STATISTICS 22S:194

Luke Tierney

Spring 2003

Week 1

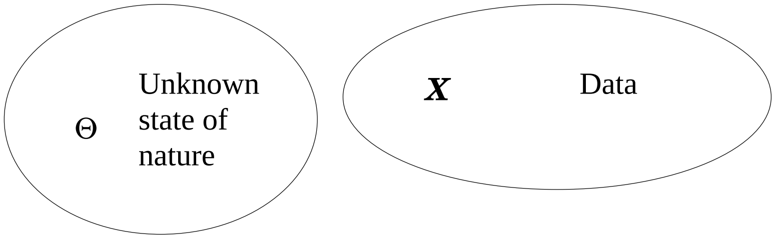
Wednesday, January 22, 2003

Review Course Outline

Review First Semester Final Exam

Statistical Inference

The basic framework:



Objectives: use data $X \in \mathscr{X}$ to learn about aspects of $\theta \in \Theta$, e.g.

- Based on X, what is best guess for θ?
- How accurate is our best guess?

Need a link between θ and X.

Frequentist Approach

- Assume $f(x|\theta)$ is known
- Develop procedures that work well on average in similar experiments.

Drawback: Don't relate directly to your experiments.

Bayesian Approach

- Assume $f(x|\theta)$ and $f(\theta)$ known.
- Compute $f(\theta|x)$

Drawbacks:

- Need $f(\theta)$
- Need to compute features of $f(x|\theta)$.

Resampling Approach

- Assume little, or limit use of assumptions to suggesting estimators
- Use resampling to assess variability

This is often very computationally intensive.

The basic $X, \theta, f(x|\theta)$ framework is quite general:

- Standard parametric model:

$$\begin{aligned} X &= (X_1, \ldots, X_n) \in \mathbb{R}^n \\ \theta &\in \mathbb{R} \\ f(x|\theta) &= i.i.d\ N(\theta, 1) \end{aligned}$$

- Nonparametric model:

$$\begin{aligned} X &= (X_1, \ldots, X_n) \in \mathbb{R}^n \\ \theta &= \text{a distribution on } \mathbb{R} \\ f(x|\theta) &= i.i.d.\ \theta \end{aligned}$$

Some approaches do not use $f(x|\theta)$ (randomization theory).

Often we are really interested in one or two aspects of θ:

$$f(x|\theta) = f(x|\mu, \sigma)$$

- might want to learn about μ
- might not be interested in σ.

Parameters not of direct interest are called *nuisance parameters*.

Friday, January 24, 2003

Suf£ciency

A £rst step in using $f(x|\theta)$ is to see what features of the data are important, what are super¤uous (formally at least).

De£nition

A statistic $T(X)$ is suf£cient for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Example

Let $X_1,\ldots,X_n$ be *i.i.d.* Bernoulli(p) and set

$$T(X) = \sum_{i=1}^{n} X_i$$

Then for $x_i = 0,1$ and $t = 0,\ldots,n$

$$\begin{aligned} f_{X,T}(x,t) &= p^{\sum x_i}(1-p)^{n-\sum x_i} 1_{\{\sum x_i = t\}} \\ &= p^t (1-p)^{n-t} 1_{\{\sum x_i = t\}} \\ f_T(t) &= \binom{n}{t} p^t (1-p)^{n-t} \end{aligned}$$

So

$$f_{X|T}(x|t) = \frac{f_{X,T}(x,t)}{f_T(t)} = \frac{1_{\{\sum x_i = t\}}}{\binom{n}{t}}$$

In words: $X|T = t$ is uniform on the $\binom{n}{t}$ vectors $(x_1,\ldots,x_n)$ with $x_i = 0,1$ and $\sum x_i = t$.

This distribution does not depend on p, so T is suf£cient.

Suppose this experiments is performed. You get to see all of $x_1,\ldots,x_n$ but I only get to see $T(x) = t$. Are you better off?

Answer: I can get data y with the same distribution as x and the same value of t by choosing y uniformly from its possible values given $T(y) = t$. So my data is equivalent to yours.

This assumes that the model is right.

Sufficiency Principle

A procedure based on assuming a particular form $f(x|\theta)$ should only depend on x through a sufficient statistic $T(x)$. Two observations x and y with $T(x) = T(y)$ where T is sufficient should result in the same actions.

Unfortunately we cannot use our definition of sufficiency with our conditional probability tools for continuous data, since $X, T(X)$ are not jointly continuous.

Instead, we will work with characterizations of sufficiency that are valid.

Halmos-Savage Factorization Theorem

If $f(x|\theta)$ is the joint PMF or PDF of X, then $T(X)$ is sufficient for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that for *all* x and *all* θ

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

Proof

This proof is only for the discrete case.

Suppose T is sufficient. Then

$$f(x|\theta) = P(X = x) = P(X = x, T(X) = T(x)) = \underbrace{f_{X|T}(x|T(x))}_{h(x)}\underbrace{f_T(T(x)|\theta)}_{g(T(x)|\theta)}$$

So a factorization exists.

For the converse, suppose

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

for some g, h. Let $A_t = \{y : T(y) = t\}$. Then

$$f_T(t) = \sum_{y \in A_t} f(y|\theta) = g(t|\theta) \sum_{y \in A_t} h(y)$$

So

$$\begin{aligned} f_{X|T}(x|t) &= \frac{f(x|\theta)1_{\{T(x)=t\}}}{f_T(t)} = \frac{g(t|\theta)h(x)1_{\{T(x)=t\}}}{g(t|\theta)\sum_{y \in A_t} h(y)} \\ &= \frac{h(x)1_{\{T(x)=t\}}}{\sum_{y \in A_t} h(y)} = \frac{h(x)1_{A_t}(x)}{\sum_{y \in A_t} h(y)} \end{aligned}$$

which does not depend on θ. □

We can use the factorization theorem to verify that a statistic is sufficient.

Examples

1. $X_1,\ldots,X_n$ are *i.i.d.* Bernoulli(θ), $T(X) = \sum X_i$. Then

$$f(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \underbrace{\theta^{T(x)}(1-\theta)^{n-T(x)}}_{g(T(x)|\theta)} \times \underbrace{1}_{h(x)}$$

2. $X_1,\ldots,X_n$ *i.i.d.* $N(\theta,1)$, $T(X) = \overline{X}$. Then

$$\begin{aligned} f(x|\theta) &= \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i-\theta)^2\right\} \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i-\overline{x})^2 - \frac{n}{2}(\overline{x}-\theta)^2\right\} \\ &= \underbrace{\frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i-\overline{x})^2\right\}}_{h(x)} \underbrace{\exp\{-\frac{n}{2}(\overline{x}-\theta)^2\}}_{g(\overline{x}|\theta)} \end{aligned}$$

So $\overline{X}$ is sufficient.

To use the factorization theorem to *find* a sufficient statistic, we need to

1. Split $f(x|\theta)$ into part that depends on θ and part that doesn't
2. Work out how the part that depends on θ depends on X.

Example

$X_1,\ldots,X_n$ *i.i.d* $N(\mu,\sigma^2)$, $\theta = (\mu,\sigma^2)$. Then

$$\begin{aligned} f(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\overline{x})^2 - \frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{n-1}{2\sigma^2}s^2 - \frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right\} \\ &= g(s^2,\overline{x}|\theta) \times 1 \end{aligned}$$

So $(S^2,\overline{X})$ is sufficient.

Note: if T is sufficient and $T(X) = H(R(X))$, then R is also sufficient (look at the factorization theorem).

So $(\sum X_i, \sum X_i^2)$ is also sufficient.

Example

Suppose $X_1,\ldots,X_n$ are *i.i.d.* from an exponential family

$$f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\right\}$$

Then

$$\begin{aligned} f(x_1,\ldots,x_n|\theta) &= \left(\prod_{i=1}^{n} h(x_i)\right) c(\theta)^n \exp\left\{\sum_{j=1}^{k} w_j(\theta)\sum_{i=1}^{n} t_j(x_i)\right\} \\ &= \left(\prod_{i=1}^{n} h(x_i)\right) c(\theta)^n \exp\left\{\sum_{j=1}^{k} w_j(\theta)T_j(x)\right\} \end{aligned}$$

with $T_j(x) = \sum_{i=1}^{n} t_j(x_i)$. So $(T_1,\ldots,T_k)$ is suf£cient for θ.

Example

$X_1,\ldots,X_n$ *i.i.d.* Poisson(λ).

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = \frac{1}{x!}e^{-\lambda}e^{x\log\lambda}$$

So $T_1 = \sum X_i$ is suf£cient.

Example

Suppose $X_1,\ldots,X_n$ are *i.i.d.* $U[0,\theta]$. Then

$$\begin{aligned} f(x_1,\ldots,x_n|\theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n} 1_{[0,\theta]}(x_i) \\ &= \left(\prod_{i=1}^{n} 1_{[0,\infty)}(x_i)\right)\frac{1}{\theta^n}\left(\prod_{i=1}^{n} 1_{(-\infty,\theta]}(x_i)\right) \\ &= \underbrace{\left(\prod_{i=1}^{n} 1_{[0,\infty)}(x_i)\right)}_{h(x)}\underbrace{\frac{1}{\theta^n}1_{(-\infty,\theta]}(x_{(n)})}_{g(x_{(n)}|\theta)} \end{aligned}$$

So $X_{(n)} = \max\{X_1,\ldots,X_n\}$ is suf£cient for θ.

Example

Suppose $X_1,\dots,X_n$ are $i.i.d.$ $U[\theta_1,\theta_2]$. Then

$$\begin{aligned} f(x_1,\dots,x_n|\theta) &= \frac{1}{(\theta_2-\theta_1)^n}\prod_{i=1}^n 1_{[\theta_1,\theta_2]}(x_i) \\ &= \frac{1}{(\theta_2-\theta_1)^n} 1_{[\theta_1,\infty)}(x_{(1)}) 1_{(-\infty,\theta_2]}(x_{(n)}) \end{aligned}$$

So $(X_{(1)},X_{(n)})$ is suf£cient.

Homework

Problem 6.3
Problem 6.6

Due Friday, January 31, 2003.

Week 2

Monday, January 27, 2003

Minimal Sufficiency

Definition

A sufficient statistic T is minimal sufficient if for any other sufficient statistic T', T is a function of T', i.e. $T = R(T')$ for some function R.

Lehman-Scheffé Theorem

Let $f(x|\theta)$ be a PMF or PDF of X and let $\mathscr{X} = \{x : f(x|\theta) > 0 \text{ for some } \theta\}$. Suppose $T(X)$ has the property that for every $x,y \in \mathscr{X}$ there exists a nonzero, finite number $k = k(x,y)$ such that

$$f(x|\theta) = k(x,y) f(y|\theta)$$

for all θ if and only if $T(x) = T(y)$. Then T is minimal sufficient.

To use this result, you need to show that

(i) If $T(x) = T(y)$ then $k(x,y)$ exists.

(ii) If $k(x,y)$ exists, then $T(x) = T(y)$.

If $\{x : f(x|\theta) > 0\}$ does not depend on θ, then we need to show that for all $x,y \in \mathscr{X}$

$$\frac{f(x|\theta)}{f(y|\theta)}$$

is constant in θ if and only if $T(x) = T(y)$. That is, we need to show

(i) If $T(x) = T(y)$ then $\frac{f(x|\theta)}{f(y|\theta)}$ is constant.

(ii) If $\frac{f(x|\theta)}{f(y|\theta)}$ is constant, then $T(x) = T(y)$.

If T is suf£cient, then $T(x) = T(y) = t$ implies

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{g(t|\theta)h(x)}{g(t|\theta)h(y)} = \frac{h(x)}{h(y)}$$

which is constant in θ. So

(i) is suf£ciency

(ii) is minimality

Examples

1. $X_1,\ldots,X_n$ *i.i.d.* $N(\theta,1)$

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{\exp\{-\frac{1}{2}\sum(x_i-\theta)^2\}}{\exp\{-\frac{1}{2}\sum(y_i-\theta)^2\}} = \exp\{\theta(\sum x_i - \sum y_i)\}k(x,y)$$

   If $\sum x_i = \sum y_i$ then this is constant in θ.
   If $\sum x_i \neq \sum y_i$ then this is not constant in θ.
   So $T(X) = \sum X_i$ is minimal suf£cient for θ.

2. $X_1,\ldots,X_n$ *i.i.d.* $N(\mu,\sigma^2)$, $\theta = (\mu,\sigma^2)$.

$$\begin{aligned}\frac{f(x|\theta)}{f(y|\theta)} &= \frac{\exp\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\}}{\exp\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\}}\\ &= \exp\left\{\frac{1}{2\sigma^2}\left(\sum y_i^2 - \sum x_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum x_i - \sum y_i\right)\right\}\end{aligned}$$

   If $\sum x_i = \sum y_i$ and $\sum x_i^2 = \sum y_i^2$, then this is constant in θ.
   If $\sum x_i \neq \sum y_i$ or $\sum x_i^2 \neq \sum y_i^2$, then this is not constant in θ.
   So $T(X) = (\sum X_i, \sum X_i^2)$ is minimal suf£cient for θ.
   So is $(\overline{X}, S^2)$.

3. $X_1,\ldots,X_n$ *i.i.d.* $U[0,\theta]$, $\Theta = (0,\infty)$.

$$f(x|\theta) = \frac{1}{\theta^n}1_{[0,\theta]}(x_{(n)})$$

   for $x \in \mathscr{X} = [0,\infty)^n$.
   If $x_{(n)} = y_{(n)}$, $x,y \in \mathscr{X}$, then $f(x|\theta) = f(y|\theta)$ for all $\theta \in \Theta$.

If $x_{(n)} \neq y_{(n)}$, say $x_{(n)} < y_{(n)}$, then for $\theta \in (x_{(n)}, y_{(n)})$ we have $f(x|\theta) > 0$ and $f(y|\theta) = 0$. No £nite, nonzero k can make these equal.

So $T(X) = X_{(n)}$ is minimal suf£cient.

4. $X_1, \ldots, X_n$ *i.i.d.* $U[\theta, \theta+1]$, $\mathscr{X} = \mathbb{R}^n$, $\Theta = \mathbb{R}$.

$$f(x|\theta) = \prod 1_{[\theta,\theta+1]}(x_i) = 1_{[\theta,\infty)}(x_{(1)}) 1_{(-\infty,\theta+1]}(x_{(n)})$$

If $x_1 = y_{(1)}$ and $x_{(n)} = y_{(n)}$, then $f(x|\theta) = f(y|\theta)$ for allθ.

If $x_{(1)} \neq y_{(1)}$ or $x_{(n)} \neq y_{(n)}$, then for some θ one of $f(x|\theta)$ and $f(y|\theta)$ is positive and the other zero, so no nonzero, £nite multiplier k can make them equal.

So $T(X) = (X_{(1)}, X_{(n)})$ is minimal suf£cient for θ.

5. $X_1, \ldots, X_n$ *i.i.d.*

$$f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{j=1}^{k} w_j(\theta) t_j(x)\right\}$$

Let $T_j(x) = \sum_{i=1}^{n} t_j(x_i)$. Then

$$\frac{f(x_1, \ldots, x_n|\theta)}{f(y_1, \ldots, y_n|\theta)} = \frac{\prod h(x_i)}{\prod h(y_i)} \exp\left\{\sum_{j=1}^{k} w_j(\theta)(T_j(x) - T_j(y))\right\}$$

If $T_j(x) = T_j(y)$ for $j = 1, \ldots, k$, then this is constant in θ.

Suppose the w_j have the property that

$$\sum_{j=1}^{k} w_j(\theta) a_i$$

is constant in θ is and only if $a_1 = \cdots = a_j = 0$. This is true if the set

$$\{(w_1(\theta), \ldots, w_k(\theta)) : \theta \in \Theta\}$$

contains an open set. Then the ratio is constant in θ only if $T_j(x) = T_j(y)$ for all j.

So under this condition on the w_j, $(T_1(X), \ldots, T_k(X))$ in minimal suf£cient for θ.

Homework

Problem 6.9
Problem 6.10

Due Friday, January 31, 2003.

Wednesday, January 29, 2003

Ancillary Statistics

Definition

A statistic is ancillary if its distribution does not depend on θ.

Example

$X_1,\ldots,X_n$ $i.i.d.$ $U[0,\theta]$. $S(X) = X_{(1)}/X_{(n)}$ is ancillary.

Example

Suppose $\theta \in \Theta = \mathbb{R}$ is a location parameter, $f(x|\theta) = f(x_1-\theta,\ldots,x_n-\theta)$, and S is location invariant, i.e.

$$S(x_1,\ldots,x_n) = S(x_1+c,\ldots,x_n+c)$$

for all c. Then S is ancillary for θ. To see this, let

$$Z \sim f(x_1,\ldots,x_n)$$

Then

$$Z+\theta = (Z_1+\theta,\ldots,Z_n+\theta) \sim X$$

and

$$S(X) = S(Z+\theta) = S(Z)$$

So the distribution of S does not depend on θ. Special cases:

$$S(X) = (X_1-\overline{X},\ldots,X_n-\overline{X})$$
$$S(X) = X_1-\widetilde{X},\ldots,X_n-\widetilde{X})$$
$$S(X) = \frac{1}{n-1}\sum(X_i-\overline{X})^2$$

Similar results hold for location-scale families. For a location-scale family,

$$\left(\frac{X_1-\overline{X}}{S},\ldots,\frac{X_n-\overline{X}}{S}\right)$$

is ancillary.

Ancillary statistics are often used for model criticism.

Completeness

Let $f(t|\theta)$ be a family of PDF's or PMF's for a statistic $T(X)$. The family is called complete if $E_\theta[|g(T)|] < \infty$ and

$$E_\theta[g(T)] = 0$$

for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. If the family of PDF's or PMF's of T is complete, then T is called complete.

Example

Suppose $T \sim \text{Binomial}(n, p)$, $0 < p < 1$. Suppose

$$E_p[g(T)] = \sum_{t=0}^{n} g(t)\binom{n}{t} p^t (1-p)^{n-t} = 0$$

for all $p \in (0,1)$. Then

$$\sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t = 0$$

for all $p \in (0,1)$, or

$$\sum_{t=0}^{n} g(t)\binom{n}{t} x^t = 0$$

for all $x > 0$. A polynomial is zero on an open interval if and only if all its coef£cients are zero. So $g(t) = 0$ for $t = 0, \ldots, n$.

Example

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $U[0,\theta]$, $T(X) = X_{(n)}$, and so

$$f(t|\theta) = \begin{cases} \frac{n}{\theta^n} t^{n-1} & 0 < t < \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose

$$\int_0^\theta \frac{n}{\theta^n} t^{n-1} g(t) dt = 0$$

for all $\theta > 0$. Then

$$\int_0^\theta t^{n-1} g(t) dt = 0$$

for all $\theta > 0$. If g is continuous, this implied that $t^{n-1}g(t) = 0$ for all $t > 0$ and hence $g(t) = 0$ for all $t > 0$. If g is not continuous but measurable, it implies that $g(t) = 0$ for "almost all" $t > 0$. So $P_\theta(g(X_{(n)}) = 0) = 1$ for all $\theta > 0$.

Example

Suppose $X_1,\dots,X_n$ are *i.i.d.* from an exponential family

$$f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\right\}$$

Suppose the set

$$\{(w_1(\theta),\dots,w_k(\theta)) : \theta \in \Theta\}$$

contains an open set in $\mathbb{R}^k$. Then $(T_1,\dots,T_k)$ with

$$T_j = \sum_{i=1}^{n} t_j(X_i)$$

is complete.

Basu's Theorem

If $T(X)$ is complete and sufficient and $S(X)$ is ancillary, then $T(X)$ and $S(X)$ are independent.

Proof

Let S be ancillary and T complete and sufficient. For any set A let

$$g(t) = P(S(X) \in A|T(X) = t) - P(S(X) \in A)$$

Since T is sufficient, $P(S(X) \in A|T(X) = t)$ does not depend on θ. Since S is ancillary, $P(S(X) \in A)$ does not depend on θ. So $g(t)$ does not depend on θ. But

$$\begin{aligned} E_\theta[g(T)] &= E[P(S \in A|T) - P(S \in A)] \\ &= E[P(S \in A|T)] - P(S \in A) \\ &= P(S \in A) - P(S \in A) = 0 \end{aligned}$$

for all θ. Since T is complete, $g(T) = 0$ almost surely, and so $P(S \in A|T) = P(S \in A)$ almost surely. This holds for all A, so S, T are independent. □

Examples

1. Suppose $X_1,\dots,X_n$ are *i.i.d.* $U[0.\theta]$, $\Theta = (0,\infty)$. Then $U_i = X_i/\theta \sim U[0,1]$. So $X_{(1)}/X_{(n)} = U_{(1)}/U_{(n)}$ is ancillary. Since $X_{(n)}$ is complete and sufficient, $X_{(1)}/X_{(n)}$ and $X_{(n)}$ are independent.

2. Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $N(\theta, 1)$ with $\Theta = \mathbb{R}$. Then $Z_i = X_i - \theta \sim N(0,1)$ and

$$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2 = \frac{1}{n-1} \sum (Z_i - \overline{Z})^2$$

   is ancillary. $\overline{X}$ is minimal suf£cient and complete. So $\overline{X}$ and S^2 are independent.

3. Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $N(\mu, \sigma^2)$. Let $Z_i = (X_i - \mu)/\sigma \sim N(0,1)$ and

$$C_i = (X_i - \overline{X})/S = (Z_i - \overline{Z})/S_Z$$

   Then $(C_1, \ldots, C_n)$ is ancillary. $\overline{X}, S^2$ is suf£cient and complete. So $(C_1, \ldots, C_n)$ is independent of $(\overline{X}, S^2)$.

Homework

Problem 6.14
Problem 6.20

Due Friday, January 31, 2003.

Friday, January 31, 2003

Likelihood

Definition

Let $f(x|\theta)$ be the joint PMF or PDF of X. Then given $X = x$ is observed, the *likelihood function* is the function of θ,

$$L(\theta|x) = f(x|\theta)$$

Informally, if $L(\theta_1|x) > L(\theta_2|x)$ then there is more support in the data for θ_1 than for θ_2.

Likelihood Principle

If x, y are such that $L(\theta|x) = c(x,y)L(\theta|y)$ for all θ and for some $c(x,y) \neq 0$, then x and y should lead to the same inferences about θ.

Stronger version: Two experiments that lead to the same likelihood function should lead to the same inferences about θ.

It can be argued that $L(\theta|x)$ is essentially a minimal sufficient statistic, or that $T(x)$ is minimal sufficient if and only if it is a one-to-one function of the likelihood function.

The likelihood principle follows from the sufficiency principle and the conditionality principle.

Conditionality Principle

Consider two situations:

1. Experiment E_1 is performed.
2. A fair coin is flipped to choose between E_1 and E_2, and E_1 is chosen and performed.

The two should lead to the same conclusions.

Examples

1. Suppose $X \sim$ Negative Binomial(r, p),

$$f(x|p) = \binom{r-1}{x-1} p^r (1-p)^{x-r}$$

for $x = r, r+1, \ldots$. Suppose $r = 4, x = 7$. Then

$$L(p|r,x) \propto p^4(1-p)^3$$

Common approach: Estimate p as

$$\hat{p} = \frac{r}{x} = \frac{4}{7}$$

and look at the sampling distribution of $\hat{p}$.

2. Suppose $X \sim \text{Binomial}(n,p)$

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, \ldots, n$. Suppose $x = 4, n = 7$. Then

$$L(p|n,x) \propto p^4(1-p)^3$$

Common approach: Estimate p as

$$\hat{p} = \frac{x}{n} = \frac{4}{7}$$

and look at the sampling distribution of $\hat{p}$.

The estimates, likelihood functions are the same. Sampling distributions of the estimators and interval estimates based on these sampling distributions are not. (They are close for large r, n.)

Some feel the conditionality principle implies that all inference should be done conditionally on any ancillary statistic (the random choice of experiment is ancillary).

There are ways of defining maximal ancillary statistics.

Many feel the conditionality and sufficiency principles are compelling.

Together they imply the likelihood principle.

Many standard frequentist methods violate the likelihood principle (often not by much, but the difference can be substantial in sequential experiments).

A fully Bayesian approach automatically satisfies the likelihood principle.

How excited should you get about these observations?

References

HELLAND, INGE S. (1995), "Simple counterexamples against the conditionality principle," *The American Statistician* 49(4), 351–356.

LIANG, K-Y, AND ZEGER, S. L. (1995) “Inference based on estimating functions in the presence of nuisance parameters,” *Statistical Science* 10(2), 158–172.

REID, N. (1995), “The roles of conditioning in inference,” *Statistical Science* 10(2), 138–157.

Week 3

Monday, February 3, 2003

Point Estimation

A standard “first stab” at fitting a model to data is to ask: What value of θ is the “best guess” for the “true” value of θ based on the data.

