REPORTING HETEROGENEITY IN SUBJECTIVE HEALTH MEASURES: AN EXTENDED LATENT CLASS APPROACH

VALENTINO DARDANONI E PAOLO LI DONNI

REPORTING HETEROGENEITY IN SUBJECTIVE HEALTH MEASURES: AN EXTENDED LATENT CLASS APPROACH

VALENTINO DARDANONI AND PAOLO LI DONNI

Abstract. There are two general sources of individual unobserved heterogeneity when subjective indicators are used to measure health status (Shmueli [20]): variations in unobservable true health and differences in self-reporting behavior for given level of "true health". In this paper we extend the empirical strategy proposed by Etile and Milcent [7] to distinguish between the effect of individual characteristics on "true" unobserved health from the effect on self reporting behavior. To this aim we use some recent developments on finite mixture models to identify two unobserved types of individuals which differ with respect to their "true" health. To estimate unobserved health we use biomarkers, which are raising a great deal of interest since they can be used to validate respondents' self-reported health measures. Our results show a positive relationships between biomarkers and true health and support the existence of self-reporting bias related to socioeconomic characteristics and individual life styles. Comparing our results with those from the generalized ordered logit we obtain an overall reduction on self-reporting bias due to individual characteristics.

JEL Classification Numbers I10; I12
Keywords Self-assessed health, Self-reporting bias, Biomarkers, Latent Class Analysis.

1. Introduction

Self-assessed health status (SAH) is a widely employed subjective health indicator in empirical research. It is based on the simple question "How is your health in general?" with a response framed in ordered categories ranging from "very good" or "excellent" to "poor" or "very poor". It is assumed that these responses are generated by a corresponding continuous latent variable representing self-perceived health. Several studies found that SAH is a good predictor of mortality, morbidity and subsequent use of health care (Idler and Benyamini [13]). Furthermore Gerdtham *et al.* [8] showed that a continuous health measure obtained from the ordinal responses of SAH is highly correlated with other individual health measures.

Date: June 15, 2009.

As subjective indicator SAH has caused some concern among researchers related to the idea that individuals may link differently the same level of true health with the SAH's categories. The existence of these differences in self-reporting behavior is convincingly supported by empirical findings (Crossley and Kennedy [5], Groot [10]). For example, Crossley and Kennedy [5] exploit a particular feature of the Australian National Survey in which SAH question was asked to respondents and again to a random subsample. Results show that the distribution of SAH for this subgroup of respondents changes significantly between the two questions and that this variation depends on age, income and occupation.

This source of measurement error in the mapping of true health into SAH that we call - following Shmueli [20] - unobserved reporting heterogeneity, has also been termed 'state-dependent reporting bias' (Kerkhofs and Lindeboom [15]), 'scale of reference bias' (Groot [10]), 'response category cut-point shift' (Sadana *et al.* [19], 2000; Murray *et al.* [17]).

To account for self-reporting behavior a possible approach is to use an ordered probit with cut point shift. This model allows the cut-points defining the mapping of latent health into the SAH's categories to depend on observable variables (Terza [21]). Although this approach allows for both index and cut off points shifts, it requires strong a priori restrictions on parameters to solve identifiability problems especially when the set of covariates on latent health and the cut-off point overlap (Lindeboom and van Doorslaer [16]).

There are many other papers that have analysed SAH. Van Doorslaer and Jones [22] use the McMaster 'Health Utility Index Mark' (HUI) to scale the intervals of SAH. They assume a stable mapping of HUI on the latent health determining SAH. Therefore the position of an individual ranked according to HUI should correspond to her rank according to SAH. They exploit this relationship between HUI and SAH to estimate an interval regression model where the upper (lower) bound of these intervals corresponds to the upper (lower) value on HUI's empirical distribution

corresponding to the empirical cumulative frequency of SAH. A second approach was proposed by Kerkhofs and Lindeboom [15] and Lindeboom and van Doorslaer [16]. They stratify the population in several groups according to some individual characteristics and then estimate an ordered response model of SAH on HUI as proxy of true health. This estimation approach allows differences both with respect to cut-points and index-shift. On the same fashion Etilè and Milcent [7] use latent class analysis to construct a synthetic measure of clinical health and estimate a generalized ordered logit to investigate the effect of socio-economic status (e.g income level) on self reporting behavior. To assess the magnitude of reporting heterogeneity related to income they assume that all the information on true health are captured by the synthetic measure of clinical health, that is, they argue that individual characteristics should be *ignorable* to predict SAH. Therefore reporting heterogeneity is tested considering whether these characteristics have a significant effect after conditioning on clinical health.

Etilè and Milcent's [7] approach assumes that no information on "true" health are contained on SAH to construct the synthetic clinical health measure. To overcome this problem a possible approach relies on the use of multiple indicators to construct an estimate of true health and then identify the variation in true health from reporting heterogeneity. Shmueli [20] estimates a structural equation model exploiting some features of multiple indicators-multiple causes (MIMIC) modelling to shape the relationships between true health and and a set of indicators (Joreskog and Goldberger [14]).

In this paper we use some recent developments on latent class analysis to provide an empirical assessment of reporting heterogeneity using a set of "manifest" (objective and subjective) health indicators in a recursive model with unobserved latent class. In particular our aim is to investigate how to disentangle the effect of individual characteristics on health production on the one hand, and its judgement effect on the other hand. Further we evaluate the magnitude of some individual

characteristics on self-reporting heterogeneity considering the residual association between self-reported indicators conditioning on true health. As mentioned above our econometric strategy relies on some recent developments on latent class analysis (LCA) which allow to model explicitly the residual association between indicators and renders the approach we follow substantially different from those relying on MIMIC (Shmueli [20]).

