Lecture 10 (or whatever): Fisher Information and the MLE

- We'll go over the univariate case of Fisher Information Today.
- we start with a 1-parameter family of densities

$$\mathcal{F} = \{ f_\theta(x), \underbrace{\theta \in \Omega}_{\text{Recall: Parameter Space}}, \underbrace{x \in \mathcal{X}}_{\text{Sample Space}} \}$$

We'll consider just the continuous case for simplicity.

- Recall that we defined the log-likelihood function as;

$$\ell_x(\theta) = \log f_\theta(x)$$

- we'll call its derivative wrt θ the score function

$$\underbrace{\dot{\ell}_x(\theta)}_{\text{How higher / lower the likelihood value of our sample gets as the true } \theta \text{ varies.}} = \frac{\partial}{\partial \theta} \log f_\theta(x) = \frac{\dot{f}_\theta(x)}{f_\theta(x)}$$

Recall (chain rule)

$$\frac{\partial}{\partial x} \log(f(x)) = \frac{f'(x)}{f(x)}$$

- First, let's compute the expectation of the score function.

Recall $E(x) = \int_x x \underbrace{f(x)}_{\text{density.}} dx$

$$E[\dot{\ell}_x(\theta)] = \int_x \dot{\ell}_x(\theta) \cdot f_\theta(x)\, dx = \int_x \dot{f}_\theta(x)\, dx = \int_x \frac{\partial}{\partial \theta} f_\theta(x)\, dx$$

- Assuming $f_\theta(x)$ continuous & continuously differentiable

$$\int_{\mathcal{X}} \frac{\partial}{\partial\theta} f_\theta(x)\,dx = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f_\theta(x)\,dx = \frac{\partial}{\partial\theta}\cdot 1 = 0$$

First result today $\boxed{E[\dot{\ell}_x(\theta)] = 0}$

- OK, got the expectation. Now, let's talk about the variance of $\dot{\ell}_x(\theta)$.

Recall $V(X) = \int_{\mathcal{X}} [x - E(x)]^2 \cdot f(x)\,dx$

Since $E[\dot{\ell}_x(\theta)] = 0$

$$\underbrace{I_\theta \doteq V(\dot{\ell}_x(\theta))} = \int_{\mathcal{X}} [\dot{\ell}_x(\theta)]^2 \cdot f_\theta(x)\cdot dx$$

Define the Fisher Information as the variance of the score function.

Now, we'll show that:

Thm. $\hat{\theta}^{MLE} \overset{approx}{\sim} N(\theta, 1/I_\theta)$

(Sketch of a) Proof: Let $\ddot{\ell}_x(\theta) = \frac{\partial^2}{\partial\theta^2} \log f_\theta(x)$

Recall: Quotient differentiation rule $\frac{\partial}{\partial x}\frac{g(x)}{h(x)} = \frac{g'(x)\cdot h(x) - h'(x)\cdot g(x)}{[h(x)]^2}$

$$= \frac{\partial}{\partial\theta}\left[\frac{\dot{f}_\theta(x)}{f_\theta(x)}\right]$$

$$= \frac{\ddot{f}_\theta(x)\cdot f_\theta(x) - \dot{f}_\theta(x)\cdot \dot{f}_\theta(x)}{[f_\theta(x)]^2} = \frac{\ddot{f}_\theta(x)}{f_\theta(x)} - \left(\frac{\dot{f}_\theta(x)}{f_\theta(x)}\right)^2$$

Now, $\ddot{\ell}_x(\theta)$ has expectation

$$E[\ddot{\ell}_x(\theta)] = \int_{\mathcal{X}} \frac{\ddot{f}_\theta(x)}{f_\theta(x)}\cdot f_\theta(x)\,dx - \int_{\mathcal{X}} \left(\frac{\dot{f}_\theta(x)}{f_\theta(x)}\right)^2 f_\theta(x)\,dx$$

$$= \int \ddot{f}_\theta(x)\,dx - \int_{\mathcal{X}} [\dot{\ell}_x(\theta)]^2 \cdot f_\theta(x)\,dx$$

$$\overbrace{x\sim}\quad\overbrace{\mathcal{X}}$$

$$= \frac{\partial^2}{\partial\theta^2}\int_{\mathcal{X}} f_\theta(x)\,dx - I_\theta$$

Definition of Fisher Information

$$= \cancel{\frac{\partial^2}{\partial\theta^2} g(x)}^{\,0} - I_\theta$$

- The expectation of the second derivative of the log-likelihood function is equal to the negative of the Fisher information.

