

Political Science 787: Multivariate Analysis

Fall 2008

Monday 11–1 (5664 HH)

Professor: Walter R. Mebane, Jr.

Office: 7735 Haven Hall (734/763-2220); email

wmebane@umich.edu

Office hours: Tue 2–4 or other times by appointment.

Course web page: <http://www.umich.edu/~wmebane/ps787.html>

asymptotic arguments

- **m-estimation**
 - maximum likelihood (ML) estimation
- **misspecified models (quasi maximum likelihood estimation)**
 - correctly specified models (ML estimators)
- **quasi-likelihood**

two convergence concepts

- $(\mathcal{X}, \mathcal{A}, P)$ is a probability space
- $\{X_n\}$ is a sequence of random variables for $n = 1, 2, \dots$
- X is a random variable
- (almost sure convergence) $\{X_n\}$ converges to X almost surely (a.s.) if

$$P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1$$

- $(\mathcal{X}_n, \mathcal{A}_n, P_n)$ is a probability space for each $n = 1, 2, \dots$
- (convergence in probability, or weak convergence) $\{X_n\}$ converges to X in probability if for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P_n(|X_n - X| > \epsilon) = 0$$

m-estimation (slightly specializing Huber 1967):

Setting:

- $(\mathcal{X}, \mathcal{A}, P)$ is a probability space
- Θ is an open subset of m -dimensional Euclidean space (\mathbb{R}^m)
- $\psi(x, \theta) : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^m$ is a function
- x_1, x_2, \dots are independent random variables with values in \mathcal{X} having common distribution P
- for each $n = 1, 2, \dots$, let $T_n : \mathcal{X}^n \rightarrow \Theta$ be a function

Task: give sufficient conditions that any sequence T_n such that

$$\frac{1}{n} \sum_{i=1}^n \psi(x_i; T_n) \rightarrow 0 \quad (1)$$

almost surely converges almost surely to some constant θ_0 (or in probability converges in probability to some constant θ_0)

Example

- let $f(x, \theta)$ be a differentiable parametric family of probability densities, $dP = f(x, \theta)d\mu$ (μ is Lebesgue [rectangle volumes] measure)
- if $\psi(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta)$, then the ML estimator satisfies equation (1):

$$\frac{1}{n} \sum_{i=1}^n \psi(x_i; T_n) \rightarrow 0$$

Example: Poisson regression model for mean function $\mu(x)$:

$$f(x, \theta) = \mathbf{Prob}[Y = y|x] = \frac{e^{-\mu(x)} \mu(x)^y}{y!}, \quad y = 0, 1, \dots$$

- **mean function:** $\mu(x_i) = \exp(x_i' \beta), \quad \theta = \beta$
- **(concentrated) log likelihood:**

$$\log f_i = y_i \log \mu(x_i) - \mu(x_i) = y_i x_i' \beta - \exp(x_i' \beta)$$

- **score:**

$$s_i = \frac{\partial}{\partial \theta} \log f(x_i, \theta) = y_i x_i - \exp(x_i' \beta) x_i$$

- **hessian:**

$$H_i = \frac{\partial}{\partial \theta'} \frac{\partial}{\partial \theta} \log f(x_i, \theta) = -\exp(x_i' \beta) x_i x_i'$$

Definition: separable (as used by Huber; technical)

- let N be a P -null set
- let $\Theta' \subset \Theta$ be a countable subset
- ψ is separable if N and Θ' exist such that for every open set $U \subset \Theta$ and every closed interval A , the sets

$$\{x | \psi(x, \theta) \in A, \forall \theta \in U\}, \quad \{x | \psi(x, \theta) \in A, \forall \theta \in U \cap \Theta'\}$$

differ by at most a subset of N

- ψ is determined (in probability) by its values on a countable dense subset
- assuming this separability ensures that various limits, infima and suprema are measurable

Assumptions:

1. for each fixed $\theta \in \Theta$, $\psi(x, \theta)$ is \mathcal{A} -measurable in x , and ψ is separable
2. ψ is a.s. continuous in θ :

$$\lim_{\theta' \rightarrow \theta} |\psi(x, \theta') - \psi(x, \theta)| = 0 \quad a.s.$$

