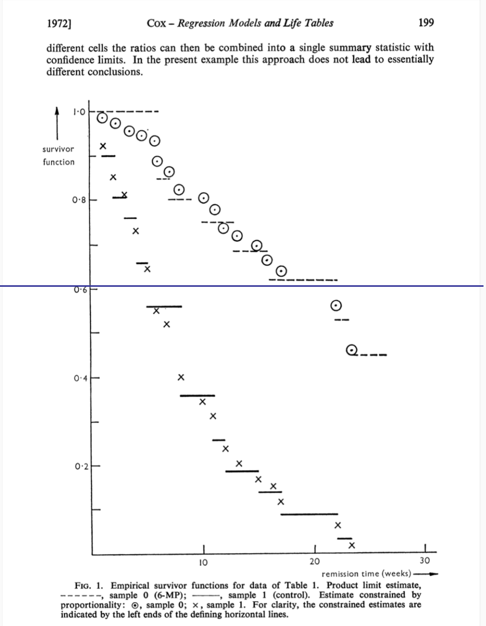
Topics in Likelihood Inference

STA4508H

Nancy Reid

University of Toronto

January 26, 2022



different cells the ratios can then be combined into a single summary statistic with confidence limits. In the present example this approach does not lead to essentially different conclusions.

FIG. 1. Empirical survivor functions for data of Table 1. Product limit estimate, ------, sample 0 (6-MP); ———, sample 1 (control). Estimate constrained by proportionality: ⊙, sample 0; ×, sample 1. For clarity, the constrained estimates are indicated by the left ends of the defining horizontal lines.

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

- presentations and report Publications of D R Cox
- Suggestions: 46 61 72 93 118 133 158 217 232 260 269 320 332 357 371 378

371. Cox, D.R. and Battey, H. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Nat. Acad. Sci.* **114**, 8592-8595.
372. Cox, D.R. and Efron, B. (2017). Statistical thinking for 21st scientists. Science Progress 3, c1700768.
373. Cox, D.R., Kartsonaki, C. and Keogh, R.H. (2018). Big data: some statistical issues. *Statistics and Probability Letters*, **136**, 111-115.
374. Battey, H. and Cox, D.R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. Roy. Soc. A*, **474**, 20170631.
375. Granchelli, A.M., Adler, F.R., Keogh, R.H., Kartsonaki, C., Cox, D.R. and Liou, T.G. (2018). Microbial interactions in the cystic fibrosis airway. *J. Clinical Mic*
376. Battey, H.S., Cox, D.R. and Jackson, M. (2019). On the linear in probability model for binary data. *Royal Society Open Science*, **6**, 190067.
377. Cox, D.R. and Kartsonaki, C. (2019). On the analysis of large numbers of p-values. *International Statistical Review*, **87**, 505-513.
378. Cox, D.R. (2020). Statistical significance. *Annual Review of Statistics and its Applicatons*, **7**, 1-10.
379. Kartsonaki, C. and Cox, D.R. (2020). Matched pairs with binary outcomes . *REVSTAT*, **18**, 581-592.
380. Cox, D.R., Kartsonaki, C. and Keogh, R.H. (2020). Statistical Science: Some Current Challenges. *Harvard Data Science Review*, **2.3**.
381. Battey, H.S. and Cox, D.R. (2020). High dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, **3**, 119-148.
382. Kartsonaki, C. and Cox, D.R. (2021). Regression reconstruction from a retrospective sample. *Econometrics and Statistics*, to appear.
383. Battey, H.S. and Cox, D.R. (2021). Some perspectives on inference in high dimensions. *Statistical Science*, to appear.
384. Battey, H.S. and Cox, D.R. (2021). Some aspects of non-standard multivariate analysis. *J. Multivariate Analysis*, to appear.

1. (Knight, 2000 Ch. 5.6; Owen, 1988). Suppose $Y_1, \dots, Y_n$ are independent and identically distributed from an unknown distribution function F. To estimate F we restrict attention to distributions putting positive probability mass only at the points $Y_1, \dots, Y_n$, assumed distinct. Knight defines the non-parametric log-likelihood function for $F(\cdot)$ as

$$L(p_1, \dots, p_n) = \sum_{i=1}^{n} \log(p_i), \quad p_i \geq 0, \Sigma p_i = 1,$$

where p_i is the probability mass at Y_i.

   (a) Show that $L(p)$ (or equivalently $\ell(p) = \log L(p)$) is maximized at $\hat{p}_i = 1/n$.

