POINT ESTIMATION AND EFFICIENCY

- Introduction

Goal of statistical inference: estimate and infer quantities of interest using experimental or observational data

- a class of statistical models used to model data generation process (statistical modeling)
- the "best" method used to derive estimation and inference (statistical inference: point estimation and hypothesis testing)
- validation of models (model selection)

- What about estimation?
  - One good estimation approach should be able to estimate model parameters with reasonable accuracy
  - should be somewhat robust to intrinsic random mechanism
  - an ideally best estimator should have no bias and have the smallest variance in any finite sample
  - alternatively, one looks for an estimator which has no bias and has the smallest variance in large sample

Probabilistic Models

A *model* $\mathcal{P}$ is a collection of probability distributions describing data generation.

Parameters of interest are simply some functionals on $\mathcal{P}$, denoted by $\nu(P)$ for $P \in \mathcal{P}$.

- Examples
  - a non-negative r.v. X (survival time, size of growing cell etc.)

    Case A. Models: $X \sim$ Exponential$(\theta), \theta > 0$

    $\mathcal{P} = \left\{ p_\theta(x) : p_\theta(x) = \theta e^{-\theta x} I(x \geq 0), \theta > 0 \right\}$ $\mathcal{P}$ is a parametric model. $\nu(p_\theta) = \theta$.

    Case B. $\mathcal{P} = \{p_{\lambda,G} : p_{\lambda,G} = \int_0^\infty \lambda \exp\{-\lambda x\} dG(\lambda)$, $\lambda \in R$, G is any distribution function$\}$. $\mathcal{P}$ is a semiparametric model. $\nu(p_{\lambda,G}) = \lambda$ or G.

    Case C. $\mathcal{P}$ consists of all distribution function in $[0, \infty)$. $\mathcal{P}$ is a nonparametric model.

    $\nu(P) = \int x dP(x)$.

– Suppose that $X = (Y, Z)$ is a random vector on $R^+ \times R^d$ (Y survival time, Z a number of covariates)
<u>*Case A*</u>. $Y|Z = z \sim \text{Exponential}(\lambda e^{\theta' z})$ A parametric model with parameter space $\Theta = R^+ \times R^d$.
<u>*Case B*</u>. $Y|Z = z \sim \lambda(y)e^{\theta' z} \exp\{-\Lambda(y)e^{\theta' z}\}$ where $\Lambda(y) = \int_0^y \lambda(y)dy$ and is unknown. A semiparametric model, the Cox proportional hazards model for survival analysis, with parameter space $(\theta, \lambda) \in R \times \{\lambda(y) : \lambda(y) \geq 0, \int_0^\infty \lambda(y)dy = \infty\}$.
<u>*Case C*</u>. $X \sim P$ on $R^+ \times R^d$ where P is completely arbitrary. This is a nonparametric model.

– Suppose $X = (Y, Z)$ is a random vector in $R \times R^d$ (Y response, Z covariates)

Case A.

$$Y = \theta' Z + \epsilon, \quad \theta \in R^d, \epsilon \sim N(0, \sigma^2).$$

This is a parametric model with parameter space $(\theta, \sigma) \in R^d \times R^+$.

Case B.

$$Y = \theta' Z + \epsilon, \quad \theta \in R^d, \epsilon \sim G \text{ independent of } Z.$$

This is a semiparametric model with parameters (θ, g).

Case C. Suppose $X = (Y, Z) \sim P$ where P is an arbitrary probability distribution on $R \times R^d$.

- A general rule for choosing statistical models
  - models should obey scientific rules
  - models should be flexible enough but parsimonious
  - statistical inference for models is feasible

Review of Estimation Methods

- Least Squares Estimation
  - Suppose n i.i.d observations (Y_i, Z_i), $i = 1, ..., n$, are generated from the distribution in Example 1.3.

$$\min_{\theta} \sum_{i=1}^{n} (Y_i - \theta' Z_i)^2, \quad \hat{\theta} = (\sum_{i=1}^{n} Z_i Z_i')^{-1} (\sum_{i=1}^{n} Z_i Y_i).$$

  - More generally, suppose $Y = g(X) + \epsilon$ where g is unknown. Estimating g can be done by minimizing $\sum_{i=1}^{n} (Y_i - g(X_i))^2$.
  - Problem with the latter: the minimizer is not unique and not applicable

UMVUE

- Ideal estimator
  - is unbiased, $E[T] = \theta$;
  - has the smallest variance among all the unbiased estimators;
  - is called the UMVUE estimator.
  - may not exist; but for some models from exponential family, it exists.

- Definition

Definition 4.1 Sufficiency and Completeness For θ, $T(X)$ is
a *sufficient statistic*, if $X|T(X)$ does not depend on θ;
a *minimal sufficient statistic*, if for any sufficient statistic U there exists a function H such that $T = H(U)$;
a *complete statistic*, if for any measurable function g, $E_\theta[g(T(X))] = 0$ for any θ implies $g = 0$, where E_θ denotes the expectation under the density function with parameter θ.

- Sufficiency and factorization

$T(X)$ is sufficient if and only if $p_\theta(x)$ can be factorized in to $g_\theta(T(x))h(x)$.

- Sufficiency in exponential family

Recall the canonical form of an exponential family:

$$p_\eta(x) = h(x)\exp\{\eta_1 T_1(x) + ...\eta_s T_s(x) - A(\eta)\}.$$

It is called full rank if the parameter space for $(\eta_1, ..., \eta_s)$ contains an s-dimensional rectangle.

Minimal sufficiency in exponential family

$T(X) = (T_1, ..., T_s)$ is minimally sufficient if the family is full rank.

Completeness in exponential Family If the exponential family is of full-rank, $T(X)$ is a complete statistic.

- Property of sufficiency and completeness

Rao-Blackwell Theorem Suppose $\hat{\theta}(X)$ is an unbiased estimator for θ. If $T(X)$ is a sufficient statistics of X, then $E[\hat{\theta}(X)|T(X)]$ is unbiased and moreover,

$$Var(E[\hat{\theta}(X)|T(X)]) \leq Var(\hat{\theta}(X)),$$

with the equality if and only if with probability 1, $\hat{\theta}(X) = E[\hat{\theta}(X)|T(X)]$.

