Higher order asymptotics in practice

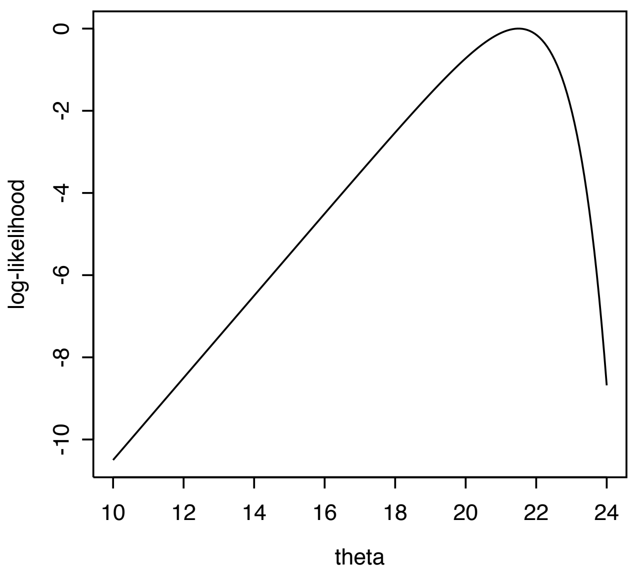
Nancy Reid, University of Toronto

www.utstat.utoronto.ca/reid/research

Case Western University, Oct 1 2004

Likelihood inference in parametric models

Example: $f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}$

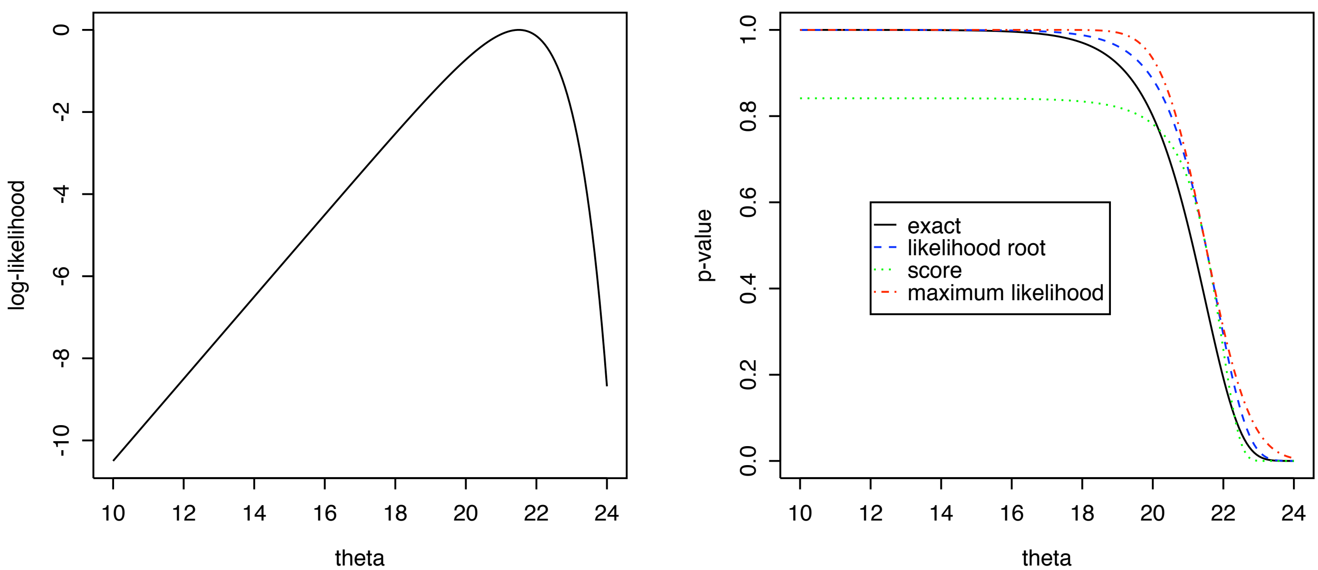


log-likelihood $\ell(\theta) = \log f(y;\theta)$

$y = y^0 = 21.5$

Likelihood inference in parametric models

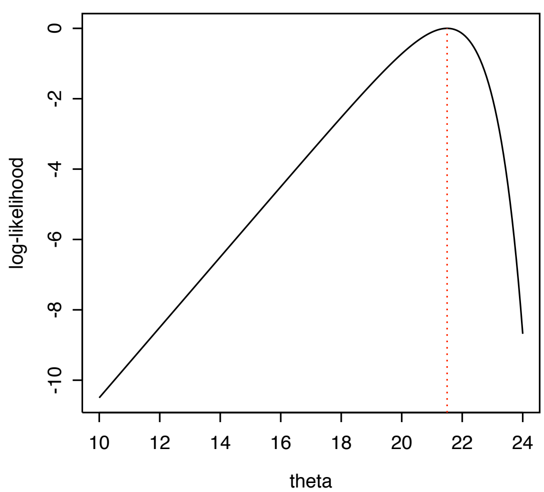
Example: $f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}$



