専門職学位課程 2 年　田島　里華

Tackling the Sparse-Data Problem with Bayesian Methods

Sparse data（イベント数に比べ説明変数の数が多い場合や、複数の説明変数で層別すると、イベント数が少ない説明変数の組み合わせパターンが生じる場合等）に対し回帰モデルを当てはめ一般的な最尤法でパラメータ推定を試みると、推定値が有限の値に収束しない、もしくは有限解が得られたとしてもバイアスが入る sparse-data problem が生じることが知られている。この原因として、漸近正規性の破綻等が考えられる[1-3]。この問題を解決するには、漸近性を仮定せず確率計算を正確に行う exact 法、尤度に罰則項を加えることで解析的に推定値を改良する罰則付き最尤法、もしくはデータに外部情報を追加することで推定の安定化を図るベイズ的方法を適用することが考えられる。

exact 法の最大の難点は膨大な計算量であり、複数の説明変数を考慮した正確な計算の実行は現時点では現実的な選択肢とは言えない[4, 5]。これに対し、罰則付き最尤法の中でも Firth の提案した補正項を加えた尤度を用いる方法は計算が容易であり、また推定値のバイアスだけでなく分散も抑えることが複数のシミュレーション研究により示されている[6, 7]。

一方ベイズ的方法というと、非常に恣意的な事前分布を入れ込むようなイメージを抱かれることもある。しかし医学研究では、研究を行う前に調べたい事柄に関して何らかの知識を既に持ち合わせていることがほとんどである。例えば疫学研究により、あるありふれた曝露の疾患に対する影響の大きさを調べる際、影響が極端に大きい（例えばリスクを 100 倍にする）、もしくは小さい（例えばリスクを 1/100 にする）といったことは現実的に考えにくい。このような言わば「常識」程度の既存の知識を事前分布として入れ込むベイズ的方法でも、sparse data 解析の結果を安定化させるのに十分役立つ、という提案もなされている[8, 9]。ベイズ的方法を適用する際には、事前分布の設定方法とその組み込み方、また事後分布の計算が大きな問題となってくる。抄読会では主に Greenland によるベイズ的方法を用いた sparse data 解析の理論の論文[8-13]や、こうした解析方法を医学に限らず人口統計やマーケティングの分野で適用した実例[3, 14, 15]の論文等をレビューする。

修士論文では以下の 3 点を中心に取り組む予定である:

1. sparse data と言ってもどれ位の sparseness でどの程度の大きさのバイアスが入るかを、研究対象者数、イベント数、説明変数の数とその影響の大きさ、といった要因を動かしてシミュレーションにより調べる
2. 1 の各要因の組み合わせパターンに対し、Firth の方法やベイズ的方法が一般的な解析方法と比べてどの程度バイアスと分散を小さくするかを、シミュレーションを通し検証する
3. クモ膜下出血の発症に対する、性・年齢ごとの飲酒および喫煙の影響を調べようとする

と、特に高齢の女性において飲酒や喫煙とイベント発症とのクロス表が往々にして sparse になるが、こうした場面においても 1 と 2 で検討した方法論を用いると一般的な解析方法と比べ推定結果が良くなるか検討する
抄読会では、1, 2 のシミュレーションの設定などについても紹介する。

References

1. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol* 2000; **151**: 531–9.
2. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409–19.
3. Hamra G, MacLehose R, Cole S. Sensitivity analyses for sparse-data problems – using weakly informative bayesian priors. *Epidemiology* 2013; **24**: 233–9.
4. Cytel Inc. LogXact® 9 User Manual. Cytel, Cambridge, MA, 2010.
5. SAS Institute Inc. SAS/STAT 9.2® User's Guide. SAS Institute Inc., Cary, NC, 2008.
6. Kessels R, Jones B, Goos P. An argument for preferring Firth bias-adjusted estimates in aggregate and individual-level discrete choice modeling. *Working papers of the Faculty of Applied Economics, University of Antwerp* 2013. [cited 2014 Apr]. Available from: http://www.itmma.ua.ac.be/main.aspx?c=*TEWENG&n=112737
7. Eyduran E. Usage of Penalized Maximum Likelihood Estimation Method in Medical Research: An Alternative to Maximum Likelihood Estimation Method. *J Res Med Sci* 2008; **13**: 325-30.
8. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol* 2006; **35**: 765-75.
9. Greenland S. Bayesian perspectives for epidemiological research: II. Regression analysis. *Int J Epidemiol* 2007; **36**: 195-202.
10. Greenland S. Prior data for non-normal priors. *Stat Med* 2007; **26** :3578–90.
11. Greenland S, Christensen R. Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Stat Med* 2001; **20**: 2421–8.
12. Greenland S. Putting background information about relative risks into conjugate prior distributions. *Biometrics* 2001; **57**: 663–70.
13. Greenland S. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics* 2003; **59**: 92–9.
14. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression. *JASA* 1991; **86**: 68-78.
15. Lenk P, Orme B. The value of informative priors in Bayesian inference with sparse data. *J Mark Res* 2009; **46**: 832–45.