Threshold regression models adapted to case-control studies. Application to lung cancer induced by occupational exposure to asbestos

Antoine Chambaz, Dominique Choudat, Catherine Huber, Jean-Claude Pairon, Mark van Der Laan

Threshold regression models adapted to case-control studies. Application to lung cancer induced by occupational exposure to asbestos

A. Chambaz* D. Choudat† C. Huber J-C. Pairon
M. J. van der Laan‡

A. Chambaz, C. Huber, MAP5, Université Paris Descartes and CNRS
D. Choudat, Assistance Publique – Hôpitaux de Paris and Université Paris Descartes
J-C. Pairon, INSERM U955 and Université Paris-Est Créteil
M. J. van der Laan, University of California, Berkeley

March 30, 2012

Abstract

We rely on a recent French matched case-control study to investigate the effect of occupational exposure to asbestos on the occurrence of lung cancer. We build a large collection of threshold regression models, data-adaptively select a better model in it by multi-fold likelihood-based cross-validation, then fit the resulting better model by maximum likelihood. A necessary preliminary step to eliminate the bias due to the case-control sampling is made possible because the conditional distribution of being a case given the matching variable and the marginal distribution of the matching variable can be computed beforehand based on two studies independent from our dataset. The implications of the fitted model in terms of expected years of life free of lung cancer lost due to the occupational exposure to asbestos are discussed.

1 Introduction

Asbestos is a powerful carcinogen [IARC, 1977]. We rely on a recent French matched case-control study on lung cancer by Pairon et al. [2009] to investigate the effect of occupational exposure to asbestos on the occurrence of lung cancer. Following a case-control study design is convenient for a rare disease like lung cancer, with a known prevalence proportion approximately equal to five cases out of 10,000 persons [Belot et al., 2008].

Logistic regression in parametric statistical models is the prevalent method of analysis in case-control study [Breslow, 1996, and references therein]. Sometimes however, as here due for instance to the nature of our exposure variable and because we do not aim for results in terms of risks, this approach would not be satisfactory. Gustavsson et al. [2002] recently undertook a case-control study on low-dose exposure to asbestos and lung cancer in Sweden with results expressed

*This collaboration took place while A. Chambaz was a visiting scholar at UC Berkeley, supported in part by a Fulbright Research Grant and the CNRS. A. Chambaz would like to thank M-L. Ting Lee for interesting discussions on threshold regression models.

†D. Choudat and J-C. Pairon were partially supported by a grant from the Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ES 2005-006).

‡The authors would like to thank two anonymous reviewers for their excellent suggestions.

in terms of risks. By following van der Laan [2008], we show how to examine the connection between occupational exposure to asbestos and lung cancer, accounting notably for tobacco use, by performing estimation based on *any parametric model* for the conditional distribution of time to incident lung cancer given the remaining information. It is possible to take up that challenge because we can estimate the conditional distribution of being a case given the matching variable and the marginal distribution of the matching variable (based on the independent study undertaken by Belot et al. [2008] and on a dataset made publicly available by the French National Institute of Statistics and Economic studies, see `http://www.insee.fr/en/`).

We model the time to incidence of lung cancer as the time until an unobservable process crosses a threshold[Lee and Whitmore, 2006]. The effect of occupational exposure to asbestos is included by accelerating the time index of the process. The effect of other covariates is not included as a time acceleration factor, but directly on the other process parameters. By separately modeling the effect of covariates and exposure in this way, we can get new insights on the effect of these covariates and also, more importantly, on the effect of exposure alone. We express the latter in terms of expected years of life lost as introduced by Robins and Greenland [1991]. Our model belongs to the family of threshold regression models, which has been playing an important role in survival analysis for some years [Lee and Whitmore, 2006, 2010, and references therein]. It is also an accelerated failure time model, with strong ties to the model developed by Oakes [1995] to study the effect of exposure to asbestos on the time until death from (not incidence of) lung cancer, as discussed later. We note that the study of a case-control dataset based on a threshold regression model is undertaken in [Lee et al., 2009], but we believe that the authors do not properly address the difficulties which stem from the type of sampling used to collect the data.

As our study involves an original qualitative description of the exposure to asbestos into 28 categories, we actually consider our model as a maximal model containing thousands of smaller models (obtained for instance by reducing the number of categories). We manage to data-adaptively select a better model in our large collection of threshold regression models by relying on multi-fold likelihood-based cross-validation [van der Vaart et al., 2006]. Then we fit the latter better model to the data by maximum likelihood, and draw our conclusions from its description.

The article is organized as follows. The dataset and the original qualitative description of the exposure to asbestos into 28 categories are described in Section 2. The case-control estimation problem is formalized in Section 3. We develop the threshold regression modeling in Section 4, showing how to derive the expected years of life free of lung cancer lost due to occupational exposure to asbestos for a case. We summarize in Section 5 the main results of the application to the dataset. A brief discussion is finally developed in Section 6. Some relevant material is gathered in the appendix. We present in Section A a few asymptotic results on the case-control weighted maximum likelihood estimator upon which our study relies. Elements of proofs are relegated to Section B. We provide additional details on the real data application in Section C. This includes the computation of the quantities required to eliminate the bias due to the case-control sampling in Section C.1, and brief presentations of *(i)* the principle of our multi-fold likelihood-based cross-validation model selection procedure in Section C.2, and *(ii)* its concrete implementation in Section C.3.

2 Dataset

In this section we describe the dataset of interest. The details of the case-control sampling scheme are given in Section 2.1. For each enrolled subject, non-professional information is collected, see Section 2.2, as well as an original history of occupational exposure to asbestos, see Section 2.3.

2.1 A matched case-control study

The matched case-control study took place between 1999 and 2002 in four Parisian hospitals. Case and control subjects were retrospectively recruited at the end of each year 1999 to 2002 among the patients of these hospitals who were free of lung cancer at the beginning of the corresponding year.

The case subjects were diagnosed with *incident* lung cancer during the period of the study. They were matched by control subjects on the basis of gender, age at end of calendar year (up to ± 2.5 years), hospital, and race. Control subjects were recruited among patients of the departments of ophthalmology, general and orthopedic surgeries, and were by definition free of lung cancer at the time of their enrollment.

The one-to-one matching (*i.e.*, the pattern of who is matched by whom) and race are not available. We propose an artificial valid matching pattern (based on gender, age and hospital) and make sure that our results do not depend on this particular choice. We exclude every subject with missing information. The resulting dataset counts $n = 860$ cases and 901 controls, hence a total of $n + 901 = 1,761$ observations.

We assume that the population sampled from during the study is stationary. Therefore, the observed data structures on experimental units made of pairs of case and matched control can be modeled as independent and identically distributed (iid) random variables. In Section 3, we invoke this fact to derive the likelihood function used in our study.

2.2 Non-professional information

Set a calendar time τ (expressed in years), and consider a subject sampled at time τ.

He/she is associated with his/her hospital of recruitment $W_0 \in \{1, 2, 3, 4\}$, gender W_1 ($W_1 = 0$ for men and $W_1 = 1$ for women), binary indicator W_2 of occurrence of lung cancer in close family ($W_2 = 0$ if no lung cancer occurred and $W_2 = 1$ otherwise), age at incident lung cancer diagnosis T ($T = \infty$ if no lung cancer occurred), and age at interview $X = X(\tau)$.

It is well known that tobacco is a serious risk factor of lung cancer [Biesalski et al., 1998]. Thus information on tobacco consumption is collected during the interview. We summarize this information by considering a discretized version of the lifetime tobacco use. Hence, he/she is also associated with his/her tobacco use W_3 ($W_3 = 0$ for never-smoker, $W_3 = 1$ for lifetime tobacco use comprised between 1 and 25 pack years, $W_3 = 2$ for lifetime tobacco use comprised between 26 and 45 pack years, $W_3 = 3$ otherwise). The boundaries are chosen to yield strata of comparable sizes (371 subjects with $W_3 = 0$, and respectively 468, 469, 453 subjects with $W_3 = 1, 2, 3$). In particular, we overlook the duration of habit and years of abstinence, although they are known to play a role in the development of lung cancer [Ruano-Ravina et al., 2003; Gustavsson et al., 2002].

We denote $W = (W_1, W_2, W_3) \in \mathcal{W} = \{0, 1\}^2 \times \{0, 1, 2, 3\}$ the explanatory covariate, $Z = \min\{T, X\}$ and $Y = \mathbf{1}\{T \leq X\}$. Note that $Y = 1$ if and only if (iff) $T = Z \leq X$ (the subject is then called a case) and $Y = 0$ iff $T > Z = X$ (the subject is then called a control).

2.3 History of occupational exposure to asbestos

In addition, the occupational history up to age X, $\bar{A}(X)$, of the latter subject sampled at τ (we justify the notation below) is determined during the interview by experts of occupational exposure to asbestos. Every employment with duration at least 6 months is associated with its start and end dates, and with an original description of the exposure to asbestos. This description was characterized by the very experts who later conducted the interviews [Pairon et al., 2009]. At the time of the characterization, the French threshold limit value for exposure to asbestos was set to $0.1 f/mL$ (f/mL stands for "asbestos fibers per milliliter").

This description is a triplet referred to as "probability/frequency/intensity", each of them taking values in $\{1, 2, 3\}$: for the considered employment, the probability of exposure, its frequency and intensity are evaluated as either low or mild or high, respectively coded by 1, 2, 3. A probability index equal to 1, 2 or 3 corresponds to a *passive exposure*, a *possible direct exposure* or a *very likely or certain direct exposure*, respectively. A frequency index equal to 1, 2 or 3 corresponds to exposures occurring *less than once a month*, *more than once a month and during less than half of the monthly working hours* or *during more than half of the monthly working hours*, respectively. An intensity index equal to 1, 2 or 3 corresponds to a concentration of asbestos fibers *less than 0.1 f/mL*, *between 0.1 and 1 f/mL* and *more than 1f/mL*, respectively. Thus, the set $\mathcal{E}$ of categories of exposure has 27+1=28 elements (we add a category 0 for no exposure), each of them corresponding

to a particular rate of exposure. Note that we will use from now on the notation $\varepsilon = \varepsilon_1\varepsilon_2\varepsilon_3$ instead of $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$.

The description is similar to that used by Gustavsson et al. [2002]. In the latter article, only the probability and intensity of exposure are considered. Four classes of probability of exposure correspond to an "estimated exposure prevalence" comprised between 0% and 20%, 20% and 50%, 50% and 85%, 85% and 100%. Four classes of intensity of exposure correspond to "time-period-specific annual arithmetic average level of exposure to asbestos" comprised between 0 and 0.03f/mL, 0.03f/mL and 0.1f/mL, 0.1f/mL and 0.3f/mL, 0.3f/mL and above. The Swedish threshold limit value for exposure to asbestos was set to 0.3f/mL in 1993.

We report in the top table of Table 1 the overall number of employments associated to each possible "probability/frequency/intensity" description. Although computed over a total of 8,432 employments, this table contains many zeros, showing that the latter description is over-parametrized. We also report in the bottom table of Table 1 the overall number of employments that feature a particular value of each coordinate of the "probability/frequency/intensity" description.

