STAT 801: Mathematical Statistics
Large Sample Theory for MLEs

We study the approximate behaviour of $\hat\theta$ by studying the random function U. Notice that U is a sum of independent random variables.

Theorem 1 *If $Y_1, Y_2, \ldots$ are iid with mean μ then*

$$\frac{\sum Y_i}{n} \to \mu.$$

This therom is called the law of large numbers. The "Strong law" is that

$$P(\lim \frac{\sum Y_i}{n} = \mu) = 1$$

while the "weak law" is that

$$\lim P(|\frac{\sum Y_i}{n} - \mu| > \epsilon) = 0.$$

For iid Y_i the first conclusion, which is stronger, holds; for our heuristics we ignore differences between these notions.

Now suppose θ_0 is true value of θ. Then

$$U(\theta)/n \to \mu(\theta)$$

where

$$\begin{aligned}\mu(\theta) =& E_{\theta_0}\left[\frac{\partial \log f}{\partial\theta}(X_i,\theta)\right] \\ =& \int \frac{\partial \log f}{\partial\theta}(x,\theta) f(x,\theta_0) dx\end{aligned}$$

Example: $N(\mu,1)$ data:

$$U(\mu)/n = \sum (X_i - \mu)/n = \bar X - \mu$$

If the true mean is μ_0 then $\bar X \to \mu_0$ and

$$U(\mu)/n \to \mu_0 - \mu$$

Consider $\mu < \mu_0$: the derivative of $\ell(\mu)$ is likely to be positive so that ℓ increases as μ increases.

For $\mu > \mu_0$: the derivative is probably negative and so ℓ tends to be decreasing for $\mu > 0$.

Hence: ℓ is likely to be maximized close to μ_0.

Now we repeat these ideas for amore general case. Study rv

$$\log[f(X_i,\theta)/f(X_i,\theta_0)].$$

You know the inequality

$$E(X)^2 \le E(X^2)$$

(difference is $\text{Var}(X) \ge 0$.)

Generalization: Jensen's inequality: for g a convex function ($g'' \ge 0$ roughly) then

$$g(E(X)) \le E(g(X))$$

Inequality above has $g(x) = x^2$. Use $g(x) = -\log(x)$: convex because $g''(x) = x^{-2} > 0$. We get

$$-\log(E_{\theta_0}[f(X_i,\theta)/f(X_i,\theta_0)] \le E_{\theta_0}[-\log\{f(X_i,\theta)/f(X_i,\theta_0)\}] .$$

But

$$\begin{aligned} E_{\theta_0}\left[\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right] &= \int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0)dx \\ &= \int f(x,\theta)dx \\ &= 1 \end{aligned}$$

We can reassemble the inequality and this calculation to get

$$E_{\theta_0}[\log\{f(X_i,\theta)/f(X_i,\theta_0)\}] \le 0$$

Fact: inequality is strict unless the θ and θ_0 densities are actually the same.

Let $\mu(\theta) < 0$ be this expected value.

Then for each θ we find

$$\begin{aligned} &\frac{\ell(\theta) - \ell(\theta_0)}{n} \\ &\quad = \frac{\sum \log[f(X_i,\theta)/f(X_i,\theta_0)]}{n} \\ &\qquad \to \mu(\theta) \end{aligned}$$

This proves likelihood probably higher at θ_0 than at any other single θ.

Idea can often be stretched to prove that the mle is **consistent**.

Definition A sequence $\hat\theta_n$ of estimators of θ is consistent if $\hat\theta_n$ converges weakly (or strongly) to θ.

Proto theorem: In regular problems the mle $\hat\theta$ is consistent.

More precise statements of possible conclusions. Use notation

$$N(\epsilon) = \{\theta : |\theta - \theta_0| \le \epsilon\} .$$

Suppose:

$\hat\theta_n$ is global maximizer of ℓ.

