HEDG Working Paper 07/06

Unravelling the influence of smoking initiation and cessation on premature mortality using a common latent factor model

Silvia Balia
Andrew M. Jones

June 2007

ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Unravelling the influence of smoking initiation and cessation on premature mortality using a common latent factor model

Silvia Balia *
Università di Cagliari
and University of York

Andrew M. Jones †
University of York

June 13, 2007

Abstract

Duration models for lifespan and smoking, that focus on the socio-economic gradient in smoking durations and length of life, are estimated controlling for individual-specific unobservable heterogeneity by means of a latent factor model. The latent factor influences the risk of starting and quitting smoking as well as the hazard of mortality. Frailty could influence smoking behaviour through two mechanisms: the effect of life expectancy on initiation of smoking and the impact of adverse health events on quitting. Our findings suggest that individual-specific preference for experimentation, which leads those people who start smoking soonest to quit early, is a potential source of spurious correlation between smoking durations. They also suggest that frailty acts according to both mechanisms, driving selection into early smoking initiation as well as selection into early smoking cessation. Overall, determinants of smoking durations and mortality hazard are largely unaffected by unobservable heterogeneity. However, the latent factor model strengthens the results of the univariate models suggesting that increasing the quitting rate and reducing the duration of smoking would decrease premature mortality. Whereas, prompting people to delay starting would shorten the length of time spent smoking.

JEL codes I1 C10 C41
Keywords: smoking; mortality; duration analysis; unobservable heterogeneity; latent factors.

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

†Department of Economics and Related Studies, University of York, York YO10 5DD, UK. *E-mail address*: amj1@york.ac.uk. *Fax*: 0044-1904-433759

The epidemiological evidence from the 1950s to date suggests that tobacco smoking is responsible for about 30 per cent of cancer deaths in developed countries and it also causes deaths from vascular, respiratory and other diseases (see Vineis et al., 2004). A host of studies show that patterns of mortality are likely to be affected by the proportion of persons who give up smoking and tobacco-related diseases account for a large proportion of all-cause mortality in Europeans countries (see e.g., Peto et al., 2005). According to the World Health Organization about half of smokers will die of a tobacco-related disease as morbidity caused by tobacco use is more and more prevalent (WHO, 2005). Tobacco consumption is has been linked to increased prevalence of illness and therefore is responsible of the resulting economic and social burden. Nonetheless, the preventability of smoking gives scope for public health interventions.

Smoking trends are often associated with socio-economic inequalities. Systematic differences in tobacco consumption exist between individuals depending on their socio-economic status. Statistics show that in the UK the highest prevalence of smoking is among the lowest socio-economic group in the population. The general evidence for the US and the EU is that the lowest groups have a higher risk of dying from smoking-related diseases then men from upper groups (Kunst et al., 2004). People who have experienced social and economic disadvantages during childhood, adolescence and adult life may run the greatest risk of becoming addicted to nicotine and smoking.

This paper investigates smoking behaviour and tries to identify the effect of both smoking initiation and smoking cessation on the hazard of mortality. Data from the British Health and Lifestyle Survey (HALS), which provide rich information about individual smoking behaviour, socio-economic characteristics and mortality, are used to study duration models for lifespan and smoking. The analysis focuses on the socio-economic gradient in smoking durations and mortality hazard and estimates the impact of smoking behaviour on mortality modelling unobservable heterogeneity.

An objective of this paper is to study the determinants of smoking cessation for individuals who started smoking at some point in their life. In particular, we look at the effect of age at starting on persistence of the smoking habit over time and we consider that age at starting is potentially endogenous in the cessation equation due to the presence of unobservable factors which influence both durations. Ignoring unobservable heterogeneity among smokers could lead to inconsistent and biased estimates of the causal effect of age at onset of smoking on the hazard of quitting. Evidence of this can be found in van Ours (2006) who emphasizes a spurious correlation between smoking initiation and cessation. He finds that, on one side, higher ages at starting increase the hazard of quitting (the causal effect) and, on the other side, the correlation between the unobserved components in the two hazards is positive, which means that those who start smoking soon are also more likely to quit soon (the selection effect). A plausible interpretation is that the effect of age at starting on smoking duration captures the fact that for some individuals starting smoking is an experiment and for them smoking is not likely to be a long lasting habit.

Therefore, the estimated causal effect of age at starting on the hazard of quitting may be spurious in the presence of unobservable or unmeasured factors that influence both starting and quitting. Unobservable heterogeneity would be reflected in a non-zero covariance between the error components in the smoking equations, therefore the genuine causal relationship between age at starting and smoking duration can be recovered only by an estimator that can isolate the effect of the unobservables.

A second objective of this paper is to investigate the relationship between smoking behaviour and the risk of mortality. Individual-specific unobservable heterogeneity is an issue in the estimation of the causal effect of health-related behaviours on health outcomes, as unobservable factors are likely to drive selection into or out of smoking as well as the hazard of dying early. The potential for selection bias to influence estimates of the impact of smoking on health and mortality is well known

(see e.g., Adda and Lechene, 2001, 2004; Mark and Robins, 1993). Smoking is a choice variable and potential endogeneity of smoking durations needs to be taken into consideration when studying the relationship with the hazard of dying. Lahiri and Song (2000) estimate a model of smoking behaviour and smoking-related morbidity. They emphasise the fact that individuals may, rationally, self-select into or out smoking behaviour on the basis of their perception, beliefs and knowledge of the risk of damaging their own health and increasing the likelihood of contracting smoking-related illnesses. Such beliefs can be based on information which evolves over time and is hidden to the econometrician. Neglecting the existence of this source of heterogeneity between smokers will bias the health and mortality effects of smoking. Smokers who choose to continue smoking can have a lower predisposition to develop a smoking-related illness and so incidence of that disease or the mortality risk among this group of people would be lower than would be found if there were random allocation to smoking.

Adda and Lechene (2001) find evidence of selection into smoking for individuals with lower life expectancy and this results is confirmed in Adda and Lechene (2004) where they show that smokers come from a population with poorer health. In particular, Adda and Lechene (2001) look at the effect of a measure of potential life expectancy on quantities smoked for current smokers and on the hazard of quitting for both current and ex-smokers: they find a negative correlation between smoking and life expectancy which is a sign of selection into smoking. Adda and Lechene (2004) investigate such selection effect using tobacco-free morbidity scores and estimate a logistic regression for the probability of starting and a Cox proportional hazard model for the hazard of quitting. They show that there is selection into smoking in both starting and quitting: less healthy individuals are more likely to start smoking at an early age, and are also less likely to give up smoking.

We can detect at least two types of unobservable heterogeneity. The first type occurs when the hazards of starting and quitting are driven by an individual-specific

preference for experimentation, as in van Ours (2006). Experimentation is hard to capture through the set of exogenous regressors and this information is perhaps unobservable to the researcher. If individuals in the population are heterogeneous due to differences in their preference for experimenting with smoking, we expect that unobservable heterogeneity accelerates time to starting as well as time to quitting. Therefore we expect a positive correlation between the two durations.

The second type of unobservable heterogeneity is known, in particular in the duration analysis literature, as unobservable frailty. The recent literature shows mixed evidence regarding the role of unobservable heterogeneity in the relationship between smoking and mortality. In work that investigates the role of lifestyles on individual mortality risk, Balia and Jones (2007) find evidence to suggest that frailer individuals tend to select into healthy behaviours such as non smoking. Here we can expect that frailer individuals do the same. Frailer individuals may consider that the loss of health due to tobacco consumption is higher because of their poor health and illnesses: for these individuals time to starting is more likely to decelerate and time to quitting to accelerate. In this case smoking behaviour can vary depending on what extent a smoker internalises the negative effects of tobacco consumption on health and correlation between smoking durations would be negative. We also expect the correlation between smoking initiation and mortality to be positive, and the correlation between smoking cessation and mortality to be negative. However, it could be the case, as in Adda and Lechene (2001, 2004), that frailer individuals tend to select into smoking. This might occurs when individuals' beliefs about their life expectancy influence the opportunity cost of smoking, the latter being reduced by lower life expectancy. Therefore, frailty can accelerate time to starting and decelerate time to quitting, meaning that individuals who expect to live shorter lives smoke longer than less frail individuals. We expect to find again a negative correlation between smoking durations, as well as between smoking initiation and the risk of dying. On the other hand, the correlation between smoking cessation

and mortality is expected to be positive.

Here we try to separate the causal effect of smoking from any selection effect into smoking, driven by unobservable characteristics, using a simultaneous model of equations for lifespan and smoking. The model, which consists of structural form equations for mortality and smoking cessation, and a reduced-form equation for smoking initiation, has a triangular recursive structure and allows for correlation between the error component of the three duration equations. However, correlation between errors complicates the maximisation of the likelihood because the integral has no closed-form solution. Parametric and semi-parametric methods based on approximation or simulation can be used to evaluate the integral when closed-form solutions do not exist.

We find evidence of both types of unobservable heterogeneity described above. In particular, estimates suggest that due to individual-specific preference for experimentation those people who start smoking soonest quit earliest, in line with van Ours (2006). Furthermore, our results seem to confirm both Adda and Lechene (2001, 2004) and Balia and Jones (2007)'s arguments on the effect of individual-specific unobservable frailty on smoking and mortality. We find that unobservable frailty drives selection into early smoking initiation (Adda and Lechene's argument) as well as selection into early smoking cessation (Balia and Jones's argument). Overall, the estimated causal effect of smoking on mortality is largely unaffected by the presence of unobservable heterogeneity.

1. Survival Data in HALS

The Health and Lifestyles Survey (HALS) data contain information about time to death, time to starting tobacco consumption and time to quitting for a representative sample of the British population. This information is exploited to construct duration variables that indicate the time elapsed before each particular event occurs.

Table 1
Flagging status in April 2005

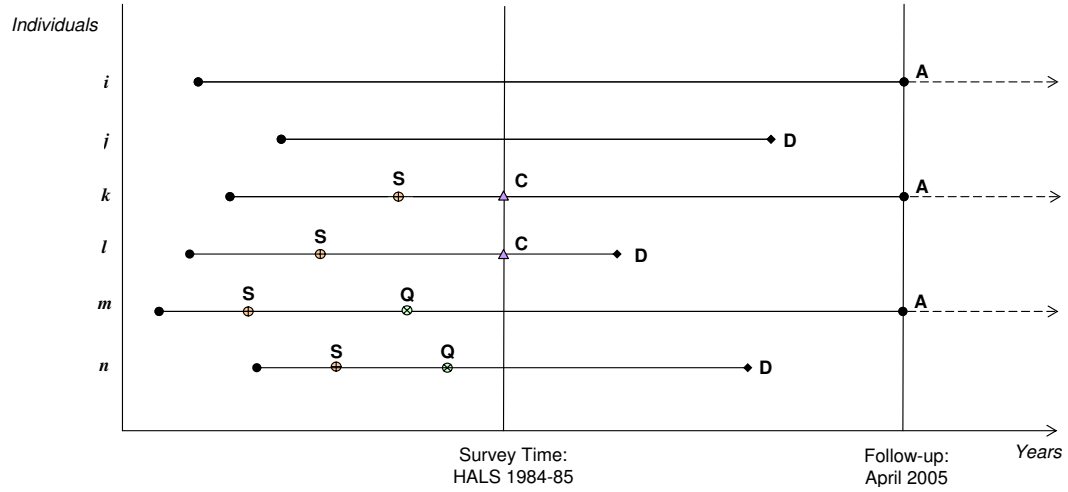
Flagging Status	Frequency	%
On file[a]	6248	69.40
Not NHS registered[b]	85	0.94
Deceased[c]	2431	27.00
Reported dead to HALS not on NHS Register[d]	1	0.01
Embarked - abroad[e]	42	0.47
Not yet flagged[f]	196	2.18

Notes:
[a] Currently alive and flagged on the NHS Register.
[b] But not known to be dead.
[c] Known dead and death certificate information recorded on file.
[d] May be alive.
[e] Identified on NHS Register but currently out of country.
[f] Not currently flagged for various reasons (no name etc.).

The HALS was carried out between Autumn 1984 and Summer 1985, in two home visits (the second one by a nurse). The questionnaire was designed and piloted by a study team at the University of Cambridge School of Clinical Medicine and funded by the Health Promotion Research Trust. The HALS was designed as a representative survey of adults in Great Britain (see Cox et al., 1987, 1993). The population surveyed was individuals aged 18 and over living in private households. 9,003 interviews were completed. This is a response rate of 73.5 per cent. HALS respondents have been tracked on the NHS registers on a regular basis to provide reliable information about individual mortality. The latest deaths data were released in June 2005. This allows us to investigate survival up to April 2005. As shown in Table 1, up to April 2005, 97.8 per cent of the original sample has been flagged and 27 per cent of the respondents had died.

The HALS questionnaire was designed to provide comprehensive information about risky behaviours. In particular, the survey data contain retrospective information on smoking. Self-reported variables describing age at starting smoking, current smoking status and how long ago the respondent stopped smoking are used to derive two duration variables which can be used to study the hazard of starting

Figure 1
Study time and survival time for individuals type in HALS



smoking and the hazard of quitting smoking.[1]

Figure 1 illustrates the study time and the survival times for the respondents of the HALS. The study time is the calendar time period that each respondent spent in the study since the time of entry, her birth, indicated by a full dot "•". Dots that are closer to the y-axis represent older individuals, dots closer to the vertical line indicating the time of the interview represent younger individuals. For respondents who died by April 2005 survival time is the period of time elapsed from the survey time to death (**D**). Respondents who are still alive at the follow-up (**A**) have a censored survival time. Respondents type i and j never smoked, so they do not have survival times for smoking. Respondents type k to n started smoking (**S**). In particular, k and l are current smokers at the time of the interview (they might have stopped smoking at some point in the future), so they have a complete survival time for starting using tobacco but a censored survival time for quitting smoking (**C**); while m and n are ex-smokers (**Q**) and, in fact, have a complete survival time for

[1]In particular, the variables in the dataset are *agestrt*, *exfag*, *exfagan*, *regfag*. They describe respectively age at starting smoking, whether or not the individual is an ex-smoker, how long ago they stopped smoking, and whether or not they smoke regularly at least one cigarette per day.

both starting and quitting smoking.

