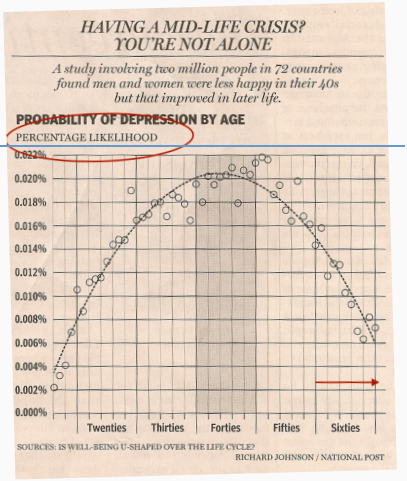
Mathematical Statistics II

STA2212H S LEC9101

Week 1

January 10 2023



Today

Start Recording

1. Course Overview
2. Review of Likelihood MS Ch 5
3. Upcoming seminars of interest

January 16 3.30 – 4.30 Nicholas Horton Details

“Teaching reproducibility and responsible workflows ”

Course Overview

Link

STA 2212S: Mathematical Statistics II
Tuesday, 10.00-13.00 Eastern January 10 – April 4 2023
OI 5150

From the calendar:

This course is a continuation of STA2112H. It is designed for graduate students in statistics and biostatistics. Topics include: Likelihood inference, Bayesian methods, Significance testing, Linear and generalized linear models, Goodness-of-fit, Computational methods
Prerequisite: STA2112H

I will definitely cover the first 3 topics, and the 5th, and we'll see how time goes for the others. "Computational methods" was probably meant to be shorthand for "bootstrap" and "MCMC", and will be touched on in the other topics.

STA 2212S: Mathematical Statistics II
Syllabus

Link

Spring 2023

Week	Date	Methods	References
1	Jan 10	Likelihood inference: review of ML estimation; mis-specified models; computation; nonparametric mle	MS §§5.1–7, SM Ch 4
2	Jan 17	Bayesian estimation; Bayesian inference	MS §5.8; AoS §§ 11.1–4; SM §§11.1,2
3	Jan 24	Optimality in estimation	MS Ch 6; AoS Ch 12; SM §7.1, 11.5.2
4	Jan 31	Interval estimation; Confidence bands	MS §§7.1,2; AoS Ch 7; SM §7.1.4

Link

HW Question Week 1 STA 2212S 2023

Due January 16 *MS, Exercise 5.2*

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent pairs of random variables where X_i and Y_i are i.i.d. $N(\mu_i, \sigma^2)$ random variables.

(a) Find the maximum likelihood estimators of $\mu_1, \ldots, \mu_n$ and σ^2.

(b) Show that the maximum likelihood estimator of σ^2 is not consistent. Does this contradict the theory we have established regarding the consistency of maximum likelihood estimators? Why or why not?

(c) Suppose we observe only $Z_1, \ldots, Z_n$, where $Z_i = X_i - Y_i$. Find the maximum likelihood estimator of σ^2 based on $Z_1, \ldots, Z_n$ and show that it is consistent.

Link

STA2212: Inference and Likelihood

A. Notation

One random variable: Given a model for X which assumes X has a density $f(x;\theta)$, $\theta \in \Theta \subset \mathbb{R}^k$, we have the following definitions:

likelihood function	$L(\theta;x) = c(x)f(x;\theta)$	$\mathcal{L}(\theta)$
log-likelihood function	$\ell(\theta;x) = \log L(\theta;x) = \log f(x;\theta) + a(x)$	
score function	$u(\theta) = \partial\ell(\theta;x)/\partial\theta$	$\ell'(x;\theta)$
observed information function	$j(\theta) = -\partial^2\ell(\theta;x)/\partial\theta\partial\theta^T$	$J(\theta) = \mathrm{E}_\theta\{j(\theta)\}$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}$[1]	$I(\theta)$ (p.245)

Independent observations: When we have X_i independent, identically distributed from $f(x_i;\theta)$, then, denoting the observed sample $\boldsymbol{x} = (x_1,\ldots,x_n)$ we have:

Important Definitions

For $X \sim f(x;\theta)$, we define the following:

- ⋆ Score function: $U(\theta) = \frac{d\ell(x;\theta)}{d\theta}$
- ⋆ Observed information function: $J(\theta) = -\frac{d^2\ell(x;\theta)}{d\theta^2} = -\frac{dU(\theta)}{d\theta}$
- ⋆ Expected information function:

$$I(\theta) = E_\theta[U^2(X;\theta)]$$

Note: Sometimes we add a subscript 1 to these functions to indicate they are based on a single observation. $I_1(\theta)$ $J_1(\theta)$ $U_1(\theta)$

