二項回帰モデルの不均衡極限と変形指数型分布族

慶應義塾大学　清 智也

1　設定

$\{(X_i, Y_i)\}_{i=1}^m$ を独立な $\mathbb{R}^p \times \{0,1\}$ 値の確率変数とし，X_i の周辺分布は $F(dx)$，Y_i の条件付き分布は**二項回帰モデル**

$$P(Y_i = 1 \mid X_i, a_m, b_m) = G(a_m + b_m' X_i), \quad (a_m, b_m) \in \mathbb{R} \times \mathbb{R}^p, \quad i \in \{1, \cdots, m\}, \qquad (1)$$

に従うとする．ただし G は $\mathbb{R}$ 上の累積分布関数である．逆関数 G^{-1} はリンク関数である．

ここでは真のパラメータ (a_m, b_m) がサンプルサイズ m に依存し，m が増大するとともに式 (1) の確率が 0 に近づく状況を考える．その極限を**不均衡極限**ということにする．

2　主結果

極値理論によれば，多くの分布関数 G は次の仮定を満たす（[3], Theorem 1.1.2, 1.1.3）．

仮定　ある実数 q と数列 $c_m \in \mathbb{R}$，$d_m > 0$ が存在して，任意の $z \in \mathbb{R}$ に対して

$$G(c_m + d_m z) = \frac{1}{m} \exp_q(z) + o(m^{-1}), \quad m \to \infty,$$

が成り立つ．ただし $\exp_q(z) = \max(1 + (1-q)z, 0)^{1/(1-q)}$（$q \neq 1$），$\exp_1(z) = e^z$ とする．

この仮定の下で，二項回帰モデルの不均衡極限は次のようになる（[7]）．

定理　真のパラメータが $a_m = c_m + d_m \alpha$，$b_m = d_m \beta$，$(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^p$，で与えられるとき，集合 $\{X_i \mid 1 \leq i \leq m,\ Y_i = 1\}$ は $m \to \infty$ の下でポアソン点過程に分布収束し，その強度は

$$\lambda(dx) = \exp_q(\alpha + \beta' x) F(dx), \quad x \in \mathbb{R}^p, \qquad (2)$$

となる．

式 (2) を**変形指数型分布族**またはアルファ分布族（アルファ$= 2q - 1$）という（[5], [1]）．

例　G^{-1} が logit，probit，complementary log-log リンクの場合は $q = 1$ である（[2], [6], [8]）．また，G がコーシー分布の場合は $q = 2$，一様分布の場合は $q = 0$ である．

例　t-logistic 回帰は $G(z) = \exp_t(z - \gamma_t(z))$，$\sum_{y \in \{0,1\}} \exp_t(yz - \gamma_t(z)) = 1$ で定義される（[4]）．この場合，$q = \max(t, 0)$ である．

参考文献

[1] Amari, S. (1985). *Differential-geometrical methods in statistics*, Springer.

[2] Baddeley, A. et al. (2010). *Electron. J. Statist.*, **4**, 1151–1201.

[3] de Haan, L. & Ferreira, A. (2006). *Extreme value theory, an introduction*, Springer.

[4] Ding, N. et al. (2011). *J. Mach. Learn. Res.*, **12**, 1–55.

[5] Naudts, J. (2002). *Physica A*, 316, 323–334.

[6] Owen, A. B. (2007). *J. Mach. Learn. Res.*, **8**, 761–773.

[7] Sei, T. (2013). preprint, arXiv:1303.7297.

[8] Warton, D. I. & Shepherd, L. C. (2010). *Ann. Appl. Statist.*, **4**, 1383–1402.