p-value function $p(\theta) = F(y^0;\theta)$

Likelihood inference in parametric models

Example: $f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}$

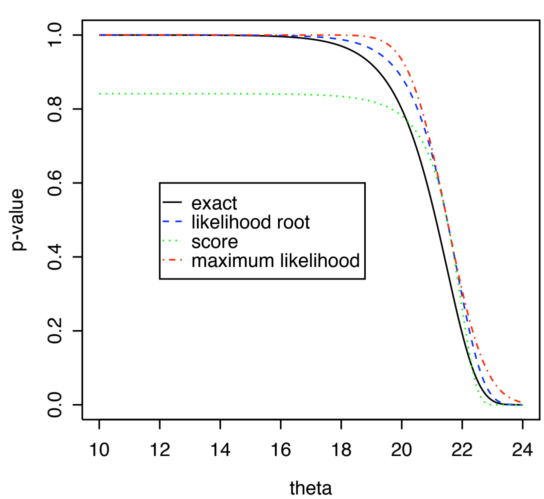


log-likelihood function $\ell(\theta) = \log f(y;\theta)$

score function $\ell'(\theta)$

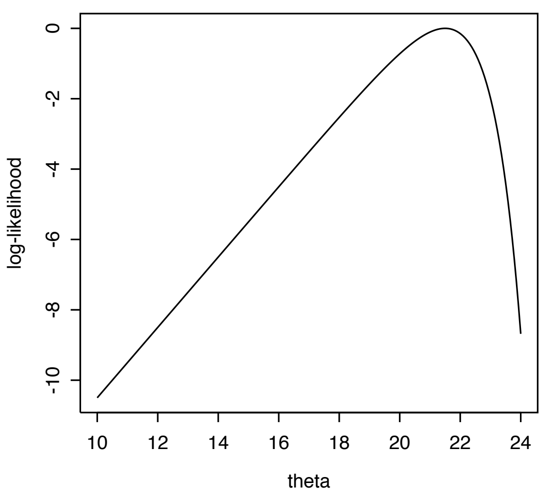
maximum likelihood estimate $\widehat{\theta}$

observed information $j(\widehat{\theta}\,) = -\ell''(\widehat{\theta}\,)$



Likelihood inference in parametric models

Example: $f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}$

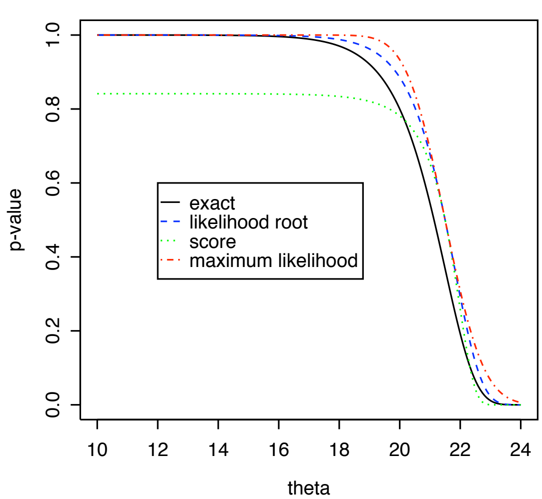


log-likelihood function $\ell(\theta) = \log f(y;\theta)$

score function $\ell'(\theta)$

maximum likelihood estimate $\hat{\theta}$

observed information $j(\hat{\theta}) = -\ell''(\hat{\theta})$



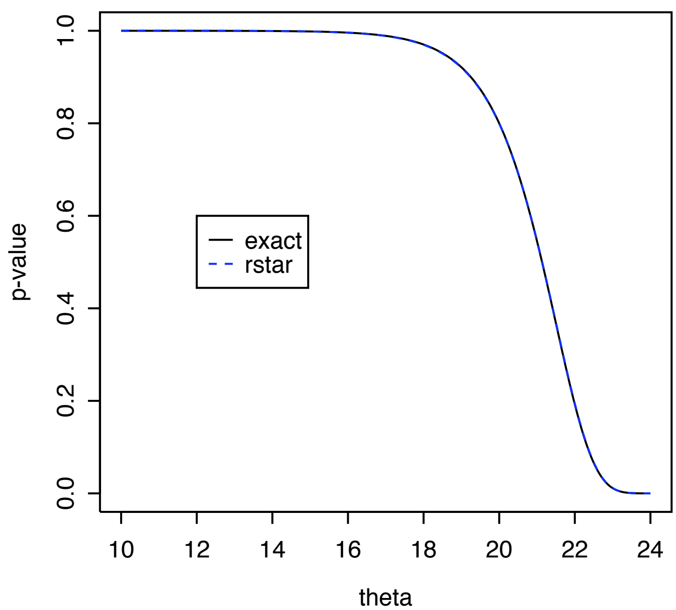
score $s(\theta) = \ell'(\theta)\{j(\hat{\theta})\}^{-1/2}$