Proof

$E[\hat{\theta}(X)|T]$ is clearly unbiased.

By Jensen's inequality,

$$Var(E[\hat{\theta}(X)|T]) = E[(E[\hat{\theta}(X)|T])^2] - E[\hat{\theta}(X)]^2$$

$$\leq E[\hat{\theta}(X)^2] - \theta^2 = Var(\hat{\theta}(X)).$$

The equality holds if and only if $E[\hat{\theta}(X)|T] = \hat{\theta}(X)$ with probability 1.

- Ancillary statistics

A statistic V is called *ancillary* if V's distribution does not depend on θ.

Basu's Theorem If T is a complete sufficient statistic for the family $\mathcal{P} = \{p_\theta, \theta \in \Omega\}$, then for any ancillary statistic V, V is independent of T.

Proof

For any $B \in \mathcal{B}$, let $\eta(t) = P_\theta(V \in B|T = t)$.

$\Rightarrow$ $E_\theta[\eta(T)] = P_\theta(V \in B) = c_0$ does not depend on θ.

$\Rightarrow$

$$E_\theta[\eta(T) - c_0] = 0 \Rightarrow \eta(T) = c_0.$$

$\Rightarrow$ $P_\theta(V \in B|T = t)$ is independent of t.

- UMVUE based on complete sufficient statistics

Proposition 4.1 Suppose $\hat{\theta}(X)$ is an unbiased estimator for θ; i.e., $E[\hat{\theta}(X)] = \theta$. If $T(X)$ is a sufficient statistic of X, then $E[\hat{\theta}(X)|T(X)]$ is unbiased. Moreover, for any unbiased estimator of θ, $\tilde{T}(X)$,

$$Var(E[\hat{\theta}(X)|T(X)]) \leq Var(\tilde{T}(X)),$$

with the equality if and only if with probability 1, $\tilde{T}(X) = E[\hat{\theta}(X)|T(X)]$.

Proof

For any unbiased estimator for θ, $\tilde{T}(X)$,
$\Rightarrow$ $E[\tilde{T}(X)|T(X)]$ is unbiased and

$$Var(E[\tilde{T}(X)|T(X)]) \leq Var(\tilde{T}(X)).$$

$E[E[\tilde{T}(X)|T(X)] - E[\hat{\theta}(X)|T(X)]] = 0$ and $E[\tilde{T}(X)|T(X)]$ and $E[\hat{\theta}(X)|T(X)]$ are independent of θ.

The completeness of $T(X)$ gives that

$$E[\tilde{T}(X)|T(X)] = E[\hat{\theta}(X)|T(X)].$$

$\Rightarrow$ $Var(E[\hat{\theta}(X)|T(X)]) \leq Var(\tilde{T}(X))$.
The above arguments show such a UMVUE is unique.

- Two methods in deriving UMVUE

Method 1:

- find a complete and sufficient statistics $T(X)$;
- find a function of $T(X)$, $g(T(X))$, such that $E[g(T(X))] = \theta$.

Method 2:

- find a complete and sufficient statistics $T(X)$;
- find an unbiased estimator for θ, denoted as $\tilde{T}(X)$;
- calculate $E[\tilde{T}(X)|T(X)]$.

- Example
  - $X_1, ..., X_n$ are i.i.d $\sim U(0, \theta)$. The joint density of $X_1, ..., X_n$:

$$\frac{1}{\theta^n} I(X_{(n)} < \theta) I(X_{(1)} > 0).$$

    $X_{(n)}$ is sufficient and complete (check).
  - $E[X_1] = \theta/2$. A UMVUE for $\theta/2$ is given by

$$E[X_1 | X_{(n)}] = \frac{n+1}{n} \frac{X_{(n)}}{2}.$$

– The other way is to directly find a function $g(X_{(n)}) = \theta/2$ by noting

$$E[g(X_{(n)})] = \frac{1}{\theta^n}\int_0^\theta g(x)nx^{n-1}dx = \theta/2.$$

$$\int_0^\theta g(x)x^{n-1}dx = \frac{\theta^{n+1}}{2n}.$$

$$\Rightarrow g(x) = \frac{n+1}{n}\frac{x}{2}.$$

Other Estimation Methods

- Robust estimation
  - (least absolute estimation) $Y = \theta' X + \epsilon$ where $E[\epsilon] = 0$.
    LSE is sensitive to outliers. One robust estimator is to minimize $\sum_{i=1}^n |Y_i - \theta' X_i|$.
  - A more general objective function is to minimize
    $$\sum_{i=1}^n \phi(Y_i - \theta' X_i),$$
    where $\phi(x) = |x|^k, |x| \le C$ and $\phi(x) = C^k$ when $|x| > C$ (Huber estimators).

- Estimating functions (equations)
  - The estimator solves an equation

$$\sum_{i=1}^{n} f(X_i; \theta) = 0.$$

  - $f(X; \theta)$ satisfies $E_\theta[f(X; \theta)] = 0$.
    Rationale: $n^{-1} \sum_{i=1}^{n} f(X_i; \theta) \to_{a.s.} E_\theta[f(X; \theta)]$.

- Examples
  - In a linear regression example, for any function $W(X)$, $E[XW(X)(Y-\theta' X)]=0$. Thus an estimating equation for θ can be constructed as

$$\sum_{i=1}^{n} X_i W(X_i)(Y_i-\theta' X_i)=0.$$

  - Still in the regression example but we now assume the median of ϵ is zero. It is easy to see that $E[XW(X)sign(Y-\theta' X)]=0$. Then an estimating equation for θ can be constructed as

$$\sum_{i=1}^{n} X_i W(X_i)sign(Y_i-\theta' X_i)=0.$$

- Maximum likelihood estimation (MLE)
  - MLE is the most commonly use estimator;
  - it is likelihood-based;
  - it possesses a nice asymptotic optimality.

