STAT 583 Exam Guide

June 2023

Contents

1 Chapter 1

1.1 Weak convergence

Definition 1.1: Weak Convergence

Let $(\mathbb{D}, d)$ denote a generic metric space. We say that a $\mathbb{D}-$valued sequence $\{X_n\}_{n=1}^{\infty}$ converges weakly to X if, for all bounded continuous $f : \mathbb{D} \mapsto \mathbb{R}$,

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$$

Theorem 1.2: Portmanteau

Let $\{X_n\}_{n=1}^{\infty}$ denote a sequence of $\mathbb{D}$-valued RVs and X is an RV. Then the following are equivalent:

- $X_n \rightsquigarrow X$
- $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous functions
- $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, Lipschitz-continuous functions
- $limsup_n \mathbb{E}[f(X_n)] \leq \mathbb{E}[f(X)]$ for every upper semicontinuous f that is bounded above
- $liminf_n \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ for every lower semicontinuous f that is bounded below
- $limsup_n P\{X_n \in F\} \leq P\{X \in F\}$ for all closed F, where $P\{X_n \in F\} := \mathbb{E}\mathbb{I}\{X_n \in F\}$
- $liminf_n P\{X_n \in U\} \geq P\{X \in U\}$ for all open U

Theorem 1.3: Continuous mapping for metric spaces

Let $(\mathbb{D}, d)$ and $(\mathbb{E}, e)$ be metric spaces. Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of $\mathbb{D}$-valued RVs and X is $\mathbb{D}_0$-valued where $\mathbb{D}_0 \subseteq \mathbb{D}$. Let $f : \mathbb{D} \mapsto \mathbb{E}$ be continuous on $\mathbb{D}_0$. Then $X_n \rightsquigarrow X$ implies $f(X_n) \rightsquigarrow f(X)$.

Lemma 1.4: Partial Slutsky's for weak convergence

Let $X_1, ...$ and $Y_1, ...$ be 2 sequences of $\ell^\infty(\mathscr{F})$-valued RVs for some function class $\mathscr{F}$. Suppose $||X_n - Y_n||_{\mathscr{F}} = o_P(1)$. Then if $X_n \rightsquigarrow X$ in $\ell^\infty(\mathscr{F})$ relative to $||\cdot||_{\mathscr{F}}$ for some $\ell^\infty(\mathscr{F})$-valued RV X, then $Y_n \rightsquigarrow X$ in $\ell^\infty(\mathscr{F})$ relative to $||\cdot||_{\mathscr{F}}$.

Definition 1.5: Asymptotically ρ-equicontinuous

Let ρ be a pseudometric on $\mathscr{F}$ and, for an $\delta > 0$, $\mathscr{F}(\delta) = \{(f_1, f_2) \in \mathscr{F}^2 : \rho(f_1, f_2) < \delta\}$. Then a stochastic process $\{X_n\}_{n=1}^{\infty}$ is asymptotically ρ-equicontinuous if

$$sup_{(f_1,f_2)\in\mathscr{F}(\delta_n)} |X_n(f_1) - X_n(f_2)| = o_P(1)$$

for all positive sequences $\delta_n \to 0$

Definition 1.6: Tightness

We say that an $\ell^\infty(\mathscr{F})$-valued RV is tight is $\forall \epsilon > 0$, there exists a compact set $K \subseteq \ell^\infty(\mathscr{F})$ s.t. $P(X \in K) \geq 1 - \epsilon$

Theorem 1.7: Equivalent characterization of weak convergence in $\ell^\infty(\mathscr{F})$

X_n converges weakly to a tight random variable X in $\ell^\infty(\mathscr{F})$ if and only if both of the following hold:

1. (Convergence in distribution of marginals): For each finite collection $\{f_1, ..., f_k\} \subseteq \mathscr{F}$, it holds that $\{X_n(f_j) : j = 1, ..., k\} \Rightarrow \{X(f_j) : j = 1, ..., k\}$
2. (Existence of a suitable pseudometric): There exists a pseudometric ρ on $\mathscr{F}$ s.t. X_n is asymptotically uniformly ρ-equicontinuous and $N(\epsilon, \mathscr{F}, \rho) < \infty$ for all $\epsilon > 0$

Lemma 1.8

Let $\mathscr{F}$ denote a collection of functions mapping from $X \mapsto \mathbb{R}$ and $\mathbb{G}_n := \{\sqrt{n}(P_n - P)f : f \in \mathscr{F}\}$. Then $\mathbb{G}_n$ is $\ell^\infty(\mathscr{F})$-valued when $\mathscr{F}$ has a finite and P_0-integrable envelope function, $\bar{F}$, i.e. $sup_{f\in\mathscr{F}}|f(x)| \leq \bar{F}(x)$

1.2 P_0-Donsker

Definition 1.9: P_0-Donsker

We say that $\mathscr{F}$ is P_0-Donsker if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathscr{F})$ for some tight weak limit $\mathbb{G}$.

