Données longitudinales et modèles de survie

5. Modèles de régression en temps discret

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

Données

- Nous disposons de données observées en temps discret avec une périodicité fixe : jour, mois année, ...

$$t_k, k = 1, 2, \ldots, m$$

  Si l'unité de temps est le mois, $t_k = k^{\text{ème}}$ mois.
- Nous connaissons la situation de chaque individu au début et à la fin de chaque période, mais nous ne savons pas à quel moment précis de la période un événement, ou un changement de la valeur d'une covariable, est intervenu.
- Nous voulons exprimer le risque instantané ou la fonction de survie en fonction de facteurs explicatifs x :

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

Exemple : Données biographiques allemandes

- 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 n=201 personnes ayant eu de 1 à 9 emplois.
- 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 ?
- On souhaite aussi contrôler les effets par cohorte.

Temps discret

- La situation de chaque individu à chaque date t_k où il est exposé au risque de connaître l'événement étudié (quitter ses parents, mariage, fin d'emploi, ...) constitue une observation.
- Pour chaque individu, il y a autant d'observations que de dates t_k (mois, années, ...) où il est exposé au risque.
  $\longrightarrow$ On est donc amené à travailler avec un fichier de données "personne-période".
- Par exemple, si on s'intéresse au divorce, chaque année de mariage constitue une observation.
- Pour simplifier, on omet l'indice k et on note $t = 1, 2, \ldots, m$.

Données personnes-périodes

- La principale difficulté de la modélisation en temps discret réside dans l'organisation des données sous forme personnes-périodes.
- Le fichier résultant a aussi le problème de la taille : Le nombre de lignes se calcule comme

$$\sum_{i=1}^{n} n_i$$

où n_i est le nombre d'observations de la i-ième personne.

- En revanche, l'utilisation de données sous forme personnes-périodes permet de pouvoir ensuite estimer le modèle sans procédure particulière.

Exemple : Données biographiques allemandes

- Le fichier original comporte **600 lignes**.
- Le fichier personnes-périodes comporte **40782 lignes** qui correspondent chacune à 1 mois d'observation d'un individu.
- Le fichier personnes-périodes comporte deux variables supplémentaires par rapport au fichier original :
  - **desti** : variable indiquant les mois (lignes du fichier) durant lesquels l'événement étudié (fin de l'emploi) est survenu ;
  - **month** : variable numérotant les mois de chaque observation originale et représentant donc le temps écoulé depuis le début de l'emploi.

Plan du cours

Statut et risque instantané

- A chaque observation personne-période on associe une variable de statut (**desti** dans l'exemple des données biographiques allemandes) qui prend la valeur
  - 1 pour la dernière observation de chaque individu (celle où l'événement survient) ;
  - 0 pour les autres observations du même individu.
- Le risque instantané est défini comme

$$h_t = P(\text{statut}_t = 1 | \text{statut}_s = 0, \quad \forall\, s < t)$$

- Cela correspond à la probabilité que le statut prenne la valeur 1 en t parmi les observations personnes-périodes.

Estimation du risque instantané

- A partir d'un échantillon de données, le risque instantané est estimé par

$$\hat{h}_t = \frac{d_t}{R_t}$$

avec

  - d_t : nombre d'événements en t
  - R_t : nombre de cas à risque en t (= nombre d'observations à la date t)

- L'objectif est de modéliser le risque instantané h_t.

Principaux modèles en temps discrets

- Les modèles en temps discrets expriment une transformation du risque h_t comme une fonction linéaire $(x'\beta)$ des variables indépendantes.
- Il existe plusieurs types de modélisation sur des données personnes-périodes dont les principaux sont
  - Régression logistique : $\text{logit}(h_t) = \log\left(h_t/(1-h_t)\right)$
  - Régression log-log du complémentaire : $\log\left(-\log(1-h_t)\right)$ (complementary log-log)
  - Modèle log-taux de type log-linéaire : $\log(h_t) = \log(d_t/R_t)$
- Nous nous concentrerons ici sur l'approche par la régression logistique.

