Topics in Likelihood Inference

STA4508H

Nancy Reid

University of Toronto

February 2, 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

Presentations

Feb 9	Shiki: Solomon and Cox (1992)
Feb 16	Angela: Robert: Barndorff-Nielsen and Cox (1979)
Feb 23	Hengchao: Siyue: De Stavola and Cox (2008) Manuel: Battey and Cox (2018) Ziang: Cox (1975)

Feb 9 and Feb 16 in SS 2120;
Feb 23 online

Recap: nuisance parameters

profile, marginal, conditional, saddlepoint, Laplace

Recap: nuisance parameters

profile, marginal, conditional, saddlepoint, Laplace

... Nuisance parameters

- partition score vector: $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}$; $\quad \frac{1}{\sqrt{n}} U_\psi(\theta) \xrightarrow{d} N_q\{0, i_{1\psi\psi}(\theta)\}$
- partition information matrix: $i_1(\theta) = \begin{pmatrix} i_{1\psi\psi} & i_{1\psi\lambda} \\ i_{1\lambda\psi} & i_{1\lambda\lambda} \end{pmatrix} \qquad i_1^{-1}(\theta) = \begin{pmatrix} i_1^{\psi\psi} & i_1^{\psi\lambda} \\ i_1^{\lambda\psi} & i_1^{\lambda\lambda} \end{pmatrix}$

$$i^{\psi\psi} = \quad (i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi})^{-1}$$

$$\sqrt{n}(\hat{\psi} - \psi) \doteq \frac{1}{\sqrt{n}} (i_1^{\psi\psi})^{-1} (U_\psi - i_{\psi\lambda} i_{\lambda\lambda}^{-1} U_\lambda) \qquad \sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N_q\{0, i_1^{\psi\psi}(\theta)\}$$

$$2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \doteq (\hat{\psi} - \psi)^{\mathrm{T}} i^{\psi\psi} (\hat{\psi} - \psi) \qquad 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \xrightarrow{d} \chi^2_q$$

$$\sqrt{n}(\hat{\theta} - \theta) \doteq \frac{1}{\sqrt{n}} i_1^{-1}(\theta) U(\theta)$$

column vectors

Linear exponential families

- conditional density free of nuisance parameter
- $f(y_i;\psi,\lambda) = \exp\{\psi^T s(y_i) + \lambda^T t(y_i) - k(\psi,\lambda)\}h(y_i)$
- $f(y;\psi,\lambda) = \exp\{\psi^T \Sigma s(y_i) + \lambda^T \Sigma t(y_i) - nk(\psi,\lambda)\}\Pi h(y_i)$

  Let $s = \Sigma s(y_i), t = \Sigma t(y_i)$
- $f(s,t;\psi,\lambda) = \exp\{\psi^T s + \lambda^T t - nk(\psi,\lambda)\}\tilde{h}(s)$

$$
\begin{aligned}
f(s \mid t;\psi) &= \frac{f(s,t;\psi,\lambda)}{\int f(s,t;\psi,\lambda)ds} \\
&= \frac{\exp\{\psi^T s + \lambda^T t - nk(\psi,\lambda)\}\tilde{h}(s)}{\int \exp\{\psi^T s + \lambda^T t - nk(\psi,\lambda)\}\tilde{h}(s)ds} \\
&= \frac{\exp\{\psi^T s\}\tilde{h}(s)}{\int \exp\{\psi^T s\}\tilde{h}(s)ds} \\
&= \exp\{\psi^T s - n\tilde{k}_t(\psi)\}\tilde{h}_t(s)
\end{aligned}
$$

$\tilde{k}_t$, $\tilde{h}_t$ just convenient notation for integral of denominator

Approximate conditional and marginal inference

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

- $\ell_m(\psi) \doteq \ell_p(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|$

