Società italiana di economia pubblica
XX CONFERENZA

ECONOMIA DELLA TASSAZIONE
Sistemi tributari, pressione fiscale, crescita

Pavia, Università, 25 - 26 settembre 2008

REPORTING EXPECTED LONGEVITY AND SMOKING: EVIDENCE FROM THE SHARE

SILVIA BALIA

pubblicazione internet realizzata con contributo della

COMPAGNIA di San Paolo

Reporting expected longevity and smoking: evidence from the SHARE.

Silvia Balia*

Università di Cagliari and University of York†

July 2007

Abstract

This paper investigates formation of expected longevity in an elderly population. We use Italian data from the early (2004) release of the Survey of Health, Ageing and Retirement in Europe (SHARE). The SHARE provides a numerical measure for subjective survival probability (SSP). To assess internal consistency and investigate validity of SSP as a proxy of actual mortality, we compare SSP to lifetables and look at the variation with health, smoking and socio-economic variables. In a multivariate framework, we propose a recursive model for expected longevity, self-assessed health and smoking duration, where health and smoking variables are potentially endogenous. Unobservable individual-specific heterogeneity is considered by estimating a finite mixture model via the EM algorithm, which allows division of the population according to different latent classes and estimation of class membership probabilities. Our mixture model fits the data better than the single class model and provides evidence of individual unobserved heterogeneity in the formulation of survival expectations. Expectations are shown to vary most with health status, socio-economic characteristics, parental mortality and age. Two-types of individuals in the population are identified, that differ in terms of unobservable frailty and rationality in addiction. We also find differences between current and former smokers in the way they discount future consequences of tobacco consumption on health and mortality risk. Our findings suggest caution in the use of SSP as a proxy of actual mortality.

JEL codes I12 C0 C30 C41

Keywords: subjective survival probability; smoking; beta regression; duration analysis; unobservable heterogeneity; mixture model; EM algorithm.

---

*The author wishes to thank Andrew Jones, Casey Queen, Teresa Bago d'Uva, Martin Forster, Rinaldo Brau, Elisabetta Strazzera and the HEDG group for their suggestions and comments. Data from the early release 1 of the Survey of Health, Ageing and Retirement in Europe (SHARE) 2004 were supplied by the CentERdata.

†Dipartimento di Ricerche Economiche e Sociali, Universitá di Cagliari, viale S. Ignazio, 78, 09123 Cagliari, Italia. *E-mail address*: silvia.balia@unica.it

Beliefs about future events play a central role in the utility maximizing behaviour of rational forward-looking individuals. In the decision-making process, expectations are used to make decisions about savings and investments. In health economics, expectations about survival could be used to study risk-behaviours such as smoking and drug use.

Elicitation of expectations has been used in market research of consumer behaviour and purchasing probability since the Sixties. Juster (1966) finds that survey consumer intentions to buy are inefficient predictors of the actual purchase rate, because the adjectival scale reduces the accuracy of the probability judgment. A quantitative scale is, instead, more precise and the mean value of the distribution of probabilities is a good predictor of future purchasing behaviour. Income and return expectations have been used more recently in models of consumption and saving (Guiso et al., 1992, 2002; Dominitz and Manski, 1997, 2005). According to Dominitz and Manski (1997), the elicitation of probabilistic expectations should be preferred to qualitative questions: problems may arise in the interpretation of qualitative responses, since they are subject to large variation between individuals.

Subjective probabilities have, however, many interesting features. First, the metric has the advantage of providing a numerical scale for responses. This makes it easier to compare responses across individuals relative to the more standard approach, which is based on qualitative judgements. Survey questionnaires that elicit self-reported measures of health, satisfaction or well-being and subjective probability often use a list of adjectives that describe the response (i.e., a range from "very likely" to "somewhat unlikely" can be used in the case of probabilities). Responses may depend on cognitive, linguistic and cultural differences and are usually affected by response bias. This bias is due to a systematic tendency to respond to a range of questionnaire items on some basis other than that which the items were designed to measure. It is generally accepted that the quantitative approach overcomes this limitation. Secondly, with a quantitative measure of expectations it is possible to

assess the internal consistency about different event probabilities, and the external accuracy of the responses. For example, income expectations could be compared with actual income and subjective risk of mortality could be compared either with life tables or observed mortality data to prove external consistency. Several studies show that subjective probabilities have more predictive power than qualitative responses (Juster, 1966; Dominitz and Manski, 1997).

Life-cycle models of consumption usually employ hazard rates from life tables to show how mortality risks influence savings and investment. However, life tables tend to understate survivor probability because they do not consider that health improvements can increase longevity. Subjective survival probability is broadly considered a better predictor of future mortality than objective life table hazard rates. Hurd and McGarry (1995) show that subjective survival probabilities (SSP) in cross-sectional data from the Health and Retirement Survey (HRS) are internally consistent: they find the same average survival than in life tables. They also find that SSP covaries well with socio-economic variables and risk factors as actual mortality does; it is highly correlated with health and death of a parent. Interestingly, they find that education, income, wealth and smoking have a stronger effect on self-assessed health which, in turn, has a notable effect on expected longevity. Moreover they find only little systematic variation of SSP with respect to covariates.

In subsequent work Hurd et al. (1999), using the Asset and Health Dynamics Study (AHEAD) and Hurd and McGarry (2002), using the HRS panel, stress the role of unobservable heterogeneity in SSP rather then life tables, which only capture the effect of observed factors on mortality. They claim that, despite being correlated with health and the onset of diseases, SSP is not simply an alternative measure of overall health status: it has an element of expectations that accounts for most of the unobservable heterogeneity, understanding which would help in the estimation of life-cycle models. Using HRS data, Smith et al. (2001) confirm that individual survival probability reflects health changes due to health shocks and the onset

of new limitations; it also responds to events that, according to epidemiologists, increase the probability of dying. Observed deaths would be "signalled" through lower expectations, even after controlling for health.

This paper aims to assess internal consistency and investigate validity of SSP, elicited in the Survey of Health, Ageing and Retirement in Europe (SHARE), as a proxy of actual mortality in the elderly population. Data from the early (2004) release of the wave 1 of the SHARE are used.[1] The first wave of the SHARE, in fact, does not contain any information about deaths, but has the interesting feature of providing information about the expected longevity. SSP will be compared to lifetables and variation with health, smoking and socio-economic variables will be explored. In a multivariate framework, we propose an empirical model for expected longevity that investigates the determinants of subjective survival expectations.

The main focus of interest is to explain formation of expected longevity looking at the effect that smoking behaviour might have on people's perception of risk. Recent works have used survey data to measure the effect of smoking cigarettes on perceived risk of onset of diseases such as lung cancer (see e.g., Viscusi, 1990). Little evidence exists on the impact of smoking prevalence and intensity on perception of mortality risk. Using the HRS data, Schoenbaum (1997) studies the effect of being a smoker (distinguishing between never, former, current light and current heavy smokers) on individuals expectations of reaching age 75. His main finding, in contrast with Viscusi's results, is that at heavy smokers tend to underestimate the negative effect of smoking intensity (in terms of number of cigarettes smoked) on expected longevity. Here, we distinguish between never, current and former smokers

[1]This release is preliminary and may contain errors that will be corrected in later releases. The SHARE data collection has been primarily funded by the European Commission through the 5th framework programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life). Additional funding came from the US National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064). Data collection in Austria (through the Austrian Science Fund, FWF), Belgium (through the Belgian Science Policy Office) and Switzerland (through BBW/OFES/UFES) was nationally funded. The SHARE data set is introduced in Börsch-Supan et al. (2005); methodological details are contained in Börsch-Supan and Jürges (2005).

as well, and exploit the nature of the dataset to measure the effect of the numbers of years spent smoking on reporting SSP.

1. The framework

It is a common belief that ageing shrinks socio-economic differences because genetic and biological factors dominate other determinants of survival probability. Despite a downward trend in observed aggregate mortality however, observed individual mortality in the "oldest" old is still under the influence of socio-economic and lifestyle differences. For example, the impact of unhealthy behaviours, such as heavy smoking and drinking, can be stronger at older ages, when the negative effects on health are more likely to manifest themselves. Empirical research shows that unhealthy individual lifestyles are usually positively related to the onset of chronic diseases and largely explain the observed variation in health and longevity in the population.[2] Here we try to see how this reflects in formation of individual expectations about survival. In particular, we want to see to what extent expected longevity can be explained by smoking behaviour.

The health economics of smoking explain smoking behaviour according to two alternative theories. One theory defines smokers as irrational or myopic (see Thaler and Sheffrin, 1981; Winston, 1980). When optimizing their utility, individuals make short-term plans, since they care more about present satisfaction than about the future. For myopic individuals, tobacco consumption is a commodity that enhances instantaneous utility: negative effects on future health and well-being are not internalized. According to the other theory, smokers are instead rational and forward-looking (Becker and Murphy, 1988), taking into account future effects of their deci-

[2]Everyday behaviours, such as smoking, drinking, eating and sleeping, have been shown to be important determinants of individual morbidity and mortality (see Balia and Jones, 2007). The epidemiological evidence suggest that there is a strong linkage between tobacco consumption, cancers, vascular and respiratory diseases, and mortality (see e.g., Vineis et al., 2004; Peto et al., 2005).

sions. This means that the detrimental effects of smoking on health are internalized. In the "rational addiction model", the leading economic model on smoking habits, Becker and Murphy argue that rational smokers decide to smoke if the benefits outweigh the costs of smoking, and present-oriented (myopic) individuals are potentially more addicted than others.

More recently, Arcidiacono et al. (2007) have investigated whether models of forward-looking behaviour explain heavy smoking and drinking better than models of myopic behaviour in the elderly, taking into account unobservable heterogeneity. Assuming that the myopic model can be simply nested within the forward-looking model, they provide a description of the profile of each behaviour. They find that, overall, both models predict decreasing smoking rates: smoking is less attractive as individuals age, because more illnesses occur and health worsens. Sharp declines are predicted by the fully rational model. However, between the age of 50 up to the age of 62 fully rational individuals smoke more than myopic individuals; after the age of 62 and up to the age of 80, smoking rates are higher in the myopic model. At that cut-off age of 80 there is, according to their simulations, an upward trend in smoking behaviour for forward-looking individuals. Fully rational individuals feel that they are at the end of the life-cycle, therefore they decide to enjoy the time left smoking. This is what Becker and Murphy would defined as a rationally myopic attitude of older people. They are less concerned with the future effects of smoking on health. They also find that this "end-of-life effect" as well as the degree of sensitiveness to medical care improvements can influence the risk of dying. In particular, they find that death rates increase at older ages, in the forward-looking model, even after medical advances, due to the higher smoking rates.

Furthermore, smoking behaviour can have different dynamics depending on age. The youngsters, for example, are more likely to experiment with smoking and are often subject to strong peer influences (Orphanides and Zervos, 1995; Kremers et al., 2004). In addition to that, for them there is still a perception of long lifespan during

which they can compensate for the effect of smoking by diversifying their investments in health.

Hence, heterogeneity within the group of smokers needs to be taken into account to have a better understanding of how individuals form their expectations about longevity. Our analysis is an attempt to consider these differences in a model that exploits information about expected longevity, self-reported general health status and duration of smoking.

We propose a structural equation model for expected longevity where health and smoking are potentially endogenous. Subjective probability of surviving to some future age is explained by observed individual characteristics and the presence of unobservable individual-specific heterogeneity is taken into account, in order to give a consistent measure of the impact of health, smoking behaviour and socio-economic factors on expected longevity. Factors such as genetics, past experiences, tastes, risk aversion and individual rates of time preference, which are usually unknown or not revealed in survey questionnaires, could influence the duration of smoking, the probability of being in good or excellent health as well as the perception of the chances of living longer. Unobservable heterogeneity is also a special matter of concern in duration analysis (see van den Berg, 2001).[3] Neglecting heterogeneity leads to biased estimates of the aggregate hazard function. In a duration model for smoking, negative duration dependence would be overestimated, meaning that the hazard of quitting is estimated to fall faster then the actual hazard rate. This motivates modelling unobservable heterogeneity.

Finite mixture models allow control over unobservable heterogeneity in the underlying population, relying on very simple assumptions. The sample of respondents

[3]Heterogeneity exists at the beginning of the observed period and falls to zero over time, because only people with low heterogeneity stay in the state of smokers (individuals with high heterogeneity will quit sooner). Alternatively, unobservable heterogeneity can be very low or null at the beginning of the observational time and then it can expand over time. Frijters et al. (2005) suggest that unobservable heterogeneity follows a precise path for each individual according to his health shocks. Each individual is subject, during her lifespan, to persistent health deteriorations which vary and tend to cumulate over time. These health shocks are often unmeasured or unobservable.

is assumed to be drawn from a population that consists of a finite number of sub-populations, or latent classes, from which each data point is drawn randomly. Class membership is not observed, so the heterogeneity problem is simplified to the omission of an indicator of membership to a population type and the genuine impact of determinants of expectations can be recovered. Indeed, the main advantage of the finite mixture approach is that the underlying continuous mixing distribution does not need to be parametrically specified. Heterogeneity is, in fact, approximated by a discrete distribution. We use the expectation-maximisation (EM) algorithm, which deals in particular with missing and incomplete data, to estimate our recursive model for expected longevity, health and smoking using maximum likelihood (see Schafer, 1997, for an extensive review of the method).

2. Data and variables

The SHARE is part of a new project recently designed in response to the Special EU Council in Lisbon, March 2000. The main objective of the project is to create a large European longitudinal survey that gives the scope for studying in depth how to cope with ageing in different cultures, economies, welfare and health care systems in Europe. One of the most interesting feature of this survey is that it provides an indicator of subjective survival probability, which has so far been absent from European surveys. This work uses the Italian cross-sectional data from wave 1 of the SHARE. Italy is one of the European countries with the highest ageing rate and old age dependency ratio. Cardiovascular diseases represent the main cause of mortality, being responsible of about 44 per cent of total deaths. This makes it particularly interesting to see how survival expectations of the oldest Italians are formed and influenced by past and current smoking behaviours.

