DURATION MODELS

Lecture Notes 2015-2016

Part 2

Parametric and semi-parametric models

Frederic PLANCHET / Olivier DURAND

Version 3.4

December 2016

Table of content

1. Taking into account censoring in duration models 4
- 1.1. Type I censoring: fixed censoring 4
  - 1.1.1. Point estimation 5
  - 1.1.2. Interval estimation 6
- 1.2. Type III censoring: random censoring 7
  - 1.2.1. The case of an iid sample 7
  - 1.2.2. Taking into account covariables 8
- 1.3. Another type of censoring: “stop at rth death” (type II censoring) 9
- 1.4. Truncation 11
  - 1.4.1. Definition 11
  - 1.4.2. Type III truncation and censoring 12

2. Latent and observable likelihood in the presence of censoring 13
- 2.1. Application of the maximum-likelihood method 15
  - 2.1.1. General information 15
  - 2.1.2. Latent likelihood and observable likelihood 16
- 2.2. The case of duration models 16
- 2.3. Example: Weibull model 17
  - 2.3.1. Parameters estimation 17
  - 2.3.2. Numerical illustration 18
- 2.4. Numerical algorithms for likelihood maximisation 19
  - 2.4.1. Newton-Raphson algorithm 19
  - 2.4.1. Expectation-Maximisation algorithm 20
  - 2.4.2. Other methods 21

3. Proportional hazard models 21
- 3.1. Case where the basic hazard function is known 23
  - 3.1.1. Likelihood equations 23
  - 3.1.2. Fisher Information 24
- 3.2. Case of a parametric basic hazard: Weibull model 24
  - 3.2.1. General presentation 25
  - 3.2.2. Particular case of the exponential model 26
- 3.3. Case where the basic hazard function is not specified: Cox model 26
  - 3.3.1. Estimate of the parameters 27
  - 3.3.2. Model tests 29

4. Tests based on likelihood 30
  - 4.1.1. Ratio of likelihood maxima 30
  - 4.1.2. Wald test 30
  - 4.1.3. The score test 30

5. Adjustment of raw mortality rates 31
- 5.1. Discretised maximum-likelihood 32
- 5.2. Application: Makeham model 33
  - 5.2.1. Curve fitness to Makeham model 34
  - 5.2.2. Fitting though maximum-likelihood method 35
- 5.3. Thatcher model 36
- 5.4. Raw rates fitting based on Logits 37
  - 5.4.1. The logistic function 37

5.4.2. Logistic fittings 38
5.4.3. Parameters estimation 40
5.5. Confidence intervals for raw rates 40
5.5.1. Asymptotic confidence intervals 41
5.5.2. Finite distance confidence intervals 42
6. References 43

1. Taking into account censoring in duration models

The object of this section is to determine the general form of likelihood of a censored duration model depending on the type of censoring and to illustrate, in the case of the exponential distribution, the impact of censoring[1] on likelihood.

In practice, one can be confronted to right-censoring (if X is the variable of interest, the observation of censoring C indicates that $X \geq C$) or left-censoring (the observation of censoring C indicates that $X \leq C$); the two types of censoring can be observed simultaneously. The traditional example is set by the following situation: one wants to know at which age X *the* children of a given group are able to carry out a certain task. When the experiment begins, some children of age C are already able to achieve the task, and for them $X \leq C$: this is left-censoring; at the end of the experiment, some children are still not able to achieve the task and for them $X \geq C$: this is right-censoring.

1.1. Type I censoring: fixed censoring

Let us consider a sample of survival durations $(X_1, \ldots, X_n)$ as well as $C > 0$ fixed; the likelihood function of the model associated with the observations $(T_1, D_1), \ldots, (T_n, D_n)$ with:

$$T_i = X_i \wedge C \text{ and } D_i = \begin{cases} 1 \, si \, X_i \leq C \\ 0 \, si \, X_i > C \end{cases}$$

holds a continuous component and a discrete component; it is written:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(T_i)^{D_i} S_\theta(C)^{1-D_i}$$

in other words, when one observes exit before censoring, it is the density term that intervenes in likelihood, and in the contrary case one finds the discrete term, the value of which is the survival function at the date of censoring. The distribution is thus continuous with respect to T_i and discrete with respect to D_i.

To demonstrate this formula, one only needs to calculate $P(T_i \in [t_i, t_i + dt_i], D_i = d_i)$. Since D_i can only take values 0 and 1, one calculates, on $[0, C]$:

$$\begin{aligned} P(T_i \in [t_i, t_i + dt_i], D_i = 1) &= P(X_i \wedge C \in [t_i, t_i + dt_i], X_i \leq C) \\ &= P(X_i \in [t_i, t_i + dt_i]) = f_\theta(t_i) dt_i \end{aligned}$$

(one can always assume dt_i sufficiently small so that $t_i + dt_i \leq C$) and

$$\begin{aligned} P(T_i \in [t_i, t_i + dt_i], D_i = 0) &= P(X_i \wedge C \in [t_i, t_i + dt_i], X_i \geq C) \\ &= P(X_i \geq C) = S_\theta(C) \end{aligned}$$

[1] And, marginally, truncation, which will be mentioned but not developed.

These two cases can be summarised as follows:

$$P\left(T_i \in [t_i, t_i + dt_i], D_i = d_i\right) = f_\theta(t_i)^{d_i} S_\theta(C)^{1-d_i}$$

One can also find this expression by observing that:

$$P\left(T_i > t_i, D_i = 1\right) = P\left(X_i > t_i, X_i \le C\right) = \int_{t_i}^{C} f_\theta(u)\,du$$

and in the case where $D_i = 0$ since then $T_i = C$ there is no density, but simply the probability of this event is equal to $S_\theta(C)$. Since for a censored observation, by definition, $T_i = C$, the expression above can be rewritten:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(T_i)^{D_i} S_\theta(T_i)^{1-D_i}$$

By remembering that the probability density function can be written as a function of hazard function and survival function $f_\theta(t) = h_\theta(t) S_\theta(t)$ one can also write likelihood in the following form (except for a multiplicative constant):

$$L(\theta) = \prod_{i=1}^{n} S_\theta(T_i) h_\theta(T_i)^{D_i}$$

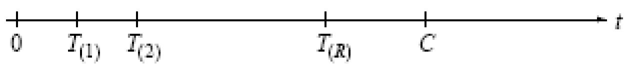
This expression is thus simply the product of the values of the survival function (which translates the fact that the individuals are observed at least until T_i), weighed for non-censored exits by the value of the hazard function (which translates the fact that for these observations the exits indeed take place at time T_i). One generally uses log-likelihood, which is, except for an additive constant:

$$\ln L(\theta) = \sum_{i=1}^{n} \left[D_i \ln\left(h_\theta(T_i)\right) + \ln\left(S_\theta(T_i)\right) \right].$$

As an illustration, the cases of point estimation and interval estimation in the context of exponential distribution are detailed hereafter.

1.1.1. Point estimation

Let us now consider the case where the underlying distribution is exponential, of parameter θ ; let $R = \sum_{i=1}^{n} D_i$ be the number of observed deaths:



Since $f_\theta(t) = \theta e^{-\theta t}$, the likelihood is thus written $L(\theta) = \prod_{i=1}^{n} \left(\theta e^{-\theta T_i}\right)^{D_i} \left(e^{-\theta C}\right)^{1-D_i}$, which becomes:

$$L(\theta) = \theta^R \exp\left(-\theta \sum_{i=1}^{n} T_i\right)$$

One can incidentally notice that the distribution of R is discrete, and is a binomial distribution of parameters $\left(n, 1-e^{-\theta C}\right)$: the number of non-censored exits corresponds to draws from n values, the chances of success being equal to $1-e^{-\theta C} = P_\theta(T \le C)$.

If $T = \sum_{i=1}^{n} T_i$ indicates the total "risk exposure"[2], one can write $T = \sum_{i=1}^{R} T_{(i)} + (n-R)C$; by cancelling the first derivative of the log-likelihood $l(\theta) = R\ln(\theta) - \theta\left(\sum_{i=1}^{R} T_{(i)} + (n-R)C\right)$ with respect to θ, one finds that the maximum-likelihood estimator (MLE) of θ is $\hat{\theta} = \frac{R}{T}$. The exhaustive statistic is thus two-dimensional (T, R).

The estimator of θ is therefore the ratio of observed deaths to risk exposure; in a non-censored model (obtained as the limit case of a censored model when $C \to +\infty$), the expression $\hat{\theta} = \frac{R}{T}$ becomes $\hat{\theta} = \frac{1}{\overline{X}}$; indeed, all deaths are observed in that case, and the estimator is the traditional "multiplicative inverse of the empirical average of lifetimes".

1.1.2. Interval estimation

One can use the asymptotic efficiency of the maximum-likelihood estimator to determine a confidence interval for the estimator. In the case of exponential distribution one can also notice that, provided $m_C(\theta)$ and $\sigma_C(\theta)$ are the mean and standard deviation of T, then using the central limit theorem one gets $\sqrt{n}\frac{T - m_C(\theta)}{\sigma_C(\theta)}$ which converges in probability to a standard normal distribution. Indeed, the random variables $T_i = X_i \wedge C$ are iid, since X_i are iid. The expressions of $m_C(\theta)$ and $\sigma_C(\theta)$ can be obtained by a few calculations:

✓ $m_C(\theta) = \int_0^C u\theta e^{-\theta u} du + Ce^{-\theta C} = \frac{1-e^{-\theta C}}{\theta}$

[2] T is sometimes called « global working time during trials ».

✓ $\sigma_C^2(\theta) = E\left(T_i^2\right) - \left(m_C(\theta)\right)^2 = \frac{1}{\theta^2}\left(1 - 2\theta C e^{-\theta C} - e^{-2\theta C}\right)$

Under the assumption that the duration of experiment C is small compared to the *a priori* lifetime of each individual $\frac{1}{\theta}$, one has θC which is small compared to 1 and one can thus make a limited development of exponentials to the order 3 in θC, which leads to: $\sigma_C^2(\theta) = \frac{\theta C^3}{3}$. One thus obtains a relatively simple form of confidence interval for the parameter θ.

1.2. Type III censoring: random censoring[3]

1.2.1. The case of an iid sample

Type III censoring generalises type I censoring to the case where the censoring date is a random variable; more precisely, let $(X_1, \ldots, X_n)$ be a sample of survival durations, and $(C_1, \ldots, C_n)$ be a second independent sample composed of positive variables; it is said that there is type III censoring for that sample if instead of directly observing $(X_1, \ldots, X_n)$ one observes $(T_1, D_1), \ldots, (T_n, D_n)$ with:

$$T_i = X_i \wedge C_i \text{ and } D_i = \begin{cases} 1 \ si \ X_i \le C_i \\ 0 \ si \ X_i > C_i \end{cases}$$

The likelihood of the sample $(T_1, D_1), \ldots, (T_n, D_n)$ is written, with obvious notations:

$$L(\theta) = \prod_{i=1}^{n} \left[f_X(T_i, \theta) S_C(T_i, \theta) \right]^{D_i} \left[f_C(T_i, \theta) S_X(T_i, \theta) \right]^{1-D_i}$$

The form of the above likelihood can be deduced, for example, from the fact that $(T_1, \ldots, T_n)$ is a sample of the distribution $S_T(\theta,.)$ with:

$$S_T(\theta, t) = P_\theta(T_i > t) = P_\theta(X_i \wedge C_i > t) = P_\theta(X_i > t) P_\theta(C_i > t) = S_X(t, \theta) S_C(t, \theta).$$

One writes, as in 1.1 above, that:

$$\begin{aligned} P\left(T_i \in [t_i, t_i + dt_i], D_i = 1\right) &= P\left(X_i \wedge C_i \in [t_i, t_i + dt_i], X_i \le C_i\right) \\ &= P\left(X_i \in [t_i, t_i + dt_i], t_i < C_i\right) = f_X(\theta, t_i) S_C(\theta, t_i) dt_i \end{aligned}$$

and

[3] These models can be analyzed as models with 2 independent competing risks.

$$P\left(T_i \in \left[t_i, t_i + dt_i\right], D_i = 0\right) = P\left(X_i \wedge C_i \in \left[t_i, t_i + dt_i\right], X_i \geq C_i\right)$$
$$= P\left(C_i \in \left[t_i, t_i + dt_i\right], X_i > t_i\right) = S_X(\theta, t_i) f_C(\theta, t_i) dt_i$$

These expressions are directly obtained from those seen in 1.1, in conditioning with respect to censoring, then integrating with respect to its distribution. More precisely, one writes:

$$P\left(T_i > t_i, D_i = 1\right) = P\left(X_i \wedge C_i > t_i, X_i \leq C_i\right) = P\left(t_i < X_i \leq C_i\right)$$
$$= \int_{t_i}^{+\infty} P\left(t_i < X_i \leq c\right) f_C(\theta, c) dc = \int_{t_i}^{+\infty} \left(\int_{t_i}^{c} f_X(\theta, x) dx \right) f_C(\theta, c) dc$$

then by Fubini one reverses integrals to obtain:

$$P\left(T_i > t_i, D_i = 1\right) = \int_{t_i}^{+\infty} f_X(\theta, x) \left(\int_{x}^{+\infty} f_C(\theta, c) dc \right) dx$$
$$= \int_{t_i}^{+\infty} f_X(\theta, x) S_C(\theta, x) dx$$

and finally $P\left(T_i \in \left[t_i, t_i + dt_i\right], D_i = 1\right) = -\frac{d}{dt_i} P\left(T_i > t_i, D_i = 1\right) = f_X(\theta, t_i) S_C(\theta, t_i) dt_i$. The assumption is then made that the censoring is non-informative, i.e. that the censoring distribution is independent from the parameter θ. Likelihood is in this case of the form:

$$L(\theta) = const \prod_{i=1}^{n} f_\theta(T_i)^{D_i} S_\theta(C_i)^{1-D_i}$$

The *const* term gathers information coming from the censoring distribution, which does not depend on the parameter. This last expression can be written as in 1.1 above:

$$L(\theta) = \prod_{i=1}^{n} S_\theta(T_i) h_\theta(T_i)^{D_i}$$

One can observe the fact that fixed censoring is simply a particular case of non-informative random censoring in which the censoring distribution is one of Dirac at point *C*. The expression established in the particular case of fixed censoring can therefore easily be generalised.

