Le modèle de Poisson avec sur-représentation de zéros
Une application démographique

Rencontres méthodes et recherche

UNIL - FORS

Reto Schumacher

Institut d'études démographiques et du parcours de vie,
Université de Genève

27 septembre 2011

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Institut d'études démographiques et du parcours de vie

plan

introduction

- présentation basée sur une contribution au XVe Colloque national de démographie 2010 (Conférence universitaire de démographie et d'étude des populations - CUDEP)
- objectif : proposer un outil statistique pour détecter un comportement fécond particulier : le comportement d'arrêt de la fécondité (birth stopping) dans une population transitionnelle
- cette présentation focalise sur les modèles de comptage et le problème de surdispersion

qu'est-ce qu'un taux démographique ?

- le **taux** est un terme à significations diverses :
  - ▷ taux de chômage : **proportion**
  - ▷ taux de change : **rapport, ratio**
  - ▷ taux d'inflation : **variation**
- en **démographie** : rapport entre un nombre d'événements (démographiques) et un nombre d'unités de temps exposées au risque de l'événement ("**occurence/exposure rate**")
- le taux démographique mesure l'**intensité** d'un événement, fréquence relative compte tenu de la durée d'exposition

  6 naissances de 10 femmes observées pendant 2 ans =
  3 naissances de 10 femmes observées pendant 1 an =
  3 naissances de 5 femmes observées pendant 2 ans <
  3 naissances de 5 femmes observées pendant 1 an =
  3 naissances de 10 femmes observées pendant 6 mois

notion de taux

- la période de risque souvent délimitée par d'autres événements
  - ▷ le "risque" d'une naissance limité aux femmes en âge de procréer (en démographie : 15-49)
  - ▷ le risque d'une 2e naissance limité aux mères d'un enfant
  - ▷ le risque d'un divorce limité aux personnes mariées
- on peut distinguer différents types d'événement :
  - ▷ événement **fatal** $\rightarrow$ décès
  - ▷ événement **non fatal** $\rightarrow$ mariage, divorce, naissance ...
  - ▷ événement **unique** $\rightarrow$ décès, 1er mariage, 1ère naissance ...
  - ▷ événement **répétable** $\rightarrow$ naissance, mariage, divorce ...

analyse démographique longitudinale

- trois perspectives d'analyse longitudinale en démographie :
  - ▷ calendrier des événements (**timing** → when ?)

    analyse de survie (time to occurrence)
  - ▷ ordre des événements (**sequencing** → in which order ?)

    analyse des sequences (event sequences)
  - ▷ fréquence des événements (**quantum** → how many ?)

    analyse des taux (occurrene/exposure, count models)

    → la fréquence des événements est souvent modélisée par **un modèle de Poisson**

distribution de Poisson

- la **loi de Poisson** décrit la distribution d'événements (rares) survenant dans un intervalle de temps ou dans une unité de surface de façon aléatoire et indépendante

$$P(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

- μ est la moyenne de la distribution
- μ est aussi la variance : $\text{Var}(y) = \mu \rightarrow$ **equidispersion**
  souvent $\text{Var}(y) > \mu \rightarrow$ **surdispersion**
- quand μ augmente, $P(y = 0)$ diminue
  le nombre de 0 souvent plus important que prédit

le modèle de Poisson

- modèle de comptage (count model)
- modèle linéaire généralisé (GLM) avec

  distribution de probabilité : **Poisson**
  fonction de lien : **log**

$$
\begin{aligned}
\mathrm{E}(\mathbf{Y}) &= \boldsymbol{\mu} &&= \exp(\mathbf{X}\boldsymbol{\beta}) \\
\mathrm{E}(y_i|X_i) &= \mu_i &&= \exp(X_i\beta) \\
\ln y_i &= && \textstyle\sum_k \beta_k X_{ki} \\
&= && \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \\
\hat{y}_i &= && \exp\left(\textstyle\sum_k \beta_k X_{ki}\right) \\
\exp\beta &= && \mathrm{IRR}
\end{aligned}
$$

le modèle de Poisson avec offset

- modéliser des taux plutôt que des fréquences :
- "pondérer" les nombres d'événements par la durée d'exposition au risque en définissant un **offset** :

$$\begin{aligned} \mu_i &= t_i \lambda_i \\ \ln(\lambda_i) &= \ln\left(\frac{y_i}{t_i}\right) = \sum_k \beta_k X_{ki} \\ \ln(y_i) &= \underbrace{\ln(t_i)}_{\text{offset}} + \underbrace{\sum_k \beta_k X_{ki}}_{\ln(\lambda_i)} \end{aligned}$$

t_i	durée d'exposition au risque
λ_i	taux
$\ln(t_i)$	offset (pondération) = variable dont coefficient fixé à 1

problème de surdispersion

- **surdispersion** (overdispersion) : $\mathrm{Var}(Y) > \mu$

  causes : hétérogénéité non observée,
  sur-représentation de zéros
  conséquences : sous-estimation des erreurs standars,
  problèmes d'inférence (cf. hétéroscedasticité)

