Topics in Likelihood Inference

STA4508H

Nancy Reid
University of Toronto

January 19, 2022

Various 'types' of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood — misspecified models
4. empirical likelihood, penalized likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

OH Mon 7-8 pm
Tue 5-6 pm

Various 'types' of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood ⊕ Cox 1972 proportional hazards regression
3. quasi-likelihood, composite likelihood — misspecified models
4. empirical likelihood, penalized likelihood — Cox regression
5. simulated likelihood, indirect inference
6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Various ‘types’ of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood — misspecified models
4. empirical likelihood, penalized likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Various ‘types’ of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood misspecified models
4. empirical likelihood, penalized likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Various ‘types’ of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood misspecified models
4. empirical likelihood, penalized likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Various 'types' of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood misspecified models
4. empirical likelihood, penalized likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, h-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Recap

- likelihood function is proportional to the probability density of the observed data or observable data

- need to assume a probability model in order to write down a likelihood function

- these models are usually parametric, i.e. a class of models that vary with a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$

- but are sometimes non-parametric, in the sense that Θ might be an infinite-dimensional space

  e.g. the class of all twice-differentiable functions

  e.g. the intensity function for a Poisson process

- random effects model: why do we integrate out the random effects?

$$y_{ij} = \underline{\mu + x_i^T\beta} + \varepsilon_{ij}$$

$j = 1, \dots, n_i$ $\quad \varepsilon_{ij} \overset{\text{ass}}{\sim} \text{iid } N(0, \sigma^2)$

$i = 1, \dots, n$

$$y_{ij} = \mu + x_i^T\beta + b_i + \varepsilon_{ij}$$

$j = 1, \dots, n_i$ $\quad b_i \sim N(0, \sigma_b^2)$

$i = 1, \dots, n$ $\quad \perp \quad \varepsilon_{ij} \sim N(0, \sigma^2)$

$\cancel{\ell(\beta, \sigma^2) = y_{ij}} \mid x$

$$\ell(\beta, \sigma; y) = \prod_i \prod_j f(y_{ij}; \beta, \sigma^2) \mid x$$

$$= \prod \prod \frac{e^{-\frac{1}{2\sigma^2}(y_{ij} - x_i^T\beta)^2}}{\sqrt{2\pi}\sigma}$$

$y_{ij} \mid b_i$ ind't over j

$\sim N(\mu + x_i^T\beta, \sigma^2)$

$$f(y_i \mid x_i, b_i; \beta, \sigma^2)$$

$$= \prod_j f(y_{ij} \mid x_i, b_i; \beta, \sigma^2)$$

$$\prod_{i=1}^{n} f(y_i \mid x_i; \beta, \sigma^2, \sigma_b^2) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} \underbrace{f(y_{ij} \mid b_i; \beta, \sigma^2)}_{\text{Bernoulli}} \cdot f(b_i; \sigma_b^2)\, db_i$$

$$y \sim MVN(-, -) = L(\beta, \sigma^2, \sigma_b^2; y)$$

$\mathrm{var}(y_{ij}) = \sigma_b^2 + \sigma^2$

$E(y_{ij}) = \mu + x_i^T\beta$

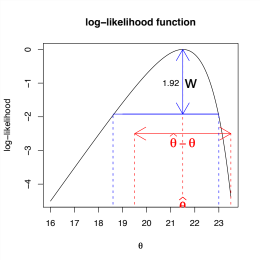
$\mathrm{cov}(y_{ij}, y_{ij'}) = \sigma_b^2$

... Recap

SModels §9.4.1

$= cor(b_i + \varepsilon_{ij}, b_i + \varepsilon_{ij'})$

- several examples: regression, time series, continuous time processes, correlated binary data, etc.
- several examples of likelihood functions that involve integration — complicated
- an example where the likelihood function can't be written down completely — Ising model
- these examples meant to motivate variations on the usual likelihood function — to come
- notation and derived quantities: score function, observed and expected Fisher information, maximum likelihood estimate, likelihood ratio statistic



Don't forget the handout

STA 4508: Likelihood and derived quantities January 2022

Given a model for Y which assumes Y has a density $f(y;\theta)$, $\quad \theta \in \Theta \subset \mathbb{R}^d$, we have the following definitions:

observed likelihood function	$L(\theta; y) = c(y) f(y;\theta)$
log-likelihood function	$\ell(\theta; y) = \log L(\theta; y) = \log f(y;\theta) + a(y)$
score function	$U(\theta) = \partial \ell(\theta; y)/\partial\theta$
observed information function	$j(\theta) = -\partial^2 \ell(\theta; y)/\partial\theta\partial\theta^T$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta U(\theta) U(\theta)^T$ (called $i_1(\theta)$ in CH)

When we have Y_i independent, identically distributed from $f(y_i;\theta)$, then, denoting the observed sample $y = (y_1, \ldots, y_n)$ we have:

log-likelihood function	$\ell(\theta) = \ell(\theta; y) + a(y)$	$O_p(n)$
maximum likelihood estimate	$\hat\theta = \hat\theta(y) = \arg\sup_\theta \ell(\theta)$	$\theta + O_p(n^{-1/2})$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta) = U_+(\theta)$	$O_p(n^{1/2})$
observed information function	$j(\theta) = -\ell''(\theta) = -\ell(\theta; Y)$	$O_p(n)$
observed (Fisher) information	$j(\hat\theta)$	
expected (Fisher) information	$i(\theta) = E_\theta\{U(\theta)U(\theta)^T\} = n i_1(\theta)$	$O(n)$,

where with the risk of some confusion we use the same notation. Sometimes the expected Fisher information is defined instead as $i(\theta) = \mathrm{E}_\theta\{-\partial U(\theta; Y)/\partial\theta^T\}$ (e.g.