3. the expected value $\lambda(\theta) = E\psi(x, \theta) = \int \psi(x, \theta)dP(x)$ exists for all $\theta \in \Theta$, and there is a unique value $\theta = \theta_0$ such that $\lambda(\theta_0) = 0$

4. there is a continuous function $b(\theta)$ that is bounded away from zero, $b(\theta) \geq b_0 > 0$, such that

(a) $\sup_{\theta} \frac{|\psi(x, \theta)|}{b(\theta)}$ is integrable

(b) $\liminf_{\theta \rightarrow \infty} \frac{|\lambda(\theta)|}{b(\theta)} \geq 1$

(c) $E \left[\limsup_{\theta \rightarrow \infty} \frac{|\psi(x, \theta) - \lambda(\theta)|}{b(\theta)} \right] < 1$

assumption 4(c) allows assumption 2 to be strengthened to

5. as the neighborhood U of θ shrinks to $\{\theta\}$,

$$E \left[\sup_{\theta' \in U} |\psi(x, \theta) - \lambda(\theta)| \right] \rightarrow 0$$

assumption 5 implies that λ is continuous. if a function b exists that satisfies Assumption 4, then we can use

$$b(\theta) = \max(|\lambda(\theta)|, b_0)$$

m-estimator consistency theorem

- **Lemma m.1:** if assumptions 1 and 4 hold, then there is a compact set $C \subset \Omega$ such that any sequence T_n satisfying equation (1) a.s. (or in probability) ultimately stays in C
- **Theorem m.1:** if assumptions 1, 3 and 5 hold, then every sequence T_n satisfying equation (1) and Lemma m.1 converges to θ_0 almost surely (in probability)

proof of Lemma m.1: Huber (1967) uses two general results from probability theory

- **dominated convergence theorem: if Y is integrable and $|X_n| \leq Y$ almost everywhere, and if $X_n \rightarrow X$ almost surely (or in probability), then $E(|X_n - X|) \rightarrow 0$ uniformly in the relevant measurable set (Loève 1977, 126)**
- **strong law of large numbers: if the sequence $b_n \uparrow \infty$ and the series $\sum \frac{E|X_n|}{b_n} < \infty$, then $\frac{1}{b_n} \sum_{k=1}^n (X_k - EX_k) \rightarrow 0$ (Loève 1977, 253)**

for the proof see Huber (1967, 225)

m-estimator asymptotic normality

Setting:

- $(\mathcal{X}, \mathcal{A}, P)$ is a probability space
- Θ is an open subset of m -dimensional Euclidean space (\mathbb{R}^m)
- $\psi(x, \theta) : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^m$ is a function
- x_1, x_2, \dots are independent random variables with values in \mathcal{X} having common distribution P
- for each $n = 1, 2, \dots$, let $T_n : \mathcal{X}^n \rightarrow \Theta$ be a function

Task: give sufficient conditions that every sequence T_n that satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i; T_n) \rightarrow 0 \quad (2)$$

in probability is asymptotically normal (assume T_n is consistent)

Assumptions

1. for each fixed $\theta \in \Theta$, $\psi(x, \theta)$ is \mathcal{A} -measurable in x , and ψ is separable.

define

$$\lambda(\theta) = E\psi(x, \theta)$$

$$u(x, \theta, d) = \sup_{|\tau - \theta| \leq d} |\psi(x, \tau) - \psi(x, \theta)|$$

always take expectations with respect to the true P

2. there is a θ_0 such that $\lambda(\theta_0) = 0$

3. for $|\theta|$ denoting any norm equivalent to the Euclidean norm, there are strictly positive numbers a, b, c, d_0 such that

(a) $|\lambda(\theta)| \geq a|\theta - \theta_0|$, for $|\theta - \theta_0| \leq d_0$

(b) $Eu(x, \theta, d) \leq bd$, for $|\theta - \theta_0| + d \leq d_0$

(c) $E[u(x, \theta, d)^2] \leq cd$, for $|\theta - \theta_0| + d \leq d_0$

4. $0 < E(|\psi(x, \theta_0)|^2) < \infty$

m-estimator asymptotic normality theorem

- **Theorem m.2: if assumptions 1 to 4 hold and T_n satisfies equation (2), and if $P(|T_n - \theta_0| \leq d_0) \rightarrow 1$, then**

