# Données longitudinales et modèles de survie

## 1. Introduction

André Berchtold

Département des sciences économiques, Université de Genève

Cours de Master

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES
ÉCONOMIQUES ET SOCIALES
Département des sciences
économiques

# Plan du cours

# Plan du cours

# Introduction

- Les données longitudinales (séquences, séries temporelles), par opposition au données transversales, sont des données observées au fil du temps.
- On peut observer une même information à plusieurs reprises (température chaque jour à midi à Genève) ou plusieurs informations se succédant (état-civil d'une personne année après année).
- Les données peuvent avoir été enregistrées à intervalles fixes (chaque jour par exemple) ou irréguliers (chaque fois qu'un changement d'état-civil intervient).
- Il est toujours possible de travailler en temps discret ou en temps continu.

# Exemple (1)

- Le comportement d'un jeune singe a été observé durant 51 sessions de socialisation avec des congénères. Les périodes d'observation durent généralement 5 minutes, mais parfois moins.
- Source : Berchtold & Sackett (2002).
- Il y a 4 comportements possibles :
  1. Passivité
  2. Exploration
  3. Peur/Aggressivité
  4. Jeu
- On connaît seconde après seconde le comportement de l'animal.

# Exemple (2)

- Nous disposons des variables suivantes :
  - session : numéro de la session (1 à 51)
  - temps : temps écoulé en secondes depuis le début de la session
  - age : âge en jour au moment de chaque session (41 à 261)
  - age_3c : âge recodé en 3 catégories (1 : 41-100, 2 : 101-200, 3 : 201-261)
  - sujet : comportement du singe observé (1 à 4)
  - interacteur : comportement du singe en interaction avec le sujet observé (1 à 4, codé 1 : Passivité en cas d'absence d'interaction)
  - anormal : nombre de singes anormaux dans le groupe de socialisation (1 ou 3)

# Plan du cours

# Evénement

- Lors de l'analyse de données longitudinales, nous nous attendons à observer des transitions entre différentes situations.
- L'objet principal de l'analyse est alors un **événement** représentant le passage entre deux situations :
  - Premier mariage, naissance du premier enfant (analyse des biographies, event history analysis).
  - Décès consécutif à une certaine maladie (analyse de survie, survival analysis).
  - Panne d'une machine (analyse des défaillances, failure time analysis).
- Notre objectif est de représenter, modéliser et analyser la probabilité que cet événement survienne au fil du temps.

# Censure

- Il n'est pas du tout certain que l'événement d'intérêt soit observé chez chaque personne étudiée :
  - Une personne peut très bien quitter l'étude avant que l'événement ne survienne (exemple : déménagement hors de la zone géographique de l'étude).
  - L'événement étudié peut ne jamais survenir chez une personne (exemple : une personne n'a jamais d'enfants).
- On parle alors de données **censurées**.
- Une variable dichotomique 0-1 est généralement créée afin de différencier les données censurées (codées 0) des données non-censurées (1).

# Deux fonctions importantes

- La fonction de survie ou de séjour, notée $S(t)$, est la fonction représentant la probabilité que l'événement observé ne se soit pas encore produit. Si l'on note $T$ le moment de survenance de l'événement, alors

$$S(t) = P(T > t)$$

- La fonction de hasard ou risque instantané, $h(t)$ ou $\lambda(t)$, décrit la probabilité que l'événement se produise au temps $t$. Elle représente la proportion de personnes susceptibles d'avoir l'événement au temps $t$ et qui l'ont effectivement.
- On préfère souvent représenter la fonction de hasard cumulé au fil du temps, $H(t)$ ou $\Lambda(t)$, qui s'interprète comme le nombre moyen d'événements qui serait observé si le sujet était constamment exposé au risque.

# Exemple

- Nous définissons comme événement d'intérêt le fait que le singe observé et son interacteur jouent simultanément.
- La fonction de survie décrit la probabilité que les deux singes n'aient pas encore joué en même temps.
- La fonction de hasard décrit la probabilité au fil du temps que l'événement se produise.
- Durant certaines sessions de 5 minutes, le sujet observé et son interacteur n'ont jamais joué simultanément. Ce sont des données censurées.