... Recap: inference based on likelihood

- "pure likelihood": values of θ are plausible if $L(\hat{\theta})/L(\theta)$ not too large

  or $L(\theta)/L(\hat{\theta})$ not too small

- Bayesian inference: posterior $\propto$ Likelihood $\times$ prior $\quad \pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$

  $\int \pi(\theta | y) d\theta = 1$ $\qquad = L(\theta; y)\pi(\theta)/m(y)$

- frequentist: quantities derived from the likelihood function have "good" properties

  behave well when we have large samples from the model

- also frequentist: pivotal quantities derived from the likelihood function can be used to construct p-value functions

  also called significance functions

- p-value functions provide nested sets of confidence intervals

  if monotone in θ

$y_1, \dots, y_n$ iid $f(y;\theta)$ $\theta \in \mathbb{R}$

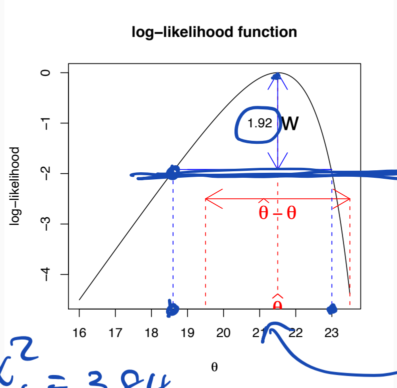
1. (*) $\frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N\{0, i_1(\theta)\}$ CLT

2. $\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N\{0, i_1^{-1}(\theta)\}$ Taylor

3. $= 2\{\ell(\hat\theta) - \ell(\theta)\} \xrightarrow{d} \chi^2_1$ "

$w(\theta) \dot\sim \chi^2_1$

95% pt of $\chi^2_1 = 3.84$



$U(\theta) \dot\sim N(0, i_n(\theta))$

$\hat\theta \dot\sim N(\theta, 1/i_n(\theta))$

Leading to a trio of approximate confidence intervals:

1. $\boxed{\{\theta : |U(\theta) i^{-1/2}(\theta)| \le z_{1-\alpha/2}\}}$
2. $\{\theta : |(\hat\theta - \theta) i^{1/2}(\hat\theta)| \le z_{1-\alpha/2}\}$ ✓
3. $\{\theta : 2[\ell(\hat\theta) - \ell(\theta)]\} \le \chi^2_{1,1-\alpha}$ ✓

$\hat\theta \pm 1.96 \sqrt{\frac{1}{i_n(\hat\theta)}}$

$\sim$ 95% CI for θ

symmetric about $\hat\theta$

captures the asymmetry

p-value functions of θ

- or leading to a trio of approximate pivotal quantities

$$U(\theta) \dot{\sim} N(0, \ldots)$$

$$r_u(\theta) \dot{\sim} N(0,1)$$

$$
\begin{aligned}
1.\ \ r_u(\theta) &= U(\theta)j^{-1/2}(\hat\theta) \dot{\sim} N(0,1) \\
2.\ \ \underline{r_e(\theta)} &= \underline{(\hat\theta-\theta)j^{1/2}(\hat\theta)}, \ \dot{\sim} N(0,1) \\
3.\ \ \underline{r(\theta)} &= \text{sign}(\hat\theta-\theta)[\underbrace{2\{\ell(\hat\theta)-\ell(\theta)\}}_{w(\theta)\ \sim\chi^2_1}]^{1/2} \ \dot{\sim} N(0,1)
\end{aligned}
$$

piv. q: $f(y;\theta)$ with a known dist^{n}

$$C_{.95} \doteq \{\theta : -1.96 \le \underset{\approx}{r}(\theta) \le 1.96\} \Longleftrightarrow \{\theta : \underline{\theta_L \le \theta \le \theta_u}\}$$

pivoting

$$\{\theta : -1.96 \le r_e(\theta) \le 1.96\} \Longleftrightarrow \{\theta : \cancel{\theta}\ -1.96 \le (\hat\theta-\theta)j^{1/2}(\hat\theta) \le 1.96\}$$

p-value functions of θ

$\Longleftrightarrow \{ \hat{\theta} \pm 1.96 \cdot j^{-1/2}(\hat{\theta}) \}$

- or leading to a trio of approximate pivotal quantities

$$
\begin{aligned}
1.\quad r_u(\theta) &= U(\theta) j^{-1/2}(\hat{\theta}) \stackrel{\cdot}{\sim} N(0,1) \\
2.\quad r_e(\theta) &= (\hat{\theta} - \theta) j^{1/2}(\hat{\theta}), \\
3.\quad r(\theta) &= \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2}
\end{aligned}
$$

- $\Pr\{r_u(\theta) \le r_u^0(\theta)\} \doteq \Phi\{r_u^0(\theta)\}$

under sampling from the model $f(y;\theta) = f(y_1, \ldots, y_n; \theta)$

p-value functions of θ

- or leading to a trio of approximate pivotal quantities

$$
\begin{aligned}
1.\ r_u(\theta) &= U(\theta)j^{-1/2}(\hat\theta) \mathrel{\dot\sim} N(0,1)\\
2.\ r_e(\theta) &= (\hat\theta-\theta)j^{1/2}(\hat\theta),\\
3.\ r(\theta) &= \text{sign}(\hat\theta-\theta)[2\{\ell(\hat\theta)-\ell(\theta)\}]^{1/2}
\end{aligned}
$$

- $\Pr\{r_u(\theta) \le r_u^0(\theta)\} \doteq \Phi\{r_u^0(\theta)\}$ under sampling from the model $f(y;\theta) = f(y_1,\ldots,y_n;\theta)$