Survival time data give additional information relative to binary variables describing the occurrence of an event (e.g., death) or the choice of participation (e.g., starting or quitting). Here we use continuous time data assuming that the transition event may occur at any instant in time. We define the length of a spell for an individual in the sample as the realisation of a continuous random variable, T, that has the following cumulative distribution function (cdf) or failure function:

$$F(t) = P(T \leq t)$$

The complement function is the survivor function, which indicates the probability of surviving up a specific point in time t. The probability of survival is equal to 1 at entry in the state of interest and can be defined as:

$$S(t) = 1 - F(t) = P(T > t) \qquad \text{where} \quad 0 \leq S(t) \leq 1$$

The slope of the failure function, the density function, indicates the concentration of failure times along the time axis, and is expressed by:

$$f(t) = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t}$$

In particular, we are interested in the hazard function which represents the instantaneous rate of failing per unit of time, conditional on individual survival up to that instant:

$$h(t) = \lim_{\Delta t \to \infty} \frac{\partial F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{1 - F(t)} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Survival analysis uses variables indicating the length of time a person stays in the state of interest. Usually respondents in population samples are asked about

the date of entry and exit from the state, as in the case of our data. Individuals are assumed to enter the state at time 0 and leave it at some time t, when the event of failure occurs. If entry and failure are observed, it will be possible to measure a complete spell. While, if only entry is observed and exit will eventually occur at some time T in the future, the spell will be incomplete. Such incomplete durations are known as *right-censored* spells, where censoring is at the time of the last observation, and we only know that the complete duration will be $T > t$, as in the case of respondents type i, k and m for time to death, and k and l for time to quitting.[2] Depending on the state of interest, only those individuals who have survived for a minimum amount of time in the state are included in the sample or, putting the problem another way, individuals who fail before the time of observation will not be included. Hence, the remaining observed survival times are said to be *left-truncated*.[3]

In order to analyse smoking initiation we define a time variable *starting* which represents the number of years elapsed before someone starts smoking. For consistency reason, individuals who claimed to be current smokers but whose age at starting was zero are eliminated. The variable *starting* measures a complete duration if an individual had started smoking and an incomplete, or censored, duration if they had not. For those who started smoking, *starting* is equal to the age at starting, and it is right-censored at the age at the time of the interview for those who had not started by then. A binary indicator, *start*, indicates whether or not an individual started smoking.

Smoking cessation is described by the time variable *sm_years* which indicates the number of years a person smoked. To construct this variable we use information about the date at onset of smoking, how long ago an individual became an ex-

[2]When the date of entry is unknown, the exact length of the spell cannot be measured and the survival time is said to be *left-censored* but this is never the case with the data we use.

[3]An option that we do not pursue is to impose *right-truncation* to the data, this would restrict the population sample just to individuals who failed by the observation time, thus eliminating all longer survival times.

smoker and the date of the interview. For current smokers smoking duration is right-censored at the time of the interview, for quitters the true duration is recovered from a direct question on the questionnaire. A binary indicator, *quit*, indicates whether or not a smoker had stopped smoking completely.

As shown by Figure 1, the HALS data give us the scope to investigate the hazard of death. A complete duration is observed for those who died before the follow-up period, while an incomplete duration is associated with individuals who were still alive in April 2005. We construct a time variable *lifespan* which is simply equal to the age at death for those who died by April 2005, and to the age at the time of the follow-up, for those who were still alive. The variable *lifespan* assumes that the measure of duration begins at birth and measures the full lifespan. As an alternative one could assume that individuals enter the initial state only when they participate in the survey process, so that the entry date would be the seen date at HALS (see Cheung, 2000). The advantage of defining *lifespan* in the way that we do is that we are able to measure length of survival from birth, conditional on survival up to the time of the survey, in which case the distribution of the survival time is said to be left-truncated.

2. Methods

A duration model with unobservable heterogeneity can be written as a mixture model, which specifies the distribution of survival time, t_i, as conditional on a vector of observed individual characteristics, $\mathbf{x}_i$, and an individual-specific random effect which represents individual unobservable heterogeneity, ν_i. The hazard and survival functions conditional on the unobserved ν_i can be written as:

$$h(t_i|\mathbf{x}_i, \nu_i) = \nu_i h(t_i|\mathbf{x}_i) \tag{1}$$

$$S(t_i|\mathbf{x}_i, \nu_i) = S(t_i|\mathbf{x}_i)^{\nu_i} \tag{2}$$

where $h(t_i|\mathbf{x}_i)$ and $S(t_i|\mathbf{x}_i)$ can be the hazard and survival functions from either a proportional hazard or an accelerated failure time regression model.

The population (or unconditional) hazard and survival function are obtained by integrating out the unobservable heterogeneity term. This implies calculating either an integral when ν_i has a continuous mixing distribution with density $g(\nu)$:

$$h(t) = \int_0^\infty h(t|\mathbf{x}, \nu)\, g(\nu)\, d(\nu) \tag{3}$$

$$S(t) = \int_0^\infty S(t|\mathbf{x}, \nu)\, g(\nu)\, d(\nu) \tag{4}$$

or a sum when ν_i is discrete and each realisation ν_k is observed with unknown probability π_k:

$$h(t) = \sum_{k=1}^{K} h(t|\mathbf{x}, v_k)\, \pi_k. \tag{5}$$

$$S(t) = \sum_{k=1}^{K} S(t|\mathbf{x}, v_k)\, \pi_k. \tag{6}$$

Equations (3) and (4) imply that a parametric distribution for the random effect needs to be specified. The variance of ν_i is the additional parameter that needs to be estimated in a frailty model.[4]

Parameters estimates rely not only on the correct specification of the baseline hazard but also of the random effect. However, often, parametric distributions for

[4] A convenient normalisation used to allow identification of the parameters of the heterogeneity distribution is to impose the restriction $E(\nu_i) = 1$.

the heterogeneity term are not correctly specified and this may affect inference on the hazard function and the efficiency of the estimator. Heckman and Singer (1984) suggest a non-parametric specification of heterogeneity where a discrete distribution $\pi_k(\nu_k)$ approximates the continuous distribution $g(\nu)$ and the population survival and hazard take the form of equations (5) and (6). The mass points of the discrete distribution are realisations of ν_i and there is a finite number K of them with associated probability π_k. The semi-parametric approach is more flexible and avoids inference problems arising from the wrong choice of functional form for the distribution of the heterogeneity.

2.1. *A latent factor model for smoking and mortality*

In this paper mortality risk and smoking are modelled taking into account unobservable heterogeneity and potential selection effects in the data. To measure the causal effect of smoking on mortality we need to consider that a spurious correlation between smoking behaviour and mortality hazard can be induced by the endogeneity bias of the smoking duration variables in the mortality equation and that age at starting can be endogenous in the hazard of quitting as well. Endogeneity arises because of the presence of unmeasured or unobservable factors that influence time to starting, time to quitting and time to death. These factors, which are specific to each individual and are not captured by the observed covariates in the model, make the population heterogenous in the hazard of dying as well as in the smoking hazard rates.

In a similar framework, Lillard and Panis (1996) estimate a joint duration model for the hazard of dying, health status, marriage formation and dissolution, to identify the protective effect of marriage and disentangle adverse selection from positive selection into marriage. Marriage durations and individual health are potentially endogenous in the mortality equation and unobservable heterogeneity is controlled by assuming that the heterogeneity terms reflecting unobserved factors are distributed

as a multivariate normal. Gauss-Hermite quadrature is used to approximate the likelihood function and get full-information maximum likelihood (FIML) estimates, and correlation coefficients are estimated as additional parameters.

We propose a trivariate duration model with unobservable heterogeneity where the hazard of mortality depends on observed variables, such as individual socio-economic characteristics and demographics, smoking initiation and cessation; in turn, the hazard of quitting smoking depends on observed characteristics and age at starting, and ultimately the hazard of starting smoking is explained only by exogenous variables. Therefore, the model that we estimate is a system of simultaneous hazard regressions with triangular form, where a reduced-form hazard for smoking initiation is defined first. This reflects the chronology of events, as starting smoking precedes quitting and quitting smoking precedes death but not viceversa.

The hazard and survival functions in equations (1) and (2) vary between individuals depending on their observed and unobservable characteristics. If the hazard regression is expressed in accelerated failure time (AFT) metric, and this is will be shown to be the case for smoking durations in HALS, a linear relationship between the logarithm of survival time T_i and individual characteristics is assumed.[5] In the AFT models covariates act addictively on the log of survival time and the acceleration factor can be written as a generic $\lambda = exp(-\beta'\mathbf{x})$, while in proportional hazard (PH) model covariates act multiplicatively on the hazard function and $\lambda = exp(\beta'\mathbf{x})$. In our model:

$$\lambda_{1i} = exp\big(- \beta_1'\mathbf{W}_i - \theta_1'\mathbf{Z}_{1i} - \nu_{1i}\big), \tag{7}$$

$$\lambda_{2i} = exp\big(- \alpha_2' ln(t_{1i}) - \beta_2'\mathbf{W}_i - \theta_2'\mathbf{Z}_{2i} - \nu_{2i}\big), \tag{8}$$

$$\lambda_{3i} = exp\big(\alpha_3' ln(t_{1i}) + \delta_3' ln(t_{2i}) + \beta_3'\mathbf{W}_i + \nu_{3i}\big) \tag{9}$$

[5]The AFT model can be written as $ln(T_i) = x_i\beta + \sigma u_i$, where T_i is latent survival time σ is a scale factor depending on the shape parameter and u_i is the disturbance term.

where t_1 is the duration variable *starting* for smoking initiation , t_2 is *sm_years* for smoking cessation and t_3 is *lifespan* for length of life. $\mathbf{W}_i$ is a matrix of exogenous variables that affect the three time-to-failure responses, $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ are matrices of different sets of factors which have a direct effect on starting and quitting respectively, but do not affect the other outcome variables. The trivariate model in equations (7)-(9) is identified when the three hazard functions share the same covariates and this depends on non-linearity of the functional form. However, estimators that rely on functional form for identification are usually unstable: stronger identification restrictions are recommended to achieve more reliable estimates. Therefore, we decided to set some exclusion restrictions: the regressors in $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ have been chosen as instrumental variables in the smoking duration equations which need to be excluded from the lifespan equation.

The model assumes that the smoking variables, t_{i1} and t_{i2}, have a causal effect on mortality, t_{i3}, and that t_{i1} has a causal effect on the risk of quitting smoking, t_{i2}. It also assumes that t_{i1} and t_{i2} are potentially endogenous. We assume that endogeneity depends on an unobserved random effect which is common to the three equations and is distributed independently of the response and the observed covariates. Therefore the error structure in the model is characterised by a latent factor which reflect the fact that the hazard of quitting depends on unobservables affecting also age at starting (for example, propensity to experiment) and the hazard of dying depends on unobservables affecting selection into smoking duration (for example, individual frailty).

Latent factor models have been employed to account for endogeneity of regressors in simultaneous equations models and selection bias due to unmeasured variables. Many studies of health care utilisation take into account the endogeneity of the insurance status chosen by the patient and estimate models for binary, continuous or count outcome variables and endogenous binary regressors (see e.g., Goldman, 1995; Mello et al., 2002; Deb and Trivedi, 2006).

Unobservable heterogeneity is integrated out either by specifying a continuous distribution for heterogeneity and evaluating the integrals in the likelihood by quadrature or simulation methods, or by approximating the heterogeneity distribution by discrete mass-points (see e.g., Heckman and Singer, 1984; Mroz, 1999). Mroz (1999) proposes a discrete factor approximation in simultaneous equation models to estimate the impact of a binary regressor on a continuous outcome. In the context of quality of care in hospitals Picone et al. (2003) use a latent factor model where intensity of treatment and length of stay are treated as endogenous, and estimate a model for a binary dependent variable and two continuous endogenous variables. Original applications of the latent factor model using a discrete approximation of the heterogeneity distribution can be found also in van Ours (2003, 2004, 2006). Studying the dynamics in the use of drugs and wage effects, he estimates bivariate duration models for the duration of non-use of two drugs and, more recently, for starting and quitting using a specific drug. Any parametric assumption on heterogeneity is relaxed and the joint model is estimated in a semi-parametric framework. The advantage of using this methodology is that it gives consistent estimates under any distributional assumption about the error term, while maximum likelihood estimators strictly rely on joint normality or other parametric assumptions.

Assuming that our recursive model can be described by a one-latent factor model, the error process depends on a common shock that affects smoking behaviours and mortality:

$$\begin{aligned} \nu_{1i} &= \varrho_1 l_i + u_{1i} \\ \nu_{2i} &= \varrho_2 l_i + u_{2i} \\ \nu_{3i} &= \varrho_3 l_i + u_{3i} \end{aligned} \tag{10}$$

where the error components u_i are independent in the three duration equations and l_i

is the unobservable heterogeneity term that can be considered as an approximation for the unmeasured or missing variables. The ϱ_s are factor loadings and can be interpreted as coefficients for the unobserved variables.

