Problem set 2

Notation y_i $i = 1, \ldots, N$

(scalar)

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{pmatrix}$$

p.d.f. $f(y_i | x_i, \theta)$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{pmatrix}$$

$$L(\theta) = \prod_{i=1}^{N} f(y_i | x_i, \theta)$$

likelihood

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ln f(y_i | x_i, \theta)$$

log likelihood

$$f_i(\theta) \equiv f(y_i | x_i, \theta)$$

$$l_i(\theta) \equiv \ln f_i(\theta)$$

Then $\mathcal{L}(\theta) = \sum_{i=1}^{N} l_i(\theta)$

$$\mathcal{L}(\theta) \longrightarrow \max_{\theta \in \Theta}$$

FOC

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1} \stackrel{!}{=} 0$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_2} \stackrel{!}{=} 0$$

$$\vdots$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_k} \stackrel{!}{=} 0$$

del

$$\nabla_\theta \mathcal{L}(\theta) = \left[\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1} \cdots \frac{\partial \mathcal{L}(\theta)}{\partial \theta_k} \right]$$

$$= \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$$

$$\nabla_\theta \left(\sum l_i \right) = \sum \underbrace{\nabla_\theta l_i(\theta)}_{\text{score for ith observation}}$$

$$s(y_i|x_i,\theta) = s_i(\theta) = \nabla_\theta \ell_i(\theta)^{(1)} \quad (1)\ \text{transpose}$$

$k \times 1$

$$= \nabla_\theta \ln f_i(\theta)' = \frac{1}{f_i(\theta)} \nabla f_i(\theta)' \stackrel{!}{=} 0$$

$$= \frac{1}{f_i(\theta)} \begin{bmatrix} \partial f_i(\theta)/\partial\theta_1 \\ \vdots \\ \partial f_i(\theta)/\partial\theta_k \end{bmatrix}$$

Why ∇f is a row?

$$f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0)$$

Taylor expansion

$$E[s_i(\theta)|x_i] = E\left(\frac{1}{f_i(\theta)} \nabla_\theta f_i(\theta)\right)$$

$$= \int_{-\infty}^{+\infty} \frac{1}{f_i(\theta)} \nabla_\theta f_i(\theta) \cdot f_i(\theta_0) dy_i$$

Note: $f_i(\theta_0)$ is the true density function. We take expectation wrt to true density

So for $s_i(\theta_0)$ we have

$$E(s_i(\theta_0)|x_i) = \int_{-\infty}^{\infty} \frac{1}{f_i(\theta_0)} \nabla_\theta f_i(\theta_0) f_i(\theta_0) dy$$

$$= \int_{-\infty}^{\infty} \nabla_\theta f_i(\theta_0) dy_i$$

For discrete y_i

Example: $y_i = \begin{cases} 1 & \text{w.p. } p(x_i, \theta_0) \\ 0 & \text{w.p. } 1 - p(x_i, \theta_0) \end{cases}$

Then

$$E(s_i(\theta_0)|x_i) = \frac{1}{p(x_i,\theta_0)} \nabla_\theta p(x_i,\theta_0) \cdot p(x_i,\theta_0) + \frac{1}{1-p(x_i,\theta_0)} \left(-\nabla_\theta p(x_i,\theta_0)\right) \times \left(1 - p(x_i,\theta_0)\right)$$

$$E(s_i(\theta_0)|x_i) = \int_{-\infty}^{+\infty} \nabla_\theta f_i(\theta_0)\, dy$$

OR

$$E(s_i(\theta_0)|x_i) = \sum_j \nabla_\theta p_j(\theta_0) = \sum_j \nabla_\theta p_j(y_i|x_i,\theta_0)$$

j indexes possible values of y_i

Notice that $\int_{-\infty}^{\infty} f_i(\theta)\, dy = 1$

$\sum_j p_j(\theta) = 1$ for any θ

$$\nabla_\theta \int f_i(\theta)\, dy = \nabla_\theta(1) = 0$$

If $f_i(\cdot)$ is "good", can bring ∇ under $\int$

$$\int_{-\infty}^{\infty} \nabla_\theta f_i(\theta_0)\, dy = 0 = E(s_i(\theta_0)|x_i)$$

$E(s_i(\theta_0)|x_i)$ does not depend on x_i

$$\Rightarrow \boxed{E(s_i(\theta_0)) = 0} \quad (1)$$

Second derivative

$$\nabla_\theta^2 \mathcal{L}(\theta) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_1^2} & \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_1 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_k^2} \end{pmatrix}$$

$$= \nabla_\theta \left(\nabla_\theta \mathcal{L}(\theta)'\right) = \nabla_\theta \left(\sum s_i(\theta)\right)$$

$$= \sum_{i=1}^{N} \underbrace{\nabla s_i(\theta)}_{H_i(\theta)}$$

$H_i(\theta)$ ← Hessian matrix for i^{th} obs.

$$H_i(\theta) = \nabla_{\theta'}\left[\frac{1}{f_i(\theta)}\nabla_\theta f_i(\theta)\right]$$