The target population of the SHARE is made up of non-institutionalised individuals 50 years of age and over.[4] The first wave of the survey was carried out

[4]Only those individuals that survive at age 50 in 2004 are included in the analysis. This could

in 2004, between April and October.[5] For Italy, 2559 individuals were interviewed including 1132 males and 1427 females. Respondents in the SHARE are all household members aged 50 and over, plus their spouses, who may be younger. Only 53 individuals below 50 years old were interviewed. The individual response rate was 79.9%. The questionnaire is divided into 20 sections and offers a picture of individuals' health, lifestyles, families, social networks and economic situation. A specific agency for each country worked on data collection, while the programming of the individual instruments was done centrally by CentERdata, a survey research institute affiliated with Tilburg University in the Netherlands. The data were collected using a computer assisted personal interviewing program (CAPI), supplemented by a self-completion paper and pencil questionnaire.[6] For the statistical analysis in this study, the original sample has been reduced to 1837 individuals, including primary and secondary respondents. The sample has been first reduced according to item non response: only individuals who answered all the questions relevant for the analysis are in sample.[7]

To elicit SSP the SHARE questionnaire uses a format analogous to the one used in the HRS. This has the advantage of helping recovery of the probability distribution of the uncertain event "time to death". The question is as follows: "What are the chances that you will live to be age T or more?". A different age T is proposed to each age class of the respondents. In particular, as reported in Table 1, the target ages are such that the distance from the current age is between 6 and 25 years. As an example, we can consider a person aged 58 in 2004. For this individual we

introduce mortality selection bias, with those with higher heterogeneity dying before the survey, leading to a more homogeneous population. If heterogeneity in the population is proven to exist, then mortality selection bias should not be a matter of concern.

[5]So far the SHARE data consists of one wave relative to 2004. The next waves of the survey will provide information about exit from the sample of those individuals interviewed in 2004. The panel structure of the data will give scope for longitudinal and lifespan duration analysis.

[6]More information about the design of the survey and some cross-countries statistics are available in the SHARE Project webpage http://www.share-project.org/ where official documentation can be found.

[7]Unreliable income records, i.e. 6 observations that report an equivalised household income higher than 1,150,000 Euro, have been eliminated.

Table 1
Target ages in the subjective survival probability question

Age class of the respondent	Target age
50-55	75
56-60	75
61-65	75
66-70	80
71-75	85
76-80	90
81-85	95
86-95	100

will know the subjective probability of surviving up to age 75. However, the Italian data present some incongruity with that: people of the same age class are asked to evaluate survival probability at different target ages. This could be a problem in the interpretation of the distribution of probabilities. Hence, those observations (87 individuals) that do not correspond to the age - target age association have been eliminated. An additional concern is that the SHARE sample is not asked to evaluate the chances of surviving for the same number of years, which would have made interpretation of probabilities easier. Instead, the difference between current and target age decreases for older ages. This requires conditioning of the distribution of survival probabilities on age and target ages.

The SHARE collects individual perceptions of risk for a battery of future events. Beside the risk of mortality, these events include the future value of retirement (in terms of pension and retirement age), the risk of receiving or leaving bequests and the quality of life in the future. Respondents are asked to assess the subjective probability of some events occurring.Questions are accompanied by cards, with a numerical sequence of probability, which are shown to the respondent. The answer to the questions is a number from 0 to 100. The card shows that a value of 0 means *absolutely no chance* and a value of 100 means *absolutely certain*. The question on SSP is the ninth of 11 questions about expectations. This is encouraging because previous questions should make answers more correct by introducing respondents

to the probabilistic format. To ensure reliability of the responses at the question on SSP, however, individuals whose answers to the whole battery of expectation questions were unclear have been eliminated using a criterion based on some of the other questions about expectations in the SHARE questionnaire. We focus in particular on two questions which ask the respondents to self-evaluate the chance that the standard living will be better and the chance that the standard living will be worse. A subjective probability of 0 for both events means that there's a high expected probability that the standard of living will be unchanged. Some doubts about the coherence of the two answers may arise if the probabilities sum to greater than 1.[8] This would indicate that the respondent is unreliable in their self-assessment of probability, generally, including survival. Assuming that respondents can still give coherent subjective probability even if they do not sum to 1, due to some problem in using a numerical scale or misunderstanding of the questions, we use a tolerance level of 0.10 that allows us to drop 111 individuals for whom the sum of the probabilities is higher than 1.10.

The SHARE data offers a wide range of indicators of physical health, all self-reported. In the empirical literature, a commonly used measure of general health status is self-assessed health (SAH, Deaton and Paxson, 1998; Smith, 1999). Self-assessment of health is known to be a good indicator of morbidity and a powerful predictor of future health and mortality (Idler and Kasl, 1995; Idler and Benyamini, 1997; van Doorslaer and Gerdtham, 2003). Individuals are asked to assess their health status according to a qualitative scale. In the SHARE, respondents were initially randomised to answer the SAH item either at the beginning or at the end of the physical health section of the questionnaire. The question is "How is your health?".[9] Health can be very good, good, fair, bad, very bad. We use a

[8]For example, it is not possible that the probability of being better next year is 50% and the probability of being worse is 50% as well, or even more.

[9]The question is very simple and, with respect to many similar questions in other surveys as the British Health and Lifestyle Survey or the British Households Panel Survey, it does not require comparison of your own health to the health of a person of you age.

Table 2
Variable definitions

Variable name	Variable definition
ssp	subjective survival probability, range 0-100
sah	1 if self-assessed health is good or very good, 0 if less than good
chronic	1 if has 2 or more chronic diseases, 0 if has less
ever started	1 if did smoke for at least one year, 0 never smoked
quitter	1 if did quit smoking, 0 otherwise
smy	numbers of years smoking, 0 if never smoked
healthy	1 if prudent drinker, does exercise, is not obese, 0 otherwise
income	household income divided by household size
low quartile	household income lowest quartile
low2 quartile	household income second lowest quartile
above median	household income above median level
education	number of years of education
single	1 if single, 0 if living with spouse or partner
retired	1 if retired, 0 otherwise
employed	1 if worker, 0 otherwise
unemployed	1 if unemployed, 0 otherwise
homemaker	1 if housekeeper, 0 otherwise
sick	1 if absent from work due to sickness, 0 otherwise
house owner	1 if own house, 0 otherwise
household size	number of other people in the house
mother dead	1 if mother died, 0 otherwise
father dead	1 if father died, 0 otherwise
agemth	age mother died
agefth	age father died
male	1 if male, 0 otherwise
age	age in years
age50-65	individuals aged 50 to 65
age66-70	individuals aged 66 to 70
age71-75	individuals aged 71 to 75
age76-80	individuals aged 76 to 80
age81-85	individuals aged 81 to 85
age86 more	individuals aged 86 or over
target age (T)	target age in the subjective survival probability question
(T - age)	distance between current age of the respondent and target age

Table 3
Sample means in subgroups of subjective survival probability

Variable	Full sample	if ssp<50	if ssp=50	if ssp>50
ssp	65.763	18.288	50.000	87.119
sah	0.512	0.238	0.511	0.601
chronic	0.439	0.625	0.427	0.385
ever started	0.454	0.395	0.432	0.481
quitter	0.566	0.544	0.525	0.586
smy	29.821	32.978	29.608	29.066
habits	0.459	0.401	0.465	0.475
income	22168.946	21102.984	22639.947	22326.620
low quartile	0.247	0.288	0.267	0.226
low2 quartile	0.250	0.270	0.239	0.248
above median	0.503	0.442	0.494	0.526
education	7.181	6.006	7.217	7.543
single	0.186	0.250	0.212	0.155
retired	0.555	0.590	0.573	0.537
employed	0.198	0.090	0.177	0.240
unemployed	0.019	0.026	0.014	0.019
homemaker	0.217	0.267	0.227	0.197
sick	0.011	0.026	0.010	0.007
house owner	0.541	0.456	0.582	0.552
household size	2.555	2.424	2.547	2.601
mother dead	0.788	0.872	0.771	0.768
father dead	0.918	0.951	0.928	0.904
agemth[a]	74.945	74.350	74.548	75.316
agefth[a]	71.062	70.544	71.141	71.205
male	0.465	0.422	0.461	0.480
age	63.913	67.529	63.950	62.741
age50-65	0.613	0.424	0.611	0.674
age66-70	0.162	0.169	0.158	0.161
age71-75	0.120	0.186	0.131	0.094
age76-80	0.070	0.131	0.060	0.055
age81-85	0.025	0.052	0.031	0.014
age86 more	0.010	0.038	0.010	0.002
target age (T)	78.819	81.657	78.854	77.896
(T - age)	14.905	14.128	14.905	15.155
sample size[b]	1837	344	419	1074

Notes:

[a] Individuals whose mother died are 1448; individuals whose father died are 1687.

[b] The proportion of quitter is calculated for the subsample of individuals who ever started smoking.

dichotomised version of SAH that takes the value 1 if health is good or very good, and 0 otherwise.[10]

The SHARE data provides information about health behaviours, including tobacco smoking habit (cigarettes, pipe, cigars, cigarillos). For each individual it is possible to know if they have ever smoked at least for a year, whether they are current smokers, have stopped smoking and the number of years for which they smoked. The latter smoking indicator gives us the scope for employing duration analysis techniques.

Table 2 shows variable definitions and Table 3 reports sample means. About 47 per cent of the respondents are male. The mean age is around 64 years. The mean income is 71,200 euros but half of the sample has an income lower than 14,595 euros.[11] On average individuals have studied for 7 years. Half of the respondents feel they are in good or very good health and around 46 per cent of them have an healthy style of life (i.e., drink prudently, are not obese and usually do sport). However 45 per cent individuals did start smoking and smoked for at least one year. On average they smoked for nearly 30 years. Around 19 per cent of the individuals live alone, either because single, widowed, separated or divorced. On average each household has 2.5 components. As for parental mortality, 79 per cent of respondents' mothers and 92 per cent of respondents' fathers are already dead. On average age at death of the mother is higher than age at death of the father, and this reflects healthy life expectancy at birth.

[10] In a work on French data, Etilè and Milcent (2006) find that a way to reduce reporting heterogeneity in SAH, so that SAH can be used as a reliable approximation of true health, is to convert the ordered variable in a binary variable. In this particular case, they suggest to use poor SAH as the reference category.

[11] In order to adjust for the household size, household income is scaled by dividing the value of income by the square root of the number of persons in the household.

3. The distribution of subjective survival probability

Besides the advantages of using subjective survival probabilities as a proxy for mortality risk, including its appeal as an indicator when actual data on deaths are missing, there are a few drawbacks. Subjective survival probability is related strictly to one's ability to think in a probabilistic manner. Individuals might internally associate a qualitative value to an interval of numbers and express their expectation in relation to that, for example numbers smaller than some threshold can be associated to a very low chance of realization of that event. For those who find it easier to think in terms of lower or higher probabilities however, probability and a wide numerical scale provide an incentive to choose a point when an interval would be preferred.

The section of the SHARE questionnaire dedicated to expectations includes a warm-up question, asking "What do you think the chances are that it will be sunny tomorrow?". This should help respondents feel at ease with the numerical scale used in the whole set of subjective expectations questions (see pp. 332-338, Börsch-Supan et al., 2005). Since the survival probability question comes after eight questions about expectations, respondents should be already familiar with thinking probabilistically. Two other issues arise when dealing with numerical answers. The first is known as rounding to focal values. The evidence is that respondents are more likely to choose values that ends with a zero. In particular, responses usually heap at focal values as 0, 50 and 100. Bruine de Bruin et al. (2002) argue that a 50 per cent response may reflect "epistemic uncertainty", where the respondent does not have any expectations at all and the response would be equivalent to "don't know". However, and this is the second issue, Hill et al. (2005) claim that this should not preclude the possibility that an answer of 50 per cent is a true probability. The event of dying can be equally likely to occur or not to occur.[12]

[12]Here it can be appropriate to assume that only rational individuals can understand the probabilistic questions and give answers that respect the metric. Otherwise the warm-up question and the design of the questionnaire would not be of help per se. Rationality is also at the basis of the 50 per cent response since it is the natural expression of rational uncertainty according to the

The use of a self-reported indicator of survival probability requires some particular attention. When the response depends upon a numerical scale, there can be a problem of rounding to focal values. The indicator of SSP in the SHARE takes 29 different values in the range (0, 100). Table 4 shows the frequency of responses. As expected responses are heaped at some foci. Around 3.54 per cent of the sample give a 0 per cent survival probability, 22.73 per cent respond 50 per cent and 23.98 per cent respond 100 per cent. This could represent a lack of variability in the distribution of SSP.

People who give a 50 per cent probability of surviving to some target age and beyond might, in fact, be more oriented towards a "don't know" answer: 50 is the number that better represent uncertainty in a numerical scale. Assuming rationality, the individual who is uncertain would consider the chances to live longer equal to the chances to die and would tend to answer 50 per cent. Comparing the subgroups of individuals who tend to give probabilities lower than 50, equal to 50 and higher than 50 could help to find systematic differences in the responses between individuals. Table 3 shows sample means in these three subgroups. 60 per cent of individuals who self-report a survival probability higher than 50 per cent, also report a good or very good health status. Self-assessed health is lower for subgroups who report lower probabilities and the proportion of respondents with good or very good health decreases monotonically moving from the highest probabilities to the lowest. This trend is confirmed by number of chronic diseases and symptoms in the three subgroups. Concerning risk factors, the table show that the proportion of individuals who undertake three or more healthy behaviours is smaller in the group with lower expected probability of survival.

Interestingly, Table 3 shows that smokers are less concentrated in the low survival group and are more concentrated in the high survival one but the duration of smoking is longer in the lower survival probability group. This apparently coun-

classical utility theory.

