

# ***Infinitely Imbalanced Logistic Regression***

Art B. Owen

Journal of Machine Learning Research, April 2007

Presenter: Ivo D. Shterev

# *Outline*

- Motivation
- Introduction
- Numerical Examples
- Notation
- Silvapulle's Results
- Overlap Conditions
- Technical Lemmas
- Main Results
- Example

# *Motivation*

- Binary classification problems.
- Two imbalanced classes - one class is very rare compared to the other.
- Applications include fraud detection, drug discovery, modeling of rare events in political sciences, etc.

# Introduction

- We have the observation

$$Y \in \{0, 1\}$$

- where 0 is in the common case, and 1 is in the rare case.
- The limiting logistic regression coefficient  $\beta(N)$  satisfies

$$\bar{x} = \frac{\int \exp(x' \beta) x dF_0(x)}{\int \exp(x' \beta) dF_0(x)}$$

- where  $F_0$  is the distribution of  $X$  given  $Y = 0$ ,  $\bar{x}$  is the average of the sample values  $x_i$  for which  $y = 1$ .
- when  $F_0 = \mathcal{N}(\mu_0, \Sigma_0)$ , then

$$\lim_{N \rightarrow \infty} \beta(N) = \Sigma_0^{-1}(\bar{x} - \mu_0)$$

## Numerical examples

- Suppose there are  $N$  observations  $(X|Y = 0) \sim \mathcal{N}(0, 1)$ , and a single observation  $(x|y = 1) = 1$ . The logistic regression for this case is

$$(x_i, y_i) = \left( \Phi^{-1}\left(\frac{i - \frac{1}{2}}{N}\right), 0 \right), \text{ for } i = 1, \dots, N$$

$$(x_{N+1}, y_{N+1}) = (1, 1)$$

- where  $\Phi(\cdot)$  is the cumulative distribution function of  $X$ .
- as  $N$  increases the problem becomes more imbalanced.

# Numerical Examples

N	$\alpha$	$Ne^\alpha$	$\beta$
10	-3.19	0.4126	1.5746
100	-5.15	0.5787	1.0706
1,000	-7.42	0.6019	1.0108
10,000	-9.71	0.6058	1.0017
100,000	-12.01	0.6064	1.0003

Table 1: Logistic regression intercept  $\alpha$  and coefficient  $\beta$  for imbalanced data described in the text. There are  $N$  observations with  $Y = 0$  and stratified  $X \sim N(0, 1)$  and one observation with  $Y = 1$  and  $X = 1$ .

● it seems that

$$\lim_{N \rightarrow \infty} \alpha = -\log N$$

$$\lim_{N \rightarrow \infty} \beta = 1$$

## Numerical Examples

N	$\alpha$	$Ne^\alpha$	$\beta$	$Ne^\beta$
10	-2.36	0.94100	0.1222260	1.2222
100	-4.60	0.99524	0.0097523	0.9752
1,000	-6.90	0.99953	0.0009537	0.9536
10,000	-9.21	0.99995	0.0000952	0.9515
100,000	-11.51	0.99999	0.0000095	0.9513

Table 2: Logistic regression intercept  $\alpha$  and coefficient  $\beta$  for imbalanced data described in the text. There are  $N$  observations with  $Y = 0$  and stratified  $X$  from the standard Cauchy distribution, and one observation with  $Y = 1$  and  $X = 1$ .

● it seems that

$$\alpha(N) = \text{const} - \log N$$

$$\lim_{N \rightarrow \infty} \beta = 0$$

## Numerical Examples

N	$\alpha$	$Ne^\alpha$	$\beta$	$e^\beta/N$
10	-3.82	0.2184	2.85	1.74
100	-7.13	0.0804	4.19	0.66
1,000	-10.71	0.0223	5.82	0.34
10,000	-14.52	0.0050	7.62	0.20
100,000	-18.49	0.0009	9.54	0.14

Table 3: Logistic regression intercept  $\alpha$  and coefficient  $\beta$  for imbalanced data described in the text. There are  $N$  observations with  $Y = 0$  and stratified  $X \sim U(0, 1)$  and two observations with  $Y = 1$ , one with  $X = 1/2$ , the other with  $X = 2$ .

