Markov Chain Monte Carlo Exact Inference for Binomial & Multinomial Logistic Regression Models

Jon Forster, Mac McDonald & Peter Smith

University of Southampton, UK

bigmac@soton.ac.uk

http://www.maths.soton.ac.uk/staff/JJForster/paper.html

THE PROBLEM

Exact conditional inference

- tests, particularly goodness-of-fit (GOF) tests
- residual analysis
- confidence intervals

for binomial & multinomial logistic regression models

BINOMIAL LOGISTIC REGRESSION

$$Y_i \sim \text{binomial}(m_i, \pi_i) \qquad\qquad i = 1, \ldots, n$$

Compare

$$M_0 : \text{logit}(\boldsymbol{\pi}) = X\boldsymbol{\beta}$$

with

$$M_1 : \text{logit}(\boldsymbol{\pi}) = X\boldsymbol{\beta} + Z\boldsymbol{\gamma}$$

M_1 is saturated (GOF test) if rank$(X, Z) = n$

EXACT CONDITIONAL INFERENCE FOR γ

Based on the distribution of (or some 1-dim function of)

$$Z^T\boldsymbol{y}|X^T\boldsymbol{y} = X^T\boldsymbol{y}_{obs}$$

a margin of

$$\boldsymbol{y}|X^T\boldsymbol{y} = X^T\boldsymbol{y}_{obs}$$

$$f(\text{successes} \mid \text{sufficient statistics for } \boldsymbol{\beta})$$

Conditional distribution of the vector of responses $\boldsymbol{y}$, given $\boldsymbol{X}^T\boldsymbol{y}$, the vector of sufficient statistics for $\boldsymbol{\beta}$, is

$$f(\boldsymbol{y}|\boldsymbol{X}^T\boldsymbol{y}=\boldsymbol{X}^T\boldsymbol{y}_{obs};\boldsymbol{\gamma}) \propto \exp(\boldsymbol{\gamma}^T\boldsymbol{Z}^T\boldsymbol{y}) \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix}$$

Inference by

- enumeration
- (Markov Chain) Monte Carlo sampling

followed by marginalization

CONDITIONAL DISTRIBUTION FOR INFERENCE

- uniform when $m_i = 1$ for all i
- GOF test for pure binary data is not sensible
- degenerate for continuous covariates, since only $\boldsymbol{y}_{obs}$ satisfies the conditioning constraints
- not usually degenerate when covariate values are integer or evenly spaced

METROPOLIS-HASTINGS SAMPLING

1. Given current value $\boldsymbol{y}$, generate a new value $\boldsymbol{y}'$ from some probability distribution $q(\boldsymbol{y}, \boldsymbol{y}')$

2. Accept $\boldsymbol{y}'$ as the next realization of the chain with probability $a(\boldsymbol{y}, \boldsymbol{y}')$, where

$$a(\boldsymbol{y}, \boldsymbol{y}') \quad = \quad \min\left\{\frac{f(\boldsymbol{y}')q(\boldsymbol{y}', \boldsymbol{y})}{f(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{y}')}, 1\right\}$$

   otherwise, retain $\boldsymbol{y}$

Provided q is chosen appropriately, then f is the stationary distribution distribution for this chain

SIMPLE LINEAR LOGISTIC REGRESSION

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i \qquad\qquad i = 1, \ldots, 6$$

Covariate with integer values: $x_i = i$

Sufficient statistics $\quad \boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{y}_{obs} = (s_0, s_1)^T$

- $s_0 = \Sigma y_i$ for β_0
- $s_1 = \Sigma x_i y_i$ for β_1

Let $\boldsymbol{y}' = \boldsymbol{y} + \boldsymbol{v}$ such that $\Sigma y_i' = s_0$ & $\Sigma x_i y_i' = s_1$

Let $\boldsymbol{y}' = \boldsymbol{y} + \boldsymbol{v}$ such that $\Sigma y_i' = s_0$ & $\Sigma x_i y_i' = s_1$

x_i	m_i	y_i	v_i	y_i'
1	20	1	-1	0
2	20	4	2	6
3	20	9	-1	8
4	20	13	0	13
5	20	18	0	18
6	20	20	0	20

thus maintaining the sufficient statistics

PROPOSED METROPOLIS-HASTINGS ALGORITHM

- enumerate all integer $\boldsymbol{v}$ that satisfy $\boldsymbol{X}^T\boldsymbol{v} = \boldsymbol{0}$ and $\Sigma|v_i| \leq r$ for some even r, typically 4, 6 or 8

