MSF100
Statistical Inference Principles - Lecture 4

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

January 26, 2012

1 The Score function

Let's revisit the likelihood function. We have that $L(\theta|\tilde{x})$ is a function of θ. A natural estimator for θ is the maximizer of this function: $\hat{\theta}_n = \arg\max_\theta L(\theta|\tilde{x})$.

Properties of the MLE

- Consistent: $\hat{\theta}_n \to \theta$ in probability.
- Equivariant: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $\tau = g(\theta)$
- Asymptotically normally distributed
- Asymptotically most efficient (minimum variance among unbiased estimators)

In the upcoming lectures we will go through these properties.

Let us first establish the notation.

- The likelihood function $L(\theta|\tilde{x}) = L_n(\theta)$. We write with subscript n to emphasize that the likelihood depends on the sample.
- The loglikelihood function $l_n(\theta) = \log L_n(\theta)$.
- The *Score* function $S_n(\theta) = \frac{d}{d\theta} l_n(\theta)$ (often written without the n, just remember it depends on the sample).
- The MLE (maximum likelihood estimator) is given by the zero-crossing of the score function: $\hat{\theta}_n = \{\theta : S_n(\theta) = 0\}$

Now, under so-called *regularity conditions* (such that we can change the order of integration and derivation, puts some smoothness constraints on the pdf $f_\theta(x)$) we have

$$E_\theta[S_n(\theta)] = 0$$

This follows from

$$E_\theta[S_n(\theta)] = \int \cdots \int \frac{d \log f(x_i|\theta)}{d\theta} dx_1 \cdots dx_n = \int \cdots \int \frac{\frac{df(\tilde{x}|\theta)}{d\theta}}{f(\tilde{x}|\theta)} f(\tilde{x}|\theta) dx_1 \cdots dx_n = \frac{d}{d\theta} \int \cdots \int f(\tilde{x}|\theta) dx_1 \cdots dx_n = 0$$

So, the expected value of the score is 0 if we take the expectation at the same θ as we compute the score at. This motivates using the zero-crossing of the observed score as our estimator for θ.

As we discussed in previous lectures, it is not only the maximizer of the likelihood function that is of interest, but also its shape since the shape tells us how "easy" or "difficult" the estimation problem is.

How can we summarize the shape? The complete shape is only captured if we look at $l(\theta)$ or $L(\theta)$ in its entirety. However, locally near the maximum peak of $l(\theta)$ we can summarize the function by the *curvature* $l''(\theta)$. The curvature of the loglikelihood function is

$$l''(\theta) = \frac{d^2}{d\theta^2} \log f(\tilde{x}|\theta).$$

We evaluate this at the maximum and denote this quantity as the *observed Fisher information*

$$i(\hat{\theta}) = l''(\hat{\theta})$$

The *expected* value of the curvature is called the *Fisher information*:

$$FI(\theta) = I_n(\theta) = E_\theta[i(\theta)]$$

We have $E_\theta[S(\theta)] = 0$ and so it follows that

$$Var_\theta[S(\theta)] = E_\theta[S(\theta)^2] = E_\theta[(\frac{d}{d\theta} \log f(\tilde{x}|\theta))^2]$$

We can write this on a form that is easier to compute:

$$\frac{d^2 \log f(\tilde{x}|\theta)}{d\theta^2} = \frac{d}{d\theta}(\frac{\frac{d}{d\theta} f(\tilde{x}|\theta)}{f(\tilde{x}|\theta)}) = \frac{f(\tilde{x}|\theta)\frac{d^2}{d\theta^2} f(\tilde{x}|\theta) - (\frac{d}{d\theta} f(\tilde{x}|\theta))^2}{(f(\tilde{x}|\theta))^2} =$$

$$= \frac{\frac{d^2}{d\theta^2} f(\tilde{x}|\theta)}{f(\tilde{x}|\theta)} - (\frac{\frac{d}{d\theta} f(\tilde{x}|\theta)}{f(\tilde{x}|\theta)})^2$$

Take the expected value and we have that

$$E_\theta[-\frac{d^2 l(\theta)}{d\theta^2}] = E_\theta[(S(\theta))^2] - E_\theta[\frac{\frac{d^2}{d\theta^2} f(\tilde{x}|\theta)}{f(\tilde{x}|\theta)}] = E_\theta[(S(\theta))^2]$$

It is often easier to compute the expected value of $l''(\theta)$ than the expected value of the square of the score.

Now, we have that $\hat{\theta}$ is the zero-crossing of the score function. We also know that the score has expected value 0 if we match the θ of derivation and expectation (i.e. no model misspecification). In addition, we have that the variance of the score function is $E_\theta[-\frac{d^2 l(\theta)}{d\theta^2}]$. We will see later that these properties of the score function directly translates to properties of the MLE.

