

Political Science 787: Multivariate Analysis

Fall 2007

Monday 4–6 (7603 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>

- **linear regression**
 - **computation**
 - **influence**

- **linear regression computation tidbits: orthogonal decomposition**

- **problem: minimize sum of the squares of the residuals**

$$z = y - Xb$$

for $N \times 1$ matrix y , $N \times p$ matrix X and coefficient matrix b (mini quiz: what are the dimensions of b and of z)

- **problem in terms of the Euclidean vector norm**

$$\|u\|^2 = u'u = \sum_i u_i^2: \text{minimize}$$

$$\begin{aligned} S^2 &= \|z\|^2 \\ &= \|y - Xb\|^2 \end{aligned}$$

- **linear regression computation: orthogonal decomposition**
 - **conventionally stated solution to the problem minimize $S^2 = \|y - Xb\|^2$ is the normal equations:**

$$\hat{b} = (X'X)^{-1}X'y$$

but this approach is numerically less stable than an alternative

- **numerical stability: floating point arithmetic**
 - the numerical stability issue relates to using floating-point arithmetic and difficulties from computing the inverse $(X'X)^{-1}$
 - floating-point arithmetic uses a fixed number of binary bits to represent all numbers (64 bits for double precision arithmetic), some for an exponent and some for a mantissa (in IEEE floating point, 11 bits for exponent and 52 for mantissa)
 - so there is a fixed number of significant figures, and the gaps between representable small numbers are smaller than the gaps between representable large numbers
 - numbers in computation are not continuous

- **linear regression computation: orthogonal decomposition**
 - the normal equation approach is numerically less stable than using orthogonal decomposition
 - **QR (orthogonal-triangular) decomposition:** for $N \times p$ matrix X , find $N \times r$ matrix Q and $r \times p$ matrix R such that

$$X = QR$$

with Q having unit orthogonal (orthonormal) columns,

$$Q'Q = I$$

where I is the $r \times r$ identity matrix

- **linear regression computation: orthogonal decomposition**
 - **QR decomposition: for $N \times p$ matrix X , find $N \times r$ matrix Q and $r \times p$ matrix R such that**

$$X = QR$$

with Q having unit orthogonal (orthonormal) columns,

$$Q'Q = I$$

where I is the $r \times r$ identity matrix

- **Q is a basis for X**
- **the smallest r for which such a Q exists is the rank of X**

- **linear regression computation: orthogonal decomposition**
 - if Q is a basis for X , then there exists a vector c such that

$$Xb = Qc$$

- hence

$$\begin{aligned} S^2 &= \|y - Xb\|^2 \\ &= \|y - Qc\|^2 \end{aligned}$$

and S^2 is minimized by

$$c = Q'y$$

- to find the coefficients b , solve

$$Rb = c$$

(by back-substitution when R is upper triangular)

- **linear regression computation: orthogonal decomposition**
- **proof that $c = Q'y$ is a least-squares solution**
 - **let $Q_* = [Q \ Q_0]$ be an $N \times N$ orthogonal matrix (i.e., $Q_*Q_*' = I$)**
 - **then**

$$\begin{aligned} y &= Q_*Q_*'y \\ &= QQ'y + Q_0Q_0'y \end{aligned}$$

- **substituting into S^2 :**

$$\begin{aligned} S^2 &= \|y - Qc\|^2 \\ &= \|Q_0Q_0'y + Q(Q'y - c)\|^2 \\ &= \|Q_0Q_0'y\|^2 + \|Q'y - c\|^2 \end{aligned}$$

$\|Q_0Q_0'y\|^2$ is independent of c , and $\|Q'y - c\|^2 = 0$ if $c = Q'y$

- **linear regression computation: orthogonal decomposition**
- **computing QR decompositions**
 - **Gram-Schmidt algorithm**
 - **Householder transformation**
 - **Givens transformation**
- **for details see work on matrix computations**

- **linear regression: influence**
- **start with the distribution of the residuals relative to the following model: X is an $n \times p$ full-rank matrix of known constants, y is an n -vector of observed responses, β is a p -vector of unknown constant parameters, ε is an n -vector of unobservable errors and**

$$y = X\beta + \varepsilon$$

$$E(\varepsilon) = 0$$

$$\text{Var}(\varepsilon) = \sigma^2 I$$

- **linear least-squares residuals: for $V = (v_{ij}) = X(X'X)^{-1}X'$ and \hat{y} the vector of fitted values,**

$$e = (e_i) = y - \hat{y}$$

$$= (I - V)y$$

- linear regression: influence
- linear least-squares residuals:

$$\begin{aligned}e &= (e_i) = y - \hat{y} \\ &= (I - V)y\end{aligned}$$

- relationship between e and ε :

$$\begin{aligned}e &= (I - V)y \\ &= (I - V)(X\beta + \varepsilon) \\ &= (I - V)\varepsilon\end{aligned}$$

or in scalar form,

$$e_i = \varepsilon_i - \sum_{j=1}^n v_{ij}\varepsilon_j$$

- the size of v_{ij} is crucial

- **linear regression: influence**
- **the hat matrix**
 - V is symmetric ($V = V'$) and idempotent ($VV = V$)
 - V represents the linear transformation that projects any n -vector onto the space spanned by the columns of X (the column space of X)
 - V was dubbed the “hat” matrix because $\hat{y} = Vy$
 - from idempotency and symmetry we have

$$\text{trace}(V) = \text{rank}(V) = p$$

$$\sum_j v_{ij}^2 = v_{ii}$$

- V is invariant under nonsingular linear reparameterizations (i.e., for nonsingular matrix A , X and XA have the same V)