- generate a $\boldsymbol{v}$ uniformly & generate integer d from

$$g_D(d|\boldsymbol{v}) \propto \exp(\boldsymbol{\gamma}^T\boldsymbol{Z}^T\{\boldsymbol{y} + d\boldsymbol{v}\}) \prod_{i=1}^{n} \binom{m_i}{y_i + dv_i} \quad (1)$$

- set $\boldsymbol{y}' = \boldsymbol{y} + d\boldsymbol{v}$, where $0 \leq y_i' \leq m_i$ for all i

- most v_i are zero so (1) a product of at most $r + 1$ terms, so the support of d is typically small

DOSE-RESPONSE DATA

log-dose	m_i	y_i	x_i	
0.301	19	19	1	
0.000	20	18	0	
-0.301	19	19	-1	
-0.602	21	14	-2	
-0.903	19	15	-3	
-1.208	20	4	-4	-4ϵ
-1.509	16	0	-5	-4ϵ
-1.807	19	0	-6	$-\epsilon$
-2.108	40	0	-7	$-\epsilon$
-2.710	81	2	-9	$-\epsilon$

DOSE-RESPONSE DATA

- taken from Bedrick and Hill (1990)
- tumorigenicity of benzepyrene in mice
- doses *almost* equally spaced on log-dose scale
- exact results available for comparison

Bedrick, EJ & Hill, JR (1990) Outlier tests for logistic regression: A conditional approach *Biometrika* **77** 815–827

Test statistics & p-values for dose-response data

	observed value	df	asymptotic p-value	estimated exact p-values	exact p-value
L^2	26.679	8	0.001	0.0064 ± 0.0008	0.006
X^2	32.096	8	0.000	0.0116 ± 0.0009	0.013

- estimated p-values based on sample of one million
- approximate 99% CI used method of batch means
- exact p-values from Bedrick and Hill
- MCMC estimates in good agreement with the exact p-values

IRREDUCIBILITY

How to choose $\boldsymbol{v}$ so that any $\boldsymbol{y}$ satisfying conditioning constraints can be reached by the chain?

- results for special cases, e.g. equally-spaced covariate
- Gröbner basis approach (Diaconis & Sturmfels, 1998)
  - ◇ sufficient set of moves
  - ◇ computationally demanding

How important is irreducibility in practice?

GREYING OF HAIR AND MORTALITY

- 469 adult Mexicans scored on hair greyness in 1948:

  1 - none, 2 - slight, 3 - moderate, 4 - general

- age groups: 17–24, 25–29, . . . , 70–74, 75+

- 65 distinct covariate patterns

- response: natural death between 1948 and 1969

Lasker, GW & Kaplan, B (1974) Graying of the hair and mortality *Social Biology* **21** 290–295 males

greyness	none		slight		moderate		general	
age	y_i	m_i	y_i	m_i	y_i	m_i	y_i	m_i
1	1	46	0	1				
2	1	29						
3	3	23	0	3				
4	4	33	3	7				
5	2	12	3	12	0	2		
6	1	12	5	15	3	7	0	2
7	1	1	3	16	0	1	5	8
8	1	2	5	6	1	4	3	9
9	0	3	1	4	3	6	3	6
10					2	3	3	5
11			1	1	2	2	3	4
12					2	2	3	3

females

greyness	none		slight		moderate		general	
age	y_i	m_i	y_i	m_i	y_i	m_i	y_i	m_i
1	2	34						
2	0	21	0	1				
3	1	13						
4	0	23	0	5	0	1		
5	0	11	0	2	1	1	1	1
6	0	8	4	7			0	3
7	0	3	1	7	0	2	1	4
8	1	2	0	6	1	4	3	7
9	1	1	0	2			1	1
10			0	1	0	2	0	1
11			1	1			2	2
12			1	1	1	1		

GREYING OF HAIR AND MORTALITY . . .

- age: equally-spaced covariate (1 to 12)
- greyness score: equally-spaced covariate (1 to 4)
- estimated exact p-values suggest better fit than do asymptotic p-values
- reject the SEX+AGE+GREY model at the 5% level using the asymptotic p-value for L^2, but not using the estimated exact p-value

GREYING OF HAIR AND MORTALITY . . .