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i, \theta_0) + \sqrt{n} \lambda(T_n) \rightarrow 0$$

in probability

- **the conclusion of Theorem m.2 states a condition sufficient for the Lindeberg-Feller central limit theorem to hold**

m-estimator asymptotic normality theorem

- **assumption Λ** : the expectation λ has a nonsingular derivative matrix Λ at θ_0 : i.e., $|\lambda(\theta) - \lambda(\theta_0) - \Lambda \cdot (\theta - \theta_0)| = o(|\theta - \theta_0|)$
- **Corollary m.2**: if the assumptions of Theorem m.2 and assumption Λ hold, then $\sqrt{n}(T_n - \theta_0)$ is asymptotically normal with mean 0 and covariance matrix $\Lambda^{-1}C(\Lambda')^{-1}$, where C is the covariance matrix of $\psi(x, \theta_0)$
- $-\Lambda$ is the observed information (a.k.a. the Hessian matrix)

m-estimator asymptotic normality theorem: ML special case

- **assume** $dP = f(x, \theta)d\mu$ **and** $\psi(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta)$
- **let assumptions 1, 3 and 4 hold locally uniformly in θ_0**
- **assume the ML estimator that satisfies $\frac{1}{n} \sum_{i=1}^n \psi(x_i; T_n) \rightarrow 0$ in probability is consistent**
- **assume the Fisher information matrix**

$$I(\theta) = \int \psi(x, \theta)\psi(x, \theta)' f(x, \theta)d\mu$$

is continuous at θ_0

- **Proposition m.3: under the stated assumptions, we have $\lambda(\theta_0) = 0$, $\Lambda = -C = -I(\theta_0)$ and, in particular, $\Lambda^{-1}C(\Lambda')^{-1} = I(\theta_0)^{-1}$**
- **$I(\theta_0)$ is the expected information (a.k.a. the OPG)**

m-estimator asymptotic normality theorem: ML special case

- **assume** $dP = f(x, \theta)d\mu$ **and** $\psi(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta)$
- **the expectation** λ **has a nonsingular derivative matrix** Λ **at** θ_o : **i.e.**, $|\lambda(\theta) - \lambda(\theta_o) - \Lambda \cdot (\theta - \theta_o)| = o(|\theta - \theta_o|)$ (**observed information**)
- **Fisher information matrix (expected information):**

$$I(\theta) = \int \psi(x, \theta)\psi(x, \theta)' f(x, \theta)d\mu$$

is continuous at θ_o

- $-\Lambda = I(\theta_o)$ **states the information matrix equality**

misspecified models: quasi-maximum likelihood estimator (QMLE) (White 1994)

- **Assumption 2.1:** The observed data are a realization of a stochastic process $X = \{X_t : \Omega \rightarrow \mathbb{R}^\nu, \nu \in \mathbb{N}, t = 1, 2, \dots\}$ on a complete probability space $(\Omega, \mathcal{F}, P_o)$, where $\Omega = \mathbb{R}^{\nu\infty} \equiv \times_{t=1}^{\infty} \mathbb{R}^\nu$ and $\mathcal{F} = \mathcal{B}^{\nu\infty} \equiv \mathcal{B}(\mathbb{R}^{\nu\infty})$.
- **Assumption 2.3:** The functions $f_t : \mathbb{R}^{\nu t} \times \Theta \rightarrow \mathbb{R}^+$ are such that $f_t(\cdot, \theta)$ is measurable- $\mathcal{B}^{\nu t}$ for each θ in Θ , a compact subset of \mathbb{R}^p , $p \in \mathbb{N}$, and $f_t(X^t, \cdot)$ is continuous on Θ a.s.- P_o , i.e. $f_t(x^t, \cdot)$ is continuous on Θ for all x^t in some $F_t \in \mathcal{B}^{\nu t}$, $P_o^t[F_t] = 1$, $t = 1, 2, \dots$.
- **Assumption 3.1:** (a) For each θ in Θ , $E(\log f_t(X^t, \theta))$ exists and is finite, $t = 1, 2, \dots$; (b) $E(\log f_t(X^t, \cdot))$ is continuous on Θ , $t = 1, 2, \dots$; and (c) $\{\log f_t(X^t, \theta)\}$ obeys the strong (weak) uniform law of large numbers

misspecified models: QMLE (White 1994)

