Chapter 1
Acceleration, due to occupational exposure, of time to onset of a disease

A. Chambaz, D. Choudat, C. Huber

Abstract Occupational exposure to pollution may accelerate or even induce the onset of specific diseases. A pecuniary compensation, to be paid by the state or the company, is then due to the exposed worker. The computation of the amount of this compensation is based on the so-called "expected number of years of disease-free life" lost by the worker due to their occupational exposure. In order to estimate this number of years, we propose a method based on the threshold regression, also known as first hitting time model (FHT model). This model was initially developed for cohort studies. As our motivating example is a case-control study conducted in France to evaluate the link between lung cancer occurrence and occupational exposure to asbestos, we define a FHT model, adapt it to case-control data, and finally derive, for each worker in the study, the estimated expected number of disease-free years lost due to their occupational exposure to asbestos.

1.1 Introduction

Quality of life is a central concern in medicine. Thus, the challenge of defining and estimating the expected number of years of disease-free life lost due to an occupational exposure frequently arises for the sake of characterizing the amount of a pecuniary compensation due to a worker, on a case-by-case basis. Our motivating

A. Chambaz
Modal'X, Université Paris Ouest Nanterre, 92001 Nanterre, France, e-mail: achambaz@u-paris10.fr,

D. Choudat
Département de médecine du travail, Assistance Publique-Hôpitaux de Paris, Université Paris Descartes, Sorbonne Paris Cité, 75014 Paris, France,e-mail: dominique.choudat@parisdescartes.fr

C. Huber
MAP5, Université Paris Descartes, Sorbonne Paris Cité, 75270 Paris Cedex 06, France, e-mail: catherine.huber@parisdescartes.fr

example is a French case-control study on the occurrence of lung cancer for workers exposed to asbestos (J.C. Pairon et al, 2009). As our objective is not to evaluate the risk of developing the disease, the logistic model, which is typically used for case-control studies, would not be the proper choice here. Several other models, though, can be used in order to solve our problem. One of the simplest is the Cox model involving the occupational exposure as a covariate together with other risk factors that could also induce lung cancer, like family history of cancer and tobacco consumption. However we choose to adapt the threshold regression model initially developed for cohort studies (M.L.T. Lee and G.A. Whitmore, 2006) to the case control study. This model allows us to deal with the occupational exposure as an accelerator of the time leading possibly to a quicker onset of the disease, while the other covariates are divided into two classes, depending on how they act on the time to onset: the initial factors, like gender, past family history and genetic factors, and the life long ones, like biological and lifestyle covariates. The expected number of disease-free years lost due to occupational exposure to asbestos is then derived from the model by replacing for each exposed subject his or her time to onset by the decelerated time he or she would have had when exposure is removed and all other factors in the model remain the same (A. Chambaz et al, 2013).

1.2 Motivation of the choice of FHT model

1.2.1 Preliminary studies

We start with a preliminary non parametric study of the data. The Kaplan-Meier estimators of the survival functions of subsets of the data based on high or low occupational exposures may show a possible influence of the amount of occupational exposures and other factors.
Then, we could consider a Cox model that puts all covariates $Z = (Z_1, \cdots, Z_k)$ including exposure covariates as well as initial and life long covariates alike, on the same level. The model reads

$$\lambda(t|Z) = \lambda_0(t) \times \exp(< \theta, Z >) \tag{1.1}$$

where λ is the incidence rate at time t, λ_0 a baseline incidence rate and θ is a k-dimensional real parameter.
But, considering that, actually, the covariates are not all of the same kind, we choose a FHT model that enables us to separate the covariates into three different kinds based on their actions on the health status of the patient.

1.2.2 FHT model

When estimating the influence of an occupational exposure on the onset of a disease, three different types of covariates are distinguished by considering how they act on (or account for), the decrease of the latent "amount of health". Specifically, the three types of covariates are as follows
- the initial covariates which act on the initial amount of health of the patient, including genetic factors, gender and past family disease history;
- the life long covariates which act on (or account for) the "decrease" of the initial amount of health, including biological, environmental and lifestyle covariates such as, for example, cholesterol level and tobacco consumption;
- the occupational exposure under study which may accelerate the time to onset of the considered disease.