- and a trio of p-value functions of θ, for fixed data

$H_0: \theta = \theta_0$ $\quad$ p-value $= \Pr_{\theta_0}\{(\underset{\text{r.v.}}{\hat\theta}-\theta_0)j^{1/2}(\hat\theta) \ge (\underset{\text{obs value}}{\hat\theta^0}-\theta_0)j^{1/2}(\hat\theta^0)\}$

$= \Pr_{\theta_0}\{r_e(\theta_0) \ge r_e^{obs}(\theta_0)\}$

$= \Pr_{\theta_0}\{\hat\theta \ge \hat\theta^{obs}\}$

p-value functions of θ

- or leading to a trio of approximate pivotal quantities

Efron & Hinkley '78

$$
\begin{aligned}
1.\quad \underline{r_u(\theta)} &= U(\theta)\underline{j^{-1/2}(\hat\theta)} \,\dot\sim\, \underline{N(0,1)} \\
2.\quad \underline{r_e(\theta)} &= (\hat\theta-\theta)\underline{j^{1/2}(\hat\theta)}, \quad \dot\sim N(0,1) \\
3.\quad r(\theta) &= \text{sign}(\hat\theta-\theta)[2\{\ell(\hat\theta)-\ell(\theta)\}]^{1/2} \quad \sim N(0,1)
\end{aligned}
$$

Sometimes use $i(\theta)$ inst. of j.

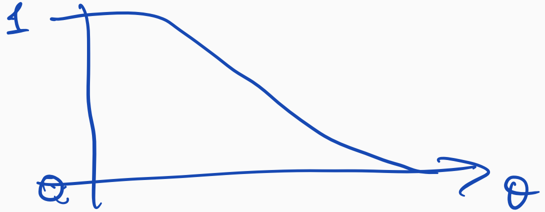
- $\Pr\{r_u(\theta) \le r_u^0(\theta)\} \doteq \Phi\{r_u^0(\theta)\}$

under sampling from the model $f(y;\theta) = f(y_1,\dots,y_n;\theta)$

- and a trio of p-value functions

of θ, for fixed data

- similarly



$$
\begin{aligned}
1.\quad p_u(\theta) &= \Phi\{r_u^0(\theta)\}, \\
2.\quad p_e(\theta) &= \Phi\{r_e(\theta)\} \\
3.\quad p_r(\theta) &= \Phi\{r(\theta)\}
\end{aligned}
$$

$$\Phi(x) = \int^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz$$

Observed and expected Fisher information

$$\frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N(0, \cancel{} \; i_1(\theta)) \quad \text{under } f(y;\theta)$$

iid sampling

app. 1 $\quad \cancel{} \dfrac{U(\theta)}{\sqrt{i_n(\theta)}} \sim N(0,1)$ $\qquad i_n(\theta) = n i_1(\theta)$

2 $\quad \dfrac{U(\theta)}{\sqrt{i_n(\hat\theta)}} \sim N(0,1)$ $\qquad \dfrac{U(\theta)}{\sqrt{i_n(\theta)}} \cdot \dfrac{\sqrt{i_n(\theta)}}{\sqrt{i_n(\hat\theta)}}$

$\downarrow$ 1 $\qquad \downarrow p$ 1 $\qquad$ Slutsky

$$\boxed{\begin{array}{l} X_n \xrightarrow{d} X \\ Y_n \xrightarrow{p} a \end{array} \Rightarrow X_n Y_n \xrightarrow{d} aX}$$

$\times$ 1

3. $\quad \dfrac{U(\theta)}{\sqrt{j_n(\hat\theta)}} \sim N(0,1)$

$j_n(\hat\theta) = -\ell''(\hat\theta)$ $\qquad i_n(\theta) = E\{-\ell''(\theta)\}$

Example: Exponential

$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 \pm \cdots$

- $f(y_i;\theta) = \theta e^{-y_i\theta}, \quad i = 1, \dots, n$
- $\ell(\theta) =$ $n\log\theta - \theta\sum y_i$
- $\ell'(\theta) =$ $n/\theta - \sum y_i$ $\Rightarrow$ $n/\hat\theta = \sum y_i \Rightarrow \hat\theta = \frac{n}{\sum y_i} = \frac{1}{\bar y}$
- $\ell''(\theta) =$ $-n/\theta^2$ $\quad i_n(\theta) = n/\theta^2 \quad i_1(\theta) = 1/\theta^2 \quad j(\theta) = n/\theta^2 \quad j(\hat\theta) = n/\hat\theta^2$
- $r_u(\theta) =$ $u(\theta)\, j(\hat\theta)^{-1/2} = \left(\frac{n}{\theta} - \frac{n}{\hat\theta}\right)\left(\frac{n}{\hat\theta^2}\right)^{-1/2} = \left(\frac{n}{\theta} - \frac{n}{\hat\theta}\right)\frac{\hat\theta}{\sqrt n} = \sqrt n\left(\frac{\hat\theta}{\theta} - 1\right) \leftarrow$
- $r_e(\theta) =$ $(\hat\theta - \theta)\, j^{+1/2}(\hat\theta) = (\hat\theta - \theta)\frac{\sqrt n}{\hat\theta} = \sqrt n\left(1 - \frac{\theta}{\hat\theta}\right) \uparrow$
- $r(\theta) =$ $\left[2\left\{n\log\hat\theta - \frac{\hat\theta n}{\hat\theta} - n\log\theta + \theta n/\hat\theta\right\}\right]^{1/2}$