Table 4
Subjective survival probability

Response	Frequency	Percentage
0	65	3.54
1	5	0.27
2	1	0.05
3	1	0.05
5	7	0.38
7	2	0.11
8	3	0.16
9	2	0.11
10	66	3.59
15	3	0.16
20	69	3.76
25	2	0.11
30	67	3.65
40	50	2.72
45	1	0.05
50	419	22.81
51	1	0.05
55	1	0.05
60	73	3.97
65	1	0.05
70	133	7.24
75	7	0.38
80	260	14.15
85	4	0.22
90	129	7.02
95	10	0.54
98	1	0.05
99	18	0.98
100	436	23.73

terintuitive result might reflect cognitive dissonance, that is, smokers reduce their belief about the negative effect of smoking on health (Chapman et al., 1983). Therefore, further investigation on the effect of smoking on expected longevity is needed. Socio-economic variables, in particular income, being employed and years of education indicate that socio-economic status has a negative impact on perceived mortality risk. Retirement seems to capture the effect of age on the risk of mortality: in the group of low survival probabilities about 60 per cent of individuals are retired against about 54 per cent in the group with high probabilities. The size of the household is assumed to be an indicator of social capital: larger families may

ensure more support and care for the elderly and sick. However, the composition of our three subgroups does not vary much in relation to the size of the household. Unsurprisingly, the proportion of mothers who have already died is smaller than the proportion of fathers and individuals who report a SSP lower than 50 per cent have, overall, a higher proportion of mothers and fathers who have already died with respect to those whose SSP is higher than 50 per cent. Sample means seem to suggest that a 50 per cent response to the survival expectation question is a rational answer since the subgroups differ according to health status, risk factors, socio-economic characteristics and demographics, with increasing SSP for better health, lifestyle and socio-economic status.

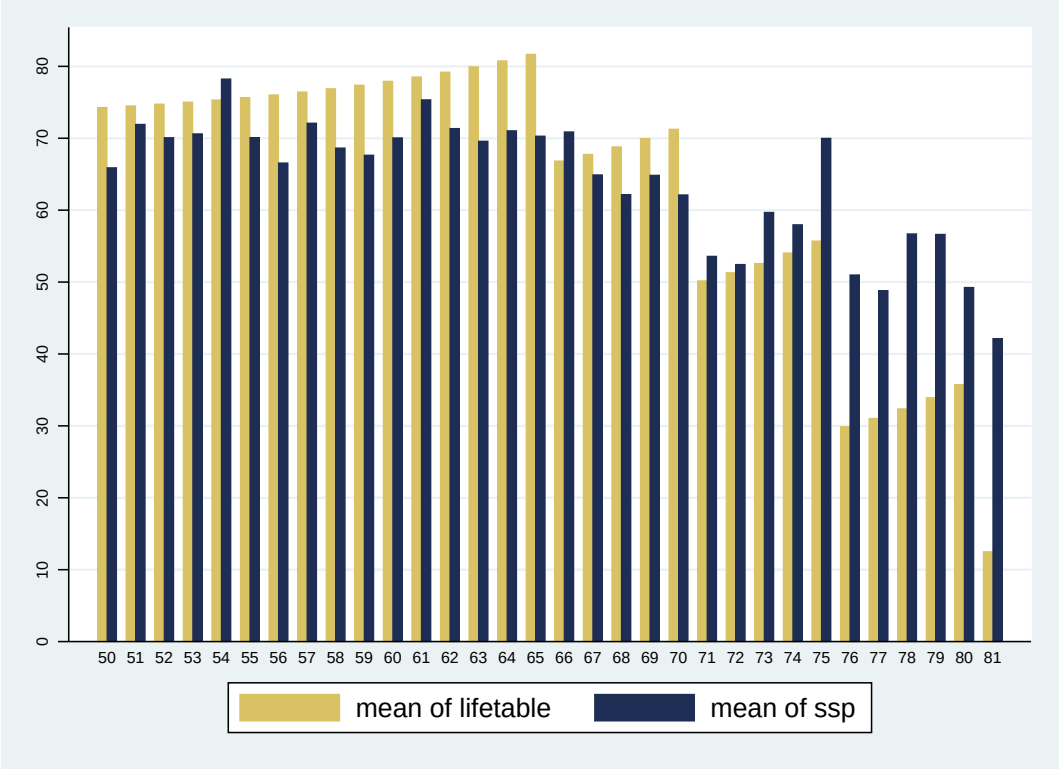
3.1. *Comparison with life tables*

In order to assess the internal consistency of SSP we compare it with survival probabilities from life tables. Data on observed mortality rates for 2004 are not yet available. The Human Mortality Database (HMD) provides death rates and life tables for many countries, including Italy, according to the year of death, the cohort of birth and for each age per year up to 2001.[13] We use life tables for the period 1990-2001 to construct the population counterpart of SSP in the SHARE sample. Mean SSP for each age in the sample is compared to the mean survival probability to the corresponding target age from the life tables over the last 12 years, as in Börsch-Supan et al. (2005).[14] Individuals older than 81 are excluded because of the low frequency in the sample. The histogram in Figure 1 shows that, until age 74, subjective survival probabilities correspond very well with life tables. For ages 56 to 65 there is still a clear correspondence, although the subjective assessments are slightly lower relative to life tables survival probabilities.

[13]Free access to the data is possible from the website http://www.mortality.org.

[14]Since the HMD calculates number of survivors at each age as recent as 2001, we have decided to construct the probability of surviving to each target age for individuals of the same age as the SHARE sample across the last 12 years because they are close to 2004 in term of mortality risk and health shocks. A better comparison would need the 2004 life table.

Figure 1
Subjective survival probability and life tables

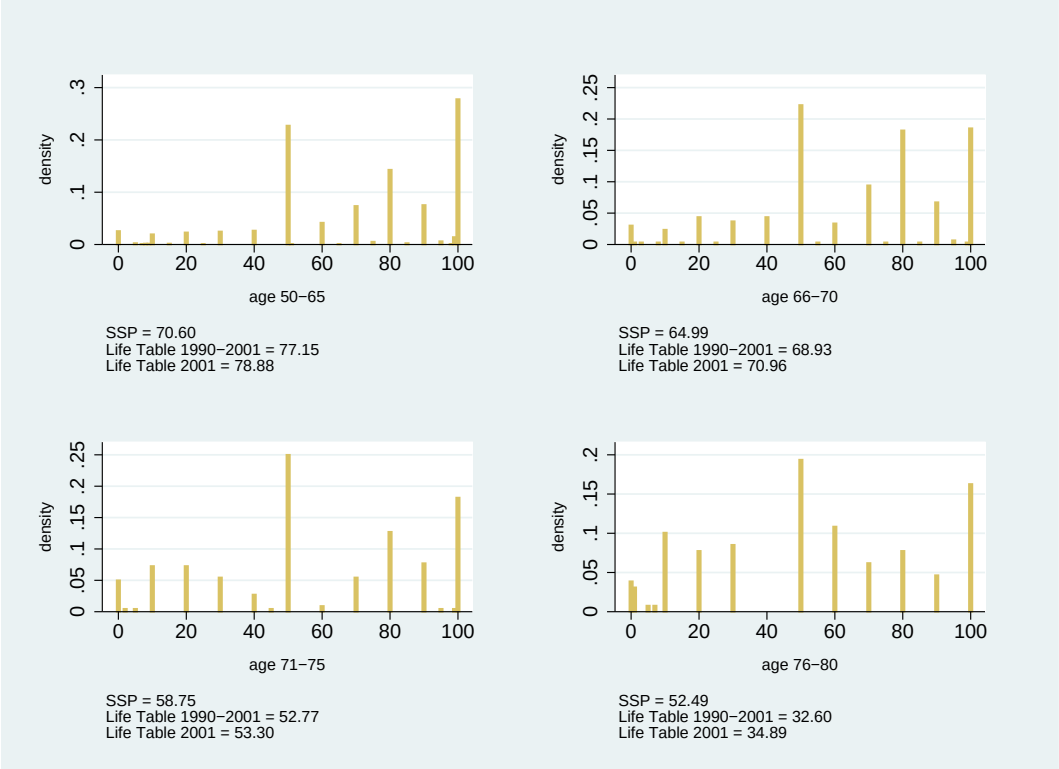


Survival probabilities computed using life tables still give an aggregate measure of survival that does not take account of observable and unobservable heterogeneity. For ages 71 to 81, subjective probabilities of surviving are higher than predictions from life tables, especially for ages older than 75. This might reflect positive health shocks and increasing life expectancy of the elderly in the most recent years. Life tables may not approximate well subjective survival probability of a cohort because probability of surviving increases due to improvements in mortality rates (Hurd and McGarry, 1995). Furthermore, the SHARE might accidentally have collected information about very healthy individuals who have, indeed, a high perception of their survival. This would explain why life tables tend to underestimate survival probability at older ages. The distribution of subjective probabilities also depends upon observable and unobservable differences in mortality risk factors; it may be interesting to see how SSP covaries with health and risk factor in a regression framework.

Figure 1 also shows that the shape of the distribution of SSP reflects the pattern

Figure 2

Subjective survival probability by age classes



of target ages. The jumps in the distribution correspond to the age classes to which different target ages are proposed. This suggests that subjective survival probability is conditional on a function of age and target age. This pattern is better illustrated in Figure 2, where histograms of responses are reported for age classes according to the target age proposed. For each age class the average SSP is compared with the relative average survival probability from the life tables for period 1990-2001 as well as for 2001 separately since this is the most recent. The histograms show a high frequency of the 50 per cent response, especially for individuals aged 71 to 75 years old. However, the average probability dramatically decreases with age, from about 71 per cent for age 50-65, to 65 per cent for age 66-70, to 59 per cent for age 71-75 to 53 per cent for age 76-80. As for the comparison with life tables, the average values of SSP are quite close with life table survival probabilities for ages 66-70 and 71-75. For the first age group, 50-65, SSP tends to overestimate slightly mortality risk. This might be due to the fact that, for this age class, the difference between

Table 5
Subjective survival probability and health by age groups

Self-assessed health	Average response
very good	
age50-65	77.081
age66-70	74.737
age71-75	68.000
age76-80	80.000
good	
age50-65	74.623
age66-70	74.590
age71-75	68.826
age76-80	59.857
fair	
age50-65	66.227
age66-70	61.440
age71-75	57.367
age76-80	56.607
bad	
age50-65	51.362
age66-70	50.868
age71-75	38.194
age76-80	33.600
very bad	
age50-65	47.125
age66-70	28.500
age71-75	54.286
age76-80	32.625

current age and target age ranges from 10 to 25, while for the other classes they are much closer. Instead, survival from life tables appears to be much lower at older ages, particularly for ages 76-80.

3.2. *Variation with health, risk factors and socio-economic variables*

SSP varies according to health status, risk factors, socio-economic characteristics and demographics. Observed variation in subjective survival probability should be in line with epidemiological evidence on the effect of health status and risk factors on mortality risk.

In Table 5 the sample is divided in groups according to individual health status. Average SSP is reported for each age class, but for individuals aged 80 or over

Table 6
Subjective survival probability and smoking by age groups

Smoking status	Average response
current smokers	
age50-65	70.942
age66-70	51.738
age71-75	64.839
age76-80	50.111
formerly smokers	
age50-65	74.858
age66-70	67.539
age71-75	58.517
age76-80	52.367
never smokers	
age50-65	68.104
age66-70	66.525
age71-75	56.504
age76-80	52.803

because of low cell frequency. For each age class there is a dramatic change in expected longevity depending on whether health status is reported as very good or very bad. As expected average SSP is higher for younger individuals who report very good health. Individuals aged 76 to 80, who report a very good health, also report the highest SSP in the sample. This suggests that individuals who have survived to old ages and have no health problems are optimistic about their survival to older ages.[15] It is not clear why individuals in good health of age 71-75 report higher SSP with respect to individuals of the same age but in better health, and why individuals of the same age class in very bad health report the highest SSP in the group of individuals in the same SAH category.

For risk factors, Table 6 shows that average SSP is higher for those who have never smoked relative to current smokers, if they are aged 66 to 80 and only slightly higher than for formerly smokers older than 76 years. Generally, former smokers tend to report higher expected probability of survival at all ages below 71, suggesting that further investigation is needed.

[15]This also could be interpreted as evidence of survivor bias, which we would expect.

Comparing SSP in different socio-economic groups, as in Table 7, suggests that expected longevity is positively related to socio-economic status. Higher income individuals report higher expected longevity. For the most educated individuals, in terms of years spent at school, SSP is higher than for non educated individuals in each age class. However, there is not a clear gradient within the education groups. Education could therefore capture the ability to understand the question. This would partially explain the gradient in the less educated group.

Overall, descriptive analysis of the data shows a quite clear positive relationship between socio-economic factors and expected longevity. The link between self-reported health, and smoking, and subjective survival probability, however, is less clear. This justifies the use of multivariate analysis to better investigate this link. Controlling for unobservable heterogeneity would be desirable to recover the genuine relationship between SSP and its determinants.

3.3. *Variation with parental mortality*

Investigating the validity and the predictive power of subjective survival probability, Hurd and McGarry (1995, 2002) find that self-reported survival probability declines with the death of a parent. Mortality experiences of the parents, however, do not have a direct effect on self-reported health status, rather they can have an effect through deaths caused by genetic diseases.

We look at variation of average SSP with parental mortality in the SHARE data. Table 8 shows that average perceived survival is higher if parents are still alive and if they have died between age 61 and 80. This results may at first seem puzzling. One hypothetical explanation is that parents' early and late deaths are not related to the survival of the children. They depend, in turn, on accidents and the ageing process.