We will look at

1. Methods for finding estimators.
2. Methods for deciding how good an estimator is.

For now, a *point estimator* of θ is any statistic $T(X)$ you decide you want to use to produce a guess for the value of θ.

Calling a statistc a point estimator says *nothing* about its quality or appropriateness.

Method of Moments

The oldest method of finding point estimators is the method of moments.

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $f(x|\theta_1, \ldots, \theta_k)$ and that we have k functions $M_1, \ldots, M_k$ such that

$$\mu_{M_j} = E[M_j(X)] = \mu_{M_j}(\theta_1, \ldots, \theta_k)$$

are known. Let

$$m_j = \frac{1}{n}\sum_{i=1}^{n} M_j(X_i)$$

By the Law of Large Numbers, $m_j \approx \mu_{M_j}$ for large n.

Method of Moments: Set

$$\begin{aligned} m_1 &= \mu_{M_1}(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu_{M_k}(\theta_1, \ldots, \theta_k) \end{aligned}$$

and solve for $\theta_1, \ldots, \theta_k$ to get $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k$.

Usually we try to use

$$M_j(x) = x^j$$

This choice leads to the traditional method of moments.

But sometimes other choices are used.

Examples

1. Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $N(\mu, \sigma^2)$. Then

$$\begin{aligned} \frac{1}{n}\sum X_i &= \mu \\ \frac{1}{n}\sum X_i^2 &= \mu^2 + \sigma^2 \end{aligned}$$

produces

$$\begin{aligned} \widetilde{\mu} &= \overline{X} \\ \widetilde{\sigma}^2 &= \frac{1}{n}\sum(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \end{aligned}$$

This is reasonable; $\widetilde{\sigma}^2$ may be a bit different from what one might expect.

2. Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $U[0, \theta]$. Then

$$\overline{X} = \frac{\theta}{2}$$

yields $\widetilde{\theta} = 2\overline{X}$.

Problem: Can have $X_{(n)} > 2\overline{X}$—in this case we know $\widetilde{\theta}$ is too small.

A better estimator would insure that this kind of inconsistency does not occur.

The method of moments is often easy to use.

The choice of $M_1, \ldots, M_k$ is arbitrary; the best choice is not obvious.

The estimators produced are often not ideal.

The basic idea is not easy to extend to non-$i.i.d$ data.

The method of moments is often useful as a £rst step.

Homework

Problem 7.6
Problem 7.11

Due Friday,February 7 , 2003.

Wednesday, February 5, 2003

Maximum Likelihood

Definition

Let $L(\theta|x) = f(x|\theta)$ be the likelihood function for an observed $X = x$. For each x, let $\widehat{\theta}(x)$ be the value that maximizes $L(\theta|x)$ as a function of θ with x held fixed. Then $\widehat{\theta}(x)$ is a maximum likelihood estimator of θ.

Notes:

- $\widehat{\theta} \in \Theta$ by construction.
- If $L(\theta'|x) = f(x|\theta') = 0$, then $\widehat{\theta} \neq \theta'$.
- $\widehat{\theta}$ may not exist.
- $\widehat{\theta}$ may not be unique.
- $\widehat{\theta}$ may exist and be unique but be hard to find.

Often we can find the MLE by

- differentiating and finding roots
- checking for a global maximum

It is almost always easier to maximize

$$\log L(\theta|x)$$

instead of $L(\theta|x)$ (and equivalent). As a convention, $\log 0 = -\infty$.

Examples

1. Suppose $X_1, \ldots, X_n$ are *i.i.d.* $N(\theta, 1)$ Then

$$\begin{aligned} L(\theta|x) &= \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i-\theta)^2\right\} \\ \log L(\theta|x) &= \text{const} - \frac{1}{2}\sum(x_i-\theta)^2 \\ \frac{d}{d\theta}\log L(\theta|x) &= \sum(x_i-\theta) = \sum x_i - n\theta = n(\overline{x}-\theta) \end{aligned}$$

The unique root is $\widehat{\theta} = \overline{x}$.

The likelihood is continuously differentiable, and $\theta \to \pm\infty$ implies $\log L(\theta|x) \to -\infty$. Therefore a global maximum exists, every global maximum is an interior local maximum and thus a root of the derivative. Since there is only one such root, $\widehat{\theta} = \overline{X}$ is the unique global maximizer.

Alternative:

$$\frac{d^2}{d\theta^2}\log L = -n < 0$$

for *all* θ, so $\log L$ is strictly concave and a zero of the derivative is a unique global maximum.

2. Suppose $X_1,\dots,X_n$ are *i.i.d.* $N(\mu,\sigma^2) = N(\theta_1,\theta_2)$.

$$\begin{aligned}
L(\theta|x) &= \frac{1}{(2\pi\theta_2)^{n/2}}\exp\left\{-\frac{1}{2\theta_2}\sum(x_i-\theta_1)^2\right\}\\
\log L(\theta|x) &= \text{const} - \frac{n}{2}\log\theta_2 - \frac{1}{2\theta_2}\sum(x_i-\theta_1)^2\\
\frac{\partial}{\partial\theta_1} &= \frac{1}{\theta_2}\sum(x_i-\theta_1)\\
\frac{\partial}{\partial\theta_2} &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum(x_i-\theta_1)^2
\end{aligned}$$

Likelihood equations:

$$\begin{aligned}
0 &= \frac{1}{\theta_2}\sum(x_i-\theta_1)\\
0 &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum(x_i-\theta_1)^2
\end{aligned}$$

Solution:

$$\begin{aligned}
\widehat{\theta}_1 &= \overline{x}\\
\widehat{\theta}_2 &= \frac{1}{n}\sum(x_i-\overline{x})^2 = \frac{n-1}{n}s^2
\end{aligned}$$

To see that this is a global maximum:

- for each θ_2, $\widehat{\theta}_1 = \overline{x}$ maximizes $L(\theta_1,\theta_2|x)$ over θ_1.
- for $\theta_1 = \overline{x}$, $L(\overline{x},\theta_2|x)$ is strictly concave.

Global second derivative conditions are harder.

More MLE Examples

Examples

1. $X_1,\dots,X_n$ *i.i.d.* Bernoulli(p).

$$\begin{aligned} L(p|x) &= p^{\sum x_i}(1-p)^{n-\sum x_i} \\ \log L(p|x) &= \sum x_i \log p + (n - \sum x_i)\log(1-p) \end{aligned}$$

   Differentiate, set to zero for $p \in (0,1)$:

$$\begin{aligned} 0 &= \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} && \text{or} \\ 0 &= (1-p)\sum x_i - p(n-\sum x_i) && \text{or} \\ 0 &= \sum x_i - np && \text{so} \\ \hat{p} &= \frac{1}{n}\sum x_i \end{aligned}$$

   This is an interior local maximum if $0 < \sum x_i < n$ and a global maximum.

   If $\sum x_i = 0$, then $L(p|x)$ is decreasing, so $\hat{p} = 0$.

   If $\sum x_i = n$, then $L(p|x)$ is increasing, so $\hat{p} = 1$.

   If $\Theta = (0,1)$, then $\hat{p}$ does not exist in these boundary cases.

   If $\Theta = [0,1]$, then $\hat{p}$ exists for all samples, and in all cases $\hat{p} = \frac{1}{n}\sum x_i$.

2. $X_1,\dots,X_n$ *i.i.d.* $U[0,\theta]$.

$$L(\theta|x) = \frac{1}{\theta^n} 1_{[0,\theta]}(x_{(n)})$$

   This is maximized at $\theta = x_{(n)}$, so the MLE is $\widehat{\theta} = X_{(n)}$.

   This is a better estimator than the MM estimator, but we know it has to be a bit too small.

   Suppose we use $U(0,\theta)$ instead. Then the MLE, strictly speaking, does not exist:

MLE Invariance

Suppose we are interested in a function $\tau(\theta)$ and $\widehat{\theta}$ is the MLE of θ. Is $\tau(\widehat{\theta})$ the MLE of $\tau(\theta)$?

If τ is one-to-one, then the answer is yes: We can write

$$L^*(t|x) = L(\tau^{-1}(t)|x)$$

and if $\widehat{\theta}$ maximizes L, then $\widehat{t} = \tau(\widehat{\theta})$ maximizes L^*.

If τ is not one-to-one, then it is not clear what "the MLE of $\tau(\theta)$" really means—MLE's are de£ned assuming θ uniquely identi£es $f(x|\theta)$. If τ is not one-to-one, then we may have several θ's, with possibly different values of $L(\theta|x)$, that have the same value of $\tau(\theta)$.

Solution: *De£ne* $L^*(t|x)$, the induced (or pro£le) likelihood, as

$$L^*(t|x) = \sup\{L(\theta|x) : \tau(\theta) = t\}$$

Now let $\widehat{t}$ be the value that maximizes L^*. Then

$$\begin{aligned} L^*(\widehat{t}|x) &= \sup_t \{L(\theta|x) : \tau(\theta) = t\} \\ &= \sup_\theta L(\theta|x) \\ &= L(\widehat{\theta}|x) \end{aligned}$$

and

$$\begin{aligned} L(\widehat{\theta}|x) &= \sup\{L(\theta|x) : \tau(\theta) = \tau(\widehat{\theta})\} \\ &= L^*(\tau(\widehat{\theta})|x) \end{aligned}$$

So $\tau(\widehat{\theta})$ is an MLE of $\tau(\theta)$ *based on this de£nition*.

The property that

$$\widehat{\tau(\theta)} = \tau(\widehat{\theta})$$

is called the invariance property of the MLE.

Homework

Problem 7.13
Problem 7.14

Due Friday, February 7, 2003.

Friday, February 7, 2003

Bayes Estimators

The Bayesian approach uses a prior distribution and a likelihood to compute a posterior distribution and bases all inferences on the posterior distribution.

We can also use a posterior distribution to produce point estimators.

The posterior mean is a common choice.

The median is another possibility.

Example

Let $X_1,\ldots,X_n$ be $i.i.d.$ Bernoulli(p). Suppose we use a prior that is Beta(α,β). Then the posterior is

$$\begin{aligned} f(p|x) &= \frac{f(x|p)f(p)}{f(x)} \propto f(x|p)f(p) \\ &\propto p^{\sum x_i}(1-p)^{n-\sum x_i} p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+\sum x_i-1}(1-p)^{\beta+n-\sum x_i-1} \\ &\sim \text{Beta}\left(\alpha+\sum x_i, \beta+n-\sum x_i\right) \end{aligned}$$

So

$$E[p|x] = \frac{\alpha+\sum x_i}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{\alpha+\beta+n} + \frac{1}{n}\sum x_i \frac{n}{\alpha+\beta+n}$$

For $\alpha,\beta \approx 0$, $E[p|x] \approx \frac{1}{n}\sum x_i$.

For $\alpha,\beta > 0$, $0 < E[p|x] < 1$.

Conjugate Families

Let $\mathscr{F} = \{f(x|\theta) : \theta \in \Theta\}$ be a class of PMF's or PDF's. A collection Π of prior distributions on Θ is conjugate for $\mathscr{F}$ if the posterior distribution is in Π for any prior distribution in Π and any $x \in \mathscr{X}$.

The family $\Pi = \{f(p) = \text{Beta}(\alpha,\beta) : \alpha,\beta > 0\}$ is conjugate for

$$\begin{aligned} \mathscr{F} &= \{n\ i.i.d.\ \text{Bernoulli}(p)\} \\ &= \{n\ i.i.d.\ \text{Geometric}(p)\} \\ &= \{\text{Binomial}(n,p)\} \\ &= \{\text{Negative Binomial}(n,p)\} \end{aligned}$$

Examples

1. $X_1,\ldots,X_n$ *i.i.d.* Poisson(λ), $\lambda \sim \text{Gamma}(\alpha,\beta)$

$$\begin{aligned} f(\lambda|x) &\propto f(x|\lambda)f(\lambda) \\ &\propto \lambda^{\sum x_i} e^{-n\lambda} \lambda^{\alpha-1} e^{\lambda/\beta} \\ &= \lambda^{\alpha+\sum x_i - 1} e^{-\lambda(n+1/\beta)} \\ &\sim \text{Gamma}(\alpha + \sum x_i, (n+1/\beta)^{-1}) \end{aligned}$$

So

$$E[\lambda|x] = \frac{\alpha+\sum x_i}{n+1/\beta} = \alpha\beta\frac{1}{1+n\beta} + \bar{x}\frac{n\beta}{1+n\beta}$$

The family $\Pi = \{f(\lambda) = \text{Gamma}(\alpha,\beta) : \alpha,\beta > 0\}$ is conjugate for

$$\begin{aligned} \mathscr{F} &= \text{Poisson } i.i.d. \\ &= \text{Poisson Process} \\ &= \text{Exponential } i.i.d., \text{ mean } 1/\lambda \end{aligned}$$

2. $X_1,\ldots,X_n$ *i.i.d.* $N(\theta,\sigma^2)$, σ^2 *known*, $\theta \sim N(\mu,\tau^2)$.

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)f(\theta) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\ &\propto \exp\left\{-\frac{n}{2\sigma^2}\theta^2 - \frac{1}{2\tau^2}\theta^2 + \frac{\theta}{\sigma^2}\sum x_i + \frac{\theta}{\tau^2}\mu\right\} \end{aligned}$$

This is of the form

$$\exp\left\{-\frac{1}{2}\frac{(\theta-a)^2}{b}\right\}$$

with

$$\begin{aligned} \frac{a}{b} &= \frac{1}{\sigma^2}\sum x_i + \frac{\mu}{\tau^2} \\ \frac{1}{b} &= \frac{n}{\sigma^2} + \frac{1}{\tau^2} \end{aligned}$$

So $f(\theta|x) \sim N(a,b)$, with

$$\begin{aligned} a &= \frac{\frac{1}{\sigma^2}\sum x_i + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} = \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu \\ b &= \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n} \end{aligned}$$

Homework

Problem 7.22
Problem 7.23

Due Friday, February 14, 2003.

Week 4

Monday, February 10, 2003

Methods of Evaluating Estimators

Mean Squared Error

A useful measure of the quality of an estimator W of a quantity $\tau(\theta)$ is the *mean square error (MSE)*:

$$\text{MSE}(W,\theta) = E_\theta[(W-\tau(\theta))^2]$$

Notes:

MSE(W,θ) measures the average error.

Other "loss functions" are possible but are less convenient.

MSE(W,θ) is a function of θ.

We can decompose MSE(W,θ) into

$$\begin{aligned}\text{MSE}(W,\theta) = E_\theta[(W-\tau(\theta))^2] &= \text{Var}_\theta(W) + (E_\theta[W]-\tau(\theta))^2\\ &= \text{Var}_\theta(W) + \text{Bias}(W,\theta)^2\end{aligned}$$

Bias

The bias of W is

$$\text{Bias}(W,\theta) = E_\theta[W]-\tau(\theta)$$

W is called unbiased if Bias$(W,\theta)=0$ for all θ.

So there are two components to the MSE: bias and variance. Sometimes we can trade off one against the other.

Example

Let $X_1, \ldots, X_n$ be *i.i.d.* $N(\mu, \sigma)$.

$\overline{X}$ and S^2 are unbiased for μ and σ^2. So

$$\begin{aligned}\text{MSE}(\overline{X}, \mu, \sigma^2) &= \text{Var}(\overline{X}) = \frac{\sigma^2}{n}\\ \text{MSE}(S^2, \mu, \sigma^2) &= \text{Var}(S^2) = \frac{2\sigma^4}{n-1}\end{aligned}$$

The MLE is $\widehat{\mu} = \overline{X}, \widehat{\sigma}^2 = \frac{n-1}{n}S^2$. The MSE of $\widehat{\sigma}^2$ is

$$\begin{aligned}\text{MSE}(\widehat{\sigma}^2, \mu, \sigma^2) &= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} + \frac{1}{n^2}\sigma^4\\ &= \frac{\sigma^4}{n^2}(2(n-1)+1) = \frac{\sigma^4}{n^2}(2n-1)\end{aligned}$$

But

$$\frac{2n-1}{n^2} < \frac{2}{n-1}$$

for $n \geq 1$, so

$$\text{MSE}(\widehat{\sigma}^2) < \text{MSE}(S^2)$$

Often a variance-bias tradeoff is useful.

Finding Optimal Estimators?

Ideally, we would like to £nd an estimator W^* such that

$$\text{MSE}(W^*, \theta) \leq \text{MSE}(W, \theta)$$

for all θ and all other estimators W.

Unfortunately, this is usually impossible. Take

$$W \equiv 7$$

Then

$$\begin{aligned}\text{MSE}(W, \theta) &= \text{Var}(W) + (E[W] - \tau(\theta))^2\\ &= 0 + (7 - \tau(\theta))^2 = (7 - \tau(\theta))^2\end{aligned}$$

which is zero if $\tau(\theta) = 7$.

This is not a “reasonable” estimator.

“Reasonable” estimators will have $\text{Var}_\theta(W) > 0$ for most if not all θ.

We need to restrict ourselves to “reasonable” estimators to develop a nice theory. “Reasonable” means the estimator must make some effort to “track” the target. This needs to be given a precise de£nition to be useful.

A few possibilities:

Unbiasedness—require $E[W] = \tau(\theta)$ for all θ.

Invariance—shifting $\tau(\theta)$ by a shifts W by a.

Consistency—$W_n \xrightarrow{P} \tau(\theta)$ for all θ.

The cleanest theory is available for unbiased estimation.

Requiring (exact) unbiasedness can be very restrictive. It can (though it usually doesn’t) lead to really stupid estimators. An example where this is the case:

Example

Suppose $X \sim \text{Poisson}(\theta)$ and

$$\tau(\theta) = e^{-2\theta}$$

Suppose W is unbiased for $\tau(\theta)$. Then

$$e^{-2\theta} = \sum_{k=0}^{\infty} w(k) \frac{\theta^k}{k!} e^{-\theta}$$

for all $\theta > 0$, or

$$e^{-\theta} = \sum_{k=0}^{\infty} w(k) \frac{\theta^k}{k!}$$

But

$$e^{-\theta} = \sum_{k=0}^{\infty} (-1)^k \frac{\theta^k}{k!}$$

and power series are unique on their radius of convergence. So we must have $w(k) = (-1)^k$. Thus the *only* unbiased estimator of $\tau(\theta) = e^{-2\theta}$ is

$$W = \begin{cases} -1 & \text{if } X \text{ is odd} \\ +1 & \text{if } X \text{ is even} \end{cases}$$

This is a pretty silly estimator.

Homework

Problem 7.33

Due Friday, February 14, 2003.

Wednesday, February 12, 2003

Definition

An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_\theta[W^*] = \tau(\theta)$ for all θ, and for any other estimator W with $E_\theta[W] = \tau(\theta)$ for all θ we have

$$\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$$

for all θ. W^* is also called a uniformly minimum variance unbiased estimator (UMVUE).

Finding UMVUE's by trial and error is hard. We will look at two approaches:

1. Find a lower bound on the best possible variance (Cramér-Rao lower bound). If an estimator W^* achieves this lower bound, then it must be UMVUE. (We can characterize when this is possible.)
2. Show that there is a restricted class $\mathscr{C}$ of estimators such that for any unbiased W there is a $W' \in \mathscr{C}$ that is at least as good.

   Show that under some conditions $\mathscr{C}$ has only one element.

   Then if $W \in \mathscr{C}$ is that element, W must be the UMVUE (Lehmann-Scheffé approach)

Cramer-Rao Lower Bound

Let $X_1, \ldots, X_n$ have joint PDF $f(x|\theta)$ for $\theta \in \Theta$, an open subset of $\mathbb{R}$, and let W be any estimator such that $E_\theta[W]$ is differentiable with respect to θ over Θ. Suppose that $f(x|\theta)$ satisfies

$$\frac{d}{d\theta}\int\cdots\int h(x) f(x|\theta) dx_1 \cdots dx_n = \int\cdots\int h(x)\frac{\partial}{\partial\theta} f(x|\theta) dx_1 \cdots dx_n$$

for any $h(x)$ with $E_\theta[|h(X)|] < \infty$ for all θ. Then

$$\mathrm{Var}_\theta(X) \geq \frac{\left(\frac{d}{d\theta}E_\theta[W]\right)^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}$$

Variations:

For discrete data, replace $\int$ by $\sum$.

For Θ an open subset of $\mathbb{R}^m$ and W real-valued,

$$\mathrm{Var}_\theta(X) \geq \underbrace{\nabla E_\theta[W]}_{1\times m} \underbrace{\left(E_\theta\left[\frac{\partial}{\partial\theta_i}\log f(X|\theta)\frac{\partial}{\partial\theta_j}\log f(X|\theta)\right]\right)_{ij}^{-1}}_{m\times m} \underbrace{\nabla E_\theta[W]^T}_{m\times 1}$$

Cramer-Rao Lower Bound

Proof

The proof uses the Cauchy-Schwartz inequality in the form

$$\text{Var}(X) \geq \frac{\text{Cov}(X,Y)^2}{\text{Var}(Y)}$$

with $X = W$ and $Y = \frac{\partial}{\partial\theta}\log f(X|\theta)$. First,

$$\begin{aligned}
E_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] &= E_\theta\left[\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right] \\
&= \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)dx \\
&= \int \frac{\partial}{\partial\theta} f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \frac{\partial}{\partial\theta}1 = 0
\end{aligned}$$

Similarly,

$$\begin{aligned}
\text{Cov}\left(W, \frac{\partial}{\partial\theta}\log f(X|\theta)\right) &= \int W(x)\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)dx \\
&= \int W(x)\frac{\partial}{\partial\theta} f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}\int W(x) f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}E_\theta[W]
\end{aligned}$$

and

$$\text{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$$

So from the Cauchy-Schwartz inequality,

$$\text{Var}_\theta(W) \geq \frac{\left(\frac{\partial}{\partial\theta}E_\theta[W]\right)^2}{E\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}$$

□

The quantity in the denominator is called the *Fisher information* for θ,

$$\begin{aligned} I_n(\theta) &= E\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] & \theta \in \mathbb{R} \\ &= \left(E\left[\frac{\partial}{\partial\theta_i}\log f(X|\theta)\frac{\partial}{\partial\theta}\log f(X|\theta)\right]\right)_{ij} & \theta \in \mathbb{R}^m \end{aligned}$$

If $X_1,\dots,X_n$ are $i.i.d.$ from f, then

$$\log f(x_1,\dots,x_n|\theta) = \sum f(x_i|\theta)$$

and therefore

$$I_n(\theta) = nI_1(\theta)$$

Equality in the CRLB

Equality in the CRLB occurs if and only if there is equality in the Cauchy-Schwartz inequality. This happens if and only if

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = a(\theta) + b(\theta)W(x)$$

for some $a(\theta), b(\theta)$. This implies

$$\begin{aligned} \log f(x|\theta) &= C(x) + B(\theta)W(x) + A(\theta) \\ f(x|\theta) &= \exp\{C(x)\}\exp\{A(\theta)\}\exp\{B(\theta)W(x)\} \end{aligned}$$

So f is an exponential family with suf£cient statistic $W(X)$.

Conversely, if

$$f(x|\theta) = c(\theta)h(x)\exp\{t(x)w(\theta)\}$$

then

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = \frac{c'(\theta)}{c(\theta)} + t(x)w'(\theta)$$

So

$$E[t(X)] = -\frac{c'(\theta)}{c(\theta)w'(\theta)}$$

and $t(X)$ is a UMVUE for $-\frac{c'(\theta)}{c(\theta)w'(\theta)}$.

An unbiased estimator is called *ef£cient* if it achieves the CRLB for all θ.

Computing the Fisher Information

Suppose

$$\int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x|\theta)dx = 0$$

for all θ. Then

$$\begin{aligned} E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] &= -\int \frac{ff''-f'f'}{f^2} f dx \\ &= -\int f'' dx + \int \left(\frac{f'}{f}\right)^2 f dx \\ &= I(\theta) \end{aligned}$$

or

$$I(\theta) = -\left(E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X|\theta)\right]\right)_{ij}$$

Differentiability assumptions hold for all exponential families.

Examples

1. $X_1,\dots,X_n$ *i.i.d.* $N(\theta,1)$.

$$\begin{aligned} \log f(x|\theta) &= \text{const} - \frac{1}{2}\sum(x_i-\theta)^2 \\ \frac{\partial}{\partial\theta}\log f(x|\theta) &= \sum(x_i-\theta) = n(\overline{x}-\theta) \\ I(\theta) &= E[(n(\overline{X}-\theta))^2] = n^2\text{Var}(\overline{X}) = n \\ -\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) &= n \end{aligned}$$

So for W that are unbiased for θ, $\text{Var}(W) \geq 1/n$. So $\overline{X}$ is UMVUE.

