Efficient MCMC Estimation of Binomial Logit Models

Agnes Fussl

In collaboration with
Sylvia Frühwirth-Schnatter (WU) and **Rudolf Frühwirth** (ÖAW)

Österreichische Statistiktage
September 8^{th}, 2011

Outline

Binary data

- sequence $y_1, \ldots, y_N$ of binary data
- $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept
- $\boldsymbol{\beta}$ is an unknown regression parameter of dimension d

Binary logit regression model

$$\Pr(y_i = 1|\boldsymbol{\beta}) = \pi_i(\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})}$$

RUM and dRUM representation

- to perform bayesian inference with data augmentation the logit model can be rewritten as ***random utility model*** (RUM) introduced by McFadden (1974) or as ***difference RUM*** (dRUM)
- ***RUM:*** y_{ki}^u is the utility of choosing category $k = 0, 1$, which is assumed to depend on covariates $\mathbf{x}_i$; category 1 is observed, if $y_{1i}^u > y_{0i}^u$
- ***dRUM***: choose category 0 as baseline and assume the differences of the utilities $y_{1i}^u - y_{0i}^u$ to depend on some covariates $\mathbf{x}_i$; category 1 is observed, if $y_{1i}^u - y_{0i}^u > 0$
- for both representations the logit regression model results as marginal distribution of y_i

MCMC for Binary Logit Model

- Frühwirth-Schnatter and Frühwirth (2010) show for the binary logit model that MCMC estimation based on the dRUM representation is much more efficient than MCMC estimation based on the RUM representation
- **How can we improve MCMC sampling for the binomial logit model?**

Binomial data

- $\mathbf{y} = (y_1, \ldots, y_N)$ are conditionally independent data from a binomial distribution with known repetition parameter N_i

Binomial logit regression model

$$y_i|\pi_i \sim \text{BiNom}\left(N_i, \pi_i\right), \quad \log \frac{\pi_i}{1-\pi_i} = \log \lambda_i = \mathbf{x}_i\boldsymbol{\beta} \qquad (1)$$

- $y_i|\pi_i \sim \text{BiNom}\left(N_i, \pi_i\right)$ is considered as the marginal distribution of an augmented model involving latent variables

Example

Titanic passenger data (Hilbe, 2007)

yi	Ni	intercept	child	female	class 3	class 2
14	31	1	1	1	1	0
13	48	1	1	0	1	0
76	165	1	0	1	1	0
80	93	1	0	1	0	1
140	144	1	0	1	0	0
75	462	1	0	0	1	0
14	168	1	0	0	0	1
57	175	1	0	0	0	0

y_i . . . number of survived passengers
N_i . . . number of exposed passengers in each group

$\sum N_i = 1286$ observations $\Rightarrow$ reduced to $N = 8$ covariate patterns

Individual RUM representation

- Frühwirth-Schnatter and Frühwirth (2007) consider each observation y_i as the aggregated number of successes of N_i independet binary experiments with outcomes $z_{1i}, \ldots, z_{N_i,i}$
- z_{ni} follows a binary logit model with the same log odds ratio as in (1), i.e. $\Pr(z_{ni} = 1|\pi_i) = \pi_i$
- the binary outcomes $z_{1i}, \ldots, z_{N_i,i}$ can be reconstructed easily from the binomial observation y_i

Individual RUM representation

- for each binary observation z_{ni} we introduce the utilities $y^u_{0,ni}$ and $y^u_{1,ni}$ of choosing category 0 or 1 as latent variables:

Individual RUM version

$$y^u_{0,ni} = \epsilon_{0,ni}, \qquad \epsilon_{0,ni} \sim \mathcal{EV}, \tag{2}$$

$$y^u_{1,ni} = \log \lambda_i + \epsilon_{1,ni}, \qquad \epsilon_{1,ni} \sim \mathcal{EV} \tag{3}$$

$$z_{ni} = I\{y^u_{1,ni} > y^u_{0,ni}\},$$

- where $i = 1, \ldots, N$ and $n = 1, \ldots, N_i$ with independent extreme value distributed errors $\epsilon_{0,ni}$, $\epsilon_{1,ni}$
- ***disadvantage:*** very high-dimensional latent variable