- **solutions** :
  - ▷ régression binomiale négative (**NBR1 et 2**)
  - ▷ **hurdle models** : logit + Poisson tronqué (ZTP) :
    zéros et nombres non nuls générés par processus distincts
  - ▷ régression de **Poisson avec sur-représentation de zéros**
    (Zero Inflated Poisson) : zéros générés par processus distincts

modèle de Poisson avec sur-représentation de zéros

- le modèle ZIP suppose la présence de groupes latents :
  1. **groupe** A : toujours 0
  2. **groupe** $\overline{A}$: ne pas toujours 0
- le modèle estime simultanément l'appartenance au groupe A ($\mu_i = 0$) par un modèle binaire (logit ou probit) et le nombre d'événements ($\mu_i \geq 0$) pour les membres du groupe $\overline{A}$ par un modèle de Poisson

$$\begin{aligned} P(A_i = 1) &= \phi_i \\ P(A_i = 0) &= 1 - \phi_i \\ P(Y_i = 0 | A_i = 1, X, Z) &= 1 \\ P(Y_i = 0 | A_i = 0, X, Z) &= \frac{e^{-\mu_i} \mu_i^0}{0!} \end{aligned}$$

une application démographique du ZIP

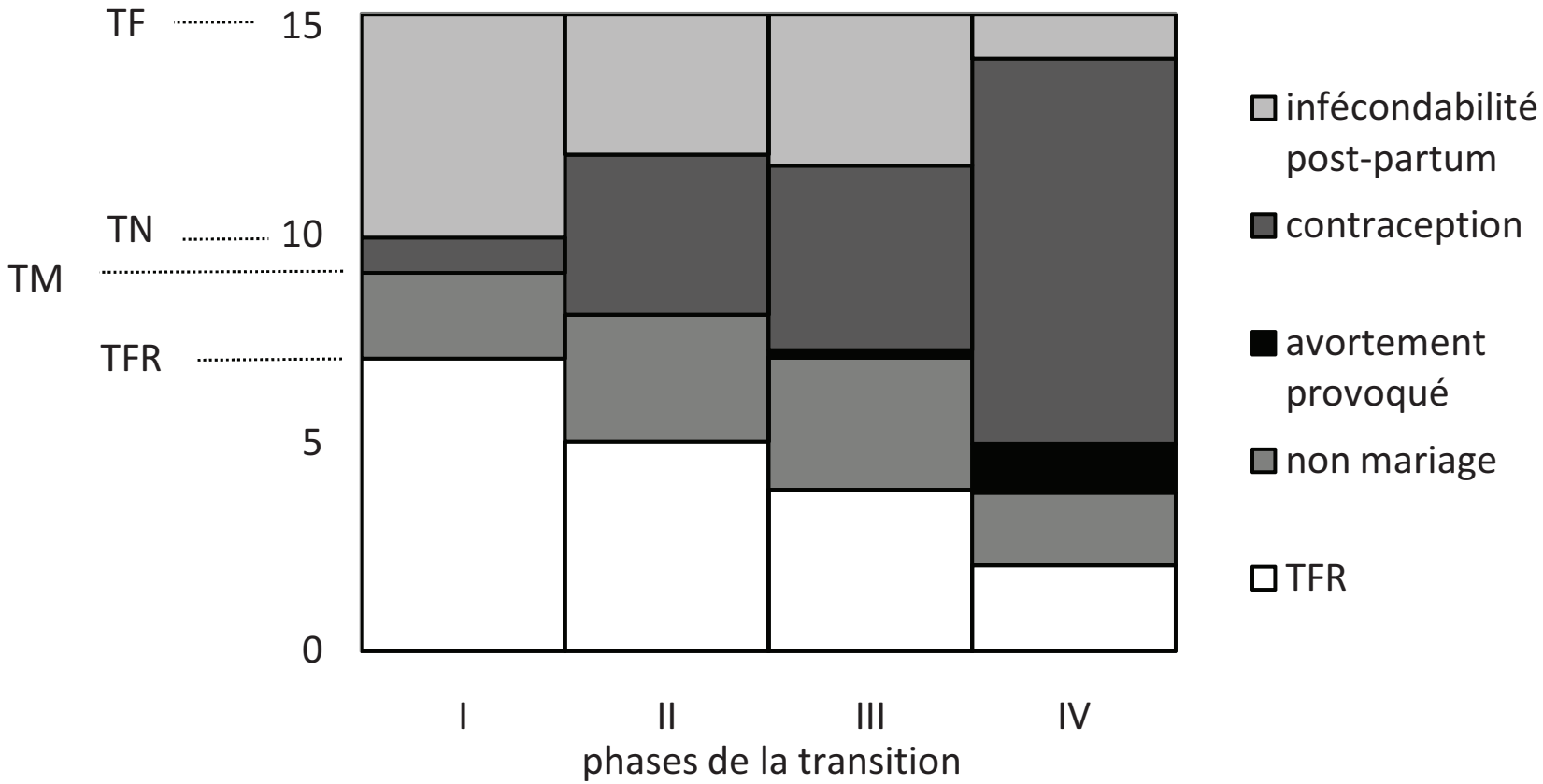
- **application du modèle ZIP** :

