Événements récurrents, risques concurrents et causes manquantes

Jean-Yves Dauxois (Univ. Franche-Comté)

Travaux en collaboration avec Laurent Bordes (Univ. Pau), Pierre Joly (Univ. Bordeaux II et INSERM), Sophie Sencey (INSEE)

27 Août 2008

Plan

Plan

Durée de vie: temps d'attente de la guérison ou de la mort d'un patient, temps d'attente de la panne d'un matériel, durée du chômage, durée de vie des entreprises, durée de fidélité des clients...

Modélisation par une v.a. T de support $\mathbb{R}^+$, de f.d.r. F, de fonction de survie $\bar{F} = 1 - F$ et de fonction de risque cumulé définie par :

$$\Lambda(t) = \int_{]0,t]} \frac{dF(s)}{\bar{F}(s^-)},$$

pour tout t positif.

Dans le cas où T est de loi absolument continue, on peut définir la fonction de risque instantané, en tout t positif, par :

$$\lambda(t) = \lim_{h \longrightarrow 0^+} \frac{P(T \in [t, t+h[| T \geq t)}{h} = \frac{f(t)}{\bar{F}(t)},$$

où f est la densité de la loi de T.

Quand on observe un échantillon $T_1, \ldots, T_n$ de la durée de vie, on sait bien sûr faire...

Quand on observe un échantillon $T_1, \ldots, T_n$ de la durée de vie, on sait bien sûr faire...

Observations censurées

Souvent on observe seulement un échantillon $(X_i, \delta_i)_{i=1,\ldots,n}$ des v.a.

$$\left\{ \begin{array}{rcl} X & = & T \wedge C \\ \delta & = & I(\{T \leq C\}) \end{array} \right. .$$

La v.a. C, supposée indépendante de T, modélise la censure. Soit G sa f.d.r. et $\bar{G}$ sa fonction de survie.

Estimation non paramétrique

Définissons les processus de comptage $N_.$ et du nombre à risque Y, en tout t positif, respectivement par :

$$N_.(t) = \sum_{i=1}^{n} I(\{X_i \leq t, \delta_i = 1\})$$

et

$$Y(t) = \sum_{i=1}^{n} I(\{X_i \geq t\}).$$

L'estimateur de Nelson-Aalen $\widehat{\Lambda}$ de la fonction de risque cumulé est alors donné, pour tout t positif, par

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN_\cdot(s)}{Y(s)} = \sum_{i:X_{(i)} \leq t} \frac{\Delta N_\cdot(X_{(i)})}{Y(X_{(i)})},$$

où $X_{(1)} \leq \cdots \leq X_{(n)}$ sont les n statistiques d'ordre.

L'estimateur de Kaplan-Meier $\widehat{\bar{F}}$ de la fonction de survie est lui défini par :

$$\widehat{\bar{F}}(t) = \prod_{s \leq t} (1 - \widehat{\lambda}(s)) = \prod_{i:X_{(i)} \leq t} \left(1 - \frac{\Delta N_\cdot(X_{(i)})}{Y(X_{(i)})}\right).$$

Plan

Ici, on suppose qu'il y a p ($p \geq 2$) causes différentes de mort (panne), notées $D_1, \ldots, D_p$.

Ici, on suppose qu'il y a p ($p \geq 2$) causes différentes de mort (panne), notées $D_1, \ldots, D_p$.

On note T_j, pour $j = 1, \ldots, p$, la durée de vie de l'individu dans l'hypothétique situation où seulement la cause D_j agirait.

Ici, on suppose qu'il y a p ($p \geq 2$) causes différentes de mort (panne), notées $D_1, \ldots, D_p$.

On note T_j, pour $j = 1, \ldots, p$, la durée de vie de l'individu dans l'hypothétique situation où seulement la cause D_j agirait.

L'observation de la durée de vie pour un individu est donnée par

$$\left\{ \begin{array}{rcl} T & = & \min_{1 \leq j \leq p} T_j \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \end{array} \right. ,$$

où d indique donc la cause ayant entraîné la mort (panne) de l'individu (du matériel).

Sans hypothèse supplémentaire, les fonctions λ_{T_j} et F_{T_j}, pour $j = 1, \ldots, p$, ne sont pas estimables.

Sans hypothèse supplémentaire, les fonctions λ_{T_j} et F_{T_j}, pour $j = 1, \ldots, p$, ne sont pas estimables.

Deux solutions (parmi d'autres)

Sans hypothèse supplémentaire, les fonctions λ_{T_j} et F_{T_j}, pour $j = 1, \ldots, p$, ne sont pas estimables.