$= \left[2n\left\{\log(\hat\theta/\theta) - (1 - \theta/\hat\theta)\right\}\right]^{1/2}$ $\quad \log\left(\frac{\hat\theta}{\theta}\right) = \log\left(\frac{\hat\theta}{\theta} - 1 + 1\right)$

expand $\log(\theta\bar y)$ around 1 to get asymptotic equivalence to r_e, r_u

Example: Exponential

- $f(y_i;\theta) = \theta e^{-y_i\theta}, \quad i = 1, \ldots, n$
- $\ell(\theta) = n\log\theta - n\theta\bar{y}$
- $\ell'(\theta) = \frac{n}{\theta} - n\bar{y}$

$\hat{\theta} = \bar{y}^{-1}$

- $\ell''(\theta) = -\frac{n}{\theta^2}$

$\frac{1}{\theta y} - 1$

$y \sim \exp(\theta)$

- $r_u(\theta) = \frac{1}{\sqrt{n}}\ell'(\theta)j^{-1/2}(\hat{\theta}) = \sqrt{n}(\frac{1}{\theta\bar{y}} - 1)$

$\hat{\theta} = \frac{n}{\Sigma y_i} = \frac{1}{\bar{y}}$

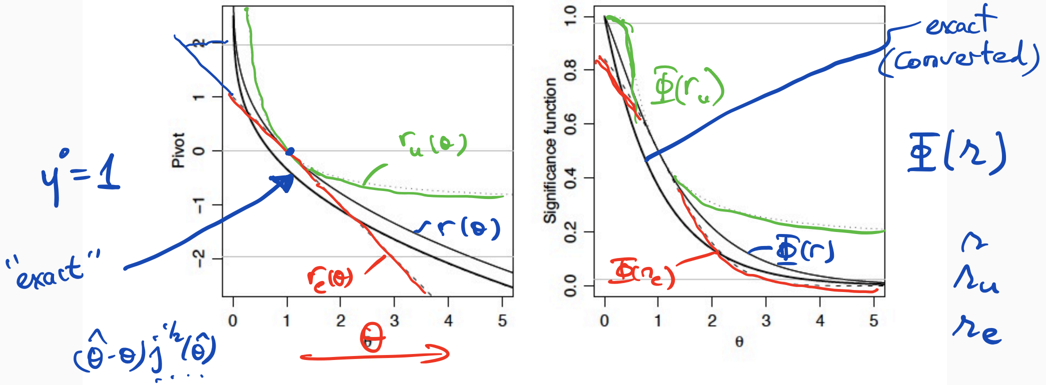
- $r_e(\theta) = (\hat{\theta} - \theta)j^{1/2}(\hat{\theta}) = \sqrt{n}(1 - \bar{y}\theta)$ $1 - y\theta$
- $r(\theta) = \sqrt{(2n)}\{\theta\bar{y} - 1 - \log(\theta\bar{y})\}^{1/2}$

$\sqrt{2}\,(\theta y - 1 - \log\theta y)^{1/2}$

expand $\log(\theta\bar{y})$ around 1 to get asymptotic equivalence to r_e, r_u

… Example: Exponential

14 *CHAPTER 2. UNCERTAINTY AND APPROXIMATION*



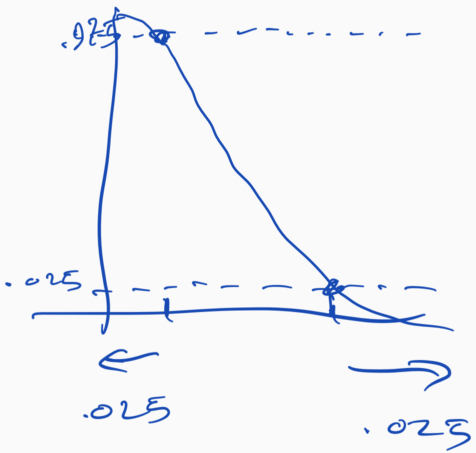


Figure 2.2: Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_j$ (dot-dash). The modified likelihood root is indistinguishable from the exact pivot. The horizontal lines are at $0, \pm 1.96$. Right: corresponding significance functions, with horizontal lines at 0.025 and 0.975.

Aside

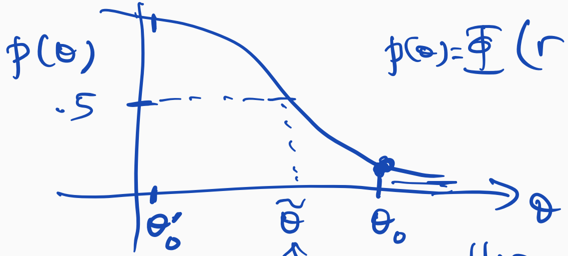
- for inference re θ, given y, plot $p(\theta)$ vs θ

- for p-value for $H_0 : \theta = \theta_0$, compute $p(\theta_0)$

- for checking whether, e.g. $\Phi\{r_e(\theta)\}$ is a good approximation,
  - compare $p(\theta) = \Phi\{r_e(\theta)\}$ to $p_{exact}(\theta)$, as a function of θ, fixed y
  - or compare $p(\theta_0)$ to $p_{exact}(\theta_0)$ as a function of y

- if $p_{exact}(\theta)$ not available, simulate

- if θ is a vector, choose one component at a time



$p(\theta_0) = \Phi(r(\theta_0))$

estimate of θ

$H: \theta = \theta_0$?

often mle

$\leftarrow$ we fix θ, see if under sampling $y_1, \dots, y_n$ this r.v. has 'good' properties

