Modèles de régression
en présence de compétition

Aurélien Latouche
direction : Sylvie Chevret
co-direction : Raphaël Porcher

20 Décembre 2004
Plan

1. Problématique : Risques compétitifs
2. Modélisation par temps latents, identifiabilité
3. Le modèle de Fine & Gray
   - Présentation
   - Inclusion de covariables dépendant du temps
4. Stratégies de modélisation
   - Calcul d’un nombre de sujets nécessaire
   - Modèle mal spécifié
5. Conclusions et Perspectives
1) Problématique

- Analyse de survie : Modèle à 2 états

- Compétition : Cas particulier des modèles multi-états
Exemples

Grossesse unique

- accouchement par voie basse ($\epsilon = 1$)
- césarienne ($\epsilon = 2$)

Hématologie maligne

- vivant en rémission ($\epsilon = 1$)
- décès en rémission ($\epsilon = 2$)
2) Modélisation par temps latents

- On considère \( K \) temps latents
- Les observations consistent en \((T, \varepsilon)\), où
  - \( T = \min(T_1, \ldots, T_K) \) et
  - \( \varepsilon = k \) si \( T = T_k \)
- Cadre théorique
  - Quantités nettes : seule une cause d’intérêt agirait
  - Quantités brutes : en présence des \( K \) causes

Dans la suite, on considère que \( \varepsilon = 1 \) est la cause d’intérêt
Quantités brutes

- Fonction de risque instantané cause-spécifique pour l’événement de type 1

\[
\lambda_1(t) = \lim_{h \to 0} \frac{\Pr(t \leq T \leq t + h, \varepsilon = 1|T \geq t)}{h}
\]

- Fonction d’incidence cumulée (fonction de sous-répartition, \textit{subdistribution}) pour l’événement de type 1

\[
F_1(t) = \Pr(T \leq t, \varepsilon = 1)
\]

Remarque : \(\lim_{t \to \infty} F_1(t) = \Pr(\varepsilon = 1)\)

\[\Rightarrow\] Contrairement au cas de l’analyse de survie (\textit{i.e.} \(K = 1\))

\(F_1\) dépend aussi des \(\lambda_k\) (\(k \neq 1\))
Illustration du problème d'identifiabilité a

Soit la fonction de survie conjointe

\[ H(t_1, t_2) = \exp \left[ 1 - a_1 t_1 - a_2 t_2 - \exp \{a_{12}(a_1 t_1 + a_2 t_2)\} \right] \]

où \( a_1, a_2 > 0 \) et \( a_{12} > -1 \)

Alors

\[ \lambda_k(t) = a_k \left[ 1 + a_{12} \exp \{a_{12}(a_1 + a_2)t\} \right], \quad k = 1, 2 \]

Si \( a_{12} = 0 \), les temps latents sont indépendants

Cependant, soit \( H^* \), définie par

\[ H^*(t_1, t_2) = \exp \left[ 1 - a_1 t_1 - a_2 t_2 - \sum_{k=1}^{2} a_k \exp \{a_{12}(a_1 + a_2) t_k\} / (a_1 + a_2) \right] \]

\( \forall a_{12} \), les temps latents sont indépendants et conduisent aux mêmes \( \lambda_k \).

aPrentice et al, 1978
Inférence

- Sous l’hypothèse de temps latents indépendants
  - Inférence possible sur les quantités nettes et brutes
  - Mais ...
    - Hypothèse impossible à tester (Tsiatis, 1975)
  → Problèmes d’identifiabilité

- En absence d’hypothèse d’indépendance des temps latents
  - Seules les quantités brutes sont identifiables
  - Alternative : modéliser $\lambda_1$ ou $F_1$
Modèles de régression

- Risques proportionnels
- Choix de la fonction de risque ?
  - Fonction de risque cause spécifique, $\lambda_1$
  - Fonction de risque associée à $F_1$, $\alpha_1$

- Modèle de Cox : modéliser $\lambda_1$ conditionnellement à une covariable $Z$
  - $\lambda_1(t|Z) = \lambda_0(t) \exp(\beta Z)$
3) Le modèle de Fine & Gray

La fonction de risque associée à la fonction d’incidence cumulée (ou fonction de risque de sous-répartition (*subdistribution hazard*))

\[ \alpha_1(t) = -\frac{d}{dt} \log\{1 - F_1(t)\} \]