max. lik. $q(\theta) = (\hat{\theta} - \theta)\{j(\hat{\theta})\}^{1/2}$

likelihood root $r(\theta) = \sqrt{[2\{\ell(\hat{\theta}) - \ell(\theta)\}]}$

$$\xrightarrow{d} N(0,1)$$

Third order approximation



$$r^* = r + \frac{1}{r}\log\frac{s}{r}$$

$$r^* \xrightarrow{d} N(0,1)$$

θ	14.59	16.20	16.90	17.82	18.53	20.00	20.47	20.83
exact	0.999	0.995	0.990	0.975	0.950	0.800	0.700	0.600
$\Phi(r^*)$	0.999	0.995	0.990	0.974	0.949	0.799	0.700	0.600

θ	21.41	21.69	21.98	22.60	22.81	23.03	23.17	23.43
exact	0.400	0.300	0.200	0.050	0.025	0.010	0.005	0.001
$\Phi(r^*)$	0.401	0.302	0.202	0.051	0.025	0.010	0.005	0.001

Likelihood statistics as pivots

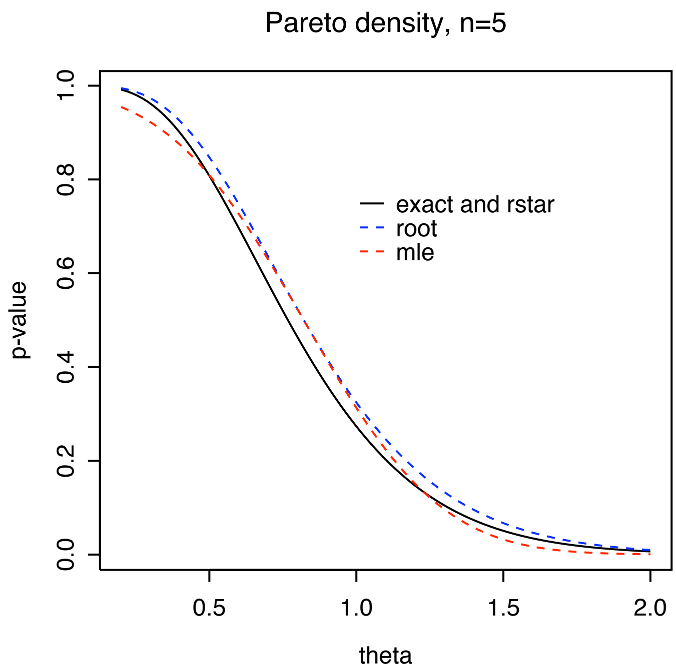
score statistic	$s(\theta) = \ell'(\theta)\{j(\widehat{\theta})\}^{-1/2}$
standardized m.l.e.	$q(\theta) = (\widehat{\theta} - \theta)\{j(\widehat{\theta})\}^{1/2}$
likelihood root	$r(\theta) = \surd[2\{\ell(\widehat{\theta}) - \ell(\theta)\}]$

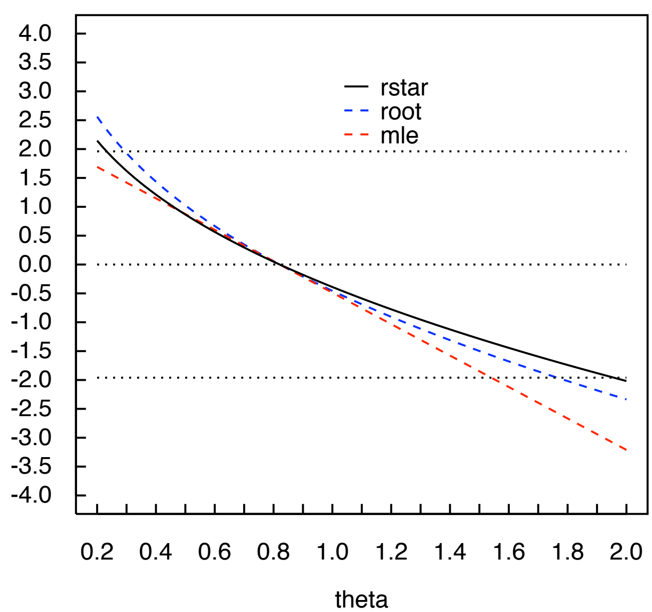
First order p-values

$$\begin{aligned} p(\theta) &\doteq \Phi\{r(\theta)\} \\ &\doteq \Phi\{q(\theta)\} \\ &\doteq \Phi\{s(\theta)\} \end{aligned}$$