- linear regression: influence
- the hat matrix: size of v_{ii}
 - if the model contains a constant, then $v_{ii} \geq 1/n$
 - the upper bound on v_{ii} depends on the number of times row i of X (x_i) is replicated; let c denote this repetition count
 - if $x_i = x_j$, then $v_{ij} = v_{ii}$ and

$$v_{ii} = \sum_{j=1}^n v_{ij}v_{ji} = \sum_{j=1}^n v_{ij}^2 \geq cv_{ii}^2$$

therefore

$$1/n \leq v_{ii} \leq 1/c$$

- $v_{ii} = 1$ only if $c = 1$ and $v_{ij} = 0$ for all $j \neq i$, in which case $\hat{y}_i = y_i$ (rare and pathological)

- linear regression: influence
- the hat matrix: size of v_{ii}
 - for a model with an intercept, let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_{p-1}$ denote the eigenvalues of the crossproduct of the matrix of centered X variable, let p_1, \dots, p_{p-1} denote the corresponding eigenvectors, and let θ_{li} denote the angle between p_l and x_i
 - then

$$\cos(\theta_{li}) = \frac{p_l' x_i}{(x_i' x_i)^{1/2}}$$

and

$$v_{ii} = \frac{1}{n} + x_i' x_i \sum_{l=1}^{p-1} \frac{\cos^2(\theta_{li})}{\mu_l}$$

- linear regression: influence
- the hat matrix: computing v_{ii}
 - e.g., if q_i is row i of QR decomposition matrix Q , then

$$v_{ij} = q_i' q_j$$

- there are other ways as well

- **linear regression: influence**
- **the hat matrix: uses**
 - **cases remote in the factor space ($x_i'x_i$ large and x_i in the direction of a small eigenvector) will have large values of v_{ii}**
 - **$\text{var}(\hat{y}_i) = v_{ii}\sigma^2$ and $\text{var}(e_i) = (1 - v_{ii})\sigma^2$**
 - **if $\max(v_{ii})$ is not much smaller than 1, inspecting the residuals may be unlikely to detect an outlier**
 - **if $\max(v_{ii})$ is near 1, then robust regression may not work (due to the difficulty of having residuals to downweight)**
 - **asymptotics: $\max(v_{ii}) \rightarrow 0$ as $n \rightarrow \infty$ is necessary and sufficient for linear least squares estimates to be asymptotically normal**

- linear regression: influence
- studentized residuals: scale invariant and unrelated to X
- “internally” studentized: the scale for observation i is estimated using the data for i
 - let $\hat{\sigma}^2 = \sum_{i=1}^n e_i^2 / (n - p)$ denote the residual mean square and define

$$r_i = \frac{e_i}{\hat{\sigma}(1 - v_{ii})^{1/2}}$$

- for $m < n - p$, let $J = (j_1, \dots, j_m)'$ index m studentized residuals of interest, let $e_J = (e_{j_1}, \dots, e_{j_m})'$, and let V_J be the matrix containing the corresponding subset of rows and columns of V
- if $\varepsilon \sim N(0, \sigma^2 I)$, then $e \sim N(0, \sigma^2(I - V_J))$

- **linear regression: influence**
- **studentized residuals: scale invariant and unrelated to X**
- **“internally” studentized:**
 - **let $R_J = (r_{j_1}, \dots, r_{j_m})'$**
 - **density of R_J : for $m > 1$, the density $f(r)$ is an inverted Student function (complicated); but**

$$\text{cov}(r_i, r_j) = -v_{ij} / [(1 - v_{ii})(1 - v_{jj})]^{1/2}, \quad i \neq j$$

- **for $m = 1$, using $\nu = (n - p - m)/2$, the density of r_i is**

$$f(r) = \frac{\Gamma(\nu + \frac{1}{2})}{\Gamma(\nu)\Gamma(\frac{1}{2})(n - p)^{1/2}} \left(1 - \frac{r^2}{n - p}\right)^{\nu-1},$$

$$|r| \leq (n - p)^{1/2}$$

**hence $r_i^2/(n - p) \sim B(1/2, (n - p - 1)/2)$ so that $E(r_i) = 0$
and $\text{var}(r_i) = 1$**

- linear regression: influence
- studentized residuals: scale invariant and unrelated to X
- “externally” studentized: the scale estimator for observation i is independent of e_i
 - the residual mean square error computed without case i is

$$\begin{aligned}\hat{\sigma}_{(i)}^2 &= \frac{(n-p)\hat{\sigma}^2 - e_i^2/(1-v_{ii})}{n-p-1} \\ &= \hat{\sigma}^2 \left(\frac{n-p-r_i^2}{n-p-1} \right)\end{aligned}$$

- under normality, $\hat{\sigma}_{(i)}^2$ and e_i are independent, and the externally studentized residual is

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}(1-v_{ii})^{1/2}}$$

with $t_i \sim t(n-p-1)$

- **linear regression: influence**
- **studentized residuals: scale invariant and unrelated to X**
- **internal versus external**
 - t_i and r_i relate as follows:

$$t_i = r_i \left(\frac{n - p - 1}{n - p - r_i^2} \right)^{1/2}$$

- linear regression: influence
- studentized residuals: use to detect simple outliers
- mean shift outlier model

$$y = X\beta + d_i\phi + \varepsilon$$

$$E(\varepsilon) = 0$$

$$\text{var}(\varepsilon) = \sigma^2 I$$

where d_i is a vector with element i equal to one and the rest equal to zero

- if $\phi \neq 0$ then observation i is an outlier
- this model encompasses outliers due to either y_i or x_i
- assuming $\varepsilon \sim N(0, \sigma^2 I)$, the t -statistic to test $\phi = 0$ is t_i
- when $\phi \neq 0$, t_i^2 is distributed as noncentral F with noncentrality parameter $\phi^2(1 - v_{ii})/\sigma^2$

