Lecture 6: Maximum likelihood estimation for logistic regression

Newton's method for logistic regression

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

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

Hessian $= \frac{\partial u}{\partial \beta} = H(\beta)$

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

3) Stop when algorithm converges

(usually when $\ell(\beta)$ stops changing, or changes by a very small amount)

Example

multiplication: %o * %o

inverse: solve(...)

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).

$\beta^{(r+1)} = ?$

$$\beta^{(t+1)} = \begin{bmatrix} -3.1 \\ 0.9 \end{bmatrix} + \begin{bmatrix} 17.834 & 53.218 \\ 53.218 & 180.718 \end{bmatrix}^{-1} \begin{bmatrix} 9.16 \\ 31.91 \end{bmatrix}$$

$$= \begin{bmatrix} -3.21 \\ 1.11 \end{bmatrix}$$

(Actual MLE (from glm function) is $\begin{bmatrix} -3.36 \\ 1.17 \end{bmatrix}$,

So we got closer!)

Newton's method for logistic regression

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

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

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & x_{21} & \ddots \\ 1 & \vdots & \end{bmatrix}$$

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

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

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

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

$$= \underbrace{\left(-\frac{\partial p}{\partial \beta}\right)}_{\downarrow W} \cdot X$$

Want $\frac{\partial p}{\partial \beta} = \left[\frac{\partial p_1}{\partial \beta} \quad \frac{\partial p_2}{\partial \beta} \quad \cdots \quad \frac{\partial p_n}{\partial \beta} \right] \in \mathbb{R}^{(k+1) \times n}$

($k+1$: # βs; n: # observations)

Columns of $\frac{\partial p}{\partial \beta}$

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

$$g'(u) = \frac{e^u}{(1+e^u)^2} \qquad \frac{\partial}{\partial \beta} \beta^T x_i = x_i$$

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

$$= p_i(1-p_i) x_i$$

Chain rule for matrix derivatives:
If x is a vector, and $h(x) \in \mathbb{R}$ and $g(h(x)) \in \mathbb{R}$, then

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

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

$$= \boxed{-X^T W X}$$

$W = \text{diag}(p_i(1-p_i))$

Comparison:

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

(σ^2 ↑ $Var(\varepsilon)$)

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

$W = diag(\lambda_i)$

Preview to GLMs:

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

$W = diag\left(\frac{Var(Y_i)}{\phi^2}\right)$

ϕ = dispersion parameter

Bernoulli: $\phi = 1$

Poisson: $\phi = 1$

Normal: $\phi = \sigma^2$

Checking the solution is a unique maximum

Newton's method finds β^* s.t. $U(\beta^*) = 0$

How do we know that β^* maximizes the likelihood?

Important property: If $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable then x^* is a global minimizer of f if and only if

$$\frac{\partial f}{\partial x}\Big|_{x=x^*} = 0$$

$\Rightarrow$ if $-\ell(\beta|X,Y)$ is a convex function of β (so $\ell(\beta|X,Y)$ is concave)

then β^* maximizes $\ell(\beta|X,Y)$

if & only if $U(\beta^*) = 0$

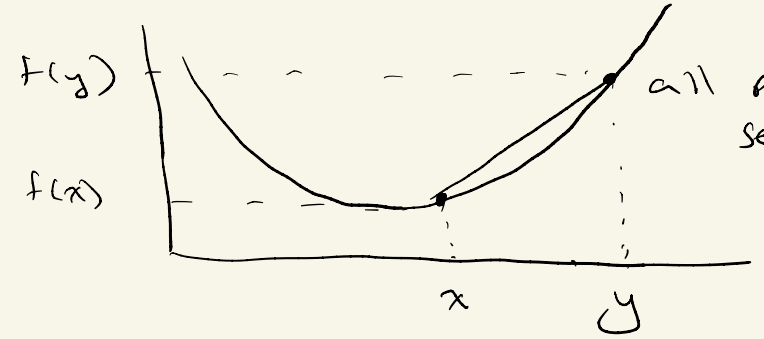
$\Rightarrow$ Newton's method gives us the MLE if $-\ell(\beta|X,Y)$ is convex!

Convex functions

Def: $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\forall\, x, y \in \mathbb{R}^n$ and $\forall\ 0 \le \lambda \le 1$,

$$f(\underbrace{\lambda x + (1-\lambda) y}_{\text{convex combination}}) \le \lambda f(x) + (1-\lambda) f(y)$$

e.g.



all points on the line segment lie above the graph of f

Theorem: A twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the Hessian $H(x)$ is positive-semidefinite $\forall\, x \in \mathbb{R}^n$

i.e. $y^T H(x)\, y \ge 0 \quad \forall\, y \in \mathbb{R}^n$

Claim: for logistic regression, $-H(\beta)$ is positive semi-definite

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

Let $v \in \mathbb{R}^{k+1}$ ($\beta \in \mathbb{R}^{k+1}$)

$$v^T X^T W X v = (Xv)^T W (Xv)$$

Let $s = Xv \in \mathbb{R}^n$

$$s^T W s = \sum_{i=1}^{n} \underbrace{p_i(1-p_i)}_{\in (0,1)} s_i^2 \qquad p_i \in (0,1)$$

$s_i^2 \geq 0$

$$\geq 0$$

$\Rightarrow -H(\beta)$ is PSD, so $-\ell(\beta|X,Y)$ is convex

$\Rightarrow \ell(\beta|X,Y)$ is concave $\Rightarrow$ Solving $u(\beta)$ gives MLE //