The joint distribution of the errors is:

$$f(\nu_{1i}, \nu_{2i}, \nu_{3i}) = \int_{-\infty}^{\infty} f(\nu_{1i}, \nu_{2i}, \nu_{3i}|l) \cdot dG(l) \tag{11}$$

where $f(\nu_{1i}, \nu_{2i}, \nu_{3i}|l)$ is the joint distribution conditional on l and $G(l)$ is the marginal continuous distribution function of l. Given independence of u_{1i}, u_{2i}, u_{3i}, equation (11) is equivalent to:

$$f(\nu_{1i}, \nu_{2i}, \nu_{3i}) = \int_{-\infty}^{\infty} f(\nu_{1i}|l)\, f(\nu_{2i}|l)\, f(\nu_{3i}|l) \cdot g(l)\, dl \tag{12}$$

This specification allows us to integrate out heterogeneity and implies specifying a parametric distribution for the density $g(l)$. In the case of our model, as specified in (7), the contribution to the sample likelihood of individual i is:

$$\begin{aligned} L_i = \int_{-\infty}^{\infty} \Big\{ & f(t_1|x_1, l; \beta_1, \theta_1, \varrho_1) \\ & h(t_2|x_2, l; \alpha_2, \beta_2, \theta_2, \varrho_2)^q \cdot S(t_2|x_2, l; \alpha_2, \beta_2, \theta_2, \varrho_2) \\ & h(t_3|x_3, l; \alpha_3, \delta_3, \beta_3, \theta_3, \varrho_3)^d \cdot \frac{S(t_{i3}|x_3, l; \alpha_3, \delta_3, \beta_3, \theta_3, \varrho_3)}{S(\tau_3|x_3, l; \alpha_3, \delta_3, \beta_3, \theta_3, \varrho_3)} \Big\} \cdot g(l)\, dl \end{aligned} \tag{13}$$

The first component of the likelihood, $f(t_1)$, is the density function for time spent in the status of non smoking, for every individual, conditional on a vector of observed regressors affecting smoking initiation, x_1, and the latent factor l. Ex-smokers, $(q = 1)$, for whom a complete spell of smoking years is observed, contribute with both the hazard and survival functions, $h(t_2), S(t_2)$, while current smokers, $(q = 0)$, who have a censored spell, contribute only with the survival function, $S(t_2)$. The hazard of dying is observed for both ex-smokers and current smokers: those who have a complete spell $(d = 1)$ are represented by the hazard function

$h(t_3)$, while those who are still alive at the time of the follow-up are represented by the left-truncated survival function $\frac{S(t_3)}{S(\tau_3)}$ where τ_3 is the truncation variable, age at the time of the first interview.

The error variance-covariance matrix deriving from the error structure in equation (10) depends on the distribution of the latent factor l_i, and the distribution of the error terms in each equation:[6]

$$cov \begin{pmatrix} \nu_{1i} \\ \nu_{2i} \\ \nu_{3i} \end{pmatrix} = \begin{bmatrix} \varrho_1^2 V_{(l_i)} + V_{(u_{1i})} & \varrho_1 \varrho_2 V_{(l_i)} & \varrho_1 \varrho_3 V_{(l_i)} \\ ... & \varrho_2^2 V_{(l_i)} + V_{(u_{2i})} & \varrho_2 \varrho_3 V_{(l_i)} \\ ... & ... & \varrho_3^2 V_{(l_i)} + V_{(u_{3i})} \end{bmatrix} \tag{14}$$

where $V_{(l_i)}$ is the variance of the latent factor and $V_{(u_i)}$ on the diagonal are the variances of the independent error components.

The effect of the latent factor l_i on smoking durations and mortality hazard is captured by the factor loadings, the cross-product of which determines the association between the errors of the three durations models. As a result, we look at the signs of the elements of matrix (14) to test hypotheses on the effect of unobserved heterogeneity. The covariance between errors in the smoking initiation and cessation models is measured by $\varrho_1 \varrho_2 V_{(l_i)}$, which is expected to be positive in order to reflect smoking experimentation; $\varrho_1 \varrho_3 V_{(l_i)}$ is the covariance between errors of the smoking initiation and lifespan model and a negative value is expected from Adda and Lechene (2001)'s argument that frailty is a source of selection into smoking; finally $\varrho_2 \varrho_3 V_{(l_i)}$, which is the covariance between the errors of the smoking cessation and lifespan model, is expected to be negative according to Balia and Jones (2007)'s argument that of frailty is a source of selection out of smoking.

[6]Analysis needed to be done in order to find the best distribution fitting the data. We model u_{1i} using a log-logistic density, u_{2i} using a Weibull density and u_{3i} using a Gompertz density. This is explained in the next section.

2.2. *Estimation techniques*

For simplicity, equation (13) can be written as:

$$L_i = \int_{-\infty}^{\infty} \Im(\cdot)\, g(l)\, dl \tag{15}$$

To evaluate the integral in (15), which does not exist in closed form, parametric and semiparametric estimation methods can be used. We use two parametric procedures, Gauss-Hermite Quadrature (GHQ) and Maximum Simulated Likelihood (MSL), and a semi-parametric approach, the Discrete Factor Model (DFM).

GHQ evaluates the integral in (15) using numerical integration by quadrature. The likelihood is approximated to a weighted sum of densities evaluated at different points:

$$L_i \approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^{M} \omega_j\, \Im(\sqrt{2}\varrho a_j) \tag{16}$$

The weights (ω_j) and the abscissas (a_j) are tabulated in Abramowitz and Stegun (1965), M is the number of elements in the vector of weights or abscissas and ϱ is the factor loading. The accuracy of GHQ depends on the number of evaluation points used to approximate the integral. The log-likelihood of the model and the estimated coefficients from different numbers of evaluation points need to be compared.

As an alternative MSL can be used. This means maximising a log-likelihood based on a simulated estimate of the density, where the simulator of the density is obtained by Monte Carlo integration. In this case, a finite number of draws from $g(l)$ is necessary to approximate the density and the individual contribution to the simulated likelihood function becomes a sample average over the number of draws (R):

$$\widehat{L}_i = \frac{1}{R} \sum_{r=1}^{R} \Im(\varrho z^r) \tag{17}$$

where z^r are draws from a standard normal distribution. An asymptotic property

of MSL is that for larger R and N the estimator is unbiased and consistent, but the literature suggests that increasing the number of draws is more computationally intensive. Compared to GHQ, however MSL technique has the adavantage of being more usable in the case of high-dimensional integrals.

The density can be computed either using random draws from a given $g(l)$, pseudo-random numbers, antithetic draws $(-z_j, z_j)$ or Halton sequences. The last two types of draws allow for negative correlation over observations which reduces the variance of the simulated density function. We use Halton draws, which are based on a non-random selection of points within the domain of integration and are known to cover well the domain of the sampling distribution (see Train, 2003).

The GHQ and the MSL estimations are based on normality of the heterogeneity term: the latent factor has a standard normal distribution, $l_i \sim N(0,1)$. The DFM offers a semi-parametric alternative which has the advantage of reducing the bias in the identification of the distribution of the latent factor when it is non-normal. Heckman and Navarro (2007) discuss semi-parametric identification of the distribution of unobservable heterogeneity in models with treatment times: they adopt a common factor specification as well.

A finite density estimator can be derived that approximates the unknown density $g(l)$ using a step function based on a set of mass points, η_k. The likelihood becomes:

$$L_i = \sum_{k=1}^{K} \pi_k \, \Im(\varrho \eta_k) \tag{18}$$

where $\pi_k = Pr(l = \eta_k)$ is a probability weight and η_k is the mass point, each $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. To ensure that the probability weights sum up to one, we give the probabilities π_k a logistic form, $\pi_k = \frac{e^{\zeta_j}}{1+\sum_{j=1}^{K-1} e^{\zeta_j}}$, where ζ need to be estimated to recover the probability weights. Mean and variance of the discrete distribution are calculated as $m_{(l)} = \frac{1}{K} \sum_{k=1}^{K} \pi_k \eta_k$ and $V_{(l)} = \sum_{k=1}^{K} \pi_k (\eta_k - m_{(l)})^2$.

As in the GHQ and MSL procedures, the factor loadings (ϱ) are estimated to-

gether with the other coefficients, the mass points and the probability weights are additional parameters to estimate.

Since the model includes an intercept for each equation, the location of the distribution of l is arbitrary; also the scale of l is arbitrary and undetermined (see Mroz, 1999). Therefore, identification of the DFM requires some normalisations. We impose two identification restrictions: we fix the first and last mass point respectively to 0 and 1, this means restricting both mean and variance of the discrete distribution which only depend on $K-1$ probability weights and $K-2$ mass points.[7]

3. Results

3.1. *The sample and the shape of the hazard and survival functions*

This section shows simple statistics that describe the raw data as well as the main results of the non-parametric and parametric analysis carried out on the univariate durations models. This is useful to identify the best distributions for our survival variables.

[7]The existing literature on the DFM offers a range of equivalent strategies to identify the additional parameters of the discrete distribution by fixing the scale and/or the location of the distribution. If both are fixed, one of the factor loading is set to 1 and either one of the η_k is set to 0 (see Mroz, 1999) or the mean of the discrete distribution is restricted to be null with the result that one of η_k can be expressed as a function of the others (Kan et al., 2003, see e.g.,). If only the location is fixed, the first and the last mass points are set respectively to 0 and 1 (this strategy is used by Mroz (1999) when $k > 2$). Other applications also impose that the middle mass points follow a logistic distribution such that $\eta_k \in (0,1)$ (see Mello et al., 2002; Picone et al., 2003). The π_k can be parameterised using various distributions as the logistic, the normal or the sine function such that each π_k is between 0 and 1 and they sum up to 1. We have compared our approach to identification to these other approaches to check robustness of the results.

Table 2

Variable definitions and summary statistics

Variable name	Variable definition	Mean	S.D.
Survival time variables			
and censoring indicators			
starting	number of years elapsed before starting	33.184	21.543
sm_years	number of years a person smoked	32.237	14.008
lifespan	number of years lived by April 2005	73.999	9.624
start	1 if started smoking before the HALS, 0 otherwise	0.624	0.484
quit	1 if quitted smoking before the HALS, 0 otherwise	0.503	0.500
death	1 if has died by April 2005, 0 alive	0.428	0.495
Observed characteristics			
sc1	1 if professional/student or managerial/intermediate, 0 otherwise	0.300	0.458
sc2	1 if skilled or armed service, 0 otherwise	0.474	0.499
sc3	1 if partly skilled, unskilled, unclass. or never occupied, 0 otherwise	0.226	0.418
degree	1 if University degree, 0 otherwise	0.117	0.321
hvqA	1 if higher vocational qualifications or A level or equivalent, 0 otherwise	0.116	0.321
O-cse	1 if O level/CSE, 0 otherwise	0.091	0.288
no edu.	1 if no qualification, 0 otherwise	0.625	0.484
other edu.	1 if other vocational/professional qualifications, 0 otherwise	0.050	0.218
married	1 if married, 0 otherwise	0.752	0.432
widow	1 if widow, 0 otherwise	0.129	0.335
sepdiv	1 if separated or divorced, 0 otherwise	0.056	0.229
single	1 if single, 0 otherwise	0.063	0.244
full time	1 if full time worker or student, 0 otherwise		
part time	1 if part time worker, 0 otherwise	0.129	0.335
unemployed	1 if the individual unemployed, 0 otherwise	0.031	0.174
sick	1 if absent from work due to sickness, 0 otherwise	0.033	0.178
retired	1 if retired, 0 otherwise	0.356	0.479
housekeeper	1 if housekeeper, 0 otherwise	0.097	0.296
rural	1 if lives in the countryside, 0 otherwise	0.214	0.410
suburb	1 if lives in the suburbs of the city, 0 otherwise	0.462	0.499
household size	number of other people in the house	1.602	1.253
mother smoked	1 if only mother smoked, 0 otherwise	0.046	0.209
father smoked	1 if only father smoked, 0 otherwise	0.580	0.494
both smoked	1 if both parents smoked, 0 otherwise	0.228	0.419

continued on next page

Table 2 – *continued from previous page*

Variable name	Variable definition	Mean	S.D.
others smoked	1 if anyone else in house smoked, 0 otherwise	0.345	0.475
cohsmo	1 if person started smoking after 1954 , 0 otherwise	0.242	0.428
male	1 if male, 0 otherwise	0.454	0.498
age	age in years	58.046	11.754

For the purpose of our analysis, the sample from the HALS has been reduced according to item non-response in the variables of interest. Furthermore, the cumulative distribution of age at death suggests restricting the analysis to individuals older than 40.[8] The remaining sample consists of 4646 individuals of whom about 45 per cent are males and the mean age is 58 years. About 43 per cent of the respondents had died by April 2005 and the mean lifespan is 74 years. Those who started smoking at some point in their life account for the 62 per cent of the sample among whom about 50 per cent had stopped smoking at the time of the interview in the HALS (1984-85). On average, current and ex-smokers in the sample have smoked for around 32 years. Complete summary statistics are reported in Table 2.

All the individuals in the sample are at risk of starting smoking from time 0 (i.e., there is no left-censoring) and the data contain a record of *starting* for each subject at risk: the median survival before starting smoking is 21 as shown in Table 3. Here the analysis looks at both starters and never smokers, hence censored survival times, which are longer than completed spells, are considered. Non-parametric estimation of the survivor, hazard and cumulative hazard functions are obtained using the Kaplan-Meier (KM) or product-limit estimator. This implies ordering the observed failure times from a sample of censored survival data as times $t_1 < t_2 < ... < t_j <$

[8]Only 1 per cent of the sample died before age 40, so this small part of the sample is not retained in the analysis. This allows us to avoid confounding mortality with accidents, injuries or a genetic predisposition towards early death not related to smoking.

