Last two weeks

- $y = (y_1, \ldots, y_n) \sim f(y; \theta), \quad \theta = (\psi, \lambda), \quad \psi \in \mathbb{R}$

- two steps in constructing a pivot for (non-Bayesian) inference about ψ:
  1. find a distribution on $\mathbb{R}^d$ that captures all the information about θ
  2. find a distribution on $\mathbb{R}$ that captures all the information about ψ
- in an exponential family model,
  $f(y; \theta) \propto \exp\{\psi s_1(y) + \lambda^T s_2(y) - c(\psi, \lambda)$,
  step 1. follows by marginalizing to the distribution of the sufficient statistic: $s = (s_1, s_1) \in \mathbb{R}^d$
- in a regression-scale model, for example
  $f(y; \theta) \propto \sigma^{-n} \prod f_0\{y_i - x_i^T \beta)/\sigma\}$,
  step 1. follows by conditioning on the residuals $(y_i - x_i^T \hat{\beta}/\hat{\sigma})$, which are ancillary for $\theta = (\beta, \sigma)$.

... last week

- $y = (y_1, \ldots, y_n) \sim f(y; \theta), \quad \theta = (\psi, \lambda), \quad \psi \in \mathbb{R}$

- two steps in constructing a pivot for (non-Bayesian) inference about ψ:
  1. find a distribution on $\mathbb{R}^d$ that captures all the information about θ
  2. find a distribution on $\mathbb{R}$ that captures all the information about ψ
- in an exponential family model,
  $f(y; \theta) \propto \exp\{\psi s_1(y) + \lambda^T s_2(y) - c(\psi, \lambda)$,
  step 2. follows by conditioning:
  $f(s_1 \mid s_2; \psi, \lambda) = f(s_1 \mid s_2; \psi)$
- in a regression-scale model, for example
  $f(y; \theta) \propto \sigma^{-n} \prod f_0\{y_i - x_i^T \beta)/\sigma\}$,
  step 2. follows by marginalizing: $f_m\{(\hat{\beta}_1 - \beta_1)/\hat{\sigma}\}$ has a distribution free of $(\beta_{(2)}, \sigma)$ Week 3 handout.

... last week

- in both cases, exponential family models and regression-scale models, the resulting pivot is

$$r^* = r + \frac{1}{r}\log(\frac{Q}{r}) \dot{\sim} N(0,1)$$

- only the form of Q changes (slightly)
- suggests that the result can be applied more generally
- use an approximate transformation model (ancillary) to reduce from n to d
- use an approximate exponential family model to reduce from d to 1
- these are both embedded in the tangent exponential model

- Bayesian posterior distribution functions have the same approximation, with $Q = q_B$ depending on the prior

This week

1. a few hoa examples
2. composite likelihood
3. estimating functions
4. generalized estimating equations (GEE)

Several 2×2 tables.

Institution	y_1	m_1	y_2	m_2	Institution	y_1	m_1	y_2	m_2
1	3	4	1	3	12	2	2	0	2
2	3	4	8	11	13	1	4	1	5
3	2	2	2	3	14	2	3	2	4
4	2	2	2	2	15	2	4	4	6
5	2	2	0	3	16	4	12	3	9
6	1	3	2	3	17	1	2	2	3
7	2	2	2	3	18	3	3	1	4
8	1	5	4	4	19	1	4	2	3
9	2	2	2	3	20	0	3	0	2
10	0	2	2	3	21	2	4	1	5
11	3	3	3	3					

Lipsitz et al. 1988: Biometrics; BDR p.64

... matched pairs

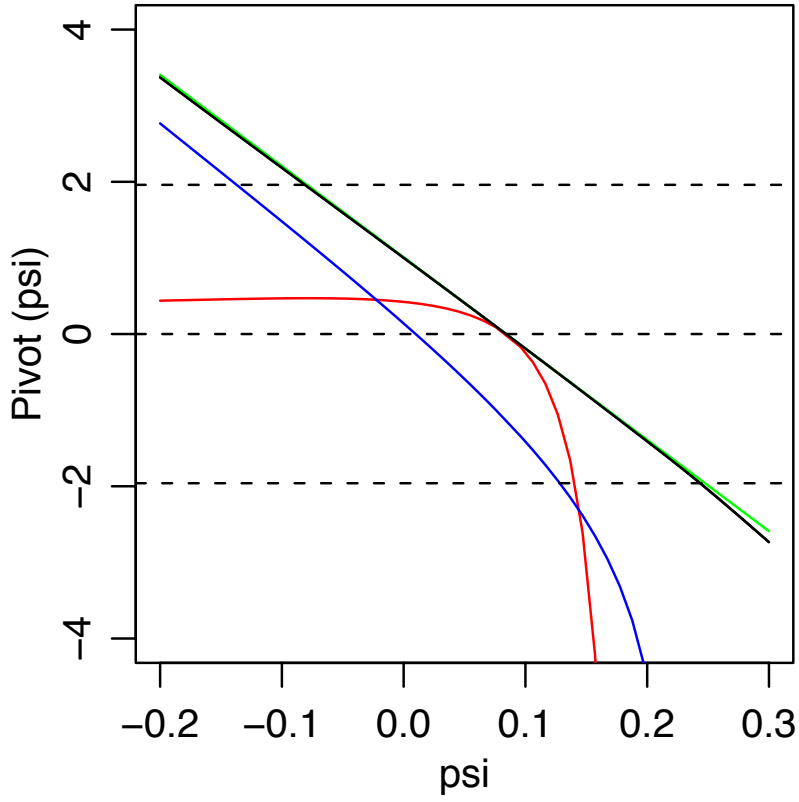
Model: $Y_{1i} \sim$ Binomial (m_{1i}, p_{1i}) $\quad Y_{2i} \sim$ Binomial (m_{2i}, p_{2i})