- $\ell_c(\psi) \doteq \ell_p(\psi) + \frac{1}{2}\log|j_{\eta\eta}(\psi, \hat\eta_\psi)|$ $\qquad \exp\{\psi^{\mathrm{T}}s + \eta^{\mathrm{T}}t - c(\psi, \eta)\}$

- adjusted profile log-likelihood

$$\ell_A(\psi) = \ell_p(\psi) + A(\psi)$$

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

- generic form is $A_{FR}(\psi) = +\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| - \log|\frac{d(\lambda)}{d\hat\lambda_\psi}|$ $\qquad$ Fraser 03

- closely related $A_{BN}(\psi) = -\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| + \log|\frac{d\hat\lambda}{d\hat\lambda_\psi}|$ $\qquad$ SM §12.4.1

BN 83

Semi-parametric models

- Recall: $y_1, \ldots, y_n$ jumps of a Poisson process

- rate function $\lambda(\cdot)$ observed on $(0, \tau)$
- events at $0 < y_1 < \cdots < y_n < \tau$ SM §6.5.1
- likelihood function

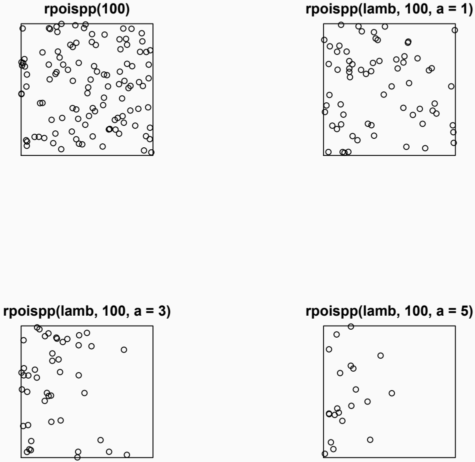
$$L\{\lambda(\cdot); y\} = \left\{\prod_{i=1}^{n} \lambda(y_i)\right\} \exp\{-\int_0^{\tau} \lambda(u) du\}$$

- log-likelihood function

$$\ell\{\lambda(\cdot); y\} = \sum_{i=1}^{n} \log \lambda(y_i) - \int_0^{\tau} \lambda(u) du$$

- in space:

$$\ell\{\lambda(\cdot); y\} = \sum_{i=1}^{n} \log \lambda(y_i) - \int_{\mathcal{S}} \lambda(u) du$$



$$\lambda(y_1, y_2) = 100 \exp(-a y_1)$$

Survival data

- Example: Survival data $(y_i, d_i), i = 1, \dots, n$
  - $y_i = \min(y_i^o, c_i)$ $\quad y_i^o \sim F(\cdot;\theta);\ c_i \sim G;\ y_i^o$ independent of c_i
  - $d_i = 1\{y_i = y_i^o\}$ $\quad$ uncensored observation
  - $f(y_i, d_i; \theta) = [f(y_i;\theta)\{1 - G(y_i)\}]^{d_i}[\{1 - F(y_i;\theta)\}g(y_i)]^{1-d_i}$

joint density

$$\ell(\theta) = \sum_{i=1}^{n}[d_i \log f(y_i;\theta) + (1 - d_i)\log\{1 - F(y_i;\theta)\}]$$

$$+ \text{ terms depending on } G$$

$$= \sum\{d_i \log \lambda(y_i;\theta) - \Lambda(y_i;\theta)\}$$

$\Lambda(y;\theta) = -\log\{1 - F(y;\theta)\}; \quad \lambda(y;\theta) = f(y;\theta)/\{1 - F(y;\theta)\}$