- If $\mathscr{F}$ is P_0-Donsker, then $\mathbb{G}$ is a mean-zero Gaussian process with covariance function $(f_1, f_2) \mapsto \mathbb{E}[\mathbb{G}(f_1)\mathbb{G}(f_2)] = P_0(f_1 f_2) - P_0 f_1 P_0 f_2$

Remark 1.1. *: For a Donsker class, the pseudometric guaranteed to exist by Theorem 1.7 take the form* $\rho_0(f_1, f_2) = sd_{\P_0}[f_1(x) - f_2(x)]$

Theorem 1.10: Permanence properties of Donsker classes

- If $\mathscr{F}$ and $\mathscr{G}$ are Donsker classes, then $\mathscr{F} + \mathscr{G}$, $-\mathscr{F}$, and $\mathscr{F} \cup \mathscr{G}$ are Donsker classes as well
- Let $\mathscr{F}_1, ..., \mathscr{F}_k$ be P_0-Donsker with $||P_0||_{\mathscr{F}_j} < \infty$ for all j and let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ be Lipschitz. Then $\phi \circ (\mathscr{F}_1, ..., \mathscr{F}_k)$ is P_0-Donsker, provided $x \mapsto \phi(f_1(x), ..., f_k(x))$ is P_0-square integrable $\forall f_j \in \mathscr{F}_j, j \in \{1, ..., k\}$
- If $\mathscr{F}$ is Donsker, then $\mathscr{G} \subseteq \mathscr{F}$ is also Donsker

Theorem 1.11: Sufficient conditions for class to be Donsker

A class $\mathscr{F}$ of functions with a square integrable envelope function is Donsker if either of the following hold:

- $J_{[]}(1, \mathscr{F}, L^2(P_0)) < \infty$ where $J_{[]}(\delta, \mathscr{F}, L^2(P_0)) = \int_0^\delta \sqrt{logN_{[]}(\epsilon, \mathscr{F}, L^2(P_0))}d\epsilon$
- $\int_0^\infty sup_Q \sqrt{logN(\epsilon, \mathscr{F}, L^2(Q))} < \infty$

Remark 1.2. *: If a class is Donsker, it must be Glivenko-Cantelli.*

Example 1.1. (Using results for STAT 582)*:*

1. *Let $\mathscr{F}$ be a class of functions whose total variation is bounded by 1. Then $N_{[]}(\epsilon, \mathscr{F}, L^2(P)) \leq \frac{K}{\epsilon}$. This implies $J_{[]}(1, \mathscr{F}, L^2(P_0)) < \infty$, so $\mathscr{F}$ is Donsker.*

   This also implies the following collections of functions are Donsker:

   - *all univariate uniformly bounded and monotone functions*
   - *all differentiable univariate functions defined over a bounded region with a uniformly bounded derivative*

2. *Let $\mathscr{F}$ be a collection of indicators for half-lines $(-\infty, x]$ for $x \in \mathbb{R}$. Then $N_{[]}(\epsilon, \mathscr{F}, L^2(P)) \leq \frac{1}{\epsilon}$, so $\mathscr{F}$ is Donsker.*

3. *Let $\mathscr{F}$ be a collection of uniformly bounded and monotone functions. Then $N_{[]}(\epsilon, \mathscr{F}, L^2(P)) \leq \frac{1}{\epsilon}$, so $\mathscr{F}$ is Donsker.*

Example 1.2. (Confidence band of CDF)*: We are interested in obtaining a point and interval estimate for the CDF, $F_0 := P_0\{x \leq t\}$.*

Let $\mathscr{H} = \{x \mapsto \mathbb{I}\{x \leq t\} : t \in \mathbb{R}\}$ and observe $\{\sqrt{n}[F_n(t) - F_0(t)] : t \in \mathbb{R}\} = \{\mathbb{G}_n h : h \in \mathscr{H}\}$. Observe also, $\mathscr{H}$ has envelope function $x \mapsto 1$, so G_n is $\ell^\infty(\mathscr{H})$-valued.