Deux solutions (parmi d'autres)

- Soit on s'intéresse à l'estimation des fonctions d'incidence cumulée (CIF)

$$F_j(t) = P(T \leq t, d = j),\ j = 1, \ldots, p; 0 < t < +\infty.$$

ou aux fonctions de risque spécifique

$$\lambda_j(t) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h[, d = j | T \geq t)}{h},\ j = 1, \ldots, p.$$

Sans hypothèse supplémentaire, les fonctions λ_{T_j} et F_{T_j}, pour $j = 1, \ldots, p$, ne sont pas estimables.

Deux solutions (parmi d'autres)

- Soit on s'intéresse à l'estimation des fonctions d'incidence cumulée (CIF)

$$F_j(t) = P(T \leq t, d = j),\ j = 1, \ldots, p; 0 < t < +\infty.$$

ou aux fonctions de risque spécifique

$$\lambda_j(t) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h[, d = j | T \geq t)}{h},\ j = 1, \ldots, p.$$

- Soit on suppose les durées de vie $T_1, \ldots, T_p$ indépendantes et on peut alors estimer λ_{T_j} et F_{T_j}, pour $j = 1, \ldots, p$. C'est l'hypothèse que l'on fera dans la dernière partie de cet exposé.

Souvent on suppose également la présence d'une censure à droite C indépendante de f.d.r. G.
C'est à dire qu'au lieu d'observer (T, d), on ne peut observer que

$$\left\{ \begin{array}{lcl} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{array} \right. ,$$

où C est une v.a. indépendante des $T_1, \ldots, T_p$.

Souvent on suppose également la présence d'une censure à droite C indépendante de f.d.r. G.
C'est à dire qu'au lieu d'observer (T, d), on ne peut observer que

$$\left\{ \begin{array}{lcl} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{array} \right. ,$$

où C est une v.a. indépendante des $T_1, \ldots, T_p$.
On peut montrer que l'on a dans ce cas, pour tout $t \geq 0$:

$$\Lambda_j^C(t) = \Lambda_j(t),$$

ce qui permet d'estimer Λ_j, les fonctions F_j, pour $j = 1, \ldots, K$, et $\bar{F}$ la fonction de survie de T.

Estimations non paramétrique

Estimations non paramétrique

Notant, pour $j = 1, \ldots, p$,

$$N_j(t) = \sum_{i=1}^{n} I\{X_i \leq t, d_i = j\} \text{ et } Y(t) = \sum_{i=1}^{n} I\{X_i \geq t\},$$

les estimateurs de Nelson-Aalen des fonctions de risque cumulé spécifique sont donnés, pour $j = 1, \ldots, p$, par :

$$\hat{\Lambda}_j(t) = \int_0^t \frac{dN_j(u)}{Y(u)}.$$

Estimations non paramétrique

Notant, pour $j = 1, \ldots, p$,

$$N_j(t) = \sum_{i=1}^{n} I\{X_i \leq t, d_i = j\} \text{ et } Y(t) = \sum_{i=1}^{n} I\{X_i \geq t\},$$

les estimateurs de Nelson-Aalen des fonctions de risque cumulé spécifique sont donnés, pour $j = 1, \ldots, p$, par :

$$\hat{\Lambda}_j(t) = \int_0^t \frac{dN_j(u)}{Y(u)}.$$

Enfin, les f.i.c. peuvent être estimées par les estimateurs de Aalen-Johansen, pour $j = 1, \ldots, p$:

$$\widehat{F}_j(t) = \int_0^t \widehat{\bar{F}}(u^-) \frac{dN_j(u)}{Y(u)}$$

Plan

Plan

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation service between 1995 and 1999.

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation service between 1995 and 1999.

- There are 67 different types of infection. The most frequent are: urinary tracts, pneumonias or septicaemias and herpes.

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation service between 1995 and 1999.

- There are 67 different types of infection. The most frequent are: urinary tracts, pneumonias or septicaemias and herpes.
- Each patient can develop several infections, the maximum number observed being 13, each type of infection being able to occur several times.

Data set on nosocomial infections

We dispose of data reporting nosocomial infections of 7867 patients hospitalized in a French reanimation service between 1995 and 1999.

- There are 67 different types of infection. The most frequent are: urinary tracts, pneumonias or septicaemias and herpes.
- Each patient can develop several infections, the maximum number observed being 13, each type of infection being able to occur several times.
- The end of hospitalization in the reanimation service can be due to death or censoring, death being clearly dependent of the experienced infections.