Important Definitions

For $X_1, \ldots, X_n \sim^{iid} f(x;\theta)$, we define the following:

- ⋆ Score function: $U_n(\theta) = \frac{d\ell_n(x_i;\theta)}{d\theta}$
- ⋆ Observed information function: $J_n(\theta) = -\frac{d^2\ell_n(x_i;\theta)}{d\theta^2} = -\frac{dU_n(\theta)}{d\theta}$
- ⋆ Observed (Fisher) information: $J_n(\hat{\theta}_n)$
- ⋆ Expected (Fisher) information function:

$$I_n(\theta) = E_\theta[U_n^2(X_i;\theta)] = nI_1(\theta)$$

Note: Typically the subscript n is removed, but is used to emphasize that we are considering a random sample of size n.

Vector parameters

- model $X \sim f(x;\theta), \theta \in \mathbb{R}^p$ θ is a column vector
- $L(\theta; x)$ map from $\mathbb{R}^p \to \mathbb{R}$
- $\ell'(\theta; x)$ $p \times 1$ vector
- $-\ell''(\theta; x)$ $p \times p$ matrix

Vector parameters

- model $X \sim f(x;\theta), \theta \in \mathbb{R}^p$ θ is a column vector
- $L(\theta; x)$ map from $\mathbb{R}^p \to \mathbb{R}$
- $\ell'(\theta; x)$ $p \times 1$ vector
- $-\ell''(\theta; x)$ $p \times p$ matrix

Properties of maximum likelihood estimators

- maximum likelihood estimators are equivariant
- maximum likelihood estimators are biased — special exceptions
- maximum likelihood estimators have no explicit formula — in general

Asymptotic properties of maximum likelihood estimators

- maximum likelihood estimators are (i) consistent, (ii) asymptotically normal
- (ii) TS expansion

p.256

Your friend the delta-method

MS Th.3.4 and p.148

Suppose $\theta \in \mathbb{R}^p$, $\boldsymbol{x} = (x_1, \ldots, x_p)$

$$a_n(\boldsymbol{x} - \theta) \xrightarrow{d} \boldsymbol{Z},$$

and $g(\boldsymbol{x})$ is continuously differentiable at θ, then $\{g_1(\boldsymbol{x}), \ldots g_k(\boldsymbol{x})\}$

$$a_n\{g(\boldsymbol{x}) - g(\theta)\} \xrightarrow{d} D(\theta)\boldsymbol{Z}$$

where $D(\theta) =$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\{0, I_n^{-1}(\theta)\}$$

$$\sqrt{n}\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} N\{0, g'(\theta)^T I_n(\theta)^{-1} g'(\theta)\}$$

See also AoS §9.9

Example

$X_1, \ldots, X_n$ i.i.d. Gamma (α, λ)

$$f(x_i; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_i^{\alpha-1} \exp(-\lambda x_i)$$

... Example

find a.var($\hat{\mu}$) via mv delta method

Newton-Raphson:

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)$$
$$\hat{\theta} \approx \theta_0 - \{\ell''(\theta_0)\}^{-1}\ell'(\theta_0)$$

- suggests iteration

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \{-\ell''(\hat{\theta}^{(k)})\}^{-1}\ell'(\hat{\theta}^{(k)}) = \hat{\theta}^{(k)} + \frac{S(\hat{\theta}^{(k)})}{H(\hat{\theta}^{(k)})}$$

MS p.270; note change in notation

- requires reasonably good starting values for convergence
- need $-\ell''(\hat{\theta}^{(k)})$ to be non-negative definite
- Fisher scoring replaces $-\ell''(\cdot)$ by its expected value $J(\cdot)$
- N-R and F-S are gradient methods; many improvements have been developed
- solution is a global max only if $\ell(\theta)$ is concave

E-M algorithm: procedure

- complete data $\boldsymbol{X} = (X_1, \ldots, X_n)$, X_i i.i.d. $f_X(x; \theta)$
- observed data $y = (y_1, \ldots, y_m)$, with $y_i = g_i(\boldsymbol{x})$ many-to-one
- joint density $f_Y(y; \theta) = \int_{A(y)} f_X(x; \theta)$ $A(y) = \{x; y_i = g_i(x), i = 1, \ldots, m\}$
- algorithm:
  1. (E step) estimate the complete data log-likelihood function for θ using current guess $\hat{\theta}^{(k)}$
  2. (M step) maximize that function over θ and update to $\hat{\theta}^{(k+1)}$ usually by N-R or Fisher scoring
- likelihood function increases at each step
- can be implemented in complex models
- doesn't automatically provide an estimate of the asymptotic variance