# Deux méthodes d'estimation

- Il existe plusieurs méthodes permettant d'estimer la fonction de survie empirique à partir d'un ensemble de données. Les plus connues sont
  - La méthode actuarielle (temps discret).
  - La méthode de Kaplan-Meier (temps continu).
- L'application de ces méthodes nécessite l'utilisation d'une base de données dans laquelle chaque séquence temporelle correspond à une ligne. De plus, il faut disposer des deux variables suivantes :
  - Durée : Variable indiquant la durée de temps avant que l'événement ne survienne ou, dans le cas de données censurées, la durée totale d'observation.
  - Censure : Variable codée 1 si l'événement est effectivement survenu et zéro sinon (données censurées).

# Exemple : Fichier de données

| Session | Age | Age_3c | Anormal | Duree | Censure |
|---|---|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 31 | 182 | 2 | 1 | 7 | 1 |
| 32 | 187 | 2 | 1 | 41 | 1 |
| 33 | 191 | 2 | 1 | 310 | 0 |
| 34 | 194 | 2 | 1 | 83 | 1 |
| 35 | 196 | 2 | 1 | 81 | 1 |
| 36 | 203 | 3 | 3 | 96 | 0 |
| 37 | 204 | 3 | 3 | 100 | 0 |
| 38 | 211 | 3 | 3 | 75 | 0 |
| 39 | 216 | 3 | 1 | 215 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# Exemple : Kaplan-Meier (1)