2 Regular likelihood function, quadratic approximation

Let us again consider the likelihood function in its total. Now, *if* $l(\theta)$ is quadratic, then the location of the peak and the curvature at the peak is sufficient to summarize the entire function. Even if the entire function is not quadratic, it could be that $l(\theta)$ is near quadratic near its peak. If so, we call the loglikelihood function *regular* and we can derive many results useful for inference in this case.

How well the likelihood function can be approximated by a quadratic depends on the underlying model family, the sample size and the particular outcome. In Figure 1 we see some examples. Locally, all of these likelihood functions are quite well approximated by a quadratic, especially when $x = 5$. What am I comparing the loglikelihood functions to in the figures? Well, I use a Taylor expansion of the loglikelihood function hear $\hat{\theta}$ to obtain

$$\log L(\theta) \simeq \log L(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2.$$

This follows from the definition of the score and information function (note, here the *observed* information). Since $\hat{\theta}$ is obtained at the zero-crossing of the score function, the linear term goes away and we have

$$\log \frac{L(\theta)}{L(\hat{\theta})} \simeq -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2.$$

For the Binomial example, we have that

$$S(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}, \ \hat{\theta} = x/n$$

and

$$\frac{d^2 l}{d\theta^2} = \frac{-x}{\theta^2} + \frac{n-x}{(1-\theta)^2}$$

and so the observed information is

$$I(\hat{\theta}) = \frac{n}{\hat{\theta}} - \frac{n}{1-\hat{\theta}} = \frac{n}{\hat{\theta}(1-\hat{\theta})}$$

I plot $\log L(\theta)/L(\hat{\theta})$ (solid lines) and $(-1/2) * I(\hat{\theta})(\theta - \hat{\theta})^2$ in the figure (dashed line). (Try this at home with a smaller or larger sample size and for different x.)

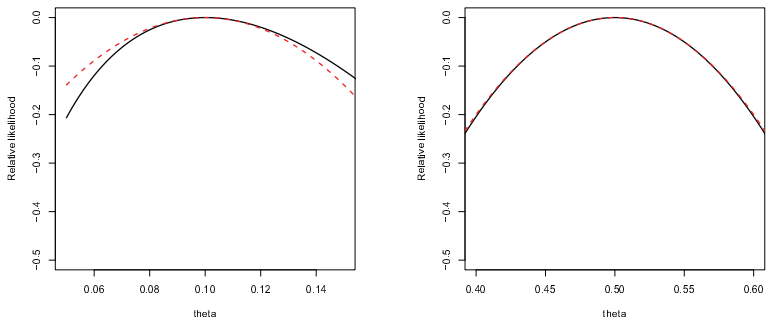

Figure 1: loglikelihood functions for $X \sim Bin(n, \theta)$. (Left): $n = 10$ and $x = 1$ (black), quadratic approximation (red). (Right): $n = 10$ and $x = 5$ (black) and quadratic approximation (red).

Now, let us take the derivative of the relative loglikelihood and the quadratic approximation and equate the two parts:

$$\frac{d}{d\theta} \log \frac{L(\theta)}{L(\hat{\theta})} = S(\theta) \simeq I(\hat{\theta})(\hat{\theta} - \theta).$$

We have that $E_\theta[S(\theta)] = 0$ and $Var_\theta[S(\theta)] = E_\theta[I(\theta)]$.
I use these properties to standardize the score function on the LH above (note, here I am replacing the expected value by its observed - we will return to check if this is OK later):

$$I^{-1/2}(\hat{\theta}) S(\theta) \simeq I^{1/2}(\hat{\theta})(\hat{\theta} - \theta)$$

The right hand (RH) is of great interest to us: from this we should be able to say something about the properties of $\hat{\theta}$. The LH side has now variance 1 and so it follows that the variance of $\hat{\theta}$ is $I^{-1}(\hat{\theta})$. What does this tell us? Well, the curvature of the likelihood is measured by the information and the sharper the curvature the smaller the estimation variance for $\hat{\theta}$!

Example:

- Let X_i be iid $f(x|\theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}} I_{\{\nu, \infty\}}(x)$

- It follows that the likelihood function is

$$L(\theta,\nu)=\frac{\theta^n\nu^{n\theta}}{(\prod_i x_i)^{\theta+1}}I\{x_{(1)}>\nu\}$$

- We see that $X_{(1)},\prod_i X_i$ are the sufficient statistics for θ,ν.
- If θ is known, the loglikelihood for ν is increasing in ν and gives $\hat{\nu}=x_{(1)}$
- Given ν and $x_{(1)}\geq\nu$ we have the score for θ

$$S(\theta)=\frac{n}{\theta}+n\log(\nu)-\sum_i\log(x_i)$$

from which we find $\hat{\theta}=\frac{n}{\sum_i\log(x_i)-n\log(\nu)}$.