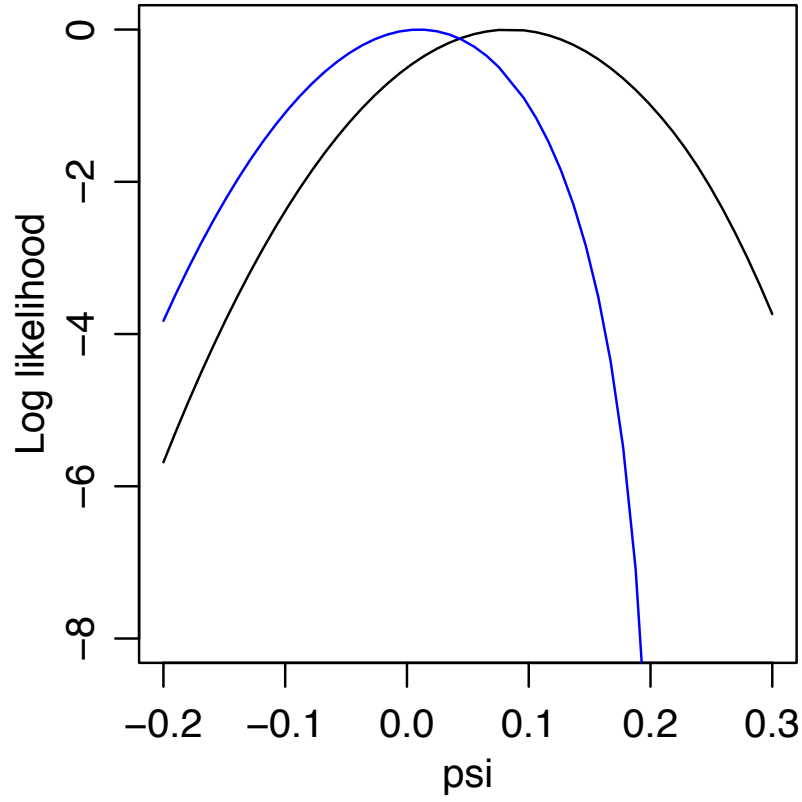
parameter of interest $\psi = p_{2i} - p_{1i}$
nuisance parameters $p_{1i}, i = 1, \ldots 21$

inference for ψ :

	lower	upper	point estimate	p-value for $\psi = 0$
$\Phi(r)$	-0.081	0.243	0.10	0.16
$\Phi(r^*)$	-0.137	0.126	0.01	0.45

$$\varphi(\theta) = \Big[\sum_{i=1}^{n} m_{2i} \log\{p_{1i}/(1-p_{1i})\},$$
$$m_{2i} \log\{p_{2i}/(1-p_{2i}) + m_{1i} \log\{p_{1i}/(1-p_{1i})\}, i = 1, \ldots, n\Big]$$





likelihood root, r^*, q

profile log likelihood,
modified profile

Example G: Cox & Snell, 1980

	cost	date	T1	T2	cap	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0
13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0

$n = 32, d = 8$

Linear regression, non-normal error

- Model $Y_i = \beta_0 + x_i^T \beta + \sigma \epsilon_i$
- $\epsilon \sim N(0,1)$ or $\epsilon \sim t_\nu$

	Normal		t_4, first order		t_4, third order	
	Est (SE)	z	Est (SE)	z	Est (SE)	z
Constant	−13.26 (3.140)	−4.22	−11.30 (3.67)	−3.01	−11.86 (3.70)	−3.21
`date`	0.212 (0.043)	4.91	0.191 (0.048)	3.97	0.196 (0.049)	4.02
`log(cap)`	0.723 (0.119)	6.09	0.648 (0.113)	5.71	0.682 (0.129)	5.31
`NE`	0.249 (0.074)	3.36	0.242 (0.077)	3.12	0.239 (0.080)	2.97
`CT`	0.140 (0.060)	2.32	0.144 (0.054)	2.68	0.143 (0.063)	2.26
`log(N)`	−0.088 (0.042)	−2.11	−0.060 (0.043)	−1.40	−0.072 (0.048)	−1.51
`PT`	−0.226 (0.114)	−1.99	−0.282 (0.101)	−2.80	−0.265 (0.110)	−2.42

ν	`log(N)`		`PT`	
	First order	Third order	First order	Third order
4	0.162	0.151	0.005	0.024
6	0.110	0.116	0.007	0.032
8	0.081	0.098	0.009	0.036
10	0.064	0.086	0.011	0.038
20	0.036	0.064	0.016	0.045
40	0.025	0.053	0.029	0.050
100	0.020	0.047	0.022	0.053
∞	0.035	0.045	0.046	0.057

```
library(marg)
# part of package 'hoa' on cran-r
data(nuclear)

# Fit normal-theory linear model and examine its contents:

nuc.norm <- lm( log(cost) ~ date + log(cap) + NE + CT + log(N) + PT,
+                  data = nuclear )
summary(nuc.norm)

# Fit linear model with t errors and 4 df and examine its contents:

nuc.t4 <- rsm( log(cost) ~ date + log(cap) + NE + CT + log(N) + PT,
+                 data = nuclear, family = student(4) )
summary(nuc.t4)
plot(nuc.t4)

# Conditional analysis for partial turnkey guarantee:

nuc.t4.pt <- cond( nuc.t4, offset = PT )
summary(nuc.t4.pt)
plot(nuc.t4.pt)

# For conditional analysis for other covariates, replace pt by
# log(N), ...
```

