A pseudo-marginal Metropolis-Hastings algorithm for estimating logistic regression models in the presence of missing data

Taylor R. Brown

Timothy L. McMurry, Alexander Langevin

Department of Statistics, School of Medicine, and Department of Systems and Information Engineering

Problem Description

Dependent variable: $y_i \in \{0, 1\}$
whether a car crash injury was "substantial"

Our predictors: $\mathbf{x}_i^T$

1. driver characteristics (e.g. age, sex, body mass index)
2. vehicle attributes (e.g. model year, type of vehicle)
3. direction of force (driver side or passenger side)
4. change in velocity at impact
5. was the knee airbag deployed (our variable of interest)

Issues:

1. missing predictor data
2. lots of prior knowledge (from Tim)

For more details see Brown, McMurry, and Langevin, 2019.

Our Model

Conditional likelihood

$$p(\mathbf{y} \mid \mathbf{x}, \beta) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{1-y_i}$$

Our Model

Conditional likelihood

$$p(\mathbf{y} \mid \mathbf{x}, \beta) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{1-y_i}$$

Missing predictors

$$p(\mathbf{x}_{\text{mis}} \mid \mathbf{x}_{\text{obs}}, \alpha)$$

Our Model

Conditional likelihood

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right)^{1-y_i}$$

Missing predictors

$$p(\mathbf{x}_{\text{mis}} \mid \mathbf{x}_{\text{obs}}, \boldsymbol{\alpha})$$

Missingness-mechanism $m_{ij} \in \{0, 1\}$

$$p(\mathbf{m} \mid \mathbf{y}, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \phi)$$

Our Model

Conditional likelihood

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right)^{1-y_i}$$

Missing predictors

$$p(\mathbf{x}_{\text{mis}} \mid \mathbf{x}_{\text{obs}}, \boldsymbol{\alpha})$$

Missingness-mechanism $m_{ij} \in \{0, 1\}$

$$p(\mathbf{m} \mid \mathbf{y}, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \boldsymbol{\phi})$$

The prior:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi})$$

Sampling the Posterior

Sample a Markov chain that targets the posterior by proposing new

$$\theta' \sim q_{\text{MH}}(\cdot \mid \theta^{i-1})$$

and accept the draw with probability

$$a(\theta, \theta') = \min\left[1, \frac{\hat{p}(\mathbf{m}, \mathbf{y} \mid \mathbf{x}_{\text{obs}}, \theta')p(\theta')q_{\text{MH}}(\theta \mid \theta')}{\hat{p}(\mathbf{m}, \mathbf{y} \mid \mathbf{x}_{\text{obs}}, \theta)p(\theta)q_{\text{MH}}(\theta' \mid \theta)}\right] \tag{1}$$

$\hat{p}(\mathbf{m}, \mathbf{y} \mid \mathbf{x}_{\text{obs}}, \theta')$ is an unbiased, nonnegative, computationally expensive, importance sampling estimate of the intractable integral:

$$\int p(\mathbf{m}, \mathbf{y}, \mathbf{x}_{\text{mis}} \mid \mathbf{x}_{\text{obs}}, \theta')d\mathbf{x}_{\text{mis}}$$

References

Brown, Taylor R., Timothy L. McMurry, and Alexander Langevin (2019). *A Pseudo-Marginal Metropolis-Hastings Algorithm for Estimating Generalized Linear Models in the Presence of Missing Data*. eprint: arXiv:1907.09090.