Furthermore, our approach differs from the previous literature regarding the type of indicators exploited to measure "true" health. In fact "true" health is constructed using both subjective (SAH and other self reported health conditions) and objective indicators (biomarkers). On the one hand, this avoids the arbitrariness of excluding SAH itself from the set of measures indicating clinical health. On the other hand, there is a great deal of interest among researchers on biological measures for several reasons. Biomarkers can be used not only to validate respondents' self-reported health measures but also to identify true health status and compare different groups of individuals (Banks *et al.* [2]); biological measures allow to take into account the preclinical levels of disease even when the respondents may not have been aware.

We identify two unobserved classes representing people in good health and those in ill-health respectively. Our main finding provides further evidence of heterogeneity in self reporting behavior. In fact after conditioning on unobserved individual 'true' health-types personal characteristics cannot be ignored to predict SAH. In particular for a given level of "true" health people with higher income, better education and living in less deprived area tend to report systematically better health. Moreover there is evidence that individual characteristics affect differently the reporting behavior in each category of SAH.

The paper is organized as follows. In the next section we describe the data we use on our analysis. The following section explains the methods and the empirical strategy we exploit. Empirical findings are found in section 4. The last section discusses the results and concludes.

2. Data and variable definition

We use cross-sectional data from two waves (2003 and 2004) of Health Survey for England (HSE). This is a large survey covering a wide range of fields related to socioeconomic status, health and life-style. In 2003 the major focus of the survey is cardiovascular disease, which is (including heart attacks and strokes) one of the largest single cause of death in England. Even when this type of disease is not fatal, it brings ill-health and disability which might deeply affect individuals' life. Therefore it is extremely important to obtain objective measures of cardiovascular risk in order to asses individual "true" health. For this reason HSE is very suitable for our aims because it contains some biological measures obtained from a blood sample. The same biological measures are available for all the sample aged 16 years or over in 2003, while only for a subgroup which represents a minority ethnic groups in 2004.

Health Survey for England is composed by two parts, an interviewer-administered interview (Stage 1), and a visit by a nurse to carry out measurements and take a blood sample (Stage 2). At each stage participants are asked to decide whether to proceed with the following stage or not. Therefore someone may agree to take part at Stage 1 but decide not to continue to Stage 2.

In the first stage individual questionnaire is administered in order to collect information on general health, eating habits, physical activity, smoking, drinking, family cardiovascular disease history and socioeconomic status (e.g. income, employment status, educational background). At the end of this stage respondents are asked to proceed with stage 2 and in case to make an appointment with a qualified nurse. In this second stage nurses ask more information on health and health care utilization and to provide only for those older than 35 in 2003 and 16 in 2004 a fasting blood sample and a blood pressure measurement.

For our analysis we consider an homogeneous sample of individuals aged 30 or over excluding cases with incomplete or inconsistent information on the relevant

socioeconomic, demographic, health and life-style variables. The remaining sample size consists of 3,381 observations.

There are three important sets of variables relevant for our analysis (see tables 1 and 2). The first set includes three binary objective health indicators obtained considering only valid measures of the lab tests and excluding all the cases in which lab test results have been affected by individual behavior (e.g. people that have smoked or eaten before the nurse visit, etc.). The first indicator (BPN) takes 1 if individual has normal blood pressure measured by a qualified nurse with Dinamap and Omron measures. Our second objective indicator (CHL) is a binary measure derived from the total cholesterol/high density lipoprotein ratio in the blood. This ratio is more indicative of cardiovascular disease than total cholesterol since it consider both high and low density lipoprotein cholesterol. Then the variable CHL takes 1 if individual has the cholesterol ratio below a sex-adjusted threshold indicating a low risk of cardiovascular disease. In particular for men an acceptable ratio of total cholesterol/high density lipoprotein is 4.5 or below, and for women is 4.0 or below. Finally our last objective indicator is based on the c-reactive protein (CRP) blood test. CRP may be used to screen apparently healthy people for cardiovascular disease (CVD). If the CRP level in the blood drops, it means that individual are getting better and CVD risk factor is being reduced. The CRP indicator takes 1 if individual have a lab test score lower 3.0 mg/L, associated to low chance of having a sudden heart problem.

Our second set of variables includes two subjective health indicators: SAH and self-reported limiting longstanding illness (LLI). This latter variable is available directly from the survey and it was derived considering whether individual has longstanding illness and whether daily activities are limited due to this illness. LLI takes 1 if individual has no chronical limitations on daily activities.

The responses categories for SAH were very bad, bad, fair, good and very good. We combine three SAH's categories (very bad, bad and fair) in one category representing poor health, because only a relatively small fraction (15.5%) of the sample reports poor health. The SAH variable thus consists of three ordered categories: poor, good and very good health.

Finally the last set of variables provides information on socio-economic status, demographic characteristics and life-style. Socioeconomic status is measured using the equivalised income provide and an overall index of multiple deprivation (IMD2004). This is a composite index of relative deprivation at small area level, based on seven domains of deprivation involving for example income, employment, health deprivation and disability, education, crime and living environment.[1] This survey is also rich of information on individual life-style. These variables offer a good opportunity to better identify individual health. In particular there are detailed information on past and present smoking behavior as well as physical activities, sport intensity and the daily number of portion of fruits and vegetables. HSE provides also a three levels fat score ranged from "low fat" to "high fat" eating habits derived considering the consumption of cheese, fish, fried food, meat, etc.