Type II censored data

40 units on test, 28 failures at (log) times

0.0507	0.0579	0.0784	0.0954	0.1376	0.2249	0.2362	0.2481
0.2501	0.2811	0.3027	0.3091	0.4296	0.5379	0.5621	0.5781
0.7811	0.8228	0.9455	0.9871	1.0060	1.0335	1.0377	1.0471
1.0876	1.2473	1.2776	1.3445				

Weibull model: $f(y;\mu,\sigma) = e^{(y-\mu)/\sigma} \exp\{-e^{(y-\mu)/\sigma}\}$

90% confidence intervals

	μ	σ
$\Phi(r)$	$(-0.116, 0.476)$	$(0.700, 1.217)$
$\Phi(r^*)$	$(-0.107, 0.510)$	$(0.743, 1.320)$
Exact (num. int.)	$(-0.11, 0.51)$	$(0.724, 1.277)$

Lawless 2003 Ch.5; Wong & Wu 2003

Vector parameter of interest

- use tangent exponential model

  or usual exponential family model

- construct a scalar parameter of interest

  direction in sample space

- apply higher order approximation

- Example $y \sim N(\mu, \Sigma)$; $H_0 : \Sigma^{-1}$ is tri-diagonal
- First-order Markov dependence in a graphical model

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5
First order	5.5	10.5	17.0	27.0	48.7	73.0	89.5	96.7	98.5	99.4
Second order	1.1	2.6	5.0	10.1	24.8	49.8	74.9	89.9	94.9	97.4

Composite likelihood

- Vector observation: $Y \sim f(y;\theta), \quad Y \in \mathcal{Y} \subset \mathbb{R}^m, \quad \theta \in \mathbb{R}^d$
- Set of events: $\{\mathcal{A}_k, k \in K\}$

- Composite Likelihood: (Lindsay, 1988)

$$CL(\theta; y) = \prod_{k \in K} L_k(\theta; y)^{w_k}$$

- $L_k(\theta; y) = f(\{y \in \mathcal{A}_k\}; \theta)$ likelihood for an event
- $\{w_k, k \in K\}$ a set of weights

Examples

- Composite Conditional Likelihood: (Besag, 1974)

$$\mathcal{L}_C(\theta; y) = \prod_{s \in \mathcal{S}} f_{s|s^c}(y_s \mid y_{s^c})^{w_s},$$

and variants by modifying events

- Composite Marginal Likelihood:

$$CML(\theta; y) = \prod_{s \in \mathcal{S}} f_s(y_s; \theta)^{w_s},$$

$f_s(y_s; \theta)$: marginal density of the subvector y_s induced by f

- Independence Likelihood:

- Pairwise Likelihood:

Derived quantities

- log composite likelihood: $c\ell(\theta; y) = \log CL(\theta; y)$
- score function: $U(\theta; y) = \nabla_\theta c\ell(\theta; y) = \sum_{s \in \mathcal{S}} w_s U_s(\theta; y)$
- $$U_s(\theta; y) = \nabla_\theta f_s(y_s; \theta)$$
- variability matrix: $J(\theta) = var_\theta\{U(\theta; Y)\}$
- sensitivity matrix: $H(\theta) = \mathrm{E}_\theta\{-\nabla_\theta U(\theta; Y)\}$
- Godambe information (or sandwich information):

$$G(\theta) = H(\theta) J(\theta)^{-1} H(\theta)$$

Inference

- Sample: $Y_1, \ldots, Y_n$, i.i.d., $CL(\theta; \underline{y}) = \prod_{i=1}^{n} CL(\theta; y_i)$
- $$\sqrt{n}(\hat{\theta}_{CL} - \theta) \dot{\sim} N\{0, G^{-1}(\theta)\} \qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$$

... inference

- $w(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \dot{\sim} \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$
- $\mu_1, \ldots, \mu_d$ eigenvalues of $J(\theta)H(\theta)^{-1}$

... inference

- $w(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \dot{\sim} \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$
- $\mu_1, \ldots, \mu_d$ eigenvalues of $J(\theta)H(\theta)^{-1}$

$$w(\theta) \doteq (\hat{\theta}_{CL} - \theta)\{nH(\theta)\}(\hat{\theta}_{CL} - \theta)$$

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

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

- constrained estimator: $\tilde{\theta}_\psi = \sup_{\theta=\theta(\psi)} c\ell(\theta; y)$
- $\sqrt{n}(\hat{\psi}_{CL} - \psi) \dot{\sim} N\{0, G^{\psi\psi}(\theta)\}$ $\qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- $w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_\psi)\} \dot{\sim} \sum_{a=1}^{d_0} \mu_a Z_a^2$
- $\mu_1, \ldots, \mu_{d_0}$ eigenvalues of $(H^{\psi\psi})^{-1} G^{\psi\psi}$

