Données longitudinales et modèles de survie

7. Développements

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

Définition

- Parfois, un même événement peut se produire à plusieurs reprises au fil du temps pour un même sujet. On parle alors d'événements à répétition.
- Exemples :
  - Mariage, naissance d'un enfant, changement d'emploi.
  - Panne d'une machine, révision d'une machine.
  - Hospitalisation, début de la prise d'un médicament.

Plusieurs approches

- Il existe au moins deux approches pour traiter les événements à répétition :
  1. La période observée pour un sujet est décomposée en autant de sous-intervalles qu'il y a d'événements observés (ou de censure dans le cas du dernier intervalle).
     $\Longrightarrow$ Counting process (processus de dénombrement).
  2. Un modèle différent est considéré pour chaque sous-intervalle en tenant compte de l'ordre de ceux-ci.
     $\Longrightarrow$ Modèles indépendants ou modèle stratifié.

Principe

- Soit un individu observé de t=0 à t=151 et ayant subi 3 événements aux temps 12, 36 et 129.
- Cet individu peut être décomposé sous la forme de 4 épisodes :

Intervalle	Début	Fin	Evénement
1	0	12	1
2	13	36	1
3	37	129	1
4	130	151	0

- Le principe de l'approche par dénombrement consiste à traiter ces différents épisodes comme s'ils concernaient des sujets différents.

Hypothèses (1)

- Pour que cette approche soit valide, il faut faire deux hypothèses :
  1. Les épisodes successifs d'une même personne peuvent être considérés comme indépendants les uns des autres.
  2. La distribution du hasard doit être la même pour chaque épisode.
- Par ailleurs, il faut considérer le fait que les épisodes d'un même sujet se succèdent au fil du temps et il faut donc tenir compte des dates de début et de fin de chaque épisode.

Hypothèses (2)

- Il est possible de limiter les problèmes causés par le non-respect des hypothèses de la façon suivante :
  1. Inclure comme variables explicatives des données concernant les épisodes précédents, comme par exemple le nombre d'épisodes précédents ou la durée de l'épisode précédent (zéro pour le premier intervalle).
  2. Ajuster les variances et écarts-types en fonction du nombre n de sujets plutôt que du nombre total N d'épisodes. De cette façon, des données fortement corrélées en provenance d'un même sujet n'auront pas plus d'importance que s'il n'y avait qu'un seul épisode.

Exemple : Données biographiques allemandes

- C'est exactement l'approche qui a été utilisée précédemment pour ces données !

```
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    0.106486  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
```

Modèles indépendants pour chaque événement

- La première possibilité consiste à calculer autant de modèles indépendants que d'épisodes observés au maximum chez un sujet.
- Par exemple, si dans la base de données un sujet est décomposé au maximum en 5 épisodes, alors 5 modèles séparés seront calculés :
  - Modèle 1 pour le premier épisode de chaque sujet.
  - Modèle 2 pour le deuxième épisode de chaque sujet (lorsque cet épisode existe).
  - Modèle 3 ...
- Cette approche peut être très fastidieuse et les résultats difficiles à interpréter.

Exemple : Données biographiques allemandes (1)

- Distribution du nombre d'épisodes :

1	2	3	4	5	6	7	8	9
201	162	107	62	32	20	11	4	1

- Plus on avance dans les épisodes, moins de données sont à disposition pour l'estimation.

Exemple : Données biographiques allemandes (2)

- Episode 1 :

```
Call:
coxph(formula = survie_emploi_D1 ~ D1$edu + D1$lfx + D1$pnoj +
    D1$pres + D1$coho2 + D1$coho3)

  n= 201, number of events= 185

              coef exp(coef)  se(coef)      z Pr(>|z|)
D1$edu    0.064077  1.066174  0.032715  1.959  0.05016 .
D1$lfx          NA        NA  0.000000     NA       NA
D1$pnoj         NA        NA  0.000000     NA       NA
D1$pres  -0.012192  0.987882  0.008948 -1.362  0.17305
D1$coho2  0.535879  1.708949  0.192461  2.784  0.00536 **
D1$coho3  0.428609  1.535121  0.183240  2.339  0.01933 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (3)

- Episode 2 :

