Données longitudinales et modèles de survie

4. Le modèle de Cox

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

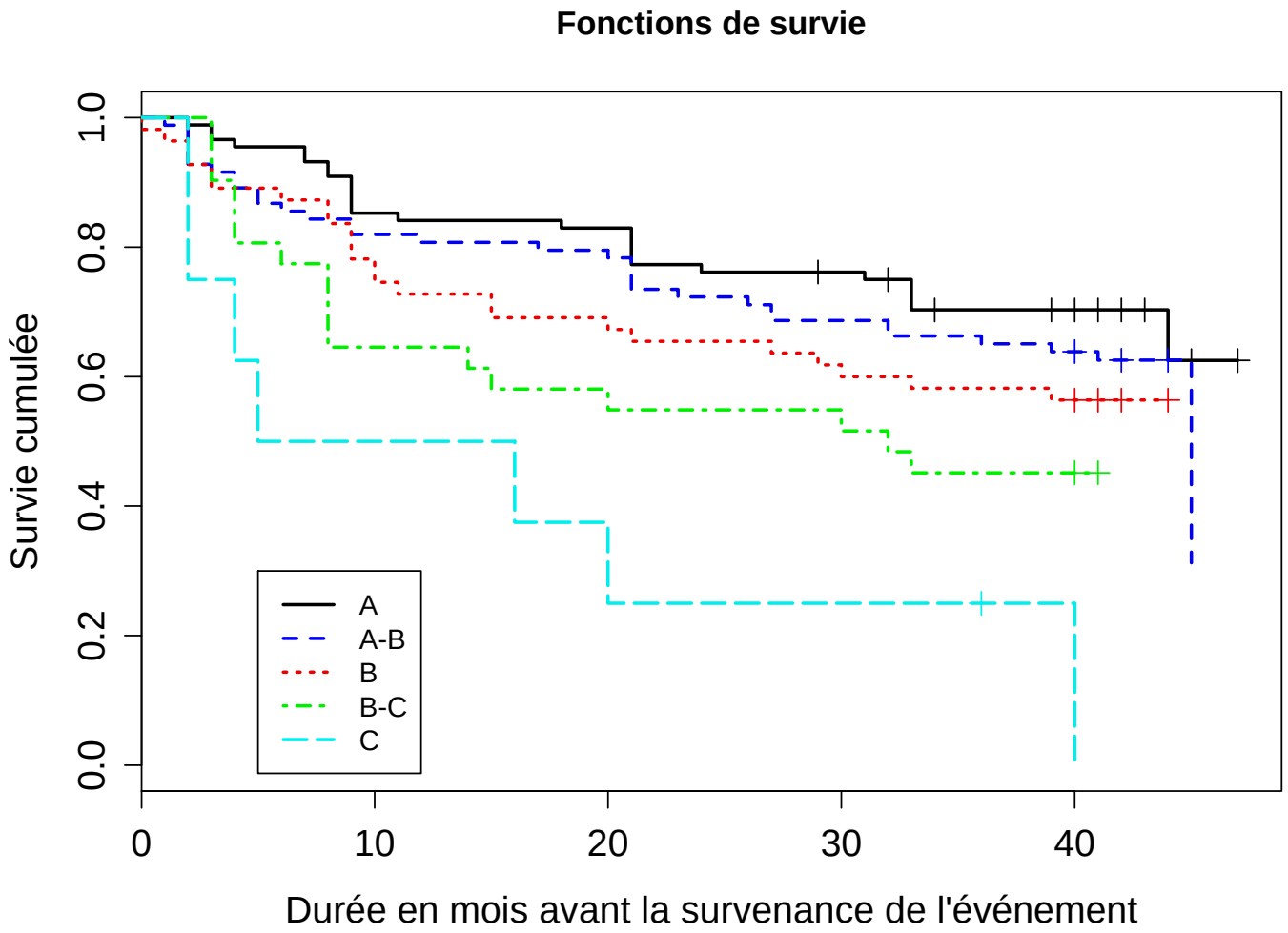
Plan du cours

Plan du cours

Influence de facteurs explicatifs

- La méthode actuarielle et la méthode de Kaplan-Meier permettent de construire la courbe de survie d'un ensemble d'individus pour un événement particulier.
- Il est aussi possible de construire différentes courbes de survie correspondant à différents sous-groupes d'individus.
- Problèmes :
  - Difficile d'évaluer précisément l'influence (significative ou non) de l'appartenance à un sous-groupe plutôt qu'à un autre.
  - Que faire avec les facteurs continus ?
  - Que faire s'il y a beaucoup de facteurs à prendre en compte simultanément ?

Exemple Yamaguchi (1)



Exemple Yamaguchi (2)

- On dispose en fait des variables suivantes :

dur	durée jusqu'à l'obtention du diplôme ou l'abandon
evt	abandon (1, ou 0 sinon)
sex	genre : homme (0) ou femme (1)
grd	note moyenne à l'école secondaire : A (meilleure) à C
prt	études à temps partiel (1, ou 0 sinon)
lag	temps entre fin école secondaire et début études (en mois)
mrg	temps jusqu'au mariage depuis début 80 (en mois)
tms	temps jusqu'au début des études depuis début 80 (en mois)

- On peut faire l'hypothèse que le risque d'abandonner ses études est influencé par plusieurs facteurs simultanément :
  - genre ;
  - note moyenne obtenue à l'école secondaire ;
  - intervalle de temps ayant séparé la fin de l'école secondaire de l'entrée à l'université ;
  - ...

Principe

- Exprimer le risque instantané $h(t)$ ou la fonction de survie $S(t)$ en fonction de facteurs explicatifs x (covariables).

$$\begin{aligned} h(t,x) &= h(t, \beta_1 x_1 + \cdots + \beta_k x_k(t) + \cdots) = h(t, x'\beta) \\ S(t,x) &= S(t, \beta_1 x_1 + \cdots + \beta_k x_k(t) + \cdots) = S(t, x'\beta) \end{aligned}$$

- On distingue entre
  - modèles en temps continu semi-paramétriques et paramétriques ;
  - modèles en temps discret.
- On distingue aussi entre
  - modèles sans facteurs explicatifs dépendant du temps ;
  - modèles avec facteurs explicatifs dépendant du temps.

Plan du cours

Introduction

- Le modèle de Cox, ou "modèle continu semi-paramétrique à risques proportionnels", est un modèle de régression en temps continu.
- L'objectif est de modéliser le logarithme du rique instantané en fonction d'un ensemble de variables explicatives x dont la valeur peut éventuellement varier au fil du temps :

$$\ln h(t,x) = \beta_0(t) + \sum_k \beta_k x_k(t) = \beta_0(t) + x'\beta$$

- De façon équivalente :

$$h(t,x) = h_0(t) \exp\left(\sum_k \beta_k x_k(t)\right) = h_0(t) \exp\left(x'\beta\right)$$

Modèle semi-paramétrique

- Le terme $h_0(t)$ est un risque de base indépendant des facteurs explicatifs du modèle.
- Aucune hypothèse n'est faite sur la distribution des durées, c'est-à-dire sur la forme de $h(t)$ ou $S(t)$.
- Le modèle de Cox est capable d'approximer correctement des modèles paramétriques (Weibull, exponentiel, ...).
- Si le vrai modèle paramétrique est inconnu, Cox est une bonne alternative. Sinon, il vaut mieux utiliser le modèle paramétrique.

Construction

- Comme pour une courbe de survie, il est nécessaire de disposer de deux variables particulières pour estimer un modèle de Cox :
  - une variable indiquant la durée de temps jusqu'à la survenance de l'événement ou jusqu'à la fin de la période d'observation dans le cas de données censurées ;
  - une variable codée 1 si l'événement a eu lieu et zéro sinon.
- Lorsque toutes les variables explicatives sont invariantes dans le temps, le modèle peut être calculé directement sur le jeu de données, alors que lorsque certains facteurs évoluent au cours du temps, une procédure particulière est appliquée au préalable sur les données.
- S'il n'y a pas de facteurs explicatifs évoluant dans le temps, le modèle s'apparente à une "simple" régression linéaire.

Risques proportionnels

- Pour tout t et toute paire d'individus $\{i, j\}$, le rapport des risques $h_i(t) = h(t, x_i)$ et $h_j(t) = h(t, x_j)$ reste indépendant du temps

$$\frac{h_i(t, x_i)}{h_j(t, x_j)} = c$$

- Ce rapport est indépendant du risque de base :

$$\begin{aligned} \frac{h_i(t, x_i)}{h_j(t, x_j)} &= \frac{h_0(t) \exp\left(x_i' \beta\right)}{h_0(t) \exp\left(x_j' \beta\right)} \\ &= \frac{\exp\left(x_i' \beta\right)}{\exp\left(x_j' \beta\right)} = \exp\left((x_i' - x_j')\beta\right) \end{aligned}$$

Exemple : Données biographiques allemandes (1)

- Données extraites de l'enquête biographique allemande réalisée entre 1981 et 1983 (Mayer & Brückner, 1989) et utilisées notamment par (Blossfeld & Rohwer, 2002).
- Trois cohortes de naissance : 1929-1931 (coho1), 1939-1941 (coho2), 1949-1951 (coho3).
- Echantillon de 201 personnes pour lesquelles on dispose d'informations concernant 1 à 9 emplois pour un total de n=600 observations (= emplois).
- *Remarque : Dans cet exemple, les différents observations d'un même individu (c'est-à-dire ses différents emplois successifs) sont considérées comme étant indépendantes les unes des autres, ce qui n'est certainement pas très réaliste ...*

Exemple : Données biographiques allemandes (2)

- Comment le niveau d'éducation (*edu*), l'expérience sur le marché du travail (*lfx*), le nombre d'emplois précédents (*pnoj*) et le prestige de l'emploi (*pres*) influencent-ils
  - le risque de terminer un emploi ?
  - la durée de l'emploi ?
- D'autres variables sont disponibles, comme le prestige du prochain emploi (*presn*) ou la date du mariage (*tmar*).
- On souhaite aussi contrôler les effets par cohorte.

Vraisemblance d'un modèle

- La vraisemblance d'un modèle est la probabilité que ce dernier ait pu générer les données observées.
- Trouver la combinaison des valeurs des paramètres qui maximise la vraisemblance revient à trouver le meilleur modèle par rapport aux données observées.
- La vraisemblance est généralement infinitésimale et on préfère donc travailler avec son logarithme, la log-vraisemblance.
- Dans le cas du modèle de Cox, on parle de vraisemblance **partielle**, car seuls les sujets subissant l'événement étudiés entrent dans le calcul, les sujets censurés n'étant considérés qu'indirectement.

Construction de la vraisemblance partielle

- Sous l'hypothèse d'absence d'égalités (toutes les durées sont différentes), on peut ordonner les durées observées comme

$$t_{(1)} < t_{(2)} < \cdots < t_{(q)}$$

certaines de ces durées correspondant à des événements et les autres à des censures.

- Supposons que la durée $t_{(i)}$ corresponde à un événement. La vraisemblance partielle considère alors la probabilité qu'en $t_{(i)}$, l'individu i subisse l'événement plutôt qu'un autre individu exposé au risque au même instant

$$LP_i = \frac{h_{(i)}}{\sum_{j \geq i} h_{(j)}} = \frac{\exp(x_i' \beta)}{\sum_{j \geq i} \exp(x_j' \beta)}$$

Maximisation

- La vraisemblance partielle que l'on maximise est donc :

$$LP = \prod_{i \in I_{\text{non censuré}}} \frac{\exp(x_i' \beta)}{\sum_{j \geq i} \exp(x_j' \beta)}$$

$$\ln LP = \sum_{i \in I_{\text{non censuré}}} \left(x_i' \beta - \ln \left(\sum_{j \geq i} \exp(x_j' \beta) \right) \right)$$

- La vraisemblance partielle est indépendante du risque de base $h_0(t)$ et donc de ses paramètres qui ne peuvent alors pas être estimés par cette approche.