1.2.2. Taking into account covariables

When the model holds *p* explanatory variables (covariables) $Z = \left(Z_1, \ldots, Z_p\right)$, the assumption is made that the conditional distribution of *X* knowing *Z* depends on a parameter θ.

The observed sample becomes a sequence of triplets $\left(T_i, D_i, Z_i\right)$; one gets back to the assumption of non-informative censoring; one assumes moreover that *X* and *C* are independent conditionally to *Z*, and that *C* is non-informative for the parameters of the conditional distribution of X knowing Z. One finally assumes that Z admits a probability density which depends on a parameter ϕ, $f_Z(z,\phi)$.

Under these conditions, the expression of likelihood seen in 1.2 above becomes:

$$L(\theta) = \prod_{i=1}^{n} h_{\theta|Z}(T_i)^{D_i} S_{\theta|Z}(T_i) f_Z(Z_i,\phi)$$

When the distribution of T knowing Z and the distribution of Z have no common parameter, one simply finds the expression of 1.2, in which the distribution of X is replaced by the conditional distribution of X knowing Z. This reasoning can be generalised without difficulty to the case of covariables depending on time.

1.3. Another type of censoring: "stop at rth death" (type II censoring)

Let us now consider the case where the date of observation end is not defined in advance, observation instead stopping at the time of the rth exit. The end date of the experiment is therefore random and is equal to $X_{(r)}$.

In a more formal way, let us consider a sample of survival durations $(X_1,\dots,X_n)$ and $r>0$ fixed; it is said that there is type II censoring for this sample if, instead of directly observing $(X_1,\dots,X_n)$, one observes $(T_1,D_1),\dots,(T_n,D_n)$ with:

$$T_i = X_i \wedge X_{(r)} \text{ and } D_i = \begin{cases} 1\ si\ X_i = T_i \\ 0\ si\ X_i \neq T_i \end{cases}$$

$X_{(r)}$ being the rth order statistic of sample $(X_1,\dots,X_n)$. The definition of the censoring indicator can be rewritten $D_i = \begin{cases} 1\ si\ X_i \leq X_{(r)} \\ 0\ si\ X_i > X_{(r)} \end{cases}$, which is a form similar to the case of fixed censoring with $C = X_{(r)}$.

The likelihood has a form close to that of the type I censoring case; for its formulation one notices that, in the discrete part of the distribution, the times of the r exits are chosen among the n observations. Which results in:

$$L(\theta) = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r} f_\theta\left(X_{(i)}\right)\right] S_\theta\left(X_{(r)}\right)^{n-r}$$

$$= \frac{n!}{(n-r)!}\prod_{i=1}^{n} f_\theta(T_i)^{D_i} S_\theta(T_i)^{1-D_i} = \frac{n!}{(n-r)!}\prod_{i=1}^{n} h_\theta(T_i)^{D_i} S_\theta(T_i)$$

If the reference distribution is the exponential distribution, one therefore finds that:

$$L(\theta) = \frac{n!}{(n-r)!}\theta^r \exp(-\theta T)$$

with $T = \sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}$; the statistic T is therefore exhaustive for the model. The maximum-likelihood estimator can easily be deducted from the above expression: $\hat{\theta} = \frac{r}{T}$. In fact, in this case one can completely determine the distribution of T; precisely:

<u>Proposal</u>: $2\theta T$ follows a Khi-2 distribution with $2r$ degrees of freedom or, in an equivalent way, T follows a $\gamma(r,\theta)$ distribution, since the Khi-2 distribution with $2r$ degrees of freedom is a Gamma distribution of parameters $\left(r, \frac{1}{2}\right)$

<u>Demonstration</u>: One wants to show that $P(T \leq x) = P(\chi^2_{2r} \leq 2\theta x)$; since the distribution of Khi-2 with $2r$ degrees of freedom is a Gamma distribution of parameter $\left(r, \frac{1}{2}\right)$, its density is:

$$f(x) = \frac{1}{2^r \Gamma(r)} x^{r-1} e^{-\frac{x}{2}}.$$

One writes:

$$P(T \leq x) = \frac{n!}{(n-r)!} \theta^r \int_{A_x} \mathbf{exp}\left(-\theta\left(\sum_{i=1}^{r} t_i + (n-r)t_r\right)\right) dt_1 \ldots dt_r ,$$

with $A_x = \left\{0 < t_1 \ldots < t_r \,/ \sum_{i=1}^{t} t_i + (n-r)t_r \leq x\right\}$. The following variable shift is made:

$$t_1 = u_1; t_2 = u_1 + u_2 ; \ldots ; t_{r-1} = u_1 + \ldots + u_{r-1} ; \sum_{i=1}^{r-1} t_i + (n-r+1)t_r = u .$$

It is verified that the determinant of the Jacobian matrix of generic term $\frac{\partial t_i}{\partial u_j}$ is $\frac{1}{n-r+1}$, which leads to:

$$P(T \leq x) = \frac{n!}{(n-r)!} \theta^r \int_{B_x} \frac{1}{n-r+1} e^{-\theta u} du_1 .. du_{r-1} du$$

with $B_x = \left\{u_1 > 0, \ldots, u_{r-1} > 0 ; \sum_{i=1}^{t-1} (r-i)u_i < u \leq x\right\}$. Another change of variable:

$$v_i = (r-i) \times u_i , 1 \leq i \leq r-1 ; v = u$$

finally allows to obtain:

$$P(T \leq x) = \frac{n!}{(n-r+1)!} \theta^r \int_0^x \left[\int_{C_v} \frac{1}{(r-1)!} dv_1 .. dv_{r-1} \right] e^{-\theta v} dv$$

with $C_v = \left\{ v_1 > 0,..,v_{r-1} > 0 \,;\, \sum_{i=1}^{t-1} v_i \le v \right\}$; by observing that the multiple integer on C_x is of the form $cste \times x^{t-1}$, one finally concludes that:

$$P(T \le x) = \frac{1}{\Gamma(r)} \int_0^{\theta x} u^{r-1} e^{-u} du = P\left(\chi^2_{2r} \le 2\theta x\right).$$

One deduces in particular from this proposal that the MLE estimator is biased and that $E\left(\hat{\theta}\right) = \frac{r}{r-1}\theta$: indeed, if T follows a gamma distribution of parameter (r,λ), then $E\left(T^p\right) = \lambda^{-p} \frac{\Gamma(r+p)}{\Gamma(r)}$ for all $p > -r$ and thus:

$$E\left(\hat{\theta}\right) = 2\theta r E\left(\frac{1}{2\theta T}\right) = 2\theta r \frac{1}{2} \frac{\Gamma(r-1)}{\Gamma(r)} = \theta \frac{r}{r-1}.$$

The best unbiased estimator for θ is therefore $\tilde{\theta} = \frac{r-1}{T}$. One can also demonstrate that the variance of $\tilde{\theta}$ is $V\left(\tilde{\theta}\right) = \frac{\theta^2}{r-2}$.

This result can be obtained more simply. One uses for that the fact that the joint distribution of the order statistics $\left(X_{(1)},\dots,X_{(n)}\right)$ is $\bar{f}(x_1,\dots,x_n) = n! \prod_{i=1}^{n} f(x_i) 1_{\{x_1<\dots<x_n\}}$. Through a shift of variable, it is demonstrated that the random variables $Y_i = (n-i+1)\left(X_{(i)} - X_{(i-1)}\right)$ are independent and of common distribution the exponential distribution of parameter θ.

Since $T = \sum_{i=1}^{r} Y_i$ one immediately obtains the result by observing that the sum of r exponential random variables of parameter θ follows a $\gamma(r,\theta)$ distribution. The average duration of the experiment can also easily be deducted: since $T_{(r)} = \sum_{i=1}^{r} \frac{Y_i}{n-i+1}$, one has $E\left(T_{(r)}\right) = \frac{1}{\theta} \sum_{i=1}^{r} \frac{1}{n-i+1}$.

1.4. Truncation

1.4.1. Definition

It is said that there is left-truncation (resp. right-truncation) when the variable of interest is not observable when it is lower than a threshold $c > 0$ (resp. higher than a threshold $C > 0$).

The phenomenon of truncation is very different from censoring, since in this case all information on observations outside of the interval are lost: in the case of censoring, one is aware of the fact that there is information, but one does not know its precise value, simply the fact that it exceeds a threshold; in the case of truncation one does not have this information.

The distribution observed in this case is therefore the conditional distribution with respect to the event $\{c < T < C\}$. The truncated survival function is therefore written:

$$S(t | c < T < C) = \begin{cases} 1 \; si \; t < c \\ \dfrac{S(t) - S(C)}{S(c) - S(C)} \; si \; c \le t \le C \\ 0 \; si \; t > C \end{cases}$$

The hazard function also has $\{c < t < C\}$ as support and is written $h(t | c < T < C) = h(t) \dfrac{S(t)}{S(t) - S(C)}$, which shows that the expression of h does not depend on c. Right-truncation increases the hazard function, and if there is only left-truncation $(C = +\infty)$ then the hazard function is not modified.

Truncation can for example be observed in the case of an IT/data migration process during which only claims in progress would have been included – claims or cases closed being left behind, and the corresponding information being lost. Truncation is also observed in the case of a sick leave policy with a deductible: the absences of durations lower than the deductible are not observed, and one therefore does not have any information about these claims.

1.4.2. Type III truncation and censoring

Most of the time, individuals are not observed since origins, but rather since the age (or the seniority) reached at the beginning of the observation period, noted E_i. Censoring C_i can be lower than the age reached at the end of the observation period if the exit takes place in an anticipated way (for example cancellation). Under these conditions, the expression of the likelihood of the model is:

$$L(\theta) = \prod_{i=1}^{n} h_{\theta|Z,E}(t_i)^{d_i} \, S_{\theta|Z,E}(t_i) \, f_Z(z_i, \phi)$$

When the distribution of T knowing Z and the distribution of Z have no common parameter, one finds the following expression:

$$\ln L(\theta) = cste + \sum_{i=1}^{n} d_i \ln\left(h_{\theta|Z,E}(t_i)\right) + \ln S_{\theta|Z,E}(t_i).$$

As $h_{\theta|Z,E}(t_i) = h_{\theta|Z}(t_i)$ and $S_{\theta|Z,E}(t_i) = \dfrac{S_{\theta|Z}(t_i)}{S_{\theta|Z}(e_i)}$ and one finally gets:

$$\ln L(\theta) = cste + \sum_{i=1}^{n} \left\{ d_i \ln\left(h_{\theta|Z}(t_i)\right) + \ln S_{\theta|Z}(t_i) - \ln S_{\theta|Z}(e_i) \right\}.$$

Of course if all individuals are observed since the origin, $e_i = 0$ and the traditional expression is found:

$$\ln L(\theta) = cste + \sum_{i=1}^{n} d_i \ln\left(h_{\theta|Z}(t_i)\right) + \ln S_{\theta|Z}(t_i).$$

Example: let us consider the Weibull's proportional hazard model (*cf* 3.2) in which:

$$h(x|z;\theta,\alpha) = \exp(-z'\theta)\,\alpha x^{\alpha-1}.$$

The log-likelihood of this model is written according to the general expression pointed out *supra*:

$$\ln L(y|z;\theta,\alpha) = d\ln(\alpha) + (\alpha-1)\sum_{i=1}^{n} d_i \ln(t_i) - \sum_{i=1}^{n} d_i z_i'\theta - \sum_{i=1}^{n} \exp\left(-z_i'\theta\right)(t_i - e_i)^{\alpha}$$

where one noted $d = \sum_{i=1}^{d} d_i$ the number of non-censored exits.