Kent, 1982

Model selection

- Akaike's information criterion Varin and Vidoni, 2005

$$AIC = -2\mathrm{c}\ell(\hat{\theta}_{CL}; y) - 2\,\dim(\theta)$$

- Bayesian information criterion Gao and Song, 2009

$$BIC = -2\mathrm{c}\ell(\hat{\theta}_{CL}; y) - \log n\,\dim(\theta)$$

- effective number of parameters

$$\dim(\theta) = \mathrm{tr}\{H(\theta)G^{-1}(\theta)\}$$

- these criteria used for model averaging
  Hjort and Claeskens, 2008
- or for selection of tuning parameters Gao and Song, 2009

Example: symmetric normal

- $Y_i \sim N(0, R)$, $var(Y_{ir}) = 1$, corr $(Y_{ir}, Y_{is}) = \rho$
- compound bivariate normal densities to form pairwise likelihood

$$c\ell(\rho; y_1, \ldots, y_n) = -\frac{nm(m-1)}{4}\log(1-\rho^2) - \frac{m-1+\rho}{2(1-\rho^2)}SS_w - \frac{(m-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_b}{m}$$

$$SS_w = \sum_{i=1}^{n}\sum_{s=1}^{m}(y_{is} - \bar{y}_{i.})^2, \quad SS_b = \sum_{i=1}^{n} y_{i.}^2$$

$$\ell(\rho; y_1, \ldots, y_n) = -\frac{n(m-1)}{2}\log(1-\rho) - \frac{n}{2}\log\{1+(m-1)\rho\} - \frac{1}{2(1-\rho)}SS_w - \frac{1}{2\{1+(m-1)\rho\}}\frac{SS_b}{m}$$

... symmetric normal

- a. $\text{var}(\hat{\rho}) = \dfrac{2}{nm(m-1)} \dfrac{\{1+(m-1)\rho\}^2(1-\rho)^2}{1+(m-1)\rho^2}$
- a. $\text{var}(\hat{\rho}_{CL}) = \dfrac{2}{nm(m-1)} \dfrac{(1-\rho)^2 c(m,\rho)}{(1+\rho^2)^2}$
- $c(m,\rho) =$
  $(1-\rho)^2(3\rho^2+1) + m\rho(-3\rho^3+8\rho^2-3\rho+2) + m^2\rho^2(1-\rho)^2$

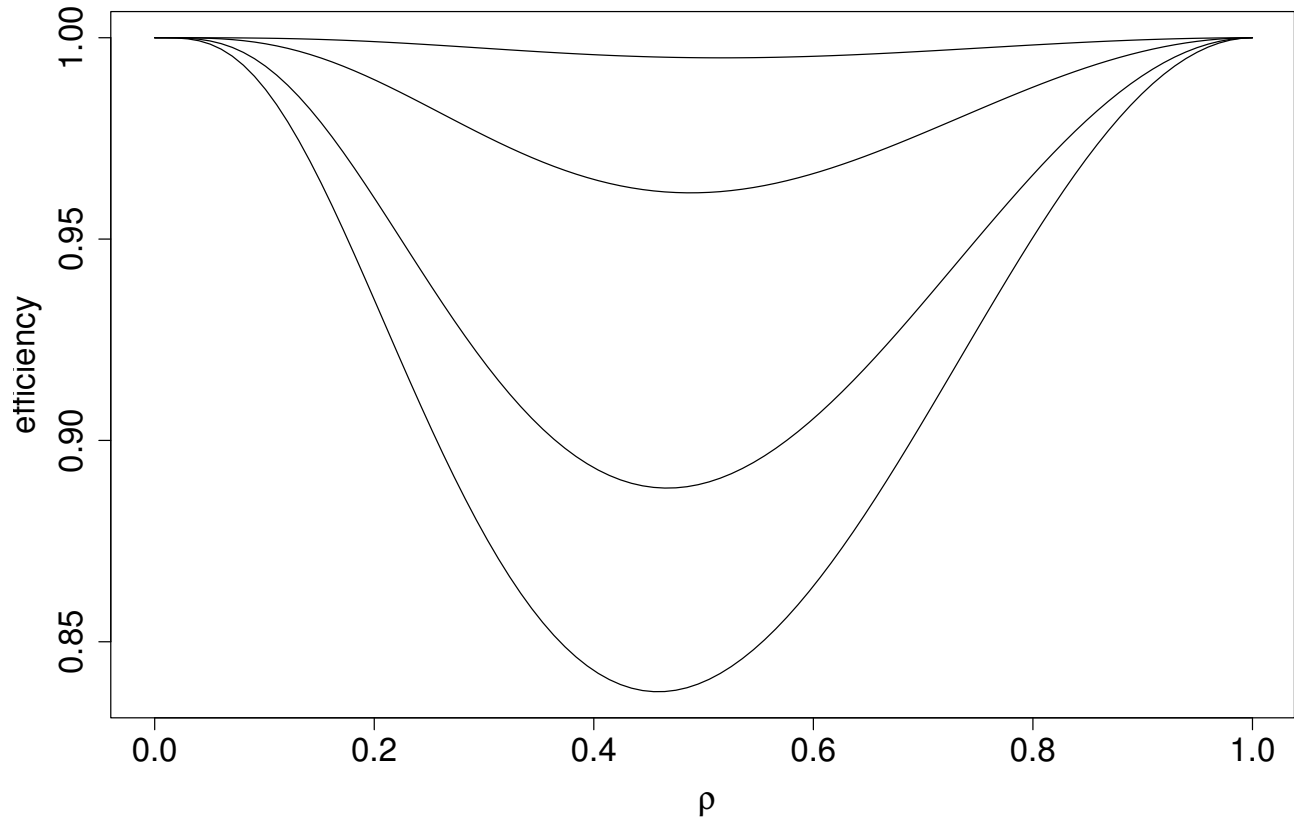
$$\text{a.var}(\hat{\rho}_{CL}) = \frac{2}{nm(m-1)} \frac{(1-\rho)^2}{(1+\rho^2)^2} c(m,\rho)$$

$$O(\tfrac{1}{n}) \qquad\qquad O(1)$$
$$n \longrightarrow \infty \qquad\qquad m \longrightarrow \infty$$