Third order p-values

$$\begin{aligned} p(\theta) &\doteq \Phi\{r^*(\theta)\} \\ &\doteq \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{Q}\right) \\ r^*(\theta) &= r + \frac{1}{r}\log\frac{Q}{r} \\ Q &= \{\ell_{;\widehat{\theta}}(\widehat{\theta}) - \ell_{;\widehat{\theta}}(\theta)\}j(\widehat{\theta})^{-1/2} \end{aligned}$$


Pareto density, n=5
p-value
theta
exact and rstar
root
mle


theta
rstar
root
mle

Example: $Y \sim Po(\theta), \theta > b$

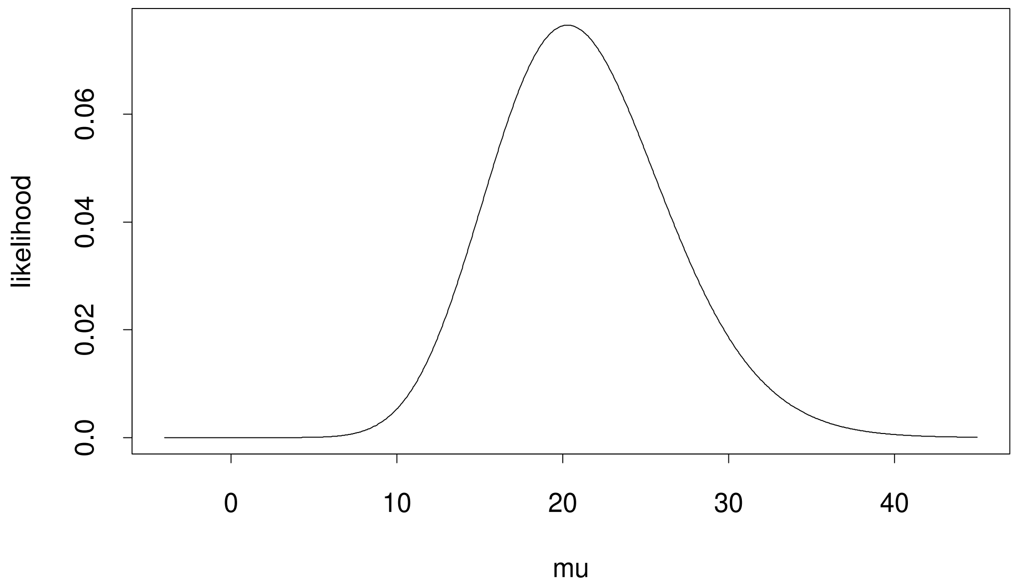
$$f(y;\theta) = \frac{1}{y!}\theta^y e^{-\theta} = \frac{1}{y!}(\mu + b)^y e^{-(\mu+b)}$$

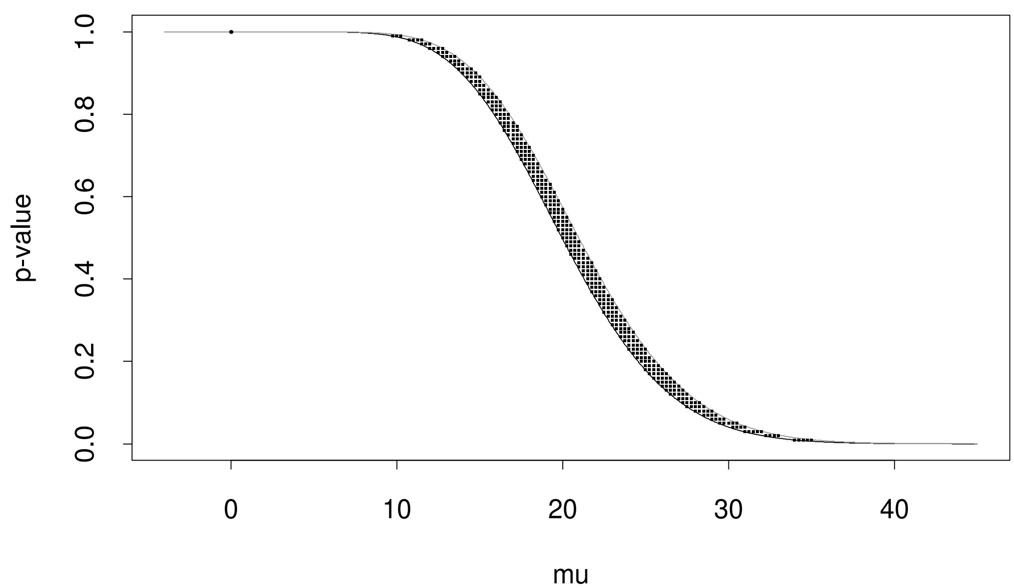
b is the background, $\mu > 0$ corresponds to a new signal