ε	nb. of emp.	ε	nb. of emp.	ε	nb. of emp.
111	213	211	53	311	138
112	167	212	64	312	105
113	3	213	6	313	24
121	150	221	59	321	136
122	46	222	36	322	189
123	3	223	3	323	22
131	0	231	2	331	1
132	0	232	0	332	3
133	0	233	0	333	0

	1	2	3
probability	582	223	618
frequency	773	644	6
intensity	752	610	61

Table 1: *Top:* Overall number of employments associated to each possible "probability/frequency/intensity" description. The total number of employments is 8,432. Only 1,423 of them feature a description in $\mathcal{E} \setminus \{0\}$. *Bottom:* Overall number of employments that feature a particular value of each coordinate of the "probability/frequency/intensity" description.

The generic longitudinal history of occupational exposure to asbestos is denoted by $\bar{a}$. It belongs to $\bar{\mathcal{E}}$, the set of functions from the nonnegative real line to $\mathcal{E}$ such that $a(t) = 0$ for t small or large enough (before the age at first employment or when no further information is available; this constraint is just a convenience, as we will make clear in Section 4). It is understood that the value of $\bar{a}$ at t is denoted by $a(t)$ while $\bar{a}(t)$ stand for the restrictions of $\bar{a}$ to $[0, t]$. Thus $a(t) = a$ correspond to an occupational position held at age t and characterized by asbestos exposure a.

One of the central issues we deal with in this article is how to associate each description in $\mathcal{E}$ with a rate of exposure. We propose an original solution which heavily exploits the underlying multiplicative nature of the "probability/frequency/intensity" encoding. Indeed, it is the product of "probability", "frequency" and "intensity" which is relevant in terms of rate of exposure.

3 Formulation of the case-control estimation problem

We show in three steps how to examine the connection between occupational exposure to asbestos and lung cancer, accounting notably for tobacco use, by performing (appropriately weighted) maximum likelihood estimation based on *any parametric model* (*i.e.*, in particular, not necessarily

a logistic model) for the conditional distribution of time to incident lung cancer given the remaining information. First, we derive from the description of our case-control study the characterization of the representative sampling scheme one may have relied upon, had the probability of being an incident case of lung cancer not been so small. This mainly amounts to defining an observed data structure $O^\star$ under representative sampling, whose distribution $P_0^\star$ presents features of interest, see Section 3.1. Second, we characterize the observed data structure O under matched case-control sampling and its distribution P_0 in terms of $O^\star$ and $P_0^\star$, see Section 3.2. Third, we show how to make inference on the features of $P_0^\star$ from data sampled under P_0, see Section 3.3. This three-step procedure follows [van der Laan, 2008; Rose and van der Laan, 2008].

3.1 Representative sampling scheme

We explained in Section 2.1 that sampling occurred at times $\tau_0, \tau_1 = \tau_0 + 1, \tau_2 = \tau_0 + 2, \tau_3 = \tau_0 + 3$ (where τ_0 stands for the initial sampling date, January 1st, 2000). Had the representative sampling been carried out, we would have observed n_0 (respectively n_1, n_2, n_3) independent observed data structures $O_i^\star$ sampled at time point τ_0 (respectively τ_1, τ_2, τ_3) under, say, $P^\star(\tau_0)$ (respectively $P^\star(\tau_1), P^\star(\tau_2), P^\star(\tau_3)$). We make the following *stationarity assumption*:

$$\forall 1 \leq k \leq 3, P^\star(\tau_k) = P^\star(\tau_0) \equiv P_0^\star. \tag{1}$$

This assumption is justified by the influx and outflow experienced by the population of the Parisian region over the period of investigation. Thus, had the representative sampling been carried out, one would have finally collected $N = n_0 + n_1 + n_2 + n_3$ independent copies $(O_1^\star, \ldots, O_N^\star)$ of $O^\star \sim P_0^\star$, with

$$O^\star = (W, X, \bar{A}(X), Y, Z). \tag{2}$$

Note that the likelihood of $O^\star$ under $P_0^\star$ writes as

$$\begin{aligned} P_0^\star(O^\star) = P_0^\star(W) P_0^\star(X|W) P_0^\star(\bar{A}(X)|X,W) \\ \times dP_0^\star(Z = T | T \geq X-1, \bar{A}(X), X, W)^Y \\ \times P_0^\star(T > X | T \geq X-1, \bar{A}(X), X, W)^{1-Y}, \end{aligned} \tag{3}$$

where $dP_0^\star(t|T \geq X-1, \bar{A}(X), X, W)$ is the conditional density of T at time t given the event $[T \geq X-1, \bar{A}(X), X, W]$.

3.2 Matched case-control sampling

Such a representative sampling would have been impractical and ineffective because the probability $P_0^\star(Y = 1)$ of being an incident case of lung cancer is very small, see Section C.1. In order to recruit some cases in the sample, one would have to sample a huge number of observations. This is the main motivation for using a case-control sampling scheme.

We describe now what is our observed data structure in this framework. Introduce the categorical matching variable $V \in \mathcal{V}$ obtained by concatenating W_0 (subject's hospital when sampled), W_1 (subject's gender) and a discretized version of the age at sampling X over bins of length five years. In the sequel, we repeatedly use the convenient (though redundant) notation (V, W).

The matched case-control sampling scheme can be described as follows:

- One first samples a case by sampling

  $$(V^1, O^{1\star}) = (V^1, W^1, X^1, \bar{A}^1(X^1), Y^1 = 1, Z^1)$$

  from the conditional distribution of $(V, O^\star)$ given $Y = 1$ (the superscript "1" refers to the fact that $Y^1 = 1$).

- Subsequently, one samples J controls

$$(V^{0,j}, O^{0,j\star}) = (V^{0,j}, W^{0,j}, X^{0,j}, \bar{A}^{0,j}(X^{0,j}), Y^{0,j} = 0, Z^{0,j})$$

from the conditional distribution of $(V, O^\star)$ given $Y = 0, V^{0,j} = V^1$ for all $j \leq J$ (the superscript "0" refers to the fact that $Y^{0,j} = 0$ for all $j \leq J$).

Conditional on $V^1 = v \in \mathcal{V}$, the ratio of the number of controls for one case, $J/1$, is much smaller than the ratio $P_0^\star(Y = 0|V = v)/P_0^\star(Y = 1|V = v)$ one would get in the population. This sampling scheme results in the observed data structure

$$O = ((V^1, O^{1\star}), (V^{0,j}, O^{0,j\star}), j = 1, \ldots, J) \sim P_0$$

whose true distribution P_0 can be deduced from $P_0^\star$ and the two-step description above.

The method naturally allows to consider the case that J is random and thus varies per experimental unit. This permits to exploit all our observations, even though we have less cases than controls. Note that each control is only taken into account once.

3.3 Case-control weighting of the log-likelihood loss function developed for representative sampling

It is remarkable that the log-likelihood loss function developed for the representative sampling scheme can be adapted to the case-control sampling scheme by appropriate weighting. This weighting relies on the *prior knowledge* of the joint distribution of (V, Y), hence of the following probabilities: for each $(y, v) \in \{0, 1\} \times \mathcal{V}$,

$$q_0 = P_0^\star(Y = 1), \tag{4}$$
$$q_0(y|v) = P_0^\star(Y = y|V = v), \tag{5}$$
$$q_0(v|y) = P_0^\star(V = v|Y = y), \tag{6}$$

or, namely, the marginal probability of being a case (4), the conditional probabilities of being a case or a control given matching variable at level v (5), and the conditional probabilities of observing level v for the matching variable given being a case or a control (6). We refer to Section C.1 for the computation of the latter key quantities.

Consider a model $\mathcal{P}^\star$ for $P_0^\star$. We assume without serious loss of generality that there exists a common dominating measure for all $P^\star \in \mathcal{P}^\star$ and $P_0^\star$, so that we can refer to the densities $p^\star$ and $p_0^\star$ of $P^\star \in \mathcal{P}^\star$ and $P_0^\star$. With a slight abuse of notation, $\mathcal{P}^\star$ denotes both the model and the corresponding set of densities.

Define for all $v \in \mathcal{V}$ the quantities

$$\bar{q}_0(v) = q_0 \frac{P_0^\star(Y = 0|V = v)}{P_0^\star(Y = 1|V = v)} = q_0 \frac{q_0(0|v)}{q_0(1|v)}, \tag{7}$$

and introduce the following case-control weighted log-likelihood loss function for the density $p_0^\star$ of $P_0^\star$ *under sampling of* $O \sim P_0$: for all $p^\star \in \mathcal{P}^\star$,

$$\ell(p^\star)(O) = q_0 \log p^\star(V^1, O^{1\star}) + \bar{q}_0(V^1) \frac{1}{J} \sum_{j=1}^{J} \log p^\star(V^1, O^{0,j\star}). \tag{8}$$

Even though q_0 appears in both terms in (8), we prefer to consider $\ell(p^\star)(O)$ as defined above rather than $q_0^{-1}\ell(p^\star)(O)$. This choice guarantees that the weighted log-likelihood $\ell(p^\star)(O)$ is on the same scale as the log-likelihood $\log P_0^\star(O^\star)$ under representative sampling. The weighted log-likelihood loss function is adapted to the case-control sampling scheme in the following sense (the proof is relegated to Section B):

Proposition 1. *Let us assume that model $\mathcal{P}^\star$ for $p_0^\star$ is such that $\int \log p^\star(o^\star) dP_0^\star(o^\star, Y = y)$ are properly defined for all $p^\star \in \mathcal{P}^\star$ and $y = 0, 1$. If $\mathcal{P}^\star$ is well-specified (*i.e.*, if $p_0^\star \in \mathcal{P}^\star$), then the density that maximizes the expectation under P_0 of the weighted loss function (8) over $\mathcal{P}^\star$,*

$$\arg\max_{p^\star \in \mathcal{P}^\star} E_{P_0} \ell(p^\star)(O),$$

is unique and coincides with $p_0^\star$.

In words, this tells us that it is possible to draw inference about $P_0^\star$ based on data sampled from P_0 by (appropriately weighted) maximum likelihood.

In the next section, we propose a parametric model for the conditional distribution of $O^\star$ given $(W, X, \bar{A}, Y)$, that is the conditional distribution of Z given $(W, X, \bar{A}, Y)$. The parametric model is sound in the sense that the conditional distribution of Z given $(W, X, \bar{A}, Y)$ only depends on $\Omega = (W, X, \bar{A}(X), Y)$. The latter parametric model is combined with a nonparametric model for the conditional distribution of $(W, X, \bar{A})$ given Y, both yielding a semiparametric model $\mathcal{P}^\star = \{P_\theta^\star : \theta \in \Theta\}$ for $P_0^\star$ because we know beforehand q_0, the true probability of being a case. Formally, the likelihood of $O^\star = (\Omega, Z)$ under $\theta \in \Theta$ writes as

$$P_\theta^\star(O^\star) = p_\theta^\star(Z|\Omega)\eta(\Omega),$$

$\eta(\Omega)$ being the likelihood of Ω, which we assume without serious loss of generality to be bounded away from 0. If we set

$$\begin{aligned} \Omega^1 &= (W^1, X^1, \bar{A}^1(X^1), Y^1 = 1), \\ \Omega^{0,j} &= (W^{0,j}, X^{0,j}, \bar{A}^{0,j}(X^{0,j}), Y^{0,j} = 0), \end{aligned}$$

hence $O^{1\star} = (\Omega^1, Z^1)$ and $O^{0,j\star} = (\Omega^{0,j}, Z^{0,j\star})$, then the weighted log-likelihood loss function for $p_0^\star$ under sampling of $O \sim P_0$ satisfies, for all $\theta \in \Theta$,

$$\begin{aligned} \ell(p_\theta^\star)(O) &= q_0 \log p_\theta^\star(V^1, O^{1\star}) + \bar{q}_0(V^1) \frac{1}{J} \sum_{j=1}^{J} \log p_\theta^\star(V^1, O^{0,j\star}) \\ &= q_0 \log p_\theta^\star(Z^1|\Omega^1, V^1) + \bar{q}_0(V^1) \frac{1}{J} \sum_{j=1}^{J} \log p_\theta^\star(Z^{0,j}|\Omega^{0,j}, V^1) \\ &\quad + \text{rem}(O), \end{aligned} \tag{9}$$

where $\text{rem}(O)$ is a random term independent of θ. We set $\tilde{\ell}(\theta)(O) = \ell(p_\theta^\star)(O) - \text{rem}(O)$ for all $\theta \in \Theta$.