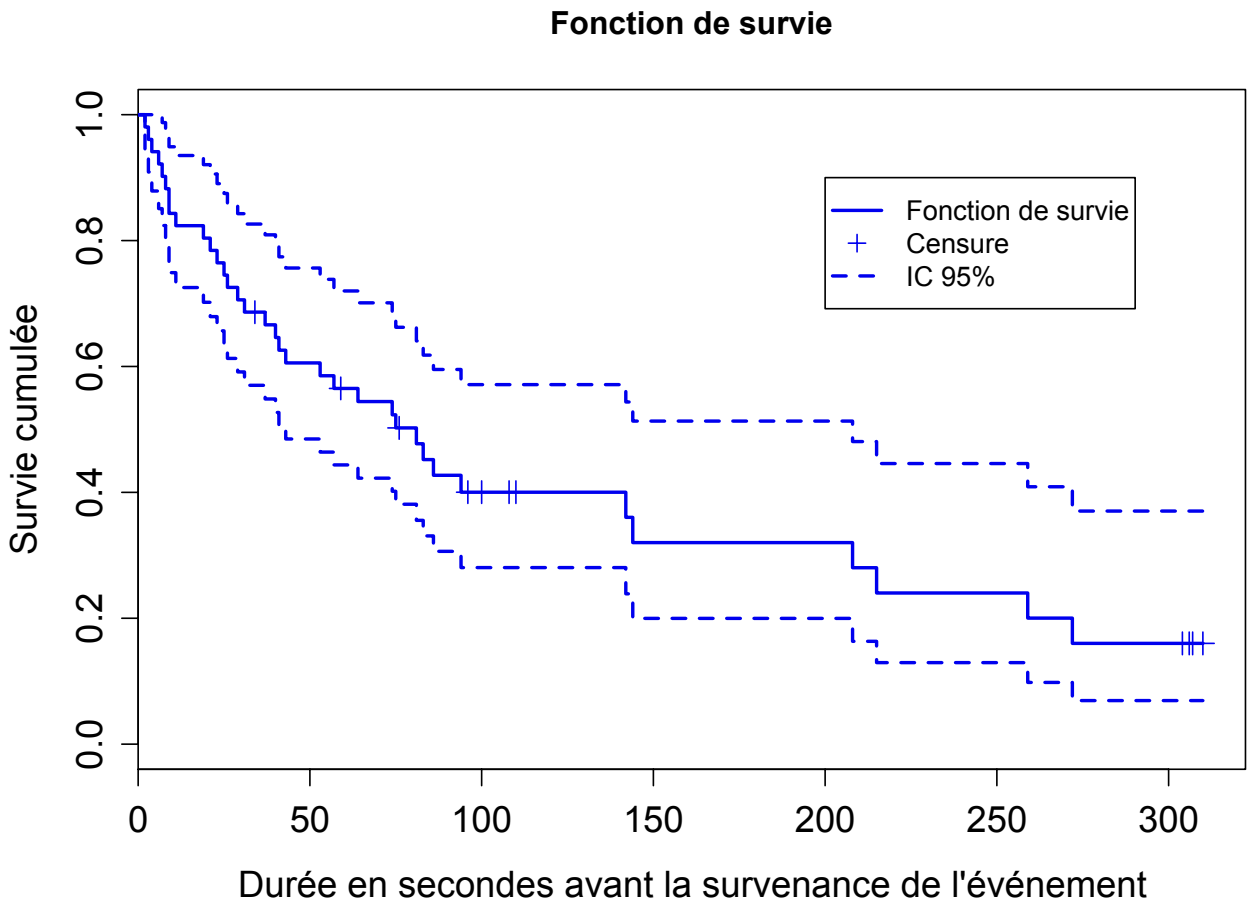
| time | n.risk | n.event | survival | std.err | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|
| 2 | 51 | 1 | 0.980 | 0.0194 | 0.9431 | 1.000 |
| 3 | 50 | 1 | 0.961 | 0.0272 | 0.9090 | 1.000 |
| 4 | 49 | 1 | 0.941 | 0.0329 | 0.8788 | 1.000 |
| 6 | 48 | 1 | 0.922 | 0.0376 | 0.8507 | 0.998 |
| 7 | 47 | 1 | 0.902 | 0.0416 | 0.8239 | 0.987 |
| 8 | 46 | 1 | 0.882 | 0.0451 | 0.7982 | 0.975 |
| 9 | 45 | 2 | 0.843 | 0.0509 | 0.7490 | 0.949 |
| 11 | 43 | 1 | 0.824 | 0.0534 | 0.7253 | 0.935 |
| 19 | 42 | 1 | 0.804 | 0.0556 | 0.7020 | 0.921 |
| 21 | 41 | 1 | 0.784 | 0.0576 | 0.6792 | 0.906 |
| 23 | 40 | 1 | 0.765 | 0.0594 | 0.6567 | 0.890 |
| 25 | 39 | 1 | 0.745 | 0.0610 | 0.6346 | 0.875 |
| 26 | 38 | 1 | 0.725 | 0.0625 | 0.6128 | 0.859 |
| 29 | 37 | 1 | 0.706 | 0.0638 | 0.5913 | 0.843 |
| 31 | 36 | 1 | 0.686 | 0.0650 | 0.5700 | 0.826 |
| 37 | 34 | 1 | 0.666 | 0.0661 | 0.5483 | 0.809 |
| 40 | 33 | 1 | 0.646 | 0.0671 | 0.5269 | 0.792 |
| 41 | 32 | 1 | 0.626 | 0.0680 | 0.5057 | 0.774 |
| 43 | 31 | 1 | 0.606 | 0.0687 | 0.4848 | 0.756 |