Example: let us consider n individuals for which the assumption is made that the underlying hazard function is constant on the interval $[x, x+1[$; using the above one finds that the log-likelihood of the model is, with the exception of a constant:

$$\ln L(\theta) = \sum_{i=1}^{n} \left[d_i \ln(\theta) + \theta \times (t_i - e_i) \right] = d_x \times \ln(\theta) + \theta \times E_x$$

with $d_x = \sum_{i=1}^{d} d_i$ and $E_x = \sum_{i=1}^{d} (t_i - e_i)$. It can be noted that it is as if the variable D_x which counts the number of exits on the interval $[x, x+1[$ followed a Poisson distribution of parameter $\theta \times E_x$; indeed, in this case $\ln\left(P(D_x = d)\right) = cste + d_x \times \ln(\theta) - \theta \times E_x$.

2. Latent and observable likelihood in the presence of censoring

In this paragraph, one considers observations of durations $(t_1,\dots,t_n)$, censored by a type I (fixed) or type III (random, non-informative) censoring, depending on the observation[4]; it is indeed a kind of censoring which one often meets in insurance problems. Let $(c_1,\dots,c_n)$ be the observed values of censoring. It is assumed that the observed lifetimes also depend on p explanatory variables[5] $(z_1,\dots,z_p)$. In the previous chapter the form of general likelihood was determined, and one now wishes to estimate the parameters by maximising this likelihood,

[4] Equivalent to a random censoring, reasoning conditionally to the censoring value.

[5] z_j is therefore a vector composed of the n values of the explanatory variable for the sample's individuals.

while taking into account these explanatory variables. One will therefore look into expressing the relationship between latent score and observable score, and obtaining the observable model's Fisher information.

Similarly to what came previously, one therefore observes:

$$T_i = X_i \wedge C_i \text{ and } D_i = \begin{cases} 1 \ si \ X_i \leq C_i \\ 0 \ si \ X_i > C_i \end{cases}$$

and the variables $Y_i = (T_i, D_i)$ are independent. When censoring is known, Y_i is a function of the latent variable X_i; the observable model is therefore a model which provides incomplete information on X_i. This functional relationship between latent and observable variables has consequences on the form of observable likelihood. More precisely, there is a functional relationship of the form $Y = \phi(X)$; the respective probability densities of Y and X are noted[6] $l(\theta)$ and $l^*(\theta)$; the observation of Y provides information on the distribution of X, and it is natural to be interested in the conditional distribution of $X \,|Y = y$; one has:

$$l^*(x,\theta) = l(y,\theta)\, l(x\,|y,\theta)$$

and shifting to log-likelihood one can write:

$$\mathbf{ln}\, l^*(x,\theta) = \mathbf{ln}\, l(y,\theta) + \mathbf{ln}\, l(x\,|y,\theta)$$

By differentiating this expression with respect to θ, and integrating with respect to the distribution of $X \,|Y = y$, one finds[7]:

$$E\left[\frac{\partial \mathbf{ln}\, l^*(x,\theta)}{\partial\theta}\Big|y\right] = \frac{\partial \mathbf{ln}\, l(y,\theta)}{\partial\theta} + E\left[\frac{\partial \mathbf{ln}\, l(x\,|y,\theta)}{\partial\theta}\Big|y\right].$$

But $E\left[\frac{\partial \mathbf{ln}\, l(x\,|y,\theta)}{\partial\theta}\Big|y\right] = \int \frac{\partial l(x\,|y,\theta)}{\partial\theta} dx$ since the conditional distribution of $X \,|Y = y$ has as for density $l(x\,|y,\theta)$; by reversing differentiation and integral, since the density integral is equal to one, one finds that $\int \frac{\partial l(x\,|y,\theta)}{\partial\theta} dx = 0$, and the score is therefore written:

$$\frac{\partial \mathbf{ln}\, l(y,\theta)}{\partial\theta} = E\left[\frac{\partial \mathbf{ln}\, l^*(x,\theta)}{\partial\theta}\Big|y\right]$$

Hence the observable score is the best prediction of the latent score, conditionally to the observations. By differentiating the log-likelihood expression twice, one obtains in the same way:

---

[6] Likelihood for an observation will be noted *l*, while likelihood for a sample will be noted *L*.

[7] Expectancies depend on parameter θ which is taken off notations for better clarity.

$$-\frac{\partial^2 \ln l^*(x,\theta)}{\partial\theta\partial\theta'} = -\frac{\partial^2 \ln l(y,\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 \ln l(x|y,\theta)}{\partial\theta\partial\theta'}$$

then by taking the expectation one finds that the Fisher informations of both latent and observable models are bound by the following relationship:

$$I_X^*(\theta) = I_Y(\theta) + E\left[E\left(-\frac{\partial^2 \ln l(x|y,\theta)}{\partial\theta\partial\theta'}|y\right)\right]$$

Note: the notation $\frac{\partial^2 f(\theta)}{\partial\theta\partial\theta'}$ indicates the Hessian matrix associated with *f*, of common term $\frac{\partial^2 f(\theta)}{\partial\theta_i\partial\theta_j}$.

2.1. Application of the maximum-likelihood method

This section presents the relationships between observable and latent likelihood in a general model, and then specifies the case of a duration model.

2.1.1. General information

One assumes the independence of observations conditionally to the explanatory variables as well as censoring; the log-likelihood of the model is written:

$$\ln L(y|z,c;\theta) = \sum_{i=1}^{n} \ln l(y_i|z_i,c_i;\theta)$$

and since log-likelihood is differentiable, the maximum-likelihood estimator cancels the vector of scores: $\frac{\partial \ln L(y|z,c;\hat{\theta})}{\partial\theta} = 0$.

Under regularity technical requirements, most of the time satisfied in practice, one knows that there exists a local maximum of log-likelihood, almost surely converging towards the true parameter value and that, moreover, the maximum-likelihood estimator is asymptotically efficient and Gaussian, i.e.:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to N\left(0, I(\theta)^{-1}\right)$$

with the Fisher information defined by $I(\theta) = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} E\left(-\frac{\partial^2 \ln l(y_i|z_i,c_i)}{\partial\theta\partial\theta'}|z_i,c_i\right)$, the limit being in probability. The asymptotic variance of the estimator can be estimated by:

$$\hat{V}(\hat{\theta}) = -\left[\frac{\partial^2 \ln L(y|z,c;\hat{\theta})}{\partial\theta\partial\theta'}\right]^{-1}$$

One thus disposes of a general framework to estimate the parameter though maximum-likelihood in the presence of censoring[8] and explanatory variables.

2.1.2. Latent likelihood and observable likelihood

Likelihood of the complete model, latent, is not observable; there is however a simple relationship between latent and observable scores, in that observable score is the optimal forecast of latent score based on observable variables, hence in a formal way:

$$\frac{\partial \ln L(y|z,c;\theta)}{\partial \theta} = E\left[\frac{\partial \ln L^*(x|z,c;\theta)}{\partial \theta}|y,z,c\right]$$

This property directly comes from the relationship established for an observation while introducing: $\frac{\partial \ln l(y,\theta)}{\partial \theta} = E\left[\frac{\partial \ln l^*(x,\theta)}{\partial \theta}|y\right]$.

With regard to the Fisher information, the information of the latent model can be broken down into the sum of the observable model's information, and a term measuring the loss of information due to the presence of censoring. Which yields:

<u>Proposal</u>: $I^*(\theta) = I(\theta) + J(\theta)$, with:

$$J(\theta) = \lim_{n\to+\infty} E\left[\frac{1}{n}\sum_{i=1}^{n} V\left(\frac{\partial \ln l^*(x_i|z_i,c_i;\theta)}{\partial \theta}|y_i,z_i,c_i\right)|z,c\right],$$

the limit being taken in probability.

To prove this result one applies the equation of decomposition of the variance $V[A] = E(V[A|B]) + V(E[A|B])$ to $A = \frac{\partial \ln l^*(x_i|z_i,c_i;\theta)}{\partial \theta}|z_i,c_i$ and $B = Y$.

2.2. The case of duration models

In the case of a duration model, likelihood is calculated as a function of hazard rate and survival function, rather than density; since $f(t) = S(t)h(t)$, one obtains:

$$\ln L^*(x|z;\theta) = \sum_{i=1}^{n} \ln h(x_i|z_i;\theta) + \sum_{i=1}^{n} \ln S(x_i|z_i;\theta)$$

Observable log-likelihood is calculated conditionally to (z,c) and is expressed by[9]:

$$\ln L(y|z,c;\theta) = \sum_{i=1}^{n} d_i \ln h(t_i|z_i;\theta) + \sum_{i=1}^{n} \ln S(t_i|z_i;\theta)$$

One therefore finds, as established in 1.2 above, that observable log-likelihood is written in the same manner as in the latent model, replacing real duration by truncated duration and only keeping hazard function for complete information (where $d_i = 1$).

---

[8] The form of likelihood in the case of a duration model is set out in 2.2.

[9] Cf. 1.1.

However, the likelihood equations do not have a simple expression in the general case; usual algorithms will be used to determine the MLE by approximation: Newton-Raphson, BHHH (Berndt, Hall, Hall, Hausman) and EM algorithm – the latter being particularly well suited to the case of incomplete data. These algorithms are presented in 2.4 *infra*.

However, for some model classes a direct approach remains possible: that is in particular the case of proportional hazard models, studied hereafter.

2.3. Example: Weibull model

The introduction showed the estimation of the parameters of the Weibull model in the non-censored case. We will now see as an example the case of right-censoring. The following model is therefore considered:

$$f(x) = \frac{\alpha}{l^{\tau}} x^{\alpha-1} \mathbf{exp}\left\{-\left(\frac{x}{l}\right)^{\alpha}\right\},\ S(x) = \mathbf{exp}\left\{-\left(\frac{x}{l}\right)^{\alpha}\right\}$$

for which the following censored sample is observed $(t_i, d_i)_{i\in\{1,\ldots,n\}}$ where $d_i = \begin{cases} 1 & \text{si } t_i = x_i \\ 0 & \text{si } t_i < x_i \end{cases}$ is the indicator of non-censored information.

2.3.1. Parameters estimation

The likelihood of this model is written:

$$L(\alpha, l) \propto \prod_{i=1}^{n} f(t_i)^{d_i} S(t_i)^{1-d_i}$$

While noting $d_{\bullet} = \sum_{i=1}^{n} d_i$ the number of non-censored exits observed, there comes:

$$L(\alpha, l) \propto \left(\frac{\alpha}{l^{\alpha}}\right)^{d_{\bullet}} \prod_{i=1}^{n} t_i^{(\alpha-1)d_i} \mathbf{exp}\left\{-d_i\left(\frac{t_i}{l}\right)^{\alpha}\right\} \mathbf{exp}\left\{-(1-d_i)\left(\frac{t_i}{l}\right)^{\alpha}\right\},$$

$$L(\alpha, l) \propto \left(\frac{\alpha}{l^{\alpha}}\right)^{d_{\bullet}} \mathbf{exp}\left\{-l^{-\alpha}\sum_{i=1}^{n} t_i^{\alpha}\right\} \mathbf{exp}\left\{(\alpha-1)\sum_{i=1}^{n} d_i \,\mathbf{ln}\, t_i\right\}$$

from which the following log-likelihood is deduced:

$$\mathbf{ln}\, L(\alpha, l) = \mathbf{ln}\, k + d_{\bullet}(\mathbf{ln}\,\alpha - \alpha\,\mathbf{ln}\, l) - l^{-\alpha}\sum_{i=1}^{n} t_i^{\alpha} + (\alpha-1)\sum_{i=1}^{n} d_i \,\mathbf{ln}\, t_i$$

Partial differential equations are thus written:

$$\begin{cases} \dfrac{\partial}{\partial l} \ln L(\alpha, l) = -\dfrac{d_\bullet}{l} + \alpha l^{-\alpha-1} \displaystyle\sum_{i=1}^{n} t_i^{\alpha} \\ \dfrac{\partial}{\partial \alpha} \ln L(\alpha, l) = d_\bullet \left(\dfrac{1}{\alpha} - \ln l \right) + l^{-\tau} \left(\ln l \displaystyle\sum_{i=1}^{n} t_i^{\alpha} - \sum_{i=1}^{n} t_i^{\alpha} \ln t_i \right) + \displaystyle\sum_{i=1}^{n} d_i \ln t_i \end{cases}$$

Hence one seeks solutions for the following system:

$$\begin{cases} l = \left(\dfrac{1}{d_\bullet} \displaystyle\sum_{i=1}^{n} t_i^{\alpha} \right)^{1/\alpha} \\ \dfrac{1}{\alpha} = \dfrac{\displaystyle\sum_{i=1}^{n} t_i^{\alpha} \ln t_i}{\displaystyle\sum_{i=1}^{n} t_i^{\alpha}} - \dfrac{1}{d_\bullet} \displaystyle\sum_{i=1}^{n} d_i \ln t_i \end{cases}$$

The second equation defines an algorithm which converges towards $\hat{\alpha}$, provided it is not given a too distant initial seed. In practice, this value can be the estimator obtained by the method of quantiles applied on the full complete observations (*cf.* introduction lecture). Once $\hat{\alpha}$ is obtained, $\hat{l}$ is deduced through the first equation.