Traitement des égalités

- En pratique, il peut y avoir des égalités entre les durées et plusieurs méthodes peuvent alors être utilisées :
  1. La méthode **exacte** considère que les durées sont en fait toutes différentes et que leur apparente égalité est due au manque de précision (discrétisation) des mesures. Il faut alors considérer dans le calcul toutes les permutations possibles des égalités, ce qui peut devenir très compliqué.
  2. La méthode de **Breslow** suppose qu'en cas d'événements simultanés, tous les événements ont le même risque que le premier d'entre eux.
  3. La méthode d'**Efron** calcule en revanche un risque moyen pour tous les événements survenus simultanément.
- En pratique, la méthode d'**Efron** est souvent plus performante que **Breslow**, particulièrement lorsque de nombreux événements surviennent simultanément.

Exemple : Données biographiques allemandes (1)

- On se propose d'estimer le modèle qui exprime le facteur de proportionnalité des risques de terminer son emploi en termes de niveau d'éducation (*edu*), d'expérience sur le marché du travail (*lfx*), du nombre d'emplois précédents (*pnoj*), du prestige de l'emploi (*pres*) et de la cohorte de naissance.
- La cohorte étant catégorielle, on choisit la première (1929-31) comme référence et on retient donc *coho2* (1939-41) et *coho3* (1949-51) comme covariables.
- *tfp* est la durée de l'emploi et *des* la variable de censure (1=événement, 0=censure).

Exemple : Données biographiques allemandes (2)

```
Call:
coxph(formula = survie_emploi ~ edu + lfx + pnoj + pres + coho2 +
    coho3)

  n= 600, number of events= 458

             coef  exp(coef)   se(coef)      z Pr(>|z|)
edu     0.0677638  1.0701125  0.0249198  2.719 0.006543 **
lfx    -0.0040101  0.9959979  0.0009327 -4.299 1.71e-05 ***
pnoj    0.0690427  1.0714820  0.0441730  1.563 0.118051
pres   -0.0265141  0.9738343  0.0055056 -4.816 1.47e-06 ***
coho2   0.4156215  1.5153123  0.1153732  3.602 0.000315 ***
coho3   0.3089053  1.3619334  0.1219682  2.533 0.011320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (3)

```
      exp(coef) exp(-coef) lower .95 upper .95
edu      1.0701     0.9345    1.0191    1.1237
lfx      0.9960     1.0040    0.9942    0.9978
pnoj     1.0715     0.9333    0.9826    1.1684
pres     0.9738     1.0269    0.9634    0.9844
coho2    1.5153     0.6599    1.2086    1.8998
coho3    1.3619     0.7343    1.0723    1.7297

Rsquare= 0.12   (max possible= 1 )
Likelihood ratio test= 76.93  on 6 df,   p=1.532e-14
Wald test            = 67.78  on 6 df,   p=1.168e-12
Score (logrank) test = 69.08  on 6 df,   p=6.316e-13
```

Interprétation des paramètres (1)

- Les paramètres β (*coef*) mesurent l'effet des variables correspondantes sur le logarithme du risque ($\ln h(t)$).
- On peut tester leur significativité à l'aide de la statistique de Wald qui est égale au carré du rapport entre un coefficient et son erreur standard (*se(coef)*) :

$$\left(\frac{\beta_i}{\sigma_{\beta_i}}\right)^2 \sim \chi^2_{d\ell}$$

- Lorsqu'un seul paramètre est testé, cela équivaut à

$$\frac{\beta_i}{\sigma_{\beta_i}} \sim N(0,1)$$

- Le nombre d'emplois précédents (*pnoj*) n'a pas d'effet significatif, au contraire de toutes les autres variables.

Interprétation des paramètres (2)

- Les valeurs *exp*(β) (*exp(coef)*) donnent le facteur de proportionnalité entre risques pour deux individus qui diffèrent par une unité de la variable considérée.
- Par exemple, pour *coho2*, la valeur 1.5153 indique que les individus de la cohorte 2 (39-41) ont, toutes choses étant égales par ailleurs, un risque de terminer leur emploi environ 1.5 fois plus élevé que ceux de la cohorte de référence (29-31).
- La valeur de référence (pas de différence) est la valeur 1.

Vraisemblance du modèle (1)

- La log-vraisemblance partielle du modèle estimé n'est pas donnée par défaut, mais il est possible de la retrouver en faisant afficher une table dans laquelle l'amélioration de la log-vraisemblance due à l'introduction successive de chaque paramètre est affichée, ainsi que la log-vraisemblance du modèle "NULL" ou modèle "naïf", c'est-à-dire le modèle ne comportant qu'une constante et aucun facteur explicatif. La dernière ligne donne la log-vraisemblance du modèle d'intérêt.
- Alternativement, il est possible de calculer explicitement le modèle NULL, puis de le comparer au modèle d'intérêt dans une table.

Vraisemblance du modèle (2)

```
       loglik   Chisq Df Pr(>|Chi|)
NULL  -2580.6
edu   -2580.6  0.0393  1   0.842931
lfx   -2560.1 41.0822  1  1.460e-10 ***
pnoj  -2558.8  2.4390  1   0.118350
pres  -2549.2 19.2307  1  1.158e-05 ***
coho2 -2545.3  7.7824  1   0.005276 **
coho3 -2542.2  6.3578  1   0.011687 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 Model 1: ~ 1
 Model 2: ~ edu + lfx + pnoj + pres + coho2 + coho3
   loglik Chisq Df P(>|Chi|)
1
2 -2542.2        6
```

Déviance

- Ici, la log-vraisemblance partielle de notre modèle vaut donc -2542.2.
- En pratique, on utilise plus fréquemment une statistique appelée **déviance**, valant -2 log-vraisemblance ($-2LL$), et qui représente une "distance" entre le modèle estimé et les données.
- Plus cette valeur est petite, meilleur est l'ajustement du modèle.
- La déviance étant distribuée selon une loi du chi-2, elle permet d'effectuer des tests d'hypothèses.
- Avec une valeur de 5084.4 contre 5161.2 pour le modèle naïf, le modèle estimé semble offrir un meilleur ajustement aux données.

Les statistiques globales (omnibus)

- Il existe plusieurs tests permettant de comparer globalement le modèle d'intérêt au modèle naïf.
- Formellement, nous considérons l'hypothèse nulle suivante :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

  Cette hypothèse est rejetée dès lors qu'au moins un des coefficients est significativement non-nul.
- Asymptotiquement, les tests *Likelihood ratio test*, *Wald test* et *Score (logrank) test* sont tous équivalents.
- Ici, pour 6 degrés de liberté (il y a 6 coefficients dans le modèle) ces distances sont significativement non nulles (p<0.001), ce qui indique que les covariables ont globalement un effet significatif sur le risque.

Comparaison de modèles emboîtés (1)

- Il est souvent nécessaire de comparer différents modèles emboîtés les uns dans les autres et différant par une ou plusieurs variables.
- Chaque modèle est calculé séparément, puis ils sont comparés à l'aide de tests ou de critères d'information.
- ATTENTION : Pour que les comparaisons soient valides, chaque modèle doit avoir été calculé exactement sur les mêmes observations. Il faut donc supprimer au préalable les observations comportant des données manquantes sur des variables n'apparaissant pas dans tous les modèles.
- Par exemple, pour tester l'effet cohorte, on compare le modèle complet calculé précédemment avec un modèle dans lequel les variables *coho2* et *coho3* ont été supprimées.

Comparaison de modèles emboîtés (2)

```
Call:
coxph(formula = survie_emploi ~ edu + lfx + pnoj + pres)

  n= 600, number of events= 458

           coef  exp(coef)   se(coef)      z Pr(>|z|)
edu   0.0541904  1.0556856  0.0240897  2.250   0.0245 *
lfx  -0.0045494  0.9954609  0.0008942 -5.088 3.62e-07 ***
pnoj  0.0865808  1.0904394  0.0431500  2.007   0.0448 *
pres -0.0236985  0.9765801  0.0053936 -4.394 1.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rsquare= 0.099   (max possible= 1 )
Likelihood ratio test= 62.79  on 4 df,   p=7.508e-13
Wald test            = 55.44  on 4 df,   p=2.628e-11
Score (logrank) test = 56.34  on 4 df,   p=1.701e-11
```

Comparaison de modèles emboîtés (3)

```
Analysis of Deviance Table
 Cox model: response is  survie_emploi
 Model 1: ~ 1
 Model 2: ~ edu + lfx + pnoj + pres
 Model 3: ~ edu + lfx + pnoj + pres + coho2 + coho3
   loglik  Chisq Df P(>|Chi|)
1
2 -2549.2          4
3 -2542.2 14.140   2 0.0008501 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- La différence entre les deux modèles est significative, ce qui implique que le modèle le plus simple (sans *coho2* et *coho3*) est rejeté au profit du modèle complet.

Critères d'information (1)

- Lorsque le nombre de données est grand, les tests basés sur la déviance perdent de leur intérêt car ils sont pratiquement toujours significatifs.
- Une alternative consiste alors à utiliser des critères d'information :
  - Akaike : $\text{AIC} = -2\text{LL}(M) + 2k$
  - Bayes (Schwarz) : $\text{BIC} = -2\text{LL}(M) + \ln(n)k$

  où k est le nombre de paramètres (facteurs explicatifs) et n le nombre d'événements observés (Raftery, 1995).
- Plus un coefficient est proche de zéro, meilleur il est.

Critères d'information (2)

- Il n'est pas possible de tester formellement la valeur du BIC, mais Raftery propose un ordre de grandeur pour la comparaison de deux modèles.
- Soit M_1 et M_2, deux modèles tels que M_2 est emboîté dans M_1 (ie : il a été obtenu en supprimant une ou plusieurs variables de M_1 et il est donc plus simple). Alors :

$BIC(M_1) - BIC(M_2)$	Evidence en faveur du modèle M_2
< 0	négative (support de M_1)
0 to 2	à peine digne d'être mentionnée
2 to 5	positive
5 to 10	forte
> 10	très forte

Pseudo-R^2

- Une autre approche consiste à calculer des pseudo-R^2 :
  - Cox & Snell :

$$R^2_{CS} = 1 - \exp\left(-\frac{-2LL_0 - (-2LL_M)}{n}\right)$$

  - Nagelkerke :

$$R^2_N = \frac{R^2_{CS}}{1 - \exp\left(\frac{2LL_0}{n}\right)}$$

- Comme pour le R^2 d'une régression linéaire, ceux-ci prennent des valeurs comprises entre 0 et 1, les plus grandes valeurs correspondant à un meilleur ajustement aux données. Attention : $max(R^2_{CS}) < 1$.

Exemple : Données biographiques allemandes

- Ici, $n = 458$ événements ont été observés parmi les 600 emplois et $\ln(458) = 6.13$.
- Comparaison des modèles :

Facteurs	-2LL	k	AIC	BIC	R^2_N
-	5161.2	0	5161.2	5161.2	0
edu lfx pnoj pres	5098.4	4	5106.4	5122.9	0.099
+ coho2 coho3	5084.4	6	5096.4	5121.2	0.120

- Globalement, aucun des 3 modèles n'est vraiment satisfaisant (cf. R^2_N).
- Le modèle complet est le meilleur modèle au sens de BIC, mais il ne se distingue pas réellement du modèle sans l'effet de la cohorte.

Principe

- Tout comme pour les autres types de modèles de régression (linéaire, logistique, ...), il est possible d'appliquer des procédures de sélection automatique des variables explicatives de manière à obtenir un modèle final dans lequel toutes les variables sont significatives.
- Ces méthodes sont toutefois à utiliser avec précaution, car
  - elles correspondent à une approche totalement empirique de la modélisation, sans réflexion théorique sous-jacente ;
  - les modèles finaux obtenus peuvent correspondre à une situation de sur-apprentissage par rapport à l'échantillon utilisé et ainsi être difficiles à généraliser. Des procédures basées sur le bootstrap peuvent alors se révéler plus efficientes.

Procédures

- Il y a deux grands types de procédures :
  - Forward : On part d'un modèle ne contenant qu'une constante et on rajoute une à une les variables explicatives jusqu'à ce qu'il ne soit plus possible d'en ajouter une autre qui soit significative.
  - Backward : On part d'un modèle contenant toutes les variables explicatives et on les supprime une à une jusqu'à ce qu'il ne reste que des variables significatives.
- Les procédures "stepwise" permettent de plus de retester toutes les variables (inclues ou non dans le modèle) à chaque étape et d'entrer à nouveau ou de supprimer à nouveau une variable qui aurait été supprimée ou entrée précédemment dans le modèle, mais dont la significativité a changé depuis lors.