- Example
  - Suppose $X_1, ..., X_n$ are i.i.d. observations from $\exp(\theta)$.

$$L_n(\theta) = \theta^n \exp\{-\theta(X_1 + ... + X_n)\}.$$

$\Rightarrow \hat{\theta} = \bar{X}$.

- Suppose $(Y_1, Z_1), ..., (Y_n, Z_n)$ are i.i.d with density function

$$\lambda(y)e^{\theta' z}\exp\{-\Lambda(y)e^{\theta' z}\}g(z),$$

where $g(z)$ is the known density function of $Z = z$.

$$L_n(\theta, \lambda) = \prod_{i=1}^{n}\left\{\lambda(Y_i)e^{\theta' Z_i}\exp\{-\Lambda(Y_i)e^{\theta' Z_i}\}g(Z_i)\right\}.$$

- The maximum likelihood estimators for (θ, λ) do not exist.

- One way is to let Λ be a step function with jumps at $Y_1, ..., Y_n$ and let $\lambda(Y_i)$ be the jump size, denoted as p_i. Then the likelihood function becomes

$$L_n(\theta, p_1, ..., p_n) = \prod_{i=1}^{n} \left\{ p_i e^{\theta' Z_i} \exp\{-\sum_{Y_j \leq Y_i} p_j e^{\theta' Z_i}\} g(Z_i) \right\}.$$

- The maximum likelihood estimators for $(\theta, p_1, ..., p_n)$ are given as: $\hat{\theta}$ solves the equation

$$\sum_{i=1}^{n} \left[Z_i - \frac{\sum_{Y_j \geq Y_i} Z_j e^{\theta' Z_j}}{\sum_{Y_j \geq Y_i} e^{\theta' Z_j}} \right] = 0$$

and

$$p_i = \frac{1}{\sum_{Y_j \geq Y_i} e^{\theta' Z_j}}.$$

- Bayesian estimation
  - The parameter θ in the model distribution $\{p_\theta(x)\}$ is treated as a random variable with some prior distribution $\pi(\theta)$.
  - The estimator for θ is defined as a value depending on the data and minimizing the expected loss function or the maximal loss function, where the loss function is denoted as $l(\theta, \hat{\theta}(X))$.
  - The usual loss function includes the quadratic loss $(\theta - \hat{\theta}(X))^2$, the absolute loss $|\theta - \hat{\theta}(X)|$, etc.
  - It often turns out that $\hat{\theta}(X)$ can be determined from the posterior distribution $P(\theta|X) = P(X|\theta)P(\theta)/P(X)$.

- Example
  - Suppose $X \sim N(\mu, 1)$. μ has an improper prior distribution and is uniform in $(-\infty, \infty)$. It is clear that the estimator $\hat{\theta}(X)$, minimizing the quadratic loss $E[(\theta - \hat{\theta}(X))^2]$, is the posterior mean $E[\theta|X] = X$.

- Non-exhaustive list of estimation methods
  - Other likelihood based estimation: partial likelihood estimation, conditional likelihood estimation, profile likelihood estimation, quasi-likelihood estimation, pseudo-likelihood estimation, penalized likelihood estimation
  - Other non-likelihood based estimation: rank-based estimation (R-estimation), L-estimation, empirical Bayesian estimation, minimax estimation, estimation under invariance principle

- A brief summary
  - no clear distinction among all the methods
  - each method has its own advantage
  - two points should be considered in choosing which method (estimator):
    (a) nice theoretical property, for example, unbiasedness (consistency), minimal variance, minimizing some loss function, asymptotic optimality
    (b) convenience in numerical calculation

Cramér-Rao Bounds for Parametric Models

A simple case: one-dimensional parametric model

$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R$.

Question: how well can one estimator be?

- Some basic assumptions
  - $X \sim P_\theta$ on $(\Omega, \mathcal{A})$ with $\theta \in \Theta$.
  - $p_\theta = dP_\theta/d\mu$ exists where μ is a σ-finite dominating measure.
  - $T(X) \equiv T$ estimates $q(\theta)$ and has $E_\theta[|T(X)|] < \infty$; set $b(\theta) = E_\theta[T] - q(\theta)$.
  - $q'(\theta) \equiv \dot{q}(\theta)$ exists.

- C-R information bound

Theorem 4.1 Information bound, Cramér-Rao Inequality Suppose:

(C1) Θ is an open subset of the real line.

(C2) There exists a set B with $\mu(B) = 0$ such that for $x \in B^c$, $\partial p_\theta(x)/\partial\theta$ exists for all θ. Moreover, $A = \{x : p_\theta(x) = 0\}$ does not depend on θ.

(C3) $I(\theta) = E_\theta[\dot{l}_\theta(X)^2] > 0$ where $\dot{l}_\theta(x) = \partial \log p_\theta(x)/\partial\theta$. Here, $I(\theta)$ is the called the *Fisher information* for θ and $\dot{l}_\theta$ is called the *score function* for θ.

(C4) $\int p_\theta(x)d\mu(x)$ and $\int T(x)p_\theta(x)d\mu(x)$ can both be differentiated with respect to θ under the integral sign.

(C5) $\int p_\theta(x)d\mu(x)$ can be differentiated twice under the integral sign.