The time T to occurrence of the disease is modeled by a stochastic process $X(t)$ which represents the amount of health of the subject at time t: the disease occurs when this amount of health hits the boundary 0 for the first time (hence the expression FHT). Let B be a Brownian motion. For any real numbers $h > 0$ and $\mu \leq 0$, the process $X(t)$ is defined as

$$X(t) = h + \mu t + B(t) \tag{1.2}$$

where h plays the role of an initial amount of health relative to the disease, and μ a rate of decay of the amount of health. The value of h depends on the initial covariates while the value of μ depends also on the life long covariates. Then

$$T(h,\mu) = \inf\{t \geq 0 : X(t) \leq 0\}, \tag{1.3}$$

is the first time the drifted Brownian motion $X(t)$ hits 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)

$$F(t|h,\mu) = 1 + e^{-2h\mu}\Phi\left((\mu t - h)t^{-1/2}\right) - \Phi\left((\mu t + h)t^{-1/2}\right), \tag{1.4}$$

where Φ is the standard normal cdf.

As $\mu \leq 0$, the drifted Brownian motion $X(t)$ will almost surely reach the boundary (i.e. $T(h,\mu) < \infty$). Therefore $T(h,\mu)$ is also characterized by its density

$$f(t|h,\mu) = \frac{h}{(2\pi t^3)^{1/2}} \exp\left(-\frac{(h - |\mu| t)^2}{2t}\right). \tag{1.5}$$

$T(h,\mu)$ has mean $h/|\mu|$ whenever $\mu < 0$.

The effect of occupational exposure is taken into account through an acceleration function R that is nondecreasing and continuous on $\mathbb{R}^+$ such that $R(t) \geq t$ for all t. The acceleration function R depends on the occupational exposure and, given R, we define

$$T(h,\mu,R) = \inf\{t \geq 0 : h + \mu R(t) + B(R(t)) \leq 0\}, \tag{1.6}$$

the first time the drifted Brownian motion $(X(B(R(t))))$ hits 0 along the modified time scale derived from R, so that the cdf of $T(h,\mu,R)$ at t is $F(R(t);h,\mu)$, and its density at t is $R'(t)f(R(t);h,\mu)$ as long as R is differentiable.
Conditional on $[T \geq x-1]$, the survival function and density of T at $t \geq x-1$ are respectively:

$$G(t|h,\mu,R) = \frac{1-F(R(t)|h,\mu)}{1-F(R(x-1)|h,\mu)} \ , \tag{1.7}$$

$$g(t|h,\mu,R) = \frac{R'(t)f(R(t)|h\mu)}{1-F(R(x-1)|h,\mu)} \ . \tag{1.8}$$

1.3 The data set

1.3.1 Description of the data set

The matched case-control study took place between 1999 and 2002 at four Parisian hospitals and consisted of $n = 1761$ patients, among which 860 were cases and 901 were controls. The non-occupational information on each patient comprised six covariates, the hospital, $W_0 \in \{1,2,3,4\}$, the gender $W_1 \in \{0,1\}$ (0 for men, 1 for women), the occurrence of lung cancer in close family, $W_2 \in \{0,1\}$, 1 for occurrence, and the tobacco consumption: $W_3 \in \{0,1,2,3\}$ respectively for pack-year $\in \{0,[1\ ;25],[26\ ;45],>45\}$, the age at interview $X(\tau)$ where τ is calendar time, the age at incidence of lung cancer T, with convention $T=\infty$ if no lung cancer occurred yet. The indicator of a case, equal to 1 for cases and 0 for controls, is thus

$$Y = 1\{T \leq X\}.$$

Matching was done based on hospital, gender and age at recruitment ± 2.5years. In the sequel, we denote

$$V = (W_0, W_1, X)$$

the matching variable.