3. The Model

Our aim is to study the association between SAH and "true" health, using recent developments on latent class analysis, which allow covariates to affect latent class membership, and possibly residual association among indicators after conditioning on latent health-types (Huang and Bandeen-Roche [12], Bartolucci and Forcina [3] and Dardanoni, Forcina and Modica [6]).

Let U be a latent discrete variable with two categories representing individuals in good ($U = 0$) and bad ($U = 1$) health. The main problem when one wants to study

[1]Equivalised income variable is provided by the HSE. It is computed using the McClement score for each household (dependent on number, age and relationships of adults and children in the household), and then dividing the total household income by this score to get an equivalised household income.

the relationship between true health and self-reported health is to disentangle the effect that some personal characteristics have on "true" health's variations from the effect that the same variables have on self-reporting.

Etilè and Milcent [7] suggest the following strategy to distinguish between these two effects. They assume that "true" health is entirely captured by a synthetic measure of clinical health (which they denote H^0) for which the following ignorability condition holds (see Wooldridge [24], p. 63):

$$Pr(Y_{sah} = i \mid H^0, \boldsymbol{z}) = Pr(Y_{sah} = i \mid H^0) \qquad i = 1, 2, 3 \tag{1}$$

where H^0 was obtained using a latent class model with self-reported health measures as indicators. The assumption above relies on the fact that the effect of covariates $\boldsymbol{z}$ on "true" health is entirely captured by H^0. Thus, if SAH is a reliable indicator of individual health, then any differences on personal characteristics should not affect the distribution of SAH after conditioning on H^0, which means that $\boldsymbol{z}$ is ignorable to predict SAH. Therefore if the assumption above holds, a test of self-reporting behavior can be easily performed regressing Y_{sah} on H^0 and $\boldsymbol{z}$ and testing whether parameters of personal characteristics are still significant conditioning on the synthetic measure of clinical health.

Our approach can be considered an extension of Etilè and Milcent's [7] test from two points of view. First, we estimate endogenously individual "true" health U by taking information both from subjective (SAH, LLI) and objective indicators (BPN, CRP, CHL), so that all available information from health indicators, including SAH, are used to estimate individual "true" health. Second we allow for residual association between health indicators in order to capture any *adaptation effect* of individuals to their own health condition. This means that the effect of subjective health measure, such as Y_{lli}, on SAH status is not only driven by U but it could also affect *indirectly* self-reporting behavior of SAH itself - for example people may

adapt to a chronical limitation status measured by Y_{lli} and report systematically better health.

Let $\boldsymbol{Y} = (Y_{sah}, Y_{lli}, Y_{bpn}, Y_{chl}, Y_{crp})$ be the vector of observable response variables. As a simple starting point, consider the traditional latent class analysis (see e.g. Goodman [9]), which implies the existence of a discrete U such that these observables $\boldsymbol{Y}$ are independent conditionally on U. This is also named "local independence" and is expressed as:

$$Pr(Y_{sah}, Y_{lli}, Y_{bpn}, Y_{chl}, Y_{crp}) = \sum_{u=0}^{1} Pr(Y_{sah} \mid u) \cdots Pr(Y_{crp} \mid u) Pr(u)$$

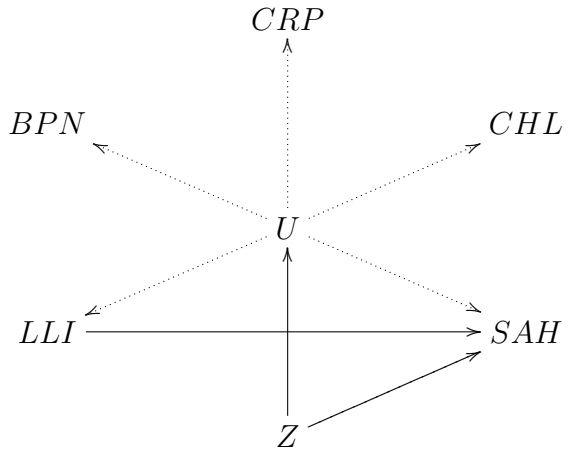
Clearly a U that makes these responses conditionally independent captures elements of individuals' "true" health. However, the local independence assumption is too restrictive for our purposes since it does not allow responses and latent health-types to depend on covariates, and does not allow residual association between any response after conditioning on U.

Our model assumes that the joint distribution of responses $(U, \boldsymbol{Y})$ conditional on the set of observable covariates $\boldsymbol{z}$ (describing demographic, socioeconomic and life-style individual characteristics) is fully determined by the following set of conditional distributions of observables, and by the marginal distribution of U:

$$\begin{aligned} &Pr(U = 1 \mid \boldsymbol{z}), \\ &Pr(Y_{sah} = i \mid Y_{lli}, \boldsymbol{z}, U) \\ &Pr(Y_{lli} = 1 \mid U) \\ &Pr(Y_{bpn} = 1 \mid U) \\ &Pr(Y_{chl} = 1 \mid U) \\ &Pr(Y_{crp} = 1 \mid U) \end{aligned} \tag{2}$$

which can be equivalently formulated in terms of the following *directed acyclic graph* (see Pearl [18]):