$\hat\theta_{n,\delta}$ maximizes ℓ over $N(\delta) = \{|\theta - \theta_0| \le \delta\}$.

$$A_\epsilon = \{|\hat\theta_n - \theta_0| \le \epsilon\}$$

$$B_{\delta,\epsilon} = \{|\hat\theta_{n,\delta} - \theta_0| \le \epsilon\}$$

$$C_L = \{\exists!\theta \in N(L/n^{1/2}) : U(\theta) = 0, U'(\theta) < 0\}$$

Theorem:

1. Under conditions **I** $P(A_\epsilon) \to 1$ for each $\epsilon > 0$.
2. Under conditions **II** there is a $\delta > 0$ such that for all $\epsilon > 0$ we have $P(B_{\delta,\epsilon}) \to 1$.
3. Under conditions **III** for all $\delta > 0$ there is an L so large and an n_+0 so large that for all $n \ge n_0$, $P(C_L) > 1 - \delta$.
4. Under conditions **III** there is a sequence L_n tending to ∞ so slowly that $P(C_{L_n}) \to 1$.

Point: conditions get weaker as conclusions get weaker. Many possible conditions in literature. See book by Zacks for some precise conditions.

Asymptotic Normality

Study shape of log likelihood near the true value of θ.

Assume $\hat\theta$ is a root of the likelihood equations close to θ_0.

Taylor expansion (1 dimensional parameter θ):

$$\begin{aligned} U(\hat\theta) =& 0 \\ =& U(\theta_0) + U'(\theta_0)(\hat\theta - \theta_0) \\ &+ U''(\tilde\theta)(\hat\theta - \theta_0)^2/2 \end{aligned}$$

for some $\tilde\theta$ between θ_0 and $\hat\theta$.

WARNING: This form of the remainder in Taylor's theorem is not valid for multivariate θ. Derivatives of U are sums of n terms.

So each derivative should be proportional to n in size.

Second derivative is multiplied by the square of the small number $\hat\theta - \theta_0$ so should be negligible compared to the first derivative term.

Ignoring second derivative term get

$$-U'(\theta_0)(\hat\theta - \theta_0) \approx U(\theta_0)$$

Now look at terms U and U'.

Normal case:

$$U(\theta_0) = \sum (X_i - \mu_0)$$

has a normal distribution with mean 0 and variance n (SD $\sqrt{n}$).

Derivative is

$$U'(\mu) = -n\,.$$

Next derivative U'' is 0.

Notice: both U and U' are sums of iid random variables.

Let

$$U_i = \frac{\partial \log f}{\partial \theta}(X_i, \theta_0)$$

and

$$V_i = -\frac{\partial^2 \log f}{\partial \theta^2}(X_i, \theta)$$

In general, $U(\theta_0) = \sum U_i$ has mean 0 and approximately a normal distribution.

Here is how we check that:

$$\begin{aligned}
E_{\theta_0}(U(\theta_0)) &= nE_{\theta_0}(U_1) \\
&= n\int \frac{\partial \log(f(x,\theta_0))}{\partial \theta} f(x,\theta_0)dx \\
&= n\int \frac{\partial f(x,\theta_0)/\partial \theta)}{f(x,\theta_0)} f(x,\theta_0)dx \\
&= n\int \frac{\partial f}{\partial \theta}(x,\theta_0)dx \\
&= n\frac{\partial}{\partial \theta}\int f(x,\theta)dx\bigg|_{\theta=\theta_0} \\
&= n\frac{\partial}{\partial \theta}1 \\
&= 0
\end{aligned}$$

Notice: interchanged order of differentiation and integration at one point.

This step is usually justified by applying the dominated convergence theorem to the definition of the derivative.

Differentiate identity just proved:

$$\int \frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta)dx = 0$$

Take derivative of both sides wrt θ; pull derivative under integral sign:

$$\int \frac{\partial}{\partial \theta}\left[\frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta)\right] dx = 0$$

Do the derivative and get

$$\begin{aligned}
&-\int \frac{\partial^2 \log(f)}{\partial \theta^2} f(x,\theta)dx \\
&\qquad = \int \frac{\partial \log f}{\partial \theta}(x,\theta)\frac{\partial f}{\partial \theta}(x,\theta)dx \\
&\qquad\qquad = \int \left[\frac{\partial \log f}{\partial \theta}(x,\theta)\right]^2 f(x,\theta)dx
\end{aligned}$$

Definition: The **Fisher Information** is

$$I(\theta) = -E_\theta(U'(\theta)) = nE_{\theta_0}(V_1)$$

We refer to $\mathcal{I}(\theta_0) = E_{\theta_0}(V_1)$ as the information in 1 observation.