2. $X_1,\dots,X_n$ *i.i.d.* Poisson(θ).

$$\begin{aligned} \log f(x|\theta) &= \text{const} + n\overline{x}\log\theta - n\theta \\ \frac{\partial}{\partial\theta}\log f(x|\theta) &= \frac{n\overline{x}}{\theta} - n = \frac{n}{\theta}(\overline{x}-\theta) \\ I(\theta) &= \frac{n^2}{\theta^2}E[(\overline{X}-\theta)^2] = \frac{n^2}{\theta^2}\frac{\theta}{n} = \frac{n}{\theta} \\ -\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) &= \frac{n\overline{x}}{\theta^2} \\ I(\theta) &= \frac{n}{\theta^2}E[\overline{X}] = \frac{n}{\theta} \end{aligned}$$

So if W is unbiased for θ, then $\text{Var}(W) \geq \theta/n$. So $\overline{X}$ is UMVUE of θ.

3. $X_1, \ldots, X_n$ *i.i.d.* $U[0, \theta]$

$$\frac{\partial}{\partial \theta} \int_0^\theta h(x) \frac{1}{\theta} dx \neq \int_0^\theta h(x) \left(-\frac{1}{\theta^2} \right) dx$$

   for all $h(x)$. So the CRLB does not apply.

4. Suppose we want an unbiased estimator of θ^2 for Poisson data. The lower bound is

$$\mathrm{Var}(W) \geq 4\theta^2 \frac{\theta}{n} = \frac{4}{n}\theta^3$$

   Is this attainable? No!

Homework

Problem 7.38
Problem 7.39

Due Friday, February 14, 2003.

Friday, February 14, 2003

Finding Best Unbiased Estimators

Rao-Blackwell Theorem

Let W be an unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Let $\phi(T) = E[W|T]$. Then $\phi(T)$ is an unbiased estimator of $\tau(\theta)$ and

$$\text{Var}_\theta(\phi(T)) \leq \text{Var}_\theta(W)$$

for all θ.

Proof

Since T is sufficient, $E[W|T]$ does not depend on θ. So $\phi(T)$ is a statistic. Furthermore,

$$E_\theta[\phi(T)] = E_\theta[E[W|T]] = E_\theta[W] = \tau(\theta)$$

So $\phi(T)$ is unbiased for $\tau(\theta)$. Finally,

$$\begin{aligned}
\text{Var}_\theta(W) &= \text{Var}_\theta(E[W|T]) + E_\theta[\text{Var}(W|T)] \\
&\geq \text{Var}_\theta(E[W|T]) \\
&= \text{Var}_\theta(\phi(T))
\end{aligned}$$

□

Example

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ Geometric(p). Want a good estimator of p.

An unbiased estimator of p is

$$W = \begin{cases} 1 & \text{if } X_1 = 1 \\ 0 & \text{if } X_1 \neq 1 \end{cases}$$

$T = \sum X_i$ is sufficient.

$$\begin{aligned}
E[W|T=t] &= P(W=1|T=t)\\
&= \frac{P(W=1, \sum_2^n X_i = t-1)}{P(\sum_1^n X_i = t)}\\
&= \frac{p\binom{t-2}{n-2} p^{n-1}(1-p)^{t-n}}{\binom{t-1}{n-1} p^n (1-p)^{t-n}}\\
&= \frac{\binom{t-2}{n-2}}{\binom{t-1}{n-1}}\\
&= \frac{(t-2)!(n-1)!}{(t-1)!(n-2)!}\\
&= \frac{n-1}{t-1}
\end{aligned}$$

So

$$\phi(\sum X_i) = \frac{n-1}{\sum X_i - 1}$$

is unbiased for p and better than W.

It is in fact the UMVUE.

Lehmann-Scheffé Theorem

Let T be a complete, sufficient statistic for θ, and let $\phi(T)$ have expectation $\tau(\theta)$ for all θ. Then $\phi(T)$ is the only function of T with expectation $\tau(\theta)$ for all θ, and it is the UMVUE of $\tau(\theta)$.

Proof

Suppose ϕ' is another function with $E_\theta[\phi'(T)] = \tau(\theta)$. Then

$$E_\theta[\phi(T) - \phi'(T)] = 0$$

for all θ, and so by completeness

$$P_\theta(\phi(T) = \phi'(T)) = 1$$

for all θ.

If W is unbiased for $\tau(\theta)$, then $\phi'(T) = E[W|T]$ is at least as good. But $\phi'(T)$ is unbiased, so $\phi' = \phi$, and thus $\phi(T)$ is at least as good as any unbiased estimator W. □

Examples

1. Suppose $X_1,\ldots,X_n$ are $i.i.d.$ $U[0,\theta]$. Then $W = \frac{n+1}{n}X_{(n)}$ is unbiased for θ. Since it is a function of a complete, sufficient statistic, it is the UMVUE.

2. Suppose $X_1,\ldots,X_n$ are $i.i.d.$ Poisson(θ) and $\tau(\theta) = \theta^2$. Then

$$W = \overline{X}^2 - \frac{\overline{X}}{n} = \overline{X}(\overline{X} - 1/n)$$

is unbiased for $\tau(\theta)$. Since $\overline{X}$ is complete and sufficient, W is the UMVUE. Note that $W < 0$ is possible.

3. Suppose $X_1,\ldots,X_n$ are $i.i.d.$ Bernoulli(p) and $\tau(p) = p(1-p)$. An unbiased estimator is given by

$$W = X_1(1-X_2)$$

$T = \sum X_i$ is complete and sufficient, so

$$\phi(T) = E[W|\sum X_i]$$

is the UMVUE. Now

$$\begin{aligned}
E[W|\sum X_i = t] &= P(X_1 = 1, X_2 = 0|\sum X_i = t) \\
&= \frac{P(X_1 = 1, X_2 = 0, \sum_3^n X_i = t-1)}{P(\sum_1^n X_i = t)} \\
&= \begin{cases} 0 & t = 0 \\ \frac{p(1-p)\binom{n-2}{t-1}p^{t-1}(1-p)^{n-t-1}}{\binom{n}{t}p^t(1-p)^{n-t}} & t = 1,\ldots,n \end{cases} \\
&= \begin{cases} 0 & t = 0 \\ \frac{\binom{n-2}{t-1}}{\binom{n}{t}} & t = 1,\ldots,n \end{cases} \\
&= \begin{cases} 0 & t = 0 \\ \frac{t(n-t)}{n(n-1)} & t = 1,\ldots,n \end{cases} \\
&= \frac{t(n-t)}{n(n-1)}
\end{aligned}$$

So the UMVUE of $\tau(p) = p(1-p)$ is

$$\phi(T) = \frac{\sum X_i(n - \sum X_i)}{n(n-1)}$$

Homework

Problem 7.44
Problem 7.48

Due Friday, February 21, 2003.

Week 5

Monday, February 17, 2003

Loss Function Optimality

A general framework: the usual three,

Θ	parameter space
$\mathscr{X}$	sample space
$f(x\|\theta)$	model

and

$\mathscr{A}$	action space
$L(\theta,a)$	loss function
$\delta(x)$	decision rules

Loss function:

$$L(\theta,a) = \text{loss when action } a \text{ is taken and state of nature is } \theta$$

Decition rule $\delta(x) : (X) \to \mathscr{A}$:

$$\delta(x) = \text{action to take if } x \text{ is observed}$$

Examples

1. Point estimation with squared error loss:

$$\begin{aligned} \Theta &= \mathbb{R} \\ \mathscr{A} &= \Theta = \mathbb{R} \\ L(\theta,a) &= (\theta-a)^2 \end{aligned}$$

   Decision rules are estimators.

2. Hypothesis tests:

$$\Theta = \Theta_0 \cup \Theta_1, \text{ with } \Theta_0, \Theta_1 \text{ disjoint}$$
$$\mathscr{A} = \{\text{Reject } H_0, \text{Accept } H_0\}$$
$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \text{ and } a = \text{Reject } H_0 \\ 1 & \text{if } \theta \in \Theta_1 \text{ and } a = \text{Accept } H_0 \\ 0 & \text{otherwise} \end{cases}$$

Decision rules are test criteria.

Loss functions used for estimation usually satisfy $L(\theta,a) \geq 0$ and $L(\theta,a) = 0$ if and only if $\theta = a$.

A number of different loss functions can be used for estimation problems:

1. Squared error loss
$$L(\theta,a) = (\theta - a)^2$$
2. Absolute error loss
$$L(\theta,a) = |\theta - a|$$
3. Asymmetric loss
$$L(\theta,a) = \begin{cases} c(\theta - a) & \text{if } \theta \geq a \\ d(a - \theta) & \text{otherwise} \end{cases}$$
4. Bounded loss
$$L(\theta,a) = \frac{(\theta - a)^2}{1 + (\theta - a)^2}$$

The actual loss incurred by using decitin rule δ when the state of nature is θ and X is observed is the random variable

$$\text{actual loss} = L(\theta, \delta(X))$$

We compare decision rules in terms of the expected loss, also called the *risk function*:

De£nition

The risk function of a decision rule δ is

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))]$$

We want to £nd decision rules with low risk. But risk depends on θ. Often risk functions cross:

Example

Suppose $X \sim N(\theta,1)$, $L(\theta,a)$ is squared error loss, $\delta_1(X) = X$, and $\delta_2(X) = 3$. Then

$$\begin{aligned} R(\theta,\delta_1) &= 1 \\ R(\theta,\delta_2) &= (\theta-3)^2 \end{aligned}$$

When risk functions do cross they are not comparable. If the do not cross we can compare them:

De£nition

A decision rule δ_1 is as good as, or at least as good as, a decision rule δ_2 is $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ for all θ.

A decision rule δ_1 is better than δ_2 if it is as good as $|delta_2$ and $R(\theta,\delta_1) < R(\theta,\delta_2)$ for some θ.

A decision rule δ is admissible if no better decision rule exists.

Example

Let $X_1,\ldots,X_n$ be $i.i.d.$ $N(\mu,\sigma^2)$. Want to estimate σ^2 with squared error loss. Consider estimators of the form

$$\delta_b(X) = bS^2$$

Now

$$\begin{aligned} R((\mu,\sigma^2),\delta_b) &= \text{Var}(bS^2) + (E[bS^2]-\sigma^2)^2 \\ &= b^2\frac{2\sigma^4}{n-1} + (b-1)^2\sigma^4 \\ &= \left[\frac{2b^2}{n-1} + (b-1)^2\right]\sigma^4 \end{aligned}$$

The value $b = (n-1)/(n+1)$ minimizes the risk for all σ^2, so $\delta_{(n-1)/(n+1)}(X) = \frac{n-1}{n+1}S^2$ is the best estimator in this class.

Wednesday, February 19, 2003

Loss Function Optimality

Many papers are written on £nding admissible estimators.

Many (not all) standard estimators are admissible.

For $X \sim N(\theta, 1)$ and square error loss the estimator $\delta(X) = X$ is admissible.

For $X_1 \sim N(\theta_1, 1)$, $X_2 \sim N(\theta_2, 1)$, X_1, X_2 independent, and loss function

$$L(\theta, a) = (\theta_1 - a_1)^2 + (\theta_2 - a_2)^2$$

the decision rule $\delta(X) = (X_1, X_2)$ is admissible.

For $X_1, \dots, X_n$ independent, $X_i \sim N(\theta_i, 1)$, and loss function

$$L(\theta, a) = \sum (\theta_i - a_i)^2$$

the decision rule $\delta(X) = (X_1, \dots, X_n)$ is *not* admissible if $n \geq 3$. Shrinkage estimators can beat it. This is known as *Stein's paradox*.

Bayes Risk and Bayes Rules

If a prior distribution $\pi(\theta)$ is available then the average risk, or Bayes risk, can be used to compare decision rules:

De£nition

The Bayes risk for a decision rule δ and a prior π is

$$B(\pi, \delta) = E_\pi[R(\theta, \delta)] = \int_\Theta R(\theta, \delta) d\theta$$

The Bayes rule δ^π is the decision rule that minimizes the Bayes risk.

The Bayes risk can be written as

$$B(\pi, \delta) = E[R(\theta, \delta)] = E[E[L(\theta, \delta(X))|\theta]] = E[E[L(\theta, \delta(X))|X]]$$

Suppose we de£ne a decision rule δ^* as

$$\delta^*(x) = \underset{a}{\operatorname{argmin}} E[L(\theta, a)|X = x]$$

Then for any decision rule δ

$$B(\pi, \delta) = E[E[L(\theta, \delta(X))|X]] \geq E[E[L(\theta, \delta^*(X))|X]] = B(\pi, \delta^*)$$

So δ^* is a Bayes rule.

Examples

1. For estimation with squared error loss the Bayes rule, often called the Bayes estimator, is the posterior mean
$$\delta^{\pi}(X) = E[\theta|X]$$
2. For estimation with absolute error loss the Bayes rule is the posterior median.

So if $X_1, \ldots, X_n$ are $i.i.d.$ Bernoulli(p) wnd the prior distribution on p is Beta(α, β), then the Bayes rule for squared error loss is

$$\delta^{\pi}(X) = E[p|X] = \frac{\sum X_i + \alpha}{\alpha + \beta + n}$$

Bayes Estimators Are Not Unbiased

Suppose $W = E[\theta|X]$ is a Bayes estimator, i.e. a Bayes rule under squared error loss, and is unbiased. Then

$$\begin{aligned}
E[(W-\theta)^2] &= E[E[(W-\theta)^2|\theta]] \\
&= E[E[W^2 - 2W\theta + \theta^2|\theta]] \\
&= E[W^2] - 2E[\theta E[W|\theta]] + E[\theta^2] \\
&= E[W^2] - E[\theta^2]
\end{aligned}$$

On the other hand,

$$\begin{aligned}
E[(W-\theta)^2] &= E[E[(W-\theta)^2|X]] \\
&= E[W^2] - 2E[WE[\theta|X]] + E[\theta^2] \\
&= E[\theta^2] - E[W^2]
\end{aligned}$$

So we must have $E[W^2] = E[\theta^2]$ and thus

$$E[(W-\theta)^2] = 0$$

So W can only be unbiased if it is perfect! (Assumes $E[W^2] < \infty$.)

Homework

Problem 7.62
Problem 7.63
Problem 7.64

Due Friday, February 21, 2003.

Friday, February 21, 2003

Hypothesis Testing

A hypothesis is a statement about a parameter.

In a testing problem, there are two hypotheses:

H_0 : the null hypothesis

H_1 : the alternative hypothesis

Usually these are complementary, i.e. one and only one of H_0 and H_1 is true.

Examples:

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned} \qquad \text{or} \qquad \begin{aligned} H_0 &: \theta \leq \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$

Less common forms:

$$\begin{aligned} H_0 &: \theta \neq \theta_0 \\ H_1 &: \theta = \theta_0 \end{aligned} \qquad \text{or} \qquad \begin{aligned} H_0 &: \theta \notin \theta_0 \pm \delta \\ H_1 &: \theta \in \theta_0 \pm \delta \end{aligned}$$

The null hypothesis often corresponds to a claim that a treatment has no effect.

The alternative then usually says that the treatment has *some* effect ($\theta \neq \theta_0$) or an effect in a particular direction ($\theta > \theta_0$).

A hypothesis testing procedure is a rule for determining, based on data X, whether to reject H_0 in favor of H_1 or not.

The set of X values for which H_0 is rejected is called the *critical region* R, or the rejection region, of the test.

A hypothesis test can also be expressed in term of a test function,

$$\phi(X) = \begin{cases} 1 & \text{if } X \text{ rejects } H_0 \\ 0 & \text{if } X \text{ does not reject } H_0 \end{cases}$$

A test function corresponding to a rejection region R takes on only the values 0 or 1. In fact,

$$\phi(X) = 1_R(X)$$

As a technical device it is useful to allow other values in $[0,1]$; then

$$\phi(X) = P(\text{reject } H_0 | \text{observe } X)$$

i.e. you flip a coin with success probability $\phi(X)$ if you observe X.

Most hypothesis tests are developed in terms of a test statistic $W = W(X)$.

The corresponding rejection region then looks something like

$$R = \{X : W(X) > c\}$$

for some choice of c.

Examples:

$$\begin{aligned} H_0 &: \mu = 3 \\ H_1 &: \mu \neq 3 \end{aligned}$$

$$\begin{aligned} W &= |\overline{X} - 3| \\ R &= \{W > 0.5\} \end{aligned}$$

or

$$\begin{aligned} H_0 &: \sigma = 2 \\ H_1 &: \sigma > 2 \end{aligned}$$

$$\begin{aligned} W &= S/2 \\ R &= \{W > 1.5\} \end{aligned}$$

A nice feature about hypothesis tests is that the errors you can make are easy to think about:

	H_0	H_1
Reject H_0	Type I Error	OK
Don't Reject H_0	OK	Type II Error

We want test procedures that make both errors have small probability.

For the moment we will look at ways of coming up with classes of tests, or test statistics, like

$$\begin{aligned} \text{Reject}H_0 &: \mu = \mu_0 \\ \text{in favor of}H_1 &: \mu > \mu_0 \end{aligned}$$

if $\overline{X}$ is too large, i.e.

$$R = \{\overline{X} > c\}$$

for some c.

Choosing c and n affects our error probabilities.

After looking at ways of generating such families of tests, we will look at ways of comparing them.

Example

Suppose λ is the mean of a Poisson population.

$H_0 : \lambda = 7$	$R_1 = \{\overline{X} > c_1\}$	reject if $\overline{X}$ is large
$H_1 : \lambda > 7$	$R_2 = \{S^2 > c_2\}$	reject if S^2 is large

Which is better? (R_1 is.)

Week 6

Monday, February 24, 2003

Methods for Constructing Tests

Likelihood Ratio Tests

The likelihood ratio test statistic for testing

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta \setminus \Theta_0 \end{aligned}$$

is

$$\Lambda(X) = \frac{\sup_{\Theta_0} L(\theta|X)}{\sup_{\Theta} L(\theta|X)}$$

(I use Λ, the text uses λ.)

A likelihood ratio test is any test that has critical region equivalent to

$$\{x : \Lambda(x) \leq c\}$$

for some c.

Rationale:

> numerator is maximum over Θ_0 only; denominator is unrestricted maximum.
>
> mathematically, denominator $\geq$ numerator
>
> If denominator is much larger than the numerator, then there is strong evidence against H_0 in favor of H_1.
>
> If the denominator and the numerator are close, then there is little evidence against H_0 in favor of H_1.

Example

$X_1,\ldots,X_n$ *i.i.d.* $N(\theta,1)$.

$$\begin{aligned}H_0 &: \theta = \theta_0\\ H_1 &: \theta \neq \theta_0\end{aligned}$$

$$\begin{aligned}\Lambda(x) &= \frac{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum(x_i-\theta_0)^2\right\}}{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum(x_i-\overline{x})^2\right\}}\\ &= \exp\left\{\frac{1}{2}\sum(x_i-\overline{x})^2 - \frac{1}{2}\sum(x_i-\theta_0)^2\right\}\\ &= \exp\left\{-\frac{n}{2}(\overline{x}-\theta_0)^2\right\}\end{aligned}$$

since

$$\sum(x_i-\theta_0)^2 = \sum(x_i-\overline{x})^2 + n(\overline{x}-\theta_0)^2$$

So

$$\{\Lambda(x) < c\} = \left\{|\overline{x}-\theta_0| > \sqrt{\frac{2\log c}{n}}\right\}$$

or

$$\{|\overline{x}-\theta_0| > c\}$$

is a likelihood ratio test.

It is usually useful to try to simplify the LRT in this way, mainly because we will need to pick a particular c or think about different values of c.

Theorem

If T is a suf£cient statistic, then the LRT only depends on the data through T. Furthermore, the LRT based on the distribution of T is equivalent to the LRT based on the full data.

Proof

Let $f(x|\theta)$ be the PDF or PMF of X, $q(t|\theta)$ the PMF or PDF of T. Then from results related to the factorization theorem, there exist g, h_1 and h_2 such that

$$\begin{aligned}f(x|\theta) &= g(T(x)|\theta)h_1(x)\\ q(t|\theta) &= g(t|\theta)h_2(t)\end{aligned}$$

So

$$\Lambda(x) = \frac{\sup_{\Theta_0} f(x|\theta)}{\sup_{\Theta} f(x|\theta)} = \frac{\sup_{\Theta_0} g(T(x)|\theta)}{\sup_{\Theta} g(T(x)|\theta)}$$
$$\Lambda^*(T(x)) = \frac{\sup_{\Theta_0} q(T(x)|\theta)}{\sup_{\Theta} q(T(x)|\theta)} = \frac{\sup_{\Theta_0} g(T(x)|\theta)}{\sup_{\Theta} g(T(x)|\theta)}$$

□

Bayes Tests

In the Bayesian framework we have

likelihood

prior

from which we compute a posterior distribution.

In particular, if our hypotheses are

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \notin \Theta_0$$

then we can compute $P(\theta \in \Theta_0|X)$.

A formal test can be constructed as

$$R = \{x : P(\theta \in \Theta_0|X = x) < c\}$$

Possible values of c might be

$$c = 1/2$$
$$c = 0.05$$

Example

Suppose $X_1, \ldots, X_n|\theta$ are $i.i.d.$ $N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$, with μ, σ^2, τ^2 known.

Then

$$\theta|X = x \sim N\left(\frac{n\tau^2\overline{x} + \sigma^2\mu}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\right)$$

Suppose we use $c = 0.05$, $\Theta_0 = (-\infty, \theta_0]$. Then

$$R = \left\{x : \frac{\theta_0 - \frac{n\tau^2\overline{x} + \sigma^2\mu}{n\tau^2 + \sigma^2}}{\sigma\tau/\sqrt{n\tau^2 + \sigma^2}} < -z_{0.05}\right\}$$

where z_α is such that $P(Z > z_\alpha) = \alpha$ if $Z \sim N(0,1)$.

So

$$R = \left\{x : \frac{n\tau^2\overline{x} + \sigma^2\mu}{n\tau^2+\sigma^2} > \theta_0 + \frac{\sigma\tau}{\sqrt{n\tau^2+\sigma^2}} z_{0.05}\right\}$$

If τ is very large, then

$$R \approx \{x : \overline{x} > \theta_0 + \sigma z_{0.05}/\sqrt{n}\}$$

This is the standard frequentist test.

It is much harder to obtain standard two-sided tests as approximate Bayesian tests.

Union-Intersection and Intersection-Union Tests

Sometimes we can write

$$H_0 : \theta \in \bigcap_{\gamma\in\Gamma} \Theta_\gamma$$

for some index set Γ, £nite or in£nite.

If we have tests with critical regions R_γ for

$$\begin{aligned} H_0 &: \theta \in \Theta_\gamma \\ H_1 &: \theta \notin \Theta_\gamma \end{aligned}$$

for each γ, then we can form a critical region for the intersection H_0 as

$$R = \bigcup_{\gamma\in\Gamma} R_\gamma$$

Two examples:

$$\begin{array}{rcr} H_0 : \theta = \theta_0 & \leftrightarrow & \{\theta \le \theta_0\} \cap \{\theta \ge \theta_0\} \\ H_0 : \theta(y) = \theta_0(y) \forall y & \leftrightarrow & \bigcap_y \{\theta(y) = \theta_0(y)\} \end{array}$$

Similarly, if H_0 can be written as

$$H_0 : \theta \in \bigcup_{\gamma\in\Gamma} \Theta_\gamma$$

and we have critical regions R_γ for each subproblem, then we can form a critical region for the union null hypothesis as

$$R = \bigcap_{\gamma\in\Gamma} R_\gamma$$

Example

Often a material is only acceptable if several parameters are within speci£ed limits, say $\theta_1 > \theta_{1,0}$ and $\theta_2 > \theta_{2,0}$. Often this will be set up as the alternative hypothesis, with H_0 corresponding to failure to meet the standard, i.e.

$$H_0 : \theta_1 \leq \theta_{1,0} \text{ or } \theta_2 \leq \theta_{2,0}$$

Homework

Problem 8.5
Problem 8.6

Due Friday, February 28, 2003.

Wednesday, February 26, 2003

First Midterm Exam

The exam will cover the material covered in readings, in class and in assignments from Chapters 6 and 7.

The exam is closed book.

The exam will include some information on distributions along the lines of the **Table of Common Distributions** in the text.

Friday, February 28, 2003

Evaluating Test Procedures

	H_0	H_1
Reject H_0	Type I Error	OK
Don't Reject H_0	OK	Type II Error

For $\theta \in \Theta_0$

$$P(\text{Type I Error}|\theta) = P(X \in R|\theta)$$

For $\theta \notin \Theta_0$

$$P(\text{Type II Error}|\theta) = P(X \notin R|\theta)$$

Switching between R, R^c is a bit awkward, so we arbitrarily choose one of them to work with: The *power function* of a test with rejection region R is

$$\beta(\theta) = P(X \in R|\theta)$$

In terms of test functions ϕ,

$$\beta(\theta) = E[\phi(X)|\theta]$$

Some use $1-\beta(\theta)$ instead. This is called the *operating characteristic* (OC) function.