Assuming that the true density $p_0^\star$ is "projected" (in terms of Kullback-Leibler divergence) onto $p_{\theta_0}^\star$ for some $\theta_0 \in \text{int}(\Theta)$, or in other terms that the mapping $\theta \mapsto \text{KL}(p_0^\star, p_\theta^\star)$ achieves a unique minimum at the unique $\theta_0 \in \text{int}(\Theta)$, we focus hereafter on the maximum likelihood estimation of θ_0. Following the lines of the proof of Proposition 1, the case-control weighted maximum likelihood estimator

$$\theta_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \tilde{\ell}(\theta)(O_i) \tag{10}$$

does estimate θ_0. We briefly consider some asymptotic properties (including consistency and asymptotic normality) in Section A.

4 Threshold regression parametric modeling

We model the time to incidence of lung cancer as the time until an unobservable process crosses a threshold, the effect of asbestos exposure being included by accelerating the time index of the process, see Section 4.1. This is typical of threshold regression modeling [Lee and Whitmore,

2006, 2010, and references therein], with an original acceleration tailored to the description of occupational exposure to asbestos (as presented in Section 2.3) that we introduce in Section 4.2. We complete the description of our model in Section 4.3, where we also specify the form of the resulting case-control weighted log-likelihood loss function. Finally, we show in Section 4.4 how our model yields insights on the association between asbestos exposure *alone* and risk of developing lung cancer.

4.1 Time to incident lung cancer

Let $\mathbb{B}$ be a Brownian motion. For any real numbers $h > 0$ and $\mu \leq 0$, define

$$T\{h, \mu\} = \inf\{t \geq 0 : h + \mu t + \mathbb{B}_t \leq 0\}, \tag{11}$$

the first time the drifted Brownian motion $(h + \mu t + \mathbb{B}_t, t \geq 0)$ crosses the threshold 0. The distribution of $T\{h, \mu\}$ is known as the inverse Gaussian distribution with parameter (h, μ). It is characterized by its cumulative distribution function (cdf): for all $t > 0$,

$$F(h, \mu)(t) = 1 + e^{-2h\mu}\Phi\left((\mu t - h)t^{-1/2}\right) - \Phi\left((\mu t + h)t^{-1/2}\right),$$

where Φ is the standard normal cdf. It is well known [Chhikara and Folks, 1989] that $T\{h, \mu\} < \infty$ almost surely eventually because $\mu \leq 0$. Therefore the distribution of $T\{h, \mu\}$ is also characterized by its density

$$f(h, \mu)(t) = \frac{h}{(2\pi t^3)^{1/2}} \exp\left(-\frac{(h - |\mu|t)^2}{2t}\right)$$

(all $t > 0$). Finally, $T\{h, \mu\}$ has mean $h/|\mu|$ whenever $\mu < 0$.

Here, $(h + \mu t + \mathbb{B}_t, t \geq 0)$ models the amount of health relative to lung cancer *in absence of occupational exposure to asbestos*. Describing what happens *in presence of exposure to asbestos* involves the introduction of an acceleration function R which summarizes the effects of this exposure. An acceleration function R is a nondecreasing continuous function on the nonnegative real line such that $R(t) \geq t$ for all $t \geq 0$. Given such a function R, we define

$$T\{h, \mu, R\} = \inf\{t \geq 0 : h + \mu R(t) + \mathbb{B}_{R(t)} \leq 0\}, \tag{12}$$

the first time the drifted Brownian motion $(h + \mu t + \mathbb{B}_t, t \geq 0)$ crosses the threshold 0 *along the modified time scale* derived from R. Obviously, $T\{h, \mu, R\} = T\{h, \mu\}$ when R is the identity, but in general $T\{h, \mu, R\} \leq T\{h, \mu\}$. Furthermore,

$$T\{h, \mu, R\} \geq t \quad \text{iff} \quad T\{h, \mu\} \geq R(t), \tag{13}$$

so that the cdf of $T\{h, \mu, R\}$ at $t > 0$ is $F(h, \mu)(R(t))$, and its density at $t > 0$ is $R'(t)f(h, \mu)(R(t))$ as soon as R is differentiable. Consequently, the conditional survival function and density of $T\{h, \mu, R\}$ at $t \geq x - 1$ given $[T\{h, \mu, R\} \geq x - 1]$ are respectively

$$G(h, \mu, R)(t) = \frac{1 - F(h, \mu)(R(t))}{1 - F(h, \mu)(R(x - 1))} \tag{14}$$

and

$$g(h, \mu, R)(t) = \frac{R'(t)f(h, \mu)(R(t))}{1 - F(h, \mu)(R(x - 1))}, \tag{15}$$

two important equalities in view of the factorization of the likelihood shown in (3).

Parameters h, μ and $T\{h, \mu\}$ are respectively interpreted as an initial amount of health relative to lung cancer, a rate of decay of this amount of health in absence of occupational exposure to asbestos, and the time to incident lung cancer in absence of occupational exposure to asbestos. As for $T\{h, \mu, R\}$, it is interpreted as the time to incident lung cancer for a history of occupational exposure to asbestos summarized by acceleration function R. Admitting that the reference time scale (that is that of the Brownian motion $\mathbb{B}$) corresponds to *chronological/calendar* time scale, the new time scale formed by the acceleration function R may be understood as a *biological* time scale. This interpretation acknowledges the fact that the ageing phenomenon related to lung cancer is stronger in presence of noxious occupational exposure to asbestos than in its absence.

4.2 Calendar versus biological ages: modeling the ageing acceleration due to occupational exposure to asbestos

We present now an original class of acceleration functions tailored to our particular description of occupational exposures to asbestos. Let us define

$$\mathcal{M} = \Big\{(M_0, (M_{k,l})_{k,l\leq 3}) \in \mathbb{R}_+ \times \mathbf{M}_{3,3}(\mathbb{R}_+) : 0 \leq M_{k,1} \leq M_{k,2} \leq M_{k,3} = 1, k = 1,2,3\Big\}. \quad (16)$$

Then the rate yielded by description $\varepsilon = \varepsilon_1\varepsilon_2\varepsilon_3 \in \mathcal{E} \setminus \{0\}$ for $M \in \mathcal{M}$ writes as

$$M(\varepsilon) = 1 + M_0 \times M_{1,\varepsilon_1} \times M_{2,\varepsilon_2} \times M_{3,\varepsilon_3}, \quad (17)$$

and that of $\varepsilon = 0$ is set to $M(0) = 1$. Notably, M_0 is the factor of acceleration of time for the higher exposure, which we recall is encoded by $\varepsilon = 333$. Rates $M(\varepsilon)$ range from 1 to $M(333) = 1 + M_0$ and (with convention $0/0 = 1$)

$$\frac{M(\varepsilon) - 1}{M_0} = M_{1,\varepsilon_1} \times M_{2,\varepsilon_2} \times M_{3,\varepsilon_3} :$$

an exposure characterized by "probability/frequency/intensity" description $\varepsilon = \varepsilon_1\varepsilon_2\varepsilon_3$ achieves a fraction $M_{1,\varepsilon_1} \times M_{2,\varepsilon_2} \times M_{3,\varepsilon_3}$ of the maximal acceleration.

Note that we only need 7 parameters in order to fully describe the 28 possibly different rates of acceleration. Furthermore, it is easily seen that this parametrization is identifiable: if $M, M' \in \mathcal{M}$ satisfy $M(\varepsilon) = M'(\varepsilon)$ for all $\varepsilon \in \mathcal{E}$ then $M = M'$.

Consider $M \in \mathcal{M}$ and a generic longitudinal history $\bar{\varepsilon}$ as presented in Section 2. The mapping $M(\varepsilon(\cdot)) : t \mapsto M(\varepsilon(t))$ is piecewise constant but it is more convenient to consider a continuous approximation to it, which we denote by $r(M, \bar{\varepsilon})$. (Formally, one may rely on convolution to derive $r(M, \bar{\varepsilon})$ from $M(\varepsilon(\cdot))$.) Finally, every pair $(M, \bar{\varepsilon})$ gives rise to the acceleration function $R(M, \bar{\varepsilon})$ characterized by

$$R(M, \bar{\varepsilon})(t) = \int_0^t r(M, \bar{\varepsilon})(s)ds \quad (18)$$

(all $t \geq 0$). The quantity $R(M, \bar{\varepsilon})(t)$ is a summary measure of the history of occupational exposure to asbestos, and can be interpreted as the cumulative occupational exposure to asbestos up to age t. In particular if $\varepsilon(t) = 0$ for all $t \geq 0$ (*i.e.*, in absence of occupational exposure to asbestos throughout lifetime) or if $M_0 = 0$ (*i.e.*, assuming that exposure to asbestos has no effect on the risk of developing lung cancer), then $R(M, \bar{\varepsilon})(t) = t$ for all $t \geq 0$: in other words, the chronological and biological time scales coincide. Note that $R(M, \bar{\varepsilon})$ is *differentiable* because $r(M, \bar{\varepsilon})$ is continuous.

Note that $R(M, \bar{\varepsilon})(t)$ is very close to a linear combination of the times spent in each job category (if larger than 6 months) up to calendar time t, where the coefficients of the combination depend through M on the job categories. In this view, our acceleration function is classical [Lee and Whitmore, 2006]. What makes it original though is how M maps each description $\varepsilon \in \mathcal{E}$ to its coefficient $M(\varepsilon)$. We comment further on our acceleration function in Section 4.3.

Now, given parameters $h > 0$, μ, $M \in \mathcal{M}$ and covariate $\bar{a}$, we obtain $R_a = R(M, \bar{a})$, which yields in turn the time to incident lung cancer $T\{h, \mu, R_a\}$ for the history of occupational exposure to asbestos summarized by R_a.

4.3 The parametric model, and related case-control weighted log-likelihood loss function

We complete the characterization of our model, and derive the related log-likelihood loss function. By Section 3, we know that it suffices to model the distribution of the observed data structure $O^\star$ under representative sampling. As explained in Section 3, we wish to model parametrically the conditional distribution of $O^\star$ given $\Omega = (W, X, \bar{A}(X), Y)$, *i.e.*, the conditional distribution of Z given Ω, leaving the conditional distribution of Ω given Y unspecified.