Fraser, Reid, Wong 2004

p-values:

upper	0.0005993
lower	0.0002170
mid	0.0004081
r^*	0.0003779
r	0.0004416
$\widehat{\theta}$	0.0062427


likelihood
0.0
0.02
0.04
0.06
0
10
20
30
40
mu


p-value
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
mu

Example: 2×2 table

Employment of men and women at the Space Telescope Science Institute, 1998–2002 (from *Science* magazine, Volume 299, page 993, 14 February 2003).

	Left	Stayed	Total
Men	1	18	19
Women	5	2	7
Total	6	20	26

Example: 2×2 table

Employment of men and women at the Space Telescope Science Institute, 1998–2002 (from *Science* magazine, Volume 299, page 993, 14 February 2003).

	Left	Stayed	Total
Men	1	18	19
Women	5	2	7
Total	6	20	26

Model: $Y_1 \sim Bin(19, p_1), Y_2 \sim Bin(7, p_2)$
$\theta = (\psi, \lambda)$
$\psi = \log\{p_1(1 - p_2)/p_2(1 - p_1)\}$
$H_0 : p_1 = p_2$, or $\psi = 0$

Example: 2×2 table

Employment of men and women at the Space Telescope Science Institute, 1998–2002 (from *Science* magazine, Volume 299, page 993, 14 February 2003).

	Left	Stayed	Total
Men	1	18	19
Women	5	2	7
Total	6	20	26

Model: $Y_1 \sim Bin(19, p_1), Y_2 \sim Bin(7, p_2)$

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

$\psi = \log\{p_1(1-p_2)/p_2(1-p_1)\}$

$H_0 : p_1 = p_2$, or $\psi = 0$

p-value:	$\Phi(q)$	0.002
	$\Phi(r)$	0.0003
	$\Phi(r^*)$	0.0005

```
> library("cond")
> astro.glm <- glm(astro~gender,family=binomial)
> astro.cond <- cond.glm(astro.glm,offset=gender)
> summary(astro.cond)
 FORMULA:  astro ~ gender
 FAMILY :  binomial
 OFFSET :  gender

 COEFFICIENTS
          Estimate   Std. Error
uncond.    -3.80666      1.32449
cond.      -3.55074      1.23712

 CONFIDENCE INTERVALS
 level = 95 %
                                                  lower  two-sided     upper
 MLE normal approximation                      -6.40261             -1.210710
 CMLE normal approximation                     -5.97544             -1.126030
 Directed deviance                             -7.06294             -1.525520
 Modified directed deviance                    -6.37130             -1.315430
 Modified directed deviance (cont. corr.)  -7.71524             -0.881396

 DIAGNOSTICS:
  INF    NP
0.126 0.234

 Approximation based on 20 points
> summary(astro.cond,test=0,digits=4)


 FORMULA:  astro ~ gender
 FAMILY :  binomial
 OFFSET :  gender

 COEFFICIENTS
          Estimate   Std. Error
```

```
uncond.      -3.807        1.324
cond.        -3.551        1.237

 HYPOTHESIS TESTING
 hypothesis : coef( gender ) = 0
                                            statistic   tail prob.
 MLE normal approximation                      -2.874    0.0020260
 CMLE normal approximation                     -2.870    0.0020510
 Directed deviance                             -3.447    0.0002838
 Modified directed deviance                    -3.298    0.0004872
 Modified directed deviance (cont. corr.)      -2.903    0.0018450

 DIAGNOSTICS:
  INF    NP
0.126 0.234

 Approximation based on 20 points

 > plot(astro.cond)
```

Nuisance parameters: $\theta = (\psi, \lambda)$

score statistic	$s(\psi) = \ell_p'(\psi)\{j_p(\hat{\psi})\}^{-1/2}$
standardized m.l.e.	$q(\psi) = (\hat{\psi} - \psi)\{j_p(\hat{\psi})\}^{1/2}$
likelihood root	$r(\psi) = \surd[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]$

First order p-values

$$
\begin{aligned}
p(\psi) &\doteq \Phi\{r(\psi)\} \\
&\doteq \Phi\{q(\psi)\} \\
&\doteq \Phi\{s(\psi)\}
\end{aligned}
$$