Example

Suppose $X_1,\ldots,X_n$ are $i.i.d.$ $N(\theta,1)$,

$$\begin{aligned} H_0 &: \theta \leq \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$

and

$$R = \{x : \overline{x} > \theta_0 + c/\sqrt{n}\}$$

Then

$$\begin{aligned} \beta(\theta) &= P(\overline{X} > \theta_0 + c/\sqrt{n}|\theta) \\ &= P(Z > c + \sqrt{n}(\theta_0 - \theta)) \end{aligned}$$

Ideally, we want

$$\begin{aligned} \beta(\theta) &= 0 && \text{if } \theta \leq \theta_0 \\ \beta(\theta) &= 1 && \text{if } \theta > \theta_0 \end{aligned}$$

Increasing n improves β for a fixed c and $\theta \neq \theta_0$

Changing c shifts the whole curve to the right or left

- this improves one error at the expense of the other
- you can't argue in general that one c is better than another.

To compare different tests, it is useful to fix one of the error probabilities.

Casella and Berger define:

1. size of a test:
$$\sup_{\theta \in \Theta_0} \beta(\theta)$$
2. a test is a level α test, $0 \leq \alpha \leq 1$, if its size is at most α.

Ideally, we would like to fix the size at α *and* fix $\beta(\theta_1)$ for some interesting θ_1 as θ.

We can usually only do this if we control n.

This is a major consideration in designing experiments.

Without control of n, we usually make H_0 be the hypothesis whose incorrect rejection probability we most want to control.

A research hypothesis we want to “prove” is usually set up as H_1. That way,

$$H_0 : \text{the research hypothesis is false}$$

has the bene£t of the doubt.

Homework

Problem 8.14
Problem 8.17

Due Friday, March 7, 2003.

Week 7

Monday, March 3, 2003

Most Powerful Tests

Consider testing

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta \setminus \Theta_0$$

De£nition

A test in a class $\mathscr{C}$ of possible tests is uniformly most powerful of class $\mathscr{C}$ if its power function $\beta(\theta)$ satis£es

$$\beta(\theta) \geq \beta'(\theta)$$

for all $\theta \in \Theta \setminus \Theta_0$ and all β' that are power functions for tests in $\mathscr{C}$.

Usually the class $\mathscr{C}$ involves a constraint on the size of the tests.

Neyman-Pearson Lemma

Consider testing

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

The data X have PMF or PDF $f(x|\theta_i), i = 0, 1$. De£ne a rejection region so that

$$\begin{array}{ll} x \in R & \text{if } f(x|\theta_1) > kf(x|\theta_0) \\ x \notin R & \text{if } f(x|\theta_1) < kf(x|\theta_0) \end{array}$$

for some $k \geq 0$ (what happens at equality is unspeci£ed). Let

$$\alpha = P(X \in R|\theta = \theta_0)$$

Then

(a) Any test of this form is UMP level α

(b) If there exists a test of this form with $k > 0$ then every UMP level α test is a size α test, and and every UMP level α test is of this form (except for a set of probability zero under $\theta = \theta_0$ and $\theta = \theta_1$).

Example

$X_1, \dots, X_n$ *i.i.d.* $N(\theta, 1)$. Consider $\theta_0 = 0$ and $\theta_1 = 1$. Then

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} = \frac{\exp\left\{-\frac{1}{2}\sum x_i^2 + \sum x_i - n/2\right\}}{\exp\left\{-\frac{1}{2}\sum x_i^2\right\}} = \exp\{n\overline{x} - n/2\}$$

So

$$R = \{x : \overline{x} > c\} = \{x : f(x|\theta_1) > e^{nc-n/2} f(x|\theta_0)\}$$

This test is UMP size $\alpha = P(z > \sqrt{n}c)$

This is true for *any* $\theta_1 > 0$.

Proof

Look at the continuous case–discrete case is analogous. If

$$\alpha = P(X \in R|\theta_0)$$

then the test has size α and hence is a level α test.

Let ϕ be a test function of the specified form and let ϕ' be any other level α test. Then

$$(\phi(x) - \phi'(x))(f(x|\theta_1) - kf(x|\theta_0)) \geq 0$$

for all x. So

$$\begin{aligned} 0 &\leq \int (\phi(x) - \phi'(x))(f(x|\theta_1) - kf(x|\theta_0))dx \\ &= \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)) \end{aligned}$$

To prove (a), note that since ϕ' is level α, we have

$$\beta'(\theta_0) \leq \alpha = \beta(\theta_0)$$

Since $k \geq 0$, this implies

$$\beta(\theta_1) \geq \beta'(\theta_1)$$

So ϕ is at least as powerful as ϕ', and hence ϕ is UMP level α.

To prove (b), suppose ϕ' is UMP level α. Since ϕ is also UMP level α, we must have

$$\beta(\theta_1) = \beta'(\theta_1)$$

Since $k > 0$, this implies

$$\beta'(\theta_0) = \beta(\theta_0) = \alpha$$

So ϕ' is size α. Furthermore,

$$\int (\phi(x) - \phi'(x))(f(x|\theta_1) - kf(x|\theta_0))dx = 0$$

implies that $\phi(x) = \phi'(x)$ for almost all x where $f(x|\theta_1) \neq kf(x|\theta_0)$. □

Homework

Problem 8.15
Problem 8.25

Due Friday, March 7, 2003.

Wednesday, March 5, 2003

More Most Powerful Tests

Corollary

Suppose T is sufficient for θ with $f(x|\theta) = g(T(x)|\theta)h(x)$. Let R be defined in terms of a subset S of the range of $T(x)$ as

$$R = \{x : T(x) \in S\}$$

where

$$\begin{aligned} &\alpha = P(T \in S|\theta_0) \\ &t \in S \qquad \text{if } g(t|\theta_1) > kg(t|\theta_0) \\ &t \notin S \qquad \text{if } g(t|\theta_1) < kg(t|\theta_0) \end{aligned}$$

for some $k \geq 0$. Then this test is UMP level α

Proof

This test is a Neyman-Pearson test. □

Corollary

Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \setminus \Theta_0$. Suppose a test based on a sufficient statistic T satisfies

(i) the test is a level α test

(ii) for some $\theta_0 \in \Theta_0$ we have $P(T \in S|\theta_0) = \alpha$.

(iii) for this θ_0 and each $\theta' \in \Theta \setminus \Theta_0$ there exists a $k' \geq 0$ such that

$$\begin{aligned} &t \in S \text{ if } g(t|\theta') > k'g(t|\theta_0) \\ &t \notin S \text{ if } g(t|\theta') < k'g(t|\theta_0) \end{aligned}$$

Then this test is UMP level α for H_0 against H_1.

Proof

Let ϕ^* be any other level α test.

Fix $\theta' \in \Theta \setminus \Theta_0$.

Then ϕ^* is a level α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta'$. By the Neyman-Pearson lemma,

$$\beta(\theta') \geq \beta^*(\theta')$$

Since θ' was arbitrary, this shows that $\beta(\theta) \geq \beta^*(\theta)$ for all $\theta \in \Theta \setminus \Theta_0$ □

Example

Let $X_1, \ldots, X_n$ be *i.i.d.* $N(\theta, 1)$.

$H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

$R = \{\overline{X} > c\}$

Set $\alpha = P(\overline{X} > c | \theta = \theta_0)$.

For $\theta < \theta_0$,

$$P(\overline{X} > c|\theta) < \alpha$$

so this is a size α test. Now

$$g(t|\theta) = \text{const} \times \exp\left\{-\frac{n}{2}(t-\theta)^2\right\}$$

Look at

$$\frac{g(t|\theta')}{g(t|\theta_0)} = \exp\left\{\frac{n}{2}[\theta_0^2 - \theta'^2 + 2t(\theta' - \theta_0)]\right\}$$

for $\theta' > \theta_0$. This is strictly increasing in t, so

$$t > c \qquad \Leftrightarrow \qquad g(t|\theta') > k' g(t|\theta_0)$$

with

$$k' = \exp\left\{\frac{n}{2}(\theta_0^2 - \theta'^2 + 2c(\theta' - \theta_0))\right\}$$

Homework

Problem 8.28
Problem 8.33

Due Friday, March 7, 2003.

Friday, March 7, 2003

Monotone Likelihood Ratio

Suppose T is a univariate sufficient statistic for θ, a real-valued parameter. Then $\{f(x|\theta) : \theta \in \Theta\}$ has *monotone likelihood ratio* (MLR) if for every $\theta_1 < \theta_2$

$$\frac{f(x|\theta_2)}{f(x|\theta_1)} = \frac{g(T(x)|\theta_2)}{g(T(x)|\theta_1)}$$

is a non-decreasing function of $T(x)$ over the set

$$\mathscr{T} = \{t : g(t|\theta_1) > 0 \text{ or } g(t|\theta_2)\}$$

(If you get non-increasing, just use $-T(X)$.)

Karlin-Rubin Theorem

Consider testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Suppose T is sufficient and $f(x|\theta)$ has MLR. Then for any c a test with $R = \{T > c\}$ is UMP level α for $\alpha = P(T > c|\theta = \theta_0)$.

Proof

(i) The power function is increasing (H.W.)

(ii) The test has power α by construction.

(iii)
$$k' = \inf_{t \in \mathscr{T}} \frac{g(t|\theta')}{g(t|\theta_0)}$$
where
$$\mathscr{T} = \{t : t > c \text{ and } g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$$

□

Examples

1. $X_1, \ldots, X_n$ *i.i.d.* $N(\theta, 1)$, $T = \overline{X}$.

$$\frac{g(t|\theta_2)}{g(t|\theta_1)} = \exp\{n\overline{X}(\theta_2 - \theta_1)\} \times \text{const}$$

2. $X_1,\ldots,X_n$ *i.i.d.* Poisson(θ), $T = \overline{X}$

$$\frac{g(t|\theta_2)}{g(t|\theta_1)} = \left(\frac{\theta_2}{\theta_1}\right)^{n\overline{X}} \times \text{xonst}$$

3. $X_1,\ldots,X_n$ *i.i.d.* $g(t|\theta) = \exp\{w(\theta)t\}c(\theta)$

$$\frac{g(t|\theta_2)}{g(t|\theta_1)} = \exp\{t(w(\theta_2) - w(\theta_1))\} \times \text{const}$$

   has MLR if w is non-decreasing.

Unbiased Tests

It is not always possible to £nd UMP tests.

Example

$X_1,\ldots,X_n$ *i.i.d.* $N(\theta,1)$, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

For a given α and a given $\theta_1 > \theta_0$,

$$R_1 = \{\overline{X} > \theta_0 + z_\alpha/\sqrt{n}\}$$

is UMP level α for $H_1' : \theta = \theta_1$. Furthermore, any test with the same size and power must be essentially the same.

But for $\theta_2 < \theta_0$ the same argument shows that the UMP test has to be

$$R_2 = \{\overline{X} < \theta_0 - z_\alpha/\sqrt{n}\}$$

These cannot both hold, so there is no UMP test.

Neither R_1 nor R_2 are very good for $H_1 : \theta \neq \theta_0$ since each has low power on its “blind” side.

To reduce the class of tests we consider to “reasonable” ones, we can require that our test be “unbiased’.”

A test with power function β is unbiased if

$$\sup_{\theta\in\Theta_0} \beta(\theta) \leq \inf_{\theta\in\Theta\backslash\Theta_0} \beta(\theta)$$

We need some additional tools to deal with this restriction.

Generalized Neyman-Pearson Lemma

Let $c_1, \ldots, c_m$ be constants, $f_1(x), \ldots, f_{m+1}(x)$ real-valued functions, and $\mathscr{C}$ a class of functions $\phi(x)$ with $0 \leq \phi(x) \leq 1$ for all x and

$$\int \phi(x) f_i(x) dx = c_i$$

for $i = 1, \ldots, m$. If $\phi^* \in \mathscr{C}$ satis£es

$$\begin{aligned} \phi^*(x) &= 1 && \text{if } f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \\ \phi^*(x) &= 0 && \text{if } f_{m+1}(x) < \sum_{i=1}^m k_i f_i(x) \end{aligned}$$

for some $k_1, \ldots, k_m$, then ϕ^* maximizes $\int \phi(x) f_{m+1}(x) dx$ over $\mathscr{C}$.

Proof

Since $0 \leq \phi \leq 1$ for all x and all $\phi \in \mathscr{C}$,

$$(\phi^*(x) - \phi(x))(f_{m+1}(x) - \sum_{i=1}^m k_i f_i(x)) \geq 0$$

for all x and all $\phi \in \mathscr{C}$. So

$$\begin{aligned} 0 &\leq \int (\phi^*(x) - \phi(x))(f_{m+1}(x) - \sum_{i=1}^m k_i f_i(x)) dx \\ &= \int \phi^*(x) f_{m+1} dx - \int \phi(x) f_{m+1}(x) dx \\ &+ \sum_{i=1}^m k_i \left(\int \phi^*(x) f_i(x) dx - \int \phi(x) f_i(x) dx \right) \\ &= \int \phi^*(x) f_{m+1} dx - \int \phi(x) f_{m+1}(x) dx \end{aligned}$$

□

Example

$X_1, \ldots, X_n$ $i.i.d.$ $N(\theta, 1)$. Want to test

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

For any test ϕ the power function β is continuously differentiable.

For any test to be unbiased it is necessary (but not suf£cient) that $\beta'(\theta_0) = 0$.

We can use the generalized Neyman-Pearson lemma to £nd a most powerful test of

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

such that $\beta(\theta_0) = \alpha$ and $\beta'(\theta_0) = 0$: Take

$$\begin{aligned} f_3(\overline{x}) &= f(\overline{x}|\theta_1) & & \\ f_2(\overline{x}) &= \frac{\partial}{\partial\theta} f(\overline{x}|\theta_0) & c_2 &= 0 \\ f_1(\overline{x}) &= f(\overline{x}|\theta_0) & c_1 &= \alpha \end{aligned}$$

The most powerful test with these restrictions rejects if

$$f_3(\overline{x}) > k_1 f_1(\overline{x}) + k_2 f_2(\overline{x})$$

for some k_1 and k_2 that satisfy the two restrictions.

Now

$$f_3(\overline{x}) > k_1 f_1(\overline{x}) + k_2 f_2(\overline{x})$$

means

$$\exp\left\{-\frac{n}{2}(\overline{x}-\theta_1)^2\right\} > k_1 \exp\left\{-\frac{n}{2}(\overline{x}-\theta_0)^2\right\} + k_2 n(\overline{x}-\theta_2) \exp\left\{-\frac{n}{2}(\overline{x}-\theta_0)^2\right\}$$

or

$$\exp\left\{-\frac{n}{2}(\theta_1^2-\theta_0^2) + n\overline{x}(\theta_1-\theta_0)\right\} \geq k_1 + k_2 n(\overline{x}-\theta_0)$$

The exponential term can be increasing or decreasing.

We can get R to be one-sided or two-sided.

To get $\beta'(\theta_0) = 0$ we need two-sided, symmetric about θ_0. With this choice R is also unbiased.

To get $\beta(\theta_0) = \alpha$, we need

$$R = \{\overline{x} : \overline{x} < \theta_0 - z_{\alpha/2}/\sqrt{n} \text{ or } \overline{x} > \theta_0 + z_{\alpha/2}/\sqrt{n}\}$$

For each $\alpha' < \alpha$ the UMP size α test is the same shape but less powerful.

So this is a UMPU level α test.

Similar ideas work with many one-parameter exponential families.

Nuisance parameters can sometimes be handled in this way as well.

Homework

Problem 8.31
Problem 8.34

Due Friday, March 14, 2003.

Week 8

Monday, March 10, 2003

P-Values

In a research setting it is usual to give not

"the test rejected H_0 at the 0.1 level"

but to compute and report the

$$\begin{aligned} p\text{-value} &= \text{smallest level where test would reject} \\ &= \text{largest level where test would not reject} \end{aligned}$$

$p = 0.049$ and $p = 0.007$ both reject at the $\alpha = 0.05$ level, but suggest a difference in the strength of evidence.

Some unfortunate terminology:

$p \leq 0.05$	"statistically significant"
$p \leq 0.01$	"highly statistically significant"

Older programs would mark these as * and **.

This is the reason for occasional comments about "star gazing".

Even p-values do not tell the whole story:

- if p is small, you need to worry if the results are of practical significance.
- if p is large, you need to think about whether it could have been otherwise (was there any power at plausible alternatives?)

Another way to look at p-values is provided by

De£nition

A p-value $p(X)$ is a statistic such that $0 \leq p(X) \leq 1$ for all X and small values of $p(X)$ give evidence in favor of H_1. A p-value is *valid* if

$$P_\theta(p(X) \leq \alpha) \leq \alpha$$

for all $\theta \in \Theta_0$ and all $\alpha \in [0,1]$.

If $p(X)$ is a valid p-value, then the rejection region

$$R = \{x : p(x) \leq \alpha\}$$

is a level α test.

Usually $p(X)$ is de£ned in terms of a test statistic:

Theorem

Suppose $W(X)$ is a test statistic such that large values of $W(X)$ are exvidence for H_1. De£ne

$$p(x) = \sup_{\theta \in \Theta_0} P_\theta(W(X) \geq W(x))$$

Then $p(X)$ is a valid p-value.

Proof

Let $p_\theta(x) = P_\theta(W(X) \geq W(x))$ and let F_θ be the CDF of $-W(X)$. Then

$$p_\theta(x) = P_\theta(-W(X) \leq -W(x)) = F_\theta(-W(x))$$

and

$$P_\theta(p_\theta(X) \leq \alpha) = P_\theta(F_\theta(-X(X)) \leq \alpha) \leq \alpha$$

[If F_θ is continuous then equality holds by the probability integral transform; in general, this inequality holds.] For $\theta \in \Theta_0$ we have $p_\theta(X) \leq p(X)$, and therefore

$$P_\theta(p(X) \leq \alpha) \leq P_\theta(p_\theta(X) \leq \alpha) \leq \alpha$$

□

Example

Suppose $X_1, \ldots, X_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution and we want to test

$$\begin{aligned} H_0 :& \mu \leq \mu_0 \\ H_1 :& \mu > \mu_0 \end{aligned}$$

The LRT is the t test which rejects H_0 when

$$W(X) = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

is large. For $\mu \leq \mu_0$ and any $\sigma > 0$

$$\begin{aligned} p_\theta(x) &= P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq W(x)\right) \\ &= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq W(x) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\ &\leq P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq W(x)\right) \\ &= P(T_{n-1} \geq W(x)) \end{aligned}$$

The maximum always occurs at the boundary value $\mu = \mu_0$.

A Graphical representation: We can plot the CDF's of p(X) for different θ values.

Often there is a boundary value θ_0 for which $p(X)$ is uniformly distributed.

If the test provided by $W(X)$ is unbiased for all choices of α, then we have

$$P\theta(p(X) \leq \alpha) \geq \alpha$$

for all $\theta \in \Theta_1$.

Homework

Problem 8.49
Problem 8.54

Due Friday, March 14, 2003.

Wednesday, March 12, 2003

Testing as a Decision Problem

$\Theta, \mathscr{X}, f$ as usual.

$\mathscr{A} = \{a_0, a_1\} = \{\text{accept } H_0, \text{reject } H_0\}$

Some loss functions:

$$\begin{aligned} L(\theta,a) &= \begin{cases} 0 & \text{if } \theta \in \Theta_0 \text{ and } a = a_0 \text{ or } \theta \in \Theta_1 \text{ and } a = a_1 \\ 1 & \text{otherwise} \end{cases} \\ &= \text{zero-one loss} \\ L(\theta,a) &= \begin{cases} c & \text{if } \theta \in \Theta_0 \text{ and } a = a_1 \\ d & \text{if } \theta \in \Theta_1 \text{ and } a = a_0 \\ 0 & \text{otherwise} \end{cases} \\ &= \text{generalized zero-one loss} \end{aligned}$$

For $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ we could use

$$\begin{aligned} L(\theta,a_0) &= \begin{cases} 0 & \theta \leq \theta_0 \\ c(\theta-\theta_0) & \theta > \theta_0 \end{cases} \\ L(\theta,a_1) &= \begin{cases} 0 & \theta > \theta_0 \\ d(\theta_0-\theta) & \theta \leq \theta_0 \end{cases} \end{aligned}$$

Relation to power:

$$\begin{aligned} \beta_\delta(\theta) &= P_\theta(\delta(X) = a_1) \\ R(\theta,\delta(X)) &= L(\theta,a_0)P_\theta(\delta(X) = a_0) + L(\theta,a_1)P_\theta(\delta(X) = a_1) \\ &= L(\theta,a_0)(1-\beta_\delta(\theta)) + L(\theta,a_1)\beta_\delta(\theta) \\ &= L(\theta,a_0) + (L(\theta,a_1) - L(\theta,a_0))\beta_\delta(\theta) \end{aligned}$$

For generalized zero-one loss, the posterior expected losses are

$$E[L(\theta,a)|X] = \begin{cases} dP(\theta \in \Theta_1|X) & a = a_0 \\ cP(\theta \in \Theta_0|X) & a = a_1 \end{cases}$$

So the Bayes rule chooses a_1 if

$$cP(\theta \in \Theta_0|X) < dP(\theta \in \Theta_1|X)$$

or if

$$\text{posterior odds of } \Theta_1 \textit{ vs } \Theta_0 = \frac{P(\theta \in \Theta_1|X)}{P(\theta \in \Theta_0|X)} > \frac{c}{d}$$

Locally Most Powerful Tests

If we can't £nd a UMP test we can look for a test ϕ^* such that for some Δ and all θ within Δ of Θ_0 $\beta^*(\theta) \geq \beta(\theta)$ for all other tests ϕ. Such tests are called *locally most powerful* (LMP).

This makes sense since we are usually most concerned about a test sort of near Θ_0

For $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, LMP means maximize $\beta'(\theta_0)$.

For $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, LMP means maximize $\beta''(\theta_0)$.

Generalized NP lemma helps here too.

Cautions on Testing

If p-values is small, make sure differences are of practical importance.

If p-value is not small, think about power at plausible alternatives.

Setting up H_1 as a research hypothesis, only rejecting if evidence is strong is a good strategy.

But understanding differences can be hard.

Often we understand what $\theta = 0$ means but not how to think about $\theta \neq 0$.

Sometimes we would like to use

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta \neq 0 \end{aligned}$$

as a pre-test for checking assumptions.

This can be very dangerous unless there is strong prior information in favor of H_0.

Homework

Problem 8.55
Problem 8.56

Due Friday, March 14, 2003.

Friday, March 3, 2000

Interval and Set Estimation

Motivation

In point estimation we give a single guess for θ or $\tau(\theta)$.

This is useful when we need a single number (e.g. to set an instrument).

But a point estimate is almost surely wrong.

Moreover, some estimators are better than others.

There are two approaches for for dealing with this issue:

> Informal approach:
>
> > In a frequentist analysis, always report an estimate *and* a standard error (estimated SD of sampling distribution).
> >
> > In a Bayesian analysis, always report a summary of location *and* spread of the posterior distribution.
>
> Formal approach:
>
> > Use the data X to determine a set $C(X) \subset \Theta$ of values that are supported by the data in some formally defined sense.

Possible Shapes

In one dimension, set estimators are often restricted to produce intervals,

$$C(X) = [L(X), U(X)]$$

It is sometimes useful to allow open, half-open, or half-infinite intervals.

In higher dimensions, there is no clear natural shape to require—one could ask for connectedness, convexity, a rectangle, etc..

The set or interval produced by a set estimator is a set-valued random variable, or a random set.

Objectives

There are two conflicting objectives:

We want the set to be small, to make a precise statement.

We want the set to be "right," i.e. to contain θ.

It is fairly clear what we mean by an interval being small—we look at its length.

The length might be random, so we can take its expected value,

$$E_\theta[U(X) - L(X)]$$

(at least for bounded intervals this is sensible).

What about being "right?"

The frequentist approach: For each θ we can compute

$$\begin{aligned} P_\theta(\text{interval covers } \theta) &= P_\theta(L \leq \theta \text{ and } U \geq \theta) \\ &= P_\theta(L \leq \theta \leq U) \\ &= \text{coverage probability} \end{aligned}$$

This may depend on θ, so we look at the worst case:

$$\text{Confidence Coefficient} = \inf_\theta P_\theta(C(X) \text{ covers } \theta)$$

Example

$X_1, \ldots, X_n$ *i.i.d.* $N(\theta, \sigma^2)$, σ^2 known.

$\overline{X}$ estimates θ.

$\text{SE}(\overline{X}) = \sigma/\sqrt{n}$

Often we report "$\overline{X}$, give or take $\sigma/\sqrt{n}$ or two."