Table 7
Subjective survival probability and socio-economic variables by age groups

Socio-economic status	Average response
lowest income quartile	
age50-65	66.183
age66-70	61.103
age71-75	56.629
age76-80	47.682
second lowest income quartile	
age50-65	69.205
age66-70	67.277
age71-75	57.029
age76-80	57.216
above median income	
age50-65	72.789
age66-70	64.911
age71-75	60.225
age76-80	53.146
max no. of years of education	
age50-65	76.900
age66-70	83.125
age71-75	72.500
age76-80	76.000
no education	
age50-65	57.692
age66-70	62.667
age71-75	70.455
age76-80	46.667

Table 8
Subjective survival probability and parents' mortality

Mother's mortality	Average response	Father's mortality	Average response
alive	71.753	alive	73.160
dead	64.154	dead	65.106
age <=50	64.402	age <=50	65.314
age 51-60	66.395	age 51-60	68.160
age 61-70	63.174	age 61-70	64.725
age 71-80	62.420	age 71-80	64.089
age 81-90	65.238	age 81-90	67.062
age 90 more	64.725	age 90 more	68.380

4. The model

Determinants of expected longevity are investigated by estimation of a recursive system of equations that describe individuals survival probability, health status and

smoking behaviour. The model is defined as:

$$t_{sm} = f(\mathbf{X}_{sm}, \boldsymbol{\mu}), \quad (1)$$

$$H = f(\mathbf{t}_{sm}, \mathbf{X}_H, \boldsymbol{\mu}), \quad (2)$$

$$SSP = f(\mathbf{t}_{sm}, H, \mathbf{X}_{SSP}, \boldsymbol{\mu},). \quad (3)$$

where t_{sm} is the duration of smoking, H is general health, SSP is expected longevity, X are exogenous variables and $\boldsymbol{\mu}$ is some unobservable component.

4.1. *A model for subjective survival probability*

Equation (3) describes the probability of surviving at future ages using a structural equation. Due to the nature of the variable SSP, which is continuous and bounded (between 0 and 100), and to the violation of normality of the errors arising from the presence of spikes in the distribution at certain response foci, we assume that a beta distribution can be used to better fit the data. We estimate a beta regression model by Maximum Likelihood (ML), where the likelihood function for each individual i can be written as:[16]

$$L_i = \frac{\Gamma(\omega + \tau)}{\Gamma(\omega)\Gamma(\tau)} y_i^{(\omega-1)} (1 - y_i)^{(\tau-1)}$$

where y_i is rescaled SSP so that it lies in the interval $(0, 1)$, Γ is the gamma distribution, and ω and τ are shape parameters, which need to be expressed as a function of observed covariates to be interpreted in terms of conditional expectations.

Therefore, ML estimation depends upon the reparametrisation of the likelihood function such that a location sub-model and a precision sub-model can be estimated (see the Appendix for details). The location sub-model is based on a logistic representation for the expected value of the dependent variable, but other link functions can be used. And this has the advantage of interpreting estimated coefficients as

[16]For the sake of simplicity, the expression for the likelihood does omit the unobservable heterogeneity component.

log-odds ratios.The expected value is a function of observed covariates , x_i, that include general health status, smoking status, smoking duration, income, education, occupational status, house ownership, household size, gender and age. Parental mortality and differences between current and target age are used as a control for influences in the formulation of expectations. The set of smoking indicators includes a dummy variable for individuals who ever started smoking, a dummy variable for those who eventually quit, a continuous measure of smoking duration as number of years spent smoking and an interaction term between quitting and smoking duration.[17] The dispersion sub-model is based on a log function and allows modelling the dispersion of the dependent variable as a function of observed covariates. Alternatively, a constant-dispersion sub-model can be estimated, where the precision parameter does not depend on any explanatory variable, and this is the way we proceed here.

Results from the ML estimation of the beta regression model, when unobservable heterogeneity is neglected, are reported in Table 9.[18] Here we rescale the dependent variable to lie in the (0, 1) interval and reduced the interval to avoid zeros and ones using the following transformation:

$$y_i = \frac{\frac{SSP}{100}(N-1)+a}{N}$$

where N is the sample size and a is some constant, in this case 0,5 (see Ferrari and Cribari-Neto, 2004).[19] Reported SSP is higher for individuals in good or very good

[17]This allows calculation of the slope for current smokers and former smokers separately. If the relationship between SSP and smoking duration can be expressed as $\hat{\beta}_1 quitter + \hat{\beta}_2 log(smy) + \hat{\beta}_3 log(smy) * quitter$, then for current smoker the slope is $\hat{\beta}_2$ and for former smokers, the slope is $(\hat{\beta}_2+\hat{\beta}_3)$ as *quitter* equals 1 for those who quit and 0 for those who are still smoking. The coefficient $\hat{\beta}_1$ goes in the intercept.

[18]Benitez-Silva and Ni (2005) use pooled OLS to estimate a model for expected longevity and claim that, with a fairly large sample size, a BLU estimator, such as ordinary least squares, should guarantee consistency of the parameter estimates even though there can be a loss of efficiency. If the location sub-model is based on the logit link function, parameters can be estimated using OLS in a model where the dependent variable is $log\frac{SSP}{1-SSP}$, but standard errors will be inefficient.

[19]This transformation allows avoiding the logarithm of zeros and ones.

Table 9
Results from a beta regression model for expected longevity

Variable	Coeff.	S.E.
sah	0.441**	0.062
ever started	0.208	0.466
quitter	−0.421	0.521
log(smy)	−0.061	0.130
log(smy)*quitter	0.145	0.150
log(income)	0.013	0.022
education	0.000	0.008
retired	0.253**	0.084
employed	0.444**	0.104
house owner	0.088	0.058
household size	0.005	0.030
mother dead	−0.156	0.189
agemth	0.001	0.002
father dead	0.092	0.191
agefth	−0.002	0.002
male	−0.006	0.069
log(age)	−2.789**	0.431
log(T-age)	−0.919**	0.204
constant	14.156**	2.222
ϕ	0.997**	0.028

Notes:
Significance levels: † : 10% ∗ : 5% ∗∗ : 1%

health: people in this group have an odds ratio of 1,55 relative to those in poorer health. Reported SSP is higher for retired and employed people than for those who are unemployed or housekeepers, and diminishes as age increases. This result holds also after controlling for difference between target and current age. Smoking behaviour and parental mortality are not important determinants of expectations in the regression model.

4.2. *A probit model for self-assessed health*

Equation (2) is the function for general health. We use a probit model for the probability of being in good or very good health.[20] The contribution of individual i

[20]The estimation of an ordered probit for SAH, as an ordinal variable with five outcomes, is feasible. However, in the context of the mixture model, there are many convergence problems that make this specification of the mixture impracticable.

Table 10
Results from probit model for general health status

Variable	Coeff.	S.E.
ever started	−0.236	0.494
quitter	0.909	0.561
log(smy)	0.052	0.137
log(smy)*quitter	−0.298†	0.162
log(income)	0.059**	0.022
education	0.056**	0.008
retired	0.090	0.087
employed	0.261*	0.110
house owner	−0.008	0.062
household size	−0.022	0.031
male	0.212**	0.074
log(age)	−1.987**	0.309
cons	7.216**	1.308

Notes:
Significance levels: † : 10% * : 5% ** : 1%

to the sample likelihood is:

$$L_i = \Phi(k_i \mathbf{x_i} \beta)$$

where y_i is SAH, $k_i = 2y_i - 1$ is an indicator of sign and x_i are observed exogenous variables as in the equation for expected longevity but excluding parental mortality variables and target age.

Results from the ML probit model are reported in Table 10. As expected income, education and employment status are positively associated to health. Among the smoking indicators, only the interaction term is statistically significant: for former smokers, smoking longer, has a negative effect on perceived health status. Men are more likely to be in better health than females and, not surprisingly, ageing has a negative impact on health.

4.3. *A duration model for smoking behaviour*

Equation (1) represents smoking duration. Smoking duration is observed for those who ever started to smoke. Assuming that the population of interest entered the

state of smoker at some starting age in the past, the exit event is represented by the decision to quit.[21] For those who quit a complete spell of smoking is observed. For those who are current smokers at the time of the survey, the spell is censored. Survival time among current smokers is right-censored at the number of years the individual has smoked by the time of the survey. The censoring variable (q_i) is a binary indicator that takes value 1 if the individual quit and 0 if they are currently smoking. The contribution of individual i to the conditional likelihood is:

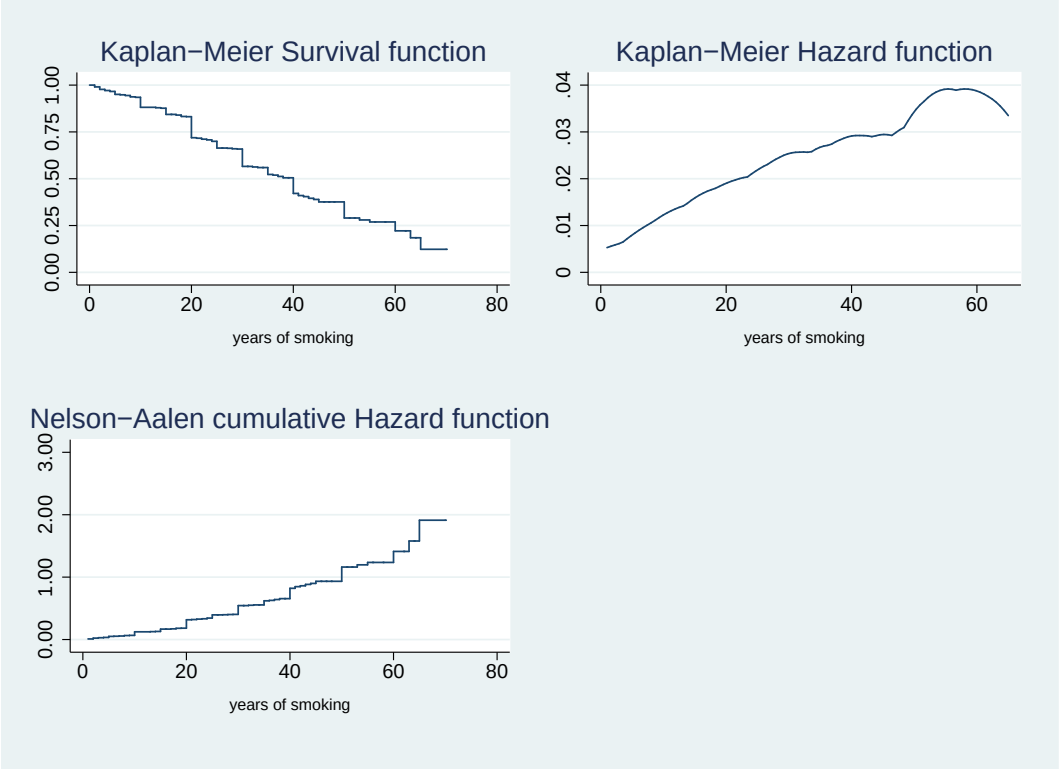
$$L_i = f(t_i|\mathbf{x_i}, \beta)^{q_i} \cdot [S(t_i|\mathbf{x_i}, \beta)]^{(1-q_i)} \tag{4}$$

where t_i is the number of years a person has smoked, q_i is the censoring binary variable, $f(t_i)$ is the density function and $S(t_i|\mathbf{x_i}, \beta_i) = [1 - F(t_i|\mathbf{x_i}, \beta_i)]$ is the survival function. When $q_i = 1$ equation (4) corresponds to the contribution of the complete spell, that is the density function of the duration variable; when $q_i = 0$ it corresponds to the contribution of the censored spell, that is the survivor function. The true duration is assumed to be independent of the onset of smoking and the censoring time, but it is conditional on a vector of exogenous variables, $\mathbf{x}_i$. Observable characteristics include an indicator of healthy habits, income, education, occupational status, house ownership, household size, marital status, gender and age. Marital status is an indicator of social support and therefore it is likely to be correlated with the event of quitting smoking. Smokers can receive advice on quitting from their partners and they can be concerned also about the positive externality on the spouse if they quit (see e.g., Hanson et al., 1990; Lindström et al., 2000).

A simple non-parametric analysis of smoking duration, carried out on the subsample of individuals who ever started smoking in their life, allows calculation of the Kaplan-Meier estimates for the survival and the hazard function from the data. The

[21] Unfortunately, the SHARE does not collect any information about the onset of smoking, thus limiting the analysis of smoking behaviour. It would have been interesting to analyse also the duration in the state of non-smoker, where the failure is the onset of smoking, trying to correlate these two durations since they might be influenced by common observable and unobservable factors (van Ours, 2005).

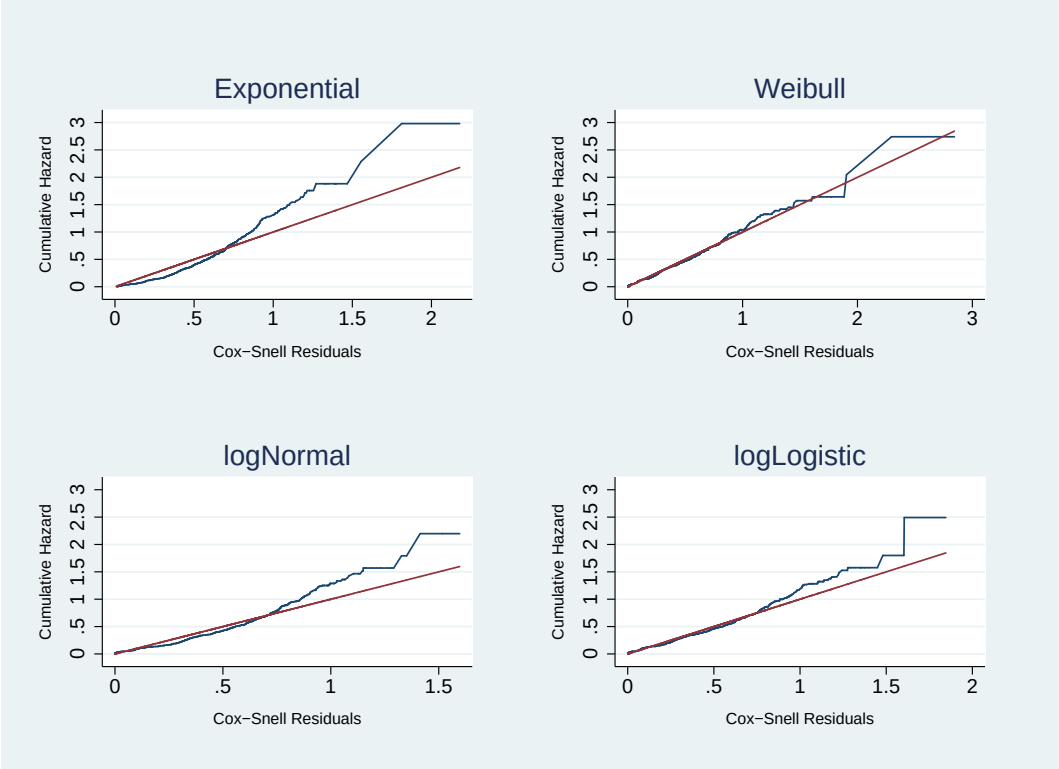
Figure 3
Smoking duration



upper left graph in Figure 3 shows the probability of remaining a smoker by duration of smoking. This probability decreases as the length of time spent smoking increases. The upper right graph shows the hazard of quitting by duration of smoking. The hazard increases with the length of time spent smoking, but for old ages the shape of the curve suggests that it increases at a decreasing rate. The lower graph shows the Nelson-Aalen estimate of the cumulative hazard function for quitting smoking. This suggests positive duration dependence in the general behaviour of the hazard function: the hazard of quitting should increase with time.