The other items observed on the patients deal with informations on occupational exposure up to the time of interview. The occupational history up to age X is measured on each of the successive jobs by its duration and three indicators of the exposure to asbestos: its probability, frequency and intensity of exposure, each with 3 levels $(1,2,3)$. A probability index equal to $1,2$ or 3 corresponds respectively to a passive exposure, a possible direct exposure or a very likely or certain direct exposure. A frequency index equal to $1,2$ or 3 corresponds respectively 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. An intensity index equal to $1,2$ or 3 corresponds respectively to a concentration of asbestos fibers less than 0.1 f/mL, between 0.1 and 1 f/mL and more than 1

Table 1.1 Number of jobs for each possible "probability/frequency/intensity" description.

exposure	count	exposure	count	exposure	count
111	213	211	53	311	138
112	167	212	6	312	105
113	3	213	6	313	24
121	150	221	5	321	136
122	46	222	3	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

f/mL. Adding a category $0 = (0,0,0)$ for no exposure at all, the set $\mathscr{E}$ of categories of exposure has 27+1=28 elements $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$.

Among the 8432 jobs held by the participants, 7009 were without any significant exposure. But although this leaves 1423 jobs featuring a significant exposure, it can be seen in table 1.1, which contains many 0's, that several profiles in $\mathscr{E}$ are not represented, which gives evidence of an over-parametrization.
Let $a_i(t)$ be the exposure of subject i at time t, $\widetilde{a}_i(t) = a_i|_0^t$ be the exposure from time 0 to time t of subject i and $\widetilde{a}$ their history of exposure along their lifetime up to the occurrence of cancer if he or she is a case or to the time of interview if he or she is a control. The function $\widetilde{a}$ is piecewise constant. For example, let subject i be a patient who had his or her first job at age 20 during 15 years with an occupational exposure $\varepsilon = (2,1,3)$. Then he or she had a second job during 10 years with an occupational exposure $\varepsilon = (2,2,3)$. Finally, he or she is diagnosed with lung cancer at age $T_i = 45$, at which point he or she becomes a case. Then, for $0 < t < 20$, $a_i(t) = 0$, for $20 \leq t < 35$, $a_i(t) = (2,1,3)$, and for $t \geq 35$, $a_i(t) = (2,2,3)$; moreover $\widetilde{a}_i(30) = a_i|_0^{30}$, and $\widetilde{a}_i = a_i|_0^{45}$.

1.3.2 Specific problems due to the data set

Several problems arise due to the way the data set was collected:

First of all, the data set contained information pertaining to occupational exposures, like silica and aromatic hydrocarbons. However, the preliminary non parametric analysis using Kaplan-Meier estimates for sub-samples of the data set revealed that these two factors did not have much influence on the age at onset. Moreover, very few people were exposed to silica and/or aromatic hydrocarbons and in very small quantities. Thus we decided to restrict our attention to asbestos.

Second, the actual matching pattern is not available. What is known is only on which covariates the pairing was done. Instead of giving up on the matching, we choose to artificially determine a valid matching pattern, and also make sure that

our results are preserved when using several different valid patterns.

Third, the FHT model, initially developed for cohort data, has to be adapted to a case-control survey via a weighing of the log-likelihood.

Fourth and finally, we have already emphasized that the current model is over-parametrized. We tackle the issue of reducing the excessive number of parameters in a sensible way as explained in section (1.4.2).

