Last weeks

- likelihood
- marginal and conditional likelihood
- profile likelihood
- adjusted profile likelihood
- composite likelihood

This week

- semiparametric likelihoods
- nonparametric likelihoods
- consistency of maximum likelihood estimators
- comments on problem sets

Survival Data: single sample

- Model: $f(t), h(t), 1 - F(t), H(t)$
  density, hazard, survivor function, cumulative hazard
- Data: $(t_1, \delta_1), \ldots, (t_n, \delta_n)$
  - t_i an observed time
  - $\delta_i = 1$ if t_i a true failure time; 0 if t_i is a [illegible]
- random censorship assumption
- parametric inference:

$$\ell(\theta; t, \delta) = \sum_{i=1}^{n} \delta_i \log h(t_i; \theta) - H(t_i; \theta)$$

- examples:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

$$H(t) = \int^{t} h(u)\,du$$

Survival Data: single sample

- Model: $f(t), h(t), 1 - F(t), H(t)$
  density, hazard, survivor function, cumulative hazard
- Data: $(t_1, \delta_1), \ldots, (t_n, \delta_n)$
  - t_i an observed time
  - $\delta_i = 1$ if t_i a true failure time, 0 if t_i is a censoring time
- random censorship assumption
- parametric inference:

$$L(\theta; \underline{t}, \underline{\delta}) = \sum_{i=1}^{n} \delta_i \log h(t_i; \theta) - H(t_i; \theta)$$

- examples:

Survival Data: single sample

- Model: $f(t), h(t), 1 - F(t), H(t)$
  density, hazard, survivor function, cumulative hazard
- Data: $(t_1, \delta_1), \ldots, (t_n, \delta_n)$
  - t_i an observed time
  - $\delta_i = 1$ if t_i a true failure time, 0 if t_i is a censoring time
- random censorship assumption
- parametric inference:

$$L(\theta; \underline{t}, \underline{\delta}) = \sum_{i=1}^{n} \delta_i \log h(t_i; \theta) - H(t_i; \theta)$$

$$L \quad \prod_{i=1}^{n} f(t_i; \theta)^{\delta_i} \{1 - F(t_i; \theta)\}^{1-\delta_i}$$

Parametric regression models

- Data: $(t_i, \delta_i, \underline{x}_i), \ldots, i = 1, \ldots, n$

- Likelihood function:

$$L(\theta; \underline{t}, \underline{\delta}) = \sum_{i=1}^{n} \delta_i \log h(t_i; \theta) - H(t_i; \theta)$$

- Example: Exponential distribution
  - $h(t; \beta) = \exp(x_i^T \beta)$, for example
  - $\ell(\beta) = \sum_{i=1}^{n} \delta_i x_i^T \beta - \exp(x_i^T \beta) t_i$
  - usual maximum likelihood theory applies

- Example: Weibull distribution

Parametric regression models

- Data: $(t_i, \delta_i, \underline{x}_i), \ldots, i = 1, \ldots, n$

- Likelihood function:

$$L(\theta; \underline{t}, \underline{\delta}) = \sum_{i=1}^{n} \delta_i \log h(t_i; \theta) - H(t_i; \theta)$$

- Example: Exponential distribution
  - $h(t; \beta) = \exp(x_i^T \beta)$, for example
  - $\ell(\beta) = \sum_{i=1}^{n} \delta_i x_i^T \beta - \exp(x_i^T \beta) t_i$
  - usual maximum likelihood theory applies

- Example: Weibull distribution

Parametric regression models

- Data: $(t_i, \delta_i, \underline{x}_i), \ldots, i = 1, \ldots, n$

- Likelihood function:

$$L(\theta; \underline{t}, \underline{\delta}) = \sum_{i=1}^{n} \delta_i \log h(t_i; \theta) - H(t_i; \theta)$$

- Example: Exponential distribution
  - $h(t; \beta) = \exp(x_i^T \beta)$, for example
  - $\ell(\beta) = \sum_{i=1}^{n} \delta_i x_i^T \beta - \exp(x_i^T \beta) t_i$
  - usual maximum likelihood theory applies

- Example: Weibull distribution
  - $h(t; \theta) = h(t; \beta, \alpha) = \exp(x_i^T \beta) t^{\alpha}$
  - $\theta = (\beta, \alpha)$
  - usual maximum likelihood theory applies

Semi-parametric regression models

- proportional hazards model:

$$h(t; x, \beta) = h_0(t)\exp(x^T\beta)$$

- $h_0(t)$ unknown
-

$$\frac{h(t;x)}{h(t;0)} = \exp(x^T\beta), \quad \text{does not depend on } t$$

-

$$1 - F(t;x) = \{1 - F_0(t)\}^{\exp(x^T\beta)}$$

- survivor functions can never cross
- $x^T\beta = x_1\beta_1 + \cdots + x_p\beta_p$: no constant term

Cox, 1972; SM, §10.8

Semi-parametric regression models

- proportional hazards model:

$$h(t; x, \beta) = h_0(t) \exp(x^T \beta)$$

- $h_0(t)$ unknown
-

$$\frac{h(t; x)}{h(t; 0)} = \exp(x^T \beta), \quad \text{does not depend on } t$$

-

$$1 - F(t; x) = \{1 - F_0(t)\}^{\exp(x^T \beta)}$$

- survivor functions can never cross
- $x^T \beta = x_1 \beta_1 + \cdots + x_p \beta_p$: no constant term