If (C1)-(C4) hold, then

$$Var_\theta(T(X)) \geq \frac{\{\dot{q}(\theta) + \dot{b}(\theta)\}^2}{I(\theta)},$$

and the lower bound is equal to $\dot{q}(\theta)^2/I(\theta)$ if T is unbiased. Equality holds for all θ if and only if for some function $A(\theta)$, we have

$$\dot{l}_\theta(x) = A(\theta)\{T(x) - E_\theta[T(X)]\}, \quad a.e. \mu.$$

If, in addition, (C5) holds, then

$$I(\theta) = -E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right\} = -E_\theta[\ddot{l}_\theta(X)].$$

Proof

Note

$$q(\theta) + b(\theta) = \int T(x) p_\theta(x) d\mu(x) = \int_{A^c \cap B^c} T(x) p_\theta(x) d\mu(x).$$

$\Rightarrow$ from (C2) and (C4),

$$\dot{q}(\theta) + \dot{b}(\theta) = \int_{A^c \cap B^c} T(x) \dot{l}_\theta(x) p_\theta(x) d\mu(x) = E_\theta[T(X)\dot{l}_\theta(X)].$$

$\int_{A^c \cap B^c} p_\theta(x) d\mu(x) = 1 \Rightarrow$

$$0 = \int_{A^c \cap B^c} \dot{l}_\theta(x) p_\theta(x) d\mu(x) = E_\theta[\dot{l}_\theta(X)].$$

$\Rightarrow$

$$\dot{q}(\theta) + \dot{b}(\theta) = Cov(T(X), \dot{l}_\theta(X)).$$

By the Cauchy-Schwartz inequality, $\Rightarrow$

$$|\dot{q}(\theta) + \dot{b}(\theta)| \leq Var(T(X))Var(\dot{l}_\theta(X)).$$

The equality holds if and only if

$$\dot{l}_\theta(X) = A(\theta)\left\{T(X) - E_\theta[T(X)]\right\}, a.s.$$

If (C5) holds, differentiate

$$0 = \int \dot{l}_\theta(x)p_\theta(x)d\mu(x)$$

$\Rightarrow$

$$0 = \int \ddot{l}_\theta(x)p_\theta(x)d\mu(x) + \int \dot{l}_\theta(x)^2 p_\theta(x)d\mu(x).$$

$\Rightarrow I(\theta) = -E_\theta[\ddot{l}_\theta(X)].$

- Examples for calculating bounds
  - Suppose $X_1, ..., X_n$ are i.i.d $Poisson(\theta)$.

$$\dot{l}_\theta(X_1, ..., X_n) = \frac{n}{\theta}(\bar{X}_n - \theta).$$

$$I_n(\theta) = n^2/\theta^2 Var(\bar{X}_n) = n/\theta.$$

Note $\bar{X}_n$ is the UMVUE of θ and $Var(\bar{X}_n) = \theta/n$. We conclude that $\bar{X}_n$ attains the lower bound.

However, although $T_n = \bar{X}_n^2 - n^{-1}\bar{X}_n$ is UMVUE of θ^2, we find $Var(T_n) = 4\theta^3/n + 2\theta^2/n^2 >$ the Cramér-Rao lower bound for θ^2. In other words, some UMVUEs attain the lower bound but some do not.

– Suppose $X_1, ..., X_n$ are i.i.d with density $p_\theta(x) = g(x-\theta)$ where g is a known density. This family is the one-dimensional location model. Assume g' exists and the regularity conditions in Theorem 3.1 are satisfied. Then

$$I_n(\theta) = nE_\theta[\frac{g'(X-\theta)}{g(X-\theta)}^2] = n\int \frac{g'(x)^2}{g(x)}dx.$$

Note the information does not depend on θ.

– Suppose $X_1, ..., X_n$ are i.i.d with density $p_\theta(x) = g(x/\theta)/\theta$ where g is a known density function. This model is a one-dimensional scale model with the common shape g. It is direct to calculate

$$I_n(\theta) = \frac{n}{\theta^2} \int (1 + y\frac{g'(y)}{g(y)})^2 g(y)dy.$$

Generalization to Multi-parameter Family

$\mathcal{P} = \{P_\theta : \theta \in \Theta \subset R^k\}$.

- Basic assumptions

Assume that P_θ has density function p_θ with respect to some σ-finite dominating measure μ; $T(X)$ is an estimator for $q(\theta)$ with $E_\theta[|T(X)|] < \infty$ and $b(\theta) = E_\theta[T(X)] - q(\theta)$ is the bias of $T(X)$; $\dot{q}(\theta) = \nabla q(\theta)$ exists.

- Information bound

Theorem 4.2 Information inequality Suppose that
(M1) Θ an open subset in R^k.
(M2) There exists a set B with $\mu(B) = 0$ such that for $x \in B^c$, $\partial p_\theta(x)/\partial\theta_i$ exists for all θ and $i = 1, ..., k$. The set $A = \{x : p_\theta(x) = 0\}$ does no depend on θ.
(M3) The $k \times k$ matrix
$I(\theta) = (I_{ij}(\theta)) = E_\theta[\dot{l}_\theta(X)\dot{l}_\theta(X)'] > 0$ is positive definite, where

$$\dot{l}_{\theta_i}(x) = \frac{\partial}{\partial\theta_i} \log p_\theta(x).$$

Here, $I(\theta)$ is called the Fisher information matrix for θ and $\dot{l}_\theta$ is called the score for θ.
(M4) $\int p_\theta(x)d\mu(x)$ and $\int T(x)p_\theta(x)d\mu(x)$ can both be

differentiated with respect to θ under the integral sign.
(M5) $\int p_\theta(x)d\mu(x)$ can be differentiated twice with respect to θ under the integral sign.
If (M1)-(M4) holds, than

$$Var_\theta(T(X)) \geq (\dot{q}(\theta) + \dot{b}(\theta))' I^{-1}(\theta)(\dot{q}(\theta) + \dot{b}(\theta))$$

and this lower bound is equal $\dot{q}(\theta)' I(\theta)^{-1}\dot{q}(\theta)$ if $T(X)$ is unbiased. If, in addition, (M5) holds, then

$$I(\theta) = -E_\theta[\ddot{l}_{\theta\theta}(X)] = -\Big(E_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right\}\Big).$$

Proof Under (M1)-(M4),

$$\dot{q}(\theta) + \dot{b}(\theta) = \int T(x)\dot{l}_\theta(x)p_\theta(x)d\mu(x) = E_\theta[T(x)\dot{l}_\theta(X)].$$

From $\int p_\theta(x)d\mu(x) = 1$, $0 = E_\theta[\dot{l}_\theta(X)]$.