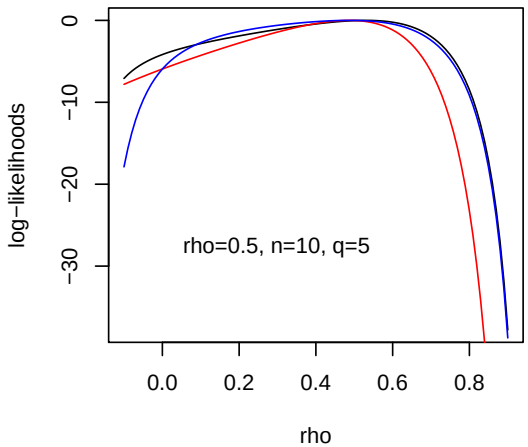
... symmetric normal

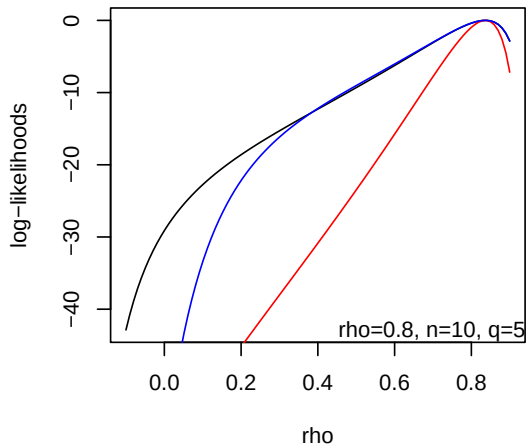
$$\frac{a.var(\hat{\rho}_{CL})}{a.var(\hat{\rho}\,)}, \quad m = 3, 5, 8, 10$$

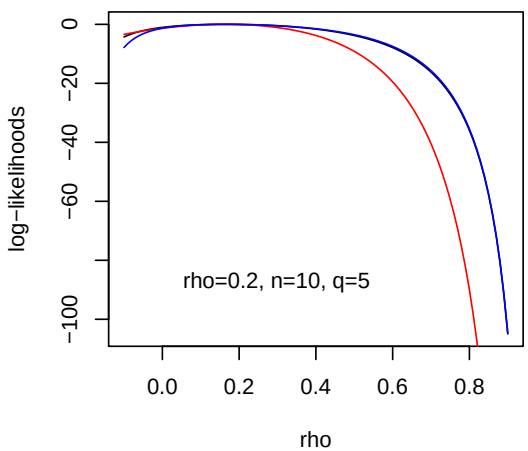
(Cox & Reid, 2004)

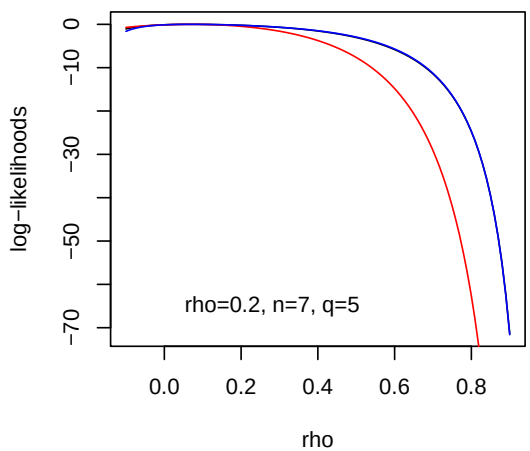


Likelihood ratio test









... symmetric normal +

- $Y_i \sim N(\mu \underline{1}, \sigma^2 R)$ $\qquad R_{st} = \rho$
- $\hat{\mu} = \hat{\mu}_{CL}, \quad \hat{\sigma}^2 = \hat{\sigma}^2_{CL}, \quad \hat{\rho} = \hat{\rho}_{CL}$
- $G(\theta) = H(\theta)J(\theta)^{-1}H(\theta) = i(\theta)$ expected Fisher information
- pairwise likelihood is fully efficient
- also true for $Y_i \sim N(\mu, \Sigma)$
  (Mardia, Hughes, Taylor 2007; Jin 2009)
- because $U_{CL}(\theta) = J(\theta)H(\theta)U_{full}(\theta)$ Pagui; Pace et al., 2011