2.3.2. Numerical illustration

An illustration is proposed in which 1'000 observations were simulated, of which 47 % censored.

A first estimate of the parameters was carried out based on the non-censored 1'000 observations of the principal risk, in order to obtain estimates which will be used as a standard to compare estimates obtained in the censored case.

Note: It is necessary to define a stop criterion for the algorithms, making it possible to obtain the mle $\hat{\alpha}$. In this illustration, the algorithm was stopped when the relative variation of the output of an iteration became lower (in absolute value) than 0.01 %.

One should notice that using the selected stop criterion, the algorithm which provides $\hat{\alpha}$ is materially faster (factor 10 of iteration count) with complete data, compared to censored data.

The following table shows the various estimates of parameters and indicates the expectancy and the variance corresponding to these estimates. Simulations were carried out taking as theoretical value for the parameters $\alpha = 2,5$ and $l = 45$.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} \mathbb{1}_{]t;\infty[}(T_x)$$

	Complete data	Incomplete data	
		Not taking censoring into account	Taking censoring into account
α	2.43	2.26	2.48
L	44.78	38.31	44.65
Average	39.71	33.93	39.61
Variance	302.97	252.29	291.72

The following table shows the relative errors when compared to the situation in which all observations are complete.

	Not taking censoring into account	Taking censoring into account
α	-7.1 %	1.9 %
L	-14.5 %	-0.3 %
Average	-14.5 %	-0.2 %
Variance	-16.7 %	-3.7 %

The use of *all* available data, even incomplete, proves to be essential. In particular, not taking into account censoring results in underestimating the survival duration by 15 %. In the same way, in the presence of type I or II censoring, not taking into account the full available observations results in estimating a model in which the maximum survival duration is the level of censoring.

2.4. Numerical algorithms for likelihood maximisation

As seen in 2.2 above, the analytical expression of log-likelihood only seldom makes it possible to perform a direct calculation of the maximum-likelihood estimator. Of course, standard algorithms such as Newton-Raphson can be used in this context. However, specific methods can prove better suited.

The reader interested by an introduction to numerical methods for optimisation may refer to Ciarlet [1990].

2.4.1. Newton-Raphson algorithm

To solve the equation $f(x_0) = 0$ an algorithm is used, built from a linearisation near the solution, based on the development of Taylor to the order one; noting that

$f(x_{k+1}) = f(x_k) + (x_{k+1} - x_k)\frac{df}{dx}(x_k) + o(x_{k+1} - x_k)$, one thus proposes the recurrence defined by $f(x_{k+1}) = 0$, which leads to:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

In the case of a duration model, one uses as function *f* the derivative of log-likelihood with respect to the parameter (score), which leads to the expression:

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2}{\partial\theta\partial\theta'}\ln L(y|z,c;\theta_k)\right]^{-1} \frac{\partial \ln L(y|z,c;\theta_k)}{\partial\theta}.$$

The above writing is matrix calculus, valid for a multidimensional θ.

In order for the algorithm to converge it is advisable to start from an initial value “near” the theoretical value. It has an interesting property: provided we have a convergent estimator, not necessarily asymptotically efficient, it can be sued as the initial value for the Newton-Raphson algorithm. Asymptotic efficiency is then obtained from the first iteration[10].

There exists an alternative to the Newton-Raphson algorithm, called BHHH algorithm (Berndt, Hall, Hall, Hausman), which consists in replacing in the above iterative expression the Fischer information matrix by its expression only based on the first derivative of log-likelihood. One obtains as follows:

$$\theta_{k+1} = \theta_k - \left[\sum_{i=1}^{n} \frac{\partial \ln l(y_i|z_i,c_i;\theta_k)}{\partial\theta}\frac{\partial \ln l(y_i|z_i,c_i;\theta_k)}{\partial\theta'}\right]^{-1} \sum_{i=1}^{n} \frac{\partial \ln l(y_i|z_i,c_i;\theta_k)}{\partial\theta}$$

This version of the Newton-Raphson algorithm has the same properties as the previous one.

2.4.1. Expectation-Maximisation algorithm

This algorithm was imagined more specifically within the framework of incomplete data; it is based on the remark that, if the variables $(x_1,\ldots,x_n)$ were observable, the estimate would simply be carried out by maximising latent log-likelihood $\ln L(x|z,c;\theta)$; since these observations are not available, the idea is to replace the target function by its best approximation knowing the observable variables $(y_1,\ldots,y_n)$. This was initially suggested by Dempster and *al.* [1977].

One introduces, for $(\theta,\hat{\theta})$ fixed, the function $q(\theta,\hat{\theta}) = E_{\hat{\theta}}\left[\ln L^*(x|z,c;\theta)|y,z,c\right]$; the EM algorithm is then defined by the repetition of the following steps:

- calculation of $q(\theta,\theta_k)$;
- maximisation in θ of $q(\theta,\theta_k)$, the solution of which is θ_{k+1}

[10] In this case, the obtained estimator is not maximum-likelihood, but it is nonetheless asymptotically efficient.

In practice this algorithm is interesting when the calculation of $q(\theta, \theta_k)$ is materially simpler than the direct calculation of $\mathbf{ln}\, L(y|z, c; \theta)$; in the contrary case, one may end up using the Newton-Raphson algorithm for the optimisation stage of $q(\theta, \theta_k)$, which makes the approach heavier.

EM algorithm presents under certain regularity conditions, which will not be detailed here, the following "good properties":

Proposal: EM algorithm is increasing, that is $\mathbf{ln}\, L(y|z, c; \theta_{k+1}) \geq \mathbf{ln}\, L(y|z, c; \theta_k)$; moreover, any limit θ_∞ of a sequence of solutions (θ_k) satisfies the following first order condition:

$$\frac{\partial\, \mathbf{ln}\, L(y|z, c; \theta_\infty)}{\partial \theta} = 0$$

Demonstration: cf. Droesbeke and *Al* [1989].

2.4.2. Other methods

Other methods can prove useful in the case of strongly censored samples; indeed in this case, the usual "frequential" estimation used up to that point may not be well suited; one can then turn to Bayesians weighed sampling algorithms, in particular MCMC algorithms.

This situation, not very common in insurance, will not be developed here; the interested reader can refer to Robert [1996].

3. Proportional hazard models

In these models the hazard function is written $h(x|z;\theta) = \mathbf{exp}(-z'\theta) h_0(x)$ with h_0 the basic hazard function, which is given. This situation occurs for example when one wants to position the mortality of a specific group compared to a reference mortality, known, represented by h_0. One can for example imagine that the mortality of an important group was adjusted using a Makeham[11] model and that one is interested in positioning the mortality of some subpopulations: men/women, smokers/non-smokers, *etc.* This approach will primarily focus on defining the positioning of a population compared to another, without always seeking the absolute risk level. The expression of the hazard function of a proportional model can be written:

$$\mathbf{ln} \frac{h(x|z;\theta)}{h_0(x)} = -z'\theta,$$

which expresses that the logarithm of the instantaneous rate of risk, expressed compared to a basic rate, is a linear function of the explanatory variables. There are *p* explanatory variables, which implies that $z'\theta = \sum_{i=1}^{p} z_i \theta_i$. It may easily be verified that the survival function of the model is of the form:

[11] Cf. section 5.

$$S(x|z;\theta) = \mathbf{exp}(-\mathbf{exp}(-z'\theta)H_0(x)),$$

with H_0 the basic cumulated hazard function[12]. Taking into account the form of the survival function, it is natural to be interested in the transformed variable $V = \mathbf{ln}(H_0(X))$; indeed if the following model is considered:

$$v = z'\theta + \varepsilon$$

(in other words one poses $\varepsilon = v - z'\theta$) it is found that

$$P(\varepsilon > t|z;\theta) = P(\mathbf{ln}\, H_0(x) - z'\theta > t|z;\theta) = P(H_0(x) > \mathbf{exp}(z'\theta)\mathbf{exp}(t)|z;\theta),$$

hence:

$$P(\varepsilon > t|z;\theta) = S\left(H_0^{-1}\left[\mathbf{exp}(z'\theta)\mathbf{exp}(t)\right]|z;\theta\right) = \mathbf{exp}(-\mathbf{exp}(t))$$

The distribution (conditional) of the residue ε is thus a Gumbel[13] distribution, which verifies $E(\varepsilon) = -\gamma$ and $V(\varepsilon) = \frac{\pi^2}{6}$, γ being the Euler[14] constant.

One recognizes in the equation $v = z'\theta + \varepsilon$ a formulation formally equivalent to that of a linear model, in which the residues are however neither Gaussian, nor centered, since $E(\varepsilon) = -\gamma$:

$$E(V|z;\theta) = -\gamma + z'\theta$$

The important point here is that the distribution of ε does not depend on the parameter. If one wishes to obtain a model with standard residues, one should consider the transformation $V = H_0(X)$. One has $P(V > t) = P\left(X > H_0^{-1}(t)\right) = S\left(H_0^{-1}(t)\right)$ and thus:

$$P(V > t) = \mathbf{exp}(-\mathbf{exp}(-z'\theta) \times t).$$

V thus follows an exponential distribution of parameter $\mathbf{exp}(-z'\theta)$, which results in posing the nonlinear model:

$$v = \mathbf{exp}(z'\theta) + \varepsilon$$

with $E(\varepsilon) = 0$, $V(\varepsilon) = \mathbf{exp}(2z'\theta)$ and $E(V|z;\theta) = \mathbf{exp}(z'\theta)$. It is noted that the residues of this model are heteroscedastic.

One can note that the rate of death of a subpopulation is expressed simply using the basic rate of death:

$$q(x|z;\theta) = 1 - \left(\frac{S(x+1|z;\theta)}{S(x|z;\theta)}\right) = 1 - \left(\frac{S_0(x+1)}{S_0(x)}\right)^{\mathbf{exp}(-z'\theta)} = 1 - (1 - q_0(x))^{\mathbf{exp}(-z'\theta)}.$$

[12] Using the relationship $S(t) = \mathbf{exp}\left(-\int_0^t h(s)ds\right)$.

[13] *Cf.* introduction focused on Weibull distribution, and http://fr.wikipedia.org/wiki/Loi_de_Gumbel

[14] The value of which is approximatively 0,577215665.

When $q_0(x)$ is small one finds as one could expect it:

$$q(x|z;\theta) \approx q_0(x) \times \mathbf{exp}(-z'\theta).$$

3.1. Case where the basic hazard function is known[15]

One is initially interested in the case of non-censored data within the framework of the linear model defined above.

One seeks to estimate θ while assuming H_0 known; the above equation can be used to build a convergent estimator of the parameter, but this estimator is non-asymptotically efficient; one can imagine to use it as initialisation value for a log-likelihood maximisation algorithm. However, the expression of the model in the form of a linear model naturally results in proposing the estimator of ordinary least squares (OLS):

$$\hat{\theta}_{MCO} = \left[\sum_{i=1}^{n} z_i' z_i \right]^{-1} \sum_{i=1}^{n} z_i' \,\mathbf{ln}\, H_0(x_i).$$

In the writing above $z_i = (z_{i1}, \ldots, z_{ip})$ is the vector of size p made up of the values of the explanatory variables for the individual nb i. If it is assumed that the model integrates a constant and that it is the first component of θ, then one can show that $\hat{\theta} - (\gamma, 0, \ldots, 0)'$ is a convergent estimator of θ. The direct transposition of the linear model case thus leads to a simple to calculate estimator, presenting *a priori* “good properties” for θ.

Within the framework of the model $v = \mathbf{exp}(z'\theta) + \varepsilon$, which has the advantage of having standard residues, MLE estimator is solution of the following nonlinear least squares program:

$$\mathbf{Min} \sum_{i=1}^{n} \left[H_0(x_i) - \mathbf{exp}(z_i'\theta) \right]^2 .$$

This estimator can easily be calculated; however, the estimators above are usable for complete data, but not in the case of censored data.

Indeed, in the presence of censoring, the estimator $\hat{\theta}_{MCO}$ restricted to complete data is asymptotically biased. However, the bias being not too large in practice, this estimator can be used as initial value for numerical algorithms.

In the presence of incomplete data, one returns to the model’s likelihood equations.

