Infinitely Imbalanced Logistic Regression

Art B. Owen

Stanford University
owen@stat.stanford.edu

12th annual winter workshop
January 2010
University of Florida
Imbalanced data

Setting:

- Data are \((X, Y)\) pairs,
Imbalanced data

Setting:

- Data are \((X, Y)\) pairs,
- Predictors \(X \in \mathbb{R}^d\)

Examples,

- for: active drug ad gets clicked, rare disease, war/coup/veto, citizen seeks elected office, non-spam in spam bucket
Imbalanced data

Setting:

- Data are \((X, Y)\) pairs,
- Predictors \(X \in \mathbb{R}^d\)
- Binary response variable \(Y \in \{0, 1\}\)
Imbalanced data

Setting:

- Data are \((X, Y)\) pairs,
- Predictors \(X \in \mathbb{R}^d\)
- Binary response variable \(Y \in \{0, 1\}\)
- Sample has lots of \(Y = 0\),
Imbalanced data

Setting:

- Data are \((X, Y)\) pairs,
- Predictors \(X \in \mathbb{R}^d\)
- Binary response variable \(Y \in \{0, 1\}\)
- Sample has lots of \(Y = 0\), very few \(Y = 1\)

Examples,
- Active drug ad gets clicked
- Rare disease
- War/coup/veto
- Citizen seeks elected office
- Non-spam in spam bucket
Imbalanced data

Setting:

- Data are \((X, Y)\) pairs,
- Predictors \(X \in \mathbb{R}^d\)
- Binary response variable \(Y \in \{0, 1\}\)
- Sample has lots of \(Y = 0\), very few \(Y = 1\)

Examples, \(Y = 1\) for:

- active drug
- ad gets clicked
- rare disease
- war/coup/veto
- citizen seeks elected office
- non-spam in spam bucket
(Why) does imbalance matter?

Irony:

500 1s and 500 0s $\implies$ OK
500 1s and 500,000 0s $\implies$ trouble
(Why) does imbalance matter?

Ironic:

500 1s and 500 0s $\implies$ OK
500 1s and 500,000 0s $\implies$ trouble

Issues:

1. It is hard to beat the rule that predicts $Y = 0$ always
2. Few $Y = 1$ cases constitute a low effective sample size
(Why) does imbalance matter?

Irony:

\[ 500 \text{ 1s and 500 0s} \implies \text{OK} \]
\[ 500 \text{ 1s and 500,000 0s} \implies \text{trouble} \]

Issues:

1. It is hard to beat the rule that predicts \( Y = 0 \) always
2. Few \( Y = 1 \) cases constitute a low effective sample size

Approaches:

1. So take account of priors and/or loss asymmetry (assuming implicit/explicit probability estimates)
2. Effective sample size really is \# of \( Y = 1 \)s
How to deal with imbalanced data:

Coping strategies:

1. Downsample the 0s (adjust prior accordingly)
2. Upsample the 1s:
   - Repeat some (or upweight them)
   - Add synthetic 1s
3. One class prob.: find small ellipsoid holding the $x_i$ for $y_i = 1$
How to deal with imbalanced data:

Coping strategies:

1. Downsampling the 0s (adjust prior accordingly)
2. Upsampling the 1s:
   - Repeat some (or upweight them)
   - Add synthetic 1s
3. One class prob.: find small ellipsoid holding the $x_i$ for $y_i = 1$

Workshops on imbalanced data:

- AAAI 2000
- ICML 2003

They prefer “imbalanced” to “unbalanced”
Is it even a problem?

Suppose data are

For $y = 1$: \[ x_{1i}, \quad i = 1, \ldots, n_1 \equiv n \]

For $y = 0$: \[ x_{0i}, \quad i = 1, \ldots, n_0 \equiv N \quad N \gg n \]

Fit logistic regression

\[
Pr(Y = 1 | X = x) = \frac{e^{\alpha + x' \beta}}{1 + e^{\alpha + x' \beta}}
\]

Let $N \to \infty$ with $n$ fixed

Expect $\hat{\alpha} \to -\infty$ like $-\log(N)$

But $\hat{\beta}$ can have a useful limit

and $\hat{\beta}$ is of most interest

$N/n \to \infty$ not necessarily so bad (for logistic regression).
Is it even a problem?