Sélection des variables

- Plusieurs tests peuvent être utilisés afin de déterminer à chaque étape quelle est la variable à ajouter ou à supprimer du modèle :
  - Wald : On se base sur le test de significativité individuel de chaque coefficient.
  - LR (rapport de vraisemblance) : On se base sur le test du rapport de vraisemblance entre deux modèles emboîtés successifs.
  - Conditionnel : Même chose que LR, mais la procédure d'estimation des paramètres β du modèle est basée sur un calcul conditionnel plutôt que sur le maximum de la log-vraisemblance partielle.
  - Critère d'information : On se base sur la valeur de AIC ou BIC pour comparer les modèles successifs.

Exemple : Données biographiques allemandes (1)

- Trois procédures de sélection backward ont été appliquées à partir du modèle complet :
  1. Une procédure "manuelle" utilisant le test de significativité individuel de chaque coefficient.
  2. Une procédure automatique utilisant le critère AIC (fonction *stepAIC* du package *MASS*).
  3. Une procédure automatique utilisant le critère BIC (fonction *stepAIC* du package *MASS* avec l'option *k=log(458)*).
- Les première et troisième procédures aboutissent à un modèle dans lequel la variable *pnoj* a été supprimée.
- La seconde procédure aboutit à conserver le modèle complet.

Exemple : Données biographiques allemandes (2)

```
Call:
coxph(formula = survie_emploi ~ edu + lfx + pnoj + pres + coho2 +
    coho3)

          coef exp(coef) se(coef)      z       p
edu    0.06776     1.070 0.024920  2.72 6.5e-03
lfx   -0.00401     0.996 0.000933 -4.30 1.7e-05
pnoj   0.06904     1.071 0.044173  1.56 1.2e-01
pres  -0.02651     0.974 0.005506 -4.82 1.5e-06
coho2  0.41562     1.515 0.115373  3.60 3.2e-04
coho3  0.30891     1.362 0.121968  2.53 1.1e-02

Likelihood ratio test=77  on 6 df, p=1.53e-14  n= 600, number of events= 458


Call:
coxph(formula = survie_emploi ~ edu + lfx + pres + coho2 + coho3)

          coef exp(coef) se(coef)      z       p
edu    0.06628     1.069 0.024770  2.68 7.5e-03
lfx   -0.00308     0.997 0.000681 -4.52 6.3e-06
pres  -0.02583     0.975 0.005484 -4.71 2.5e-06
coho2  0.43159     1.540 0.115062  3.75 1.8e-04
coho3  0.33247     1.394 0.121290  2.74 6.1e-03

Likelihood ratio test=74.5  on 5 df, p=1.15e-14  n= 600, number of events= 458
```

Exemple : Données biographiques allemandes (3)

```
Start:  AIC=5096.33
survie_emploi ~ edu + lfx + pnoj + pres + coho2 + coho3

        Df    AIC
<none>     5096.3
- pnoj   1 5096.7
- coho3  1 5100.7
- edu    1 5101.2
- coho2  1 5107.2
- lfx    1 5115.4
- pres   1 5117.5

Call:
coxph(formula = survie_emploi ~ edu + lfx + pnoj + pres + coho2 +
    coho3)

          coef exp(coef) se(coef)      z       p
edu    0.06776     1.070 0.024920  2.72 6.5e-03
lfx   -0.00401     0.996 0.000933 -4.30 1.7e-05
pnoj   0.06904     1.071 0.044173  1.56 1.2e-01
pres  -0.02651     0.974 0.005506 -4.82 1.5e-06
coho2  0.41562     1.515 0.115373  3.60 3.2e-04
coho3  0.30891     1.362 0.121968  2.53 1.1e-02

Likelihood ratio test=77  on 6 df, p=1.53e-14  n= 600, number of events= 458
```

Exemple : Données biographiques allemandes (4)

```
Start:  AIC=5121.09
survie_emploi ~ edu + lfx + pnoj + pres + coho2 + coho3

        Df    AIC
- pnoj   1 5117.3
<none>     5121.1
- coho3  1 5121.3
- edu    1 5121.8
- coho2  1 5127.8
- lfx    1 5136.0
- pres   1 5138.1

Step:  AIC=5117.35
survie_emploi ~ edu + lfx + pres + coho2 + coho3

        Df    AIC
<none>     5117.3
- edu    1 5117.9
- coho3  1 5118.7
- coho2  1 5125.1
- pres   1 5133.4
- lfx    1 5133.8
```

Exemple : Données biographiques allemandes (5)

```
Call:
coxph(formula = survie_emploi ~ edu + lfx + pres + coho2 + coho3)


          coef exp(coef) se(coef)     z       p
edu    0.06628     1.069 0.024770  2.68 7.5e-03
lfx   -0.00308     0.997 0.000681 -4.52 6.3e-06
pres  -0.02583     0.975 0.005484 -4.71 2.5e-06
coho2  0.43159     1.540 0.115062  3.75 1.8e-04
coho3  0.33247     1.394 0.121290  2.74 6.1e-03

Likelihood ratio test=74.5  on 5 df, p=1.15e-14  n= 600, number of events= 458
```

Remarques

- Les différentes procédures de sélection peuvent parfois aboutir à des modèles différents. Cela peut permettre d'identifier plusieurs groupes possibles de facteurs explicatifs, correspondant par exemple à plusieurs théories distinctes, mais plus généralement cela peut être vu comme un indicateur du manque de robustesse de ces procédures.
- Lorsque certains facteurs explicatifs sont des variables muettes correspondant à un facteur catégoriel (par exemple *coho2* et *coho3* qui sont des indicatrices de *cohort*), il est alors préférable d'ajouter ou d'enlever simultanément toutes ces variables du modèle. Dans R, cela se fait très simplement en utilisant directement la variable complète plutôt que les variables muettes.

Exemple : Données biographiques allemandes

```
Start:  AIC=5096.33
survie_emploi ~ edu + lfx + pnoj + pres + cohort

         Df    AIC
<none>      5096.3
- pnoj    1 5096.7
- edu     1 5101.2
- cohort  2 5106.5
- lfx     1 5115.4
- pres    1 5117.5

Call:
coxph(formula = survie_emploi ~ edu + lfx + pnoj + pres + cohort)

                    coef exp(coef) se(coef)     z       p
edu              0.06776     1.070 0.024920  2.72 6.5e-03
lfx             -0.00401     0.996 0.000933 -4.30 1.7e-05
pnoj             0.06904     1.071 0.044173  1.56 1.2e-01
pres            -0.02651     0.974 0.005506 -4.82 1.5e-06
cohort1939-1941  0.41562     1.515 0.115373  3.60 3.2e-04
cohort1949-1951  0.30891     1.362 0.121968  2.53 1.1e-02

Likelihood ratio test=77  on 6 df, p=1.53e-14  n= 600, number of events= 458
```

Estimation de la fonction de survie (1)

- La fonction de survie se calcule comme

$$\boxed{S(t,x)} = \exp(-H(t)) = \exp\left(-\int_0^t h(u,x)du\right)$$

$$= \exp\left(-\int_0^t h_0(u)\,\exp(x'\beta)du\right)$$

$$= \exp\left(-\int_0^t h_0(u)du\right)^{\exp(x'\beta)}$$

$$= \exp(-H_0(t))^{\exp(x'\beta)}$$

$$= \boxed{S_0(t)^{\exp(x'\beta)}}$$

Estimation de la fonction de survie (2)

- L'estimation du modèle de Cox nous donne une estimation de β. Il reste ensuite à estimer la fonction de survie de base $S_0(t)$.
- Principe : Le risque $h_0(t)$ est supposé constant entre deux temps successifs $t_{(i)}$ et $t_{(i+1)}$:

$$\begin{aligned}\frac{S(t_{(i+1)},x)}{S(t_{(i)},x)} &= \frac{S_0(t_{(i+1)})^{\exp(x'\beta)}}{S_0(t_{(i)})^{\exp(x'\beta)}} = \left(\frac{S_0(t_{(i+1)})}{S_0(t_{(i)})}\right)^{\exp(x'\beta)} \\ &= \left(1-h_0(t_{(i)})\right)^{\exp(x'\beta)} = \alpha_i^{\exp(x'\beta)}\end{aligned}$$

Estimation de la fonction de survie (3)

- On estime les α_i par une procédure de maximum de vraisemblance conditionnelle aux estimations $\hat{\beta}$, d'où

$$\hat{S}_0(t_{(k)}) = \prod_{i<k} \hat{\alpha}_i = \prod_{i<k} \left(1 - \hat{h}_0(t_{(i)})\right)$$

- Les estimations du maximum de vraisemblance des α_i sont les solutions du système

$$\sum_{j \in D_i} \frac{\exp(x_j' \hat{\beta})}{1 - \alpha_i^{\exp(x_j' \hat{\beta})}} = \sum_{j \in R_i} \exp(x_j' \hat{\beta})$$

D_i est l'ensemble des cas dont le temps de survie observé est $t_{(i)}$ et R_i l'ensemble des cas exposés au risque en $t_{(i)}$.

Estimation de la fonction de survie (4)

- En l'absence d'égalité, ces solutions s'écrivent :

$$\hat{\alpha}_i = \left(1 - \frac{\exp(x_i'\hat{\beta})}{\sum_{j \in R_i} \exp(x_j'\hat{\beta})}\right)^{\exp(-x_i'\hat{\beta})}$$

  Sinon, il faut utiliser un algorithme itératif.
- Certains logiciels utilisent l'approximation suivante (Breslow) où d_i est le nombre d'événements en $t_{(i)}$. Cette approximation permet de limiter le nombre de calculs lorsqu'il y a beaucoup d'égalités.

$$\tilde{\alpha}_i = \exp\left(\frac{-d_i}{\sum_{j \in R_i} \exp(x_j'\hat{\beta})}\right)$$

