規模の異なる変量群をもつロジスティック回帰モデルの係数 2 段階推定

上智大学大学院理工学研究科数学領域・院　竹下　佳宏

1　研究の背景

21 世紀に DNA の全塩基配列情報を把握することが可能となり，ゲノムワイド関連解析（以下，GWAS）が行われるようになった [1]．GWAS では一塩基多型（以下，SNP）と呼ばれる数十万単位ある特定の塩基配列情報に着目して，生物の性質をひもとく．遺伝情報の解析手法では，古典的な統計解析が親しまれているが，機械学習の技術が応用されるなど，最近では新しい技術も活用されている [2]．本研究では，遺伝子情報の解析に用いられるロジスティック回帰モデルの新たな係数推定方法を提案する．

2　従来の研究手法

ある i 番目の個体に対し，病気である $(y_i = 1)$ かそうではない $(y_i = 0)$ かという 2 値の応答変数を，遺伝子の情報や，その個体の環境情報を考慮して説明するモデルとして，ロジスティック回帰モデルがある．一般には病気の有無以外の 2 値の応答変数にもこのモデルは用いられるが，本研究では病気の有無についてのみ言及する．

m_e 個の環境要因 $\{E_{i,k}\}_{k=1}^{m_e}$ と m_s 個の遺伝要因である SNP 情報 $\{SNP_{i,j}\}_{j=1}^{m_s}$ のうち j 番目の SNP のみを用いて，個体 i が病気にかかる確率 $p_{i,j}$ を次のロジスティック回帰モデルで表現する [2].

$$\log\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \alpha_{0,j} + \sum_{k=1}^{m_e} \alpha_{k,j} E_{i,k} + \beta_{snp,j} SNP_{i,j} \qquad (1)$$

SNP 情報は数十万単位あるため，一度に SNP をすべて解析するのではなく，1 つの SNP 情報に対してその都度モデルを構築し、その中から病気を十分に説明できるようなモデルを考えることで，病気に関連を持つ SNP を見つける．実際はモデル式 (1) の回帰係数に推定値を代入して $p_{i,j}$ について解き，その値があるしきい値を上回れば病気 $(y = 1)$，下回れば健康 $(y = 0)$ と予測する．

従来の研究手法においては，以下の点で問題があった．

- SNP ごとにすべての回帰係数を推定し直すが，本来は環境要因は SNP から独立しているため，環境要因の回帰係数が j 番目の SNP に依存していることが不自然である
- SNP の数だけ環境要因の回帰係数を推定するのは計算効率が悪い
- 1 つのモデルにつき 1 つの SNP しか考慮しないため，複数の SNP の交互作用が説明できない

これらの問題を解決する回帰係数の推定を本研究で行った．

3　回帰係数の 2 段階推定

3.1　推定方法の概要

ロジスティック回帰モデルは，応答変数が 0 か 1 かの 2 値の他に，応答変数が生起または反応の確率である場合にも利用できる [3]．現実には生起の確率そのものが応答変数として観測されることはほとんどないため，標本全体に対する生起数の割合で応答変数の値を推定する．

この考え方を利用して，2 段階で回帰係数の推定を行う．

1. 標本の個体 $i \in \{1, 2, \ldots, n\}$ について，i と同じ環境下で罹患している標本数を求め，標本全体に対する割合で罹患確率 $\widehat{p}_{e,i}$ を推定
2. 次のような環境要因だけのロジスティック回帰モデルを考え，係数 $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{m_e})$ を推定

$$\log\left(\frac{\widehat{p}_{e,i}}{1-\widehat{p}_{e,i}}\right) = \alpha_0 + \sum_{k=1}^{m_e} \alpha_k E_{i,k} \qquad (2)$$

3. 回帰係数の推定値 $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_k)_{0 \le k \le m_e}$ を用いて，個体 i の罹患確率の推定値 $\widetilde{p}_{e,i}$ を次の式で計算

$$\widetilde{p}_{e,i} = \frac{\exp(\widehat{\boldsymbol{\alpha}} \boldsymbol{E}_i')}{1 + \exp(\widehat{\boldsymbol{\alpha}} \boldsymbol{E}_i')}$$

ここで，$\boldsymbol{E}_i = (1, E_{i,1}, \ldots, E_{i,m_e})$

4. $\widehat{p}_{e,i}$ と $\widetilde{p}_{e,i}$ の差，すなわち，環境要因だけでは説明できない罹患確率の部分を SNP 情報を用

いて説明．回帰係数の問題としては，次のモデル式で回帰係数 $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{m_s})$ を推定

$$\log\left(\frac{\widehat{p}_{e,i}}{1-\widehat{p}_{e,i}}\right) - \log\left(\frac{\widetilde{p}_{e,i}}{1-\widetilde{p}_{e,i}}\right) = \beta_0 + \sum_{j=1}^{m_s} \beta_j SNP_{i,j} \tag{3}$$

2 段階に分けることにより，環境要因の回帰係数は SNP の影響を受けずに推定できる．そのため, (2) の回帰係数は (1) 式とは異なり, 添字 j が消去される．

