#2 − F2017

Likelihood and Maximum Likelihood Estimates

- A r.v. Y_i has a density $f(y_i;\theta) = f(y_i|X_i^T,\theta)$, where X_i is deterministic.
- The joint density of $\mathbf{Y} = (Y_1, \quad, Y_n)$ is

$$f(\mathbf{y};\theta) = f(\mathbf{y}|\mathbf{X},\theta) = \prod_{i=1}^{n} f(y_i|X_i^T,\theta).$$

- The likelihood function of θ is denoted

$$L(\theta) = L(\theta|\mathbf{Y}) = \prod_{i=1}^{n} f(y_i|X_i^T,\theta).$$

- The log-likelihood is

$$l(\theta) = l(\theta|\mathbf{Y}) = \log L(\theta|\mathbf{Y}).$$

- NOTE: If X_i is random, need to consider a distribution of X_i too.

Maximum Likelihood Estimators (MLE)

- An *Maximum Likelihood Estimator* (MLE) is an maximizer of the likelihood function $L(\theta|\mathbf{Y})$, denoted as $\hat{\theta}$, i.e.

$$L(\hat{\theta}) = \max_{\theta\in\Theta} L(\theta),$$

  where Θ is a parameter space.

  - NOTE: MLE is also an maximizer of the log-likelihood, $l(\theta)$.
  - How to compute MLE? Solve the following equations:

$$\frac{\partial}{\partial\theta_j} l(\theta) = 0, \quad , j = 1, \quad , p,$$

  where $\theta = (\theta_1, \quad, \theta_p)$; or

$$\frac{\partial}{\partial\theta} l(\theta) = 0.$$

MLE: Example I

- $Y_i \sim Exp(\lambda)$, where $f(y_i;\lambda) = \lambda e^{-\lambda y_i}$ for $i = 1, \quad , n$. Find an MLE of λ.

For $Y_i \sim Exp(\lambda)$, $y_i \geq 0$

Likelihood function $L(\lambda) = \prod_{i=1}^{n} f(y_i;\lambda)$

$$\ell(\lambda) = \log L(\lambda) = \sum_{i=1}^{n} \log f(y_i;\lambda)$$
$$= \sum_{i=1}^{n} (\log\lambda - \lambda y_i)$$
$$= n\log\lambda \quad \lambda \sum_{i=1}^{n} y_i .$$

$$\frac{\partial \ell(\lambda)}{\partial\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i = 0 .$$

MLE: $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\bar{y}}$.

Let $\theta = \frac{1}{\lambda}$, then $E(Y_i) = \theta$, MLE of θ: $\hat{\theta} = \frac{1}{\hat{\lambda}} = \bar{y}$.

MLE: Example II

- Data: Tropical cyclones, see D&B, Table 1.2 on page 15, 3rd Edtion.
- The table shows the number of tropical cyclones in Northeastern Australia in 13 successive seasons (1956-7 through 1968-9).

Season:	1	2	3	4	5	6	7	8	9	10	11	12	13
Cyclones	6	5	4	6	6	3	12	7	4	2	6	7	4

- Let Y_i denote the number in season i, $i = 1, \quad , 13$. Suppose $Y_i \sim \text{Poi}(\theta)$. Then the log-likelihood function is

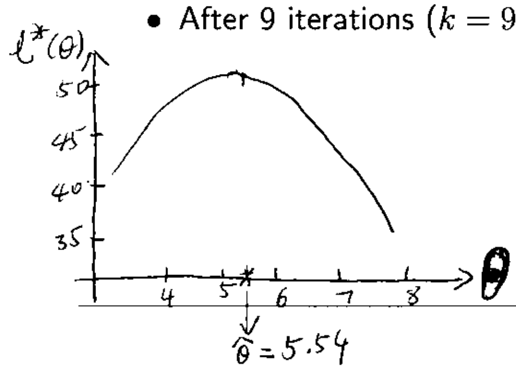
$$l(\theta) = \sum_{i=1}^{13} l_i = \sum_{i=1}^{13}(y_i \log\theta - \theta - \log y_i!) \text{ or } l^*(\theta) = \sum_{i=1}^{13}(y_i \log\theta - \theta),$$

The MLE of θ is $\hat{\theta} = \bar{y} = 72/13 = 5.538$.

- An alternative approach is to use numerical methods such as Newton-Raphson or bisection methods. See SAS code `Table1_3 sas`.