- **Definition 3.3: (Identifiable Uniqueness):** Let $\bar{Q}_n : \Theta \rightarrow \bar{\mathbb{R}}$ be continuous on Θ , a compact subset of \mathbb{R}^p , $p \in \mathbb{N}$, and let Θ_n be a nonempty compact subset of Θ , $n = 1, 2, \dots$. Suppose that $\bar{Q}_n(\theta)$ has a maximum on Θ_n at θ_n^* , $n = 1, 2, \dots$. Let $\mathcal{S}_n(\varepsilon)$ be an open sphere in \mathbb{R}^p centered at θ_n^* with fixed radius $\varepsilon > 0$. For each $n = 1, 2, \dots$, define the neighborhood $\eta_n(\varepsilon) = \mathcal{S}_n(\varepsilon) \cap \Theta_n$ with compact complement $\eta_n^c(\varepsilon)$ in Θ_n . The sequence of maximizers $\theta^* \equiv \{\theta_n^*\}$ is said to be **identifiably unique on $\{\Theta_n\}$** if either for all $\varepsilon > 0$ and all n , $\eta_n^c(\varepsilon)$ is empty, or for all $\varepsilon > 0$

$$\limsup_{n \rightarrow \infty} \left[\max_{\theta \in \eta_n^c(\varepsilon)} \bar{Q}_n(\theta) - \bar{Q}_n(\theta_n^*) \right] < 0, .$$

misspecified models: QMLE (White 1994)

- **Theorem 3.4:** Let (Ω, \mathcal{F}, P) be a complete probability space, let Θ be a compact subset of \mathbb{R}^p , $p \in \mathbb{N}$, and let $\{\Theta_n\}$ be a sequence of compact subsets of Θ . Let $\{Q_n\}$ be a sequence of random functions continuous on Θ a.s.- P and let $\hat{\theta}_n = \operatorname{argmax}_{\Theta_n} Q_n(\cdot, \theta)$ a.s.- P . If $Q_n(\cdot, \theta) - \bar{Q}_n(\theta) \rightarrow 0$ as $n \rightarrow \infty$ a.s.- P (prob- P) uniformly on Θ and if $\{\bar{Q}_n : \Theta \rightarrow \bar{\mathbb{R}}\}$ has identifiably unique maximizers θ^* on $\{\Theta_n\}$, then $\hat{\theta}_n - \theta_n^* \rightarrow 0$ as $n \rightarrow \infty$ a.s.- P (prob- P).
- **Definition 2.4: (Correctly Specified Probability Model):** The probability model \mathcal{P} is correctly specified for X if \mathcal{P} contains P_o , the data generating mechanism of Assumption 2.1. Otherwise, \mathcal{P} is misspecified for X .

misspecified models: QMLE consistency (White 1994)

- $\bar{L}_n(\theta) \equiv E(n^{-1} \sum_{t=1}^n \log f_t(X^t, \theta))$
- **Assumption 3.2:** $\{\bar{L}_n\}$ has identifiably unique maximizers $\theta^* \equiv \{\theta_n^*\}$ on Θ .
- **Theorem 3.5:** Let Assumptions 2.1, 2.3, 3.1 and 3.2 hold, and let $\hat{\theta}$ be generated by $\mathcal{S} = \{f_t\}$. Then $\hat{\theta}_n - \theta_n^* \rightarrow 0$ as $n \rightarrow \infty$ a.s.- P_o (prob- P_o).

misspecified models: QMLE asymptotic normality (White 1994)

- **supposition B:** there exists a nonstochastic sequence of $p \times p$ matrices $\{B_n^*\}$ that is $O(1)$ and uniformly positive definite such that

$$B_n^{*-1/2} \sqrt{n} \nabla Q_n^* \rightarrow N(0, I_p),$$

where $\nabla Q_n^* \equiv \nabla Q_n(\cdot, \theta_n^*)$.