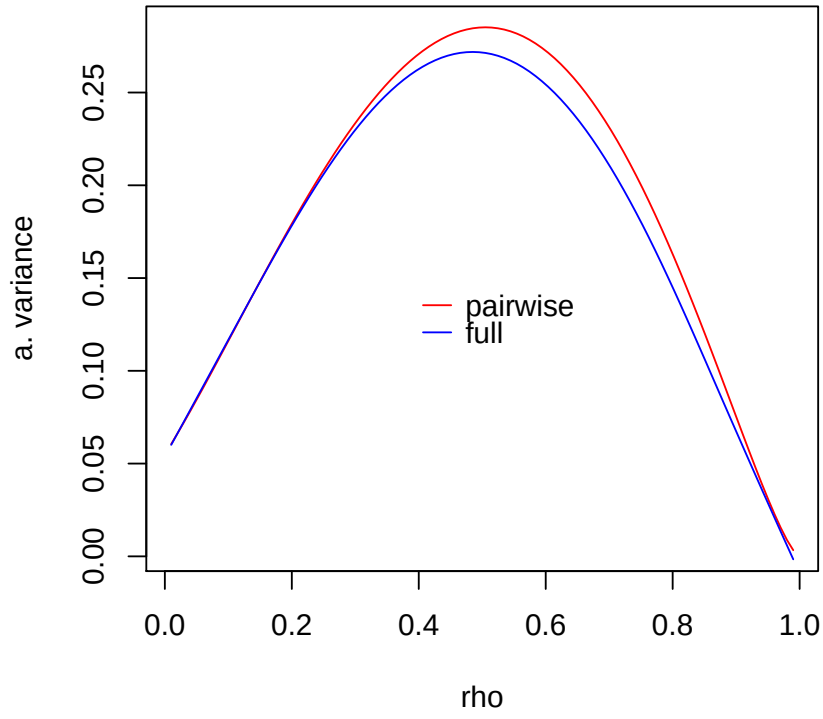
Example: dichotomized MV Normal

$Y_{ir} = 1\{Z_{ir} > 0\}$ $\quad Z \sim N(0, R)$ $\quad r = 1, \ldots, m; i = 1, \ldots, n$

$$\ell_2(\rho) = \sum_{i=1}^{n} \sum_{s<r} \{y_{ir} y_{is} \log P(y_r = 1, y_s = 1) + y_{ir}(1 - y_{is}) \log P_{10} + (1 - y_{ir}) y_{is} \log P_{01} + (1 - y_{ir})(1 - y_{is}) \log P_{00}\}$$

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{1}{n} \frac{4\pi^2}{m^2} \frac{(1 - \rho^2)}{(m-1)^2} \text{var}(T) \quad T = \sum_i \sum_{s<r} (2 y_{ir} y_{is} - y_{ir} - y_{is})$$

$$\text{var}(T) = nm^4 (p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + m^3(-6p_{1111} \ldots) + m^2(\ldots) + m(\ldots)$$



ρ	0.02	0.05	0.12	0.20	0.40	0.50
ARE	0.998	0.995	0.992	0.968	0.953	0.968
ρ	0.60	0.70	0.80	0.90	0.95	0.98
ARE	0.953	0.903	0.900	0.874	0.869	0.850

Example: multi-level probit model

- latent variable: $z_{ir} = x'_{ir}\beta + b_i + \epsilon_{ir}, \quad \epsilon_{ir} \sim N(0,1)$
- binary observations: $y_{ir} = 1(z_{ir} > 0); \quad r = 1, \ldots m_i; i = 1, \ldots n$
- probit model: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x'_{ir}\beta + b_i); \quad b_i \sim N(0, \sigma_b^2)$
- likelihood

$$L(\beta, \sigma_b) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{r=1}^{m_i} \Phi(x'_{ir}\beta + b_i)^{y_{ir}} \{1 - \Phi(x'_{ir}\beta + b_i)\}^{1-y_{ir}} \phi(b_i, \sigma_b^2) db_i$$

- pairwise likelihood

$$CL(\beta, \sigma_b) = \prod_{i=1}^{n} \prod_{r<s} P_{11}^{y_{ir}y_{is}} P_{10}^{y_{ir}(1-y_{is})} P_{01}^{(1-y_{ir})y_{is}} P_{00}^{(1-y_{ir})(1-y_{is})}$$

- each $Pr(y_{ir} = j, y_{is} = k)$ evaluated using $\Phi_2(\cdot, \cdot; \rho_{irs})$

(Renard et al., 2004)

... multi-level probit (Renard et al. 2004)

- computational effort doesn't increase with the number of random effects
- pairwise likelihood numerically stable
- efficiency losses, relative to maximum likelihood, of about 20% for estimation of β
- somewhat larger for estimation of σ_b^2