- **linear regression: influence**
- **studentized residuals: use to detect simple outliers**
- **mean shift outlier model**
- **assuming $\varepsilon \sim N(0, \sigma^2 I)$, the t -statistic to test $\phi = 0$ is t_i**
- **when $\phi \neq 0$, t_i^2 is distributed as noncentral F with noncentrality parameter $\phi^2(1 - v_{ii})/\sigma^2$**
- **when it is not known which observation may be an outlier, one can do a multiple comparisons test of the largest t_i^2 values**
 - **Bonferroni test level (α) adjustment: reject if $\max_i |t_i| > t(\alpha/n; n - p - 1)$ or (alternative) reject if $\max_i |t_i| > t(v_{ii}\alpha/p; n - p - 1)$**

- **linear regression: influence**
- **studentized residuals: use to detect simple outliers**
- **multiple case mean shift outlier model**

$$y = X\beta + D\phi + \varepsilon$$

$$E(\varepsilon) = 0$$

$$\text{var}(\varepsilon) = \sigma^2 I$$

where D is $n \times m$ with columns d_i , $i \in J = (j_1, \dots, j_m)'$ and ϕ is an m -vector

- **assuming $\varepsilon \sim N(0, \sigma^2 I)$, a statistic to test $\phi = 0$ is**

$$t_J^2 = \frac{(e'_J(I - V_J)^{-1}e_J)(n - p - m)}{((n - p)\hat{\sigma}^2 - e'_J(I - V_J)^{-1}e_J)(m)}$$

- **when $\phi = 0$, $t_J^2 \sim F(m, n - p - m)$**

- **linear regression: influence**
- **the stability question: how do results vary when the problem definition is modified?**
- **ideal: small changes in or deviations from assumptions should produce small changes in results**
- **linear least squares fails this criterion**
 - **breakdown point of zero: changing one observation ($1/n$ of the data) can completely destroy the estimator; “breakdown zero” arises upon letting $n \rightarrow \infty$**
 - **influence function: high sensitivity to distributional assumptions**

- **linear regression: influence**
- **influence curve**
 - let T_n be a vector-valued statistic of length k based on iid sample $(z_1, \dots, z_n)'$ from cdf F defined on R^m
 - how does T_n change when an aspect of the problem is changed?
 - let \hat{F} denote the empirical cdf based on $(z_1, \dots, z_n)'$
 - find a functional T that maps cdfs onto R^k such that $T(\hat{F}) = T_n$
 - given such a functional, model the properties of T_n by studying how $T(F)$ or $T(\hat{F})$ behave when F or \hat{F} is perturbed

- **linear regression: influence**
- **influence curve: example functional**

- **let $m = k = 1$ and $T_n = n^{-1} \sum_{i=1}^n z_i = \bar{z}$**
- **the corresponding functional is**

$$T(F) = \int z dF(z)$$

and

$$T(\hat{F}) = \int z d\hat{F} = \bar{z}$$

- **if small changes in F or \hat{F} do not produce big changes in $T(F)$ or $T(\hat{F})$, then $T_n = \bar{z}$ would be considered robust**

- **linear regression: influence**

- **influence curve**

- let δ_z denote the cdf giving mass 1 to point z in R^m

- the influence curve of T at F is

$$ICF_{T,F}(z) = \lim_{\varepsilon \rightarrow 0} \frac{T[(1 - \varepsilon)F + \varepsilon\delta_z] - T(F)}{\varepsilon}$$

as long as the limit exists for all $z \in R^m$

- $ICF_{T,F}(z)$ measures how T changes when an observation is added at z as $n \rightarrow \infty$

- **gross error sensitivity:**

$$\gamma^* = \sup_z ICF_{T,F}(z)$$

- **linear regression: influence**
- **influence curve example: sample average**
 - **let $k = m = 1$ and $\mu = T(F) = \int z dF$**
 - **the influence curve of T at F is**

$$IC(z) = \lim_{\varepsilon \rightarrow 0} \frac{(1 - \varepsilon)\mu + \varepsilon z - \mu}{\varepsilon} = z - \mu$$

- **$IC(z)$ is unbounded**

- **linear regression: influence**
- **empirical influence curve (“infinitely large” sample):**

$$EIC(x, y) = n(X'X)^{-1}x(y - x'\hat{\beta})$$

or

$$EIC(x_i, y_i) = n(X'X)^{-1}x_i e_i$$

- **case-deleted empirical influence curve:**

$$\begin{aligned} EIC_{(i)}(x, y) &= (n - 1)(X'_{(i)}X_{(i)})^{-1}x_i(y_i - x'_i\hat{\beta}_{(i)}) \\ &= (n - 1)(X'X)^{-1}x_i e_i / (1 - v_{ii}) \end{aligned}$$

- **linear regression: influence**
- **sample influence curve (finite sample): set $z'_i = (x'_i, y_i)$ and $\varepsilon = -1/(n - 1)$ and use the empirical cdf $\hat{F}_{(i)} = (1 - \varepsilon)\hat{F} + \varepsilon\delta_{z_i}$:**

$$\begin{aligned}
 SIC_i &= -(n - 1)(T(\hat{F}_{(i)}) - T(\hat{F})) \\
 &= (n - 1)(\hat{\beta} - \hat{\beta}_{(i)}) \\
 &= \frac{(n - 1)(X'X)^{-1}x_i e_i}{1 - v_{ii}}
 \end{aligned}$$

or

$$EIC(x_i, y_i) = n(X'X)^{-1}x_i e_i$$

- linear regression: influence
- a diagnostic statistic based on *SIC*: Cook's distance

$$D_i = r_i^2 \frac{v_{ii}}{(1 - v_{ii})p}$$

- notice

$$\frac{v_{ii}}{(1 - v_{ii})} = \left[\sum_j \mathbf{var}(x'_j \hat{\beta}_{(i)}) - \sum_j \mathbf{var}(x'_j \hat{\beta}_i) \right] / \sigma^2$$