Suppose data are

For $y = 1$: $x_{1i}, i = 1, \ldots, n_1 \equiv n$
For $y = 0$: $x_{0i}, i = 1, \ldots, n_0 \equiv N \quad N \gg n$

Fit logistic regression

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + x'\beta}}{1 + e^{\alpha + x'\beta}}$$
Is it even a problem?

Suppose data are

For \( y = 1 \): \( x_{1i}, \; i = 1, \ldots, n_1 \equiv n \)
For \( y = 0 \): \( x_{0i}, \; i = 1, \ldots, n_0 \equiv N \quad N \gg n \)

Fit logistic regression

\[
\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + x'\beta}}{1 + e^{\alpha + x'\beta}}
\]

Let \( N \to \infty \) with \( n \) fixed

Expect \( \hat{\alpha} \to -\infty \) like \( -\log(N) \)
But \( \hat{\beta} \) can have a useful limit
and \( \hat{\beta} \) is of most interest
Is it even a problem?

Suppose data are

For $y = 1$: $x_{1i}$, $i = 1, \ldots, n_1 \equiv n$
For $y = 0$: $x_{0i}$, $i = 1, \ldots, n_0 \equiv N \quad N \gg n$

Fit logistic regression

$$Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + x'\beta}}{1 + e^{\alpha + x'\beta}}$$

Let $N \to \infty$ with $n$ fixed

Expect $\hat{\alpha} \to -\infty$ like $-\log(N)$

But $\hat{\beta}$ can have a useful limit

and $\hat{\beta}$ is of most interest

$N/n \to \infty$ not necessarily so bad (for logistic regression).
Main result

Suppose

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_1 i \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0$$

Let $\alpha(N)$ and $\beta(N)$ be logistic regression estimates
Main result

Suppose

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_{1i} \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0 \]

Let \( \alpha(N) \) and \( \beta(N) \) be logistic regression estimates

Under mild conditions

\[ N e^{\alpha(N)} \rightarrow A \in \mathbb{R} \quad \text{and} \quad \beta(N) \rightarrow \beta \in \mathbb{R}^d \]
Main result

Suppose

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_{1i} \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0 \]

Let \( \alpha(N) \) and \( \beta(N) \) be logistic regression estimates

Under mild conditions

\[ Ne^{\alpha(N)} \rightarrow A \in \mathbb{R} \quad \text{and} \quad \beta(N) \rightarrow \beta \in \mathbb{R}^d \]

where \( \beta \) solves

\[ \bar{x} = \frac{\int x e^{x'\beta} dF_0(x)}{\int e^{x'\beta} dF_0(x)} \]
Interpretation

We have

\[ \bar{x} = \frac{\int x e^{x' \beta} dF_0(x)}{\int e^{x' \beta} dF_0(x)} \]

\( \beta \) is the \textit{exponential tilt} to take \( E_{F_0}(X) \) onto \( \bar{x} \)
Interpretation

We have

\[ \bar{x} = \frac{\int x e^{x' \beta} dF_0(x)}{\int e^{x' \beta} dF_0(x)} \]

\( \beta \) is the \textit{exponential tilt} to take \( E_{F_0}(X) \) onto \( \bar{x} \)

For \( F_0 = N(\mu_0, \Sigma_0) \)

\[ \beta = \Sigma_0^{-1}(\bar{x} - \mu_0) \]
Interpretation

We have

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

$\beta$ is the \textit{exponential tilt} to take $E_{F_0}(X)$ onto $\bar{x}$

For $F_0 = N(\mu_0, \Sigma_0)$

$$\beta = \Sigma_0^{-1}(\bar{x} - \mu_0)$$

Compare

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0) \text{ for } X \sim N(\mu_j, \Sigma) \text{ given } Y = j \in \{0, 1\}$$
Suppose $\beta$ solves

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

Then only $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $F_0$ matter. Clearly $n$ is the effective sample size.
Suppose $\beta$ solves

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

Then only $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $F_0$ matter

Clearly $n$ is the effective sample size

We could:

replace $(x_{1i}, 1)$ for $i = 1, \ldots, n$

by just one point $(X, Y) = (\bar{x}, 1)$

and get the same $\beta$ as $N \to \infty$
Suppose $\beta$ solves

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

Then only $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $F_0$ matter

Clearly $n$ is the effective sample size

We could:

replace $(x_{1i}, 1)$ for $i = 1, \ldots, n$

by just one point $(X, Y) = (\bar{x}, 1)$

and get the same $\beta$ as $N \to \infty$

Upshot:

IILR downsamples the rare case to a single point

Whether logistic works well or badly on given problem

Other classifiers (e.g. CART) would be different
Uses

The predictions are trivial

\[ \Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d \]
Uses

The predictions are trivial

\[
\Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d
\]

But ratios are informative and simple

\[
\frac{\Pr(\tilde{Y} = 1 \mid X = \tilde{x})}{\Pr(Y = 1 \mid X = x)} \to e^{(\tilde{x} - x)'}\beta
\]
Uses

The predictions are trivial

\[ \Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d \]

But ratios are informative and simple

\[ \frac{\Pr(\tilde{Y} = 1 \mid X = \tilde{x})}{\Pr(Y = 1 \mid X = x)} \to e^{(\tilde{x} - x)'\beta} \]

For fraud or active learning, obtain \( Y \) corresponding to largest

\[ e^{x'\beta} \quad \text{(best chance to see a 1)} \]
Uses

The predictions are trivial

\[ \Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d \]

But ratios are informative and simple

\[
\frac{\Pr(\tilde{Y} = 1 \mid X = \tilde{x})}{\Pr(Y = 1 \mid X = x)} \to e^{(\tilde{x} - x)'\beta}
\]

For fraud or active learning, obtain \( Y \) corresponding to largest

- \( e^{x'\beta} \) (best chance to see a 1)
- \( v e^{x'\beta} \) (when case has value \( v \))
Uses

The predictions are trivial

\[ \Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all } x \in \mathbb{R}^d \]

But ratios are informative and simple

\[ \frac{\Pr(\tilde{Y} = 1 \mid X = \tilde{x})}{\Pr(Y = 1 \mid X = x)} \to e^{(\tilde{x} - x)'\beta} \]

For fraud or active learning, obtain \( Y \) corresponding to largest

- \( e^{x'\beta} \) (best chance to see a 1)
- \( v e^{x'\beta} \) (when case has value \( v \))
- \( v e^{x'\beta} / c \) (and investigative cost \( c \))
Logistic regression

Log likelihood (with $x_i \equiv x_{1i}$)

$$\sum_{i=1}^{n} \left\{ \alpha + x_i' \beta - \log(1 + e^{\alpha + x_i' \beta}) \right\} - \sum_{i=1}^{N} \left\{ \log(1 + e^{\alpha + x_{0i} \beta}) \right\}$$
Logistic regression

Log likelihood (with \( x_i \equiv x_{1i} \))

\[
\sum_{i=1}^{n} \left\{ \alpha + x_i' \beta - \log(1 + e^{\alpha + x_i' \beta}) \right\} - \sum_{i=1}^{N} \left\{ \log(1 + e^{\alpha + x_{0i} \beta}) \right\}
\]

For large \( N \)

\[
\sum_{i=1}^{N} \left\{ \log(1 + e^{\alpha + x_{0i} \beta}) \right\} \approx N \int \log(1 + e^{\alpha + x' \beta}) \, dF_0(x)
\]
Centering data

With foresight, center data at $\bar{x}$

$$
\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}}
$$
Centering data

With foresight, center data at $\bar{x}$

$$
Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}}
$$

Centered log likelihood $\ell(\alpha, \beta)$

$$
n\alpha - \sum_{i=1}^{n} \log \left(1 + e^{\alpha + (x_i - \bar{x})' \beta}\right) - N \int \log \left(1 + e^{\alpha + (x - \bar{x})' \beta}\right) dF_0(x)
$$
Centering data

With foresight, center data at $\bar{x}$

$$
\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}}
$$

Centered log likelihood $\ell(\alpha, \beta)$

$$
n\alpha - \sum_{i=1}^{n} \log \left( 1 + e^{\alpha + (x_i - \bar{x})' \beta} \right) - N \int \log \left( 1 + e^{\alpha + (x - \bar{x})' \beta} \right) dF_0(x)
$$

Because $\sum_{i=1}^{n} (\alpha + (x_i - \bar{x})' \beta) = n\alpha$
Sketch of the proof