1.4 Data analysis

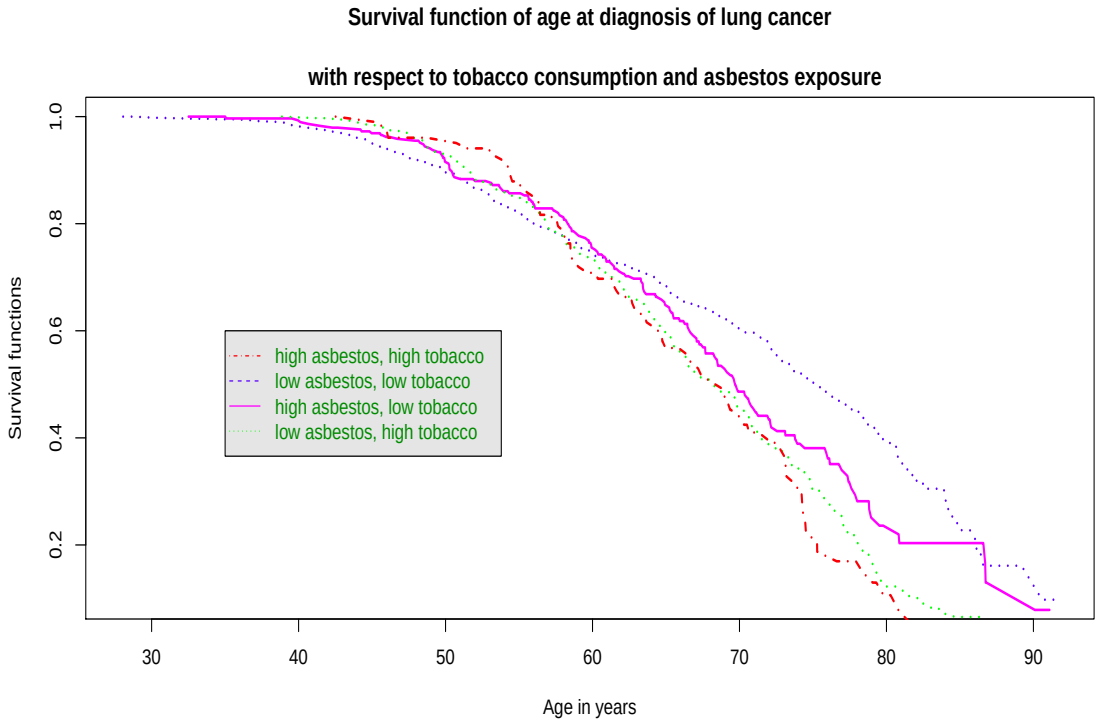
1.4.1 Preliminary studies

We first apply the non parametric Kaplan-Meier method to estimate the survival functions of four sub-samples having low/high tobacco consumption and low/high asbestos exposure. The cutting points are the respective medians. We obtain the corresponding survival functions for the time to onset of the lung cancer (see figure 1.1). We see from this figure that there is probably an impact on lung cancer occurrence of both tobacco consumption and asbestos exposure. Applying the same process to the other exposures present in the data set, like silica and aromatic hydrocarbons, gives no such evidence. This may be due to the fact that there are very few jobs featuring a significant exposure to silica or aromatic hydrocarbons.
A naive application of an FHT model to those data would consist in defining $\log(h)$ as a linear function of gender (W_1) and past family history of lung cancer (W_3), μ as a linear function of (W_1), (W_3)and also tobacco consumption (W_2) and the acceleration $R(t)$ as $\sum_{j=1}^{J} m_j \times a_j(t)$ for a patient having experienced J jobs, where m_j is the acceleration parameter attached to the category ε_j. The dimension of this naive model equals $3+4+28=35$, and the underlying assumption of linearity of $\log(h)$ and μ seems quite restrictive. In contrast, we build a more general though slightly more economic FHT model of dimension 27, and consider it as a maximal model containing simpler models among which we select, based on our data, a better model (1.4.4).

1.4.2 Acceleration due to occupational exposure

First, a class of acceleration functions tailored to our description of occupational exposures is built. The acceleration depends on three variables,ε_1, ε_2 and ε_3). We replace this function of three variables, each of them having three values, by the product of three functions of one variable, $M_1(\varepsilon_1)$, $M_2(\varepsilon_2)$ and $M_3(\varepsilon_3)$. Each of these three functions, M_1 for probability, M_2 for frequency and M_3 for intensity is