FIGURE 1. *Directed Acyclic Graph of model* (2)



We then model these conditional probabilities as linear functions of the covariates $\boldsymbol{z}$ using a logit link to form a multivariate regression system of logit equations:

$$\begin{aligned}
&\Pr\left(U=1\left|\boldsymbol{z}\right.\right)=\Lambda\left(\alpha_u\left(u\right)+\boldsymbol{z}'\boldsymbol{\gamma}\right)\\
&\Pr\left(Y_{sah}>1\left|u,y_{lli},\boldsymbol{z}\right.\right)=\Lambda\left(\alpha_1\left(u,y_{lli}\right)+\boldsymbol{z}'\boldsymbol{\beta}_1\right)\\
&\Pr\left(Y_{sah}>2\left|u,y_{lli},\boldsymbol{z}\right.\right)=\Lambda\left(\alpha_2\left(u,y_{lli}\right)+\boldsymbol{z}'\boldsymbol{\beta}_2\right)\\
&\Pr\left(Y_{lli}=1\left|u\right.\right)=\Lambda\left(\alpha_3\left(u\right)\right)\\
&\Pr\left(Y_{bpn}=1\left|u\right.\right)=\Lambda\left(\alpha_4\left(u\right)\right)\\
&\Pr\left(Y_{crp}=1\left|u\right.\right)=\Lambda\left(\alpha_5\left(u\right)\right)\\
&\Pr\left(Y_{chl}=1\left|u\right.\right)=\Lambda\left(\alpha_6\left(u\right)\right)
\end{aligned} \tag{3}$$

where Λ is a logit link function $\Lambda = e^t/(1+e^t)$. Note that $\alpha_i\left(u,y_{lli}\right)$, $i=1,2$, represents all the possible combinations between U and Y_{lli}; since U and Y_{lli} are binary this means we have 4 α parameters in the second and third equation that could be alternatively expressed as $a_1+a_2U+a_3Y_{lli}+a_4Y_{lli}U$. Notice that for sake of generality we assume that the same set of covariates affecting "true" unobserved health may also potentially affect reporting behavior.

Parameters in model (3) are estimated by the EM algorithm.[2] In the E step the so called posterior probability of latent class U given the observed configuration $\boldsymbol{y}$

[2]We are grateful to Antonio Forcina for kindly providing the Matlab code for the estimation.

is computed. The M-step maximizes a likelihood function that is further refined in each iteration by the E-step. Details on estimation and identification of model (3) can be derived by looking at the Appendix of Dardanoni, Forcina and Modica [6] and at Bartolucci and Forcina [3].

We propose two tests of reporting heterogeneity. The first test is nothing but the ignorability condition of $\boldsymbol{z}$ in the equation determining SAH (compare with Etilè and Milcent's [7] equation (1) above), that is:

$$Pr(Y_{sah} = i \mid U, Y_{lli}, \boldsymbol{z}) = Pr(Y_{sah} = i \mid U, Y_{lli}) \qquad i = 1, 2, 3 \tag{4}$$

which can be performed by testing whether $\boldsymbol{z}$ has a significant influence on SAH in model (3), that is, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}$. Our second test is more specific and is focused on whether individual characteristics affect different parts of SAH distribution after conditioning on U and Y_{lli}, that is, testing whether $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Both tests are performed by estimating a restricted model and computing a LR-test which has a chi-square asymptotic distribution.

4. Results

4.1. **Generalized ordered logit results.** As benchmark of our analysis we use the results obtained by a generalized ordered logit model of SAH on the set of indicators and individual characteristics:

$$\begin{aligned} \Pr\left(Y_{sah} > 1 \,|\boldsymbol{z}, \boldsymbol{w}\right) &= \Lambda\left(\alpha_1 + \boldsymbol{z}'\boldsymbol{\eta}_1 + \boldsymbol{w}'\boldsymbol{\theta}\right) \\ \Pr\left(Y_{sah} > 2 \,|\boldsymbol{z}, \boldsymbol{w}\right) &= \Lambda\left(\alpha_2 + \boldsymbol{z}'\boldsymbol{\eta}_2 + \boldsymbol{w}'\boldsymbol{\theta}\right) \end{aligned} \tag{5}$$

(compare with the second and third equations of system (3)), where $\boldsymbol{z}$ is the vector of socio-economic, demographic and life styles characteristics as above, and $\boldsymbol{w}$ is a vector of health variables. Notice that following Etilè and Milcent [7] (and to make the regression systems (3) and (5) directly comparable), we assume that only the coefficients of $\boldsymbol{z}$ are allowed to vary across the categories of SAH; while health variables are assumed to affect uniformly the SAH's distribution.

We first tested the null hypothesis of parallel lines - called the *proportional odds assumption* in the statistical literature (Agresti [1], p. 275) - by imposing the restriction that $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$; the likelihood ratio test is equal to 63.25 with 37 df (p-value .0045). Thus, the hypothesis of parallel lines is rejected.

Table (4) shows the estimated coefficients from the generalized ordered logit. Results show that with this specification there are several individual characteristics that affect reporting behavior. In particular people with higher income who live in less deprived area tend to report better health. This result confirms what obtained in many other papers about the existence of income-related reporting heterogeneity (Hernandez-Quevedo *et al.* [11], Etilè and Milcent [7], Lindeboom and van Doorslaer [16]). Moreover health related life variables, such as sport and physical activities, also increase the probability to report good health. Interestingly those who eat more portions of fruits and vegetables per day are more likely to report very good health than just at least good health.