# Exemple : Kaplan-Meier (2)

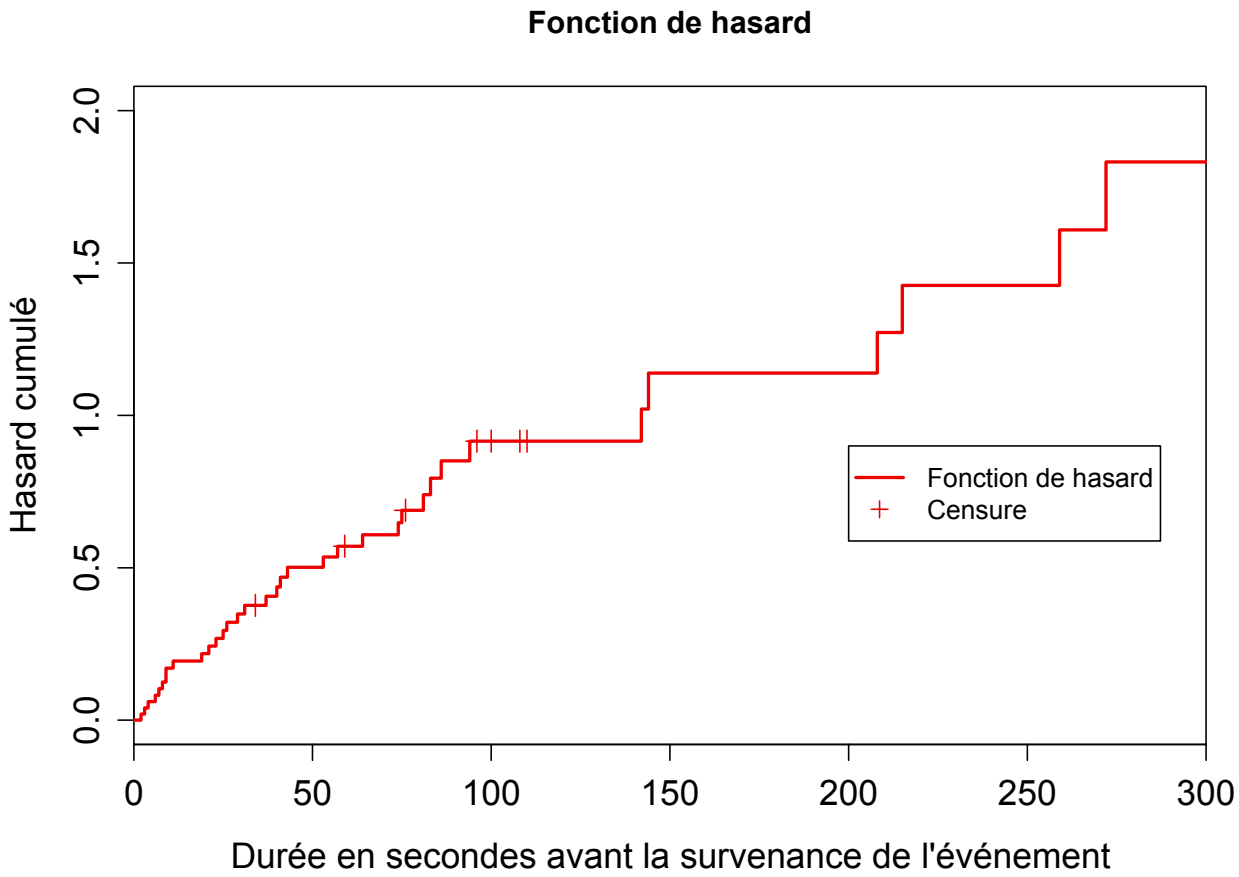
| time | n.risk | n.event | survival | std.err | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|
| 53 | 30 | 1 | 0.585 | 0.0693 | 0.4641 | 0.738 |
| 57 | 29 | 1 | 0.565 | 0.0698 | 0.4436 | 0.720 |
| 64 | 27 | 1 | 0.544 | 0.0703 | 0.4225 | 0.701 |
| 74 | 26 | 1 | 0.523 | 0.0707 | 0.4016 | 0.682 |
| 75 | 25 | 1 | 0.502 | 0.0709 | 0.3810 | 0.662 |
| 81 | 20 | 1 | 0.477 | 0.0716 | 0.3556 | 0.640 |
| 83 | 19 | 1 | 0.452 | 0.0721 | 0.3307 | 0.618 |
| 86 | 18 | 1 | 0.427 | 0.0724 | 0.3063 | 0.595 |
| 94 | 16 | 1 | 0.400 | 0.0726 | 0.2806 | 0.571 |
| 142 | 10 | 1 | 0.360 | 0.0756 | 0.2388 | 0.544 |
| 144 | 9 | 1 | 0.320 | 0.0771 | 0.1999 | 0.513 |
| 208 | 8 | 1 | 0.280 | 0.0771 | 0.1634 | 0.481 |
| 215 | 7 | 1 | 0.240 | 0.0758 | 0.1294 | 0.446 |
| 259 | 6 | 1 | 0.200 | 0.0730 | 0.0980 | 0.409 |
| 272 | 5 | 1 | 0.160 | 0.0685 | 0.0693 | 0.370 |

| records | n.max | n.start | events | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|
| 51 | 51 | 51 | 35 | 81 | 43 | 208 |

