Fitting logistic regression models

Fisher scoring for logistic regression

Practice question: Fisher scoring

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

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

$$\mathcal{I}(\beta^{(r)}) = \begin{bmatrix} 17.834 & 53.218 \\ 53.218 & 180.718 \end{bmatrix}$$

> Use the Fisher scoring algorithm to calculate $\beta^{(r+1)}$ (you may use R or a calculator, you do not need to do the matrix arithmetic by hand). Take 2--3 minutes, then we will discuss.

Alternative to Fisher scoring: gradient ascent

$$Y_i \sim Bernoulli(p_i)$$

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

Choose $\beta = (\beta_0, \ldots, \beta_k)^T$ to maximize $L(\beta)$.

Gradient ascent:

Motivation for gradient ascent: walking uphill

Practice question: gradient ascent

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

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

- Use gradient ascent with a learning rate (aka step size) of $\gamma = 0.01$ to calculate $\beta^{(r+1)}$.
- The actual maximum likelihood estimate is $\widehat{\beta} = (-3.360, 1.174)$. Does one iteration of gradient ascent or Fisher scoring get us closer to the optimal $\widehat{\beta}$?
- Discuss in pairs for 2--3 minutes.

Gradient ascent vs. Fisher scoring

Special topic: Feedforward neural networks

Fitting neural networks: stochastic gradient descent