Lecture 5: Maximum likelihood estimation

Ciaran Evans

Recap: maximum likelihood estimation

Definition: Let $\mathbf{Y} = (Y_1, ..., Y_n)$ be a sample of n observations, and let $f(\mathbf{y}|\theta)$ denote the joint pdf or pmf of $\mathbf{Y}$, with parameter(s) θ. The *likelihood function* is

$$L(\theta|\mathbf{Y}) = f(\mathbf{Y}|\theta)$$

Definition: Let $\mathbf{Y} = (Y_1, ..., Y_n)$ be a sample of n observations. The *maximum likelihood estimator* (MLE) is

$$\widehat{\theta} = \text{ argmax}_\theta \; L(\theta|\mathbf{Y})$$

Example: $N(\mu, \sigma^2)$

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

$$Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

$$\mu \in (-\infty, \infty)$$

$$\sigma^2 > 0$$

$$L(\theta \mid Y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \right\}$$

$$\Rightarrow \ell(\theta \mid Y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2$$

Plan: maximize wrt μ, then maximize wrt σ^2

$$\frac{\partial}{\partial \mu} \ell(\theta \mid Y) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu) \overset{set}{=} 0$$

$$= \sum_{i=1}^{n} Y_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$= \bar{Y}$$

For any value of σ^2:

$$L(\theta|Y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2\right\}$$

$$\leq (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}$$

Now we want to maximize

$$(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}$$

$$l^*(\sigma^2|Y) = \underbrace{-\frac{n}{2}\log(2\pi\sigma^2)}_{-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

$$\frac{dl^*}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \overset{set}{=} 0$$

$$\Rightarrow \boxed{\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \\ \hat{\mu} &= \bar{Y}\end{aligned}}$$

Linear regression with normal errors

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

Suppose we observe independent samples $(X_1, Y_1), ..., (X_n, Y_n)$.
Write down the likelihood function

$$L(\beta|\mathbf{X}, \mathbf{Y}) \propto \prod_{i=1}^{n} f(Y_i|\beta, X_i)$$

$$L(\beta \mid X, Y) = f(X, Y \mid \beta) = \prod_{i=1}^{n} f(X_i, Y_i \mid \beta) \quad \text{(independence)}$$

(X_i, Y_i) are iid from the same joint distribution

but $Y_i \mid X_i$ is different depending on X_i

X_i doesn't depend on β

β describes $Y_i \mid X_i$

$$= \prod_{i=1}^{n} \underbrace{f(X_i \mid \beta)}_{f(X_i)} f(Y_i \mid X_i, \beta)$$

$$= \left(\prod_{i=1}^{n} f(X_i)\right)\left(\prod_{i=1}^{n} f(Y_i \mid X_i, \beta)\right)$$

$$\Rightarrow L(\beta \mid X, Y) \propto \prod_{i=1}^{n} f(Y_i \mid X_i, \beta)$$

$$f(Y_i \mid X_i, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \mu_i)^2 \right\}$$

μ_i can be different for each observation

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \beta^T X_i)^2 \right\}$$

$\mu_i = \beta_0 + \beta_1 X_{i1} + \ldots$

$= \beta^T X_i \qquad X_i = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix}$

$$L(\beta \mid X, Y) \propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 \right\}$$

maximize L wrt β: minimize $\sum_i (Y_i - \beta^T X_i)^2$ (SSE again!)

Maximizing L wrt β means

minimize $\sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 = \|Y - X\beta\|^2$

$= (Y - X\beta)^T (Y - X\beta)$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

Let $u = Y - X\beta$

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & & X_{nk} \end{pmatrix}$$

(design matrix)

Minimize $u^T u$

$\frac{\partial}{\partial \beta} u^T u$

Matrix chain rule:

$$\frac{\partial}{\partial \beta} u^T u = \underbrace{\left(\frac{\partial u}{\partial \beta} \right)}_{\frac{\partial}{\partial \beta}(Y - X\beta)} \underbrace{\left(\frac{\partial}{\partial u} u^T u \right)}_{2u}$$

$= -\frac{\partial}{\partial \beta}(X\beta) = -X^T$

$= -2X^T (Y - X\beta) \overset{\text{set}}{=} 0$

$\Rightarrow X^T (Y - X\beta) = 0$

$\Rightarrow X^T Y = X^T X \beta$

$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$

Logistic regression

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

Suppose we observe independent samples $(X_1, Y_1), ..., (X_n, Y_n)$.
Write down the likelihood function

$$L(\beta|\mathbf{X},\mathbf{Y}) \propto \prod_{i=1}^{n} f(Y_i|\beta, X_i)$$