Third order p-values

$$
\begin{aligned}
p(\psi) &\doteq \Phi\{r^*(\psi)\} \\
&\doteq \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{Q}\right) \\
r^*(\psi) &= r + \frac{1}{r}\log\frac{Q}{r} \\
Q &=
\end{aligned}
$$

Steps

– Reduce from dimension n to p by conditioning, as in location family model

– Reduce from dimension p to 1 by marginalizing, as in exponential family model

– Replace exact ancillary by ancillary directions V

– Use saddlepoint approximation at the marginalization step

The algorithm

Have: – data $y \in R^n$, parameter $\theta \in R^p$, parameter of interest $\psi \in R$

– likelihood function at the observation $\ell(\theta; y^0)$

Goal: – scalar function of y that measures ψ

– e.g. a quantity $r(y, \psi)$ with a fixed distribution

– the distribution $\Pr\{r(Y, \psi) \geq r(y^0, \psi)\}$ to be known exactly or to a high order of approximation

Add: – $z_i = z_i(y_i, \theta)$, a pivotal statistic

Example: $y_i = x_i'\beta + \sigma e_i, \qquad z_i = e_i$

Example: $F(y_i; \theta) \sim U(0, 1)$

Compute: – the likelihood function and its derivative in the *sample space*

– $\ell(\theta; y^0)$

– $\varphi(\theta) = \ell_{;V}(\theta; y^0) = \sum \ell_{;y_i}(\theta; y^0) V_i$

$$V_i = -\left(\frac{\partial z_i}{\partial y_i}\right)^{-1} \left(\frac{\partial z_i}{\partial \theta}\right)\Bigg|_{\widehat{\theta}^0}$$

– $\varphi(\theta)$ is the canonical parameter for an approximating exponential family model

– this exponential family model has p-dimensional sample space

Compute: – the constrained m.l.e. $\widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$

– the likelihood root $r = \pm\sqrt{[2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\}]}$

– an improved version $r^* = r + \frac{1}{r} \log \frac{Q}{r}$

r^* is the scalar variable we seek, $r^* \dot\sim N(0, 1)$

exponential family

$f(y_1, \ldots, y_n; \theta) = \exp\{\theta' y_+ - nc(\theta) - \sum d(y_i)\}$

....

$$
\begin{aligned}
Q &= \{\nu(\widehat{\theta}) - \nu(\widehat{\theta}_\psi)\}/\widehat{\sigma}_\nu \\
&= \frac{|\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \quad \varphi_{\lambda'}(\widehat{\theta}_\psi)|}{|\varphi_{\theta'}(\widehat{\theta})|} \frac{|j_{\theta\theta}(\widehat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\widehat{\theta}_\psi)|^{1/2}}
\end{aligned}
$$

$$
\varphi(\theta) = \ell_{;V}(\theta; y^0) = \sum \ell_{;y_i}(\theta; y^0) V_i
$$

$j_{\theta\theta}(\widehat{\theta})$ from fitting full model

$j_{\lambda\lambda}(\widehat{\theta}_\psi)$ from fitting reduced model (ψ fixed)

derviatives can be computed numerically

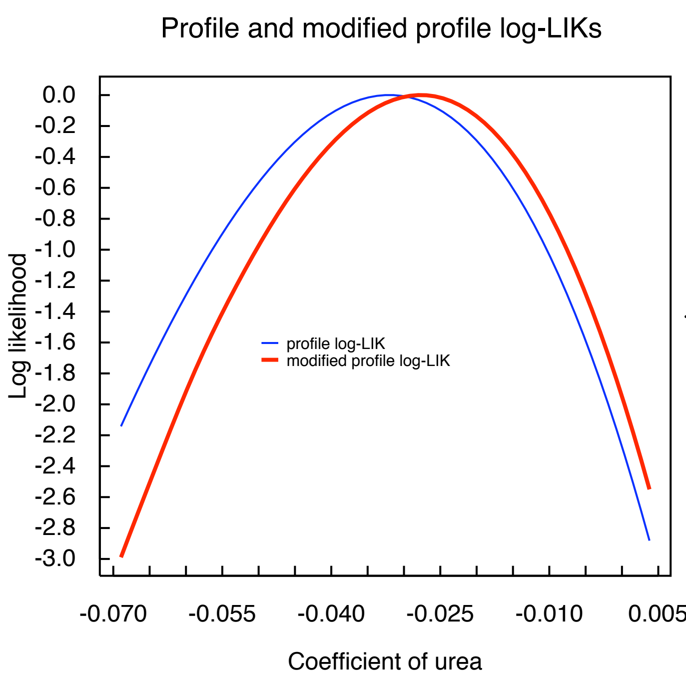
Example: Logistic regression (Brazzale, 2000)

– data from Davison and Hinkley (1997, Ch.7)