- it seems that  $\beta$  does not converge to a useful limit.

# Notation

- Data  $(\mathbf{x}, y)$ , where  $\mathbf{x} \in \mathbb{R}^d$  and  $y \in \{0, 1\}$ . Observations  $n$  with  $y = 1$ , and  $N$  with  $y = 0$ , where  $n \ll N$ . Denote  $(\mathbf{x}|y = 1) = (\mathbf{x}_{11}, \dots, \mathbf{x}_{1n})$ ,  $(\mathbf{x}|y = 0) = (\mathbf{x}_{01}, \dots, \mathbf{x}_{0N})$ .
- the logistic regression model is

$$Pr(Y = 1 | \mathbf{X} = \mathbf{x}) = \frac{\exp(\alpha + \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x}'\boldsymbol{\beta})}$$

- where  $\alpha \in \mathbb{R}$  and  $\boldsymbol{\beta} \in \mathbb{R}^d$ . The log-likelihood in logistic regression is

$$l(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^n \left( \alpha + \mathbf{x}'_{1i}\boldsymbol{\beta} - \log(1 + \exp(\alpha + \mathbf{x}'_{1i}\boldsymbol{\beta})) \right) - \sum_{i=1}^N \log(1 + \exp(\alpha + \mathbf{x}'_{0i}\boldsymbol{\beta}))$$

## Notation

- centering the logistic regression around  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{1i}$ , gives

$$\begin{aligned} l(\alpha, \boldsymbol{\beta}) &= n\alpha + \sum_{i=1}^n (\mathbf{x}_{1i} - \bar{\mathbf{x}})' \boldsymbol{\beta} \\ &\quad - \sum_{i=1}^n \log \left( 1 + \exp \left( \alpha + (\mathbf{x}_{1i} - \bar{\mathbf{x}})' \boldsymbol{\beta} \right) \right) \\ &\quad - N \int \log \left( 1 + \exp \left( \alpha + (\mathbf{x} - \bar{\mathbf{x}})' \boldsymbol{\beta} \right) \right) dF_0(\mathbf{x}) \end{aligned}$$

- the study is focussed on the maximum likelihood estimate (MLE)  $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ .
- the centered  $\hat{\alpha}_0 = \hat{\alpha} + \bar{\mathbf{x}}' \hat{\boldsymbol{\beta}}$ , while  $\hat{\boldsymbol{\beta}}$  stays the same.
- $\lim_{N \rightarrow \infty} \hat{\alpha} = -\infty$ , but  $\hat{\boldsymbol{\beta}}$  does not necessarily diverge.

# Silvapulle's Results

- Theorem 1: For  $y = 1$  let  $\mathbf{z}_{1i} = (1, \mathbf{x}'_{1i})'$  for  $i = 1, \dots, n_1$ . For  $y = 0$  let  $\mathbf{z}_{0i} = (1, \mathbf{x}'_{0i})'$  for  $i = 1, \dots, n_0$ . Let  $\theta = (\alpha, \beta')$ .
- then the logistic regression model has

$$Pr(Y = y | \mathbf{X} = \mathbf{x}) = \frac{\exp(\mathbf{z}'\theta)}{1 + \exp(\mathbf{z}'\theta)}$$

- employ two convex cones

$$C_j = \sum_{i=1}^{n_j} k_{ji} \mathbf{z}_{ji} | k_{ji} > 0, j \in \{0, 1\}.$$

- assume that the  $n_0 + n_1$  by  $d + 1$  matrix with rows taken from  $\mathbf{z}_{ji}$ , has rank  $d + 1$ . Iff  $C_0 \cap C_1 \neq \emptyset$ , then a unique finite MLE  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}')$  exists.