Cox, 1972; SM, §10.8

Semi-parametric regression models

- proportional hazards model:

$$h(t; x, \beta) = h_0(t) \exp(x^T \beta)$$

- $h_0(t)$ unknown
- $$\frac{h(t; x)}{h(t; 0)} = \exp(x^T \beta), \quad \text{does not depend on } t$$
- $$1 - F(t; x) = \{1 - F_0(t)\}^{\exp(x^T \beta)}$$
- survivor functions can never cross
- $x^T \beta = x_1 \beta_1 + \cdots + x_p \beta_p$, no constant term

Cox, 1972; SM, §10.8

Semi-parametric regression models

- proportional hazards model:

$$h(t; x, \beta) = h_0(t) \exp(x^T \beta)$$

- $h_0(t)$ unknown
- $$\frac{h(t; x)}{h(t; 0)} = \exp(x^T \beta), \quad \text{does not depend on } t$$
- $$1 - F(t; x) = \{1 - F_0(t)\}^{\exp(x^T \beta)}$$
- survivor functions can never cross
- $x^T \beta = x_1 \beta_1 + \cdots + x_p \beta_p$, no constant term

Cox, 1972; SM, §10.8

Estimation of β

- partial likelihood

$$L_{part}(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(x_i^T\beta)}{\sum_{k \in \mathcal{R}_i} \exp(x_k^T\beta)} \right)^{\delta_i}$$

- $\mathcal{R}_i$ risk set at time t_i^-; number of units with $t_k \geq t_i$

- derived in SM §10.8 as approximately a profile likelihood ($h_0(\cdot)$ maximized out)

- $\hat{\beta}$ estimated by maximizing partial log-likelihood $\ell_{part}(\beta) = \log L_{part}(\beta)$
- estimated standard error from $-\ell''_{part}(\hat{\beta})$

Estimation of β

- partial likelihood

$$L_{part}(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(x_i^T \beta)}{\sum_{k \in \mathcal{R}_i} \exp(x_k^T \beta)} \right)^{\delta_i}$$

- $\mathcal{R}_i$ risk set at time t_i^-; number of units with $t_k \geq t_i$

- derived in SM §10.8 as approximately a profile likelihood ($h_0(\cdot)$ maximized out)

- $\hat{\beta}$ estimated by maximizing partial log-likelihood $\ell_{part}(\beta) = \log L_{part}(\beta)$
- estimated standard error from $-\ell''_{part}(\hat{\beta})$ part

... partial likelihood

- usual asymptotic theory applies:

$$\hat{\beta}_{part} \dot{\sim} N[\beta, \{-\ell''_{part}(\hat{\beta}_{part})\}^{-1}]$$

- special property of this model: components of the score vector are uncorrelated
- no need to compute analogue of Godambe information
- there could be loss of efficiency in estimating β; this loss has been shown to be small in a wide range of settings
- general treatment of likelihood inference for semi-parametric models Murphy and van der Waart, 2000
- this model is particularly easy to handle

Cox, 1975; 2006, §7.6.5

... partial likelihood

- usual asymptotic theory applies:

$$\hat{\beta}_{part} \dot{\sim} N[\beta, \{-\ell''_{part}(\hat{\beta}_{part})\}^{-1}]$$

- special property of this model: components of the score vector are uncorrelated
- no need to compute analogue of Godambe information
- there could be loss of efficiency in estimating β; this loss has been shown to be small in a wide range of settings
- general treatment of likelihood inference for semi-parametric models Murphy and van der Waart, 2000
- this model is particularly easy to handle

Cox, 1975; 2006, §7.6.5

Semi-parametric regression models

- for example, $E(y_i) = \mu_i(\theta) = x_i^T\beta + m(t_i), \quad \text{Var}(y_i) = \sigma^2$
- $m(\cdot)$ a 'smooth' function of covariates t

- least squares

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2$$

- without constraint on $m(\cdot)$, minimum will be 0, thus

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2 - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$

- equivalent to

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2 + \lambda m^T K m$$

for suitable $n \times n$ matrix K

Semi-parametric regression models

- for example, $E(y_i) = \mu_i(\theta) = x_i^T\beta + m(t_i), \quad \text{Var}(y_i) = \sigma^2$
- $m(\cdot)$ a 'smooth' function of covariates t

- least squares

$$\min_{\beta,m(\cdot)} \sum_{i=1}^{n}\{y_i - x_i^T\beta - m(t_i)\}^2$$

- without constraint on $m(\cdot)$, minimum will be 0, thus

$$\min_{\beta,m(\cdot)} \sum_{i=1}^{n}\{y_i - x_i^T\beta - m(t_i)\}^2 - \frac{1}{2}\lambda\int\{m''(t)\}^2 dt$$

- equivalent to

$$\min_{\beta,m(\cdot)} \sum_{i=1}^{n}\{y_i - x_i^T\beta - m(t_i)\}^2 + \lambda m^T K m$$

for suitable $n \times n$ matrix K

Semi-parametric regression models

- for example, $E(y_i) = \mu_i(\theta) = x_i^T\beta + m(t_i), \quad \text{Var}(y_i) = \sigma^2$
- $m(\cdot)$ a 'smooth' function of covariates t

- least squares

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2$$

- without constraint on $m(\cdot)$, minimum will be 0, thus

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2 - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$

- equivalent to

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2 + \lambda m^T K m$$

for suitable $n \times n$ matrix K

Semi-parametric regression models

- for example, $E(y_i) = \mu_i(\theta) = x_i^T\beta + m(t_i), \quad \text{Var}(y_i) = \sigma^2$
- $m(\cdot)$ a 'smooth' function of covariates t

- least squares

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2$$

- without constraint on $m(\cdot)$, minimum will be 0, thus

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2 - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$

