# Maximum Likelihood

Recall that if we have data $X = (X_1, \dots, X_n)$ with $X_i \sim f_\theta$ then the joint likelihood is written as

$$f(x_1, \dots, x_n; \theta).$$

One estimate for $\theta$ is

$$\hat{\theta}_{MLE} \equiv \underset{\theta \in \Theta}{\operatorname{argmax}} \; f(x_1, \dots, x_n; \theta)$$

To stress that we maximize the joint distribution in terms of $\theta$ we define the (log-) likelihood function

$$\mathcal{L}(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta)$$

$$\ell(\theta | X) \doteq \log \mathcal{L}(\theta; x_1, \dots, x_n)$$

<u>Rmk</u>: $\mathcal{L} \geq 0$ $\log$ monotone implies that

[illegible] $\ell(\theta | X) = \hat{\theta}$

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta \mid x) = \hat{\theta}_{MLE}$$

So we can find $\hat{\theta}_{MLE}$ by solving

$$\ell'(\hat{\theta}_{MLE}) = 0$$

<u>Rmk:</u> We normally say $\ell'(\theta)$ is the <u>score</u> function and under regularity conditions

$$I(\theta) = \mathbb{E}\left(\ell'(\theta)\,\ell'(\theta)^T\right)$$

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

which we call the <u>Fisher Information</u>.

Under further regularity conditions

$$\mathbb{E}[\ell'(\theta)] = 0 \qquad I(\theta) = \operatorname{Var}[\ell'(\theta)]$$

$$\hat{\theta}_{MLE} \xrightarrow{D} N\left(\theta, I(\theta)^{-1}\right)$$

Ex: $X \sim N(\mu, \sigma^2 I)$, $\sigma^2$ known

Ex. $y \sim N(X\beta, \sigma^2 I_n)$ $\quad \sigma^2$ known

$$\mathcal{L}(\beta; Y, X) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right\}$$

$$\ell(\beta; Y, X) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)$$

$$\propto \frac{-1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)$$

$$\hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}} \frac{-1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)$$

$$= \underset{\beta}{\operatorname{argmin}} (Y-X\beta)^T(Y-X\beta) \qquad \text{(OLS criterion)}$$

$$= \hat{\beta}_{OLS}$$

Now recall that

$$\frac{\partial a^T x}{\partial x} = a \qquad \frac{\partial x^T a}{\partial x} = a \qquad \frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

So regarding the form of the solution.

$$\frac{d\ell}{d\beta} = \frac{-1}{2\sigma^2}\,\partial\left\{Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta\right\}$$

$$\partial\beta \qquad \frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}\left(Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta\right)$$

$$= \frac{-1}{2\sigma}\left\{-X^TY - X^TY + 2X^TX\beta\right\}$$

$$= \frac{1}{\sigma^2}\left\{X^TY - X^TX\beta\right\} \overset{}{\underset{\text{set}}{=}} 0$$

$$X^TX\beta = X^TY \qquad \text{(normal equations)}$$

$$\hat{\beta}_{MLE} = (X^TX)^{-1}X^TY$$

<u>Ex</u>: Suppose $X_1, \ldots, X_n \sim \text{Pois}(\lambda)$

$$\mathcal{L}(\lambda; X) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum x_i}\, e^{-\lambda n}}{\prod_{i=1}^{n} x_i!}$$

$$\ell(\lambda; X) = \sum x_i \log\lambda - n\lambda - \log\prod x_i!$$

$$\frac{\partial\ell(\lambda;x)}{\partial\lambda} = \frac{\sum x_i}{\lambda} - n \overset{\text{set}}{=} 0$$

$$\hat{\lambda}_{MLE} = \overline{X}$$

**Ex:** $X_1, \dots, X_n \overset{iid}{\sim} \Gamma(\alpha, \beta)$

$$\mathcal{L}(\alpha, \beta; x) = \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}$$

$$\ell(\alpha, \beta; x) = \sum_{i=1}^{n} \left\{ \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x_i - \beta x_i \right\}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left\{ \frac{\alpha}{\beta} - x_i \right\} \overset{set}{=} 0 \implies \hat{\beta}_{MLE} = \frac{\hat{\alpha}_{MLE}}{\bar{X}}$$

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log x_i \right\} \overset{set}{=} 0$$

$$n \log \hat{\beta} - \psi(\alpha) + \sum \log x_i = 0$$

$$\log \hat{\beta} = \frac{\psi(\hat{\alpha})}{n} - \overline{\log x}$$

$$\log \hat{\alpha} - \psi(\hat{\alpha}) = \log \bar{x} - \overline{\log x}$$

Can't solve this analytically so we need a numerical method.

# Newton - Raphson Algorithm

* Numerical method for finding roots

Use it for $\ell'(\theta)=0$

With current estimate $\theta^{(t)}$ use the first order Taylor Expansion around $\theta^{(t)}$

$$\ell'(\theta) \simeq \ell'(\theta^{(t)}) + \ell''(\theta^{(t)})(\theta - \theta^{(t)})$$

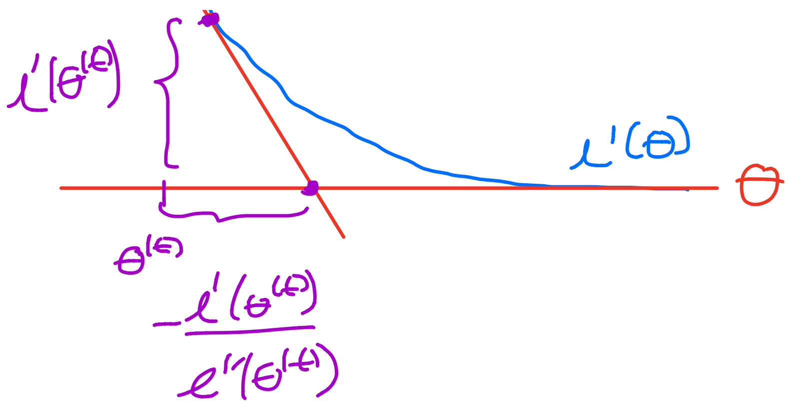
Goal: Solve $\ell'(\theta)=0$ by iteratively solving the first order problem

Solving for $\theta$

$$0 = \ell'(\theta^{(t)}) + \ell''(\theta^{(t)})(\theta - \theta^{(t)})$$

gives the update

$$\boxed{\theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})}}$$




The multivariate version

→ likelihood is still a number so Taylor works

→ $\ell'(\theta^{(t)}) \in \mathbb{R}^n$ gradient

→ $\ell''(\theta^{(t)}) \in \mathbb{R}^{n \times n}$ Hessian

$$\theta^{(t+1)} = \theta^{(t)} - \left[\ell''(\theta^{(t)})\right]^{-1} \ell'(\theta^{(t)})$$

$$\ell'(\theta) = \nabla_\theta \ell = \begin{bmatrix} \partial \ell / \theta_1 \\ \vdots \\ \partial \ell / \theta_m \end{bmatrix}$$

$$\ell''(\theta) = \frac{\partial \ell}{\partial \theta \partial \theta^T} = \left( \frac{\partial \ell}{\partial \theta_i \partial \theta_j} \right)_{1 \le i,j \le m}$$