Various ‘types’ of likelihood

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

October 30

- HW 1 comments
- presentations
- nuisance parameters: Bayes marginal posterior
- adjusted profile likelihood
- semi-parametric likelihood

1. SM 4.9.2 Let $\eta(\theta)$ be a 1-1 transformation of θ, and consider a model with log-likelihoods $\ell(\theta)$ and $\ell^*(\eta) = \ell(\theta(\eta))$ in the two parametrizations respectively. Assume that $\ell(\cdot)$ has a unique maximum at which the score equation is satisfied.

   (a) Show that the Fisher information transforms as

   $$i^*(\eta) = \frac{\partial\theta^{\mathrm{T}}}{\partial\eta} i(\theta) \frac{\partial\theta}{\partial\eta},$$

   and that a similar equation holds for observed information $j(\hat{\theta})$, but not for the observed information function $j(\theta)$.

   (b) Show that the log-likelihood ratio statistic $w(\theta)$ and the standardized score statistic $w_u(\theta)$ are parametrization invariant, but that the standardized maximum likelihood statistic $w_e(\theta)$ is not, where

   $$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\},$$
   $$w_u(\theta) = U(\theta)^{\mathrm{T}} j^{-1}(\hat{\theta}) U(\theta),$$
   $$w_e(\theta) = (\hat{\theta} - \theta)^{\mathrm{T}} j(\hat{\theta})(\hat{\theta} - \theta).$$

   n ot true, actually, only $U^{\mathrm{T}}(\theta)i^{-1}(\theta)U(\theta)$ works, not with $j(\hat{\theta})$. But $U_\psi^{\mathrm{T}}(\psi, \hat{\lambda}_\psi)(\tilde{j}^{\psi\psi})^{-1}U_\psi(\psi, \hat{\lambda}_\psi)$ is okay.
   re $w_e(\psi)$, note that we would need $\hat{\theta} - \theta = (\hat{\eta} - \eta)\ \partial\theta/\partial\eta|_{\hat{\theta}}$ to get invariance.

2. Suppose that $y_1, \ldots, y_n$ are independent, identically distributed random variables from an exponential family distribution

   $$f(y_i; \theta) = \exp\{\theta^{\mathrm{T}} s(y_i) - c(\theta) - d(y_i)\}.$$

   Show that the distribution of $S = \sum_{i=1}^n s(y_i)$ follows an exponential family distribution, with log-likelihood function

   $$\ell(\theta; s) = \theta^{\mathrm{T}} s - nc(\theta),$$

   and that the maximum likelihood estimate of θ is given by

   $$\mathrm{E}_{\hat{\theta}}(S) = s.$$

3. This is a journal reading exercise.should have specified 2018!

   (a) Find a paper in one of the theoretical statistics journals (e.g. *Annals of Statistics, Biometrika, JASA Theory & Methods, JRSS B, Bernoulli*) with "likelihood" as one of the key words. Explain how likelihood is used in the paper.

   (b) Find a paper in one of the applied statistics journals (e.g. *Annals of Applied Statistics, Biometrics, JASA applications, JRSS C*) with "likelihood" as one of the key words. Explain briefly how likelihood is used in the paper.

Marginal posterior

$$\begin{aligned}
\pi(\theta \mid y) &= \frac{\exp\{\ell(\theta)\}\pi(\theta)}{\int \exp\{\ell(\theta)\}\pi(\theta)d\theta} \\
&\doteq \frac{\exp\{\ell(\theta)\}\pi(\theta)}{\exp\{\ell(\hat\theta)\}(2\pi)^{p/2}|j(\hat\theta)|^{-1/2}\pi(\hat\theta)} \\
&\doteq \frac{1}{(2\pi)^{d/2}}|j(\hat\theta)|^{+1/2}\exp\{\ell(\theta;y)-\ell(\hat\theta;y)\}\frac{\pi(\theta)}{\pi(\hat\theta)}
\end{aligned}$$

