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
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 \(\theta\) to maximize \(l_Y(\theta; y | y)\) (condition on the observed data values \(y\))
statistics (see Cox and Hinkley 1974)

- let $y = (y_1, \ldots, y_n)$ 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 $\theta$ if $f_{Y|S}(y|s; \theta)$ does not depend on $\theta$

- let $S$ be minimal sufficient for $\theta$ with $\text{dim}(S) > \text{dim}(\theta)$; if $S = (T, C)$ with the marginal density of $C$ independent of $\theta$, 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 $\theta$, 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 \( \phi = g \ast \theta \) 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 \( \theta \) should satisfy \( t(g(y)) = g \ast t(y) \)

- 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 \( \theta \) as the value of a random variable \( \Theta \) with a know 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 \( \Theta \) 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 \( y = (y_1, \ldots, y_n) \) 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 \( t = t_{\text{obs}} = t(y) \):

\[
p_{\text{obs}} = \text{pr}(T \geq t_{\text{obs}}; H_0)
\]

- 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: $\log[f_g(y)/g(y)]; \text{difficult}$
  - 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:

  \[ \text{pr}(T^\alpha \geq \theta; \theta) = 1 - \alpha \]

  if \( \alpha_1 > \alpha_2 \) and \( T^{\alpha_1} \) and \( T^{\alpha_2} \) are both defined, then

  \[ T^{\alpha_1} \leq T^{\alpha_2} \]

  \( T^\alpha \) is a \( 1 - \alpha \) upper confidence limit for \( \theta \)

- lower confidence limit:

  \[ \text{pr}(T_\alpha \leq \theta; \theta) = 1 - \alpha \]

- conservative confidence limits:

  \[ \text{pr}(T^\alpha \geq \theta; \theta) \geq 1 - \alpha, \quad \text{pr}(T_\alpha \leq \theta; \theta) \geq 1 - \alpha \]
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 \(\alpha_1\) and \(\alpha_2\) with \(\alpha_1 + \alpha_2 = \alpha\) will define a \(1 - \alpha\) confidence interval
test statistics (examples, no nuisance parameters)

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

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

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

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

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

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

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

  $$w_P = (\hat{\theta} - \theta_0)^T i(\theta_0)(\hat{\theta} - \theta_0)$$
confidence intervals by inversion

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

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

where $\chi^2_{q,1-\alpha}$ is the tabulated $1 - \alpha$ point of $\chi^2_q$. 
confidence intervals by inversion

- the Wald statistic often has a $\chi_q^2$ distribution, so for a $1 - \alpha$ confidence region choose $\theta$ 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 $\chi_q^2$

- 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 $\theta$)

- the profile likelihood $L_P(\psi)$ for $\psi$ is

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

- profile log-likelihood: $\ell_P = \log L_P$

- the maximum profile likelihood estimate of $\psi$ equals $\hat{\psi}$ (the MLE)

- profile log-likelihood statistic tests $\psi = \psi_0$, i.e., for $\theta = (\psi, \chi)$,

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

- a profile likelihood region $\{\ell_P(\hat{\psi}) - \ell_P(\psi) < c\}$ is, generally, an approximate confidence region for $\psi$
higher-order asymptotic theory: Bartlett adjustment

- for random vector $y$, probability density $f(y; \omega)$ with parameter $\omega$, 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 $\chi$, 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 $n \to \infty$, $w$ converges to $\chi^2_d$, the chi-squared distribution with $d$ degrees of freedom, where $d$ is the dimension of, respectively, $\omega_0$ or $\psi_0$
higher-order asymptotic theory: Bartlett adjustment

- let \( q_d(x) \) denote the density of \( \chi_d^2 \)
- 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 \( \chi_d^2 \) 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**

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

**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 \( \omega \) (see Barndorff-Nielsen and Cox 1984)
bootstrap

- some notation

- data: $y_1, \ldots, y_n$ are iid random variables $Y_1, \ldots, Y_n$

- pdf is $f$ and cdf is $F$

- population characteristic: $\theta$

- statistic: $T$ estimates $\theta$, 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, $x^*_1, \ldots, x^*_B$, evaluate $\hat{\theta}^*(b) = s(x^*_b)$ and compute the average

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

then

$$\text{bias}_B = \hat{\theta}^*(\cdot) - t(\hat{F})$$
bootstrap distribution estimates: general theory

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

$$H(x) = P(U \leq x)$$

$$= \Phi(x/\sigma) + n^{-1/2}p(x/\sigma)\phi(x/\sigma) + O(n^{-1})$$

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

$$\hat{H}(x) = P(U^* \leq x|\mathcal{X})$$

$$= \Phi(x/\hat{\sigma}) + n^{-1/2}\hat{p}(x/\hat{\sigma})\phi(x/\hat{\sigma}) + O_p(n^{-1})$$

- 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 $\sigma^2$ is the asymptotic variance of $\hat{\theta}$), then for some polynomial $q(x)$

$$G(x) = P(T \leq x)$$

$$= \Phi(x) + n^{-1/2}q(x)\phi(x) + O(n^{-1})$$

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

$$\hat{G}(x) = P(T^* \leq x|\mathcal{X})$$

$$= \Phi(x) + n^{-1/2}\hat{q}(x)\phi(x) + O_p(n^{-1})$$

- 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 $\sigma$

• in some applications $\sigma$ may be difficult to estimate in a stable way, or $\sigma$ 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 \( \alpha \)-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 $\Phi$ to denote the standard normal CDF and $z^{(\alpha)}$ for the $100\alpha$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”

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