- **supposition A:** There exists a sequence $\{A_n : \Theta \rightarrow \mathbb{R}^{p \times p}\}$ such that $\{A_n\}$ is continuous on Θ uniformly in n , $\nabla^2 Q_n(\cdot, \theta) - A_n(\theta) \rightarrow 0$ as $n \rightarrow \infty$ prob- P uniformly on Θ and $\{A_n^* \equiv A_n(\theta_n^*)\}$ is $O(1)$ and uniformly nonsingular (i.e., $|\det A_n^*| > 0$ for almost all n).

misspecified models: QMLE asymptotic normality (White 1994)

- **Theorem 6.2:** Let (Ω, \mathcal{F}, P) be a complete probability space, let Θ be a compact subset of \mathbb{R}^p ($p \in \mathbb{N}$) with nonempty interior and let $Q_n : \Omega \times \Theta \rightarrow \mathbb{R}$ be a random function continuously differentiable of order 2 on Θ a.s.- P , $n = 1, 2, \dots$. Let $\hat{\theta}_n : \Omega \rightarrow \Theta$ be measurable- \mathcal{F} , $n = 1, 2, \dots$, such that $\hat{\theta}_n = \operatorname{argmax}_{\Theta} Q_n(\cdot, \theta)$ a.s.- P and $\hat{\theta}_n - \theta_n^* \rightarrow 0$ as $n \rightarrow \infty$ **prob- P** , where $\{\theta_n^*\}$ is interior to Θ uniformly in n . **Assume suppositions B and A. Then**

$$\sqrt{n}(\hat{\theta}_n^* - \theta_n^*) = -A_n^{*-1} \sqrt{n} \nabla Q_n^* + o_P(1)$$

$$B_n^{*-1/2} A_n^* \sqrt{n}(\hat{\theta}_n^* - \theta_n^*) \rightarrow N(0, I_p).$$

misspecified models: QMLE asymptotic normality (White 1994)

- applicability to of Theorem 6.2 to QMLE: choose $Q_n(\omega, \theta) = L_n(X^n(\omega), \theta)$, where the log-quasilielihood L_n satisfies regularity conditions.
- $s_t^* \equiv \nabla \log f_t(X^t, \theta_n^*)$
- **Theorem 6.4 (QMLE Asymptotic Normality):** Given Assumptions 2.1, 2.3, 3.1, 3.2', 3.6, 3.7(a), 3.8, 3.9 and 6.1,

$$\sqrt{n}(\hat{\theta}_n^* - \theta_n^*) = -A_n^{*-1} \sqrt{n} \nabla L_n^* + o_{P_o}(1)$$

and

$$B_n^{*-1/2} A_n^* \sqrt{n}(\hat{\theta}_n^* - \theta_n^*) \rightarrow N(0, I_p)$$

where $A_n^* \equiv \nabla^2 \bar{L}_n(\theta_n^*) = E(\nabla^2 L_n^*)$ and

$B_n^* \equiv \text{var}[n^{-1/2} \sum_{t=1}^n s_t^*]$, so that $\text{avar} \theta_n^* = A_n^{*-1} B_n^* A_n^{*-1}$

misspecified models: QMLE asymptotic normality with correct specification (White 1994)

- $A_n^o \equiv E(n^{-1} \nabla^2 \log f^n(X^n, \theta_o))$
- $B_n^o \equiv \mathbf{var}(n^{-1/2} \nabla \log f^n(X^n, \theta_o))$
- **Theorem 6.5: (i) Given Assumptions 2.1-2.3, 3.1(a,b), 3.2, 3.4, 3.6 and 6.2, if the model specification is correct in its entirety at Θ_o which is in the interior of Θ , then $\theta_n^* = \theta_o$ and**

$$A_n^o = -B_n^o, \quad n = 1, 2, \dots$$

(ii) If Assumptions 3.1(c), 3.7(a), 3.8 and 6.1 also hold, then the conclusions of Theorem 6.4 hold with

$$\mathbf{avar} \hat{\theta}_n = -A_n^{o-1} = B_n^{o-1}$$

- **this is the information matrix equality**

misspecified models: information matrix equality test (White 1982)