Suppose we use

$$[L, U] = \overline{X} \pm 2\sigma/\sqrt{n}$$

Then

$$\begin{aligned} P_\theta(L \leq \theta \leq U) &= P_\theta(\overline{X} - 2\sigma/\sqrt{n} \leq \theta \leq \overline{X} + 2\sigma/\sqrt{n}) \\ &= P_\theta\left(\sqrt{n}\frac{\overline{X} - \theta}{\sigma} \leq 2, \sqrt{n}\frac{\overline{X} - \theta}{\sigma} \geq -2\right) \\ &= P_\theta\left(-2 \leq \sqrt{n}\frac{\overline{X} - \theta}{\sigma} \leq 2\right) \\ &= P(-2 \leq Z \leq 2) \approx 0.95 \end{aligned}$$

The coverage probability is ≈ 0.95 for all θ, so the confidence coefficient is ≈ 0.95.

Suppose $n = 4$, $\bar{x} = 3.7$, $\sigma = 1$. Then

$$[\ell, u] = 3.7 \pm 2 \times 1/2 = 3.7 \pm 1 = [2.7, 4.7]$$

is an observed 95% CI for θ.

It is *not* true that $P(\theta \in [2.7, 4.7]) = 0.95$.

It *looks* like this is what is being said, but it is not.

Homework

Problem 9.1
Problem 9.2

Due Friday, March 28, 2003.

Week 9

Monday, March 24, 2003

Inverting Tests

Suppose $X_1, \dots, X_n$ are *i.i.d* $N(\theta, \sigma^2)$ with σ^2 known.

For any θ_0, a UMPU test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is

$$R = \{x : |\bar{x} - \theta_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$$

This test has size α, so

$$P(\overline{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \theta_0 \leq \overline{X} + z_{\alpha/2}\sigma/\sqrt{n} | \theta = \theta_0) = 1 - \alpha$$

for any θ_0. So

$$P_\theta(\overline{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \theta \leq \overline{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$$

and so $\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$ is a $1-\alpha$-level CI for μ.

Inverting a test requires a family of tests, one for each $\theta_0 \in \Theta$.

The set estimate obtained by inverting a family of tests is the set of all θ that would not be rejected by the corresponding tests.

Theorem

For each $\theta_0 \in \Theta$ let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $x \in \mathscr{X}$ define $C(x)$ as

$$C(x) = \{\theta_0 : x \in A(\theta_0)\}$$

Then the random set $C(X)$ is a confidence set with confidence coefficient at least $1-\alpha$. Conversely, let $C(X)$ be a confidence set with confidence coefficient at least $1-\alpha$. For any $\theta_0 \in \Theta$ define

$$A(\theta_0) = \{x : \theta_0 \in C(x)\}$$

Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$ against, say, $H_0 : \theta \neq \theta_0$ for each $\theta_0 \in \Theta$.

Proof

Suppose $\{A(\theta) : \theta \in \Theta\}$ are acceptance regions of level α tests. Then

$$1-\alpha \le P_\theta(X \in A(\theta)) = P_\theta(\theta \in C(X))$$

For the converse, if $C(X)$ is a con£dence set with con£dence level at least $1-\alpha$, then

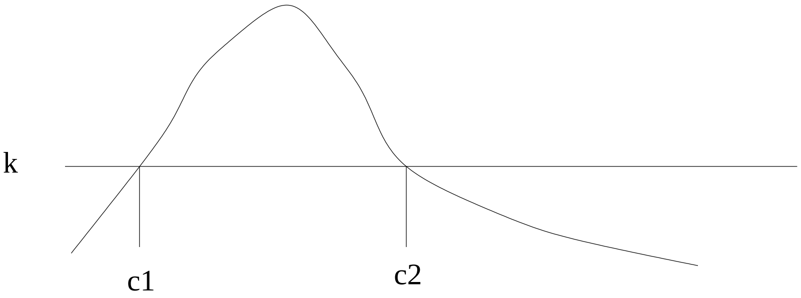
$$\alpha \ge P_{\theta_0}(C(X) \text{ does not cover } \theta_0) = P_{\theta_0}(X \notin A(\theta_0))$$

so $A(\theta_0)$ is the acceptance region of a level α test. □

Examples

1. Suppose $X_1, \ldots, X_n$ are *i.i.d.* $N(\mu, \sigma^2)$ and we want an interval estimate for σ^2. The likelihood ratio statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$ is

$$\Lambda = \frac{\left(\frac{1}{\sigma_0}\right)^{n/2} e^{-n\widehat{\sigma}^2/(2\sigma_0)}}{\left(\frac{1}{\widehat{\sigma}^2}\right)^{n/2} e^{-n/2}} = f\left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\right)$$



For $R = \{\Lambda < k\}$ use

$$R = \left\{\widehat{\sigma}^2 < \sigma_0^2 \frac{1}{n}\chi^2_{1-\alpha_1} \text{ or } \widehat{\sigma}^2 > \sigma_0^2 \frac{1}{n}\chi^2_{\alpha_2}\right\}$$

where $\alpha_1 + \alpha_2 = \alpha$ and

$$f\left(\frac{1}{n}\chi^2_{1-\alpha_1}\right) = f\left(\frac{1}{n}\chi^2_{\alpha_2}\right)$$

So

$$A(\sigma_0^2) = \{S^2 : \sigma_0^2\chi^2_{1-\alpha_1} \le (n-1)S^2 \le \sigma_0^2\chi^2_{\alpha_2}\}$$

and therefore

$$\begin{aligned} C(X) &= \{\sigma^2 : S^2 \in A(\sigma^2)\} \\ &= \{\sigma^2 : (n-1)S^2 \ge \sigma^2\chi^2_{1-\alpha_1} \text{ and } (n-1)S^2 \le \sigma^2\chi^2_{\alpha_2}\} \\ &= [(n-1)S^2/\chi^2_{\alpha_2}, (n-1)S^2/\chi^2_{1-\alpha_1}] \end{aligned}$$

Usually we cheat and use $\alpha_1 = \alpha_2 = \alpha/2$, which is not quite right.

2. Suppose $X_1,\ldots,X_n$ are *i.i.d.* Poisson(λ) and we want a lower con£dence limit on λ. Look at the LR test for

$$\begin{aligned} H_0 &: \lambda = \lambda_0 \\ H_1 &: \lambda > \lambda_0 \end{aligned}$$

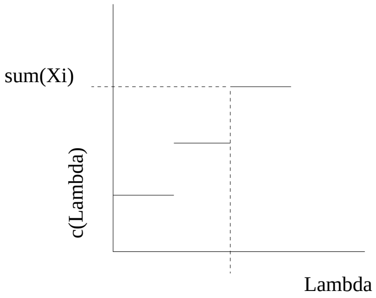
The test statistic is

$$\Lambda = \begin{cases} 1 & \overline{X} \leq \lambda_0 \\ \left(\frac{\lambda_0}{\overline{X}}\right)^{n\overline{X}} e^{n(\overline{X}-\lambda_0)} & \overline{X} > \lambda_0 \end{cases}$$

$\Lambda > k$ if and only if $\sum X_i > c$ for some c.

For each λ_0, £nd the smallest integer $c(\lambda_0)$ such that

$$P_{\lambda_0}\left(\sum X_i \geq c(\lambda_0)\right) \leq \alpha$$



The smallest λ with $c(\lambda) \geq \sum X_i$ is a lower con£dence limit.

Using the CLT:

$$\sqrt{\overline{X}} \sim \text{AN}\left(\sqrt{\lambda}, \frac{1}{4n}\right)$$

$$c(\lambda) \approx n\left(\sqrt{\lambda} + \frac{1}{2\sqrt{n}} z_\alpha\right)^2$$

so

$$\begin{aligned} \sum X_i \geq c(\lambda) &\Leftrightarrow \sqrt{\overline{X}} \geq \sqrt{\lambda} + \frac{1}{2\sqrt{n}} z_\alpha \\ &\Leftrightarrow \sqrt{\overline{X}} - \frac{1}{2\sqrt{n}} z_\alpha \geq \sqrt{\lambda} \\ &\Leftrightarrow \left(\sqrt{\overline{X}} - \frac{1}{2\sqrt{n}} z_\alpha\right)^2 \geq \lambda \end{aligned}$$

Can also solve quadratic for the usual normal approximation.

Homework

Problem 9.4

Due Friday, March 28, 2003.

Wednesday, March 26, 2003

Pivotal Quantities

A random variable $Q(X,\theta)$ is a *pivotal quantity*, or a *pivot*, if its distribution is independent of all unknown parameters.

Examples

1. $X_1,\ldots,X_n$ *i.i.d.* $N(\theta,1)$, $Q=\sqrt{n}(\overline{X}-\theta)\sim N(0,1)$.
2. $X_1,\ldots,X_n$ *i.i.d.* $N(\theta,\sigma^2)$, $Q=\sqrt{n}(\overline{X}-\theta)/S\sim t_{n-1}$.
3. $X_1,\ldots,X_n$ *i.i.d.* $N(\mu,\sigma^2)$, $Q=(n-1)S^2/\sigma^2\sim \chi^2_{n-1}$.

Pivotal quantity method:

1. Choose a set A such that
$$P(Q(X,\theta)\in A)=1-\alpha$$
2. Let $C(x)=\{\theta : Q(x,\theta)\in A\}$

Then $C(X)$ is a $(1-\alpha)$-level con£dence set.

Usually there is a “reasonable” choice of A based on monotonicity ideas.

Example

Suppose $X_1,\ldots,X_n$ are *i.i.d.* $N(\mu,\sigma^2)$ and we want a con£dence set for σ^2.

Let

$$\begin{aligned}Q&=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}\\ A&=[\chi^2_{n-1,1-\alpha/2},\chi^2_{n-1,\alpha/2}]\end{aligned}$$

Then

$$\begin{aligned}C(x)&=\{\sigma^2:\chi^2_{n-1,1-\alpha/2}\leq(n-1)S^2/\sigma^2\leq\chi^2_{n-1,\alpha/2}\}\\ &=\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}},\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]\end{aligned}$$

Bayesian Intervals

Suppose

$$\begin{aligned} \theta &\sim f(\theta) \\ X|\theta &\sim f(x|\theta) \end{aligned}$$

Then

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$

is the PDF of the posterior distribution.

Given the posterior distribution and a level $1-\alpha$, we can compute sets of posterior probability $1-\alpha$.

Such sets are called *credible sets*.

Intervals are called *credible intervals*.

A credible region's probability of containing θ is a posterior probability, not a coverage probability based on conceptual repetitions of the experiment.

There is a relation:

$$\begin{aligned} E[P(\theta \in C(X)|X)] &= \iint 1_{C(x)}(\theta)f(\theta|x)d\theta f(x)dx \\ &= \iint 1_{C(x)}(\theta)f(\theta,x)d\theta dx \\ &= \iint 1_{C(x)}(\theta)f(x|\theta)dx f(\theta)d\theta \\ &= \int P_\theta(\theta \in C(X))f(\theta)d\theta \end{aligned}$$

So $P(\theta \in C(X)|X=x) \geq 1-\alpha$ for all x implies

$$\int P_\theta(\theta \in C(X))f(\theta)d\theta \geq 1-\alpha$$

But $P_\theta(\theta \in C(X)) \ll 1-\alpha$ for *some* θ is possible.

Example

Suppose $X_1,\ldots,X_n$ are *i.i.d.* $N(\theta,\sigma^2)$, $\theta \sim N(\mu,\tau^2)$, and μ,σ^2,τ^2 are known.

We know that

$$\theta|X=x \sim N\left(\frac{n\tau^2}{n\tau^2+\sigma^2}\overline{X}+\frac{\sigma^2}{n\tau^2+\sigma^2}\mu, \left(\frac{\sigma\tau}{\sqrt{n\tau^2+\sigma^2}}\right)^2\right)$$

So a lower $1-\alpha$ level credible bound is

$$\frac{n\tau^2}{n\tau^2+\sigma^2}\overline{X}+\frac{\sigma^2}{n\tau^2+\sigma^2}\mu - z_\alpha\frac{\sigma\tau}{\sqrt{n\tau^2+\sigma^2}}$$

i.e.

$$P\left(\theta > \frac{n\tau^2}{n\tau^2+\sigma^2}\overline{X}+\frac{\sigma^2}{n\tau^2+\sigma^2}\mu - z_\alpha\frac{\sigma\tau}{\sqrt{n\tau^2+\sigma^2}}\middle| X=x\right)=1-\alpha$$

A two-sided $1-\alpha$ credible interval is

$$\frac{n\tau^2}{n\tau^2+\sigma^2}\overline{X}+\frac{\sigma^2}{n\tau^2+\sigma^2}\mu \pm z_{\alpha/2}\frac{\sigma\tau}{\sqrt{n\tau^2+\sigma^2}}$$

If τ is very large, then $\theta|C=x$ is approximately

$$N(\bar{x},\sigma^2/n)$$

So for a vague prior, the "usual" CI's are credible intervals with

$$\text{confidence level} = \text{posterior probability of containment}$$

Choosing the Smallest Credible Set

How should you choose a credible interval/set for a given probability level?

Suppose f is a PDF. For a given α, we can choose C such that

$$\int_C f dx = 1-\alpha$$

and

$$\text{area of } C = \int_C dx$$

is minimized. Equivalently, we want to maximize $-\int_C dx$.

Use the generalized Neyman-Pearson lemma:

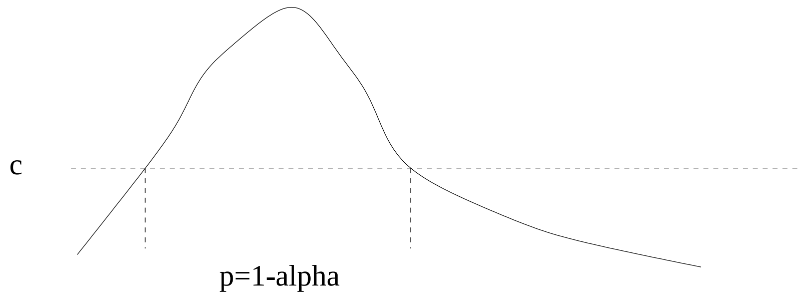
$$\begin{aligned} f_2 &\equiv -1\\ f_1 &= f \end{aligned}$$

The maximal negative area occurs if C is described by ϕ with

$$\begin{aligned}\phi(x) &= \begin{cases} 1 & \text{if } -1 > kf(x)\\ 0 & \text{if } -1 < kf(x)\end{cases}\\ &= \begin{cases} 1 & \text{if } f(x) > c\\ 0 & \text{if } f(x) < c\end{cases}\end{aligned}$$

for some k, which has to be be negative, or some $c=-1/k$.

This says: choose the *highest posterior density region*:



Homework

Problem 9.12
Problem 9.13

Due Friday, March 28, 2003.

Friday, March 28, 2003

Evaluating Con£dence Sets

Minimizing Interval Length

One approach is to ask for minimum (expected) length given the con£dence level.

For $X_1,\dots,X_n$ *i.i.d.* $N(\mu,\sigma^2)$

$$[\overline{X}-bS/\sqrt{n},\overline{X}-aS/\sqrt{n}]$$

with $P(a<T_{n-1}<b)=1-\alpha$ has expected length

$$E[\text{length}]=E[(b-a)S/\sqrt{n}]=(b-a)\sigma\times\text{const}(n)$$

We minimize $b-a$ subject to $P(a<T_{n-1}<b)=1-\alpha$ by choosing a,b at contour levels, i.e. $a=-t_{n-1,\alpha/2}, b=t_{n-1,\alpha/2}$.

This criterion is useful in principle for choosing tail allocations.

It is a bit messy as a theoretical criterion.

It depends on the measurement scale.

It also does not work for one-sided intervals.

Exploiting Relations to Testing

Alternative approach: try to exploit the relation to testing.

We have an optimality theory for testing; let's map it to con£dence sets.

In testing we have $\Theta_0,\beta(\theta),\Theta_1=\Theta_0^c$.

We constrain β on Θ_0, optimize it on $\Theta_1=\Theta_0^c$.

In con£dence sets, we have a family of tests with a family of θ_0's and a family of β_{θ_0}'s.

We consider

acceptance regions $A(\theta_0)$

alternate hypotheses $\Theta_1(\theta_0)$

power functions $\beta_{\theta_0}(\theta)$

De£ne for θ, θ' the probability of false coverage as

$$P_\theta(\theta' \in C(X)) \qquad \theta \in \Theta_1(\theta')$$

For two-sided intervals:

$$P_\theta(\theta' \in [L(X), U(X)]) \qquad \theta \neq \theta'$$

One-sided:

$$\begin{array}{rr} P_\theta(\theta' \in [L(X), \infty)) & \theta' < \theta \\ P_\theta(\theta' \in (-\infty, U(X)] & \theta' > \theta \end{array}$$

Relation to power:

$$1 - \beta_{\theta'}(\theta) = P_\theta(\text{false coverage of } \theta')$$

A $1-\alpha$ con£dence set that minimizes the probability of false coverage among a class of such sets is called *uniformly most accurate*, UMA.

A $1-\alpha$ con£dence set is unbiased if

$$P_\theta(\theta' \in C(X)) \leq 1 - \alpha$$

when $\theta \in \Theta_1(\theta')$.

Theorem

Let $X \sim f(x|\theta)$. For each $\theta_0 \in \Theta$ let $A^*(\theta_0)$ be the acceptance region of a UMP level α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1(\theta_0)$. Let $C^*(X)$ be the $1-\alpha$ con£dence set obtained by inverting the tests. Then for any other $1-\alpha$ con£dence set C,

$$P_\theta(\theta' \in C^*(X)) \leq P_\theta(\theta' \in C(X))$$

for all θ, θ' with $\theta \in \Theta_1(\theta')$. That is, C^* is UMA level $1-\alpha$.

Proof

Suppose θ, θ' satisfy $\theta \in \Theta_1(\theta')$. Let $A(\theta')$ be the acceptance region of the $1-\alpha$ level test from inverting $C(X)$. Since $A^*(\theta')$ is UMP,

$$\begin{aligned} P_\theta(\theta' \in C^*(X)) &= P_\theta(X \in A^*(\theta')) = 1 - \beta^*_{\theta'}(\theta) \\ &\leq 1 - \beta_\theta(\theta) = P_\theta(X \in A(\theta')) \\ &= P_\theta(\theta' \in C(X)) \end{aligned}$$

□

Suppose UMP tests are not available.

Then we can look at UMPU tests.

Unbiased tests correspond to unbiased intervals/sets.

If we have a family of UMPU tests, then they invert to UMAU confidence sets.

Relating False Coverage to Length

Theorem

Let X be real-valued, $X \sim f(x|\theta)$, with θ real-valued. Let $C(X) = [L(X), U(X)]$ be a CI for θ. If $L(x), U(x)$ are both strictly increasing in x, then for any θ^*

$$E_{\theta^*}[U(X) - L(X)] = \int_{\theta \neq \theta^*} P_{\theta^*}(L(X) \leq \theta \leq U(X)) d\theta$$

Proof

$$\begin{aligned}
E_{\theta^*}[U(X) - L(X)] &= \int_{\mathscr{X}} [U(x) - L(x)] f(x|\theta^*) dx \\
&= \int_{\mathscr{X}} \int_{L(x)}^{U(x)} d\theta dx \\
&= \int_{\Theta} \int_{U^{-1}(\theta)}^{L^{-1}(\theta)} f(x|\theta^*) dx d\theta \\
&= \int_{\Theta} P_{\theta^*}(U^{-1}(\theta) \leq X \leq L^{-1}(\theta)) d\theta \\
&= \int_{\Theta} P_{\theta^*}(L(X) \leq \theta \leq U(X)) d\theta \\
&= \int_{\theta \neq \theta^*} P_{\theta^*}(L(X) \leq \theta \leq U(X)) d\theta
\end{aligned}$$

□

Examples

1. $X_1, \ldots, X_n$ $i.i.d.$ $N(\theta, 1)$. $\overline{X} - z_\alpha/\sqrt{n}$ is a UMA lower confidence bound for θ.
2. $X_1, \ldots, X_n$ $i.i.d.$ $N(\theta, 1)$. $\overline{X} \pm z_{\alpha/2}/\sqrt{n}$ is a UMAU confidence interval for θ.
3. $X_1, \ldots, X_n$ $i.i.d.$ $N(\theta, \sigma^2)$. $\overline{X} \pm t_{n-1,\alpha/2} S/\sqrt{n}$ is a UMAU confidence interval for θ.

Homework

Problem 9.12
Problem 9.13

Due Friday, March 28, 2003.

Week 10

Monday, March 31, 2003

Consistency

Often an estimator W is described by a rule that can be applied to any sample size.

We can capture the idea that W is "reasonable" by looking at a sequence W_n as $n \to \infty$ and requiring that W "do the right thing" if n is large.

De£nition

A sequence W_n of estimators of $\tau(\theta)$ is (weakly) consistent if $W_n \xrightarrow{P} \tau(\theta)$ as $n \to \infty$ for all θ. W_n is strongly consistent if $W_n \xrightarrow{a.s.} \tau(\theta)$.

From our study of convergence in probability, we know that if

$$\text{MSE}(W_n, \theta) = E_\theta[(W_n - \tau(\theta))^2] \to 0$$

then W_n is consistent for $\tau(\theta)$.

Since $\text{MSE} = \text{Var} + \text{Bias}^2$, if

$$\text{Var}(W_n) \to 0$$

and

$$\text{Bias}(W_n) \to 0$$

then W_n is consistent.

Examples

1. $\overline{X}$ is consistent for μ.
2. S^2 is consistent for σ^2.
3. S is consistent for σ.

4. $\overline{X}^2$ is consistent for μ^2.

Theorem

Under suitable regularity conditions the MLE $\widehat{\theta}$ is consistent for θ.

To get a feel for why this is so, suppose $X_1,\ldots,X_n$ are *i.i.d.* from $f(x|\theta_0)$, i.e. θ_0 is the "true" parameter value. Look at

$$g(\theta) = E_{\theta_0}\left[\log\left(\frac{f(X|\theta)}{f(X|\theta_0)}\right)\right]$$

By Jensen's inequality,

$$\begin{aligned} g(\theta) &\leq \log E_{\theta_0}\left[\frac{f(X|\theta)}{f(X|\theta_0)}\right] \\ &= \log\int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)dx \\ &= \log\int f(x|\theta)dx = 0 \end{aligned}$$

with equality if and only if

$$P_{\theta_0}(f(X|\theta) = f(X|\theta_0)) = 1$$

i.e. if and only if $\theta = \theta_0$ for an identi£able θ.

So $g(\theta)$ has a strict global maximum at θ_0 with $g(\theta_0) = 0$.

Now look at the average log likelihood:

$$\frac{1}{n}(\log L_n(\theta|X) - \log L(\theta_0|X)) = \frac{1}{n}\ell_n(\theta|X) = \frac{1}{n}\sum \log\frac{f(X_i|\theta)}{f(X_i|\theta_0)}$$

Then

$$E_{\theta_0}\left[\frac{1}{n}\ell_n(\theta|X)\right] = g(\theta)$$

and by the strong law of large numbers,

$$\frac{1}{n}\ell_n(\theta|X) \overset{a.s.}{\rightarrow} g(\theta) < 0$$

for all $\theta \neq \theta_0$ and

$$\frac{1}{n}\ell_n(\theta|X) \overset{a.s.}{\rightarrow} 0$$

for $\theta = \theta_0$.

So for all large n, all θ other than θ_0 are eventually ruled out in pairwise comparisons.

This proves consistency if Θ is a finite set.

It can be made to work if Θ is compact and g, ℓ_n are continuous.

Dropping compactness is hard.

Homework

Problem 10.1

Due Friday, April 4, 2003.

Wednesday, April 2, 2003

Second Midterm Exam

The exam will cover the material covered in readings, in class and in assignments from Chapters 8 and 9.

The exam is closed book.

The exam will include some information on distributions along the lines of the **Table of Common Distributions** in the text.