Survival data

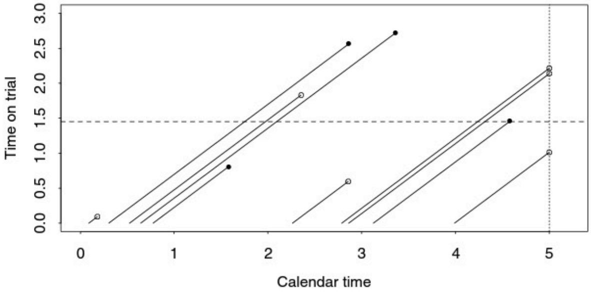


Figure 5.8 Lexis diagram showing typical pattern of censoring in a medical study. Each individual is shown as a line whose x coordinates run from the calendar time of entry to the trial to the calendar time of failure (blob) or censoring (circle). Censoring occurs at the end of the trial, marked by the vertical dotted line, or earlier. The vertical axis shows time on trial, which starts when individuals enter the study. The risk set for the failure at calendar time 4.5 comprises those individuals whose lines touch the horizontal dashed line; see page 543.

thus we study events on the vertical axis. Calendar time may be used to account for changes in medical practice over the course of a trial.

In applications the assumption that C_j and Y_j^0 are independent is critical. There would be serious bias if the illest patients drop out of a trial because the treatment makes them feel even worse, thereby inducing association between survival and censoring variables because patients die soon after they withdraw.

The examples above all involve *right-censoring*. Less common is left-censoring, where the time of origin is not known exactly, for example if time to death from a disease is observed, but the time of infection is unknown.

In practice a high proportion of the data may be censored, and there may be a

0+	1+	1+	3+	3+	7	10+	11+	12+	12+	15+	18+
20+	22+	22+	24+	25+	26+	31+	36+	36+	36	38	40
47+	47+	49+	53+	53+	55+	56+	57+	61+	67+	67+	70
73	75+	77+	83+	84+	88+	89+	99	121+	122+	123+	141+
0+	0+	2+	2+	2+	2+	3	3+	4+	5+	9+	10+
11	12+	13	13+	18+	22+	22+	24+	24+	24+	25+	26+
27	28	32+	35+	36	40+	43+	50+	54			

Table 5.3 Blalock–Taussig shunt data (Oakes, 1991). The table gives survival time of shunt (months after operation) for 48 infants aged over one month at time of operation, followed by times for 33 infants aged 30 or fewer days at operation. Infants whose shunt has not yet failed are marked +.

Proportional hazards regression

- semi-parametric model: $\lambda(y; x, \beta) = \lambda(y) \exp(x^{\mathrm{T}} \beta)$

- log-likelihood function

$$\ell(\beta, \lambda; y, d) = \sum_{i=1}^{n} d_i \log\{\lambda(y_i; x_i, \beta)\} - \Lambda(y_i; x_i, \beta)$$

$$= \sum_{i=1}^{n} [d_i \{x_i^{\mathrm{T}} \beta + \log \lambda(y_i)\} - \Lambda(y_i) \exp(x_i^{\mathrm{T}} \beta)]$$

- partial log-likelihood function

$$\ell_{part}(\beta; y, d) = \sum_{i=1}^{n} d_i \{x_i^{\mathrm{T}} \beta - \log \sum_{j \in \mathcal{R}_i} \exp(x_j^{\mathrm{T}} \beta)\}$$

- $y_1 < \cdots < y_n; \quad \mathcal{R}_i = \{j; y_j \geq y_i\}$

... PH regression

$$\ell_{part}(\beta; y, d) = \sum_{i=1}^{n} d_i\{x_i^{\mathrm{T}}\beta - \log\sum_{j\in\mathcal{R}_i}\exp(x_j^{\mathrm{T}}\beta)\}$$

$$= \sum_{i=1}^{n} d_i\{x_i^{\mathrm{T}}\beta - \log A_i(\beta)\}$$

$$\frac{\partial \ell_{part}(\beta)}{\partial\beta} = \sum_{i=1}^{n} d_i\left\{x_i - \frac{A_i'(\beta)}{A_i(\beta)}\right\}$$

$$-\frac{\partial^2 \ell_{part}(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}} = \sum_{i=1}^{n} d_i\left\{\frac{A_i''(\beta)}{A_i(\beta)} - \frac{A_i'(\beta)A_i'(\beta)^{\mathrm{T}}}{A_i(\beta)^2}\right\}$$

notation is a bit careless

... PH regression

- partial log-likelihood function

$$\ell_{part}(\beta; y, d) = \sum_{i=1}^{n} d_i \{x_i^{\mathrm{T}}\beta - \log \sum_{j \in \mathcal{R}_i} \exp(x_i^{\mathrm{T}}\beta)\}$$