```
Call:
coxph(formula = survie_emploi_D2 ~ D2$edu + D2$lfx + D2$pnoj +
    D2$pres + D2$coho2 + D2$coho3)

  n= 162, number of events= 126

              coef exp(coef)  se(coef)      z Pr(>|z|)
D2$edu    0.109260  1.115453  0.054259  2.014 0.044042 *
D2$lfx   -0.006540  0.993481  0.001847 -3.541 0.000398 ***
D2$pnoj         NA        NA  0.000000     NA       NA
D2$pres  -0.046276  0.954779  0.011503 -4.023 5.74e-05 ***
D2$coho2  0.226766  1.254537  0.218187  1.039 0.298655
D2$coho3  0.064858  1.067007  0.236691  0.274 0.784071
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (4)

- Episode 4 :

```
Call:
coxph(formula = survie_emploi_D4 ~ D4$edu + D4$lfx + D4$pnoj +
    D4$pres + D4$coho2 + D4$coho3)

  n= 62, number of events= 38

              coef exp(coef)  se(coef)      z Pr(>|z|)
D4$edu    0.097376  1.102275  0.134992  0.721  0.47070
D4$lfx   -0.005709  0.994308  0.002654 -2.151  0.03147 *
D4$pnoj         NA        NA  0.000000     NA       NA
D4$pres  -0.071334  0.931151  0.021185 -3.367  0.00076 ***
D4$coho2  0.784214  2.190685  0.403677  1.943  0.05206 .
D4$coho3 -0.026548  0.973801  0.472170 -0.056  0.95516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Données biographiques allemandes (5)

- Episode 8 :

```
Call:
coxph(formula = survie_emploi_D8 ~ D8$edu + D8$lfx + D8$pnoj +
    D8$pres + D8$coho2 + D8$coho3)

  n= 4, number of events= 1

              coef exp(coef) se(coef)      z Pr(>|z|)
D8$edu      0.0000    1.0000   0.0000     NA       NA
D8$lfx     -0.3847    0.6807 730.7812 -0.001        1
D8$pnoj     0.0000    1.0000   0.0000     NA       NA
D8$pres    -0.0112    0.9889   0.0000   -Inf   <2e-16 ***
D8$coho2    0.0000    1.0000   0.0000     NA       NA
D8$coho3    0.0000    1.0000   0.0000     NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle stratifié

- La seconde possibilité consiste à traiter chaque épisode comme une strate différente d'un même modèle.
- En pratique, il suffit de disposer d'une variable numérotant par ordre chronologique les épisodes d'un même sujet et de stratifier par rapport à celle-ci.

Exemple : Données biographiques allemandes

- Stratification par rapport à la variable *sn* :

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

  n= 600, number of events= 458

           coef exp(coef)  se(coef)      z Pr(>|z|)
edu    0.068147  1.070523  0.025069  2.718 0.006560 **
lfx   -0.003993  0.996015  0.000990 -4.033 5.51e-05 ***
pnoj         NA        NA  0.000000     NA       NA
pres  -0.026432  0.973914  0.005565 -4.750 2.04e-06 ***
coho2  0.413156  1.511580  0.116587  3.544 0.000394 ***
coho3  0.293924  1.341682  0.123473  2.380 0.017291 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plan du cours

Définition

- Parfois, le phénomène étudié se compose de plusieurs événements distincts que l'on souhaite étudier conjointement. On parle alors d'événements multiples.
- Exemples :
  - Fin de scolarité, premier emploi, second emploi, mariage, naissance du premier enfant, naissance du deuxième enfant, veuvage, décès.
  - Mise en production d'une nouvelle machine, panne de la machine, révision de la machine.
  - Déclaration d'une maladie, début d'un premier traitement, hospitalisation, arrêt du premier traitement, début d'un deuxième traitement.

Complexité

- Le traitement de ce type d'événements peut être très complexe, notamment en raison des liens possibles entre les événements eux-mêmes (successifs, mutuellement exclusifs, ...).
- De nombreuses approches ont été développées, aucune n'étant une panacée.
- En pratique, peu d'approches sont vraiment implémentées sur les logiciels statistiques courants.

