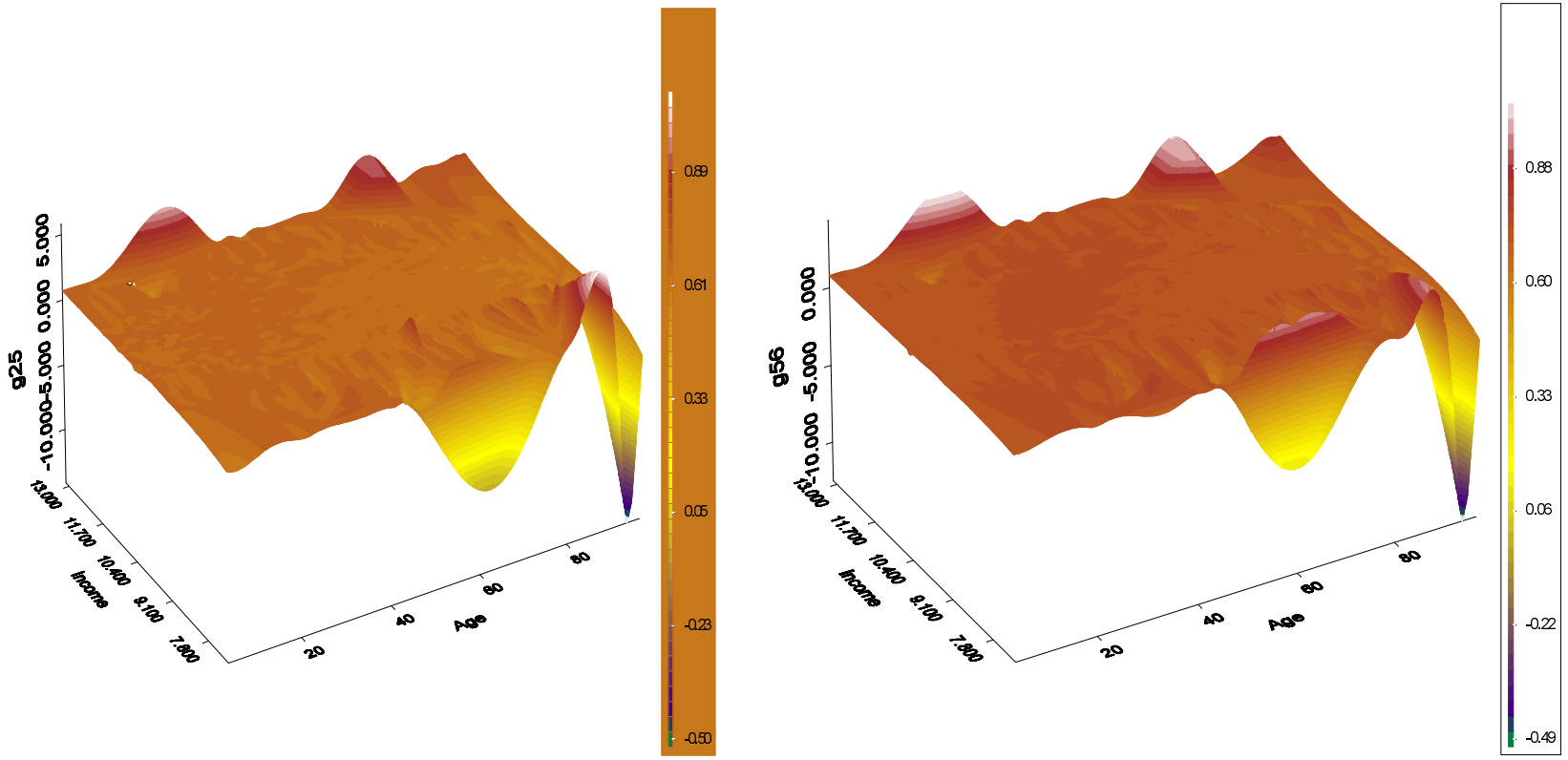
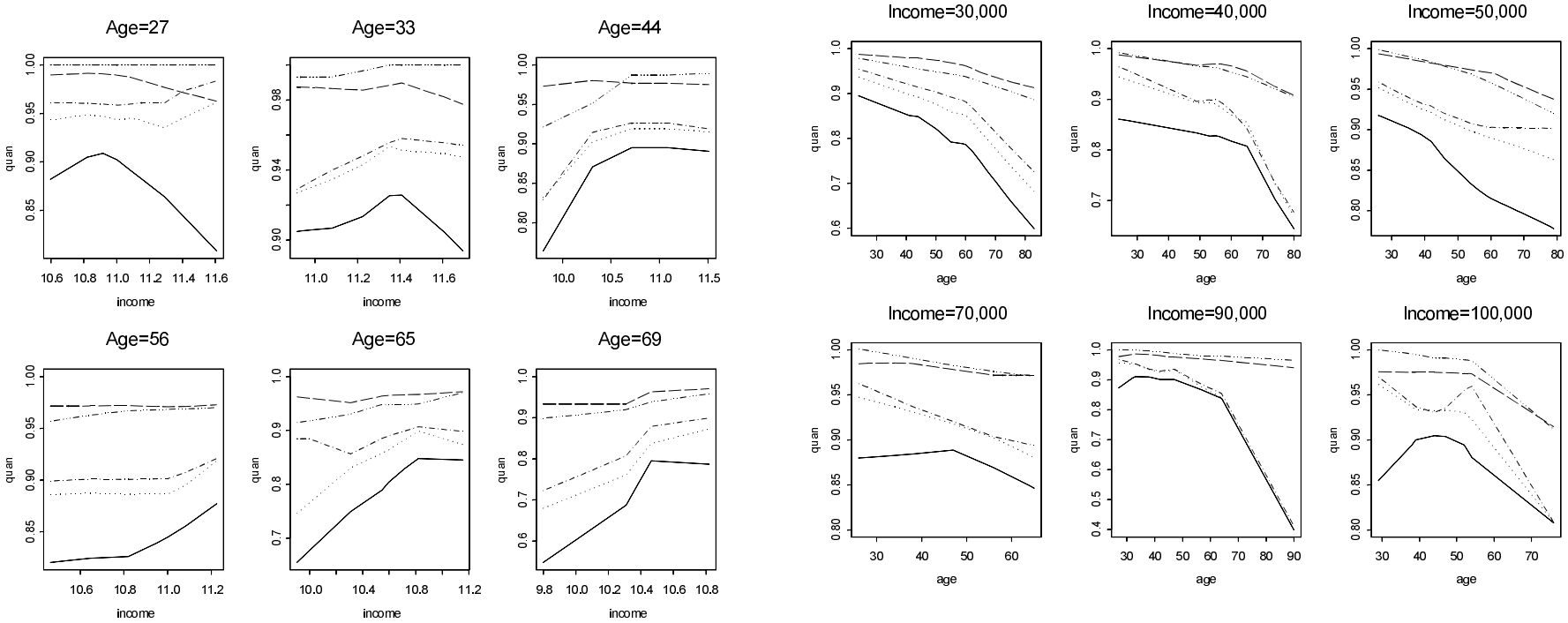


Figure 3. Fitted health status for (Male, LVG2)



Note: From bottom to top, α = 0.2, 0.25, 0.30, 0.56, and 0.65, respectively.