# Silvapulle's Results

- Lemma 2: Define the convex hull of the  $\mathbf{x}$ 's for  $y = 0$  and  $y = 1$

$$H_j = \sum_{i=1}^{n_j} \lambda_{ji} \mathbf{x}_{ji} \mid \lambda_{ji} > 0, \sum_{i=1}^{n_j} \lambda_{ji} = 1$$

- $C_0 \cap C_1 \neq \emptyset$  is equivalent to  $H_0 \cap H_1 \neq \emptyset$ .
- Proof: Suppose that  $\mathbf{x}_0 \in H_0 \cap H_1$ , then  $\mathbf{z}_0 = (1, \mathbf{x}_0')' \in C_0 \cap C_1$ . Conversely, suppose that  $\mathbf{z}_0 \in C_0 \cap C_1$ . Then we can write

$$\mathbf{z}_0 = \sum_{i=1}^{n_0} k_{0i} \begin{pmatrix} 1 \\ \mathbf{x}_{0i} \end{pmatrix} = \sum_{i=1}^{n_1} k_{1i} \begin{pmatrix} 1 \\ \mathbf{x}_{1i} \end{pmatrix}$$

- find  $K$ , the common positive value for  $\sum_{i=1}^{n_0} k_{0i}$  and  $\sum_{i=1}^{n_1} k_{1i}$ .  
Choose  $\lambda_{ji} = k_{ji}/K$ , then  
 $\mathbf{x}_0 = \sum_{i=1}^{n_0} \lambda_{0i} \mathbf{x}_{0i} = \sum_{i=1}^{n_1} \lambda_{1i} \mathbf{x}_{1i} \in H_0 \cap H_1$ .

# Overlap Conditions

- Assume some overlap between  $x_1$  and the distribution  $F_0$ , to get interesting results.
- let  $\Omega = \{\omega \in \mathbb{R}^d \mid \omega' \omega = 1\}$  be the unit sphere in  $\mathbb{R}^d$ .
- Definition 3:  $F$  on  $\mathbb{R}^d$  has the point  $x_*$  surrounded if

$$\int_{(x-x_*)' \omega > \epsilon} dF(x) > \delta$$

holds for some  $\epsilon > 0$ , some  $\delta > 0$  and all  $\omega \in \Omega$ .

- If  $F$  has the point  $x_*$  surrounded, then there exist  $\eta$  and  $\gamma$  satisfying

$$\inf_{\omega \in \Omega} \int_{(x-x_*)' \omega > 0} dF(x) \geq \eta > 0$$

# Overlap Conditions

• and

$$\inf_{\omega \in \Omega} \int [(\mathbf{x} - \mathbf{x}_*)' \omega]_+ dF(\mathbf{x}) \geq \gamma > 0$$

- where  $Z_+ = \max(Z, 0)$ .
- For Theorem 1 (Lemma 2) to hold, is it sufficient that there is some point  $\mathbf{x}_*$  that is surrounded by both  $F_0$  and  $F_1$ .
- In the infinitely imbalanced case it is expected that  $F_0$  will surround almost all  $\mathbf{x}$ , but it is not required. It is not sufficient that  $F_0$  surrounds only one  $\mathbf{x}_*$ , but it is sufficient that  $F_0$  surrounds  $\bar{\mathbf{x}}$ .
- It is not necessary that  $F_1$  surrounds  $\bar{\mathbf{x}}$ .

## Technical Lemmas

- Lemma 4: For  $\alpha, z \in \mathbb{R}$

$$\begin{aligned}\exp(\alpha + z) &\geq \log(1 + \exp(\alpha + z)) \\ &\geq \left[ \log(1 + \exp(\alpha)) + \frac{z \exp(\alpha)}{1 + \exp(\alpha)} \right]_+ \\ &\geq \left[ \frac{z \exp(\alpha)}{1 + \exp(\alpha)} \right]_+ \\ &= \frac{z_+ \exp(\alpha)}{1 + \exp(\alpha)}\end{aligned}$$

- Lemma 5: Let  $n \geq 1$  and  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  be given. Assume that  $F_0$  surrounds  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{1i}$  and that  $0 < N < \infty$ . Then  $l(\alpha, \beta)$  has a unique maximizer  $(\hat{\alpha}, \hat{\beta})$ .

