Probability-Based Optimal Design with Separation Problems

Mohammad Lutfor Rahman,Queen Mary, University of London and
Steven G. Gilmour, University of Southampton

Queen Mary University of London

Introduction

Separation is a common problem in models with binary responses when one or more co-variates perfectly predicts some binary outcome. The separation problem leads convergence difficulties as well as non-existence of likelihood estimates of model parameters. Some methods have been proposed in literature to deal with the separation problems. In this endevour we propose a new probability-based optimality criterion (P_s) that would reduce the problem of separation and thus maximize the probability of the existence of likelihood estimates.

Types of Separation

Albert and Anderson (1984) classify logistic regression data sets into three mutually exclusive and exhaustive categories: complete separation, quasi-complete separation, and overlap. The maximum likelihood estimates exists only for overlapped data.

Complete separation occurs whenever there exists some vector of coefficients α such that $Y_i = 1$ if $\alpha \boldsymbol{X_i} > \mathbf{0}$ and $Y_i = 0$ if $\alpha \boldsymbol{X_i} \leq \mathbf{0}$. In other words, complete separation occurs whenever a linear function of $\mathbf{X}$ can generate perfect predictions of Y.

Quasi-complete separation occurs when there exists some coefficient vector α such that $Y_i = 1$ if $\alpha \boldsymbol{X_i} \geq \mathbf{0}$ and $Y_i = 0$ if $\alpha \boldsymbol{X_i} \leq \mathbf{0}$, and equality holds for at least one case in each category of the dependent variable.

Data which are neither completely or quasi-completely separated are called overlapped. A certain degree of overlap is a necessary and sufficient condition for the existence of maximum likelihood estimates for the binomial response (Silvapulle, 1981).

Hypothetical Example of Separation and Existing Solutions of Separation

Consider two hypothetical examples of separation. The complete separation is shown in Table 1 and Table 2 displays a data set regarding quasi-complete separation. What distinguishes the data set in Table 2 is that there are two additional observations, each with x values of 0 but having different values of y.

Table: Data exhibiting complete separation

Y	X
0	-5
0	-4
0	-3
0	-2
0	-1
1	1
1	2
1	3
1	4
1	5

Table: Data exhibiting quasi-complete separation

Y	X
0	-5
0	-4
0	-3
0	-2
0	-1
0	0
1	0
1	1
1	2
1	3
1	4
1	5

Quasi-complete separation can be dealt with by data analytic methods such as deletion of the problem variable, combining categories, reporting likelihood ratio chi-square, Bayesian estimation, and penalized likelihood estimation, but it is difficult to deal with complete separation (Allison, 2008). Having some pitfalls in all the existing methods, we propose optimal design techniques to reduce the probability of separation problem in this study.

Notations: Separation Probability

Let Y_i be a response corresponding to ith observation and associated covariate is X_i , $i = 1, 2, \ldots, n$. The distribution of Y_i is a Bernoulli distribution with probability π_i. In matrix notations

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

For simplicity in defining separation probability let us describe the Table 3 with Y_i and X_i values. In the X_i column there is a cut off point X^* which separates the outcome Y into two groups namely 'success' (S_i) defined by $Y_i = 1|X_i$ and 'failure' (F_i) defined by $Y_i = 0|X_i$.

Table: Design Matrix

Y_i	X_i
Y_1	X_1
Y_2	X_2
$\vdots$	$\vdots$
Y_k	X_k
Y_{k+1}	X_{k+1}
$\vdots$	$\vdots$
Y_n	X_n

$$Y_i = \begin{cases} 0 \text{ if } X_i \leq X^*, & \text{i=1, 2, \ldots k} \\ 1 \text{ if } X_i > X^*, & \text{i=k+1, k+2, \ldots n} \end{cases} \tag{1}$$

Probability of Complete and Quasi-complete Separations

The probability of complete separation would be

$$\begin{aligned} \text{P(Complete Separation)} &= \sum_{k=1}^{n-1} P\left[\left(F_1 \bigcap F_2 \bigcap \ldots \bigcap F_k \bigcap S_{k+1} \bigcap \ldots \bigcap S_n\right) \bigcup \right. \\ &\left. \left(S_1 \bigcap S_2 \bigcap \ldots \bigcap S_k \bigcap F_{k+1} \bigcap \ldots \bigcap F_n\right)\right] \\ &= \sum_{k=1}^{n-1} \left[\prod_{i=1}^{k} P(F_i) \prod_{i=k+1}^{n} P(S_i) + \prod_{i=1}^{k} P(S_i) \prod_{i=k+1}^{n} P(F_i)\right] \end{aligned} \tag{2}$$

The conditions described in 1 should be modified slightly for quasi-complete separation. In this situation equality holds for at least one for each of the categories i.e.