# Exemple : Kaplan-Meier (3)



# Exemple : Kaplan-Meier (4)



# Exemple : Kaplan-Meier (5)

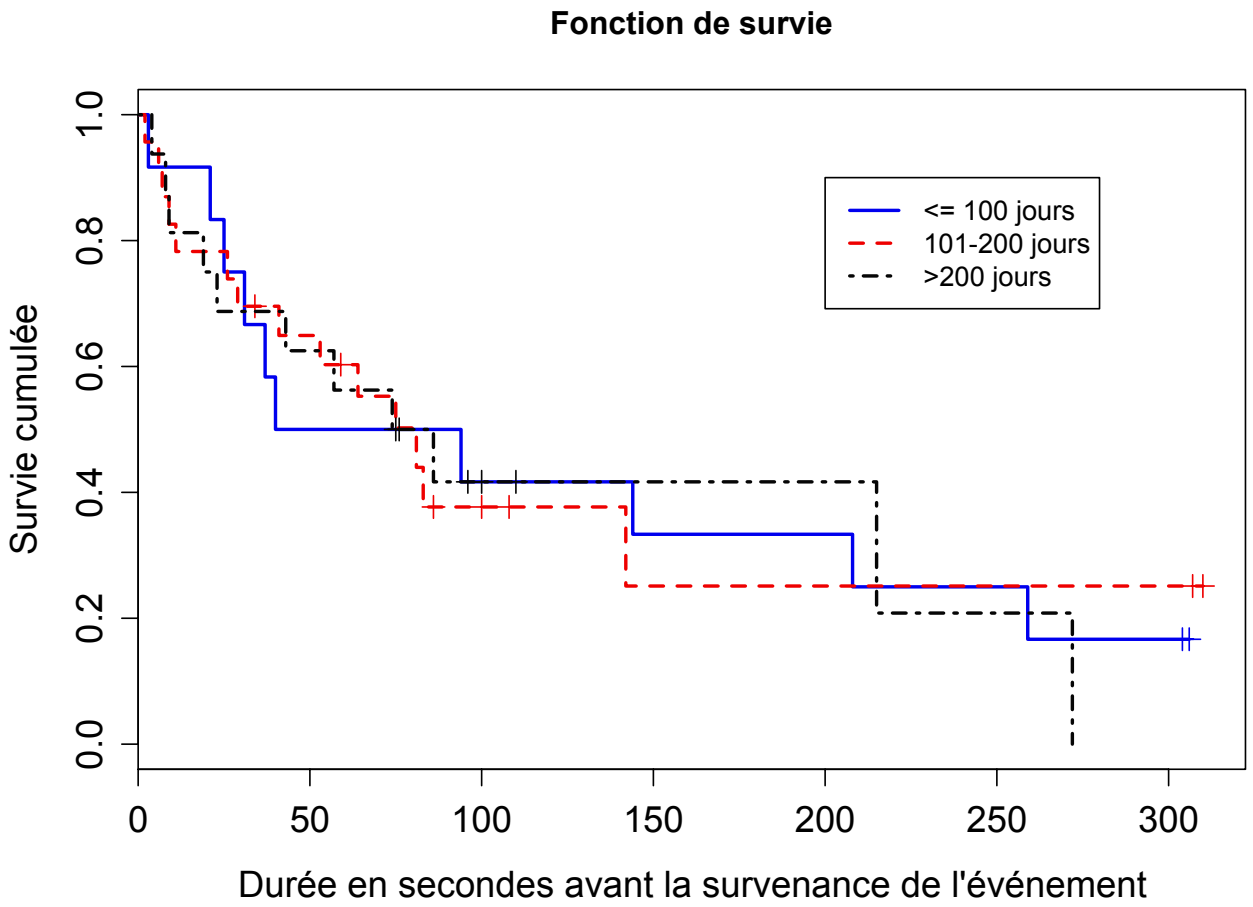
- On se demande maintenant dans quelle mesure la fonction de survie dépend de l'âge.
- On calcule trois courbes de survie pour chaque catégorie d'âge définie par la variable age_3c.
- On représente sur un même graphique les 3 courbes obtenues.
- On effectue également un test de l'égalité des 3 courbes.