From Lemma 10.15 in VdV, we know that $sup_Q log N(\epsilon, \mathscr{H}, L^2(Q)) < \infty$, so $\mathscr{H}$ is P-Donsker. Then by CMT, $||\mathbb{G}_n||_{\mathscr{H}} \rightsquigarrow ||\mathbb{G}||_{\mathscr{H}}$.

Then it follows that a suitable $(1-\alpha)$-level confidence band for F_0 is $F_n(t) \pm \frac{c}{\sqrt{n}}$ where c is the $(1-\alpha)$ quantile of $||\mathbb{G}||_{\mathscr{H}}$

Lemma 1.12: Evaluation of the empirical process on a random function

Let $\mathscr{F} \subseteq L^2(P)$ be a P-Donsker class satisfying $sup_{f \in \mathscr{F}} \rho_P(f) < \infty$, where $\rho_P : f \mapsto \sqrt{P(f - Pf)^2}$. Let $h_1, ...$ be a sequence of random functions in $L^2(P)$ such that $P(h_n \in \mathscr{F}) \to 1$ and $P(h_n - h_0)^2 = o_P(1)$ for some $h_0 \in \mathscr{F}$. Then $\mathbb{G}_n(h_n - h_0) = o_P(1)$.

2 Chapter 2

2.1 Asymptotic Linearity

Definition 2.1: Asymptotically linear

An estimator ψ_n of ψ_0 is called asymptotically linear if

$$\psi_n - \psi_0 = \frac{1}{n}\sum_i^n \phi_{P_0}(x_i) + o_P(n^{-1/2})$$

where the influence function, phi_{P_0}, has P_0-mean zero and is P_0-square integrable

Remark 2.1. *If ψ_n is an asymptotically linear estimator of ψ_0, then, by CLT and Slutsky's Theorem,*

$$\sqrt{n}(\psi_n - \psi_0) \rightsquigarrow N(0, \sigma^2_{P_0})$$

where $\sigma^2_{P_0} = P_0\phi^2_{P_0}$

Theorem 2.2: Delta Method

Suppose that ψ_n is an estimator of $\psi_0 \in \mathbb{R}^d$ s.t.

$$\sqrt{n}(\psi_n - \psi_0) \rightsquigarrow N(0, \Sigma)$$

Then if $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable

$$f(\psi_n) - f(\psi_0) = \langle \psi_n - \psi_0, \nabla f(\psi_0)\rangle + o_P(n^{-1/2})$$

Corollary 2.1. (for influence functions): *Suppose that ψ_n is an AL estimator of $\psi_0 \in \mathbb{R}^d$ with influence function ϕ_{P_0} and that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable. Then $f(\psi_n)$ is an AL estimator for $f(\psi_0)$ with influence function $x \mapsto \langle \nabla f(\psi_0), \phi_{P_0}(x)\rangle$*

Example 2.1. (Sample variance)*: Suppose we want to estimate $\sigma_0^2 = Var_{P_0}(X)$ with $\sigma_n^2 = \frac{1}{n}\sum_i^n (x_i - \frac{1}{n}\sum_i^n X_i)^2$. Observe σ_n^2 is an AL estimator for σ_0^2 with influence function $x \mapsto [x-\mu_0]^2 - \sigma_0^2$, where μ_0 is the true mean.*

Example 2.2. (Z-estimator)*: Suppose we want to estimate some $\psi_0 \in \mathbb{R}$, which is the unique solution to $P_0U(\psi) = 0$. An estimator ψ_n, defined as a solution to $P_nU(\psi) = 0$, is an AL estimator for ψ_0 with influence function $x \mapsto (-\frac{\partial}{\partial\psi}P_0U(\psi)|_{\psi=\psi_0})^{-1}U(\psi_0)(x)$.*

Example 2.3. (Average absolute deviation from the mean)*: Suppose we want to estimate $\psi_0 := \int |x - \mathbb{E}_{P_0}[X]| dP_0(x)$. Let $f_n(x) = |X - \bar{X}_n|$ and $f_0(x) = |x - \mu_0|$ and consider the estimator $\psi_n := Pnf_n$. Observe $\psi_n - \psi_0 = (P_n - P_0)f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)$.*