- equivalent to

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T\beta - m(t_i)\}^2 + \lambda m^T K m$$

for suitable $n \times n$ matrix K

... semi-parametric regression models

- $$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T \beta - m(t_i)\}^2 - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$
- extend to generalized linear model

$$h\{\mathrm{E}(y_i)\} = x_i^T \beta + m(t_i) = \eta_i$$

- penalized log-likelihood

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \ell_i(\eta_i) - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$

Green, 1987; Green & Silverman, 1994; SM, §10.7

... semi-parametric regression models

- $$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \{y_i - x_i^T \beta - m(t_i)\}^2 - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$
- extend to generalized linear model

$$h\{\mathrm{E}(y_i)\} = x_i^T \beta + m(t_i) = \eta_i$$

- penalized log-likelihood

$$\min_{\beta, m(\cdot)} \sum_{i=1}^{n} \ell_i(\eta_i) - \frac{1}{2}\lambda \int \{m''(t)\}^2 dt$$

Green, 1987; Green & Silverman, 1994; SM, §10.7

Nonparametric likelihood

- likelihood functions for infinite-dimensional parameters can be tricky
  pause
- for example, given $y_1, \ldots, y_n$ i.i.d. with distribution function $F(\cdot)$ and density function $f(\cdot)$
- the nonparametric maximum likelihood estimator of $F(\cdot)$ is

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq t), \quad t \in \mathbb{R}$$

- this is a cumulative distribution function, although discrete
- the nonparametric maximum likelihood estimator of $f(\cdot)$ is not a density function
- unless we put some constraints on the class of densities over which we maximize
- for example, might require $f(x)$ to be log concave: $f(x) = \exp\{\eta(x)\}, \eta$ concave Balabdaoui et al, 2009

Nonparametric likelihood

- likelihood functions for infinite-dimensional parameters can be tricky
  pause
- for example, given $y_1, \ldots, y_n$ i.i.d. with distribution function $F(\cdot)$ and density function $f(\cdot)$
- the nonparametric maximum likelihood estimator of $F(\cdot)$ is

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \le t), \quad t \in \mathbb{R}$$

- this is a cumulative distribution function, although discrete
- the nonparametric maximum likelihood estimator of $f(\cdot)$ is not a density function
- unless we put some constraints on the class of densities over which we maximize
- for example, might require $f(x)$ to be log concave: $f(x) = \exp\{\eta(x)\}, \eta$ concave Balabdaoui et al, 2009

Nonparametric likelihood

- likelihood functions for infinite-dimensional parameters can be tricky
  pause
- for example, given $y_1, \ldots, y_n$ i.i.d. with distribution function $F(\cdot)$ and density function $f(\cdot)$
- the nonparametric maximum likelihood estimator of $F(\cdot)$ is

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq t), \quad t \in \mathbb{R}$$

- this is a cumulative distribution function, although discrete
- the nonparametric maximum likelihood estimator of $f(\cdot)$ is not a density function
- unless we put some constraints on the class of densities over which we maximize
- for example, might require $f(x)$ to be log concave: $f(x) = \exp\{\eta(x)\}, \eta$ concave

Balabdaoui et al, 2009

Empirical likelihood

- $y_1, \ldots, y_n$ i.i.d. with distribution function $F_0(\cdot)$
- define

$$L(F) = \prod_{i=1}^{n} \{F(y_i) - F(y_i^-)\}$$

- maximized at F_n, empirical c.d.f.
- empirical likelihood ratio

$$R(F) = \frac{L(F)}{L(F_n)}$$

- suppose $T(F_0)$ is a function of interest, e.g. $\mu = \int x dF_0(x)$
- maximizing $R(F)$, subject to μ fixed, is equivalent to

$$\max_{w_1,\ldots,w_n} \prod_{i=1}^{n} w_i, \text{ subject to } \sum_{i=1}^{n} w_i y_i = \mu, \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall i$$

Owen, 1988; 2001

Empirical likelihood

- $y_1, \ldots, y_n$ i.i.d. with distribution function $F_0(\cdot)$
- define

$$L(F) = \prod_{i=1}^{n} \{F(y_i) - F(y_i^-)\}$$

- maximized at F_n, empirical c.d.f.
- empirical likelihood ratio

$$R(F) = \frac{L(F)}{L(F_n)}$$

- suppose $T(F_0)$ is a function of interest, e.g. $\mu = \int x dF_0(x)$
- maximizing $R(F)$, subject to μ fixed, is equivalent to

$$\max_{w_1,\ldots,w_n} \prod_{i=1}^{n} w_i, \text{ subject to } \sum_{i=1}^{n} w_i y_i = \mu, \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall i$$

Owen, 1988; 2001

... empirical likelihood

- $$\max_{w_1,\ldots,w_n} \prod_{i=1}^{n} w_i, \text{ subject to } \sum_{i=1}^{n} w_i y_i = \mu, \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall i$$

- likelihood ratio confidence intervals are valid

$$-2 \log R(F_0) \xrightarrow{\mathcal{L}} \chi^2_1, \quad n \to \infty$$

- parameter of interest, $\mu \in \mathbb{R}$
- nuisance parameter $w = (w_1, \ldots, w_n)$
- generalized to many more complex situations

Hjort et al. 2009

... empirical likelihood

- $$\max_{w_1,\ldots,w_n} \prod_{i=1}^{n} w_i, \text{ subject to } \sum_{i=1}^{n} w_i y_i = \mu, \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall i$$
- likelihood ratio confidence intervals are valid

$$-2 \log R(F_0) \xrightarrow{\mathcal{L}} \chi_1^2, \quad n \to \infty$$

- parameter of interest, $\mu \in \mathbb{R}$
- nuisance parameter $w = (w_1, \ldots, w_n)$
- generalized to many more complex situations