Set \( \frac{1}{N} \frac{\partial}{\partial \beta} \ell(\alpha, \beta) = 0 \)

\[
0 = -\frac{1}{N} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) e^{\alpha + (x_i - \bar{x})'\beta}}{1 + e^{\alpha + (x_i - \bar{x})'\beta}} - \int \frac{(x - \bar{x}) e^{\alpha + (x - \bar{x})'\beta}}{1 + e^{\alpha + (x - \bar{x})'\beta}} dF_0(x)
\]
Sketch of the proof

Set \( \frac{1}{N} \frac{\partial}{\partial \beta} \ell(\alpha, \beta) = 0 \)

\[
0 = -\frac{1}{N} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) e^{\alpha + (x_i - \bar{x})' \beta}}{1 + e^{\alpha + (x_i - \bar{x})' \beta}} - \int \frac{(x - \bar{x}) e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}} dF_0(x)
\]

\( N \to \infty \), so ignore the first sum:

\[
0 = \int \frac{(x - \bar{x}) e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}} dF_0(x)
\]
Sketch of the proof

Set \( \frac{1}{N} \frac{\partial}{\partial \beta} \ell(\alpha, \beta) = 0 \)

\[
0 = - \frac{1}{N} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) e^{\alpha + (x_i - \bar{x})' \beta}}{1 + e^{\alpha + (x_i - \bar{x})' \beta}} - \int \frac{(x - \bar{x}) e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}} dF_0(x)
\]

\( N \to \infty \), so ignore the first sum:

\[
0 = \int \frac{(x - \bar{x}) e^{\alpha + (x - \bar{x})' \beta}}{1 + e^{\alpha + (x - \bar{x})' \beta}} dF_0(x)
\]

If \( \alpha \to -\infty \), denominator \( \to 1 \), and so \( \beta \) solves:

\[
\int (x - \bar{x}) e^{\alpha + (x - \bar{x})' \beta} dF_0(x) = 0 \quad \Box
\]
Example: $F_0 = N(0, 1), \bar{x} = 1, n = 1, N \rightarrow \infty$

Common values:

$x_{0i} \sim N(0, 1)$

Rare value

$n = 1$

$x_{11} = 1$
Example: \( F_0 = N(0, 1), \bar{x} = 1, n = 1, N \to \infty \)

Common values:
\[ x_{0i} \sim N(0, 1) \]

Rare value
\[ n = 1 \]
\[ x_{11} = 1 \]
Example: $F_0 = N(0, 1)$, $\bar{x} = 1$, $n = 1$, $N \to \infty$

Common values:
\[ x_{0i} \sim N(0, 1) \]

Rare value
\[ n = 1 \]
\[ x_{11} = 1 \]

We should see $\beta \to \Sigma_0^{-1}(\bar{x} - \mu_0) = 1^{-1}(1 - 0) = 1$
Example: \( F_0 = N(0, 1), \bar{x} = 1, n = 1, N \to \infty \)

For \( Y = 0 \) and \( i = 1, \ldots, N \) take

\[
x_{0i} = \Phi^{-1}\left(\frac{i - 1/2}{N}\right)
\]

We should see \( \beta \to \Sigma_0^{-1}(\bar{x} - \mu_0) = 1^{-1}(1 - 0) = 1 \)

Logistic regression results

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \alpha )</th>
<th>( Ne^\alpha )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-3.19</td>
<td>0.4126</td>
<td>1.5746</td>
</tr>
<tr>
<td>100</td>
<td>-5.15</td>
<td>0.5787</td>
<td>1.0706</td>
</tr>
<tr>
<td>1,000</td>
<td>-7.42</td>
<td>0.6019</td>
<td>1.0108</td>
</tr>
<tr>
<td>10,000</td>
<td>-9.71</td>
<td>0.6058</td>
<td>1.0017</td>
</tr>
<tr>
<td>100,000</td>
<td>-12.01</td>
<td>0.6064</td>
<td>1.0003</td>
</tr>
<tr>
<td>( \infty )</td>
<td></td>
<td>( \infty )</td>
<td>1</td>
</tr>
</tbody>
</table>
We will need conditions for the exponential tilting to work. One counterexample has a Cauchy distribution. The other a uniform.
Example: now \( F_0 = \text{Cauchy} \)