$$\begin{aligned}
\pi_m(\psi \mid y) &= \int \pi(\psi,\lambda \mid y)d\lambda \\
&\doteq \frac{\int \exp\{\ell(\psi,\lambda)\}\pi(\psi,\lambda)d\lambda}{\exp\{\ell(\hat\theta)\}(2\pi)^{p/2}|j(\hat\theta)|^{-1/2}\pi(\hat\theta)} \\
&\doteq \frac{\exp\{\ell(\psi,\hat\lambda_\psi)\}(2\pi)^{(p-d)/2}|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{-1/2}\pi(\psi,\hat\lambda_\psi)}{\exp\{\ell(\hat\theta)\}(2\pi)^{p/2}|j(\hat\theta)|^{-1/2}\pi(\hat\theta)} \\
&\doteq \frac{1}{(2\pi)^{d/2}}\exp\{\ell(\psi,\hat\lambda_\psi)-\ell(\hat\psi,\hat\lambda)\}\frac{|j(\hat\psi,\hat\lambda)|^{1/2}}{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}\frac{\pi(\psi,\hat\lambda_\psi)}{\pi(\hat\psi,\hat\lambda)}
\end{aligned}$$

... marginal posterior

$$\pi_{\mathrm{m}}(\psi \mid y) \doteq \frac{1}{(2\pi)^{d/2}} \exp\{\ell_{\mathrm{p}}(\psi) - \ell_{\mathrm{p}}(\hat{\psi})\} |j(\hat{\psi}, \hat{\lambda})|^{1/2} |j_{\lambda\lambda}(\psi, \hat{\lambda}_{\psi})|^{-1/2} \frac{\pi(\psi, \hat{\lambda}_{\psi})}{\pi(\hat{\psi}, \hat{\lambda})}$$

$$\doteq \frac{1}{(2\pi)^{d/2}} \exp\{\ell_{\mathrm{p}}(\psi) - \ell_{\mathrm{p}}(\hat{\psi})\} j_{\mathrm{p}}^{1/2}(\hat{\psi}) \left(\frac{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_{\psi})|} \right)^{1/2} \frac{\pi(\psi, \hat{\lambda}_{\psi})}{\pi(\hat{\psi}, \hat{\lambda})}$$

$$\pi(\theta \mid y) \doteq \frac{1}{(2\pi)^{p/2}} \exp\{\ell(\theta) - \ell(\hat{\theta})\} |j(\hat{\theta})^{1/2} \pi(\hat{\theta})$$

$$\log \pi_m(\psi \mid y) = \ell_{\mathrm{p}}(\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_{\psi})| + \log\{\pi(\hat{\lambda}_{\psi} \mid \psi)\} + \log\{\pi(\psi)\} + c(y)$$

Posterior marginal cdf, $d = 1$

$$\int_{-\infty}^{\psi_0} \pi_{\mathrm{m}}(\psi \mid y) d\psi \doteq \int_{-\infty}^{\psi_0} \frac{1}{(2\pi)^{1/2}} \exp\{\ell_{\mathrm{p}}(\psi) - \ell_{\mathrm{p}}(\hat{\psi})\} |j_p(\hat{\theta})|^{1/2} \frac{\tilde{\pi}}{\hat{\pi}} \left(\frac{|\hat{j}_{\lambda\lambda}|}{|\tilde{j}_{\lambda\lambda}|} \right)^{1/2}$$

$$\vdots$$

$$= \Phi(r) + \phi(r) \left(\frac{1}{r} - \frac{1}{q_{\mathrm{m}}} \right)$$

$$r = \pm\surd[2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\}^{1/2}]$$

$$q_{\mathrm{m}} = -\ell_{\mathrm{p}}'(\psi) j_{\mathrm{p}}^{-1/2}(\hat{\psi}) \frac{\hat{\pi}}{\tilde{\pi}} \left(\frac{|\tilde{j}_{\lambda\lambda}|}{|\hat{j}_{\lambda\lambda}|} \right)^{1/2}$$