Table 3
Median survival times

	starting	starting	quitting	lifespan
actual	21	17	43	80.2
predicted	34.293	17.356	41.639	78.355

$...t_k < \infty$. Each time interval contains a death time and this death time is assumed to occur at the start of the interval, while the censored times fall into the intervals. The empirical survivor function, from which it is possible to derive the estimated failure function and the integrated hazard function is as follows:

$$\hat{S(t)} = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} \tag{19}$$

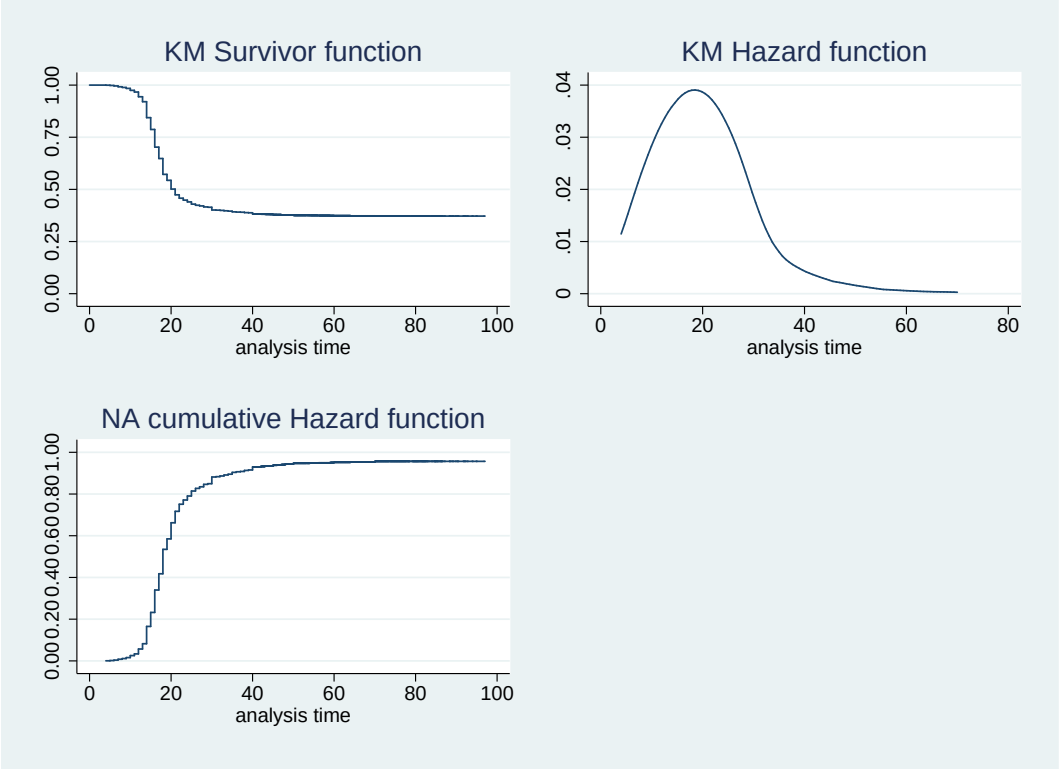
Where n_j and d_j are respectively the number of persons at risk of making a transition and the number of persons for which exit is observed. Therefore, the estimated survivor function is the product, for each time j, of the ratio between those who survive and the total number of persons at risk; it has the shape of a step function with origin at $t = 0$ and at each t_j the height is equal to $S(t_j)$. The cumulative hazard function is estimated using the Nelson-Aalen (NA) estimator that behaves better than the KM estimator in small samples:

$$\hat{H(t)} = \sum_{j=1}^{k} \frac{d_j}{n_j} \tag{20}$$

Figure 2 shows the survivor function, with survival diminishing faster for individuals aged 15 to 20 years and then less than proportionally. In fact, the hazard of starting smoking increases up to age 20 and then falls closer to 0 for individuals of 35 years or older.

About half of the 2910 individuals who started smoking claimed to be ex-smokers, with year of starting ranging from 1906 to 1982 and year of quitting from 1915 to 1986. The median duration of smoking is 43 years; 25 per cent of the sample survived for 26 years and 75 per cent for as long as 59 years. Non-parametric estimates for

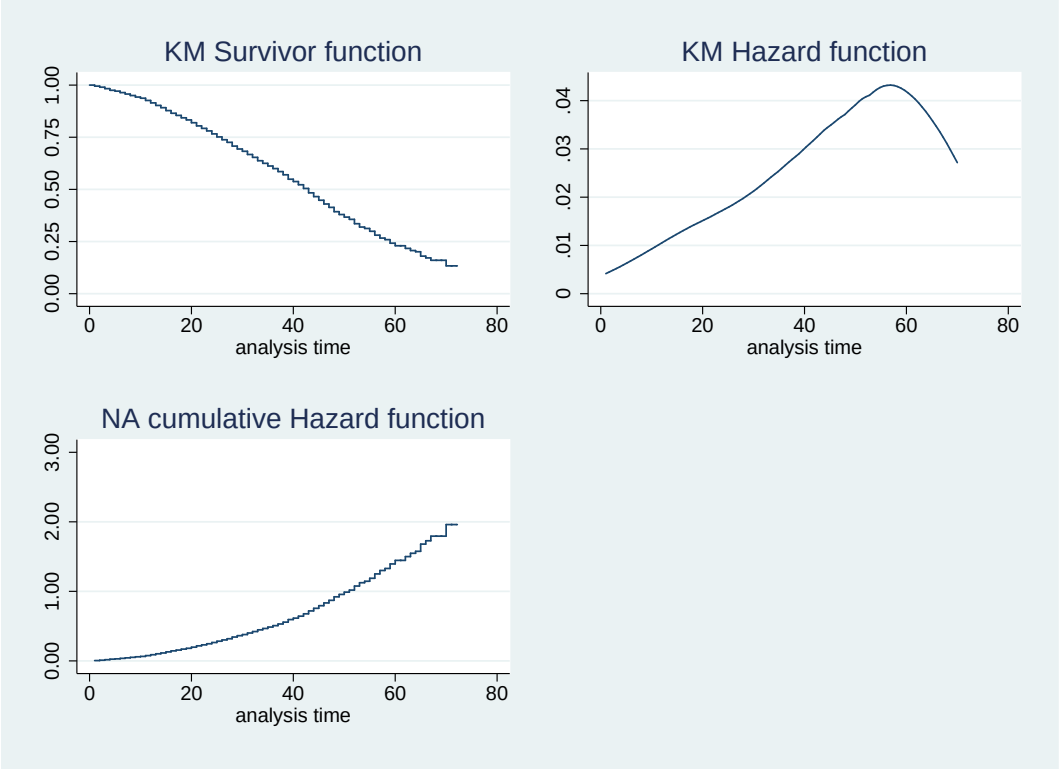
Figure 2
Non-parametric functions for smoking initiation



sm_years are shown in Figure 3 and show a decreasing pattern of survival, which diminishes less than proportionally for the first 40 years of the analysis time. In fact, the probability of surviving as a current smoker is still high (between 1 and 0.50) for durations as long as 40 years, after which it starts decreasing faster. The shape of the hazard function is increasing, with a peak between 50 and 60 years of smoking, and then decreases. This is confirmed by the large jump in the cumulative hazard function at the highest survival times.

To analyse $lifespan$ we need to consider that individuals who died before the observation time in the HALS (1984-85) are not surveyed but are excluded from the sample. The idea of exclusion must not be confused with the problem of missing observations, but simply refers to the fact that the HALS sample is the result of a selection process, conditional on the event of death having not occurred prior to the survey time. We could think of the HALS sample as made up of individuals who have a relatively lower hazard of dying, since individuals with a higher hazard will have

Figure 3
Non-parametric functions for smoking cessation



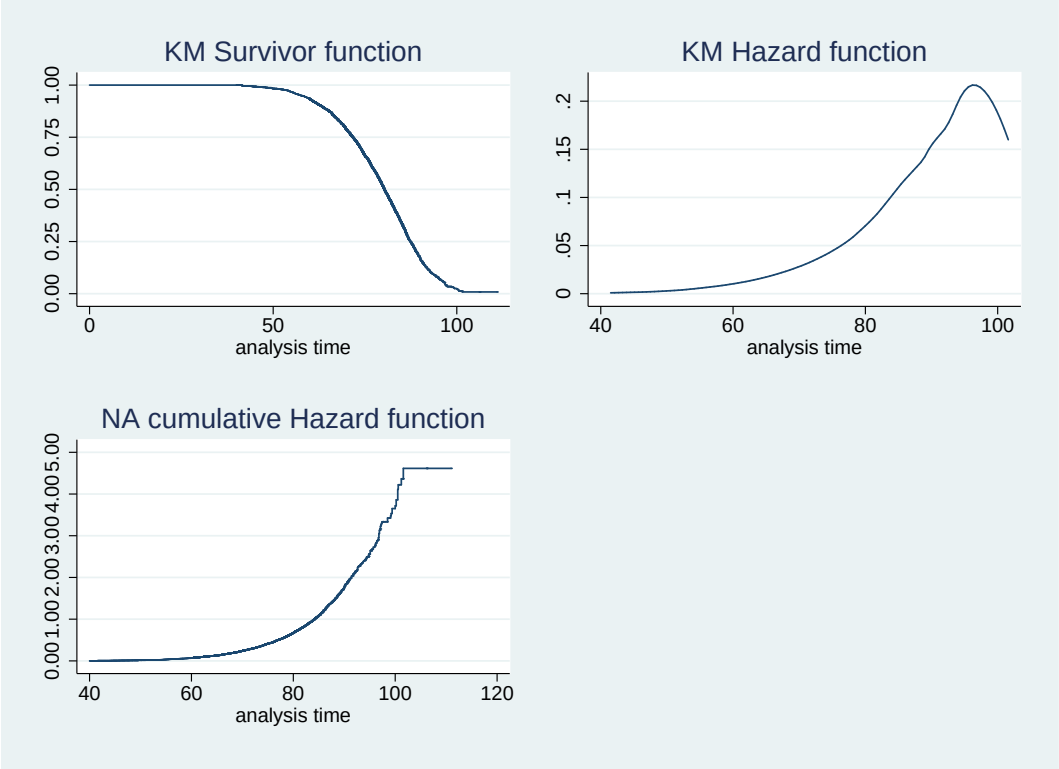
died before any information had been collected about them. Hence, the remaining sample would have a lower hazard than the truncated part of the population. The median survival time is 80 years. Figure 4 shows that the survivor function has a decreasing pattern and the hazard function is increasing: the latter is steeper for durations longer than 80 years and shows a decreasing pattern after the peak, around age 95, where the cumulative hazard function flatters out.

3.2. *Results from univariate duration models*

This section illustrates results from univariate duration models for smoking and lifespan. Regression models are estimated under the restrictive assumption of exogeneity, that will be relaxed in the next step of the analysis.

The regression approach to duration analysis requires some testing in order to specify the parametric distribution which best fits the data. We use the cumulative Cox-Snell residuals as well as statistical tests that penalise the log-likelihood,

Figure 4
Non-parametric functions for lifespan



like the Akaike and Bayesian information criteria (AIC and BIC), to discriminate among distributions that represent smoking durations and lifespan. We compare the exponential, Weibull, log-normal and log-logistic distributions.[9]

Ideally, to analyse smoking dynamics, information recorded at the time when the individual started and quit smoking (e.g., price of tobacco, family situation, social context, peer influences and so on) would be desirable. Unfortunately this is not available in the dataset we use. In this work, smoking initiation and cessation are specified as a flexible function of age and depend on demographics, socio-economic controls (social class, education level, occupational status), geographical variables

[9]The exponential is the most basic model: it describes a flat hazard function, that is the hazard function is constant over time, and requires no additional parameter to be estimated (the shape parameter is set equal to one). The Weibull has a more general form of the hazard function, which can be monotonically increasing or decreasing depending on the shape parameter. If the shape parameter equals 1, then the Weibull reduces to the exponential distribution. The log-logistic and the log-normal represent the logarithm of time using a logistic and a normal distribution respectively. They tend to produce similar results: the shape parameter can describe either a monotonically decreasing hazard rate or a first increasing and then decreasing hazard rate.

and marital status, as in 1984-85, under the assumption that they reflect past social and economic background.[10]

Smoking initiation may be influenced by past parental smoking: this is taken into account including dummy variables for parents' smoking. We also suppose that smoking behaviours (both initiation and cessation) are influenced by awareness of the the health consequences of smoking. This can be captured by defining cohorts according to a particular historical period.[11] In the UK the first scientific report was published in 1954 (Doll and Hill, 1954) and in the same year, the Minister of Health reported on the findings of a Government-approved scientific committee which had been investigating possible links between smoking and lung cancer. The committee concluded that there exist a relationship between smoking and lung cancer and that risk increases particularly with the amount of cigarettes smoked. We include an indicator to control for the effect of the first dissemination of smoking effects on health: individuals who started smoking after 1954 are assumed to be more aware about the health risks of smoking, therefore they might start later and quit earlier.