– L’effet d’une covariable sur cette fonction de risque se traduit directement sur la fonction de sous-répartition
Effets des covariables sur les deux fonctions

- $\lambda_{11} = 3$ et $\lambda_{12} = 2$
- $\lambda_{21} = 3$ et $\lambda_{22} = 1$
Le modèle de Fine & Gray (2)

\[ \alpha_1(t) = \lim_{h \to 0} \frac{\Pr\{t \leq T \leq t + h, \varepsilon = 1 | (T \geq t) \cup (T \leq t \cap \varepsilon \neq 1)\}}{h} \]

- C’est la fonction de risque instantané de la pseudo-variable aléatoire

\[ T^* = 1_{[\varepsilon = 1]} \times T + 1_{[\varepsilon \neq 1]} \times \infty \]
Le modèle de Fine & Gray (3)
- Le modèle à risques proportionnels s’écrit

\[ \alpha_1(t; Z) = \alpha_0(t) \exp(\gamma Z) \]

où \( \exp(\gamma) = \theta \), dit *Subdistribution hazard ratio*
- Propriétés asymptotiques de l’estimateur de \( \gamma \):
  - En l’absence de censure ou présence de censure à droite administrative : se déduisent de celles d’un modèle de Cox (vraisemblance partielle)
  - Dans les autres situations de censure à droite : Estimateur pondéré (*Inverse Probability Weighting*)
- En l’absence de censure, on définit

\[ \mathcal{R}_i = \{ j : (T_i < T_j) \cup (T_i \geq T_j \cap \varepsilon_j \neq 1) \} \]

- Il se compose à l’instant \( t = T_i \) (date d’événement d’intérêt du sujet \( i \))

1. Individus \( j \) n’ayant expérimenter aucun événement avant \( t \) (idem Cox)

2. Individus \( j \) ayant expérimenter un événement en compétition avant \( t \)
Cox model

In the risk set
Out the risk set
Time
Failure of interest
No failure
Other failure

Fine and Gray model

In the risk set
Out the risk set
Time
Failure of interest
No failure
Other failure
Inclusion de covariables dépendant du temps
- Exemples nombreux en recherche clinique
  - Données longitudinales (CD4, taux d’immunoglobuline)
  - Covariables binaires, souvent indicatrices de survenue d’un événement (survenue de GvHD après allogreffe de moelle)
- Problème : Observation de $X(t)$ peut être empêchée par la survenue d’un événement en compétition (par exemple si absorbant)

"A note on including time-dependent covariates in regression model for competing risks data", A. Latouche et al, soumis
Exemple

- 2 risques en compétition : rechute et décès en rémission
  - **Objectif** : Estimer l’effet de la survenue de aGvHD sur la rechute
  - Introduction de la covariable $X(t) = 1_{\{T_{aGvHD} \leq t\}}$
- Analyse naïve : effet significatif pour $F$ & $G$, non significatif pour Cox
- Étude de simulation : biais dans l’estimation de l’effet $\gamma$ de $X(t)$
$X(t)$ observable

\begin{align*}
\text{(a)} &\quad r_2/r_1 \\
\text{(b)} &\quad p
\end{align*}
$X(t)$ non observable

(a)

(b)

$E(\hat{\gamma})$

$r_2/r_1$

$\gamma$

$r_1=0.005$

$r_1=0.01$

$r_1=0.02$

$E(\hat{\gamma})$

$p$

$\gamma$

$r_1=0.005$

$r_1=0.01$

$r_1=0.02$
4) Stratégies de modélisation

- En pratique, que faire ?
- Opposer ces modèles ?
- Quantités d’intérêt différentes et/ou complémentaires
Calcul d’un nombre de sujets nécessaire
- En présence de compétition, développer une formule pour des covariables binaires potentiellement corrélées
- Extension des formules Schoenfeld (86) et Schmoor (99) dans le cadre du modèle de Cox en absence de compétition

On considère

$$\alpha_1(t; X, Y) = \alpha_0(t) \exp(\gamma X + bY)$$

- $X$ et $Y$ binaires (éventuellement corrélées, $\rho$)
- $H_0 : \{\gamma = 0\}$
- Statistique (partielle) de Wald : $\sqrt{n} \hat{\gamma}$