1 • $\underline{U}(\theta)$

$$\frac{1}{\sqrt{n}} \underline{U}(\underline{\theta}) \xrightarrow{d} N_p(0, i_1(\underline{\theta})_{p\times p})$$

$$\begin{pmatrix} u_1(\theta) \\ \vdots \\ u_p(\theta) \end{pmatrix} = \begin{pmatrix} \partial \ell / \partial \theta_1 \\ \vdots \\ \partial \ell / \partial \theta_p \end{pmatrix}$$

$$\underline{U}(\theta) \dot\sim N_p(\underline{0}, j_n(\hat\theta)) \dot\sim N(0, i_n(\hat\theta)) \dot\sim N(0, i_n(\theta))$$

$$i(\theta) = E\left(-\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}\right)_{p\times p}$$

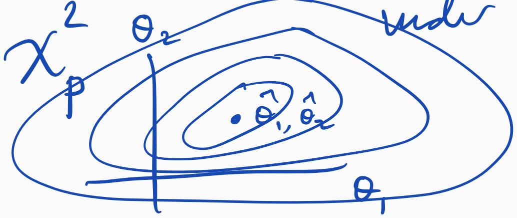
matrix

2 • $\hat{\theta}$

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N_p(0, i_1^{-1}(\theta))$$

$$\hat\theta \dot\sim N_p(\theta, i_n^{-1}(\theta)) \dot\sim N_p(\theta, i_n^{-1}(\hat\theta)) \dot\sim N_p(\theta, j^{-1}(\hat\theta))$$

3 • $\underline{2\{\ell(\hat{\theta}) - \ell(\theta)\}}$ $\xrightarrow{d} \chi^2_p$ under $y_1, \dots, y_n$ iid $f(y; \underline{\theta})$



Parameter of interest and nuisance parameter

- $\theta = (\psi, \lambda) =$ $(\psi_1, \dots, \psi_q, \lambda_1, \dots, \lambda_{p-q})$ $\vdots$ $\psi \in \mathbb{R}$ usually
  $q = 1$

Parameter of interest and nuisance parameter

- $\theta = (\psi, \lambda) =$

$$\ell(\theta; y) = \ell(\psi, \lambda; y)$$

- $U(\theta) = \begin{bmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial \ell}{\partial \psi}(\psi, \lambda) \\ \frac{\partial \ell}{\partial \lambda}(\psi, \lambda) \end{bmatrix} \dot\sim N_p\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, i(\theta)\right)$

$$i(\theta) = \begin{bmatrix} i_{\psi\psi}(\theta) & i_{\psi\lambda}(\theta) \\ i_{\lambda\psi}(\theta) & i_{\lambda\lambda}(\theta) \end{bmatrix}$$

$$i_{\psi\psi} = E\left(-\frac{\partial^2 \ell}{\partial \psi^2}\right) \qquad i_{\psi\lambda} = E\left(-\frac{\partial^2 \ell}{\partial \psi \partial \lambda}\right)$$

Parameter of interest and nuisance parameter

- $\theta = (\psi, \lambda) =$
- $U(\theta) =$
- $i(\theta) =$

$$j(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\lambda\psi}(\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix}$$

$$j_{\psi\psi} = -\frac{\partial^2 \ell(\psi,\lambda)}{\partial\psi^2}$$

Parameter of interest and nuisance parameter

$$\hat{\theta} \,\dot{\sim}\, N(\theta, j^{-1}(\hat{\theta}))$$

$$\hat{\psi} \,\dot{\sim}\, N(\psi, \{j^{-1}(\hat{\theta})\}_{\psi\psi})$$

- $\theta = (\psi, \lambda) =$
- $U(\theta) =$
- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}$ $\qquad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}$ $\qquad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}$

$$j^{-1}(\theta) \cdot j(\theta) = I$$

Parameter of interest and nuisance parameter

- $\theta = (\psi, \lambda) =$
- $U(\theta) =$
- $i(\theta) =$ $\quad$ $j(\theta) =$
- $i^{-1}(\theta) =$ $\quad$ $j^{-1}(\theta) =$
- $i^{\psi\psi}(\theta) =$

Parameter of interest and nuisance parameter

- $\theta = (\psi, \lambda) =$
- $U(\theta) =$
- $i(\theta) =$ $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1}$ $j(\theta) =$
- $i^{-1}(\theta) =$ $= \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ $\to \frac{1}{a-bd^{-1}c}$ $j^{-1}(\theta) =$
- $i^{\psi\psi}(\theta) =$ $\left(i_{\psi\psi} - i_{\psi\lambda}\, i_{\lambda\lambda}^{-1}\, i_{\lambda\psi}\right)^{-1}$? ntbc ? $j^{\psi\psi}$ same
- $\ell_{\mathrm{P}}(\psi) =$ $\ell(\psi, \hat\lambda_\psi)$ $\cancel{j_{\mathrm{P}}(\psi) =}$ $\ell_P'(\psi) = 0$ $\frac{d}{d\psi}\ell(\psi, \hat\lambda_\psi) = 0$

Profile log-likelihood : $\hat\lambda_\psi$ maximizer $\ell(\psi, \lambda)$ wrt λ $\frac{\partial}{\partial\psi}\ell(\psi,\lambda) + \frac{\partial}{\partial\lambda}(\psi,\lambda)\frac{d\hat\lambda_\psi}{d\psi}$

$\hat\psi \;\dot\sim\; N(\psi, [i^{-1}(\theta)]_{\psi\psi})$

$\dot\sim\; N(\psi, i^{\psi\psi}(\theta))$

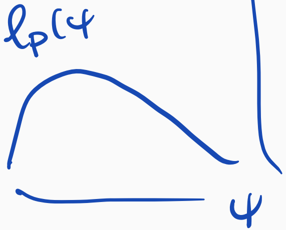
Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \ldots, \psi_q, \lambda_1, \ldots, \lambda_{d-q})$

$$\ell_p'(\psi) = \frac{\partial \ell}{\partial \psi}(\psi, \hat\lambda_\psi) + \underbrace{\frac{\partial \ell}{\partial \hat\lambda_\psi}(\psi, \hat\lambda_\psi)}_{=0}\frac{\partial \hat\lambda_\psi}{\partial \psi}$$