# Main Results

- Lemma 6: Under the conditions of Lemma 5, let  $\hat{\alpha}$  and  $\hat{\beta}$  maximize  $l$ . Let  $\eta$  satisfy  $\int_{(\mathbf{x}-\mathbf{x}_*)'\boldsymbol{\omega}>\epsilon} dF(\mathbf{x}) \geq \eta > 0$ . Then, for  $N \geq \frac{2n}{\eta}$  we have  $\exp(\hat{\alpha}) \leq \frac{2n}{N\eta}$ .
- Lemma 7: Under the conditions of Lemma 5, let  $\hat{\alpha}$  and  $\hat{\beta}$  maximize  $l$ . Then  $\limsup_{N \rightarrow \infty} \|\hat{\beta}\| < \infty$ .
- Theorem 8: Let  $n \geq 1$  and  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  and suppose that  $F_0$  satisfies the tail condition  $\int \exp(\mathbf{x}'\boldsymbol{\beta})(1 + \|\mathbf{x}\|)dF_0(\mathbf{x}) < \infty$  for  $\forall \boldsymbol{\beta} \in \mathbb{R}^d$  ( $F_0$  has too heavy tails), and surrounds  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ . Then the maximizer  $(\hat{\alpha}, \hat{\beta})$  of  $l$  satisfies

$$\lim_{N \rightarrow \infty} \frac{\int \exp(\mathbf{x}'\hat{\boldsymbol{\beta}})\mathbf{x}dF_0(\mathbf{x})}{\int \exp(\mathbf{x}'\hat{\boldsymbol{\beta}})dF_0(\mathbf{x})} = \bar{\mathbf{x}}$$

## Main Results

- In the limit  $N \rightarrow \infty$  the logistic regression ( $\hat{\beta}$ ) depends on the  $\mathbf{x}_1, \dots, \mathbf{x}_n$  only through  $\bar{\mathbf{x}}$ .

Method	$\alpha$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$
Original	-3.707	4.629	4.807	0.398	0.594	0.170	0.130
Single $y = 1$	-10.116	4.623	4.984	0.397	0.595	0.193	0.182
$x_{1j} = \bar{x}$	-3.701	4.765	5.136	0.410	0.614	0.204	0.190
SE	0.041	0.696	0.851	0.040	0.130	0.299	0.413

Table 4: This table shows logistic regression coefficients for the chemical compound data set described in the text. The top row shows ordinary logistic regression coefficients. The second row shows the coefficients when the cases with  $y = 1$  are deleted and replaced by a single point  $(\bar{x}, 1)$ . The third row shows the coefficients when all 608 cases with  $y = 1$  are replaced by  $(\bar{x}, 1)$ . The fourth row shows standard errors for the ordinary logistic regression coefficients in the top row.

## Example

- Suppose  $F_0$  is a Gaussian mixture

$$F_0 = \sum_{k=1}^K \lambda_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \lambda_k > 0 \text{ and } \sum_{k=1}^K \lambda_k = 1$$

- if at least one of  $\boldsymbol{\Sigma}_k$  has full rank, then  $F_0$  will surround the point  $\bar{\boldsymbol{x}}$  and the solution to  $\boldsymbol{\beta}$  is defined through

$$\bar{\boldsymbol{x}} = \frac{\sum_{k=1}^K \lambda_k (\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \boldsymbol{\beta}) \exp(\boldsymbol{\beta}' \boldsymbol{\mu}_k + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_k \boldsymbol{\beta})}{\sum_{k=1}^K \lambda_k \exp(\boldsymbol{\beta}' \boldsymbol{\mu}_k + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_k \boldsymbol{\beta})}$$

- or

$$\mathbf{0} = \sum_{k=1}^K \lambda_k (\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \boldsymbol{\beta} - \bar{\boldsymbol{x}}) \exp(\boldsymbol{\beta}' \boldsymbol{\mu}_k + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_k \boldsymbol{\beta})$$

## *Example*

- Solving the equation can be done by convex optimization, using Newton's method with  $\mathcal{O}(d^3)$  computations per iteration.