Done with the definitions. Now suppose $\underline{X} = (x_1, \ldots, x_n)$ is sample from $f_\theta(x)$. Then

$$\dot{\ell}_{\underline{x}}(\theta) = \sum_{i=1}^{n} \dot{\ell}_{x_i}(\theta)$$

Because $\dot{\ell}_{\underline{x}}(\theta)$ is in log-space

Further, $-\ddot{\ell}_{\underline{x}}(\theta) = -\sum_{i=1}^{n} \ddot{\ell}_{x_i}(\theta)$

- Since $\hat{\theta}^{MLE}$ for the full sample $\underline{x}$ satisfies maximizing condition $\dot{\ell}_{\underline{x}}(\hat{\theta}^{MLE})$. Then, we can get Taylor series approximation...

Recall this is just a local linearization

$$0 = \dot{\ell}_{\underline{x}}(\hat{\theta}^{MLE}) \approx \dot{\ell}_{\underline{x}}(\theta) + \ddot{\ell}_{\underline{x}}(\theta)\cdot(\hat{\theta}^{MLE} - \theta)$$

$$\Rightarrow \theta\,\ddot{\ell}_{\underline{x}}(\theta) \simeq \dot{\ell}_{\underline{x}}(\theta) + \ddot{\ell}_{\underline{x}}(\theta)\cdot\hat{\theta}^{MLE}$$

$$\Rightarrow \hat{\theta}^{MLE} \simeq \frac{\theta\cdot\ddot{\ell}_{x}(\theta) - \dot{\ell}_{\underline{x}}(\theta)}{\ddot{\ell}_{\underline{x}}(\theta)}$$

$$\Rightarrow \hat{\theta}^{MLE} \simeq \theta - \frac{\dot{\ell}_{\underline{x}}(\theta)}{\ldots}$$

$\ell_{\underline{x}}(\theta)$

Finally, recall $\dot{\ell}_{\underline{x}}(\theta) = \sum \dot{\ell}_{x_i}(\theta)$, so it's a sum of a function of random variables.

While we like sums of random variables, we love means of random variables (why?)

Central Limit Theorem (ghetto version): If you take the mean of a large # of random variables, it will approximately follow a Normal distr.

By CLT... $$\frac{\dot{\ell}_{\underline{x}}(\theta)}{n} \sim N(0, \underbrace{I_\theta/n}_{\text{variance}})$$

Then, it is convenient to write

$$\hat{\theta}^{MLE} \approx \theta - \frac{\dot{\ell}_{\underline{x}}(\theta)/n}{\ddot{\ell}_{\underline{x}}(\theta)/n} \sim N\left(0, \frac{1}{nI_\theta}\right)$$

As $n \to \infty$, we know the numerator $\sim N(0, \frac{I_\theta}{n})$

Further $$\frac{\ddot{\ell}_{\underline{x}}(\theta)}{n} \to \frac{E_\theta[\ddot{\ell}_{\underline{x}}(\theta)]}{n} = -I_\theta$$

. Using the denominator as a constant and recalling $V(c \cdot X) = c^2 \cdot V(X)$

we get that $$\frac{\dot{\ell}_{x}(\theta)/n}{\ell_{\underline{x}}(\theta)/n} \sim N\left(0, \frac{I_\theta/n}{I_\theta^2}\right)$$

$$\sim N\left(0, \frac{1}{n \cdot I_\theta}\right)$$

Thus, $$\hat{\theta}^{MLE} \overset{Approx}{\sim} N\left(\theta, \frac{1}{nI_\theta}\right)$$

This is the Fisher Information

for n iid observations.

$$E\left(\theta - \frac{\dot{\ell}_{\underline{x}}(\theta)}{\ddot{\ell}_{\underline{x}}(\theta)}\right) = E(\theta) - E\left(\frac{\dot{\ell}_{\underline{x}}(\theta)}{\ddot{\ell}_{\underline{x}}(\theta)}\right)$$

$$= \theta - 0 = \theta$$

Equivalent to writing

$$\boxed{\hat{\theta}^{MLE} \sim N\left(\theta, \frac{1}{I_\theta(n)}\right)}$$