Note that the first term is linear and Lemma 1.12 can be used to show that $(P_n - P_0)(f_n - f_0) = o_P(n^{-1/2})$, so $\psi_n - \psi_0 = \frac{1}{n}\sum_i^n (f_0 - P_0f_0) + P_0(f_n - f_0) + o_P(n^{-1/2})$

Then let $h(u) = \int |x - \mu| dP_0(x)$ and note the second term, $P_0(f_n - f_0) = h(\bar{X}_n) - h(\mu_0) = \frac{1}{n}\sum_i^n h'(\mu_0)(x_i - \mu_0) + o_P(n^{-1/2}) = \frac{1}{n}\sum_i^n (2F_0(\mu_0) - 1)(x_i - \mu_0) + o_P(n^{-1/2})$.

Then it follows that ψ_n is an AL estimator for ψ_0 with influence function $x \mapsto |x - \mu_0| - \psi_0 + (2F_0(\mu_0) - 1)(x - \mu_0)$

Example 2.4. (Sample coefficient of variation)*: Suppose we want to estimate $c_0 = \sigma_0/\mu_0$. We can use the delta method to show $c_n = \sigma_n/\mu_n$ is an AL estimator for c_0 with influence function $x \mapsto c_0[\frac{1}{2}(\frac{x-\mu_0}{\sigma_0})^2 - \frac{x}{\mu_0} + \frac{1}{2}]$.*

Example 2.5. (Sample quantile)*: Suppose we want to estimate $Q_0(p)$, the p^{th} quantile of P_0. We can show that $Q_n(p) := inf\{y : F_n(y) \geq p\}$ is an AL estimator for $Q_0(p)$ with influence function $x \mapsto \frac{F_0(Q_0(p)) - \mathbb{I}(x \leq Q_0(p))}{f_0(Q_0(p))}$.*

2.2 V- and U-Statistics

Definition 2.3: V-Statistics

A V-statistic takes the form

$$V_0(p) := P_0^m H = \int ... \int H(x_1, ..., x_m) dP(x_1)...dP(x_m)$$

Then the resulting plug-in estimator for V_0 is

$$V_n := Pn^m H = \frac{1}{n^m} \sum_{i_1}^{n} ... \sum_{i_m}^{n} H(x_{i_1}, ..., x_{i_m})$$

Lemma 2.4: Linearization of V-statistic

If H is symmetric, then $V_n - V_0 = (P_n^m - P_0^m)H = \sum_k^m \binom{m}{k}(P_n - P_0)^k H_k$, where $H_k := P_0^{m-k} H$.

Note: If H is not symmetric, we can replace it with the average of evaluations of H over permutations of its arguments.

Corollary 2.2. *: We have shown $V_n - V_0 = m(P_n - P_0)H_1 + \sum_{k=2}^m \binom{m}{k}(P_n - P_0)^k H_k$. We also know*

1. $\sqrt{n} m(P_n - P_0)H_1 \rightsquigarrow \mathcal{N}(0, \sigma^2)$, *where* $\sigma^2 := m^2 Var_{P_0}(H_1(x))$
2. $\sum_{k=2}^m \binom{m}{k}(P_n - P_0)^k H_k = o_P(n^{-1/2})$

So it follows that V_n is an AL estimator for V_0 with influence function $x \mapsto m[H_1(x) - V_0]$.

Remark 2.2. *V_n is generally a biased estimator for V_0.*

Definition 2.5: U-statistics

An alternative estimator for V_0, is called the U-statistic and takes the form

$$U_n = \binom{n}{m}^{-1} \sum_{\bar{i}_m \in D_{m,n}} H(x_{i_1}, ..., x_{i_m})$$

where $D_{m,n} := \{\bar{i}_m := (i_1, ..., i_m) : 1 \leq i_1 < ..., i_m \leq n\}$

Remark 2.3. *U_n and V_n are asymptotically equivalent when there is no degeneracy, i.e. $\tau_1^2 := m^2 Var_{P_0}(H_1(x)) > 0$*

Lemma 2.6: Finite sample variance of a U-statistic

$Var(U_n) = \sum_k^m \binom{n}{m}^{-1} \binom{m}{k} \binom{n-m}{m-k} \tau_k^2$ where $\tau_k^2 := Var(H_k(x_1, ..., x_k))$

Theorem 2.7: Asymptotic distribution of 1-degenerate U- and V-statistics

If H is a symmetric kernel with $m \geq 2$ and $\tau_2^2 > \tau_1^2 = 0$, then

$$n(U_n - V_0) \rightsquigarrow \sum_k^\infty \lambda_k (Z_k^2 - 1)$$

where $Z_1, ... \sim N(0, 1)$ and $\lambda_1, ...$ are the eigenvalues of a certain linear operator.