For this purpose, we state that under $\theta = (\alpha, \beta, M) \in \Theta = \mathbb{R}^4 \times \mathbb{R}^{16} \times \mathcal{M}$, the conditional distribution of T (the possibly unobserved time to incident lung cancer of the subject associated with $O^\star$) given Ω is that of $T\{h, \mu, R_a\}$ with

$$\log h = \alpha_{\text{index}(W_1, W_2)} \tag{19}$$

(each level of (W_1, W_2) is associated with a unique positive initial health h),

$$\log(-\mu) = \beta_{\text{index}(W_1, W_2, W_3)} \tag{20}$$

(each level of (W_1, W_2, W_3) is associated with a unique negative drift μ), and

$$R_a = R(M, \bar{A}(X)).$$

Therefore, it holds that $\log p^\star_\theta(Z|\Omega) = Y \log g(\theta)(Z) + (1 - Y) \log G(\theta)(Z)$, with convention $G(\theta)(Z) = G(h, \mu, R_a)(Z)$ and $g(\theta)(Z) = g(h, \mu, R_a)(Z)$ (see (14) and (15) for the definitions of G and g). Finally, the relevant part of the resulting case-control weighted log-likelihood at $\theta \in \Theta$ writes as

$$\sum_{i=1}^{n} \left\{ q_0 \log g(\theta)(Z_i^1) + \bar{q}_0(V_i^1) \frac{1}{J_i} \sum_{j=1}^{J_i} \log G(\theta)(Z_i^{0,j}) \right\} = P_n \tilde{\ell}(\theta), \tag{21}$$

where $\tilde{\ell}(\theta)(O) = \ell(p^\star_\theta)(O) - \text{rem}(O)$ (see equation (9)) and $P_n = \sum_{i=1}^n \delta_{O_i}$ denotes the empirical measure.

Comments. Oakes [1995] carried out the study of the effect of exposure to asbestos on time until death from (not incidence of) lung cancer. Although the path that leads us to our model is quite different from his (asbestos exposure are measured differently, more covariates are accounted for here than there), the conditional cdf at $t \geq 0$ of time until the event of interest given the rest ends up in both cases being of the form $F(R_a(t))$. Here $F = F(h, \mu)$ and $R_a = R(M, \bar{A}(X))$ whereas there F is a Weibull cdf and $R_a : t \mapsto t + \rho \text{cum}(t)$, $\text{cum}(t)$ representing the cumulative exposure to asbestos at age t in 100 million particles per cubic foot (measured in mppcf-years), and ρ an "equivalence parameter". Hence both model are accelerated failure time models of the simple collapsible form (since $F(R_a(t))$ depends on the cumulated exposure at time t and not on the entire exposure history up to t; see [Duchesne and Rosenthal, 2003] for theoretical conditions under which an accelerated failure time model turns out to be collapsible). Note that the expression "equivalence parameter" conveys the notion that a cumulative exposure of $1/\rho$ mppcf-years of asbestos dust has the same effect on the cumulative lung cancer risk as an additional year of life. In view of (17) and (18), the two acceleration functions are very similar, the component M_0 of $M = (M_0, (M_{k,l})_{k,l \leq 3})$ playing the same role as ρ, and $(M_{k,l})_{k,l \leq 3}$ serving as a mean to associate a concentration of exposure to asbestos to every $\varepsilon \in \mathcal{E}$.

Furthermore, we emphasize that our case-control weighted log-likelihood differs from the log-likelihood developed in [Lee et al., 2009] even though the latter article also relies on a case-control study. Lee et al. [2009] justify the form of their log-likelihood and discuss on the limitations of the inferences they subsequently draw from it. However, we believe that the difficulties that stem from the fact that a case-control sampling is performed are not fully addressed by them.

4.4 Expected years of life free of lung cancer lost due to occupational exposure to asbestos

Equivalence (13) has an important consequence: given parameters $h > 0$, $\mu \leq 0$, $M \in \mathcal{M}$ and history of occupational exposure to asbestos $\bar{A}(X)$, $T\{h, \mu\} = R_a(T\{h, \mu, R_a\})$. In words, all things (gender, occurrence of lung cancer in close family, lifetime tobacco use) being equal, the time to incident lung cancer *in the absence of occupational exposure to asbestos* can be deduced

deterministically from the (observed) age at incident lung cancer and history of occupational exposure to asbestos of a case. The nonnegative quantity

$$R_a(T\{h,\mu,R_a\}) - T\{h,\mu,R_a\}$$

(with convention $R_a(\infty) = \infty$ and $\infty - \infty = 0$) can be interpreted as a number of years of life free of lung cancer that the case could have enjoyed had he/she not been exposed to asbestos. Note that $R_a(T\{h,\mu,R_a\}) - T\{h,\mu,R_a\}$ is different from the remaining number of years of life free of lung cancer, as death may occur anytime after $T\{h,\mu,R_a\}$ even in the absence of occupational exposure to asbestos.

There is a strong connection between $R_a(T\{h,\mu,R_a\}) - T\{h,\mu,R_a\}$ and *expected years of life lost* as introduced by Robins and Greenland [1991]. Following [Robins and Greenland, 1991], consider the expected years of life lost $\Delta(t)$ for a randomly sampled person among subjects *(i)* who share the same characteristics (gender, occurrence of lung cancer in close family, lifetime tobacco use, hence the common parameter (h,μ)), *(ii)* who develop an incident lung cancer at age t, and *(iii)* whose histories of occupational exposure to asbestos up to age t coincide (hence the same acceleration function R_a, at least up to time t). Then under some assumptions on how the exposure affects health [Robins and Greenland, 1991, Theorem 3], it holds that $\Delta(t) = S_0^{-1} \circ S_1(t) - t$, where $S_0 = 1 - F(h,\mu)$ and $S_1 = (1 - F(h,\mu)) \circ R_a$, so that $\Delta(t) = R_a(t) - t$. For this reason, we decide to refer to $R_a(T\{h,\mu,R_a\}) - T\{h,\mu,R_a\}$ as expected years of life (free of lung cancer) lost (due to occupational exposure to asbestos).

5 Results

We present here the inferences we draw from our dataset, model, and resulting case-control weighted log-likelihood loss function as described in Sections 2 and 4.3. In Section 5.1 we comment on the fitted model, then we focus in Section 5.2 on its implications in terms of expected years of life lost.

5.1 Fitting the best model

It makes no doubt that the model we have built so far is over-dimensioned. The "probability/frequency/intensity" description with its 28 different levels is itself certainly too rich (see Table 1), or at least difficult to establish and prone to errors. We rather consider the model $\{P_\theta^\star : \theta \in \Theta\}$ described so far as a "maximal" model giving rise to a large collection of sub-models $\{P_\theta^\star : \theta \in \Theta_k\}$ obtained by adding constraints on the "maximal" parameter $\theta = (\alpha, \beta, M) \in \Theta$. The number of such sub-models is large indeed: there are $(1 + 7^3) = 344$ sub-models defined by adding only constraints on M (of the type $M_0 = 0$, or $M_0 > 0$ and for any $k = 1, 2, 3$, $0 = M_{k,1}$ or $M_{k,1} = M_{k,2}$ or $M_{k,2} = 1$ or $0 = M_{k,1} = M_{k,2}$ or $M_{k,1} = M_{k,2} = 1$ or $(0 = M_{k,1}, M_{k,2} = 1)$), hence the total number of sub-models equals $2^2 \times 2^3 \times 344 = 11,008$. It is out of question to explore the whole collection of sub-models. Instead, we propose to

(i) define a large collection $\{\Theta_k : k \in \mathcal{K}\}$ of interest,

(ii) let the data select a better $\Theta_{\hat{k}}$ in the latter collection based on a multi-fold likelihood-based cross-validation criterion.

van der Vaart et al. [2006] show that, under mild assumptions, the multi-fold likelihood-based cross-validation criterion will select a better model comparing favorably with the oracle model of the collection (whose definition involves the true distribution of the data). By this we mean that the likelihood risk of the better model will not be much bigger than that of the oracle model. Although we cannot invoke rigorously this remarkable property here, it motivates the procedure that we describe in Sections C.2 and C.3.

The best model $\{P_\theta^\star : \theta \in \Theta_{\hat{k}}\}$ is described in Section C.3. Its characterization teaches us that neither the initial health nor the drift parameter depend on the indicator of occurrence of lung

cancer in close family. Moreover, *(i)* a passive exposure (probability index equal to 1) is the same as no exposure at all, and *(ii)* being exposed less than once a month (frequency index equal to 1) or between once a month and during less than half of the monthly working hours (frequency index equal to 2) have the same effect.

We first fit the best model in terms of maximum likelihood on the whole dataset. Regarding the derivation of confidence intervals, we decide to rely on the bootstrap instead of a central limit theorem (such as Proposition 3 in Section A). The particulars of the bootstrap procedure follow. We set $\alpha = 2.5\%$, $B = 1,000$ and $p = 5\%$, then for b ranging from 1 to B, we repeatedly resample without replacement $n(1-p) = 817$ observed data structures, yielding the bootstrapped empirical measure $P^b_{n(1-p)}$, in order to compute and store the corresponding maximum likelihood estimate $\theta_{n(1-p),\hat{k}}(P^b_{n(1-p)})$ of $\theta \in \Theta_{\hat{k}}$. The mean and median values of $\boldsymbol{\theta}^B_{n,\hat{k}} = \{\theta_{n(1-p),\hat{k}}(P^b_{n(1-p)}) : b \leq B\}$ only very slightly differ from each other. Moreover, they are very close to the maximum likelihood estimate $\theta_{n,\hat{k}}(P_n)$ computed on the whole dataset. The componentwise $\alpha/16$- and $(1-\alpha/16)$-quantiles of $\boldsymbol{\theta}^B_{n,\hat{k}}$ are used as lower- and upper-bounds of confidence intervals, which simultaneously provide a $(1-2\alpha) = 95\%$-coverage by the applied Bonferroni correction. Specifically, we obtain:

- *initial health*:

W_1	h
0	23.82 [23.42; 24.13]
1	25.09 [24.86; 25.40]

  It is seen in particular that women are associated with a significantly larger initial health than men.

- *drift*:

	-100μ	
W_3	$W_1 = 0$	$W_1 = 1$
0	0.69 [0.08; 1.46]	0.02 [0.01; 0.03]
1	7.70 [6.91; 8.28]	6.63 [5.73; 7.68]
2	13.89 [13.25; 14.46]	10.55 [9.63; 11.80]
3	17.67 [17.11; 18.38]	14.79 [13.65; 17.77]

  Two main features arise:

  – For each level of lifetime tobacco use, the absolute value of the drift is significantly larger for men than for women (actually, the confidence intervals for $W_3 = 3$ slightly overlap). Combined with the already mentioned fact that women are associated with a larger initial health, this implies that *for any given history of exposure to asbestos* and for every level of lifetime tobacco use, the distribution of time to incident lung cancer in women is stochastically dominated by the distribution of time to incident lung cancer in men. In other words, given a man and a woman sharing the same history of exposure to asbestos and lifetime tobacco use, given an age t, the man is more likely to have developed an incident lung cancer at age t than the woman.