3.1.1. Likelihood equations

According to the general equations determined in 2.1.2 above, one has:

$$\mathbf{ln}\, L^*(x|z;\theta) = \sum_{i=1}^{n} \left(-z_i'\theta + \mathbf{ln}\, h_0(x_i) \right) - \sum_{i=1}^{n} \mathbf{exp}\left(-z_i'\theta\right) H_0(x_i)$$

[15] In Cox model, the basic hazard function is assumed unknown, when it is here assumed known.

for latent likelihood, and:

$$\ln L(y|z,c;\theta) = \sum_{i=1}^{n} d_i \left(-z_i'\theta + \ln h_0(t_i)\right) - \sum_{i=1}^{n} \exp\left(-z_i'\theta\right) H_0(t_i)$$

for observable likelihood. By differentiation one finds the vector of latent scores:

$$\frac{\partial \ln L^*(x|z;\theta)}{\partial \theta} = -\sum_{i=1}^{n} z_i' + \sum_{i=1}^{n} z_i' \exp\left(-z_i'\theta\right) H_0(x_i) = \sum_{i=1}^{n} z_i' \exp\left(-z_i'\theta\right) \varepsilon_i$$

The latent score is therefore the scalar product between errors $\varepsilon_i = H_0(x_i) - \exp\left(z_i'\theta\right)$ and the explanatory variables, for the metric defined by the weights $\exp\left(-z_i'\theta\right)$. With regard to the observable vector of scores, one has:

$$\frac{\partial \ln L(y|z,c;\theta)}{\partial \theta} = \sum_{i=1}^{n} z_i' \exp\left(-z_i'\theta\right) \tilde{\varepsilon}_i$$

with $\tilde{\varepsilon}_i = E\left(\varepsilon_i | y_i, z_i, c_i\right)$. As the residue of the non-censored model is defined by $\varepsilon_i = H_0(x_i) - \exp\left(z_i'\theta\right)$, it is thus a question of showing that $E\left(\varepsilon_i | y_i, z_i, c_i\right) = H_0(t_i) - d_i \exp\left(z_i'\theta\right)$.

The likelihood equations are thus equivalent to a condition of orthogonality between explanatory variables and expected errors, like in the case of a traditional linear model.

3.1.2. Fisher Information

The Fisher information has a particularly simple expression here:

$$I(\theta) = \sum_{i=1}^{n} z_i' z_i p_i$$

with $p_i = E\left(d_i | z_i, c_i\right) = P\left(X_i < c_i\right)$ the probability that the observation is complete. For which one writes that $\frac{\partial^2 \ln L(y|z,c;\theta)}{\partial\theta\partial\theta'} = -\sum_{i=1}^{n} z_i' z_i \exp\left(-z_i'\theta\right) H_0(t_i)$ then one takes the expectancy by observing that the vector of scores is, in this model, standard. The decomposition of the Fisher information introduced in 2.1.2 above is written here:

$$\sum_{i=1}^{n} z_i' z_i = \sum_{i=1}^{n} z_i' z_i p_i + \sum_{i=1}^{n} z_i' z_i (1 - p_i)$$

3.2. Case of a parametric basic hazard: Weibull model

One examined into 2.3 Weibull model without explanatory variables; one wishes here to generalise this model within the framework of a proportional hazard model. The basic hazard function is not supposed to be known anymore, and is supposed to follow a distribution of Weibull; it depends on a parameter, which will therefore have to be estimated, and the model thus comprises an additional parameter compared to the previous version.

3.2.1. General presentation

This model is defined by the specification[16]:

$$h(x|z;\theta,\alpha) = \mathbf{exp}(-z'\theta)\,\alpha x^{\alpha-1}$$

According to the above, the model log-likelihood is written[17]:

$$\mathbf{ln}\, L(y|z,c;\theta,\alpha) = d\,\mathbf{ln}(\alpha) + (\alpha-1)\sum_{i=1}^{n} d_i\,\mathbf{ln}(t_i) - \sum_{i=1}^{n} d_i z_i'\theta - \sum_{i=1}^{n} \mathbf{exp}(-z_i'\theta)\, t_i^{\alpha}$$

where $d = \sum_{i=1}^{d} d_i$ is the number of non-censored exits. The likelihood equations are therefore:

$$\frac{\partial\,\mathbf{ln}\, L(y|z,c;\hat{\theta},\hat{\alpha})}{\partial\theta} = -\sum_{i=1}^{n} d_i z_i' + \sum_{i=1}^{n} z_i'\,\mathbf{exp}(-z_i'\hat{\theta})\, t_i^{\hat{\alpha}} = 0$$

$$\frac{\partial\,\mathbf{ln}\, L(y|z,c;\hat{\theta},\hat{\alpha})}{\partial\alpha} = \frac{d}{\hat{\alpha}} + \sum_{i=1}^{n} d_i\,\mathbf{ln}(t_i) + \sum_{i=1}^{n} \mathbf{exp}(-z_i'\hat{\theta})\, t_i^{\hat{\alpha}}\,\mathbf{ln}(t_i) = 0.$$

Like in the case where the basic hazard function is known, the first equation is interpreted like a scalar product, between the explanatory variables and the generalised residues $\tilde{\varepsilon}_i = t_i^{\hat{\alpha}} - d_i\,\mathbf{exp}(z_i'\hat{\theta})$, like in 3.1.1 above, but after estimation of the basic hazard function. The second equation does not present any particular interpretation.

These equations must be solved through numerical methods.

The terms of the Fisher information matrix are obtained by differentiating another time, and one finds:

$$\frac{\partial^2\,\mathbf{ln}\, L(y|z,c;\theta,\alpha)}{\partial\theta^2} = \sum_{i=1}^{n} z_i' z_i\,\mathbf{exp}(-z_i'\theta)\, t_i^{\alpha}$$

$$\frac{\partial^2\,\mathbf{ln}\, L(y|z,c;\theta,\alpha)}{\partial\theta\partial\alpha} = \sum_{i=1}^{n} z_i'\,\mathbf{exp}(-z_i'\theta)\, t_i^{\alpha}\,\mathbf{ln}(t_i)$$

$$\frac{\partial^2\,\mathbf{ln}\, L(y|z,c;\hat{\theta},\hat{\alpha})}{\partial\alpha^2} = -\frac{d}{\alpha^2} - \sum_{i=1}^{n} \mathbf{exp}(-z_i'\theta)\, t_i^{\alpha}\,(\mathbf{ln}(t_i))^2$$

[16] The scale parameter of Weibull distribution is fixed to 1.
[17] This expression can be related to that in 2.3 in the model without explanatory variables.

3.2.2. Particular case of the exponential model

When in the previous model the parameter α is set to 1, one obtains the case of an exponential basic hazard function, with parameter noted λ (equal to the value of the hazard function[18]). In 1.1.1 above one studied this case and showed that the maximum likelihood estimator was equal[19] to $\dfrac{d}{\sum_{i=1}^{n} d_i t_i + (n-d)c}$.

Let us now take for parameter $\theta = \dfrac{1}{\lambda}$; in the non-censored case, the estimator of θ is the empirical average of the sample, which is without bias. In the presence of censoring, MLE estimator of θ is the reverse of the estimator above (by functional invariance of MLE) $\hat{\theta} = \dfrac{\sum_{i=1}^{n} d_i t_i + (n-d)c}{d}$, which is a biased estimator. The existence of censoring thus introduces bias into the model. One can show[20] that this bias has for expression:

$$E\left(\hat{\theta}\right) - \theta = \frac{c\,\mathbf{exp}\left(-\frac{c}{\theta}\right)}{n\left[1 - \mathbf{exp}\left(-\frac{c}{\theta}\right)\right]^2} + O\left(n^2\right),$$

and that the asymptotic variance is written:

$$V\left(\hat{\theta}\right) \approx \frac{\theta^2}{n\left[1 - \mathbf{exp}\left(-\frac{c}{\theta}\right)\right]}.$$

The usual normal approximation is deduced.

3.3. Case where the basic hazard function is not specified: Cox model[21]

The constraint of a particular form for the basic hazard function is now lifted; it therefore becomes a nuisance parameter, of infinite size.

Indeed, completely specifying a parametric model can prove too restrictive in some cases; moreover, one can be interested in the sole measurement of the effect of covariables, in which case the specification of the basic hazard function does not bring anything to the model (except for constraints). In other words, the context is one where the objective is to position various populations with respect to each other, without consideration for the absolute level of the risk. That justifies the interest for a partial specification, studied here.

---

[18] In other words the scale parameter which was ignored in Weibull model is reintroduced here.
[19] Assuming all censoring equal to *c*.
[20] See BARTHOLOMEW [1957] and BARTHOLOMEW [1963].
[21] For a detailed presentation of Cox model, see DUPUY [2002], whose notations and presentation logics are used here.

Let us start with the following formulation:

$$h(x|z;\theta) = \mathbf{exp}(-z'\theta)h_0(x)$$

with h_0 an unknown factor.

3.3.1. Estimate of the parameters

To carry out the statistical inference in this model, Cox [1972] proposed to use a partial likelihood in which the parameter of nuisance h_0 does not intervene. This approach is a particular case of a more general approach consisting in determining a partial likelihood when the model contains a parameter of nuisance of great dimension. The principle of this approach, described in Cox [1975], is presented hereafter, and then applied to the case of the Cox model.

Let us consider a vector *X* of density $f_X(x,\beta)$. It is supposed possible to break down *X* into a pair (V,W) such as:

$$f_X(x,\beta) = f_{W|V}(w|v,\beta) f_V(v,\beta)$$

An example of such a decomposition is provided by the vector *V* of the values of *X* ordered by ascending order and *W* the vector of the ranks. It is also assumed that the parameter β is of the form $\beta = (\theta, h_0)$, θ being the parameter of interest. The idea is that, if, in the decomposition above, one of the terms of does not depend on h_0, it can be used to estimate θ. The simplification caused by this approximation must compensate for the loss of information.

Reminder – the basic model under consideration is still the following:

$$T_i = X_i \wedge C_i \text{ and } D_i = \begin{cases} 1 \ si \ X_i \le C_i \\ 0 \ si \ X_i > C_i \end{cases}$$

with $h(x|z;\theta) = \mathbf{exp}(-z'\theta)h_0(x)$. According to the general expression of the likelihood of a censored model in the presence of covariables (*cf* 1.2.2 above), one can write the complete likelihood of Cox model:

$$L(\theta, h_0) =$$

$$\prod_{i=1}^{n} \left[h_0(t_i)\mathbf{exp}(-\theta' z_i)\mathbf{exp}\left(-H_0(t_i)\mathbf{exp}(-\theta' z_i)\right)\right]^{d_i} \left[\mathbf{exp}\left(-H_0(t_i)\mathbf{exp}(-\theta' z_i)\right)\right]^{1-d_i}$$

In the expression above, the basic hazard function intervenes in two manners: directly, and through the cumulated hazard function H_0. One can show that there does not exist a maximum for likelihood unless a restriction is not imposed on the basic hazard function.

By breaking down likelihood in a way that isolates the incidence of the basic hazard function in a term which will be neglected, one obtains (after a series of tedious developments which are not written here, *cf.* Dupuy [2002]) the following expression of partial likelihood (valid with or without *ex-æquo*):

$$L_{Cox}(\theta) = \prod_{i=1}^{n} \left[\frac{\exp(-\theta' z_i)}{\sum_{j=1}^{n} \exp(-\theta' z_j) 1_{\{T_i \le T_j\}}} \right]^{d_i}$$

One can however give a simple heuristic justification of the above formula; it is observed indeed that in the denominator of the above fraction intervenes $R_i = \sum_{j=1}^{n} 1_{\{T_i \le T_j\}}$, which is none other than the population subjected to the risk exposure at the time of the exit of individual i (if it is observed). Conditioning by the timing of occurrence of deaths $0 < t_1 \ldots < t_k$ (with thus $k \le n$ corresponding to the non-censored exits), one considers the following events (ordered): C_i is the set of the censorings occurring between t_{i-1} and t_i and D_i the set of the non-censored exits (death) occurring in t_i. Which sums up to a problem of combinative analysis consisting in counting the exits configurations leading to the observed sequence, the dates of death being known. In other words, one is not interested by the absolute claim level, but simply by the positioning of individuals with respect to one another, according to the values taken by each explanatory variable. One can then break down the probability of observing the sequence (C_i, D_i) into:

$$P\left[(C_i, D_i), 1 \le i \le k\right] \\ = \prod_{i=1}^{k} P\left(D_i \mid C_1 \ldots C_i D_1 \ldots D_{i-1}\right) \times \prod_{i=1}^{k} P\left(C_i \mid C_1 \ldots C_{i-1} D_1 \ldots D_{i-1}\right).$$

By gathering the events related to deaths on the one hand, and those relating to censoring on the other hand, the above expression becomes:

$$P\left[(C_i, D_i), 1 \le i \le k\right] \\ = \prod_{i=1}^{k} P\left(D_i \mid C_1 \ldots C_i D_1 \ldots D_{i-1}\right) \times \prod_{i=1}^{k} P\left(C_i \mid C_1 \ldots C_{i-1} D_1 \ldots D_{i-1}\right)$$

One notices the analogy of the above formula with the general expression of likelihood given *supra*. One can then note R_i the complementary of $\left\{C_1 \ldots C_i D_1 \ldots D_{i-1}\right\}$ to describe the population under risk right before the moment t_i. The basic idea of Cox's partial likelihood consists in ignoring in the likelihood the term associated with censoring, to only keeping:

$$P\left[(C_i, D_i), 1 \le i \le k\right] \approx \prod_{i=1}^{k} P\left(D_i \mid R_i\right).$$

What remains is the evaluation of $P(D_i \mid R_i)$; for simplification purposes the assumption is made of absence of *ex-aequo*, which amounts to saying that the set D_i is a singleton: $D_i = \{j_i\}$. Which then yields:

$$P\left(D_i \middle| R_i\right) = \frac{h\left(t_i, z_{j_i}\right)}{\sum_{j \in R_i} h\left(t_i, z_j\right)} = \frac{\mathbf{exp}\left(-\theta' z_{j_i}\right)}{\sum_{j \in R_i} \mathbf{exp}\left(-\theta' z_j\right)},$$

And finally leads to the expression we were looking for.