Hjort et al. 2009

Those pesky regularity conditions

- two proofs of the consistency of the maximum likelihood estimator
- Wald, 1949 – the log-likelihood is maximized in expectation at the true value; apply Jensen's inequality to conclude $\hat{\theta}$ must converge to the true value
- requires the parameter space to be compact

- Cramer, 1946 – there exist solutions to the score equation that are consistent
- Taylor series expansion of $\log f(y;\theta)$
- if the likelihood function is maximized in the interior of the parameter space, the m.l.e. is one of these solutions
- if the score equation has only one root, the m.l.e. is consistent

Those pesky regularity conditions

- two proofs of the consistency of the maximum likelihood estimator
- Wald, 1949 – the log-likelihood is maximized in expectation at the true value; apply Jensen's inequality to conclude $\hat{\theta}$ must converge to the true value
- requires the parameter space to be compact

- Cramer, 1946 – there exist solutions to the score equation that are consistent
- Taylor series expansion of $\log f(y;\theta)$
- if the likelihood function is maximized in the interior of the parameter space, the m.l.e. is one of these solutions
- if the score equation has only one root, the m.l.e. is consistent

max. lik est^n Scholz ESS (Wiley)

Non-standard cases

- true parameter θ_0 on the boundary of the parameter space
- example: $y_{ij} = \mu + b_i + \epsilon_{ij}, \quad b_i \sim N(0, \sigma_b^2), \epsilon_{ij} \sim N(0, \sigma^2)$
- if $\sigma_b^2 = 0$, no difference between groups; this is a boundary point of the parameter space

- non-identifiability; two different θ_1, θ_2 for which $f(y; \theta_1) = f(y; \theta_2)$
- example $f(y; \theta) = pN(\mu_1, 1) + (1 - p)N(\mu_2, 1)$
- if $\mu_1 = \mu_2$, then p is not identifiable
- if $p = 0$,then μ_1 is not identifiable
- likelihood ratio test of, e.g. $H_0 : p = 0$ will not be asymptotically χ^2

Non-standard cases

- true parameter θ_0 on the boundary of the parameter space
- example: $y_{ij} = \mu + b_i + \epsilon_{ij}, \quad b_i \sim N(0, \sigma_b^2), \epsilon_{ij} \sim N(0, \sigma^2)$
- if $\sigma_b^2 = 0$, no difference between groups; this is a boundary point of the parameter space

- non-identifiability; two different θ_1, θ_2 for which $f(y; \theta_1) = f(y; \theta_2)$
- example $f(y; \theta) = pN(\mu_1, 1) + (1 - p)N(\mu_2, 1)$
- if $\mu_1 = \mu_2$, then p is not identifiable
- if $p = 0$,then μ_1 is not identifiable
- likelihood ratio test of, e.g. $H_0 : p = 0$ will not be asymptotically χ^2

... non-standard cases

- multi-modal log-likelihoods
- in principle, find all the stationary points, and choose that corresponding to the maximum
- in practice, may not be feasible
- example: feed-forward neural networks

- support of the distribution depends on the parameter
- example $U(0, \theta)$; $n(y_{(n)} - \theta) \xrightarrow{\mathcal{L}}$ Exponential

- example $f(y; \theta) = \lambda \exp\{-\lambda(y - \mu)\}$

SM, §4.6; BNC94, §3.8; Cox, Ch. 7

... non-standard cases

- multi-modal log-likelihoods
- in principle, find all the stationary points, and choose that corresponding to the maximum
- in practice, may not be feasible
- example: feed-forward neural networks

- support of the distribution depends on the parameter
- example $U(0, \theta)$; $n(y_{(n)} - \theta) \xrightarrow{\mathcal{L}}$ Exponential $\hat{\theta} = y_{(n)}$

  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N$
- example $f(y; \theta) = \lambda \exp\{-\lambda(y - \mu)\}$

SM, §4.6; BNC94, §3.8; Cox, Ch. 7

... non-standard cases

- multi-modal log-likelihoods
- in principle, find all the stationary points, and choose that corresponding to the maximum
- in practice, may not be feasible
- example: feed-forward neural networks

- support of the distribution depends on the parameter
- example $U(0, \theta)$; $n(y_{(n)} - \theta) \xrightarrow{\mathcal{L}}$ Exponential

- example $f(y; \theta) = \lambda \exp\{-\lambda(y - \mu)\}$

SM, §4.6; BNC94, §3.8; Cox, Ch. 7

... non-standard cases

- singular information matrix: $\text{var}_{\theta_0}\{U(\theta_0)\} \equiv 0$
- usual Taylor series expansions do not apply; need to go to higher order terms
- might be fixable by re-parameterization

- Example: skew-normal distribution
- $Z \sim SKN(\alpha) : f_Z(z;\alpha) = 2\phi(z)\Phi(\alpha z)$
- three-parameter version: $Y = \xi + \omega Z$
- information matrix is singular, at $\alpha = 0$
- can be fixed by reparametrization to (μ, σ, α) Azzalini, 1999; 2011
- Example: informative non-response Rotnitzky et al., 2000