In this framework it is hard to distinguish the effect of personal characteristics on"true" health from the effect on self-reporting behavior. For this reason we estimate model (3) where observables are allowed to affect both individuals' health and reporting heterogeneity.

4.2. **Results from model** (3)**.**

4.2.1. *Intercepts.* Since U is binary, table (5) shows $2^*4 + 2^*2^2 + 1 = 17$ estimated intercepts α. In particular there is 1 parameter to describe the class membership probability, 2 parameters for each of the 4 health indicators, and 4 parameters to describe the effect of U and Y_{LLI} on SAH for people who report at least good or very good health.

A glance at the table reveals that people with good health ($U = 0$) are much more likely to have desirable lab test scores and no limiting longstanding illness compared to people with ill-health ($U = 1$). Furthermore, it is easily checked that people with $U = 0$ are also much more likely to report at least good or very good

health conditional on $Y_{lli} = 0, 1$. Regarding the effect of Y_{lli} on SAH, it is also easily checked that the probability to report at least good or very good health is also increasing in Y_{lli} conditional on $U = 0, 1$.[3] This result provide further evidence on the existence of heterogeneity in self-reporting behavior related to differences in self-perceived limiting illness.

4.2.2. *Variations in the unobserved "true" health U.* The first column of table (6) reports the estimated parameters $\boldsymbol{\gamma}_u$ of both health related variables and socioeconomic characteristics:

- the effects of demographics characteristics on unobserved "true" health have the expected sign. In fact, ill-health is positive and statistically related to age, but negatively with sex and ethnicity. As expected women tend generally to have better health than men (Wingard [23]);
- socio-economic characteristics play an important role in health determination. In particular those with higher education have lower probability of being classified with poor health than those with no qualification. Health status is also strongly and positively correlated with income and social class as showed by estimated coefficients of equivalised income and social class. Another important role on determining individual health is also played by the index of multiple deprivation. Individuals who live in highly deprived area register a lower level of health than those living in less deprived area, although the effect seems to statistically vanishes as the deprivation decreases. Finally there is a small and negative statistical significant effect on health of the number of months individual lived in the same area;
- unobserved health is also related with individual life-style characteristics; individual who are no smokers who practice sport regularly with a moderate physical activity and follow a diet with a low fat content have a greater

[3]Just as an example, the probability that people with $U = 0$ report very good health is on average .52 if $Y_{lli} = 0$ while it is .97 if $Y_{lli} = 1$.

probability of having good unobserved health; the opposite holds for individuals who are obese with cardiovascular conditions in the family. Finally ill-health seems to be negatively correlated with the parents' age and with to the number of units drank in the heaviest day in the last seven days. This last result is clearly unexpected, although it could be related to a sort of measurement error in the drinking-unit variable.

4.2.3. *Variations in reporting behavior.* The discussion above shows significant variation in unobservable health status by personal characteristics, representing a considerable source of unobserved heterogeneity in health production which should be taken into account. In the present section, we analyse the direct relationships between self-reported health and observable characteristics conditional on true unobserved health status and no limiting longstanding illness. Recall that for the sake of generality we have assumed (see (3)) that the set of covariates $\boldsymbol{z}$ affecting "true" unobserved health may also potentially affect reporting behavior.

We first tested the parallel line assumption by estimating a restricted model in which individual characteristics have the same effect on different categories of the SAH distribution. The value of log-likelihood for the restricted and unrestricted model is equal to -10979 and -10953. The value of likelihood ratio test is 53.72 which is clearly rejected with 37 d.f (p-value = .037). Results on unobserved heterogeneity in self-reported behavior are reported in the last four columns of table (6). They differ slightly with respect of SAH classes and can be summarized as:

- after conditioning on U and Y_{lli} a wide set of variables (such as ethnicity, education and individual life styles) are not anymore statistically significant in model (3) in both SAH's categories as compared with the results obtained in the generalized order logit model discussed above. In fact only some variables which appeared to significantly affect both SAH'categories in model (5) are also significant in both at least good and very good health in model (3);

- individual with higher income tend to self report better health status. On the contrary people living in the most deprived area and with low educational attainment report systematically a worse level of health. The magnitude of the effect does not differ significantly among SAH's classes;
- physical activities and sports increases the probability to report "very good" health but not at least "good" health. In particular this effect seems to increase as the physical activity is more vigourous and performed regularly.
- age affects self-reporting behavior. In particular elders tend to report better health than expected. This result seems plausible with previous findings (Lindeboom and van Doorslaer [16], Groot [10]) and confirms the apparently puzzle between self reported health and age, which is related to the existence of individual adaption to chronical ill conditions;
- individual's life-styles tend also to affect differently reporting behavior. This may be connected to the fact that individual with healthy life may over estimate their health status and then tend to over report subjective health. This effect is also increasing in the effort required by the activity itself, but is smaller compared with those obtained with the generalized ordered logit.

5. Discussion and final remarks

The present study explores the relationships between socio-economic, demographic and life-style personal characteristics and health. In particular we test the existence of reporting heterogeneity on SAH implementing the approach proposed by Etilè and Milcent [7]. Our empirical strategy is innovative in two ways. First we use some recent developments on LCA to disentangle the effect of personal characteristics on self-reporting behavior from the effect on heterogeneity in health production, and we allow residual association between self-reported indicators in order to capture differences related to reporting heterogeneity after conditioning on latent "true" health. Second, to identify unobserved individual latent health we use not only subjective measures, but also objective indicators, such as biological measures, which help to

validate respondents' self reports and to identify individual health by taking into account the pre-clinical levels of disease even when the respondents may not have been aware (Banks *et al.* [2]).