– binary response (presence of calcium oxalate crystals in urine)

– 77 observations, 6 covariates

– although there are 7 observations per parameter, higher order approximations make a difference



First order p-value : 0.01648

Third order p-value: 0.02664

Model

$\text{logit}(p_i) = \beta_0 + \beta_1\texttt{gravity} + \beta_2\texttt{ph} + \beta_3\texttt{osmo} + \beta_4\texttt{cond} + \beta_5\texttt{calc} + \beta_6\texttt{urea}$

$y_i \sim Bin(1, p_i)$

$\beta_0, \ldots, \beta_5$ can be exactly eliminated by conditioning on $\sum y_i$, $\sum y_i\texttt{gravity}_i$, etc.

method outlined above is equivalent to 3rd order

A Bayesian version

$\mathrm{Pr_m}(\Psi \geq \psi \mid y) \doteq \Phi(r^*)$

$$
\begin{aligned}
r^* &= r + \frac{1}{r}\log\frac{q_B}{r} \\
r &= \mathrm{sign}(\widehat{\psi} - \psi)[2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\}]^{1/2} \\
q_B &= \ell_\theta(\widehat{\theta}_\psi)\left\{\frac{|j_{\lambda\lambda}(\widehat{\theta}_\psi)|}{|j_{\theta\theta}(\widehat{\theta})|}\right\}^{1/2}\frac{\pi(\widehat{\theta})}{\pi(\widehat{\theta}_\psi)}.
\end{aligned}
$$

Example: simple logistic regression
$\mathrm{logit}(p_i) = \alpha + \beta x_i$

Comparison of 95% confidence intervals for β (Hosmer & Lemeshow, Ex. 1.1; y=chd, x=age, n=100)

	Lower endpoint	Upper endpoint
First order	0.06377	0.1581
Bayesian r^*	0.06680	0.1598
Frequentist r^*	0.06524	0.1596

Choice of prior?

– noninformative prior $\pi(\theta) \propto i_{\beta\beta}^{1/2}(\theta) g(\eta)$
– g arbitrary
– η orthogonal to β with respect to expected Fisher information
– $\eta = \eta(\alpha, \beta) = \Sigma n_i p_i(\alpha, \beta)$
– invariant to choice of $g(\cdot)$ in this setting

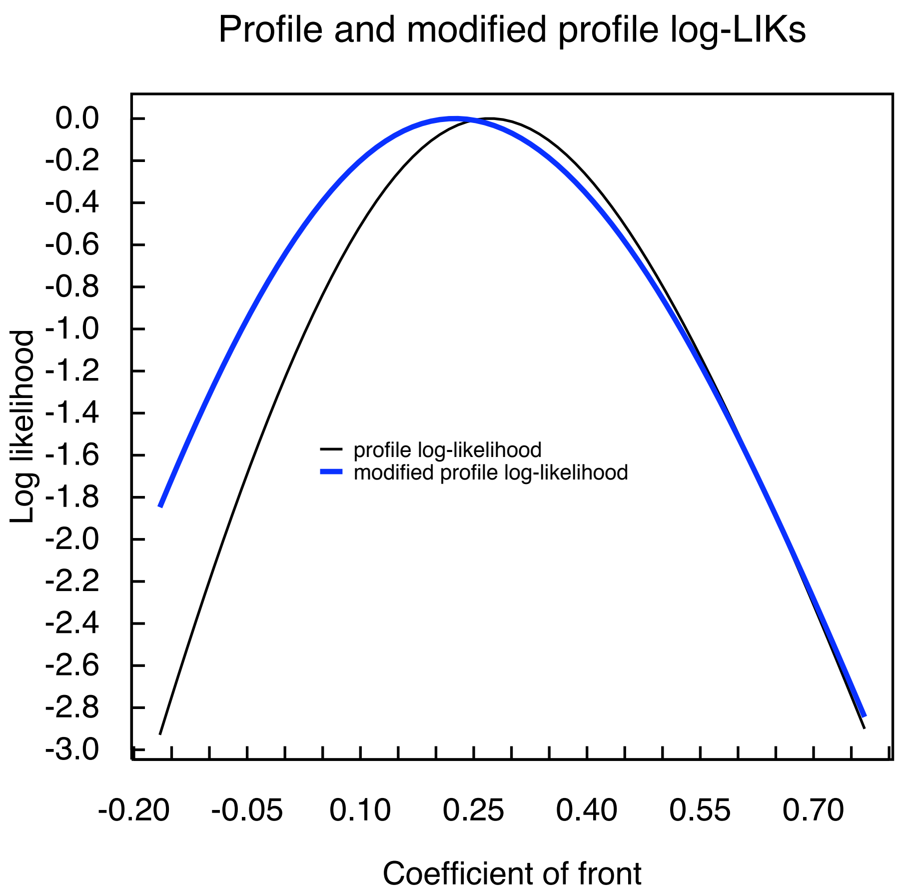
Example: linear regression, non-normal error

