Maximum likelihood estimation for logistic regression

Invariance of the MLE

Last time: $Y_1, \dots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Q: what if we want $\hat{\sigma}$? Take the square root!

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Theorem (invariance of the MLE): (see Theorem 7.2.10 in CB)

Let $\hat{\theta}$ be the MLE of θ. For any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$

Maximum likelihood estimation for logistic regression

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

Suppose we observe independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. Write down the likelihood function

$$L(\beta | \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} f(Y_i | \beta, X_i)$$

for the logistic regression problem.

$$L(\beta \mid X, Y) = \prod_{i=1}^{n} f(y_i \mid \beta, x_i) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}} \right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta^T x_i}} \right)^{1-y_i} \qquad x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix} \in \mathbb{R}^{(k+1)}$$

$$\Rightarrow \quad \ell(\beta \mid X, Y) = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right) + (1-y_i) \log\left(\frac{1}{1 + e^{\beta^T x_i}} \right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right) \right\}$$

$$\ell(\beta \mid X, Y) = \sum_{i=1}^{n} \left\{ Y_i \beta^T X_i - \log\left(1 + e^{\beta^T X_i}\right) \right\}$$

Rules for matrix derivatives

$$\frac{\partial}{\partial \beta} \beta^T X_i = X_i$$

$$\frac{\partial \ell}{\partial \beta} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_K} \end{pmatrix}$$

$$X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}$$

$$\Rightarrow X^T = [X_1 \ X_2 \cdots X_n]$$

$$\Rightarrow X^T(Y - p) = \sum_{i=1}^{n} X_i (Y_i - p_i)$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left\{ \underbrace{\frac{\partial \ell}{\partial \beta} Y_i \beta^T X_i}_{Y_i X_i} - \underbrace{\frac{\partial}{\partial \beta} \log\left(1 + e^{\beta^T X_i}\right)}_{\frac{1}{1+e^{\beta^T X_i}} \cdot e^{\beta^T X_i} \cdot X_i} \right\}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left(Y_i X_i - \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} X_i \right) = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} \right) X_i$$

p_i

$$= X^T (Y - p) \overset{\text{set}}{=} 0$$

solve for β.... hard!

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1K} \\ 1 & X_{21} & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & & X_{nK} \end{bmatrix} \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

Iterative methods for maximizing likelihood

$$\underbrace{U(\beta)}_{\text{score function}} = \frac{\partial \ell}{\partial \beta} = X^T(Y - p) = 0$$

score function

want β^* st $U(\beta^*) = 0$

but no closed-form solutions

Idea:
1) Start w/ an initial guess $\beta^{(0)}$
2) update guess to $\beta^{(1)}$, which is (hopefully) closer to β^*
3) iterate!

Newton's method

want β^* st $u(\beta^*) = 0$, given initial guess $\beta^{(0)}$

First-order Taylor expansion around $\beta^{(0)}$

$$\underbrace{u(\beta^*)}_{=\,0} \approx u(\beta^{(0)}) + \frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} (\beta^* - \beta^{(0)})$$

$$\Rightarrow u(\beta^{(0)}) + \frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} (\beta^* - \beta^{(0)}) \approx 0$$

$$\Rightarrow \beta^* \approx \underbrace{\beta^{(0)} - \left(\frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} \right)^{-1} u(\beta^{(0)})}_{\text{we can evaluate this!}}$$

$$\beta^* \approx \beta^{(0)} - \left(\frac{\partial U(\beta^{(0)})}{\partial \beta^{(0)}} \right)^{-1} U(\beta^{(0)})$$

what is $\dfrac{\partial \overbrace{U(\beta)}^{\text{vector}}}{\partial \beta}$? $\Rightarrow$ matrix of partial derivatives

vector

$$\frac{\partial U(\beta)}{\partial \beta} = \frac{\partial^2 \ell}{\partial \beta^2} = \begin{bmatrix} \frac{\partial^2 \ell(\beta)}{\partial \beta_0^2} & \frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_k} \\ \vdots & & & \\ \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_0} & \cdots & & \frac{\partial^2 \ell(\beta)}{\partial \beta_k^2} \end{bmatrix}$$

$$U(\beta) = \frac{\partial \ell}{\partial \beta}$$

Hessian of log likelihood $H(\beta)$

Newton's method

1) Initial guess $\beta^{(0)}$

2) Update: $\beta^{(r+1)} = \beta^{(r)} - \left(H(\beta^{(r)})\right)^{-1} U(\beta^{(r)})$

3) Stop when $\beta^{(r)} \approx \beta^{(r+1)}$

Logistic regression: $U(\beta) = X^T(Y-p)$

$$H(\beta) = \frac{\partial U(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} X^T(Y-p) = \frac{-\partial}{\partial \beta} X^T p$$

- Under certain conditions (which hold for logistic regression)

$\beta^{(r)} \to \beta^*$

if $\beta^{(0)}$ is sufficiently close to β^*

Newton's method for logistic regression

Example

Suppose that $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$, and we have

$$\beta^{(r)} = \begin{bmatrix} -3.1 \\ 0.9 \end{bmatrix}, \qquad U(\beta^{(r)}) = \begin{bmatrix} 9.16 \\ 31.91 \end{bmatrix},$$

$$\mathbf{H}(\beta^{(r)}) = -\begin{bmatrix} 17.834 & 53.218 \\ 53.218 & 180.718 \end{bmatrix}$$

Use Newton's method to calculate $\beta^{(r+1)}$ (you may use R or a calculator, you do not need to do the matrix arithmetic by hand).