Our results confirm the existence of self-reporting behavior phenomenon which is reported in many other empirical investigations, which concerns the existence of individuals who tend to under(over) report individual true health. Interestingly after conditioning on individual unobserved health-type, several individual characteristics have a statistically significant effect on self-reporting behavior limited to SAH's categories as compared with the generalized order logit which does not distinguish explicitly between heterogeneity in health and reporting behavior.

Appendix A. Tables

Table 1. Variable Definitions

Variable	Definition
chl	Total/high density lipoprotein cholesterol (1 = if lab test score is good, 0 otherwise)
crp	C-reactive protein (1 = if lab test score is lower than 3 mg/L, 0 otherwise)
bpn	Blood pressure (1 = normotensive with Dinamap and Omron readings), 0 otherwise)
lli	Limiting Longstanding Illness (1 = if no Limiting Longstanding Illness, 0 otherwise)
sah	Self-Assessed Health status (1 = "poor health", 2 = "good", 3 = "very good")
mar	1 = if individual is married, 0 otherwise
age	age of individuals
women	1 = female, 0 otherwise
black	1 = black, 0 otherwise
white	1 = white, 0 otherwise
noqual	1 = no qualification, 0 otherwise
eduh	1 = second level or higher, 0 otherwise
scl2	1 = social class for skilled non-manual and skilled manual
scl3	1 = social class professional and managerial technical
eqvinc	Equivalised income
imd3	1 = third quintile of Overall Index of Multiple Deprivation
imd4	1 = fourth quintile of Overall Index of Multiple Deprivation
imd5	1 = fifth quintile of Overall Index of Multiple Deprivation (most deprived)
hse04	1 = if individual belongs to HSE 2004, 0 otherwise
bmil	value of Body Mass Index if it is lower than 18.5, 0 otherwise
bmih	value of Body Mass Index if it is higher than 29.9, 0 otherwise

Table 2. Variable Definitions

Variable	Definition
drinkun	# of drinking units in the heaviest day
agema	age of mother
agepa	age of father
demam	1 = whether the mother is dead
depa	1 = whether the father is dead
famcvd	1 = whether there are cardiovascular conditions in the family history
smacc	1 = if someone smokes in the accommodation
smkc	1 = if individual smokes currently
smkevr	1 = if individual has ever smoked
smkex	1 = if individual is an ex smoker
smkoc	1 = if individual smokes occasionally
sportm	1 = moderate sport activity
sportr	1 = regular sport activity
hrsspt	# of hours of sport per week
phy2	1 = medium physical activity level
phy3	1 = high physical activity level
veg	# of portions of fruits and vegetables per day
fatt2	1 = if individual's diet has a medium fat score
fatt3	1 = if individual's diet has a high fat score
livehm	# of months individual has lived in this local year
urban	1 = if individual lives in an urban area

TABLE 3. Descriptive Statistics

	Mean	S.D		Mean	S.D
chl	0.5927	0.4913	drinkun	3.1656	2.4252
crp	0.7403	0.4385	agema	70.1709	11.9926
bpn	0.6613	0.4733	agepa	68.9665	11.5362
lli	0.5862	0.4925	demam	0.4285	0.4949
sah	4.1963	0.8085	depa	0.5894	0.4920
mar	0.7071	0.4551	famcvd	0.1230	0.3285
age	49.3688	13.0599	smacc	0.8041	0.3968
women	0.5211	0.4996	smkc	0.1736	0.3788
black	0.0337	0.1805	smkevr	0.3921	0.4883
white	0.8550	0.3520	smkex	0.2729	0.4455
noqual	0.1878	0.3906	smkoc	0.1401	0.5106
eduh	0.7749	0.4176	sportm	0.2691	0.4435
scl2	0.3620	0.4806	sportr	0.1227	0.3281
scl3	0.5034	0.5000	hrsspt	1.2022	3.0489
eqvinc	3.1720	2.7412	phy2	0.4046	0.4908
imd3	0.2037	0.4028	phy3	0.3022	0.4593
imd4	0.1768	0.3816	veg	3.8218	2.4619
imd5	0.1336	0.3403	fatt2	0.1520	0.3590
hse04	0.1685	0.3744	fatt3	0.0301	0.1710
bmil	0.5025	3.0435	livehm	160.0535	208.1209
bmih	18.8296	14.3018	urban	0.1792	0.3836