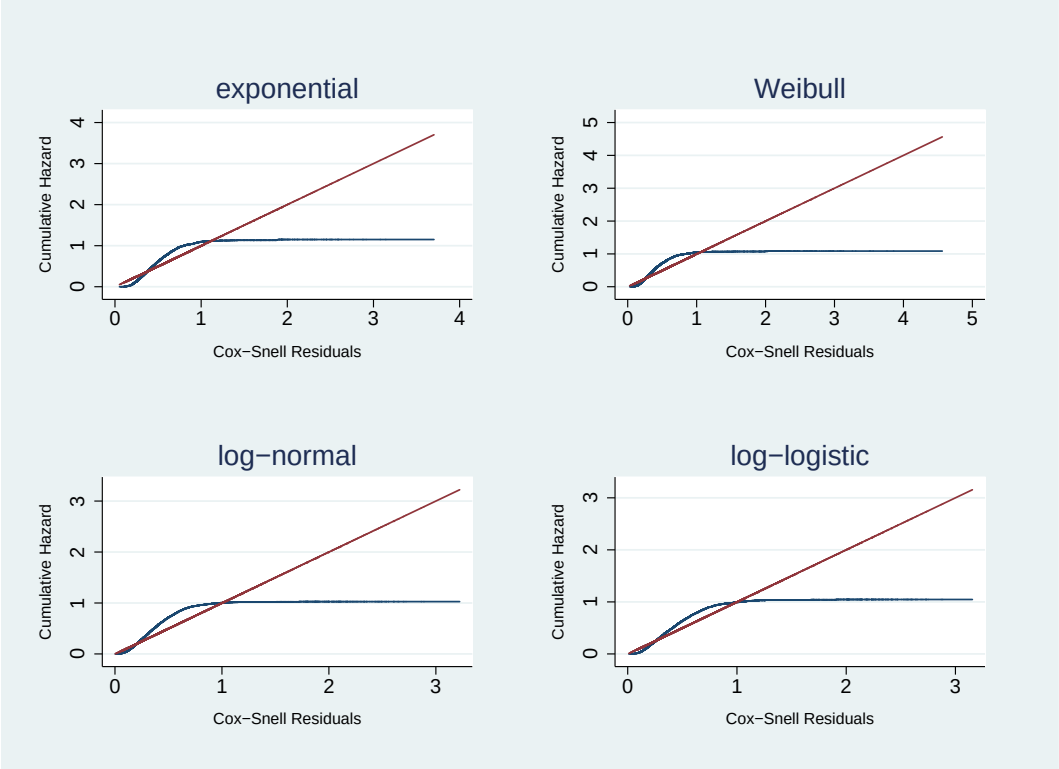
The Cox-Snell residuals test and the information criteria, as reported in Figure 5 and Table 4, suggest that the distributions with the best fit are the log-normal and the log-logistic. We present results from the estimation of the log-logistic model which has been shown to best fit the HALS data in Forster and Jones (2001).

In a first run of the model we exploited information from the full sample of respondents, however, predicted median age of starting were far too high when compared to the actual values, and this gives an indication that the model is inadequate. This is shown in Table 3. Standard duration models assume that eventually everyone fails. In this case, everyone would eventually start smoking. This seems to be

[10]Education could also be interpreted as a signal of individual ability which would explain differences in the onset of smoking for individuals with different ability. In this view, education is exogenous in the determination of smoking duration and does not represent educational investments. van Ours (2006) use education as an indicator of ability in models of dynamics drug use.

[11]In a study on the US population, Farrell and Fuchs (1982) assume that the youngest cohort are more aware about the health consequences of smoking and use the Second World War, the year of publication of the first article linking smoking to lung cancer, and the publication of the first official report on smoking and health as critical years to define cohorts.

Figure 5
Cox-Snell residuals test for smoking initiation



an implausible assumption, and models based on this assumption do not do a good job of fitting the observed data. An alternative is to use a split population model. This augments the standard duration analysis by adding a splitting mechanism. So, for example, a logit or probit specification could be added to model the probability that somebody will eventually start smoking. When this splitting mechanism is added to the duration model, it does a far better job of explaining the observed data on age of starting than models that omit a splitting mechanism (see Forster and Jones, 2001). The results of Forster and Jones suggest that a simplified version of the split population model, which can be viewed as a two-part specification of the duration model, will work well with the HALS data. This uses a standard binary choice model, such as a logit or probit, for the indicator of whether an individual has started and then applies the duration model only to the starters in the sample. As a result, we select a sub-sample of starters only and replicate the non-parametric

Table 4
Comparison between distributions for duration models

Information criterium	exponential	Weibull	log-normal	log-logistic	Gompertz
Smoking Initiation (N = 4646)					
AIC	11671.870	11497.430	10469.330	10621.200	
BIC	11826.520	11658.530	10630.430	10782.300	
Smoking Initiation (N=2901)					
AIC	6033.338	612.578	340.021	137.149	
BIC	6176.685	761.898	489.341	286.469	
Smoking Cessation (N=2901)					
AIC	5766.691	5085.732	5444.553	5203.633	
BIC	5904.065	5229.079	5587.900	5346.980	
Lifespan (N=2901)					
AIC	−723.702	−1003.105	−860.944	−881.075	−1004.825
BIC	−604.245	−877.676	−735.5147	−755.6456	−879.396

analysis as well as the estimation of the regression model.[12] The KM estimate of the survival function on the sub-sample, as reported in Figure 6, shows a steeper decrease around the age of 17-20 than the same curve for the full sample; the KM hazard function, in the same figure, is first increasing and then decreasing but the decrease starts later and is not so steep towards zero as in the full sample.

Estimation of the duration model for the sub-sample of starters means that each individual has a complete spell for *starting* and the contribution to the sample likelihood for individual i is simply given by their density function:[13]

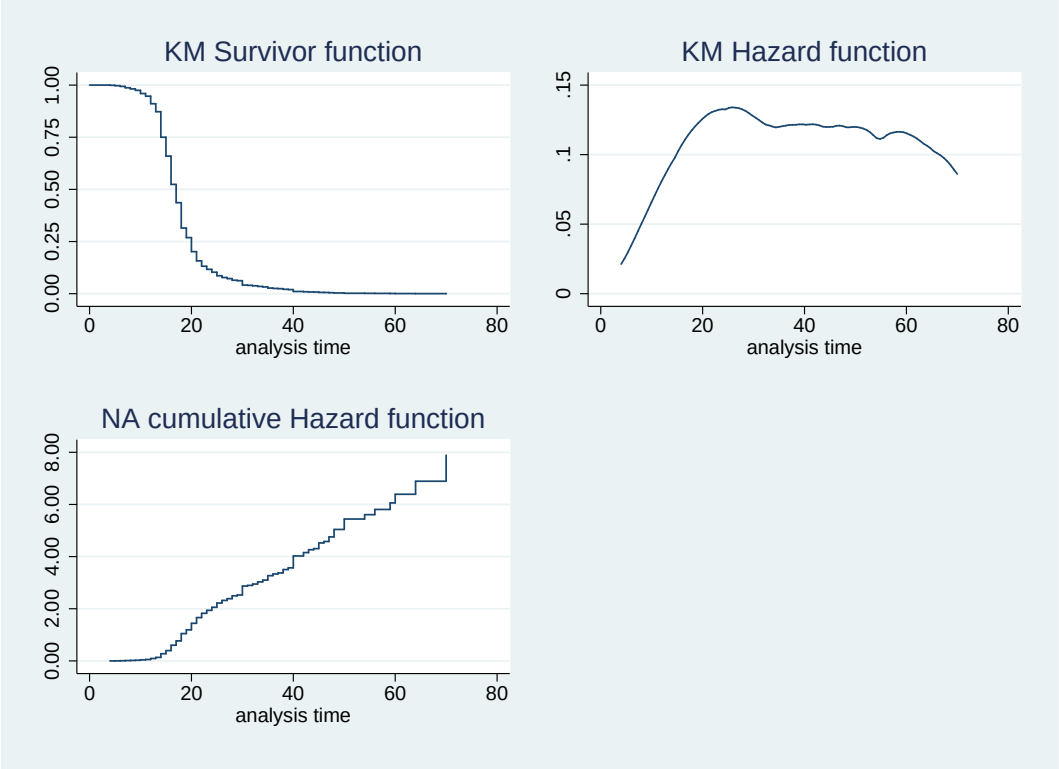
$$L_i = \frac{\phi_i^{\frac{1}{\gamma}} t_i^{\frac{1}{\gamma}} - 1}{\gamma[1 + (\phi_i t_i)^{\frac{1}{\gamma}}]^2}$$

Where $\phi_i = exp(-x_i\beta)$ is a non-negative function that depends on observed characteristics and whose shape is given by the ancillary parameter γ: if $\gamma \geq 1$ the

[12]The choice of the log-logistic distribution is based again on the comparison of different distributions.

[13]Note that the censoring indicator disappears. everybody in the sub-sample did start smoking at some point in life. When the model is estimated over the full sample, the censoring indicator helps to identify individuals for whom we observe a complete spell (non censoring) and individuals for whom the survival time is censored to the last observation.

Figure 6
Non-parametric functions for smoking initiation - only starters



hazard rate is monotonically increasing; if $\gamma < 1$ the hazard first rises and then decreases monotonically. The log-logistic model has an AFT metric.

Graphical analysis from the fitted model, reported in Figure 7, shows that survival declines rapidly from ages 17-18; the hazard is predicted to rise and then fall monotonically. Table 5 reports results from the regression model: time to starting is predicted to accelerate for men, individuals in the bottom social classes, with no education, part-time workers, housekeepers, retired and workers in sick leave as well as for individuals whose father or both parents used to smoke. Survival time is predicted to be longer for older individuals and for those who started smoking after 1954, when knowledge of health risks began to be disseminated. Predicted median age at starting is about 17, very close to the actual value as reported in Table 3.

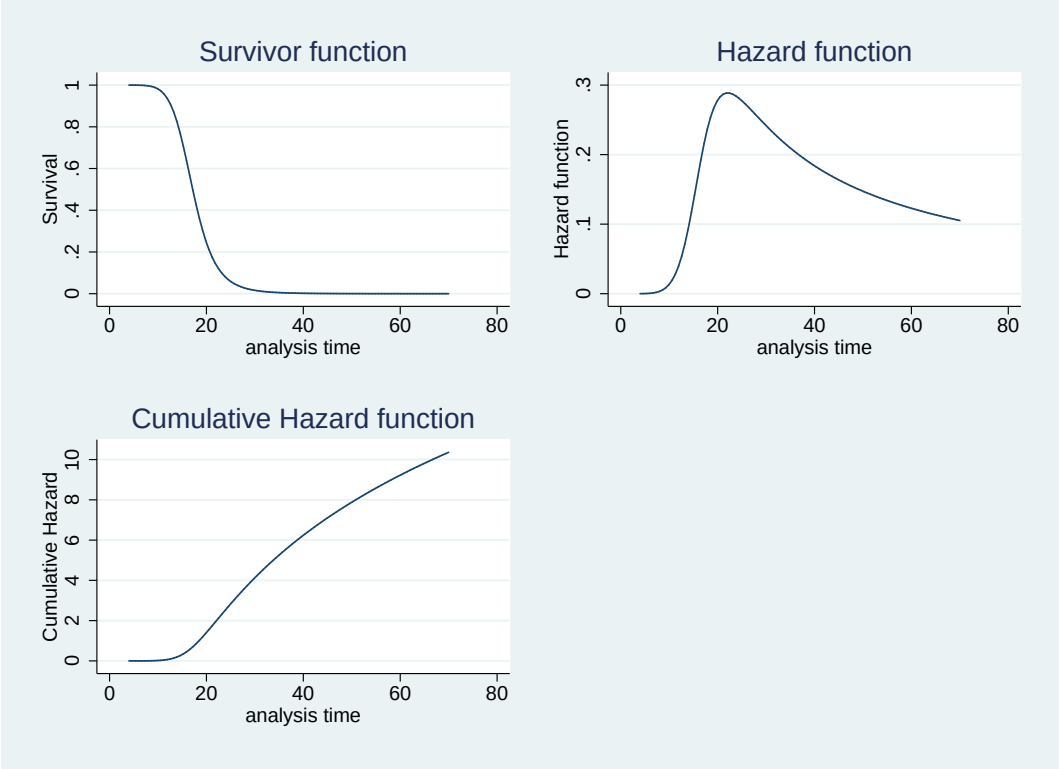
To estimate the hazard of quitting we exclude past parental smoking variables as they are likely to have a direct effect on age at starting but not on smoking cessation, and we assume that smoking behaviours in the household at the time of the survey

Table 5
Results from univariate duration models

	Smoking initiation (AFT metric)		*Smoking cessation* (AFT metric)		*Lifespan* (PH metric)	
Variable	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
ln(starting)			-0.481^{**}	0.052	-0.232^{*}	0.099
quit					-0.222^{**}	0.066
ln(sm_years)					0.250^{**}	0.064
sc1	0.006	0.011	-0.083^{*}	0.033	-0.101	0.078
sc3	-0.024^{*}	0.011	0.027	0.034	0.126^{*}	0.064
degree	0.020	0.020	-0.082	0.059	0.032	0.153
hvqA	0.021	0.020	0.028	0.058	-0.041	0.151
no edu.	-0.035^{*}	0.016	$0.090^{\dagger}$	0.049	0.180	0.119
other edu.	-0.010	0.024	$0.128^{\dagger}$	0.072	$0.280^{\dagger}$	0.159
part time	-0.067^{**}	0.016	0.000	0.050	0.031	0.122
unemployed	-0.012	0.023	0.186^{*}	0.084	0.492^{**}	0.150
sick	-0.068^{**}	0.024	0.096	0.073	0.638^{**}	0.128
retired	-0.106^{**}	0.016	0.065	0.046	-0.003	0.098
housekeeper	-0.055^{**}	0.018	0.036	0.061	$0.259^{\dagger}$	0.145
rural	0.012	0.012	$-0.067^{\dagger}$	0.037	-0.006	0.078
suburb	$0.019^{\dagger}$	0.010	$-0.055^{\dagger}$	0.030	0.007	0.061
household size	0.002	0.005	-0.019	0.015	-0.038	0.032
male	-0.181^{**}	0.011	-0.131^{**}	0.033	0.349^{**}	0.069
ln(age)	0.794^{**}	0.049	0.496^{**}	0.146	$0.673^{\dagger}$	0.382
widow	0.010	0.016	0.095^{*}	0.045		
sepdiv	0.014	0.020	0.325^{**}	0.076		
single	0.043^{*}	0.019	0.145^{*}	0.060		
cohsmo	0.328^{**}	0.015	-0.223^{**}	0.048		
mother smoked	-0.026	0.024				
father smoked	-0.040^{**}	0.015				
both smoked	-0.062^{**}	0.016				
others smoked			0.390^{**}	0.032		
cons	-0.249	0.199	3.141^{**}	0.571	-12.565^{**}	1.303
γ	0.136^{**}	0.002				
p			2.036^{**}	0.049		
φ					0.084^{**}	0.005