Our aim

Two questions are of interest in this area.

Our aim

Two questions are of interest in this area.

- Is the occurrence rate of a given type of events increasing with time?

Our aim

Two questions are of interest in this area.

- Is the occurrence rate of a given type of events increasing with time?
- Is the instantaneous probability of experiencing an event of a given type always greater than the one of an other type?

Plan

Notation

We consider a setting of competing risks for recurrent events in presence of a terminal event. That is, each type is concurring with the others and we focus on the frequency of each type.

Notation

We consider a setting of competing risks for recurrent events in presence of a terminal event. That is, each type is concurring with the others and we focus on the frequency of each type.

To simplify notation we restrict our attention to the case where only two different types of event are concurring.

Notation

We consider a setting of competing risks for recurrent events in presence of a terminal event. That is, each type is concurring with the others and we focus on the frequency of each type.

To simplify notation we restrict our attention to the case where only two different types of event are concurring.

Write $N_j^*(t) =$ total number of events of type j experienced by an individual up to time t, for $j = 1, 2$.

Individuals are subject to a terminal event at time D, such that they cannot experience any further event after. This time, with survival function $S(t) = \mathbb{P}(D > t)$, is supposed to be dependent from the recurrent event processes $N_j^*(t)$, $j = 1, 2$.

Individuals are subject to a terminal event at time D, such that they cannot experience any further event after. This time, with survival function $S(t) = \mathbb{P}(D > t)$, is supposed to be dependent from the recurrent event processes $N_j^*(t)$, $j = 1, 2$.

Finally, the presence of an independent random right-censoring mechanism is also allowed and denoted by the random variable C.

This means that the observation are:

$$\begin{aligned} N_j(t) = & \quad N_j^*(t \wedge C), \\ X = & \quad D \wedge C, \end{aligned}$$

for $j = 1, 2$, and

$$\delta = I(D \leq C).$$

This means that the observation are:

$$\begin{aligned} N_j(t) = & \quad N_j^*(t \wedge C), \\ X = & \quad D \wedge C, \end{aligned}$$

for $j = 1, 2$, and

$$\delta = I(D \leq C).$$

The overall recurrent event process $N = N_1 + N_2$ counts the number of events of all types experienced by a subject up to time t.

Now, let us define the specific (resp. overall) mean function, for $j = 1, 2$, by:

$$\mu_j(t) = \mathbb{E}(N_j^*(t)) \text{ (resp. } \mu(t) = \mathbb{E}(N^*(t))).$$

One can define the specific mean frequency functions by

$$\begin{aligned} \mu_j'(t) &= \lim_{\Delta_t \to 0} \frac{1}{\Delta_t} \mathbb{P}\left(N_j^*(t+\Delta_t) - N_j^*(t) = 1\right) \\ &= \text{infinitesimal probability of observing an event} \\ &\quad \text{of type } j \text{ at time } t \end{aligned}$$

And the overall mean frequency function is $\mu'(t) = \mu_1'(t) + \mu_2'(t)$

The probability of observing an event of type j at time t, $j = 1, 2$, given that an event occurs at time t, is given by

$$p_j(t) = \frac{\mu_j'(t)}{\mu'(t)}.$$

Finally note that we have:

$$\mu_j(t) = \int_0^t S(u-)dR_j(u), \tag{1}$$

where $dR_j(t) = \mathbb{E}(dN_j^*(t)|D \geq t)$, for $j = 1, 2$.

Estimators and their weak convergence

Suppose that we observe a sample $(N_{1i}, N_{2i}, X_i, \delta_i)$, $i = 1, \ldots, n$.

The survival function of D is easily estimated by the Kaplan-Meier estimator $\widehat{S}$.

$$\widehat{S}(t) = \prod_{i:X_{(i)} \leq t} \left(1 - \frac{\sum_{j=1}^{n} \mathbb{1}_{\{X_j = X_{(i)}, \delta_j = 1\}}}{\bar{Y}(X_{(i)})} \right),$$

where

$$\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t) \equiv \sum_{i=1}^{n} I(X_i \geq t).$$

Moreover, a Nelson-Aalen type estimator of R_j is given by

$$\widehat{R}_j(t) = \sum_{i=1}^{n} \int_0^t J(u) \frac{dN_{ji}(u)}{\bar{Y}(u)},$$

where $J(t) = I(\bar{Y}(t) > 0)$.