    Note that there is no clear consensus in the literature on whether there exist differences in lung cancer risk between men and women or not (for instance, Zang and L. [1996] argue that women are more susceptible to tobacco carcinogens, but Haiman et al. [2006] show that men *or* women are more susceptible to tobacco carcinogens, depending on ethnic and racial group).

  – Both in men and women, the absolute value of the drift significantly increases with lifetime tobacco use. This implies that, both in men and women, *for any given history of exposure to asbestos* and for every $0 \leq w < w' \leq 3$, the distribution of time to incident lung cancer for lifetime tobacco use equal to w is stochastically dominated by

ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$
111	0	211	0.026 [0.000; 0.171]	311	0.026 [0.000; 0.173]
112	0	212	0.092 [0.001; 0.530]	312	0.094 [0.001; 0.537]
113	0	213	1.078 [0.297; 1.939]	313	1.108 [0.309; 1.964]
121	0	221	0.026 [0.000; 0.171]	321	0.026 [0.000; 0.173]
122	0	222	0.092 [0.001; 0.530]	322	0.094 [0.001; 0.537]
123	0	223	1.078 [0.297; 1.939]	323	1.108 [0.309; 1.964]
131	0	231	0.027 [0.000; 0.174]	331	0.028 [0.000; 0.176]
132	0	232	0.099 [0.001; 0.539]	332	0.101 [0.001; 0.546]
133	0	233	1.159 [0.330; 1.971]	333	1.192 [0.344; 1.998]

Table 2: Estimated values (precision 10^{-3}) of the factor of acceleration of time $(M(\varepsilon)-1)$ and related confidence intervals for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$. Recall that $M(0)=1$.

the distribution of time to incident lung cancer for lifetime tobacco use equal to w'. In other words, given two persons sharing the same gender and history of exposure to asbestos, the person with the larger lifetime tobacco use is more likely to have developed an incident lung cancer at age t than the other.

This is in agreement with the general scientific consensus [Biesalski et al., 1998].

- *exposure to asbestos*:

M_0: 1.19 [0.34; 2.00]		
$M_{1,1}=0$	$M_{1,2}$: 0.97 [0.96; 0.99]	$M_{1,3}=1$
$M_{2,1}=M_{2,2}$	$M_{2,2}$: 0.93 [0.90; 0.98]	$M_{2,3}=1$
$M_{3,1}$: 0.02 [0.00; 0.09]	$M_{3,2}$: 0.09 [0.00; 0.27]	$M_{3,3}=1$

We notably derive from the above table the values of $(M(\varepsilon)-1)$ (which can be interpreted as a factor of acceleration of time due to an exposure of level ε, see (17)) and related confidence intervals for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$, see Table 2.

5.2 Application to the expected years of life free of lung cancer lost due to occupational exposure to asbestos

In view of Section 4.4, the results of the previous section provide us with a way of evaluating the expected years of life (free of lung cancer) lost (due to occupational exposure to asbestos) on a case by case basis. Say that we mostly care for a pointwise estimation of, and confidence lower-bound on, the expected years of life lost. A confidence upper-bound could be derived similarly. We compute a counterpart of Table 2 based on the componentwise $2\alpha/5$-quantiles of $\boldsymbol{\theta}^B_{n,\hat{k}}$. They simultaneously provide $(1-2\alpha)=95\%$-coverage for parameter M *on its own* by the applied Bonferroni correction, since M has 5 degrees of freedom, see Table 3.

Elementary algebra permits to compute an evaluation $\delta(t,\bar{a}(t))$ of, and confidence lower-bound $\delta^-(t,\bar{a}(t))$ on, the expected years of life lost for any couple $(t,\bar{a}(t))$ of age t at incident lung cancer and history $\bar{a}(t)$ of occupational exposure to asbestos till t. Let us consider three examples:

- Consider a case of incident lung cancer at age t who spent, till that age, 30 years with an occupational exposure to asbestos $\varepsilon=332$: one evaluates the 95%-confidence lower bound $\delta^-(t,\bar{a}(t)) = 30\times 0.004 = 0.09$ expected years of life lost (approximately 44 days) and $\delta(t,\bar{a}(t))=30\times 0.101=3.03$ expected years of life lost.

  This is quite an extreme case, since 3 out of the 8,432 employments described in the dataset achieve the description $\varepsilon=332$.

ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$	ε	$M(\varepsilon)-1$
111	0	211	0.026 $[0.001;\infty)$	311	0.026 $[0.001;\infty)$
112	0	212	0.092 $[0.004;\infty)$	312	0.094 $[0.004;\infty)$
113	0	213	1.078 $[0.374;\infty)$	313	1.108 $[0.389;\infty)$
121	0	221	0.026 $[0.001;\infty)$	321	0.026 $[0.001;\infty)$
122	0	222	0.092 $[0.004;\infty)$	322	0.094 $[0.004;\infty)$
123	0	223	1.078 $[0.374;\infty)$	323	1.108 $[0.389;\infty)$
131	0	231	0.027 $[0.001;\infty)$	331	0.028 $[0.002;\infty)$
132	0	232	0.099 $[0.004;\infty)$	332	0.101 $[0.004;\infty)$
133	0	233	1.159 $[0.414;\infty)$	333	1.192 $[0.431;\infty)$

Table 3: Estimated values (precision 10^{-3}) of the factor of acceleration of time ($M(\varepsilon)-1$) and related *right* confidence intervals for each level of exposure $\varepsilon \in \mathcal{E} \setminus \{0\}$. Recall that $M(0)=1$. A Bonferroni correction ensures that the confidence regions simultaneously guarantee $(1-2\alpha)=95\%$-coverage (for $\{M(\varepsilon)-1 : \varepsilon \in \mathcal{E}\}$ on its own).

- Consider a case of incident lung cancer at age t who spent, till that age, 10 years (then later 5 years and 2 years) with an occupational exposure to asbestos $\varepsilon = 322$ (then later $\varepsilon = 121$ and $\varepsilon = 222$): one evaluates the 95%-confidence lower bound $\delta^-(t, \bar{a}(t)) = 10 \times 0.004 + 5 \times 0 + 2 \times 0.004 = 0.048$ years of life lost (approximately 17.5 days) and $\delta(t, \bar{a}(t)) = 10 \times 0.094 + 5 \times 0 + 2 \times 0.092 = 1.124$ expected years of life lost.

  Note that 150, 36 and 189 out of the 8,432 employments described in the dataset achieve the descriptions $\varepsilon = 121$, $\varepsilon = 222$ and $\varepsilon = 322$.

- Consider a case of incident lung cancer at age t who spent, till that age, 10 years (then later 15 years) with an occupational exposure to asbestos $\varepsilon = 213$ (then later $\varepsilon = 223$): one evaluates the 95%-confidence lower bound $\delta^-(t, \bar{a}(t)) = 10 \times 0.374 + 15 \times 0.374 = 9.350$ expected years of life lost and $\delta(t, \bar{a}(t)) = 10 \times 1.078 + 15 \times 1.078 = 26.95$ expected years of life lost.

  This is quite an extreme case, since only 6 and 3 out of the 8,432 employments described in the dataset achieve the descriptions $\varepsilon = 213$ and $\varepsilon = 223$.

Among the $n = 860$ cases of our dataset, only 259 (*i.e.*, 30%) cases are associated with positive expected years of life lost. We report in Table 4 the quartiles, mean and extreme values of expected years of life lost as computed on those 259 cases.

- The maximum value is reached by a male who accumulated through his professional life a total of 33 years with occupational exposure to asbestos equal to $\varepsilon = 313$ and was diagnosed a lung cancer at 70 years old. Although this is not relevant as far as the evaluation of the expected years of life lost is concerned, his lifetime tobacco equals 45 pack years.

- The minimum value is reached by 4 women who accumulated through their professional lives a total of 1 year with occupational exposure to asbestos $\varepsilon \in \{211, 221\}$ and were diagnosed a lung cancer at 51 (for two of them), 59 and 68 years old. Although this is not relevant as far as the evaluation of the expected years of life lost is concerned, their lifetime tobacco uses equal 25, 30, 32 and 55 pack years).

- The median value is reached by a man who accumulated through his professional life a total of 4 years (respectively, 5 and 7) with occupational exposure to asbestos equal to $\varepsilon = 111$ (respectively, $\varepsilon = 211$ and $\varepsilon = 212$) and was diagnosed a lung cancer at 71 years old. Although this is not relevant as far as the evaluation of the expected years of life lost is concerned, his lifetime tobacco equals 55 pack years.

We represent in Figure 1 the empirical cdf of the expected years of life lost (and corresponding 95%-confidence lower bounds) for the 259 cases for whom it is positive.

	min.	25%	50%	mean	75%	max.
expected years of life lost	0.026	0.289	0.769	2.467	2.408	36.577
95%-lower bound	0.001	0.014	0.037	0.555	0.102	12.832

Table 4: Quartiles, mean and extreme values of the expected years of life lost and corresponding 95%-confidence lower-bound (precision 10^{-3}), as computed on those 259 cases (*i.e.*, 30% of all cases) for whom the evaluated expected years of life lost is positive.

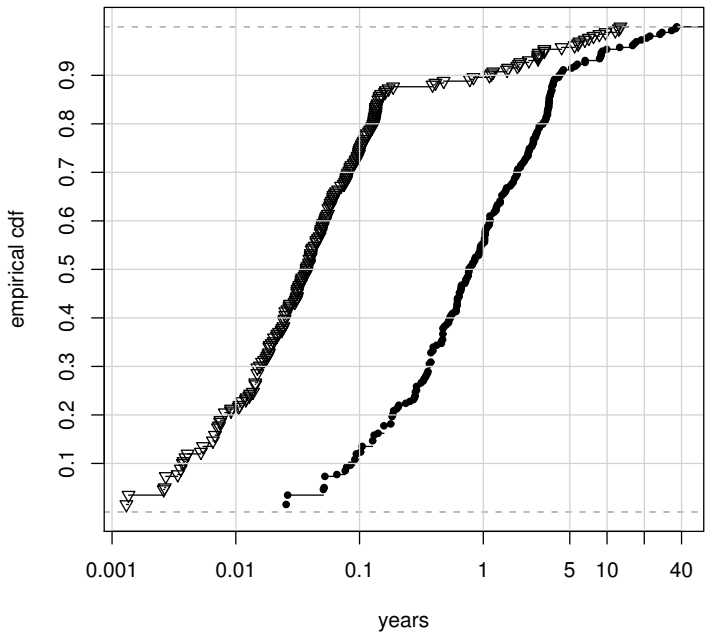

Figure 1: **Empirical distributions of expected years of life lost and related confidence lower-bound.** The rightmost curve with bullets (respectively leftmost curve with triangles) represents the empirical cdf of the expected years of life lost (respectively of the 95%-confidence lower bound on that number) of those cases for whom it is positive, that is the empirical cdf of $\{\delta(T_i^1, \bar{A}^1(T_i^1)) : \delta(T_i^1, \bar{A}^1(T_i^1)) > 0, i \leq n\}$ (respectively $\{\delta^-(T_i^1, \bar{A}^1(T_i^1)) : \delta(T_i^1, \bar{A}^1(T_i^1)) > 0, i \leq n\}$). Only 30% of the cases are concerned. The x-axis scale is logarithmic.