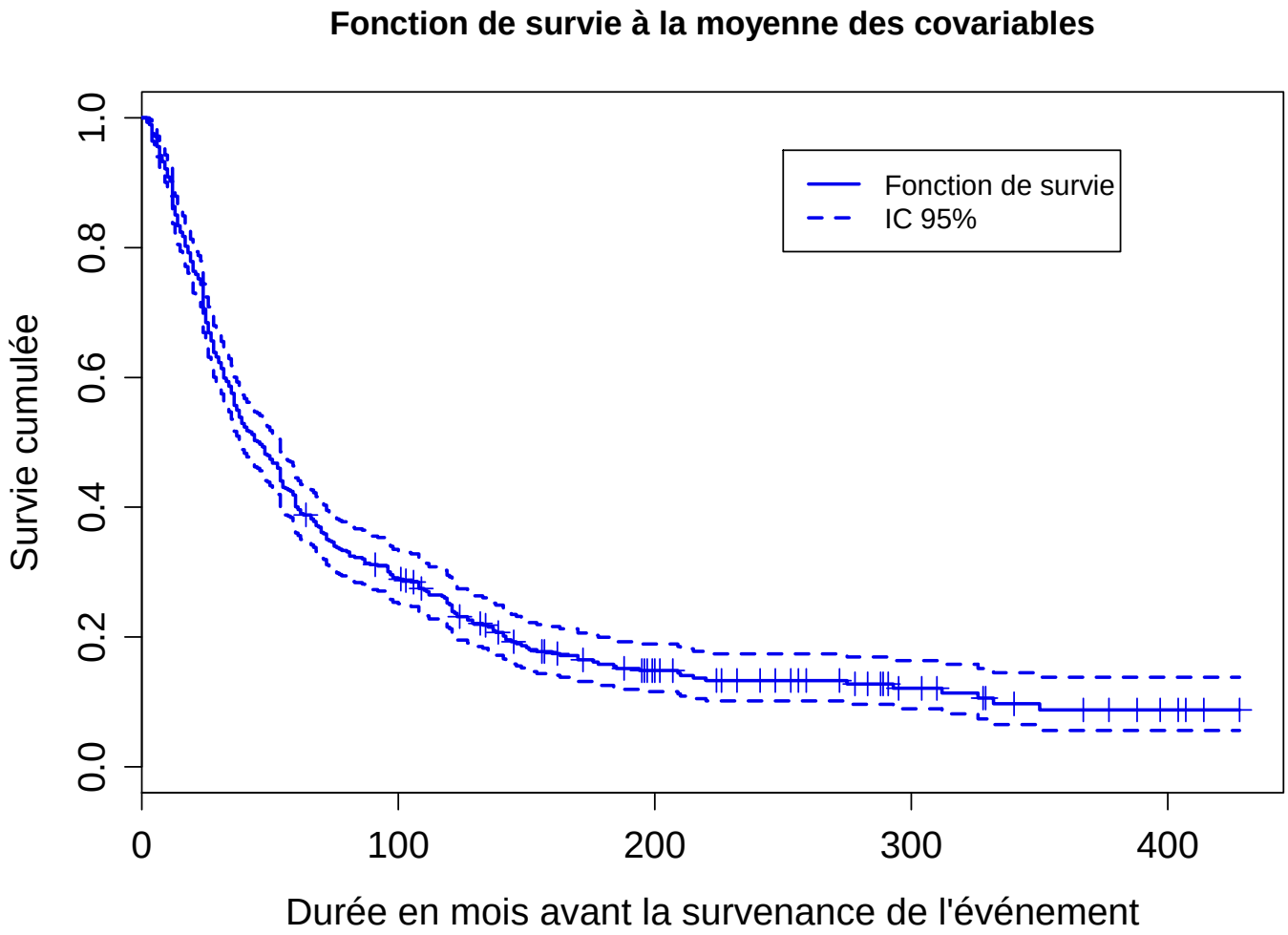
Affichage de la courbe de survie

- Par défaut, les fonctions de survie $S(t, x)$ et de risque cumulé $H(t, x)$ sont construites pour le profil moyen des covariables : $x = \bar{x}$.
- Problème : Ce profil moyen n'est souvent pas interprétable. Par exemple, pour une covariable comme le sexe qui serait codée 0 pour les hommes et 1 pour les femmes, la courbe de survie obtenue ne correspondrait ni aux hommes ni aux femmes, mais à la proportion de femmes dans l'échantillon.
- Heureusement, il est aussi possible de choisir des valeurs précises pour chaque covariable. Cette approche permet ainsi de visualiser l'impact des différentes valeurs possibles d'une covariable sur la courbe de survie.

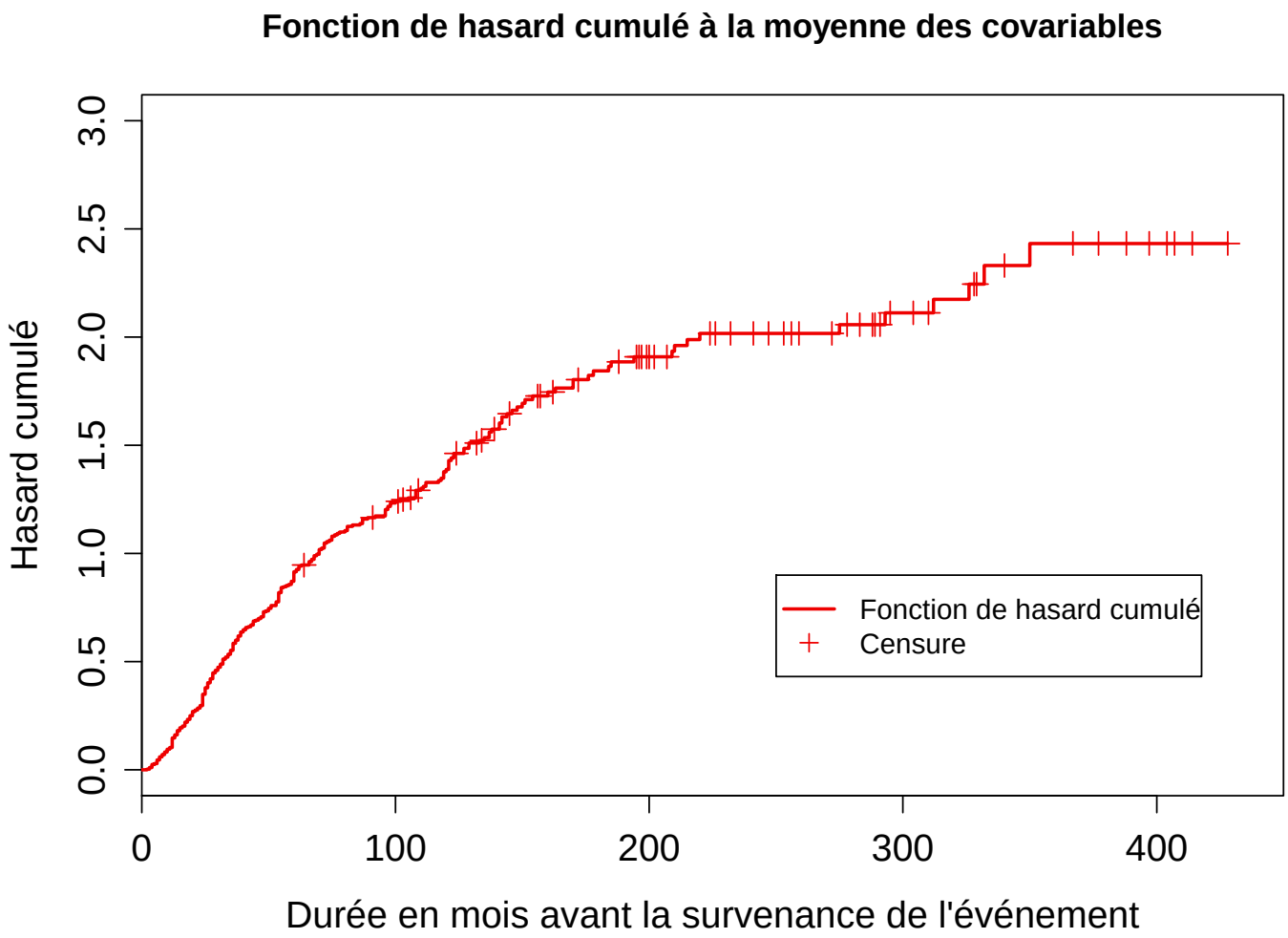
Exemple : Données biographiques allemandes (1)

- On repart du modèle complet incluant 4 covariables (edu, lfx, pnoj, pres) et l'effet cohorte.
- On construit les courbes suivantes :
  - courbes de survie et de hasard cumulé correspondant à la moyenne de chaque covariable ;
  - courbe de survie pour une personne sans expérience sur le marché du travail (lfx=0) ;
  - 3 courbes de survie correspondant aux 3 cohortes.

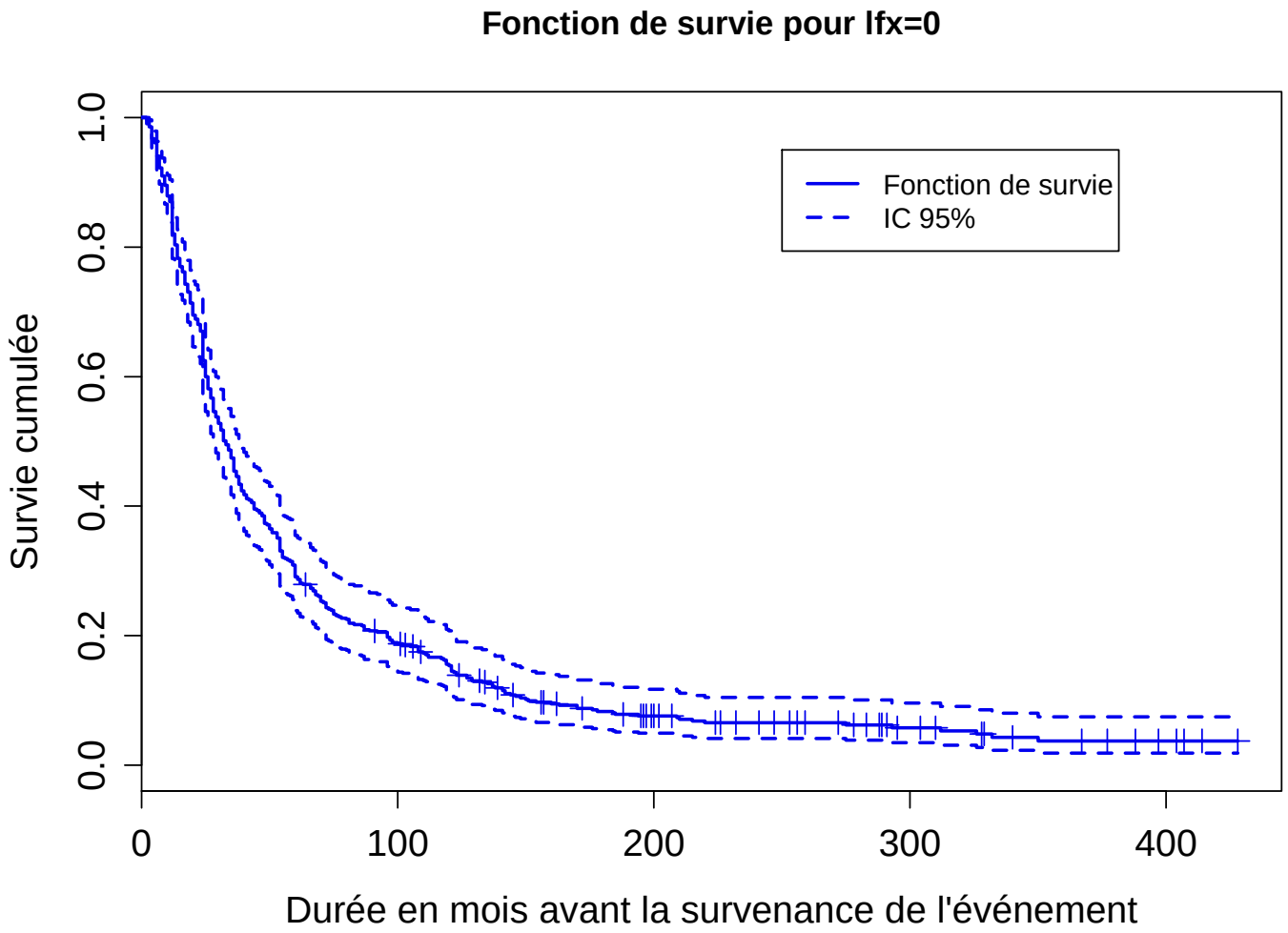
Exemple : Données biographiques allemandes (2)