The expression of partial likelihood is generalised without difficulty to the case of time dependent covariables; in the case of fixed covariables, one can show (*cf.* Fleming and Harrington [1991]) that this expression is equal to the distribution of the rank vector associated with $(T_1,...,T_n)$. In practice the resolution of the system of equation $\frac{\partial}{\partial \theta_i} \mathbf{ln}\, L_{Cox}(\theta) = 0$ is carried out *via* a numerical algorithm (*cf. infra*).

The interest of the obtained estimator $\hat{\theta}$ is legitimated by the fact that it is convergent and asymptotically normal, like an estimator of the standard maximum-likelihood[22].

3.3.2. Model tests

Two types of tests can be carried out within the framework of the Cox model:

- Validation of the assumption of proportional hazard;
- The total nullity of coefficients, *i.e.* $\theta = 0$

The global model validation can be carried out using a test, the principle of which is studied in detail by Therneau and Grambsch [2000], based on the residues of Schoenfeld. The latter are defined for each individual *i* and each covariable *j* as the difference between the value, at the date T_i of exit of *i*, of the covariable for this individual $z_i = \left(z_{i1}, \ldots, z_{ip}\right)$, and its expected value:

$$r_i = d_i \times \left(z_i - \frac{\sum_{j \in R_i} \mathbf{exp}\left(-\theta' z_j\right) z_j}{\sum_{j \in R_i} \mathbf{exp}\left(-\theta' z_j\right)} \right).$$

By then introducing the product of the inverse of the variance-covariance matrix of Schoenfeld's residues for individual *i* with the vector of these same residues, called reduced Schoenfeld residue, one can build a test of the assumption of proportional hazard. This test will be studied in detail later on.

The total nullity of the coefficients can be tested *via* a traditional test – say Wald or score (*cf.* section 4).

[22] This result is demonstrated by ANDERSEN and GILL [1982].

4. Tests based on likelihood

One proposes here to test an assumption of the form $g(\theta)=0$, where g is a function with values in $\Re^r$, against the alternative $g(\theta)\neq 0$. Three asymptotic tests based on likelihood are classically used: the ratio of likelyhood maxima, Wald test, and the score test. It can be demonstrated that they are equivalent, in the sense that the associated statistics differ from an infinitely small amount in probability. One will thus choose that whose implementation is simplest.

One notes $\hat{\theta}$ the maximum-likelihood estimator in the non-constrained model and $\hat{\theta}^0$ its equivalent in the constrained model. $g(\theta)$ is a vector of dimension r (a (r,1) matrix) and it is supposed that the matrix $\frac{\partial g'}{\partial \theta} = \left(\frac{\partial g_j}{\partial \theta_i}\right)$ which is of dimension (p,r) is of rank r.

4.1.1. Ratio of likelihood maxima

The idea here is to compare constrained and non-constrained likelihoods and to accept the null hypothesis if these two values are close. The following statistic is therefore used:

$$\xi^R = 2\left(\mathbf{ln}\, L\left(\hat{\theta}\right) - \mathbf{ln}\, L\left(\hat{\theta}^0\right)\right)$$

which converges under the null hypothesis towards $\chi^2(r)$, hence a test whose critical region is given by $W=\left\{\xi^R > \chi^2_{1-\alpha}(r)\right\}$.

4.1.2. Wald test

The idea of Wald test is that, if $g\left(\hat{\theta}\right)\approx 0$, then the null hypothesis is accepted. In a formal way, the following statistic:

$$\xi^W = ng'\left(\hat{\theta}\right)\left[\frac{\partial g\left(\hat{\theta}\right)}{\partial \theta'} I\left(\hat{\theta}\right)^{-1} \frac{\partial g'\left(\hat{\theta}\right)}{\partial \theta}\right]^{-1} g\left(\hat{\theta}\right)$$

converges under the null hypothesis towards $\chi^2(r)$, hence a test whose critical region is given by $W=\left\{\xi^W > \chi^2_{1-\alpha}(r)\right\}$.

4.1.3. The score test

One is interested here in the first order condition of the constrained model, which reveals the Lagrangian $\mathbf{ln}\, L(\theta) + g'(\theta)\lambda$. The first order condition is thus written:

$$\frac{\partial\, \mathbf{ln}\, L\left(\hat{\theta}^0\right)}{\partial \theta} + \frac{\partial g'\left(\hat{\theta}^0\right)}{\partial \theta}\hat{\lambda} = 0$$

and the following statistic is used:

$$\xi^S = \frac{1}{n} \frac{\partial \ln L(\hat{\theta}^0)}{\partial \theta'} I(\hat{\theta}^0)^{-1} \frac{\partial \ln L(\hat{\theta}^0)}{\partial \theta}$$

which converges under the null hypothesis towards $\chi^2(r)$, hence a test whose critical region is given by $W = \{\xi^S > \chi^2_{1-\alpha}(r)\}$.

5. Adjustment of raw mortality rates

In this paragraph one illustrates the implementation of a parametric approach in the case of the construction of a mortality table. For various ages $x_0 \leq x \leq x_1$, observations are available, made up on the one hand of populations under risk at the beginning of period[23], noted N_x and, on the other hand, of the deaths observed over the reference period, D_x. The number of deaths at age x is a binomial random variable of parameters N_x and q_x, where q_x is the mortality rate at age x. It is natural to estimate this rate by the empirical estimator $\hat{q}_x = \frac{D_x}{N_x}$, which is unbiased, convergent and asymptotically normal[24]. It will be supposed that one has sufficient data to consider that the Normal approximation is valid. One will for example use Cochran criterion, which consists in checking that $N_x \times \hat{q}_x \geq 5$ and $N_x \times (1-\hat{q}_x) \geq 5$.

According to the above, the most direct method to estimate the parameters of a parametric model in this context consists, once the form of the hazard function is fixed, in writing log-likelihood:

$$\ln L(y_1,\ldots,y_n;\theta) = \sum_{i=1}^{n} d_i \times \ln h_\theta(t_i) + \sum_{i=1}^{n} \ln S_\theta(t_i) - \sum_{i=1}^{n} \ln S_\theta(e_i)$$

then to solve the normal equations $\frac{\partial}{\partial \theta} \ln L(y_1,\ldots,y_n;\theta) = 0$. It is what was carried out in example 1.1.1 above. However, in practice these equations can be tricky to solve. Therefore, if one wishes to use Makeham model, the log-likelihood of a censored sample[25] has the following form:

$$\ln L(y_1,\ldots,y_n;\theta) = \sum_{i=1}^{n} d_i \times \ln\left(a + b \times c^{t_i}\right) + \sum_{i=1}^{n} \left(-at_i - \frac{b}{\ln(c)}\left(c^{t_i} - 1\right)\right).$$

Solving the system of equations $\frac{\partial}{\partial a} \ln L = 0$, $\frac{\partial}{\partial b} \ln L = 0$, $\frac{\partial}{\partial c} \ln L = 0$, is tedious, when possible. Indeed, the additions at play in the expressions above comprise a potentially very alarge number of terms. Thus, one is led to propose a two-steps approach:

[23] Generally, the time period is one year.
[24] In practical terms, a raw mortality rate will often be obtained through a non-parametric framework (Kaplan-Meier) and then the risk exposure of the rate as well as of the observed number of deaths at a given age.
[25] Assumed non-left-truncated to make writing simpler.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} \mathbb{1}_{]t;\infty[}(T_x)$$

- ✓ one starts by calculating raw rates of deaths $\hat{q}_x$ through a method that takes into account potential censoring (also taking into account the degree of accuracy associated with the individual data),
- ✓ then one adjusts the chosen parametric model to these raw rates. For that purpose one uses the "passage formula" between the expression of the continuous time model and the following raw rates:

$$q_x = 1 - \mathbf{exp}\left[-\int_x^{x+1} \mu(y)\, dy \right]$$

This relationship between the discrete mortality rate q_x and the hazard function[26] μ_x simply expresses the fact that the probability of survival between x and $x+1$, conditionally to the fact that the individual is alive at age x, is equal to $\frac{S(x+1)}{S(x)}$.

Curve fitting/smoothing is justified by the fact that the curve of raw rates presents irregularities with age, and that it can be assumed that these abrupt variations are not due to variations of the real risk incidence, but rather due to data insufficiency. Fitting through a function that models the underlying risk is a way of smoothing these sampling[27] fluctuations. Among the distributions most often used, one finds the Makeham distribution, which will be applied below, after having presented the general approach.

5.1. Discretised maximum-likelihood

Within the framework of the binomial model[28], the number of deaths observed at age x, D_x, follows a binomial distribution of parameters $(N_x, q_x(\theta))$ and the likelihood associated with the realisation of a number of deaths d_x is therefore equal to:

$$P(D_x = d_x) = C_{N_x}^{d_x}\, q_x^{d_x}\, (1-q_x)^{N_x - d_x}.$$

For all observations one thus obtains the following log-likelihood (except for a constant independent of the parameter):

$$\mathbf{ln}\, L(\theta) = \sum_x d_x\, \mathbf{ln}\, q_x(\theta) + \sum_x (N_x - d_x)\, \mathbf{ln}(1 - q_x(\theta)).$$

This expression is not very easy to handle (for example within the framework of the Makeham model one will show that $q_x(\theta) = 1 - s \times g^{c^x(c-1)}$), however, numerically the search for a maximum does not pose major problems. In order to get to a problem of weighed least squares, the approximation is generally made that $\hat{q}_x$ follows a normal distribution:

$$\hat{q}_x \approx N\left(q_x(\theta); \sigma^2(\theta) = \frac{q_x(\theta)(1 - q_x(\theta))}{N_x} \right).$$

---

[26] Hazard function h is traditionally noted μ in demographics.

[27] For more developments, cf. lecture on « fitting and smoothing ».

[28] One can often practically get back to this model through proper determination of the population under risk.

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} \mathbb{1}_{]t;\infty[}(T_x)$$

The likelihood function is then written, by making the assumption of independence between ages:

$$L(\theta) = \prod_x \frac{1}{\sigma(\theta)\sqrt{2\pi}} \mathbf{exp}\left(\frac{1}{2} \frac{(q_x(\theta) - \hat{q}_x)^2}{\sigma^2(\theta)} \right),$$

hence the log-likelihood:

$$\mathbf{ln}(L(\theta)) = \sum_x \mathbf{ln}\left(\frac{1}{\sigma(\theta)\sqrt{2\pi}} \right) - \sum_x \frac{1}{2} \frac{(q_x(\theta) - \hat{q}_x)^2}{\sigma^2(\theta)}.$$

The target function is still complex and the parameter comes into the normal distribution's expectancy as well as its variance; that can generate instability in searching for the optimum through algorithms; one will therefore use approached likelihood in which the theoretical variance is replaced by the estimated variance. The maximisation of likelihood is then equivalent to the minimisation of:

$$\sum_x \frac{1}{2} \frac{(q_x(\theta) - \hat{q}_x)^2}{\hat{\sigma}^2} = \sum_x \frac{N_x}{\hat{q}_x(1-\hat{q}_x)} (q_x(\theta) - \hat{q}_x)^2.$$

The problem is reduced to a problem of weighed least squares in the nonlinear case; it can be solved numerically by most specialised statistical software.

However, one has to correctly specify what reference population N_x is used for the binomial experiment. It appears reasonable to wish that on average the model is unbiased, which translates into $E(D_x) = q_x \times N_x$. Without any truncation or censoring, one thus chooses $N_x = S(x)$. In the presence of truncation and/or censoring, it is necessary to take into account these phenomena in the calculation. One can show that it is then reasonable to retain the risk exposure $N_x = E_x$ where $E_x = \sum_{i \in I} d_i(x)$ with $d_i(x)$ the duration of presence under risk of individual *i*. This result will be justified in the lecture on mortality tables.

5.2. Application: Makeham model

The Makeham distribution verifies the relation: $\mu_x = a + b \times c^x$ where μ_x is the instantaneous rate of death at age *x*. The parameter *a* can be interpreted like an accidental incidence; the coefficient $b \times c^x$, corresponding to an ageing of the population, has the rate of death growing in an exponential way. Taking into account the growth of rates of death with age, one must have a constant *c* higher than 1 as well as a positive *b*. One then obtains:

$$p_x = \mathbf{exp}\left[-\int_x^{x+1} \mu_y dy \right] = \mathbf{exp}\left[-\int_x^{x+1} (a + b \times c^y) dy \right] = \mathbf{exp}(-a)\mathbf{exp}\left[-\frac{b}{\mathbf{ln}(c)} c^x (c-1) \right].$$

Let us write $s = \mathbf{exp}(-a)$ and $g = \mathbf{exp}\left(-\frac{b}{\mathbf{ln}(c)} \right)$, the function used for the fitting/smoothing the rates of discrete deaths is thus:

$$q_x = 1 - p_x = 1 - s \times g^{c^x(c-1)}.$$

From now on, we shall use this discretised version of the model.