Thus, estimators of the specific mean functions are, for $j = 1, 2$:

$$\widehat{\mu}_j(t) = \int_0^t \widehat{S}(u-) d\widehat{R}_j(u).$$

Let $\pi(t) = \mathbb{P}(X > t)$, $t \geq 0$, and define τ such that $\pi(\tau-) > 0$.

Theorem

As $n \to \infty$,

$$n^{1/2}\left(\widehat{\mu}_1 - \mu_1, \widehat{\mu}_2 - \mu_2\right) \xrightarrow{\mathcal{D}} (G_1, G_2)$$

in $D^2[0, \tau]$, *where* $G = (G_1, G_2)$ *is a mean-zero Gaussian process with covariance function* $\xi_{jk}(s,t) = cov(G_j(s), G_k(t))$

The covariance function $\xi_{jk}(s,t) = cov(G_j(s), G_k(t))$ of the limiting process can be consistently estimated.

Plan

In this section, our aim is to derive a test of

$$H_0 : p_1(t) \text{ is constant}$$

against

$$H_1 : p_1(t) \text{ is an increasing function of } t.$$

The alternative hypothesis

$$H_2 = \bar{H}_0 : p_1(t) \text{ is not constant}$$

will be also considered.

One can show that

- for $t \in [0, \tau]$,

$$U(t) = \frac{\mu_1(t)\mu_2(t)}{2} - \int_0^t \mu_1(s) d\mu_2(s).$$

is a measure of the deviation from the null hypothesis H_0 on the interval $[0, t]$, when H_1 is in mind.

One can show that

- for $t \in [0, \tau]$,

$$U(t) = \frac{\mu_1(t)\mu_2(t)}{2} - \int_0^t \mu_1(s) d\mu_2(s).$$

  is a measure of the deviation from the null hypothesis H_0 on the interval $[0, t]$, when H_1 is in mind.
- A measure of deviation from H_0 can be defined by

$$\sup_{u \in [0,t]} |U(u)|$$

  when the alternative H_2 is in mind.

A plug in estimator of U is given by

$$\widehat{U}(t) = \left[\frac{\widehat{\mu}_1(t)\widehat{\mu}_2(t)}{2} - \int_0^t \widehat{\mu}_1(s) d\widehat{\mu}_2(s) \right], \ t \geq 0.$$

Theorem

For τ with $\pi(\tau) > 0$,

$$\sqrt{n}\left(\widehat{U}(.) - U(.)\right) \xrightarrow{\mathcal{D}} W_U$$

in $D[0,\tau]$ where W_U is a mean-zero Gaussian process defined, for all t, by:

$$\begin{aligned} W_U(t) &= \frac{\mu_2(t)G_1(t)}{2} - \frac{\mu_1(t)G_2(t)}{2} \\ &+ \int_0^t G_1(s)d\mu_2(s) - \int_0^t G_2(s-)d\mu_1(s). \end{aligned}$$

Our first test statistic, testing H_0 against H_1 (i.e. detecting if p_1 is increasing), is:

$$T_{1n} = \sqrt{n}\widehat{U}(\tau).$$

Since $U(\tau)$ is null under H_0, Theorem 2.2 leads to

$$T_{1n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{V}(W_U(\tau))).$$

Our second test statistic, detecting if p_1 is constant or not is defined by

$$T_{2n} = \sqrt{n} \sup_{t \in [0,\tau]} \left| \widehat{U}(t) \right|.$$

Still under H_0, one obtains

$$T_{2n} \xrightarrow{\mathcal{D}} \sup_{t \in [0,\tau]} |W_U(t)|.$$

Application to the nosocomial infections data set

Recall: 7867 patients hospitalized in a reanimation service between 1995 and 1999. They may have developed nosocomial infections of different importance, as septicemia, pneumonia, herpes or urinary tract infections.

Our aim is to detect if the probability of contracting a particular type of event, knowing that the patient experienced an event at time t, is decreasing with time.

If septicemia is the cause of primary interest, we obtain a p-value of nearly 0 in the test of $H_0 : p_{sept}\ is\ constant$ against $H_3 : p_{sept}\ is\ decreasing$

If septicemia is the cause of primary interest, we obtain a p-value of nearly 0 in the test of $H_0 : p_{sept}\ is\ constant$ against $H_3 : p_{sept}\ is\ decreasing$

If we consider pneumonia, our test of $H_0 : p_{pneu}\ is\ constant$ against $H_3 : p_{pneu}\ is\ decreasing$ gives a p-value of 0.001.