- **m-estimators for robustness: bounding influence**
- **the influence function:**

$$IC(x; F, T) = \lim_{s \rightarrow 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}$$

- **let $f_\theta(x) = f(x; \theta)$ be a family of probability densities indexed by real parameter (vector) θ**
- **plan: estimate θ by an m-estimate $T = T(F)$ where the functional T is defined implicitly by**

$$\int \psi(x; T(F)) F(dx) = 0$$

- **impose two constraints:**
 - **Fisher consistency: $T(F_\theta) = \theta$**
 - **a bound $k(\theta)$ on the gross error sensitivity:**

$$|IC(x; F_\theta, T)| \leq k(\theta), \quad \text{for all } x$$

- **m-estimators for robustness: bounding influence**
- **plan: estimate θ by an m-estimate $T = T(F)$ where**

$$\int \psi(x; T(F)) F(dx) = 0$$

- **two constraints:**
 - **Fisher consistency: $T(F_\theta) = \theta$**
 - **a bound $k(\theta)$ on the gross error sensitivity:**

$$|IC(x; F_\theta, T)| \leq k(\theta), \quad \text{for all } x$$

- **solution:**

$$\psi(x; \theta) = [g(x; \theta) - a(\theta)]_{-b(\theta)}^{+b(\theta)}$$

where

$$g(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$$

with $b(\theta) > 0$ and $[x]_u^v = \max[u, \min(v, x)]$

- **m-estimators for robustness: bounding influence**
- **bounded influence solution:**

$$\psi(x; \theta) = [g(x; \theta) - a(\theta)]_{-b(\theta)}^{+b(\theta)}$$

where

$$g(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$$

with $b(\theta) > 0$ **and** $[x]_u^v = \max[u, \min(v, x)]$

- **maximum likelihood:**

$$\sum_i g(x_i; \theta) = 0$$

or (conceptually)

$$\int g(x; \theta) F(dx) = 0$$

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

- bounded influence estimator:

$$\psi(x; \theta) = [g(x; \theta) - a(\theta)]_{-b(\theta)}^{+b(\theta)}$$

where

$$g(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$$

with $b(\theta) > 0$ and $[x]_u^v = \max[u, \min(v, x)]$

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

- 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 $\lambda(\theta) = E\psi(x, \theta)$ 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) $E u(x, \theta, d) \leq bd,$ **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-quasiliikelihood 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)

- $A_n^o \equiv E(n^{-1} \nabla^2 \log f^n(X^n, \theta_o))$
- $B_n^o \equiv \text{var}(n^{-1/2} \nabla \log f^n(X^n, \theta_o))$
- **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

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

- **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:

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

is distributed asymptotically as χ_q^2

Poisson regression model

- **Poisson model density for mean function $\mu(x)$:**

$$\text{Prob}[Y = y|x] = \frac{e^{-\mu(x)} \mu(x)^y}{y!}$$

- **mean function:**

$$\mu(x_i) = \exp(x_i' \beta)$$

- **(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:**

$$s = \sum_{i=1}^n [y_i x_i - \exp(x_i' \beta) x_i]$$

- **hessian:**

$$H = - \sum_{i=1}^n \exp(x_i' \beta) x_i x_i'$$

Poisson regression model

- influence function (Hampel et al. 1986):

$$\begin{aligned} IF(x; T, F) &= \lim_{t \downarrow 0} \frac{\sum_{k=0}^{\infty} k[(1-t)f(k) + t1_{\{x\}}(k)] - \sum_{k=0}^{\infty} kf(k)}{t} \\ &= \lim_{t \downarrow 0} \frac{t \sum_{k=0}^{\infty} k1_{\{x\}}(k) - t \sum_{k=0}^{\infty} kf(k)}{t} \\ &= x - \mu(x) \end{aligned}$$

- variance of influence function:

$$\begin{aligned} \mathbf{var}_{IF}(T, F) &= \int_{\mathcal{X}} IF(x; T, F)^2 dF(x) \\ &= \sum_{k=0}^{\infty} (k - \mu(x))^2 f(k) \\ &= \mu(x) \end{aligned}$$

Negative binomial regression model

- negbin from Poisson-gamma mixture
- conditional Poisson model distribution

$$f(y_i|\theta_i) = \frac{\exp(-\theta_i)\theta_i^{y_i}}{y!}, \quad y = 0, 1, \dots$$

- for random ε_i , let

$$\begin{aligned}\theta_i &= \exp(\beta_0 + x_i'\beta_1 + \varepsilon_i) \\ &= \mu_i\nu_i\end{aligned}$$

with $\mu_i = \exp(\beta_0 + x_i'\beta_1)$ **and** $\nu_i = \exp(\varepsilon_i)$

- use the distribution $g(\nu_i)$ to integrate out ν_i and get the marginal distribution of y :

$$h(y_i|\mu_i) = \int f(y_i|\mu_i, \nu_i)g(\nu_i)d\nu_i$$

Negative binomial regression model

- let $g(\nu_i)$ be a two-parameter gamma distribution $g(\nu; \delta, \phi)$,

$$g(\nu, \delta, \phi) = \frac{\delta^\phi}{\Gamma(\delta)} \nu^{\delta-1} e^{-\nu\phi}, \quad \delta, \phi > 0$$

which implies $E(\nu) = \delta/\phi$, $\mathbf{Var}(\nu) = \delta/\phi^2$

- for identification, set $\delta = \phi$
- changing variables by $\theta = \mu\nu$ gives

$$g(\theta|\mu, \delta) = \frac{(\delta/\mu)^\delta}{\Gamma(\delta)} \theta^{\delta-1} e^{-\delta\theta/\mu}$$