Sample Size=3,381

TABLE 4. Results for the Ordered Logit Model

	β_1	S.E.	β_2	S.E.
mar	0.198*	0.11	0.063	0.09
age	0.000	0.01	0.003	0.01
women	0.105	0.11	-0.035	0.08
black	0.313	0.28	0.405	0.25
white	0.777**	0.22	0.689**	0.17
noqual	-0.862**	0.31	-0.594**	0.22
eduh	-0.695	0.31	-0.281	0.21
scl2	0.066	0.14	-0.202	0.13
scl3	0.219	0.16	0.060	0.13
eqvinc	0.711**	0.26	0.397**	0.16
imd3	-0.076	0.14	-0.139	0.11
imd4	-0.254*	0.15	-0.227*	0.11
imd5	-0.798**	0.16	-0.429**	0.14
hse04	-0.058	0.18	-0.005	0.13
bmil	0.010	0.01	-0.011	0.01
bmih	-0.002	0.00	-0.008**	0.01
drinkun	0.022	0.02	0.017	0.02
agema	-0.004	0.00	-0.001	0.01
agepa	0.009*	0.00	0.002	0.01
demam	-0.231	0.14	0.099	0.11
depa	0.014	0.14	0.101	0.10
famcvd	-0.243	0.16	-0.105	0.13
smacc	0.276*	0.16	0.054	0.12
smkc	-0.271	0.25	-0.237	0.18
smkevr	-0.164	0.20	0.141	0.14
smkex	-0.045	0.21	-0.034	0.15
smkoc	0.007	0.14	0.008	0.09
sportm	0.405**	0.15	0.230**	0.06
sportr	0.522*	0.27	0.419**	0.16
hrsspt	-0.013	0.02	0.026**	0.01
phy2	0.622**	0.12	0.269**	0.10
phy3	0.715**	0.15	0.436**	0.11
veg	0.004	0.02	0.034	0.01
fatt2	0.209	0.15	0.016	0.10
fatt3	-0.032	0.28	0.084	0.23
livehm	-0.000	0.01	-0.000	0.00
urban	0.201	0.14	0.073	0.11
intercept	-0.426**	0.66	-2.761**	0.50

Parameters of vector $\boldsymbol{w}$

	θ	S.E.
lli	1.589**	.07
crp	0.162*	.08
chl	0.172	.09
bpn	0.157**	.07

** Significant at the 5% level;
* Significant at the 10% level.

TABLE 5. Estimated intercepts of model 3

	α	S.E.	Prob.
$U = 1$	0.1615	0.1249	0.54
$chl \mid U = 0$	1.6278	0.0853	0.84
$chl \mid U = 1$	-0.5323	0.0596	0.37
$crp \mid U = 0$	1.8694	0.0852	0.86
$crp \mid U = 1$	0.5100	0.0539	0.62
$bpn \mid U = 0$	1.9212	0.0934	0.87
$bpn \mid U = 1$	-0.1277	0.0556	0.47
$lli \mid U = 0$	0.8031	0.0599	0.69
$lli \mid U = 1$	-0.0378	0.0521	0.49
$sah > 1 \mid U = 0, lli = 0$	-0.1717	0.1109	0.46
$sah > 1 \mid U = 0, lli = 1$	2.5915	0.1517	0.93
$sah > 1 \mid U = 1, lli = 0$	-1.6320	0.1219	0.16
$sah > 1 \mid U = 1, lli = 1$	0.6901	0.1063	0.67
$sah > 2 \mid U = 0, lli = 0$	0.0806	0.1097	0.52
$sah > 2 \mid U = 0, lli = 1$	3.4650	0.2195	0.97
$sah > 2 \mid U = 1, lli = 0$	-1.3091	0.1478	0.21
$sah > 2 \mid U = 1, lli = 1$	1.3354	0.1652	0.79

TABLE 6. Estimated covariates' coefficients of model 3

	γ_u	S.E.	β_1	S.E.	β_2	S.E
mar	-0.1079	0.2216	-0.0159	0.1303	0.1689	0.1172
age	0.1493**	0.0183	-0.0070	0.0103	0.0130**	0.0067
women	-2.5172**	0.2734	-0.0290	0.1366	-0.0849	0.1199
black	-2.2479**	0.5418	0.6449	0.3430	0.1330	0.3403
white	-0.1400	0.3893	1.1140**	0.2352	0.3628	0.2324
noqual	-0.2231	0.6467	-0.5668	0.3942	-0.7306**	0.2591
eduh	-1.0248	0.6119	-0.3372	0.3477	-0.4912*	0.2532
scl2	-0.5605	0.3170	0.0747	0.2156	-0.2078	0.1485
scl3	-0.3788	0.3267	0.1987	0.2198	0.1065	0.1601
eqvinc	-0.8082*	0.2967	0.5116**	0.2334	0.5070**	0.2174
imd3	0.3715	0.2514	-0.1228	0.1463	-0.0855	0.1373
imd4	0.4941*	0.2775	-0.0789	0.1684	-0.3462**	0.1472
imd5	1.5698**	0.3452	-0.8177**	0.2251	-0.4448**	0.1659
hse04	0.2451	0.3162	0.1559	0.1990	-0.2741	0.1878
bmil	-0.0721	0.0468	0.0042	0.0134	-0.0510	0.0399
bmih	0.1355**	0.0119	-0.0083*	0.0046	0.0044	0.0057
drinkun	-0.1297**	0.0456	0.0364	0.0284	0.0057	0.0236
agema	-0.0137	0.0108	0.0083	0.0076	-0.0078	0.0048
agepa	-0.0183*	0.0106	-0.0027	0.0069	0.0082	0.0046
demam	0.0627	0.2542	0.0705	0.1646	-0.0412	0.1427
depa	-0.0534	0.2495	0.0554	0.1482	0.1146	0.1529
famcvd	0.5728*	0.3313	-0.2160	0.2311	-0.1146	0.1499
smacc	-0.6574**	0.3137	-0.1505	0.1971	0.2991**	0.1556
smkc	0.8732**	0.4396	-0.7655**	0.2664	0.2281	0.2349
smkevr	-0.0649	0.3346	0.1928	0.2082	-0.0301	0.1961
smkex	0.0471	0.3558	-0.0078	0.2282	-0.0309	0.1929
smkoc	0.0710	0.2354	-0.0269	0.1363	0.0546	0.1340
sportm	-1.1110**	0.2469	-0.1289	0.1425	0.4725**	0.1519
sportr	-1.1047**	0.3918	-0.1735	0.2262	0.8126**	0.2994
hrsspt	0.0487	0.0324	0.0706**	0.0281	-0.0432	0.0288
phy2	-0.5596**	0.2572	0.5506**	0.1708	0.3364**	0.1229
phy3	-1.3925**	0.2989	0.5664**	0.1809	0.6059**	0.1489
veg	0.0005	0.0400	0.0270	0.0245	0.0344	0.0219
fatt2	-0.0002	0.2652	-0.2242	0.1615	0.2951**	0.1393
fatt3	0.7564	0.5752	0.6370	0.4567	-0.2132	0.2617
livehm	-0.0011**	0.0005	-0.0005	0.0004	-0.0001	0.0002
urban	0.3106	0.2605	0.0596	0.1620	0.1475	0.1448