Exemple : Divorce

- Nous considérons un fichier personnes-périodes construit à partir de variables issues de 11 vagues du Panel Suisse des Ménages, de 1999 à 2009.
- Nous ne considérons que des personnes qui étaient mariées en 1999.
- L'objectif est d'expliquer le divorce en fonction du sexe et de l'âge de la personne.
- Un second événement est considéré : avoir vu tous ses enfants quitter le domicile familial. Il est représenté par une variable *NoKid* codée 1 dès lors que tous les enfants ont quitté le domicile familial et zéro sinon.

Relation entre les événements

- Lorsque les événements étudiés sont liés les uns aux autres, il est souvent possible d'inclure dans l'analyse d'un événement B une ou plusieurs variables dépendant de la réalisation ou non d'un événement A.
- Par exemple, dans une étude sur la criminalité, il est possible d'inclure comme facteur prédictif d'une arrestation pour attaque à main armée une variable indiquant si l'on a déjà été arrêté pour d'autres types de vols.
- De même, le mariage peut être utilisé comme facteur prédictif de la naissance d'un enfant, de l'achat d'une villa, etc ...

Exemple : Divorce

- L'événement *divorce* est expliqué par le sexe, l'âge et la variable NoKid.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.02957    0.52409  -5.781 7.44e-09 ***
DM$SEXwoman -0.02440    0.23623  -0.103    0.918
DM$AGE      -0.06874    0.01168  -5.884 4.00e-09 ***
DM$NoKid    -0.37625    0.72711  -0.517    0.605
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1053.8  on 36655  degrees of freedom
Residual deviance: 1008.7  on 36652  degrees of freedom
AIC: 1016.7

Number of Fisher Scoring iterations: 10
```

Evénements complexes

- Dans le cas où le phénomène d'intérêt est la co-occurence, simultanée ou non, de différents événements, il est possible de définir un événement complexe à partir de plusieurs événements simples et de faire porter l'analyse sur la survenance ou non de cet événement complexe.
- Par exemple, on peut être intéressé par la conjonction entre une arrestation pour vol et une condamnation effective pour ce délit.
- Dans le domaine médical, il peut être intéressant d'étudier les facteurs prédisant qu'une sortie d'hôpital soit suivie par une réadmission pour le même motif dans le mois suivant.

Exemple : Divorce (1)

- L'événement étudié est la co-occurence simultanée du *divorce* et de NoKid==1.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -25.99379 2627.33039  -0.010    0.992
DM$SEXwoman   17.41156 2627.32893   0.007    0.995
DM$AGE        -0.01106    0.05667  -0.195    0.845

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.265  on 36655  degrees of freedom
Residual deviance: 40.453  on 36653  degrees of freedom
AIC: 46.453

Number of Fisher Scoring iterations: 25
```

Exemple : Divorce (2)

- Dans ce cas, le modèle ne converge pas, car le nombre d'événements observés (2) est trop faible :

```
> table(DM$statut,DM$NoKid)

       0     1
  0 34325  2258
  1    71     2
```

Evénements indépendants

- Lorsque les différents événements possibles ne sont absolument pas liés les uns aux autres, chaque événement peut être étudié à l'aide d'un modèle séparé.
- Ces différents modèles peuvent très bien utiliser des facteurs explicatifs différents.
- Il est possible de montrer alors que la vraisemblance des données peut être factorisée en autant de vraisemblances qu'il y a d'événements.

Exemple : Divorce et départ des enfants (1)

- Modèle pour le divorce seulement :

```
Call:
glm(formula = DM$statut ~ DM$SEX + DM$AGE, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.99133    0.52204  -5.730 1.00e-08 ***
DM$SEXwoman -0.02718    0.23619  -0.115    0.908
DM$AGE      -0.06993    0.01157  -6.046 1.49e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1053.8  on 36655  degrees of freedom
Residual deviance: 1009.0  on 36653  degrees of freedom
AIC: 1015

Number of Fisher Scoring iterations: 10
```

Exemple : Divorce et départ des enfants (2)

- Modèle pour le départ des enfants seulement :

```
Call:
glm(formula = DM2$NoKid ~ DM2$SEX + DM2$time, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.19491    0.10970 -47.354   <2e-16 ***
DM2$SEXwoman  0.01694    0.09140   0.185    0.853
DM2$time      0.17379    0.01427  12.181   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5136.1  on 34883  degrees of freedom
Residual deviance: 4989.8  on 34881  degrees of freedom
AIC: 4995.8

Number of Fisher Scoring iterations: 7
```