# Exemple : Kaplan-Meier (6)

```
                                      records n.max n.start events median 0.95LCL 0.95UCL
strata(age_3c)=age_3c=<= 100 jours         12    12      12     10     40      31      NA
strata(age_3c)=age_3c=101-200 jours        23    23      23     14     81      41      NA
strata(age_3c)=age_3c=> 200 jours          16    16      16     11     80      23      NA
```

```
                          N Observed Expected (O-E)^2/E (O-E)^2/V
age_3c=<= 100 jours      12       10     10.1   0.00025  0.000368
age_3c=101-200 jours     23       14     14.4   0.01175  0.020358
age_3c=> 200 jours       16       11     10.5   0.02021  0.029218

 Chisq= 0  on 2 degrees of freedom, p= 0.984
```

# Exemple : Kaplan-Meier (7)



# Plan du cours

# Temps discret ou continu

- La courbe de survie empirique n'a qu'un but descriptif. Pour expliquer la probabilité que l'événement d'intérêt survienne, il est nécessaire d'introduire des covariables au moyen d'un modèle de régression.
- Il est possible de travailler en temps discret ou en temps continu :
  - En temps discret, le modèle utilisé est en fait une régression logistique.
  - En temps continu, le modèle le plus courant est le modèle semi-paramétrique de Cox.
- Les deux approches donnent généralement des résultats semblables, mais l'une ou l'autre peut se révéler plus simple à utiliser en fonction des données analysées.

# Covariables

- Deux types de covariables peuvent coexister dans un même problème :
  - Des covariables fixes au cours du temps (sexe, pays de naissance, ...).
  - Des covariables dont la valeur peut évoluer au fil du temps (nombre d'enfants, état de santé, ...).
- Le second type de covariable est beaucoup plus difficile à traiter que le premier et des procédures spéciales d'estimation ont dû être développées.

# Principe

- Le modèle continu semi-paramétrique à risques proportionnels, plus souvent appelé modèle de Cox, est un modèle de régression en temps continu.
- L'objectif est de modéliser le logarithme du risque instantané de voir l'événement d'intérêt se produire en fonction d'un ensemble de covariables.
- Les coefficients du modèle mesurent l'effet des covariables sur le logarithme du risque, alors que l'exponentielle d'un coefficient représente un effet de proportionnalité entre les risques de deux individus ayant une différence d'une unité sur la covariable.

