MLE's Bias Pathology, Model Updated MLE and MME, and Wallace's Minimum Message Length (MML) method

• We start with an examination of parametric statistical problems. A brief description of MLE and MME, examples showing that they are not always unbiased and that the MLE can be inconsistent.

• Then we will show why MLE may not be unbiased and how to correct it by **updating the model providing the Likelihood equations**.

• More general results will follow.

PARAMETER ESTIMATION PROBLEMS

The set-up: Data $X_1, \ldots, X_n$ $i.i.d.$ r.v.'s,
$X_1 \sim f(x, \theta), \theta \in \Theta (\subseteq R^p$ usually),
f has known form, θ is unknown.

• Estimate: any function of the data only,
$T_n = T(X_1, \ldots, X_n)$.

• T_n is unbiased for θ when $ET_n = \theta \; \forall \; \theta \in \Theta$.

• The Aims:
Find an estimate, $T_n = T(X_1, \ldots, X_n)$, of θ, evaluate the estimation error and/or an estimate of T_n's variance. Determine T_n's distribution.

• Think of estimation as "separation" of the true θ from $R^p - \{\theta\}$ with some error around θ.

• We solve a STOCHASTIC OPTIMIZATION PROBLEM. It differs from minimizing OR maximizing an objective function in Calculus. The Objective Function is determined by the Data.

• Often estimates are obtained using data model f and a Method $\mathcal{M}$ in steps, e.g. solving equations:

$$DATA \overset{EVOLVES\ VIA\ f\ AND\ \mathcal{M}}{\Longrightarrow} EQUATIONS$$

• $DATA\ X_1, \ldots, X_n$ is not touched again after finding the $EQUATIONS$.

• The $EQUATIONS$ include estimates and

parameters.

• Estimates are the different coefficients that will be known because of the observed sample: $X_1 = x_1, \ldots, X_n = x_n$.

Questions for you: I guess you had a first course in Math Stat. Think of MLE and MME.

a) Do the $EQUATIONS$ evolve?
b) Does $DATA$ evolve?
c) What are the implications in the $EQUATIONS$?
d) Is there new information in the process?
e) Should we use this new information before solving the $EQUATIONS$?
f) What $EQUATIONS$?

Let us try to see these questions in things you have seen before.

MOMENTS ESTIMATION METHOD

$DATA$: $X_1, \ldots, X_n$ i.i.d. r.vs. X_1 follows $f(x, \theta); \theta \in R^p, p \geq 1$.

• The $\underline{k\text{-th moment}}$ of X_1,

$$\mu_k = E(X_1^k), \text{ i.e.}$$

$$\mu_k = \int x^k f(x,\theta)dx = g_k(\theta),$$

$\underline{\text{is function of } p \text{ unknown parameter(s)}}$,
$\theta = (\theta_1, \ldots, \theta_p), k = 1, 2, \ldots$.

To determine the p unknown parameters need p-equations.

If there are known expressions for p moments of X_1 use them. Which p moments? The lower p-moments, for $k = 1, \ldots, p$. Here is "How" assuming those are the first p-moments: replace μ_k by its sample estimate

$$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_j^k,$$

and solve the system of $EQUATIONS$

$$\frac{1}{n}\sum_{i=1}^{n} X_j^k = g_k(\theta_1, \ldots, \theta_p), k = 1, \ldots, p, \quad (1)$$

obtaining the Moments Estimates (ME) $\tilde{\theta}_k, k = 1, \ldots, p$.

Example 1. $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2 both unknown, $\theta = (\mu, \sigma^2)$. Find the moments estimates of μ, σ^2.

Step 1: $p = 2$ so we need two equations.
Step 2: $\mu_1 = EX_1 = \mu,\ \mu_2 = EX_1^2 = \sigma^2 + \mu^2$.
Step 3: Replace the population moments on the left sides of the equalities by their sample counterparts to obtain $EQUATIONS$ in μ, σ^2.

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j = \mu, \tag{2}$$

$$\frac{1}{n}\sum_{j=1}^{n} X_j^2 = \mu^2 + \sigma^2 \tag{3}$$

NOTE: $X_1, \ldots, X_n$ is history. What we have now is (2) and (3).
Questions: What is random in (2)?
What is random in (3)?

Do they satisfy the Method of Moments approach, i.e. E of left side equals right side?
Which one we solve first?
I guess you will say the fist equation, i.e. (2), obtaining $\tilde{\mu}_{ME} = \bar{X}$ which is unbiased for μ.
Replacing in (3) we get:

$$\frac{1}{n}\sum_{j=1}^{n} X_j^2 = \bar{X}_n^2 + \sigma^2 \text{ or } \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 = \sigma^2. \tag{4}$$

Then, (4) is solved for σ^2, obtaining

$$\tilde{\sigma}^2_{ME} = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \bar{X}_2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2,$$

which **is NOT unbiased** for σ^2.