$$Y_i = \begin{cases} 0 & \text{if } X_i < X^*, \quad \text{i=1, 2, \ldots k-1} \\ 0 \quad \text{or} \quad 1 & \text{if } X_i = X^*, \quad \text{i=k, k+1} \\ 1 & \text{if } X_i > X^*, \quad \text{i=k+2, k+3, \ldots n} \end{cases} \tag{3}$$

Probability of Complete and Quasi-complete Separations

The probability of quasi-complete separation would be

$$\begin{aligned} \text{P(QCS)} = & \left[\{P(F_1)P(S_2) + P(S_1)P(F_2)\}\left\{\prod_{i=3}^{n} P(S_i) + \prod_{i=3}^{n} P(F_i)\right\}\right] \\ & + \sum_{k=2}^{n-2} [\{P(F_k)P(S_{k+1}) + P(S_k)P(F_{k+1})\} \\ & \left\{\prod_{i=1}^{k-1} P(F_i) \prod_{i=k+2}^{n} P(S_i) + \prod_{i=1}^{k-1} P(S_i) \prod_{i=k+2}^{n} P(F_i)\right\}] \\ & + \left[\{P(F_{n-1})P(S_n) + P(S_{n-1})P(F_n)\}\left\{\prod_{i=3}^{n-2} P(F_i) + \prod_{i=3}^{n-2} P(S_i)\right\}\right] \end{aligned} \tag{4}$$

Lemma and Numerical Illustrations

Lemma

Given a design $X_1 \leq X_2 \leq \ldots \leq X_n$ with $X_k = X_{k+1} = X^*$ for some k, the probability of separation is reduced by changing X_k and X_{k+1} to $X_k^* = X^* - \delta$ and $X_{k+1}^* = X_{k+1} + \delta$ for a small $\delta > 0$

The followings are two numerical illustrations supporting this lemma.

Table: Probability of Complete Separation at $X_3 = X^*$

Y_i	X_i	X_i^*.
0	-1	-1
0	0	-0.05
0	0	0.05
1	1	1
1	2	2
1	3	3
P(CS)	0.11220	0.11213

Table: Probability Complete and Quasi-complete Separations at $X_3 = X^*$

Y_i	X_i	X_i^*.
0	-2	-2
0	-1	-1
0	0	-0.01
1	0	0.01
1	1	1
1	2	2
P(CS)	-	0.10495
P(QCS)	0.20783	-

Probability-based Optimality Criteria

The general form of proposed P_s-optimality criterion, analogous to McGree and Eccleston (2008), could be

$$\psi^{(P_s)} = \text{f(P(Separation))= f}(F_i, S_i) \tag{5}$$

where

$$\text{P}(\text{S}_i) = \text{P}(\text{Y}_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \quad \text{and} \quad \text{P}(\text{F}_i) = 1 - \text{P}(\text{S}_i) \tag{6}$$

A combined criterion can be developed considering both D- and P_s- optimality and, to achieve the dual goals of efficient parameter estimation and reducing probability of separation, a compromise is necessary. Therefore, a compound criterion may be defined as a product of efficiencies of design ξ with respect to D- and P_s-optimality, weighted by a pre-defined mixing constant $0 \leq \rho \leq 1$.

$$\psi^{(DP_s)} = [D_{eff}(\xi)]^{\rho}[P_{seff}(\xi)]^{1-\rho} \tag{7}$$

where

$$D_{eff}(\xi) = \left(\frac{|M(\xi)|}{|M(\xi_D^*)|}\right)^{1/q} \quad \text{and} \quad P_{seff}(\xi) = \left(\frac{P(\xi)}{P(\xi_{P_s}^*)}\right) \tag{8}$$

Conclusions and Future Work

- We have defined probability of complete and quasi-complete separations
- We have proposed probability-based criteria that might minimize separation problems
- We will check the existence of maximum likelihood estimates (MLEs) in designed experiments
- We will see the possible configuration of response vector and MLEs existence with a given covariate vector
- We will find local optimal designs with respect to proposed probability-based criteria
- We will find global optimal designs with respect to proposed probability-based criteria

References

- Albert, A. and Anderson, J. A. (1984) "On the existence of maximum likelihood estimates in logistic regression models", Biometrika, **71**, 1-10.
- Silvapulle, M. J. (1981). "On the existence of maximum likelihood estimates for the binomial response models", Journal of Royal Statistical Society, Series B, **43**, 310-13
- McGree, J. M. and Eccleston, J. A. (2008). "Probability-based optimal designs", Australian Newzealand Journal of Statistics, **50**, 13-28.
- Allison, P. D. (2008). "Convergence failures in logistic regression", SAS Global Forum, Paper 360-2008, http://www2.sas.com/proceedings/forum2008/360-2008.pdf
- Hamada, M. and Tse, S. K. (1996). "The existence of maximum likelihood estiamtes from designed experiments", Journal of Quality Technology, **28**, 244-254.

DAE Conference 2012, The University of Georgia, Athens, USA

m.rahman@qmul.ac.uk/ s.g.gilmour@qmul.ac.uk