Example:

- Let X_i iid double-exponential pdf: $f(x|\theta)=\frac{1}{2}e^{-|x-\theta|}$.
- The likelihood function is $L(\theta)=\frac{1}{2^n}e^{-\sum_i|x_i-\theta|}$.
- $L(\theta)$ is maximized when $\sum_i|x_i-\theta|$ is minimized
- To simplify this expression, consider the case $x_{(j)}\leq\theta\leq x_{(j+1)}$. We can then divide the sum above into two parts:

$$\sum_i|x_i-\theta|=\sum_{i=1}^{j}\theta-x_{(i)}+\sum_{i=j+1}^{n}x_{(i)}-\theta$$

$$=(2j-n)\theta-\sum_{i=1}^{j}x_{(i)}+\sum_{i=j+1}^{n}x_{(i)}$$

- This is linear in θ and decreases for j less than $n/2$ and increases for j greater than $n/2$
- If n is even, the likelihood is constant between $x_{(n/2)}$ and $x_{(n/2+1)}$ and so any θ in this interval is the MLE, we usually take the midpoint.
- If n is odd, the likelihood is maximized for $\hat{\theta}=x_{(n+1/2)}$
- This is the *median*

Example:

- Let X_i be iid $f(x|\theta)=\log(\theta)\theta^x/(\theta-1)$, $0<x<1$, $\theta>1$
- We compute the score function:

$$\frac{d}{d\theta}\log L(\theta)=\frac{d}{d\theta}\sum_i[\log\log\theta-\log(\theta-1)+x_i\log(\theta)]=$$

$$=\sum_i(\frac{1}{\theta\log(\theta)}-\frac{1}{\theta-1})+\frac{1}{\theta}\sum_i x_i=\frac{n}{\theta\log(\theta)}-\frac{n}{\theta-1}+\frac{n\bar{x}}{\theta}=\frac{n}{\theta}[\bar{x}-(\frac{\theta}{\theta-1}-\frac{1}{\log(\theta)})]$$

- Remember the Cramer-Rao bound is achieved for estimators W of $\tau(\theta)$ for which $W-\tau(\theta)$ is directly proportional to the score (linear in the score).
- Therefore, $\bar{x}$ is an unbiased estimator of $(\frac{\theta}{\theta-1}-\frac{1}{\log(\theta)})$ and achieves the CR bound for this function of θ.

Example:

- X_i are independently distributed $f_{X_i}(x|\lambda)=\lambda w_i e^{-\lambda w_i x},\ x>0,\lambda>0$ with known constants w_i

- The likelihood function is $L(\lambda) = \lambda^n(\prod_i w_i)e^{-\lambda\sum_i w_i x_i}1\{x_{(1)} > 0\}$
- This is an exponential family, so $T = \sum_i w_i X_i$ is a complete sufficient statistic for λ
- The loglikelihood is $l(\lambda) = n\log(\lambda) + \log(\prod_i w_i) - \lambda\sum_i w_i x_i$
- The score function is $S(\lambda) = l'(\lambda) = \frac{n}{\lambda} - \sum_i w_i$
- We obtain the MLE as $\hat{\lambda} = \frac{\sum_i w_i X_i}{n}$
- Does it achieve the CR bound?
- Well, the score can be written as $S(\lambda) = -n(\frac{\sum_i w_i x_i}{n} - \frac{1}{\lambda})$ and so, by the corollary of the CR theorem, $W(\tilde{X}) = \frac{\sum_i w_i x_i}{n}$ achieves the CR bound if it is an unbiased estimator of $\frac{1}{\lambda}$
- We have that $E[W] = \frac{1}{n}\sum_i w_i E[X_i] = \frac{1}{\lambda}$ so W achieves the CR bound for estimating $\frac{1}{\lambda}$. The MLE for $\frac{1}{\lambda}$ is W (since $\frac{1}{\hat{\lambda}} = W$)
- The MLE for λ is not of the required form for achieving the CR bound, but it does so asymptotically.
- What are the CR bounds for λ and $1/\lambda$ respectively?
- For $1/\lambda$ we can get to it in two ways (since we know W achieves the CR bound). First, we compute $E[-l''(\lambda)] = E[\frac{n}{\lambda^2}]$ and so we get that the CR bound for estimating λ is λ^2/n.
- The expected value of W is $\tau(\lambda) = 1/\lambda$ and so the CR bound is $\frac{(\tau'(\lambda))^2}{n/\lambda^2} = \frac{1}{n\lambda^2}$
- We know from above that W achieves this bound. We can also compute this directly in this case: $V(W) = \sum_i w_i^2 V(X_i)/n^2$, where $V(X_i) = E(X_i^2) - (E(X_i))^2 = \frac{2}{\lambda^2 w_i^2} - (\frac{1}{\lambda w_i})^2 = \frac{1}{\lambda^2 w_i^2}$ and so $V(W) = \frac{1}{n\lambda^2}$