The hazard of quitting smoking is estimated using a Weibull model, which is commonly reckoned to be adequate to describe the hazard of quitting smoking (see Douglas, 1998). In order to select the distribution that fits the survey data on smoking duration, the generalized Cox-Snell residuals for alternative parametric distributions are computed. Residuals are plotted against the cumulative hazard estimated using Kaplan-Meier. A graphical comparison between the plots for the

Figure 4
Distributions comparison for smoking duration



Exponential, the Weibull, the log-Normal and the log-Logistic distribution, Figure 4, suggests that the Weibull distribution fits the data best. Information criteria also favour the Weibull distribution, as shown in Table 11.[22]

In the Weibull model the hazard, the survival and the density function are parameterized as follows:

$$\begin{aligned} h(t_i|\mathbf{x}_i;\beta) &= \lambda\alpha t_i^{\alpha-1}, \\ S(t_i|\mathbf{x}_i;\beta) &= exp(-\lambda t^{\alpha}), \\ f(t_i|\mathbf{x}_i;\beta) &= \lambda\alpha t_i^{\alpha-1} exp(-\lambda t_i^{\alpha}). \end{aligned}$$

where λ is a non-negative function that depends on the observed characteristics, $\lambda = exp(-\alpha\mathbf{x}_i\beta)$; $\alpha t^{\alpha-1}$ is the baseline hazard whose shape depends on the ancillary

[22] I calculate the Akaike information criterion (AIC) as $-2logL + 2q$ where q is the number of parameters. The Bayesian information criterion (BIC) is calculated as $-2logL + log(N)q$ where N is the sample size.

Table 11
Comparison between distributions for smoking duration

Information criterium	Exponential	Weibull	log-Normal	log-Logistic
AIC	1866.080	1754.532	1847.373	1784.190
BIC	1908.616	1801.795	1894.636	1831.453
N	834	834	834	834

parameter α. The Weibull model can yield a monotonic increasing, constant, or decreasing hazard of quitting. Regarding this, the sign of the shape parameters needs to be interpreted. If $\alpha = 1$ the Weibull equals the Exponential, with $h(t) = \lambda$. If $\alpha > 1$ the hazard function is monotonically increasing; if $\alpha < 1$ the hazard function is monotonically decreasing. The last two cases are known as positive and negative duration dependence.

Substituting the Weibull density and the survival function into equation (4), the individual likelihood function becomes:

$$L_i = [\lambda\alpha t^{\alpha-1} exp(-\lambda t^{\alpha})]^{q_i} \cdot [exp(-\lambda t^{\alpha})]^{(1-q_i)} \tag{5}$$

Results from the ML estimation of the Weibull model are reported in Table 12. The Weibull model is estimated in the Accelerated Failure Time (AFT) metric, which emphasizes the time to failure.[23] For high income individuals and better educated individuals the time to failure is predicted to accelerate: survival time is shorter. Also, individuals with a healthy lifestyle tend to survive for a shorter period if smoking, they are more likely to quit sooner. The shape parameter α suggests that there is positive duration dependence. Figure 5 shows the decreasing pattern of survival. Survival diminishes less than proportionally after 30 years of smoking. The

[23]This means that the model can be written as $ln(T_i) = \mathbf{x}_i\beta + \sigma u_i$ where T_i is latent survival time, σ is the scale factor and is measured as the inverse of the shape parameter and u_i is the error term. The AFT metric assumes a linear relationship between the log of survival time and characteristics $\mathbf{x}_i$. Interpretation is straightforward if the model is written as $ln(T_i\psi) = \sigma u_i$ where $\psi = exp(-\mathbf{x}_i\beta)$ is a survival time scaling factor.

Table 12
Results from a AFT Weibull model for smoking duration

Variable	Coeff.	S.E.
log(income)	-0.110^{**}	0.031
healthy	$-0.102^{\dagger}$	0.059
log(education)	$-0.098^{\dagger}$	0.053
retired	-0.031	0.108
employed	-0.052	0.121
single	0.037	0.082
male	-0.026	0.073
log(age)	-0.191	0.285
cons	5.944^{**}	1.192
α	1.611^{**}	0.066
$\sigma = 1/\alpha$	0.621^{**}	0.025

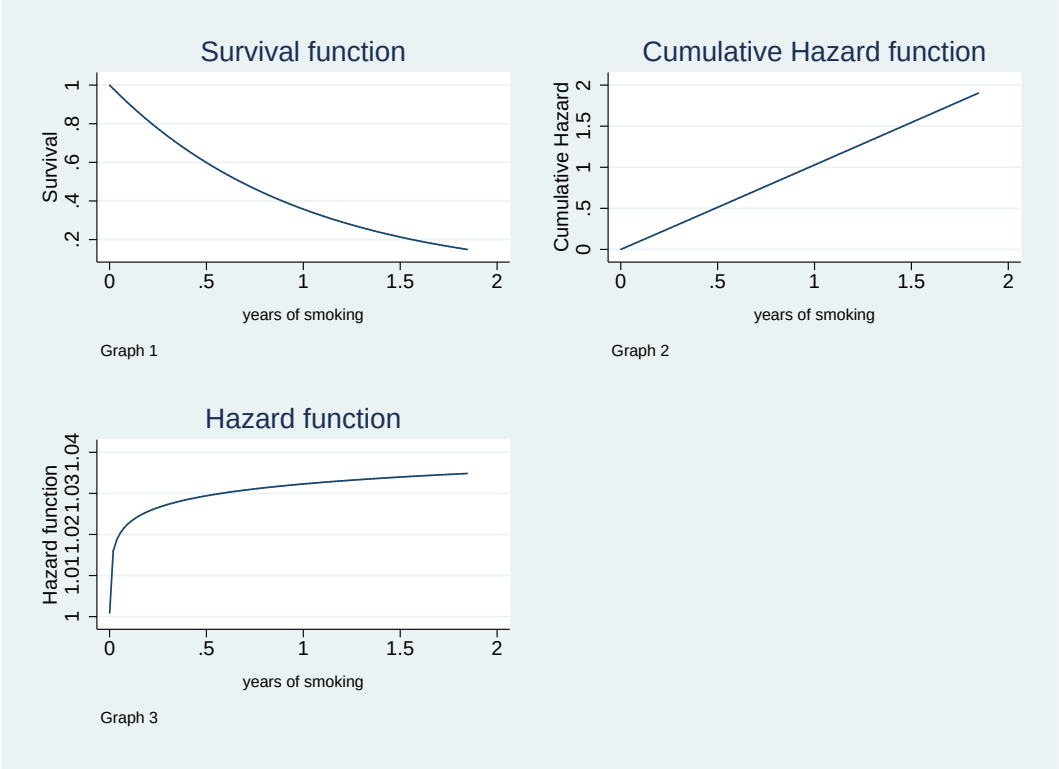
Notes:
Significance levels: $\dagger$: 10% $*$: 5% $**$: 1%

hazard of quitting increases with time spent smoking but less than proportionally. Positive duration dependence is likely to be underestimated in this model because it does not control for unobservable heterogeneity.

4.4. *The likelihood for the recursive model*

The expression of the likelihood for the model described by equations (1) to (3) requires to ensure a contribution from each individual in the sample. The likelihood in equation (5) is restricted to individuals who ever started smoking. Those who never started, will never quit, and for these individuals smoking duration is not observed so that standard duration analysis for smoking behaviour may not be appropriate. To allow for every individual's contribution in the model for quitting, we split the sample according to starting. Douglas and Hariharan (1994), Douglas (1998), Forster and Jones (2001) apply a full split population model to the onset of smoking. Here the idea is to use a similar approach for the quitting hazard. Usually split population models are such that the hazard depends upon a selection mechanism, in this case the decision of becoming a smoker at some point in life.

Figure 5
Weibull model for smoking duration



The separation model is written as:

$$[f(t_i|\mathbf{x}_i,\beta)\Phi(\mathbf{x}_i\beta)]^{s_i\cdot q_i}\cdot[S(t_i|\mathbf{x}_i,\beta)\Phi(\mathbf{x}_i\beta)]^{s_i\cdot(1-q_i)}\cdot[1-\Phi(\mathbf{x}_i\beta)]^{(1-s_i)} \tag{6}$$

where s_i is a binary indicator that takes value 1 when the individual ever started smoking and $\Phi(\mathbf{x}_i\beta)$ is the probability of ever being a smoker.

This way of writing the smoking duration model divides the population in 3 groups.[24] The first group consists of individuals that started smoking and quitted, for which $s_i = 1$ and $q_i = 1$. The second group is made up of those who started and are current smokers at the time of the survey but never quitted, for which $s_i = 1$ and $q_i = 0$. The third group includes those who never started smoking and therefore miss an observation for the duration of smoking, for which $s_i = 0$ and q_i is missing.

[24]Techniques for splitting the population are typically used to allow for heterogeneity between groups and heterogeneity is supposed to depend only on observed characteristics. In this framework, we assume that unobservable factors influence the choice of starting smoking as well as the hazard of quitting.

Table 13
Results from a probit model for smoking participation

Variable	Coeff.	S.E.
log(income)	0.071^{**}	0.023
education	0.032^{**}	0.008
male	0.954^{**}	0.063
log(age)	-1.075^{**}	0.252
cons	2.981^{**}	1.058

Notes:
Significance levels: † : 10% ∗ : 5% ∗∗ : 1%

Only variables measured prior to decision of starting should be used to explain smoking participation, which was a past choice. These include income, education, gender and age. We use education and income in 2004 as indicator of past socio-economic status, assuming that social rank is stable over time. Education can also be considered as a signal of ability as well as pure investment in health.[25] As reported in Table 13, the propensity to smoke increases with income and education, is higher for men and decreases with age.

Given equation (6), the contribution to the sample likelihood for individual i is:

$$L_i = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)\Gamma(\tau)}\, y_i^{(\omega-1)}\,(1-y_i)^{(\tau-1)} \cdot [\Phi(k_i \mathbf{x}_i\beta)] \qquad (7)$$
$$[\lambda\alpha t_i^{\alpha-1} exp(-\lambda t_i^{\alpha})\Phi(\mathbf{x}_i\beta)]^{s_i q_i} \cdot [exp(-\lambda t^{\alpha}))\Phi(\mathbf{x}_i\beta)]^{s_i(1-q_i)} \cdot [1-\Phi(\mathbf{x}_i\beta)]^{(1-s_i)}.$$

4.5. *A finite mixture model for longevity and smoking*

Mixture models are used to account for unobservable heterogeneity in the data. By assuming that the sample is drawn from a finite mixture distribution it is possible to define two, or more, types of the population from which the observed data come from (McLachlan and Peel, 2000). In a complex likelihood function like equation (7), a mixture model would reduce the unobservable individual heterogeneity to an

[25] From this point of view, different levels of education depend upon individual ability and, in turn, differences in individual ability would reflect differences in the probability of starting smoking.

unobservable factor that is common to the phenomena considered (survival probability, health and smoking behaviour). Although based on a simple assumption of heterogeneity, mixture models do allow control over self-selection.

A popular way to estimate mixture models is the EM algorithm (Dempster et al., 1977). The EM algorithm can be defined as a general method of finding maximum likelihood estimates when the data are incomplete or have missing values. The EM algorithm assumes the existence of additional missing parameters and estimates likelihood functions that are otherwise analytically intractable.

Using general notation, the probabilistic model of equation (7) can be written as:

$$f(y_i|\mathbf{x}_i;\Theta) = \sum_{k=1}^{K} p_k \cdot f_k(y_i|\mathbf{x}_i;\theta_k)$$

where $\Theta = (p_1, ..., p_k, \theta_1, ..., \theta_k)$ and f_k is a density function parameterised by θ_k. The expression above assumes that there are K component densities mixed together with K mixing coefficients p_k. The sample log-likelihood expression for the incomplete data is given by:[26]

$$logL(y_i|\mathbf{x}_i;\Theta) = log \prod_{i=1}^{N} f(y_i|\mathbf{x}_i;\Theta) = \sum_{i=1}^{n} log\big(\sum_{k=1}^{K} p_k \cdot f_k(y_i|\mathbf{x}_i;\theta_k)\big) \qquad (8)$$

The (unconditional) likelihood function above is difficult to maximize because it contains the logarithm of a sum, and the values of the missing variables are unknown. However, by writing the likelihood for the complete data and assuming the values of the mixing proportions and the probabilities of belonging to a mixture component, it is possible to proceed in the maximization process. Each mixing proportion can be thought of as the prior probability that a data point is randomly drawn from a mixture component k of the population, where $0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$. This implies that the free mixing parameters are $K - 1$.