but methods exist to obtain this as a side-product

Example

- $f_X(x_i; \lambda, \mu, \theta) = \alpha \dfrac{e^{-\lambda} x^{\lambda}}{x!} + (1-\alpha) \dfrac{e^{-\mu} x^{\mu}}{x!}, \quad x = 1, 2, \ldots; \lambda, \mu > 0, 0 < \theta < 1$
- Observed data: $x_1, \ldots, x_n$
- Complete data: $(x_1, y_1), \ldots, (x_n, y_n)$; $y_i \sim Bernoulli(\theta)$
- Complete data log-likelihood function:

$$\ell_c(\alpha, \lambda, \mu; y, x) = \sum_{i=1}^{n} y_i \{\log(\alpha) + x_i \log(\lambda) - \lambda\} + \sum_{i=1}^{n} (1 - y_i) \{\log(1-\theta) + x_i \log(\mu) - \mu\}$$

-

$$\mathrm{E}_{\hat{\theta}^{(k)}} \{\ell_c(\alpha, \lambda, \mu; y, x) \mid x\} = \sum_{i=1}^{n} \hat{y}_i \{\log(\alpha) + x_i \log(\lambda) - \lambda\} + \sum_{i=1}^{n} (1 - \hat{y}_i) \{\log(1-\alpha) + x_i \log(\mu) - \mu\}$$

- $\hat{y}_i = \mathrm{E}(Y_i \mid x_i; \hat{\theta}^{(k)})$ see p.280 for exact value
- maximizing values of α, λ, μ can be obtained in closed form p.281

AoS likes to work with $\log \mathcal{L}_n(\theta) / \mathcal{L}_n(\hat{\theta}^{(k)})$

... Example

General-purpose Optimization

Description

General-purpose optimization based on Nelder–Mead, quasi-Newton and conjugate-gradient algorithms. It includes an option for box-constrained optimization and simulated annealing.

Usage

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN",
                 "Brent"),
      lower = -Inf, upper = Inf,
      control = list(), hessian = FALSE)

optimHess(par, fn, gr = NULL, ..., control = list())
```

Arguments

Notes on optimization: Tibshirani, Pena, Kolter CO 10-725 CMU

- Goal: $\max_\theta \ell(\theta; \boldsymbol{x})$
- Solve: $\ell'(\hat{\theta}; \boldsymbol{x}) = 0$
- Iterate: $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \{j(\hat{\theta}^{(t)})\}^{-1}\ell'(\hat{\theta}^{(t)})$
- Rewrite: $j(\hat{\theta}^{(t)})(\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}) = \ell'(\hat{\theta}^{(t)})$ $B\Delta\theta = -\nabla\ell(\theta)$
- Quasi-Newton:
  - approximate $j(\hat{\theta}^{(t)})$ with something easy to invert
  - use information from $j(\hat{\theta}^{(t)})$ to compute $j(\hat{\theta}^{(t+1)})$
- optimization notes add a step size to the iteration $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \epsilon_t\{j(\hat{\theta}^{(t)})\}^{-1}\ell'(\hat{\theta}^{(t)})$

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent"),
      lower = -Inf, upper = Inf, control = list(), hessian = FALSE)
```

Regularity conditions

MS Thm 5.4

- (B1) The parameter space Θ is an open subset of $\mathbb{R}^p$
- (B2) The set $A = \{x : f(x;\theta) > 0\}$ does not depend on θ
- (B3) $\ell(\theta)$ is three times continuously differentiable on A
- (B4) $\mathrm{E}_\theta\{\ell'(\theta; X_i)\} = 0 \forall \theta$ and $\mathrm{Cov}\{\ell'(\theta; X_i)\} = I(\theta)$ is positive definite $\forall \theta$
- (B5) $\mathrm{E}_\theta\{-\ell''(\theta; X_i)\} = J(\theta)$ is positive definite $\forall \theta$
- (B6) For each $\theta, \delta > 0, 1 \leq j, k, l, \leq p$,

$$\left| \frac{\partial^3 \ell(\theta^*; x_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} \right| \leq M_{jkl}(\theta^*),$$

for $||\theta - \theta^*|| \leq \delta$, where $\mathrm{E}_\theta\{M_{jkl}(X_i)\} < \infty$

Regularity conditions

MS Thm 5.4

- (B1) The parameter space Θ is an open subset of $\mathbb{R}^p$
- (B2) The set $A = \{x : f(x;\theta) > 0\}$ does not depend on θ
- (B3) $\ell(\theta)$ is three times continuously differentiable on A
- (B4) $\mathrm{E}_\theta\{\ell'(\theta; X_i)\} = 0 \forall \theta$ and $\mathrm{Cov}\{\ell'(\theta; X_i)\} = I(\theta)$ is positive definite $\forall \theta$
- (B5) $\mathrm{E}_\theta\{-\ell''(\theta; X_i)\} = J(\theta)$ is positive definite $\forall \theta$
- (B6) For each $\theta, \delta > 0, 1 \leq j, k, l, \leq p$,

$$\left| \frac{\partial^3 \ell(\theta^*; x_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} \right| \leq M_{jkl}(\theta^*),$$

for $||\theta - \theta^*|| \leq \delta$, where $\mathrm{E}_\theta\{M_{jkl}(X_i)\} < \infty$