With μ known the Moments Estimate of σ^2 is the unbiased

$$\tilde{\sigma}^2_{ME} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2.$$

What is random in equation (4), i.e. in

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 = \sigma^2.$$

The left side of the equation.

Did the equation evolve? YES.

Does the equation satisfy the Method of Moments approach, i.e. E of left side equals right side? NO.

What if we will respect it?

The only data we have is the left side. We need to estimate one parameter, so take its first moment:

$$E\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 = \frac{n-1}{n}\sigma^2.$$

Use the Method of Moments Approach equat-

ing the left side with the evolved data,

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 = \frac{n-1}{n}\sigma^2.$$

Solving it we get the Model Updated Moments Estimate which is unbiased for σ^2 :

$$\tilde{\sigma}^2_{MUME} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2.$$

We will see that Model Updated MLE makes often MLE unbiased. It will be proved why this happens.

THE MAXIMUM LIKELIHOOD METHOD

• It is the most common estimation method.

Assume that the data $\mathbf{X} = (X_1, \ldots, X_n)$ have joint density or probability

$$f(x_1, \ldots, x_n|\theta),\ \theta \in \Theta \subset R^p.$$

Given the observed data values $X_1 = x_1, \ldots,$ $X_n = x_n$, we want to see how much each of the models determined by Θ "supports" the data.

We do that by looking at $f(x_1, \ldots, x_n|\theta)$ as function of $\theta(\in \Theta)$.

We may also use $f(X_1, \ldots, X_n|\theta)$ instead, that makes it a random variable.

The likelihood of θ (or of each model) as a function of $X_1, \ldots, X_n$, is

$$\text{lik}(\theta) = f(X_1, \ldots, X_n|\theta). \tag{5}$$

The Maximum Likelihood Estimate (MLE) $\hat{\theta}$ of θ is that value in Θ that maximizes (5).
(Assume uniqueness ...)

When the data $X_1, \ldots, X_n$ are i.i.d. observations with density f then the likelihood in (5) is

$$\text{lik}(\theta) = \Pi_{j=1}^{n} f(X_j|\theta). \tag{6}$$

Instead of maximising the product $\text{lik}(\theta)$ in θ it is easier to maximise its logarithm, called

log likelihood, that is the sum

$$l(\theta) = \sum_{j=1}^{n} \log f(X_j|\theta). \qquad (7)$$

• When the domain of $f(x|\theta)$ depends on θ you have to maximise directly either lik(θ) or $l(\theta)$.

• When the domain of the density $f(x|\theta)$ is independent of θ and f is "smooth" (i.e. has derivatives with respect to θ), to maximise (7) equate its derivative to zero, and solve it to obtain $\hat{\theta}$. Show the sign of the second derivative at $\hat{\theta}$ is negative, or that $l(\theta)$ is concave function of θ.

The steps to obtain the MLE $\hat{\theta}$:

1) If $X_1, \ldots, X_n$ are i.i.d. $f(x|\theta)$,

$$\text{lik}(\theta) = \Pi_{j=1}^{n} f(X_j|\theta).$$

2) $l(\theta) = \sum_{j=1}^{n} \log f(X_j|\theta)$

3) Solve the equation

$$0 = \frac{dl(\theta)}{d\theta}|_{\theta=\hat{\theta}}$$

to obtain $\hat{\theta}$. (May need iterative methods.)

4) Check $\hat{\theta}$ is a maximum.

Potential problem: Many solutions in **3)**.

• ML ESTIMATION WHEN $\theta \in R^p$

When $\theta = (\theta_1, \ldots, \theta_p)$ replace **3)** by a system of p-equations with p unknowns

3*) $0 = \frac{\partial l(\theta_1,\ldots,\theta_p)}{\partial \theta_i}|_{\theta_1=\hat{\theta}_1,\ldots,\theta_p=\hat{\theta}_p}, \ i = 1, \ldots, p.$

Example 2. $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, μ, σ both unknown. Find the MLE $\hat{\mu}, \ \hat{\sigma}^2$.