6 Discussion

We have developed a collection of threshold regression models (see Section 4.3), and have data-adaptively selected a better model in it by relying on multi-fold likelihood-based cross-validation (see Sections C.2 and C.3 for the descriptions of the model selection procedure and derived better model). The latter better threshold regression model has been fitted by maximum likelihood, and bootstrapped confidence intervals have been obtained (see Section 5.1). The statistical procedure has been adjusted in order to eliminate the bias induced by the matched case-control sampling design used to collect the dataset (see Sections 3.3 and 4.3). This necessary preliminary step was made possible because the joint distribution of (V, Y) in the population of interest can be computed beforehand independently from our dataset (see Section C.1). We have discussed the implications of the fitted threshold regression model in terms of expected years of life (free of lung cancer) lost (due to occupational exposure to asbestos) which is naturally attached to it (see Section 5.2).

We finally acknowledge a limitation of the approach undertaken in this article: The link between the occupational exposure to asbestos and age at incident lung cancer is well-defined in the context of the proposed threshold regression models, but we do not extend it beyond. The parameter we aim for is therefore difficult to comprehend (it is related to the Kullback-Leibler projection of the true distribution of the data onto a threshold regression model), and the inference procedure certainly fails to estimate optimally/efficiently what we really care for, which would be a measure of the strength of the link between the occupational exposure to asbestos and age at incident lung cancer defined non- or semiparametrically. We intend to build on the present study and go further in that direction in future work.

A Asymptotic properties of the case-control weighted maximum likelihood estimator

We recall that $\mathcal{F} = \{\ell(p^\star_\theta) : \theta \in \Theta\}$ is P_0-Glivenko-Cantelli if $\sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^n \ell(p^\star_\theta)(O_i) - E_{P_0}\ell(p^\star_\theta)(O)| = o_{P_0}(1)$. The following classical consistency result holds (see Theorem 5.7 and Example 19.8 in [van der Vaart, 1998]).

Proposition 2. *Assume that $E_{P_0^\star} \log p^\star_0(O)$ is well-defined and that the mapping $\theta \mapsto \mathrm{KL}(p^\star_0, p^\star_\theta)$ from Θ to the nonnegative real numbers attains its minimum uniquely at $\theta_0 \in \mathrm{int}(\Theta)$ ($p^\star_{\theta_0}$ is the Kullback-Leibler projection of $p^\star_0$ upon $\{p^\star_\theta : \theta \in \Theta\}$). If $\mathcal{F}$ is P_0-Glivenko-Cantelli then θ_n converges in probability to θ_0. This is for instance the case if Θ is a compact metric space, if $\theta \mapsto \ell(p^\star_\theta)(o^\star)$ is continuous for every $o^\star$ and if $\mathcal{F}$ admits an integrable envelope function with respect to P_0.*

We also derive an asymptotic normality result (inspired by classical results of asymptotic normality, see Theorem 5.23 in [van der Vaart, 1998]; we omit the measurability conditions).

Proposition 3 (first part)**.** *In the context of Propositions 1 and 2, assume in addition that $\theta \mapsto \log p^\star_\theta(Z|\Omega)$ is twice differentiable at θ_0 $P^\star_0$-almost surely with first and second derivatives $\dot\ell^\star_{\theta_0}(O^\star)$ and $\ddot\ell^\star_{\theta_0}(O^\star)$ such that $\int l(o^\star)dP^\star_0(o^\star, Y = y)$ are properly defined for $l = \dot\ell^\star_{\theta_0}, \ddot\ell^\star_{\theta_0}$ and $y = 0, 1$, and introduce accordingly the weighted versions*

$$\begin{aligned}
\dot{\tilde\ell}(\theta_0)(O) &= q_0\dot\ell^\star_{\theta_0}(V^1, O^{1\star}) + \bar q_0(V^1)\frac{1}{J}\sum_{j=1}^J \dot\ell^\star_{\theta_0}(V^1, O^{0,j\star}),\\
\ddot{\tilde\ell}(\theta_0)(O) &= q_0\ddot\ell^\star_{\theta_0}(V^1, O^{1\star}) + \bar q_0(V^1)\frac{1}{J}\sum_{j=1}^J \ddot\ell^\star_{\theta_0}(V^1, O^{0,j\star}).
\end{aligned}$$

Suppose also that, for every θ_1, θ_2 in a neighborhood of θ_0 and a function $\dot m$ such that $E_{P_0^\star}\dot m(O^\star)^2 < \infty$, $P^\star_0$-almost surely

$$|\log p^\star_{\theta_1}(Z|\Omega) - \log p^\star_{\theta_2}(Z|\Omega)| \le \dot m(O^\star)\|\theta_1 - \theta_2\|.$$

Furthermore, assume that $\theta \mapsto E_{P_0}\ell(p^\star_\theta)(O) = E_{P_0^\star}\log p^\star_\theta(O^\star)$ *(by virtue of Proposition 1) admits a second-order Taylor expansion at* θ_0 *with nonsingular symmetric second derivative matrix* $S_{\theta_0} = E_{P_0}\ddot{\ell}(\theta_0)(O)$*. Then* $S_{\theta_0} = E_{P_0^\star}\ddot{\ell}^\star_{\theta_0}(O^\star)$ *and*

$$\sqrt{n}(\theta_n - \theta_0) = -S_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}(\theta_0)(O_i) + o_{P_0}(1). \tag{22}$$

In particular, the sequence $\sqrt{n}(\theta_n - \theta_0)$ *is asymptotically Gaussian with mean zero and covariance matrix* $\Sigma = S_{\theta_0}^{-1}E_{P_0}[\dot{\ell}(\theta_0)(O)\dot{\ell}(\theta_0)(O)^\top]S_{\theta_0}^{-1}$*.*

We purposely do not use the same convention to denote the first and second order derivatives at θ_0 of $\theta \mapsto \log p^\star_\theta(Z|\Omega)$ (respectively $\dot{\ell}^\star_{\theta_0}(O^\star)$ and $\ddot{\ell}^\star_{\theta_0}(O^\star)$) and the derivatives of $\theta \mapsto \tilde{\ell}(\theta)(O)$ (respectively $\dot{\tilde{\ell}}(\theta_0)(O)$ and $\ddot{\tilde{\ell}}(\theta_0)(O)$). We intend to stress that the former are related to the representative sampling observed data structure $O^\star$ whereas the latter are related to the case-control sampling observed data structure O.

A natural question arises: How does the asymptotic covariance matrix Σ compare with the asymptotic covariance matrix one would have got under representative sampling? We give in the second part of Proposition 3 a very simple answer, but for a representative sampling under a modified version of $P_0^\star$. Introduce for clarity of exposition the notation

$$\bar{q}_0(o^\star) = q_0\frac{q_0(y|v)}{q_0(1|v)} \tag{23}$$

such that $\bar{q}_0(o^\star) = q_0$ if $y = 1$ ($o^\star$ corresponds to a case) and $\bar{q}_0(o^\star) = \bar{q}_0(v)$ if $y = 0$ ($o^\star$ corresponds to a control). By setting

$$\frac{dP_1^\star}{dP_0^\star}(o^\star) = \frac{1}{2\bar{q}_0(o^\star)}, \tag{24}$$

we define a probability distribution $P_1^\star$ for the observed data structure $O^\star$ (indeed, $o^\star \mapsto \bar{q}_0(o^\star)$ is positive and $E_{P_0^\star}(2\bar{q}_0(O^\star))^{-1} = 1$). Moreover:

- under $P_1^\star$, being a case is as likely as being a control (*i.e.*, $P_1^\star(Y = 1) = \frac{1}{2}$);
- the marginal distribution of the matching variable V under $P_1^\star$ equals the conditional distribution of V under $P_0^\star$, conditionally on being a case (*i.e.*, $P_1^\star(V = v) = q_0(v|1)$ for all $v \in \mathcal{V}$);
- given (V, Y), $O^\star$ has the same distribution under $P_1^\star$ as under $P_0^\star$ (indeed, $\bar{q}_0(o^\star)$ depends on $o^\star$ through (v, y) only).

Furthermore, since obviously

$$2E_{P_1^\star}\bar{q}_0(O^\star)\log p^\star_\theta(O^\star) = E_{P_0^\star}\log p^\star_\theta(O^\star),$$

$\theta \mapsto \bar{q}_0(O^\star)\log p^\star_\theta(O^\star)$ is a proper loss function for the purpose of estimating θ_0 under representative sampling of $O^\star \sim P_1^\star$.

Proposition 3 (second part)**.** *Define* $S'_{\theta_0} = E_{P_1^\star}\bar{q}_0(O^\star)\ddot{\ell}^\star_{\theta_0}(O^\star)$*. Suppose that the model is well-specified (or equivalently that* $\mathrm{KL}(p_0^\star, p^\star_{\theta_0}) = 0$*). Assume in addition that for all* $o^{1\star} = (z^1, \omega^1)$*, the class of derivatives of* $p^\star_\theta(z^1|\omega^1)$ *with respect to* θ *is uniformly bounded (in* θ*) by an integrable function (of* z^1*). Then the covariance matrix* Σ *satisfies*

$$\Sigma = \frac{1}{2}{S'_{\theta_0}}^{-1}E_{P_1^\star}[\bar{q}_0(O^\star)^2\dot{\ell}^\star_{\theta_0}(O^\star)\dot{\ell}^\star_{\theta_0}(O^\star)^\top]{S'_{\theta_0}}^{-1}.$$

In particular, 2Σ can be interpreted as the asymptotic covariance matrix of the M-estimator of θ_0 based on the loss function $\theta \mapsto \bar{q}_0(O^\star)\log p^\star_\theta(O^\star)$ and n iid observations drawn from $P_1^\star$. The n observations under P_0-case-control sampling correspond to $2 \times n$ observations under $P_1^\star$-representative sampling, each of them in the former counting for two in the latter. Elements of proof are relegated to Section B.