- the marginal distribution of y is

$$h(y|\mu, \delta) = \int \frac{\exp(-\theta)\theta^y}{y!} \frac{(\delta/\mu)^\delta}{\Gamma(\delta)} \theta^{\delta-1} e^{-\delta\theta/\mu} d\theta$$

Negative binomial regression model

- using some facts about the gamma function (see e.g. Cameron and Trivedi 1998, 101)

$$h(y|\mu, \delta) = \frac{\Gamma(\delta + y)}{\Gamma(\delta)\Gamma(y + 1)} \left(\frac{\delta}{\delta + \mu}\right)^\delta \left(\frac{\mu}{\mu + \delta}\right)^y$$

- this gives the first two moments

$$E(y|\mu, \delta) = \mu$$

$$\mathbf{Var}(y|\mu, \delta) = \mu(1 + \mu/\delta)$$

Negative binomial regression model

- to compute the score, use the digamma function

$$\psi(x) = \partial \log \Gamma(x) / \partial x = \Gamma'(x) / \Gamma(x)$$

- elements of the score for $\mu_i = \exp(x_i' \beta)$

$$\frac{\partial \log h(y_i | \mu_i, \delta)}{\partial \delta} = \psi(\delta + y_i) - \psi(\delta) + \log \left(\frac{\delta}{\delta + \mu_i} \right) + \frac{\mu_i - y_i}{\delta + \mu_i}$$

$$\frac{\partial \log h(y_i | \mu_i, \delta)}{\partial \beta} = \frac{(y_i - \mu_i) \delta}{\mu_i (\mu_i + \delta)} \frac{\partial \mu_i}{\partial \beta} = \frac{y_i - \mu_i}{1 + \mu_i / \delta} x_i$$

for slightly different formulations see Cameron and Trivedi (1998, 71)

- for second derivatives use the trigamma function

$$\psi'(x) = \partial \psi(x) / \partial x$$

Poisson model: regression-based tests for overdispersion

- **mean:** $\mu_i = \exp(x_i'\beta)$
- **Poisson variance:** $\text{Var}(y_i|x_i) = \mu_i$
- **alternative variance:** $\text{Var}(y_i|x_i) = \mu_i + \alpha g(\mu_i)$, with, e.g.,
 $g(\mu_i) = \mu_i$ or $g(\mu_i) = \mu_i^2$
- **the alternative implies** $E[(y_i - \mu_i)^2 - y_i|x_i] = \alpha g(\mu_i)$ **while Poisson variance implies** $\alpha = 0$
- **define** $\omega_i = \omega(\mu_i) = \text{Var}[(y_i - \mu_i)^2 - y_i|x_i]$
- **under the Poisson model,** $\omega_i = 2\mu_i^2$
- **test** $\alpha = 0$ **by testing for** $\alpha = 0$ **in the (weighted) least squares regression**

$$\sqrt{\hat{\omega}_i}((y_i - \hat{\mu}_i)^2 - y_i) = \alpha\sqrt{\hat{\omega}_i} + u_i$$

Poisson model: regression-based tests for overdispersion

- a t test (one DF) statistic

$$T_1 = \left[s^2 \sum_{i=1}^n \hat{\omega}_i^{-1} g^2(\hat{\mu}_i) \right]^{-1/2} \sum_{i=1}^n \hat{\omega}_i^{-1} g^2(\hat{\mu}_i) ((y_i - \hat{\mu}_i)^2 - y_i)$$

where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n \hat{\omega}_i^{-1} ((y_i - \hat{\mu}_i)^2 - y_i - g^2(\hat{\mu}_i) \hat{\alpha})^2$$

- under the Poisson model, $\text{plim } s^2 = 1$ (since $\alpha = 0$), so asymptotically equivalent is

$$T_2 = \left[\sum_{i=1}^n \hat{\omega}_i^{-1} g^2(\hat{\mu}_i) \right]^{-1/2} \sum_{i=1}^n \hat{\omega}_i^{-1} g^2(\hat{\mu}_i) ((y_i - \hat{\mu}_i)^2 - y_i)$$

Poisson model: regression-based tests for overdispersion

- **for more about these tests see Cameron and Trivedi (1998, 170ff)**

Poisson model: regression-based tests for overdispersion

- the alternative also implies $E[(y_i - \mu_i)^2 - \mu_i | x_i] = \alpha g(\mu_i)$
- define $\omega_i = \omega(\mu_i) = \mathbf{Var}[(y_i - \mu_i)^2 - y_i | x_i]$
- test for $\alpha = 0$ in the (weighted) least squares regression

$$\sqrt{\hat{\omega}_i}((y_i - \hat{\mu}_i)^2 - \hat{\mu}_i) = \alpha \sqrt{\hat{\omega}_i} + u_i$$

- let w_{ij} be the ij entry of $W = [D - \Delta(\Delta' D_\mu^{-1} \Delta)^{-1} \Delta']^{-1}$ where D and D_μ are $n \times n$ diagonal matrices with respective entries $2\mu_i^2 + \mu_i$ and μ_i , and Δ is $n \times k$ with rows $\partial \mu_i / \partial \beta'$
- t test (one DF) is

$$T_3 = \left[\sum_{i=1}^n \sum_{j=1}^n \hat{w}_{ij} g(\hat{\mu}_i) g(\hat{\mu}_j) \right]^{-1/2} \sum_{i=1}^n \sum_{j=1}^n \hat{w}_{ij} g(\hat{\mu}_i) ((y_j - \hat{\mu}_j)^2 - \hat{\mu}_j)$$