Friday, April 4, 2003

Approximate Normality

Suppose n is large enough so that $\widehat{\theta}$ is close to θ_0. Then for θ near θ_0,

$$\frac{1}{n}\ell_n(\theta|X) \approx \frac{1}{n}\ell_n(\theta_0|X) + \frac{1}{n}\frac{\partial}{\partial\theta}\ell_n(\theta_0|X)(\theta-\theta_0) + \frac{1}{2n}\frac{\partial^2}{\partial\theta^2}\ell_n(\theta_0|X)(\theta-\theta_0)^2$$

Maximize this quadratic to get the approximate MLE:

$$\widehat{\theta} - \theta_0 = -\frac{\frac{1}{n}\ell_n'(\theta_0|X)}{\frac{1}{n}\ell_n''(\theta_0|X)}$$

Now

$$\frac{1}{n}\ell''(\theta_0|X) = \frac{1}{n}\sum\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0) \overset{a.s.}{\to} -I_1(\theta_0)$$

by the strong law of large numbers. Furthermore,

$$\frac{1}{n}\ell'(\theta_0|X) = \frac{1}{n}\sum\frac{\partial}{\partial\theta}\log f(X_i|\theta_0) = \frac{1}{n}\sum Y_i$$

with

$$\begin{aligned} E[Y_i] &= 0 \\ \mathrm{Var}(Y_i) &= I_1(\theta_0) \end{aligned}$$

So by the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum Y_i \overset{\mathscr{D}}{\to} N(0, I_1(\theta_0))$$

By Slutsky's theorem,

$$\begin{aligned} \sqrt{n}(\widehat{\theta}-\theta_0) &\approx -\frac{\frac{1}{\sqrt{n}}\ell_n'(\theta_0|X)}{\frac{1}{n}\ell_n''(\theta|X)} \\ &\overset{\mathscr{D}}{\to} N\left(0, \frac{I_1(\theta_0)}{I_1(\theta_0)^2}\right) = N(0, I_1(\theta_0)^{-1}) \end{aligned}$$

or

$$\widehat{\theta} \sim \mathrm{AN}(\theta_0, I_n(\theta_0)^{-1})$$

This holds in m dimensions as well.

So under suitable regularity conditions (similar to the ones needed for the CRLB) the MLE is asymptotically normal.

If $\widehat{\theta}$ is the MLE and we want to estimate $\tau(\theta)$, then

$$\tau(\widehat{\theta}) \sim \text{AN}\left(\tau(\theta_0), \frac{\tau'(\theta_0)^2}{I_n(\theta_0)}\right)$$

These results also hold, under some conditions, in some non-*i.i.d.* situations.

The expected information $I_n(\theta_0)$ can be approximated by the *observed information*

$$\widehat{I}_n(\widehat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log L(\widehat{\theta}|X)$$

Examples

1. $X_1,\ldots,X_n$ *i.i.d.* Poisson(λ)

$$\begin{aligned}
\widehat{\lambda} &= \overline{X}\\
\frac{\partial}{\partial\lambda}\log L(\lambda|X) &= \frac{\partial}{\partial\lambda}(\sum X_i \log\lambda - n\lambda) = n\left(\frac{\overline{X}}{\lambda}-1\right)\\
\frac{\partial^2}{\partial\lambda^2}\log L(\lambda|X) &= -\frac{n\overline{X}}{\lambda^2}
\end{aligned}$$

So $I_n(\lambda) = \frac{n}{\lambda}$, $\widehat{I}_n(\widehat{\lambda}) = \frac{n}{\overline{X}}$, and

$$\widehat{\lambda} = \overline{X} \sim \text{AN}(\lambda, \lambda/n)$$

and

$$\frac{\widehat{\lambda}-\lambda}{\sqrt{\overline{X}}/\sqrt{n}} \xrightarrow{\mathscr{D}} N(0,1)$$

2. $X_1,\ldots,X_n$ *i.i.d.* Gamma(α, 1).

$$\log L(\alpha|X) = \text{const} - n\log\Gamma(\alpha) + (\alpha-1)\sum\log X_i$$

Closed form of the MLE is not available. The method of moments estimator

$$\widetilde{\alpha} = \overline{X}$$

is a good initial guess; we can £nd $\widehat{\alpha}$ numerically by solving

$$-\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \frac{1}{n}\sum\log X_i = 0$$

We can approximate the distribution of $\widehat{\alpha}$ as $N(\alpha, I_n(\alpha)^{-1})$, and $I_n(\alpha)$ is approximately

$$\widehat{I}_n(\widehat{\alpha}) = \left[\frac{\Gamma''(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} - \left(\frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})}\right)^2\right]$$

3. $X_1, \ldots, X_n$ *i.i.d.* Geometric(p).

$$\begin{aligned}
f(x|p) &= p^n(1-p)^{\Sigma x_i - n}\\
\log L(p|X) &= n \log p + \left(\sum X_i - n\right) \log(1-p)\\
\frac{\partial}{\partial p} \log L(p|X) &= n\left(\frac{1}{p} - \frac{\overline{X}-1}{(1-p)}\right)\\
\frac{\partial^2}{\partial p^2} \log L(p|X) &= -n\left(\frac{1}{p^2} + \frac{\overline{X}-1}{(1-p)^2}\right)
\end{aligned}$$

So $\widehat{p} = 1/\overline{X}$ and

$$\begin{aligned}
I_n(p) &= \frac{n}{p^2} + \frac{n/p - n}{(1-p)^2} = \frac{n}{p^2(1-p)}\\
\widehat{T}_n(\widehat{p}) &= \frac{n}{\widehat{p}^2} + \frac{n/\widehat{p} - n}{(1-\widehat{p})^2} = \frac{n}{\widehat{p}^2(1-\widehat{p})}
\end{aligned}$$

Homework

Problem 10.3
Problem 10.9 (but only for $e^{-\lambda}$; do not do $\lambda e^{-\lambda}$)

Due Friday, April 11, 2003.

Week 11

Monday, April 7, 2003

Asymptotic Efficiency

Definition

A sequence of estimators W_n is asymptotically efficient for $\tau(\theta)$ if $\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{\mathscr{D}} N(0, v(\theta))$ with

$$v(\theta) = \frac{[\tau'(\theta)]^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}$$

Thus the MLE is asymptotically efficient.

Is this definition reasonable?

Theorem

Suppose a sequence of estimators W_n satisfies is $\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{\mathscr{D}} N(0, v(\theta))$ with $v(\theta)$ continuous. Then, under suitable regularity conditions,

$$v(\theta) \geq \frac{[\tau'(\theta)]^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}$$

The following example shows that the continuity requirement on $v(\theta)$, or something like it, is needed:

Example

Let $X_1,\ldots,X_n$ be *i.i.d.* $N(\theta,1)$ and let

$$W_n = \begin{cases} \overline{X} & \text{if } |\overline{X}| > n^{1/4} \\ a\overline{X} & \text{if } |\overline{X}| \le n^{1/4} \end{cases}$$

Then $\sqrt{n}(W_n-\theta) \xrightarrow{\mathscr{D}} N(0,v(\theta))$ with $v(\theta)=1$ for $\theta \neq 0$ and $v(\theta)=a^2$ for $\theta = 0$. If $a<1$ (e.g. $a=0$) then this estimator is *superefficient*.

Is this example entirely artificial?

Suppose the sequence of estimators W_n of θ satisfies $\sqrt{n}(W_n-\theta) \xrightarrow{\mathscr{D}} N(0,v(\theta))$ for some $v(\theta)$. We can constuct a new sequence V_n by taking a single Newton step from W_n towards the MLE:

$$V_n = W_n - \frac{\ell_n'(W_n)}{\ell_n''(W_n)}$$

This new sequence is asymptotically efficient (under suitable regularity conditions):

$$\begin{aligned}
\sqrt{n}(V_n-\theta) &= \sqrt{n}(W_n-\theta) - \sqrt{n}\frac{\ell_n'(W_n)}{\ell_n''(W_n)} \\
&\approx \sqrt{n}(W_n-\theta) - \sqrt{n}\frac{\ell_n'(\theta)}{\ell_n''(W_n)} - \sqrt{n}\frac{(W_n-\theta)\ell_n''(\theta)}{\ell_n''(W_n)} \\
&= \sqrt{n}\frac{\ell_n'(\theta)}{\ell_n''(\theta)} + \sqrt{n}\frac{\ell_n'(\theta)}{n}\left(\frac{n}{\ell'(W_n)} - \frac{n}{\ell'(\theta)}\right) - \sqrt{n}(W_n-\theta)\left(1-\frac{\ell_n''(\theta)}{\ell_n''(W_n)}\right) \\
&\approx \sqrt{n}\frac{\ell_n'(\theta)}{\ell_n''(\theta)} \xrightarrow{\mathscr{D}} N(0,I(\theta)^{-1})
\end{aligned}$$

with

$$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$$

since $\ell_n''(\theta)/n \xrightarrow{P} -I(\theta)$ and $\ell_n''(W_n)/n \xrightarrow{P} -I(\theta)$.

Non-Normal Limiting Distributions

Some MLE's have non-normal limiting distributions:

Example

Suppose $X_1,\ldots,X_n$ are *i.i.d.* $U[0,\theta]$. Then the MLE is $\widehat{\theta}_n = X_{(n)}$. Now for $x < n\theta$

$$P(n(\theta-\widehat{\theta}) > x) = \left(1-\frac{x}{n\theta}\right)^n \to e^{-x/\theta}$$

So the limiting distribution of $n(\theta - \widehat{\theta})$ is exponential with mean one.

Variance-Stabilizing Transforms

For constructing CI's, it is useful to have normal approximations with variances that do not depend on the paameter.

Supose $W_n \sim \mathrm{AN}(\theta, \sigma_W^2(\theta)/n)$. Then for a smooth function g

$$g(W_n) \sim \mathrm{AN}(g(\theta), g'(\theta)^2 \sigma_W^2(\theta))$$

Suppose $g'(\theta)^2 \sigma_W^2(\theta) \equiv 1$, say. Then

$$g'(\theta) = \frac{1}{\sqrt{\sigma_W^2(\theta)}}$$

and thus

$$g(\theta) = \int \frac{1}{\sqrt{\sigma_W^2(\theta)}}$$

Examples

1. If $X_1, \ldots, X_n$ are *i.i.d*. Poisson(λ), then $W_n = \overline{X}_n \sim \mathrm{AN}(\lambda, \lambda/n)$. So $\sigma_W^2(\lambda) = \lambda$, and

   $$g(\lambda) = \int \frac{1}{\sqrt{\lambda}} d\lambda = 2\sqrt{\lambda}$$

   So $2\sqrt{\overline{X}} \sim \mathrm{AN}(2\sqrt{\lambda}, 1/n)$.

2. If $X_n \sim \mathrm{Binomial}(n, p)$, then $W_n = X_n/n \sim \mathrm{AN}(p, p(1-p)/n)$. So $\sigma_W^2(p) = p(1-p)$, and

   $$\begin{aligned} g(p) &= \int \frac{1}{\sqrt{p(1-p)}} dp \\ &= \int \frac{2}{\sqrt{1-y^2}} dy && p = y^2 \\ &= 2\sin^{-1}(y) \\ &= 2\sin^{-1}(\sqrt{p}) \end{aligned}$$

   So $2\sin^{-1}(\sqrt{X_n/n}) \sim \mathrm{AN}(2\sin^{-1}(\sqrt{p}), 1/n)$.

Homework

Problem: Find the approximate joint distribution of the maximum likelihood estimators in problem 7.14 of the text.

Due Friday, April 11, 2003.

Wednesday, April 9, 2003

Approximating Posterior Distributions

The posterior distribution of θ is given by

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$

Let $\widehat{\theta}$ be the MLE and set $T = \sqrt{n}(\theta - \widehat{\theta})$. Then the density of $T|X$ is

$$f(t|x) \propto \frac{f(x|\widehat{\theta}+t/\sqrt{n})f(\widehat{\theta}+t/\sqrt{n})}{f(x|\widehat{\theta})f(\widehat{\theta})}$$

Note that θ and T are random variables; the conditioning makes x and hence $\widehat{\theta}$ constants.

Now take logs and expand around $\widehat{\theta}$:

$$\begin{aligned}
\log f(t|x) &\approx 0 + \frac{t}{\sqrt{n}}\frac{\partial}{\partial\theta}\log f(x|\widehat{\theta}) + \frac{t^2}{2n}\frac{\partial^2}{\partial\theta^2}\log f(x|\widehat{\theta}) + \log\frac{f(\widehat{\theta}+t/\sqrt{n})}{f(\widehat{\theta})}\\
&= 0 + 0 + \frac{t^2}{2n}\frac{\partial^2}{\partial\theta^2}\log f(x|\widehat{\theta}) + \log\frac{f(\widehat{\theta}+t/\sqrt{n})}{f(\widehat{\theta})}\\
&= -\frac{t^2}{2n}\widehat{I}_n(\widehat{\theta}) + \log\frac{f(\widehat{\theta}+t/\sqrt{n})}{f(\widehat{\theta})}\\
&\approx -\frac{t^2}{2n}\widehat{I}_n(\widehat{\theta})
\end{aligned}$$

If this were exact, we would have

$$\begin{aligned}
&T|X \sim N(0, n/\widehat{I}_n(\widehat{\theta})) && \text{or}\\
&\theta|X \sim N(\widehat{\theta}, \widehat{I}_n(\widehat{\theta})^{-1})
\end{aligned}$$

Under suitable regularity conditions, the postarior distribution of θ is approximately

$$N(\widehat{\theta}, \widehat{I}_n(\widehat{\theta})^{-1})$$

for 1-dimensional and m-dimentional θ.

Some notes:

1. This is a legitimate distributional statement, since $\widehat{\theta}$ and $\widehat{I}_n(\widehat{\theta})$ are £xed conditional on X.

2. The prior has been neglected here. It could be included by using the posterior mode and second derivative at the postarior mode instead of the MLE.

3. The observed information

$$\widehat{I}_n(\widehat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log L(\widehat{\theta}|X)$$

is the right thing to use—it is *not* being used to approximate the expected information $I_n(\theta_0)$.

4. Results are based on the law of large numbers, not the CLT.

Examples

1. $X_1,\ldots,X_n$ *i.i.d.* Bernoulli(p). The prior distribution of p is assumed smooth.

$$\begin{aligned}
\log L(p|x) &= \sum x_i \log p + (n - \sum x_i)\log(1-p)\\
\frac{\partial}{\partial p}\log L(p|x) &= \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p}\\
\widehat{p} &= \overline{x}\\
\frac{\partial^2}{\partial p^2}\log L(p|x) &= -\frac{\sum x_i}{p^2} - \frac{n-\sum x_i}{(1-p)^2}
\end{aligned}$$

So

$$\widehat{I}_n(\widehat{p}) = \frac{n}{\widehat{p}(1-\widehat{p})}$$

and $p|X$ is approximately $N(\widehat{p},\widehat{p}(1-\widehat{p})/n)$.

Supose $n = 100, \sum x_i = 46$. What is $P(p < 0.5|X)$?

$$\begin{aligned}
\text{SD}(p|X) &\approx \sqrt{0.46 \times 0.54/100} \approx 0.05\\
P(p < 0.5|X) &= P\left(\frac{p-0.46}{0.05} < \frac{0.04}{0.05}\right)\\
&\approx P(Z < 0.8) = 0.79
\end{aligned}$$

Similarly,

$$P(0.36 < p < 0.56|X) \approx 0.95$$

2. $X_1,\ldots,X_n$ *i.i.d* $N(\mu,\sigma^2)$, prior on (μ,σ^2) is smooth.

$$\begin{aligned}
\log L(\mu,\sigma^2|x) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2\\
\frac{\partial}{\partial\mu}\log L(\mu,\sigma^2|x) &= \frac{1}{\sigma^2}\sum(x_i-\mu)\\
\frac{\partial}{\partial\sigma^2}\log L(\mu,\sigma^2|x) &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum(x_i-\mu)^2\\
\frac{\partial^2}{\partial\mu^2}\log L(\mu,\sigma^2|x) &= -\frac{n}{\sigma^2}\\
\frac{\partial^2}{(\partial\sigma^2)^2}\log L(\mu,\sigma^2|x) &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum(x_i-\mu)^2\\
\frac{\partial^2}{\partial\mu\partial\sigma^2}\log L(\mu,\sigma^2|x) &= -\frac{1}{2\sigma^4}\sum(x_i-\mu)
\end{aligned}$$

Now $\widehat{\mu}=\overline{x}$, $\widehat{\sigma}^2=\frac{1}{n}\sum(x_i-\overline{x})^2$. So

$$\widehat{I}_n(\widehat{\mu},\widehat{\sigma}^2)=\begin{bmatrix}\frac{n}{\widehat{\sigma}^2} & 0\\ 0 & \frac{n}{2\widehat{\sigma}^4}\end{bmatrix}$$

and thus

$$\widehat{I}_n(\widehat{\mu},\widehat{\sigma}^2)^{-1}=\begin{bmatrix}\widehat{\sigma}^2/n & 0\\ 0 & 2\widehat{\sigma}^4/n\end{bmatrix}$$

So $\mu,\sigma^2|X$ is approximately

$$N\left(\begin{bmatrix}\widehat{\mu}\\ \widehat{\sigma}^2\end{bmatrix},\begin{bmatrix}\widehat{\sigma}^2/n & 0\\ 0 & 2\widehat{\sigma}^4/n\end{bmatrix}\right)$$

Homework

Problem: In the setting of problem 7.14 of the text, suppose $n=100$, $\sum W_i=71$, and $\sum Z_i=7802$. Also assume a smooth, vague prior distribution. Find the posterior probability that $\lambda>100$.

Due Friday, April 11, 2003.

Friday, April 11, 2003

Limiting Distribution of Order Statistics

Suppose Y_n has a Beta(α_n, β_n) distribution, $\alpha_n \to \infty$, $\beta_n \to \infty$, and $p_n = \alpha_n/(\alpha_n + \beta_n) \to p \in (0,1)$. Then

$$\sqrt{\alpha_n + \beta_n}(Y_n - p_n) \xrightarrow{\mathscr{D}} N(0, p(1-p))$$

This can be shown using the central limit theorem for Gamma variables and the bivariate delta method.

Suppose F is continuous with positive density at the p-th population quantile $F^{-1}(p)$. Let $X_1, \ldots, X_n$ be a random sample from F and $U_i = F(X_i)$. Then $U_i \sim U[0,1]$, $X_{(k)} = F^{-1}(U_{(k)})$, and $U_{(k)} \sim$ Beta$(k, n-k+1)$. So for $p \in (0,1)$

$$\sqrt{n}(X_{(\{np\})} - F^{-1}(p)) \approx \sqrt{n}\frac{1}{f(F^{-1}(p))}(U_{(\{np\})} - p) \xrightarrow{\mathscr{D}} N\left(0, \frac{p(1-p)}{f(F^{-1}(p))^2}\right)$$

by the delta method.

Example

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $N(\mu, \sigma^2)$ and let $\widetilde{X}_n$ be the sample median. Then

$$\sqrt{n}(\widetilde{X}_n - \mu) \xrightarrow{\mathscr{D}} N\left(0, \frac{1/4}{1/(\sqrt{2\pi}\sigma)^2}\right) = N\left(0, \frac{\pi}{2}\sigma^2\right)$$

Asymptotic Relative Efficiency

We can compare two asymptotically normal estimators using their asymptotic reative efficiency:

Definition

Suppose $\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{\mathscr{D}} N(0, \sigma_W^2)$ and $\sqrt{n}(V_n - \tau(\theta)) \xrightarrow{\mathscr{D}} N(0, \sigma_V^2)$. Then the asymptotic relative efficiency of V_n to W_n is

$$\text{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}$$

Example

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ $N(\mu, \sigma^2)$. Then the asymptotic relative ef£ciency of the sample median to the sample mean is

$$\text{ARE}(\widetilde{X}_n, \overline{X}_n) = \frac{\sigma^2}{\frac{\pi}{2}\sigma^2} = \frac{2}{\pi} = 0.6366$$

So using the mean we need only 64% as many observatons to achieve the same accuracy as the median.

Example

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ Gamma$(\alpha, 1)$. The method of moments estimator of α is $\overline{X} \sim \text{AN}(\alpha, \alpha/n)$. The maximum likelihood estimator must be calculated numerically, or we can use a one step Newton approximation starting from the MM estimator. The negative second derivative of the single observation log likelihood is

$$-\frac{\partial^2}{\partial\alpha^2}(-\log\Gamma(\alpha) - (\alpha-1)\log x - x) = \frac{d^2}{d\alpha^2}\log\Gamma(\alpha)$$

So the asumptotic relative ef£ciency of the MM estimator to the MLE is

$$\text{ARE}(\overline{X}_n, \widehat{\alpha}_n) = \left[\alpha\frac{d^2}{d\alpha^2}\log\Gamma(\alpha)\right]^{-1}$$

α	0.5	1	2	5	10	100
ARE$(\overline{X}_n, \widehat{\alpha}_n)$	0.4053	0.6079	0.7753	0.9037	0.9509	0.9950

The function $\psi_1(\alpha) = \frac{d^2}{d\alpha^2}\log\Gamma(\alpha)$ is known as the *trigamma function*.

Homework

Problem: Let $X_1, \ldots, X_n$ be a random sample from a Pareto$(1, \beta)$ distribution with density $f(x|\beta) = \beta/x^{\beta+1}$ for $x \geq 1$. Find the asymptotic relative ef£ciency of the method of moments estimator of β to the MLE of β.

Due Friday, April 18, 2003.

Week 12

Monday, April 14, 2003

The Bootstrap

Suppose $X_1, \ldots, X_n$ is a random sample from $F(x|\theta)$ and $W = W(X_1, \ldots, X_n)$ is an estimator of $\tau(\theta)$. There are two forms of boostrap:

- Parametric Bootstrap:
  1. Estimate θ by $\widehat{\theta}$.
  2. Compute $E^*[W] = E[W|\theta = \widehat{\theta}]$ and $\text{Var}^*(W) = \text{Var}(W|\theta = \widehat{\theta})$.
- Nonparametric Bootstrap:
  1. Estimate $F(x|\theta)$ by the empirical distribution F_n.
  2. Compute $E^*[W] = E[W|F = F_n]$ and $\text{Var}^*(W) = \text{Var}(W|F = F_n)$.

Boostrap theory says that, under suitable conditions, $E^*[W] \approx E[W]$ and $\text{Var}^*(W) \approx \text{Var}(W)$ for large n.

Often bootsrtap approximations are more accurate than ones based on the delta method.

How do we compute $E^*[W]$ and $\text{Var}^*(W)$? In some cases we can do this analytically:

Example

Suppose $X_1, \ldots, X_n$ are *i.i.d* $N(\mu, \sigma^2)$ and $W = S^2$. We can use $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma}^2 = S^2$ in a parametric bootstrap. Then

$$E^*[S^2] = \sigma^2\Big|_{\sigma^2 = S^2} = S^2$$
$$\text{Var}^*[S^2] = \frac{2\sigma^4}{n-1}\bigg|_{\sigma^2 = S^2} = \frac{2S^4}{n-1}$$

So the analytic version of the parametric boostrap involves computing $\text{Var}(W|\theta)$ as a function of θ and plugging in an estimate $\widehat{\theta}$ to obtain the parametric bootstrap variance $\text{Var}^*(W|\widehat{\theta})$.

Usually bootstrap variances are computed using computer simulation: Consider the same setting as in the previous example. Given the estimates $\widehat{\mu}$ and $\widehat{\sigma}^2$ we draw a sample $X_1^*,\dots,X_n^*$ from a $N(\widehat{\mu},\widehat{\sigma}^2)$ distribution and compute $W_1^* = W(X_1^*,\dots,X_n^*)$. Repeat this B times to obtain $W_1^*,\dots,W_B^*$. Then approximate $E^*[W]$ and $\text{Var}^*(W)$ by

$$\begin{aligned} E_B^*[W] &= \overline{W}^* \\ \text{Var}_B^*(W) &= \frac{1}{B-1}\sum_{i=1}^{B}(W_i^* - \overline{W}^*)^2 \end{aligned}$$

The law of large numbers implies that $E_B^*[W] \xrightarrow{P} E^*[W]$ and $\text{Var}_B^*(W) \xrightarrow{P} \text{Var}^*(W)$ as $B \to \infty$.

The nonparametric bootstrap uses the same idea, except each sample is drawn from the empirical distribution F_n:

- Draw $X_1^*,\dots,X_n^*$ from F_n
- Compute $W_1^* = W(X_1^*,\dots,X_n^*)$.
- Repeat B times to get $W_1^*,\dots,W_B^*$.

Drawing a random sample from F_n means sampling the observed values of the data with replacement.

Example

Times between failures of air conditioning units, in hours, are

```
> ac
 [1]   3   5   7  18  43  85  91  98 100 130 230 487
```

The sample standard deviation is

```
> sd(ac)
[1] 136.2321
```

Using the `boot` package we can obtain bootstrap estimates of the bias and standard deviation:

```
> boot(ac, function(d, i) sd(d[i]), 1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = ac, statistic = function(d, i) sd(d[i]), R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 136.2321 -14.96460    48.22422
```

Some notes:

- Bootstrapping can be applied to any estimator.
- Bootstrapping requires computing the estimator many times.
- Regression problems can be bootstrapped several ways (cases, residuals, ...)

The nonparametric bootstrap is a shift in philosophy:

- use a model to suggest an estimator
- do not use the model to assess how well the estimator works.

The boostrap uses asymptotics in two ways:

- The data sample size n has to be large for $\text{Var}^*(W)$ to be close to $\text{Var}(W)$.
- The bootstrap sample size B has to be large for $\text{Var}^*_B(W)$ to be close to $\text{Var}^*(W)$.

