Maximum likelihood estimation for regression models

Ciaran Evans

Warmup

Work on the warmup activity (handout), then we will discuss as a class.

Warmup

Suppose that we have independent observations $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ from the model

$$Y_i | \mathbf{x}_i \sim N(\mathbf{x}_i^T \beta, \sigma_i^2).$$

Suppose that the variances $\sigma_1^2, \ldots, \sigma_n^2$ are known. Show that the maximum likelihood estimator of β minimizes the weighted sum of squares

$$WSS(\beta) = \sum_{i=1}^{n} w_i (Y_i - \mathbf{x}_i^T \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta)$$

$$L(\beta | y, X) \propto \prod_{i=1}^{n} f(y_i | x_i, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{1}{2\sigma_i^2} (y_i - x_i^T \beta)^2 \right\}$$

$$= \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - x_i^T \beta)^2 \right\}$$

$$= \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2 \right\} \qquad w_i = \frac{1}{\sigma_i^2}$$

(inverse variance weighting)

$\Rightarrow$ maximizing L $\Leftrightarrow$ minimizing WSS!

Maximum likelihood estimation and logistic regression

Let $(\mathbf{x}_1, Y_1), ..., (\mathbf{x}_n, Y_n)$ be iid samples from the model

$$Y_i|\mathbf{x}_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\beta$$

where the distribution of $\mathbf{x}_i$ does not depend on β.

$$\ell(\beta|\mathbf{y}, \mathbf{X}) \overset{\text{(up to a constant)}}{=} \sum_{i=1}^{n}\left\{Y_i\mathbf{x}_i^T\beta - \log(1+e^{\mathbf{x}_i^T\beta})\right\}$$

score function $\rightsquigarrow$ $U(\beta) = \frac{\partial\ell}{\partial\beta} = \mathbf{X}^T(\mathbf{y}-\mathbf{p})$

Newton's method

- Want β^* such that $U(\beta^*) = \mathbf{0}$
- Begin with initial estimate $\beta^{(0)}$
- Iterative updates:

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

$\beta = (\beta_0, \ldots, \beta_K)^T \in \mathbb{R}^{K+1}$

$u(\beta) = \frac{\partial \ell}{\partial \beta} = X^T(Y - p) \in \mathbb{R}^{K+1}$

$H(\beta) = \frac{\partial^2 \ell}{\partial \beta^2} \in \mathbb{R}^{(K+1)\times(K+1)}$

Hessian

The Hessian

$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \in \mathbb{R}^n$$

$$U(\beta) = \frac{\partial}{\partial \beta} \ell(\beta | \mathbf{y}, \mathbf{X}) = \mathbf{X}^T (\mathbf{y} - \mathbf{p})$$

$$\mathbf{H}(\beta) = \frac{\partial}{\partial \beta} U(\beta) = \frac{\partial}{\partial \beta} \mathbf{X}^T (\mathbf{y} - \mathbf{p}) = \left(-\frac{\partial \mathbf{p}}{\partial \beta} \right) \mathbf{X}$$

$$\frac{\partial \mathbf{p}}{\partial \beta} = \begin{bmatrix} \frac{\partial p_1}{\partial \beta} & \frac{\partial p_2}{\partial \beta} & \cdots & \frac{\partial p_n}{\partial \beta} \end{bmatrix} \in \mathbb{R}^{(k+1) \times n}$$

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

Chain rule: $\frac{\partial}{\partial \beta} g(h(\beta)) = g'(h(\beta)) \frac{\partial h(\beta)}{\partial \beta}$

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

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

The Hessian

$$U(\beta) = \frac{\partial}{\partial \beta} \ell(\beta | \mathbf{y}, \mathbf{X}) = \mathbf{X}^T (\mathbf{y} - \mathbf{p})$$

$$\mathbf{H}(\beta) = \frac{\partial}{\partial \beta} U(\beta) = \frac{\partial}{\partial \beta} \mathbf{X}^T (\mathbf{y} - \mathbf{p}) = \left(-\frac{\partial \mathbf{p}}{\partial \beta} \right) \mathbf{X}$$

$$\frac{\partial \mathbf{p}}{\partial \beta} = \begin{bmatrix} \frac{\partial p_1}{\partial \beta} & \frac{\partial p_2}{\partial \beta} & \cdots & \frac{\partial p_n}{\partial \beta} \end{bmatrix} \in \mathbb{R}^{(k+1) \times n}$$

$$\frac{\partial p_i}{\partial \beta} = p_i (1 - p_i) x_i$$

$$\Rightarrow \frac{\partial p}{\partial \beta} = \begin{bmatrix} p_1 (1-p_1) x_1 & p_2 (1-p_2) x_2 & \cdots & p_n (1-p_n) x_n \end{bmatrix}$$

$$= X^T W \qquad W = \begin{bmatrix} p_1 (1-p_1) & & & 0 \\ & p_2 (1-p_2) & & \\ & & \ddots & \\ 0 & & & p_n (1-p_n) \end{bmatrix}$$

(diagonal matrix)

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

Putting everything together

Want to maximize the log likelihood $\ell(\beta|\mathbf{y}, \mathbf{X})$.

- begin with initial guess $\beta^{(0)}$
- update:

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

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

$$p^{(r)} = (p_1^{(r)}, \cdots, p_n^{(r)})^T$$

$$W^{(r)} = \text{diag}\left(p_i^{(r)}(1 - p_i^{(r)})\right)$$

Class activity

Work on the class activity:

https://sta711-s26.github.io/class_activities/ca_07_2.html

Submit your work on Canvas.