If septicemia is the cause of primary interest, we obtain a p-value of nearly 0 in the test of $H_0 : p_{sept}\ is\ constant$ against $H_3 : p_{sept}\ is\ decreasing$

If we consider pneumonia, our test of $H_0 : p_{pneu}\ is\ constant$ against $H_3 : p_{pneu}\ is\ decreasing$ gives a p-value of 0.001.

Finally, if we consider urinary tract infections (UTI), we obtain a p-value of nearly 0 in our test of $H_0 : p_{UTI}\ is\ constant$ against $H_3 : p_{UTI}$ *is increasing*

Plan

Now our aim is to derive a test of

$$H_0 : \mu_1'(t) = \mu_2'(t), \ \forall t$$

against

$$H_1 : \mu_1'(t) \geq \mu_2'(t), \ \forall t.$$

With the notation

$$\bar{F}_j = \exp(-\mu_j),$$

for $j = 1, 2$, we can reformulate our hypotheses as

$$H_0 : \frac{\bar{F}_1(t)}{\bar{F}_2(t)} = 1$$

and

$$H_1 : \frac{\bar{F}_1(t)}{\bar{F}_2(t)} \text{ is an increasing function.}$$

Let

$$q(k_1, k_2) = \psi_1(k_1)\psi_2(k_2) - \psi_1(k_2)\psi_2(k_1),$$

with

$$\psi_j(k_i) = \int_0^\tau k_i(s)\bar{F}_j(s)ds$$

where k_1 and k_2 are positive weight function such that k_1/k_2 is an increasing function.

One can show that $q(k_1, k_2)$ is a measure of the deviation from H_0. Indeed the quantity is null under H_0 and positive under H_1.

Let K_i be an estimator of k_i, and let us estimate $\bar{F}_j(t)$ by

$$\widehat{\bar{F}}_j(t) = \exp(-\widehat{\mu}_j(t)).$$

We can then estimate ψ_{ji} by

$$\widehat{\psi}_j(K_i) = \int_0^\tau K_i(s)\widehat{\bar{F}}_j(s)ds.$$

Then a plug in estimator of the non-proportionality measure q is given by

$$\widehat{Q}(K_1, K_2) = \widehat{\psi}_1(K_1)\widehat{\psi}_2(K_2) - \widehat{\psi}_1(K_2)\widehat{\psi}_2(K_1).$$

Theorem

Assume that

$$\sup_{t\in[0,\tau]} |K_i(t) - k_i(t)| \xrightarrow{\mathbb{P}} 0.$$

Then

$$\sqrt{n}(\widehat{Q}(K_1, K_2) - q(k_1, k_2)) \xrightarrow{\mathcal{D}} N$$

in $D[0,\tau]$, where N is a mean-zero Gaussian random variable

From this theorem one can use the statistic

$$T_{3n} = \sqrt{n}\widehat{Q}(K_1, K_2)$$

for testing H_0 against H_1.

Plan

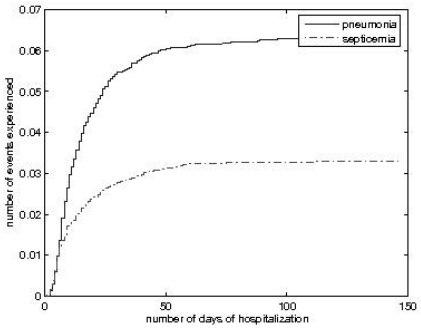

Figure: Plot of μ_{pneum} and μ_{septi}

Plan

Plan

On suppose que

- Les durées $T_1, \ldots, T_p$ sont indépendantes.

On suppose que

- Les durées $T_1, \ldots, T_p$ sont indépendantes.
- Modélisation semiparamétrique : modèle à hazard additif. Pour $j = 1, \ldots, p$, la loi de T_j, conditionnelle à un vecteur de covariables $Z \in \mathbb{R}^k$, est définie par la fonction de risque instantanée

$$\lambda_j(t|Z) = \lambda_{0j}(t) + \beta_j^T Z, \quad t \geq 0,$$

où λ_{0j} est la fonction de risque instantané de base de la $j$ème cause, et $\beta_j \in \mathbb{R}^k$ est le vecteur des paramètres de régression associé à la $j$ème cause (sous la contrainte d'avoir $\lambda_j(t|Z) \geq 0$).

Plan

On suppose de plus

- La présence d'une censure à droite C. C'est à dire qu'au lieu d'observer (T, d), on ne peut observer que

$$\left\{ \begin{array}{lcl} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{array} \right. ,$$

où C est une v.a. indépendante (conditonnellement à Z) des $T_1, \ldots, T_p$.

