Multi-state Markov model of smoking behaviour: estimation of uptake and cessation from multiple surveys [Penalized purged Markov models]

Mark Clements, Theodore Holford

August 15, 2015

Abstract

We describe a model for population-level smoking behaviour based on a multi-state Markov model. We give simple mathematical relationships between the transition intensities of uptake, cessation and mortality with (i) transition probabilities for the underlying population, using the Kolmogorov differential equations, and (ii) transition intensities and transition probabilities for cross-sectional survey data, using the theory of purged Markov chains (Hoem, 1969). A preliminary implementation of the model is provided using maximum penalised likelihood estimation, with the transition probabilities calculated using an ordinary differential equation solver in C code. The advantages of this approach include the provision of a simple, integrated formulation of smoking behaviour, combined with an elegant statistical formulation of current status and retrospective data from multiple cross-sectional surveys. The model allows for estimation of ever smoker reclassification from multiple surveys, and can readily be used to calculate smoking projections under different scenarios for future smoking uptake and cessation.

1 Introduction

Motivation: fit a model for US smoking patterns based on survey data.

Issues: differential survival due to smoking status (Harris, 1983); mis-classification of former smokers as never smokers (van de Mheen and Gunning-Schepers, 1994).

Related literature: Titman (2011) suggested using B-splines to represent log-hazards for modeling continuous time Markov models for panel data. Titman further suggested calculating transition probability matrix from time s to time t using probability matrices from time t_0 and using matrix inversion, such that $\boldsymbol{P}(s,t) = \boldsymbol{P}^{-1}(t_0,s)\boldsymbol{P}(t_0,t)$.

Hoem (1969) introduced purged Markov processes, where some states are not observable. The use of purged Markov processes for modeling retrospective data has received relatively little attention. Notable exceptions include: Andersen and Green (1985: TODO) on the incidence of diabetes where observations are conditional on not emigrating;

Outline: smoothing using natural splines and P-splines; ordinary differential equations; purged Markov processes.

2 Underlying multi-state models for smoking behaviour

In the following, we consider a birth cohort born in year c with time scale t, which could be either age or calendar period. Consider a system with four live states and a death state, where state 1 is for never smokers, state 2 is for current smokers, state 3 is for former smokers, state 4 represents reclassified smokers (that is, former smokers who report themselves as being never smokers), and state 0 represents death. Let $\alpha_{ij}(t)$ be the transition intensities from state i to state j at time t. As a model simplification, we ignore any dependence of the transitions intensities on duration in state, such that the model has one primary time scale and is a *continuous time multi-state Markov model*.

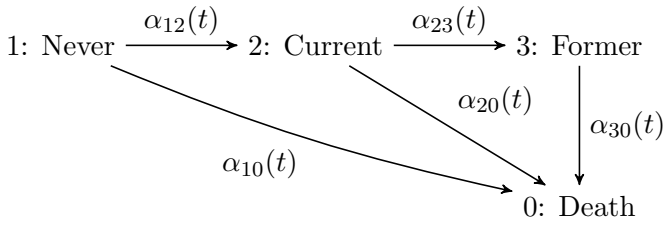

Figure 1: Model A for smoking behaviour

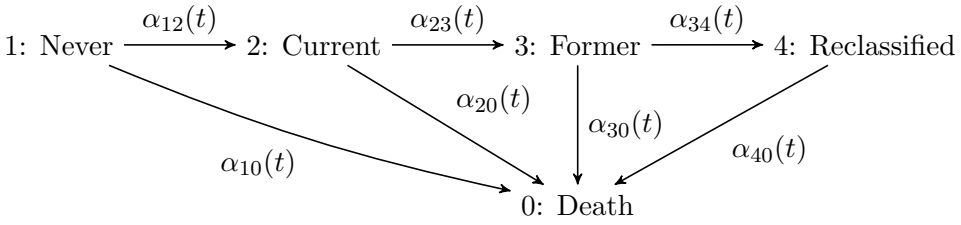

Figure 2: Model B for smoking behaviour

Let the probability of moving from state i at time s to state j at time t be $P_{ij}(s,t)$, and define $\alpha_{ii}(s) = -\sum_{j\neq i}\alpha_{ij}(s)$. Then we have the Kolmogorov forward differential equations:

$$P_{ij}(s,s) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$
$$\frac{\partial P_{ij}(s,t)}{\partial t} = \sum_k P_{ik}(s,t)\alpha_{kj}(t), \qquad s < t$$

Given the transition intensities $\alpha_{ij}(t)$ as continuous functions of time t and initial values $P_{ij}(s,s)$, the transition probabilities can be calculated using ordinary differential equations as an initial value problem. Notably, we propose continuous-time functions for the transition intensities, rather than piecewise-constant functions or non-parametric baseline hazard functions that are typically used for multi-state Markov models.

3 Survey sampling with recall of behaviour

If a survey is undertaken at time τ and an analysis is undertaken retrospectively, then we have a model that is conditional on being observable, with observed transition intensity $\lambda_{ij}(t;\tau)$ from state i to state j at time t, with observed transition probabilities $Q_{ij}(s,t;\tau)$ from state i at time s to state j at time t.

Figure 3: Observed model for smoking behaviour

The framework of purged Markov chains by Hoem (1969) provides theory for when some states are "purged" and not observed. From Hoem and others, it is well known that the observed transition rates are biased by differential survival. Let $\boldsymbol{K}$ represent the states that are observable from the survey, and let $P_{i\boldsymbol{K}}(t,\tau) = \sum_{j\in\boldsymbol{K}} P_{ij}(t,\tau)$ be the probability of

being in state i at time t and being in an observable state at time τ. Then

$$\lambda_{ij}(t;\tau) = \alpha_{ij}(t)\frac{P_{j\boldsymbol{K}}(t,\tau)}{P_{i\boldsymbol{K}}(t,\tau)}$$

and

$$Q_{ij}(s,t;\tau) = P_{ij}(s,t)\frac{P_{j\boldsymbol{K}}(t,\tau)}{P_{i\boldsymbol{K}}(s,\tau)}$$

This result generalises the approach used by Harris (1983) to adjust for the effect of differential survival when reconstructing smoking prevalence.

The reclassification of former smokers as never smokers is awkward, as the observed and modelled states are no longer one-to-one: observed never smokers comprise never and reclassified smokers in the underlying model. This requires that we differentiate between the observed states and the states for the underlying model; for observed never smokers, the probability is the sum of probabilities for never smokers and reclassified smokers (that is, $Q_{11}(s,t;\tau) + Q_{14}(s,t;\tau)$). The uptake of smoking can now be measured in terms of those who have never been smokers, which is the modelled $\alpha_{12}(s)$, but which is not observable, or in terms of those who presently identify as never being smokers, which would be expressed as $\alpha_{12}(s)\frac{P_{11}(0,s)}{P_{11}(0,s)+P_{14}(0,s)}$. At younger ages, $P_{14}(0,s)$ will be small and the two transition rates will be similar.

Given values of the transition intensities $\alpha_{ij}(t)$, we are able to calculate $P_{ij}(s,t)$, $P_{i\boldsymbol{K}}(t,\tau)$, $\lambda_{ij}(t;\tau)$ and $Q_{ij}(s,t;\tau)$ for any values of i, j, s, t and τ. In particular, given the intensities for the underlying model, we are able to calculate the transition intensities and transition probabilities for the observed survey data, which supports estimation of the underlying intensities using the likelihood for the observed data.

3.1 Some theoretical considerations

The proposed Model A is progressive and is a Coxian phase-type distribution with nonhomogeneous (time-dependent) transition intensities.