$\ell_p'(\psi) = 0$ defines an est. $\hat\psi$ ⇐ it is indeed mle

$$\hat\theta : \begin{bmatrix} \left.\frac{\partial \ell(\psi,\lambda)}{\partial \psi}\right|_{\hat\theta} \\ \left.\frac{\partial \ell(\psi,\lambda)}{\partial \lambda}\right|_{\hat\theta} \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$



profile curve $\lambda = \hat\lambda_\psi$

Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \dots, \psi_q, \lambda_1, \dots, \lambda_{d-q})$

- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}, \qquad U_\lambda(\psi, \hat{\lambda}_\psi) = 0$

- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$

- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$

- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i^{-1}_{\lambda\lambda}(\theta) i_{\lambda\psi}(\theta)\}^{-1},$

Nuisance parameters

Limit dist:

- $\theta = (\psi, \lambda) = (\psi_1, \dots, \psi_q, \lambda_1, \dots, \lambda_{d-q})$
- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}$, $\quad U_\lambda(\psi, \hat\lambda_\psi) = 0$
- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$
- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i^{-1}_{\lambda\lambda}(\theta) i_{\lambda\psi}(\theta)\}^{-1}$,
- $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat\lambda_\psi)$, $\quad j_{\mathrm{P}}(\psi) = -\ell''_{\mathrm{P}}(\psi)$

$$\frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N_p(0, i(\theta))$$

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N_p(0, i^{-1}(\theta))$$

$$2\{\ell(\hat\theta) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

$$\begin{pmatrix} \hat\psi - \psi \\ \hat\lambda - \lambda \end{pmatrix} \mathrel{\dot\sim} N_p\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} i^{\psi\psi}(\theta) & i^{\psi\lambda}(\theta) \\ i^{\lambda\psi}(\theta), & i^{\lambda\lambda}(\theta) \end{pmatrix}\right]$$

Approximations from limiting distributions, nuisance parameters

$(\hat\psi - \psi) \dot\sim N_q(0, j^{\psi\psi}(\hat\theta))$ $\quad \hat\theta = (\hat\psi, \hat\lambda)$: $\quad (\psi, \hat\lambda_\psi) \doteq \hat\theta_\psi$

$\left.\frac{\partial \ell(\psi,\lambda)}{\partial \psi}\right|_{\lambda = \hat\lambda_\psi}$

$$w_u(\psi) = U_\psi(\psi, \hat\lambda_\psi)^T \{i^{\psi\psi}(\psi, \hat\lambda_\psi)\} U_\psi(\psi, \hat\lambda_\psi) \quad \dot\sim \quad \chi^2_q$$

$$w_e(\psi) = (\hat\psi - \psi)\{i^{\psi\psi}(\hat\psi, \hat\lambda)\}^{-1}(\hat\psi - \psi) \quad \dot\sim \quad \chi^2_q$$

$$w(\psi) = 2\{\ell(\hat\psi, \hat\lambda) - \ell(\psi, \hat\lambda_\psi)\} = 2\{\ell_P(\hat\psi) - \ell_P(\psi)\} \quad \dot\sim \quad \chi^2_q;$$

$q \geq 1$

$q = 1 \quad w_e \sim \chi^2_1 \qquad (\hat\psi - \psi)^2 \{i^{\psi\psi}(\hat\psi, \hat\lambda)\}^{-1}$

$\pm\sqrt{\ } \dot\sim N(0,1)$

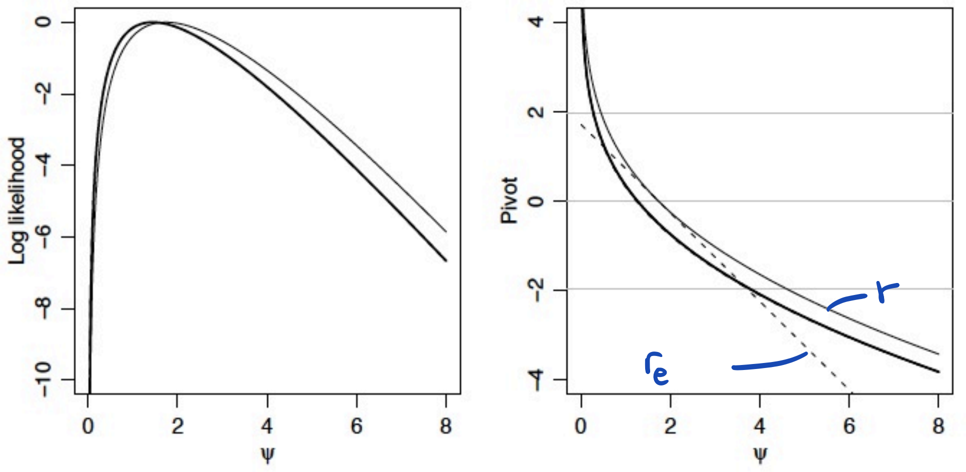
Approximate Pivots, $q = 1$

$$\pm\sqrt{w_u} = r_u(\psi) = \ell_P'(\psi) j_P(\hat\psi)^{-1/2} \dot\sim N(0,1),$$

$$\pm\sqrt{w_e} = r_e(\psi) = (\hat\psi - \psi) j_P(\hat\psi)^{1/2} \dot\sim N(0,1),$$

$$\pm\sqrt{w} = r(\psi) = \text{sign}(\hat\psi - \psi)[2\{\ell_P(\hat\psi) - \ell_P(\psi)\}]^{1/2} \dot\sim N(0,1)$$

