

STAT 830

Likelihood Asymptotics

Richard Lockhart

Simon Fraser University

STAT 830 — Fall 2011



Purposes of These Notes

- Discuss the behaviour of mles in large samples.
- Show log-likelihood is nearly quadratic.
- Emphasize local rather than global behaviour.
- Give sequence of examples.



Large Sample Theory

- Study approximate behaviour of $\hat{\theta}$ by studying the function U .
- Notice U is sum of independent random variables.

Theorem

If Y_1, Y_2, \dots are iid with mean μ then

$$\frac{\sum Y_i}{n} \rightarrow \mu$$

- Law of large numbers. Strong law

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

and the weak law that

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

- For iid Y_i the stronger conclusion holds; for our heuristics ignore differences between these notions.



Score function at true value of θ

- Now suppose θ_0 is true value of θ .
- Then

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

where

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



Normal example

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

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

- If the true mean is μ_0 then $\bar{X} \rightarrow \mu_0$ and

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

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



Same ideas in more general case

- Study rv

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

- You know the inequality

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

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

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

$$g(E(X)) \leq E(g(X))$$



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

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

- But

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

- Reassemble the inequality and this calculation to get

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



- 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

$$\frac{\ell(\theta) - \ell(\theta_0)}{n} = \frac{\sum \log[f(X_i, \theta)/f(X_i, \theta_0)]}{n} \rightarrow \mu(\theta)$$

- 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**; need **uniform** convergence in θ .



- **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

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

- Suppose: $\hat{\theta}_n$ is global maximizer of ℓ .
- $\hat{\theta}_{n,\delta}$ maximizes ℓ over $N(\delta) = \{|\theta - \theta_0| \leq \delta\}$.

$$A_\epsilon = \{|\hat{\theta}_n - \theta_0| \leq \epsilon\}$$

$$B_{\delta,\epsilon} = \{|\hat{\theta}_{n,\delta} - \theta_0| \leq \epsilon\}$$

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



Some precision

Theorem

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

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):

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

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 .



Asymptotic normality continued

- 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' .



Asymptotic normality continued

- 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:

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



- 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

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



- **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

$$I(\theta) = \text{Var}_{\theta}(U(\theta)) = n\mathcal{I}(\theta)$$

- 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 .

- Have shown $U = \sum U_i$ 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

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

- 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

$$\mathcal{I}_{ij} = E_{\theta_0} \{V_{1,ij}(\theta_0)\}$$

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

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

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

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

in probability as $n \rightarrow \infty$.



Summary Continued

- Also

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

in probability as $n \rightarrow \infty$.

- $n^{1/2}(\hat{\theta} - \theta_0) - \{\mathcal{I}(\theta_0)\}^{-1}U(\theta_0) \rightarrow 0$ in probability as $n \rightarrow \infty$.
- $n^{1/2}(\hat{\theta} - \theta_0) \Rightarrow MVN(0, \mathcal{I}^{-1})$.
- In general (not just iid cases)

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

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

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

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

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

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

- Similar interval can be derived from 3rd expression, though this is much more complicated.
- Usual summary: mle is consistent and asymptotically normal with 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.
- 2 See homework for Neyman Scott example where MLE is not consistent.
- 3 Multiple roots of the likelihood equations: you must choose the right root.
- 4 Start with different, consistent, estimator; apply iterative scheme like Newton Raphson to likelihood equations to find MLE.
- 5 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

$$E(X^r)$$

- (**Central** moments are

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

and

$$E[(X - \mu)^r] .$$



Method of moments continued

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

$$\mu_1 = \bar{X}$$

$$\mu'_2 = \overline{X^2}$$

and so on to

$$\mu'_p = \overline{X^p}$$

- Remember that population moments μ'_k are formulas involving the parameters.



Gamma Example

- The Gamma(α, β) density is

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

and has

$$\mu_1 = \alpha\beta$$

and

$$\mu'_2 = \alpha(\alpha + 1)\beta^2.$$

- This gives the equations

$$\alpha\beta = \bar{X}$$

$$\alpha(\alpha + 1)\beta^2 = \overline{X^2}$$

or

$$\alpha\beta = \bar{X}$$

$$\alpha\beta^2 = \overline{X^2} - \bar{X}^2.$$



Gamma continued

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

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

- Then from the first equation get

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

- Method of moments equations much easier to solve than likelihood equations which involve *digamma* ftn

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

- Score function has components

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

and

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



Gamma continued

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

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

- To use Newton Raphson on this you begin with the preliminary estimate $\hat{\alpha}_1 = \tilde{\alpha}$ and then compute iteratively

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

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, β column vector.



GLM: “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 (ignoring irrelevant factorials)

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

- 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.
- Key observation: no need to believe Y_i has Poisson distribution to make solving equation $U = 0$ sensible.
- Suppose only that $\log(E(Y_i)) = x_i \beta$.
- Then we have assumed that $E_\beta(U(\beta)) = 0$.
- Key condition to prove existence of consistent root of likelihood equations; here needed, roughly, to prove equation $U(\beta) = 0$ has 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 mle case had identities relating expectation of V to 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

- If w_i is any set of deterministic weights (possibly depending on μ_i) then could define

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

- Can 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.

Def'n: An estimating equation is unbiased if

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



Unbiased estimating equations

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

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

and let

$$A(\theta) = \text{Var}_{\theta}(U(\theta)) \text{ and } C(\theta) = B(\theta)A^{-1}(\theta)B(\theta).$$

Then

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

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

Extras

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