Bisection Algorithm

- Step 1. Take $\theta^{(1)} = 5$ and $\theta^{(2)} = 6$ as initial values.
- Step 2: Take approximations $\theta^{(k)}$ for $k = 3, 4,$ are the average values of the two previous estimates of θ with the largest value of $l^*(\theta)$. e.g., $\theta^{(6)} = \frac{1}{2}(\theta^{(5)} + \theta^{(3)})$.
- Step 3: Repeat Step 2 until the algorithm converges. For example, if $|\theta^{(k)} - \theta^{(k-1)} < 0.01$ (correct to 2 decimal places), stop.
- After 9 iterations ($k = 9$), $\hat{\theta} \approx 5.54$, and $l^*(\theta^{(k)}) = 51.24$



$l^*(\theta)$ differs from $l(\theta)$ by a Constant.
For Figure on the left, see Fig.1.2, page 16 in textbook and Figure1-2.sas
Study the SAS code
1 proc IML
2. SAS Macro Variable, e.g., see SAS-Macro-Var.pdf

k	$\theta^{(k)}$	$l^*(\theta)$
1	5	50.878
2	6	51.007
3	5.5	51.242 *
4	5.75	51.192
5	5.625	51.235 *
6	5.5625	51.243 *
7	5.5313	51.24354
8	5.5469	51.24352
9	5.5391	51.24361

Program: See Table1-3.R or Table1-3.sas.

Exampl to calculate $\theta^{(6)}$

$\leftarrow 5.75 \left[= \frac{6 + 5.5}{2} = 5.75 = \theta^{(4)}\right]$

$\leftarrow 5.5625 \left[= \theta^{(6)} = \frac{1}{2}(\theta^{(5)} + \theta^{(3)}) = \frac{5.5 + 5.625}{2} = 5.5625\right]$

In SAS interface,
To see the results, → Left Panel
→ Results
→ HTML or Text Format
To see the generated data sets:
→ Left panel
→ Explorer
→ Work (default folder)
→ e.g., Fig1-1

MLE (continued)

- **Score function**:

$$U(\theta) = \frac{\partial}{\partial\theta} l(\theta).$$

- **Score equation**:

$$U(\theta) = 0.$$

  - Very often MLE is the root of score equations.
  - Suppose $U(\hat{\theta}) = 0$. Then the variance of MLE $\hat{\theta}$ can be estimated by the inverse of

$$-\frac{\partial^2}{\partial\theta\partial\theta^T} l(\theta)\,|_{\theta=\hat{\theta}}$$

  - Note: MLE may be found at the boundary of Θ, and we may not have the nice results listed above. But in this course, all the (log) likelihood functions are *concave*, i.e., the 2nd derivative of $l(\theta)$ (Hessian matrix H) is negative definite (or $-H$ positive definite).

MLE (continued)

- **Information Matrix** for θ, if Y_i's are iid,

$$\begin{aligned} I_n(\theta) &= E[U(\theta)U(\theta)^T] = \sum_{i=1}^n E[U_i(\theta)U_i(\theta)^T] \\ &= nE[U_1(\theta)U_1(\theta)^T] = nI(\theta), \\ &\text{or} \\ &= -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} l(\theta)\right] = -\sum_{i=1}^n E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} l_i(\theta)\right] \\ &= -nE\left[\frac{\partial^2}{\partial\theta\partial\theta^T} l_1(\theta)\right], \end{aligned}$$

$\Rightarrow I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} l(\theta)\right]$

where $I(\theta)$ is called the information matrix for a single observation.

It is seen that

$E[U_1(\theta)\, U_1(\theta)^T] = E\left[\frac{\partial l_1(\theta)}{\partial\theta}\left(\frac{\partial l_1(\theta)}{\partial\theta}\right)^T\right] = -E\left\{\frac{\partial^2 l_1(\theta)}{\partial\theta\,\partial\theta^T}\right\}$

MLE (continued)

- **Observed Information Matrix**:

$$\hat{I}_n(\theta) = \sum_{i=1}^{n} U_i(\theta) U_i(\theta)^T$$

or

$$= -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} l_i(\theta),$$

$\rightarrow$ This is the observed information matrix for all observations

and $\hat{I}(\theta) = \frac{1}{n}\hat{I}_n(\theta)$. $\rightarrow$ This is an average observed information for all observations

MLE (continued)

- Some properties:
  - $E[U(\theta)] = 0$.
  - Suppose Y_i are iid and $I(\theta)$ exists. Then

    Average of all observed information $\leftarrow$ $\hat{I}(\theta) \rightarrow_p I(\theta)$. $\rightarrow$ Expectation of a single observation
  - Under regularity conditions, the MLE $\hat{\theta}$ has the following properties:
    - $\hat{\theta} \rightarrow_p \theta$.
    - $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d MVN(0, I^{-1}(\theta))$.