  détecter des sous-groupes pratiquant le comportement d'arrêt de la fécondité dans la population genevoise du 19e siècle
- modéliser conjointement, à chaque parité de naissance atteinte, la probabilité d'appartenir au groupe A des contracepteurs et l'intensité de la fécondité dans le groupe $\overline{A}$
- analyse des taux de fécondité légitime par classe d'âge quinquennale et par rang de naissance atteint

some background information

- la première transition de la fécondité européenne résulte du passage d'un comportement reproductif indépendant du rang de naissance atteint vers un **comportement dépendant du nombre d'enfants déjà nés**
  - ▷ de la fécondité naturelle à la fécondité dirigée [Henry 1961a, Henry 1961b]
  - ▷ from non-parity related to parity related fertility control [Coale 1986]
  - ▷ from birth spacing to birth stopping [Guinnane et al. 1994, Knodel 1987]

la transition des comportements reproductifs



détecter des comportements d'arrêt [van Bavel 2004b]

- **modèle de Coale et Trussell** : calendrier par âge [Coale et Trussell 1974, Coale et Trussell 1978]
- **modèle de Page** : calendrier par âge et durée de mariage [Page 1977, Rodriguez et Cleland 1988]
- **modèle de McDonald** : démarrage, espacement, arrêt [McDonald 1984]
- **l'analyse des parités des cohortes (CPA)** : répartition par parité [David et al. 1988]
- **l'analyse démographique des biographies** : analyse multivariée des intervalles intergénésiques [Alter 1988, Gutmann et Alter 1993]

critique des méthodes classiques

- confusion des effets de comportements d'espacement et des effets de comportements d'arrêt [Okun 1994]
- limitation aux histoires génésiques complètes
- confusion des effets d'intensité et de calendrier
- difficulté d'identifier les sous-groupes pratiquant la contraception d'arrêt dans une population transitionelle [Schumacher 2010]

$\rightarrow$ proposition d'une alternative : **analyse des taux de fécondité par âge et par parité atteinte à l'aide d'un modèle de Poisson avec sur-représentation des zéros**

exemple d'application : Genève au 19e siècle

- **contexte démographique particulier** :
  - ▷ déclin précoce de la fécondité dès la fin du 17e siècle au sein de l'élite [Henry 1956]
  - ▷ diffusion du contrôle de la fécondité des classes supérieures aux classes ouvrières dès la 2ème moitié du 18e siècle [Perrenoud 1990]
  - ▷ recomposition importante de la population au 19e siècle en raison de flux migratoires soutenus [Schumacher 2010]
- **échantillon** alphabétique de plus de **2000 histoires génésiques** (complètes ou tronquées à gauche)
  - ▷ couples mariés à Genève entre 1800 et 1880
  - ▷ distinction de trois classes sociales et trois groupes d'origine

résultats : parité 2

facteur	poisson	logit
classe sociale		
classe supérieure	⊖	⊖
classe moyenne	ref	ref
classe ouvrière		⊖
origine		
couples natifs		
couples mixtes	ref	ref
couples immigrés		

naissances : 2050 , femmes-années : 17'688
coefficients ajustés pour l'âge (classes d'âge quinquennales)
erreurs standards robustes pour données hiérarchiques

résultats : parité 3

facteur	poisson	logit
classe sociale		
classe supérieure	⊖⊖	
classe moyenne	ref	ref
classe ouvrière	⊖	⊖
origine		
couples natifs		
couples mixtes	ref	ref
couples immigrés		⊖

naissances : 1188 , femmes-années : 10'600
coefficients ajustés pour l'âge (classes d'âge quinquennales)
erreurs standards robustes pour données hiérarchiques

résultats : parité 4

facteur	poisson	logit
classe sociale		
classe supérieure	$\ominus\ominus$	
classe moyenne	ref	ref
classe ouvrière	$\ominus$	$\ominus$
origine		
couples natifs		
couples mixtes	ref	ref
couples immigrés	$\oplus$	$\ominus$

naissances : 683 , femmes-années : 5896
coefficients ajustés pour l'âge (classes d'âge quinquennales)
erreurs standards robustes pour données hiérarchiques