On suppose de plus

- La présence d'une censure à droite C. C'est à dire qu'au lieu d'observer (T, d), on ne peut observer que

$$\left\{ \begin{array}{lcl} X & = & \min(T, C) = \min(T_1, \ldots, T_p, C) \\ \delta & = & I(T \leq C) \\ d & = & \sum_{j=1,\ldots,p} j I(\{T = T_j\}) \text{ si } \delta \neq 0 \end{array} \right. ,$$

  où C est une v.a. indépendante (conditonnellement à Z) des $T_1, \ldots, T_p$.
- La cause de décès (ou de mort) d est parfois manquante (même quand la durée de vie T n'est pas censurée).

Hypothèses sur l'absence de marque

Hypothèses sur l'absence de marque

- Le mécanisme d'absence de la cause d est modélisé par une v.a. binaire

$$M = \begin{cases} 1 & \text{quand la cause } d \text{ est observée} \\ 0 & \text{sinon} \end{cases}$$

Hypothèses sur l'absence de marque

- Le mécanisme d'absence de la cause d est modélisé par une v.a. binaire

$$M = \begin{cases} 1 & \text{quand la cause } d \text{ est observée} \\ 0 & \text{sinon} \end{cases}$$

- On suppose que :

$$P(M = 1|X, Z, \delta = 1) = P(M = 1|\delta = 1) = \alpha \in [0, 1]$$

et

$$P(M = 0|X, Z, \delta = 0) = P(M = 0|\delta = 0) = 1,$$

Finalement les observations sont (X, δ, D, Z) où

$$D = \delta M \sum_{j=1}^{p} j 1(T_j = T).$$

Le vecteur aléatoire (X, δ, D), peut être vu, conditionnellement à Z, comme la réalisation d'un processus de Markov inhomogène à (p+3) états dont tous sauf un sont absorbants.

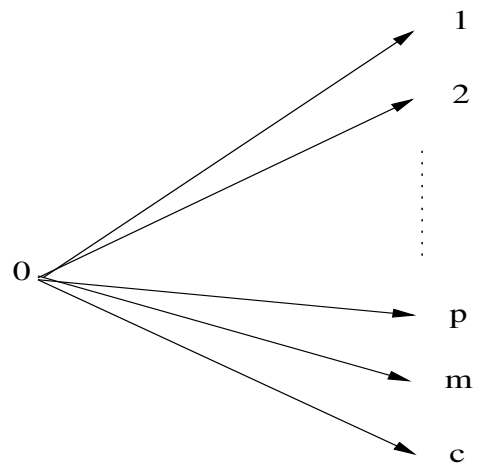

Figure: Graphe de markov associé à (X, δ, D)

En notant $\bar{\lambda}_{0x}(\cdot|Z)$ le taux de transition, conditionnellement à Z, pour la transition $0 \to x$ ($x \in \{1, \ldots, p, m, c\}$) de ce processus, nous obtenons :

$$\left\{ \begin{array}{lcl} \bar{\lambda}_{0j}(t) & = & \alpha\lambda_j(t|Z) = \alpha(\lambda_{0j}(t) + \beta_j^T Z) \text{ for } j \in \{1, \ldots, p\}, \\ \bar{\lambda}_{0m}(t) & = & (1-\alpha)\sum_{j=1}^p \lambda_j(t|Z) = (1-\alpha)\left(\lambda_m(t) + \beta_m^T Z\right), \\ \bar{\lambda}_{0c}(t) & = & \lambda_C(t), \end{array} \right.$$

où $\lambda_m = \sum_{j=1}^p \lambda_{0j}$ et $\beta_m = \sum_{j=1}^p \beta_j$.

Au paramètre α ou $1-\alpha$ près, les taux de transitions restent additifs (sauf pour la transition $0 \to c$).

Plan

- On note $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ les n replications i.i.d de (X, δ, D, Z).

- On note $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ les n replications i.i.d de (X, δ, D, Z).
- Pour $j \in \{1, \ldots, p, m\}$, on définit les processus de comptage :

$$\begin{aligned} N_{ij}(t) &= 1(X_i \leq t, D_i = j) \text{ for } j \neq m, \\ N_{im}(t) &= 1(X_i \leq t, \delta_i = 1, D_i = 0). \end{aligned}$$

- On note $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ les n replications i.i.d de (X, δ, D, Z).
- Pour $j \in \{1, \ldots, p, m\}$, on définit les processus de comptage :

$$\begin{aligned} N_{ij}(t) &= 1(X_i \leq t, D_i = j) \text{ for } j \neq m, \\ N_{im}(t) &= 1(X_i \leq t, \delta_i = 1, D_i = 0). \end{aligned}$$

- Pour simplifier les notations on écrit $m \equiv p + 1$ et on note Y_i le processus à risque défini par $Y_i(t) = 1(X_i \geq t)$.

