Maximum-likelihood fitting of logistic regression models

STAT 206
19 Mar 21

The model is

Extra Lecture

$$(Y_i \mid p_i, \mathcal{B}) \overset{I}{\sim} \text{Bernoulli}(p_i) \quad (i = 1, \dots, n) \qquad (1)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j = \beta_0 + (\underset{\sim}{X}\underset{\sim}{\beta})_i$$

[LR] = logistic regression model

$\underset{\sim}{Y} = (Y_1, \dots, Y_n)$ observed

$\underset{\sim}{X} = (X_1, \dots, X_n)$

in which

$$\underset{n \times k}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

is the matrix of predictor values (1)

$$(\underset{\sim}{\beta})_{k \times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

is the vector of logistic regression coefficients

$$(\underset{\sim}{\theta})_{k+1 \times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

is the vector of unknown parameters

joint

The sampling distribution is

$$p(\underset{\sim}{y} \mid \beta_0, \underset{\sim}{\beta}, [LR], \mathcal{B}) =$$

$$\prod_{i=1}^{n} p(y_i \mid \beta_0, \underset{\sim}{\beta}, [LR], \mathcal{B})$$

$\{\underset{\sim}{\theta}\} = k+1$ unknowns

$\dim(\underset{\sim}{\theta}) = k+1$ ($n > k+1$)

and this is

$$p(y \mid \beta_0, \beta [LR] B) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$ ②

expressed in terms of the probabilities $p_i = P(Y_i = 1 \mid B)$

---

so the likelihood function is ↓ $\beta = (\beta_1, \ldots, \beta_n)$

$$\ell(\beta \mid y [LR] B) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} p_i^{y_i} (1-p_i)(1-p_i)^{-y_i} = \prod_{i=1}^{n} (1-p_i) \left(\frac{p_i}{1-p_i} \right)^{y_i}$$

---

and the log likelihood function is then

$$\ell\ell(\beta \mid y [LR] B) =$$

$$\left[\sum_{i=1}^{n} \log(1-p_i) \right] + \left[\sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i} \right) \right]$$

$$= \left[\sum_{i=1}^{n} \log(1-p_i) \right] + \left\{ \sum_{i=1}^{n} y_i \left[\beta_0 + (X\beta)_i \right] \right\}$$

$$= \left[\sum_{i=1}^{n} \log(1-p_i) \right] + \left(\beta_0 \sum_{i=1}^{n} y_i \right) + \left(\sum_{i=1}^{n} y_i \sum_{j=1}^{k} x_{ij} \beta_j \right)$$

interchanging the order of summation in the last term gives

$$\sum_{i=1}^{n} y_i \sum_{j=1}^{k} x_{ij}\beta_j = \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij} y_i \beta_j$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij} y_i \beta_j = \sum_{j=1}^{k} \beta_j \boxed{\left(\sum_{i=1}^{n} x_{ij} y_i\right)}$$

so we "see" that there is a $(k+1)$ dimensional set of minimal sufficient statistics (MSS), namely

$$\left\{ \underbrace{\left(\sum_{i=1}^{n} y_i\right)}_{S_y}, \underbrace{\left(\sum_{i=1}^{n} x_{ij} y_i\right)}_{(S_{xy})_j} \text{ for each } j = 1, \ldots, k \right\} \hat{} \quad (*)$$

but wait: what about the annoying $\sum_{i=1}^{n} \log(1-p_i)$ term? Not so fast with the MSS claim, although (*) was a highly useful observation: those $(k+1)$

quantities can be computed once, before the iterative search for the MLEs begins

---

To deal with the $\sum_{i=1}^{n} \log(1-p_i)$ term, we need to solve the equation $\log\left(\frac{p_i}{1-p_i}\right) = \ell_i$ (ℓ for logit not likelihood) backwards for p_i:

$$\frac{p_i}{1-p_i} = e^{\ell_i} \rightarrow p_i = (1-p_i)e^{\ell_i} = e^{\ell_i} - p_i e^{\ell_i}$$

$$p_i(1+e^{\ell_i}) = e^{\ell_i} \rightarrow p_i = \frac{e^{\ell_i}}{1+e^{\ell_i}}$$