- $d_{lt}(\theta) = \frac{\partial \log f_t}{\partial \theta_i} \frac{\partial \log f_t}{\partial \theta_j} + \frac{\partial^2 \log f_t}{\partial \theta_i \partial \theta_j}$, $l = 1, \dots, p(p+1)/2$,
 $i, j = 1, \dots, p$
- “indicators” (elements of $A + B$): $D_l(\hat{\theta}) = n^{-1} \sum_{t=1}^n d_{lt}(\hat{\theta})$
- let d_t denote a vector containing some subset of $q \leq p(p+1)/2$ of the d_{lt} values, and let $D(\hat{\theta}) = n^{-1} \sum_{t=1}^n d_t(\hat{\theta})$ denote the corresponding subset vector of “indicators”
- define $q \times q$ Jacobian matrices

$$\nabla D_n(\theta) = \left\{ n^{-1} \sum_{t=1}^n \frac{\partial d_{lt}(\theta)}{\partial \theta_k} \right\}$$
$$\nabla D(\theta) = \{E(\partial d_{lt}/\partial \theta_k)\}$$

misspecified models: information matrix equality test (White 1982)

- **variance estimator:**

$$V_n(\hat{\theta}) = n^{-1} \sum_{t=1}^n [d_t(\hat{\theta}) - \nabla D_n(\hat{\theta}) A_n(\hat{\theta})^{-1} \nabla f_t(\hat{\theta})] \cdot [d_t(\hat{\theta}) - \nabla D_n(\hat{\theta}) A_n(\hat{\theta})^{-1} \nabla f_t(\hat{\theta})]'$$

where $A_n(\hat{\theta})$ denotes the observed information

- **test statistic:**

$$\mathcal{J}_n = n D_n(\hat{\theta})' V_n(\hat{\theta})^{-1} D_n(\hat{\theta})$$

is distributed asymptotically as χ_q^2

profile likelihood

- let $\psi = \psi(\theta)$ be a subparameter (or a function of the parameter θ)
- the profile likelihood $L_P(\psi)$ for ψ is

$$L_P(\psi) = \max_{\theta|\psi} L(\theta)$$

- profile log-likelihood: $l_P = \log L_P$
- the maximum profile likelihood estimate of ψ equals $\hat{\psi}$ (the MLE)
- profile log-likelihood statistic tests $\psi = \psi_0$, i.e., for $\theta = (\psi, \chi)$,

$$2\{l_P(\hat{\psi}) - l_P(\psi)\} = 2\{\ell(\hat{\psi}, \hat{\chi}) - \ell(\psi_0, \hat{\chi}_{\psi_0})\}$$

- a profile likelihood region $\{l_P(\hat{\psi}) - l_P(\psi) < c\}$ is, generally, an approximate confidence region for ψ

higher-order asymptotic theory: Bartlett adjustment

- for random vector y , probability density $f(y; \omega)$ with parameter ω , a test of the null hypothesis $\omega = \omega_0$ may be based on the likelihood ratio $L(\hat{\omega})/L(\omega)$ or

$$w = 2\{l(\hat{\omega}) - l(\omega_0)\}$$

where $L(\omega)$ is the likelihood, $l(\omega)$ is the log-likelihood and $\hat{\omega}$ is the MLE

- with parameter partitioning $\omega = (\chi, \psi)$ with null hypothesis $\psi = \psi_0$ and nuisance parameter χ , the test statistic becomes

$$w = 2\{l(\hat{\chi}, \hat{\psi}) - l(\hat{\chi}_0, \psi_0)\}$$

where $\hat{\chi}_0$ is the profile MLE given $\psi = \psi_0$

- regularity conditions: as $n \rightarrow \infty$, w converges to χ_d^2 , the chi-squared distribution with d degrees of freedom, where d is the dimension of, respectively, ω_0 or ψ_0

higher-order asymptotic theory: Bartlett adjustment

- let $q_d(x)$ denote the density of χ_d^2
- if, under the null hypothesis,

$$E(w) = d\{1 + b/n + O(n^{-3/2})\}$$

where b is either constant or can be estimated consistently,
then

$$w' = (1 + b/n)^{-1}w$$

has an expected value closer to that of χ_d^2 than has w

- $(1 + b/n)^{-1}$ is the Bartlett adjustment factor
- covariance matrix proportionality example

