STAT 801 Lecture 12

Goals of Today's Lecture:

- Develop large sample theory for the mle.
- Define and interpret Fisher information.
- Extend the ideas to estimating equations.

Today's notes

For $X_1, \ldots, X_n$ iid log likelihood is

$$\ell(\theta) = \sum \log(f(X_i, \theta)) .$$

The score function is

$$U(\theta) = \sum \frac{\partial \log f}{\partial \theta}(X_i, \theta) .$$

MLE $\hat\theta$ maximizes ℓ. If maximum occurs in interior of parameter space and the log likelihood continuously differentiable then $\hat\theta$ solves the likelihood equations

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

Examples

$\mathbf{N}(\mu, \sigma^2)$

Unique root of likelihood equations is a global maximum.

[Remark: Suppose we called $\tau = \sigma^2$ the parameter. Score function still has two components: first component same as before but second component is

$$\frac{\partial}{\partial \tau}\ell = \frac{\sum (X_i - \mu)^2}{2\tau^2} - \frac{n}{2\tau}$$

Setting the new likelihood equations equal to 0 still gives

$$\hat\tau = \hat\sigma^2$$

General **invariance** (or **equivariance**) principal: Suppose $\phi = g(\theta)$ is some reparametrization of a model (a one to one relabelling of the parameter values). Then $\hat\phi = g(\hat\theta)$. Does not apply to other estimators.]

Cauchy: location θ

At least 1 root of likelihood equations but often several more. One root is a global maximum; others, if they exist may be local minima or maxima.

Binomial(n, θ)

If $X = 0$ or $X = n$: no root of likelihood equations; likelihood is monotone. Other values of X: unique root, a global maximum. Global maximum at $\hat{\theta} = X/n$ even if $X = 0$ or n.

The 2 parameter exponential

The density is

$$f(x; \alpha, \beta) = \frac{1}{\beta} e^{-(x-\alpha)/\beta} 1(x > \alpha)$$

Log-likelihood is $-\infty$ for $\alpha > \min\{X_1, \ldots, X_n\}$ and otherwise is

$$\ell(\alpha, \beta) = -n \log(\beta) - \sum (X_i - \alpha)/\beta$$

Increasing function of α till α reaches

$$\hat{\alpha} = X_{(1)} = \min\{X_1, \ldots, X_n\}$$

which gives mle of α. Now plug in $\hat{\alpha}$ for α; get so-called profile likelihood for β:

$$\ell_{\text{profile}}(\beta) = -n \log(\beta) - \sum (X_i - X_{(1)})/\beta$$

Set β derivative equal to 0 to get

$$\hat{\beta} = \sum (X_i - X_{(1)})/n$$

Notice mle $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ does *not* solve likelihood equations; we had to look at the edge of the possible parameter space. α is called a *support* or *truncation* parameter. ML methods behave oddly in problems with such parameters.

Three parameter Weibull

The density in question is

$$\begin{aligned} f(x; \alpha, \beta, \gamma) = & \frac{1}{\beta} \left(\frac{x - \alpha}{\beta} \right)^{\gamma - 1} \\ & \times \exp[-\{(x - \alpha)/\beta\}^\gamma] 1(x > \alpha) \end{aligned}$$

Three likelihood equations: Set β derivative equal to 0; get

$$\hat{\beta}(\alpha,\gamma) = \left[\sum (X_i - \alpha)^\gamma / n\right]^{1/\gamma}$$

where $\hat{\beta}(\alpha,\gamma)$ indicates mle of β could be found by finding the mles of the other two parameters and then plugging in to the formula above. It is not possible to find explicitly the remaining two parameters; numerical methods are needed. However, you can see that putting $\gamma < 1$ and letting $\alpha \to X_{(1)}$ will make the log likelihood go to ∞. The mle is not uniquely defined, then, since any $\gamma < 1$ and any β will do.

If the true value of γ is more than 1 then the probability that there is a root of the likelihood equations is high; in this case there must be two more roots: a local maximum and a saddle point! For a true value of $\gamma > 0$ the theory we detail below applies to the local maximum and not to the global maximum of the likelihood equations.

Large Sample Theory

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

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

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

This is called the law of large numbers. The strong law says

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

and the weak law that

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

For iid Y_i the stronger conclusion holds but for our heuristics we will ignore the differences between these notions.

