Lecture 8: Fisher Information

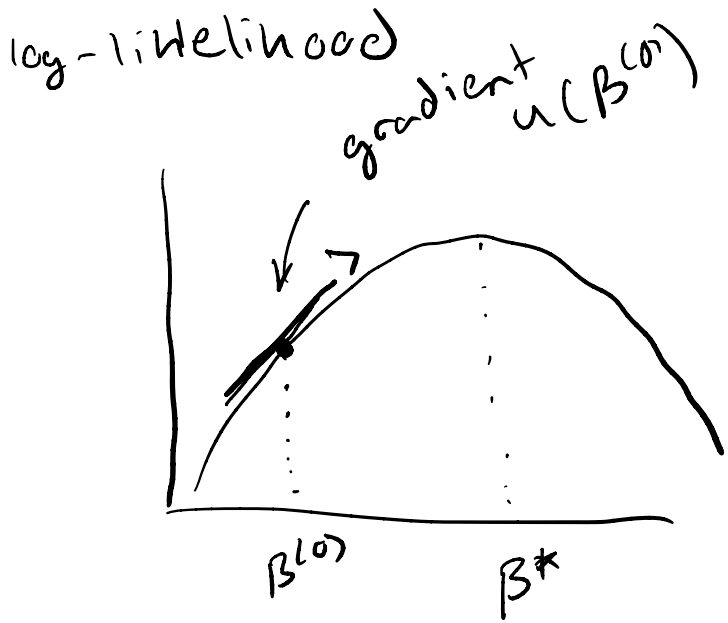
Ciaran Evans

Recap: Newton's method

To find β^* such that $U(\beta^*) = 0$, when there is no closed-form solution we use Newton's method:

- Begin with an initial guess $\beta^{(0)}$
- Iteratively update: $\beta^{(r+1)} = \beta^{(r)} - \mathbf{H}^{-1}(\beta^{(r)})U(\beta^{(r)})$
- Stop when the algorithm converges

Intuition:



use gradient to walk uphill

Hessian tells us how far along gradient to move

Some intuition about Hessians

Example: Suppose that $\beta = (\beta_0, \beta_1)^T \in \mathbb{R}^2$, and

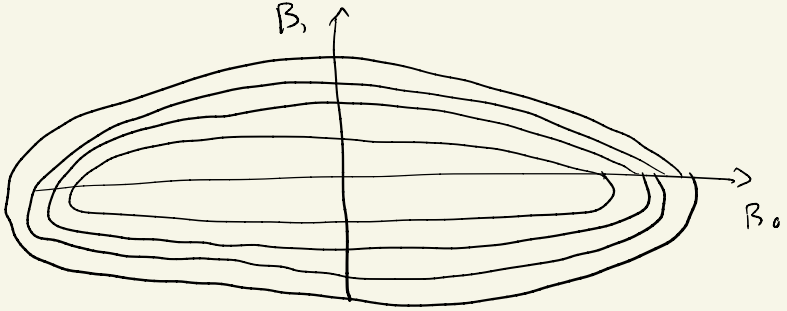
$$\ell(\beta) = -\beta_0^2 - 100\beta_1^2$$

Calculate

$$U(\beta) = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \end{bmatrix} \qquad \mathbf{H}(\beta) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} \end{bmatrix}$$

$$= \begin{bmatrix} -2\beta_0 \\ -200\beta_1 \end{bmatrix} \qquad \begin{bmatrix} -2 & 0 \\ 0 & -200 \end{bmatrix}$$

Level curves $\ell(\beta) = -\beta_0^2 - 100\beta_1^2$ (maximum at $(0,0)$)



a small change in β_1 makes a bigger difference to $\ell(\beta)$ than a small change in β_0

Two different approaches to moving along gradient

gradient ascent: step size

$$\beta^{(r+1)} = \beta^{(r)} + \alpha \nabla(\beta^{(r)})$$

Example: $\beta^{(r)} = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$

$$\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} + \alpha \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix}$$

$\alpha = \frac{1}{200}$

$$\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} + \frac{1}{200} \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix}$$

$$= \begin{bmatrix} 0.1 - 0.001 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} + \alpha \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix}$$

