専門職学位課程 1 年　田島　里華

Dealing with Separation in Logistic Regression for Rare Events

前回の抄読会では、JALS 0 次データの解析中に、イベント数の少ない 1 つのサブグループについてロジスティック回帰によりリスク推定を試みたところ､SAS で警告が表示され、推定値の信頼性が十分でなかったという体験を契機に、イベント発症が稀なサブグループについて、限られた量のデータからより正確かつ精密なリスク推定を行う方法論に興味を持った、という紹介をした。その警告は、具体的にはデータの「準完全分離」を知らせていた。ロジスティック回帰での完全分離とは、0, 1 の 2 値をとる $y_i\left(i=1\sim n\right)$ と説明変数 $\boldsymbol{\beta}$、回帰係数 $\boldsymbol{x}_i$ について、すべての $\boldsymbol{x}_i$ に対し

$$\begin{cases}\boldsymbol{\beta}'\boldsymbol{x}_i<0\left(y_i=0\right)\\ \boldsymbol{\beta}'\boldsymbol{x}_i>0\left(y_i=1\right)\end{cases} \tag{1}$$

なる $\boldsymbol{\beta}$ が存在する状態をいい､(1)式の不等号にイコールが入る場合を準完全分離という[1]。完全／準完全分離状態の下で、ロジスティック回帰の当てはめを試み、一般的な最尤法によりパラメータを推定しようとすると、最尤推定量が一意に定まらないことが知られている[2]。言い換えると、（説明変数）×（結果変数）の分割表を書いたとき、カウントが 0 であるセルが 1 つでもあればロジスティック回帰パラメータの最尤推定量は存在しない[3]。稀なイベントについて複数の調整因子で層別した場合や、イベント自体が稀でなくとも、説明変数の端のカテゴリーにおいては、完全／準完全分離状態が生じやすいことが予測される。

完全／準完全分離状態でのロジスティック回帰の推定能力を改善させる方法として、

1. ベイズ的方法
2. penalized likelihood を用いる方法
3. exact 法

が提案されている[3-5]。1 つ目のベイズ的方法は、原理的には、分離したデータに事前情報を加えることで一種のスムージングを施し、推定値が有限の値に収束するよう修正する。事前情報の定め方が自由であるため柔軟なモデリングが可能となるが、その反面恣意性も問題になりやすい。2 つ目の penalized likelihood においては、penalty の定め方が複数提案されている中でも、特に有用性が期待されているのが Firth’s penalized likelihood である。これは対数尤度関数を $l\left(\theta\right)$、フィッシャー情報量行列を $i\left(\theta\right)$ として、

$$l\left(\theta\right)+\frac{1}{2}\log\left|i\left(\theta\right)\right|$$

という形で対数尤度関数を補正する[6]。この Firth の修正を行った対数尤度関数を用いてロジスティック回帰パラメータを推定すると、完全／準完全分離の場合でも最尤推定量が一意に定まるようになる。また、この修正はベイズの枠組みにおける Jeffreys の事前分布を用いることと同等である[4, 7]。最後に 3 つ目の exact 法について、理論は古くは Fisher の時代から提案されていたものの、変数の数や対象者数が多い場合に計算量が膨大となるため実行は困難を極めていた。しかしコンピュータの計算能力が向上し、その上新たな計算アルゴリズムも提案されたことによって、回帰分析においても十分統計量の分布について正確な確率計算を行うことが可能となりつつある。この方法を用いると、完全／準完全分離の下で最尤推定量が一意に定まらない場合であっても median unbiased estimate を計算することが可能となる。さらに、近年では SAS の PROC LOGISTIC や、exact な回帰分析に特化したソフトウェアである LogXact を用いて exact 法を適用できる範囲が拡大している。

本抄読会では、完全／準完全分離状態のデータに対応すべくロジスティック回帰を改良するこれらの方法について、互いを比較しつつ概要を紹介することで、イベントが稀なサブグループにおけるリスク評価の方法論を考える上での足掛かりにしたい。

5. References

1. 大倉征幸、鎌倉稔成. 精確ロジスティック回帰の近似推定値. *応用統計学* 2007; **36**: 87-98.
2. Albert A, Anderson A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; **71**: 1-10.
3. Allison PD. Convergence Failures in Logistic Regression. *SAS Global Forum* 2008; **360**: 1-11.
4. Agresti A. *Categorical Data Analysis, 3rd Edition.*Wiley, Hoboken, NJ, 2013.
5. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409-19.
6. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**: 27-38.
7. Cytel Inc. *LogXact® 9 User Manual*. Cytel, Cambridge, MA, 2010.
8. Mehta CR, Patel NR. Exact Logistic Regression: Theory and Examples. *Stat Med* 1995; **14**: 2143-60.