Aggregated representations

- instead of the whole sequence $y^u_{1,1i}, \ldots, y^u_{1,N_i,i}$ we introduce a ***single aggregated latent*** $y_i^\star$ for each binomial observation y_i
- properties of an aggregated representation:
  - **(1)** latent equation should take the form of a regression-type model

$$y_i^\star = \log \lambda_i + \epsilon_i$$

  - **(2)** error ϵ_i in the model has a distribution which depends on no or only few parameters so that it can be approximated easily
  - **(3)** it should be easy to simulate from the posterior $y_i^\star | y_i, \lambda_i$
- Frühwirth-Schnatter et al. (2009): **aggregated RUM representation** $\rightarrow$ the aggregation step is only applied to equation (3), modeling the utility of choosing 1

An alternative aggregated representation

- the aggregation step is also applied to equation (2), modeling the utility of choosing 0
- we aggregate the individual utilities for each category:

$$e^{-y_{0i}^{\star}} = \sum_{n=1}^{N_i} \exp(-y_{0,ni}^{u}), \qquad e^{-y_{1i}^{\star}} = \sum_{n=1}^{N_i} \exp(-y_{1,ni}^{u}),$$

where $e^{-y_{0i}^{\star}}, e^{-y_{1i}^{\star}} | \lambda_i$ are independent apriori, each following a Gamma distribution

Aggregation steps

- latent equations in form of a regression-type model:

$$y_{0i}^{\star} = \epsilon_{0i},$$
$$y_{1i}^{\star} = \log \lambda_i + \epsilon_{1i},$$

where $\epsilon_{ki} = -\log \xi_{ki}$ with $\xi_{ki} \sim \mathcal{G}(N_i, 1)$ follows the negative log-Gamma distribution with shape parameter N_i for $k = 0, 1$

- the single aggregated latent variable is then defined as

$$y_i^{\star} = y_{1i}^{\star} - y_{0i}^{\star}$$

Aggregated dRUM representation

Aggregated dRUM version

$$y_i^\star = \log \lambda_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{LG}(N_i), \tag{4}$$

- where $\varepsilon_i = \epsilon_{1i} - \epsilon_{0i}$ and $\mathcal{LG}(\alpha)$ is the Type III generalized logistic distribution with parameter α (see Balakrishnan, 1992)
- the first two moments are given by:

$$\mathrm{E}(\varepsilon_i|N_i) = 0, \qquad \mathrm{V}(\varepsilon_i|N_i) = 2\psi'(N_i)$$

Data-augmented independence MH

- error ε_i in (4) is approximated by the normal distribution $\mathcal{N}(0, 2\psi'(N_i))$
- the resulting posterior of $\boldsymbol{\beta}$ is used as proposal
- the proposal density is independent of the previous draw $\boldsymbol{\beta}^{\text{old}}$, but depends on the latent variable $\mathbf{z} = (y_1^\star, \ldots, y_N^\star)$
- high acceptance rate as the distribution of ε_i is approximately normal $\mathcal{N}(0, 2N_i)$ for large N_i

Auxiliary mixture sampler

- error ε_i in (4) is approximated by a scale mixture of normal distributions, all component means equal to 0:

$$f_{\mathcal{LG}}(\alpha) = \frac{\Gamma(2\alpha)e^{-\alpha\varepsilon}}{\Gamma(\alpha)^2(1+e^{-\varepsilon})^{2\alpha}} \approx q_\alpha = \sum_{r=1}^{R(\alpha)} w_r(\alpha)\,\varphi(0, s_r^2(\alpha)),$$

  where $\varphi(0, s^2)$ denotes a normal density with mean 0 and variance s^2

- the number of components $R(\alpha)$, the weights $w_r(\alpha)$ and the variances $s_r^2(\alpha)$ depend on $\alpha = N_i$

Hybrid auxiliary mixture (HAM) sampling

- combining both data-augmented MH and auxiliary mixture sampling
- in cases where the ratio y_i/N_i is neither close to 0 nor close to 1, the normal approximation in the MH algorithm will give a contribution α_i to the acceptance rate $\alpha = \prod_{i=1}^{t} \alpha_i$ close to 1
- for extreme ratios $y_i/N_i \leq c_{low}$ and $y_i/N_i \geq c_{up}$ (e.g. $c_{low} = 0.05$, $c_{up} = 0.95$) α_i will be considerably smaller
- ***idea:*** use the mixture approximation only for extreme ratios of y_i/N_i and the normal approximation of the MH sampler otherwise