3.2 環境要因の回帰係数の推定

3.2.1 応答変数の補正

現実のデータでは，$\widehat{p}_{e,i} = 0, 1$ もあり得る．その場合，(2) と (3) が定義できない．そこで，十分小さい $\delta > 0$ を用いて，$\widehat{p}_{e,i}$ を次のように補正する．

$$\widehat{p}_{e,i}(\delta) = \begin{cases} \delta & : \ \widehat{p}_{e,i} = 0 \\ \widehat{p}_{e,i} & : \ 0 < \widehat{p}_{e,i} < 1 \\ 1-\delta & : \ \widehat{p}_{e,i} = 1 \end{cases} \tag{4}$$

$\widehat{p}_{e,i}$ を $\widehat{p}_{e,i}(\delta)$ で置き換えて，3.2.2 節以降で説明する回帰係数の推定を行う．

3.2.2 最尤法による推定

環境要因の回帰係数 $\boldsymbol{\alpha}$ は，$\{\widehat{p}_{e,i}(\delta)\}_{i=1}^n$ を応答変数の値として用いた最尤法で推定する．実際には最尤解は陽な形で得ることができないので，ニュートン法によって数値的に求める．

$\{\widehat{p}_{e,i}(\delta)\}$ にもとづく最尤解 $\widehat{\boldsymbol{\alpha}}_\delta$ は，補正量 $\delta > 0$ の影響を受けており，本来の $\{\widehat{p}_{e,i}\}_{i=1}^n$ にもとづく解 $\widehat{\boldsymbol{\alpha}}$ とはずれがある．けれども，次の命題から，$\widehat{\boldsymbol{\alpha}}_\delta$ が $\widehat{\boldsymbol{\alpha}}$ の近似解となっていることがわかる．

命題 $\widehat{\boldsymbol{\alpha}}$ を補正がない場合 ($\delta = 0$) の $\{\widehat{p}_{e,i}\}_{i=1}^n$ による最尤推定量とするとき，

$$\lim_{\delta \to +0} \mathbb{E}\left\{(\widehat{\boldsymbol{\alpha}}_\delta - \widehat{\boldsymbol{\alpha}})^2\right\} = 0$$

よく知られているように，最尤推定量 $\widehat{\boldsymbol{\alpha}}$ は漸近的に正規分布に従う．命題の結果から，十分小さな δ について，$\widehat{\boldsymbol{\alpha}}_\delta$ も近似的に正規分布に従う．よって，$\widehat{\boldsymbol{\alpha}}_\delta$ について正規性にもとづく有意性の検定ができる．

3.3 遺伝要因の回帰係数の推定

3.1 節の (3) 式にもとづいて，SNP の回帰係数 $\boldsymbol{\beta}$ を推定する．回帰モデルとしては，重回帰モデルにしたがって最小自乗解を得る．以下，環境要因で説明し残した部分を次のように定める．

$$\boldsymbol{r} = (r_1, \ldots, r_n)'$$
$$r_i = \log\left\{\frac{\widehat{p}_{e,i}(\delta)}{1-\widehat{p}_{e,i}(\delta)}\right\} - \log\left\{\frac{\widehat{p}_{e,i}}{1-\widehat{p}_{e,i}}\right\}$$

また，SNP 情報を $n \times (m_s + 1)$ 行列 $\boldsymbol{S}$ として表すとき，最小自乗解は次のように表される．

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{r} \tag{5}$$

ただし，行列 $\boldsymbol{S}$ の 1 列目は切片項に対応するため，成分が 1 だけの列とする．

ところが，一般に SNP の情報はサンプル数 n よりも SNP 数 m_s の方が多いため，$\boldsymbol{S}$ にはランク落ちが生じる．そこで，(5) ではなく，$\boldsymbol{\beta}$ は $\boldsymbol{S}$ のムーア・ペンローズ型一般逆行列 $\boldsymbol{S}^+$ を用いた次の形で推定する．

$$\widehat{\boldsymbol{\beta}}_+ = \boldsymbol{S}^+\boldsymbol{r}$$

推定量 $\hat{\boldsymbol{\beta}}_+$ は次の性質を持つ [4].

1. $\widehat{\boldsymbol{\beta}}_+ = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{r} - \boldsymbol{S}\boldsymbol{\beta}\|$
2. $\hat{\boldsymbol{\beta}}_+$ は不偏推定量
3. 任意の不偏推定量 $\tilde{\boldsymbol{\beta}}$ に対して，

$$V\left(\widehat{\boldsymbol{\beta}}_+\right) \leq V\left(\widetilde{\boldsymbol{\beta}}\right)$$

ここで，$\boldsymbol{V}(\cdot)$ は分散共分散行列で，不等号は 2 次形式に関する大小関係を表す．

4 数値実験

上記の推定方法を実際のデータに適用し，従来の推定方法との比較を行った．数値実験の結果はシンポジウム当日に発表する．

参考文献

[1] 鎌谷直之．遺伝統計学入門, 岩波書店, 2007.

[2] 上辻茂男．”ゲノムワイド関連研究に学ぶ遺伝統計学”, 計算機統計学，25 巻，第 1 号, 17-39 (2012).

[3] 中村永友．多次元データ解析法 (R で学ぶデータサイエンス 2), 共立出版, 2010.

[4] 柳井晴夫，竹内啓．射影行列・一般逆行列・特異値分解, 東京大学出版, 1983.