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

Figure 7
Log-logistic functions for smoking initiation

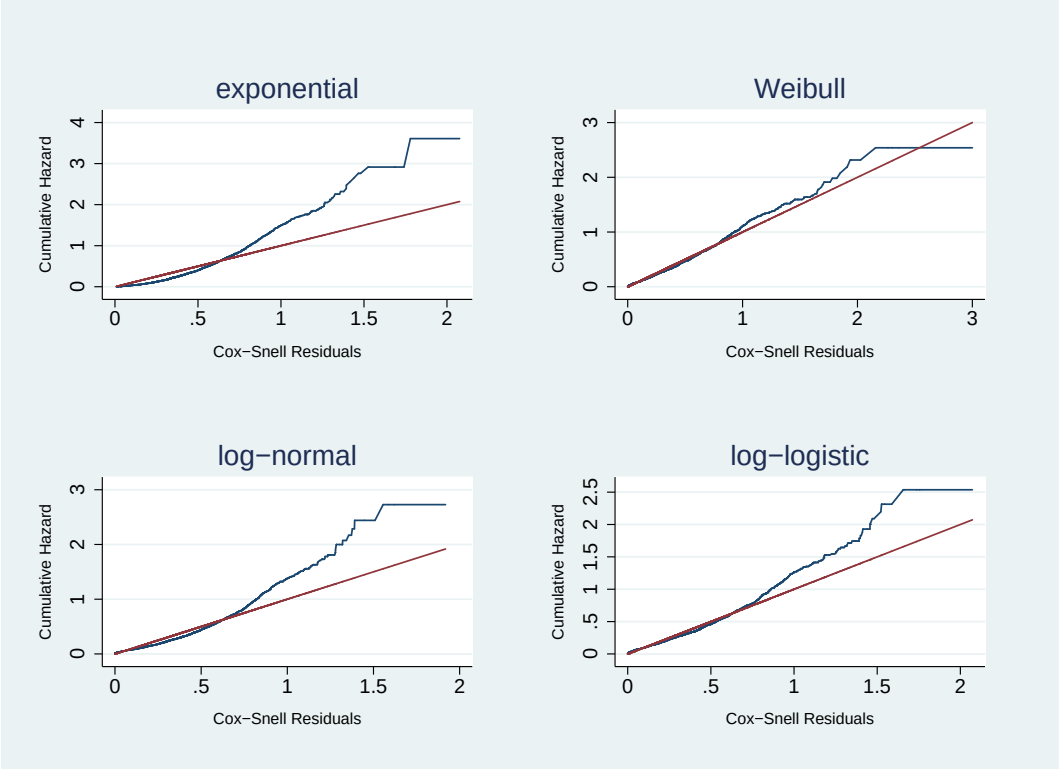


can influence time to quit smoking. The logarithm of age at starting is included in the model, under the restrictive assumption of exogeneity, that will be relaxed in the next step of the analysis. The graphical analysis in Figure 8 shows that, except for the exponential, all the distributions fit the data quite well. The information criteria help to discriminate among these and favour the Weibull distribution (Table 4). The sample likelihood to maximise is the product of the density and the survival functions:

$$L_i = [\lambda_i p t_i^{p-1} exp(-\lambda_i t_i^p)]^{q_i} \cdot [exp(-\lambda_i t_i^p)]^{1-q_i}$$

where q_i is the censoring indicator that separates the contribution of ex-smokers from the contribution of current smokers: complete spells are recorded for ex-smokers ($q_i = 1$) while censored spells are measured for current smokers ($q_i = 0$). λ_i is a non-negative function that depends on observed characteristics, $\lambda_i = exp(-px_i\beta)$;

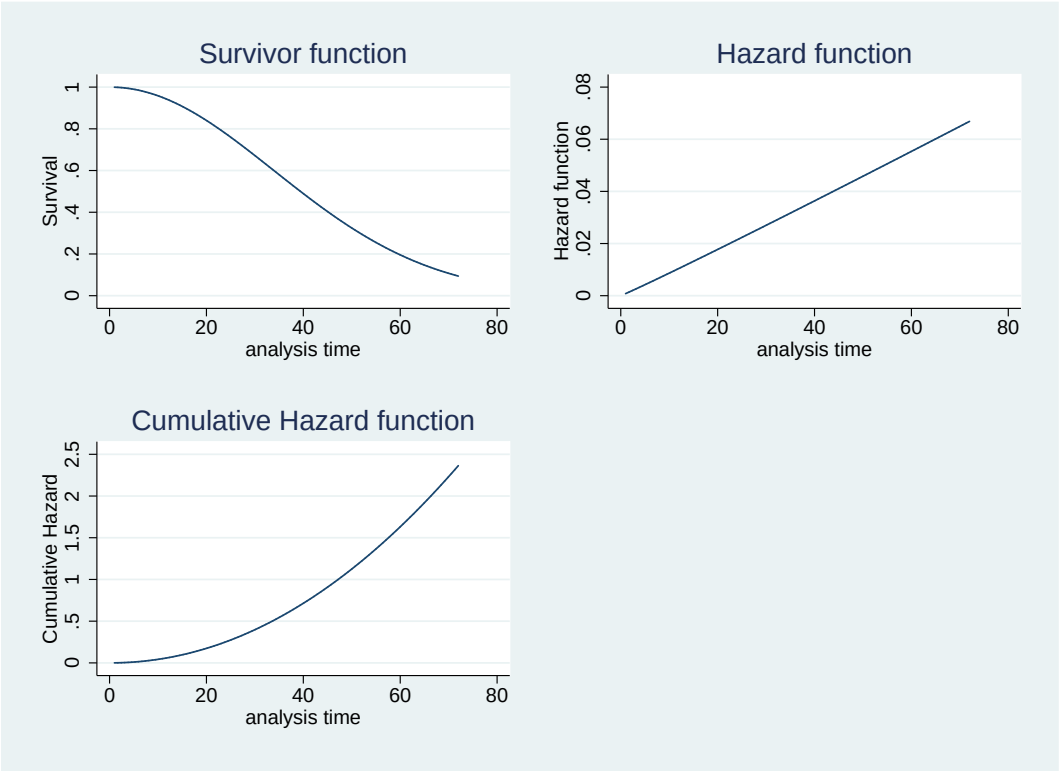
Figure 8
Cox-Snell residuals test for smoking cessation



pt_i^{p-1} is the baseline hazard whose shape depends on the ancillary parameter p. The Weibull model can yield a monotonic increasing or decreasing hazard of quitting: if $p = 1$ the Weibull equals the exponential, with $h(t) = \lambda$; if $p > 1$ the hazard function is monotonically increasing and when $p < 1$ the hazard function is monotonically decreasing, representing respectively positive and negative duration dependence. The Weibull model is estimated using the AFT metric.

The graphical analysis allows comparison of the fitted survivor and hazard functions with the non-parametric functions in Figure 3, and shows that the Weibull survivor function is a good match for the Kaplan-Meier survivor function: the fitted hazard function moves away from the empirical hazard function only for the right tail of the survival time distribution (Figure 9). Table 5 shows that time to quitting is predicted to be shorter for higher age at starting: starting smoking later in life is likely to accelerate time to quitting, meaning that late starters smoke relatively less. Shorter durations of smoking are predicted for men, individuals in the upper socio-

Figure 9
Weibull functions for smoking cessation

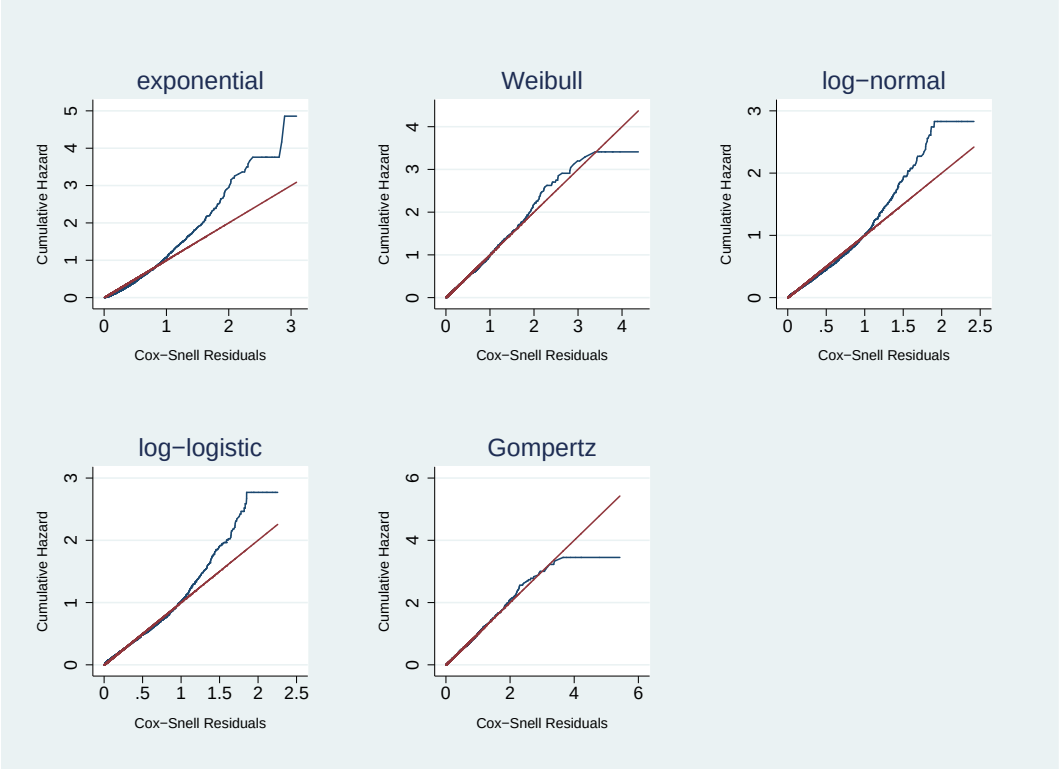


economic class, and for those who started smoking after 1954. Longer durations are predicted for the unemployed and not married individuals. Ageing and other smokers in the household decelerate time to quitting, so that duration of smoking is predicted to be longer. The model predicts positive duration dependence and a median time to quitting of about 42 years, very close to the actual value.

The hazard of dying is assumed to be a function of the same observed characteristics that are included in the smoking equations, however dummy variables for parents and others' smoking behaviours and dissemination effect are excluded, as well as marital status.[14] The logarithm of *starting* and *sm_years* and the censoring indicator for quitting are included as regressors. At this point, we do not attempt to deal with potential endogeneity of these variables, so estimates should be treated with caution if they are to be interpreted as causal effects of smoking on lifespan.

[14]Marital status is assumed to have a direct effect on lifestyles such as smoking rather than on the hazard of dying, and marital status as observed in 1984 for individuals aged 40 or over reflects marital status over a longer period of time, hence explaining smoking durations.