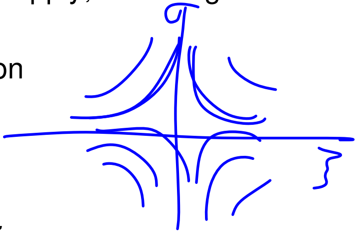
... non-standard cases

- singular information matrix: $\text{var}_{\theta_0}\{U(\theta_0)\} \equiv 0$
- usual Taylor series expansions do not apply; need to go to higher order terms
- might be fixable by re-parameterization

- Example: skew-normal distribution
- $Z \sim SKN(\alpha) : f_Z(z; \alpha) = 2\phi(z)\Phi(\alpha z)$
- three-parameter version: $Y = \xi + \omega Z$
- information matrix is singular, at $\alpha = 0$
- can be fixed by reparametrization to (μ, σ, α) Azzalini, 1999; 2011
- Example: informative non-response Rotnitzky et al., 2000

... non-standard cases

- singular information matrix: $\text{var}_{\theta_0}\{U(\theta_0)\} \equiv 0$
- usual Taylor series expansions do not apply; need to go to higher order terms
- might be fixable by re-parameterization



- Example: skew-normal distribution
- $Z \sim SKN(\alpha) : f_Z(z; \alpha) = 2\phi(z)\Phi(\alpha z)$
- three-parameter version: $Y = \xi + \omega Z$
- information matrix is singular, at $\alpha = 0$
- can be fixed by reparametrization to (μ, σ, α) Azzalini, 1999; 2011
- Example: informative non-response Rotnitzky et al., 2000

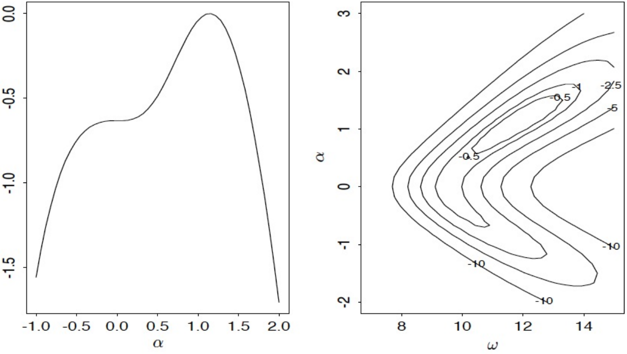


Figure 2: *Twice relative profile loglikelihood of* α *(left) and contour levels of the similar function of* (ω, α) *(right) for the Otis data, when the direct parametrization is used*

Azzalini, 1999

... non-standard cases

- informative non-response

Rotnitzky et al., 2000; Cox, 2009 Example 7.6

- observation (R_i, Y_i): $R_i = \mathbf{1}(Y_i \text{ observed })$
-

$$Y_i \sim N(\mu, \sigma^2), \quad \Pr(R_i = 1) = \exp\{H(\alpha_0 + \alpha_1(y_i - \mu)/\sigma)\}$$

$$\ell(\theta; y, r) = \sum_{i=1}^{n} -r_i \log\sigma - r_i(y_i - \mu)^2/(2\sigma^2) + r_i H\{\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$
$$+(1 - r_i)\log E[1 - \exp\{H(\alpha_0 + \alpha_1(Y_i - \mu)/\sigma)\}]$$

singular

information matrix at $\alpha = 0 \equiv$ missing at random

if, e.g., μ and σ^2 both unknown, sampling fluctuations in $\hat{\alpha}_1$ are $O_p(n^{-1/2})$

... non-standard cases

- informative non-response

  Rotnitzky et al., 2000; Cox, 2009 Example 7.6
- observation (R_i, Y_i): $R_i = \mathbf{1}(Y_i \text{ observed })$
- $$Y_i \sim N(\mu, \sigma^2), \quad \Pr(R_i = 1) = \exp\{H(\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$

$$\ell(\theta; y, r) = \sum_{i=1}^{n} -r_i \log\sigma - r_i(y_i - \mu)^2/(2\sigma^2) + r_i H\{\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$
$$+(1 - r_i)\log e[1 - \exp\{H(\alpha_0 + \alpha_1(Y_i - \mu)/\sigma)\}]$$ singular

information matrix at $\alpha = 0 \equiv$ missing at random

if, e.g., μ and σ^2 both unknown, sampling fluctuations in $\hat{\alpha}_1$ are $O_p(n^{-1/2})$

... non-standard cases

- informative non-response

  Rotnitzky et al., 2000; Cox, 2009 Example 7.6

- observation (R_i, Y_i): $R_i = \mathbf{1}(Y_i \text{ observed })$
-

$$Y_i \sim N(\mu, \sigma^2), \quad \Pr(R_i = 1) = \exp\{H(\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$

$$\ell(\theta; y, r) = \sum_{i=1}^{n} -r_i \log \sigma - r_i(y_i - \mu)^2/(2\sigma^2) + r_i H\{\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$
$$+(1 - r_i) \log e[1 - \exp\{H(\alpha_0 + \alpha_1(Y_i - \mu)/\sigma)\}] \text{ singular}$$

information matrix at $\alpha = 0 \equiv$ missing at random

if, e.g., μ and σ^2 both unknown, sampling fluctuations in $\hat{\alpha}_1$ are $O_p(n^{-1/2})$

... non-standard cases

- informative non-response

  Rotnitzky et al., 2000; Cox, 2009 Example 7.6

- observation (R_i, Y_i): $R_i = \mathbf{1}(Y_i \text{ observed })$
-

$$Y_i \sim N(\mu, \sigma^2), \quad \Pr(R_i = 1) = \exp\{H(\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$

$$\ell(\theta; y, r) = \sum_{i=1}^{n} -r_i \log\sigma - r_i(y_i - \mu)^2/(2\sigma^2) + r_i H\{\alpha_0 + \alpha_1(y_i - \mu)/\sigma\}$$
$$+(1 - r_i)\log e[1 - \exp\{H(\alpha_0 + \alpha_1(Y_i - \mu)/\sigma)\}]$$ singular

information matrix at $\alpha_1 = 0 \equiv$ missing at random
if, e.g., μ and σ^2 both unknown, sampling fluctuations in $\hat{\alpha}_1$ are $O_p(n^{-1/6})$