The idea is that I is a measure of how curved the log likelihood tends to be at the true value of θ. Big curvature means precise estimates. Our identity above is

$$I(\theta) = Var_\theta(U(\theta)) = n\mathcal{I}(\theta)$$

Now we return to our Taylor expansion approximation

$$-U'(\theta_0)(\hat{\theta} - \theta_0) \approx U(\theta_0)$$

and study the two appearances of U.

We have shown that $U = \sum U_i$ is a sum of iid mean 0 random variables. The central limit theorem thus proves that

$$n^{-1/2}U(\theta_0) \Rightarrow N(0, \sigma^2)$$

where $\sigma^2 = \text{Var}(U_i) = E(V_i) = \mathcal{I}(\theta)$.

Next observe that

$$-U'(\theta) = \sum V_i$$

where again

$$V_i = -\frac{\partial U_i}{\partial \theta}$$

The law of large numbers can be applied to show

$$-U'(\theta_0)/n \to E_{\theta_0}[V_1] = \mathcal{I}(\theta_0)$$

Now manipulate our Taylor expansion as follows

$$n^{1/2}(\hat{\theta} - \theta_0) \approx \left[\frac{\sum V_i}{n}\right]^{-1} \frac{\sum U_i}{\sqrt{n}}$$

Apply Slutsky's Theorem to conclude that the right hand side of this converges in distribution to $N(0, \sigma^2/\mathcal{I}(\theta)^2)$ which simplifies, because of the identities, to $N\{0, 1/\mathcal{I}(\theta)\}$.

Summary

In regular families: assuming $\hat{\theta} = \hat{\theta}_n$ is a consistent root of $U(\theta) = 0$.

- $n^{-1/2}U(\theta_0) \Rightarrow MVN(0, \mathcal{I})$ where

  $$\mathcal{I}_{ij} = \text{E}_{\theta_0}\left\{V_{1,ij}(\theta_0)\right\}$$

  and

  $$V_{k,ij}(\theta) = -\frac{\partial^2 \log f(X_k, \theta)}{\partial \theta_i \partial \theta_j}$$

- If $\mathbf{V}_k(\theta)$ is the matrix $[V_{k,ij}]$ then

$$\frac{\sum_{k=1}^n \mathbf{V}_k(\theta_0)}{n} \to \mathcal{I}$$

- If $\mathbf{V}(\theta) = \sum_k \mathbf{V}_k(\theta)$ then

$$\{\mathbf{V}(\theta_0)/n\} n^{1/2}(\hat\theta - \theta_0) - n^{-1/2}U(\theta_0) \to 0$$

  in probability as $n \to \infty$.

- Also
$$\{\mathbf{V}(\hat\theta)/n\} n^{1/2}(\hat\theta - \theta_0) - n^{-1/2}U(\theta_0) \to 0$$

  in probability as $n \to \infty$.

- $n^{1/2}(\hat\theta - \theta_0) - \{\mathcal{I}(\theta_0)\}^{-1}U(\theta_0) \to 0$ in probability as $n \to \infty$.

- $n^{1/2}(\hat\theta - \theta_0) \Rightarrow MVN(0, \mathcal{I}^{-1})$.

- In general (not just iid cases)

$$\begin{aligned}
\sqrt{I(\theta_0)}(\hat\theta - \theta_0) &\Rightarrow N(0,1) \\
\sqrt{I(\hat\theta)}(\hat\theta - \theta_0) &\Rightarrow N(0,1) \\
\sqrt{V(\theta_0)}(\hat\theta - \theta_0) &\Rightarrow N(0,1) \\
\sqrt{V(\hat\theta)}(\hat\theta - \theta_0) &\Rightarrow N(0,1)
\end{aligned}$$

  where $V = -\ell''$ is the so-called *observed information*, the negative second derivative of the log-likelihood.

Note: If the square roots are replaced by matrix square roots we can let θ be vector valued and get $MVN(0, I)$ as the limit law.