Homework

Problem: Let $X_1, \ldots, X_n$ be *i.i.d*. Poisson(λ) and let $W = e^{-\overline{X}}$. Find the parametric bootstrap variance $\text{Var}^*(W)$ and show that $\text{Var}^*(W)/\text{Var}(W) \xrightarrow{P} 1$ as $n \to \infty$.

Due Friday, April 18, 2003.

Wednesday, April 16, 2003

Estimating Equations

Many estimators W_n are de£ned by an *estimating equation*

$$\sum_{i=1}^{n} h(X_i, W_n) = 0$$

for some well-behaved function h.

Example

In maximum likelihood estimation

$$h(x,t) = \frac{\partial}{\partial\theta} \log f(x|\theta)\bigg|_{\theta=t}$$

What does W_n estimate? Suppose t^* satis£es

$$E[h(X,t^*)] = 0$$

Generally we will then have $W_n \xrightarrow{P} t^*$.

Expanding the estimating equation around t^* gives

$$0 = \sum h(X_i,t^*) + \sum \frac{\partial}{\partial t} f(X_i,t^*)(W_n - t^*) + \dots$$

and so

$$\begin{aligned}\sqrt{n}(W_n - t^*) &\approx -\frac{\frac{1}{\sqrt{n}}\sum h(X_i,t^*)}{\frac{1}{n}\sum \frac{\partial}{\partial t} h(X_i,t^*)}\\ &\xrightarrow{\mathscr{D}} N\left(0, \frac{E[h(X,t^*)^2]}{(E[\frac{\partial}{\partial t}h(X,t^*)])^2}\right)\end{aligned}$$

We can estimate the asymptotic variance by

$$\widehat{\text{Var}}(\sqrt{n}(W_n - t^*)) = \frac{\frac{1}{n}\sum h(X_i,W_n)^2}{(\frac{1}{n}\sum \frac{\partial}{\partial f} h(X_i,W_n))^2}$$

This is sometimes called the *sandwich estimator*. To see why, we need to look at the multidimensional version.

If θ is $m \times 1$ then we need m equations, so $h(x,t)$ is $m \times 1$ and $\frac{\partial}{\partial t}h(X,t^*)$ is $m \times m$. The covariance matrix of $h(X,t^*)$ is

$$C_h = E[h(X,t^*)h(X,t^*)^T]$$

and

$$\sqrt{n}(W_n - t^*) \xrightarrow{\mathscr{D}} N(0, A_h C_h A_h^T)$$

with

$$A_h = E[\frac{\partial}{\partial t}h(X,t^*)]^{-1}$$

The corresponding estimated asymptotic covariance matrix is $\widehat{A}_h\widehat{C}_h\widehat{A}_h^T$ with $\widehat{A}_h$ and $\widehat{C}_h$ the empirical analogs of A_h and C_h. So $\widehat{C}_h$ is *sandwiched* between $\widehat{A}_h$ and $\widehat{A}_h^T$.

MLE's Using an Incorrect Model

Suppose $X_1,\dots,X_n$ are $i.i.d.$ from g. We use a model $g(x) = f(x|\theta)$ to obtain an estimator W_n. This "MLE" will be consistent for the value θ^* that solves

$$E_g\left[\frac{\partial}{\partial\theta}\log f(X|\theta^*)\right] = 0$$

or

$$\begin{aligned}\theta^* &= \underset{\theta}{\text{argmax}}\, E_g[\log f(X|\theta)]\\ &= \underset{\theta}{\text{argmax}}\, E_g\left[\log\frac{f(X|\theta)}{g(X)}\right]\\ &= \underset{\theta}{\text{argmin}}\int\log\frac{g(x)}{f(x|\theta)}g(x)dx\\ &= \underset{\theta}{\text{argmin}}\,\text{KL}(g(\cdot), f(\cdot|\theta))\end{aligned}$$

$\text{KL}(g,f)$ is the *Kullback-Liebler divergence* from g to f. $\text{KL}(g,f) \geq 0$ for all g,f with equality only if $g = f$ almost everywhere.

If $g(x) = f(x|\theta_0)$ for some θ_0, then $\theta^* = \theta_0$ if the parameter is identi£able. Otherwise, θ^* corresponds to the model in the family $\{f(x|\theta) : \theta\in\Theta\}$ that is closest to $g(x)$ in Kullback-Liebler divergence.

The limiting distribution of W_n is

$$\sqrt{n}(W_n - \theta^*) \xrightarrow{\mathscr{D}} N\left(0, \frac{E_g[(\frac{\partial}{\partial\theta}\log f(X|\theta^*))^2]}{(E_g[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta^*)])}\right)$$

and the asymptotic variance can be estimated by

$$\widehat{\text{Var}}(\sqrt{n}(W_n - \theta^*)) = \frac{\frac{1}{n}\sum(\frac{\partial}{\partial\theta}\log f(X_i|W_n))^2}{[\frac{1}{n}\sum(\frac{\partial^2}{\partial\theta^2}\log f(X_i|W_n))]^2}$$

In the spirit of the nonparametric bootstrap some prefer to use the sandwich estimator to estimate the variance of a maximum likelihood estimator.

In some settings the speci£cation of a mean structure may be easier to justify than the rest of a model. MLE's may then be consistent for the parameters of the mean structure even if the rest of the model is wrong; the sandwich estimator of the variance will then also be consistent.

Homework

1. Let $X_1,\ldots,X_n$ be a random sample that may come from a Poisson distribution with mean λ. Find the sandwich estimator of the asymptotic variance of the MLE $\widehat{\lambda} = \overline{X}$.
2. Let $g(x) = e^{-x}$ for $x > 0$ be an exponential density with mean one and let $f(x|\theta)$ be a $N(\theta,1)$ density. Find the value θ^* corresponding to the density of the form $f(x|\theta)$ that is closest to g in Kullback-Liebler divergence.

Due Friday, April 18, 2003.

Friday, April 18, 2003

Robust Estimators

Many estimators are derived based on an assumed model. If the model is not correct, these estimators may not work very well at all.

Ideally we would like something along these lines:

- optimal or near optimal performance if the model is correct
- small deviations from the model should reduce the performance only a little.
- slightly larger deviations should not cause disasters

Breakdown

One way to think about "no disasters" is *breakdown*:

> Breakdown is the largest fraction of data that can be moved to in£nity before the estimator is pulled to in£nity.

For the mean $\overline{X}$ the breakdown is 0.

For the median the breakdown is 50%.

For the α-trimmed mean

$$\frac{1}{n(1-2\alpha)}\sum_{k=\{\alpha n\}}^{\{(1-\alpha)n\}} X_{(k)}$$

the breakdown is α.

M-Estimators

Many estimators are de£ned as mimimizers of a criterion,

$$\widehat{\theta}_M = \underset{a}{\operatorname{argmin}} \sum \rho(X_i - a)$$

For location models $X_i \sim f(x-\theta)$ taking $\rho(x) = -\log f(x)$ makes $\widehat{\theta}_M$ the MLE. Estimators of this form are therefore called *M-estimators*.

Huber proposed this class and a particular member,

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le k \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| > k \end{cases}$$

Generally the M-estimator $\widehat{\theta}_M$ also solves

$$\sum \psi(X_i - \widehat{\theta}_M) = 0$$

with $\psi = \rho'$. For the Huber M-estimator

$$\psi(x) = \begin{cases} -k & \text{if } x < -k \\ x & \text{if } |x| \leq k \\ k & \text{if } x > k \end{cases}$$

k is a tuning constant; it is sometimes chosen based on a robust measure of scale, such as the IQR.

Suppose θ_0 satisfies $E[\psi(X - \theta_0)] = 0$. Then $\widehat{\theta}_M$ is generally consistent for θ_0 and

$$\sqrt{n}(\widehat{\theta}_M - \theta_0) \xrightarrow{\mathscr{D}} N\left(0, \frac{E[\psi(X_i - \theta_0)^2]}{(E[\psi'(X_i - \theta_0)])^2}\right)$$

One advantage of the M-estimator formulation is that it can be extended to regression settings.

Influence Functions

The influence function (or influence curve) is a useful tool for thinking about the robustness of estimators. To define the influence function, think of an estimator as a functional $T(F_n)$ of the empirical distribution F_n. The corresponding population characteristic is $T(F)$.

The influence function is based on thinking about small "contaminations" in which a point mass of probability δ is added at a point x. That is, $X \sim F_\delta$ means

$$X \sim \begin{cases} F & \text{with probability } 1 - \delta \\ x & \text{with probability } \delta \end{cases}$$

The influence function measures the rate of change of T as the amount of contamination δ changes:

$$\text{IF}(T, x) = \lim_{\delta \downarrow 0} \frac{1}{\delta}(T(F_\delta) - T(F))$$

The influence function is essentially a directional derivative.

For the sample mean

$$T(F_\delta) = (1 - \delta)\mu + \delta x$$

and

$$\frac{1}{\delta}(T(F_\delta) - \mu) = x - \mu$$

so the influence function of the sample mean is

$$\text{IF}(\overline{X}, x) = x - \mu$$

The influence function of an M-estimator is

$$\text{IF}(\widehat{\theta}_M, x) = \frac{\psi(x - \theta_0)}{E[\psi'(X - \theta_0)]}$$

This is bounded for Huber's M-estimator. Bounded influence is a characteristic of robust methods.

For the α-th sample quantile the influence function is

$$\text{IF}(X_{(\{\alpha n\})}, x) = \begin{cases} \frac{\alpha}{f(F^{-1}(\alpha))} & \text{if } x > F^{-1}(\alpha) \\ \frac{\alpha - 1}{f(F^{-1}(\alpha))} & \text{if } x < F^{-1}(\alpha) \end{cases}$$

A useful general result:

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathscr{D}} N(0, E[\text{IF}(T, X)^2])$$

There is a relation between the influence function and the breakdown value of an estimator; the homework problem explores this.

Computing Influence Functions

A variety of techniques are available for computing influence functions. If T is defined by an equation, then implicit differentiation is often a useful approach.

Example

Suppose $T = F^{-1}(\alpha)$ is the α-th population quantile, and suppose F has density f with $f(T) > 0$. Then T satisfies

$$F(T) = \alpha$$

Now

$$F_\delta(T_\delta) = (1 - \delta)F(T_\delta) + \delta 1_{[x,\infty)}(T_\delta) = g(\delta, T_\delta)$$

with

$$g(u, v) = (1 - u)F(v) + u1_{[x,\infty)}(v)$$

The partial derivatives of g are, for $v \neq x$,

$$\frac{\partial}{\partial u} g(u, v) = 1_{[x,\infty)}(v) - F(v)$$
$$\frac{\partial}{\partial v} g(u, v) = (1 - u)f(v)$$

Differentiating the de£ning equation for T with respect to δ and evaluating at $\delta = 0$ produces

$$\begin{aligned}
0 &= \left(\frac{\partial}{\partial u} g(\delta, T_\delta) + \frac{\partial}{\partial v} g(\delta, T_\delta) \frac{d}{d\delta} T_\delta\right)\bigg|_{\delta=0}\\
&= \left(1_{[x,\infty)}(T_\delta) - F(T_\delta) + (1-\delta) f(T_\delta) \frac{d}{d\delta} T_\delta\right)\bigg|_{\delta=0}\\
&= 1_{[x,\infty)}(T) - \alpha + f(T)\mathrm{IF}(T,x)
\end{aligned}$$

and therefore

$$\mathrm{IF}(T,x) = \frac{\alpha - 1_{[x,\infty)}(T)}{f(T)}$$

Homework

Problem 10.30 (b)

Due Friday, April 25, 2003.

Week 13

Monday, April 21, 2003

Large Sample Hypothesis Tests

Informal Methods

Suppose we want to test

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \qquad \text{or one-sided if } \theta \text{ is real-valued} \end{aligned}$$

Suppose W_n is an estimator of θ and W_n is $\sqrt{n}$-consistent and asymptotically normal, i.e. under H_0

$$W_n \sim \text{AN}(\theta_0, \frac{1}{n}\sigma_W^2)$$

Then we can use as a test statistic

$$\frac{W_n - \theta_0}{\sigma_W/\sqrt{n}}$$

which is approximately $N(0,1)$ if $\theta = \theta_0$.

If $\sigma_W = \sigma_W(\theta)$ depends continuously on θ, then

$$\frac{W_n - \theta_0}{\sigma_W(\theta_0)/\sqrt{n}} \sim \text{AN}(0,1)$$

and

$$\frac{W_n - \theta_0}{\sigma_W(W)/\sqrt{n}} \sim \text{AN}(0,1)$$

if $\theta = \theta_0$, so either can be used as the basis for a test.

In some cases we can find a variance stabilizing transformation g such that

$$\sqrt{n}(g(W_n) - g(\theta_0)) \sim \text{AN}(0,1)$$

if $\theta = \theta_0$.

If $\sigma_W = \sigma_W(\psi)$ depends continuously on another parameter ψ, and if V_n is a consistent estimator of ψ, then

$$\frac{W_n - \theta_0}{\sigma_W(V_n)/\sqrt{n}} \sim \text{AN}(0,1)$$

if $\theta = \theta_0$.

Example

Suppose $X_1, \ldots, X_n$ are *i.i.d.* Poisson(λ) and we wish to test the hypotheses

$$\begin{aligned} H_0 &: \lambda = \lambda_0 \\ H_1 &: \lambda \neq \lambda_0 \end{aligned}$$

Then $W_n = \overline{X}_n \sim \text{AN}(\lambda, \lambda/n)$. The variance stabilizing transformation is $g(x) = 2\sqrt{x}$, so $2\sqrt{\overline{X}_n} \sim \text{AN}(\sqrt{\lambda}, 1)$. Thus a test can be based on any one of the statistics

$$\begin{aligned} Z_{n,1} &= \sqrt{n}(\overline{X}_n - \lambda_0)/\sqrt{\lambda_0} \\ Z_{n,2} &= \sqrt{n}(\overline{X}_n - \lambda_0)/\sqrt{\overline{X}_n} \\ Z_{n,3} &= \sqrt{n}\left(2\sqrt{\overline{X}_n} - 2\sqrt{\lambda_0}\right) \end{aligned}$$

in each case rejecting if $|Z_{n,k}|$ is larger than $z_{\alpha/2}$.

If θ is m-dimensional and $W_n \sim \text{AN}(\theta, \Sigma_W)$ with Σ_W nonsingular, then, viewing θ and W_n as $m \times 1$ column vectors,

$$Y_n = n(W_n - \theta_0)^T \Sigma_W^{-1}(W_n - \theta_0) \xrightarrow{\mathscr{D}} \chi^2_m$$

if $\theta = \theta_0$. An approximate level α test is therefore obtained by rejecting H_0 if $Y_n > \chi^2_{m,\alpha}$.

To see why the limiting distribution is approximately χ^2_m suppose $Y \sim N(0, \Sigma)$ with Σ nonsingular, and let A be such that $\Sigma = AA^T$. Such matricies A exist and are nonsingular. Let $Z = A^{-1}Y$. Then

$$Y^T\Sigma^{-1}Y = Y^T(AA^T)^{-1}Y = Y^TA^{-T}A^{-1}Y = (A^{-1}Y)^T(A^{-1}Y) = Z^TZ = \sum_{i=1}^m Z_i^2$$

and

$$Z \sim N(0, A^{-1}\Sigma A^{-T}) = N(0, A^{-1}AA^TA^{-T}) = N(0, I)$$

So $Z_1, \ldots, Z_m$ are *i.i.d.* standard normal and $\sum Z_i \sim \chi^2_m$.

Likelihood Ratio Tests for Large Samples

For many problems where optimal tests can be found, LR tests turn out to be optimal.

Suppose we cannot £nd optimal tests. The LRT may still be a good test to use.

But we may not be able to £nd the distribution of Λ, or a function of Λ, under H_0.

Fortunately, a general result can often be applied.

Suppose

Θ is m-dimensional

Θ_0 is $k < m$-dimensional

Under suitable regularity conditions, $-2\log\Lambda$ is approximately a χ^2_{m-k} random variable if H_0 is true.

To see where this comes from, look at $\Theta_0 = \{\theta_0\}$, $k = 0$. Then let

$$V_n(\theta_0) = \left(\frac{\partial}{\partial\theta}\log f(X|\theta_{0_i})\right)_{i=1,\dots,m}$$

viewed as a column vector. The function $V_n(\theta)$ is called the *score function*, and if $\theta = \theta_0$ then

$$V_n(\theta_0) \sim \mathrm{AN}(0, I_n(\theta_0))$$

Based on a two term Taylor expansion around θ_0 the maximized log likelihood is approximately

$$\log L(\widehat{\theta}) \approx \log L(\theta_0) + \frac{1}{2}V_n(\theta_0)^T I_n(\theta_0)^{-1} V_n(\theta_0)$$

and therefore

$$\Lambda = \frac{L(\theta_0)}{L(\widehat{\theta})} \approx \exp\left\{-\frac{1}{2}V_n(\theta_0)^T I_n(\theta_0)^{-1} V_n(\theta_0)\right\}$$

So if $\theta = \theta_0$, then

$$-2\log\Lambda \approx V_n(\theta_0)^T I_n(\theta_0)^{-1} V_n(\theta_0) \xrightarrow{\mathscr{D}} \chi^2_m$$

The regularity conditions require both restricted and unrestricted MLE problems to be nice:

differentiability

no boundaries—θ_0 must be interior to Θ_0 and Θ.

This rules out one-sided situations like

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

Often we have Θ_0 described by constraints,

$$\Theta_0 = \{\theta : g(\theta) = 0\}$$

for some $g : \mathbb{R}^m \to \mathbb{R}^p$. Then usually $\dim(\Theta_0) = m - p$ and so $-2\log\Lambda$ is approximately χ^2_p.

Example

(N_0, N_1, N_2) are multinomial (n, p_0, p_1, p_2).

$$H_0 : p_i = \binom{2}{i} p^i (1-p)^{2-i}, 0 \le p \le 1$$

i.e. H_0 is that the p_i correspond to a Binomial(2,p) distribution for some p. This might be the case if a particular genetic model is true.

Θ is 2-dimensional (since $p_0 + p_1 + p_2 = 1$).

Θ_0 is 1-dimensional.

$$\Lambda = \frac{(1-\widehat{p})^{2N_0}(2\widehat{p}(1-\widehat{p}))^{N_1}\widehat{p}^{2N_2}}{\widehat{p}_0^{N_0}\widehat{p}_1^{N_1}\widehat{p}_2^{N_2}}$$

with

$$\begin{aligned}\widehat{p} &= \frac{N_1 + 2N_2}{2n}\\ \widehat{p}_i &= \frac{N_i}{n}\end{aligned}$$

Then

$$\begin{aligned}-2\log\Lambda &= 2N_0 \log\frac{\widehat{p}_0}{(1-\widehat{p})^2} + 2N_1 \log\frac{\widehat{p}_1}{2\widehat{p}(1-\widehat{p})} + 2N_2 \log\frac{\widehat{p}_2}{\widehat{p}^2}\\ &= G^2 \text{ statistic}\end{aligned}$$

which is related to the χ^2 statistic.

Homework

Problem: Consider the setting of Problem 10.31. Derive an expression for $-2\log\Lambda$, where Λ is the likelihood ratio test statistic, and £nd the approximate distribution of this quantity under the null hypothesis.

Due Friday, April 25, 2003.

Wednesday, April 23, 2003

Other Likelihood-Based Methods

We saw earlier that under some conditions

$$\widehat{\theta} \sim \text{AN}(\theta, I_n(\theta)^{-1})$$

From this it follows that

$$\frac{\widehat{\theta} - \theta_0}{\sqrt{I_n(\theta_0)^{-1}}} \sim \text{AN}(0,1)$$
$$\frac{\widehat{\theta} - \theta_0}{\sqrt{\widehat{I}_n(\widehat{\theta})^{-1}}} \sim \text{AN}(0,1)$$

if $\theta = \theta_0$ and θ is a scalar, or

$$(\widehat{\theta} - \theta_0)^T I_n(\theta_0)(\widehat{\theta} - \theta_0) \xrightarrow{\mathscr{D}} \chi^2_m$$
$$(\widehat{\theta} - \theta_0)^T \widehat{I}_n(\widehat{\theta})(\widehat{\theta} - \theta_0) \xrightarrow{\mathscr{D}} \chi^2_m$$

if $\theta = \theta_0$ and θ is m-dimensional.

Tests based on these statistics, in particular the second form (for our text), are called *Wald tests*.

A test can also be based on the score function

$$V_n(\theta_0) = \left(\frac{\partial}{\partial \theta} \log f(X|\theta_{0_i})\right)_{i=1,\ldots,m}$$

If $\theta = \theta_0$ then

$$V_n(\theta_0) \sim \text{AN}(0, I_n(\theta_0))$$

and so

$$\frac{V_n(\theta_0)}{\sqrt{I_n(\theta_0)}} \sim \text{AN}(0,1)$$
$$\frac{V_n(\theta_0)}{\sqrt{\widehat{I}_n(\widehat{\theta})}} \sim \text{AN}(0,1)$$

for scalar θ and

$$V_n(\theta_0)^T I_n(\theta_0)^{-1} V_n(\theta_0) \xrightarrow{\mathscr{D}} \chi^2_m$$
$$V_n(\theta_0)^T \widehat{I}_n(\widehat{\theta})^{-1} V_n(\theta_0) \xrightarrow{\mathscr{D}} \chi^2_m$$

for m-dimentional θ.

Tests based on these statistics, in particular the first form, are called *score tests*. One advantage of the first form in particular is that it does not require computation of the MLE.

Example

Suppose $X_1, \ldots, X_n$ are $i.i.d.$ Bernoulli(p) and we want to test the hypotheses

$$\begin{aligned} H_0 &: p = p_0 \\ H_1 &: p \neq p_0 \end{aligned}$$

The score function is

$$\begin{aligned} V_n &= \frac{\partial}{\partial p}\left(\sum X_i \log p + \left(n - \sum X_i\right)\log(1-p)\right) = \frac{\sum X_i}{p} - \frac{1-\sum X_i}{1-p} \\ &= n\left(\frac{\widehat{p}}{p} - \frac{1-\widehat{p}}{1-p}\right) = n\frac{\widehat{p}-p}{p(1-p)} \end{aligned}$$

and the expected and observed information are

$$\begin{aligned} I_n(p) &= \frac{n}{p(1-p)} \\ \widehat{I}_n(p) &= \frac{n}{\widehat{p}(1-\widehat{p})} \end{aligned}$$

The score test statistic is

$$\frac{V_n(p_0)}{\sqrt{I_n(p_0)}} = n\frac{\widehat{p}-p_0}{p_0(1-p_0)}\bigg/\sqrt{\frac{n}{p_0(1-p_0)}} = \sqrt{n}\frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)}}$$

The Wald test statistic is

$$\frac{\widehat{p}-p_0}{\sqrt{\widehat{I}_n(\widehat{p})}} = \sqrt{n}\frac{\widehat{p}-p_0}{\widehat{p}(1-\widehat{p})}$$

If $I_n(p_0)$ is used in the Wald statistic then the Wald and score statistica are identical.

Some notes:

- For discrete data, approximations can sometimes be improved by using continuity corrections.
- For simple null hypotheses, exact p-values can sometimes be computed by simulation.

Homework

Problem 10.38

Due Friday, April 25, 2003.

Friday, April 25, 2003

Approximate Con£dence Sets

Suppose the usual regularity conditions hold. If Θ is k-dimensional, then the LRT for

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

rejects if

$$-2\log\frac{f(x|\theta_0)}{f(x|\widehat{\theta})} > \chi^2_{k,\alpha}$$

(approximately). Inverting this produces

$$C(X) = \left\{\theta : \log\frac{f(x|\theta)}{f(x|\widehat{\theta})} \geq -\frac{1}{2}\chi^2_{k,\alpha}\right\}$$

Likelihood contours are con£dence sets.

Similarly, $(\widehat{\theta}-\theta_0)^T\widehat{I}(\widehat{\theta})(\widehat{\theta}-\theta_0)$ is approximately $\chi^2_{k,\alpha}$ (Wald test). Inverting, or using as a pivot, gives

$$\begin{aligned} C(X) &= \{\theta : (\widehat{\theta}-\theta)^T\widehat{I}(\widehat{\theta})(\widehat{\theta}-\theta) \leq \chi^2_{k,\alpha}\} \\ &= \text{ellipse (ellipsoid)} \end{aligned}$$

Score tests can also be inverted.

$\widehat{I}(\widehat{\theta})$ can be replaced by $I(\theta)$. This makes things more complicated but is sometimes usable.

If $W \sim \text{AN}(\theta, \sigma^2_W/n)$, σ^2_W known, then

$$\frac{W-\theta}{\sigma_W/\sqrt{n}} \sim \text{AN}(0,1)$$

is an approximate pivotal, and

$$W \pm \frac{\sigma^2_W}{\sqrt{n}} z_{\alpha/2}$$

is an approximate $1-\alpha$ level CI.