Theorem 1. *For a purged Markov model, consider an observation that starts in state u_1 at time t_1 and is observed in state v_I at time τ with intervals $[t_i, t_{i+1})$ being in state u_i at time t_i and state v_i at time $(t_{i+1}-)$ with an observed transition from state v_i to state u_{i+1} at time t_{i+1}. Then the density of the observation is*

$$\prod_{i=1}^{I-1}\left\{Q_{u_i,v_i}(t_i,t_{i+1};\tau)\lambda_{v_i,u_{i+1}}(t_{i+1};\tau)\right\}Q_{u_I,v_I}(t_I,\tau;\tau) = \frac{\prod_{i=1}^{I-1}\left\{P_{u_i,v_i}(t_i,t_{i+1})\alpha_{v_i,u_{i+1}}(t_{i+1})\right\}P_{u_I,v_I}(t_I,\tau)}{P_{u_1,K}(t_1,\tau)}$$

Proof.

$$\begin{aligned}
&\prod_{i=1}^{I-1}\left\{Q_{u_i,v_i}(t_i,t_{i+1};\tau)\lambda_{v_i,u_{i+1}}(t_{i+1};\tau)\right\}Q_{u_I,v_I}(t_I,\tau;\tau)\\
=&\prod_{i=1}^{I-1}\left\{P_{u_i,v_i}(t_i,t_{i+1})\frac{P_{v_i,K}(t_{i+1},\tau)}{P_{u_i,K}(t_i,\tau)}\alpha_{v_i,u_{i+1}}(t_{i+1})\frac{P_{u_{i+1},K}(t_{i+1},\tau)}{P_{v_i,K}(t_{i+1},\tau)}\right\}\frac{P_{u_I,v_I}(t_I,\tau)}{P_{u_I,K}(t_I,\tau)}\\
=&\prod_{i=1}^{I-1}\left\{P_{u_i,v_i}(t_i,t_{i+1})\alpha_{v_i,u_{i+1}}(t_{i+1})\right\}\frac{P_{u_I,v_I}(t_I,\tau)}{P_{u_I,K}(t_I,\tau)}\prod_{i=1}^{I-1}\frac{P_{u_{i+1},K}(t_{i+1},\tau)}{P_{u_i,K}(t_i,\tau)}\\
=&\prod_{i=1}^{I-1}\left\{P_{u_i,v_i}(t_i,t_{i+1})\alpha_{v_i,u_{i+1}}(t_{i+1})\right\}\frac{P_{u_I,v_I}(t_I,\tau)}{P_{u_1,K}(t_1,\tau)}
\end{aligned}$$

□

4 Data sources

Individual-level data were available from respondents to smoking questions from the National Health Interview Surveys from 1965 to 2010. In total, there were 1,000,387 respondents.

The main data requirements are accurate estimates of the mortality rate functions by age, sex, calendar period and smoking status[1] . The current smoking history generator only includes differential mortality by ever versus never smokers, while the proposed model would need these rates for never, current and former smokers.

The current smoking history generator is also based on unweighted survey data. Survey weights and any other survey data, such as primary sampling unit, would improve the validity of the estimates.

