MLE IV: Multivariate Case and more on the Fisher Approximation Theorem

STAT 24400
Lecture 12
February 16, 2016

As you now know by heart,

θ is the parameter, a state of nature.

random vars $X_1, X_2, \ldots X_n$ data
$L(\theta) = f(x|\theta)$ or $f(x_1, \ldots x_n|\theta)$
the likelihood function
(model as a function of θ)

We want $\hat{\theta} = \hat{\theta}(x_1, \ldots x_n)$, a random variable, to estimate θ. $\hat{\theta}$ maximizes $L(\theta)$.

In real life, of course, we frequently have multidimensional θ and X.

Ex. Normal Case

n measurements $X_1, \ldots, X_n$

Model: X_i's indep, each $N(\mu, \sigma^2)$

Data: $X = (X_1, \ldots, X_n)$

Parameter: $\theta = \vec{\theta} = (\mu, \sigma^2)$

$$L(\theta) = f(x|\theta) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right)$$

$$= (2\pi\phi)^{-\frac{n}{2}} e^{-\frac{1}{2\phi}\sum(x_i-\mu)^2} \quad (\phi = \sigma^2)$$

$$\log(L(\theta)) = -\frac{n}{2}\log(2\pi\phi) - \frac{1}{2\phi}\sum(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu}\log(L(\theta)) = \frac{-1}{\phi}\sum(x_i-\mu) = \frac{n}{\phi}(\bar{x}-\mu)$$

$$\frac{\partial}{\partial\phi}\log(L(\theta)) = \frac{-n}{2\phi} + \frac{1}{2\phi^2}\sum(x_i-\mu)^2$$

Set $= 0$:

$$\hat{\mu} = \bar{X}, \quad \hat{\phi} = \frac{1}{n}\sum(x_i-\hat{\mu})^2$$

$$= \frac{1}{n}\sum(x_i-\bar{x})^2$$

To show this is a max, need $2^{\underline{nd}}$ derivs:

Let $$l_{11}(\theta) = \frac{\partial^2}{\partial \mu^2} \log L(\theta)$$

$$l_{22}(\theta) = \frac{\partial^2}{\partial \phi^2} \log L(\theta)$$

$$l_{12}(\theta) = \frac{\partial^2}{\partial \mu \partial \theta} \log L(\theta)$$

Enough to show

$$\Delta = (l_{12}(\hat{\theta}))^2 - l_{11}(\hat{\theta}) l_{22}(\hat{\theta}) < 0$$

and $l_{11}(\hat{\theta}) < 0$

Now, $l_{12}(\theta) = \frac{-n}{\phi^2}(\bar{x} - \mu)$ so $l_{12}(\hat{\theta}) = 0$

$$l_{11}(\theta) = \frac{-n}{\phi}, \text{ so } l_{11}(\hat{\theta}) < 0$$

$$l_{22}(\theta) = \frac{n}{2\phi^2} - \frac{1}{\phi^3}\sum(x_i - \mu)^2$$

$$\Rightarrow l_{22}(\hat{\theta}) = \frac{-n}{2\phi^2}$$

so $$\Delta = 0 - \left(\frac{-n}{\phi}\right)\left(\frac{-n}{2\phi^2}\right) < 0.$$

OK. MLE's are $\hat{\mu} = \bar{X}$ and

$$\hat{\sigma}^2 = \hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Bias: $E(\hat{\mu}) = E(\bar{X}) = \mu$ (unbiased)

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \quad \text{(after some algebra)}$$

$$\boxed{\text{Use } E(W^2) = Var(W) + [E(W)]^2}$$

so:

$$E(\bar{X}^2) = \frac{E(\Sigma X_i)^2}{n^2} = \frac{Var(\Sigma X_i) + [E(\Sigma X_i)]^2}{n^2}$$

$$= \frac{n\sigma^2 + [n\mu]^2}{n^2} = \frac{\sigma^2}{n} + \mu^2$$

$$E\left(\frac{1}{n}\Sigma X_i^2\right) = \frac{1}{n}\Sigma(X_i^2) = \frac{1}{n}\Sigma(\sigma^2 + \mu^2)$$

$$= \sigma^2 + \mu^2$$

so

$$E(\hat{\sigma}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \left(\frac{n-1}{n}\right)\sigma^2$$

(biased!)

