Maximum likelihood estimation for logistic regression

Recap: Newton's method

Want to find β^* st $U(\beta^*) = 0$

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

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

$= \text{Hessian} = \frac{\partial U(\beta^{(r)})}{\partial \beta^{(r)}}$

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

Example

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

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

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

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).

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

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

Actual MLE : $\begin{bmatrix} -3.36 \\ 1.17 \end{bmatrix}$

So we got closer!

Newton's method for logistic regression

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

$$u(\beta) = X^T(Y-p) \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots \\ 1 & x_{12} & \\ 1 & \vdots & \end{bmatrix}$$

$$\Rightarrow u(\beta^{(t)}) = X^T(Y-p^{(t)}) \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

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

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

$$= \frac{\partial^2 \ell(\beta)}{\partial \beta^2} = -\frac{\partial}{\partial \beta} X^T p$$

$$-\frac{\partial}{\partial \beta} X^T p = -\frac{\partial p}{\partial \beta} X \qquad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

$$\frac{\partial p}{\partial \beta} = \begin{bmatrix} \frac{\partial p_1}{\partial \beta} & \frac{\partial p_2}{\partial \beta} & \cdots & \frac{\partial p_n}{\partial \beta} \end{bmatrix}$$

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

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

$$f(\beta) = \beta^T X_i$$

$$\frac{\partial f}{\partial \beta} = X_i$$

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

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

$$= p_i(1-p_i) X_i$$

Rules for matrix derivatives

- if x and $v(x)$ are vectors, and A is a matrix which does not depend on x, then

$$\frac{\partial A v(x)}{\partial x} = \frac{\partial v(x)}{\partial x} A^T$$

- If x is a vector, $f(x) \in \mathbb{R}$, and $g(f(x)) \in \mathbb{R}$, then

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

$$H(\beta) = -\left[p_1(1-p_1) X_1 \quad p_2(1-p_2) X_2 \quad \cdots \quad p_n(1-p_n) X_n \right] X$$

$$= -\left[X_1 \cdots X_n \right] \begin{bmatrix} p_1(1-p_1) & & & \\ & p_2(1-p_2) & & \\ & & \ddots & \\ & & & p_n(1-p_n) \end{bmatrix} X = \boxed{-X^T W X}$$

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

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

$\sigma^2 \uparrow \text{Var}(\varepsilon)$

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

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

Newton's method for logistic regression

$U(\beta) = X^T(Y-p)$ $\quad$ $H(\beta) = -X^T W X$

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

2) $\beta^{(r+1)} = \beta^{(r)} + (X^T W^{(r)} X)^{-1} X^T (Y - p^{(r)})$

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

Checking the solution is a unique maximum

> Newton's method finds β^* such that $U(\beta^*) = 0$. 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 (so $\ell(\beta|X,Y)$ is concave), then β^* maximizes $\ell(\beta|X,Y)$ if and only if $U(\beta^*) = 0$

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

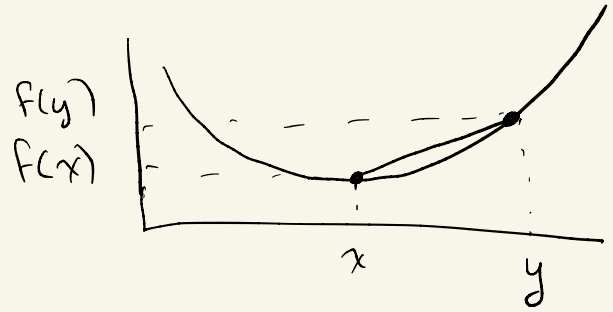
Convex functions

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

$$f(\underbrace{\lambda x + (1-\lambda) y}) \le \lambda f(x) + (1-\lambda) f(y)$$

convex combination

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

Pf: $H(\beta) = -X^T W X$ $\qquad W = \begin{bmatrix} p_1(1-p_1) & & & \\ & p_2(1-p_2) & & \\ & & \ddots & \\ & & & p_n(1-p_n) \end{bmatrix}$

$\Rightarrow -H(\beta) = X^T W X \in \mathbb{R}^{(k+1)\times(k+1)}$

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

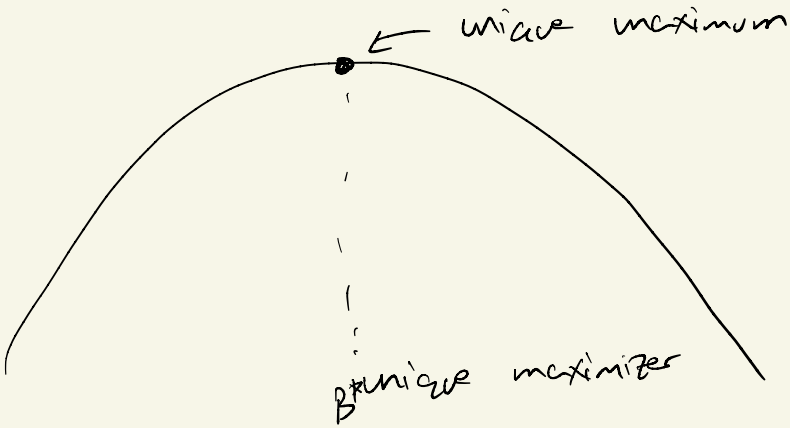
$-v^T H(\beta) v = v^T X^T W X v = (Xv)^T W (Xv)$

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

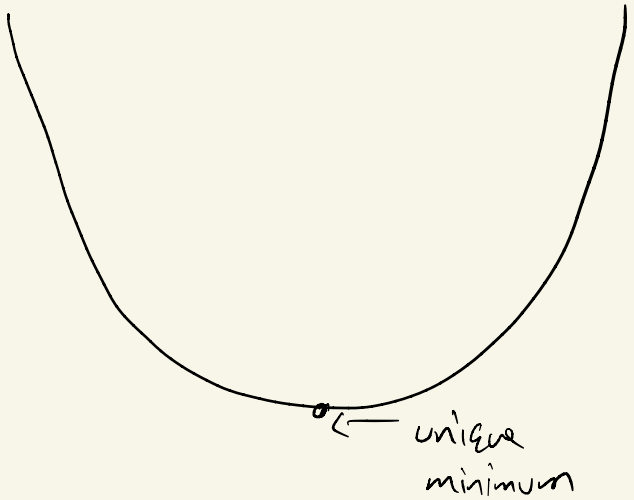
$\Rightarrow -v^T H(\beta) v = s^T W s$ $\qquad [s_1\ s_2 \cdots s_n] \begin{bmatrix} p_1(1-p_1) & & \\ & \ddots & \\ & & p_n(1-p_n) \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}$

$= \sum_{i=1}^{n} p_i(1-p_i) s_i^2 \geq 0$ //

$\Rightarrow$ $-\ell(\beta|X,Y)$ is convex, so solving $U(\beta) = 0$ gives us MLE of β



Concave function



convex function

$-\ell(\beta|X,Y)$

Some intuition about Hessians

Fisher information

Properties

Example