   (b) Suppose that $\mu = E(Y_i) = \int y dF(y)$ is the parameter of interest, with $F(\cdot)$ as a nuisance parameter. The profile likelihood is obtained by maximizing

$$L(p_1, \dots, p_n), \text{ subject to } p_i \geq 0, \Sigma p_i = 1, \Sigma p_i Y_i = \mu,$$

   where there is now an additional constraint on the vector p. Show that the solution to the maximization problem is given by

$$\begin{aligned} \hat{p}_i(\mu) &= \frac{1}{n} \frac{1}{1 + \lambda(Y_i - \mu)}, \text{ where } \lambda \text{ solves} \\ 0 &= \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i - \mu}{1 + \lambda(Y_i - \mu)}. \end{aligned}$$

2. Choose a paper for your report and presentation, and provide the complete citation and a one-sentence description of the paper.

   You should plan for a 15 minute presentation followed by 5 minutes of questions. The presentation can be either on slides or presented live on a tablet/ipad. My guideline for number of slides is one per minute.

Nuisance parameters

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

... Nuisance parameters

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

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

$$\sqrt{n}(\hat{\psi} - \psi) \doteq \frac{1}{\sqrt{n}} (i_1^{\psi\psi})^{-1} (U_\psi - i_{\psi\lambda} i_{\lambda\lambda}^{-1} U_\lambda)$$

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

$$2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\} \doteq (\hat{\psi} - \psi)^{\mathrm{T}} i^{\psi\psi} (\hat{\psi} - \psi)$$

$$2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\} \xrightarrow{d} \chi_q^2$$

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

column vectors

... Nuisance parameters

$$
\begin{aligned}
w_u(\psi) = U_\psi(\psi,\hat\lambda_\psi)^T\{i^{\psi\psi}(\psi,\hat\lambda_\psi)\}U_\psi(\psi,\hat\lambda_\psi) &\quad \dot\sim \quad \chi^2_q \\
w_e(\psi) = (\hat\psi-\psi)\{i^{\psi\psi}(\hat\psi,\hat\lambda)\}^{-1}(\hat\psi-\psi) &\quad \dot\sim \quad \chi^2_q \\
w(\psi) = 2\{\ell(\hat\psi,\hat\lambda)-\ell(\psi,\hat\lambda_\psi)\} = 2\{\ell_{\mathrm{P}}(\hat\psi)-\ell_{\mathrm{P}}(\psi)\} &\quad \dot\sim \quad \chi^2_q;
\end{aligned}
$$

Approximate Pivots, $q = 1$

$$
\begin{aligned}
r_u(\psi) &= \ell_{\mathrm{P}}'(\psi)j_{\mathrm{P}}(\hat\psi)^{-1/2} \;\dot\sim\; N(0,1), \\
r_e(\psi) &= (\hat\psi-\psi)j_{\mathrm{P}}(\hat\psi)^{1/2} \;\dot\sim\; N(0,1), \\
r(\psi) &= \mathrm{sign}(\hat\psi-\psi)[2\{\ell_{\mathrm{P}}(\hat\psi)-\ell_{\mathrm{P}}(\psi)\}]^{1/2} \;\dot\sim\; N(0,1)
\end{aligned}
$$

Profile LRT

$$w(\psi) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\} = 2\{\ell_{\mathrm{P}}(\hat{\psi}) - \ell_{\mathrm{P}}(\psi)\} \,\dot{\sim}\, \chi^2_q$$

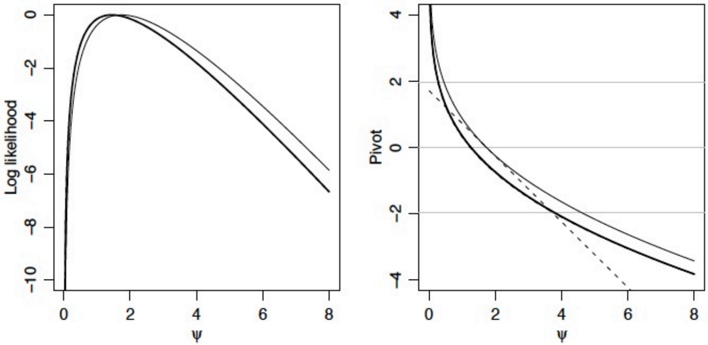


Figure 2.3: Inference for shape parameter ψ of gamma sample of size $n = 5$. Left: profile log likelihood ℓ_{p} (solid) and the log likelihood from the conditional density of u given v (heavy). Right: likelihood root $r(\psi)$ (solid), Wald pivot $t(\psi)$ (dashes), modified likelihood root $r^*(\psi)$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