\[
f_0(x) = \frac{1}{\pi} \frac{1}{1 + x^2}
\]

\[
x_{0i} = F_0^{-1}\left(\frac{i - 1/2}{N}\right), \quad i = 1, \ldots, N
\]

\[
x_{1i} = 1, \quad i = 1 \quad \text{only}
\]
Example: now $F_0 = \text{Cauchy}$

$$f_0(x) = \frac{1}{\pi} \frac{1}{1 + x^2}$$

$$x_{0i} = F_0^{-1}\left(\frac{i - 1/2}{N}\right), \quad i = 1, \ldots, N$$

$$x_{1i} = 1, \quad i = 1 \quad \text{only}$$

Logistic regression results

<table>
<thead>
<tr>
<th>N</th>
<th>$\alpha$</th>
<th>$Ne^\alpha$</th>
<th>$\beta$</th>
<th>$Ne^\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>−2.36</td>
<td>0.94100</td>
<td>0.1222260</td>
<td>1.2222</td>
</tr>
<tr>
<td>100</td>
<td>−4.60</td>
<td>0.99524</td>
<td>0.0097523</td>
<td>0.9752</td>
</tr>
<tr>
<td>1,000</td>
<td>−6.90</td>
<td>0.99953</td>
<td>0.0009537</td>
<td>0.9536</td>
</tr>
<tr>
<td>10,000</td>
<td>−9.21</td>
<td>0.99995</td>
<td>0.0000952</td>
<td>0.9515</td>
</tr>
<tr>
<td>100,000</td>
<td>−11.51</td>
<td>0.99999</td>
<td>0.0000095</td>
<td>0.9513</td>
</tr>
</tbody>
</table>
Example: now $F_0 = \text{Cauchy}$

$$f_0(x) = \frac{1}{\pi} \frac{1}{1 + x^2}$$

$$x_{0i} = F_0^{-1}\left(\frac{i - 1/2}{N}\right), \quad i = 1, \ldots, N$$

$$x_{1i} = 1, \quad i = 1 \quad \text{only}$$

Logistic regression results

<table>
<thead>
<tr>
<th>N</th>
<th>$\alpha$</th>
<th>$Ne^\alpha$</th>
<th>$\beta$</th>
<th>$Ne^\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-2.36</td>
<td>0.94100</td>
<td>0.1222260</td>
<td>1.2222</td>
</tr>
<tr>
<td>100</td>
<td>-4.60</td>
<td>0.99524</td>
<td>0.0097523</td>
<td>0.9752</td>
</tr>
<tr>
<td>1,000</td>
<td>-6.90</td>
<td>0.99953</td>
<td>0.0009537</td>
<td>0.9536</td>
</tr>
<tr>
<td>10,000</td>
<td>-9.21</td>
<td>0.99995</td>
<td>0.0000952</td>
<td>0.9515</td>
</tr>
<tr>
<td>100,000</td>
<td>-11.51</td>
<td>0.99999</td>
<td>0.0000095</td>
<td>0.9513</td>
</tr>
</tbody>
</table>

$\beta(N) \rightarrow 0$  Cauchy has no mean to tilt onto $\bar{x}$!
Example: now $F_0 = U[0, 1]$ and $n_1 = 2$

Common values:
$x_{0i} \sim U(0, 1)$

Rare values:

$n = 2$
$x_{11} = 0.5$
$x_{12} = 2.0$
Example: now \( F_0 = U[0, 1] \) and \( n_1 = 2 \)

Common values:
\[ x_{0i} \sim U(0, 1) \]

Rare values:
\[ n = 2 \]
\[ x_{11} = 0.5 \]
\[ x_{12} = 2.0 \]
Example: now $F_0 = U[0, 1]$ and $n_1 = 2$

Common values:
\[ x_{0i} \sim U(0, 1) \]

Rare values:
\[ n = 2 \]
\[ x_{11} = 0.5 \]
\[ x_{12} = 2.0 \]

We can’t tilt $U(0, 1)$ to have mean $\bar{x} = 1.25$
Example: now $F_0 = U[0, 1]$ and $n_1 = 2$

\[
x_{0i} = \frac{i - 1/2}{N}, \quad i = 1, \ldots, N
\]

\[
x_{11} = \frac{1}{2}, \quad x_{12} = 2 \quad \text{only}
\]