Why all these different forms? Use limit laws to test hypotheses and compute confidence intervals. Test $H_o : \theta = \theta_0$ using one of the 4 quantities as test statistic. Find confidence intervals using quantities as *pivots*. E.g.: second and fourth limits lead to confidence intervals

$$\hat\theta \pm z_{\alpha/2}/\sqrt{I(\hat\theta)}$$

and
$$\hat\theta \pm z_{\alpha/2}/\sqrt{V(\hat\theta)}$$

respectively. The other two are more complicated. For iid $N(0, \sigma^2)$ data we have
$$V(\sigma) = \frac{3\sum X_i^2}{\sigma^4} - \frac{n}{\sigma^2}$$

and

$$I(\sigma) = \frac{2n}{\sigma^2}$$

The first line above then justifies confidence intervals for σ computed by finding all those σ for which

$$\left|\frac{\sqrt{2n}(\hat\sigma - \sigma)}{\sigma}\right| \le z_{\alpha/2}$$

Similar interval can be derived from 3rd expression, though this is much more complicated.

Usual summary: mle is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information.

Problems with maximum likelihood

1. Many parameters lead to poor approximations. MLEs can be far from right answer. See homework for Neyman Scott example where MLE is not consistent.

2. Multiple roots of the likelihood equations: you must choose the right root. Start with different, consistent, estimator; apply iterative scheme like Newton Raphson to likelihood equations to find MLE. Not many steps of NR generally required if starting point is a reasonable estimate.

Finding (good) preliminary Point Estimates

Method of Moments

Basic strategy: set sample moments equal to population moments and solve for the parameters.

Definition: The r^{th} sample moment (about the origin) is

$$\frac{1}{n}\sum_{i=1}^n X_i^r$$

The r^{th} population moment is

$$\mathrm{E}(X^r)$$

(**Central** moments are

$$\frac{1}{n}\sum_{i=1}^n (X_i - \bar X)^r$$

and

$$\mathrm{E}\left[(X-\mu)^r\right] .$$

If we have p parameters we can estimate the parameters $\theta_1, \ldots, \theta_p$ by solving the system of p equations:

$$\mu_1 = \bar X$$

$$\mu_2' = \overline{X^2}$$

and so on to

$$\mu_p' = \overline{X^p}$$

You need to remember that the population moments μ_k' will be formulas involving the parameters.

Gamma Example

The Gamma(α,β) density is

$$f(x;\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\frac{x}{\beta}\right]1(x>0)$$

and has

$$\mu_1 = \alpha\beta$$

and

$$\mu_2' = \alpha(\alpha+1)\beta^2$$

This gives the equations

$$\begin{aligned}\alpha\beta &= \overline{X} \\ \alpha(\alpha+1)\beta^2 &= \overline{X^2}\end{aligned}$$

or

$$\begin{aligned}\alpha\beta &= \overline{X} \\ \alpha\beta^2 &= \overline{X^2} - \overline{X}^2\end{aligned}$$

Divide the second equation by the first to find the method of moments estimate of β is

$$\tilde\beta = (\overline{X^2} - \overline{X}^2)/\overline{X}\,.$$

Then from the first equation get

$$\tilde\alpha = \overline{X}/\tilde\beta = (\overline{X})^2/(\overline{X^2} - \overline{X}^2)\,.$$

The method of moments equations are much easier to solve than the likelihood equations which involve the function

$$\psi(\alpha) = \frac{d}{d\alpha}\log(\Gamma(\alpha))$$

called the digamma function.

Score function has components

$$U_\beta = \frac{\sum X_i}{\beta^2} - n\alpha/\beta$$

and

$$U_\alpha = -n\psi(\alpha) + \sum\log(X_i) - n\log(\beta)\,.$$

You can solve for β in terms of α to leave you trying to find a root of the equation

$$-n\psi(\alpha) + \sum \log(X_i) - n \log(\sum X_i/(n\alpha)) = 0$$

To use Newton Raphson on this you begin with the preliminary estimate $\hat\alpha_1 = \tilde\alpha$ and then compute iteratively

$$\hat\alpha_{k+1} = \frac{\overline{\log(X)} - \psi(\hat\alpha_k) - \log(\overline{X})/\hat\alpha_k}{1/\alpha - \psi'(\hat\alpha_k)}$$

until the sequence converges. Computation of ψ', the trigamma function, requires special software. Web sites like *netlib* and *statlib* are good sources for this sort of thing.

