Mixture modelling of recurrent event times with long-term survivors: Analysis of Hutterite birth intervals

John W. ‘Mac’ McDonald &
Alessandro Rosina

Quantitative Methods in the Social Sciences
Seminar - Analysis of Longitudinal Data

Hutterite data

- natural high fertility population
- no use of contraception
- first birth interval
  - 432 closed birth intervals
  - 18 open birth intervals
- 2+ birth intervals
  - 2,544 closed birth intervals
  - 301 open birth intervals

Analysis of Hutterite birth intervals

- use complete birth histories
  - simultaneously analyse all intervals
  - including open intervals
- simultaneously model
  - waiting time to conception
  - primary sterility
  - secondary sterility

New estimation approach

- Bayesian approach
- Gibbs sampling
  - flexible
  - but computationally demanding

Why mixture models for survival data?

- **problem of long-term survivors**
  - time to event of interest equals infinity
  - e.g. sterile never close birth interval
  - bias if problem ignored
- **use mixture model**
  - proper inference
  - few demographic applications to date

Mixture model for two populations

- $1:$ those eventually having the event
- $0:$ those at zero risk of having the event

$$Y = \begin{cases} 1 & \text{if eventually having the event} \\ 0 & \text{if zero risk, i.e., long-term survivor} \end{cases}$$

Mixture model for two populations . . .

- **Y is partially observable**
  - known for uncensored observation
  - missing for right-censored observation
- **logistic model for distribution of Y**

$$\text{logit}(\text{pr}(Y = 1)) = X'\beta$$

Generic survival model

- individual with vector of covariates X
- **uncensored observation at time t:** likelihood term is the probability of experiencing the event at time t,

$$\text{pr}(Y = 1 \mid X)\, f(t \mid Y = 1, X) \tag{1}$$

Generic survival model . . .

- **censored observation at time t:**

  likelihood term is the probability of being a long-term survivor plus the probability of experiencing the event after time t

$$\text{pr}(Y = 0 \,|\, X) + \text{pr}(Y = 1 \,|\, X)\, S(t \,|\, Y = 1, X) \quad (2)$$

- likelihood is product of (1) & (2) terms

Alternative models for a single event

- **Weibull**
  - Farewell (1982)
- **accelerated failure time**
  - Yamaguchi (1992)
  - Yamaguchi and Ferguson (1995)
- **piecewise constant hazard**
  - Li and Choe (1997)
  - Wang and Murphy (1998)

Logistic-normal-geometric model

- **discrete-time survival model**
- model for a single event or recurrent events
- including unobserved heterogeneity
- can be extended to account for long-term survivors
- McDonald and Rosina (2001)

Mixture modelling of recurrent events

- **simultaneous estimation of logistic model for long-term survivors**
  - estimates effects of covariates on the probability of the event occurring
- **simultaneous estimation of survival model**
  - estimates effects of covariates on the timing of the event, given that it will occur

Algorithms for model fitting

- **Expectation-Maximization algorithm**
  - commonly used
  - simple, but slow
  - standard errors not obtained
- **Markov chain Monte Carlo methods, e.g. Gibbs sampling**

Discrete-time event history model

- **individual event history is a series of Bernoulli or success/failure trials** at each discrete time point, e.g., month
- let T **= waiting time to conception (success) in months**, $t = 1, 2, 3, \ldots$
- **discrete hazard** is $h_t \equiv \text{pr}(T = t \mid T \geq t)$

Discrete-time event history model . . .

- **probability of conception at time t** is

$$\text{pr}(T = t) = (1 - h_1)\,(1 - h_2) \cdots (1 - h_{t-1})\,h_t$$

- **geometric distribution has constant hazard**, i.e., $h_t = p$, and

$$\text{pr}(T = t) = q^{t-1}\,p \quad \text{where} \quad q \equiv 1 - p$$

Geometric distributed waiting times

- **uncensored observation at time t:**
  - likelihood term is

$$\text{pr}(T = t) = q^{t-1}\, p$$

  - or 1 success out of t trials
- **censored observation at time t:**
  - likelihood term is

$$\text{pr}(T \geq t) = q^{t-1}$$

Geometric distributed waiting times . . .