Under regularity conditions,

$$n(V_n - V_0) \rightsquigarrow \sum_k^\infty \lambda_k Z_k^2$$

2.3 Functional Differentiation

Goal: Derive an ALE for functional, $\Psi(P_0)$, by defining some version of the delta method that works for functionals.

Definition 2.8: ρ-continuous

The functional $\Psi : P \mapsto \mathbb{R}$ is **ρ-continuous** if $\forall$ sequences $\{F_1, F_2, ...\} \subseteq P$ such that $\rho(\tilde{F_k} - F) \to^{k\to\infty} 0$ for some F,

$$\Psi(F_n) \to^{k\to\infty} \Psi(F_0)$$

Theorem 2.9: CMT for functionals

If Ψ is ρ-continuous at F_0 and $\rho(F_n - F_0) \to^p 0$, then $\Psi(F_n) \to^p \Psi(F_0)$

Definition 2.10: Gateaux differentiability

The functional $\Psi : P \mapsto \mathbb{R}$ is **Gateaux differentiable** at F_0 in the direction $h \in Q(F_0) := \{c(F - F_0) : c \in \mathbb{R}, F \in P\}$ if both hold:

1. the differential $\dot{\Psi}(F_0, h) := \frac{d}{d\epsilon}\Psi(F_0 + \epsilon h)|_{\epsilon=0}$ is well-defined
2. $h \mapsto \dot{\Psi}(F_0, h)$ is linear (in h)

Remark 2.4. *The lack of uniformity over h of the limit defining Gateaux differentiability means that this notion of differentiability is too weak to prove a functional version of the delta method. However, by applying a third condition, as below, we get a strong enough notion of differentiability.*

Definition 2.11: Hadamard differentiability

The functional Ψ is called Hadamard differentiable relative to ρ on $Q(F_0)$ if both hold:

1. Ψ is Gateaux differentiable at F_0
2. Let $R_{F_0,\epsilon}(h) := \frac{\Psi(F_0+\epsilon h)-\Psi(F_0)}{\epsilon} - \dot{\Psi}(F_0, h)$. Then for any sequence $\epsilon_n \to 0$ and $\{h_1, ...\} \subseteq Q(F_0)$ such that $\rho(h_n, h) \to^{n\to\infty} 0$ for some $h \in Q(F_0)$ and $F_0 + \epsilon_n h_n \in P$ for all n,

   $$lim_{n\to\infty} R_{F_0,\epsilon_n}(h_n) = 0$$

   or, equivalently, for all compact subsets H of $Q(F_0)$,

   $$sup_{h\in H}|R_{F_0,\epsilon}(h)| \to^{\epsilon\to 0} 0$$

Remark 2.5. *Often we take $\rho(h_1, h_2) := sup_x|h_1(x) - h_2(x)|$*

Theorem 2.12: Functional delta method

If Ψ is Hadamard differentiable at F_0 relative to $\rho = ||\cdot||_\infty$, then

$$\Psi(F_n) - \Psi(F_0) = \dot{\Psi}(F_0; F_n - F_0) + o_P(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}(F_0; \mathbb{I}\{x_i \leq \cdot\} - F_0(\cdot)) + o_P(n^{-1/2})$$

or, in other words, $\Psi(F_n)$ is an ALE for $\Psi(F_0)$ with influence function

$$x \mapsto \dot{\Psi}(F_0; \mathbb{I}\{x \leq \cdot\} - F_0(\cdot))$$

Note: $\int g(x) d\mathbb{I}\{X_i \leq \cdot\} = g(X_i)$

Theorem 2.13: Integration by parts

Let $g : [a, b] \mapsto \mathbb{R}$ and $h : [a, b] \mapsto \mathbb{R}$ be cadlag functions (continuous from the right with limits from the left, e.g. CDF) with bounded variation. Then

$$\int_{(a,b]} g(u)dh(u) + \int_{(a,b]} h(u-)dg(u) = g(b)h(b) - g(a)h(a)$$

where $h(u-) = lim_{v \to u-} h(v)$

If at least one of the 2 functions is continuous then,

$$\int_{(a,b]} g(u)dh(u) + \int_{(a,b]} h(u)dg(u) = g(b)h(b) - g(a)h(a)$$

3 Chapter 3

Suppose $X_1, .., X_n \sim P_0 \subseteq \mathscr{M}$ and consider the functional, $\Psi : \mathscr{M} \mapsto \mathbb{R}$.
Goal: Find a lower bound on the asymptotic variance, $v_0^*(\mathscr{M})$, for an estimator of $\Psi_0 := \Psi(P_0)$.