... Example

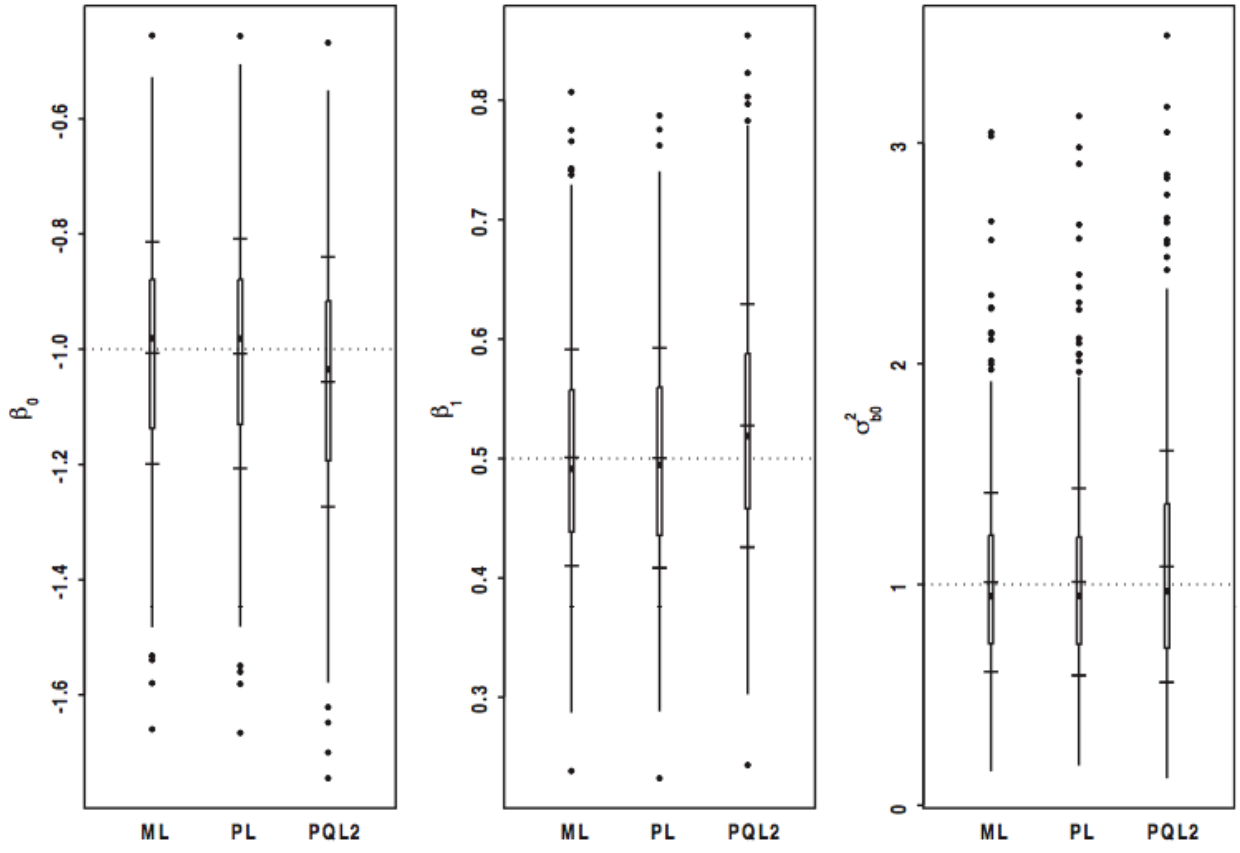


Fig. 5. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (10) with random intercept.

Markov chains Hjort and Varin, 2008

- comparison of likelihood

$$L(\theta; y) = \prod \mathrm{pr}(Y_r = y_r \mid Y_{r-1} = y_{r-1}; \theta)$$

- adjoining pairs CML

$$CML(\theta; y) = \prod \mathrm{pr}(Y_r = y_r, Y_{r-1} = y_{r-1}; \theta)$$

- composite conditional likelihood (= Besag's PL)

$$CCL(\theta; y) = \prod \mathrm{pr}(Y_r = y_r \mid \text{ neighbours } ; \theta)$$

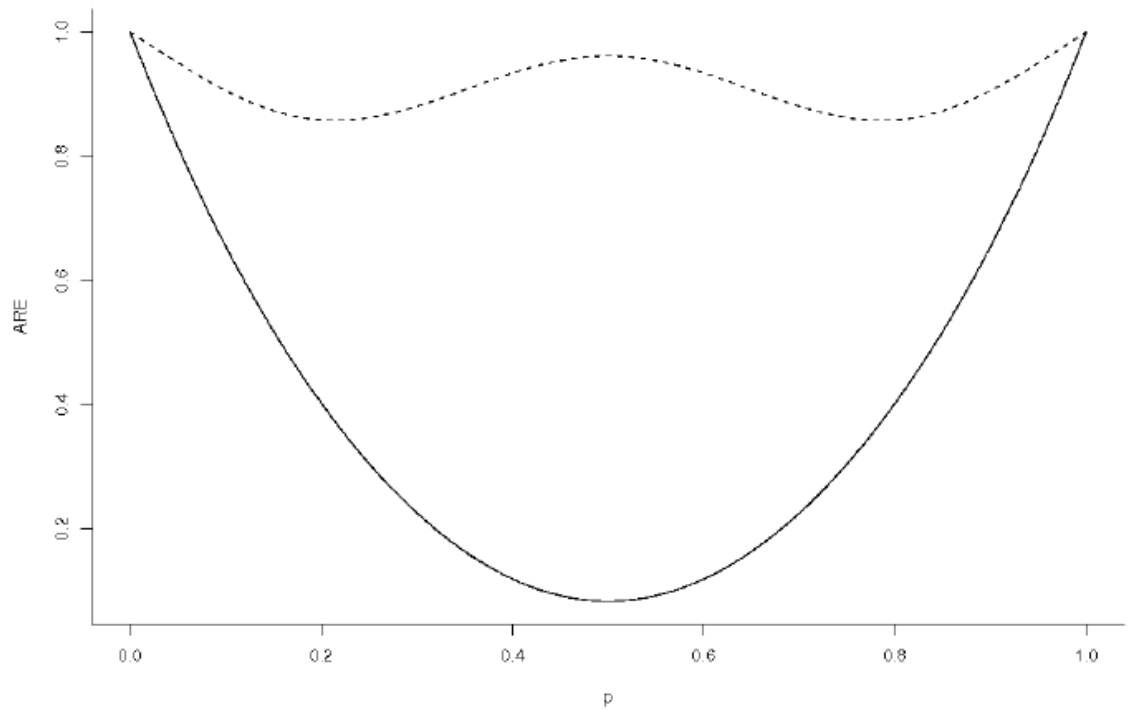
... Markov chain example

- Random walk with p states and two reflecting barriers
- Transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 1-\rho & 0 & \rho & 0 & \dots & 0 \\ 0 & 1-\rho & 0 & \rho & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 & 0 \end{pmatrix}$$