Fig. 1.1 Survival functions of age at diagnosis of lung cancer for high or low tobacco consumption and exposure to asbestos



assumed to be non negative, non decreasing and having 1 as maximum value. In this view, define

$$\mathscr{M} = \Big\{ (M_0, (M_k(l))_{k,l\leq 3} \in \mathbb{R}^+ \times (\mathbb{R}^+)^{2\times 3} \\ 0 \leq M_k(1) \leq M_k(2) \leq M_k(3) = 1, k = 1,2,3 \Big\}. \quad (1.9)$$

Then the rate yielded by description $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathscr{E} \setminus \{0\}$ is expressed as

$$M(\varepsilon) = 1 + M_0 \times M_1(\varepsilon_1) \times M_2(\varepsilon_2) \times M_3(\varepsilon_3)$$

with convention $M(0) = 1$. Note that $M(0) = 1 \leq M(\varepsilon) \leq M(3,3,3) = 1 + M_0$. Exposure $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ can be written as a fraction M_ε of the maximal acceleration, where

$$M_\varepsilon = \frac{M(\varepsilon) - 1}{M_0} = M_1(\varepsilon_1) \times M_2(\varepsilon_2) \times M_3(\varepsilon_3).$$

This parametrization is identifiable and reduces the number of parameters needed to associate every category of exposure with an acceleration rate, from 28 to 7.
Set $M \in \mathscr{M}$ and a generic longitudinal description $\widetilde{a}$ be as presented in Section 1.3. For convenience, we consider a continuous approximation to the piecewise constant function $t \mapsto M(\varepsilon(t))$, which we denote by $r(M,\widetilde{a})$ (see A. Chambaz et al (2013) for details). Then every pair $(M,\widetilde{a})$ thus gives rise to the nondecreasing and differentiable acceleration function

$$R(M,\widetilde{a})(t) = \int_0^t r(M,\widetilde{a})(s)ds \leq t. \tag{1.10}$$

1.4.3 The case-control weighed log*-likelihood*

Let us recapitulate the parametrization of our FHT model:

$$\begin{array}{lll} \log(h) & = \alpha(W_1,W_2) & \in \mathbb{R}^4, \\ \log(-\mu) & = \beta(W_1,W_2,W_3) & \in \mathbb{R}^{16}, \\ R & = R(M,\widetilde{A}(X)) \ , & M \in \mathscr{M}, \\ \theta & = (\alpha,\beta,M) & \in \Theta. \end{array}$$

Recall that V and Y are respectively the matching variable and the case indicator, and define $Z = min(T,X)$. We rely on weights whose characterization requires the prior knowledge of the joint probability of (V,Y) which implies the knowledge of the conditional probabilities

$$\begin{array}{l} q_v^*(y) := P(Y = y|V = v), \\ q_y(v) := P(V = v|Y = y)). \end{array}$$

Recall the definitions of G and g from (1.7) and (1.8), and let case i be matched by J_i controls. Then the weighed log-likelihood is

$$loglik(\theta) = \sum_{i=1}^n \left\{ q_1(V_i) \log g(Z_i|\theta) + q_0(V_i) \frac{1}{J_i} \sum_{j=1}^{J_i} \log G(Z_i|\theta) \right\}.$$

Asymptotic properties of the resulting estimators are derived in (A. Chambaz et al, 2013).

1.4.4 Model selection by cross validation

The maximal model Θ gives rise to a collection of sub-models Θ_k obtained by adding constraints on the maximal parameter $\theta = (\alpha,\beta,M) \in \Theta$. We define a large collection $\{\Theta_k : k \in \mathscr{K}\}$ of sub-models of interest. Then we let the data select a better sub-model $\Theta_{\hat{k}}$ based on a multi-fold likelihood cross validation criterion. The sample is divided into ten sub-samples of equal size. The initialization goes like this:

in turn, we exclude one of the ten sub-samples and use the nine others to estimate the parameter of the maximal model, then we compute the likelihood of the tenth sub-sample under the estimated value of the parameter. The average likelihood, L_0, is finally computed. We then consider all one-step sub-models obtained by excluding W_1 or W_2 from the parametrization of h and μ, or by putting additional constraints on M. We compute their cross-validated scores, say L_1, in the same manner as L_0 was computed. If one sub-model at least satisfies $L_1 - L_0 > g$ for some pre-specified $g > 0$, then the model yielding the largest increase is selected. The above process is repeated with the best sub-model in place of the maximal model, until no gain is observed of if the gain is no larger than $c \times g$ for some pre-specified c.

