専門職学位課程 2 年　田島 里華

ベイズ的手法を活用したスパースな疫学研究データの解析

ある疾患の有無に対する曝露因子の影響の大きさを交絡因子の影響を調整しロジスティック回帰で推定する解析は広く一般的に行われている。研究の対象者数もイベント発症数も多く、調整因子の層ごとに（疾患の有無）×（興味のある曝露因子の有無）の分割表を書いたとき各セルに一定量以上の人数が存在するならば、ロジスティック回帰の結果は十分に信用できる。しかし先のような分割表を書いた際に人数のカウントが極端に小さなセルが存在してしまうようなデータはスパースデータと呼ばれ、通常のロジスティック回帰を行うと推定結果の統計的性質が悪いという問題が指摘されている。データがスパースになる原因としては、研究の対象者数が少ない、集団全体でのイベント割合が小さい、共変量について数が多い・曝露割合が小さい・共変量同士の相関が強い等複数の要因が複合的に影響している場合が多く、このような傾向を持つデータの解析には、通常の方法よりもスパースさに頑健な方法を用いた方がより優れた結果が得られることが期待される。

現在スパースデータの解析に適した方法として、exact 法、Firth の罰則付き尤度を用いた方法、ベイズ的方法が知られているが、対象者数や調整したい交絡因子の数が多い疫学研究の場面には、計算量の多い exact 法は適さない。一方 Firth の方法は、計算量も少なく一般的な統計解析ソフトウェアによる実行が容易であり、近年利用が広まりつつある。ところでこの Firth の方法は、尤度方程式を解く際に Jeffreys の無情報事前分布を用いたベイズ的方法と解釈することもできる。疫学研究の場面においては、イベントに対する曝露因子の影響の方向および大きさの範囲に関する事前の見地（例 : オッズ比はほぼ間違いなく 0.05 から 20 の間に収まるだろう、といった大まかな見積もり）が研究実施時点で存在している場合が多く、解析段階にてこのような事前の知識を活用すれば、全く事前情報を用いない Firth の方法に比べてさらに解析結果の統計的性質が改善されることが見込まれる。

よって本研究は、疫学研究の場面から得られたスパースデータの解析にベイズ的方法を適用することで、より優れた推定が可能になるか検証することを目的とする。

本研究のモチベーショナルデータは、日本動脈硬化縦断研究（JALS）の 0 次統合研究参加者のうち、女性対象者のみに限定した集団である。このデータから心筋梗塞発症に対する総コレステロール（TC）の影響の大きさを探索しようと試みると、イベント発症割合が少ないこと、調整変数の数が比較的多いこと、また曝露割合の小さい説明変数が存在すること等によりデータがスパースになっており、通常のロジスティック回帰から得られるパラメータ推定値は常識では考えられないほど極端に大きな値をとってしまう。そこで、一

つは Firth の方法、その他に心筋梗塞発症に対する TC のオッズ比の大きさに関して既に得られている見地を事前情報として用いたベイズ的方法のうち、計算方法の異なる 2 つの方法の計 3 通りの解析方法を用いることで、解析結果がデータのスパースさから受ける影響を通常のロジスティック回帰よりも減らすことができるか検討する。なお 2 通りのベイズ的方法とは、近似的ではあるもののベイズ的解析結果を MCMC なしで短時間のうちに計算可能な Data Augmentation Prior を用いる方法と、従来に比べ格段に実行速度が短くなり、計算アルゴリズムも改良され続けている MCMC を用いてより正確に事後分布に関する推測を行う方法である。

本研究では、はじめにモチベーショナルデータにおけるイベント発症割合や曝露割合、イベント発症と各説明変数とのオッズ比の大きさ等を模した n=10,000 のデータセットを 10,000 個シミュレーションにより発生させ、通常のロジスティック回帰、Firth の方法、DAP を用いる方法、MCMC を用いる方法、以上 4 つの方法により解析する。その後、バイアス、平均二乗誤差、CI が真値を含む確率、CI の平均長、パラメータ計算が収束した回数、以上の 5 つの評価基準を用いて各解析方法の性能を比較する。

つぎに、モチベーショナルデータである JALS 0 次統合研究の女性対象者において、4 つの解析方法により TC と心筋梗塞発症との関連を調べる。さらに、TC と心筋梗塞発症との関連については「やや正の関係がある」という事前分布を想定することが最も自然であると考えられるが、この他に「かなり強い正の関係がある」、「負の関係がある」、「事前情報がない」といった異なる事前分布を用いた場合の推定結果の変化についても検証する。

9 月 17 日の抄読会では、スパースデータに関する背景、本研究にて扱う各解析手法の詳細、シミュレーション実験の設定の詳細および結果について説明する。その後モチベーショナルデータ解析の進捗、また今後の予定を報告する。

参考文献

1. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409-19
2. Hamra G, MacLehose R, Cole S. Sensitivity analyses for sparse-data problems—using weakly informative bayesian priors. *Epidemiol* 2013;**24**:233–9.
3. Greenland S. Putting background information about relative risks into conjugate prior distributions. *Biometrics* 2001;**57**:663–70.
4. Sullivan SG, Greenland S. Bayesian regression in SAS software. *Int J Epidemiol* 2013;**42**:308–17.
5. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int J Epidemiol* 2007;**36**:195–202.