Lecture 9: Fisher information and Fisher scoring

Recap: Fisher information

imagine $u(\beta) = X^T(Y - p)$

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

For a given β, $u(\beta)$ will vary from dataset to dataset

Def: Let $\ell(\theta \mid Y)$ be a log-likelihood and $u(\theta) = \frac{\partial \ell}{\partial \theta}$. The Fisher information is $\mathcal{I}(\theta) = \text{Var}(u(\theta) \mid \theta)$

depends on observed $Y_1, \ldots, Y_n$ (and so $u(\theta)$ is a random variable)

Last time: Started to see how $\mathcal{I}^{-1}(\theta)$ is related to variance of MLE $\hat{\theta}$

Properties

Let Y be a r.v. w/ prob. function f

$\mathbb{E}[Y] = \int y f(y) dy \qquad \mathbb{E}[g(Y)] = \int g(y) f(y) dy$

Under appropriate regularity conditions,

$$\mathbb{E}[U(\theta)|\theta] = 0$$

Proof:

$$\mathbb{E}[U(\theta)|\theta] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ell(\theta|Y) \,\Big|\, \theta\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \log f(Y|\theta) \,\Big|\, \theta\right]$$

$$= \mathbb{E}\left[\frac{1}{f(Y|\theta)} \cdot \frac{\partial}{\partial\theta} f(Y|\theta) \,\Big|\, \theta\right] = \int \left(\frac{\partial}{\partial\theta} f(y|\theta)\right) \cdot \frac{1}{\cancel{f(y|\theta)}} \cdot \cancel{f(y|\theta)} \, dy$$

$$= \int \left(\frac{\partial}{\partial\theta} f(y|\theta)\right) dy$$

$$= \frac{\partial}{\partial\theta} \underbrace{\int f(y|\theta) dy}_{1}$$

$$= 0$$

required regularity conditions: swap derivative and integral

(see Casella & Berger 2.4)

//

Properties

$$-\mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta^2} f(Y|\theta)}{f(Y|\theta)} \,\middle|\, \theta\right] = -\int \left(\frac{\partial^2}{\partial\theta^2} f(Y|\theta)\right) \cdot \frac{1}{f(Y|\theta)} \cdot f(Y|\theta)$$

$$\overset{\text{(regularity)}}{=} -\frac{\partial^2}{\partial\theta^2}\int f(y|\theta)\,dy = 0$$

Under appropriate regularity conditions,

$$\mathrm{Var}(u(\theta)|\theta) = \mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta|\mathbf{Y}) \,\middle|\, \theta\right]$$

Proof:

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta|Y) \,\middle|\, \theta\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(Y|\theta) \,\middle|\, \theta\right]$$

$$= -\mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\log f(Y|\theta)\right) \,\middle|\, \theta\right]$$

$$= -\mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta} f(Y|\theta)}{f(Y|\theta)}\right) \,\middle|\, \theta\right]$$

$$= -\mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta^2} f(Y|\theta)}{f(Y|\theta)} - \left(\frac{\frac{\partial}{\partial\theta} f(Y|\theta)}{f(Y|\theta)}\right)^2 \,\middle|\, \theta\right]$$

$$= \underbrace{-\mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta^2} f(Y|\theta)}{f(Y|\theta)} \,\middle|\, \theta\right]}_{0} + \underbrace{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(Y|\theta)\right)^2 \,\middle|\, \theta\right]}_{\mathbb{E}[(u(\theta))^2|\theta]} = \mathrm{Var}(u(\theta)|\theta) \;//$$

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

$$\mathrm{Var}(u(\theta)) = \mathbb{E}[u(\theta)^2] - \underbrace{\mathbb{E}(u(\theta))^2}_{0}$$

Fisher scoring

Newton's method: $\beta^{(r+1)} = \beta^{(r)} \underbrace{- H^{-1}(\beta^{(r)})}_{\text{replaced w/ } \mathcal{I}^{-1}(\beta^{(r)})} U(\beta^{(r)})$

under regularity conditions

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

2) Update: $\beta^{(r+1)} = \beta^{(r)} + \mathcal{I}^{-1}(\beta^{(r)}) U(\beta^{(r)})$

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

(Iteratively reweighted least squares)

IRLS for logistic regression

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

$\boxed{\text{Recall: linear regression } \hat{\beta} = (X^T X)^{-1} X^T Y}$

Rewrite: $\beta^{(r+1)} = \underbrace{(X^T W^{(r)} X)^{-1} (X^T W^{(r)} X)}_{\text{(identity matrix)}} \beta^{(r)} + (X^T W^{(r)} X)^{-1} X^T (Y - p^{(r)})$

$$= (X^T W^{(r)} X)^{-1} X^T W^{(r)} \underbrace{\left(X\beta^{(r)} + (W^{(r)})^{-1} (Y - p^{(r)}) \right)}_{Z^{(r)}}$$

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

working responses at iteration r

Actually doing weighted least squares w/ weights $W^{(r)}$, responses $Z^{(r)}$, and design matrix X

Weighted least squares

Suppose $Y = X\beta + \varepsilon$

$$\underbrace{W^{\frac{1}{2}}Y}_{} = W^{\frac{1}{2}}X\beta + W^{\frac{1}{2}}\varepsilon$$

$$Y_w = X_w\beta + \varepsilon_w$$

$$\Rightarrow \hat{\beta} = (X_w^T X_w)^{-1} X_w^T Y_w$$

$$= (X^T W X)^{-1} X^T W Y$$

Use inverse-variance weights!

usual assumption $\varepsilon \sim N(0, \sigma^2 I)$

If I have non-constant variance:

$$\varepsilon \sim N(0, W^{-1})$$

$$W = \mathrm{diag}(w_1, \dots, w_n)$$

$$\mathrm{Var}(\varepsilon_i) = \frac{1}{w_i}$$

$$W^{\frac{1}{2}} = \mathrm{diag}(\sqrt{w_1}, \dots, \sqrt{w_n})$$

$$\varepsilon_w \sim N(0, I)$$

$$\mathrm{Var}(W^{\frac{1}{2}}\varepsilon) = W^{\frac{1}{2}}W^{-1}W^{\frac{1}{2}} = I$$

$$\mathrm{Var}(AX) = A\,\mathrm{Var}(X)\,A^T$$