Application

- application of the MCMC samplers to the Titanic passenger data and two simulated data sets to compare the different approaches
- ***MCMC details:***
  - independent standard normal prior for each regression coefficient
  - starting value for β: OLS estimation of the utilities on the covariates
  - 10000 draws from the posterior distribution after a burn-in of 2000 iterations
- ***'quality criteria':*** runtime, effective sampling size, effective sampling rate ESR, acceptance rate (MH & HAM sampler)

Titanic Data ($d = 8$, including interactions) $N = 8$, min $N_i = 31$, max $N_i = 462$, $\sum N_i = 1286$				
Sampler	α (%)	**T** (sec)	**med ESS** (total draws)	**med ESR** (draws/sec)
Indiv. dRUM-MH	55.8	16.1	1321.2	82.0
Agg. RUM-MH	93.8	4.4	741.0	168.0
Agg. dRUM-MH	99.6	4.4	1485.9	340.8
Agg. dRUM-Aux		7.1	1551.7	219.2
Agg. dRUM-HAM	99.6	8.5	1463.6	171.6

- HAM sampler: $c_{low} = 0.05$ and $c_{up} = 0.95$

- $y_i \sim \mathrm{BiNom}(N_i, \pi_i)$, where N_i and π_i are independent
- balanced data set

Simulated Data A ($d = 10$) $N = 485$, $\min N_i = 1$, $\max N_i = 116$, $\sum N_i = 8731$				
Sampler	α (%)	**T** (sec)	**med ESS** (total draws)	**med ESR** (draws/sec)
Indiv. dRUM-MH	48.4	105.2	727.2	6.9
Agg. RUM-MH	73.6	11.3	367.0	32.4
Agg. dRUM-MH	94.9	12.9	889.6	69.1
Agg. dRUM-Aux		16.6	977.4	58.8
Agg. dRUM-HAM	97.2	18.6	964.0	51.9

- HAM sampler: $c_{low} = 0.01$ and $c_{up} = 0.99$

- $y_i \sim \mathrm{BiNom}(N_i, \pi_i)$, where N_i and π_i are dependent
- for π_i near 0 the group sizes N_i are small, for π_i near 1 the group sizes N_i are large $\Rightarrow$ very extreme data set

Simulated Data B ($d = 10$)
$N = 490$, $\min N_i = 1$, $\max N_i = 126$, $\sum N_i = 25803$

Sampler	α (%)	**T** (sec)	**med ESS** (total draws)	**med ESR** (draws/sec)
Indiv. dRUM-MH	0.0	273.3	–	–
Agg. RUM-MH	0.0	10.7	–	–
Agg. dRUM-MH	0.0	11.0	–	–
Agg. dRUM-Aux		16.2	889.8	54.8
Agg. dRUM-HAM	98.5	18.7	886.5	47.5

- HAM sampler: $c_{low} = 0.1$ and $c_{up} = 0.9$

Results

- the aggregation step yields a considerable reduction of computing time compared to the individual dRUM
- the modifications in the aggregated dRUM lead to a remarkable gain in efficiency
- ***augmented MH sampler:*** unbeatable concerning runtime and ESR compared to the other two methods, if it moves away from the starting values and the markov chain converges to the stationary distribution

If the augmented MH sampler doesn't work...

- ***HAM sampler:*** slightly slower than the auxiliary mixture sampler as the algorithm has to compute mixture components AND acceptance probability
- ***AuxMix sampler:*** the most time-consuming part of the algorithm (sampling of the component indicators) is coded in a quite efficient way

Outlook

- improve the samplers as well for the multinomial regression model
- modify the models and algorithms to applicate them to data sets in economics and educational sciences (dissertation)

References

Balakrishnan, N. (ed.) (1992). *Handbook of the Logistic Distribution*, New York: Marcel Dekker.

Frühwirth-Schnatter, S. and R. Frühwirth (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis 51*, 3509-3528.

Frühwirth-Schnatter, S. and R. Frühwirth (2010). Data augmentation and MCMC for binary and multinomial logit models. In Kneib, T. and Tutz, G. (Eds.): *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pp. 111-132, Heidelberg: Physica-Verlag.

Frühwirth-Schnatter, S. , R. Frühwirth, L. Held and H. Rue (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing 19*, 479-492.

Hilbe, J.M. (2007). *Negative Binomial Regression*, Cambridge: University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In Zarembka, P. (Ed.): *Frontiers of Econometrics*, pp. 105-142, New York: Academic.

Scott, S.L. (2009). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statistical Papers*, accepted for publication.

Thank you for your attention!