専門職学位課程 1 年　田島　里華

Stably Estimating the Risk of Rare Events

1. Introduction

昨年度卒業論文の執筆に当たり、JALS（Japan Arteriosclerosis Longitudinal Study; 日本動脈硬化縦断研究）の 0 次統合研究データの一部を用いて、ロジスティック回帰による女性の心筋梗塞発症リスクの推定を試みた際に計算が不可能となり、性別を無視した全対象者および男性のみについてリスク推定を行った。この原因として、女性の心筋梗塞発症数が少ないにもかかわらず、複数の説明変数を含むロジスティック回帰モデルを当てはめたためパラメータ計算が収束しなかったと考えられる。

複数の因子で層別したときに各層内でのイベント数が少なくなるような(sparse な層が多くなる）場合には、複数の層を統合して解析しリスクを求める方法が一般的である。しかしこれとは別に、層の統合を行わずイベント数の少ない層ごとのままでも安定的にリスクを推定する解析方法はないのだろうか、という疑問を抱いた。そこで本抄読会では、sparse data の解析方法として広く用いられている条件付きロジスティック回帰（conditional logistic regression）について触れた後、今後こうした場面への応用を目指して改変できるのではないかと考えた階層回帰（hierarchical regression）について詳しく紹介する。

2. Conditional Logistic Regression

通常のロジスティック回帰を用いて層内での人数が少ない sparse data を解析すると、オッズ比が null から離れる方向のバイアス（sparse-data bias）が生じてしまう。これを防ぐ方法として条件付きロジスティック回帰が一般的に用いられるが、モデル内の未知パラメータ数が多い場合などには、sparse-data bias を制御しきれない恐れがある。バイアスの減少を目指す条件付きロジスティック回帰に対して、sparse data に何らかの外部情報をベイズ的に加えることで推定の安定化を図った方法が階層回帰である[1]。

3-1. Hierarchical Regression

ある説明変数 y の期待値を、説明変数ベクトル $\mathbf{x}$ と調整変数ベクトル $\mathbf{w}$ を含むモデルにより予測する一般化線形モデルは

$$g\left[E\left(y\,|\,\mathbf{x},\mathbf{w}\right)\right]=\alpha+\mathbf{x}\boldsymbol{\beta}+\mathbf{w}\boldsymbol{\gamma}$$

（ただし $\boldsymbol{\beta},\boldsymbol{\gamma}$ は未知パラメータベクトル（うち $\boldsymbol{\gamma}$ の要素は局外母数)、g(・)はリンク関数）と書き表すことができる。y に対して $\boldsymbol{\beta}$ の要素の数が多すぎるため推定が不安定な場合に、$\boldsymbol{\beta}$ 自体に何らかの分布を仮定し、その分布をより少ない未知パラメータ数で定義することで $\boldsymbol{\beta}$ の要素同士をいわば関連づけ、それぞれの推定値を安定させることを目指したモデルが階層モ

デルと呼ばれる。$\boldsymbol{\beta}$ の分布の一例として、共変量行列を $\mathbf{Z}$、係数行列を $\boldsymbol{\pi}$、誤差 $\boldsymbol{\delta}$ の各要素が $N(0, \tau^2)$に従うとして

$$\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\pi} + \boldsymbol{\delta}$$

などを想定することができる[2]。

3-2. Adjustment of τ^2 in Hierarchical Regression

階層モデルにおける $\boldsymbol{\beta}$ の分布の決め方は様々あるが、3-1 節で例示した分布の誤差分散 τ^2 の大きさを、解析するデータ自体から見積もる方法は empirical-Bayes approach として知られている。このアプローチでは τ^2 を過度に厳しく制限してしまう場合があり、逆に $\tau^2 = \infty$ として誤差分散に何の制限も設けなければ未知パラメータ推定が不安定になりがちである。このため、両者の中間を取り、主観的情報を用いて τ^2 を適当な範囲に制限する方法を Greenland は提唱し、Semi-Bayes approach と名付けた。抄読会当日は、τ^2 の制限方法の異なる複数の階層回帰モデルに関するシミュレーション並びに実データへの当てはめ結果を通じ、それぞれのモデルの推定能力などを比較した論文も紹介する。

4. Upcoming Plans

1 節で想定したような場面に、実際階層回帰が拡張可能であるかどうか検討したい。また、両方法論について勉強を続けつつ、条件付きロジスティック回帰および階層回帰が、どのようなデータで用いると何が原因で問題が生じるのか、といった具体的な事項についてもシミュレーション等を通して検証する予定である。同時に、今回紹介した 2 つの方法論以外でも、想定したような場面に適用できる可能性のある他の方法論も探索していきたい。

5. 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. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med* 1993; **12**: 717-36.
3. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med* 1992; **11**: 219-30.