Lecture 6: Maximum likelihood estimation for logistic regression

Ciaran Evans

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 $(X_1, Y_1), ..., (X_n, Y_n)$.
Write down the likelihood function

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

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

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

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$$

$$X_i = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{id} \end{pmatrix}$$

$$p_i = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}$$

$$f(Y_i \mid X_i, \beta) = p_i^{Y_i}(1 - p_i)^{1 - Y_i}$$

$$\Rightarrow L(\beta \mid X, Y) \propto \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1 - Y_i}$$

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

(up to a constant)

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

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

$$\ell(\beta \mid X, Y) = \sum_{i=1}^{n} \left\{ Y_i \beta^T X_i - \log\left(1 + e^{\beta^T X_i}\right) \right\} \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\frac{\partial \ell}{\partial \beta} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_k} \end{pmatrix} \qquad X_i = \begin{pmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{ik} \end{pmatrix}$$

(gradient)

Rule for matrix derivatives

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

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

$$= \sum_{i=1}^{n} \left\{ Y_i X_i - \underbrace{\frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}}_{p_i} X_i \right\}$$

X = design matrix

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad p = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$$

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

Linear regression:

$$X^T(Y - \underbrace{X\beta}_{\mu}) \overset{set}{=} 0$$

$$\mu = X\beta$$

Logistic regression:

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

$$p = \frac{e^{X\beta}}{1+e^{X\beta}}$$

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

<u>Want</u>: value β^* such that $u(\beta^*) = 0$

no closed-form solution for logistic regression model

<u>Idea</u>: 1) Start with an initial guess $\beta^{(0)}$

2) Update guess to $\beta^{(1)}$, which is (hopefully!) closer to β^*

3) Iterate!

Newton's method

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

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

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

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

$\Rightarrow$ update:

$$\beta^{(1)} = \beta^{(0)} - \left(\frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} \right)^{-1} u(\beta^{(0)})$$

Sidebar: initial guess for logistic regression could be

$$\beta^{(0)} = \begin{pmatrix} \log\left(\frac{\bar{Y}}{1-\bar{Y}}\right) \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

← match intercept to observed sample mean

set other coefficients to 0

$$u(\beta) = \frac{\partial \ell}{\partial \beta} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_K} \end{pmatrix}$$

↑ vector

$$\frac{\partial u}{\partial \beta} = \frac{\partial^2 \ell}{\partial \beta^2}$$

← matrix

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

Hessian of log-likelihood

$$\beta^{(1)} = \beta^{(0)} - \left(H(\beta^{(0)})\right)^{-1} u(\beta^{(0)})$$

Newton's method for logistic regression

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

$$p = \frac{e^{X\beta}}{1+e^{X\beta}}$$

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

$$p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

$$= -\frac{\partial}{\partial \beta} X^T p$$

$$p_i = \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}$$

(chain rule)

$$= \left(-\frac{\partial p}{\partial \beta}\right) X$$

$$\frac{\partial p}{\partial \beta} = \left[\frac{\partial p_1}{\partial \beta} \quad \frac{\partial p_2}{\partial \beta} \quad \cdots \quad \frac{\partial p_n}{\partial \beta} \right]$$

$$p_i = \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}} = \underbrace{g(h(\beta))}_{\frac{e^{h(\beta)}}{1+e^{h(\beta)}}}$$

$$g(u) = \frac{e^u}{1+e^u} \qquad h(\beta) = \beta^T X_i$$

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).