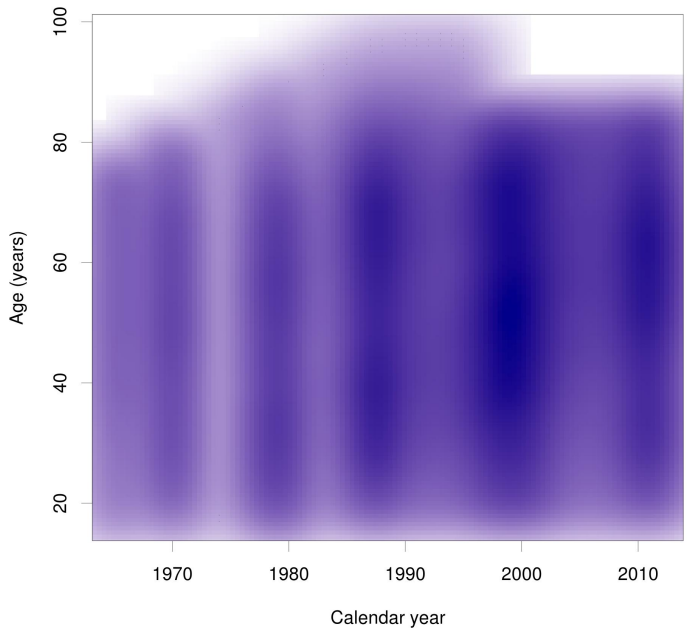

Figure 4: Density plot for respondents to the NHIS surveys, by age and calendar period, 1965–2011.

5 Likelihood

The likelihood components for observed data are available for six different forms of data: current status for (i) never, (ii) current and (iii) former smokers, and recalled behaviour for (iv) current smokers, (v) former smokers with time of uptake and cessation, and (vi) former smokers with time of cessation only.

The likelihood component L for an observed never smoker is, given the underlying system, the probability for a never smoker and the probability for a reclassified smoker, such that

$$\begin{aligned} L_1(\tau) &= Q_{11}(0,\tau;\tau) + Q_{14}(0,\tau;\tau) \\ &= P_{11}(0,\tau)\frac{P_{1\boldsymbol{K}}(\tau,\tau)}{P_{1\boldsymbol{K}}(0,\tau)} + P_{14}(0,\tau)\frac{P_{4\boldsymbol{K}}(\tau,\tau)}{P_{1\boldsymbol{K}}(0,\tau)} = \frac{P_{11}(0,\tau)+P_{14}(0,\tau)}{P_{1\boldsymbol{K}}(0,\tau)} \end{aligned}$$

In this case, the likelihood based on current status and recalled information are the same. For current smokers who have recalled information on age s that they started smoking,

$$\begin{aligned} L_{12}(s,\tau) &= Q_{11}(0,s;\tau)\lambda_{12}(s;\tau)Q_{22}(s,\tau;\tau) \\ &= P_{11}(0,s)\frac{P_{1\boldsymbol{K}}(s,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}\alpha_{12}(s)\frac{P_{2\boldsymbol{K}}(s,\tau)}{P_{1\boldsymbol{K}}(s,\tau)}P_{22}(s,\tau)\frac{P_{2\boldsymbol{K}}(\tau,\tau)}{P_{2\boldsymbol{K}}(s,\tau)} = \frac{P_{11}(0,s)\alpha_{12}(s)P_{22}(s,\tau)}{P_{1\boldsymbol{K}}(0,\tau)} \end{aligned}$$

[1] As a possible model extension, the smoking model could be expanded to include states by smoking dose.

For current smokers who only have information on their current status,

$$L_2(\tau) = Q_{12}(0,\tau;\tau) = P_{12}(0,\tau)\frac{P_{2\boldsymbol{K}}(\tau,\tau)}{P_{1\boldsymbol{K}}(0,\tau)} = \frac{P_{12}(0,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}$$

For former smokers who have recalled information on age s that they started smoking and age t that they quit smoking,

$$\begin{aligned}
L_{123}(s,t,\tau) &= Q_{11}(0,s;\tau)\lambda_{12}(s;\tau)Q_{22}(s,t;\tau)\lambda_{23}(t;\tau)Q_{33}(t,\tau;\tau) \\
&= P_{11}(0,s)\frac{P_{1\boldsymbol{K}}(s,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}\alpha_{12}(s)\frac{P_{2\boldsymbol{K}}(s,\tau)}{P_{1\boldsymbol{K}}(s,\tau)}P_{22}(s,t)\frac{P_{2\boldsymbol{K}}(t,\tau)}{P_{2\boldsymbol{K}}(s,\tau)}\alpha_{23}(t)\frac{P_{3\boldsymbol{K}}(t,\tau)}{P_{2\boldsymbol{K}}(t,\tau)}P_{33}(t,\tau)\frac{P_{3\boldsymbol{K}}(\tau,\tau)}{P_{3\boldsymbol{K}}(t,\tau)} \\
&= \frac{P_{11}(0,s)\alpha_{12}(s)P_{22}(s,t)\alpha_{23}(t)P_{33}(t,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}
\end{aligned}$$