Let $\Pi_i = \{\pi_{i1} \dots \pi_{iK}\}$ be the vector of missing data that consists of K indicators

[26]Here "log" stands for natural logarithm.

for the components membership that are identically and independently distributed as a multinomial distribution with probabilities $\{p_1 \dots p_K\}$. Using Bayes' rule, it follows that:

$$E(\pi_{ik}|y_i, \mathbf{x}_i; \Theta) = \frac{p_k \cdot f_k(y_i|\mathbf{x}_i; \theta_k)}{\sum_{k=1}^{K} p_k \cdot f_k(y_i|\mathbf{x}_i; \theta_k)} = \frac{p_k \cdot f_k(y_i|\mathbf{x}_i; \theta_k)}{f(y_i|\mathbf{x}_i; \Theta)} = \hat{\pi}_{ik} \tag{9}$$

$\hat{\pi}_{ik}$ is the posterior probability, that is the probability that a data point belongs to sub-population k. The posterior probability show which component density generated each observation in the sample.

Maximizing the sample log-likelihood in equation (8), the estimate of the unconditional probability that an individual is of component k, is the mean of the conditional probability that they belong to component k:

$$\hat{p}_k = E(\hat{\pi}_{ik}) = \frac{1}{N} \sum_{i=1}^{N} \hat{\pi}_{ik}.$$

The completed log-likelihood can be derived substituting p_k in equation (8) as $p_k = \frac{\pi_{ik} \cdot f(\mathbf{x}_i; \Theta)}{f_k(\mathbf{x}_i|\theta_k)}$, from equation (9). After manipulation, the (conditional) log-likelihood for the mixture model is:

$$logL(y_i, \Pi_i|\mathbf{x}_i; \Theta) = \sum_{i=1}^{N} \big(\sum_{k=1}^{K} \hat{\pi}_{ik} \cdot log f_k(\mathbf{x}_i|\theta_k) + \sum_{k=1}^{K} \hat{\pi}_{ik} \cdot log(p_k) \big).$$

The intractability of the log-likelihood is overcome and the function to be maximised has again an additive form. Estimates for θ_k are obtained by

$$argmax \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{\pi}_{ik} \cdot log\big(f_k(y_i|\mathbf{x}_i; \theta_k)\big).$$

and so the values of $\hat{\theta}_k$ maximise both the unconditional and the conditional likelihood.

The EM is an iterative procedure: it iterates two steps until convergence.[27] The first step, called E-step, computes the conditional expectation of the expression $log\big(f_k(x_i|\theta_k)\big)$. In the case of finite mixture models where Bayes' theorem is applied, the E-step computes the posterior probabilities according to equation (9).

The M-step maximizes the log-likelihood function in separate parts. What makes this algorithm particularly appealing is that the M-step can be implemented easily as maximum likelihood estimation of weighted models, where the weights are the posterior probabilities.

The E-step and the M-step alternate in a loop that starts with starting values for the parameters: ${\theta_k}^*$ and ${p_k}^*$. At the first iteration (i) the E-step calculates:

$$\hat{\pi}_{ik}^{i} = Pr(\Pi_{ik} = 1|y_i, x_i; \theta^*, {p_k}^*).$$

The M-step provides the updating formulas for θ_k^{i+1} and p_k^{i+1} that are used to compute $\hat{\pi}_{ik}^{i+1}$, this:

$$p_k^{i+1} = \frac{1}{N}\sum_{i=1}^{N}\pi_{ik}^{i} \qquad \text{and} \qquad \theta_k^{i+1} = argmax \sum_{i=1}^{N}\sum_{k=1}^{K} {\pi_{ik}}^{i} \cdot log\big(f_k(x_i|\theta_k^i)\big)$$

The peculiarity of the EM is that the values of the likelihood are monotonically increasing, i.e. $L(\theta^{i+1}) \geq L(\theta^i)$, with equality found if the estimates of the θs at iteration i equals estimates at iteration $(i+1)$. Under suitable regularity conditions, the sequence θ^i converges to a stationary point of $L(\theta)$. Properties of convergence, including these conditions, are well discussed in Dempster et al. (1977), McLachlan and Krishnan (1996) and Schafer (1997). The convergence criterion can be set as the difference between the value of the likelihood at the last iteration and the value at the previous iteration.[28]

[27] The algorithm for the maximisation of the mixture model for the likelihood in equation (7) has been written in Stata.

[28] A criterion of failure of convergence is given by setting the maximum number of iterations to be executed.

Usually the likelihood is not unimodal, meaning that there are several local maxima and a unique global maximum. The solution found by the EM loop can depend critically upon the set of initial values for the prior probabilities and the θs. One possibility is to run the loop several times with a different initialisation and choose the best model comparing the value of the likelihood. Starting values can be guessed, or they can be computed as a linear transformation of the parameter estimates from the single component model. We construct a grid of starting values such that, for each "guess" of the prior probability, all the elements in a sequence of values for the θs must be taken as initial values for the parameters. This grid allows the loop to run 56 times and get comparable results.

The mixture model can be compared to the single component model using information criteria, so that testing the hypothesis of unobservable heterogeneity in the population and endogeneity of health and smoking behaviour is allowed.[29] The same test statistics can be used to determine the optimal number of latent classes. However, due to the smaller sample size after mixing and because the EM is a computationally intensive procedure, we present here only results for the two-class model. Future work might be done to verify the feasibility of a three-class model.

5. Results from the finite mixture model

This section presents results from the finite mixture model for equation (7), estimated using the EM technique.

The specification of the model has been chosen looking at the statistical fit of alternative specifications. The RESET test does not reject the use of the logarithm for

[29] Mroz (1999) uses a mixture model to control for the endogeneity of a dummy regressor. The approach employed requires a discrete factor approximation to a continuous latent variable, and relies upon a step function, whose mass points and factor loadings are estimated. This methodology is shown to perform better than alternative estimators, especially in the case of non-normality. Arcidiacono et al. (2007) emphasise the fact that this can complicate the maximisation of an already complex likelihood function, because the log-likelihood would lose its additive form. Therefore they propose to use the EM algorithm, as it gives additive separability of the log-likelihood, allowing for sequential estimation of the parameters at each step.

Table 14
Testing heterogeneity in the expected longevity model

Information criterion	One-class model	Two-class model
-2logL	5926.578	5375.580
AIC	6022.578	5569.580
CAIC	6288.340	6105.621
BIC	6287.340	6104.621
parameters	48	97
N	1837	1837

smoking duration, income and age. Table 14 reports the value of the penalized information criteria both in the two-class mixture and the single class model. The AIC, BIC and CAIC are usually used to test the hypothesis of unobservable heterogeneity and in this case they all favour the two-class model against the single model.[30] This means that unobservable heterogeneity is a matter of concern and that the mixture performs better than a model that neglects heterogeneity. Moreover, endogeneity of SAH and the smoking variables is confirmed by the same tests.

The coefficients in the two classes of the mixture model suggest that expectations about survival and self-reported health are formulated differently according to variation in observed characteristics. The estimated sample proportions are 0.56 for the first latent class and 0.44 for the second one, which is evidence of a division of the population in two heterogenous sub-groups. Table 15 reports the coefficient estimates in the mixture for each equation of the recursive model.

In both classes the indicator of health status has a statistically significant and positive effect on reported survival probability. Individuals who report good or very good SAH are also more likely to report higher expected longevity, though the impact of SAH is bigger in class 1 where the increase in the odds of reporting higher SSP is about twice for individuals in good or very good health relative to those in poorer health. The odds ratio is about 1 in class 2 meaning that, although SAH is an

[30]The first information criterion used is the negative value of twice the log-likelihood. The consistent Akaike information criterion (CAIC) is calculated as $-2logL + (1 + log(N)q)$.

important determinant of SSP, the odds of reporting higher SSP is equally likely in both SAH categories. We also find that class 1 and 2 are similar in that, conditional on target age, for both, SSP decreases as people age. The effect of ageing on the odds is relatively bigger in class 2.

Classes differ in the variation of SSP to socio-economic variables and parental mortality. In the first latent class, individual expected longevity is higher for employed individuals and house owners. In the second latent class the impact of socio-economic variables is less clear and SSP decreases with income and increases with education. The odds of reporting higher SSP significantly decrease, in class 2, if the father has already died and increase with father's age at death. While, in class 1 variation of SSP is not explained by parental mortality.

The most interesting result has to do with the smoking variables, in that they do not have a statistically significant effect on reported survival probability. We speculate on the sign and dimension of the coefficients and find that there are differences between smokers and non smokers as well as among the group of smokers in both reporting SSP and SAH. The direction of causality in the relationship between smoking variables and SSP or SAH can, at first glance, appear counterintuitive.

In both classes, smokers (defined as individuals who ever started smoking) tend to report a higher SSP relative to non smokers. This result is not expected but can be interpreted in terms of myopic behaviour of smokers who do not internalise the negative effect of smoking on mortality risk.

As shown in the previous section (footnote 15), we can distinguish between current and former smokers. Current smokers, in both classes, are less optimistic about the effect of smoking duration on future survival. The odds of reporting higher SSP decreases as the number of years spent smoking increases, and is bigger in class 1 than in class 2. This suggest that current smokers internalise the negative effects of intensity of smoking as if they were aware of the fact that smoking behaviour

Table 15
Coefficients from the two-class model

		Class 1		Class 2	
Equation	Variable	Coeff.	S.E.	Coeff.	S.E.
smoking participation	log(income)	0.043	0.031	0.110**	0.038
	education	0.052**	0.010	0.003	0.012
	male	0.785**	0.083	1.201**	0.097
	log(age)	−0.808*	0.327	−1.465**	0.401
	cons	2.085	1.385	4.295**	1.661
smoking duration	log(income)	−0.151**	0.042	−0.068	0.047
	healthy	−0.138†	0.076	−0.010	0.091
	education	−0.137*	0.068	−0.075	0.082
	retired	0.099	0.134	−0.244	0.176
	employed	0.178	0.154	−0.415*	0.190
	single	−0.119	0.104	0.263*	0.129
	male	−0.215*	0.094	0.334**	0.116
	log(age)	−0.181	0.358	−0.215	0.454
	cons	6.439**	1.523	5.397**	1.859
	α	1.719**	0.095	1.544**	0.092
	$\sigma = 1/\alpha$	0.582**	0.032	0.648**	0.039
general health status	ever started	−0.252	0.614	0.017	0.859
	quitter	1.522*	0.747	0.199	0.923
	log(smy)	0.091	0.170	−0.072	0.240
	log(smy)*quitter	−0.487*	0.217	−0.077	0.263
	log(income)	0.090**	0.033	0.036	0.033
	education	0.051**	0.011	0.060**	0.013
	retired	−0.009	0.118	0.204	0.132
	employed	0.396**	0.145	0.061	0.171
	house owner	−0.050	0.083	0.034	0.094
	household size	0.041	0.044	−0.093*	0.047
	male	0.103	0.097	0.389**	0.119
	log(age)	−1.110**	0.405	−3.131**	0.489
	cons	3.148†	1.738	12.331**	2.055
expected longevity	sah	0.543**	0.085	0.262**	0.052
	ever started	0.197	0.597	0.313	0.445
	quitter	−0.362	0.691	−0.011	0.478
	log(smy)	−0.071	0.166	−0.106	0.124
	log(smy)*quitter	0.128	0.200	0.081	0.137
	log(income)	0.001	0.031	−0.046**	0.018
	education	−0.001	0.011	0.011†	0.006
	retired	0.307**	0.116	0.063	0.069
	employed	0.381**	0.141	0.092	0.090
	house owner	0.146†	0.080	0.053	0.049
	household size	0.005	0.042	−0.027	0.024
	mother dead	−0.203	0.253	−0.034	0.165
	agemth	0.002	0.003	0.001	0.002
	father dead	0.090	0.260	−0.452**	0.163
	agefth	−0.002	0.003	0.005**	0.002
	male	0.073	0.092	−0.060	0.063
	log(age)	−3.352**	0.595	−1.919**	0.355
	log(T-age)	−1.083**	0.284	−0.780**	0.168
	cons	17.055**	3.085	10.622**	1.817
	ϕ	0.666**	0.025	7.776**	0.365
	p_1	0.557			
	logL:	-2687.790			
	N:	1837			

Notes:

Significance levels: † : 10% ∗ : 5% ∗∗ : 1%

enhances the risk of mortality.[31]

Relative to current smokers, and in both classes, those who quit by the time of the interview tend to report lower SSP, while we would have expected to find that the decision of quitting has a positive effect on the perception of the chances of living longer (as well as on actual observed mortality risk). In particular, the odds of reporting higher SSP decreases more in class 1, where, however, former smokers tend to report higher SSP the longer they smoked and, therefore, appear more optimistic than current smokers. On the contrary, for former smokers in class 2 the odds of reporting higher SSP decreases as the number of years spent smoking increases. We conclude that in class 1 the benefits of quitting smoking as well as the negative effects of smoking duration are not internalised in the formulation of expected longevity, thus suggesting non-rationality of former smokers in this class.

The general health status equation shows that the binary indicator of quitting and the interaction term of quitting and number of years smoked, are statistically significant and largely explain variation in reported health. Analysis of the coefficients of the full set of smoking variables shows that smokers are less likely to report good and very good health than non smokers in class 1, while in class 2 smokers are more likely to report better health than non smokers. Discriminating among smokers allows us to find that, as smoking duration increases, current smokers tend to report higher SAH if they are in class 1 and lower SAH if they are in class 2. Former smokers, however, behave similarly in both classes. They seem to internalise the detrimental effect of smoking on health, also in terms of smoking duration: those who quit are more likely to report good or very good health than current smokers and to report worse health as smoking duration increases. The impact of quitting and duration of smoking on self-reported health is predicted to be bigger in class 1.

[31]The SHARE dataset does not provide information about intensity of smoking in terms of number of cigarettes smoked. This can be a matter of concern in the interpretation of the results. The only available indicator of intensity of smoking is the number of years spent smoking. However, for a given duration of smoking, the actual number of cigarettes smoked, either per day or per year, may be important to explain effects of smoking on health conditions.