Rappels (1)

- Objectif : Mesurer l'impact de différents facteurs sur une variable binaire (ou dichotomique).
- Soit par exemple la variable "être actuellement marié". Cette variable n'a que deux modalités : oui ou non. Soit p la probabilité d'être actuellement marié (oui) et 1-p la probabilité de ne pas l'être (non).
- Pour un échantillon de taille n, la cote (odds) associée à cette variable est le rapport entre le nombre n_{oui} de personnes mariées et le nombre n_{non} de personnes non-mariées, avec $n = n_{\text{oui}} + n_{\text{non}}$, ce qui revient à écrire

$$\frac{n_{\text{oui}}}{n_{\text{non}}} = \frac{n_{\text{oui}}/n}{n_{\text{non}}/n} = \frac{p}{1-p}$$

Rappels (2)

- Le logit est le logarithme de la cote :

$$\log\left(\frac{p}{1-p}\right)$$

- Le modèle de régression logistique est alors défini comme

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

- L'exponentiel des coefficients ($\exp(\beta)$) mesure par combien la cote de la variable expliquée est multipliée lorsque le facteur x correspondant augmente d'une unité. C'est ce que l'on appelle un *odds ratio*.

Régression logistique pour risque instantané (1)

- Au lieu d'exprimer directement le risque instantané h_t en fonction des facteurs explicatifs x, on modélise le rapport des cotes en fonction de x :

$$\begin{aligned}\gamma(x) &= \frac{h_t(x)/(1-h_t(x))}{h_t(0)/(1-h_t(0))} \\ &= \exp(x'\beta) = \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)\end{aligned}$$

- Par ailleurs,

$$\frac{h_t(x)}{1-h_t(x)} = \gamma(x)\ \frac{h_t(0)}{1-h_t(0)}$$

Régression logistique pour risque instantané (2)

- En prenant le logarithme, on obtient un modèle de régression logistique :

$$\underbrace{\log\left(\frac{h_t(x)}{1-h_t(x)}\right)}_{\text{logit}} = \beta_{0t} + \sum_{k=1}^{q} \beta_j x_j$$

avec $\beta_{0t} = \log\left(\frac{(h_t(0)}{(1-h_t(0)}\right)$.

Constance des ratios de cotes et évolution des risques

- Soit $x = (x_1, \ldots, x_k)$ le vecteur des facteurs explicatifs (covariables).
- La cote $h_t(x)/(1 - h_t(x))$ et le logit prédits restent **constants** dans le temps pour chaque individu lorsque
  - $\beta_{0t} = \beta_0$ (identique à chaque période t) et
  - x ne contient que des variables indépendantes de t.
- La cote $h_t(x)/(1 - h_t(x))$ et le logit prédits **changent** avec le temps lorsque
  - il existe t et s tels que $\beta_{0t} \neq \beta_{0s}$ ou
  - x contient des variables qui dépendent du temps t.

Modèles particuliers dépendant du temps

- Deux cas particuliers :
  1. $\beta_{0t} = \beta_0 t \Rightarrow$ Gompertz
  2. $\beta_{0t} = \beta_0 \ln(t) \Rightarrow$ Weibull

- Remarque : Le rapport des cotes

$$\frac{h_t(x_1)/(1-h_t(x_1))}{h_t(x_2)/(1-h_t(x_2))}$$

  pour 2 individus reste constant dans le temps si x_1 et x_2 ne changent pas avec t.
  $\Rightarrow$ Modèle à cotes proportionnelles.
- Attention : Ratio de cotes constant $\not\Rightarrow$ cotes constantes.