For former smokers who only have information on their current status,

$$L_3(\tau) = Q_{13}(0,\tau;\tau) = P_{13}(0,\tau)\frac{P_{3\boldsymbol{K}}(\tau,\tau)}{P_{1\boldsymbol{K}}(0,\tau)} = \frac{P_{13}(0,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}$$

For the final likelihood component, some respondents who were former smokers only had information on the age that they quit smoking, thus

$$\begin{aligned}
L_{23}(t,\tau) &= Q_{12}(0,t;\tau)\lambda_{23}(t;\tau)Q_{33}(t,\tau;\tau) \\
&= P_{12}(0,t)\frac{P_{2\boldsymbol{K}}(t,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}\alpha_{23}(t)\frac{P_{3\boldsymbol{K}}(t,\tau)}{P_{2\boldsymbol{K}}(t,\tau)}P_{33}(t,\tau)\frac{P_{3\boldsymbol{K}}(\tau,\tau)}{P_{3\boldsymbol{K}}(t,\tau)} \\
&= \frac{P_{12}(0,t)\alpha_{23}(t)P_{33}(t,\tau)}{P_{1\boldsymbol{K}}(0,\tau)}
\end{aligned}$$

The likelihood components can be recognised as the standard multi-state model likelihoods conditional on being observable. Let the data be stratified into disjoint sets Ω_k, where k represents the available data. Then the total likelihood is equal to

$$L = \prod_{i\in\Omega_1} L_1(\tau_i) \prod_{i\in\Omega_2} L_2(\tau_i) \prod_{i\in\Omega_3} L_3(\tau_i) \prod_{i\in\Omega_{12}} L_{12}(s_i,\tau_i) \prod_{i\in\Omega_{123}} L_{123}(s_i,t_i,\tau_i) \prod_{i\in\Omega_{23}} L_{23}(t_i,\tau_i)$$

6 Estimation

The mortality intensity functions $\alpha_{10}(t)$, $\alpha_{20}(t)$ and $\alpha_{30}(t)$ were available. We also assumed that the mortality rates for reclassified smokers were the same as for never smokers (that is, $\alpha_{40}(t) = \alpha_{10}(t)$). We estimated parameters for $\alpha_{12}(t)$ and $\alpha_{23}(t)$ would be estimated. Moreover, for multiple surveys, parameters for $\alpha_{34}(t)$ was estimated.

The likelihood was optimised using the Nelder-Mead algorithm, with variance estimates calculated using the inverse of the Hessian matrix.

7 Implementation

As the transition probabilities need to be recalculated for each update of parameters, we chose to implement the model using a compiled language. Specifically, we implemented the model in C, using an ordinary differential equation solver from the GNU Scientific Library (GSL; see Appendix). As the values of t are not fixed in advanced, the basis functions for $\alpha_{ij}(t)$ need to be calculated in C code. We have implemented the functional forms for the initiation and cessation rates using penalised B-splines with equidistant knots ("p-splines"). These spline functions were implemented in C using the GSL. The C code was linked to R for optimising the likelihood and for post-estimation. The implementation does not currently

offer automatic selection of the smoothing parameters, where selection would be based on an approximate cross-validation criterion (e.g. Joly et al., 2002; Cai and Betensky, 2003).