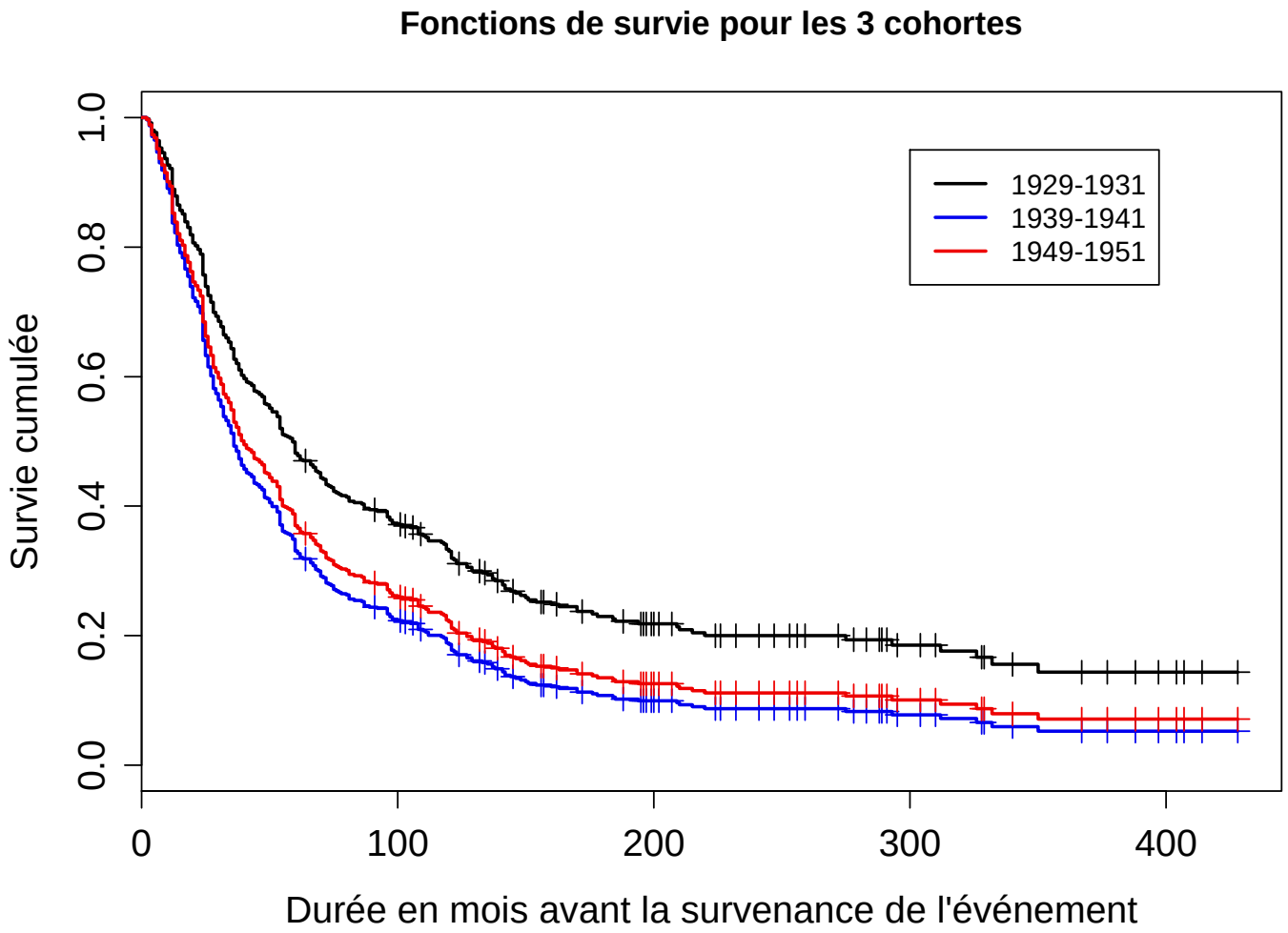
Exemple : Données biographiques allemandes (3)



Exemple : Données biographiques allemandes (4)



Exemple : Données biographiques allemandes (5)



Plan du cours

Evolution dans le temps

- Le modèle de Cox de base suppose que les caractéristiques individuelles sont constantes dans le temps (du moins durant la période observée).
- Cependant, beaucoup de caractéristiques sont susceptibles d'évoluer au fil du temps :
  - Etat-civil, nombre d'enfants, lieu d'habitation, ...
  - Etat de santé, prise ou non de médicaments, ...
  - Temps écoulé depuis la dernière révision d'une machine, ...
- Comment intégrer cette évolution dans le modèle de Cox ?

Solution avec le modèle de Cox

- Dans le cade du modèle de Cox, l'idée consiste à réévaluer la valeur des variables variant dans le temps pour chaque terme (chaque valeur $t_{(i)}$) de la vraisemblance partielle.
- Cela revient à maximiser la vraisemblance partielle généralisée suivante :

$$LP = \prod_{i \in I_{\text{non censuré}}} \frac{\exp\left(x_i(t_{(i)})'\beta\right)}{\sum_{j \geq i} \exp\left(x_j(t_{(i)})'\beta\right)}$$

Utilisation de plusieurs sous-épisodes (1)

- L'observation d'un individu peut être décomposée en différents sous-épisodes, un nouveau sous-épisode débutant à chaque changement de la valeur de la covariable.
- Par exemple, un individu sans enfants de l'année 0 à l'année 4, puis avec un enfant de 5 à 10 serait représenté sous la forme de deux épisodes :
  1. Episode commençant l'année 0 et finissant l'année 4 ; sans enfants.
  2. Episode commençant l'année 5 et finissant l'année 10 ; avec 1 enfant.

Utilisation de plusieurs sous-épisodes (2)

- Si l'on considère plusieurs covariables variant dans le temps, il faut commencer un nouveau sous-épisode chaque fois que l'une des covariables change de valeur, ce qui peut devenir très complexe.
- Cette méthode nécessite de pouvoir indiquer au logiciel utilisé non seulement la durée d'un épisode, mais aussi sa date de début.

Exemple : Données biographiques allemandes (1)

- Nous retraitons de manière différente l'exemple utilisé précédemment en considérant maintenant comme une observation les personnes et non les emplois.
- Dans ce cas, les variables tpnoj, lfx et pres, qui étaient fixes pour un emploi donné, varient maintenant au fil du temps pour une personne.
- Un nouvel épisode, avec dates de début (ts) et de fin (tf), est créé pour chaque changement de l'une de ces 3 variables, c'est-à-dire pour chaque nouvel emploi.
- En pratique, cela revient ici à utiliser la même base de données que précédemment, puisqu'elle était déjà sous la forme d'épisodes-emplois, mais les résultats sont bien entendu différents.

Exemple : Données biographiques allemandes (2)

```
Call:
coxph(formula = survie_emploi_2 ~ edu + lfx + pnoj + pres + coho2 +
    coho3)

  n= 600, number of events= 458

           coef exp(coef)  se(coef)      z Pr(>|z|)
edu    0.106486  1.112362  0.024274  4.387 1.15e-05 ***
lfx    0.001641  1.001642  0.001189  1.380 0.167634
pnoj   0.058932  1.060703  0.044100  1.336 0.181444
pres  -0.019855  0.980340  0.005523 -3.595 0.000325 ***
coho2  1.134481  3.109560  0.144893  7.830 4.88e-15 ***
coho3  1.673949  5.333189  0.208448  8.031 9.99e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (3)

```
      exp(coef) exp(-coef) lower .95 upper .95
edu      1.1124     0.8990    1.0607     1.167
lfx      1.0016     0.9984    0.9993     1.004
pnoj     1.0607     0.9428    0.9729     1.156
pres     0.9803     1.0201    0.9698     0.991
coho2    3.1096     0.3216    2.3408     4.131
coho3    5.3332     0.1875    3.5445     8.025

Rsquare= 0.199   (max possible= 0.999 )
Likelihood ratio test= 133.4  on 6 df,   p=0
Wald test            = 118.3  on 6 df,   p=0
Score (logrank) test = 126.8  on 6 df,   p=0
```

Deux situtions principales

- L'utilisation de variables dont la valeur change avec le temps implique souvent une transformation de la base de données. Nous pouvons distinguer deux cas principaux :
  1. La base de données originale comporte une ligne par individu et de nouvelles variables temporelles sont créées en combinant les variables (fixes) préexistantes. Par exemple, le nombre d'enfants peut être créé à partir des dates de naissance des enfants.
  2. La base de données originale comporte une ligne par individu et les différentes observations successives des variables évoluant dans le temps sont enregistrées dans des variables distinctes.
- Dans les deux cas, le résultat est l'obtention d'un fichier de données de type *personnes-périodes*, mais la procédure de création diffère.

Création de nouvelles variables

- Pour chaque personne (ligne) de la base de données, il faut réaliser deux étapes :
  - Calculer les différentes valeurs successives de la variable temporelle, ainsi que les temps des changements de valeur. *Ex : Nombre d'enfants et moment de naissance de chacun d'eux.*
  - Construire autant d'épisodes qu'il y a de valeurs pour cette variable, avec le temps de début et de fin de chacun d'eux. Cela se fait par duplication autant de fois que nécessaire de la ligne originale du fichier de données, puis modification des variables impliquées, y compris éventuellement la variable d'événement.

Exemple : Données biographiques allemandes (1)

- Nous voulons étudier l'influence du fait d'être oui ou non marié sur la survie.
- En utilisant la date du mariage, ainsi que les temps de début et de fin de chaque épisode, on crée pour chaque épisode une variable *married* codée 0 pour les personnes non-mariées et 1 sinon.
- Différents cas doivent être considérés selon le moment relatif du mariage par rapport à l'épisode considéré.
- Par ailleurs, dans R les épisodes doivent avoir une date de fin postérieure à la date de début, ce qui peut nécessiter quelques ajustements.

Exemple : Données biographiques allemandes (2)

- S'il n'y a pas de mariage, rien ne change dans l'épisode et married prend la valeur 0.
- Si le mariage a lieu après la fin de l'épisode, rien ne change dans l'épisode et married prend la valeur 0.
- Si le mariage a lieu avant l'épisode, alors married prend la valeur 1 pour l'épisode.
- Si le mariage a lieu durant l'épisode, alors l'épisode est scindé en deux sous-épisodes, le premier allant du mois initial au mois précédant le mariage (et married vaut 0), le second allant du mois du mariage au mois final de l'épisode (et married vaut 1). De plus, si l'événement a lieu à la fin de l'épisode originel, alors il n'a lieu qu'à la fin du second sous-épisode, le premier étant considéré comme une censure.

Exemple : Données biographiques allemandes (3)

```
Call:
coxph(formula = survie_emploi_3 ~ Data2$edu + Data2$lfx + Data2$pnoj
    + Data2$pres + Data2$married + Data2$coho2 + Data2$coho3)

  n= 747, number of events= 458

                   coef exp(coef)  se(coef)      z Pr(>|z|)
Data2$edu      0.104833  1.110525  0.024408  4.295 1.75e-05 ***
Data2$lfx      0.001598  1.001599  0.001188  1.345 0.178727
Data2$pnoj     0.060458  1.062323  0.044152  1.369 0.170901
Data2$pres    -0.019978  0.980220  0.005533 -3.610 0.000306 ***
Data2$married -0.075602  0.927185  0.123389 -0.613 0.540066
Data2$coho2    1.109379  3.032475  0.150647  7.364 1.78e-13 ***
Data2$coho3    1.629718  5.102436  0.223958  7.277 3.42e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (4)

```
               exp(coef) exp(-coef) lower .95 upper .95
Data2$edu         1.1105     0.9005    1.0587     1.165
Data2$lfx         1.0016     0.9984    0.9993     1.004
Data2$pnoj        1.0623     0.9413    0.9743     1.158
Data2$pres        0.9802     1.0202    0.9696     0.991
Data2$married     0.9272     1.0785    0.7280     1.181
Data2$coho2       3.0325     0.3298    2.2572     4.074
Data2$coho3       5.1024     0.1960    3.2896     7.914

Rsquare= 0.165   (max possible= 0.996 )
Likelihood ratio test= 134.8  on 7 df,   p=0
Wald test            = 120.2  on 7 df,   p=0
Score (logrank) test = 129.1  on 7 df,   p=0
```

Empilement des variables de chaque période

- Lorsqu'une variable distincte existe dans la base de données pour chaque nouvelle observation d'une variable temporelle, la procédure consiste à empiler ces différentes variables et à créer autant d'épisodes que de variables.
- Ce cas se rencontre principalement lorsque l'on travaille avec des données collectées en temps discret, une fois par année par exemple. Chaque variable susceptible d'évoluer au fil du temps est alors décomposée en autant de variables qu'il y a de périodes d'observation.

Exemple (1)

- Nous considérons 3 personnes interrogées 3 années successives (2008, 2009 et 2010).
- Nous disposons de l'année de naissance, du sexe et du revenu annuel brut en dizaines de milliers de francs suisses.
- Base de données originale :

ID	Naissance	Sexe	Rev 2008	Rev 2009	Rev 2010
1	1954	F	63	58	62
2	1963	F	112	115	121
3	1961	H	113	115	117

Exemple (2)

- Base de données personnes-périodes :

ID	Année	Naissance	Sexe	Revenu
1	2008	1954	F	63
1	2009	1954	F	58
1	2010	1954	F	62
2	2008	1963	F	112
2	2009	1963	F	115
2	2010	1963	F	121
3	2008	1961	H	113
3	2009	1961	H	115
3	2010	1961	H	117

Exemple (3)

- Base de données personnes-périodes pour R avec temps de début et fin de chaque épisode en mois :

ID	Année	Début	Fin	Naissance	Sexe	Revenu
1	2008	1	12	1954	F	63
1	2009	13	24	1954	F	58
1	2010	25	36	1954	F	62
2	2008	1	12	1963	F	112
2	2009	13	24	1963	F	115
2	2010	25	36	1963	F	121
3	2008	1	12	1961	H	113
3	2009	13	24	1961	H	115
3	2010	25	36	1961	H	117

Empilement de variables dans R

- La fonction *reshape* de R permet l'empilement de variables.
- En pratique, il faut tout d'abord définir quelles sont les variables originales représentant la même information au fil du temps.
- La fonction reshape crée alors automatiquement une nouvelle base de données organisée par période plutôt que par personne.
- Si on le désire, on peut ensuite trier la base de données selon les personnes et éventuellement créer des variables représentant les temps de début et de fin de chaque épisode.