- can be motivated as:
  1. marginal log-likelihood of the ranks of the failure times
  2. $\prod_{i=1}^{n} \Pr(\text{unit } i \text{ fails at } y_i \mid \text{history to } y_i^-, \text{one failure from } \mathcal{R}_i)$ CL
  3. profile log-likelihood function

- for inference, $\ell_{part}(\beta)$ has usual properties

  1. $\hat{\beta}_{part} \dot{\sim} N\{\beta, j_{part}^{-1}(\hat{\beta})\}$, $\quad 2\{\ell_{part}(\hat{\beta}_{part}) - \ell_{part}(\beta)\} \dot{\sim} \chi^2_d$ Davison §10.8; Cox 1972, 1975

- partial log-likelihood function

$$\ell_{part}(\beta; y, d) = \sum_{i=1}^{n} d_i \{x_i^{\mathrm{T}}\beta - \log \sum_{j \in \mathcal{R}_i} \exp(x_j^{\mathrm{T}}\beta)\}$$

- is also, 3. profile log-likelihood function if $\lambda(\cdot)$ is represented by a vector of values $(\lambda_1, \ldots, \lambda_n) = \{\lambda(y_1), \ldots \lambda(y_n)\}$

- why does usual likelihood inference apply?

- can be connected to theory of empirical likelihood

Murphy & van der Waart, 2000; van der Waart 1998, Ch. 25

Profile likelihood in semi-parametric models

- $\ell(\beta, \lambda; y), \beta \in \mathbb{R}^d; \lambda = \lambda(\cdot)$
- $\ell_{\mathrm{p}}(\beta; y) = \ell(\beta, \tilde{\lambda}_\beta; y); \quad \tilde{\lambda}_\beta = \arg\sup_\lambda \ell(\beta, \lambda; y)$
- example: failure times y with hazard $\lambda(y \mid x) = e^{x\beta}\lambda(y)$

PH model, no censoring

- $f(y_i; \theta, \lambda) = e^{x_i\beta}\lambda(y_i)\exp\{-e^{x_i\beta}\Lambda(y_i)\}$ $\quad \Lambda = \int \lambda$
- empirical likelihood:

$$EL(\beta, \Lambda; y) = \prod_{i=1}^{n} e^{x_i\beta}\Lambda\{y_i\}\exp\{-e^{x_i\beta}\Lambda(y_i)\}$$

- maximizing value of $\Lambda(\cdot)$ must have jumps at y_i only — replace $\Lambda(y_i)$ by sum

SM §10.8 "it suffices to estimate the baseline cum. haz. fn by a step function $\Sigma_{j:y_j \geq y}\lambda_j$"

- empirical likelihood:

$$EL(\beta, \Lambda; y) = \prod_{i=1}^{n} e^{x_i\beta} \Lambda\{t_i\} \exp\{-e^{x_i\beta}\Lambda(t_i)\}$$

- $\hat{\Lambda}_\beta\{y_i\} = \left\{\sum_{i:y_j \geq y_i} \exp(x_i\beta)\right\}^{-1}$

- profile log-likelihood

$$L_{\mathrm{p}}(\beta) = \prod_{i=1}^{n} \frac{e^{x_i\beta}}{\sum_{i:y_j \geq y_i} \exp(x_i\beta)}$$

- same as partial likelihood motivated by different arguments

- observation (D, W, Z); D and W are independent, given Z

- $\Pr(D = 0) = \{1 + \exp(\gamma + \beta e^z)\}^{-1}$

- $W \sim N(\alpha_0 + \alpha_1 z; \sigma^2)$
- $Z \sim g(\cdot)$, non-parametric