conclusion et discussion

- **résultats**
  - ▷ retard dans l'adoption du nouveau comportement reproductif parmi les ouvriers et les immigrés
  - ▷ persistance de l'ancien comportement d'espacement chez les ouvriers, pas (moins) d'espacement chez les immigrés
- **apports du ZIP**
  - ▷ adapté aux populations hétérogènes et aux histoires génésiques tronquées (populations mobiles)
  - ▷ moindre risque de confondre stopping et spacing
- **faiblesses du ZIP**
  - ▷ ne mesure pas l'intensité du birth stopping
  - ▷ multiplie les coefficients à estimer
- **à faire**
  - ▷ tester la validité du modèle avec données DHS

conclusion et discussion

merci !

références

Alter, George (1988). *Family and female life course. The women of Verviers, Belgium 1849-1880*. Madison : The University of Wisconsin Press.

Bongaarts, John et Robert C. Potter (1983). *Fertility, biology and behaviour*. New York : Academic Press.

Coale, Ansley J. (1986). The decline of fertility in Europe since the 18th century as a chapter in demographic history. In Coale, Ansley J. et Susan C. Watkins (eds.). *The decline of fertility in Europe*. Princeton : Princeton University Press, 1-30.

Coale, Ansley J. et James T. Trussell (1974). Model fertility schedules : variations in the age structure of childbearing in human populations. *Population Index* 40(2) : 185-258.

Coale, Ansley J. et James T. Trussell (1978). Technical note : finding the two parameters that specify a model schedule of marital fertility. *Population Index* 44(2) : 203-213.

David, Paul A., Mroz, Thomas A., Sanderson, Warren C., Wachter, Kenneth W. et David R. Weir (1988). Cohort parity analysis. Statistical estimates of the extent of fertility control. *Demography* 25() : 163-188.

Guinnane, Timothy W., Okun, Barbara S. et James Trussell (1994). What do we know about the timing of fertility transitions in Europe ? *Demography* 31(1) : 1-20.

Gutmann, Myron P. et George Alter (1993). Family reconstitution as event-history analysis. In Reher, David et Roger Schofield (eds.). Old and new methods in historical demography. Oxford : Clarendon Press, 159-177.

Henry, Louis (1956). *Anciennes familles genevoises. Etude démographique, 16-20e siècle*. Paris : PUF.

références

Henry, Louis (1961a). La fécondité naturelle. Observation - théorie - résultats. *Population* 4(4) : 625-636.

Henry, Louis (1961b). Some data on natural fertility. *Eugenics Quarterly* 8(2) : 81-91.

Knodel, John (1987). Starting, stopping and spacing during the early stages of fertility transition : the experience of German village populations in the 18th and 19th centuries. *Demography* 24(2) :143-162

Leridon, Henry (1989). Fécondité naturelle et espacement des naissances. *Annales de démographie historique* 1988 : 21-33.

Long, Scott J. et Jeremy Freese (2006). *Regression models for categorical dependent variables using Stata*. 2nd edition. College Station, TX : Stata Press.

McDonald, P. (1984). *Nuptiality and completed fertility : a study of starting, stopping and spacing behavior*. Voorburg,NL : International Statistical Institute.

Okun, Barbara S. (1994). Evaluating methods for detecting fertility control : Coale and Trussell's model and cohort parity analysis. *Population Studies* 48(2) : 193-222.

Perrenoud, Alfred (1990). Aspects of fertility decline in an urban setting : Rouen and Geneva. In van de Woude, Ad, de Vries, Jan et Akira Hayami (eds). *Urbanization in history. A process of dynamic interactions*. Oxford : Oxford University Press, 243-263.

Page, Hillary J. (1977). Patterns underlying fertility schedules : a decomposition by both age and marriage duration. *Population Studies* 31(1) : 85-106.

références

Rodriguez, German et John Cleland (1988). Modelling marital fertility by age and duration : an empirical appraisal of the Page model. Population Studies 42(2) : 241-257.

Santow, Gigi (1995). Coitus interruptus and the control of natural fertility. *Population Studies* 49(1) : 19-43.

Schoumaker, Bruno (2004). Une approche personnes-périodes pour l'analyse des histoires génésiques. *Population* 59(5) : 783-796.

Schumacher, Reto (2010). *Structures et comportements en transition. La reproduction démographique à Genève au 19e siècle*. Collection Population, famille et société vol. 12. Berne : Peter Lang.

van Bavel, Jan (2004a). Deliberate birth spacing before the fertility transition in Europe : evidence from nineteenth-century Belgium. *Population Studies* 58(1) :95-107

van Bavel, Jan (2004b). Distinguer contraception d'arrêt et contraception d'espacement. Revue des méthodes en démographie historique. *Population* 59(1) :119-132.