$\frac{1}{f_i(\theta)}$: scalar; $\nabla_\theta f_i(\theta)$: $K \times 1$

$$= -\frac{1}{f_i(\theta)^2}\nabla_\theta f_i(\theta)'\nabla_\theta f_i(\theta) + \frac{1}{f_i(\theta)}\nabla_\theta^2 f_i(\theta)$$

$-\frac{1}{f_i(\theta)^2}\nabla_\theta f_i(\theta)'$: $\nabla_{\theta'}\left(\frac{1}{f_i(\theta)}\right)$

$\nabla_\theta^2 f_i(\theta)$: $\nabla_{\theta'}(\nabla_\theta f_i(\theta))$

$= s_i(\theta)s_i(\theta)'$ since $s_i(\theta) = \frac{1}{f_i(\theta)}\nabla_\theta f_i(\theta)$

The results here are correct but the derivation is a bit sloppy. See the added sheet

$$H_i(\theta) = -s_i(\theta)s_i(\theta)' + \frac{1}{f_i(\theta)}\nabla_\theta^2 f_i(\theta)$$

$$E(s_i(\theta_0)s_i(\theta_0)'|x_i) = V(s_i(\theta_0)|x_i)$$ variance-covariance matrix of s_i

$$= -E(H_i(\theta_0)|x_i) + E\left[\frac{1}{f_i(\theta_0)}\nabla_\theta^2 f_i(\theta_0)|x_i\right]$$

$E\left[\frac{1}{f_i(\theta_0)}\nabla_\theta^2 f_i(\theta_0)|x_i\right] = 0$ → same trick as for $E\left[\frac{1}{f_i(\theta)}\nabla f_i(\cdot)|x_i\right] = 0$

Additional sheet

This is the rigorous and more detailed derivation of the stuff in the purple box above.

$$H_i(\theta) = \nabla_\theta \left[\frac{1}{f_i(\theta)} \nabla_\theta f_i(\theta) \right]' = \nabla_\theta \left[\nabla_\theta f_i(\theta)' \underbrace{\frac{1}{f_i(\theta)}}_{\text{scalar}} \right]$$

$$[(AB)' = B'A']$$

$$= \nabla_\theta^2 f_i(\theta) \frac{1}{f_i(\theta)} + \nabla_\theta f_i(\theta)' \nabla_\theta \left[\frac{1}{f_i(\theta)} \right]$$

$$= \nabla_\theta^2 f_i(\theta) \frac{1}{f_i(\theta)} + \nabla_\theta f_i(\theta)' \left(-\frac{1}{f_i(\theta)^2} \nabla_\theta f_i(\theta) \right)$$

chain rule

$$= \frac{1}{f_i(\theta)} \nabla_\theta^2 f_i(\theta) - \frac{1}{f_i(\theta)^2} \nabla_\theta f_i(\theta)' \nabla_\theta f_i(\theta)$$

$$= \frac{1}{f_i(\theta)} \nabla_\theta^2 f_i(\theta) - s_i(\theta) s_i(\theta)'$$

Note that the order of multiplication matters!

$$V(S_i(\theta_0)|X_i) = E(S_i(\theta_0) S_i(\theta_0)' | X_i)$$
$$= -E(H_i(\theta_0)|X_i)$$

$$E(S_i(\theta_0) S_i(\theta_0)') = -E(H_i(\theta_0)) \quad (2)$$

1(B) Why the log?

→ computational convenience

→ log ⇒ sums ⇒ can apply LLN CLT

→ It can be shown that

θ_0 maximizes $E(\mathcal{L}(\theta))$

Slide 55

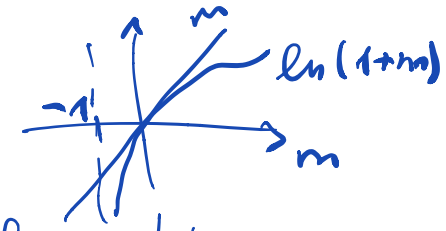
$\forall \theta \in \Theta$, $E[\ln f(y_i|x_i, \theta_0)]$

for all

$$\geq E[\ln f(y_i|x_i, \theta)]$$

To show that, we are going to use

- $\ln(1+m) \leq m$



- $f(z), g(z)$ density functions

$S_f = \{z \mid f(z) > 0\}$

$$\int_{S_f} f(z)\,dz \geq \int_{S_f} g(z)\,dz$$

$$\mathcal{I}(f,g) = \int_{S_f} \ln \frac{f(z)}{g(z)} f(z)\,dz \geq 0$$

$$= -\int_{S_f} \ln \frac{g(z)}{f(z)} f(z)\,dz$$

$$\geq -\int_{S_f} \left(\ln \frac{g(z)}{f(z)} f(z) + f(z) - g(z) \right) dz$$

$$= -\int \left[\ln\left(\frac{g}{f}\right) - \underbrace{\left(\frac{g}{f} - 1\right)}_{m} \right] dz$$

$$= -\int [\ln(1+m) - m]\, dz \geq 0$$

$J(f, g)$ is called the Kullback-Leibler distance / information criterion

$$J(f,g) = \int_{S_f} \ln \frac{f(z)}{g(z)} f(z)\, dz$$

$$= \int_{S_f} \ln f(z) \cdot f(z)\, dz - \int_{S_f} \ln g(z) \cdot f(z)\, dz$$

$$= E_f \ln f(z) - E_f(\ln g(z)) \geq 0$$

Substitute $f_i(\theta_0)$ for f

$f_i(\theta)$ for g

$$\Rightarrow E[\ln f_i(y_i | x_i, \underline{\theta_0})] \geq E[\ln f_i(y_i | x_i, \underline{\theta})]$$