$\Rightarrow$

$$\begin{aligned}
& |\left\{\dot{q}(\theta) + \dot{b}(\theta)\right\}' I(\theta)^{-1} \left\{\dot{q}(\theta) + \dot{b}(\theta)\right\}| \\
= \quad & |E_\theta[T(X)(\dot{q}(\theta) + \dot{b}(\theta))' I(\theta)^{-1}\dot{l}_\theta(X)]| \\
= \quad & |Cov_\theta(T(X), (\dot{q}(\theta) + \dot{b}(\theta))' I(\theta)^{-1}\dot{l}_\theta(X))| \\
\leq \quad & \sqrt{Var_\theta(T(X))(\dot{q}(\theta) + \dot{b}(\theta))' I(\theta)^{-1}(\dot{q}(\theta) + \dot{b}(\theta))}.
\end{aligned}$$

Under (M5), differentiate $\int \dot{l}_\theta(x)p_\theta(x)d\mu(x) = 0$

$\Rightarrow$

$$I(\theta) = -E_\theta[\ddot{l}_{\theta\theta}(X)] = -\left(E_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right\}\right).$$

- Examples
  - The Weibull family $\mathcal{P}$ is the parametric model with densities

$$p_\theta(x) = \frac{\beta}{\alpha}(\frac{x}{\alpha})^{\beta-1}\exp\left\{-(\frac{x}{\alpha})^\beta\right\} I(x \geq 0)$$

with respect to the Lebesgue measure where $\theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)$.

$$\dot{l}_\alpha(x) = \frac{\beta}{\alpha}\left\{(\frac{x}{\alpha})^\beta - 1\right\},$$

$$\dot{l}_\beta(x) = \frac{1}{\beta} - \frac{1}{\beta}\log\left\{(\frac{x}{\alpha})^\beta\right\}\left\{(\frac{x}{\alpha})^\beta - 1\right\}.$$

⇒ the Fisher information matrix is

$$I(\theta) = \begin{pmatrix} \beta^2/\alpha^2 & -(1-\gamma)/\alpha \\ -(1-\gamma)/\alpha & \left\{\pi^2/6 + (1-\gamma)^2\right\}/\beta^2 \end{pmatrix},$$

where γ is Euler's constant ($\gamma \approx 0.5777...$). The computation of $I(\theta)$ is simplified by noting that $Y \equiv (X/\alpha)^\beta \sim \text{Exponential}(x)$.

Efficient Influence Function and Score Function

- Definition
  - $T(X) = \dot{q}(\theta)' I^{-1}(\theta) \dot{l}_\theta(X)$, the latter is called the *efficient influence function* for estimating $q(\theta)$ and its variance, which is equal to $\dot{q}(\theta)' I(\theta)^{-1} \dot{q}(\theta)$, is called the *information bound* for $q(\theta)$.

- Notation

If we regard $q(\theta)$ as a function on all the distributions of $\mathcal{P}$ and denote $\nu(P_\theta) = q(\theta)$, then

  - the efficient influence function is represented as $\tilde{l}(X, P_\theta|\nu, \mathcal{P})$
  - the information bound for $q(\theta)$ is denoted as $I^{-1}(P_\theta|\nu, \mathcal{P})$

- Invariance property

Proposition 4.3 The information bound $I^{-1}(P|\nu, \mathcal{P})$ and the efficient influence function $\tilde{l}(\cdot, P|\nu, \mathcal{P})$ are invariant under smooth changes of parameterization.

Proof

Suppose $\gamma \mapsto \theta(\gamma)$ is a one-to-one continuously differentiable mapping of an open subset Γ of R^k onto Θ with nonsingular differential $\dot{\theta}$.

The model of distribution can be represented as $\{P_{\theta(\gamma)} : \gamma \in \Gamma\}$.

The score for γ is $\dot{\theta}(\gamma)\dot{l}_\theta(X) \Rightarrow$ the information matrix for γ is equal to $I(\gamma) = \dot{\theta}(\gamma)' I(\theta)\dot{\theta}(\gamma)$.

Under the new parameterization, the information bound for $q(\theta) = q(\theta(\gamma))$ is

$$(\dot{q}(\theta(\gamma))\dot{\theta}(\gamma))' I(\gamma)^{-1} (\dot{q}(\theta(\gamma))\dot{\theta}(\gamma)) = \dot{q}(\theta)' I(\theta)^{-1} \dot{q}(\theta),$$

which is the same as the information matrix for $\theta = \theta(\gamma)$.

The efficient influence function for γ is equal to

$$(\dot{\theta}(\gamma)\dot{q}(\theta(\gamma)))' I(\gamma)^{-1} \dot{l}_\gamma = \dot{q}(\theta)' I(\theta)^{-1} \dot{l}_\theta$$

and it is the same as the efficient influence function for θ.

- Canonical parameterization

$\theta' = (\nu', \eta')$ and $\nu \in \mathcal{N} \subset R^m$, $\eta \in \mathcal{H} \subset R^{k-m}$. ν can be regarded as a map mapping P_θ to one component of θ, ν, and it is the parameter of interest while η is a nuisance parameter.

Information bound in presence of nuisance parameter

Goal: want to assess the cost of not knowing η by comparing the information bounds and the efficient influence functions for ν in the model $\mathcal{P}$ (η is unknown parameter) and $\mathcal{P}_\eta$ (η is known and fixed).

Case I: η is unknown parameter

$$\dot{l}_\theta = \begin{pmatrix} \dot{l}_1 \\ \dot{l}_2 \end{pmatrix}, \quad \tilde{l}_\theta = \begin{pmatrix} \tilde{l}_1 \\ \tilde{l}_2 \end{pmatrix}$$

$$I(\theta) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where $I_{11} = E_\theta[\dot{l}_1 \dot{l}_1']$, $I_{12} = E_\theta[\dot{l}_1 \dot{l}_2']$, $I_{21} = E_\theta[\dot{l}_2 \dot{l}_1^1]$, and $I_{22} = E_\theta[\dot{l}_2 \dot{l}_2']$.

$$I^{-1}(\theta) = \begin{pmatrix} I_{11\cdot 2}^{-1} & -I_{11\cdot 2}^{-1} I_{12} I_{22}^{-1} \\ -I_{22\cdot 1}^{-1} I_{21} I_{11}^{-1} & I_{22\cdot 1}^{-1} \end{pmatrix} \equiv \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix},$$

where $I_{11\cdot 2} = I_{11} - I_{12} I_{22}^{-1} I_{21}$, $\quad I_{22\cdot 1} = I_{22} - I_{21} I_{11}^{-1} I_{12}$.