- On note $(X_i, \delta_i, D_i, Z_i)_{1 \leq i \leq n}$ les n replications i.i.d de (X, δ, D, Z).
- Pour $j \in \{1, \ldots, p, m\}$, on définit les processus de comptage :

$$\begin{aligned} N_{ij}(t) &= 1(X_i \leq t, D_i = j) \text{ for } j \neq m, \\ N_{im}(t) &= 1(X_i \leq t, \delta_i = 1, D_i = 0). \end{aligned}$$

- Pour simplifier les notations on écrit $m \equiv p+1$ et on note Y_i le processus à risque défini par $Y_i(t) = 1(X_i \geq t)$.
- Pour $1 \leq i \leq n$ et $j \in \{1, \ldots, p+1\}$ les processus M_{ij} définis par

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(s) \bar{\lambda}_{0j}(s) ds, \quad t \geq 0,$$

sont des $\mathbb{F}$-martingales par rapport à la filtration $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ définie par :

$$\mathcal{F}_t = \sigma\{N_{ij}(s), Y_i(s); s \leq t; 1 \leq i \leq n, j \in \{1, \ldots, p+1\}\}.$$

Première estimation des paramètres de régression

En notant τ la borne supérieure de l'intervalle d'estimation, on estime naturellement α par :

$$\hat{\alpha} = \hat{\alpha}(\tau) = \frac{\sum_{i=1}^{n} 1(D_i > 0; X_i \leq \tau)}{\sum_{i=1}^{n} 1(\delta_i = 1; X_i \leq \tau)} = \frac{\sum_{j=1}^{p} N_{\cdot j}(\tau)}{N_{\cdot\cdot}(\tau)},$$

Les individus effectuant une transition $0 \longrightarrow j$, pour $j = 1, \ldots, p$, nous permettent d'estimer le paramètre β_j.
En suivant Lin et Ying (1994) et puisque nous avons un modèle additif, un estimateur de β_j est obtenu par la solution (explicite !) de l'équation $\mathcal{U}_j(\beta, \tau) = 0$ où

$$\mathcal{U}_j(\beta, \tau) = \sum_{i=1}^{n} \int_0^{\tau} \left[Z_i - \bar{Z}(s) \right] \left[dN_{ij}(s) - \hat{\alpha}\beta^T Z_i Y_i(s) ds \right],$$

avec

$$\bar{Z}(s) = \frac{\sum_{i=1}^{n} Y_i(s) Z_i}{\sum_{i=1}^{n} Y_i(s)}.$$

De même, la transition de $0 \longrightarrow p+1$ étant à risques additifs, on peut estimer β_{p+1} par $\hat{\beta}_{p+1}$ solution de $\mathcal{U}_m(\beta, \tau) = 0$ où

$$\mathcal{U}_m(\beta, \tau) = \sum_{i=1}^{n} \int_0^{\tau} \left[Z_i - \bar{Z}(s) \right] \left[dN_{i,p+1}(s) - (1-\hat{\alpha}) \beta^T Z_i Y_i(s) ds \right].$$

Amélioration de l'estimation des paramètres de régression

Pour $j = 1, \ldots, p, p+1$, nous avons un estimateur $\hat{\beta}_j$ de β_j.

Amélioration de l'estimation des paramètres de régression

Pour $j = 1, \ldots, p, p+1$, nous avons un estimateur $\hat{\beta}_j$ de β_j.

Puisque $\hat{\beta}_{p+1}$ est un estimateur de $\beta_{p+1} = \beta_1 + \cdots + \beta_p$, il est naturel de vouloir l'utiliser pour améliorer l'estimation des β_j.

Amélioration de l'estimation des paramètres de régression

Pour $j = 1, \ldots, p, p+1$, nous avons un estimateur $\hat{\beta}_j$ de β_j.

Puisque $\hat{\beta}_{p+1}$ est un estimateur de $\beta_{p+1} = \beta_1 + \cdots + \beta_p$, il est naturel de vouloir l'utiliser pour améliorer l'estimation des β_j.