... Markov chain example

Reflecting barrier with five states: efficiency of pairwise likelihood (dashed line) and Besag's pseudolikelihood (solid line)



Example: longitudinal count data

- subjects $i = 1, \ldots, n$
- observations counts $y_{ir}, r = 1, \ldots m_i$
- model $y_{ir} \sim \text{Poisson}(u_{ir} x_{ir}^T \beta)$
- $u_{i1}, \ldots, u_{im_i}$ gamma-distributed random effects
- but correlated $\text{corr}(u_{ir}, u_{is}) = \rho^{|r-s|}$
- joint density has combinatorial number of terms in m_i; impractical
- weighted pairwise composite likelihood

$$\mathcal{L}_{pair}(\beta) = \prod_{i=1}^{n} \frac{1}{m_i - 1} \prod_{r=1}^{m_i} \prod_{s=r+1}^{m_i} f(y_{ir}, y_{is}; \beta)$$

- weights chosen so that $\mathcal{L}_{pair}$ = full likelihood if $\rho = 0$

Henderson & Shimura, 2003

Estimating functions

- Suppose $y = (y_1, \ldots, y_n)$ are i.i.d. from density $f(y; \theta)$, and $g(y_i, \theta)$ is a function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$, satisfying

$$\mathrm{E}\{g(y_i, \theta)\} = 0$$

- examples:
- define an estimator of θ by

$$\sum_{i=1}^{n} g(y_i, \tilde{\theta}_g) = 0$$

- under smoothness conditions on g and f, etc., we have

$$\sqrt{n}(\tilde{\theta}_g - \theta) \xrightarrow{\mathcal{L}} N\{0, G^{-1}(\theta)\}, \quad G^{-1}(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$$

- $H(\theta) = -\mathrm{E}\{\nabla_\theta g(y_1, \theta)\}, \quad J(\theta) = \mathrm{var}\{g(y_1, \theta)\}$
- composite likelihood a particular example of this

... estimating functions

- an optimal estimating function maximizes $G(\theta)$, or minimizes $G^{-1}(\theta)$ in a class of unbiased estimating functions
- suppose now that we assume instead Y_i independent, and

$$\mathrm{E}(y_i;\theta) = \mu_i(\theta), \quad \mathrm{var}(y_i;\theta) = V_i(\theta)$$

- define $g(\underline{y},\theta)$ by

$$\sum_{i=1}^{n} w_i(\theta)\{y_i - \mu_i(\theta)\}$$

- $G(\theta)$ is maximized if

$$w_i(\theta) \propto \frac{\mu_i'(\theta)}{V_i(\theta)}$$

... estimating functions

- $\sum_{i=1}^{n} w_i(\theta)\{y_i - \mu_i(\theta)\} = 0; \quad w_i \propto \frac{\mu_i'(\theta)}{V_i(\theta)}$
- special case: $h(\mu_i) = x_i^T\beta$, $V_i = \phi a_i V(\mu_i)$
- $\tilde{\beta}_g \dot{\sim} N(\beta, \dots)$
- a. $\text{var}(\tilde{\beta}_g) = H^{-1}JH^{-1} = X^T\Omega X, \quad \Omega = \text{diag}(\omega_i)$
- $\omega_i \propto \frac{1}{\phi a_i V(\mu_i) h'(\mu_i)^2}$
- If $V(\mu_i)$ is mis-specified, then

$$\text{a. var}(\tilde{\beta}_g) = H^{-1}JH^{-1} = (X^T\Omega X)^{-1}(X^T\tilde{\Omega}X)(X^T\Omega X)^{-1}$$

... estimating functions

- even if $V(\mu)$ is incorrectly specified, $\tilde{\beta}$ is still consistent
- $$\text{a.Var}(\tilde{\beta}) = (X^T \omega X)^{-1} \text{Var}\{g(y;\beta)\}(X^T \Omega X)^{-1}$$
- often is well approximated by $(X^T \Omega X)^{-1}$ in any case
- when extended to dependent data, called generalized estimating equations

Generalized estimating equations

- $y_j = (y_{j1}, \ldots, y_{jn_j})$; $\mathrm{E}(y_j) = \mu_j$; $\mathrm{var}(y_j) =$
- estimating equation for β:

$$\sum_{j=1}^{n} \left(\frac{\partial \mu_j}{\partial \beta^T} \right) V(\mu_j; \alpha)^{-1} (y_j - \mu_j) = 0$$

- multivariate version of quasi-likelihood equation
- needs some specification of $V(\cdot;\cdot)$ called "working covariance matrix"
- `gee in library(gee) offers several choices: independent, exchangeable,` $AR(p)$`, etc.`
- estimate of β is consistent, even if $V(\cdot;\cdot)$ is mis-specified
- but estimates of $\mathrm{Var}(\tilde{\beta})$ will be incorrect if