The model development has focused on cohort-specific transition rates. In practice, data from multiple birth cohorts would be modelled together across the Lexis diagram. A simple solution would be to model the rates using age-period models. A better solution would be to model the rates across the Lexis diagram using tensor product splines. The specification of the transition rates across the Lexis diagram will also allow for the investigation of future smoking prevalence under different scenarios for future smoking uptake and cessation rates.

8 Results

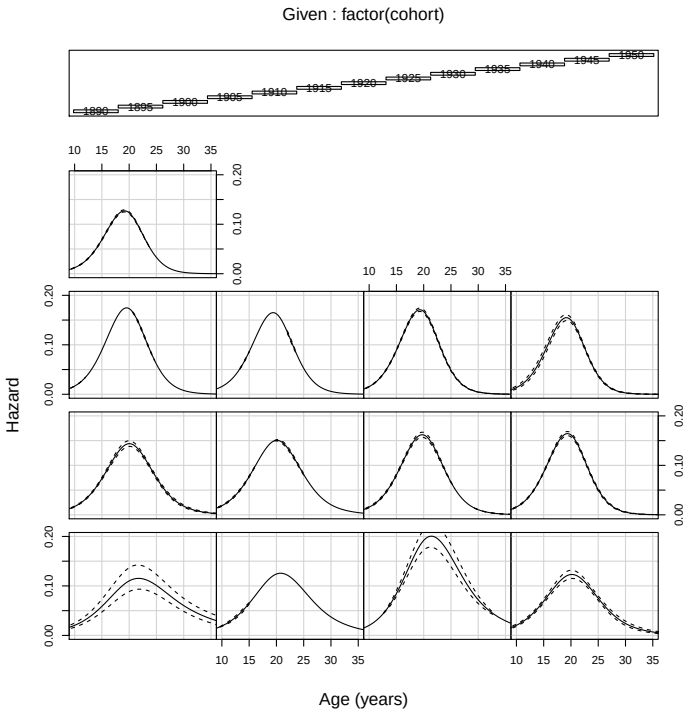

Figure 5: Smoking initiation rates, NHIS, males

References

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, 59(3):570–579.

Harris, J. E. (1983). Cigarette smoking among successive birth cohorts of men and women in the United States during 1900-80. *J Natl Cancer Inst*, 71:473–479. 3.

Hoem, J. M. (1969). Purged and partial Markov chains. *Scandinavian Actuarial Journal*, 1969(3-4):147–155.

Joly, P., Commenges, D., Helmer, C., and Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics (Oxford, England)*, 3(3):433–443.

van de Mheen, P. J. and Gunning-Schepers, L. J. (1994). Reported prevalences of former smokers in survey data: the importance of differential mortality and misclassification. *Am J Epidemiol*, 140:52–57. 1.

