2024年5月24日 統計数理研究所 オープンハウス

修正ポアソン回帰・修正最小二乗回帰におけるRidge回帰、Lasso回帰とElastic-net

北野 喬大　　総合研究大学院大学統計科学コース 博士課程4年

研究の背景

臨床研究、疫学的研究では、二値アウトカムに対してロジスティック回帰モデルにより推定されたオッズ比（OR）が報告されることが多い。アウトカムの発生頻度が低い場合、ORはリスク比(RR)を近似できることが知られている。一方で、アウトカムの発生頻度が高い場合、ORはRRを過大評価してしまうことが知られている。この場合、RRとリスク差(RD)がより適切な疫学指標となる。RR、RDの推定には二項回帰モデルが用いられるが、しばしば収束せず、RR、RDの推定値が得られないことが知られている。この問題を克服するため、修正ポアソン回帰モデルと修正最小二乗回帰モデルが提案されている（Cheung, 2007; Zou, 2004）。ロジスティック回帰モデルは、説明変数に完全分離がある場合や、変数間に強い相関がある場合（多重共線性）に推定が不安定になることが知られており、この問題はサイズの小さなデータセットで頻繁に見られることから、Small or sparse data issueと呼ばれている。この問題に対処するために、Ridge, Lasso, Elastic-netなどの縮小推定の適用が検討されている。修正ポアソン回帰モデルと修正最小二乗回帰モデルも二項回帰モデルであることから、small or sparse data issueがあると考えられるが、これらのモデルに対する縮小推定の適用を評価した先行研究はない。そこで、これらのモデルに縮小推定を適用する手法を提案し、実際の疫学研究の事例データを用いてその有用性を評価する研究を行った。

方法：修正ポアソン回帰と修正最小二乗回帰モデル

二項回帰モデルにおけるアウトカムの発生確率

$$\Pr(y_i = 1|\boldsymbol{x}_i) = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)$$

$$\Pr(y_i = 1|\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- $y_i\ (i = 1,2,\ldots,n)$ ： 二値の結果変数
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T\ (i = 1,2,\ldots,n)$：予測変数
- $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$：回帰係数（切片を含む）

回帰係数は一般化線形モデル（GLM）の枠組みにおいて最尤推定法により推定され、それぞれ対数リスク比、リスク差の推定値と解釈される。

GLMの推定方程式

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i^T V_i^{-1}(Y_i - \mu_i) = 0$$

μ_i（平均）が正しく定義されていれば、二値の結果変数にポアソン回帰モデルおよび最小二乗回帰モデルを適用した場合であっても、推定関数の不偏性が成り立ち、得られる推定量は対数リスク比、リスク差の一致推定量となる。ただし、V_i（分散）は誤特定されているため、得られる推定量はquasi-maximum likelihood estimatesとなる。

方法：Ridge, Lasso, Elastic-net

Ridge回帰 $\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$

Lasso回帰 $\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$

Elastic-net $\ell(\boldsymbol{\beta}) + \lambda \left\{(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|\right\}$

- λ：Tuning parameter(縮小の程度を決める定数)
- α：RidgeとLassoの罰則項の重みを決める定数

上記３つに手法において、対数尤度関数にλ以下の罰則項を加えた式を最小化し、回帰係数を推定する。罰則項の存在により回帰係数は縮小され、モデルの過剰適合を防ぐことができる。Ridge回帰は元々、多重共線性に対処するために考案された。Lasso回帰は、縮小推定に加え変数選択を行うこともできる（Tibshirani, 1996）。Elastic-netはRidge, Lassoの罰則項を組み合わせて構築されており、多重共線性への対処と変数選択の両方を行うことができる。

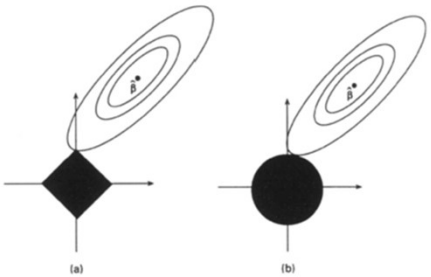

Figure 1. LassoとRidgeの最適化（Tibshirani, 1996から引用）

提案法の事例データへの適用

提案法

- 臨床研究、疫学研究における有用なRR、RDの推定方法として、修正ポアソン、修正最小二乗回帰モデルにRidge, Lasso型縮小推定を適用することを提案する。
- また、quasi-MLEの信頼区間を推定するためのBootstrapping algorithmを提案する。
- これらを円滑に実施するためのプログラムを提供する。

事例データへの適用

以下の２つの事例に提案法を適用し、解釈可能なRR, RD及びより正確な信頼区間が推定されることを確認した（Figure 2）。

- 事例１：The Retrospective Cohort Study of the Effects of Donor KIR genotype on the reactivation of cytomegalovirus after myeloablative allogeneic hematopoietic stem cell transplant（Sobecks et al., 2011）
- 事例２：National Child Development Survey (Power & Elliott, 2006)

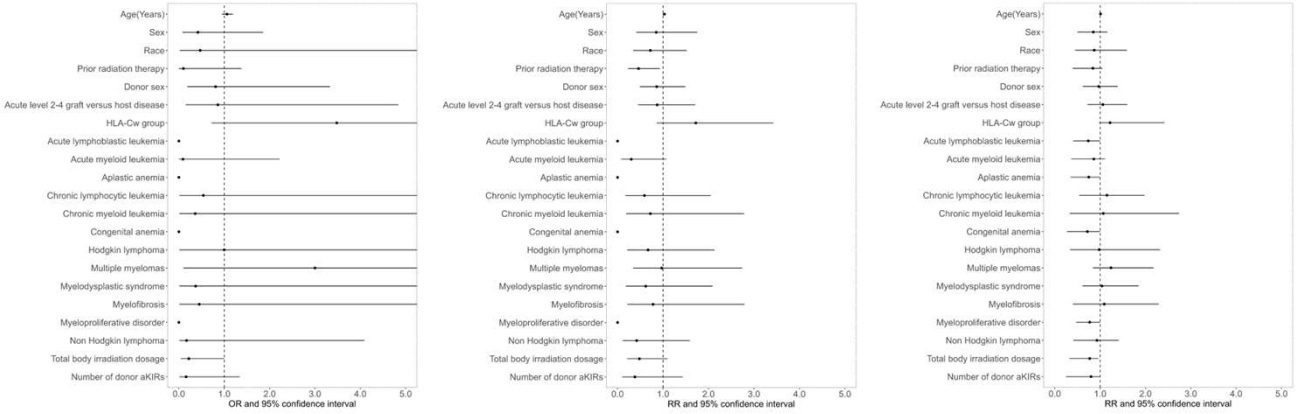

Figure 2. 事例1を用いて推定したOR, RR（quasi-MLE）, RR（Ridge regression）