- Conclusions in Case I
  - The information bound for estimating ν is equal to
$$I^{-1}(P_\theta|\nu,\mathcal{P}) = \dot{q}(\theta)' I^{-1}(\theta)\dot{q}(\theta),$$
where $q(\theta)=\nu$, and $\dot{q}(\theta) = \left(I_{m\times m} \quad 0_{m\times(k-m)} \right)$, $\Rightarrow$
$$I^{-1}(P_\theta|\nu,\mathcal{P}) = I_{11\cdot 2}^{-1} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}.$$
  - The efficient influence function for ν is given by
$$\tilde{l}_1 = \dot{q}(\theta)' I^{-1}(\theta)\dot{l}_\theta = I_{11\cdot 2}^{-1}\dot{l}_1^*,$$
where $\dot{l}_1^* = \dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2$. It is easy to check
$$I_{11\cdot 2} = E[\dot{l}_1^*(\dot{l}_1^*)'].$$
Thus, l_1^* is called the *efficient score function* for ν in $\mathcal{P}$.

Case II: η is known and fixed

- The information bound for ν is just I_{11}^{-1},
- The efficient influence function for ν is equal to $I_{11}^{-1}\dot{l}_1$.

Comparison

- knowing η increases the Fisher information for ν and decreases the information bound for ν,
- knowledge of η does not increase information about ν if and only if $I_{12} = 0$. In this case, $\tilde{l}_1 = I_{11}^{-1}\dot{l}_1$ and $l_1^* = l_1$.

Examples

– Suppose

$$\mathcal{P} = \{P_\theta : p_\theta = \phi((x-\nu)/\eta)/\eta, \nu \in R, \eta > 0\}.$$

Note that

$$\dot{l}_\nu(x) = \frac{x-\nu}{\eta^2}, \quad \dot{l}_\eta(x) = \frac{1}{\eta}\left\{\frac{(x-\nu)^2}{\eta^2} - 1\right\}.$$

Then the information matrix $I(\theta)$ is given by by

$$I(\theta) = \begin{pmatrix} \eta^{-2} & 0 \\ 0 & 2\eta^{-2} \end{pmatrix}.$$

Then we can estimate the ν equally well whether we know the variance or not.

– If we reparameterize the above model as

$$P_\theta = N(\nu, \eta^2 - \nu^2), \eta^2 > \nu^2.$$

An easy calculation shows that $I_{12}(\theta) = \nu\eta/(\eta^2 - \nu^2)^2$. Thus lack of knowledge of η in this parameterization does change the information bound for estimation of ν.

- Geometric interpretation

Theorem 4.3

(A) The efficient score function $\dot{l}_1^*(\cdot, P_\theta|\nu, \mathcal{P})$ is the projection of the score function $\dot{l}_1$ on the orthocomplement of $[\dot{l}_2]$ in $L_2(P_\theta)$, where $[\dot{l}_2]$ is the linear span of the components of $\dot{l}_2$.

(B) The efficient influence function $\tilde{l}(\cdot, P_\theta|\nu, \mathcal{P}_\eta)$ is the projection of the efficient influence function $\tilde{l}_1$ on $[\dot{l}_1]$ in $L_2(P_\theta)$.

Proof (A) The projection of $\dot{l}_1$ on $[\dot{l}_2]$ is equal to $\Sigma\dot{l}_2$ for some matrix Σ.

Since $E[(\dot{l}_1 - \Sigma\dot{l}_2)\dot{l}_2'] = 0$, $\Sigma = I_{12}I_{22}^{-1}$, and thus the projection on the orthocomplement of $[\dot{l}_2]$ is equal to

$$\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2 = \dot{l}_1^*.$$

(B)

$$\tilde{l}_1 = I_{11\cdot 2}^{-1}(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2) = (I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1})(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2)$$
$$= I_{11}^{-1}\dot{l}_1 - I_{11}^{-1}I_{12}\tilde{l}_2.$$

From (A), $\tilde{l}_2$ is orthogonal to $\dot{l}_1$, the projection of $\tilde{l}_1$ on $[\dot{l}_1]$ is equal $I_{11}^{-1}\dot{l}_1 = \tilde{l}(\cdot, P_\theta|\nu, \mathcal{P}_\eta)$.

term	**notation**	$\mathcal{P}$ (η unknown)	$\mathcal{P}_\eta$ (η known)
efficient score	$l_1^*(, P\|\nu, \cdot)$	$l_1^* = l_1 - I_{12}I_{22}^{-1}l_2$	l_1
information	$I(P\|\nu, \cdot)$	$E[l_1^*(l_1^*)'] = I_{11} - I_{12}I_{22}^{-1}I_{21}$	I_{11}
efficient influence information	$\tilde{l}_1(\cdot, P\|\nu, \cdot)$	$\tilde{l}_1 = I^{11}l_1 + I^{12}l_2 = I_{11\cdot 2}^{-1}l_1^*$ $= I_{11}^{-1}l_1 - I_{11}^{-1}I_{12}\tilde{l}_2$	$I_{11}^{-1}l_1$
information bound	$I^{-1}(P\|\nu, \cdot)$	$I^{11} = I_{11\cdot 2}^{-1}$ $= I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1}$	I_{11}^{-1}

Asymptotic Efficiency Bound

- Motivation
  - The Cramér-Rao bound can be considered as the lower bound for any unbiased estimator in finite sample. One may ask whether such a bound still holds in large sample.
  - To be more specific, we suppose $X_1, ..., X_n$ are i.i.d P_θ ($\theta \in R$) and an estimator T_n for θ satisfies that

$$\sqrt{n}(T_n - \theta) \rightarrow_d N(0, V(\theta)^2).$$

  - Question: $V(\theta)^2 \geq 1/I(\theta)$?