- (d_C, w_C, z_C) a 'complete' observation
- (d_R, w_R) has a missing covariate

- $f(x; \theta, g) = f(d_C, w_C \mid z_C; \theta) g(z_C) \int f(d_R, w_R \mid z; \theta) g(z) dz$

$x = (d_C, w_C, z_C, d_R, w_R)$

$\theta = \gamma, \beta, \alpha_0, \alpha_1, \sigma^2$

$$EL(\theta, g) = f(d_C, w_C \mid z_C; \theta) g\{z_C\} \int f(d_R, w_R \mid z) g(z) dz$$

Profile likelihood

1. $\sqrt{n}(\hat\theta - \theta_0) = \dfrac{1}{\sqrt{n}}\tilde\imath^{-1}(\theta_0)\tilde U(\theta_0) + o_p(1)$
   - $\tilde U(\theta_0) = \dfrac{\partial \ell(\theta,\lambda)}{\partial\theta} - \text{Proj}_g \dfrac{\partial \ell(\theta,\lambda)}{\partial\theta}$
   - projection of $\partial\ell_\theta$ onto the closed linear span of the score functions for $\lambda(\cdot)$
   - $\tilde\imath(\theta_0) = \text{var}\{\tilde U_j(\theta_0)\}$

$\tilde U = \sum \tilde U_j$; $\imath$ is $O(1)$

2. $\ell_p(\hat\theta) = \ell_p(\theta_0) + \dfrac{1}{2}n(\hat\theta - \theta_0)^{\mathrm{T}}\tilde\imath(\theta_0)(\hat\theta - \theta_0) + o_p(1)$
3. for any random sequence $\tilde\theta_n \xrightarrow{p} \theta_0$, plus conditions on the model,

$$\ell_p(\tilde\theta_n) = \ell_p(\theta_0) + (\tilde\theta_n - \theta_0)^{\mathrm{T}}\sum_{j=1}^n \tilde U_j(\theta_0) - \frac{1}{2}n(\tilde\theta_n - \theta_0)^{\mathrm{T}}\tilde\imath^{-1}(\theta_0)(\tilde\theta_n - \theta_0)$$
$$+ o_p(\sqrt{n}||\tilde\theta_n - \theta_0|| + 1)^2$$

... inference

- $$\ell_{\mathrm{p}}(\tilde\theta_n) = \ell_{\mathrm{p}}(\theta_0) + (\tilde\theta_n - \theta_0)^{\mathrm T} \sum_{j=1}^n \tilde U_j(\theta_0) - \frac{1}{2} n(\tilde\theta_n - \theta_0)^{\mathrm T} \tilde\imath^{-1}(\theta_0)(\tilde\theta_n - \theta_0)$$
$$+ o_p(\sqrt{n}||\tilde\theta_n - \theta_0|| + 1)^2$$
- this result (3.) gives (1.) and (2.)
- as in parametric models, lead to

$$(\hat\theta - \theta_0) \dot\sim N\{0, \tilde\imath^{-1}(\theta_0)\}$$

- and likelihood ratio test

$$2\{\ell_{\mathrm{p}}(\hat\theta) - \ell_{\mathrm{p}}(\theta_0)\} \dot\sim \chi^2_d$$

- proof uses least favourable sub-models through the true model
- effectively turns infinite-dimensional parameter finite