Figure 10
Cox-Snell residuals test for lifespan



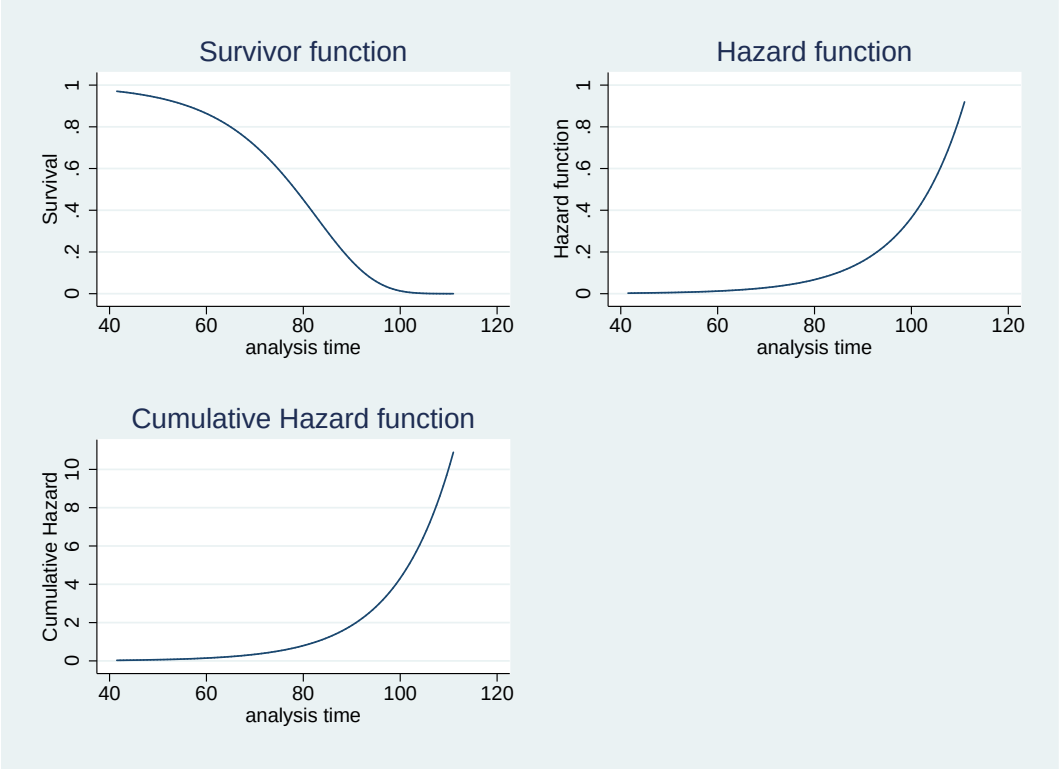
The Cox-Snell residual test and information criteria are used to compare five alternative distributions (see Figure 10 and Table 4), including the Gompertz distribution. The Gompertz is the favoured distribution.[15] The sample likelihood for a left-truncated and right-censored survival time variable is the product of the hazard function, observed only for deaths, and the survival function conditional on survival up to the interview date (where τ_i is age at the time of the interview) for every individual in the sample:

$$L_i = [\varphi_i exp(\mu t_i)]^{d_i} \cdot \frac{exp(-\frac{\varphi_i}{\mu}[exp(\mu t_i) - 1])}{exp(-\frac{\varphi_i}{\mu}[exp(\mu \tau_i) - 1])} \tag{21}$$

The Gompertz model is parameterized as a proportional hazard model or log-relative hazard form, where the baseline hazard is $exp(\mu t_i)$ and $\varphi_i = exp(x_i\beta)$

[15]The maximization of the likelihood in the Weibull model does not converge to a global maximum, so that tests based on the Weibull estimated coefficients and residuals are not reliable. The Gompertz mortality model is the most used in biology and medical modelling.

Figure 11
Gompertz functions for smoking cessation



scales the baseline hazard multiplicatively by the same amount at each instant t. If μ is positive the hazard function increases with time, if it is negative the hazard function decreases with time. The exponential hazard function is a special case of the Gompertz hazard when $\mu = 0$. The Gompertz model produces estimated functions, see Figure 11, that mimic very well the empirical functions reported in Figure 4. The hazard function increases with time and is convex. Estimated coefficients, reported in Table 5, show that the hazard of dying is significantly higher for men, it increases with age and for individuals from the bottom social class, poor education, unemployed, sick and housekeepers. The relationship between smoking behaviours and mortality is tested looking at the coefficients of the survival time variables and the censoring indicator: the hazard of dying is predicted to decrease with age at onset of smoking, meaning that starting later in life is negatively related with length of life. However, the model also predicts that ex-smokers have a lower hazard of dying, although, at the same time, the latter increases with time spent smoking.

Duration dependence is positive and median time to death is predicted to be about 78 years, quite close to the actual value.

3.3. *Results from the latent factor model*

The univariate duration models presented in this section suffer from not considering the endogeneity bias which potentially affects the estimates. Here we relax the restrictive assumption of exogeneity and estimate the model in equations (7) to (9) as a latent factor model with a common random variable representing individual-specific unobservable heterogeneity. Estimation is carried out by means of the techniques described in section 2.2. Coefficients and factor loadings are estimated simultaneously.

Tables 6 and 7 reports results from the latent factor model estimated using unidimensional GHQ and MSL. Integration by quadrature works well with 10 evaluation points. We checked sensitivity of the quadrature approximation using up to 24 evaluation points and comparing relative differences in coefficients estimates and model log-likelihood. For MSL we created 400 Halton draws per equation and dropped 10 initial draws to avoid correlation between the first elements of the sequence (see Train, 2003). Both parametric techniques lead to very similar estimated coefficients: only very small difference in the point estimates of the factor loadings and in the log-likelihood of the model can be noted.[16]

Table 8 reports results from the DFM with 3 points of support: based on a likelihood ratio (LR) test we prefer 3 to 2 points of support.[17] We tried unsuccessfully to include one additional point of support, but the algorithm does not converge

[16]To check robustness of results we estimated the model relaxing the assumption of unit variance. The new parametric distribution for the heterogeneity term is $l_i \sim N(0, \sigma^2)$, where σ^2 is the unknown variance, and one of the factor loadings need to be set to 1 for identification reasons. Results do not differ much in terms of log-likelihood and estimated coefficients even if the variance is unknown.

[17]The LR test is a $\chi^2_2 = 11.986$ with p-value= 0.003. Furthermore, coefficients from the DFM with $k = 3$ are much closer to those from the parametric estimations than those from the DFM with $k = 2$.

Table 6
Results from latent factor model - GHQ estimator

Variable	*Smoking initiation* (AFT metric) Coeff.	S.E.	*Smoking cessation* (AFT metric) Coeff.	S.E.	*Lifespan* (PH metric) Coeff.	S.E.
ln(starting)			−0.482**	0.053	−0.185	0.134
quit					−0.244**	0.080
ln(sm_years)					0.255**	0.068
sc1	0.007	0.011	−0.083*	0.033	−0.109	0.082
sc3	−0.024*	0.011	0.027	0.034	0.124†	0.067
degree	0.019	0.020	−0.082	0.059	0.033	0.160
hvqA	0.021	0.020	0.028	0.058	−0.048	0.157
no edu.	−0.035*	0.016	0.090†	0.049	0.186	0.124
other edu.	−0.010	0.024	0.128†	0.072	0.289†	0.166
part time	−0.067**	0.016	0.000	0.050	0.032	0.127
unemployed	−0.012	0.023	0.186*	0.084	0.509**	0.157
sick	−0.068**	0.024	0.096	0.073	0.680**	0.140
retired	−0.106**	0.016	0.065	0.046	0.002	0.103
housekeeper	−0.055**	0.018	0.036	0.061	0.278†	0.150
rural	0.012	0.012	−0.067†	0.037	−0.009	0.082
suburb	0.019†	0.010	−0.055†	0.030	0.008	0.064
household size	0.002	0.005	−0.019	0.015	−0.040	0.033
male	−0.181**	0.011	−0.132**	0.033	0.379**	0.079
ln(age)	0.794**	0.049	0.497**	0.146	0.418	0.453
widow	0.010	0.016	0.095*	0.045		
sepdiv	0.014	0.020	0.325**	0.076		
single	0.044*	0.019	0.145*	0.060		
cohsmo	0.328**	0.015	−0.222**	0.048		
mother smoked	−0.027	0.024				
father smoked	−0.040**	0.015				
both smoked	−0.062**	0.016				
others smoked			0.390**	0.032		
cons	−0.250	0.199	3.141**	0.571	−12.261**	1.374
γ	0.135**	0.002				
p			2.036**	0.049		
μ					0.092**	0.008
ϱ_1	−0.013	0.018				
ϱ_2	−0.006	0.049				
ϱ_3	0.334*	0.155				
logL:	-20706.918					
N:	2901					

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

Table 7
Results from latent factor model - MSL estimator

Variable	*Smoking initiation* (AFT metric) Coeff.	S.E.	*Smoking cessation* (AFT metric) Coeff.	S.E.	*Lifespan* (PH metric) Coeff.	S.E.
ln(starting)			−0.483**	0.054	−0.176	0.137
quit					−0.249**	0.082
ln(sm_years)					0.257**	0.069
sc1	0.007	0.011	−0.083*	0.033	−0.110	0.082
sc3	−0.024*	0.011	0.027	0.034	0.124†	0.068
degree	0.019	0.020	−0.082	0.059	0.034	0.161
hvqA	0.021	0.020	0.028	0.058	−0.048	0.158
no edu.	−0.035*	0.016	0.090†	0.049	0.186	0.125
other edu.	−0.010	0.024	0.128†	0.072	0.290†	0.167
part time	−0.067**	0.016	0.000	0.050	0.032	0.127
unemployed	−0.012	0.023	0.186*	0.084	0.510**	0.158
sick	−0.068**	0.024	0.096	0.073	0.685**	0.141
retired	−0.106**	0.016	0.065	0.046	0.003	0.103
housekeeper	−0.055**	0.018	0.036	0.061	0.281†	0.150
rural	0.012	0.012	−0.067†	0.037	−0.009	0.082
suburb	0.019†	0.010	−0.055†	0.030	0.008	0.065
household size	0.002	0.005	−0.019	0.015	−0.041	0.033
male	−0.181**	0.011	−0.132**	0.034	0.384**	0.080
ln(age)	0.794**	0.049	0.497**	0.146	0.387	0.456
widow	0.010	0.016	0.095*	0.045		
sepdiv	0.014	0.020	0.325**	0.076		
single	0.044*	0.019	0.145*	0.060		
cohsmo	0.328**	0.015	−0.222**	0.048		
mother smoked	−0.027	0.024				
father smoked	−0.040**	0.015				
both smoked	−0.062**	0.016				
others smoked			0.390**	0.032		
cons	−0.250	0.199	3.141**	0.571	−12.234**	1.379
γ	0.135**	0.002				
p			2.036**	0.049		
μ					0.093**	0.009
ϱ_1	−0.014	0.017				
ϱ_2	−0.009	0.049				
ϱ_3	0.355*	0.150				
logL:	-20706.622					
N:	2901					

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

Table 8
Results from latent factor model - DFM estimator (k=3)

	Smoking initiation (AFT metric)		*Smoking cessation* (AFT metric)		*Lifespan* (PH metric)	
Variable	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
ln(starting)			−0.507**	0.057	−0.034	0.163
quit					−0.349**	0.088
ln(sm_years)					0.276**	0.073
sc1	0.007	0.011	−0.083*	0.033	−0.113	0.085
sc3	−0.023*	0.011	0.027	0.034	0.120†	0.071
degree	0.019	0.020	−0.080	0.059	0.063	0.167
hvqA	0.021	0.020	0.028	0.058	−0.056	0.166
no edu.	−0.035*	0.016	0.092†	0.049	0.181	0.131
other edu.	−0.010	0.024	0.131†	0.072	0.276	0.177
part time	−0.067**	0.016	−0.001	0.050	0.012	0.133
unemployed	−0.013	0.023	0.187*	0.084	0.543**	0.167
sick	−0.070**	0.024	0.090	0.073	0.740**	0.153
retired	−0.106**	0.016	0.062	0.046	−0.002	0.109
housekeeper	−0.055**	0.018	0.035	0.061	0.292†	0.159
rural	0.012	0.012	−0.067†	0.037	−0.002	0.086
suburb	0.019†	0.010	−0.054†	0.030	0.015	0.067
household size	0.002	0.005	−0.019	0.015	−0.045	0.035
male	−0.181**	0.011	−0.135**	0.034	0.419**	0.086
ln(age)	0.792**	0.049	0.513**	0.146	0.019	0.540
widow	0.010	0.016	0.095*	0.045		
sepdiv	0.013	0.020	0.326**	0.076		
single	0.044*	0.019	0.149*	0.060		
cohsmo	0.328**	0.015	−0.216**	0.048		
mother smoked	−0.025	0.024				
father smoked	−0.039**	0.015				
both smoked	−0.061**	0.016				
others smoked			0.390**	0.032		
cons	−0.242	0.199	3.147**	0.572	−12.134**	1.455
γ	0.135**	0.002				
p			2.040**	0.049		
μ					0.106**	0.012
ρ_1	−0.062	0.045				
ρ_2	−0.128	0.114				
ρ_3	1.836**	0.406				
η_2	−2.464	1.547				
ζ_1	2.679**	0.940				
ζ_2	−0.907	0.979				
π_1	0.912**	0.095				
π_2	0.025*	0.012				
$m_{(l)}$	−0.021					
$V_{(l)}$	0.151					
logL:	-20701.663					
N:	2901					

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

suggesting that the model is no longer identified.[18] Estimates from the DFM are consistent with those from the GHQ and MSL.

The factor loadings can be interpreted as coefficients of the random effect l in the three equations. We notice that ϱ_3 has a statistically significant positive impact on lifespan, meaning that there exist unobservable factors that increase the hazard of dying. Factor loadings ϱ_1 and ϱ_2 are not statistically significant but we notice that they both have a negative sign: unobservable factors accelerate both the hazard of starting and the hazard of quitting smoking. We calculate the elements of the covariance matrix of equation (14) substituting into it the estimated factor loadings. This is shown in Table 9.

The covariance between the smoking duration equations errors, $cov(\nu_{1i}, \nu_{2i})$, is $\varrho_1\varrho_2 V_{(l_i)}$ in equation (14) and it is close to zero. Nonetheless the positive sign is in line with the hypothesis of unobservable heterogeneity as an individual preference for experimentation: those people who start smoking soonest quit early. Once this effect is taken into account what is measured by the coefficient of $ln(starting)$ is just the causal relationship between age at onset of smoking and the hazard of quitting. $\varrho_1\varrho_3 V_{(l_i)}$ measures the covariance between the error terms of the smoking initiation and lifespan equation, $cov(\nu_{1i}, \nu_{3i})$, and is negative and close to zero. Unobservable heterogeneity here seems to represent frailer individuals' selection into smoking: given the lower opportunity cost of smoking for people who are likely to die earlier, they also start smoking sooner. Finally, covariance between errors in the smoking cessation and lifespan equations, $cov(\nu_{2i}, \nu_{3i})$, is given by $\varrho_2\varrho_3 V_{(l_i)}$ and is close to zero

[18] This problem can be related to the nature of the date we use: the probability weights are very large on one mass point suggesting that increasing the number of mass points will not improve identification of the discrete factor distribution. This is in line with the findings of Monte Carlo studies which show that two to four points of support adequately model many distributions Mroz (1999).