and $1-p_i = \frac{1}{1+e^{\ell_i}}$ so the pesky term is

$$\sum_{i=1}^{n} \log(1-p_i) = \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{\ell_i}}\right) = -\sum_{i=1}^{n} \log(1+e^{\ell_i})$$

and the full log likelihood function in terms of the β_j's is

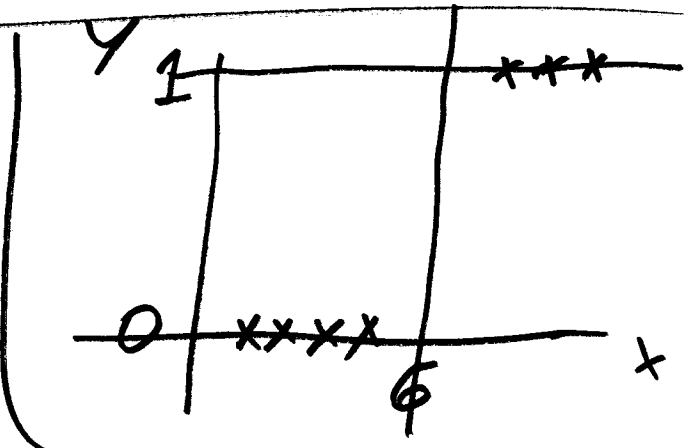
$$\ell\ell(\beta_0, \underline{\beta} \mid \underline{x}, \underline{y}) = -\sum_{i=1}^{n} \log\{1 + \exp[(\beta_0 + (\underline{x}\underline{\beta})_i]\} + \beta_0 \underbrace{\sum_{i=1}^{n} y_i}_{S_y} + \sum_{j=1}^{k} \beta_j \underbrace{\left(\sum_{i=1}^{n} x_{ij} y_i\right)}_{(S_{xy})_j}$$

Covariance information between $\underline{x}$ and $\underline{y}$

This clearly cannot be maximized in closed form; iterative numerical methods are needed to find the MLEs. The standard numerical method, Fisher scoring, is just an application of Newton-Raphson, and usually converges quickly to a (local) maximum of the log likelihood function; this method is also an instance of an approach called Iteratively Reweighted Least Squares (IRLS).

* next page

Convergence issues from ⊛ last page: The MLEs can fail to exist in this model in 2 situations: ① If $(X_{n\times(k+1)}\ \mathbf{1}_n)$ is not of full rank, i.e., if there are linear dependencies among the columns of X; R solves this by booting predictors out of the model that cause the collinearity ② Here for this dataset there's a separating hyperplane



(in this case, the vertical line $x=6$) such that $(x<6)$ perfectly predicts $y=0$ and $(x>6)$ perfectly predicts $y=1$. The log likelihood function may be maximized in this situation at $\beta=\pm\infty$, or ludicrous standard errors may emerge.

(to be continued someday)

$(Y \mid p, B) \sim \text{Bernoulli}(p)$

PMF

$p(Y = y \mid p, B) =$

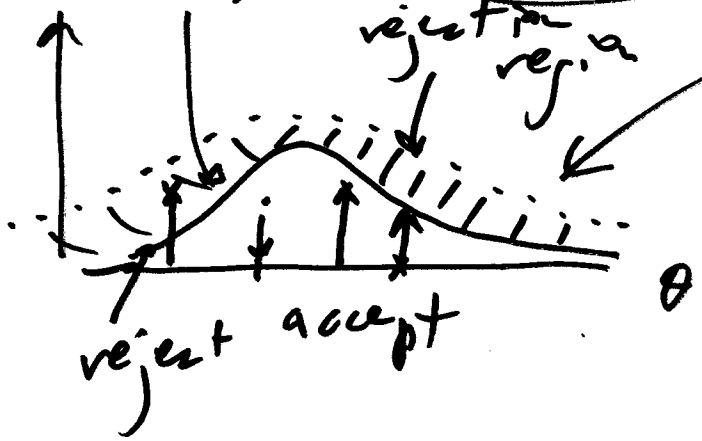
$E(Y)$

$p^y (1-p)^{1-y} \cdot \mathbb{I}_{\{0,1\}}(y) =$

$\begin{cases} p & \text{if } y = 1 \\ 1-p & 0 \\ 0 & \text{else} \end{cases}$

target PDF

$p(\theta \mid \cdots)$



envelope function $G(\theta)$

① $G(\theta) \geq p(\theta)$

② it's integrable

$\int_{\mathbb{R}} G(\theta)\, d\theta < \infty$

normalizable to $g(\theta) = \dfrac{G(\theta)}{\int_{\mathbb{R}} G(\tilde{\theta})\, d\tilde{\theta}}$