B Elements of proof

Proof of Proposition 1. On one hand, note that

$$\begin{aligned} E_{P_0} q_0 \log p^\star(V^1, O^{1\star}) &= \int q_0 \log p^\star(v^1, o^{1\star}) dP_0^\star(v^1, o^{1\star} | y = 1) \\ &= \int \log p^\star(v^1, o^{1\star}) dP_0^\star(v^1, o^{1\star}, y = 1) \\ &= \int \log p^\star(o^\star) dP_0^\star(o^\star, y = 1). \end{aligned} \tag{25}$$

On the other hand, for each $j \leq J$,

$$\begin{aligned} E_{P_0} \bar{q}_0(V^1) \log p^\star(V^1, O^{0,j\star}) &= E_{P_0} \bar{q}_0(V^1) E_{P_0} \left[\log p^\star(V^1, O^{0,j\star}) | V^1 \right] \\ &= E_{P_0} \bar{q}_0(V^1) \int \log p^\star(V^1, o^\star) dP_0^\star(o^\star | V^1, y = 0) \\ &= \int \bar{q}_0(v^1) \log p^\star(v^1, o^\star) dP_0^\star(o^\star | v^1, y = 0) dP_0(v^1). \end{aligned}$$

Furthermore, for each $v \in \mathcal{V}$, $dP_0(v) = dP_0^\star(v | y = 1) = q_0(v|1)\delta_v(v)$ (we use the same shorthand notation as in (4), (5), (6), (7)) and denote by δ_v the Dirac mass at v), hence

$$\bar{q}_0(v) dP_0(v) = q_0 \frac{q_0(0|v)}{q_0(1|v)} q_0(v|1)\delta_v = q_0(0|v) P_0^\star(v) \delta_v(v) = dP_0^\star(v, y = 0).$$

Consequently, we obtain

$$\begin{aligned} E_{P_0} \bar{q}_0(V^1) \log p^\star(V^1, O^{0,j\star}) &= \int \log p^\star(v^1, o^\star) dP_0^\star(o^\star | v^1, y = 0) dP_0^\star(v^1, y = 0) \\ &= \int \log p^\star(v^1, o^\star) dP_0^\star(v^1, o^\star, y = 0) \\ &= \int \log p^\star(o^\star) dP_0^\star(o^\star, y = 0) \end{aligned} \tag{26}$$

(which does not depend on j). Combining (25), (26) finally yields

$$E_{P_0} \ell(p^\star)(O) = \int \log p^\star(o^\star) dP_0^\star(o^\star) = E_{P_0^\star} \log p^\star(O^\star).$$

The conclusion is straightforward, because

$$E_{P_0^\star} \log p^\star(O^\star) - E_{P_0^\star} \log p_0^\star(O^\star) = -\mathrm{KL}(p_0^\star, p^\star),$$

the opposite of the Kullback-Leibler divergence between $p_0^\star$ and $p^\star$, which is positive for $p^\star \neq p_0^\star$ and equals zero otherwise. □

Proof of Proposition 3. The expansion (22) and the related distributional limit result are a consequence of [van der Vaart, 1998, Theorem 5.23]. The fact that $S_{\theta_0} = E_{P_0^\star} \ddot{\ell}_{\theta_0}^\star(O^\star)$ is obtained by adapting slightly the proof of Proposition 1. Regarding $E_{P_0}[\dot{\tilde{\ell}}(\theta_0)(O) \dot{\tilde{\ell}}(\theta_0)(O)^\top]$, let us abbreviate $xx^\top$ to x^2 and note that

$$\begin{aligned} &\dot{\tilde{\ell}}(\theta_0)(O) \dot{\tilde{\ell}}(\theta_0)(O)^\top \\ &\quad = \Big[q_0 \bar{q}_0(O^{1\star}) \dot{\ell}_{\theta_0}^\star(V^1, O^{1\star})^2 + \bar{q}_0(V^1) \Big(\tfrac{1}{J} \textstyle\sum_j \bar{q}_0(O^{0,j\star}) \dot{\ell}_{\theta_0}^\star(V^1, O^{0,j\star}) \Big)^2 \Big] \\ &\qquad + \Big[q_0 \bar{q}_0(V^1) \dot{\ell}_{\theta_0}^\star(V^1, O^{1\star}) \Big(\tfrac{1}{J} \textstyle\sum_j \bar{q}_0(O^{0,j\star}) \dot{\ell}_{\theta_0}^\star(V^1, O^{0,j\star}) \Big)^\top \\ &\qquad\qquad + q_0 \bar{q}_0(V^1) \Big(\tfrac{1}{J} \textstyle\sum_j \bar{q}_0(O^{0,j\star}) \dot{\ell}_{\theta_0}^\star(O^{0,j\star}) \Big) \dot{\ell}_{\theta_0}^\star(O^{1\star})^\top \Big]. \end{aligned}$$

The P_0-expected value of the first term between brackets is $E_{P_0^\star}\ddot{\ell}^\star_{\theta_0}(O^\star)$, as another simple adaptation of the proof of Proposition 1 straightforwardly yields. Moreover,

$$\begin{aligned}
&E_{P_0} q_0 \bar{q}_0(V^1)\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top \\
&\quad = E_{P_0}\left[q_0\bar{q}_0(V^1)E_{P_0}\left(\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top\middle| V^1\right)\right] \\
&\quad = E_{P_0}\left[q_0\bar{q}_0(V^1)E_{P_0}\left(\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})\middle| V^1\right)\right. \\
&\qquad\qquad \left. E_{P_0}\left(\Big(\tfrac{1}{J}\textstyle\sum_j \bar{q}_0(O^{0,j\star})\dot{\ell}^\star_{\theta_0}(V^1,O^{0,j\star})\Big)^\top\middle| V^1\right)\right]
\end{aligned}$$

by conditional independence. Denote by $\Pi = E_{P_0}(\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})|O^{1\star}\setminus Z^1)$ the conditional expectation of $\dot{\ell}^\star_{\theta_0}(V^1,O^{1\star})$ given every component of $O^{1\star}$ but Z^1, that is given Ω^1 (compatible with V^1). The projection Π can be written as a measurable function of Ω^1 times

$$\int \dot{\ell}^\star_{\theta_0}(z,\Omega^1)p^\star_{\theta_0}(z|\Omega^1)dz = \int \left.\frac{\partial p^\star_\theta(z|\Omega^1)}{\partial\theta}\right|_{\theta=\theta_0} dz = 0,$$

provided that the order of differentiation and integration can be reversed. This is ensured by the stated constraint on the derivatives of $p^\star_\theta(z|\Omega^1)$ with respect to θ. Consequently, the P_0-expected value of the second term between brackets in the first display is zero, hence the validity of the alternative version of Σ. The conclusion simply follows from another application of [van der Vaart, 1998, Theorem 5.23] in the classical iid framework associated with $P_1^\star$. □

C Application

C.1 Conditional distribution of being a case

Estimating the joint distribution of (V,Y) hence q_0 (4), $(q_0(1|v))_{v\in\mathcal{V}}$ (5), and $(\bar{q}_0(v))_{v\in\mathcal{V}}$ (7) is made possible thanks to [Belot et al., 2008], an independent study of cancer incidence and mortality in France over the period 1980–2005 (for the conditional distribution of Y given V), and on data made publicly available by the French National Institute of Statistics and Economic studies (for the marginal distribution of V, see `http://www.insee.fr/en/`). However, we must assume either *(i)* that the data from [Belot et al., 2008], which are collected over the whole French population, are representative of the Parisian population of interest, or *(ii)* that sampling from the four Parisian hospitals that participate to the study is stochastically equivalent to sampling from the population of France.

We first estimate these quantities for each year from 1999 to 2002 separately. In agreement with our stationarity assumption (1), we remark that the various estimates are very consistent over the years. In order to gain precision, we average the estimates over the years. The final estimates are presented in Table 5. We emphasize that the weights $(\bar{q}_0(v))_{v\in\mathcal{V}}$ are far from being homogeneous.

C.2 Multi-fold likelihood-based cross-validation

The likelihood risk of $\theta\in\Theta$ is by definition

$$\mathcal{R}(\theta) = -E_{P_0}\tilde{\ell}(\theta)(O),$$

which is closely related to minus the Kullback-Leibler divergence between the density $p_0^\star$ of $P_0^\star$ and $p_\theta^\star$, as explained in Section 3. Let us denote by $\theta_{n,k}(P_n)$ the case-control weighted maximum likelihood estimator defined in (10) with θ ranging over Θ_k. Given the collection $\{\theta_{n,k}(P_n) : k\in\mathcal{K}\}$ we wish to select the estimator $\theta_{n,\bar{k}}(P_n)$ that minimizes $\mathcal{R}$, where $\bar{k}$ itself depends on P_n. Because

$q_0 = 470.0682\text{e-}06$

a	$q_0(1\|0,a)$	$q_0(1\|1,a)$
1	2.058932e-06	1.663324e-06
2	1.859944e-05	1.460473e-05
3	6.803086e-05	4.461827e-05
4	2.586692e-04	1.184914e-04
5	6.484864e-04	1.947058e-04
6	1.192778e-03	2.542976e-04
7	1.854668e-03	3.294062e-04
8	2.331553e-03	3.588764e-04
9	2.928415e-03	4.466062e-04
10	3.686216e-03	5.312313e-04
11	3.608302e-03	5.332930e-04
12	3.636995e-03	5.395069e-04
13	2.171286e-03	3.234775e-04

a	$\bar{q}_0(0,a)$	$\bar{q}_0(1,a)$
1	228.3063171	282.6071669
2	25.2727716	32.1855444
3	6.9091613	10.5348607
4	1.8167860	3.9666376
5	0.7243998	2.4137787
6	0.3936251	1.8480261
7	0.2529813	1.4265470
8	0.2011416	1.3093632
9	0.1600496	1.0520638
10	0.1270504	0.8843954
11	0.1298040	0.8809745
12	0.1287763	0.8708223
13	0.2160229	1.4527010

Table 5: Estimating the probability distribution of being a case, based on the independent study [Belot et al., 2008]. Left: Estimates of $q_0(1|w_1, v_2)$, as defined in (5). Middle: Estimate of q_0, as defined in (4). Right: Estimates of $\bar{q}_0(w_1, v_2)$, as defined in (7). Here, $w_1 = 0$ for men and $w_1 = 1$ for women, and $v_2 = a$ if the age at sampling x belongs to $[t_a; t_{a+1})$, where $t_0 = 0$, $t_a = 30 + 5(a-1)$ for $1 \leq a \leq 12$ and $t_{13} = \infty$.

the definition of $\mathcal{R}$ involves the true distribution P_0, we must estimate $\mathcal{R}(\theta_{n,k}(P_n))$ and choose to do so by multi-fold cross-validation. Details follow.

We split the data randomly into a *training* and a *validation* samples. Given an integer V (later set to $V = 10$), each observed data structure O_i is associated with a label $\text{lab}_i = 1 + (i \mod V)$. The collection of labels $\{\text{lab}_i : i \leq n\} \subset \{1, \ldots, V\}$ is such that $\max_{l,l' \leq V} |\sum_{i=1}^n \mathbf{1}\{\text{lab}_i = l\} - \sum_{i=1}^n \mathbf{1}\{\text{lab}_i = l'\}| \leq 1$. The splitting random variable $S = (S_1, \ldots, S_n) \in \{0,1\}^n$ is drawn independently of $O_1, \ldots, O_n$ in such a way that, for each $1 \leq l \leq V$, $S = (\mathbf{1}\{\text{lab}_1 = l\}, \ldots, \mathbf{1}\{\text{lab}_n = l\})$ with probability V^{-1}. Conditionally on S, the observed data structure O_i belongs to the training sample if $S_i = 0$ (there are approximately $n(V-1)/V$ such O_i's), otherwise it belongs to the validation sample. The empirical distribution of those O_i's for which $S_i = 0$ (respectively, $S_i = 1$) is $P^0_{n,S}$ (respectively, $P^1_{n,S}$). The empirical distribution of those O_i's for which $\text{lab}_i = l$ (respectively, $\text{lab}_i \neq l$) is P^l_n (respectively, P^{-l}_n).