- **trick: software for fitting logistic models can be used to fit geometric distributed waiting times and allow for covariates!**

$$\text{logit}(\text{hazard}) = X'\beta$$

- **observed heterogeneity in risk is modelled by a geometric regression model**

Logistic-normal-geometric model

- **unobserved heterogeneity in risk is modelled by a mixed-geometric regression model for waiting times**

$$\begin{aligned} \text{logit(hazard)} &= X'\beta + Z\,\sigma \\ Z &\sim N(0,1) \end{aligned}$$

  - Z unobserved covariate value
  - σ standard deviation parameter
  - $Z\,\sigma$ random effect

Logistic-normal-geometric model . . .

- **trick: software for fitting random effects logistic models can be used to fit mixed-geometric distributed waiting times and allow for observed and unobserved heterogeneity!**

Constant hazard: recurrent events

- **T_k = time between $k-1$ & k th conception**
- $T = T_1 + \cdots + T_k$
- probability of kth conception at time t is

$$\text{pr}(T=t) = q^{t_1-1}p \times q^{t_2-1}p \times \cdots \times q^{t_k-1}p$$

- **uncensored observation at time t:**
  - likelihood term is

$$\text{pr}(T=t) = q^{t-k}p^k$$

1st birth interval/primary sterility

- **1st birth interval**
  - covariates F and effects γ
  - logit(hazard | fecund) $= F'\gamma + Z\sigma$
- **primary sterility**
  - covariates P and effects α
  - logit(primary sterility) $= P'\alpha$

2+ birth intervals/secondary sterility

- **2+ birth intervals**
  - covariates H and effects δ
  - logit(hazard | fecund) $= H'\,\delta \,+\, Z\,\sigma$
- **secondary sterility**
  - covariates S and effects β
  - logit(secondary sterility) $= S'\,\beta$

All birth intervals

- **1st birth interval**
  - logit(hazard | fecund) $= F' \gamma + Z \sigma$
- **2+ birth intervals**
  - logit(hazard | fecund) $= H' \delta + Z \sigma$
- **common unobserved** $Z \sim N(0, 1)$
- **common effect** σ
- **common normal random effect**

Logistic-normal-geometric model . . .

- logit(hazard) $= F'\gamma + Z\sigma$
- have **joint likelihood** $l(\gamma, \sigma, Z)$
  - but each Z is unknown
- obtain **marginal likelihood** $l(\gamma, \sigma)$
  - by integrating out Z over its prior $N(0, 1)$
  - exact integration impossible as logistic is a nonlinear model

Estimation methods

- **quadrature**
  - approximate using numerical integration
- **quasi-likelihood**
  - Taylor series approximation
  - nonlinear model becomes linear
- **Bayesian inference via Gibbs sampling**
  - samples [γ, σ, Z] from **joint** likelihood for inference on each parameter

Bayesian Inference

- **parameters random with distributions**
  - **prior** to data
  - **posterior** given data
- apply Bayes Theorem
- **posterior** $\propto$ **likelihood** $\times$ **prior**
- 'uninformative' locally uniform priors $\Longrightarrow$ posterior $\simeq$ standardized likelihood

Estimation by Gibbs sampling

- **sample from joint posterior by sampling univariately from full conditionals**
  - sample $\gamma^1 \sim [\gamma \mid \sigma^0, Z^0]$
  - sample $\sigma^1 \sim [\sigma \mid \gamma^1, Z^0]$
  - sample $Z^1 \sim [Z \mid \gamma^1, \sigma^1]$
- **generate a joint sample (**γ^1**,** σ^1**,** Z^1**)**

  and so on 2, 3, $\cdots$
- discard early ‘burn-in’ phase

Estimation by Gibbs sampling . . .

- **fitting mixed-geometric survival model with long-term survivors is straightforward extension**
- WinBUGS for fitting Bayesian model
- can use the univariate sample average for inference on each parameter

Hutterite data

- 724 unions
- exclusions for incomplete information
- **dataset for analysis**
  - **450 families**
  - **2,976 births**
  - **3,295 birth intervals**

Hutterite data . . .