Approximate Bayesian inference

- $\pi(\theta \mid y) = \dfrac{\exp\{\ell(\theta; y)\}\pi(\theta)}{\int \exp\{\ell(\theta; y)\}\pi(\theta)d\theta}$
- expand numerator and denominator about $\hat{\theta}$, assuming $\ell'(\hat{\theta}) = 0$
- $\pi(\theta \mid y) \doteq N\{\hat{\theta}, j^{-1}(\hat{\theta})\}$

... Approximate Bayesian inference

- $\pi(\theta \mid y) = \dfrac{\exp\{\ell(\theta; y)\}\pi(\theta)}{\int \exp\{\ell(\theta; y)\}\pi(\theta)d\theta}$
- expand numerator and denominator about $\hat{\theta}$, assuming $\ell'(\hat{\theta}) = 0$
- $\pi(\theta \mid y) \doteq N\{\hat{\theta}, j^{-1}(\hat{\theta})\}$

"data swamps the prior"

Posterior is asymptotically normal

$$\pi(\theta \mid y) \dot\sim N\{\hat\theta, j^{-1}(\hat\theta)\} \quad \theta \in \mathbb{R}, y = (y_1, \dots, y_n)$$

careful statement

Berger, Ch.4; Walker, 1969

For any $a, b \in \mathbb{R}, a < b$, let $a_n = a_n(y) = \hat\theta_n + aj^{-1/2}(\hat\theta_n)$, $b_n = b_n(y) = \hat\theta_n + bj^{-1/2}(\hat\theta_n)$, where $\hat\theta_n$ is the solution of $\ell'(\theta; y) = 0$, assumed unique, and $j(\theta) = -\ell''(\theta; y)$. Then

$$\int_{a_n}^{b_n} \pi(\theta \mid y) \longrightarrow \Phi(b) - \Phi(a), \quad n \to \infty.$$

... posterior is asymptotically normal

$$\pi(\theta \mid y) \dot{\sim} N\{\hat{\theta}, j^{-1}(\hat{\theta})\} \quad \theta \in \mathbb{R}, y = (y_1, \ldots, y_n)$$

equivalently $\pi(\theta \mid y) \doteq N\{\hat{\theta}_\pi, j_\pi^{-1}(\hat{\theta}_\pi)\}$

$\hat{\theta}_\pi$ solves $h'(\theta) = 0$; $h(\theta) = \ell(\theta) + \log \pi(\theta)$

$\hat{\theta} = \hat{\theta}_\pi + O_p(n^{-1})$

… posterior is asymptotically normal

In fact,

If $\pi(\theta) > 0$ and $\pi'(\theta)$ is continuous in a neighbourhood of θ_0, there exist constants D and n_y s.t.

$$|F_n(\xi) - \Phi(\xi)| < Dn^{-1/2}, \quad \text{for all } n > n_y,$$

on an almost-sure set with respect to the joint distribution of y, θ at θ_0, i.e. $y = (y_1, \ldots, y_n)$ is a sample from $f(y; \theta_0)$, and θ_0 is fixed value drawn from the prior density $\pi(\theta)$.

$$F_n(\xi) = \Pr\{(\theta - \hat{\theta})j^{1/2}(\hat{\theta}) \leq \xi \mid y\}$$

Johnson (1970); Datta & Mukerjee (2004)

Laplace approximation

- expand denominator only about $\hat{\theta}$

Laplace approximation

- expand denominator only about $\hat{\theta}$

- result

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

$$\pi(\theta \mid y) = \frac{1}{(2\pi)^{1/2}} |j(\hat{\theta})|^{+1/2} \exp\{\ell(\theta; y) - \ell(\hat{\theta}; y)\} \frac{\pi(\theta)}{\pi(\hat{\theta})} \{1 + O_p(n^{-1})\}$$

$$y = (y_1, \dots, y_n), \quad \theta \in \mathbb{R}^1$$

$$\pi(\theta \mid y) = \frac{1}{(2\pi)^{1/2}} |j_\pi(\hat{\theta}_\pi)|^{+1/2} \exp\{\ell_\pi(\theta; y) - \ell_\pi(\hat{\theta}_\pi; y)\} \{1 + O_p(n^{-1})\}$$