Test statistics and p-values for grey hair data

test	observed statistic	df	asymptotic p-value	estimated exact p-value
Goodness of fit of	$L^2 = 87.80$	62	0.0172	0.0487 ± 0.0059
SEX+AGE	$X^2 = 85.81$	62	0.0244	0.0518 ± 0.0054
Goodness of fit of	$L^2 = 84.01$	61	0.0270	0.0959 ± 0.0091
SEX+AGE+GREY	$X^2 = 77.05$	61	0.0806	0.0973 ± 0.0089

TEST AGAINST NON-SATURATED ALTERNATIVE

To compare the two models

1. extract from the Markov chain for the SEX+AGE model a sample of the sufficient statistic for the greyness score parameter

2. estimate exact p-value by ranking observed value of sufficient statistic, $\boldsymbol{Z}^T \boldsymbol{y}_{obs} = 235$, among this sample

Against the one-sided alternative that hair greyness is deleterious, $\widehat{p} = 0.0314 \pm 0.0068$

RESIDUAL ANALYSIS FOR AGE + SEX MODEL

Standardized deviance residuals and p-values for grey hair data

y_i	m_i	covariate values	residual	asymptotic p-value	estimated exact p-value	support points
0	3	M 9 0	-2.229	0.0258	0.0878	4
3	7	M 4 1	2.008	0.0446	0.0484	6
4	7	F 6 1	2.703	0.0068	0.0075	7
1	1	F 5 2	2.161	0.0306	0.0978	2
1	1	F 5 3	2.161	0.0306	0.1000	2

RESIDUAL ANALYSIS FOR AGE + SEX MODEL

- $\widehat{p}$ calculated using the empirical distribution of the residuals extracted from the MCMC sample used to test goodness of fit
- no $\widehat{p}$ for the 65 residuals indicates that the lack of fit is due to a small number of extreme cases
- asymptotic p-values closest to $\widehat{p}$ for the empirical distributions with largest numbers of support points

MONTE CARLO EXACT CONFIDENCE INTERVAL

Monte Carlo exact inference is based on a sample generated from

$$f(\boldsymbol{y}|\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y}_{obs}; \boldsymbol{\gamma}) \propto \exp(\boldsymbol{\gamma}^T \boldsymbol{Z}^T \boldsymbol{y}) \prod_{i=1}^{n} \begin{pmatrix} m_i \\ \\ y_i \end{pmatrix}$$

For scalar γ, an exact p-value for H_γ is estimated using a tail area of the empirical distribution of $\boldsymbol{Z}^T\boldsymbol{y}$

MONTE CARLO EXACT CI . . .

- lower (upper) end point of an exact $(1-2\alpha)$ CI for γ can be estimated by finding the value γ^0 such that the observed value of $\boldsymbol{Z}^T\boldsymbol{y}$ is the upper (lower) α-quantile of the empirical distribution
- given a sample for $\gamma = \gamma^*$, the exact p-value under H_γ: $\gamma = \gamma^0$ can be estimated by weighting the sample by $\exp\{(\gamma^0 - \gamma^*)\boldsymbol{Z}^T\boldsymbol{y}\}$

MONTE CARLO EXACT CI . . .

- in principle, a grid search for both end points of a CI may be based on a single Monte Carlo sample
- a natural choice is $\gamma^* = 0$, if a Monte Carlo test of $\gamma = 0$ has already been performed
- alternatively, $\gamma^* = \widehat{\gamma}$, the MLE, is a value which is supported by the observed data

EXACT CI FOR GREYNESS SCORE PARAMETER

- estimated exact 95% CI is
  - $\diamond$ $(-0.015, 0.613)$ using $\gamma^* = 0$
  - $\diamond$ $(-0.010, 0.600)$ using $\gamma^* = \widehat{\gamma} = 0.295$
- show relatively little sensitivity to the choice of γ^*
- are similar to the asymptotic CI $(-0.001, 0.592)$

MULTINOMIAL LOGISTIC REGRESSION

- polytomous response with categories $0, \ldots, K$
- ith observation represented by the $K+1$ counts $(y_{i0}, y_{i1}, \ldots, y_{iK}),\ i = 1, \ldots, n$,
  with the total count $m_i = \sum_{k=0}^{K} y_{ik}$, assumed fixed
- $\boldsymbol{Y} = (y_{ik})$ is $n \times K$ matrix of responses, where k runs from 1
- denote the K columns of $\boldsymbol{Y}$ by $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ with $\boldsymbol{y}_0 = \boldsymbol{m} - \sum_{k=1}^{K} \boldsymbol{y}_k$ where $\boldsymbol{m} = (m_1, \ldots, m_n)^T$

