Lecture 7: Maximum likelihood estimation for logistic regression

Ciaran Evans

Recall: Newton's method

Score equation: $u(\beta) = \frac{\partial \ell}{\partial \beta} \stackrel{\text{set}}{=} 0$

1) Initial guess $\beta^{(0)}$ Hessian $= \frac{\partial^2 \ell}{\partial \beta^2}$

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

3) Stop when algorithm converges

e.g. $\|\beta^{(r+1)} - \beta^{(r)}\| < \varepsilon$

or $\|\ell(\beta^{(r+1)}) - \ell(\beta^{(r)})\| < \delta$

Newton's method for logistic regression

X = design matrix

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

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

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

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

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

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

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

$\#\beta$s $\sim (n+1) \times n \leftarrow \#$ obs

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

So: need $\frac{\partial p_i}{\partial \beta} \in \mathbb{R}^{n+1}$

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

$\beta^T x_i \in \mathbb{R}$ (scalar) $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}$ $x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{in} \end{pmatrix}$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$

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

$u = \beta^T x_i$

$$g(u) = \frac{e^u}{1+e^u}$$

$h(\beta) = \beta^T x_i$

$$g'(u) = \frac{e^u}{(1+e^u)^2}$$

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

Chain rule for single variable calculus:

$$\frac{\partial}{\partial x} g(h(x)) = g'(h(x))\, h'(x)$$

Chain rule for matrix calculus: If x is a vector and $h(x) \in \mathbb{R}$ and $g(h(x)) \in \mathbb{R}$. Then

$$\frac{\partial g(h(x))}{\partial x} = \underbrace{g'(h(x))}_{\text{scalar piece}} \underbrace{\frac{\partial h}{\partial x}}_{\text{vector piece}}$$

$$\frac{\partial p_i}{\partial \beta} = \frac{e^{\beta^T x_i}}{(1+e^{\beta^T x_i})^2} x_i = p_i(1-p_i)\, x_i$$

$$H(\beta) = \left(-\frac{\partial p}{\partial \beta}\right) X = -\left[p_1(1-p_1)x_1 \quad p_2(1-p_2)x_2 \quad \cdots \quad p_n(1-p_n)x_n\right] X$$

$$= \boxed{-X^T W X} \qquad W = \operatorname{diag}(p_i(1-p_i))$$

$[p_1(1-p_1)X_1 \quad p_2(1-p_2)X_2 \quad \cdots \quad p_n(1-p_n)X_n] \; X$

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

$X = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}$

i^{th} column = X_i

= i^{th} observation in dataset

i^{th} row = i^{th} individual (i^{th} observation)

j^{th} row = j^{th} variable

j^{th} column = j^{th} variable

$[X_1 \quad X_2 \cdots X_n] \; ? \; = [p_1(1-p_1)X_1 \cdots p_n(1-p_n)X_n]$

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

Recall: $[X_1 \; X_2 \cdots X_n] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a_1X_1 + a_2X_2 \cdots + a_nX_n$

vectors

= linear combination of columns of matrix

$X_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{in} \end{pmatrix}$

$[X_1 \; X_2 \cdots X_n] \begin{bmatrix} p_1(1-p_1) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ vector $= p_1(1-p_1)X_1$

$$\begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} p_1(1-p_1) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = p_1(1-p_1)X_1$$

$$\begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 0 \\ p_2(1-p_2) \\ 0 \\ \vdots \\ 0 \end{bmatrix} = p_2(1-p_2)X_2$$

$$\underbrace{\begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}}_{X^T} \overbrace{\begin{bmatrix} p_1(1-p_1) & 0 & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & 0 & & 0 \\ \vdots & 0 & p_3(1-p_3) & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ & & & & p_n(1-p_n) \end{bmatrix}}^{W}$$

$$\Rightarrow X^T W X = \begin{bmatrix} p_1(1-p_1)X_1 & p_2(1-p_2)X_2 & \cdots & p_n(1-p_n)X_n \end{bmatrix}$$

Comparison:

Linear regression: $H(\beta) = -\frac{1}{\sigma^2} X^T X$

($\sigma^2 = \text{var}(\varepsilon_i)$)

Newton's method (if we used it) converges in a single step

$$-(H(\beta))^{-1} = \sigma^2 (X^T X)^{-1}$$

$$u(\beta) = \frac{1}{\sigma^2} X^T (Y - \mu) = \frac{1}{\sigma^2} X^T (Y - X\beta)$$

$$\Rightarrow -(H(\beta))^{-1} u(\beta) = \sigma^2 (X^T X)^{-1} \frac{1}{\sigma^2} X^T (Y - X\beta)$$

$$= (X^T X)^{-1} X^T (Y - X\beta)$$

- minimize $(Y - X\beta)^T (Y - X\beta)$ quadratic function of β
- $u(\beta) = \frac{1}{\sigma^2} X^T (Y - X\beta)$ linear function of β
  - $\Rightarrow$ closed form solution
  - $\Rightarrow$ first-order Taylor expansion is exact (no approximation)

$$\beta^{(0)} + (X^T X)^{-1} X^T (Y - X\beta^{(0)})$$

$$= \beta^{(0)} + (X^T X)^{-1} X^T Y - \underbrace{(X^T X)^{-1} X^T X}_{I} \beta^{(0)}$$

$$= \beta^{(0)} + (X^T X)^{-1} X^T Y - \beta^{(0)}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Linear regression: $H(\beta) = -\frac{1}{\sigma^2} X^T X$

Logistic regression: $H(\beta) = -X^T W X$

$W = \text{diag}(\underbrace{p_i(1-p_i)}_{\text{Var}(Y_i|X_i)})$

Poisson regression: $H(\beta) = -X^T W X$

$W = \text{diag}(\lambda_i)$

λ_i: $\text{Var}(Y_i|X_i)$

preview to GLMs: for many models,

$H(\beta) = -X^T W X$

W: related to variance of Y_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).

Checking the solution is a unique maximum