Marginal posterior

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

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

... 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, $d = 1$

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

$$\vdots$$

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

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

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

Eliminating nuisance parameters: nonBayesian

- Profile likelihood poor if q large; fails if $q \to \infty$
- alternative: marginal likelihood: $f(\underline{y}_n; \psi, \lambda) \propto f_{\mathrm{m}}(t_1; \psi) f_{\mathrm{c}}(t_2 \mid t_1; \psi, \lambda)$ $\qquad t_j = t_j(\underline{y})$
- Example $N(X\beta, \sigma^2 I)$: $\quad f(y; \beta, \sigma^2) \propto f_{\mathrm{m}}(RSS; \sigma^2) f_{\mathrm{c}}(\hat{\beta} \mid RSS; \beta, \sigma^2)$

$$L_m(\sigma^2) \propto f_m(RSS; \sigma^2)$$

- alternative conditional likelihood: $f(\underline{y}; \psi, \lambda) \propto f_{\mathrm{c}}(t_1 \mid t_2; \psi) f_{\mathrm{m}}(t_2; \psi, \lambda)$
- Example 2×2 tables: $f(\underline{y}; \psi, \underline{\lambda}) \propto \prod_{i=1}^{n} f_c(y_{i1} \mid y_{i1} + y_{i2}; \psi) f_m(y_{i1} + y_{i2}; \psi, \lambda_i)$

$$L_c(\psi) = \prod f_c(y_{i1} \mid y_{i1} + y_{i2}; \psi)$$

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

Logistic regression

- $y_i \sim Binom(m_i, p_i), i = 1, \ldots, n$
- $\log\{p_i/(1 - p_i)\} = x_i^{\mathrm{T}}\beta$
- $f(y; \beta) = \exp\{\beta_1 \Sigma(x_{i1} y_i) + \cdots + \beta_p \Sigma(x_{ip} y_i) - \Sigma m_i \log(1 + e^{x_i^{\mathrm{T}}\beta})\}$
- $f_c(s_5 \mid s_{-(5)}; \beta_5) \propto \exp\{\beta_5 s_5 - \tilde{k}(\beta_5)\} h(s)$

… Logistic regression

4.2. URINE DATA 57

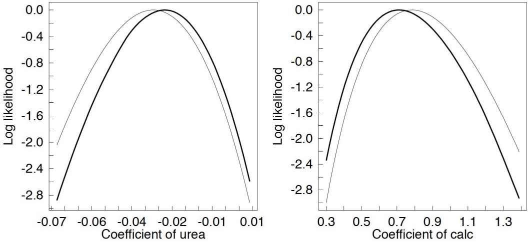


Figure 4.2: Comparison of log likelihoods for the urine data: profile log likelihood (solid line), approximate conditional log likelihood (bold line). The variables of interest are urea (left panel) and calcium concentration (right panel). The graphical output is obtained with the `plot` method of the `cond` package.

$$f_c(s_5 \mid s_{-(5)}; \beta_5) \propto \exp\{\beta_5 s_5 - \tilde{k}(\beta_5)\} h(s)$$

Summary 4.1 Approximate conditional inference for the urine data.

```
> urine.glm <- glm( formula=r~I(100*(gravity-1))+ph+osmo+conduct+urea+calc,
+                   family=binomial, data=urine )

> summary(urine.glm)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.60609    3.79582   0.160  0.87314
I(100 * (gravity - 1))  3.55944    2.22110   1.603  0.10903
ph                     -0.49570    0.56976  -0.870  0.38429
osmo                    0.01681    0.01782   0.944  0.34536
conduct                -0.43282    0.25123  -1.723  0.08493 .
urea                   -0.03201    0.01612  -1.986  0.04703 *
calc                    0.78369    0.24216   3.236  0.00121 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 105.17  on 76  degrees of freedom
Residual deviance:  57.56  on 70  degrees of freedom
AIC: 71.56

> urine.cond.urea <- cond( urine.glm, offset=urea )

> coef( urine.cond.urea )
           Estimate   Std. Error
uncond.  -0.03201315   0.01611884
cond.    -0.02759202   0.01489919

> summary( urine.cond.urea, coef=F )

Confidence intervals
--------------------
 level = 95 %
                                           lower two-sided      upper
Wald pivot                                -0.06361          -0.0004208
```

