STAT 3100 Lecture Note

Week Four (Sep 27 & 29, 2022)

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2022/Sep/25 03:40:31

Estimating equations

Parametric models

- A parametric model is a set of distributions indexed by unknown $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ with small or moderate p
  - Say $\{f(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$, where f is either a pdf or a pmf and $\boldsymbol{\Theta}$ is the set of all the possbile values of $\boldsymbol{\theta}$
- Believed that the true parameter (vector) $\boldsymbol{\theta}_0$ ($\in \boldsymbol{\Theta} \subset \mathbb{R}^p$) is fixed
  - Rather than making $\boldsymbol{\theta}_0$ random in the Bayesian philosophy

Method of moments (MOM, CB Sec 7.2.1)

- Procedure
  1. Equate raw moments to their empirical counterparts.
  2. Solve the resulting simultaneous equations for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.
- Features
  - Easy implementation
  - Start point for more complex methods
  - No constraint
  - Not uniquely defined
  - No guarantee on optimality

Exercise Lec7.1

- Let $X_1, \ldots, X_n$ iid follow the following distributions. Find MOM estimators for (θ_1, θ_2).
  a. $N(\theta_1, \theta_2)$, $(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}^+$.
  b. Binom(θ_1, θ_2) with pmf

$$p_X(x \mid \theta_1, \theta_2) = \binom{\theta_1}{x} \theta_2^x (1-\theta_2)^{\theta_1 - x} \mathbf{1}_{\{0,\ldots,\theta_1\}}(x), \quad (\theta_1, \theta_2) \in \mathbb{Z}^+ \times (0,1).$$

Exercise Lec7.2

- Let $X_1, \ldots, X_n$ iid follow pdf $f(x \mid \theta) = \theta x^{\theta-1} \mathbf{1}_{[0,1]}(x)$, $\theta > 0$.
  a. Find an MOM estimator of θ.
  b. Can we employ the second (raw) moment instead of the first one?

Maximum Likelihood Estimator (MLE, CB Sec 7.2.2)

- Likelihood function: $L : \boldsymbol{\Theta} \to \mathbb{R}$ such that, given $\boldsymbol{x}$ (a realization of $\mathbf{X}$),

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = f_{\mathbf{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}),$$

  where $f_{\mathbf{X}}$ is the joint pdf or pmf.

- For each $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be the maximizer of $L(\boldsymbol{\theta}; \boldsymbol{x})$ (or log-likelihood $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \ln L(\boldsymbol{\theta}; \boldsymbol{x})$) with respect to $\boldsymbol{\theta}$ constrained in $\boldsymbol{\Theta}$, i.e.,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}; \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}; \boldsymbol{x}).$$

  Then the statistic $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$ is the MLE for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

- Invariance property of MLE (CB Thm 7.2.10): As long as $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, for ANY function g, the $g(\hat{\boldsymbol{\theta}})$ is teh MLE of $g(\boldsymbol{\theta})$.
- If $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ is differetiable, the score funtion $\mathbf{S}$ is defined as its gradient

$$\mathbf{S}(\boldsymbol{\theta}; \boldsymbol{x}) = \left[\frac{\partial}{\partial \theta_1} \ell(\boldsymbol{\theta}; \boldsymbol{x}), \ldots, \frac{\partial}{\partial \theta_p} \ell(\boldsymbol{\theta}; \boldsymbol{x}) \right]^\top.$$

- If $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ is twice differetiable, we have hessian of $\ell(\boldsymbol{\theta}; \boldsymbol{x})$

$$\mathbf{H}(\boldsymbol{\theta}; \boldsymbol{x}) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\boldsymbol{\theta}; \boldsymbol{x}) \right]_{p \times p}.$$

- Procedure
  - A direct maximization if ℓ or L is monotonic OR
  - Solving simultaneous equations $S(\boldsymbol{\theta}; \boldsymbol{x}) = \mathbf{0}$ for $\boldsymbol{\theta}$. Specifically,
    1. Collect solutions with negative definite Hessian (indicating interior local maximizers)
    2. Compare likelihoods (or log-likelihoods) corresponding to all candidates (consisting of previously picked solutions plus boundary values of $\boldsymbol{\Theta}$)
    3. May involve discussions on different values of $\boldsymbol{x}$

Exercise Lec7.3

- Suppose $X_1, \ldots, X_n$ are iid as the following distributions. Find MLEs for corresponding parameters.
  a. $N(\mu, \sigma^2)$, $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$.
  b. Bernoulli(θ): $p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \mathbf{1}_{\{0,1\}}(x)$, $\theta \in [0, 1/2]$.
  c. Two-parameter exponential: $f(x \mid \alpha, \beta) = \beta^{-1} \exp\{-(x-\alpha)/\beta\} \mathbf{1}_{(\alpha,\infty)}(x)$, $(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^+$.