compare October 23

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

Example

- $$\ell_{BN}(\psi) = \ell_{\mathrm{p}}(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| + \log|\frac{d\hat\lambda}{d\hat\lambda_\psi}|$$
- $$y_i = x_i^{\mathrm{T}}\beta + \sigma z_i, i = 1, \ldots, n; \quad z \sim N(0, I), \quad \psi = \beta, \lambda = \sigma^2$$
- $\hat\lambda_\psi = \hat\sigma^2_\beta = \frac{1}{n}(y - X\beta)^{\mathrm{T}}(y - X\beta) = \hat\sigma^2 + (\hat\beta - \beta)^{\mathrm{T}}X^{\mathrm{T}}X(\hat\beta - \beta)$
- $$\frac{d\hat\sigma^2}{d\hat\sigma^2_\beta} = 1$$

for fixed $\hat\beta$

$\hat\beta, \hat\sigma^2$ is sufficient

- $\ell_{BN}(\beta) = \ell_{CR}(\beta) = \ell_{\mathrm{p}}(\beta) - \frac{1}{2}\log|j_{\sigma^2\sigma^2}(\beta, \hat\sigma^2_\beta)|$ Cox & R, 1987

Adjusted profile likelihood

- $\ell_{CR}(\psi) = \ell_{\mathrm{p}}(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ need $i_{\psi\lambda} = 0$
- not invariant under transformations $(\psi, \eta) \to (\psi, \lambda(\eta, \psi))$

  interest-respecting reparametrization
- $\ell_{BN}(\psi) = \ell_{\mathrm{p}}(\psi) - \dfrac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| + \log|\dfrac{d\hat{\lambda}}{d\hat{\lambda}_\psi}|$ is invariant
- calculation of 3rd term can be difficult – need to write $\hat{\lambda}_\psi(\hat{\lambda}, \hat{\psi}, a)$
- inference treating ℓ_{CR} or ℓ_{BN} as a real likelihood is only first order
- e.g. $\sqrt{n}(\hat{\psi}_{BN} - \psi) \xrightarrow{d} N(0, \nu)$, ν can be estimated from $-\ell''_{BN}$, etc.

  "the usual asymptotics"
- finite sample approximations seem to be better than those based on profile Sartori, 2003

Aside: parametric inference

- partition score vector:

$$U(\theta) = \begin{bmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{bmatrix}; \quad \frac{1}{\sqrt{n}} U_\psi(\theta) \xrightarrow{d} N_q\{0, i_{1\psi\psi}(\theta)\}$$

- partition information matrix:

$$i_1(\theta) = \begin{bmatrix} i_{1\psi\psi} & i_{1\psi\lambda} \\ i_{1\lambda\psi} & i_{1\lambda\lambda} \end{bmatrix} \qquad i_1^{-1}(\theta) = \begin{bmatrix} i_1^{\psi\psi} & i_1^{\psi\lambda} \\ i_1^{\lambda\psi} & i_1^{\lambda\lambda} \end{bmatrix}$$

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

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

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

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

column vectors

Various ‘types’ of likelihood

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

Semi-parametric models

- Recall: $y_1, \dots, y_n$ 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

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

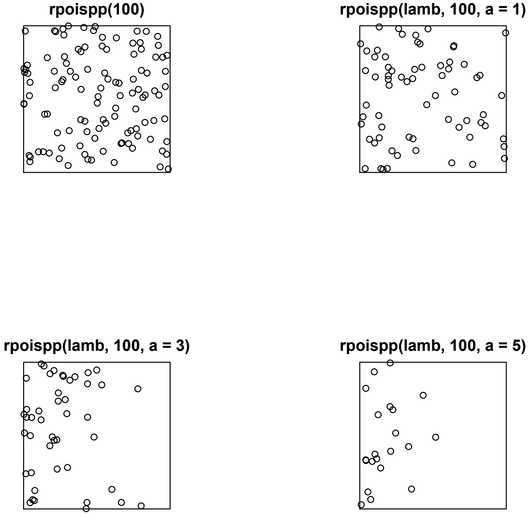
- log-likelihood function

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

- in space:

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

$$y_1, \dots, y_n \subset \mathcal{S}$$



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

Survival data

- Example: Survival data $(y_i, d_i), i = 1, \dots, n$
  - $y_i = \min(y_j^0, c_i)$ $\qquad y_i^0 \sim F(\cdot;\theta); c_i \sim G; y_i^0$ independent of c_i
  - $d_i = 1\{y_i = y_i^0\}$ $\qquad$ 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

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

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

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

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

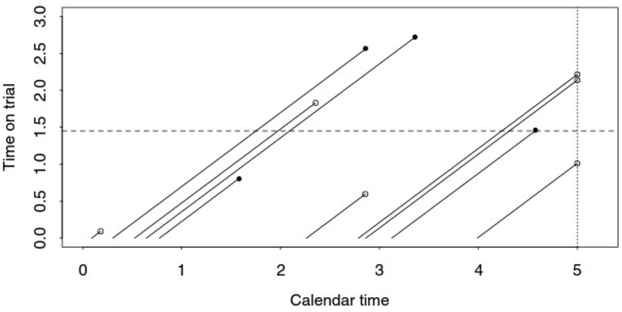


Figure 5.8 Lexis diagram showing typical pattern of censoring in a medical study. Each individual is shown as a line whose x coordinates run from the calendar time of entry to the trial to the calendar time of failure (blob) or censoring (circle). Censoring occurs at the end of the trial, marked by the 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 serious loss of efficiency if they are ignored (Example 4.20). There will also be bias

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

- for inference, $\ell_{part}(\beta)$ has usual properties
  1. $\hat{\beta}_{part} \,\dot{\sim}\, N\{\beta, j_{part}^{-1}(\hat{\beta})\}$,
  2. $2\{\ell_{part}(\hat{\beta}_{part}) - \ell_{part}(\beta)\} \,\dot{\sim}\, \chi_d^2$ Davison §10.8; Cox 1972, 1975

… semi-parametric models

- partial log-likelihood function

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

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

- 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

vdW Ch. 25.10

- $\ell(\beta, \lambda; y), \beta \in \mathbb{R}^d; \lambda = \lambda(\cdot)$
- $\ell_{\mathrm{p}}(\beta; y) = \ell(\beta, \tilde{\lambda}_\beta; y); \quad \tilde{\lambda}_\beta = \arg\sup_\lambda \ell(\beta, \lambda; y)$
- 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

  also Davison SM §10.8

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

... semi-parametric models

- 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

Example: missing covariate

- 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. $\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \tilde{\imath}^{-1}(\theta_0)\tilde{U}(\theta_0) + o_p(1)$
   - $\tilde{U}(\theta_0) = \frac{\partial \ell(\theta, \lambda)}{\partial \theta} - \text{Proj}_g \frac{\partial \ell(\theta, \lambda)}{\partial \theta}$
   - projection of $\partial \ell_\theta$ onto the closed linear span of the score functions for $\lambda(\cdot)$
   - $\tilde{\imath}(\theta_0) = \text{var}\{\tilde{U}_j(\theta_0)\}$ $\qquad$ $\tilde{U} = \sum \tilde{U}_j$; $\imath$ is $O(1)$
2. $\ell_p(\hat{\theta}) = \ell_p(\theta_0) + \frac{1}{2} n(\hat{\theta} - \theta_0)^{\mathrm{T}} \tilde{\imath}(\theta_0)(\hat{\theta} - \theta_0) + o_p(1)$
3. for any random sequence $\tilde{\theta}_n \xrightarrow{p} \theta_0$, plus conditions on the model,

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

- $$\ell_{\mathrm{p}}(\tilde{\theta}_n) = \ell_{\mathrm{p}}(\theta_0) + (\tilde{\theta}_n - \theta_0)^{\mathrm{T}} \sum_{j=1}^{n} \tilde{U}_j(\theta_0) - \frac{1}{2} n(\tilde{\theta}_n - \theta_0)^{\mathrm{T}} \tilde{\imath}^{-1}(\theta_0)(\tilde{\theta}_n - \theta_0) + o_p(\sqrt{n}||\tilde{\theta}_n - \theta_0|| + 1)^2$$
- 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, ...