Proportional hazards model

simplified

- $$\ell(\beta, \lambda(\cdot); y, d) = \sum_{i=1}^{n} [d_i\{x_i\beta + \log \lambda(y_i)\} - \Lambda(y_i)\exp(x_i\beta)]$$
- score function for β:
$$\partial\ell/\partial\beta = \sum_{i=1}^{n} \{d_i x_i - x_i e^{x_i\beta}\Lambda(y_i)\}$$
- score function for $\lambda(\cdot)$:

in the 'direction' $h(\cdot)$

$$\sum_{i=1}^{n} d_i h(y_i) - e^{x_i\beta}\int_0^{y_i} h(t)d\Lambda(t)$$
- we need to project $\partial\ell/\partial\beta$ on the space spanned by the nuisance score functions
- result: $\sum_{i=1}^{n} d_i \left(x_i - \frac{M_1}{M_0}(y_i)\right) - e^{x_i\beta}\int_0^{y_i}\left(x_i - \frac{M_1}{M_0}(t)\right)d\Lambda(t)$

Semi-parametric models

- profile log-likelihood can (often) be defined using a least favorable sub-model

finite dimensional

- standard likelihood asymptotics apply for inference based on the profile log-likelihood

- in other examples, we see that profiling out large numbers of nuisance parameters can lead to poor finite sample results

- ?does this happen in semi-parametric models?
- seems unlikely for proportional hazards regression

complete separation of the parameters?

- other examples in vdW & M include current status data, gamma frailty models, partially missing data, ...

Infinite-dimensional models

- recall that $L(\theta; y) \propto f(y; \theta)$

$f(y;\theta)$ a density w.r.t. to dominating measure

- more abstract definition:
  if a probability measure Q is absolutely continuous w.r.t. to a probability measure P, and both possess densities w.r.t. to a measure μ, then the likelihood of Q w.r.t. to P is the Radon-Nikodym derivative

$$\frac{dQ}{dP} = \frac{q}{p}, \text{ a.e. } P$$

- some semi-parametric models have a dominating measure, and a family of densities
- some can be handled by the notion of empirical likelihood
- some may use mixtures of these

- Definition: Given a measure P, and a sample $(y_1, \dots, y_n)$, the empirical likelihood function is

$$EL(P; y) = \prod_{i=1}^{n} P(\{y_i\}),$$

where $P\{y\}$ is the measure of the one-point set $\{y\}$

- Definition: Given a model $\mathcal{P}$, a maximum likelihood estimator is the distribution $\widehat{P}$ that maximizes the empirical likelihood over $\mathcal{P}$

- may or may not exist

- $\mathcal{P}$ is the set of all probability distributions on a measurable space $\{\mathcal{Y}, \mathcal{A}\}$

1-point sets are measurable

- suppose the observed values $y_1, \ldots, y_n$ are distinct
- $\{(P\{y_1\}, \ldots, P(\{y_n\}); P \in \mathcal{P}\} \iff$

$$(p_1, \ldots, p_n), p_i \geq 0, \Sigma p_i = 1)$$

- empirical likelihood maximized at

$$(\frac{1}{n}, \ldots, \frac{1}{n})$$

- empirical distribution function is the nonparametric MLE

$F_n(\cdot) = n^{-1}\Sigma 1(Y_i \leq \cdot)$

- EL is not the same as $\prod f(y_i)$, even if P has a density f

Compare Owen, Ch. 2

- for $y \in \mathbb{R}$, define $F(y) = \Pr(Y \leq y)$ and
$$F(y^-) = \Pr(Y < y)$$

- for $y_1, \ldots, y_n$ the nonparametric likelihood function is
$$L(F) = \prod_{i=1}^{n} \{F(y_i) - F(y_i^-)\},$$

- hence 0 if F is continuous

- Theorem 2.1 of Owen:
$$L(F) < L(F_n), \quad F_n(y) = \frac{1}{n} \Sigma 1\{y_i \leq y\}$$

- there is a likelihood function on the space of distribution functions for which the empirical c.d.f. is the maximum likelihood estimator

why does this fail for densities?