Estimating Equations

Same large sample ideas arise whenever estimates derived by solving some equation.

Example: large sample theory for **Generalized Linear Models**.

Suppose Y_i is number of cancer cases in some group of people characterized by values x_i of some covariates.

Think of x_i as containing variables like age, or a dummy for sex or average income or ....

Possible parametric regression model: Y_i has a Poisson distribution with mean μ_i where the mean μ_i depends somehow on x_i.

Typically assume $g(\mu_i) = \beta_0 + x_i\beta$; g is **link** function.

Often $g(\mu) = \log(\mu)$ and $x_i\beta$ is a matrix product: x_i row vector, μ column vector.

“Linear regression model with Poisson errors”.

Special case $\log(\mu_i) = \beta x_i$ where x_i is a scalar.

The log likelihood is simply

$$\ell(\beta) = \sum (Y_i \log(\mu_i) - \mu_i)$$

ignoring irrelevant factorials. The score function is, since $\log(\mu_i) = \beta x_i$,

$$U(\beta) = \sum (Y_i x_i - x_i\mu_i) = \sum x_i(Y_i - \mu_i)$$

(Notice again that the score has mean 0 when you plug in the true parameter value.) The key observation, however, is that it is not necessary to believe that Y_i has a Poisson distribution to make solving the equation $U = 0$ sensible. Suppose only that $\log(E(Y_i)) = x_i\beta$. Then we have assumed that

$$E_\beta(U(\beta)) = 0$$

This was the key condition in proving that there was a root of the likelihood equations which was consistent and here it is what is needed, roughly, to prove that the equation $U(\beta) = 0$ has a consistent root $\hat\beta$. Ignoring higher order terms in a Taylor expansion will give

$$V(\beta)(\hat\beta - \beta) \approx U(\beta)$$

where $V = -U'$. In the mle case we had identities relating the expectation of V to the variance of U. In general here we have

$$\text{Var}(U) = \sum x_i^2 \text{Var}(Y_i) \, .$$

If Y_i is Poisson with mean μ_i (and so $\text{Var}(Y_i) = \mu_i$) this is

$$\text{Var}(U) = \sum x_i^2 \mu_i \, .$$

Moreover we have

$$V_i = x_i^2 \mu_i$$

and so

$$V(\beta) = \sum x_i^2 \mu_i \, .$$

The central limit theorem (the Lyapunov kind) will show that $U(\beta)$ has an approximate normal distribution with variance $\sigma_U^2 = \sum x_i^2 \text{Var}(Y_i)$ and so

$$\hat\beta - \beta \approx N(0, \sigma_U^2 / (\sum x_i^2 \mu_i)^2)$$

If $\text{Var}(Y_i) = \mu_i$, as it is for the Poisson case, the asymptotic variance simplifies to $1/\sum x_i^2 \mu_i$.

Other estimating equations are possible, popular. If w_i is any set of deterministic weights (possibly depending on μ_i) then could define

$$U(\beta) = \sum w_i (Y_i - \mu_i)$$

and still conclude that $U = 0$ probably has a consistent root which has an asymptotic normal distribution.

Idea widely used:

Example: Generalized Estimating Equations, Zeger and Liang.

Abbreviation: GEE.

Called by econometricians Generalized Method of Moments.

An estimating equation is unbiased if

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

Theorem: Suppose $\hat\theta$ is a consistent root of the unbiased estimating equation

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

Let $V = -U'$. Suppose there is a sequence of constants $B(\theta)$ such that

$$V(\theta)/B(\theta) \to 1$$

and let

$$A(\theta) = Var_\theta(U(\theta))$$

and

$$C(\theta) = B(\theta) A^{-1}(\theta) B(\theta).$$

Then

$$\sqrt{C(\theta_0)}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{C(\hat{\theta})}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$

Other ways to estimate A, B and C lead to same conclusions. There are multivariate extensions using matrix square roots.