Exemple : Modèle de base (1)

```
Call:
glm(formula = Data_pp$desti ~ Data_pp$edu + Data_pp$lfx
       + Data_pp$pnoj + Data_pp$pres + Data_pp$coho2
       + Data_pp$coho3, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3497  -0.1944  -0.1607  -0.1393   3.2645

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.9692941  0.2378941 -20.889  < 2e-16 ***
Data_pp$edu    0.0910435  0.0216718   4.201 2.66e-05 ***
Data_pp$lfx    0.0011143  0.0006608   1.686   0.0917 .
Data_pp$pnoj   0.0350775  0.0346604   1.012   0.3115
Data_pp$pres  -0.0205894  0.0048509  -4.244 2.19e-05 ***
Data_pp$coho2  0.6410463  0.1023109   6.266 3.71e-10 ***
Data_pp$coho3  0.8900750  0.1060624   8.392  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Modèle de base (2)

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6254.0  on 40781  degrees of freedom
Residual deviance: 6153.2  on 40775  degrees of freedom
AIC: 6167.2

Number of Fisher Scoring iterations: 7

Odds ratios:
(Intercept)     0.0069
Data_pp$edu     1.0953
Data_pp$lfx     1.0011
Data_pp$pnoj    1.0357
Data_pp$pres    0.9796
Data_pp$coho2   1.8985
Data_pp$coho3   2.4353

n=600,    R2 de Nagelkerke: 0.0174,    AIC: 6167.2,    BIC: 6197.9
Statistique du rapport de vraisemblance:  100.8 (df=6)
```

Exemple : Modèle avec variable “married” (1)

```
Call:
glm(formula = Data_pp$desti ~ Data_pp$edu + Data_pp$lfx
       + Data_pp$pnoj + Data_pp$pres + Data_pp$married
       + Data_pp$coho2 + Data_pp$coho3, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3363  -0.1948  -0.1610  -0.1398   3.2601

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.9301404  0.2385659 -20.666  < 2e-16 ***
Data_pp$edu      0.0904058  0.0217074   4.165 3.12e-05 ***
Data_pp$lfx      0.0013971  0.0006887   2.029   0.0425 *
Data_pp$pnoj     0.0366093  0.0346728   1.056   0.2910
Data_pp$pres    -0.0198366  0.0048793  -4.065 4.79e-05 ***
Data_pp$married -0.1377834  0.0926682  -1.487   0.1371
Data_pp$coho2    0.6381243  0.1024194   6.231 4.65e-10 ***
Data_pp$coho3    0.9014608  0.1064731   8.467  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Modèle avec variable “married” (2)

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6254  on 40781  degrees of freedom
Residual deviance: 6151  on 40774  degrees of freedom
AIC: 6167

Number of Fisher Scoring iterations: 7

Odds ratios:
(Intercept)        0.0072
Data_pp$edu        1.0946
Data_pp$lfx        1.0014
Data_pp$pnoj       1.0373
Data_pp$pres       0.9804
Data_pp$married    0.8713
Data_pp$coho2      1.8929
Data_pp$coho3      2.4632

n=600,    R2 de Nagelkerke: 0.0175,    AIC: 6167.0,    BIC: 6202.1
Statistique du rapport de vraisemblance:  103 (df=7)
```

Plan du cours

Evaluation globale (1)

On peut considérer les statistiques suivantes :

- La déviance (-2 Log Likelihood) qui donne la “distance” entre le modèle et les observations. Elle est utile pour comparer des modèles.
- La statistique chi-2 du rapport de vraisemblance qui évalue l’amélioration de la déviance par rapport au modèle “NULL” (avec constante seulement). Si cette statistique n’est pas significative, le modèle est rejeté.
- Les degrés de liberté d de la statistique du chi-2 correspondent au nombre de paramètres supplémentaires introduits.

Evaluation globale (2)

- Les pseudo-R^2 :
  - Cox & Snell :

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

  - Nagelkerke :

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

où $-2LL_0$ est la déviance du modèle NULL, $-2LL_M$ est la déviance du modèle que l'on teste et n est le nombre de lignes du fichier personnes-périodes.

Evaluation globale (3)

- Les coefficients d'information AIC et BIC :
  - AIC = $-2LL + 2k$
  - BIC = $-2LL + k\log(n)$

  où k est le nombre de paramètres du modèle et n est le nombre d'événements observés.

Evaluation des effets de chaque variable (1)

- La significativité des coefficients détermine si l'effet correspondant est significatif ou s'il peut être supprimé du modèle.
  - Variable quantitative : C'est l'effet de la variable elle-même, car il n'y a qu'un seul coefficient.
  - Variable catégorielle : La significativité d'un coefficient indique uniquement l'effet significatif de la variable muette correspondante.
- Significativité globale d'une variable catégorielle :
  - Lorsqu'une variable catégorielle est représentée par plusieurs variables muettes dans un modèle, la statistique du chi-2 entre le modèle sans et le modèle avec la variable mesure l'amélioration engendrée par l'introduction de la variable testée et donc sa significativité globale.