Summary 4.1 Approximate conditional inference for the urine data (cont.).

```
> urine.cond.calc <- cond( urine.glm, offset=calc )

> coef( urine.cond.calc )
         Estimate   Std. Error
uncond.  0.7836913   0.2421638
cond.    0.7110584   0.2282501

> summary( urine.cond.calc, coef=F )

Confidence intervals
 level = 95 %
                                     lower two-sided upper
Wald pivot                           0.3091           1.258
Wald pivot (cond. MLE)               0.2637           1.158
Likelihood root                      0.3815           1.342
Modified likelihood root             0.3193           1.213
Modified likelihood root (cont. corr.) 0.3044         1.254

Diagnostics:
-----------
    INF      NP
0.08451 0.32878
```

Marginal and conditional likelihoods

$$L_c(\psi) = \log f_c\{s(y) \mid t(y); \psi\},$$
$$L_m(\psi) = \log f_m\{s(y); \psi\}$$

- Inference based on usual asymptotics applies, under regularity conditions on $f(y; \psi, \lambda)$
- likelihoods based on observable random variables
- Bartlett identities apply directly
- use conditional or marginal Fisher information, etc.

- might lose information in other component

$$f(y; \psi, \lambda) \propto f_m(s; \psi) f_c(t \mid s; \psi, \lambda)$$

- marginal likelihoods associated with transformation models

REML

Approximate conditional 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 $A_{FR}(\psi) = +\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| - \log|\frac{d(\lambda)}{d\hat\lambda_\psi}|$ Fraser 03
- closely related $A_{BN}(\psi) = -\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| + \log|\frac{d\hat\lambda}{d\hat\lambda_\psi}|$ 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 $0 < y_1 < \cdots < y_n < \tau$
- likelihood function

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

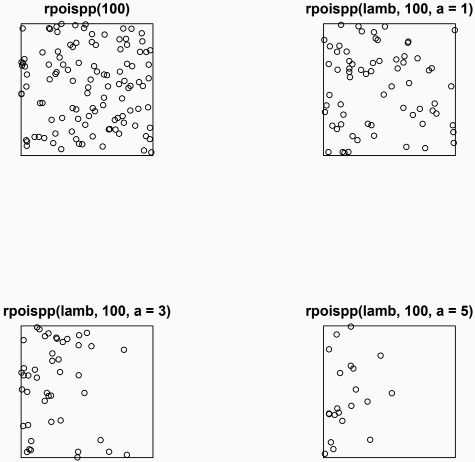
- log-likelihood function

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

- in space:

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

SM §6.5.1



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

Survival data

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

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

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

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

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

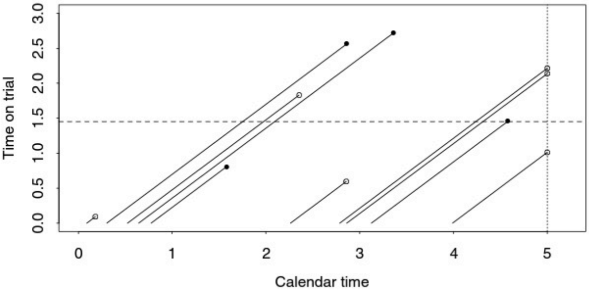


Figure 5.8 Lexis diagram showing typical pattern of censoring in a medical study. Each individual is shown as a line whose x coordinates run from the calendar time of entry to the trial to the calendar time of failure (blob) or censoring (circle). Censoring occurs at the end of the trial, marked by the 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

- 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.
- for inference, $\ell_{part}(\beta)$ has usual properties
  1. $\hat\beta_{part} \dot\sim N\{\beta, j_{part}^{-1}(\hat\beta)\}$, $2\{\ell_{part}(\hat\beta_{part}) - \ell_{part}(\beta)\} \dot\sim \chi_d^2$ 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 $(\lambda_1, \ldots, \lambda_n) = \{\lambda(y_1), \ldots \lambda(y_n)\}$
- 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

- $\ell(\beta, \lambda; y), \beta \in \mathbb{R}^d; \lambda = \lambda(\cdot)$
- $\ell_{\rm p}(\beta; y) = \ell(\beta, \tilde\lambda_\beta; y); \quad \tilde\lambda_\beta = \arg\sup_\lambda \ell(\beta, \lambda; y)$
- 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 $\Sigma_{j:y_j \geq y}\lambda_j$”

... 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

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

- 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

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

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

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

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

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 $\lambda(\cdot)$
    - $\tilde\imath(\theta_0) = \text{var}\{\tilde U_j(\theta_0)\}$

$\tilde U = \sum \tilde U_j$; $\imath$ 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,

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

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

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

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

Proportional hazards model

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