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmax}} \; E[\ln f_i(\theta)]$$

1c Asymptotic distribution of

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; \mathcal{L}(\theta)$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} l_i(\theta) = \sum_{i=1}^{N} \ln f_i(\theta)$$

If we have a sample of size N, maximization does not depend on N.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; \frac{1}{N} \sum_{i=1}^{N} l_i(\theta)$$

Our observations are i.i.d.

$$\forall \theta \quad \frac{1}{N} \sum_{i=1}^{N} l_i(\theta) \xrightarrow{P} E[l_i(\theta)] = E(\ln f_i(\theta))$$

pointwise convergence

It would seem logical

$$\operatorname{argmax}_\theta \frac{1}{n}\sum_i \ell_i(\theta) \xrightarrow{P} \operatorname{argmax}_\theta E(\ell_i(\theta))$$

Unfortunately, pointwise convergence is not enough.

Recall

$X_n \xrightarrow{P} a$ if

$$\forall \varepsilon > 0,\ P(|X_n - a| > \varepsilon) \to 0$$

Example

$$f_n(x) = n x^n (1-x),\ x \in (0,1)$$

$$\lim_{n\to\infty} \underbrace{n}_{\to\infty} \underbrace{x^n}_{\to 0}(1-x) = [\infty \cdot 0]$$

$$= \lim_{n\to\infty} \frac{n(1-x)}{\left(\frac{1}{x}\right)^n} = \left[\frac{\infty}{\infty}\right]$$

l'Hôpital's rule $\lim_{x\to\infty} \frac{h(x)\to\infty}{m(x)\to\infty}$

$$= \lim_{x\to\infty} \frac{h'(x)}{m'(x)}$$

Analogously, for sequences

$$\lim_{n\to\infty} \frac{h(n)}{m(n)} = \lim_{n\to\infty} \frac{h(n+1)-h(n)}{m(n+1)-m(n)}$$

Numerator: $n(1-x)$

$$[(n+1)-n](1-x) = 1-x$$

Denominator: $\left(\frac{1}{x}\right)^n$

$$\left(\frac{1}{x}\right)^{n+1} - \left(\frac{1}{x}\right)^n = \left(\frac{1}{x}\right)^n \left(\frac{1-x}{x}\right)$$

$$\lim_{n\to\infty} \frac{n(1-x)}{\left(\frac{1}{x}\right)^n} = \lim_{n\to\infty} \frac{1-x}{\left(\frac{1}{x}\right)^n \frac{1-x}{x}} =$$

$$= \lim_{n\to\infty} \frac{x}{(1/x)^n} = 0$$

$$f_n(x) = n x^n (1-x)$$

$$\frac{d f_n(x)}{dx} = n^2 x^{n-1}(1-x) - n x^n \stackrel{!}{=} 0$$

$$x^* = \frac{n}{n+1} \longrightarrow 1$$

What we need is uniform convergence:

$$\max_{\theta \in \Theta} \left| N^{-1} \sum l_i(\theta) - E\, l_i(\theta) \right| \xrightarrow{p} 0$$

→ Θ is closed and bounded

→ $l_i(\theta)$ continuous for all values of y and x

→ $|l_i(\theta)| < b(y, x_i)$, $b(\cdot) \geq 0$

$E(b(\cdot)) < \infty$

Uniform convergence +

identification assumption

$$\theta_0 = \arg\max_{\theta \in \Theta} E(\ln f_i(\theta))$$

θ_0 is unique

$$\Rightarrow \quad \hat{\theta} \xrightarrow{p} \theta_0$$

$\hat{\theta}$ is consistent

Asymptotic normality

Multivariate normal distr.

p.d.f.

$$g(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)\right)}{(\sqrt{2\pi})^n \sqrt{\det \Sigma^{-1}}}$$

μ is the mean

Σ is the V-Cov matrix

$$\mathcal{N}(\mu, \Sigma)$$

$$E(s_i(\theta_0)) = 0$$

$$E(s_i(\theta_0)\, s_i(\theta_0)') = -E(H_i(\theta_0))$$

Taylor expansion of $s(\cdot)$ around θ_0

$$\underbrace{\sum_{i=1}^{n} s_i(\hat{\theta})}_{=0 \text{ (FOC)}} = \sum s_i(\theta_0) + \sum H_i(\theta_0)(\hat{\theta} - \theta_0) + R(\hat{\theta})$$

$R(\hat{\theta})$ converges to 0 faster than $\hat{\theta} \to \theta_0$