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

November 6

- HW 2 comments & K-L divergence
- presentations
- semi-parametric likelihood as profile
- empirical likelihood
- composite likelihood

Exercises October 23 **STA 4508S (Fall, 2018)**

1. The Kullback-Leibler divergence from the distribution G to the distribution F is given by

$$KL(F:G) = \int \log \frac{f(y)}{g(y)} f(y) dy, \tag{1}$$

where f and g are and density functions with respect to Lebesgue measure. Note that the divergence is not symmetric in its arguments. This is called the directed information distance in Barndorff-Nielsen and Cox (1994) where the more general definition $KL(F : G) = \int \log(dF/dG)dF$ is used, assuming F and G are mutually absolutely continuous.

   (a) In the canonical exponential family model with density $f(s;\varphi) = \exp\{\varphi^T s - k(\varphi)\}h(s)$, $s \in \mathbb{R}^p$, find an expression for the KL divergence between the model with parameter φ_1 and that with parameter φ_2.

   (b) Show that for a sample of observations from a model with density $f(y;\theta)$ the maximum likelihood estimator minimizes the KL divergence from $F(\cdot;\theta)$ to $G_n(\cdot)$, where $G_n(\cdot)$ is the empirical distribution function putting mass $1/n$ at each observation y_i.

2. Suppose $y_i \sim N(\mu_i, 1/n), i = 1, \ldots, k$ and $\psi^2 = \Sigma_{i=1}^k \mu_i^2$ is the parameter of interest.[1]

   (a) Show that the marginal posterior density for $n\psi^2$, assuming a flat prior $\pi(\underline{\mu}) \propto 1$, is a non-central χ_k^2 distribution, with non-centrality parameter $n\Sigma y_i^2$.

   (b) Show that the maximum likelihood estimate of ψ^2 is $\hat{\psi}^2 = \Sigma y_i^2$, and that $n\hat{\psi}^2$ has a non-central χ_k^2 distribution with non-centrality parameter $n\psi^2$.

   (c) Compare the normal approximations to $r_u(\psi)$, $r_e(\psi)$ and $r(\psi)$ with the exact distribution of the maximum likelihood estimate.

   (d) Compare the 95% Bayesian posterior probability interval for ψ^2, based on (a) to the 95% confidence interval for ψ^2, based on (b).

[1] It will be convenient to use $\lambda_i = \mu_i/\sqrt{\Sigma\mu_i^2}$ for the nuisance parameters.

Proportional hazards regression

- partial log-likelihood function $\quad y_1 < y_2 < \cdots < y_n$

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

$\mathcal{R}_i = \{j : y_j \geq y_i\}$

set of individuals that could be observed to fail at time y_i

see SM §10.8 for treatment of ties

- can be motivated as:
  1. marginal log-likelihood of the ranks of the failure times
  2. $\prod_{i=1}^{n} \Pr(\text{unit } i \text{ fails at } y_i \mid \mathcal{R}_i, \text{ there is one failure at } y_i)$
  3. profile log-likelihood function if $\lambda(\cdot)$ is represented by a vector of values $(\lambda_1, \ldots, \lambda_n) = \{\lambda(y_1), \ldots \lambda(y_n)\}$

Inference

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

- $\tilde{U}$ is the projection of $\partial\ell/\partial\theta$ on space spanned by nuisance function
- as in parametric models, lead to

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

- and likelihood ratio test

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

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

Infinite-dimensional models

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

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

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

... infinite-dimensional models

- Definition: Given a measure P, and a sample $(y_1, \ldots, y_n)$, 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

Example: the empirical distribution

vdW 25.68

- $\mathcal{P}$ 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
- $\{(P\{y_1\}, \ldots, P(\{y_n\}); P \in \mathcal{P}\} \Longleftrightarrow$
  $(p_1, \ldots, p_n), p_i \geq 0, \Sigma p_i = 1)$

- 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, $\quad \mathrm{E}(y_i) = \theta_0, \mathrm{var}(y_i) < \infty$,
- $$-2 \log \mathcal{R}(\theta_0) \xrightarrow{d} \chi^2_1, \quad n \to \infty$$

  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

VandW Ex.25.71

- $Pr(Y = 1 \mid V, W) = \dfrac{e^{\theta V + \eta(W)}}{1 + e^{\theta V + \eta(W)}}$
- 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