higher-order asymptotic theory: Bartlett adjustment

- if the density of w is

$$\left(1 - \frac{1}{2}dbn^{-1}\right)q_d(x) + \frac{1}{2}dbn^{-1}q_{d+2}(x) + O(n^{-3/2})$$

then the density of $w' = (1 + b/n)^{-1}w$ is $q_d(x)$ with error $O(n^{-3/2})$

- that is, the density is

$$p(w'; \omega) = q_d(x) \{1 + O(n^{-3/2})\}$$

- the background theory here starts with

$$p(\hat{\omega}; \omega) = c|\hat{j}|^{1/2} \frac{L(\omega)}{L(\hat{\omega})} \{1 + O(n^{-3/2})\}$$

and involves integration over samples with respect to $\hat{\omega}$ conditioning on ω (see Barndorff-Nielsen and Cox 1984)

bootstrap

- some notation
- data: y_1, \dots, y_n are iid random variables Y_1, \dots, Y_n
- pdf is f and cdf is F
- population characteristic: θ
- statistic: T estimates θ , with sample value t
- empirical distribution: puts probability n^{-1} at each sample value y_j
- empirical distribution function (edf or empiric), \hat{F} :

$$\hat{F}(y) = \frac{\#\{y_j \leq y\}}{n}$$

bootstrap

- let $\hat{\theta}^*$ denote an estimate computed in a bootstrap resample
- plug-in principle: to find

$$\text{pr}(\hat{\theta} - \theta)$$

use

$$\text{pr}(\hat{\theta}^* - \hat{\theta})$$

bootstrap: bias estimation

- let $\theta = t(F)$ be a parameter
- let $\hat{\theta} = s(x)$ be a statistic
- bias: $E_F\{s(x)\} - t(F)$
- bootstrap estimate of bias: $E_{\hat{F}}\{s(x^*)\} - t(\hat{F})$
- Monte Carlo bootstrap estimation: using B independent bootstrap replications, x^{*1}, \dots, x^{*B} , evaluate $\hat{\theta}^*(b) = s(x^{*b})$ and compute the average

$$\hat{\theta}^*(\cdot) = \sum_{b=1}^B \hat{\theta}^*(b)/B = \sum_{b=1}^B s(x^{*b})/B$$

then

$$\widehat{\text{bias}}_B = \hat{\theta}^*(\cdot) - t(\hat{F})$$

pivots

- **a statistic is pivotal if its distribution does not depend on unknown parameters**
- **if the distribution of a statistic depends on unknown parameters, the statistic is nonpivotal**

bootstrap distribution estimates: general theory

- if U is a nonpivotal statistic with asymptotic variance σ^2 (e.g., $U = n^{1/2}(\hat{\theta} - \theta_0)$), then for some polynomial $p(x/\sigma)$

$$\begin{aligned} H(x) &= P(U \leq x) \\ &= \Phi(x/\sigma) + n^{-1/2}p(x/\sigma)\phi(x/\sigma) + O(n^{-1}) \end{aligned}$$

and the corresponding bootstrap distribution given the sample \mathcal{X} is

$$\begin{aligned} \hat{H}(x) &= P(U^* \leq x | \mathcal{X}) \\ &= \Phi(x/\hat{\sigma}) + n^{-1/2}\hat{p}(x/\hat{\sigma})\phi(x/\hat{\sigma}) + O_p(n^{-1}) \end{aligned}$$