R. et al 2000

Problems – Week 4

1. Suppose $y = y_1, \ldots, y_n$ are independent and identically distributed from a distribution with density $f(y;\theta) = \prod_{i=1}^n f_1(y_i;\theta)$, $\theta \in R$. Further let $g(y;\theta) = \sum_{i=1}^n g_1(y_i;\theta)$ be an *unbiased estimating equation* for θ, satisfying $E_\theta\{g(y_i;\theta)\} = 0$ for all θ. The estimate defined by $g(y;\tilde{\theta}_g) = 0$ has asymptotic variance $G^{-1}(\theta) = H^{-1}(\theta)J(\theta)H^{-1}(\theta)$, where $H(\theta) = -\mathrm{E}_\theta\{\nabla_\theta g(y_1;\theta)\}$ and $J(\theta) = \mathrm{var}_\theta\{g(y_1;\theta)\}$. The estimating equation is called *optimal* if it has the largest possible value of $G(\theta)$.

   Show that $G(\theta) \le i_1(\theta)$, where $i_1(\theta)$ is the expected Fisher information in a single observation. This implies that the score equation is the optimum estimating equation.

   Two fun facts that you don't need to prove:

   (a) The multivariate version of this is that $i_1(\underline{\theta}) - G(\underline{\theta})$ is non-negative definite (but you don't need to show this).

   C-S as in

   (b) In the autoregressive model

$$y_i = \theta y_{i-1} + \epsilon_i, \quad i = 1, \ldots, n$$

   where y_0 is a constant and ϵ_i are i.i.d. $N(0, \sigma^2)$, show the equation

$$\Sigma y_i y_{i-1} - \theta \Sigma y_i^2 = 0$$

   ← Score eq$^{\underline{n}}$

   is an unbiased estimating equation obtaining the lower bound.

$$\ell(\theta) = -\frac{1}{2\sigma^2} \sum (y_i - \theta y_{i-1})^2 - \frac{n}{2} \ln \sigma^2$$

$$\ell'(\theta) = 0 \implies$$

$$\sum y_i y_{i-1} = \theta \sum y_i^2$$

Problems – Week 4

evalues $(H^{-1}J)$ all of same sign

2. Suppose $Z \sim \sum_{r=1}^{m} \mu_r X_r^2$, where $X_1, \ldots, X_m$ are independent observations from a $N(0,1)$ distribution. If all the μ_r were equal, the distribution of Z would be proportional to a χ_m^2. *Satterthwaite's approximation* (Satterthwaite, 1946) to the distribution of Z is $a\chi_b^2$, where a and b are chosen so that $\mathrm{E}(Z)$ and $\mathrm{var}(Z)$ are equal to the mean and variance of a $a\chi_b^2$ random variable. This idea can also be used to approximate a non-central χ^2 distribution, and arises in the distribution of quadratic forms in unbalanced analysis of variance.

    (a) Find expressions for a and b, in terms of $\mu_1, \ldots, \mu_m$.

    (b) Illustrate the approximation numerically in a simple example with, say, $m = 5, 10$. You can choose the values of μ_r in any way you like, but one possibility is to simulate a random vector from $N(0, A)$ for some choice of $A \neq I$; then $X^T X$ will (I think), have the distribution you are looking for. The function `mvrnorm` in the `MASS` library simulates multivariate normal random variables.

$Z \sim a\chi_b^2$

$$EZ = \sum_1^r \mu_r = E(a\chi_b^2) = ab$$

$$\mathrm{var}\, Z = 2\sum_1^r \mu_r^2 = \mathrm{var}(a\chi_b^2) = a^2 2b$$

Problems – Week 3

1. Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed from a model $f(y;\theta), y \in R, \theta \in R$, and that $\pi(\theta)$ is a proper prior density (with respect to Lebesgue measure on R). Denote by $\hat{\theta}_\pi$ the posterior mode:

$$\hat{\theta}_\pi = \arg\sup_\theta \pi(\theta \mid y)$$

which we assume is obtained as the unique root of the equation

$$\frac{d}{d\theta} \log \pi(\hat{\theta}_\pi \mid y) = 0. \qquad (1)$$

Denote by $\tilde{\theta}$ the posterior mean:

$$\tilde{\theta} = \int \theta \pi(\theta \mid y) d\theta.$$

2 Laplace

Show that

$$\hat{\theta}_\pi - \hat{\theta} = O_p(\frac{1}{n}), \text{ and } \tilde{\theta} - \hat{\theta} = O_p(\frac{1}{n}),$$

(H.O. part 1 (b))

where $\hat{\theta}$ is the maximum likelihood estimator of θ.