③ g should be easy & fast to sample from

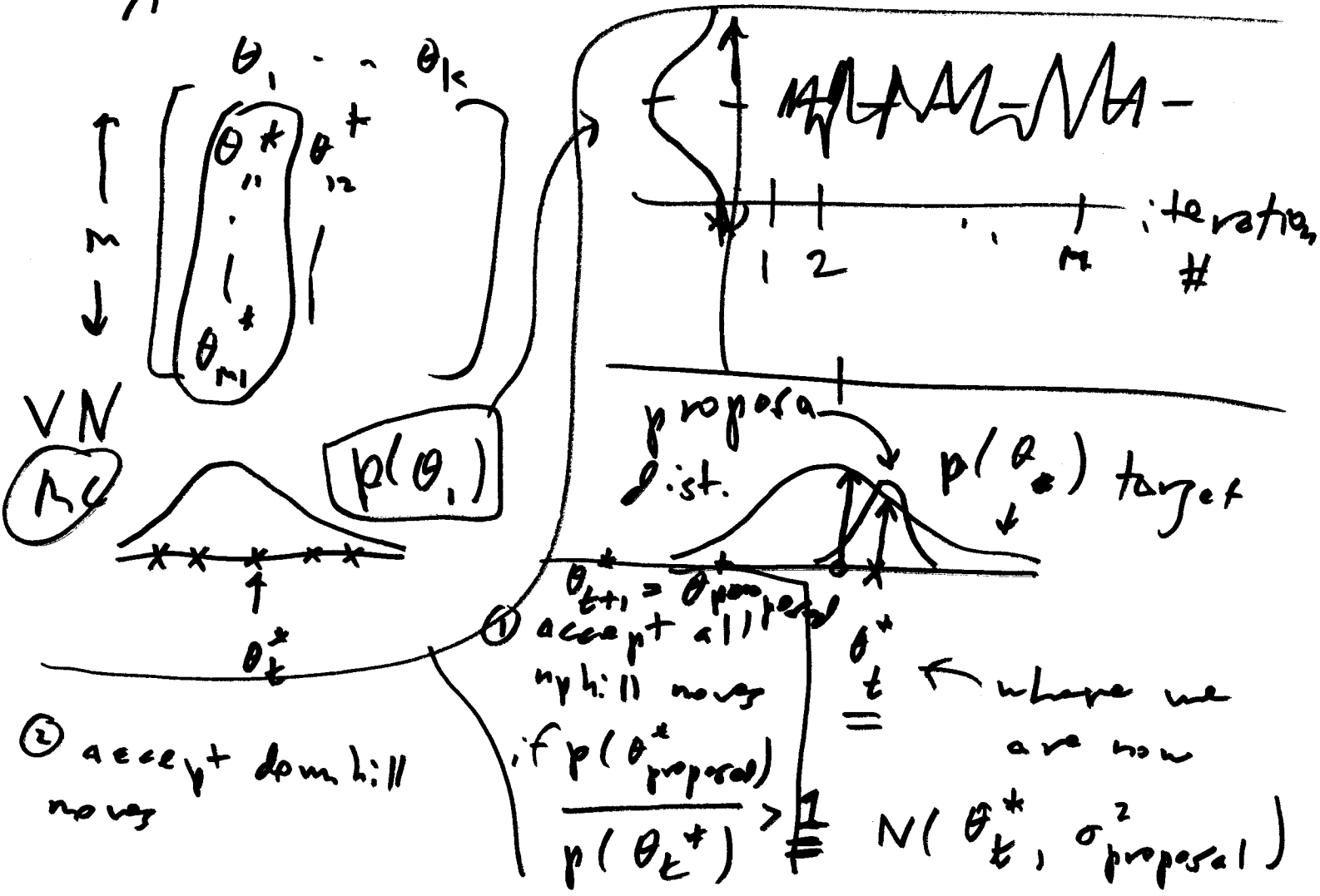
v. N: [I][D] but low Mc efficiency for large k