** Significant at the 5% level;
* Significant at the 10% level.

References

[1] Agresti, A. (1990): *Categorical Data Analysis*. New York, Wiley.

[2] Banks, J., Marmot, M., Oldfield, Z. and Smith, J. (2007): "The SES health gradient on both sides of the Atlantic." *Institute for Fiscal Studies Working Papers*, W07/04.

[3] Bartolucci, F. and Forcina, A. (2006): "A class of latent marginal models for capture-recapture data with continuous covariates", *Journal of the American Statistical Association*, 101, pp. 786-794

[4] Bartolucci, F., Colombi, R. and Forcina, A. (2007): "An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints," *Statistica Sinica*, **17**: 691-711.

[5] Crossley, T.F. and Kennedy, S., (2002): "The reliability of self-assessed health status." *Journal of Health Economics* 21, pp. 643-658.

[6] Dardanoni, V., Forcina, A., and Modica, S. (2008): "Direct causal effects in education transmission", *submitted*.

[7] Etilè, F. and Milcent, C. (2006): "Income-related reporting heterogeneity in self-assessed health: evidence from France." *Health Economics*, 15, pp. 965-981.

[8] Gerdtham, U.G., Johannesson, M., Lundberg, L. and Isacson, D. (1999): "The demand for health: results from new measures of health capital." *European Journal of Political Economy*, 15, pp. 501-521.

[9] Goodman, L. (1974): "Exploratory latent structure analysis using both identifiable and unidentifiable models", *Biometrika*, 61, pp. 215-231.

[10] Groot W. (2000): "Adaptation and scale of reference bias in self-assessment of quality of life." *Journal of Health Economics*, 19, pp. 403-420.

[11] Hernandez-Quevedo, C., Jones, A.M. and Rice, N. (2005): "Reporting bias and heterogeneity in self-assessed health. Evidence from the British Household Panel Survey." *HEDG Working Papers*, 05/04, 2005.

[12] Huang G., Bandeen-Roche K. (2004): "Building an identifiable latent variable model with covariate effects on underlying and measured variables," *Psychometrika* 69(1), pp. 5-32.

[13] Idler, E.L. and Benyamini, Y. (1997): "Self-rated health and mortality: a review of twenty-seven community studies." *Journal of Health and Social Behavior*, 38, pp. 21-37.

[14] Joreskog, K. G. and Goldberger, A. S. (1975): "Estimation of a model with multiple indicators and multiple causes of a single latent variable." *Journal of the American Statistical Association*, 10, pp. 631-639.

[15] Kerkhofs, M., Lindeboom ,M. (1995): "Subjective health measures and state dependent reporting errors." *Health Economics*, 4, pp. 221-235.

[16] Lindeboom, M. and van Doorslaer, E. (2004): "Cut-point shift and index shift in self-reported health." *Journal of Health Economics*, 23, pp. 1083-1099.

[17] Murray, C.J.L, Tandon, A., Salomon, J. and Mathers, C.D. (2001): "Enhancing cross-population comparability of survey results." *GPE Discussion Paper*, World Health Organisation, 35.

[18] Pearl, J. (2000): *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[19] Sadana, R., Mathers, C.D., Lopez, A.D., Murray, C.J.L. and Iburg, K. (2000). "Comparative analysis of more than 50 household surveys on health status." *GPE Discussion Paper*, World Health Organisation, 15.

[20] Shmueli, A. (2003): "Socio-economic and demographic variations in health and in its measure: the issue of reporting heterogeneity." *Social Science and Medicine*, 57, pp. 125-134.

[21] Terza, J.V. (1985): "Ordinal probit: a generalization." *Communications in Statistics*, 14, pp. 1-11.

[22] van Doorslaer E, Jones, A.M. (2003): "Inequalities in self-reported health: validation of a new approach to measurement." *Journal of Health Economics*, 22, pp. 61-87.

[23] Wingard, D. (1984): "The sex differential in morbidity, mortality and lifestyle." *Annual Review of Public Health*, 5, pp. 433-458.

[24] Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data, Cambridge*. The MIT Press.

Facoltà di Economia, Università di Palermo
E-mail address: vdardano@unipa.it

Department of Economics and Related Studies, University of York; Facoltà di Economia, Università di Palermo
E-mail address: plidonni@gmail.com