$$\tilde{\theta} = \frac{\int \theta \pi(\theta|y)\, d\theta}{} = \frac{\int \theta\, e^{\ell(\theta)} \pi(\theta)\, d\theta}{\int e^{\ell(\theta)} \pi(\theta)\, d\theta}$$

notes on Laplace ↑ check

$$\doteq \frac{\sqrt{2\pi}\, \hat{\theta}\, e^{\ell(\hat{\theta})} \pi(\hat{\theta}) \{-\ell''(\hat{\theta})\}^{-\frac{1}{2}}}{\sqrt{2\pi}\, e^{\ell(\hat{\theta})} \pi(\hat{\theta}) \{-\ell''(\hat{\theta})\}^{-\frac{1}{2}}} +$$

$$\left\{1 + \frac{A}{n}\right\} \Big/ \left(1 + \frac{B}{n}\right) = \hat{\theta} + O_p\left(\frac{1}{n}\right)$$

$$\frac{\left(1+\frac{A}{n}\right)}{\left(1+\frac{B}{n}\right)}$$

$$Z_n\{1+o_p(1)\} = Y_n\{1+o_p(1)\}$$

$$\Rightarrow Z_n = Y_n\{1+o_p(1)\}$$

$$\| $$

$$\frac{1+O\left(\frac{1}{n}\right)}{1+O\left(\frac{1}{n}\right)} = 1+O\left(\frac{1}{n}\right) + \cdots$$

$$= \left(1+\frac{A}{n}\right)\left(1+\frac{B}{n}\right)^{-1} = \left(1+\frac{A}{n}\right)\left(1-\frac{B}{n}+\cdots\right)$$

Problems – Week 3

Sartori 2003 Bka?

MPL w̄ nuisance

2. Consider a linear regression model

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n$$

where x_i and β are $p \times 1$ vectors, and ϵ_i are i.i.d. $N(0, \sigma^2)$. Compare the log-likelihood ratio statistics for inference about β, based on the

(a) profile log-likelihood $w(\beta) = 2\{\ell_{\mathrm{p}}(\hat{\beta}) - \ell_{\mathrm{p}}(\beta)\}$,

(b) adjusted profile log-likelihood $w_{\mathrm{A}}(\beta) = 2\{\ell_{\mathrm{A}}(\hat{\beta}_{\mathrm{A}}) - \ell_{\mathrm{A}}(\beta)\}$, and

(c) modified profile log-likelihood $w_{\mathrm{M}}(\beta) = 2\{\ell_{\mathrm{M}}(\hat{\beta}_{\mathrm{M}}) - \ell_{\mathrm{M}}(\beta)\}$,

where

$$\ell_{\mathrm{p}}(\beta) = \ell(\beta, \hat{\sigma}^2_\beta), \quad \ell_{\mathrm{A}}(\beta) = \ell_{\mathrm{p}}(\beta) - \frac{1}{2}\log|j_{\sigma^2\sigma^2}(\beta, \hat{\sigma}^2_\beta)|, \text{ and } \ell_{\mathrm{M}}(\sigma^2) = \ell_{\mathrm{A}}(\beta) + \log|\frac{d\hat{\sigma}^2}{d\hat{\sigma}^2_\beta}|,$$

and $\hat{\beta}_{\mathrm{A}}$, $\hat{\beta}_{\mathrm{M}}$ are the adjusted and modified maximum likelihood estimators, respectively.

$$\ell_p(\beta) = -\frac{n}{2}\log(y - X\beta)^T(y - X\beta) \qquad \ell_A(\beta) = -\left(\frac{n-2}{2}\right)\log RSS_\beta$$

$\sigma^2 \perp \beta \quad \Rightarrow \quad L_M = L_A$ (class)

$$\hat{\sigma}^2_\beta = \frac{1}{n}(y - X\beta)^T (y - X\beta) \qquad \ell_M, \ell_A$$

ℓ_P

$$\hat{\sigma}^2_\beta = \frac{1}{n}(y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$= \hat{\sigma}^2 + (\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta)$$

$$\frac{d\hat{\sigma}^2_\beta}{d\hat{\sigma}^2} = 1$$

Problems – Week 2

1. Suppose $Y_1, \ldots, Y_n$ are i.i.d. with density

$$f_{Y_i}(y;\mu) = \frac{1}{\mu}\exp(-\frac{y}{\mu}), y > 0, \mu > 0.$$

Show that the leading term in the saddlepoint approximation to the density of $\bar{Y} = \hat{\mu}$ reproduces the gamma density, with $\Gamma(n)$ replaced by Stirling's approximation to it. Deduce that the renormalized saddlepoint approximation is exact.

$\rho_j(\hat{\phi}) \equiv \rho_j$ $\forall j$ for gamma, N, InvG

3 1-par. families for which renormalized s-pt approx$^{\underline{n}}$ is exact

Gamma Normal Inv. Gaussian

$$f_{S_n}(s_0) \doteq \hat{f}_{S_n}\left\{1 + \frac{3\rho_4 - 5\rho_3^2}{n} + \frac{\quad}{n^2} + \cdots\right\}$$

Problems – Week 2

(a) Suppose that Y_{i1} and Y_{i2} are independent observations from exponential distributions with means $\psi\lambda_i$ and ψ/λ_i, respectively, $i = 1, \ldots, n$. Show that the maximum likelihood estimator of ψ is not consistent, but converges in probability to $(\pi/4)\psi$.

(b) A modification to the profile likelihood to account for estimation of nuisance parameters was proposed in Cox & Reid (1987):

$$\ell_m(\psi) = \ell(\psi, \hat{\lambda}_\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|,$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$ and $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimator of λ. This is to be computed using a parametrization of the nuisance parameter that is *orthogonal* to the parameter of interest ψ, with respect to expected Fisher information. (The correction term $\frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$ is not invariant to reparameterizations,) Show for the exponential case that λ is orthogonal to ψ, and that the value of ψ that solves $\ell_m'(\psi) = 0$, $\hat{\psi}_m$, say, converges to $(\pi/3)\psi$.