Logistic regression results

<table>
<thead>
<tr>
<th></th>
<th>$\alpha$</th>
<th>$N\alpha$</th>
<th>$\beta$</th>
<th>$N\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$10$</td>
<td>$-3.82$</td>
<td>$0.2184$</td>
<td>$2.85$</td>
</tr>
<tr>
<td>$100$</td>
<td>$-7.13$</td>
<td>$0.0804$</td>
<td>$4.19$</td>
<td>$0.66$</td>
</tr>
<tr>
<td>$1,000$</td>
<td>$-10.71$</td>
<td>$0.0223$</td>
<td>$5.82$</td>
<td>$0.34$</td>
</tr>
<tr>
<td>$10,000$</td>
<td>$-14.52$</td>
<td>$0.0050$</td>
<td>$7.62$</td>
<td>$0.20$</td>
</tr>
<tr>
<td>$100,000$</td>
<td>$-18.49$</td>
<td>$0.0009$</td>
<td>$9.54$</td>
<td>$0.14$</td>
</tr>
</tbody>
</table>

$\beta(N) \rightarrow \infty$ also $\bar{x} = \frac{5}{4} \notin [0, 1]$ (can't tilt mean so far)
Example: now \( F_0 = U[0, 1] \) and \( n_1 = 2 \)

\[
x_{0i} = \frac{i - 1/2}{N}, \quad i = 1, \ldots, N
\]

\[
x_{11} = \frac{1}{2}, \quad x_{12} = 2 \quad \text{only}
\]

Logistic regression results

<table>
<thead>
<tr>
<th>N</th>
<th>( \alpha )</th>
<th>( Ne^\alpha )</th>
<th>( \beta )</th>
<th>( e^\beta/N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-3.82</td>
<td>0.2184</td>
<td>2.85</td>
<td>1.74</td>
</tr>
<tr>
<td>100</td>
<td>-7.13</td>
<td>0.0804</td>
<td>4.19</td>
<td>0.66</td>
</tr>
<tr>
<td>1,000</td>
<td>-10.71</td>
<td>0.0223</td>
<td>5.82</td>
<td>0.34</td>
</tr>
<tr>
<td>10,000</td>
<td>-14.52</td>
<td>0.0050</td>
<td>7.62</td>
<td>0.20</td>
</tr>
<tr>
<td>100,000</td>
<td>-18.49</td>
<td>0.0009</td>
<td>9.54</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Example: now $F_0 = U[0, 1]$ and $n_1 = 2$

$$x_{0i} = \frac{i - 1/2}{N}, \quad i = 1, \ldots, N$$

$$x_{11} = \frac{1}{2}, \quad x_{12} = 2 \quad \text{only}$$

Logistic regression results

<table>
<thead>
<tr>
<th>N</th>
<th>$\alpha$</th>
<th>$N e^\alpha$</th>
<th>$\beta$</th>
<th>$e^{\beta}/N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$-3.82$</td>
<td>0.2184</td>
<td>$2.85$</td>
<td>$1.74$</td>
</tr>
<tr>
<td>100</td>
<td>$-7.13$</td>
<td>0.0804</td>
<td>$4.19$</td>
<td>$0.66$</td>
</tr>
<tr>
<td>1,000</td>
<td>$-10.71$</td>
<td>0.0223</td>
<td>$5.82$</td>
<td>$0.34$</td>
</tr>
<tr>
<td>10,000</td>
<td>$-14.52$</td>
<td>0.0050</td>
<td>$7.62$</td>
<td>$0.20$</td>
</tr>
<tr>
<td>100,000</td>
<td>$-18.49$</td>
<td>0.0009</td>
<td>$9.54$</td>
<td>$0.14$</td>
</tr>
</tbody>
</table>

$\beta(N) \to \infty$ also $\bar{x} = \frac{5}{4} \notin [0, 1]$ (can’t tilt mean so far)
We need conditions:

Tail of $F_0$ not too heavy

$$\int \|x\| e^{x' \beta} dF_0(x) < \infty$$

to fix problem from Cauchy example
tail weight not an issue in finite samples
We need conditions:

Tail of $F_0$ not too heavy

$$\int \|x\| e^{x' \beta} \, dF_0(x) < \infty$$

to fix problem from Cauchy example

tail weight not an issue in finite samples

Overlap between $F_0$ and $\bar{x}$

to fix problem from $U(0, 1)$ example

overlap is an issue in finite samples

but we need stronger overlap condition
Overlap conditions

\( F \) has \( x^* \in \mathbb{R}^d \) surrounded if

- For all unit vectors \( \theta \in \mathbb{R}^d \)
- \( \Pr((x - x^*)'\theta > \epsilon \mid x \sim F_0) > \delta \)
- for some \( \epsilon > 0 \) and \( \delta > 0 \)

For finite samples, Silvapulle (1981, JRSS-B)

If model has intercept and \( x \)'s are full rank

We need some \( x_0 \) surrounded by both \( \hat{F}_1 \) and \( \hat{F}_0 \)
Overlap conditions

$F$ has $x^* \in \mathbb{R}^d$ surrounded if

- For all unit vectors $\theta \in \mathbb{R}^d$
- $\Pr((x - x^*)'\theta > \epsilon \mid x \sim F_0) > \delta$
- for some $\epsilon > 0$ and $\delta > 0$

For $N \to \infty$ we need:

- $F_0$ to have $\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$ surrounded
Overlap conditions

\( F \) has \( x^* \in \mathbb{R}^d \) surrounded if

- For all unit vectors \( \theta \in \mathbb{R}^d \)
- \( \operatorname{Pr}((x - x^*)' \theta > \epsilon \mid x \sim F_0) > \delta \)
- for some \( \epsilon > 0 \) and \( \delta > 0 \)

For \( N \to \infty \) we need:

- \( F_0 \) to have \( \bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i} \) surrounded

For finite samples, Silvapulle (1981, JRSS-B)

- If model has intercept and \( x \)’s are full rank
- We need some \( x_0 \) surrounded by both \( \hat{F}_1 \) and \( \hat{F}_0 \)
Theorem
Let $n \geq 1$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed. Suppose that

1. $F_0$ surrounds $\bar{x} = \sum_{i=1}^{n} x_i/n$
2. $\int \|x\|e^{x'\beta} \, dF_0(x) < \infty \quad \forall \beta \in \mathbb{R}^d$
Theorem

Let \( n \geq 1 \) and \( x_1, \ldots, x_n \in \mathbb{R}^d \) be fixed. Suppose that

1. \( F_0 \) surrounds \( \bar{x} = \sum_{i=1}^{n} x_i / n \)
2. \( \int \|x\| e^{x'\beta} dF_0(x) < \infty \quad \forall \beta \in \mathbb{R}^d \)

Then the maximizer \((\hat{\alpha}, \hat{\beta})\) of \( \ell \) satisfies

\[
\lim_{N \to \infty} \frac{\int e^{x'\hat{\beta}} x \, dF_0(x)}{\int e^{x'\hat{\beta}} \, dF_0(x)} = \bar{x}.
\]
Theorem
Let $n \geq 1$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed. Suppose that

1. $F_0$ surrounds $\bar{x} = \sum_{i=1}^{n} x_i / n$
2. $\int \|x\| e^{x'\beta} dF_0(x) < \infty \quad \forall \beta \in \mathbb{R}^d$

Then the maximizer $(\hat{\alpha}, \hat{\beta})$ of $\ell$ satisfies

$$\lim_{N \to \infty} \frac{\int e^{x'\hat{\beta}} x \, dF_0(x)}{\int e^{x'\hat{\beta}} \, dF_0(x)} = \bar{x}.$$ 

Steps

1. show $\alpha(N)$ and $\beta(N)$ exist for each $N$
2. show $Ne^{\hat{\alpha}(N)}$ is bounded
3. show $\|\hat{\beta}\|$ is bounded
4. then take partial derivatives as before
Computation

Given an approximation to $F_0$:

Solve

\[ 0 = \int (x - \bar{x}) e^{x' \beta} dF_0(x) \]

$d$ equations

Same as

\[ 0 = g(\beta) \equiv \int (x - \bar{x}) e^{(x-\bar{x})' \beta} dF_0(x) \]

I.E. Minimize

\[ f(\beta) = \int e^{(x-\bar{x})' \beta} dF_0(x) \]