Definition 3.1: Index set of submodels

Let $H(P_0)$ is an index set for the collection of all smooth (QMD) 1-dimensional parametric submodels of $\mathscr{M}$ centered at P_0.

3.1 Tangent Spaces

Definition 3.2: Tangent sets and spaces

- $G(P_0) := \{g_h : H(P_0)\}$ is called the **tangent set**, where g_h is the score function associated with submodel, h, evaluated at $\theta = 0$.
  - Note that this implies $T_{\mathscr{M}}(P_0) \subseteq L_0^2(P_0) := \{g \in L^2(P_0) : P_0 g = 0\}$. When $T_{\mathscr{M}}(P_0) = L_0^2(P_0)$, we call $\mathscr{M}$ **locally non-parametric** at P_0.
- The **tangent space**, $T_{\mathscr{M}}(P_0)$, is the $L^2(P_0)$ closure of the linear span of $G(P_0)$. In this class, we will assume $G(P_0) = T_{\mathscr{M}}(P_0)$.

Lemma 3.3: Orthogonality of tangent space

Suppose that a model $\mathscr{M}$ for the joint distribution $X = (Z, Y)$ can be represented as $\mathscr{M}_Z \otimes \mathscr{M}_{Y|Z} = \{(P_Z, P_{Y|Z}) : P_Z \in \mathscr{M}_Z, P_{Y|Z} \in \mathscr{M}_{Y|Z}\}$. Then the tangent space of $\mathscr{M}$ at $P_0 \in \mathscr{M}$ is given by

$$T_{\mathscr{M}}(P_0) = T_{\mathscr{M}_Z}(P_0) \otimes T_{\mathscr{M}_{Y|Z}}(P_0)$$

where $T_{\mathscr{M}_Z}(P_0)$ is the tangent space generated by scores for $P_{Z,0}$ and $T_{\mathscr{M}_{Y|Z}}(P_0)$ is the tangent space generated by scores for $P_{Y|Z,0}$

3.2 Gradients

Definition 3.4: Pathwise differentiability and gradients

A parameter Ψ is called **pathwise differentiable** if there exists some P_0-mean zero and square integrable function $D(P_0)$ s.t., $\forall h \in H(P_0)$,

$$\frac{\partial}{\partial \theta} \Psi(P_{\theta,h})|_{\theta=0} = P_0[D(P_0) g_h]$$

- We call such a $D(P_0)$ a **gradient** of Ψ at P_0 relative to $\mathscr{M}$, and if $D(P_0) \in T_{\mathscr{M}}(P_0)$ we call it the **canonical gradient** of Ψ at P_0 relative to $\mathscr{M}$, and denote it $D^*(P_0)$.

Equivalently, a parameter $\Psi : \mathscr{M} \mapsto \mathbb{R}$ is **pathwise differentiable** at P_0 iff there exists a continuous linear map $\dot{\Psi}_{P_0} : L_0^2(P_0) \mapsto \mathbb{R}$ such that for all $h \in H(P_0)$, $\frac{\partial}{\partial \theta} \Psi(P_{\theta,h})|_{\theta=0} = \dot{\Psi}_{P_0}(g_h)$

Lemma 3.5: Relationship between gradients

1. Let $D(P_0)$ be a gradient of Ψ at P_0 relative to $\mathscr{M}$. Then for any $q(P_0) \in T_{\mathscr{M}}(P_0)^{\perp}$, $D(P_0) + q(P_0)$ is also a gradient.
2. Let $\mathscr{M}_0 \subseteq \mathscr{M}$ be two models and take $P \in \mathscr{M}$. Suppose $\Psi : \mathscr{M}_0 \mapsto \mathbb{R}$ is pathwise differentiable at P relative to $\mathscr{M}_0$. Then Ψ is also pathwise differentiable at P relative to $\mathscr{M}$ and $Grad_{\mathscr{M}}(P) \subseteq Grad_{\mathscr{M}_0}$.