MULTINOMIAL LOGISTIC REGRESSION

baseline-category multinomial logistic regression model, with baseline category 0, is

$$\log\left(\frac{\pi_{ik}}{\pi_{i0}}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}_k + \boldsymbol{z}_i^T \boldsymbol{\gamma}_k \tag{2}$$

When the response is ordinal, it may be more appropriate to express this model as the equivalent adjacent-category model

$$\log\left(\frac{\pi_{ik}}{\pi_{i\,k-1}}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}'_k + \boldsymbol{z}_i^T \boldsymbol{\gamma}'_k \tag{3}$$

Hirji, KF (1992) Computing exact distributions for polytomous response data. *JASA*, **87** 487-492

considered the baseline-category multinomial logistic regression model

$$\log\left(\frac{\pi_{ik}}{\pi_{i0}}\right) = \theta_k + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\gamma} \qquad (4)$$

and the adjacent-category model

$$\log\left(\frac{\pi_{ik}}{\pi_{i\,k-1}}\right) = \theta_k + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\gamma} \qquad (5)$$

which are more parsimonious than (2) and (3) as the regression parameters do not depend on the category

MULTINOMIAL LOGISTIC REGRESSION

- binomial is a special case of the multinomial where $K = 1$, $\boldsymbol{y}_1 = \boldsymbol{y}$ and $\boldsymbol{y}_0 = \boldsymbol{m} - \boldsymbol{y}$
- each proposed step of our binomial algorithm may be thought of as addition of $d\boldsymbol{v}$ to $\boldsymbol{y}_1$ together with subtraction of $d\boldsymbol{v}$ from $\boldsymbol{y}_0$

MULTINOMIAL LOGISTIC REGRESSION

- $K + 1$ vectors of outcomes, $\boldsymbol{y}_0, \boldsymbol{y}_0, \ldots, \boldsymbol{y}_K$
- a proposal is obtained by selecting at random an integer $\boldsymbol{v}$ such that $\boldsymbol{X}^T \boldsymbol{v} = \mathbf{0}$ and an integer vector

$$\boldsymbol{w} = (w_0, w_1, \ldots, w_K)^T$$

  of length $K + 1$ such that $\mathbf{1}_{K+1}^T \boldsymbol{w} = 0$, where $\mathbf{1}$ is a vector of ones of the given dimension
- $\boldsymbol{Y}' = \boldsymbol{Y} + d\boldsymbol{v}\boldsymbol{w}_{\setminus 0}^T$

MULTINOMIAL LOGISTIC REGRESSION

- for computational convenience, the set of possible $\boldsymbol{w}$ is restricted to those for which $\sum_{k=0}^{K} |w_k| = 2$
- the procedure is then equivalent to selecting at random $k_1, k_2 \in \{0, 1, \ldots, K\}$, $k_1 \neq k_2$
  - ◇ adding $d\boldsymbol{v}$ to $\boldsymbol{y}_{k_1}$
  - ◇ subtracting $d\boldsymbol{v}$ from $\boldsymbol{y}_{k_2}$
- a simple extension of the binomial algorithm

GOF test statistics and p-values for pregnancy outcome

	Observed statistic	df	Asymptotic p-value	Estimated exact p-value*
Model (4)	$L^2 = 40.00$	41	0.5150	0.8200 ± 0.0037
	$X^2 = 39.83$	41	0.5226	0.7478 ± 0.0063
Model (5)	$L^2 = 42.27$	41	0.4159	0.5293 ± 0.0170
	$X^2 = 43.11$	41	0.3811	0.3849 ± 0.0201
Model (2)	$L^2 = 32.06$	32	0.4638	0.5813 ± 0.0114
	$X^2 = 32.18$	32	0.4576	0.4633 ± 0.0128

* with approximate 99% confidence interval

DISCUSSION

- proposed MH algorithm
  - ◇ intuitive and easy to construct
  - ◇ extremely efficient for exact inference
  - ◇ however, the resulting Markov chain is not necessarily irreducible!!
- MCMC estimated p-values have been in good agreement with the enumerated p-values

DISCUSSION . . .

- Markov chains seem to mix well
- an indication of good connectivity is stable $\widehat{p}$ as the number of possible moves, determined by r, is increased
- how important is irreducibility in practice?
- if chain *not* irreducible, conditioning is *also* on being in a particular reduced component of the sample space