Aside: empirical likelihood

- profile version of empirical likelihood
- $\mathcal{R}(\theta) = \sup\left\{\frac{L(F)}{L(F_n)} \mid F \in \mathcal{F}, T(F) = \theta\right\}$

  $\mathcal{R}$ a relative likelihood, hence np_i
- example: $T(F) = \int x dF(x)$
- $\mathcal{R}(\theta) = \max\{\prod_{i=1}^n np_i \mid \sum_{i=1}^n p_i y_i = \theta, p_i \geq 0, \sum p_i = 1\}$
- For $y_1, \ldots, y_n$ i.i.d. F_0, $\quad \mathrm{E}(y_i) = \theta_0, \mathrm{var}(y_i) < \infty$,
-
$$-2\log\mathcal{R}(\theta_0) \xrightarrow{d} \chi_1^2, \quad n \to \infty$$

  Theorem 2.2 Owen
- $\hat{p}_i = \frac{1}{n}\frac{1}{\{1+\alpha(y_i-\theta_0)\}}, \quad \frac{1}{n}\sum_{i=1}^n \frac{y_i-\theta_0}{1+\alpha(y_i-\theta_0)} = 0$

Semi-parametric logistic regression

- $Pr(Y = 1 \mid V, W) = \dfrac{e^{\theta V + \eta(W)}}{1 + e^{\theta V + \eta(W)}}$
- sample $(Y_i, V_i, W_i), i = 1, \ldots, n$ independent
- $$L(\theta, \eta; \underline{Y}) \propto \prod_{i=1}^{n} \left\{ \frac{e^{\theta V_i + \eta(W_i)}}{1 + e^{\theta V_i + \eta(W_i)}} \right\}^{y_i} \left\{ \frac{1}{1 + e^{\theta V_i + \eta(W_i)}} \right\}^{1-y_i}$$
- $\tilde{\eta}(w_i) = \infty$ when $y_i = 1, \tilde{\eta}(w_i) = -\infty$ when $y_i = 0$ gives
$$L(\theta, \tilde{\eta}) \to \infty$$

we can't maximize it

- suggestion: penalized log-likelihood
$$\log L(\theta, \eta; \underline{Y}) - \hat{\alpha}_n^2 \int \{\eta^{(k)}(w)\}^2 dw$$
- needs separate analysis of properties

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 Log-Likelihood:

Lindsay, 1988

$$c\ell(\theta; y) = \sum_{k \in K} w_k \ell_k(\theta; y)$$

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

- also called:
  - pseudo-likelihood (spatial modelling)
  - quasi-likelihood (econometrics)
  - limited information method (psychometrics)

Examples of composite log-likelihood

$\sum_{r=1}^{m} w_r \log f_1(y_r;\theta)$	Independence
$\sum_{r=1}^{m} \sum_{s>r} w_{rs} \log f_2(y_r, y_s;\theta)$	Pairwise
$\sum_{r=1}^{m} w_r \log f(y_r \mid y_{(-r)};\theta)$	Conditional
$\sum_{r=1}^{m} \sum_{s>r} w_{rs} \log f(y_r \mid y_s;\theta)$	All pairs conditional
$\sum_{r=1}^{m} w_r \log f(y_r \mid y_{r-1};\theta)$	Time series
$\sum_{r=1}^{m} w_r \log f(y_r \mid \text{'neighbours' of } y_r;\theta)$	Spatial

Small blocks of observations; pairwise differences; ...
your favourite combination...

Derived quantities

single response y with density $f(y;\theta)$, $\quad y \in \mathbb{R}^m, \theta \in \mathbb{R}^d$

composite log-likelihood $\quad c\ell(\theta; y) = \log cL(\theta; y) = \sum_k w_k \ell_k(\theta; y)$

composite score function $\quad U_{CL}(\theta) = \partial c\ell(\theta; y)/\partial\theta$

sensitivity $\quad H(\theta) = \mathrm{E}_\theta\{-\partial^2 c\ell(\theta; y)/\partial\theta\partial\theta^T\}$

variability $\quad J(\theta) = \mathrm{E}_\theta\{U_{CL}(\theta)U(\theta)^T\}$

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

... derived quantities

sample $y = (y_1, \dots, y_n)$ with joint density $f(y; \theta)$, $\quad y \in \mathbb{R}^m, \theta \in \mathbb{R}^d$