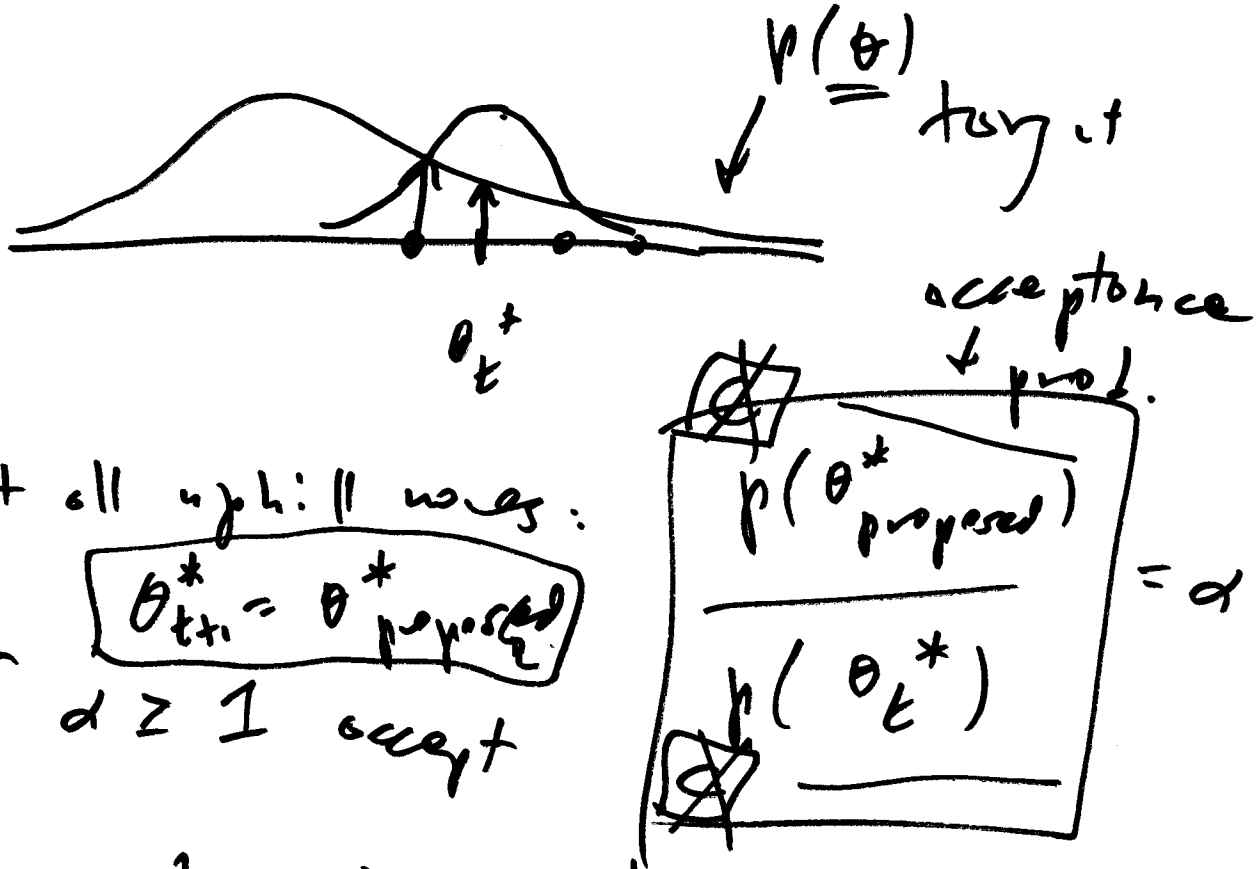
$\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$

Metr. et al. can't get rid of [D], but how about relaxing [I]?

from the target $p(\underset{\sim}{\theta})$

Constructing a 1st order Markov chain





① accept all uphill moves:

$\theta^*_{t+1} = \theta^*_{proposed}$

if $\alpha \geq 1$ accept

↔ accept with prob. $\min(\alpha, 1)$

② accept downhill moves with prob α

if reject, $\theta^*_{t+1} = \theta^*_t$

stay where you are (!)

$$p(\theta \mid -) = \boxed{c}\, p(\theta)\, \ell(\theta \mid -)$$