1. $\Pi_{j=1}^n f(X_j|\mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\sum_{j=1}^n (X_j-\mu)^2}{2\sigma^2}}$

2. $l(\mu, \sigma) = -n \log \sigma - n \log \sqrt{2\pi} - \frac{\sum_{j=1}^n (X_j-\mu)^2}{2\sigma^2}$

3^*. EQUATIONS

$$0 = \frac{\partial l(\mu,\sigma)}{\partial \mu} = \frac{\sum_{j=1}^n (X_j-\mu)}{\sigma^2}$$

$0 = \frac{\partial l(\mu,\sigma)}{\partial \sigma} = \frac{-n\sigma^2 + \sum_{j=1}^{n}(X_j - \mu)^2}{\sigma^3}$

Both equations have been obtained from the Data and the ML method.

Solving the first equation we get: $\hat{\mu}_{MLE} = \bar{X}_n$.

Then the second equation becomes,

$-n\sigma^2 + \sum_{j=1}^{n}(X_j - \bar{X})^2 = 0.$

Did the second equation evolve? YES.
The Data in it evolved? YES.
Could this equation be obtained from the Likelihood of the original Data? NO.

The MLE is biased, $\hat{\sigma}^2_{MLE} = \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{n}$.

What is the data in this last equation? $Y^* = \sum_{j=1}^{n}(X_j - \bar{X})^2$.

What is the density of Y^*/σ^2? χ^2_{n-1}. The likelihood of Y^* provides equation for σ

$$-(n-1)\sigma^2 + Y^* = 0$$

and the Model Updated MLE is unbiased,

$$\hat{\sigma}^2_{MUMLE} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Example 3. Le Cam's (1990) version of Neyman-Scott (1948) example: independent X_i, Y_i are Normal with mean μ_i and variance σ^2, i=1,. . . , n. Assuming independence of all n samples of size 2, show that MLE $\hat{\sigma}^2$ is inconsistent but a naive estimate is consistent.

The likelihood

$$\text{lik}(\mu_1, \mu_2, \ldots, \mu_n, \sigma)$$

$$= (\frac{1}{\sigma^2 2\pi})^n e^{-\frac{\sum_{j=1}^{n}[(X_j-\mu_j)^2+(Y_j-\mu_j)^2]}{2\sigma^2}},$$

the log-likelihood

$$l(\mu_1, \mu_2, \ldots, \mu_n, \sigma) = -2n\log\sigma - n\log 2\pi$$

$$-\frac{\sum_{j=1}^{n}[(X_j - \mu_j)^2 + (Y_j - \mu_j)^2}{2\sigma^2}.$$

The likelihood equations for the means are

$$0 = \frac{\partial l(\mu_1, \mu_2, \ldots, \mu_n, \sigma)}{\partial \mu_j} = \frac{X_j - \hat{\mu}_j + Y_j - \hat{\mu}_j}{\sigma^2}$$

$$\text{and } \hat{\mu}_j = \frac{X_j + Y_j}{2},\ j = 1, \ldots, n.$$

The likelihood equation for σ after replacing the μ's by their MLE's are

$$0 = \frac{\partial l(\mu_1, \mu_2, \ldots, \mu_n, \sigma)}{\partial \sigma}$$

$$= -\frac{2n}{\hat{\sigma}} + \frac{\sum_{j=1}^{n}[(X_j - \hat{\mu}_j)^2 + (Y_j - \hat{\mu}_j)^2]}{\hat{\sigma}^3}$$

and since

$$(X_j - \hat{\mu}_j)^2 = (Y_j - \hat{\mu}_j)^2 = \frac{(X_j - Y_j)^2}{4},$$

$$\text{we get } 0 = -\frac{2n}{\hat{\sigma}} + \frac{\sum_{j=1}^{n}(X_j - Y_j)^2}{2\hat{\sigma}^3}$$

$$\text{and } \hat{\sigma}^2 = \frac{1}{4n}\sum_{j=1}^{n}(X_j - Y_j)^2.$$

$$E\hat{\sigma}^2 = \frac{1}{4n}\sum_{j=1}^{n} E(X_j - Y_j)^2 = \frac{1}{4n}2n\sigma^2 = \frac{\sigma^2}{2}.$$

Naive Estimate

$W_j = X_j - Y_j \sim N(0, 2\sigma^2),\ j = 1, \ldots, n$, the W's are independent, and $EW_j^2 = 2\sigma^2$, by the WLLN

$$\frac{1}{n}\sum_{j=1}^{n} W_j^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - Y_j)^2 \xrightarrow{Prob} 2\sigma^2$$

$$\text{or } \frac{1}{2n}\sum_{j=1}^{n}(X_j - Y_j)^2 \xrightarrow{Prob} \sigma^2$$

$$\text{and } \hat{\sigma}^2 = \frac{1}{4n}\sum_{j=1}^{n}(X_j - Y_j)^2 \xrightarrow{Prob} \frac{\sigma^2}{2}.$$

Try with the Method of Moments to see what you get. You have to use all the data.