5.2.1. Curve fitness to Makeham model

Before carrying out the adjustment itself, one seeks to validate the adequacy of this kind of function to the suggested situation. For that purpose, one observes that one has $\mathbf{ln}(1-q_x) = \mathbf{ln}(s) + c^x(c-1)\mathbf{ln}(g)$. For the qx close to zero[29], one can make the approximation, $\mathbf{ln}(1-q_x) \approx -q_x$ and thus:

$$-q_x = \mathbf{ln}(s) + c^x(c-1)\mathbf{ln}(g)$$

It results from it that $q_x - q_{x+1} = c^x(c-1)^2 \mathbf{ln}(g)$, which helps noticing, by taking the logarithm of this expression, that:

$$\mathbf{ln}(q_{x+1} - q_x) = x\,\mathbf{ln}(c) + \mathbf{ln}\left((c-1)^2 \mathbf{ln}\left(\frac{1}{g}\right)\right).$$

Under the assumption that mortality rates follow a Makeham distribution, the points $(x, y = \mathbf{ln}(q_{x+1} - q_x))$ are therefore aligned on a line of slope $\mathbf{ln}(c)$. Thus the idea is to proceed to a linear regression and to produce an analysis of the regression based on the following model:

	Degree of freedom	**Sum of squares**	**Average of squares**	**F**	**Critical value of F**
Regression	1				
Residues	*n-1*				
Total	*N*				

Tab. 1 - Variance analysis

One possibly concludes with the adjustment by a line on the interval $x_0 \le x \le x_1$ by carrying out a Fisher test (with a threshold to be defined, for example 5%). Reminder: the Fisher test statistic used to test the global significance of a linear regression[30] model $y_i = \beta_0 + \beta_1 x_1 + .. + \beta_{p-1} x_{p-1} + \varepsilon_i$ is

$$F_{p-1} = \frac{R^2}{1-R^2} \frac{n-p}{p-1}$$

[29] As an indication, the mortality rate at age 60 is in France about 0,50% pour women, and 1,20% for men (source: TV/TD 99/01).

[30] To validate the fact that regression coefficients are not all zeros.

with $R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$. This statistic follows a Fisher distribution $(1, p-1)$.

5.2.2. Fitting though maximum-likelihood method

Once validated the fact that an adjustment of the Makeham type can prove to be relevant, one seeks to estimate its parameters though the method of maximum-likelyhood. It will be noted incidentally that the maximum likelihood determined in the discretised model under study is not identical to the direct maximum likelihood which is obtained based on the continuous basic model.

One notes $\theta = (s, g, c)$ the vector of parameters to be determined, and $q_x(\theta) = 1 - s \times g^{c^x(c-1)}$ the Makeham function to be fitted. One seeks the vector of parameter which provides the best possible fit to the $\hat{q}_x$ curve (raw observed incidence rates).

One can also simply use the MS Excel solver. In all cases, the algorithm converges towards the true value of the parameter only provided that the initial value θ_0 is rather close to θ.

It is thus advisable to determine acceptable initial parameters values. One can use for that purpose the property established in 5.2.1 above, about the alignment of points $(x, y = \ln(q_{x+1} - q_x))$; the ordinate at origin and the slope determine g and c, and one can solve based on the relationship[31] $\ln(p_x) - c^x(c-1)\ln(g) = \ln(s)$.

In order to test whether the coefficients of the obtained Makeham function are not significantly equal to zero, a Student test is carried out which consists in comparing the ratio (estimate/standard deviation) with a Student distribution of m degrees of freedom (m = number of ages observed - 3 estimated parameters). Finally, Khi-2 tests are carried out, based on the statistic $W = \sum N_x \frac{(\hat{q}_x - q_x)^2}{q_x}$, q_x being the theoretical rate of death of the model at age x. The asymptotic distribution of W is a $\chi^2(p-3-1)$, where p indicates the number of ages under study. It is advisable in practice to handle with precaution the Khi-2 test, the asymptotic distribution being a $\chi^2(p-k-1)$, p being the number of classes and k the number of parameters of the model, only because in that case the estimator is of maximum likelyhood. For other methods of determination of the parameter, this result is not true anymore in general (see Fischer [1924]).

The following graph shows the Makeham fitting carried out by pseudo-maximum likelihood (by standardising populations under risk at each age) on the age bracket 40-105 of the TF 00-02 table.

[31] The left hand side part of the equation must therefore only slightly depend on x.

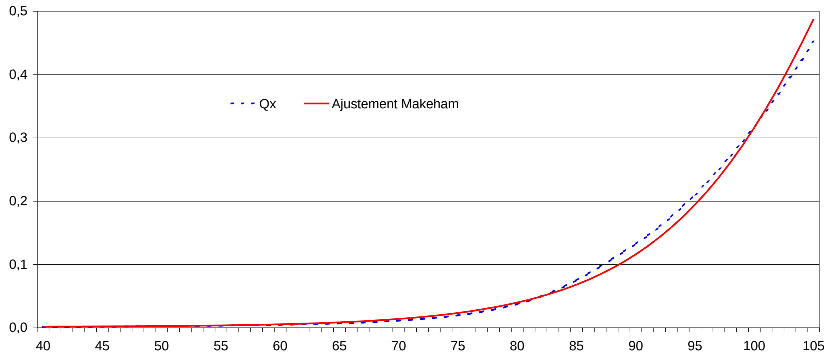


Figure 1 - *Fitting of a Makeham distribution to TH00-02*

5.3. Thatcher model

In practice, the Makeham model leads to an over-estimate of conditional rates of deaths at high ages. In order to correct this over-estimate, Thatcher [1999] proposed another (similar) model in writing $\mu(t) = \alpha + \frac{\beta \times e^{\gamma t}}{1+\beta \times e^{\gamma t}}$. While $v_{\beta,\gamma}(u) = 1 + \beta \, \mathbf{exp}(\gamma u)$ one notices that, $\frac{\beta \, \mathbf{exp}(\gamma u)}{1+\beta \, \mathbf{exp}(\gamma u)} du = \frac{1}{\gamma} \frac{dv}{v}$ which leads after some processing to $S_\theta(t) = e^{-\alpha t} v_{\beta,\gamma}(t)^{-\frac{1}{\gamma}}$. One deduces in particular:

$$E(T_\theta) = \int_0^{+\infty} e^{-\alpha t} v_{\beta,\gamma}(t)^{-\frac{1}{\gamma}} dt = \int_0^{+\infty} e^{-\alpha t} \left(1 + \beta e^{\gamma t}\right)^{-\frac{1}{\gamma}} dt .$$

Which leaves $q_x = 1 - \mathbf{exp}\left(-\int_x^{x+1} \mu(y) dy\right)$, yielding:

$$q_x = 1 - e^{-\gamma} \left(\frac{v_{\beta,\gamma}(x+1)}{v_{\beta,\gamma}(x)}\right)^{-\frac{1}{\beta}} .$$

The outcome is adjustments that are close to those obtained with the Makeham model, but with slightly lower rates:

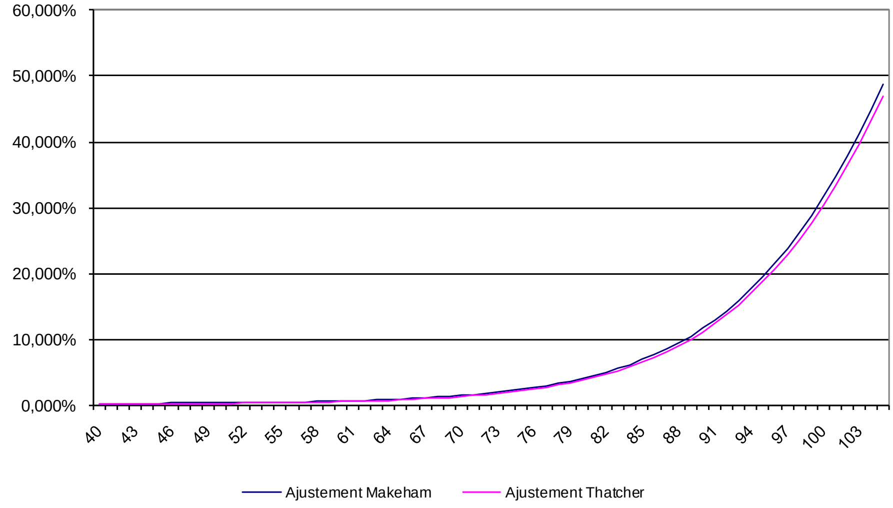


Figure 2 - *Comparison of Makeham and Thatcher fitting to TH00-02*

5.4. Raw rates fitting based on Logits

The estimate of mortality rates q_x is constrained by the fact that one must have $q_x \in [0,1]$; while posing $\mathbf{lg}(x) = \mathbf{ln}\left(q_x / (1-q_x)\right)$, the logit of the rate of death, one is brought back to a "free" value in $]-\infty,+\infty[$ and one can then use the techniques of linear regression on explanatory variables. The most simple candidates as explanatory variables can be age and logit of rates of death of a reference table.

5.4.1. The logistic function

The logistic function is by definition $\mathbf{lg}(x) = \mathbf{ln}\left(\frac{x}{1-x}\right)$ and is defined on $]0,1[$; it is increasing within this interval:

$$\frac{d}{dx}\mathbf{lg}(x) = \frac{1}{x(1-x)}.$$

Also:

$$\frac{d^2}{dx^2}\mathbf{lg}(x) = \frac{2x-1}{x^2(1-x)^2}.$$

Hence on the interval $]0,1/2[$ the function $\mathbf{lg}(x)$ is concave. Reminder: according to the Jensen inequality, if f is convex, then $\mathbf{E}\, f(X) \geq f(\mathbf{E}(X))$. This yields that, in a zone where rates of death are small, and if the rate of death was estimated by $\hat{q}_x$ assumed unbiased, then:

$$\mathbf{E}\,\mathbf{lg}(\hat{q}_x) \leq \mathbf{lg}(q_x).$$

In other words, the obtained empirical logits are negatively biased (they underestimate true logits). As the function $\mathbf{lg}(x)$ (and its inverse) is increasing, by underestimating theoretical logits, this approach underestimates theoretical rates of death. This conclusion is reversed for exit/death rates higher than ½.

Within the framework of a fitting of $\hat{y}_x = \mathbf{lg}(\hat{q}_x)$, one obtains the adjusted rates of deaths through the inverse transformation $y \to \frac{e^y}{1+e^y}$. The presence of exponentials in this expression leads to an important amplification of the bias of estimate mentionned above. In the case of mortality risk, a logits fitting model leads to underestimating rates of death in proportions which can be material (typically 5 % to 10 %).

Models using logits of rates of death must therefore be used with prudence in the context of risks in case of death. They integrate on the contrary a safety margin in the context of risks in case of survival.

The use of logistic fitting within the framework of qualitative variables is more and more "legitimated" by the following remark: the quantity $c_x = \frac{q_x}{1-q_x}$ is the ratio of the probability of "success" over the probability of "failure" within the framework of a Bernoulli experiment; an interpretation is therefore that there is "c_x time more chances for death to occur, than for it not to occur". It is then relatively natural to seek to explain the level reached by c_x using explanatory variables, and because of positivity of c_x. The simplest model that one can think of is obtained in writing $c_x = \mathbf{exp}\left({}^t\theta z_x\right)$, with z_x the vector of explanatory variables.

One then falls back on the context of a generalised linear model[32] with a logistic link function:

$$\mathbf{lg}(q_x) = {}^t\theta z_x + \varepsilon_x,$$

which allows the use the standard procedures for estimation, available in most specialised software (once the distribution of ε_x is specified). One can also note that this model can be written in the form:

$$q_x(\theta) = \frac{e^{\theta' z_x}}{1+e^{\theta' z_x}}.$$

One can thus seek the solution by the method described above, of discrete maximum likelihood.

5.4.2. Logistic fittings

The base model for logistic fitting starts from the observation that on a broad spectrum the logit of the rates of death presents a linear trend; the following modelling is proposed, the simplest version of the module presented *infra* assuming that the age constitutes a relevant explanatory variable:

$$\mathbf{lg}(\hat{q}_x) = a + bx + \varepsilon_x$$

32 See NELDER and WEDDERBURN [1972] for the original presentation, and PLANCHET et *al.* [2011] for an introduction.

Where ε is a Gaussian iid noise; one therefore simply regresses the logits from the rates of death on age. The reverse transformation of the logit being $y \to \frac{e^y}{1+e^y}$, the model $\mathbf{lg}(q_x) = a + bx$ is written in an equivalent way:

$$q_x = \frac{ce^{dx}}{1+ce^{dx}}$$

by stating that $c = e^a$ and $d = b$. An alternative approach to the linear regression $\mathbf{lg}(\hat{q}_x) = a + bx + \varepsilon$ therefore consists in carrying out an estimate by maximum likelihood in the parametric model $q_x = \frac{ce^{dx}}{1+ce^{dx}}$. This approach avoids *a priori* the underestimating effect of mortality rates associated with the linear regression approach, the rate of death being the modelled variable (but the maximum likelihood estimator has however no reason to be unbiased).