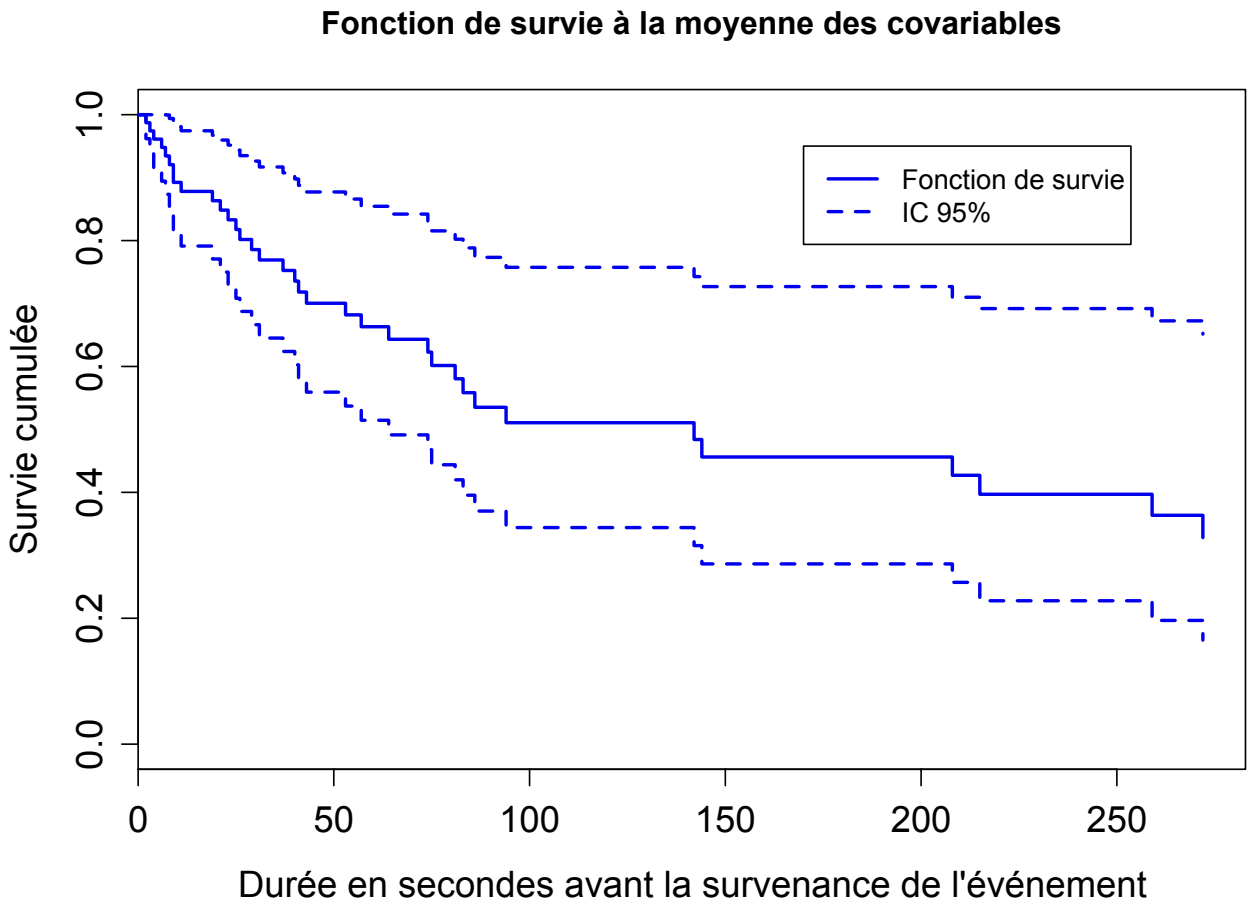
# Exemple de modèle de Cox (1)

- Nous voulons évaluer l'impact sur notre événement d'intérêt (jeu simultané du singe observé et de son interacteur) de deux covariables :
  - âge : âge du sujet en jours lors de chaque session (de 41 à 261)
  - anormal : nombre d'interacteurs anormaux dans le groupe de socialisation (1 ou 3)
- Ces deux covariables sont fixes, c'est-à-dire que leur valeur ne varie pas durant une session d'observation.
- On considère l'âge comme une variable continue, alors que *anormal* est traitée comme une variable catégorielle avec la valeur 1 comme catégorie de référence.