$E\left(\sqrt{y_{1i} y_{2i}}\right) = \ldots ? \; \frac{\pi}{4} \; ?$

Problems – Week 1

1. *Orthogonal nuisance parameters.* In a model $f(y;\theta)$ with $\theta = (\psi, \lambda)$, the component parameter ψ and λ are orthogonal (with respect to Fisher information) if $i_{\psi\lambda}(\theta) = 0$.

   (a) Suppose we have a sample $y_1, \dots, y_n$ from the density $f(y;\theta)$. Show that

   $$\hat{\lambda}_\psi = \hat{\lambda} + O_p(n^{-1/2}),$$

   whereas if ψ and λ are orthogonal that

   $$\hat{\lambda}_\psi = \hat{\lambda} + O_p(n^{-1}).$$

   (b) Assume y_i follows an exponential distribution with mean $\lambda e^{-\psi x_i}$, where x_i is known. Find conditions on the sequence $\{x_i, i = 1, \dots, n\}$ in order that λ and ψ are orthogonal with respect to expected Fisher information. Find an expression for the constrained maximum likelihood estimate $\hat{\lambda}_\psi$ and show the effect of parameter orthogonality on the form of the estimate.

$$\ell(\psi,\lambda) = \ell(\hat\psi,\hat\lambda) + \frac{1}{2}(\psi-\hat\psi)^2 \hat{\ell}_{\psi\psi} + \frac{1}{2}(\lambda-\hat\lambda)^2 \hat{\ell}_{\lambda\lambda} + (\psi-\hat\psi)(\lambda-\hat\lambda)\hat{\ell}_{\psi\lambda} + O_p\{\|\theta-\hat\theta\|^3(n)\}$$

$$l_{\psi\psi} = n i_{1,\psi\psi} + \sqrt{n} Z_{\psi\psi}$$

$$\frac{1}{n}\left(l_{\psi\psi} - i_{1,\psi\psi}\right) = \frac{Z_{\psi\psi}}{\sqrt{n}} = O_p(1) \text{ by as}$$

← use this

Problems – Week 1

2. *Sufficient statistics (CH Exercise 2.2).* Find the log-likelihood function for a sample of size n from an $AR(1)$ process:

$$y_t = \mu + \rho(y_{t-1} - \mu) + \epsilon_t, \quad \epsilon_t(i.i.d.) \sim N(0, \sigma^2), \quad t = 1, ..., n,$$

where $|\rho| < 1$, as a function of $\theta = (\mu, \sigma^2, \rho)$ and y_0. Write down the likelihood for data $y_1, \ldots, y_n$ in the cases where the initial value y_0 is

(a) a given constant;

(b) normally distributed with mean μ and variance $\sigma^2/(1-\rho^2)$;

(c) assumed equal to y_n,

and give the sufficient statistic for each case.

S has dim > 3 $\left(\sum_1^{n-1} y_t^2, \sum_1^{n-1} y_t, \sum_2^{n} y_t y_{t-1}, y_n, y_0\right)$

(b) ↔

Extra notes for HW1, 3

Notes to help

LTCC/Reid: Derivation of limiting results: scalar parameter **November 6, 2012**

Using the notation on the handout from November 5, ("week1-handout.pdf"), here is a moderately rigorous proof of the results

$$\sqrt{n}(\hat\theta - \theta) = \frac{1}{\sqrt{n}} i_1^{-1}(\theta)U(\theta)\{1 + o_p(1)\}, \quad (1)$$

$$2\{\ell(\hat\theta) - \ell(\theta)\} = (\hat\theta - \theta)^T i(\theta)(\hat\theta - \theta)\{1 + o_p(1)\}. \quad (2)$$

The vector case is unchanged, except for tedious notational changes in Taylor's theorem with remainder, although of course we need the dimension of θ fixed as $n \to \infty$.

For (1), we have

$$\ell'(\hat\theta) = \ell'(\theta) + (\hat\theta - \theta)\ell''(\theta) + \frac{1}{2}(\hat\theta - \theta)^2 \ell'''(\theta_n^*),$$

$$-\frac{\ell'(\theta)}{\ell''(\theta)} = (\hat\theta - \theta)\{1 + \frac{1}{2}(\hat\theta - \theta)\frac{\ell'''(\theta_n^*)}{\ell''(\theta)}\},$$

$$\frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\ell''(\theta)/n} \cdot \frac{i_1(\theta)}{i_1(\theta)} = \sqrt{n}(\hat\theta - \theta)\{1 - \frac{1}{2}(\hat\theta - \theta)\frac{\ell'''(\theta_n^*)/n}{-\ell''(\theta)/n}\},$$

$$\frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{i_1(\theta)}\left(\frac{i_1(\theta)}{-\ell''(\theta)/n}\right) = \sqrt{n}(\hat\theta - \theta)\{1 + Z_n\}.$$

The term in brackets on the LHS of the last line converges in probability to 1, by the WLLN, so can be written $1 + o_p(1)$. The remainder term Z_n converges in probability to 0, because we assume $\hat\theta \xrightarrow{p} \theta$, so that $\theta_n^* \xrightarrow{p} \theta$, because $|\theta_n^* - \theta| < |\hat\theta - \theta|$. Also $\frac{1}{n}\ell'''(\theta_n^*) \xrightarrow{p} E\{\ell'''(\theta; Y)\}$ which we assume is finite (p.281 of CH, for example); similarly $-\frac{1}{n}\ell''(\theta) \xrightarrow{p} i_1(\theta)$, so $Z_n = o_p(1)O_p(1) = o_p(1)$. Then we can move over the LHS term as