Other examples of estimating equations

- Least-squares estimator
- Generalized estimating equations (GEE)
- M-estimator

Evaluating estimators

Mean squared error (MSE)

- Univariate: $\mathrm{E}(\hat{\theta} - \theta_0)^2 = \{\mathrm{E}(\hat{\theta}) - \theta_0\}^2 + \mathrm{var}(\hat{\theta}_0)$

- Multivariate: $\mathrm{E}\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} = \{\mathrm{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0\}^\top \{\mathrm{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0\} + \mathrm{cov}(\hat{\boldsymbol{\theta}})$
- Best unbiased estimator (a.k.a. (uniform) minimun variance unbiased estimator, abbr. UMVUE/MVUE): if $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$ satisfies that
  - $\hat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$, i.e., $\mathrm{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$;
  - $\mathrm{var}(\hat{\boldsymbol{\theta}}) \leq \mathrm{var}(\hat{\boldsymbol{\theta}}^*)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $\hat{\boldsymbol{\theta}}^*$ such that $\mathrm{E}(\hat{\boldsymbol{\theta}}^*) = \boldsymbol{\theta}$.
- UMVUE is unique (CB Thm 7.3.19)

Cramer-Rao lower bound (CB Thm 7.3.9 & Lemma 7.3.11)

- Only consider the univariate case, i.e., one-dimensional unknown parameter θ
- Fisher information: $I(\theta) = \mathrm{var}(S(\theta; \mathbf{X})) = \mathrm{E}[\{S(\theta; \mathbf{X})\}^2] = -\mathrm{E}[\{H(\theta; \mathbf{X})\}^2]$
  - score function $S(\theta; \mathbf{X})$ and Hessian $H(\theta; \mathbf{X})$ both scalar
- Cramer-Rao lower bound: $\mathrm{var}(\hat{\theta}) \geq \{(\mathrm{d}/\mathrm{d}\theta)\mathrm{E}(\hat{\theta})\}^2 / I(\theta)$ for $\hat{\boldsymbol{\theta}}$ satisfying regularity conditions
  - Proof: Cauchy-Schwarz inequality (CB Thm 4.7.3) $\Rightarrow$ covariance inequality (CB Example 4.7.4)
- (CB Coro 7.3.15) $\hat{\theta}$ attains the lower bound $\Leftrightarrow \exists a(\theta)$ s.t. $S(\theta; \mathbf{X}) = a(\theta)\{\hat{\theta} - \mathrm{E}(\hat{\theta})\}$
- The unbiased $\hat{\theta}$ attaining the lower bound is UMVUE.

Example Lec8.1

- Find the lower bound for unbiased estimators for σ^2 in the following cases.
  a. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with known μ and unknown σ^2.
  b. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with unknown (μ, σ^2).

Sufficiency (CB Sec 6.2.1)

- A statistic $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is sufficient for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \Leftrightarrow$ the distribution of $\mathbf{X}$ conditioning on $\mathbf{T}$ and $\boldsymbol{\theta}$, say $f_{\mathbf{X}|\mathbf{T},\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{t}, \boldsymbol{\theta})$, is free of $\boldsymbol{\theta}$.
- Fisher-Neyman factorization theorem (CB Thm 6.2.6; HMC Thm 7.2.1): $\mathbf{T}$ is sufficient for $\boldsymbol{\theta} \Leftrightarrow$ the likelihood function can be factored into two parts, one of them not depending on $\boldsymbol{\theta}$, i.e.,

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = h(\boldsymbol{x}) g(\mathbf{T}(\boldsymbol{x}), \boldsymbol{\theta}),$$

  for all the possible values of $\boldsymbol{x}$ and $\boldsymbol{\theta}$.
- (HMC Thm 7.3.2) If $\mathbf{T}$ is sufficient for θ and $\hat{\theta}$ is the unique MLE of θ, then $\hat{\theta}$ must be a function of $\mathbf{T}$.

- Nonuniqueness
  - Trivial examples
    * $\mathbf{X}$ is always sufficient.
    * $(X_{(1)}, \ldots, X_{(n)})$ is always sufficient if X_i's are iid, with $X_{(1)} \leq \cdots \leq X_{(n)}$.
  - $\mathbf{T}$ is sufficient and $g(\cdot)$ is a one-to-one mapping $\Rightarrow g(\mathbf{T})$ is also sufficient.
- Minimal sufficiency: a sufficient statistc that is a function of all the other sufficient statistics.
  - How to find a minimal sufficient sufficient statistic (CB Thm 6.2.13):
    1. Find the sufficient and necessary condition for $L(\boldsymbol{\theta}; \boldsymbol{x})/L(\boldsymbol{\theta}; \boldsymbol{y})$ to be free of $\boldsymbol{\theta}$;
    2. If the condition is of the form $\mathbf{T}(\boldsymbol{x}) = \mathbf{T}(\boldsymbol{y})$, then $\mathbf{T}(\mathbf{X})$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.

Example Lec8.2

- Find the minimal sufficient statistics in the following scenarios.
    a. $X_1, \dots, X_n \sim \text{Unif}(1, \dots, \theta)$ with unknown positive integer θ.
    b. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown μ and σ^2.