“Sample size formula for proportional hazards modelling of competing risks”, A. Latouche et al, Statistics in Medicine, 2004, 23 : 3263-3274
Nombre de sujets nécessaire

\[ n = \frac{e}{\psi} = \frac{(u_{\alpha}/2 + u_{\beta})^2}{(\log \theta)^2 p (1 - p) \psi (1 - \rho^2)} \]

- \( \theta = \exp(\gamma) \)
- \( p = \Pr(X = 1) \)
- \( \psi \), proportion d’événements d’intérêt à la date d’analyse \( T_a \)
- \( \alpha \) et \( \beta \), risques d’erreur de type I et II
Expression similaire aux formules de Schoenfeld et Schmoor
- quantités identiques \((\psi, p, \rho, \alpha, \beta)\)
- mais ... rapport \(\theta\) de fonctions de risque différentes

\[
\theta = \frac{\log\{1 - F_1(t; X = 1, Y)\}}{\log\{1 - F_1(t; X = 0, Y)\}}
\]

s’obtient directement à partir d’incidences cumulées
Exemple : (re)Planification d’un essai évaluant l’efficacité d’un inducteur du travail
- Essai randomisé en double aveugle
- Critère de jugement : délai d’accouchement par voie basse

\[ \epsilon = 1 \]
\[ \epsilon = 2 \]
Calcul du nombre de sujets

<table>
<thead>
<tr>
<th>$T_a$</th>
<th>$\hat{F}_1$</th>
<th>$S$</th>
<th>$E$</th>
<th>$\theta = \frac{\log(1 - \hat{F}_E(T_a))}{\log(1 - \hat{F}_S(T_a))}$</th>
<th>$e$</th>
<th>$\psi(%)$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>12H</td>
<td>$1 - \hat{K}M$</td>
<td>0.20</td>
<td>0.30</td>
<td>1.6</td>
<td>191</td>
<td>25.0%</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>$\hat{CIF}$</td>
<td>0.20</td>
<td>0.30</td>
<td>1.6</td>
<td>191</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24H</td>
<td>$1 - \hat{K}M$</td>
<td>0.60</td>
<td>0.75</td>
<td>1.5</td>
<td>245</td>
<td>63.5%</td>
<td>364</td>
</tr>
<tr>
<td></td>
<td>$\hat{CIF}$</td>
<td>0.60</td>
<td>0.67</td>
<td>1.2</td>
<td>1160</td>
<td></td>
<td>1596</td>
</tr>
<tr>
<td>48H</td>
<td>$1 - \hat{K}M$</td>
<td>0.865</td>
<td>0.90</td>
<td>1.2</td>
<td>1357</td>
<td>77.5%</td>
<td>1546</td>
</tr>
<tr>
<td></td>
<td>$\hat{CIF}$</td>
<td>0.77</td>
<td>0.78</td>
<td>1.03</td>
<td>47339</td>
<td></td>
<td>70131</td>
</tr>
</tbody>
</table>


Modèle mal spécifié
- A ce stade, meilleure connaissance du modèle de F & G
- Choix d’un des modèles pour l’analyse
- Approche naturelle (physiologique) : modification du risque cause-spécifique

⇒ Qu’estime le modèle de Fine & Gray si l’on suppose que les données sont issues d’un modèle de Cox?

“Misspecified regression model for the cumulative incidence function”, A. Latouche et al, soumis
Hypothèses

- 2 risques en compétition \((T_1, T_2)\)
- Modèle paramétrique pour \((T_1, T_2) \sim \text{ACBVE}\)
- Paramètres ACBVE choisis pour vérifier

\[
(M) \begin{cases}
\lambda_1(t; Z) = \lambda_{10}(t) \exp(\beta_1 Z) \\
\lambda_2(t; Z) = \lambda_{20}(t) \exp(\beta_2 Z)
\end{cases}
\]

- Estimation de l'effet de \(Z\) dans un modèle de Fine et Gray

\[
\alpha_1(t; Z) = \alpha_{10}(t) \exp(\gamma Z)
\]
Expression de $\gamma$

$$\gamma(t) = \beta_1 + [a - \Psi(\beta_1, \beta_2)]t$$

$$+ \log \left[ \frac{\Psi(\beta_1, \beta_2)}{a} \times \frac{a_2 + a_1 \exp(-at)}{a_2 \exp(\beta_2) + a_1 \exp(\beta_1) \exp[-\Psi(\beta_1, \beta_2)t]} \right]$$

$\gamma$ est donc une fonction

1. compliquée!