The determination of the survival function and the hazard function, related to one another by $S(t) = \mathbf{exp}\left(-\int_0^t \mu(s)ds\right)$ requires to make assumptions. Indeed, the relationship $q(x) = 1 - \frac{S(x+1)}{S(x)}$ leads in the general case to the following constraint on the hazard function:

$$-\mathbf{ln}(1-q_x) = \int_x^{x+1} \mu(s)ds$$

In the discrete model specified until now X is *a priori* an integer. One thus needs a rule of passage of discrete time to continuous time. Various approaches may be used (Balducci, constancy of chance rates by segment, *etc.*). If one chooses the assumption of constancy of the hazard function between two integer values, one finds that the hazard function is a staircase function with, at the integer points:

$$\mu_x = \frac{cde^{dx}}{1+ce^{dx}}.$$

In practice it can appear that the curve of raw rates $\hat{q}_x$ presents a stall starting from a pivot age, which indicates an acceleration of the incidence. In this context, one is brought to seek an adjustment via models of the logistic type built on fittings of $\mathbf{ln}\left(\hat{q}_x / (1-\hat{q}_x)\right)$ on age, which will thus play the part of explanatory variable.

One seeks to adjust the raw rates on a function of the form:

$$\mathbf{ln}\left(\hat{q}_x / (1-\hat{q}_x)\right) = ax + b + c \times 0 \wedge (x - x_c)$$

$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} \mathbf{1}_{]t;\infty[}(T_x)$$

Where x_c is a "pivot age" beyond which mortality accelerates (standard *logit* model). In other words, one writes the logistic regression model according to:

$$\mathbf{ln}\left(\hat{q}_x / \left(1-\hat{q}_x\right)\right) = ax + b + c \times 0 \wedge \left(x - x_c\right)$$

where (ε X) form a Gaussian white noise. One can generalise these models by writing:

$$\mathbf{ln}\left(q_x / \left(1-q_x\right)\right) = ax + b + c \times 0 \wedge \left(x - x_c\right)^{\lambda} + \varepsilon_x$$

If one does not have sufficient data to properly structure the complete table, one can imagine using the structure of an existing reference table, simply positioning the mortality of the considered group with respect to the reference. When one wishes to position a table with respect to another one, it can appear natural to carry out the regression of logits of the raw rates on logits of the reference table, which leads to the following model:

$$\mathbf{ln}\left(\hat{q}_x / \left(1-\hat{q}_x\right)\right) = a\,\mathbf{ln}\left(q_x / \left(1-q_x\right)\right) + b + \varepsilon_x$$

5.4.3. Parameters estimation

In the case of the regression model based on age, the estimate can be carried out according to the following procedure: before the pivot age x_c, one carries out a linear regression of $\mathbf{ln}\left(\hat{q}_x / \left(1-\hat{q}_x\right)\right)$ on *x*, then beyond one makes one second regression (nonlinear) of $\mathbf{ln}\left(\hat{q}_x / \left(1-\hat{q}_x\right) - \left(ax+b\right)\right)$.

In the case of a regression of logits of the raw rates on logits of a reference table, estimation is carried out *via* a traditional ordinary least squares estimate.

5.5. Confidence intervals for raw rates

The first stage of the construction of the mortality table is the estimate of raw rates for each age. It is appropriate, beyond point estimation, to have an idea of the precision of the estimate carried out. It depends on two factors:

- ✓ risk population/exposure N_x,
- ✓ the level of mortality rate to be estimated q_x.

Indeed, the larger N_x and q_x, the better the precision. The precision will be measured by the width of the confidence interval. To determine it, two methods are available:

- ✓ the use of the Gaussian approximation, if enough observations are available;
- ✓ the calculation of the finite distance interval, which is *a priori* possible since the distribution of $\hat{q}_x$ is known.

Initially, one thus seeks which type of confidence interval shall be used. For that purpose, one notices that a relationship exists that binds the uncertainty of the estimate, the number of observations and the degree of confidence of the desired interval:

$$\Delta p = u_{a/2} \sqrt{\frac{f(1-f)}{N}}$$

where f is the value around which the interval is built. (*i.e.* f is equal to the estimated value $\hat{q}_x$) and u_p indicates the quantile of order p of the standard normal distribution.

Example

If the value to be estimated q_x is worth 0.2, if one wishes an interval at 95 % for a precision of about 0.01. It is necessary to have:

$$N_x = \frac{f(1-f)}{\Delta p^2} u_{a/2}{}^2 = \frac{0,2 \times 0,8}{0,01^2} \times 1,96^2$$

that is to say approximately:

$$N_{\approx} \approx 6\,150$$

If only 3’000 observations are available, one will turn to the finite distance confidence interval. At the 95 % level, considering the most unfavorable case of a frequency of ½, one obtains a worst-case (rather broad) necessary number of observations to obtain the precision Δp by $N = \frac{1}{\Delta p^2}$.

5.5.1. Asymptotic confidence intervals

N_x indicate the exposure to the risk at age x, D_x the number of deaths in the year of the people of age x, and q_x was estimated by $\hat{q}_x$. According to the central limit theorem:

$$\sqrt{N_x} \frac{q_x - \hat{q}_x}{\sqrt{\hat{q} \times (1-\hat{q})}} \underset{N \to \infty}{\to} N(0,1)$$

The asymptotic confidence interval of level α for q_x is thus given by:

$$I_\alpha = \left[\hat{q}_x - u_{\alpha/2} \sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{N_x}}, \hat{q}_x + u_{\alpha/2} \sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{N_x}} \right]$$

The limit of this approach is that it only allows to build confidence intervals for a specific age, but does not allow to frame the rates of death on a fixed age range with a known degree of confidence. One now wishes to frame the rates of death simultaneously on all ages x of a range of ages $[x_0, x_0 + n]$ (where n is a positive integer). The framing of rates of death thus now corresponds to a confidence “band” – gone is the confidence interval at one “point” only.

One wishes to build bands of confidence for rates of death, and not for survival functions. In practice, one seeks $t(\hat{q}_x)$ such that $P\left(q_x \in \hat{q}_x \pm t(\hat{q}_x), \forall x \in [x_0, x_0 + n]\right) = 1 - \alpha$. For this purpose, one uses the Sidak estimation method, which is based on the principle of the test threshold inflation, when the number of tests increases (*cf.* for example Abdi [2007]).

For memory, a band of confidence on the degree of confidence $1-\alpha$ on the ages range $[x_0, x_0+n]$ can be presented like a collection of confidence intervals for the various ages $x \in [x_0, x_0+n]$ built so as to have a simultaneous interval of probability equal to $1-\alpha$. That is to say $P\left(q_x \in \hat{q}_x \pm t(\hat{q}_x), x = x_0\right) = 1-\beta$ the interval of probability of level $1-\beta$ (with $\beta \in]0,1[$) for q_x at the age $x = x_0$. The simultaneous probability to frame the rates of death q_x at both ages $x = x_0$ and $x = x_0+1$is then $(1-\beta)^2$, by assuming the independent framing on these two ages. By repeating the operation so as to include all the ages of $[x_0, x_0+n]$, it appears then, always under the assumption of independence, that the simultaneous probability to frame the rates of death q_x for the various ages $x \in [x_0, x_0+n]$ is $(1-\beta)^{n+1}$.

On these bases, one can thus build a band of confidence to the threshold α on the age bracket $[x_0, x_0+n]$, by constituting specific confidence intervals for each age $x \in [x_0, x_0+n]$ with the threshold:

$$\beta = 1-(1-\alpha)^{1/(n+1)},$$

since in this case one has $(1-\alpha) = (1-\beta)^{n+1}$. Also, an approximation of the band of confidence allowing simultaneously to frame the rates of death on all the ages $[x_0, x_0+n]$ based on Sidak method is:

$$P\left(q_x \in \hat{q}_x \pm u_{\beta/2} \sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{R_x}}, \forall x \in [x_0, x_0+n] \right) = 1-\alpha,$$

with $\beta = 1-(1-\alpha)^{1/(n+1)}$. The intervals and bands of confidence above allow to frame the raw death rates under the fluctuations of sampling, respectively for a given age or on an age bracket. The bands of confidence are by construction broader than the confidence intervals.

5.5.2. Finite distance confidence intervals

Here one considers the case where N_x is not large enough to be able to use the central limit theorem. Based on the fact that $P(D_x = k) = C_{N_x}^k \times q_x^k \times (1-q_x^k)$ one calculates the exact finite distance confidence interval. One thus seeks m_α such that:

$$P\left[\hat{q}_x - m_\alpha \le q_x \le \hat{q}_x + m_\alpha\right] \ge 1-\alpha$$

By multiplying by N_x the terms of the inequality of which one wants to calculate the probability, one finds that one must have:

$$\sum_{k=\left[N(\hat{q}_x - m_a)\right]}^{\left[N(\hat{q}_x + m_a)\right]+1} P[D_x = k] \ge P\left[\hat{q}_x - m_\alpha \le q_x \le \hat{q}_x + m_\alpha\right] \ge 1-\alpha$$

One can think of an iterative procedure to find m_α :

step n°0

$P(D_x = k)$ is calculated with $k = N_x \hat{q}_x$, and compared to $1-\alpha$. If $P(D_x = k) < 1-\alpha$, go to next step.

step n°j

$\sum_{k=N\hat{q}_x - j}^{N\hat{q}_x + j} P[D_x = k]$ is calculated and compared to $1-\alpha$. If $\sum_{k=N\hat{q}_x - j}^{N\hat{q}_x + j} P[D_x = k] < 1-\alpha$, go to step j+1.

final step

When this iterative process comes to a stop, we get $m_\alpha = \frac{j}{N_x}$.

6. References

ABDI H. [2007] « Bonferroni and Sidak corrections for multiple comparisons », *N. J. Salkind (ed.). Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA: Sage.

ANDERSEN P.K, GILL R.D. [1982], « Cox's regression model for counting processes : a large sample study », *The Annals of Statistics*, 10, 1100-1120.

BARTHOLOMEW D. [1957], « A problem in life testing », *J. Amer. Statist. Assoc.*, 52, 350-355

BARTHOLOMEW D. [1963], « The sampling distribution of an estimate arising in life testing », *Technometrics*, 5, 361-374.

CIARLET P.G. [1990] *Introduction à l'analyse numérique matricielle et à l'optimisation*. Paris : Dunod.

COX D.R. [1972] « Regression models and life-tables (with discussion) ». *J. R. Statist. Soc. Ser. B*, pages 187-220.

COX D.R. [1975] « Partial likelihood ». *Biometrika*, 62, pages 269-276.

DEMPSTER, A. P., LAIRD, N. M. RUBIN, D. B. [1977], « Maximum likelihood from incomplete data via the EM algorithm ». *Journal of the Royal Statistical Society*, B, 39, 1-38.

DROESBEKE J.J., FICHET B., TASSI P. [1989], *Analyse statistique des durées de vie*, Paris : Economica

DUPUY J.F. [2002], « Modélisation conjointe de données longitudinales et de durées de vie », Université Paris V, Thèse de doctorat.

FISHER R.A. [1924], « The conditions under wich χ^2 mesasures the discrepancy between observation and hypothesis ». *J. Roy. Statist. Soc.*, 87, 442-450.

FLEMING T.R., HARRINGTON D.P. [1991] *Counting processes and survival analysis*, Wiley Series in Probability and Mathematical Statistics. New-York : Wiley.

HILL C., COM-NOUGUÉ C., KRAMAR A., MOREAU T., O'QUIGLEY J., SENOUSSI R., CHASTANG Cl. [2000] *Analyse Statistique des Données de Survie*, Paris : Flammarion Sciences

NELDER J., WEDDERBURN R. [1972] « Generalized linear models », *Journal of Roy. Stat. Soc. B*, vol. 135, 370-384.

JUILLARD M., PLANCHET F., THÉROND P.E. [2011] *Modèles financiers en assurance. Analyses de risques dynamiques - seconde édition revue et augmentée*, Paris : Economica (première édition : 2005).

ROBERT C.P. [1996] *Méthodes de Monte-Carlo par chaînes de Markov*, Paris : Economica

SAPORTA G., [1990] *Probabilités, analyse des données et statistiques*, Paris : Editions Technip

THATCHER A.R. [1999] « *The Long-term Pattern of Adult Mortality and the Highest Attained Age.* », Journal of the Royal Statistical Society 162 Part 1: 5-43.

THERNEAU T. M., GRAMBSCH P. M., FLEMING T. R. [1990] « Martingale-based residuals for survival models », Biometrika, vol. 77, n°1, pp. 147-160.

THERNEAU T. M., GRAMBSCH P. M. [1990] *Modeling Survival Data: Extending the Cox Model*, Series: Statistics for Biology and Health, New-York: Springer