score function	$U_{CL}(\theta) = \frac{\partial}{\partial\theta} c\ell(\theta; y) = \sum_{i=1}^n \frac{\partial}{\partial\theta} c\ell(\theta; y_i)$
maximum composite likelihood estimate	$\hat\theta_{CL} = \hat\theta_{CL}(y) = \arg\sup_\theta c\ell(\theta; y)$
score equation	$U_{CL}(\hat\theta_{CL}) = c\ell'(\hat\theta_{CL}) = 0$
composite LRT	$w_{CL}(\theta) = 2\{c\ell(\hat\theta_{CL}) - c\ell(\theta)\}$
Godambe information	$G(\theta) = G_n(\theta) = H_n(\theta) J_n^{-1}(\theta) H_n(\theta) = O(n)$

Inference

- Sample: $Y_1, \ldots, Y_n$, i.i.d., $CL(\theta; \underline{y}) = \prod_{i=1}^n CL(\theta; y_i)$

- $\hat\theta_{CL} - \theta \dot\sim N\{0, G^{-1}(\theta)\}$ $\qquad G_n(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$

- $U(\hat\theta_{CL}) \doteq U(\theta) + (\hat\theta_{CL} - \theta)\partial_\theta U(\theta)$ $\qquad U = U_{CL}$

- $\hat\theta_{CL} - \theta \doteq -\partial_\theta U(\theta)^{-1} U(\theta) \doteq H^{-1}(\theta)U(\theta)$

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

- $H^{-1}(\theta)U(\theta) \dot\sim N\{0, H^{-1}(\theta)J(\theta)H^{-\mathrm{T}}(\theta)\}$
- conclude

$$\sqrt{n}(\hat\theta_{CL} - \theta) \dot\sim N\{0, G^{-1}(\theta)\}$$

... 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 $J(\theta)H(\theta)^{-1}$
- $$c\ell(\hat\theta_{CL}) - c\ell(\theta) \doteq \frac{1}{2}(\hat\theta_{CL} - \theta)^{\mathrm{T}}\{-c\ell''(\hat\theta_{CL})\}(\hat\theta_{CL} - \theta)$$
- non-central χ^2 limit
- $J(\theta) = \mathrm{var}U(\theta), \qquad H(\theta) = -\mathrm{E}\partial_\theta U(\theta)$
- if $J(\theta) = H(\theta)$, $w(\theta) \dot\sim \chi^2_d$
- if $d = 1$, $w(\theta) \dot\sim \mu_1\chi^2_1 = J(\theta)H^{-1}(\theta)\chi^2_1$

H, J both scalars

Example: symmetric normal

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

$$
\begin{aligned}
c\ell(\rho; y_1, \ldots, y_n) &= -\frac{nm(m-1)}{4}\log(1-\rho^2) - \frac{m-1+\rho}{2(1-\rho^2)}SS_w \\
&\quad - \frac{(m-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_b}{m} \\
SS_w &= \sum_{i=1}^{n}\sum_{s=1}^{m}(y_{is}-\bar{y}_{i.})^2, \quad SS_b = \sum_{i=1}^{n} y_{i.}^2 \\
\ell(\rho; y_1, \ldots, y_n) &= -\frac{n(m-1)}{2}\log(1-\rho) - \frac{n}{2}\log\{1+(m-1)\rho\} \\
&\quad - \frac{1}{2(1-\rho)}SS_w - \frac{1}{2\{1+(m-1)\rho\}}\frac{SS_b}{m}
\end{aligned}
$$

... symmetric normal

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

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

$O(\frac{1}{n})$ $\quad$ $O(1)$

$n \longrightarrow \infty$ $\quad$ $m \longrightarrow \infty$

... symmetric normal

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

(Cox & Reid, 2004)