Exemple : Panel suisse des ménages (1)

- Nous disposons d'un fichier de données (*Data_SHP.rda*) extraites du panel suisse des ménages.
- Pour 3 années (1999, 2004, 2009), nous disposons des variables suivantes :
  - Etat-civil (CIVSTA99, CIVSTA04, CIVSTA09).
  - Nombre d'enfants vivant avec la personne (NBKID99, NBKID04, NBKID09).
- Par ailleurs, nous avons aussi l'ID personnel de chaque personne (IDPERS) et son sexe (SEX).

Exemple : Panel suisse des ménages (2)

- Extrait des données pour 3 personnes :

```
      IDPERS   SEX  CIVSTA99 CIVSTA04 CIVSTA09
11253  74102 woman   married  married  married
11257  85101   man separated divorced divorced
11350 295101 woman separated divorced divorced

      NBKID99 NBKID04 NBKID09
11253       3       2       2
11257       0       0       0
11350       2       2       0
```

Exemple : Panel suisse des ménages (3)

- Données personnes-périodes correspondantes après tri par personne :

	IDPERS	SEX	time	CIVSTA99	NBKID99	id
44.1	74102	woman	1	married	3	44
44.2	74102	woman	2	married	2	44
44.3	74102	woman	3	married	2	44
48.1	85101	man	1	separated	0	48
48.2	85101	man	2	divorced	0	48
48.3	85101	man	3	divorced	0	48
141.1	295101	woman	1	separated	2	141
141.2	295101	woman	2	divorced	2	141
141.3	295101	woman	3	divorced	0	141

Exemple : Panel suisse des ménages (4)

- Données personnes-périodes après tri par personne, changement des noms de variables et création des temps de début et de fin de chaque épisode en mois :

	IDPERS	SEX	time	CIVSTA	NBKID	id	time_deb	time_fin
44.1	74102	woman	1	married	3	44	1	60
44.2	74102	woman	2	married	2	44	61	120
44.3	74102	woman	3	married	2	44	121	132
48.1	85101	man	1	separated	0	48	1	60
48.2	85101	man	2	divorced	0	48	61	120
48.3	85101	man	3	divorced	0	48	121	132
141.1	295101	woman	1	separated	2	141	1	60
141.2	295101	woman	2	divorced	2	141	61	120
141.3	295101	woman	3	divorced	0	141	121	132

Plan du cours

Introduction

- Le modèle de Cox postule que les risques sont proportionnels entre individus.
- Il est nécessaire de vérifier cette hypothèse, afin de s'assurer de la fiabilité des résultats.
- Plusieurs approches sont possibles :
  1. comparaison graphique des courbes de survie ;
  2. analyse des résidus de Schoenfeld ;
  3. test de la non-interaction avec le temps.

Courbes LML (1)

- Nous considérons la transformation suivante des courbes de survie :

$$\ln(-\ln(S(t,x))$$

- Cette transformation, dite LML pour "Log Minus Log", a la propriété suivante : Si l'hypothèse de risques proportionnels est valide, alors

$$S(t,x) = S_0(t)^{\exp(x'\beta)}$$

$$\Longleftrightarrow \ln\big(-\ln S(t,x)\big) = \ln\big(-\ln S_0(t)\big) + x'\beta$$

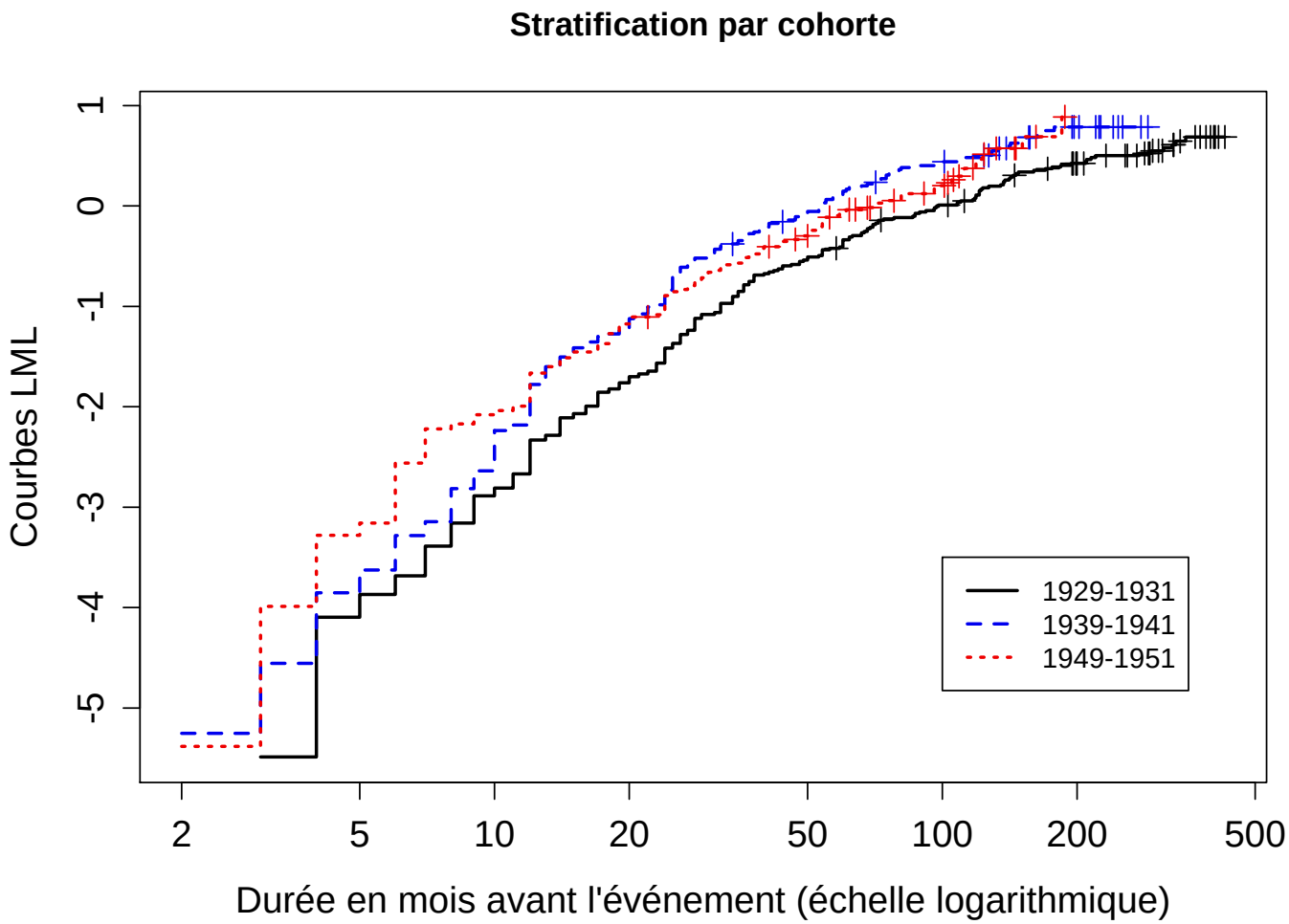
Courbes LML (2)

- Pour deux profils différents x_1 et x_2, la différence entre les courbes LML vaut
$$(x_2' - x_1')\beta$$
- Cette quantité est indépendante du temps.
- Les courbes de survie après transformation LML sont donc parallèles pour différentes valeurs de x.
- Il suffit alors de tracer les courbes LML correspondant aux différents niveaux d'une covariable, les autres covariables restant constantes, et de les comparer.
- S'il est possible de superposer les différentes courbes par simple translation, alors l'hypothèse de proportionnalité est vérifiée.

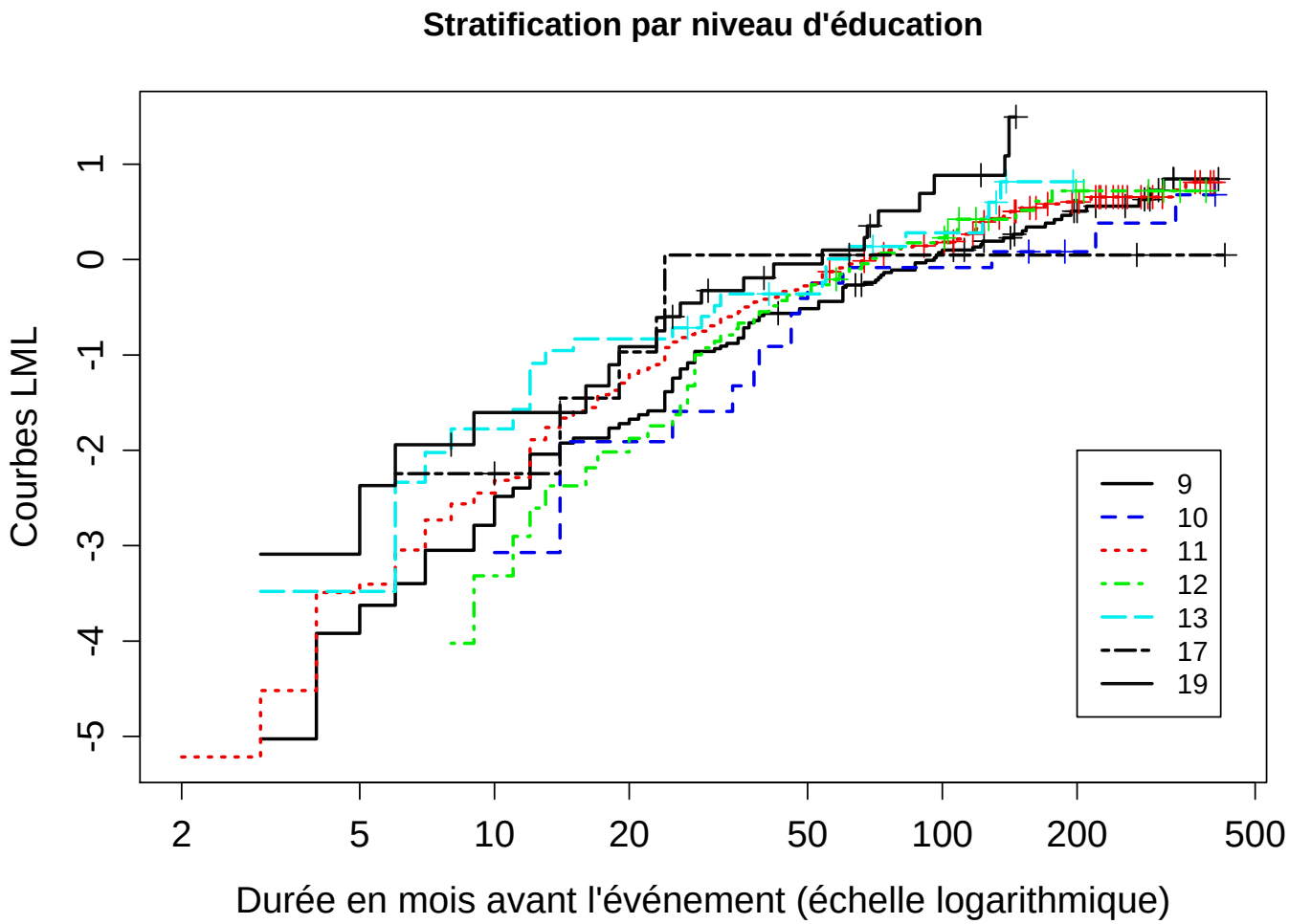
Mise en oeuvre

- La variable pour laquelle on veut vérifier qu'elle respecte l'hypothèse de proportionnalité est utilisée comme **variable de stratification**, toutes les autres variables restant telles quelles dans le modèle.
- On compare ensuite visuellement les courbes LML obtenues pour chaque catégorie de la variable de stratification.
- Remarques :
  - Dans le cas d'une variable continue, il faut tout d'abord répartir ses valeurs en un nombre fini de catégories.
  - Il n'y a pas de règle permettant de dire lorsque deux courbes sont similaires et lorsqu'elles ne le sont pas. La comparaison comporte une grande part de subjectivité !