- **cumulant generating function of the distribution of X**

$$C(t) = \log M(t) \equiv \log E(e^{tX})$$

- **derivatives of C give moments of X**

$$C' = M'/M, \quad C'' = (MM'' - (M')^2)/M^2$$

so

$$C'(0) = E(X), \quad C''(0) = \mathbf{Var}(X)$$

- **expanding C gives, in a neighborhood of 0,**

$$C(t) = \sum_{i=1}^{\infty} \frac{\kappa_i}{i!} t^i$$

where the κ_i are the cumulants of X

$$\kappa_1 = E(X), \quad \kappa_2 = E(X^2) - E(X)^2$$

sufficient statistics

- let y be a random variable Y with density $f_Y(y; \theta)$, $\theta \in \Omega_\theta$
- let s be a statistic (i.e., a function of y) with corresponding random variable S
- s is sufficient if for all s and for all $\theta \in \Omega_\theta$ the conditional density of Y given $S = s$ does not depend on θ , i.e., if

$$f_{Y|S}(y|s; \theta) = g(y, s) \quad (3)$$

sufficient statistics

- **minimal sufficient statistic: the minimal S for which (3) holds**
- **a factorization theorem: S is sufficient for θ if for all y and $\theta \in \Omega_\theta$ there are functions $g(s, \theta)$ and $h(y)$ such that**

$$f_Y(y; \theta) = g(s, \theta)h(y)$$

- **if all the densities $f_Y(y; \theta)$, $\theta \in \Omega_\theta$ have common support, then one can find the minimal sufficient statistic by choosing an arbitrary fixed $\theta_0 \in \Omega_\theta$, forming ratios**

$$L_0(\theta) = f_Y(y; \theta) / f_Y(y; \theta_0)$$

and finding s such that the ratios depend on y through s only

generalized linear models and quasi-likelihood models

- exponential family: for known ϕ ,

$$f_Y(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\theta) + c(y, \phi)\}$$

θ is the canonical parameter

- r -th order cumulants are

$$\kappa_r = a(\theta)^{2r-2} b^{(r)}(\theta)$$

e.g.,

$$\kappa_1 = b'(\theta) = E(Y), \quad \kappa_2 = a(\theta)^2 b''(\theta) = \mathbf{var}(Y)$$

- **exponential family**

$$f_Y(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\theta) + c(y, \phi)\}$$

- **normal distribution**

$$\begin{aligned} f_Y(y; \theta, \phi) &= (2\pi\sigma^2)^{-1/2} \exp\{-(y - \mu)^2 / (2\sigma^2)\} \\ &= \exp\{(y\mu - \mu^2) / 2 / \sigma^2 - (y^2 / \sigma^2 + \log(2\pi\sigma^2)) / 2\} \end{aligned}$$

hence $\theta = \mu$, $\phi = \sigma^2$ and

$$a(\phi) = \sigma, \quad b(\theta) = \theta^2 / 2, \quad c(y, \phi) = -(y^2 / \sigma^2 + \log(2\pi\sigma^2)) / 2$$

- **Poisson distribution: $\phi = 1$ and**

$$a(\phi) = 1, \quad b(\theta) = \exp(\theta), \quad c(y, \phi) = -\log y!$$

generalized linear models

- **linear predictor:** for covariates x_1, x_2, \dots, x_p and coefficients $\beta_1, \beta_2, \dots, \beta_p$, the linear predictor is $\eta = \sum_{j=1}^p x_j \beta_j$
- **link function:** for expectation $\mu = E(Y)$, the link function $g(\cdot)$ is such that $\eta_i = g(\mu_i)$
- **canonical link:** link for which there exists a sufficient statistic of dimension equal to that of β in the linear predictor, in which case $\theta = \eta$
- for canonical links the sufficient statistic is $\sum_{i=1}^n x_{ij} y_j$
- **normal distribution:** canonical link is the identity function
- **Poisson distribution:** canonical link is \log

GLMs: fitting via iteratively reweighted least squares

- **current linear predictor and fitted value:** $\hat{\eta}_0 = x' \hat{\beta}_0$,
- **current fitted value:** $\hat{\mu}_0 = g^{-1}(\hat{\eta}_0)$
- **weights:**

$$W_0^{-1} = \left(\frac{d\eta}{d\mu} \right)_0^2 V_0$$

where V_0 is the variance function evaluated at $\hat{\mu}_0$

- **adjusted dependent variable:**

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0) \left(\frac{d\eta}{d\mu} \right)_0$$

where $d\eta/d\mu$ is evaluated at $\hat{\mu}_0$

- **regress z_0 on x with weight W_0 to get $\hat{\beta}_1$, $\hat{\eta}_1$ and iterate**

GLMs: fitting via iteratively reweighted least squares

- rationale: z linearizes the link function; to first order (Taylor expansion)

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu)$$

and

$$g(\mu) + (y - \mu)g'(\mu) = \eta + (y - \mu)\frac{d\eta}{d\mu}$$

- assuming η and μ are fixed and known and ignoring the dispersion parameter ϕ , the variance of Z is W^{-1}
- starting values: $\hat{\mu}_0 = y$ (possibly with adjustments)

GLMs: fitting via iteratively reweighted least squares

- **rationale for exponential family: show that the ML equations for β_j are**

$$\sum W(y - \mu) \frac{d\eta}{d\mu} x_j = 0$$

- the ML equations for β_j are

$$\sum W(y - \mu) \frac{d\eta}{d\mu} x_j = 0$$

- log likelihood for a single observation

$$l = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)$$

- using $d\mu/d\theta = V$ and $\partial\eta_j/\partial\beta_j = x_j$ and the chain rule,