Regularity conditions

MS Thm 5.4

- (B1) The parameter space Θ is an open subset of $\mathbb{R}^p$
- (B2) The set $A = \{x : f(x;\theta) > 0\}$ does not depend on θ
- (B3) $\ell(\theta)$ is three times continuously differentiable on A
- (B4) $\mathrm{E}_\theta\{\ell'(\theta; X_i)\} = 0 \forall \theta$ and $\mathrm{Cov}\{\ell'(\theta; X_i)\} = I(\theta)$ is positive definite $\forall \theta$
- (B5) $\mathrm{E}_\theta\{-\ell''(\theta; X_i)\} = J(\theta)$ is positive definite $\forall \theta$
- (B6) For each $\theta, \delta > 0, 1 \leq j, k, l, \leq p$,

$$\left| \frac{\partial^3 \ell(\theta^*; x_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} \right| \leq M_{jkl}(\theta^*),$$

for $||\theta - \theta^*|| \leq \delta$, where $\mathrm{E}_\theta\{M_{jkl}(X_i)\} < \infty$

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x; \theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$

notation

- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$ notation
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?
- define the parameter $\theta(F)$ by

$$\int_{-\infty}^{\infty} \ell'\{x;\theta(F)\} dF(x) = 0$$

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$

- $$\sigma^2 = \frac{\int [\ell'\{x;\theta(F)\}]^2 dF(x)}{(\int [\ell''\{x;\theta(F)\}]^2 dF(x))^2}$$

Misspecified models

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$
- $$\sigma^2 = \frac{\int [\ell'\{x; \theta(F)\}]^2 dF(x)}{(\int [\ell''\{x; \theta(F)\}]^2 dF(x))^2}$$
- more generally,
$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N\{0, G^{-1}(F)\}$$
- $$G(F) = J(F) I^{-1}(F) J(F),$$
- $$J(F) = \int -\ell''\{\theta(F); x_i\} dF(x_i), \quad I(F) = \int \{\ell'(\theta(F); x_i)\}\{\ell'(\theta(F); x_i)\}^T dF(x_i)$$

Godambe information

Multi-parameter example: logistic regression

```
Boston$crim2 <- Boston$crim > median(Boston$crim) # define binary response
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
                  data = Boston) #fit logistic regression
summary(Boston.glm)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -34.103704   6.530014  -5.223 1.76e-07 ***
zn           -0.079918   0.033731  -2.369  0.01782 *
indus        -0.059389   0.043722  -1.358  0.17436
chas          0.785327   0.728930   1.077  0.28132
nox          48.523782   7.396497   6.560 5.37e-11 ***
rm           -0.425596   0.701104  -0.607  0.54383
age           0.022172   0.012221   1.814  0.06963 .
```

... Example: logistic regression

```
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
                  data = Boston) #fit logistic regression
confint(Boston.glm)
Waiting for profiling to be done...
                    2.5 %        97.5 %
(Intercept) -47.480389822 -21.699753794
zn            -0.152359922  -0.020567540
indus         -0.149113408   0.024168460
chas          -0.646429219   2.233443233
nox           34.967619055  64.088411260
rm            -1.811639107   0.950196261
age           -0.001231256   0.046865843
dis            0.280762523   1.140619391
rad            0.376833861   0.975898274
tax           -0.012038221  -0.001324887
```

Multi-parameter setting

`Waiting for profiling to be done` – what's profiling?

`Waiting for profiling to be done` – what's profiling?

Profile likelihood function

`Waiting for profiling to be done` – what's profiling?

... Profile likelihood function

`Waiting for profiling to be done` – what's profiling?

… Profile likelihood function

`Waiting for profiling to be done` – what's profiling?

Figure 4.1 Likelihoods for the spring failure data at stress 950 N/mm^2. The upper left panel is the likelihood for the exponential model, and below it is a perspective plot of the likelihood for the Weibull model. The upper right panel shows contours of the log likelihood for the Weibull model; the exponential likelihood is obtained by setting $\alpha = 1$, that is, slicing L along the vertical dotted line. The lower right panel shows the profile log likelihood for α, which corresponds to the log likelihood values along the dashed line in the panel above, plotted against α.