A modern maximum-likelihood theory for high-dimensional logistic regression

by Pragya Sur and Emmanuel J. Candés

ML inference

- Theory associated with ML estimators makes them so appealing
- Assume iid sample $y_1, \ldots, y_n$ from $f_{\boldsymbol{\theta}}(y)$
- Unknown $\boldsymbol{\theta}$: MLE $\widehat{\boldsymbol{\theta}}$ maximizes $\ell(\boldsymbol{\theta}) = \sum_i \log f_{\boldsymbol{\theta}}(y_i)$
- Other associated quantity of importance: expected Fisher information

$$\mathcal{I}_{\boldsymbol{\theta}} = \mathsf{E}_f \left\{ \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^\mathsf{T} \partial \boldsymbol{\theta}} \right\}$$

ML inference

- Theory associated with ML estimators makes them so appealing
- Assume iid sample $y_1, \ldots, y_n$ from $f_{\boldsymbol{\theta}}(y)$
- Unknown $\boldsymbol{\theta}$: MLE $\widehat{\boldsymbol{\theta}}$ maximizes $\ell(\boldsymbol{\theta}) = \sum_i \log f_{\boldsymbol{\theta}}(y_i)$
- Other associated quantity of importance: expected Fisher information
$$\mathcal{I}_{\boldsymbol{\theta}} = \mathsf{E}_f \left\{ \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^\mathsf{T} \partial \boldsymbol{\theta}} \right\}$$

  Why? Asymptotically, as $n \to \infty$
$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}_{\boldsymbol{\theta}}^{-1})$$

ML in logistic regression setting

- Logistic regression independent pairs: $(y_i, \mathbf{x}_i)$ with $y_i \in \{0, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^p$

$$\Pr\{y_i = 1 | \mathbf{x}_i\} = \rho'(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{\mathrm{e}^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_i^\top \boldsymbol{\beta}}}$$

- $\boldsymbol{\beta} \in \mathbb{R}^p$ unknown parameters: using the MLE gives approximate pivotal quantity

ML in logistic regression setting

- Logistic regression independent pairs: $(y_i, \mathbf{x}_i)$ with $y_i \in \{0,1\}$ and $\mathbf{x}_i \in \mathbb{R}^p$

$$\Pr\{y_i = 1 | \mathbf{x}_i\} = \rho'(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{\mathrm{e}^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_i^\top \boldsymbol{\beta}}}$$

- $\boldsymbol{\beta} \in \mathbb{R}^p$ unknown parameters: using the MLE gives approximate pivotal quantity

$$\sqrt{\mathcal{I}_{\boldsymbol{\beta},jj}} \left(\widehat{\beta}_j - \beta_j \right) \sim \mathcal{N}(0,1)$$

ML in logistic regression setting

- Logistic regression independent pairs: $(y_i, \mathbf{x}_i)$ with $y_i \in \{0, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^p$

$$\Pr\{y_i = 1 | \mathbf{x}_i\} = \rho'(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}) = \frac{\mathrm{e}^{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}$$

- $\boldsymbol{\beta} \in \mathbb{R}^p$ unknown parameters: using the MLE gives approximate pivotal quantity

$$\sqrt{\mathcal{I}_{\boldsymbol{\beta},jj}}\left(\widehat{\beta}_j - \beta_j\right) \sim \mathcal{N}(0, 1)$$

$$\frac{\widehat{\beta}_j - \beta_j}{\mathrm{s.e.}(\widehat{\beta}_j)} \sim \mathcal{N}(0, 1) \qquad \text{classic } z \text{ test statistic}$$

Bias

- Test usually of most interest is

$$H_0 : \beta_j = 0 \qquad H_A : \beta_j \neq 0$$

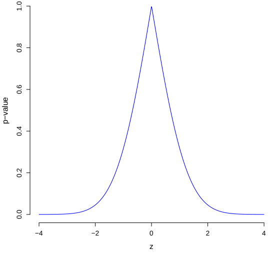
$$z = \frac{\widehat{\beta}_j}{\text{s.e.}(\widehat{\beta}_j)} \qquad \text{and} \qquad p = 2[1 - \Phi(z)],\ z > 0$$

e.g. reject when $p < 0.05$

Bias

Imagine s.e. being used is biased (deflated $\nu < 1$):

$$z_{\text{true}} = \frac{\widehat{\beta}_j}{\nu \, \text{s.e.}_{\text{true}}(\widehat{\beta}_j)}$$ i.e. bigger than should

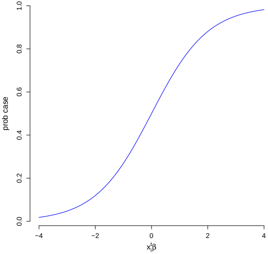


Imagine the estimate $\widehat{\boldsymbol{\beta}}$ is biased

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_{\text{true}} + \mathbf{b}$$

$$\mathbf{x}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}} = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}_{\text{true}} + \mathbf{x}_i^{\mathsf{T}} \mathbf{b}$$

i.e. pushes further out



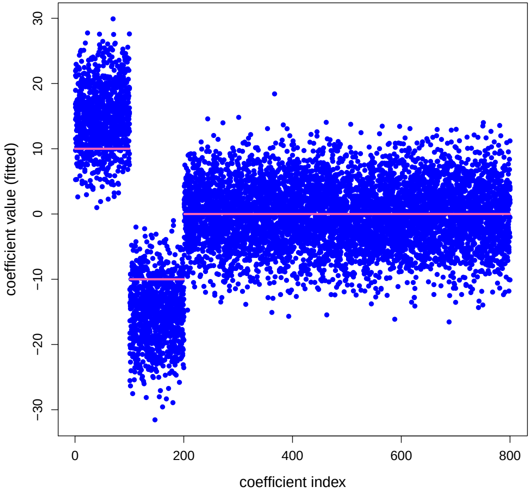
Aspect ratio asymptotics

- This paper makes the point that for *logistic regression* if $n \to \infty$ while p is fixed or remains small i.e. $p = o(n)$ then $\widehat{\beta}_j$ and
$$\frac{\widehat{\beta}_j}{\text{s.e.}(\widehat{\beta}_j)} \sim \mathcal{N}(0,1)$$
are fine.

- However, if $p/n \to \kappa$ as $n \to \infty$, that is p is non-negligible compared to n, something like $p = o(n^{1+\alpha}), 0 < \alpha < 1$, then there is bias in both the MLE and the standard error

Example for $\boldsymbol{\beta}$

$n = 4000, p = 800$, simulate with $\beta_j = 10, j = 1, \ldots, 100$, $\beta_j = -10, j = 101, \ldots, 200$, $\beta_j = 0$ otherwise.



Remedy

- To test the hypothesis

$$H_0 : \beta_j = 0 \qquad H_A : \beta_j \neq 0$$

adjust test statistic for computing the p-value

$$\frac{\widehat{\beta}_j}{\sigma^*}$$

where σ^* results from solving the non-linear system of equations.

- One might expect the bias to creep into other inferential activities as well. This is indeed the case, as the likelihood ratio statistic is shown to be an adjusted χ^2 under the null hypothesis.