Furthermore, Table 15 shows that employed, richer and more educated individuals are more likely to be in better health in class 1. Health is predicted to deteriorate with age and intensity of smoking for former smokers. In class 2, in general, more educated individuals and men are more likely report better health. Health deteriorates with age and households size.

Overall, average predicted expected longevity is higher in the first latent class: it is about 0.68 against 0.59 in the second latent class.[32] However, the predicted probability of being in good or very good health is lower in the first class, 0.49 against 0.53 per cent in the second class. This result can seem puzzling, as we would have expected to find higher predicted SSP and SAH in the same class. Putting together the estimation results from the mixture model will help us drawing a clearer picture of the latent classes.

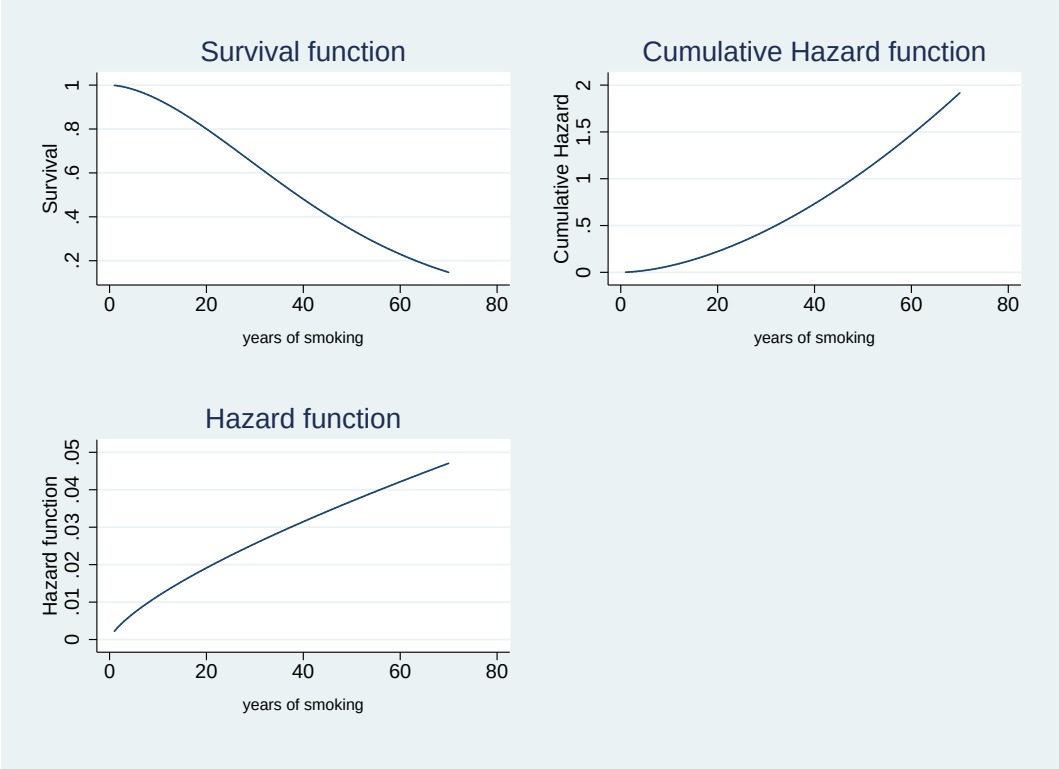
Analysis of the determinants of smoking participation helps explaining the different impact of socio-economic characteristics. In the first latent class, propensity to smoke is positively related to education and gender: men are more likely to start smoking. In the second latent class, high income individuals and men are more likely to start.[33] Ageing decreases the probability of starting in both classes. The predicted probability of starting is slightly higher in class 1, around 0.46 against 0.45 in class 2. This would confirm that individuals who are less aware of the health consequences of smoking behaviour are more concentrated in the first class.

The duration regression model by classes reflects heterogeneity in the hazard of quitting. In the first latent class time to quit is predicted to accelerate for richer and more educated individuals, those who have a healthy lifestyle, and women, meaning that these individuals quit sooner. The second latent class shows that time to quit accelerates for employed individuals and decelerates for singletons and men.

[32]The precision parameter, ϕ, is higher in the second beta regression model where predicted variance is, in fact, lower (0.027 in class 2 and 0.124 in class 1).

[33]The most recent ISTAT health survey (Indagine Multiscopo sulle Famiglie) shows that the socio-economic gradient in smoking is not very clear in particular at old ages: the frequency of smokers is higher among those with the highest educational qualifications.

Figure 6
Weibull model for smoking duration - class 1

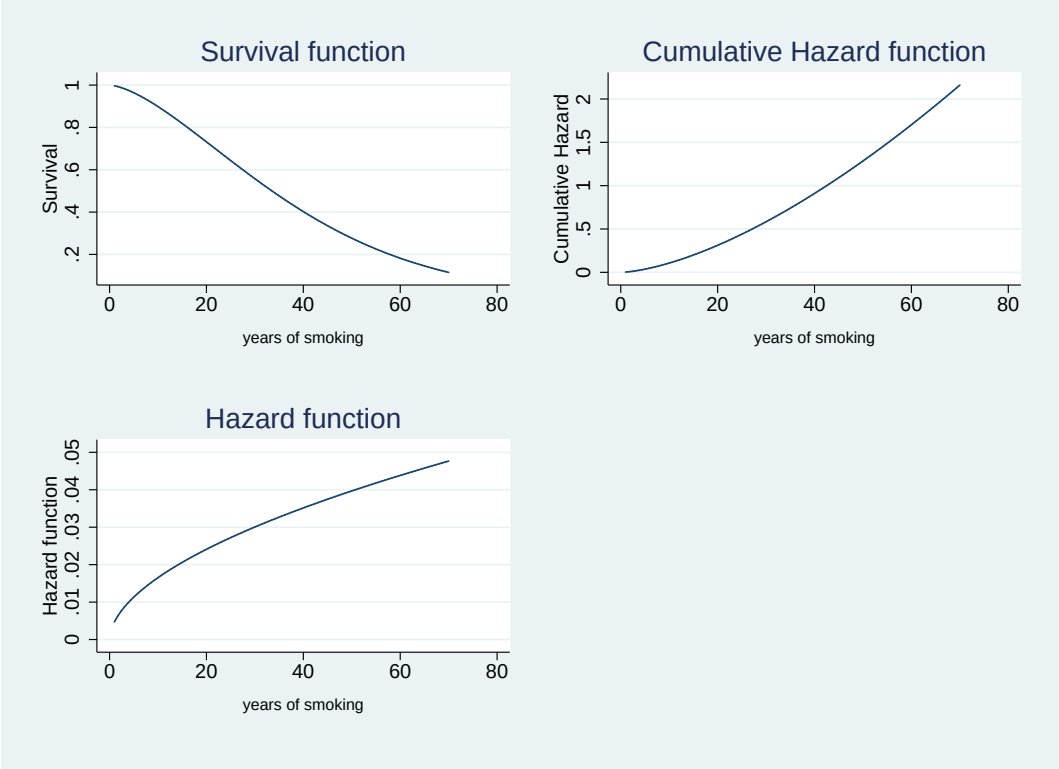


Duration dependence is positive in both classes, and therefore the hazard of quitting increases with time spent smoking, but less in the second class relative to the first.

Figures 6 and 7 show that the estimated survival and hazard functions do not differ much across classes. For low durations, 0-20 years, the probability of survival is quite high, between 0.8 and 1 in the first class, and between 0.7 and 1 in the second class. For intermediate durations, survival decreases more rapidly but and is steeper in class 2 for each survival time. The function assumes a flatter shape for the longest durations. This suggests that the decision of quitting is less and less attractive as the time spent smoking is longer than some threshold duration. The hazard function, in fact, increases less then proportionally, in particular for longer durations and seems to increase more, at least for the shortest survival times, in the second class.

The impact of the smoking variables on SAH and SSP, and the coefficients in the smoking equations of the mixture model, help us interpreting the latent classes in

Figure 7
Weibull model for smoking duration - class 2



terms of unobservable frailty and rationality in addiction. We define the first latent class as the class of "frailer" individuals who select into smoking, perhaps because individual frailty reduces the opportunity cost of smoking. In this class, in fact, smokers report poorer health. At the same time, however, smokers report higher chances of living longer than non smokers, and this can reflect myopic behaviour in discounting the future consequences of their tobacco consumption on mortality risk. The second latent class can be defined as the class of "less frail" individuals, where those who are healthier tend to select, while frailer individuals decide not to start smoking or to quit. Here, in fact, smokers report better health and higher survival probability. Both current and former smokers have a fully rational behaviour reporting SAH and SSP in the class of the less frail individuals, while in the class of frailer individuals current smokers do not evaluate the negative effect of smoking when they think to their health as well as former smokers do when they are asked to think to the chances of living longer (former smokers, however, seem to internalise

more the negative effect of smoking on health if they are frailer).

5.1. *Posterior analysis*

Individuals in the two latent classes seem to formulate survival expectations in a different way. In particular, smoking variables show that there is heterogeneity between smokers. In order to have some more insight into the determinants of class membership, it is possible to do a simple analysis using the estimated posterior probabilities from the EM. Posterior probabilities allow identification of the class to which each individual belongs. We assign each individual to the class associated to the larger posterior probability using cut-off probability of 0.5 (see Atella et al., 2004), and define a binary indicator of class membership that takes value 1 if the posterior probability for an individual is above the cut-off probability and 0 otherwise. This is equivalent to saying that individual i belongs to latent class 1 if the estimated posterior probability π_{i1} is larger than the estimated π_{i2}, since $\sum_{k=1}^{K} \pi_{ik} = 1$.

Table 16 shows sample means in each latent class. On average, the sample drawn from the first sub-population reports a higher probability of survival at some future age, about 73 per cent, relative to the second class, about 60 per cent. Around 48 per cent of the individuals in the first class report good or very good health, while the second class seems to be healthier (about 53 per cent of individuals report good or very good health). However, the second class shows a higher percentage of individuals with more than 2 chronic conditions. These findings seem to confirm that formation of expectations about survival and self-reporting own health do not follow the same path in the two sub-populations and that unobservable frailty can be used to label the latent classes.

About the same proportion of respondents in the 2 classes have ever started smoking: about 46 per cent in the first sub-sample against 45 per cent in the second sub-sample. The frequency of quitting is lower and the average time spent smoking

Table 16
Sample means in the heterogenous population

Variable	Full sample	Class 1	Class 2
ssp	0.658	0.726	0.599
sah	0.512	0.488	0.533
chronic	0.439	0.426	0.451
ever started	0.454	0.459	0.449
quitter	0.566	0.522	0.604
smy	29.821	30.910	28.870
healthy	0.459	0.439	0.476
income	22168.946	22323.151	22037.015
low quartile	0.247	0.233	0.260
low2 quartile	0.250	0.236	0.262
above median	0.503	0.531	0.479
education	7.181	7.229	7.139
single	0.186	0.181	0.190
retired	0.555	0.538	0.570
employed	0.198	0.227	0.173
unemployed	0.019	0.017	0.021
sick	0.011	0.015	0.007
homemaker	0.217	0.203	0.229
house owner	0.541	0.512	0.566
household size	2.555	2.527	2.580
mother dead	0.788	0.790	0.787
father dead	0.918	0.917	0.919
agemth	74.945	74.510	75.318
agefth	71.062	70.314	71.701
male	0.465	0.453	0.475
age	63.913	64.061	63.787
age5065	0.613	0.602	0.622
age6670	0.162	0.144	0.177
age7175	0.120	0.131	0.110
age7680	0.070	0.079	0.063
age8185	0.025	0.032	0.019
age86more	0.010	0.012	0.009
target age	78.819	79.150	78.535
sample size	1837	847	990

Table 17
Marginal effects from a probit model for class membership

Variable	dF/fx	S.E.
log(income)	$0.016^{\dagger}$	0.009
education	-0.001	0.003
retired	0.014	0.034
employed	0.117^{**}	0.042
house owner	-0.058^{*}	0.024
household size	-0.021	0.013
healthy	$-0.042^{\dagger}$	0.024
single	$-0.059^{\dagger}$	0.034
mother dead	0.079	0.076
father dead	0.125	0.073
agemth	-0.001	0.001
agefth	$-0.002^{\dagger}$	0.001
male	$-0.050^{\dagger}$	0.027
log(age)	0.581^{**}	0.172
log(T-age)	0.268^{**}	0.083

Notes:
Significance levels: $\dagger$: 10% $*$: 5% $**$: 1%

is longer in the first sub-sample. With 52 per cent of smokers who quit and an average duration of smoking of about 31 years, the first sub-sample appears to represent a population of "frailer" and "more addicted" individuals or, in other words, a population of "hard-core smokers". This sub-sample is also characterised by higher average income and education, more employed individuals, women, a smaller proportion of individuals with an healthy lifestyle and less singletons. The age composition of the sub-samples shows that the first sub-sample is relatively older, on average, as it includes more individuals aged 71 and over than the second sub-sample. The second sub-sample seems to be drawn from a population of "less frail" and "less addicted" individuals: on average, the quitting rate is higher and the duration of smoking is shorter.

A probit model can be estimated, regressing the class membership dummy variable on the exogenous variables used in the mixture model. Table 17 shows that individuals with a better socio-economic status (in terms of income and occupation) are more likely to be assigned to class 1. This might reflect the fact that smoking

is more affordable to richer individuals. Also, employees, who in the SHARE sample are relatively younger than individuals in other occupational status, are more likely to be still active smokers. The probability of being assigned to the class of the "frailest and most addicted" is lower for males, individuals who have a healthy lifestyle, singletons, and individuals whose father died at older ages. The age variables are statistically significant determinants of class membership: elderly individuals are more likely to appear in class 1, conditional on the difference between current and target age. Individuals who are approaching the end of lifespan might find smoking still attractive and decide not quit. A possible explanation is that smokers are rationally myopic in measuring the consequences of smoking longer as they believe that there is no time left for smoking to affect their mortality risk.

6. Conclusions

This paper explores formation of survival expectations in the elderly population. The aim of this work is to assess internal consistency and investigate validity of the indicator of expected longevity, SSP, collected in the SHARE. We use Italian data from the first wave of the SHARE.