- first birth interval
  - 432 closed birth intervals
  - 18 open birth intervals
- 2+ birth intervals
  - 2,544 closed birth intervals
  - 301 open birth intervals

Effective conditional fecundability

- effective $\equiv$ conception leads to live birth
- conditional on being fecund (not sterile)
- bias if open intervals excluded
  - some women may not be sterile throughout open interval

Model waiting time to conception

- **waiting time to conception**

  $\equiv$ **birth interval – 8 months**

  $= 1,\ 2,\ 3,\ \cdots$

- **discrete-time survival model**

Effects on primary sterility

		mean	2.5%	97.5%
	constant	-2.92	-4.18	-1.82
cohort	1910–19	-1.30	-3.29	0.33
	$>$1919	-0.97	-2.18	0.27
age at	< 21	-0.17	-1.50	1.17
marriage	> 23	0.70	-0.77	2.11

- cohort & age at marriage not significant

Effects on secondary sterility

		mean	2.5%	97.5%
	constant	-6.25	-8.26	-6.14
cohort	1910–19	0.10	-0.44	0.64
	>1919	1.06	0.11	1.96
	<21	-79.63	-221.80	-2.29
	24–26	0.56	-1.57	2.87
age at	27–29	0.64	-2.00	3.18
start of	30–32	0.84	-1.99	3.53
birth	33–35	1.35	-1.69	4.05
interval	36–38	1.70	-1.36	4.51
	39–41	3.35	0.27	6.29
	>41	5.99	2.95	8.88
previous child died		0.46	-0.61	1.41

- previous child died not significant
- but one cohort & late age at start significant

Effects on secondary sterility . . .

		mean	2.5%	97.5%
	3–5	-0.32	-2.22	1.53
	6–8	0.08	-1.91	2.32
duration	9–11	0.82	-1.17	3.26
of	12–14	1.15	-0.85	3.81
marriage	15–17	0.94	-1.17	3.67
	18–20	1.45	-0.63	4.15
	$>$20	1.76	-0.35	4.50

- duration of marriage not significant, but positive trend

Effects on 2+ birth spacing

		mean	2.5%	97.5%
	3–5	-0.21	-0.35	-0.06
	6–8	-0.34	-0.52	-0.15
duration	9–11	-0.33	-0.55	-0.12
of	12–14	-0.34	-0.61	-0.08
marriage	15–17	-0.42	-0.72	-0.11
	18–20	-0.39	-0.76	-0.02
	$>$20	-0.59	-1.13	-0.04
	sigma	0.104	0.005	0.256

- duration of marriage significant with negative trend

Effects on 2+ birth spacing

		mean	2.5%	97.5%
	constant	-2.37	-2.51	-2.23
cohort	1910–19	0.18	0.08	0.29
	>1919	0.18	0.07	0.29
	<21	-0.11	-0.38	0.15
	24–26	0.02	-0.13	0.17
age at	27–29	-0.02	-0.21	0.16
start of	30–32	-0.01	-0.23	0.20
birth	33–35	-0.07	-0.33	0.20
interval	36–38	-0.08	-0.37	0.21
	39–41	-0.24	-0.61	0.11
	>41	0.05	-0.58	0.66
previous child died		0.40	0.18	0.61

- previous child died significant, cohort significant & age at start not significant

Effects on first birth interval spacing

		mean	2.5%	97.5%
	constant	-1.21	-1.50	-0.92
cohort	1910–19	-0.10	-0.43	0.24
	$>$1919	0.02	-0.26	0.31
age at	$<$21	-0.16	-0.40	0.08
marriage	$>$23	-0.37	-0.70	-0.05

- cohort not significant
- last age at start significant

Results

- **effects on primary sterility**
  - cohort not significant
  - age at marrige not significant

Results . . .

- **effects on secondary sterility**
  - **age at start of interval**
    - **positive trend**
    - **39-41, 41+ significant**
  - duration of marrige not significant, but positive trend
  - **cohort: 1919+ significant**
  - previous child died not significant

Results . . .

- **effects on first birth spacing**
  - **last age at start of interval significant**
  - cohort not significant

Results . . .

- **effects on 2+ birth spacing**
  - **duration of marrige significant, negative trend**
  - age at start of interval not significant, no trend
  - **cohort: 1910-19, 1919+ significant**
  - **previous child died significant, positive**

Results . . .

- **unobserved heterogeneity σ**
- **mean .1042**
- **median .0955**
- **s.d. .0692**
- **95% CI [.0048, .2560]**

New birth interval/sterility model

- uses complete birth histories
  - simultaneously analyse all intervals
- simultaneously models determinants of
  - waiting time to conception
  - primary sterility
  - secondary sterility
- allows for zero risk subpopulations
- estimates fecundability for those fecund

Posterior probability of sterility

- posterior probability that an individual with vector of explanatory variables x comes from population $Y = 0$, given that no event has occurred by time t

$$pr(Y = 0 \mid X, T > t) = \frac{pr(Y = 0 \mid X)}{pr(Y = 0 \mid X) + pr(Y = 1 \mid X)\, S(t \mid Y = 1, X)}$$