$y_i = x_i'\beta + \sigma e_i$, $e_i \sim f(\cdot)$

Data: House prices (Sen & Srivastava)
$y =$ selling price, 4 covariates

95% confidence interval for coefficient of `frontage`:

	First order		Third order	
Student (3)	-0.07	0.65	-0.16	0.69
Student (5)	-0.09	0.65	-0.15	0.70
Student (7)	-0.08	0.66	-0.14	0.70
Normal	-0.08	0.66	-0.13	0.71


Profile and modified profile log-LIKs
Log likelihood
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
-1.8
-2.0
-2.2
-2.4
-2.6
-2.8
-3.0
profile log-likelihood
modified profile log-likelihood
-0.20
-0.05
0.10
0.25
0.40
0.55
0.70
Coefficient of front

Example: Type-2 censored data (Wong & Wu, 2000, TCS)

Weibull model : $F(t; \lambda, \beta) = 1 - \exp\{(-t/\lambda)^{\beta}\}$

data ($y_{(1)} \leq \ldots \leq y_{(r)}$, $n - r$ units still on test at end of experiment

A. C. M. WONG AND J. WU

ence the lo-
d data point
ısion of θ is

(12)

Table 1. 90% Confidence Interval for Mean, Standard Deviation, and .1 Quantile

Method	$\psi(\theta) = \mu$	$\psi(\theta) = \sigma$	$\psi(\theta) = y_1 = \mu + \log\{-\log(1 - .1)\}$
AN	(−.13, .44)	(.66, 1.16)	(−2.70, −1.49)
LR	(−.12, .48)	(.70, 1.22)	(−2.62, −1.37)
Third-order	(−.11, .51)	(.72, 1.28)	(−2.71, −1.41)
Exact	(−.11, .51)	(.72, 1.28)	(−2.71, −1.41)

Other examples

– transformed regression $y^\lambda = x'\beta + \sigma e$

– generalized linear models: Poisson, gamma, binomial

– mean of a log-normal distribution $\psi = \mu + \sigma^2/2$

– nonlinear regression with normal error, non-constant variance (Brazzale)

– linear mixed models (Brazzale & Guolo)

Extensions not (yet) available

– complex models: hierarchical models, time dependencies, spatial dependencies

– robustness or model dependence

– large datasets, especially with $p >> n$

Some references

Brazzale, A.R. (2000) Practical small-sample parametric inference. PhD thesis, EPFL.

`www.stat.unipd.it/~LIKASY`

Brazzale, A.R., Davison, A.C. and Reid, N. (200x). *Case Studies in Higher Order Aysmptotics*.

Fraser, D.A.S., Reid, N. and Wong, A. (2004). Inference for bounded parameters. *Phys. Rev. D 69, 033002*.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* 31, 1695-1731.

Wong, A.C.M. and Wu, J. (2000) Practical small-sample asymptotics for distributions used in life-data analysis. *Technometrics* 42, 149–156.

`www.utstat.utoronto.ca/reid`

`www.utstat.utoronto.ca/dfraser`

As an example consider sampling from the bivariate normal distributio
as an unknown parameter (Cox and Hinkley, 1974, Ex. 2.30; Barndorff-N
log-likelihood function is

$$\ell(\theta) = -\frac{n}{2}\log(1-\theta^2) - \frac{1}{2(1-\theta^2)}\sum(y_{1i}^2 + y_{2i}^2) + \frac{\theta}{1-\theta^2}\sum$$

		left tail					right tail	
θ	Exact	0.001	0.005	0.010	0.025	0.050	0.050	0.025
-0.9	r	0.0022	0.009	0.018	0.042	0.079	0.033	0.016
	r^*	0.0010	0.005	0.010	0.025	0.050	0.050	0.025
-0.7	r	0.0019	0.009	0.018	0.041	0.078	0.045	0.024
	r^*	0.0009	0.005	0.010	0.025	0.049	0.053	0.027
-0.5	r	0.0022	0.009	0.017	0.041	0.077	0.058	0.031
	r^*	0.0010	0.005	0.010	0.025	0.050	0.055	0.027
-0.3	r	0.0022	0.009	0.019	0.042	0.076	0.065	0.035
	r^*	0.0011	0.005	0.010	0.026	0.051	0.053	0.026
0	r	0.0022	0.010	0.018	0.040	0.074	0.072	0.039
	r^*	0.0011	0.005	0.011	0.026	0.051	0.051	0.025