Each Θ_k yields a maximum likelihood estimator $\theta_{n,k}(P^0_{n,S})$ based on the training sample only. Its risk, averaged over the splits, writes as

$$\text{crit}(k) = E_S \mathcal{R}(\theta_{n,k}(P^0_{n,S})) = -\frac{1}{V} \sum_{l=1}^{V} E_{P_0} \tilde{\ell}\left(\theta_{n,k}(P^{-l}_n)(O)\right).$$

The value $\tilde{k}$ that minimizes $k \mapsto \text{crit}(k)$ over $\mathcal{K}$ is called the *oracle* because it depends both on P_n and on P_0. In our attempt to reach that $\tilde{k}$ which is a good proxy to $\bar{k}$, we estimate $\text{crit}(k)$ by

$$\widehat{\text{crit}}(k) = -E_S E_{P^1_{n,S}} \tilde{\ell}\left(\theta_{n,k}(P^0_{n,S})(O)\right) = -\frac{1}{V} \sum_{l=1}^{V} E_{P^l_n} \tilde{\ell}\left(\theta_{n,k}(P^{-l}_n)(O)\right),$$

and propose to use the value $\hat{k}$ that minimizes $k \mapsto \widehat{\text{crit}}(k)$ over $\mathcal{K}$, whose definition is postponed to Section C.3. In conclusion, the final estimator is $\theta_{n,\hat{k}}(P_n)$.

C.3 Model selection procedure in action

We explain in Section C.2 how the best model index $\hat{k}$ (with related best model $\{P^\star_\theta : \theta \in \Theta_{\hat{k}}\}$) is obtained in a pre-determined collection $\mathcal{K}$ of sub-model indices (with related sub-models $\{P^\star_\theta :$

$\theta \in \Theta_k\}$, $k \in \mathcal{K}$). The latter collections are constructed by recursion as presented below.

We first initialize $\Theta^0 = \Theta$ and $\mathcal{K}^{-1} = \emptyset$ with convention $\max \emptyset = 0$.

At a given step $\nu \geq 0$, a sub-model Θ^ν is defined as a subset of Θ meeting ν independent one-dimensional constraints on $M \in \mathcal{M}$ (*i.e.*, constraints of the type $M_{k,l-1} = M_{k,l}$ for some $k = 1, 2, 3$ and $l = 1, 2, 3$ with convention $M_{k,0} = 0$). Start with $c(\nu + 1) = -\infty$ and $\Theta^{\nu+1} = \emptyset$. The following rule is applied to $\Theta^{\nu+1}$:

Rule 1. For every possible $\Theta' \subset \Theta^\nu$ derived from Θ^ν by adding another one-dimensional constraint on M as described above (all such models share the same dimension), evaluate the corresponding maximum log-likelihood criterion

$$\ell(\Theta') = \max_{\theta \in \Theta'} P_n \tilde{\ell}(\theta).$$

If $\ell(\Theta') \geq c(\nu + 1)$, update $c(\nu + 1) = \ell(\Theta')$ and $\Theta^{\nu+1} = \Theta'$.

Applying Rule 1 as long as possible yields 7 sets Θ^ν, $\nu = 0, \ldots, 6$. Their description is given in Table 6.

$$\begin{aligned}
\Theta^0 &= \Theta, \\
\Theta^1 &= \{\theta \in \Theta^0 : M_{1,1} = 0\}, && (27)\\
\Theta^2 &= \{\theta \in \Theta^1 : M_{2,1} = M_{2,2}\}, && (28)\\
\Theta^3 &= \{\theta \in \Theta^2 : M_{2,2} = 1\}, && (29)\\
\Theta^4 &= \{\theta \in \Theta^3 : M_{3,1} = M_{3,2}\}, && (30)\\
\Theta^5 &= \{\theta \in \Theta^4 : M_{1,2} = 1\}, && (31)\\
\Theta^6 &= \{\theta \in \Theta^5 : M_{3,1} = 0\}. && (32)
\end{aligned}$$

(27) low probability does not differ from no exposure at all;
(28) moreover, low and mild frequencies do not differ;
(29) moreover, mild and high frequencies do not differ;
(30) moreover, low and mild intensities do not differ;
(31) moreover, mild and high probabilities do not differ;
(32) moreover, low intensity does not differ from no exposure.

Table 6: Descriptions of $\Theta^0, \ldots, \Theta^6$. The collection of parameter sets is nested. For instance, Θ^3 is the set of those $\theta \in \Theta$ such that $M_{1,1} = 0$ and $M_{2,1} = M_{2,2} = 1$. Regarding dimensions, it trivially holds that, for each $0 \leq k \leq 6$, $\dim(\Theta^k) = 27 - k$.

At a given step $\nu \geq 0$, a set $\mathcal{K}^{\nu-1}$ of successive integers is defined. Start with $\mathcal{K}^\nu = \{\max K^{\nu-1} + 1\}$ (a set initially containing a single element) and define $\Theta^{\nu, \max K^{\nu-1}+1} = \Theta^\nu$. The following second rule is applied to $\mathcal{K}^\nu$:

Rule 2. For every possible constraint "$\varphi(\theta) = 0$" on $\theta \in \Theta^\nu$ of the form "α and β independent of W_l" for some $l = 1, 2, 3$ (the lth coordinate of W does not affect the value of the initial health and drift parameters h and μ, see (19) and (20)), update $\mathcal{K}^\nu = \mathcal{K}^\nu \cup \{\max K^\nu + 1\}$ and define $\Theta^{\nu, \max K^\nu} = \{\theta \in \Theta^\nu : \varphi(\theta) = 0\}$.

(Note that each Θ^ν therefore gives rise to $2^3 = 8$ sets $\Theta^{\nu,l}$.)

We apply Rule 2 for $\nu = 0, \ldots, 6$, and finally define

$$\mathcal{K} = \cup_{\nu=0}^{11} \mathcal{K}^\nu = \{1, 2, 3, \ldots, 56\}.$$

For every $k \in \mathcal{K}$ there exists a unique $\nu = 0, \ldots, 7$ such that $\Theta^{\nu,k}$ is defined: setting $\Theta_k = \Theta^{\nu,k}$ concludes the definition of the collection $\{\Theta_k : k \in \mathcal{K}\}$ of interest.

The best model $\{P_\theta^\star : \theta \in \Theta_{\hat{k}}\}$ (according to our multi-fold likelihood-based cross-validation criterion) is a subset of $\{P_\theta^\star : \theta \in \Theta^2\}$, featuring 16 degrees of freedom. Its complete description follows:

- the *initial health* parameter depends on W only through gender (hence not on the indicator of occurrence of lung cancer in close family);
- the *drift* parameter depends on W only through gender and lifetime tobacco use (hence not on the indicator of occurrence of lung cancer in close family);
- *exposure to asbestos* is significantly noxious; there is no difference between low probability and no exposure to asbestos at all (in view of (16), $M_{1,1} = 0$) and no difference either between low and mild frequencies (in terms of (16), $M_{2,1} = M_{2,2}$).

Comment on implementation. Our implementation of the model selection procedure relies on an algorithmic trick. The main difficulty arises from the combination of the following two facts:

- in order to fit a given model $\{P^\star_\theta : \theta \in \Theta'\}$ corresponding to a subset $\Theta' \subset \Theta$, it is necessary to provide the optimization algorithm with the set of constraints which characterize Θ';
- given the huge number of sub-models, it is out of question to prepare beforehand all sets of constraints for all sub-models that we may have to fit in the course of the model selection procedure.

Hence we create an algorithm which maps an explicit description of any $\Theta' \subset \Theta$ onto the corresponding set of constraints required by the optimization procedure.

References

A. Belot, P. Grosclaude, N. Bossard, E. Jougla, E. Benhamou, P. Delafosse, A. V. Guizard, F. Molinié, A. Danzon, S. Bara, A. M. Bouvier, B. Trétarre, F. Binder-Foucard, M. Colonna, L. Daubisse, G. Hédelin, G. Launoy, N. Le Stang, M. Maynadié, A. Monnereau, X. Troussard, J. Faivre, A. Collignon, I. Janoray, P. Arveux, A. Buemi, N. Raverdy, C. Schvartz, M. Bovet, L. Chérié-Challine, J. Estève, L. Remontet, and M. Velten. Cancer incidence and mortality in France over the period 1980-2005. *Rev. Epidemiol. Santé Publique*, 56(3), 2008. Detailed results and comments [online] `http://www.invs.sante.fr/surveillance/cancers/estimations_cancers/default.htm`.

H. K. Biesalski, B. B. de Mesquita, A. Chesson, F. Chytil, R. Grimble, R. J. Hermus, J. Kohrle, R. Lotan, K. Norpoth, U. Pastorino, and D. Thurnham. European consensus statement on lung cancer: risk factors and prevention. Lung cancer panel. *CA Cancer J. Clin.*, 48(3):167–176, 1998.

N. E. Breslow. Statistics in epidemiology: the case-control study. *J. Amer. Statist. Assoc.*, 91 (433):14–28, 1996.

R. S. Chhikara and J. L. Folks. *The inverse Gaussian distribution: theory, methods and applications*. Marcel Dekker: New-York, 1989.

T. Duchesne and J. S. Rosenthal. On the collapsibility of lifetime regression models. *Adv. in Appl. Probab.*, 35(3):755–772, 2003.

P. Gustavsson, F. Nyberg, G Pershagen, P Schéele, R. Jakobsson, and N. Plato. Low-dose exposure to asbestos and lung cancer: Dose-response relations and interaction with smoking in a population-based case-referent study in stockholm, sweden. *Am. J. Epidemiol.*, 155(11):1016–1022, 2002.

C. A. Haiman, D. O. Stram, L. R. Wilkens, M. C. Pike, L. N. Kolonel, B. E. Henderson, and L. Le Marchand. Ethnic and racial differences in the smoking-related risk of lung cancer. *N. Engl. J. Med.*, 354(4):333–342, 2006.

IARC. *IARC monographs on the evaluation of the carcinogenic risk of chemicals to man: asbestos*, volume 14. IARC, 1977.

M.-L. T. Lee and G. A. Whitmore. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statist. Sci.*, 21(4):501–513, 2006.

M-L. T. Lee and G. A. Whitmore. Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Anal.*, 16(2):196–214, 2010.

M-L. T. Lee, G. A. Whitmore, F. Laden, J. E. Hart, and E. Garshick. A case-control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model. *J. Stat. Plann. Inference*, 139(5):1633–1642, 2009.

D. Oakes. Multiple time scales in survival analysis. *Lifetime Data Anal.*, 1(1):7–18, 1995.

J-C. Pairon, B. Legal-Régis, J. Ameille, J-M. Brechot, B. Lebeau, D. Valeyre, I. Monnet, M. Matrat, and B. Chamming's, S. Housset. Occupational lung cancer: a multicentric case-control study in Paris area. European Respiratory Society, 19th Annual Congress, Vienna, 2009.

J. Robins and S. Greenland. Estimability and estimation of expected years of life lost due to a hazardous exposure. *Stat. Med.*, 10(1):79–93, 1991.

S. Rose and M. J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *Int. J. Biostat.*, 4:Art. 19, 24, 2008.

A. Ruano-Ravina, A. Figueiras, A. Montes-Martinez, and J-M. Barros-Dios. Dose-response relationship between tobacco and lung cancer: new findings. *Eur. J. Cancer Prev.*, 12(4):257–263, 2003.

Mark J. van der Laan. Estimation based on case-control designs with known prevalence probability. *Int. J. Biostat.*, 4:Art. 17, 58, 2008.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371, 2006.

E. A. Zang and Wynder. E. L. Differences in lung cancer risk between men and women: examination of the evidence. *J. Natl. Cancer Inst.*, 88(3-4):183–192, 1996.