By the law of large number,

$\hat{I}(\theta) = \frac{1}{n}\hat{I}_n(\theta)$

$= \frac{1}{n}\sum_{i=1}^{n} U_i(\theta) U_i^T(\theta)$

$\rightarrow E[U_i(\theta) U_i^T(\theta)]$, when $n \rightarrow \infty$

MLE (continued)

- e.g., Linear regression: Suppose that $Y_i \sim N(X_i^T\beta, \sigma^2)$ and $Y_1, \quad , Y_n$ independent. Denote $\theta = (\beta^T, \sigma^2)^T$

  i. Likelihood $f(y_i;\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right\}$

  $$L(\theta) = \prod_{i=1}^n f(y_i;\theta) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^T\beta)^2\right\},$$

  ii. Log-likelihood

  $$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(y_i;\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^T\beta)^2$$

  iii. MLEs of β and σ^2

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \begin{cases} \frac{\partial \ell(\theta)}{\partial \beta} = -\frac{1}{2\sigma^2}\sum_{i=1}^n 2(-x_i)(y_i - x_i^T\beta) = 0 \\ \frac{\partial \ell(\theta)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (y_i - x_i^T\beta)^2 = 0 \end{cases} \Rightarrow \begin{cases} n\sigma^2 = \sum_{i=1}^n (y_i - x_i^T\beta)^2 \\ \sum_{i=1}^n x_i(y_i - x_i^T\beta) = 0 \end{cases}$$

or $\left(\sum_{i=1}^n x_i x_i^T\right)\beta = \sum_{i=1}^n x_i y_i$

Then,

$$\begin{cases} \hat\beta = \left[\sum_{i=1}^n x_i x_i^T\right]^{-1} \sum_{i=1}^n x_i y_i \\ \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta)^2 . \end{cases}$$

Let $X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$, then $\hat\beta = (X^TX)^{-1} X Y$

For inference, $\hat\beta \sim N(\beta, \hat\Sigma)$, where $\hat\Sigma = \{I_n(\hat\theta)\}^{-1}_{(1)}$

$\{I_n(\theta)\}^{-1} = \begin{pmatrix} (X^TX)^{-1}\sigma^2 & 0 \\ 0 & \frac{2(\sigma^2)^2}{n} \end{pmatrix}$,

$\{I_n(\hat\theta)\}^{-1}_{(1)} = (X^TX)^{-1}\hat\sigma^2$,

Since θ or σ^2 is unknown, use $\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\hat\beta)^2$ to estimate it.

iv. Score function

$$U_i(\theta) = \begin{pmatrix} \frac{\partial \ell_i(\theta)}{\partial \beta} \\ \frac{\partial \ell_i(\theta)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} x_i (y_i - x_i^T\beta) \\ -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - x_i^T\beta)^2 \end{pmatrix}.$$

v. Observed Information matrix

$$\frac{\partial U_i(\theta)}{\partial \theta^T} = \frac{\partial^2 \ell_i(\theta)}{\partial\theta\,\partial\theta^T} = \begin{pmatrix} -\frac{1}{\sigma^2} x_i x_i^T, & -\frac{1}{(\sigma^2)^2} x_i (y_i - x_i^T\beta) \\ \frac{1}{(\sigma^2)^2} x_i^T (y_i - x_i^T\beta), & \frac{1}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}(y_i - x_i^T\beta)^2 \end{pmatrix}$$

iv. Information matrix

For all the observations, the observed information matrix is

$$\hat I_n(\theta) = \sum_{i=1}^n \frac{\partial^2}{\partial\theta\,\partial\theta^T} \ell_i(\theta)$$

$$= \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i^T, & \frac{1}{(\sigma^2)^2}\sum_{i=1}^n x_i (y_i - x_i^T\beta) \\ \frac{1}{(\sigma^2)^2}\sum_{i=1}^n x_i^T (y_i - x_i^T\beta), & \frac{-n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\sum_{i=1}^n (y_i - x_i^T\beta)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\sigma^2} X^T X, & \frac{1}{(\sigma^2)^2} X^T \varepsilon \\ \frac{1}{(\sigma^2)^2} \varepsilon^T X, & \frac{-n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\varepsilon^T\varepsilon \end{pmatrix},$$

so

$$I_n(\theta) = E[\hat I_n(\theta)] = \begin{pmatrix} \frac{1}{\sigma^2} X^T X & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{pmatrix},$$

Since consider X a constant and use $E(\varepsilon) = 0$, $E(\varepsilon^T\varepsilon) = n\sigma^2$.

Newton-Raphson Algorithm

- Step 1. Take θ^0 as initial value.
- Step 2: Update θ^k by

$$\left(\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} \Big|_{\theta = \theta^k} \right)^{-1}$$

$$\theta^{k+1} = \theta^k + [\hat{I}_n(\theta^k)]^{-1} U(\theta^k).$$

- Step 3: Repeat Step 2 until the algorithm converges.
- **HW**: redo Example II, data for tropical cyclones.