• Super-efficient estimator (Hodge's estimator) Let $X_1, ..., X_n$ be i.i.d $N(\theta, 1)$ so that $I(\theta) = 1$. Let $|a| < 1$ and define

$$T_n = \begin{cases} \bar{X}_n & \text{if} |\bar{X}_n| > n^{-1/4} \\ a\bar{X}_n & \text{if} |\bar{X}_n| \leq n^{-1/4}. \end{cases}$$

$$\begin{aligned}
\sqrt{n}(T_n - \theta) &= \sqrt{n}(\bar{X}_n - \theta) I(|\bar{X}_n| > n^{-1/4}) \\
&\quad + \sqrt{n}(a\bar{X}_n - \theta) I(|\bar{X}_n| \leq n^{-1/4}) \\
&= {}_d Z I(|Z + \sqrt{n}\theta| > n^{1/4}) \\
&\quad + \left\{ aZ + \sqrt{n}(a-1)\theta \right\} I(|Z + \sqrt{n}\theta| \leq n^{1/4}) \\
&\to_{a.s.} ZI(\theta \neq 0) + aZI(\theta = 0).
\end{aligned}$$

Thus, the asymptotic variance of $\sqrt{n}T_n$ is equal 1 for $\theta \neq 0$ and a^2 for $\theta = 0$. T_n is a superefficient estimator.

- Locally Regular Estimator

Definition 4.2 $\{T_n\}$ is a *locally regular estimator* of θ at $\theta = \theta_0$ if, for every sequence $\{\theta_n\} \subset \Theta$ with $\sqrt{n}(\theta_n - \theta) \to t \in R^k$, under P_{θ_n},

$$\text{(local regularity)} \quad \sqrt{n}(T_n - \theta_n) \to_d Z, \quad \text{as} \quad n \to \infty$$

where the distribution of Z depend on θ_0 but not on t.

- Implication of LRE
  - The limit distribution of $\sqrt{n}(T_n - \theta_n)$ does not depend on the direction of approach t of θ_n to θ_0. $\{T_n\}$ is a locally Gaussian regular if Z has normal distribution.
  - $\sqrt{n}(T_n - \theta_n) \to_d Z$ under P_{θ_n} is equivalent to saying that for any bounded and continuous function g, $E_{\theta_n}[g(\sqrt{n}(T_n - \theta_n))] \to E[g(Z)]$.
  - T_n in the first example is not a locally regular estimator.

- Hellinger Differentiability

A model $\mathcal{P} = \{P_\theta : \theta \in R^k\}$ is a parametric model dominated by a σ-finite measure μ. It is called a Hellinger-differentiable parametric model if

$$\|\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h'\dot{l}_\theta\sqrt{p_\theta}\|_{L_2(\mu)} = o(|h|),$$

where $p_\theta = dP_\theta/d\mu$.

- locally asymptotic normality (LAN)

In a model $\mathcal{P} = \{P_\theta : \theta \in R^k\}$ dominated by a σ-finite measure μ, suppose $p_\theta = dP_\theta/d\mu$. Let $l(x;\theta) = \log p(x,\theta)$ and let

$$l_n(\theta) = \sum_{i=1}^n l(X_i;\theta)$$

be the log-likelihood function of $X_1, ..., X_n$. The local asymptotic normality condition at θ_0 is

$$l_n(\theta_0 + n^{-1/2}t) - l_n(\theta_0) \rightarrow_d N(-\frac{1}{2}t'I(\theta_0)t, t'I(\theta_0)t)$$

under P_{θ_0}.

Convolution Result

Theorem 4.4 (Hájek's convolution theorem) Under three regularity conditions with $I(\theta_0)$ nonsingular, the limit distribution of $\sqrt{n}(T_n - \theta_0)$ under P_{θ_0} satisfies

$$Z =^d Z_0 + \Delta_0,$$

where $Z_0 \sim N(0, I^{-1}(\theta_0))$ is independent of Δ_0.

- Conclusion
  - the asymptotic variance of $\sqrt{n}(T_n - \theta_0)$ is larger than or equal to $I^{-1}(\theta_0)$;
  - the Cramér-Rao bound is a lower bound for the asymptotic variances of any locally regular estimator;
  - a further question is what estimator can attains this bound asymptotically (answer will be given in next chapter).

- How to check three regularity conditions?

Proposition 4.6. For every θ in an open subset of R^k let p_θ be a μ-probability density. Assume that the map $\theta \mapsto s_\theta(x) = \sqrt{p_\theta(x)}$ is continuously differentiable for every x. If the elements of the matrix $I(\theta) = E[(\dot{p}_\theta/p_\theta)(\dot{p}_\theta/p_\theta)']$ are well defined and continuous at θ, then the map $\theta \to \sqrt{p_\theta}$ is Hellinger differentiable with $\dot{l}_\theta$ given by $\dot{p}_\theta/p_\theta$.

Proof

$\dot{p}_\theta = 2 s_\theta \dot{s}_\theta$

$\Rightarrow$ $\dot{s}_\theta$ is zero whenever $\dot{p}_\theta = 0$.

$$\int \left\{ \frac{s_{\theta+th_t} - s_\theta}{t} \right\}^2 d\mu = \int \left\{ \int_0^1 (h_t)' \dot{s}_{\theta+uth} du \right\}^2 d\mu$$

$$\leq \int \int_0^1 ((h_t)' \dot{s}_{\theta+uth_t})^2 du d\mu = \frac{1}{2} \int_0^1 h_t' I(\theta + uth_t) h_t du.$$

As $h_t \to h$, the right side converges to $\int (h' \dot{s}_\theta)^2 d\mu$.