We can use this lemma to find a gradient for Ψ at P_0 relative to $\mathscr{M}$ with the following steps

1. Take a QMD parametric submodel $\{P_\theta : \theta \in [0, \delta)\} \subseteq \mathscr{M}$ with $P_{\theta=0} = P_0$ and score function $g \in T_{\mathscr{M}}(P_0)$
2. Compute $\frac{\partial}{\partial\theta}\Psi(P_\theta)|_{\theta=0}$ and write as $P_0[\tilde{D}(P_0)g]$ for some $\tilde{D}(P_0) \in L^2(P_0)$
3. Recenter $\tilde{D}(P_0)$ to obtain the gradient, $D(P_0) := x \mapsto \tilde{D}(P_0)(x) - P_0\tilde{D}(P_0)$

If we want to show that this gradient is the canonical gradient, we need to show that $D(P_0) \in T_{\mathscr{M}}(P_0)$.

- Find a QMD parametric submodel $\{P_\theta : \theta \in [0, \delta)\} \subseteq \mathscr{M}$ with $P_{\theta=0} = P_0$ which has score function $D(P_0)$

If it is not the canonical gradient, we can project $D(P_0)$ onto $T_{\mathscr{M}}(P_0)$, to obtain the canonical gradient

- Find $D^*(P) \in T_{\mathscr{M}}(P_0)$ such that $P_0[\{D(P_0) - D^*(P_0)\}g_h] = 0$ for all $h \in H(P_0)$

3.3 Derivation of generalized Cramer-Rao (GCR bound)

Equipped with the results from this chapter, we can finally give a a lower bound for the asymptotic variance.

Observe $v_0^*(\mathscr{M}) \geq v_0(\mathscr{M}_h)$ for any $h \in H(P_0)$, so then it follows that,

$$v_0^*(\mathscr{M}) \geq sup_{h\in H(P_0)} v_0(\mathscr{M}_h) = sup_{h\in H(P_0)} \frac{[\frac{\partial}{\partial\theta}\Psi(P_{\theta,h})|_{\theta=0}]^2}{P_0 g_h^2}$$

so if Ψ is pathwise differentiable at P_0 relative to $\mathscr{M}$ with some gradient, $D(P_0)$, then

$$v_0^*(\mathscr{M}) \geq sup_{g\in T_{\mathscr{M}}(P_0)} \frac{(P_0[D(P_0)g])^2}{P_0 g^2}$$

Then if we replace $D(P_0)$ with the canonical gradient, $D^*(P_0)$ (by projecting it onto $T_{\mathscr{M}}((P_0)$, then by Cauchy-Schwartz $(P_0[D^*(P_0)g])^2 = P_0(D^*(P_0)^2)P_0 g^2$, so

$$v_0^*(\mathscr{M}) \geq P_0[D^*(P_0)^2]$$

3.4 Gradients and influence functions

Lemma 3.6: Influence functions as gradients*

If Ψ_n is an ALE of $\Psi(P_0)$ with influence function ϕ_{P_0}, then

$$\psi_n \text{ is regular at } P_0 \iff \Psi \text{ is path differentiable at } P_0 \text{ and } \phi_{P_0} \text{ is a gradient of } \Psi \text{ at } P_0$$

Implication: To construct a RALE of $\Psi(P_0)$, Ψ must be pathwise differentiable at P_0 relative to $\mathscr{M}$.

Lemma 3.7: Gradients as influence functions

Under regularity conditions, for a given gradient $D(P_0)$ of a pathwise differentiable parameter Ψ,

An ALE with influence function $D(P_0)$ exists $\iff$ it is possible to estimate $D(P_0)$ in an appropriate, locally uniform sense

These two lemmas taken together with the Central Limit Theorem show that the lower bound of the asymptotic variance for model $\mathscr{M}$, $P_0[D^*(P_0)^2]$, is achievable.

Definition 3.8: Efficient influence functions

A RALE Ψ_n of $\Psi(P_0)$ is called **asymptotically efficient** if its influence function is $D^*(P_0)$, the canonical gradient of Ψ at P_0, which we also call the **efficient influence function**.

4 Chapter 4

In this chapter we focus on estimating statistical parameters with local features (density, regression function, etc.). In these cases, the bias of plug-in estimators are so large that we need to undersmooth (choose a sub-optimal bandwidth) for the bias to be $o_P(n^{-1/2})$. This means finding an ALE is not as straightforward. Below are two methods for obtaining an ALE in these situations.