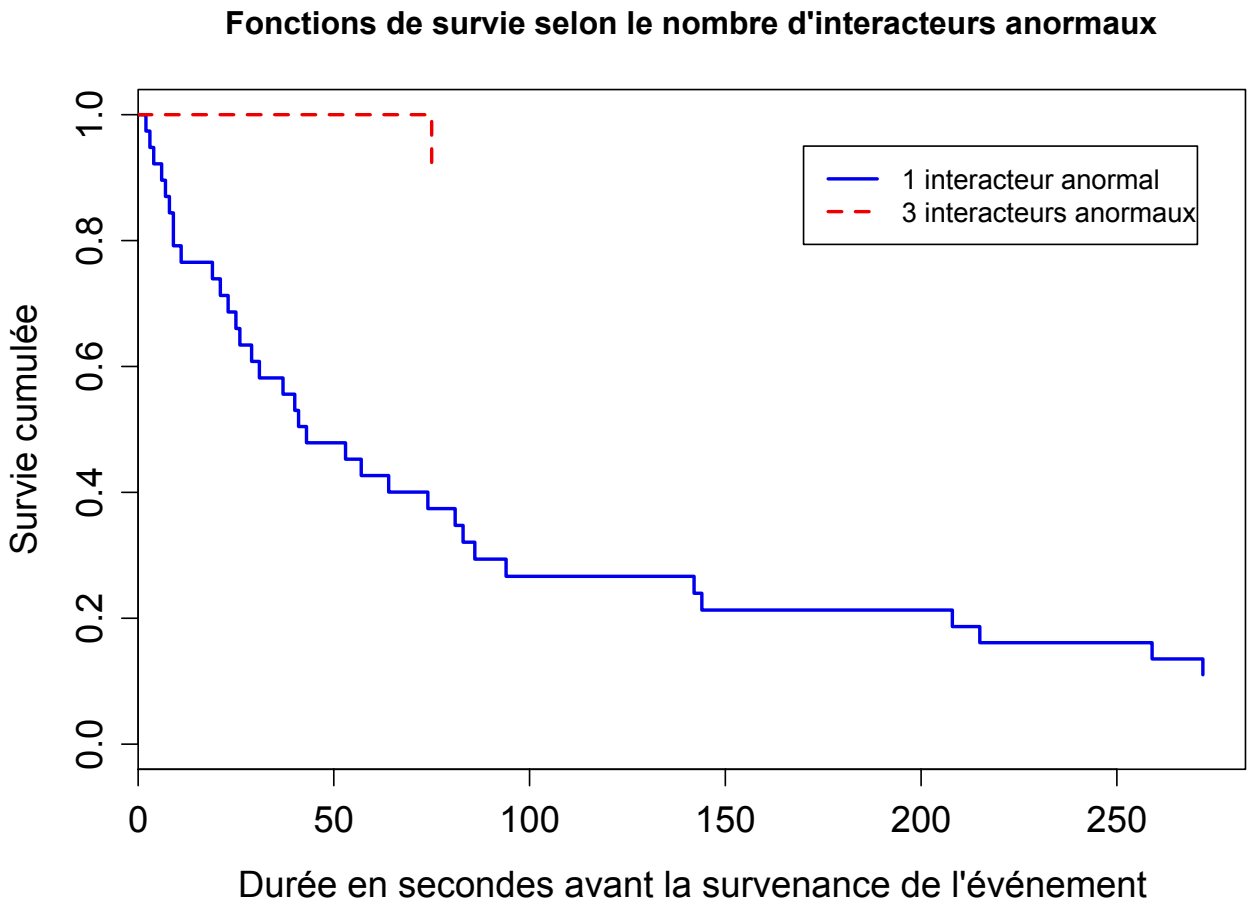
# Exemple de modèle de Cox (2)

```
  n= 51

                          coef exp(coef) se(coef)     z Pr(>|z|)
age                     0.00214   1.00214  0.00235  0.91   0.3628
as.factor(anormal)3 -2.72117   0.06580  1.02069 -2.67   0.0077 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
age                    1.0021      0.998    0.9975     1.007
as.factor(anormal)3    0.0658     15.198    0.0089     0.486

Rsquare= 0.286   (max possible= 0.989 )
Likelihood ratio test= 17.2  on 2 df,   p=0.000188
Wald test            = 7.66  on 2 df,   p=0.0217
Score (logrank) test = 12.8  on 2 df,   p=0.00167
```

# Exemple de modèle de Cox (3)



# Exemple de modèle de Cox (4)



# Bibliographie

- Berchtold A, Sackett G (2002) Markovian Models for the Developmental Study of Social Behavior. *American Journal of Primatology*, 58 (3), 149-167.