2. sens de variation non-trivial

3. dépendant des coefficients de régression pour les 2 risques en compétition
Estimateur de $\gamma$

L'estimateur du maximum de vraisemblance partielle, $\hat{\gamma}$, de $\gamma$ est un estimateur convergent de $\gamma^*$ où $\gamma^*$ est la solution de

$$\int_0^\infty E \left\{ f_1(t; Z) \left[ Z - \frac{E[Z \exp(\gamma Z)(1 - F_1(t; Z))]}{E[\exp(\gamma Z)(1 - F_1(t; Z))]} \right] \right\} dt = 0.$$

Sous certaines hypothèses on montre que

$${\gamma^*} \simeq c_0 + c_1 \beta_1$$

où $c_0 = c_0(\beta_2, Var(Z))$ et $c_1 = c_1(\beta_2, E(Z))$
Simulations \((1)^a\)

<table>
<thead>
<tr>
<th>(a_1)</th>
<th>(a_2)</th>
<th>(a_{12})</th>
<th>(\beta_1)</th>
<th>(\beta_2)</th>
<th>(\gamma^*)</th>
<th>(\text{mean}(\hat{\gamma})(se))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.296</td>
<td>-0.301 (0.186)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.2</td>
<td>0.5</td>
<td>-0.205</td>
<td>-0.213 (0.181)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.3</td>
<td>0.5</td>
<td>-0.114</td>
<td>-0.128 (0.183)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.5</td>
<td>-0.022</td>
<td>-0.038 (0.177)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.068</td>
<td>0.049 (0.174)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.296</td>
<td>-0.308 (0.188)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.296</td>
<td>-0.306 (0.189)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.296</td>
<td>-0.307 (0.189)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.296</td>
<td>-0.300 (0.189)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>2.5</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.296</td>
<td>-0.304 (0.186)</td>
</tr>
</tbody>
</table>

\(^a\)5000 \times (n = 400)
**mgus** (Therneau et Grambsch, 2000)

**Plasma cell malignancy**

- **Age < 64 ys**
- **Age >= 64 ys**

**Death**

- **Age < 64 ys**
- **Age >= 64 ys**
\( \gamma_1^* = -1.10 \), en considérant que \((a_1, a_2) = (0.0198, 0.0405)\)

<table>
<thead>
<tr>
<th>Failure cause</th>
<th>Estimated regression parameter (Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cause-specific ( \hat{\beta}_i )</td>
</tr>
<tr>
<td>Plasma cell malignancy</td>
<td>-0.225 (0.285)</td>
</tr>
<tr>
<td>Death as first event</td>
<td>1.39 (0.198)</td>
</tr>
</tbody>
</table>
5) Conclusion et Perspectives

1. Si l’intérêt premier réside dans l’effet d’une exposition sur l’incidence cumulée de l’événement
   - Intérêt en Santé publique (études épidémiologiques, médico-économiques, ...)
   - Choix : Fonctions de risque associées à la fonction de sous-répartition (Modèle de Fine and Gray)
   - Limites : Interprétation conjointe, introduction de covariables dépendant du temps?

2. Si l’intérêt premier réside dans l’effet d’une exposition sur le risque instantané de l’événement
   - Intérêt en recherche clinique (études thérapeutiques et pronostiques)
   - Choix : Fonctions de risque cause-spécifique (Modèle de Cox)
   - Limites : Pas d’interprétation secondaire en termes de probabilités d’événement
Perspectives

- Risques compétitifs : Modèle de régression pour la probabilité conditionnelle

\[ CP_1(t) = \text{Pr}(\text{Événement d'intérêt avant } t | \text{Pas d'autre événement avant } t) \]
- Risques semi-compétitifs

- Extension naturelle
- Le décès empêche la rechute ou la GvHD
- La rechute empêche la GvHD

$$\Rightarrow T_{GvHD} \leq T_{Rechute} \leq T_{Décès}$$