4.1 Estimating equations framework

Definition 4.1: Estimating equation with nuisance parameter

$U(\psi, \eta)$ is an estimating equation for ψ_0 if

$$P_0 U(\psi, \eta_0) = 0 \iff \psi = \psi_0$$

In Chapter 2, we showed that if $U(\psi, \eta)(x)$ is an estimating function for ψ_0, then the estimating equations-based estimator ψ_n of ψ_0 using an estimator η_n of η_0 is asymptotically linear with influence function:

$$-a_0^{-1}[P_n(U(\psi_0, \eta_0) + b_0(\eta_n - \eta_0))]$$

where $a_0 = \frac{\partial}{\partial\psi} P_0 U(\psi, \eta_0)|_{\psi=\psi_0}$ and $b_0 = \frac{\partial}{\partial\eta} P_0 U(\psi_0, \eta)|_{\eta=\eta_0}$

Then in this chapter we showed that when the estimating equation is a gradient of ψ at P_0, $a_0 = -1$ and $b_0 = 0$, so the estimating equations-based estimator ψ_n of ψ_0 is asymptotically linear with influence function:

$$P_n(U(\psi_0, \eta_0)) = P_n D(P_0)$$

Then the resulting ALE for $\Psi(P_0)$ is given the solution in ψ to $P_n D(\psi, \eta_n) = 0$.

We can use this strategy for any gradient, but it is particularly useful when use the canonical gradient (i.e. the EIF) because this gives an asymptotically efficient estimator.

Note: We can only use this method when the gradient depends on ψ_0, because otherwise it is not an estimating equation!

4.2 One-step estimation

As previously established, the plug-in estimator will generally have bias that is too large for it to be asymptotically linear. However, we can characterize this bias and use this to perform a correction to obtain an ALE.

Specifically, we can write

$$\Psi(\hat{P}_n) - \Psi(P_0) = (\hat{P}_n - P_0)D(\hat{P}_n) + R(\hat{P}_n, P_0) = -P_0 D(\hat{P}_n) + R(\hat{P}_n, P_0)$$

for a gradient $D(\hat{P}_n)$ of Ψ at $\hat{P}_n$ in $\mathscr{M}$ and a remainder term satisfying $\frac{R(\hat{P}_n, P_0)}{d(\hat{P}_n, P_0)} \to 0$ as $d(\hat{P}_n, P_0) \to 0$ for some discrepancy d on $\mathscr{M}$.
We can then expand the right handside by adding and subtracting $P_n D(\hat{P}_n)$ and $(P_n - P_0)D(P_0)$ to obtain

$$\Psi(\hat{P}_n) - \Psi(P_0) = (P_n - P_0)D(P_0) + (P_n - P_0)[D(\hat{P}_n) - D(P_0)] + R(\hat{P}_n, P_0) - P_n D(\hat{P}_n)$$

which implies

$$\Psi(\hat{P}_n) + P_n D(\hat{P}_n) - \Psi(P_0) = (P_n - P_0)D(P_0) + o_P(n^{-1/2})$$

if the following conditions are satisfied:

- $R(\hat{P}_n, P_0) = o_P(n^{-1/2})$
- $P_0[D(\hat{P}_n) - D(P_0)]^2 = o_P(1)$
- there is a fixed P_0-Donsker class $\mathscr{F}$ such that $D(\hat{P}_n) \in \mathscr{F}$ with probability tending to 1

5 Appendix

Properties of Hilbert Spaces: Let $H_0 \subseteq H$ be a subspace of Hilbert space, H

- The orthogonal complement of H_0 is $H_0^\perp := \{h \in H : \langle h, h_0 \rangle = 0 \quad \forall h_0 \in H_0\}$
- The projection of $h_1 \in H$ onto a closed subspace of H_0 is the element of $h_0 \in H_0$ s.t. $h_1 - h_0 \in H_0^\perp$

Jensen's Inequality: For RV X and convex function f, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

Cauchy-Schwartz Inequality: $(\int P_1(w)P_2(w))^2 \leq (\int P_1^2(w)dw)(\int P_2^2(w)dw)$

Layer Cake representation: If Z is a non-negative RV, then $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t)dt$

Lipschitz Continuity: A function $f : X \mapsto \mathbb{R}$ is L-Lipschitz continuous if $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$ for all $x_1, x_2 \in X$