$\alpha = \frac{1}{200}$

$\Rightarrow$ update $= \begin{bmatrix} 0.099 \\ 0 \end{bmatrix}$

$\alpha = \frac{1}{2}$

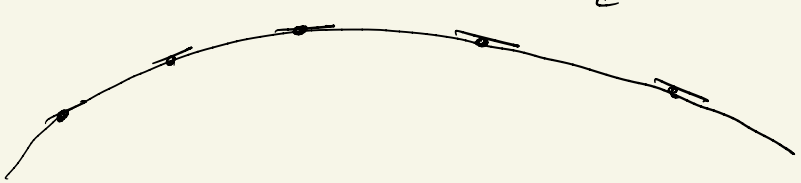
$$\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0.1 - 10 \end{bmatrix} = \begin{bmatrix} 0 \\ -9.9 \end{bmatrix}$$

Using Hessian: $H = \begin{bmatrix} -2 & 0 \\ 0 & -200 \end{bmatrix}$ $H^{-1} = \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & -\frac{1}{200} \end{bmatrix}$

$$\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} - \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & -\frac{1}{200} \end{bmatrix} \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Newton's method is faster (fewer iterations) by using info about second derivative

one more perspective



Slopes $U(\beta)$ close to 0

less variability in slopes

(most values close to 0)

Slope $U(\beta)$ are far from 0

more variability in slopes

(some values near 0, some values far from 0)

matrix A, random vector Y

$\text{Var}(AY) = A\,\text{Var}(Y)\,A^T$

$\Rightarrow$ look at variance of $U(\beta)$ (as a function of the data)

Logistic regression: $U(\beta) = X^T(Y-p)$

$$\text{Var}(U(\beta)) = \text{Var}(X^T(Y-p)) = X^T\,\text{Var}(Y-p)\,X$$

$$= X^T\,\text{Var}(Y)\,X = X^T \begin{bmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & & \\ \vdots & & \ddots & \\ 0 & & & p_n(1-p_n) \end{bmatrix} X = X^T W X$$

$\text{Var}(Y_i) = p_i(1-p_i)$

$= -H(\beta)$

Fisher information

Def : Let $\ell(\theta|Y)$ be a log-likelihood given data Y, and θ some parameter of interest.

Let $u(\theta) = \frac{\partial \ell}{\partial \theta}$

The **Fisher information** is

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

variance of score $u(\theta)$, if θ is the true parameter

Thm : Under certain regularity conditions,

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

Example: Bernoulli sample

$$\mathcal{I}(p) = \frac{n}{p(1-p)}$$

Suppose that $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(p)$.

$$L(p|Y) = p^{\sum_i Y_i} (1-p)^{n - \sum_i Y_i}$$

$$\ell(p|Y) = \left(\sum_i Y_i\right) \log p + \left(n - \sum_i Y_i\right) \log(1-p)$$

$$u(p) = \frac{\partial}{\partial p} \ell(p|Y) = \frac{\sum_i Y_i}{p} - \frac{\left(n - \sum_i Y_i\right)}{1-p}$$

$$Var(u(p)|p) = Var\left(\frac{\sum_i Y_i}{p} - \frac{\left(n - \sum_i Y_i\right)}{1-p} \right)$$

$$= Var\left(\frac{\left(\sum_i Y_i\right)(1-p)}{p(1-p)} - \frac{\left(n - \sum_i Y_i\right)(p)}{p(1-p)} \right)$$

$$= Var\left(\frac{\sum_i Y_i}{p(1-p)} \right) = \frac{1}{p^2(1-p)^2} \sum_i Var(Y_i) = \frac{n\,p(1-p)}{p^2(1-p)^2}$$

$$= \frac{n}{p(1-p)}$$

MLE: $\hat{p} = \frac{1}{n} \sum_i Y_i$

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{1}{n} \sum_i Y_i\right)$$

$$= \frac{1}{n^2} \sum_i \text{Var}(Y_i)$$

$$= \frac{1}{n^2} \cdot n\, p(1-p)$$

$$= \frac{p(1-p)}{n}$$

$$\mathcal{I}(p) = \frac{n}{p(1-p)}$$

$$\text{Var}(\hat{p}) = \mathcal{I}^{-1}(p)$$

Linear regression: $\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \mathcal{I}^{-1}(\beta)$

Preview: variance of MLEs is closely connected to $\mathcal{I}^{-1}$

Properties