Now suppose θ_0 is true value of θ. Then

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

where

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

Consider as an example the case of $N(\mu, 1)$ data where

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

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

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

If we think of a $\mu < \mu_0$ we see that the derivative of $\ell(\mu)$ is likely to be positive so that ℓ increases as we increase μ. For μ more than μ_0 the derivative is probably negative and so ℓ tends to be decreasing for $\mu > 0$. It follows that ℓ is likely to be maximized close to μ_0.

Repeat ideas for more 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 $Var(X) \ge 0$.) A generalization is called 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

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

It is possible to prove that the 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

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

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

Now let us study the shape of the log likelihood near the true value of $\hat{\theta}$ under the assumption that $\hat{\theta}$ is a root of the likelihood equations close to θ_0. We use Taylor expansion to write, for a 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 $\hat{\theta}$. (This form of the remainder in Taylor's theorem is not valid for multivariate θ.) The derivatives of U are each sums of n terms and so should be both proportional to n in size. The 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. If we ignore the second derivative term we get

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

Now let's look at the terms U and U'.

In the normal case

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

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

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

and the next derivative U'' is 0. We will analyze the general case by noticing that 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:

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

Notice that I have interchanged the 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. The same tactic can be applied by differentiating the identity which we just proved

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

Taking the derivative of both sides with respect to θ and pulling the derivative under the integral sign again gives

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

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

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

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

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

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

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:

- Under strong regularity conditions Jensen's inequality can be used to demonstrate that $\hat{\theta}$ which maximizes ℓ globally is consistent and that this $\hat{\theta}$ is a root of the likelihood equations.

- It is generally easier to study ℓ only close to θ_0. For instance define A to be the event that ℓ is concave on the set of θ such that $|\theta - \theta_0| < \delta$ and the likelihood equations have a unique root in that set. Under weaker conditions than the previous case we can prove that there is a $\delta > 0$ such that
$$P(A) \to 1$$
In that case we can prove that the root $\hat{\theta}$ of the likelihood equations mentioned in the definition of A is consistent.

- Sometimes get an even weaker conclusion. Define B to be the event that $\ell(\theta)$ is concave for $n^{1/2}|\theta - \theta_0| < L$ and there is a unique root of ℓ over this range. Again this root is consistent but there might be other consistent roots of the likelihood equations.

- Under any of these scenarios there is a consistent root of the likelihood equations which is definitely the closest to the true value θ_0. This root $\hat{\theta}$ has the property
$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, 1/\mathcal{I}(\theta))\,.$$

We usually simply say that the mle is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information. This assertion is actually valid for vector valued θ where now I is a matrix with ijth entry
$$I_ij = -E\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right)$$

Estimating Equations

Same ideas arise whenever estimates derived by solving some equation. Example: large sample theory for **Generalized Linear Models**.

Suppose that for $i = 1, \ldots, n$ we have observations of the numbers of cancer cases Y_i in some group of people characterized by values x_i of some covariates. You are supposed to think of x_i as containing variables like age, or a dummy for sex or average income or ... A parametric regression model for the Y_i might postulate that Y_i has a Poisson distribution with mean μ_i

where the mean μ_i depends somehow on the covariate values. Typically we might assume that $g(\mu_i) = \beta 0 + x_i\beta$ where g is a so-called **link** function, often for this case $g(\mu) = \log(\mu)$ and $x_i\beta$ is a matrix product with x_i written as a row vector and μ a column vector. This is supposed to function as a "linear regression model with Poisson errors". I will do as a 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

$$Var(U) = \sum x_i^2 Var(Y_i)$$

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

$$Var(U) = \sum x_i^2 \mu_i$$

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 $\sigma_U^2 = \sum x_i^2 Var(Y_i)$ and so

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

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

Notice that other estimating equations are possible. People suggest alternatives very often. If w_i is any set of deterministic weights (even possibly depending on μ_i then we 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: see, e.g., Zeger and Liang's idea of Generalized Estimating Equations (GEE) which the econometricians call Generalized Method of Moments.

Problems with maximum likelihood

1. In problems with many parameters the approximations don't work very well and maximum likelihood estimators can be far from the right answer. See your homework for the Neyman Scott example where the mle is not consistent.

2. When there are multiple roots of the likelihood equation you must choose the right root. To do so you might start with a different consistent estimator and then apply some iterative scheme like Newton Raphson to the likelihood equations to find the mle. It turns out not many steps of NR are generally required if the 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

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