We also find that the model is not always identified for every set of starting values for the additional parameters. Local maxima can be found maximising the likelihood and estimates might be sensible to changes in the starting values. From a practical point of you, this feature of the DFM may encourage practitioners to use more stable estimation methods such as the parametric approaches proposed here.

Table 9
Estimated covariances

	GHQ	MSL	DFM (k=3)
$\text{cov}(\nu_{1i}, \nu_{2i})$	0.0002	0.0001	0.0012
$\text{cov}(\nu_{1i}, \nu_{3i})$	-0.0024	-0.0049	-0.0172
$\text{cov}(\nu_{2i}, \nu_{3i})$	-0.0014	-0.0030	-0.0357

and negative.[19] In this case, the random effect would represent frailer individuals' selection out of smoking. For any age at starting, there exist people who quit sooner and die earlier.

Robustness of the estimation results to different exclusion restrictions has been checked. We also estimated a version of our trivariate model where the hazard functions depend on the same observed individual characteristics. This sensitivity analysis shows that identification of the causal effects of age at starting on quitting smoking, and of both smoking durations on the hazard of mortality is not affected by the set of instruments chosen. The duration dependence parameters as well as the additional parameters of the mixture model which indicate unobservable heterogeneity do not vary significantly. Tables with summary results are reported in the Appendix.

Overall, the latent factor model coefficients, estimated using both the semi-parametric and the parametric techniques, do not differ much from those obtained from the univariate hazard regression models. Together with the low covariances reported in table 9, this can be interpreted as low unobservable heterogeneity in the data. We are interested in the determinants of smoking initiation and cessation, and their effect on individual lifespan. The coefficients of the education and social class variables indicate that individuals in the lowest socio-economic groups tend to be younger when they start smoking than people from the middle and upper groups. Time to starting also depends on occupational status and accelerates for individuals who are not in full-time jobs: in particular, coefficients of variables indicating being in a part-time job, absent from work due to sickness, retired or housekeeper are

[19] We notice that the DFM produces bigger numbers for $cov(\nu_{1i}, \nu_{3i})$ and $cov(\nu_{2i}, \nu_{3i})$.

highly statistically significant. However, the impact of having a father (or both parents) who used to smoke is about twice the impact of being in the lower social class (or having no education). We notice that the variable *cohsmo* has the highest impact: for people who started smoking after 1954, time to starting decelerates by some 38% more than for those who started before that year. This is an indication of the strong positive effect of the dissemination of information about the health risk of smoking among the British population and the associated cultural change over time.

Social class and education are important determinants of the hazard of quitting but the variable *cohsmo* as well as marital status variables seem to explain more of the variability. The impact of being separated or divorced is twice that of singletons, and 3 times that of widows. Time to quit accelerates if smokers started after 1954 and the impact of *cohsmo* is about 2.7 times that of being in the top social class or having a degree. The coefficient of *ln(starting)* represents the elasticity of time to quit with respect to changes in age at starting and this is very close to the elasticity with respect to age. The estimated elasticity shows that the causal effect of age at starting is negative, meaning that one additional year in age at starting gives an acceleration to time to quit and that starting at young ages determines an increase in the number of years spent smoking. The positive elasticity with respect to age indicates that time to quit decelerates as smokers get older: this suggest that older smokers may believe that they would not gain much from quitting when the end of their lifespan is close.

To some extent a socio-economic gradient on the hazard of dying can be found as well, although only few variables of interest are statistically significant. Being unemployed or absent from work due to sickness matter more than other socio-economic characteristics and they do increase the hazard of dying. For the unemployed the hazard is 66 per cent higher than the hazard of employed individuals and is 4 times the hazard of a partly skilled or unskilled worker (*sc3*). The hazard of dying is

relatively more elastic to changes in the number of years spent smoking than in age at starting: the former elasticity is 8 times the elasticity with respect to age at starting. Also, the hazard of mortality for ex-smokers is 70 per cent of the hazard of current smokers.

The effect of the quitting variables (being an ex-smoker and number of years smoked) on the hazard of dying is predicted to be higher than in the univariate model for lifespan, in particular if we look at the results from the DFM. The latent factor model predicts a smaller and not statistically significant effect of age at starting on the mortality hazard rate, and this is estimated to be even smaller by the DFM. Hence, in the presence of unobservables, the elasticity of mortality hazard with respect to age at starting is overestimated if the econometric model does not include a latent heterogeneity term.

Furthermore, it is interesting to notice that the impact of unemployment (and sickness abscence) on the hazard of dying is about 1.6 (2.1) times of being a former smoker. If we look at social class and education, however, the impact of quitting is about 3 times that of being in the top social class and 5.5 times that of having a degree.

4. Conclusions

This paper investigates the relationship between smoking behaviours and individual lifespan using the British HALS. The nature of the data collection in this survey gave us the scope to exploit longitudinal information about smoking habits, as time to starting and quitting using tobacco, and length of life. Duration models are used to explore determinants of the hazard of starting, quitting and dying. Our main interest lies in the causal effect of age at starting on smoking cessation and mortality risk, and of smoking duration on mortality risk. The econometric issue concerns unobservable heterogeneity and selection bias affecting the estimates of the

causal effects as well as the duration dependence parameters.

A recursive system of three hazard regressions is estimated, which allows for a mixture of hazards functions depending on observed individual characteristics and a common latent factor representing individual time-invariant heterogeneity. Gauss-Hermite quadrature and Simulated Maximum Likelihood are used to approximate the log-likelihood of our model when parametric assumptions on the common latent factor are specified. Estimates from these two approaches are compared to a Discrete Factor Model which has the advantage of not relying upon any parametric assumption on the latent factor. Results are robust to the choice of a continuous or discrete mixture model and to changes in the exclusion restrictions.

Analysis of the covariance structure of the errors does not give very strong evidence of selection, but confirms our initial hypotheses about the role of unobservable heterogeneity in the relation between smoking and mortality. Our results suggest that unobservable heterogeneity affecting the correlation between age at starting and duration of smoking depends on individual-specific preference for experimentation: there would be smokers who started soonest and quit earliest since they are not meant to be hard-core smokers, but rather experimenters. We also find that unobserved frailty partly drives correlation between smoking and mortality in a manner that differs depending on which smoking duration variable is considered. Specifically, unobserved frailty drives selection into early smoking initiation as well as selection out of smoking or, in other words, into early smoking cessation. The recent literature gives mixed evidence about unobserved frailty: our model distinguishes between the two selection effects and suggests that they can co-exist as they act differently on smoking initiation and cessation.

Overall, investigation of the determinants of mortality hazard and smoking initiation and cessation does not seem to be substantially affected by unobservable heterogeneity and the estimated covariances show that there is not a statistically significant association between errors due to the presence of a latent common fac-

tor. However, covariances have economic significance and using a latent factor model seems to be a better choice to correct for the upward bias in the estimate of the causal effect of age at starting and the downward bias of the causal effects of quitting variables and the duration dependence parameter in the lifespan equation.

Our analysis provides additional empirical evidence on the linkage between socio-economic characteristics, lifestyles and mortality. From a policy point of view, we suggest that investments in health, in terms of opting for a less heavy smoking behaviour, can be realised by improving socio-economic conditions, diffusing knowledge about the health risk of smoking, and delaying age at onset of smoking. Furthermore, our findings show that a reduction in the hazard of mortality has to do, in turn, with improving socio-economic conditions in the population, as well as with increasing the quitting rate and shortening the the number of years spent smoking, in particular, as stressed above, by delaying time to starting smoking.

Appendix A

Table A.1
Selected coefficients in the lifespan equation from alternative GHQ models

	Model I[a]	Model II	Model III
Variable	Coeff.[b]	Coeff.	Coeff.
ln(starting)	−0.185	−0.200	−0.220
quit	−0.244**	−0.238**	−0.214**
ln(sm_years)	0.255**	0.261**	0.257**
γ	0.135**	0.135**	0.136**
p	2.036**	2.036**	2.036**
μ	0.092**	0.093**	0.092**
ϱ_1	−0.013	−0.012	−0.010
ϱ_2	−0.006	−0.005	0.005
ϱ_3	0.334*	0.341*	0.331*
logL	−20706.918	−20704.343	−20702.566

Notes:
[a] Model I is our preferred specification, Model II includes marital status in the lifespan equation, Model III does not allow for any exclusion restrictions.
[b] Significance levels: $*$: 5% $**$: 1%

Table A.2
Selected coefficients in the lifespan equation from alternative MSL models

	Model I	Model II	Model III
ln(starting)	−0.176	−0.192	−0.213
quit	−0.249**	−0.243**	−0.218**
ln(sm_years)	0.257**	0.262**	0.257**
γ	0.135**	0.135**	0.135**
p	2.036**	2.036**	2.036**
μ	0.093**	0.094**	0.093**
ϱ_1	−0.014	−0.013	−0.012
ϱ_2	−0.009	−0.007	0.003
ϱ_3	0.355*	0.361*	0.350*
logL	−20706.622	−20704.030	−20702.278

Table A.3
Selected coefficients in the lifespan equation from alternative DFM

	Model I	Model II	Model III
ln(starting)	−0.034	−0.070	−0.075
quit	−0.349**	−0.332**	−0.319**
ln(sm_years)	0.276**	0.274**	0.267**
γ	0.135**	0.135**	0.135**
p	2.040**	2.040**	2.039**
μ	0.106**	0.104**	0.105**
ϱ_1	−0.062	−0.057	−0.053
ϱ_2	−0.128	−0.122	−0.107
ϱ_3	1.836**	1.803**	1.884**
logL	−20701.663	−20698.906	−20696.469

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. National Bureau of Standards, Applied Mathematics Series n.55, Washington DC.

Adda, J. and Lechene, V. (2001). 'Smoking and endogenous mortality: Does heterogeneity in life expectancy explain differences in smoking behavior?'. Discussion Paper 77, Department of Economics, University of Oxford.

Adda, J. and Lechene, V. (2004). 'On the identification of the effect of smoking on mortality'. CeMMAP working papers CWP13/04. Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

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.

Cheung, Y. B. (2000). 'Marital status and mortality in british women: a longitudinal study'. *International Journal of Epidemiology*, vol. 29, pp. 93–99.

Cox, B., Blaxter, M., Buckle, A., Fenner, N., Golding, J., Gore, M., Huppert, F., Nickson, J., Roth, M., Stark, J., Wadsworth, M., and Whichelow, M. (1987). *The Health and Lifestyle Survey*. London: Health Promotion Research Trust.

Cox, B., Huppert, F., and Whichelow, M. (1993). *The Health and Lifestyle Survey: seven years on*. Aldershot: Dartmouth.

Deb, P. and Trivedi, P. K. (2006). 'Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care uilization'. *Econometrics Journal*, vol. 9, pp. 307–331.

Doll, R. and Hill, A. B. (1954). 'The mortality of doctors in relation to their smoking habits.'. *British Medical Journal*, vol. ii, pp. 1451–5.

Farrell, P. and Fuchs, R. V. (1982). 'Schooling and health: the cigarette connection'. *Journal of Health Economics*, vol. 1, pp. 217–230.

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.

Goldman, D. (1995). 'Managed care as a public cost-containment mechanism'. *Rand Journal of Economics*, vol. 26(2), pp. 277–295.

Heckman, J. and Singer, B. (1984). 'A method for minimizing the impact of distributional assumptions in econometric models for duration data'. *Econometrica*, vol. 52(2), pp. 271–320.

Heckman, J. J. and Navarro, S. (2007). 'Dynamic discrete choice and dynamic treatment effects'. *Journal of Econometrics,*, vol. 136, pp. 341396.

Kan, H., Goldman, D., Keeler, E., Dhanani, N., and Melnick, G. (2003). 'An analysis of unobserved selection in an inpatient diagnostic cost group model'. *Health Services and Outcomes Research Methodology*, vol. 4, pp. 71–91.

Lahiri, K. and Song, J. G. (2000). 'The effect of smoking on health using a sequential self-selection model'. *Health Economics*, vol. 9, pp. 491–511.

Lillard, L. A. and Panis, C. W. A. (1996). 'Marital status and mortality: The role of health'. *Demography*, vol. 33(3), pp. 313–327.

Mark, S. D. and Robins, J. M. (1993). 'Estimating the causal effect of smoking

cessation in the presence of confounding factors using a rank preserving structural failure time model.'. *Statistics in Medicine*, vol. 12(17), pp. 1605–1628.

Mello, M. M., Stern, S., and Norton, E. (2002). 'Do medicare hmo still reduce health service use after controlling for selection bias?'. *Health Economics*, vol. 11, pp. 323–340.

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.

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

Picone, G., Sloan, F., S.Chou, and Taylor, D. (2003). 'Does higher hospital cost imply higher quality of care?'. *The Review of Economics and Statistics*, vol. 85(1), pp. 51–62.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.

van Ours, J. C. (2003). 'Is cannabis a stepping-stone for cocaine?'. *Journal of Health Economics*, vol. 22(4), pp. 539–554.

van Ours, J. C. (2004). 'A pint a day raises a mans pay; but smoking blows that gain away'. *Journal of Health Economics*, vol. 23(5), pp. 863–886.

van Ours, J. C. (2006). 'Dynamics in the use of drugs'. *Health Economics*, vol. 15(12).

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.

WHO (2005). *The European Health Report 2005 : public health action for healthier children and populations.* The World Health Organization.