Fisher scoring

Recap: Fisher information

Def: Let $\ell(\theta|Y)$ be a log likelihood, and $u(\theta) = \frac{\partial \ell}{\partial \theta}$. The Fisher information is

$$\mathcal{I}(\theta) = \text{Var}(u(\theta)|\theta)$$

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

$$g(Y) = U(\theta)$$

Properties

Theorem: Under appropriate regularity conditions,

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

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

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

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

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

$= \frac{\partial}{\partial\theta} 1 = 0$ ✓

requires regularity: swap derivative and integral (see Casella & Berger 2.4)

$$\frac{\partial}{\partial\theta}\log f(Y|\theta) = \frac{\frac{\partial}{\partial\theta} f(Y|\theta)}{f(Y|\theta)}$$

Properties

Theorem: Under appropriate regularity conditions,

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

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

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

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

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

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

$= \mathbb{E}[(u(\theta))^2|\theta]$

$= \mathrm{Var}(u(\theta)|\theta)$ $\quad (\mathbb{E}[u(\theta)|\theta] = 0)$

Fisher information vs. Hessian

For logistic regression:

$$\mathcal{I}(\beta) = X^T W X = -H(\beta)$$

under regularity conditions, $\mathcal{I}(\beta) = -\mathbb{E}[H(\beta)]$

$\mathcal{I}$ and $-H$ are not always the same

Example: $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p)$

$$\frac{\partial^2}{\partial p^2} \ell(p|Y) = -\frac{\sum_i Y_i}{p^2} - \frac{(n - \sum_i Y_i)}{(1-p)^2}$$

← depends on the observed data

$$\mathcal{I}(p) = -\mathbb{E}\left[\frac{\partial^2}{\partial p^2} \ell(p|Y)\right] = \frac{n}{p(1-p)}$$

← does not depend on the observed data

In general, $H(\theta)$ can be more work to calculate

$\Rightarrow$ typically use $\mathcal{I}(\theta)$ in place of $H(\theta)$

Newton's method: $\beta^{(r)} - H^{-1}(\beta^{(r)})\, u(\beta^{(r)})$

Fisher scoring

replace w/ $\mathcal{I}^{-1}(\beta^{(r)})$

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

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

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

IRLS for logistic regression

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

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

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

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

↖ working responses at iteration r

Linear regression: $\hat{\beta} = (X^T X)^{-1} X^T Y$

- actually doing weighted least squares with weights $W^{(r)}$, responses $Z^{(r)}$, explanatory variables X

Suppose $Y = X\beta + \varepsilon$

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

$$Y_w \qquad X_w \qquad \varepsilon_w$$

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

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

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

$\varepsilon \sim N(0, W^{-1})$ (allow non-constant variance)

$W = \text{diag}(w_1, \ldots, w_n)$ $\quad W^{\frac{1}{2}} = \text{diag}(\sqrt{w_i})$

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

$\text{var}(W^{\frac{1}{2}} \varepsilon) = W^{\frac{1}{2}} W^{-1} W^{\frac{1}{2}} = \underline{I}$

A preview of Fisher information properties