Since $\frac{s_{\theta+th_t} - s_\theta}{t} - h' \dot{s}_\theta \to 0$, the same proof as Theorem 3.1 (E) of Chapter 3 gives

$$\int \left[\frac{s_{\theta+th_t} - s_\theta}{t} - h' \dot{s}_\theta \right]^2 d\mu \to 0.$$

Proposition 4.7 If $\{T_n\}$ is an estimator sequence of $q(\theta)$ such that

$$\sqrt{n}(T_n - q(\theta)) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}_\theta I(\theta)^{-1} \dot{l}_\theta(X_i) \to_p 0,$$

where ψ is differentiable at θ, then T_n is the efficient and regular estimator for $q(\theta)$.

Proof

"$\Rightarrow$" Let $\Delta_{n,\theta} = n^{-1/2}\sum_{i=1}^{n} \dot{l}_\theta(X_i)$. $\Rightarrow$ $\Delta_{n,\theta} \to^d \Delta_\theta \sim N(0, I(\theta))$.

From Step I of Theorem 4.4, $\log dQ_n/dP_n$ is equivalent to $h'\Delta_{n,\theta} - h'I(\theta)h/2$ asymptotically.

$\Rightarrow$ Slutsky's theorem gives that under P_θ,

$$\left(\sqrt{n}(T_n - q(\theta)), \log\frac{dQ_n}{dP_n}\right) \to_d (\dot{\psi}_\theta I(\theta)^{-1}\Delta_\theta, h'\Delta_\theta - h'I(\theta)h/2)$$

$$\sim N\left(\begin{pmatrix} 0 \\ -h'I(\theta)h/2 \end{pmatrix}, \begin{pmatrix} \dot{\psi}_\theta I(\theta)^{-1}\dot{\psi}_\theta & \dot{\psi}_\theta h \\ \dot{\psi}_\theta h' & h'I(\theta)h \end{pmatrix}\right).$$

$\Rightarrow$ From Le Cam's third lemma, under $P_{\theta+h/\sqrt{n}}$, $\sqrt{n}(T_n - q(\theta))$ converges in distribution to $N(\dot{\psi}_\theta h, \dot{\psi}_\theta I(\theta)'\dot{\psi}_\theta')$.

$\Rightarrow$ $P_{\theta+h/\sqrt{n}}$, $\sqrt{n}(T_n - q(\theta + h/\sqrt{n})) \to_d N(0, \dot{\psi}_\theta I(\theta)'\dot{\psi}_\theta')$.

- Asymptotic linear estimator

Definition 4.4 If a sequence of estimators $\{T_n\}$ has the expansion

$$\sqrt{n}(T_n - q(\theta)) = n^{-1/2} \sum_{i=1}^{n} \Gamma(X_i) + R_n,$$

where R_n converges to zero in probability, then T_n is called an *asymptotically linear estimator* for $q(\theta)$ with *influence function* Γ.

Proposition 4.3 Suppose T_n is an asymptotically linear estimator of $\nu = q(\theta)$ with influence function Γ. Then
A. T_n is Gaussian regular at θ_0 if and only if $q(\theta)$ is differentiable at θ_0 with derivative $\dot{q}_\theta$ and, with $\tilde{l}_\nu = \tilde{l}(\cdot, P_{\theta_0}|q(\theta), \mathcal{P})$ being the efficient influence function for $q(\theta)$, $E_{\theta_0}[(\Gamma - \tilde{l}_\nu)\dot{l}] = 0$ for any score $\dot{l}$ of $\mathcal{P}$.
B. Suppose $q(\theta)$ is differentiable and T_n is regular. Then $\Gamma \in [\dot{l}]$ if and only if $\Gamma = \tilde{l}_\nu$.

Proof

A. By asymptotic linearity of T_n,

$$\begin{pmatrix} \sqrt{n}(T_n - q(\theta_0)) \\ L_n(\theta_0 + t_n/\sqrt{n}) - L_n(\theta_0) \end{pmatrix}$$

$$\to_d N\left\{\begin{pmatrix} 0 \\ -t'I(\theta_0)t \end{pmatrix}, \begin{pmatrix} E_{\theta_0}[\Gamma\Gamma'] & E_{\theta_0}[\Gamma\dot{l}']t \\ E_{\theta_0}[\dot{l}\Gamma']t & t'I(\theta_0)t \end{pmatrix}\right\}.$$

From Le Cam's third lemma, $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0)) \to_d N(E_{\theta_0}[\Gamma'\dot{l}]t, E_{\theta_0}[\Gamma\Gamma']).$$

If T_n is regular, then, under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0 + t_n/\sqrt{n})) \to_d N(0, E_{\theta_0}[\Gamma\Gamma']).$$

$\Rightarrow$ $\sqrt{n}(q(\theta_0 + t_n/\sqrt{n}) - q(\theta_0)) \to E_{\theta_0}[\Gamma'\dot{l}]t.$

$\Rightarrow$ $\dot{q}_\theta = E_\theta[\Gamma'\dot{l}]$. Note $E_{\theta_0}[\tilde{l}_\nu'\dot{l}] = \dot{q}_\theta$.

To prove the other direction, since $q(\theta)$ is differentiable and under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0)) \rightarrow_d N(E_{\theta_0}[\Gamma'\dot{l}]t, E[\Gamma\Gamma'])$$

$\Rightarrow$ from Le Cam's third lemma, under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0 + t_n/\sqrt{n})) \rightarrow_d N(0, E[\Gamma\Gamma']).$$

$\Rightarrow$ T_n is Gaussian regular.

B. If T_n is regular, from A, $\Gamma - \tilde{l}_\nu$ is orthogonal to any score in $\mathcal{P}$.

$\Rightarrow$ $\Gamma \in [\dot{l}]$ implies that $\Gamma = \tilde{l}_\nu$. The converse is obvious.

READING MATERIALS: Lehmann and Casella, Sections 1.6, 2.1, 2.2, 2.3, 2.5, 2.6, 6.1, 6.2, Ferguson, Chapter 19 and Chapter 20