Lemma 1: Let $\mathbf{X} \sim f(\mathbf{x}, \theta), \theta \in R, U_\theta(\mathbf{X}, \theta) = \frac{d \ln f(\mathbf{X}, \theta)}{d\theta}$.

Assume: a) $EU_\theta(\mathbf{X}, \theta) = 0 \; \forall \; \theta$.
b) $U_{\theta\theta} = \frac{dU_\theta(\mathbf{x},\theta)}{d\theta} = C \; \forall \; \theta, C$ constant.
Then, $\hat{\theta}_{MLE}$ is unbiased for θ.

Proof:

$$U(\mathbf{x}, \hat{\theta}_{MLE}) = U(\mathbf{x}, \theta) + (\hat{\theta}_{MLE} - \theta) \cdot C,$$

$$\rightarrow E(\hat{\theta}_{MLE} - \theta) = -C^{-1} EU(\mathbf{X}, \theta) = 0.$$

Question: Do a), b) in Lemma 1 hold often?

Normal model: $\mathbf{X} = (X_1, \ldots, X_n)$ $i.i.d.$ mean θ, variance $\psi = \sigma^2$, C^*, generic constant

$$f(\mathbf{x}, \theta, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^n (x_i - \theta)^2 / 2\sigma^2},$$

$$\ln f(\mathbf{x}, \theta, \sigma) = -n \ln \sigma - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} + C^*,$$

$$U_\theta(\mathbf{x}, \theta, \sigma) = \frac{\sum_{i=1}^n (x_i - \theta)}{\sigma^2},$$

$$EU_\theta(\mathbf{X}, \theta, \sigma) = 0, \qquad U_{\theta\theta} = -n/\sigma^2.$$

$$U_\sigma(\mathbf{x}, \theta, \sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \theta)^2}{\sigma^3}$$

$$EU_\sigma(\mathbf{X}, \theta, \sigma) = 0,$$

$$U_{\sigma\sigma} = \frac{n}{\sigma^2} - \frac{3\sum_{i=1}^n (x_i - \theta)^2}{\sigma^4}$$

When θ is known,

$$U_\sigma(\mathbf{X}, \theta, \hat{\sigma}) = U_\sigma(\mathbf{X}, \theta, \sigma) + (\hat{\sigma} - \sigma)U_{\sigma\sigma},$$

otherwise

$$U_\sigma(\mathbf{X}, \hat{\theta}, \hat{\sigma}) = U_\sigma(\mathbf{X}, \hat{\theta}, \sigma) + (\hat{\sigma} - \sigma)U_{\sigma\sigma}$$

Observe we cannot draw conclusion for $\hat{\sigma}^2$ from this expansion. **Will need derivatives with respect to** σ^2.

The estimate $\hat{\sigma}$ would be unbiased from the Lemma when the expected values of the score functions $U_\sigma(\mathbf{X}, \theta, \sigma)$ (for θ known), $U_\sigma(\mathbf{X}, \hat{\theta}, \sigma)$ in the right side of the expansion have means zero AND $U_{\sigma\sigma} = C$. The ultimate does not hold though. In addition, for σ, when θ is known, since

$EU_\sigma(\mathbf{X}, \theta, \sigma) = 0$, the equation

$$EU_\sigma(\mathbf{X}, \hat{\theta}, \sigma) = 0$$

is not expected to hold.

To have the expansion for σ^2 rewrite the log-likelihood as function of $\sigma^2 = \psi$.

$\ln f(\mathbf{x}, \theta, \sigma) = -\frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} + C^*$, C^*

$= -\frac{n}{2} \ln \psi - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\psi} + C^*$,

$U_\psi = -\frac{n}{2\psi} + \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\psi^2}$,

$U_{\psi\psi} = \frac{n}{2\psi^2} - \frac{\sum_{i=1}^n (x_i - \theta)^2}{\psi^3}$

Observe

$$EU_\psi = 0, \qquad EU_{\psi\psi} = -\frac{n}{\psi^2}.$$

Suggested Exercise: For the Neyman-Scott Example presented by Le Cam with X_i, Y_i, iid Normal with mean μ_i, variance $\sigma^2, i = 1, \ldots, n$ with all observations independent, obtain the Model Updated Moments Estimates of σ^2.

References

Le Cam, L. (1990) Maximum likelihood: An introduction. *Int. Statist. Rev.*, **58**, 2, 153–171.

Neyman, J. and Scott, E. L. (1948) Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1, pp. 1–32.