Evaluation des effets de chaque variable (2)

- Les odds ratios eux-mêmes s'interprètent de la manière suivante :
  - Variable quantitative : C'est l'impact sur la cote des risques d'une augmentation de 1 unité de la variable.
  - Variable catégorielle : C'est l'impact sur la cote des risques du fait d'appartenir à la catégorie de cette variable indicatrice par rapport au fait d'appartenir à la catégorie de référence.

Exemple : Données biographiques allemandes

```
Analysis of Deviance Table

Model 1: desti ~ 1
Model 2: desti ~ edu + lfx + pnoj + pres
Model 3: desti ~ edu + lfx + pnoj + pres + coho2 + coho3
Model 4: desti ~ edu + lfx + pnoj + pres + married + coho2 + coho3

  Resid. Df Resid. Dev Df Deviance
1     40781     6254.0
2     40777     6233.7  4   20.352
3     40775     6153.2  2   80.503
4     40774     6151.0  1    2.198
```

Plusieurs constantes

- Etant donné que le risque instantané, et donc l'odds ratio correspondant, n'est pas forcément constant au fil du temps, il peut être utile de remplacer la constante unique du modèle par autant de constantes que de périodes.
- Chaque constante s'interprète alors comme la valeur de base, en l'absence d'influence des facteurs explicatifs, de l'odds ratio du risque de la période correspondante.
- Si les données sont observées sur de très nombreuses périodes, il devient difficile d'utiliser cette approche, car l'estimation du modèle devient pratiquement impossible et le nombre de constantes peut être trop grand par rapport au nombre de données disponibles.
- Si l'événement n'a pas été observé durant une période, la constante correspondante devrait être supprimée.

Exemple : Panel Suisse des ménages (1)

- 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, de l'âge et du nombre d'enfants vivant avec la personne.
- Deux modèles sont calculés, le premier avec une seule constante, le second avec 11 constantes correspondant chacune à l'une des 11 vagues.

Exemple : Panel Suisse des ménages (2)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1935  -0.0561  -0.0443  -0.0340   4.0169

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.97175    0.88337  -2.232  0.02561 *
DM$SEXwoman -0.01239    0.34758  -0.036  0.97156
DM$AGE      -0.08681    0.01716  -5.058 4.24e-07 ***
DM$NBKID    -0.54732    0.18466  -2.964  0.00304 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemple : Panel Suisse des ménages (3)

```
glm(formula = DM$statut ~ 0 + DM$X1 + DM$X2 + DM$X3 + DM$X4 +
    DM$X5 + DM$X6 + DM$X7 + DM$X8 + DM$X9 + DM$X10 + DM$X11 +
    as.numeric(DM$SEX) + DM$AGE + DM$NBKID, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2765  -0.0572  -0.0396  -0.0241   4.1277

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
DM$X1             -17.34601  814.67490  -0.021 0.983013
DM$X2              -1.14327    2.22153  -0.515 0.606810
DM$X3              -1.19351    2.24510  -0.532 0.595000
DM$X4              -0.52243    2.24468  -0.233 0.815964
DM$X5              -0.31367    2.25557  -0.139 0.889398
DM$X6              -1.23975    2.33701  -0.530 0.595774
DM$X7              -0.37337    2.29547  -0.163 0.870792
DM$X8              -1.65869    2.46141  -0.674 0.500390
DM$X9              -1.55488    2.46902  -0.630 0.528853
DM$X10             -0.39549    2.33606  -0.169 0.865561
DM$X11             -0.72301    2.37889  -0.304 0.761182
as.numeric(DM$SEX) -0.03247    0.34778  -0.093 0.925615
DM$AGE             -0.10256    0.01924  -5.331 9.75e-08 ***
DM$NBKID           -0.63145    0.18768  -3.365 0.000767 ***
```