Common to use $s^2 = \frac{n}{n-1}\hat{\sigma}^2$

$$= \frac{1}{n-1}\Sigma(X_i - \bar{X})^2$$

$$E(s^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2 \quad \text{(unbiased)}$$

Note: As long as X_i's are independent with $E(X_i) = \mu$
$Var(X_i) = \sigma^2$,

$\bar{X}$ unbiased for μ

s^2 unbiased for σ^2

('normal' not used for this)

Note 2: MLE of σ is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\tfrac{1}{n}\sum(x_i - \bar{x})^2},$$

but both $\hat{\sigma}$ and s are biased for σ a little bit, because

$$E(\sqrt{s^2}) \neq \sqrt{E(s^2)} = \sqrt{\sigma^2} = \sigma$$

In fact $E(s) = b(n)\sigma$

n	4	10	100
$b(n)$	.778	.923	.992

MSE:

$$MSE(s^2) = Var(s^2) = \frac{2\sigma^4}{(n-1)}$$ more next quarter!

$$Bias(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

$$Var(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)^2 Var(s^2) = \frac{2(n-1)}{n^2}\sigma^4$$

$$MSE(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^4$$

Since $\frac{2n-1}{n^2} < \frac{2}{n-1}$ for all $n \geq 2$

$$MSE(\hat{\sigma}^2) < MSE(s^2)$$

$$\left(\text{But } \frac{MSE(\hat{\sigma}^2)}{MSE(s^2)} = 1 - \left(\frac{3n+1}{2n^2}\right) \simeq 1\right)$$

How distributed? Next Quarter we will show that it is exactly true that:

if X_i's are normal themselves

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)s^2}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \text{ dist. Chi-Sq., } n-1 \text{ d.f.}$$

Fisher's Theorem
Multidimensional Parameter

$$L(\vec{\theta}) = f(\vec{x} \mid \vec{\theta})$$

If $\hat{\vec{\theta}}$ is found by setting derivs equal to 0, then if n large, $\hat{\vec{\theta}}$ has approx a multiple dim. normal dist. $N(\vec{\theta}, \vec{\gamma^2})$ where $\vec{\gamma^2}$ is the inverse of the matrix
$$\left[-E\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\vec{\theta}) \right) \right]$$

Idea for a Proof of Fisher's Theorem

Up to now, we found MLE's by solving

$$\frac{d}{d\theta} \log L(\theta) = 0$$

exactly. This isn't always possible. Sometimes (including numerical problems) it is useful to use an approximate method, such as Newton - Raphson.

Let

$$g(\theta) = \frac{d}{d\theta} \log L(\theta)$$

$$g'(\theta) = \frac{d^2}{d\theta^2} \log L(\theta)$$

We want to find $\hat{\theta}$ s.t. $g(\hat{\theta}) = 0$. Suppose $\hat{\theta}$ is near θ. Then the mean value theorem says that

$$g(\hat{\theta}) - g(\theta) \simeq (\hat{\theta} - \theta) g'(\theta)$$

But we supposed $g(\hat{\theta}) = 0$, so

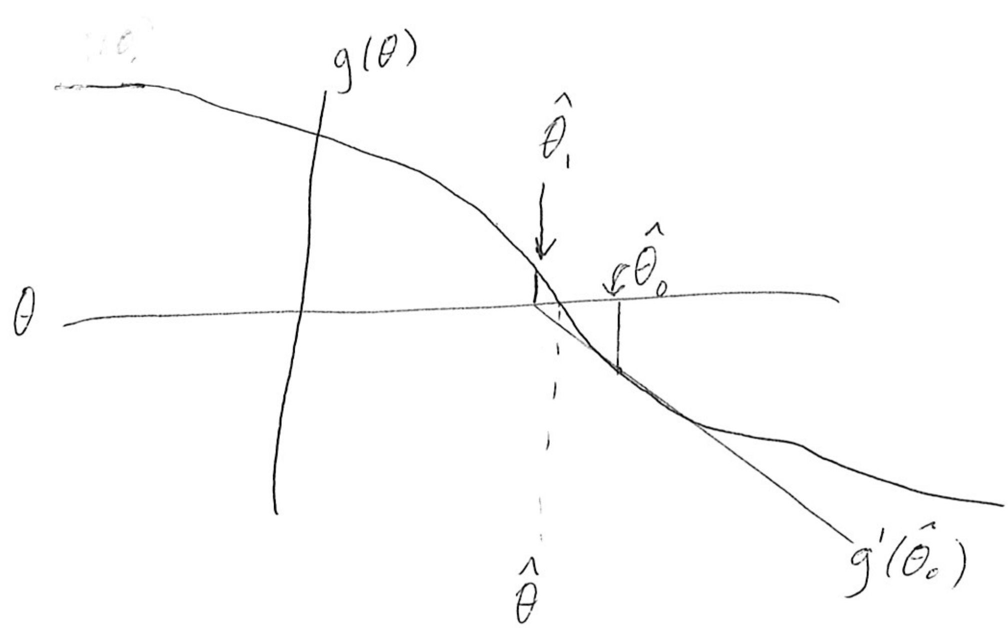
$$-g(\theta) \simeq (\hat{\theta} - \theta) g'(\theta)$$

