Lecture 5: Maximum likelihood estimation for logistic regression

Invariance of the MLE

Last time: $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Q: what if we want $\hat{\sigma}$?

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n} \sum_i (Y_i - \bar{Y})^2}$$

Theorem (invariance of the MLE): (see Thm 7.2.10 in CB)

Let $\hat{\theta}$ be the MLE of θ. For any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$

Logistic regression

$$Y_i \sim \text{Bernoulli}(p_i)$$

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

Suppose we observe independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. Write down the likelihood function

$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim}$ from joint distribution

Y_i are not iid

(b/c distribution of Y_i depends on X_i)

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

for the logistic regression problem.

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

$$= \left(\prod_{i=1}^{n} f(X_i) \right) \left(\prod_{i=1}^{n} f(Y_i \mid X_i, \beta) \right)$$

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

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

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

$$= \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i}$$

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

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

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

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

$$\frac{\partial \ell}{\partial \beta} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_K} \end{pmatrix}$$

<u>Rule for matrix derivatives</u>:

$$\frac{\partial}{\partial \beta} \beta^T X_i = X_i \qquad X_i = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{iK} \end{pmatrix}$$

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

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

$$= \sum_i (Y_i - p_i) X_i = X^T (Y - p)$$

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1K} \\ 1 & X_{21} & \cdots & X_{2K} \\ \vdots & & & \vdots \\ 1 & & & X_{nK} \end{bmatrix} \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

Aside: $X^T(Y-\mu) \overset{set}{=} 0$ (linear regression)

$\mu = X\beta$

$X^T(Y - X\beta) = 0$

$X^TY - X^TX\beta = 0 \Rightarrow (X^TX)^{-1}X^TY = \hat{\beta}$

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

$$X^T\left(Y - \frac{e^{X\beta}}{1+e^{X\beta}}\right) \overset{set}{=} 0$$

want to find β^* st $U(\beta^*) = 0$

no closed-form solution for logistic regression model

Idea: 1) Start w/ an initial guess $\beta^{(0)}$

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

3) Iterate!

Iterative methods for maximizing likelihood

Newton's method

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

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

$$\underset{\displaystyle 0}{\underset{\shortparallel}{u(\beta^*)}} \approx u(\beta^{(0)}) + \frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} (\beta^* - \beta^{(0)})$$

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

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

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

$$U(\beta) = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \end{pmatrix} = \frac{\partial \ell}{\partial \beta}$$

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

Hessian of log likelihood $H(\beta)$

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

Newton's method for logistic regression

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).