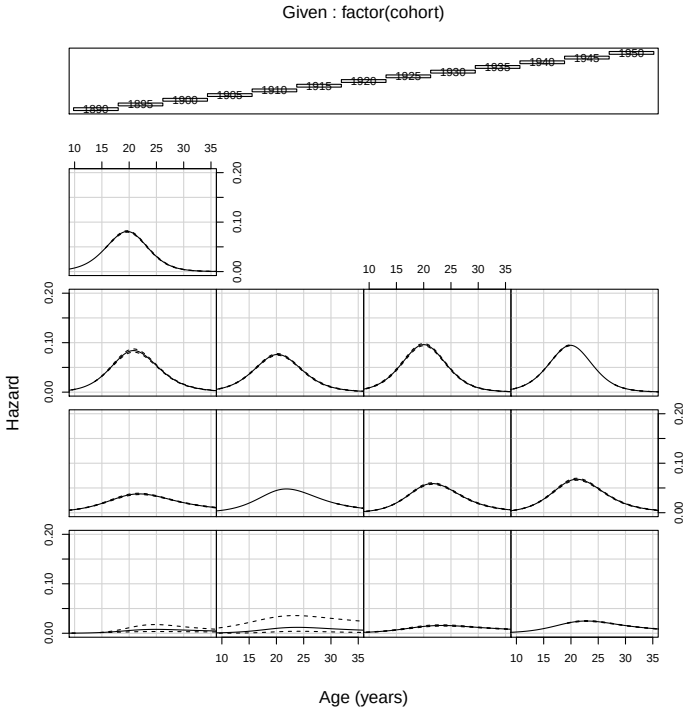

Figure 6: Smoking initiation rates, NHIS, females

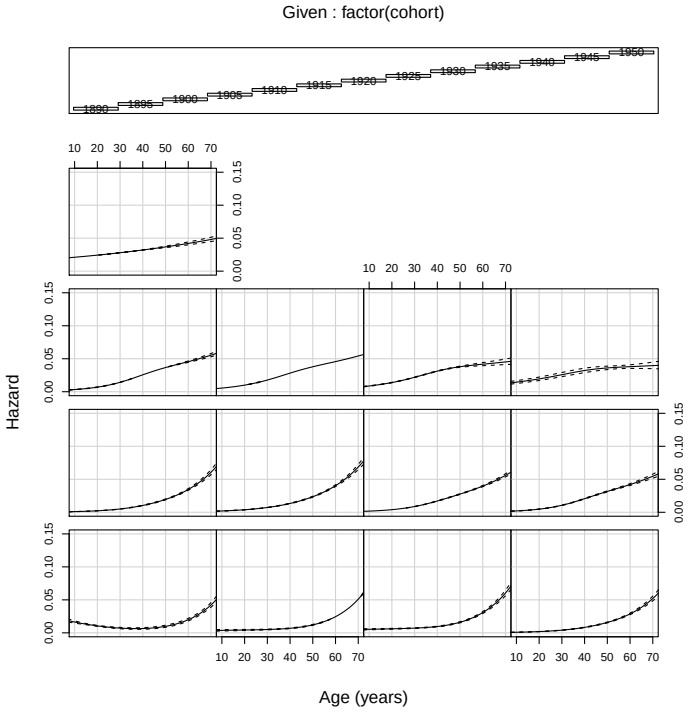

Figure 7: Smoking cessation rates, NHIS, males

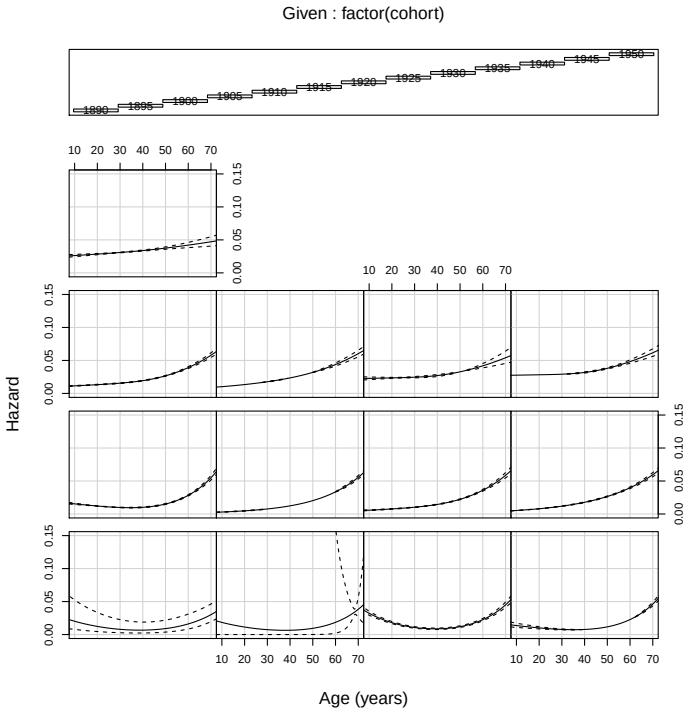

Figure 8: Smoking cessation rates, NHIS, females