Hessian is

\[ H(\beta) = \int (x - \bar{x})(x - \bar{x})' e^{(x-\bar{x})' \beta} dF_0(x) \]

convex

Newton step

\[ \beta \leftarrow \beta - H^{-1} g \]

Cost per iteration: $O(d^3)$ vs $O(N d^2)$ or $O(n d^2)$. 
Mixture of Gaussians

\[ F_0 = \sum_{k=1}^{K} \lambda_k N(\mu_k, \Sigma_k) \quad \lambda_k > 0 \quad \sum_k \lambda_k = 1 \]

Tilt a Gaussian, get a Gaussian:

\[ e^{(x-\bar{x})'\beta} e^{-\frac{1}{2} (x-\mu)'\Sigma^{-1}(x-\mu)} = e^{(\mu-\bar{x})'\beta} e^{-\frac{1}{2} (x-\mu-\Sigma\beta)'\Sigma^{-1}(x-\mu-\Sigma\beta)} \]

Newton step is

\[ \beta \leftarrow \beta - H^{-1} g \]

\[ g = \sum_{k=1}^{K} \lambda_k e^{(\mu_k-\bar{x})'\beta} \left( \tilde{\mu}_k - \bar{x} \right), \quad \tilde{\mu}_k = \mu_k + \Sigma_k \beta \]

\[ H = \sum_{k=1}^{K} \lambda_k e^{(\mu_k-\bar{x})'\beta} \left( \Sigma_k + (\bar{x} - \tilde{\mu}_k)(\bar{x} - \tilde{\mu}_k)' \right) \]
Zhu, Su, Chipman

Technometrics, 2005

\( Y = 1 \) for active drug

\( Y = 0 \) for inactive drug

\( d = 6 \) features

29,821 chemicals

only 608 active \( \approx 2\% \)

\( x_1, x_3 \) strongest

Group means plotted
Drug discovery example ctd

Fits
Plain logistic
\((608\text{ ones}), \text{ vs}\)
1 one at \(\bar{x}_1\)

Upshot
Same ordering, ROC
precision-recall
etc.
Drug discovery example ctd

ROC curves
Plain logistic
1 one at $\bar{x}_1$
Drug discovery example ctd

Fits
Plain logistic, vs,
Pretend $F_0$ is Gaussian
And use $\bar{x}_1$

Upshot
Slight difference
For easy 0s
Mixture model might improve
The drug data was not a typical example

Drug data had

- very bad separation
- Poor ROC
- $\bar{x}$ very surrounded
The drug data was not a typical example

Drug data had

- very bad separation
- Poor ROC
- \( \bar{x} \) very surrounded

**Artificial version**

\[
x_{1i} \leftarrow x_{1i} + \delta
\]

\[
\delta = \left( s/10, \ldots, s/10 \right)
\]

\[
s = 0, \ldots, 10
\]

- Original ROCs in blue
- Lumped in red
The drug data was not a typical example

Drug data had
   very bad separation
   Poor ROC
   $\bar{x}$ very surrounded

Artificial version
$x_{1i} \leftarrow x_{1i} + \delta$
$\delta = (s/10, \ldots, s/10)$
$s = 0, \ldots, 10$
Original ROCs in blue
Lumped in red
The drug data was not a typical example

Drug data had
  very bad separation
  Poor ROC
  $\bar{x}$ very surrounded

Artificial version
  $x_{1i} \leftarrow x_{1i} + \delta$
  $\delta = (s/10, \ldots, s/10)$
  $s = 0, \ldots, 10$

Original ROCs in blue
Lumped in red

Upshot
  Still only uses $\bar{x}$
Thoughts for fraud detection

Non fraud data, $Y = 0$
- Change slowly over time
- Large sample size
- So build a rich model for $F_0$
- Update rarely
Thoughts for fraud detection

Non fraud data, $Y = 0$
- Change slowly over time
- Large sample size
- So build a rich model for $F_0$
- Update rarely

Fraud data, $Y = 1$
- May change rapidly in response to detection
- May have different flavors
- Clusters appear, disappear, move, change size
- Rapidly refit model using per cluster $\bar{x}$
Acknowledgments

- Paul Louisell for comments
- NSF for funds
- Host: University of Florida
- Organizers: Agresti, Young, Daniels, Casella
- Travel help: Robyn Crawford