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \frac{\partial l}{\partial \theta} \frac{d\theta}{d\mu} \frac{d\mu}{d\eta} \frac{\partial \eta}{\partial \beta_j} \\ &= \frac{(y - \mu)}{a(\phi)} \frac{1}{V} \frac{d\mu}{d\eta} x_j \\ &= \frac{W}{a(\phi)} (y - \mu) \frac{d\mu}{d\eta} x_j \end{aligned}$$

quasi-likelihood models

- starting with the assumption that Y has mean μ and variance $\sigma^2 V(\mu)$, define the quasi-likelihood $\ell(\mu; y)$ implicitly by the system of partial differential equations

$$\frac{\partial \ell(\mu; y)}{\partial \mu} = V^{-1}(\mu)(y - \mu) = 0$$

where V^{-1} is a generalized inverse of V

- if $EY = \mu = \mu(\beta)$, i.e., the expectation is not a function of an unknown σ^2 , then generalized least squares estimators for β are

$$D'V^{-1}(y - \mu(\hat{\beta})) = 0$$

where $D = d\mu/d\beta$ is $N \times p$

- this $\hat{\beta}$ need not be an MLE

quasi-likelihood models

- let $EY = \mu = \mu(\beta)$ and $\text{Cov}(Y) = \sigma^2 V(\mu)$ with μ having bounded third derivatives, let $i_\beta = D'V^{-1}D$ and assume Y has finite third moments, then

$$n^{1/2}(\hat{\beta} - \beta) \rightarrow N(0, n\sigma^2 i_\beta^{-1}) + O_p(n^{-1/2})$$

- see McCullagh (1983) for more results

binary data

- **binomial likelihood**

$$l(\pi; y) = \sum_{i=1} n \left[y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + m_i \log(1 - \pi_i) \right]$$

- **link functions (e.g.)**

- **logit or logistic:** $g(\pi) = \log(\pi/(1 - \pi))$

- **probit or inverse normal:** $g(\pi) = \Phi^{-1}(\pi)$

- **complementary log-log:** $g(\pi) = \log(-\log(1 - \pi))$

- **deviance**

$$D(y; \hat{\pi}) = 2 \sum_i \left\{ y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{\mu}_i} \right\}$$

binary data

- **overdispersion (for $m_i > 1$):**

$$E(Y) = m\pi$$

$$\mathbf{var}(Y) = \sigma^2 m\pi(1 - \pi)$$

models

- fixed parameters and interest in distribution of a statistic
- versus random parameters and interest in posterior densities (Bayesian statistics)
- likelihood and log likelihood:

$$\text{lik}\{\theta; y\} = f(y; \theta)$$

$$l(\theta; y) = \log f(y; \theta)$$

- sampling theory: $l_Y(\theta; y)$ (data are random and parameters are fixed)
- Bayes theory: $l_\Theta(\theta; y)$ (data are fixed and parameters are random)
- maximum likelihood: choose θ to maximize $l_Y(\theta; y|y)$ (condition on the observed data values y)

statistics (see Cox and Hinkley 1974)

- **let $y = (y_1, \dots, y_n)$ be a realization of a random variable Y , and suppose we have specified a family \mathcal{F} of possible distributions**
- **a statistic is a function $T = t(Y)$**
- **a statistic S is sufficient for the family \mathcal{F} if the conditional density $f_{Y|S}(y|s)$ is the same for all distributions in \mathcal{F}**
- **with a parametric model, S is sufficient for θ if $f_{Y|S}(y|s; \theta)$ does not depend on θ**
- **let S be minimal sufficient for θ with $\dim(S) > \dim(\theta)$; if $S = (T, C)$ with the marginal density of C independent of θ , then C is an ancillary statistic (e.g., centered normal theory linear regression model)**

statistical inference: general principles (see Cox and Hinkley 1974)

- **sufficiency: given the model $f_Y(y; \theta)$ for data y and minimal sufficient statistic S for θ , identical conclusions should be drawn from data y_1 and y_2 if both data sets produce the same value of s**
- **conditionality: let C be an ancillary statistic; the conclusion about the parameter of interest is to be drawn as if C were fixed at its observed value**
- **sufficiency and conditionality: the adequacy of the model can be tested by seeing whether the data y , given $S = s$, match the known conditional distribution**

statistical inference: general principles (see Cox and Hinkley 1974)

- **invariance:**

- let the model be $f_Y(y; \theta)$
- let \mathcal{G} be a group of transformations such that if $\phi = g * \theta$ is the transformed parameter, then the distribution of the transformed random variable is $f_Y(y; \phi)$
- (point estimation invariance) any estimate $t(y)$ of θ should satisfy $t(g(y)) = g * t(y)$

- **MLE satisfies invariance**

statistical inference: selected approaches (see Cox and Hinkley 1974)

- **strong repeated sampling principle: assess statistical procedures using their behavior in hypothetical repetitions under the same conditions**
 - **interpret measures of uncertainty as hypothetical frequencies in long run repetitions**
 - **formulate optimality criteria in terms of sensitive behavior in hypothetical repetitions**
- **sampling theory: emphasize the strong repeated sampling principle**

statistical inference: selected approaches (see Cox and Hinkley 1974)

