STA 114: Statistics

Notes 8. ML Confidence Intervals based on Normal Approximation

Coverage probability calculations for non-normal models

Constructing an ML interval is conceptually simple. You can do it the moment you have got a handle on the likelihood function and chosen a threshold. But calculating the coverage probabilities, and the confidence coefficient of such an interval procedure can be a challenge. For the normal pdf model $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, the confidence coefficient of an ML interval for μ can be calculated exactly, irrespective of whether σ is a fixed variable, or an unknown model parameter. For model consisting of non-normal pdfs/pmfs, such exact calculations are rarely possible. But, astoundingly, a large number of such models can be well approximated by a normal model. This is what we shall explore today.

Asymptotic Normality of the MLE

We shall consider models of the form $X_1, \cdots, X_n \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\theta \in \Theta$, where θ is a scalar. Let $\dot{\ell}_x(\theta)$ and $\ddot{\ell}_x(\theta)$ denote the first and second order derivatives (w.r.t. θ) of the log-likelihood function $\ell_x(\theta)$.

Assume a unique MLE $\hat{\theta}_{\text{MLE}}(x)$ exists. For a fixed θ_0 inside Θ, a one term Taylor expansion of $\dot{\ell}_x(\theta_0)$ around $\hat{\theta}_{\text{MLE}}(x)$, gives

$$\dot{\ell}_x(\theta_0) = \dot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) + (\theta_0 - \hat{\theta}_{\text{MLE}}(x))\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) + R(x)$$

where $R(x)$ is the remainder term. Now, $\dot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) = 0$ and $\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) = -I_x$, so we can rearrange the above equation to write

$$\sqrt{I_x}(\hat{\theta}_{\text{MLE}}(x) - \theta_0) = \frac{\dot{\ell}_x(\theta_0)}{\sqrt{I_x}} + \tilde{R}(x)$$

for a new remainder term $\tilde{R}(x) = -R(x)/\sqrt{I_x}$.

The desired result. We will argue that when $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$, $\sqrt{I_X}(\hat{\theta}_{\text{MLE}}(X) - \theta_0)$ is approximately a $\mathsf{Normal}(0,1)$ random variable, for all large n. From the above equality, it suffices to argue that $\dot{\ell}_X(\theta_0)/\sqrt{I_X}$ is approximately $\mathsf{Normal}(0,1)$ and that $\tilde{R}(X)$ is negligible.

The crux of approximate normality. Note that

$$\dot{\ell}_X(\theta_0) = \sum_{i=1}^{n} \dot{s}_{\theta_0}(X_i)$$

where $s_\theta(x_i) = \frac{\partial}{\partial\theta}\log g(x_i|\theta)$. Therefore, $\dot{\ell}_X(\theta_0)$ is the sum of n IID random variables $s_{\theta_0}(X_i)$, and hence by CLT, is approximately $\mathsf{Normal}(na, nb^2)$ where $a = \mathrm{E}_{[X_1|\theta_0]} s_{\theta_0}(X_1)$ and $b^2 = \mathrm{Var}_{[X_1|\theta_0]} s_{\theta_0}(X_1)$. This crucial observation leads to $\dot{\ell}_X(\theta_0)/\sqrt{I_X}$ being approximately $\mathsf{Normal}(0,1)$ provided we can show $a = 0$ and $b^2 \approx I_X/n$.

Proving $a = 0$. Note that for any θ,

$$s_\theta(x_i) = \frac{\frac{\partial}{\partial\theta} g(x_i|\theta)}{g(x_i|\theta)}, \text{ and hence, } E_{[X_1|\theta]} s_\theta(X_1) = \int s_\theta(x_1) g(x_1|\theta) dx_1 = \int \frac{\partial}{\partial\theta} g(x_1|\theta) dx_1.$$

Under certain regularity conditions of the pdfs (or pmfs) $g(x_i|\theta)$, the integration and differentiation operations can be interchanged in the last term above. This gives,

$$\mathrm{E}_{[X_1|\theta]} s_\theta(X_1) = \frac{\partial}{\partial\theta}\int g(x_1|\theta) dx_1 = \frac{\partial}{\partial\theta}\{1\} = 0.$$

Because this identity holds for every θ, we conclude $a = \mathrm{E}_{[X_1|\theta_0]} s_{\theta_0}(X_1) = 0$.