The final model features that there are no constraints either on "frequency" or on "intensity", but on "probability", and the conclusion is that when $\varepsilon_1 = 1$, there is no effect of exposure and that there is no difference between $\varepsilon_1 = 2$ and $\varepsilon_1 = 3$.

1.4.5 Model estimation

First, we fit the best model by maximum likelihood on the whole data set. Then, we derive confidence intervals through percentile bootstrap with Bonferroni correction: B=1000, on 95% of the sample repeatedly re-sampled from the 860 cases together with the corresponding controls.

Table 1.2 Confidence intervals for the initial health h as a function of gender W_1 (0 for men)

W_1	h	h_{min}	h_{max}
0	23.82	23.42	24.13
1	25.09	24.86	25.40

Table 1.3 Confidence intervals for drift = - 100μ as a function of gender, W_1 and tobacco, W_3

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

Table 1.4 Confidence intervals for acceleration parameters

$M_0 = 1.19$	CI = [0.34 2.00]	
$M_1(1) = 0$		
$M_1(2) = 0.97$	CI = [0.96 0.99]	$M_1(3) = 1$
$M_2(1) = M_2(2)$		
$M_2(2) = 0.93$	CI = [0.90 0.98]	$M_2(3) = 1$
$M_3(1) = 0.02$	CI = [0.00 0.09]	
$M_3(2) = 0.09$	CI = [0.00 0.27	$M_3(3) = 1$

Table 1.5 Six examples of expected number of years free of lung cancer lost due to occupational asbestos exposure.

sex	age	asbestos	family	tobacco	years lost
0	65	228	0	1	3.1
0	57	125	0	1	2.5
0	60	25	0	1	2.7
1	41	36.0	0	1	1.6
0	66	24.0	1	1	3.0
1	61	78.0	0	0	3.4

1.5 Conclusion

This method allows us to derive, for each patient, the expected number of disease-free years of life due to occupational exposure in a simple way: once the model is estimated, the expected number of life free of lung cancer lost due to asbestos exposure of any patient i may be computed by decelerating their time T_i to onset of lung cancer by its estimated acceleration. The difference between the decelerated time and the observed time T_i is an estimation of the expected number of years free of lung cancer due to asbestos exposure. Denoting $\widehat{R}_i(t)$ the estimated acceleration function for subject i, the estimated expected number of disease-free years of life lost by him or her, denoted L_i, due to their occupational exposure is

$$L_i = \widehat{R}_i^{-1}(T_i) - T_i.$$

Examples of such values are given in table 1.5.
Although figure (1.1) supports our choice to neglect a possible interaction between tobacco consumption and occupational exposure to asbestos, future research would profitably consist in enriching the maximal model to account for this possible interaction and letting the data decide whether the added complexity is worth keeping or not.

References

1. Chambaz A., Choudat D., Huber C., Pairon J-C, van der Laan M.J. : Threshold regression models adapted to case-control studies. Application to lung cancer induced by occupational exposure. To appear in Biostatistics Journal. (2013)
2. Lee M.-L. T., Whitmore G. A. : Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Statistical Science **21(4)**, 501-513 (2006)
3. Lee M.-L. T., Whitmore G. A. : Proportional hazards and threshold regression: their theoretical and practical connections. Lifetime Data Analysis. **16(2)**, 196-214 (2010).
4. Pairon J-C, Legal-Regis B., Ameille J., Brechot J-M, Lebeau B., Valeyre D., Monnet I., Matrat M., and Chammings B., Housset S. : Occupational lung cancer: a multicentric case-control study in Paris area. European Respiratory Society, 19th Annual Congress, Vienna. (2009)