The analysis looks in particular at the relationship between SSP, SAH, smoking duration and socio-economic characteristics. A finite mixture approach is used to estimate a recursive system of equations, where unobservable individual-specific heterogeneity is taken into account. The mixture model is estimated using the EM algorithm. The mixture model is shown to fit better the data than the single class model and provides evidence of individual heterogeneity in the formulation of expected longevity and in reporting general health status.

We find that apart from ageing, individual expectations about survival are influenced by self-assessed health, socio-economic status and parental mortality. Smoking variables are not statistically significant in the SSP equation but they explain

variation in SAH, which in turn explains most of the variation in reporting SSP. Estimated coefficients and posterior analysis of class membership allows identification of two-types of individuals in the population, who differ in terms of unobservable frailty and rationality in addiction. The first type is frailer and more addicted to tobacco consumption. Such individuals seem to internalise much less the detrimental effects of smoking on health and mortality. The second type is less frail and less addicted, and seems to be more rational in evaluating health status and survival probability in terms of taking into considerations the consequences of smoking. Our findings confirm Schoenbaum (1997)'s results and provide a deeper insight into the interpretation of existing heterogeneity between smokers. We also find differences between current and former smokers in the way they discount future consequences of tobacco consumption on health and mortality risk. Overall, our findings suggest caution in the use of SSP as a proxy of actual mortality.

Appendix A The beta regression model

The SHARE indicator for subjective survival probability takes values between 0 and 100. Appropriate rescaling allows us to translate it in a new variable such that $y_i \in (0,1)$ and is distributed as a traditional beta. The beta distribution has two shape parameters $\omega, \tau > 0$ and belongs to the family of exponentials. It can be written as:

$$f(y_i;\omega,\tau) = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)\Gamma(\tau)}\, y_i^{(\omega-1)}\,(1-y_i)^{(\tau-1)}$$

The first and second moments are:

$$\begin{aligned} \boldsymbol{\mu} &= E(y) = \frac{\omega}{\omega+\tau}, \\ \sigma^2 &= Var(y) = \frac{\boldsymbol{\mu}(1-\boldsymbol{\mu})}{\omega+\tau+1} \end{aligned}$$

Using the transformation $\phi = \omega + \tau$ in the equations above, the mean and the variance can be expressed as:

$$\begin{aligned} \boldsymbol{\mu} &= \frac{\omega}{\phi}, \\ \sigma^2 &= Var(y) = \frac{\boldsymbol{\mu}(1-\boldsymbol{\mu})}{\phi+1} \end{aligned}$$

Therefore $\boldsymbol{\mu}$ is a location parameter and ϕ is a precision parameter (as precision increases the variance gets smaller). We notice that the variance depends partly on location and that we only need to estimate the location and precision parameters to recover the variance.

Maximum likelihood estimation techniques can be used to measure ω and τ (or, equivalently, $\boldsymbol{\mu}$ and ϕ). The regression model is based on the Generalised Linear Model (GLM). Therefore, sub-models for the precision and the location parameters need to be defined (see Paolino, 2001; Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006). In particular, the logit link function and the log function are used to restrict the mean to the unit interval, $\boldsymbol{\mu} \in (0,1)$, and get a strictly positive ϕ, respectively. Inverting the link functions, we get new expressions for the mean and precision parameters:

$$\begin{aligned} \boldsymbol{\mu}_i &= \frac{exp(x_i\beta)}{1+exp(x_i\beta)}, \\ \phi_i &= exp(w_i\delta) \end{aligned}$$

where β, δ are the vectors of coefficients for the covariates vectors x_i, w_i. Using some algebra we get the following identities which are needed to solve the maximisation

problem:

$$\begin{aligned}
\phi &= \omega + \tau = exp(w_i\delta) \\
\omega &= \frac{exp(x_i\beta + w_i\delta)}{1 + exp(x_i\beta)} \\
\tau &= \frac{exp(w_i\delta)}{1 + exp(x_i\beta)}
\end{aligned}$$

The log-likelihood for individual i is:

$$\begin{aligned}
&logL(\beta, \delta; y_i, W, X) = \\
&\quad log\Gamma\Big(exp(w_i\delta)\Big) - log\Gamma\Big(\frac{exp(x_i\beta + w_i\delta)}{1 + exp(x_i\beta)}\Big) - log\Gamma\Big(\frac{exp(w_i\delta)}{1 + exp(x_i\beta)}\Big) + \\
&\quad \Big(\frac{exp(x_i\beta + w_i\delta)}{1 + exp(x_i\beta)} - 1\Big)log(y_i) + \Big(\frac{exp(w_i\delta)}{1 + exp(x_i\beta) - 1}\Big)log(1 - y_i)
\end{aligned}$$

Estimation of the βs and δs allows us to recover the shape parameters (ω, τ), the expected value, $\boldsymbol{\mu}$, the precision, ϕ, and the variance, σ^2, of the dependent variable. In a less general case, the precision parameter would not be a function of explanatory variables but would be a positive constant term.

References

Arcidiacono, P., Sieg, H., and Sloan, F. (2007). 'Living rationally under the volcano? An empirical analysis of heavy drinking and smoking'. *International Economy Review*, vol. 48(1).

Atella, V., Brindisi, F., Deb, P., and Rosati, F. (2004). 'Determinants of access to physicians services in italy: a latent class seemingly unrelated probit'. *Health Economics*, vol. 13(7), pp. 657–668.

Balia, S. and Jones, A. M. (2007). 'Mortality, Lifestyle and Socio-Economic Status'. *Journal of Health Economics*. doi:10.1016/j.jhealeco.2007.03.001.

Becker, G. S. and Murphy, K. (1988). 'An theory of rational addiction'. *Journal of Political Economy*, vol. 96(4), pp. 675–700.

Benitez-Silva, H. and Ni, H. (2005). 'Health stocks and health flows in an empirical model of expected longevity'. *Risk Analysis*. Manuscript, SUNY-Stony Brook.

Börsch-Supan, A., Brugiavini, A., Jürges, H., Mackenbach, J., Siegrist, J., and Weber, G. (2005). *Health, Ageing and Retirement in Europe First Results from the Survey of Health, Ageing and Retirement in Europe*. Mannheim: MEA.

Bruine de Bruin, W., Fischbeck, P. S., Stiber, N. A., and Fischhoff, B. (2002). 'What number is "fifty-fifty"?: Redistributing excessive 50% responses in elicited probabilities'. *Risk Analysis*, vol. 922(4), pp. 713–723.

Chapman, S., Wong, W., and Smith, W. (1983). 'Self-exempting beliefs about smoking and health: differences between smokers and ex-smokers'. *American Journal of Public Health*, vol. 83(2), pp. 215–219.

Deaton, A. S. and Paxson, C. (1998). 'Ageing and inequality in income and health.'. *American Economic Review, Papers and Proceedings*, vol. 88, pp. 248–253.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). 'Maximum likelihood from incomplete data via EM algorithm.'. *Journal of the Royal Statistical Society. Series B*, vol. 34(1), pp. 1–38.

Dominitz, J. and Manski, C. F. (1997). 'Using expectation data to study subjective income expectations'. *Journal of the American Statistical Association*, vol. 92, pp. 855–867.

Dominitz, J. and Manski, C. F. (2005). 'Measuring and interpreting expectation of equity returns'. Working Paper.

Douglas, S. (1998). 'The duration of the smoking habit'. *Economic Inquiry*, vol. 36(1), pp. 49–64.

Douglas, S. and Hariharan, G. (1994). 'The hazard of starting smoking: Estimates from a split population duration model'. *Journal of Health Economics*, vol. 13(2), pp. 213–230.

Etilè, F. and Milcent, C. (2006). 'Income-related reporting heterogeneity in self-assessed health:evidence from france'. *Health Economics*, vol. 15, pp. 956–981.

Ferrari, S. and Cribari-Neto, F. (2004). 'Beta regression for modelling rates and proportions'. *Journal of Applied Statistics*, vol. 31(7), pp. 799–815.

Forster, M. and Jones, A. M. (2001). 'The role of tobacco taxes in starting and quitting smoking: duration analysis of british data'. *Journal of the Royal Statistical*

Society Series A, vol. 164, pp. 517–547.

Frijters, P., , Haisken-DeNew, J., and Shields, M. A. (2005). 'Socio-economic status, health shocks, life satisfaction and mortality: Evidence from an increasing mixed proportional hazard model'.

Guiso, L., Jappelli, T., and Terlizzese, D. (1992). 'Earnings uncertainty and precautionary saving'. *Journal of Monetary Economics*, vol. 30, pp. 307–337.

Guiso, L., Jappelli, T., and Terlizzese, D. (2002). 'An empirical analysis of earnings and employment risk'. *Journal of Business & Economic Statistics*, vol. 20, pp. 241–253.

Hanson, B. S., Isacsson, S., Janzon, L., and Lindell, S. (1990). 'Social support and quitting smoking for good. Is there an association? Results from the population study, "Men born in 1914," Malmo, Sweden.'. *Addictive Behaviors*, vol. 15(3), pp. 221–233.

Hill, D., Perry, M., and Willis, R. (2005). 'Estimating knightian uncertainty from survival probabilities on the HRS'. Manuscript, University of Michigan.

Hurd, M., McFadden, D., and Merrill, A. (1999). 'Predictors of mortality among the elderly'. NBER Working Paper 7440.

Hurd, M. and McGarry, K. (1995). 'Evaluation of the subjective probabilities of survival in the health and retirement study: Data quality and early results'. *The Journal of Human Resources*, vol. 30, pp. S268–S292.

Hurd, M. and McGarry, K. (2002). 'The predictive validity of subjective probabilities of survival'. *The Economic Journal*, vol. 112, pp. 966–985.

Idler, E. L. and Benyamini, Y. (1997). 'Self-rated health and mortality: community studies'. *Journal of Health and Social Behavior*, vol. 38(1), pp. 21–27.

Idler, E. L. and Kasl, S. V. (1995). 'Self-ratings of health: do they also predict change in functional ability?'. *Journal of Gerontology. Series B. Psychological Sciences and Social Sciences*, vol. 50(6), pp. 344353.

Juster, T. (1966). 'Consumer buying intentions and purchase probability: an experiment in survey desing'. *Journal of the American Statistical Association*, vol. 61, pp. 658–696.

Kremers, S. P., Mudde, A., and Vries, H. D. (2004). 'Model of unplanned smoking initiation of children and adolescents: an integrated stage model of smoking behavior.'. *Preventive Medicine*, vol. 38, pp. 642–650.

Lindström, M., Hanson, B., östergren, P.-O., and Berglund, G. (2000). 'Socioeconomic differences in smoking cessation: the role of social participation'. *Scandinavian Journal of Public Health*, vol. 28(3), pp. 200–208.

McLachlan, G. J. and Krishnan, T. (1996). *The EM Algorithm and Extensions*. Wiley and Sons.

McLachlan, J. and Peel, D. (2000). *Finite Mixture Models*. New York, John Wiley and Sons Ltd.

Mroz, T. A. (1999). 'Discrete factor approximations in simultaneous equation models: estimating the impact of a dummy endogenous variable on a continuous outcome'. *Journal of Econometrics*, vol. 92(2), pp. 233–274.

Orphanides, A. and Zervos, D. (1995). 'Rational addiction with learning and regret'.

Journal of Political Economy, vol. 103(4).

Paolino, P. (2001). 'Maximum likelihood estimation of models with beta-distributed dependent variables'. *Political Analysis*, vol. 9(4), pp. 325–346.

Peto, R., Lopez, A. D., Boreham, J., and Thun, M. (2005). 'Mortality from smoking in developed countries 1950-2000'. Oxford University Press, Oxford.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, volume 72. Monographs on Statistics ad Applied Probability.Chapman and Hall.

Schoenbaum, M. (1997). 'Do smokers understand the mortality effects of smoking? evidence from the health and retiremient survey'. *American Journal of Public Health*, vol. 87(5), pp. 755–759.

Smith, J. P. (1999). 'Healthy bodies and thick wallets: the dual relation between health and economic status'. *The Journal of Economic Perspective*, vol. 13(2), pp. 145–166.

Smith, V. K., Taylor, D. H., and Sloan, F. A. (2001). 'Longevity expectations and death: can people predict their own demise?'. *American Economic Review*, vol. 91, pp. 1126–1134.

Smithson, M. and Verkuilen, J. (2006). 'A better lemon squeezer? maximum likelihood regression with beta distributed dependent variables'. *Psychological Methods*, vol. 11(1), pp. 54–71.

Thaler, R. and Sheffrin, H. (1981). 'An economic theory of self-control'. *Journal of Political Economy*, vol. 89(2), pp. 392–406.

van den Berg, G. J. (2001). 'Duration models: specification, identification and multiple durations'. In Z. Griliches (ed). Handbook of Econometrics Vol(5), pp.3381-3460. Harvard University, Cambridge, MA, USA M.D. Intriligator, University of California, Los Angeles, CA, USA.

van Doorslaer, E. and Gerdtham, U.-G. (2003). 'Does inequality in self-assessed health predict inequality in survival by income? evidence from swedish data'. *Social Science and Medicine*, vol. 57, pp. 1621–1629.

van Ours, J. C. (2005). 'Dynamics in the use of drugs'. Discussion Paper n. 21. Tilburg University, Center for Economic Research.

Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y. T., Gupta, P. C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M. J., Wichmann, H. E., Wu, A. H., Zaridze, D., Peto, R., and Doll, R. (2004). 'Tobacco and cancer: Recent epidemiological evidence'. *Journal National Cancer Instute.*, vol. 96(2), pp. 99–106.

Viscusi, W. K. (1990). 'Do smokers underestimate risks?'. *The Journal of Political Economy*, vol. 98(6), pp. 1253–1269.

Winston, G. (1980). 'Addiction and backsliding: The theory of compulsive consumption'. *Journal of Economic Behaviour and Organization*, vol. 1, pp. 295–234.