Example: missing covariate

Murphy and vdW, 2000

- 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

Z is the 'gold standard' covariate, e.g. LDL cholesterol

- (d_R, w_R) has a missing covariate

W is a surrogate for Z

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

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

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

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

can be profiled, according to M&vdW

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

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
$\sum_{r=1}^{m} w_r \log f(y_r \mid$ '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	$c\ell(\theta; y) = \log cL(\theta; y) = \sum_k w_k \ell_k(\theta; y)$
composite score function	$U_{CL}(\theta) = \partial c\ell(\theta; y)/\partial\theta$
sensitivity	$H(\theta) = \mathsf{E}_\theta\{-\partial^2 c\ell(\theta; y)/\partial\theta\partial\theta^T\}$
variability	$J(\theta) = \mathsf{E}_\theta\{U_{CL}(\theta)U(\theta)^T\}$
Godambe information	$G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$

... derived quantities

sample $y = (y_1, \ldots, y_n)$ 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., $CL(\theta; \underline{y}) = \prod_{i=1}^{n} CL(\theta; y_i)$
- $\hat{\theta}_{CL} - \theta \dot{\sim} N\{0, G^{-1}(\theta)\}$ $\qquad G_n(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- $U(\hat{\theta}_{CL}) \doteq U(\theta) + (\hat{\theta}_{CL} - \theta)\partial_\theta U(\theta)$ $\qquad U = U_{CL}$
- $\hat{\theta}_{CL} - \theta \doteq -\partial_\theta U(\theta)^{-1} U(\theta) \doteq H^{-1}(\theta)U(\theta)$
- $U(\theta) \dot{\sim} N\{0, J(\theta)\}$
- $H^{-1}(\theta)U(\theta) \dot{\sim} N\{0, H^{-1}(\theta)J(\theta)H^{-T}(\theta)\}$
- 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 $J(\theta) = H(\theta)$, $w(\theta) \dot{\sim} \chi_d^2$
- if $d = 1$, $w(\theta) \dot{\sim} \mu_1 \chi_1^2 = J(\theta)H^{-1}(\theta)\chi_1^2$ H, J both scalars