$\ell_P'(\psi) = \frac{\partial \ell(\psi, \hat\lambda_\psi)}{\partial \psi} + \cancel{\frac{\partial \ell(\psi, \hat\lambda_\psi)}{\partial \lambda}} \cdot \partial \to 0$



$r(\psi)$

$r_e(\psi)$

both using

$\ell_p(\psi)$

$(\hat{\psi}-\psi)\{-\ell_p''(\hat{\psi})\}^{+1/2}$

$\parallel$

r_e

Figure 2.3: Inference for shape parameter ψ of gamma sample of size $n = 5$. Left: profile log likelihood ℓ_p (solid) and the log likelihood from the conditional density of u given v (heavy). Right: likelihood root $r(\psi)$ (solid), Wald pivot $t(\psi)$ (dashes), modified likelihood root $r^*(\psi)$ (heavy), and exact pivot overlying $r^*(\psi)$. The horizontal lines are at $0, \pm 1.96$.

$$r_u(\theta) = u(\theta)\, j^{-1/2}(\hat\theta) \quad \dot\sim\ N(0,1)$$

$$r_e(\theta) = (\hat\theta - \theta)\, j^{1/2}(\hat\theta) \quad \dot\sim\ N(0,1)$$

$$r(\theta) = \pm\sqrt{2\{\ell(\hat\theta) - \ell(\theta)\}} \quad \dot\sim\ N(0,1)$$

$$r_u(\psi) = \ell_p'(\psi)\, j_p^{-1/2}(\hat\psi) \qquad j_p(\psi) = -\ell_p''(\psi)$$

$$\ell_p(\psi) = \ell(\psi, \hat\lambda_\psi)$$

$$r_e(\psi) = (\hat\psi - \psi)\, j_p^{1/2}(\hat\psi)$$

$$r(\psi) = \pm\sqrt{2\{\ell_p(\hat\psi) - \ell_p(\psi)\}}$$

all $\dot\sim\ N(0,1)$ (limit theorem)

$$\begin{pmatrix}\hat\psi \\ \hat\lambda\end{pmatrix} \dot\sim\ N\left(\begin{pmatrix}\psi \\ \lambda\end{pmatrix}, i^{-1}(\theta)\right)$$

$$\hat\psi \ \dot\sim\ N(\psi, i^{\psi\psi}(\hat\theta))$$

pivotal q.: $$(\hat\psi - \psi)\{j^{\psi\psi}(\hat\psi, \hat\lambda_\psi)\}^{-1} \ \dot\sim\ N(0,1)$$

… Laplace approximation

marginal posterior:

$$(\hat{\psi} - \psi)\, j_p^{1/2}(\hat{\psi}) \sim N(0,1)$$

Profile likelihood: examples

- regression

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \psi = \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

$\psi = \beta_j$

$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$\hat{\beta} = (X^TX)^{-1}X^Ty$ $\quad\quad \ell(\beta, \sigma^2)$

$\ell(\sigma^2, \beta) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$

$\frac{\partial \ell}{\partial \beta} = 0 \implies -\frac{1}{2\sigma^2} \cdot 2 \cdot (y - X\beta)X^T = 0$ $\quad$ nth

$\hat{\beta} = (X^TX)^{-1}X^Ty = \hat{\beta}_{\sigma^2}$

$\frac{\partial \ell}{} = 0$

$$\frac{\partial}{\partial \sigma^2}$$

$$\ell(\sigma^2, \hat\beta_{\sigma^2}) = \ell(\sigma^2, \hat\beta)$$

$$\ell_p(\sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X\hat\beta)^T(y - X\hat\beta)$$

$$\ell_p'(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\hat\beta)^T(y - X\hat\beta)$$

$$\Big|_{\hat\sigma^2} = 0 \qquad \frac{-\cancel{4}n\sigma^{\cancel{4}2}}{2\sigma^4} + \frac{1}{2\sigma^4}(y - X\hat\beta)^T(y - X\hat\beta)$$

$$\hat\sigma^2 = \frac{(y - X\hat\beta)^T(y - X\hat\beta)}{n}$$

$$E\hat\sigma^2 = \frac{(n-p)\sigma^2}{n} \qquad \text{not} \quad \sigma^2$$

$$E\hat\sigma^2 < \sigma^2 \qquad \tilde\sigma^2 = \frac{(y - X\hat\beta)^T(y - X\hat\beta)}{n-p}$$

- yes, we do have issues w profile lik when p is large relative to n
- limit theory gives poor approximations when n is fixed & p is large

Profile likelihood: examples

- regression

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \psi = \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

- Neyman-Scott

$\hat{\mu}_i = \bar{y}_{i.}$ $\quad m \to \infty$

$$y_{ij} \sim N(\mu_i, \sigma^2), \; j = 1, \ldots, k, \; i = 1, \ldots, m$$

$$\boxed{\dim(X) = \dim(\mu) \to \infty \text{ if } m \to \infty}$$

$$\hat{\sigma}^2 = \frac{1}{mk}\sum_{i=1}^{m}(y_{ij} - \bar{y}_{i.})^2$$

too small

$\psi = \sigma^2 \quad \lambda = (\mu_1, \mu_2, \ldots, \mu_m)$

$\ell_P(\sigma^2)$ too concentrated

$E\hat{\sigma}^2 = \frac{1}{mk}\sigma^2(k-1)m$

$\hat{\sigma}^2 \to \sigma^2$ if $k \to \infty$ but not

Profile likelihood: examples

- regression

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \psi = \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