Proving $b^2 \approx I_X/n$ and the rest of the argument. Again for any θ, because $\mathrm{E}_{[X_1|\theta]} s_\theta(X_1) = 0$, we have $\mathrm{Var}_{[X_1|\theta]} s_\theta(X_1) = \mathrm{E}_{[X_1|\theta]} s_\theta^2(X_1)$. This quantity is called the (single observation) Fisher information at θ of the model under consideration, and is denoted $I_1^F(\theta)$. An interesting fact is $I_1^F(\theta) = -\mathrm{E}_{[X_1|\theta]}\frac{\partial^2}{\partial\theta^2}\log g(X_1|\theta)$. This holds because

$$\frac{\partial^2}{\partial\theta^2}\log g(x_i|\theta) = \frac{\frac{\partial^2}{\partial\theta^2} g(x_i|\theta)}{g(x_i|\theta)} - \left\{\frac{\frac{\partial}{\partial\theta} g(x_i|\theta)}{g(x_i|\theta)}\right\}^2 = \frac{\frac{\partial^2}{\partial\theta^2} g(x_i|\theta)}{g(x_i|\theta)} - s_\theta^2(X_i)$$

and hence

$$-\mathrm{E}_{[X_1|\theta]}\frac{\partial^2}{\partial\theta^2}\log g(X_1|\theta) = \mathrm{E}_{[X_1|\theta]} s_\theta^2(X_1) - \int \frac{\partial^2}{\partial\theta^2} g(x_1|\theta) dx_1 = I_1^F(\theta) - 0,$$

again, by interchanging differentiation and integration. This identity gives the following approximation via SLLN when $X_i \overset{\text{IID}}{\sim} g(x_i|\theta)$,

$$-\frac{1}{n}\ddot{\ell}_X(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} g(X_i|\theta) \approx -\mathrm{E}_{[X_1|\theta]}\frac{\partial^2}{\partial\theta^2}\log g(X_1|\theta) = I_1^F(\theta).$$

Now under some regularity conditions on the pdfs (pmfs) $g(x_i|\theta)$, for large n, $\hat{\theta}_{\text{MLE}}(X) \approx \theta_0$ when $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$, which implies

$$\frac{I_X}{n} = -\frac{1}{n}\ddot{\ell}_X(\hat{\theta}_{\text{MLE}}(X)) \approx -\frac{1}{n}\ddot{\ell}_X(\theta_0) \approx I_1^F(\theta_0) = \mathrm{Var}_{[X_1|\theta_0]} s_{\theta_0}^2(X_1) = b^2.$$

This completes the argument for an approximate $\mathsf{Normal}(0,1)$ distribution of $\dot{\ell}_X(\theta_0)/\sqrt{I_X}$. The property $\hat{\theta}_{\text{MLE}}(X) \approx \theta_0$ also implies $R(X) \approx 0$. This completes our "proof"!

The regularity conditions. The "regularity conditions" needed on the pdfs/pmfs are essentially differentiability conditions (as functions of θ). In particular, it suffices that for any x_i, the map $\theta \mapsto \log g(x_i|\theta)$ is three times differentiable and that there is a function $h(x_i)$ such that $|\frac{\partial^3}{\partial \theta^3} \log g(x_i|\theta)| < h(x_i)$ for all θ and $\mathrm{E}_{[X_1|\theta_0]} h(X_1) < \infty$. We also need that $\hat{\theta}_{\text{MLE}}(x)$ is the unique maxima of $\ell_x(\theta)$ for all x. These are known as the classic conditions (due to Crámer). Better conditions were later provided by Le Cam who requires existence of a single derivative in "quadratic mean".

Confidence coefficient of ML intervals

Now consider an ML interval $B_c(x) = \hat{\theta}_{\text{MLE}}(x) \mp c/\sqrt{I_x}$. The coverage probability at any θ_0 inside Θ is:

$$\begin{aligned}
\gamma(B_c; \theta_0) &= P_{[X|\theta_0]}(\theta_0 \in \hat{\theta}_{\text{MLE}}(X) \mp c/\sqrt{I_X}) \\
&= P_{[X|\theta_0]}\left(-c \leq \sqrt{I_X}(\hat{\theta}_{\text{MLE}}(X) - \theta_0) \leq c\right) \\
&\approx 2\Phi(c) - 1,
\end{aligned}$$

by asymptotic normality of MLE. Therefore, the confidence coefficient of B_c is approximately $2\Phi(c) - 1$. And hence an approximately $100(1-\alpha)\%$-CI intervals is given by $B_{z(\alpha)}(x) = \hat{\theta}_{\text{MLE}}(x) \mp z(\alpha)/\sqrt{I_x}$ where, as before, $z(\alpha) = \Phi^{-1}(1 - \alpha/2)$.