Example: symmetric normal

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

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

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

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

... symmetric normal

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

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

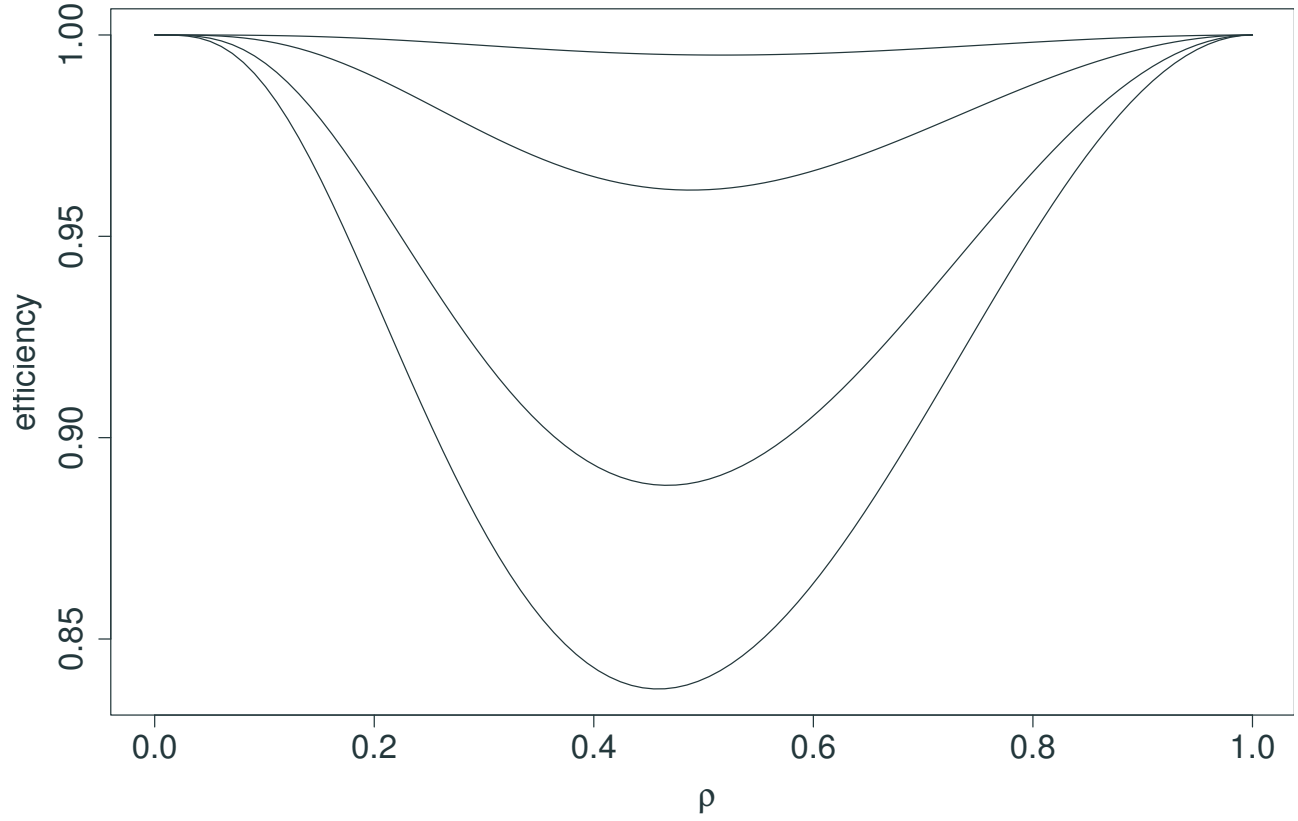
$$O(\tfrac{1}{n}) \qquad O(1)$$

$$n \longrightarrow \infty \qquad m \longrightarrow \infty$$

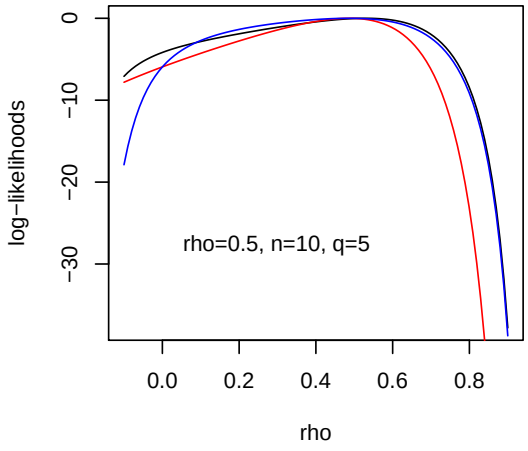
... symmetric normal

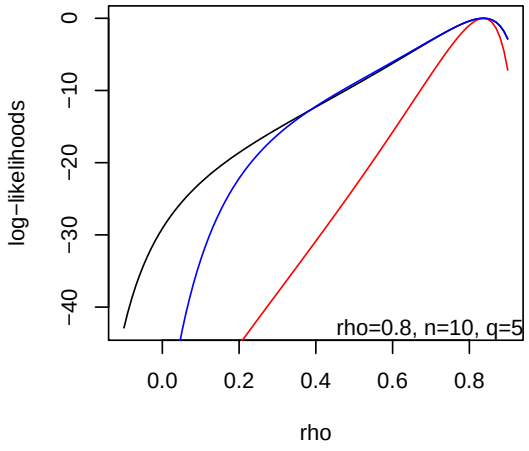
$$\frac{a.var(\hat{\rho}\,)}{a.var(\hat{\rho}_{CL}\,)}, \quad m = 3, 5, 8, 10$$

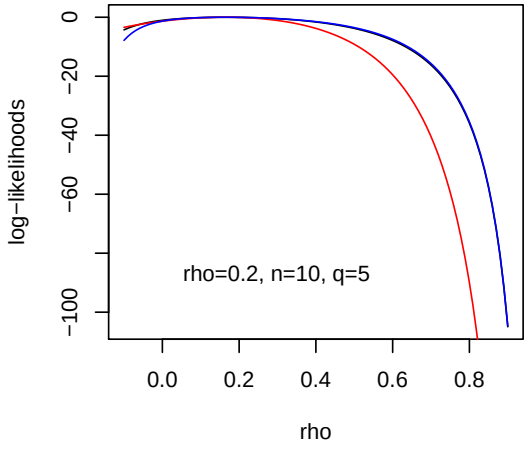
(Cox & Reid, 2004)

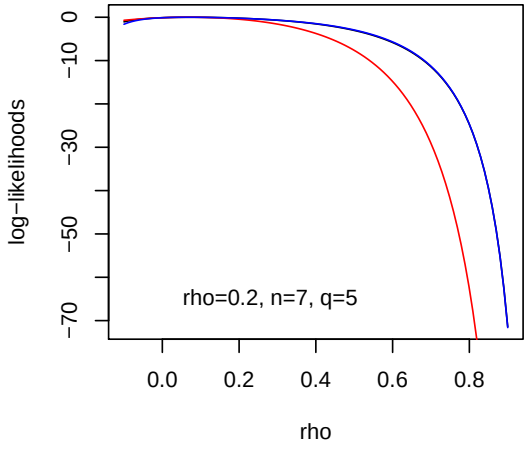


Likelihood ratio test









Example: longitudinal count data

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

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

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

Henderson & Shimura, 2003