- Neyman-Scott

$$y_{ij} \sim N(\mu_i, \sigma^2), j = 1, \dots, k; i = 1, \dots, m$$

$$\hat{\sigma}^2 = \frac{1}{mk}\sum_{i=1}^{m}(y_{ij} - \bar{y}_{i.})^2 \xrightarrow{p} \frac{k-1}{k}\sigma^2 \neq \sigma^2$$

if $m \to \infty$

- 2×2 tables

$$y_{i1} \sim Bern(p_{i1}), y_{i2} \sim Bern(p_{i2}), i = 1, \dots, n, \quad \log\{\frac{p_{i1}/(1-p_{i1})}{p_{i2}(1-p_{i2})}\} = \psi + \lambda_i$$

nuis. par.

$$\hat{\psi} \xrightarrow{p} \psi/2$$

$n \to \infty$

1 cases $y_{i1} = 1$ if case ~~is ill~~, else 0

2 controls $y_{i2} = 1$ if ~~control ill~~

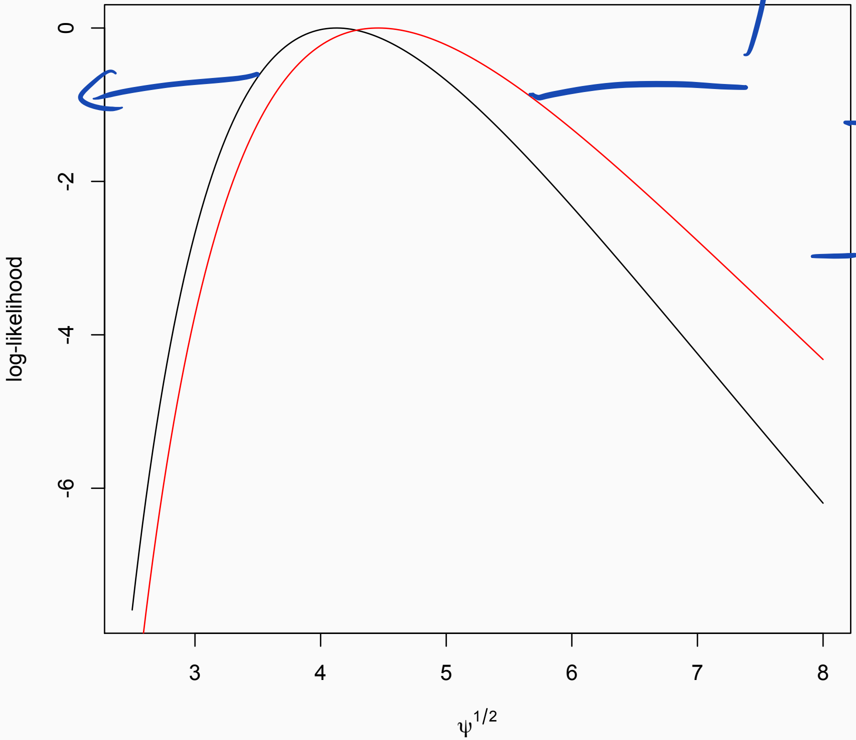
Smokers

$l_a(\sigma^2)$

This is a plot of $-n\log\sigma - (y - X\hat{\beta})^T(y - X\hat{\beta})/2\sigma^2$ (black), and $-(n-p)\log\sigma - (y - X\hat{\beta})^T(y - X\hat{\beta})/2\sigma^2$ against σ (red) for given data

profile

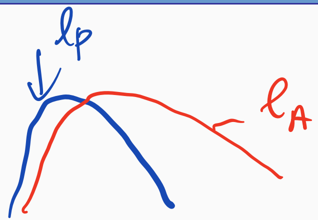
$l_p(\sigma^2)$



$-(n-p)\log\sigma$

$-\frac{1}{2\sigma^2}(y-X\hat{\beta})^T(\cdot\cdot)$

- $\ell_c(\psi) \doteq \ell_p(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|$ — $i_{\psi\lambda}(\theta) = 0$



- $\ell_m(\psi) \doteq \ell_p(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|$
- $\ell_c(\psi) \doteq \ell_p(\psi) + \frac{1}{2}\log|j_{\eta\eta}(\psi, \hat\eta_\psi)|$ — $\exp\{\psi^{\mathrm{T}} s + \eta^{\mathrm{T}} t - c(\psi, \eta)\}$
- adjusted profile log-likelihood

$$\boxed{\underline{\ell_A(\psi)} = \underline{\ell_p(\psi)} + \underline{A(\psi)}}$$

$A(\psi)$ assumed to be $O_p(1)$

- generic form is $A_{FR}(\psi) = +\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| - \log\left|\frac{d(\lambda)}{d\hat\lambda_\psi}\right|$ — Fraser 03
- closely related $A_{BN}(\psi) = -\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| + \log\left|\frac{d\hat\lambda}{d\hat\lambda_\psi}\right|$ — SM §12.4.1

BN 83

Cox & Reid adjusted likelihood
1987 showed that in general
it's a good 1st fix to

$$\ell_{CR}(\psi) = \ell_a(\psi) = \underline{\underline{\ell_p(\psi)}} - \frac{1}{2}\log\left|\underline{\underline{j}}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\right|$$

we need to work in the parametrization (ψ, λ) in which $\psi \perp \lambda$ (orthog)

meaning $i_{\psi\lambda}(\theta) \equiv 0$

exp'd F. info.

$$\left[\begin{array}{c|c} i_{\psi\psi} & 0 \\ \hline 0 & i_{\lambda\lambda} \end{array}\right]$$