- because $\hat{p} - p = O_p(n^{-1/2})$ and $\hat{\sigma} - \sigma = O_p(n^{-1/2})$,

$$\hat{H}(x) - H(x) = \Phi(x/\hat{\sigma}) - \Phi(x/\sigma) + O_p(n^{-1})$$

and $\hat{\sigma} - \sigma = O_p(n^{-1/2})$ implies $\Phi(x/\hat{\sigma}) - \Phi(x/\sigma) = O_p(n^{-1/2})$

bootstrap distribution estimates: general theory

- if T is a pivotal statistic (e.g., $T = n^{1/2}(\hat{\theta} - \theta_0)/\hat{\sigma}$ where σ^2 is the asymptotic variance of $\hat{\theta}$), then for some polynomial $q(x)$

$$\begin{aligned} G(x) &= P(T \leq x) \\ &= \Phi(x) + n^{-1/2}q(x)\phi(x) + O(n^{-1}) \end{aligned}$$

and the corresponding bootstrap distribution given the sample \mathcal{X} is

$$\begin{aligned} \hat{G}(x) &= P(T^* \leq x | \mathcal{X}) \\ &= \Phi(x) + n^{-1/2}\hat{q}(x)\phi(x) + O_p(n^{-1}) \end{aligned}$$

- because $\hat{q} - q = O_p(n^{-1/2})$,

$$\hat{G}(x) - G(x) = O_p(n^{-1})$$

bootstrap distribution estimates: general theory

- bootstrapping the distribution of a pivotal statistic typically gives an error of size n^{-1} , while bootstrapping the distribution of a nonpivotal statistic typically gives an error of size $n^{-1/2}$
- the practical downside of bootstrapping a pivotal statistic is that generally that requires having an estimate of σ
- in some applications σ may be difficult to estimate in a stable way, or σ may be too large

bootstrap confidence intervals

- **equitailed $1 - 2\alpha$ confidence interval:** for R resamples or simulations,

$$t - (t_{((R+1)(1-\alpha))}^* - t), \quad t - (t_{((R+1)\alpha)}^* - t)$$

- **studentized bootstrap:** $Z^* = (T^* - t)/V^{*1/2}$ and

$$t - v^{1/2} z_{((R+1)(1-\alpha))}^*, \quad t - v^{1/2} z_{((R+1)\alpha)}^*$$

bootstrap confidence intervals: percentile intervals

- let \hat{G} be the cumulative distribution function of the bootstrap replications $\hat{\theta}^*$
- histogram (or order statistic) motivation for the percentile interval

$$[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1 - \alpha)]$$

- “backward” relative to plug-in principle
 - $\text{pr}(\hat{\theta}^* - \hat{\theta})$ to estimate $\text{pr}(\hat{\theta} - \theta)$
 - let $\hat{H}^{-1}(\alpha)$ denote the α -percentile of $\hat{\theta}^* - \hat{\theta}$
 - what interval is implied by inverting

$$\hat{H}^{-1}(\alpha) \leq \hat{\theta} - \theta \leq \hat{H}^{-1}(1 - \alpha)$$

bootstrap confidence intervals: BC_a intervals

- for intended coverage $1 - 2\alpha$,

$$BC_a : (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)})$$

where

$$\alpha_1 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})} \right)$$
$$\alpha_2 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})} \right)$$

using Φ to denote the standard normal CDF and $z^{(\alpha)}$ for the 100α th percentile point of the standard normal distribution

- the adjustment corrects for bias and skewness

bootstrap confidence intervals: BC_a intervals

- **bias correction**

$$\hat{z}_0 = \Phi^{-1} \left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B} \right)$$

- **jackknife estimate for the “acceleration”**

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}}$$

approximations to CDFs: Cornish-Fisher expansions (or Edgeworth)

- **can be used to explain how the bootstraps work and to prove their accuracy**
- **still no guarantees**

bootstrap confidence intervals: accuracy

- **first-order accurate confidence point $\hat{\theta}[\alpha]$:**

$$\text{Prob}(\theta \leq \hat{\theta}[\alpha]) = \alpha + O(n^{-1/2})$$

- **second-order accurate confidence point $\hat{\theta}[\alpha]$:**

$$\text{Prob}(\theta \leq \hat{\theta}[\alpha]) = \alpha + O(n^{-1})$$

- **first-order correct confidence point $\hat{\theta}[\alpha]$:**

$$\hat{\theta}[\alpha] = \hat{\theta}_{\text{exact}}[\alpha] + O(n^{-1})$$

- **second-order correct confidence point $\hat{\theta}[\alpha]$:**

$$\hat{\theta}[\alpha] = \hat{\theta}_{\text{exact}}[\alpha] + O(n^{-3/2})$$

- **standard normal and Student's t points are only first-order correct**