Exemple : Panel Suisse des ménages (4)

```
Analysis of Deviance Table

Model 1: DM$statut ~ DM$SEX + DM$AGE + DM$NBKID
Model 2: DM$statut ~ 0 + DM$X1 + DM$X2 + DM$X3 + DM$X4 + DM$X5 + DM$X6 +
    DM$X7 + DM$X8 + DM$X9 + DM$X10 + DM$X11 + as.numeric(DM$SEX) +
    DM$AGE + DM$NBKID
  Resid. Df Resid. Dev Df Deviance
1     25126     491.92
2     25116     471.93 10       20
```

Construction de la courbe de survie

- Avec les estimations des coefficients de la régression, on peut estimer le logit(h) du risque pour un profil donné.
- A partir de l'estimation $\widehat{\text{logit}(h)}$, on déduit l'estimation de $\hat{h}$:

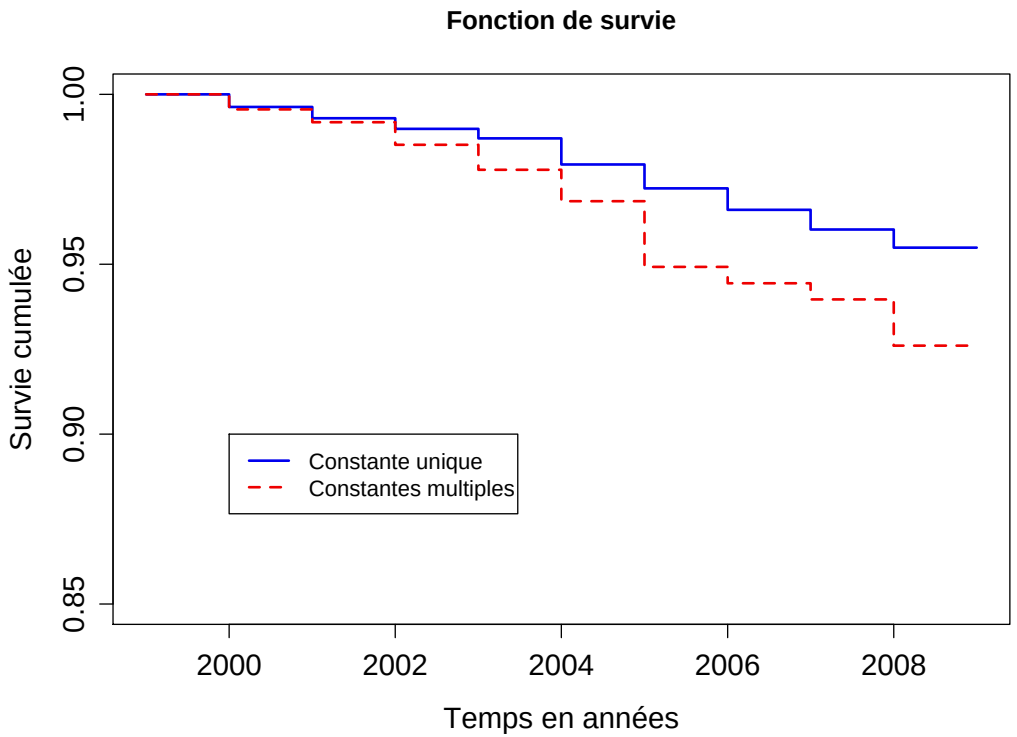
$$\hat{h} = \frac{e^{\widehat{\text{logit}(h)}}}{1 + e^{\widehat{\text{logit}(h)}}} = \frac{1}{1 + e^{-\widehat{\text{logit}(h)}}}$$

- Si le risque $\hat{h}$ varie avec t, on calcule $\hat{h}_t$ pour chaque t.
- A partir de l'estimation des risques instantanés $\hat{h}_t$, on estime la survie (en partant de $\hat{S}_0 = 1$) :

$$\hat{S}_t = \hat{S}_{t-1} \cdot (1 - \hat{h}_t)$$

Exemple

- Nous considérons une femme ayant 28 ans en 1999, vivant avec deux enfants de 1999 à 2003, puis n'ayant plus d'enfants avec elle.



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.