- **likelihood theory: use the likelihood function directly as a summary of information; likelihood ratios measure the relative plausibilities of two preassigned parameter values**
- **Bayesian theory:**
 - in addition to the pdf $f_Y(y; \theta)$, assumed to generate the data, treat the parameter θ as the value of a random variable Θ with a known marginal pdf $f_\Theta(\theta)$ (prior)
 - the data are generated from the conditional pdf $f_{Y|\Theta}(y; \theta)$
 - interest centers on the conditional distribution of Θ given $Y = y$ (posterior), which Bayes's theorem gives as

$$f_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)f_\Theta(\theta)}{\int_{\Omega} f_{Y|\Theta}(y|\theta')f_\Theta(\theta')d\theta'}$$

significance tests

- we have data $y = (y_1, \dots, y_n)$ and a hypothesis H_0 about their density $f_Y(y)$
 - a simple null hypothesis completely specifies $f_Y(y)$
 - a composite null hypothesis partially specifies $f_Y(y)$ (e.g., specifies only some of the parameters)
- null distributions: $t = t(y)$ is a function of the observations and $T = t(Y)$ is the corresponding random variable; T is a test statistic for testing H_0 if
 1. the distribution of T when H_0 is true is known at least approximately
 2. the larger the value of t , the stronger the evidence of departure from H_0

significance tests

- level of significance given $t = t_{\text{obs}} = t(y)$:

$$p_{\text{obs}} = \mathbf{pr}(T \geq t_{\text{obs}}; H_0)$$

- e.g., tests of goodness of fit
 - what is the null? what is the alternative? (generally nothing specific)
 - these are generally tests of $f_{Y|S}(y|s)$ given a minimal sufficient statistic S

significance tests

- e.g., nonnested hypothesis tests
 - likelihood ratio: $\log[f(y)/g(y)]$
 - Cox: $\log[f_g(y)/g(y)]$; difficult
 - Vuong (1989): $n^{-1} \sum_{i=1}^n \log[f(y_i)/g(y_i)]$ using information theory; gives $N(0, 1)$ (with appropriate rescaling for the variance) when the distributions are not significantly different

tests

- **distribution-free tests: the distribution of the test statistic under the null is the same for a family of densities more general than a finite parameter family**
 - **achieved by conditioning on the complete minimal sufficient statistic**
 - **regard the order statistics are fixed and use the consequence that under the null all permutations of the ordered values are equally likely**
 - **permutation tests**

confidence intervals

- **confidence limits:**

$$\mathbf{pr}(T^\alpha \geq \theta; \theta) = 1 - \alpha$$

if $\alpha_1 > \alpha_2$ and T^{α_1} and T^{α_2} are both defined, then

$$T^{\alpha_1} \leq T^{\alpha_2}$$

T^α is a $1 - \alpha$ upper confidence limit for θ

- **lower confidence limit:**

$$\mathbf{pr}(T_\alpha \leq \theta; \theta) = 1 - \alpha$$

- **conservative confidence limits:**

$$\mathbf{pr}(T^\alpha \geq \theta; \theta) \geq 1 - \alpha, \quad \mathbf{pr}(T_\alpha \leq \theta; \theta) \geq 1 - \alpha$$

confidence intervals

- $[T_., T^.]$ is a $1 - \alpha$ confidence interval if

$$\text{pr}(T_. \leq \theta \leq T^.; \theta) = 1 - \alpha$$

- (continuous case) a combination of upper and lower limits at levels α_1 and α_2 with $\alpha_1 + \alpha_2 = \alpha$ will define a $1 - \alpha$ confidence interval

test statistics (examples, no nuisance parameters)

- **background notation: for likelihood function $\ell(\theta; Y)$,**

$$u(\theta; Y) = U(\theta) = \nabla_{\theta} \ell(\theta; Y)$$

$$E\{U(\theta); \theta\} = 0$$

$$\mathbf{cov}\{U(\theta); \theta\} = E\{U(\theta)U(\theta)^{\mathbf{T}}; \theta\} = E\{-\nabla_{\theta} \nabla_{\theta}^{\mathbf{T}} \ell(\theta; Y)\} = i(\theta)$$

- **Neyman-Pearson likelihood ratio statistic:**

$$w(\theta_0) = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\}$$

- **the Wald, or maximum likelihood estimate statistic**

$$w_{\mathbf{P}} = (\hat{\theta} - \theta_0)^{\mathbf{T}} i(\theta_0) (\hat{\theta} - \theta_0)$$

confidence intervals by inversion

- the likelihood ratio statistic often has a χ_q^2 distribution, so for a $1 - \alpha$ confidence region choose θ to satisfy

$$\{\theta : 2(\ell(\hat{\theta}) - \ell(\theta)) \leq \chi_{q,1-\alpha}^2\}$$

where $\chi_{q,1-\alpha}^2$ is the tabulated $1 - \alpha$ point of χ_q^2

confidence intervals by inversion

- the Wald statistic often has a χ_q^2 distribution, so for a $1 - \alpha$ confidence region choose θ to satisfy

$$\{\theta : (\hat{\theta} - \theta)^T i(\theta) (\hat{\theta} - \theta) \leq \chi_{q,1-\alpha}^2\}$$

where $\chi_{q,1-\alpha}^2$ is the tabulated $1 - \alpha$ point of χ_q^2

- Inferior would be

$$\{\theta : (\hat{\theta} - \theta)^T i(\hat{\theta}) (\hat{\theta} - \theta) \leq \chi_{q,1-\alpha}^2\}$$

- what does this say about the usual practice of getting confidence intervals for regression model coefficients by inverting t -statistics?

profile likelihood

- let $\psi = \psi(\theta)$ be a subparameter (or a function of the parameter θ)
- the profile likelihood $L_{\mathbf{P}}(\psi)$ for ψ is

$$L_{\mathbf{P}}(\psi) = \max_{\theta|\psi} L(\theta)$$

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

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

- a profile likelihood region $\{\ell_{\mathbf{P}}(\hat{\psi}) - \ell_{\mathbf{P}}(\psi) < c\}$ 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

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

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

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

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_α intervals

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

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

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**