Exemple : Données biographiques allemandes (1)



Exemple : Données biographiques allemandes (2)



Interprétation de la notion de résidus

- Dans un modèle de régression linéaire traditionnel, les résidus mesurent la différence entre les valeurs observées de la variable dépendante et les valeurs prédites par le modèle.
- Dans le cas d'un modèle de Cox, c'est le risque instantané qui est expliqué et la notion de résidu n'a alors pas de sens, car il n'y a pas moyen de calculer une différence entre valeurs observées et prédites.
- Plusieurs autres notions de résidus ont ainsi dû être définies (Schoenfeld, déviance, martingale, score, ...). Ce sont ceux de Schoenfeld qui sont utiles ici.

Résidus partiels de Schoenfeld (1)

- Les résidus de Schoenfeld concernent avant tout les covariables et non la fonction de risque instantané. Il y en a autant que de covariables.
- Ces résidus ne concernent que les cas non-censurés.
- Le résidu lié à une covariable représente l'écart entre la valeur prise par cette covariable pour un individu au moment de la survenance de l'événement considéré et la moyenne de cette covariable parmi tous les individus exposés au risque à ce moment.

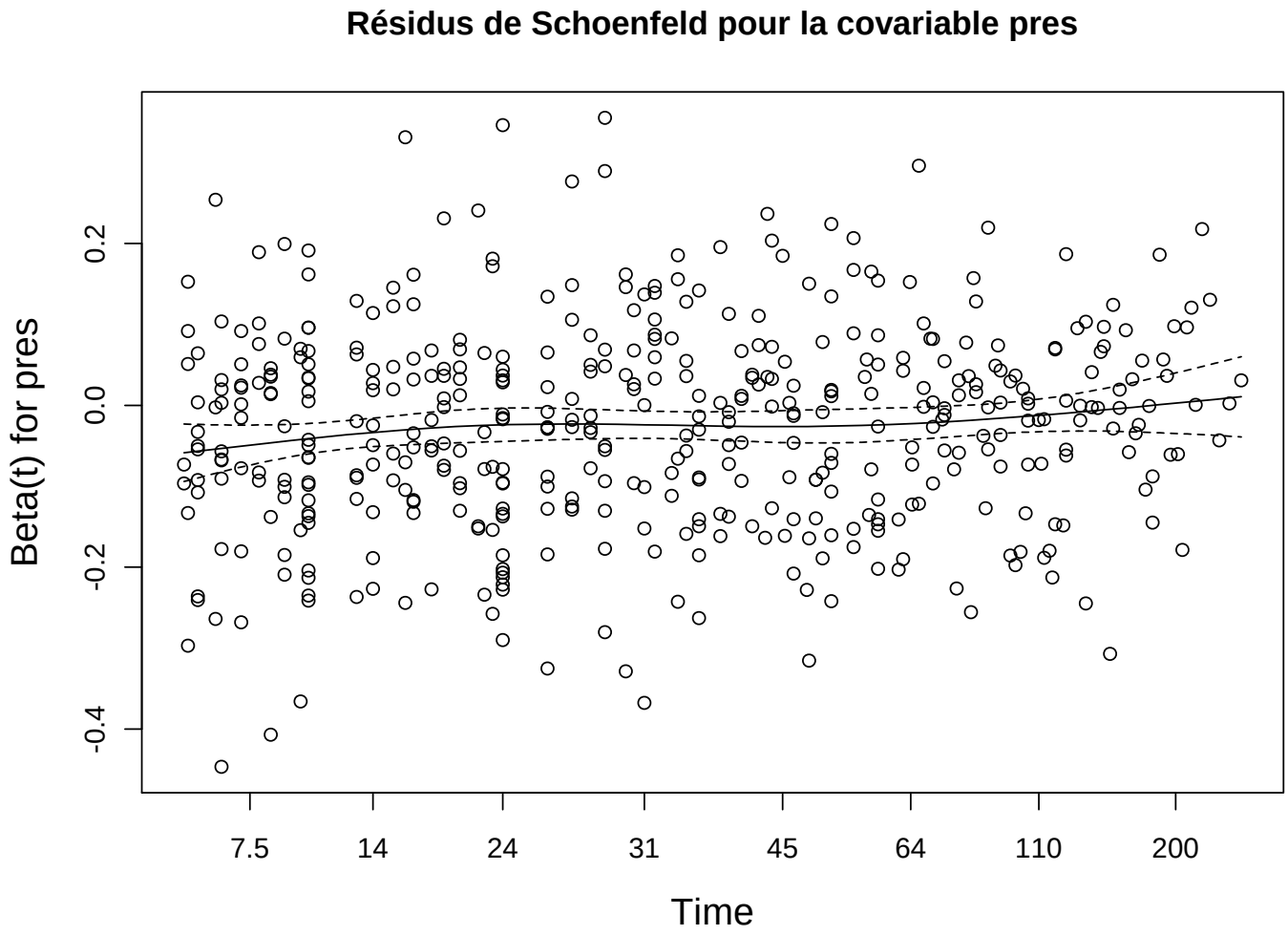
Résidus partiels de Schoenfeld (2)

- Les résidus de Schoenfeld ne sont pas une mesure de l'ajustement du modèle aux données.
- Ils s'interprètent comme une mesure de la différence de profil (par rapport aux covariables du modèle) entre un individu subissant l'événement étudié et l'ensemble des individus exposés au risque.

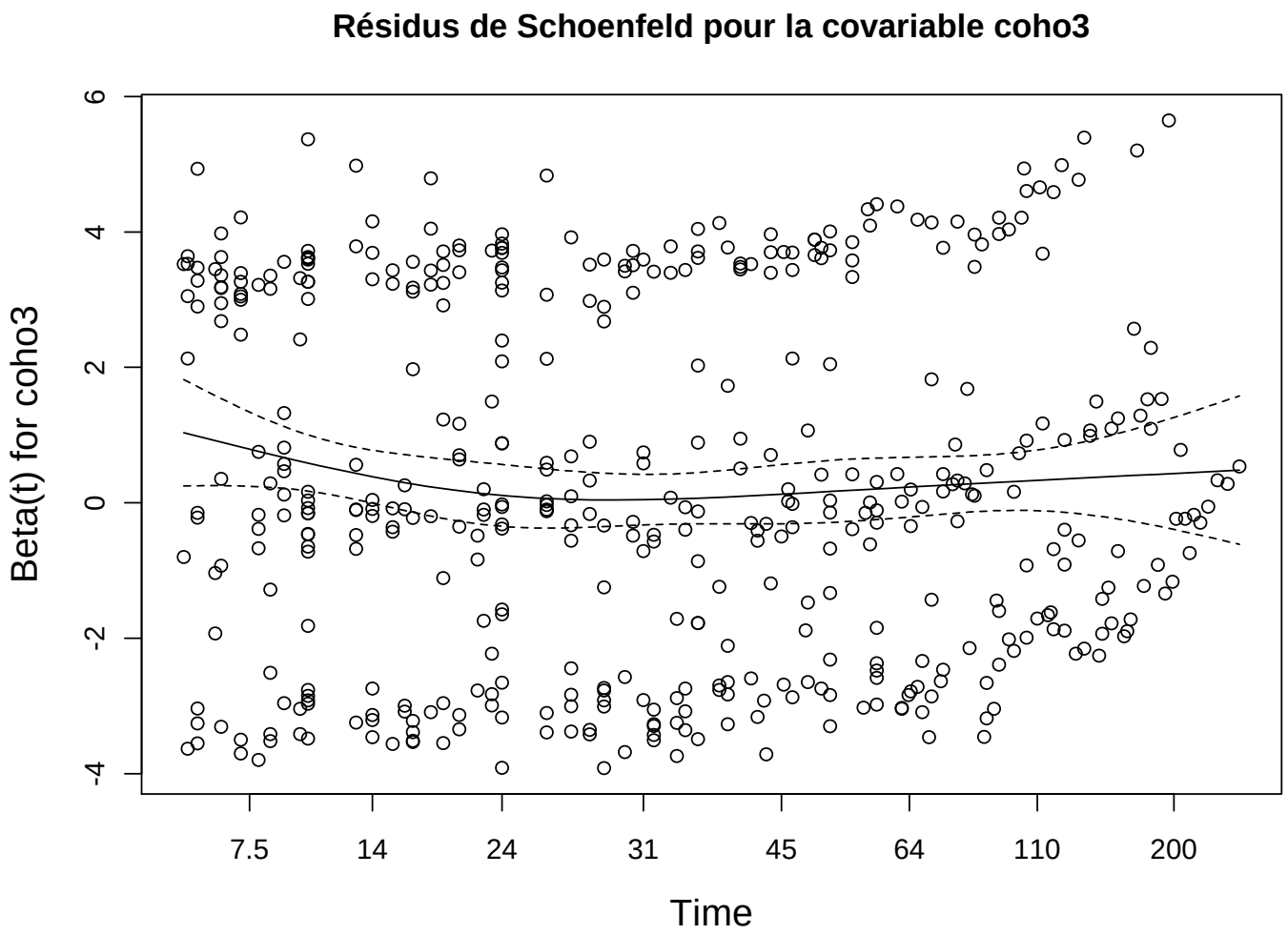
Analyse graphique des résidus

- Les résidus de Schoenfeld peuvent être analysés sur la base de graphiques, afin de détecter un éventuel non-respect de l'hypothèse de proportionnalité.
- L'idée consiste à représenter les résidus en fonction d'une transformation du temps.
- On peut aussi rajouter sur le même graphique une courbe représentant l'évolution moyenne des résidus en fonction du temps.
- Cette courbe donne la tendance générale. Toute différence par rapport à une droite horizontale représente une déviation par rapport à l'hypothèse de proportionnalité.

Exemple : Données biographiques allemandes (1)



Exemple : Données biographiques allemandes (2)



Test des résidus

- Une méthode de test des résidus consiste à calculer la corrélation entre les résidus et les durées de survie.
- Si l'hypothèse nulle d'absence de corrélation est acceptée, alors l'hypothèse de proportionnalité est vérifiée. Sinon, elle est rejetée.
- Alternativement, il est aussi possible de calculer une régression linéaire expliquant les résidus à l'aide du logarithme du temps, puis à tester si la pente de la droite de régression est bien nulle.

Exemple : Données biographiques allemandes

	rho	chisq	p
edu	-0.0701	2.275	0.1315
lfx	0.0550	1.516	0.2182
pnoj	-0.0893	3.411	0.0647
pres	0.0937	4.638	0.0313
coho2	-0.0643	1.903	0.1678
coho3	-0.0426	0.837	0.3603
GLOBAL	NA	9.703	0.1377

Dépendance au temps (1)

- Le rejet de l'hypothèse de proportionnalité des risques implique que le rôle des variables explicatives du modèle évolue au fil du temps.
- Une autre méthode de test consiste alors à introduire des coefficients de régression évoluant en fonction du temps dans le modèle de Cox et à tester leur significativité.
- S'ils sont significatifs, alors l'hypothèse de proportionnalité des risques est remise en question.
- En pratique, pour une variable x, on crée un effet d'interaction $x\_\ln(t)$ en multipliant x par $\ln(t)$.

Dépendance au temps (2)

- Dans le modèle, x et l'interaction sont introduits simultanément, d'où

$$\begin{aligned} \beta_1 x + \beta_2 x\_\ln(t) &= (\beta_1 + \beta_2 \ln(t)) x \\ &= \beta_t^* x \end{aligned}$$

- Le coefficient de régression β_t^* varie en fonction du temps.
- Si le coefficient β_2 est significativement différent de zéro, alors il y a dépendance au temps.