Cherchons un estimateur $(\tilde{\beta}_1^T, \ldots, \tilde{\beta}_p^T)^T$ tel que :

$$\begin{pmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_p \end{pmatrix} = H \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \\ \hat{\beta}_{p+1} \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} & \cdots & H_{1p+1} \\ H_{21} & H_{22} & \cdots & H_{2p+1} \\ \vdots & \vdots & & \vdots \\ H_{p1} & H_{p2} & \cdots & H_{pp+1} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \\ \hat{\beta}_{p+1} \end{pmatrix},$$

où, pour $1 \leq i \leq p$ et $1 \leq j \leq p+1$, les matrices H_{ij} sont réelles de dimension $p \times p$.

Que doit vérifier la matrice H ?

Que doit vérifier la matrice H ?

Elle doit être naturellement telle que :

$$H \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

pour tout $1 \leq i \leq p$ et $\beta_i \in \mathbb{R}^p$.

Que doit vérifier la matrice H ?

Elle doit être naturellement telle que :

$$H \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

pour tout $1 \leq i \leq p$ et $\beta_i \in \mathbb{R}^p$.

On souhaite de plus que H minimise la fonction $\hat{q}(H)$ définie par

$$\hat{q}(H) = \text{trace}(H\hat{\Sigma}H^T),$$

où $\hat{\Sigma}$ est un estimateur de la matrice de covariance asymptotique de $(\hat{\beta}_1^T, \ldots, \hat{\beta}_p^T, \hat{\beta}_{p+1}^T)^T$.

On note $\hat{H}$ la matrice minimisant $\hat{q}(H)$ et on note $\tilde{\beta}_i = \sum_{j=1}^{p+1} \hat{H}_{ij}\hat{\beta}_j$ l'estimateur final des β_i pour $i = 1, \ldots, p$.

On note $\hat{H}$ la matrice minimisant $\hat{q}(H)$ et on note $\tilde{\beta}_i = \sum_{j=1}^{p+1} \hat{H}_{ij}\hat{\beta}_j$ l'estimateur final des β_i pour $i = 1, \ldots, p$.

On montre que le problème de détermination de la matrice H se découpe en la résolution de p problèmes plus simples :
Recherche des matrices $H_{i1}, \ldots, H_{ip+1}$ telles que:

$$(P_i) \quad \begin{cases} H_{ii} + H_{ip+1} = I, \\ H_{ij} + H_{ip+1} = 0 \text{ pour } j \neq i, \\ \text{trace}(H_{i\bullet}\hat{\Sigma}H_{i\bullet}^T) \text{ soit minimale,} \end{cases}$$

où $H_{i\bullet}$ est la $i$ème ligne bloc de H.

Plan

Paramètres de régression

Theorem

Sous certaines hypothèses, le vecteur aléatoire $\sqrt{n}(\hat{\beta} - \beta)$ *est asymptotiquement gaussien et centré dont on détermine la matrice de covariance* $\Sigma(\tau)$.

Paramètres de régression

Theorem

Sous certaines hypothèses, le vecteur aléatoire $\sqrt{n}(\hat{\beta} - \beta)$ est asymptotiquement gaussien et centré dont on détermine la matrice de covariance $\Sigma(\tau)$.

Avec les notations $\tilde{\beta} = \hat{H}\hat{\beta}$ et $\beta^* = (\beta_1^T, \ldots, \beta_p^T)^T$ on a le résultat suivant pour l'estimateur amélioré.

Theorem

Sous les mêmes hypothèses, $\sqrt{n}(\tilde{\beta} - \beta^*)$ est asymptotiquement gaussien, centré et de matrice de covariance minimale au sens défini précédemment.

Paramètres fonctionnels

On sait estimer les risques cumulées et les fonctions de survie associées à chaque durée de vie T_i. Ainsi par exemple on estime le risque cumulé Λ_j par

$$\hat{\Lambda}_j(t) = \frac{1}{\hat{\alpha}} \int_0^t \frac{dN_{ij}(s)}{Y(s)}$$

et $\Lambda_m(t) = \sum_{j=1}^p \Lambda_j(t)$ par

$$\hat{\Lambda}_m(t) = \frac{1}{1-\hat{\alpha}} \int_0^t \frac{dN_{im}(s)}{Y(s)},$$

où $Y(s) = \sum_{i=1}^n 1(X_i \geq s)$ and $\hat{\alpha}$.
On a le comportement asymptotique de ces estimateurs ainsi que pour les estimateurs optimaux obtenus par le même type de méthode que celle utilisée pour les paramètres de régression.