If $\sigma_W = \sigma_W(\theta)$ is continuous, then

$$\begin{aligned} \frac{W-\theta}{\sigma_W(\widehat{\theta})/\sqrt{n}} &\sim \text{AN}(0,1) \\ \frac{W-\theta}{\sigma_W(\theta)/\sqrt{n}} &\sim \text{AN}(0,1) \end{aligned}$$

are both approximate pivotals.

The second is harder to use but may be more accurate.

Wilks argues that if

$$Q(X,\theta) = \frac{\frac{\partial}{\partial\theta}\log L(\theta|X)}{\sqrt{-E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta|X)\right]}}$$

then $Q(X,\theta) \sim \text{AN}(0,1)$ and an interval obtained by inversion is asymptotically shortest among a certain class of intervals.

It is not always possible to do the inversion.

It may be possible in a different parameterization (try a variance stabilizing transformation).

Of course, shortest in one parameterization is not necessarily shortest in another unless they are linearly related.

Example

For the binomial distribution, $\widehat{p} \sim \text{AN}(p, p(1-p)/n)$.

First approach (Wald interval):

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Second approach (Score interval):

$$\sqrt{n}\frac{\widehat{p}-p}{\sqrt{p(1-p)}} \in 0 \pm z_{\alpha/2}$$

So

$$\begin{aligned}
(\widehat{p}-p)^2 &= \frac{1}{n}p(1-p)z_{\alpha/2}\\
\widehat{p}^2 - 2p\widehat{p} + p^2 &= \frac{1}{n}z^2_{\alpha/2}(p-p^2)\\
\widehat{p}^2 - (2\widehat{p} + \frac{1}{n}z^2_{\alpha/2})p + (1+\frac{1}{n}z^2_{\alpha/2})p^2 &= 0\\
p_{1,2} = \frac{2\widehat{p} + \frac{1}{n}z^2_{\alpha/2} \pm \sqrt{(2\widehat{p}+\frac{1}{n}z^2_{\alpha/2})^2 - 4\widehat{p}^2(1+\frac{1}{n}z^2_{\alpha/2})}}{2(1+\frac{1}{n}z^2_{\alpha/2})}
\end{aligned}$$

Variation: use continuity correction.

Inverting the LRT gives

$$C = \left\{ p : -2\log\left(\frac{p^y(1-p)^{n-y}}{\widehat{p}^y(1-\widehat{p})^{n-y}}\right) \le \chi^2_{1,\alpha} \right\}$$

where $y = \sum x_i$ is the number of successes.

Other options:

- invert exact binomial test
- Agresti and Coull: Add 2 successes and 2 failures to compute $\widehat{p} = (y+2)/(n+4)$, then use Wald interval with $\tilde{n} = n+4$. (Recommended only for $\alpha = 0.05$; for other α adding $(z_{\alpha/2})^2/2$ successes and failures is recommended.)

L. D. Brown, T. T. Cai, and A DasGupta (2001), “Interval estimation for a binomial parameter (with discussion),” *Statistical Science*, **16**, 101–144.

Homework

Problem 10.41

Due Friday, May 2, 2003.

Week 14

Monday, April 28, 2003

Linear and Other Models

In many problems we want to model how a response Y is related to some explanatory variables x. Some forms of models used:

linear model:

$$\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_m + \varepsilon \\
Y &= \beta_0 + \beta_1 x + \varepsilon \\
Y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m + \varepsilon
\end{aligned}$$

nonlinear model:

$$Y = \beta_1 + \beta_2 \exp\{\beta_3 x\} + \varepsilon$$

generalized linear model:

$$\begin{aligned}
Y &\sim \text{Poisson}(\lambda = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_m\}) \\
Y &\sim \text{Bernoulli}(p = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_m))
\end{aligned}$$

with

$$\begin{aligned}
g(x) &= \exp\{x\}/(1+\exp\{x\}) && \text{Logit link} \\
g(x) &= \Phi(x) && \text{Probit link}
\end{aligned}$$

additive model:

$$Y = s_1(x_1) + \ldots s_m(x_m) + \varepsilon$$

where the s_i are "smooth" functions.

Various combinations are possible.

Simplest case: linear model.

Suppose we have n observations that can be written as

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

for $i = 1, \ldots, n$. To include a constant term take $x_i \equiv 1$. The x_{ij} are viewed as fixed constants.

Linear model assumptions:

1. $E[\varepsilon_i] = 0$ for all i.
2. The ε_i are uncorrelated.
3. The ε_i have the same variance, σ^2.
4. The ε_i are jointly normally distributed.

If all of these assumptions hold then the likelihood function for the data is

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\beta))^2\right\}$$

with

$$\mu_i(\beta) = \sum_{j=1}^{p} \beta_j x_{ij}$$

So the maximum likelihood estimator of β is

$$\begin{aligned}\widehat{\beta} &= \operatorname*{argmin}_{\beta} \sum_{i=1}^{n}(y_i - \mu_i(\beta))^2 \\ &= \text{least squares estimator}\end{aligned}$$

and the MLE of σ^2 is

$$\begin{aligned}\widehat{\sigma}^2 &= \operatorname*{argmax}_{\sigma^2} \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\widehat{\beta}))^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i(\widehat{\beta}))^2 \\ &= \frac{1}{n}(\text{sum of squared residuals})\end{aligned}$$

The partial derivatives of the sum of squared deviations are

$$\frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}(y_i - \mu_i(\widehat{\beta}))^2 = -2\sum_{i=1}^{n}(y_i - \mu_i(\widehat{\beta}))\frac{\partial}{\partial \beta_k}\mu_i(\beta)$$

and so the least squares estimators satisfy

$$\sum_{i=1}^{n} \mu_i(\widehat{\beta}) \frac{\partial}{\partial \beta_k} \mu_i(\beta) = \sum_{i=1}^{n} y_i \frac{\partial}{\partial \beta_k} \mu_i(\beta)$$

for $k = 1, \ldots, p$. There are called the *normal equations*.

For the linear model

$$\frac{\partial}{\partial \beta_k} \mu_i(\beta) = x_{ik}$$

Using matrix notation, with

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{1,1} & \ldots & x_{1,p} \\ & \vdots & \\ x_{n,1} & \ldots & x_{n,p} \end{bmatrix} \qquad \mu(\beta) = \begin{bmatrix} \mu_1(\beta) \\ \vdots \\ \mu_n(\beta) \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

we have $\mu(\beta) = X\beta$, and the normal equations can be written as

$$X^T X \beta = X^T y$$

So if the matrix X is of sull rank, then

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

Assuming only that $E[\varepsilon] = 0$ we get

$$\begin{aligned} E[\widehat{\beta}] &= E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E[Y] \\ &= (X^T X)^{-1} X^T X \beta = \beta \end{aligned}$$

So the least squares estimators are unbiased.

If we also assume that $\text{Cov}(\varepsilon) = \sigma^2 I$, then

$$\begin{aligned} \text{Cov}(\widehat{\beta}) &= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

Homework

Problem: Let $x_1, \ldots, x_n$ be constants, and suppose

$$Y_i = \beta_1 (1 - e^{-\beta_2 x_i}) + \varepsilon_i$$

with the ε_i independent $N(0.\sigma^2)$ ramdom variables.

a. Find the normal equations for the least squares estimators of β_1 and β_2.

b. Suppose β_2 is known. Find the least squares estimator for β_1 as a function of the data and β_2.

Due Friday, May 2, 2003.

Wednesday, April 30, 2003

Example

In *simple linear regression* there is a single predictor variable x, so

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon$$

The X matrix is therefore

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and

$$X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \qquad\qquad X^T y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

The inverse of the matrix $X^T X$ is

$$(X^T X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} = \frac{1}{n\sum (x_i - \overline{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

The least squares estimate of the slope β_2 is therefore

$$\widehat{\beta}_2 = \frac{-\sum x_i \sum y_i + n\sum x_i y_i}{n\sum (x_i - \overline{x})^2} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

and the least squares estimate of the intercept β_1 is

$$\widehat{\beta}_1 = \frac{\sum x_i \sum y_i - \sum x_i \sum x_i y_i}{n\sum (x_i - \overline{x})^2} = \overline{y} - \widehat{\beta}_2 \overline{x}$$

If the ε_i are uncorrelated and have common variance σ^2, then the covariance matrix of the least squares estimators is

$$\text{Cov}(\widehat{\beta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} \frac{\sum x_i^2}{n\sum (x_i - \overline{x})^2} & -\frac{\overline{x}}{\sum (x_i - \overline{x})^2} \\ -\frac{\overline{x}}{\sum (x_i - \overline{x})^2} & \frac{1}{\sum (x_i - \overline{x})^2} \end{bmatrix}$$

Some notes:

- The intercept and slope estimates are negatively correlated if $\overline{x} > 0$.
- If we can choose x values within an interval $[a, b]$ and want to obtain the most accurate estimate of the slope, then we would want to take half the observarions at $x = a$ and the other half at $x = b$.

Example

Suppose we have measurements of responses to k different treatments

Treatments				
1	2	3	…	k
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
$\vdots$	$\vdots$	y_{3n_3}	$\ldots$	$\vdots$
y_{1n_1}	$\vdots$			$\vdots$
	y_{2n_2}			y_{kn_k}

The mean responses for the treatments are $\beta_1, \ldots, \beta_k$. So

$$Y_{ij} = \beta_i + \varepsilon_{ij}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. The ε_{ij} are usually assumed to be uncorrelated with mean zero and common variance σ^2. This is called a *one way analysis of variance model*.

This model is a special case of a linear model:

$$Y = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

The $X^T X$ matrix and $X^T y$ vector are very simple:

$$X^T X = \begin{bmatrix} n_1 & 0 & 0 & \ldots & 0 \\ 0 & n_2 & 0 & \ldots & 0 \\ 0 & 0 & & & \vdots \\ \vdots & \vdots & & & 0 \\ 0 & 0 & 0 & \ldots & n_k \end{bmatrix} \qquad X^T y = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{j=1}^{n_k} y_{kj} \end{bmatrix} = \begin{bmatrix} y_{1+} \\ y_{2+} \\ \vdots \\ y_{k+} \end{bmatrix}$$

The least squares estimators are therefore

$$\widehat{\beta} = \begin{bmatrix} \frac{1}{n_1} Y_{1+} \\ \frac{1}{n_2} Y_{2+} \\ \vdots \\ \frac{1}{n_k} Y_{k+} \end{bmatrix} = \begin{bmatrix} \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \vdots \\ \overline{Y}_{k+} \end{bmatrix}$$

If the ε_{ij} are uncorrelated and have common variance σ^2 then the least squares estimators $\widehat{\beta}_1, \ldots, \widehat{\beta}_k$ are uncorrelated and

$$\text{Var}(\widehat{\beta}_i) = \frac{\sigma^2}{n_i}$$

Combinations are also possible:

$$Y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$$
$$Y_{ij} = \mu_i + \beta_i x_{ij} + \varepsilon_{ij}$$

These are sometimes called *anamysis of covariance* models.

Homework

Problem: Let $x_1, \ldots, x_n$ be constants, and suppose

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Let y^* be a constant and let let x^* satisfy

$$y^* = \beta_0 + \beta_1 x^*$$

that is, x^* is the value of x at which the mean response is y^*.

a. Find the maximum likelihood estimator $\widehat{x}^*$ of x^*.

b. Use the delta method to £nd the approximate sampling distribution of $\widehat{x}^*$.

Due Friday, May 2, 2003.

Friday, May 2, 2003

Optimality Properties of Least Squares Estimators

Unbiasedness

If $E[\varepsilon] = 0$ then the least squares estomators are unbiased:

$$E[\widehat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X\beta = \beta$$

Best Linear Unbiased Estimators (BLUE)

The least squares estimators are linear in the data. Suppose $\widetilde{\beta} = AY$ where A is a $p \times n$ matrix of constants. This is a linear estimator. Suppose $\widetilde{\beta}$ is unbiased, i.e.

$$E[\widetilde{\beta}] = AE[Y] = AX\beta = \beta$$

for all β. This means that the $p \times p$ matrix AX is the $p \times p$ identity matrix.

To compare the covariance matrices of $\widetilde{\beta}$ and $\widehat{\beta}$ we need a lemma:

Lemma

Let $U = BY$ and $V = CY$ and let $\text{Cov}(U,V)$ be the matrix of covariances $\text{Cov}(U_i,V_j)$. Then $\text{Cov}(U,V) = B\text{Cov}(Y)C^T$.

The proof involves writing out the sums for U_i and V_j, using bilinearity of the covariance, and recognizing the matrix products in the results.

The covariance matrix of $\widetilde{\beta}$ can be written as

$$\begin{aligned}\text{Cov}(\widetilde{\beta}) &= \text{Cov}((\widetilde{\beta} - \widehat{\beta}) + \widehat{\beta})\\ &= \text{Cov}(\widetilde{\beta} - \widehat{\beta}) + \text{Cov}(\widetilde{\beta} - \widehat{\beta}, \widehat{\beta}) + \text{Cov}(\widehat{\beta}, \widetilde{\beta} - \widehat{\beta}) + \text{Cov}(\widehat{\beta})\end{aligned}$$

Using the lemma and assuming $\text{Cov}(Y) = \sigma^2 I$,

$$\begin{aligned}\text{Cov}(\widetilde{\beta} - \widehat{\beta}, \widehat{\beta}) &= \text{Cov}((AY - (X^T X)^{-1} X^T Y), (X^T X)^{-1} X^T Y)\\ &= \text{Cov}((A - (X^T X)^{-1} X^T)Y, (X^T X)^{-1} X^T Y)\\ &= \sigma^2 (A - (X^T X)^{-1} X^T) X (X^T X)^{-1}\\ &= \sigma^2 (AX - (X^T X)^{-1} X^T X)(X^T X)^{-1}\\ &= \sigma^2 (I - I)(X^T X)^{-1}\\ &= 0\end{aligned}$$

So

$$\text{Cov}(\widetilde{\beta}) = \text{Cov}(\widetilde{\beta} - \widehat{\beta}) + \text{Cov}(\widehat{\beta})$$

The matrix $\text{Cov}(\widetilde{\beta} - \widehat{\beta})$ is a covariance matrix and therefore positive semidetinite. This means

$$\text{Var}(\widetilde{\beta}_k) \geq \text{Var}(\widehat{\beta}_k)$$

for $k = 1, \ldots, p$, or more generally, that for any linear combination $\sum c_k \beta_k = c^T \beta$ the estimator $c^T \widetilde{\beta} = c^T AY$ is linear, unbiased, and has variance no smaller than the corresponding least squares estimator $c^T \widehat{\beta}$:

$$\begin{aligned} E[c^T \widetilde{\beta}] &= c^T E[\widetilde{\beta}] = c^T \beta \\ \text{Var}(c^T \widetilde{\beta}) &= c^T \text{Cov}(\widetilde{\beta}) c \\ &= c^T \text{Cov}(\widetilde{\beta} - \widehat{\beta}) c + c^T \text{Cov}(\widehat{\beta}) c \\ &= \text{Var}(c^T \widetilde{\beta} - c^T \widehat{\beta}) + \text{Var}(c^T \widehat{\beta}) \\ &\geq \text{Var}(c^T \widehat{\beta}) \end{aligned}$$

Ef£ciency, UMVUE

Suppose the ε_i are $i.i.d$ $N(0, \sigma^2)$ and suppose σ^2 is known. Then

$$-\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log L(\beta, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik} = \frac{1}{\sigma^2} (X^T X)_{jk}$$

and therefore the £sher information for β is

$$I_n(\beta) = \frac{1}{\sigma^2} X^T X$$

Since $\text{Cov}(\beta) = \sigma^2 (X^T X)^{-1} = I_n(\beta)^{-1}$, the least squares estimators attain the CRLB and are ef£cient and hence they ae UMVUE's. Since the least squares estimators do not depend on σ^2 they are UMVUS's for unknown σ as well.

Alternative argument: The statistics $X^T Y$ and $\sum Y^2$ are minimal suf£cient suf£cient and $\widehat{\beta}$ is unbiased and depends on the data only through $X^T Y$.

Residuals and the Hat Matrix

The least squares residuals can be written as

$$Y - X\widehat{\beta} = Y - X(X^T X)^{-1} X^T Y = (I - X(X^T X)^{-1} X^T) Y = (I - H) Y$$

where $H = X(X^T X)^{-1} X^T$ is sometimes called the *hat matrix*.

The hat matrix has a number of useful properties:

- It is symmetric.
- It is *idempotent*:

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

- It leaves columns of the X matrix invariant:

$$HX = X(X^TX)^{-1}X^TX = X$$

- It has rank p and trace p:

$$\text{tr}(H) = \text{tr}(X(X^TX)^{-1}X^T) = \text{tr}((X^TX)^{-1}X^TX) = \text{tr}(I_{p\times p}) = p$$

As a result, the residuals can be written as

$$(I-H)Y = (I-H)(X\beta+\varepsilon) = (I-H)\varepsilon$$

and $I-H$ is also idempotent:

$$(I-H)^2 = I-H-H+H^2 = I-2H+H = I-H$$

The trace of $I-H$ is

$$\text{tr}(I-H) = \text{tr}(I) - \text{tr}(H) = n-p$$

Unbiased Estimation of σ^2

Suppose $\text{Cov}(\varepsilon) = \sigma^2 I$. Then

$$\sum(Y_i - \mu_i(\widehat{\beta}))^2 = (Y - X\widehat{\beta})^T(Y - X\widehat{\beta}) = Y^T(I-H)(I-H)Y = \varepsilon^T(I-H)\varepsilon$$

and

$$\begin{aligned} E[\sum(Y_i - \mu_i(\widehat{\beta}))^2] &= E[\varepsilon^T(I-H)\varepsilon] = E[\text{tr}(\varepsilon^T(I-H)\varepsilon)] = E[\text{tr}((I-H)\varepsilon\varepsilon^T)] \\ &= \text{tr}((I-H)E[\varepsilon\varepsilon^T]) = \sigma^2\text{tr}(I-H) = \sigma^2(n-p) \end{aligned}$$

So an unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n-p}(\text{sum of squared residuals})$$

Joint Distribution of Least Squares Estimators and Residuals

Suppose $\text{Cov}(\varepsilon) = \sigma^2 I$. Then residuals and least squares estimators are uncorrelated:

$$\text{Cov}((I-H)Y, (X^TX)^{-1}X^TY) = \sigma^2(I-H)X(X^TX)^{-1} = \sigma^2(X-HX)(X^TX)^{-1} = 0$$

As a result, if errors are jointly normal then residuals and least squares estimators are independent.

The *spectral theorem* states that any symmetric matrix A can be written as $A = UDU^T$ where D is diagonal and U is orthogonal, i.e. $UU^T = U^TU = I$.

For $I-H = UDU^T$ the fact that $I-H$ is idempotent means that $D^2 = D$. So the elements on the diagonal of D satisfy

$$x^2 = x$$

or

$$x^2 - x = x(x-1) = 0$$

So the diagonal elements of D must be zero or one. Since $\text{tr}(I-H) = n-p$ and

$$\text{tr}(I-H) = \text{tr}(UDU^T) = \text{tr}(DU^TU) = \text{tr}(D)$$

there are $n-p$ ones and p zeros.

Suppose the ε_i are $i.i.d$ $N(0,\sigma^2)$. Let

$$Z = \frac{1}{\sigma}U^T\varepsilon$$

Then

$$\text{Cov}(Z) = \frac{1}{\sigma^2}U^T\text{Cov}(\varepsilon)U = U^TU = I$$

and

$$\frac{1}{\sigma^2}(\text{sum of squared residuals}) = \frac{1}{\sigma^2}\varepsilon^T(I-H)\varepsilon = \frac{1}{\sigma^2}\varepsilon^TUDU^T\varepsilon = Z^TDZ = \sum d_iZ_i^2$$

This is the sum the squares of $n-p$ independent standard normals, so

$$\frac{1}{\sigma^2}(\text{sum of squared residuals}) \sim \chi^2_{n-p}$$

Likelihood Ratio Tests

Suppose the ε_i are $i.i.d.$ $N(0,\sigma^2)$ and that we want to test hypotheses about the mean,

$$\begin{aligned} &H_0 : \mu(\beta) \text{ satis£es some restriction} \\ &H_1 : H_0 \text{ is false} \end{aligned}$$

The likelihood ratio statistic will be of the form

$$\Lambda = \left(\frac{\mathrm{SSR}_{\Theta}}{\mathrm{SSR}_{\Theta_0}}\right)^{n/2} = \left(\frac{1}{1+(\mathrm{SSR}_{\Theta_0}-\mathrm{SSR}_{\Theta})/\mathrm{SSR}_{\Theta}}\right)^{n/2}$$

where

$$\begin{aligned}\mathrm{SSR}_{\Theta_0} &= \text{sum of squared residuals for restricted model}\\ \mathrm{SSR}_{\Theta} &= \text{sum of squared residuals for unrestricted model}\end{aligned}$$

So the likelihood ratio test rejects H_0 if

$$\frac{\mathrm{SSR}_{\Theta_0}-\mathrm{SSR}_{\Theta}}{\mathrm{SSR}_{\Theta}}$$

is large.

If the model is linear and H_0 is a *linear hypothesis*, i.e.

$$H_0 : C\beta = b$$

for some $k \times p$ matrix C and k-vector b, then $\mathrm{SSR}_{\Theta_0}-\mathrm{SSR}_{\Theta}$ and SSR_{Θ} are independent. If the rank of C is k and H_0 is true, then

$$\frac{1}{\sigma^2}(\mathrm{SSR}_{\Theta_0}-\mathrm{SSR}_{\Theta}) \sim \chi^2_k$$

So, under H_0,

$$F = \frac{(\mathrm{SSR}_{\Theta_0}-\mathrm{SSR}_{\Theta})/k}{\mathrm{SSR}_{\Theta}/(n-p)} \sim F_{k,n-p}$$

Several alternate forms of the numerator sum of squares difference are available. Let $\|y\|^2 = \sum y_i^2$ and let $\widehat{Y}$ and $\widehat{Y}_0$ be the £tted values under the unrestricted model and the model restricted to satisfy a linear null hopothesis. Then

$$\begin{aligned}\mathrm{SSR}_{\Theta_0} &= \|Y-\widehat{Y}_0\|^2\\ \mathrm{SSR}_{\Theta} &= \|Y-\widehat{Y}\|^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{SSR}_{\Theta_0}-\mathrm{SSR}_{\Theta} &= \|Y-\widehat{Y}\|^2-\|Y-\widehat{Y}_0\|^2\\ &= \|\widehat{Y}-\widehat{Y}_0\|^2\end{aligned}$$

Example

Consider a simple linear regression model

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Suppose we want to test the null hypothesis $H_0 : \beta_2 = 0$. Then

$$\|\widehat{Y} - \widehat{Y}_0\|^2 = \sum(\widehat{\beta}_1 + \widehat{\beta}_2 x_i - \overline{y})^2 = \sum(\widehat{\beta}_2(x_i - \overline{x}))^2 = \widehat{\beta}_2^2 \sum(x_i - \overline{x})^2$$

The F statsitic is therefore

$$F = \frac{\widehat{\beta}_2^2 \sum(x_i - \overline{x})^2}{S^2} = \left(\frac{\widehat{\beta}_2}{\widehat{\text{SE}}(\widehat{\beta}_2)}\right)$$

This is the square of the usual t statistic for testing whether the slope is zero, and the null distribution is $F_{1,n-2}$.

Example

Consider again the simple linear regression model and the linear null hypothesis

$$H_0 : \beta_1 + \beta_2 \overline{x} = a \text{ and } \beta_2 = b$$

for some constants a and b. Then

$$\begin{aligned}\|\widehat{Y} - \widehat{Y}_0\|^2 &= \sum(\widehat{\beta}_1 + \widehat{\beta}_2 x_i - a - b(x_i - \overline{x})^2 = \sum(\widehat{\beta}_1 + \widehat{\beta}_2 \overline{x} - a + (\widehat{\beta}_2 - b)(x_i - \overline{x}))^2 \\ &= n(\widehat{\beta}_1 + \widehat{\beta}_2 \overline{x} - a)^2 + (\widehat{\beta}_2 - b)^2 \sum(x_i - \overline{x})^2\end{aligned}$$

The F statistic is therefore

$$F = \frac{n(\widehat{\beta}_1 + \widehat{\beta}_2 \overline{x} - a)^2 + (\widehat{\beta}_2 - b)^2 \sum(x_i - \overline{x})^2}{2S^2}$$

and a $1 - \alpha$ level con£dence set for $\beta_1 + \beta_2 \overline{x}$ and β_2 is

$$C = \left\{(a,b) : n(\widehat{\beta}_1 + \widehat{\beta}_2 \overline{x} - a)^2 + (\widehat{\beta}_2 - b)^2 \sum(x_i - \overline{x})^2 \le 2S^2 F_{2,n-2,\alpha}\right\}$$