$$\hat{\theta} - \theta = -\frac{g(\theta)}{g'(\theta)}$$

$$\hat{\theta} = \theta - \frac{g(\theta)}{g'(\theta)}$$

an approximation for $\hat{\theta}$. For numerical work, we can let this approximation be our next guess for $\hat{\theta}$, so if the first guessed θ is $\hat{\theta}_0$, we have just found $\hat{\theta}_1$. Then we can continue by taking

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{g(\hat{\theta}_n)}{g'(\hat{\theta}_n)}$$



This is also the basis of a proof of Fisher's Theorem.

Let's make the X_i iid.

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$\log L(\theta) = \sum \log f(x_i|\theta)$$

Let

$$g(\theta) = \frac{d}{d\theta} \log L(\theta),$$ and

define

$$Z_i(\theta) = \frac{d}{d\theta} \log f(x_i|\theta)$$

$$= \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)}$$ (we'll need this below)

so

$$g(\theta) = \sum_{i=1}^{n} Z_i(\theta)$$

is a sum of indep. random vars.

Let's calculate $E(Z_i(\theta))$.

$$E(z_i(\theta)) = \int_{-\infty}^{\infty} z_i(\theta) f(x/\theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{\frac{d}{d\theta} f(x/\theta)}{f(x/\theta)} f(x/\theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{d}{d\theta} f(x/\theta)\, dx$$

$$= \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x/\theta)\, dx$$

$$= \frac{d}{d\theta} \circ (\text{const} = 1)$$

$\int_{-\infty}^{\infty} f(x/\theta)\, d\theta$ is a density

$$= 0.$$

$$Var(z_i(\theta)) = E[(z_i(\theta))^2] - E[z_i(\theta)]^2$$

$$= E\left[\left(\frac{d}{d\theta} \log f(x_i/\theta)\right)^2\right]$$

but we know that

$$E\left[\left(\frac{d}{d\theta} \log f(x_i/\theta)\right)^2\right] = -E\left[\frac{d^2}{d\theta^2} \log f(x_i/\theta)\right] = \frac{1}{\tau^2(\theta)}$$

But remember:

$$g(\theta) = \sum_{i=1}^{n} Z_i(\theta).$$

$E(Z) = 0$ and $Var(Z) = \frac{1}{\gamma^2}$, so the CLT says that

$$\frac{g(\theta)}{\sqrt{n}} \text{ is distributed approx } N\left(0, \frac{1}{\gamma^2(\theta)}\right)$$

Also,

$$g'(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} Z_i(\theta) = \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i|\theta)$$

is a sum of indep random vars, so the Law of Large Numbers says that as $n \to \infty$,

$$\frac{1}{n} g'(\theta) \xrightarrow{P} E\left(\frac{d}{d\theta} Z_i(\theta)\right) = E\left(\frac{d^2}{d\theta^2} \log f(x_i|\theta)\right)$$

$$= -\frac{1}{\gamma^2(\theta)}$$

So now,

$$\sqrt{n}\left(\frac{-g(\theta)}{g'(\theta)}\right) = \frac{g(\theta)/\sqrt{n}}{(g'(\theta)/n)}$$

$$\cong \frac{g(\theta)/\sqrt{n}}{1/\tau^2(\theta)}$$

$$= \tau^2(\theta) \cdot \frac{g(\theta)}{\sqrt{n}}$$

will have an approximate distribution of $\mathcal{N}\left(0, (\tau^2(\theta))^2 \cdot \frac{1}{\tau(\theta)}\right)$, or

$$\mathcal{N}(0, \tau^2(\theta)).$$

Hence $\frac{-g(\theta)}{g'(\theta)}$ is approximately

distributed $\mathcal{N}(0, \tau^2(\theta))$.

But we have approximated $\log L(\theta)$ by $\frac{-g(\theta)}{g'(\theta)}$ near $\hat{\theta}$, which is what we sought to ~~prove~~ argue for.