Données biographiques allemandes

- Ajout d'une interaction entre le temps et pres :

```
            coef exp(coef)  se(coef)       z Pr(>|z|)
edu     0.106776  1.112685  0.025436   4.198 2.69e-05 ***
lfx    -0.001192  0.998809  0.001016  -1.174    0.241
pnoj    0.257260  1.293382  0.043219   5.953 2.64e-09 ***
pres    1.460636  4.308697  0.066736  21.887  < 2e-16 ***
coho2  -2.323775  0.097903  0.200028 -11.617  < 2e-16 ***
coho3  -4.767052  0.008505  0.320715 -14.864  < 2e-16 ***
pres_T -0.253205  0.776309  0.011516 -21.987  < 2e-16 ***
```

- Le coefficient de l'interaction étant significatif, on peut penser que l'effet du prestige évolue (ici : diminue) au fil du temps.

Limitations

- Les différentes méthodes présentées ici pour tester la proportionnalité des risques doivent avant tout être utilisées pour établir un faisceau de présomptions en faveur ou en défaveur du respect de cette hypothèse.
- Aucune méthode ne doit être considérée comme parfaitement fiable.
- Par ailleurs, toutes les méthodes présentées cherchent à mettre en évidence une non-proportionnalité évoluant linéairement en fonction du temps. D'éventuels phénomènes non-linéaires pourraient donc ne pas être détectés.

En cas de non-proportionnalité

- Dans le cas où la proportionnalité des risques doit manifestement être rejetée par rapport à une ou plusieurs covariables, deux options sont possibles :
  1. Des effets d'interaction entre ces covariables et le temps sont introduits explicitement dans le modèle de Cox.
  2. On utilise le modèle de Cox stratifié
- Attention : De légères violations de l'hypothèse de proportionnalité ne sont pas très graves. Le modèle de Cox reste quand même une bonne approximation de la réalité.

Plan du cours

Utilité

- La stratification consiste à calculer un modèle de Cox en attribuant une valeur différente du risque de base $h_0(t)$ à chaque catégorie de la variable de stratification. En revanche, l'influence des variables explicatives, et donc les valeurs estimées des paramètres β, est commune à toutes les catégories.
- Cette méthode permet d'inclure dans un modèle de Cox une variable ne satisfaisant pas à l'hypothèse de proportionnalité des risques.
- Il est aussi possible de stratifier en fonction de plusieurs variables simultanément.

Hypothèses

- Soit une variable de stratification avec catégories indicées $s = 1, 2, \ldots$.
- Le modèle de Cox est estimé en admettant un risque de base $h_{s,0}(t)$ différent pour chaque strate.
- Les risques sont toujours supposés proportionnels pour individus d'une même strate, mais pas entre les strates.
- Les effets des covariables sont identiques dans toutes les strates.
  $\Longleftrightarrow$ Il n'y a pas d'effet d'interaction entre la variable de stratification et les variables explicatives du modèle.

Vraisemblance stratifiée

- La vraisemblance partielle que l'on maximise s'écrit

$$LP_{str} = \prod_{s} \prod_{i \in I_{s,\text{non censuré}}} \frac{\exp(x_i'\beta)}{\sum_{j \geq i} \exp(x_j'\beta)}$$

$$\ln LP_{str} = \sum_{s} \sum_{i \in I_{s,\text{non censuré}}} \left(x_i'\beta - \ln \left(\sum_{j \geq i} \exp(x_j'\beta) \right) \right)$$

- Les estimations sont différentes de celles obtenues sans stratification.

Données biographiques allemandes (1)

- Stratification selon le sexe.
- Le nombre d'emplois précédents (pnoj) devient significatif.

```
  n= 600, number of events= 458

             coef  exp(coef)   se(coef)      z Pr(>|z|)
edu     0.0799709  1.0832556  0.0247786  3.227  0.00125 **
lfx    -0.0040443  0.9959639  0.0009311 -4.344 1.40e-05 ***
pnoj    0.0897184  1.0938662  0.0447269  2.006  0.04487 *
pres   -0.0261239  0.9742143  0.0054548 -4.789 1.67e-06 ***
coho2   0.4017003  1.4943635  0.1155822  3.475  0.00051 ***
coho3   0.2816484  1.3253126  0.1225962  2.297  0.02160 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

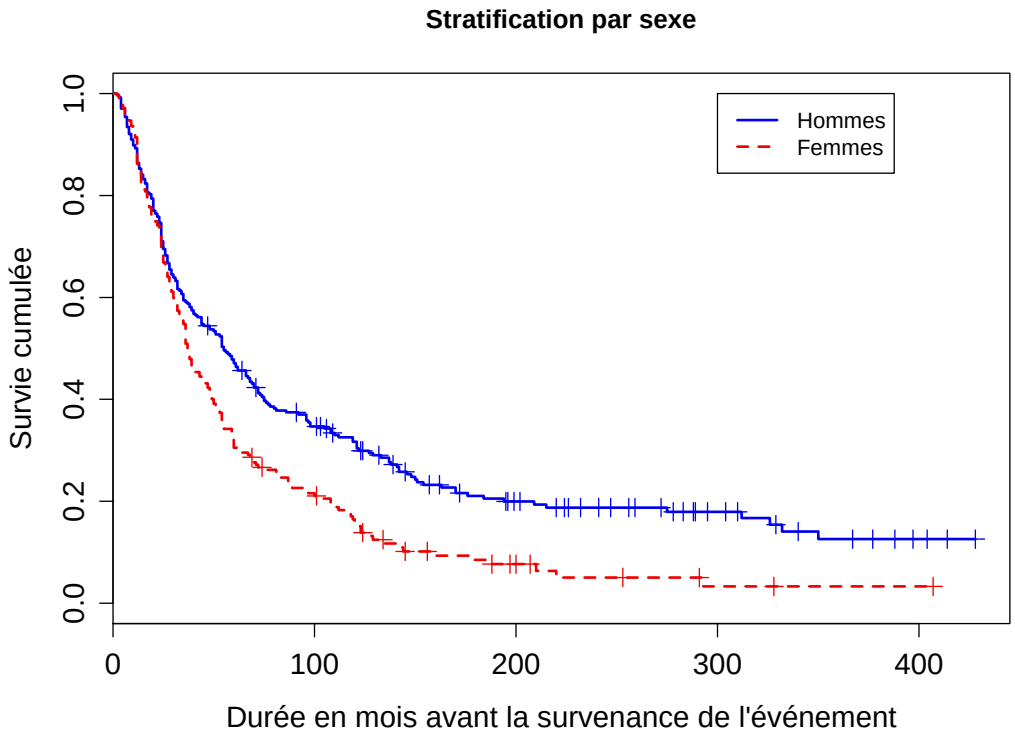
Données biographiques allemandes (2)

```
      exp(coef) exp(-coef) lower .95 upper .95
edu      1.0833     0.9231    1.0319    1.1372
lfx      0.9960     1.0041    0.9941    0.9978
pnoj     1.0939     0.9142    1.0021    1.1941
pres     0.9742     1.0265    0.9639    0.9847
coho2    1.4944     0.6692    1.1914    1.8743
coho3    1.3253     0.7545    1.0422    1.6853

Rsquare= 0.112   (max possible= 0.999 )
Likelihood ratio test= 71.08  on 6 df,   p=2.454e-13
Wald test            = 63.36  on 6 df,   p=9.326e-12
Score (logrank) test = 64.31  on 6 df,   p=5.953e-12
```

Fonctions de survie

- Mêmes β, mais risques $h_{s,0}(t)$ différents
  $\Rightarrow$ fonctions de survie $S_s(t, x)$ différentes



Comparaison avec modèles indépendants

- Le modèle de Cox stratifié ne revient pas à calculer des modèles séparément pour chaque niveau de la variable de stratification !
- Dans le cas de modèles séparés, les coefficients peuvent être différents d'un modèle à l'autre, alors que ce n'est pas le cas avec la stratification.

Données biographiques allemandes

- Hommes :

```
  n= 348, number of events= 245
          coef exp(coef)  se(coef)      z Pr(>|z|)
edu    0.057848  1.059554  0.033638  1.720 0.085482 .
lfx   -0.005355  0.994659  0.001408 -3.803 0.000143 ***
pnoj   0.087635  1.091590  0.057472  1.525 0.127300
pres  -0.015556  0.984565  0.008133 -1.913 0.055804 .
coho2  0.395812  1.485590  0.167169  2.368 0.017897 *
coho3  0.516832  1.676708  0.163724  3.157 0.001596 **
```

- Femmes :

```
  n= 252, number of events= 213
          coef exp(coef)  se(coef)      z Pr(>|z|)
edu    0.113008  1.119641  0.037958  2.977  0.00291 **
lfx   -0.003513  0.996493  0.001311 -2.680  0.00735 **
pnoj   0.116788  1.123881  0.076407  1.528  0.12639
pres  -0.033670  0.966891  0.007604 -4.428 9.52e-06 ***
coho2  0.387815  1.473757  0.159612  2.430  0.01511 *
coho3 -0.041775  0.959086  0.190908 -0.219  0.82679
```

Principe

- Le modèle stratifié part de l'hypothèse de non-interaction entre la variable de stratification et les variables explicatives du modèle.
- Pour tester l'hypothèse de non-interaction, plutôt que de calculer des modèles séparément pour chaque niveau de la variable de stratification, il est aussi possible de partir du modèle stratifié et de rajouter des termes d'interaction entre la variable de stratification et chaque variable explicative.
- On effectue ensuite un test du rapport de vraisemblance entre le modèle stratifié de départ et le modèle stratifié avec interactions.

Exemple : Données biographiques allemandes (1)

- Modèle stratifié avec interactions :

```
                      coef exp(coef)  se(coef)      z Pr(>|z|)
edu                0.057848  1.059554  0.033638  1.720 0.085482 .
lfx               -0.005355  0.994659  0.001408 -3.803 0.000143 ***
pnoj               0.087635  1.091590  0.057472  1.525 0.127300
pres              -0.015556  0.984565  0.008133 -1.913 0.055804 .
coho2              0.395812  1.485590  0.167169  2.368 0.017897 *
coho3              0.516832  1.676708  0.163724  3.157 0.001596 **
edu:sexwomen       0.055161  1.056710  0.050718  1.088 0.276776
lfx:sexwomen       0.001842  1.001843  0.001924  0.957 0.338382
pnoj:sexwomen      0.029152  1.029581  0.095609  0.305 0.760434
pres:sexwomen     -0.018114  0.982049  0.011135 -1.627 0.103772
coho2:sexwomen    -0.007997  0.992035  0.231131 -0.035 0.972400
coho3:sexwomen    -0.558607  0.572005  0.251498 -2.221 0.026343 *
---
Rsquare= 0.131   (max possible= 0.999 )
Likelihood ratio test= 84.21  on 12 df,   p=6.465e-13
```

Exemple : Données biographiques allemandes (2)

- Comparaison avec le modèle stratifié sans interactions :

```
Analysis of Deviance Table
 Cox model: response is  survie_emploi
 Model 1: ~ 1
 Model 2: ~ edu + lfx + pnoj + pres + coho2 + coho3 + strata(sex)
 Model 3: ~ edu + lfx + pnoj + pres + coho2 + coho3
            + sex:edu + sex:lfx + sex:pnoj + sex:pres + sex:coho2
            + sex:coho3 + strata(sex)
   loglik  Chisq Df P(>|Chi|)
1
2 -2219.8         6
3 -2213.2 13.125  6   0.04109 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (3)

- La statistique de test vaut

$$LR = 13.125 \sim \chi^2_{(6)}$$

- Les degré de liberté (6) correspondent au nombre d'interactions ajoutées dans le modèle.
- Pour un risque $\alpha = 5\%$, le seuil de rejet du test vaut 12.59 et la *p*-valeur vaut 0.04109.
- L'égalité des deux modèles est rejetée (mais de très peu) et l'hypothèse de non-interaction est donc aussi rejetée.
- Il serait préférable sur cet exemple de calculer un modèle incluant des interactions, ou alors de calculer des modèles séparés pour les femmes et les hommes.

Bibliographie

- Blossfeld HP, Rohwer G (2002) *Techniques of Event History Modeling, New Approaches to Causal Analysis* (2nd ed.). Mahwah NJ : Lawrence Erlbaum.
- Mayer KU, Brückner E (1989) *Lebensverläufe und Wohlfahrtsentwicklung. Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929-1931, 1939-1941, 1959-1951.* Materialen aus der Bildungsforschung. Berlin : Max-Planck Institut für Bildungsforschung.
- Raftery AE (1995) Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25 : 111-163.