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. likelihood inference for $p > n$

7. bootstrap likelihood, h-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

November 20

- HW2 and HW4 notes – please see updates on web page
- Godambe information
- quasi-likelihood
- indirect inference
- high-dimensional inference

1. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i are independent normal random variables with expected value zero and variance $\sigma_i^2 = \sigma^2(1 + \gamma x_i^2), i = 1, \ldots, n$. Simulate 1000 datasets of length $n = 50$ with parameters $\beta_0 = 1, \beta_1 = 1, \sigma^2 = 3, \gamma = 2$ and covariate x_i simulated from a $U(-1, 1)$.

   (a) Fit each dataset with a simple linear regression model (assuming $\gamma = 0$), and ~~compare~~compute the simulation mean and variance of $\hat\beta_1$ ~~to that computed from the fitted model with $\gamma = 0$~~.

   (b) Compare the true and estimated sandwich variance of $\hat\beta_1$ based on the Godambe information matrix to the naive estimate from (a) (from the regression output).

   (c) The true $\text{var}(\hat\beta_1)$ can be computed (tediously) from the appropriate element of G^{-1}, where $G(\cdot)$ is the Godambe information. Somewhat confusingly, this is a 3×3 matrix, since the fitted model has just 3 parameters, but it depends on $(\beta_0, \beta_1, \sigma^2, \gamma)$ (and these values are known since we are simulating). (Royal "we")

   The estimated value of the Godambe information is less clear, because we have no estimate of γ. However, if we compute the 3×1 score vector for each observation, say $U_i(\beta_0, \beta_1, \sigma^2)$ we can estimate $\mathrm{E}(UU^{\mathrm{T}})$ by

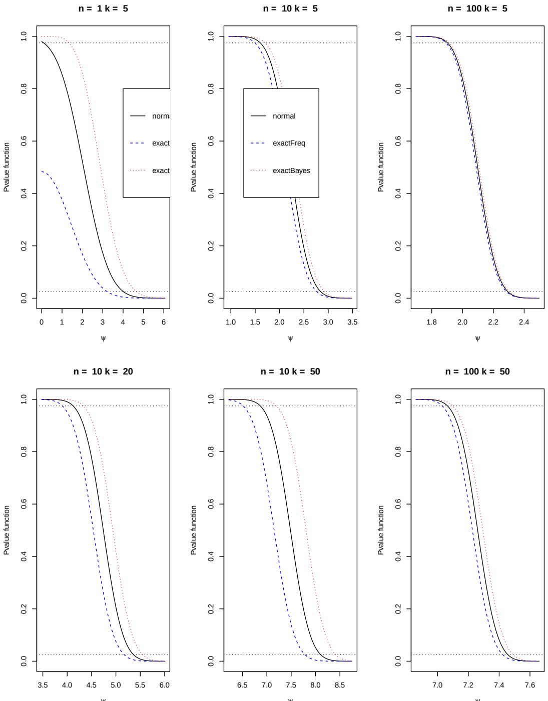
   $$\frac{1}{n}\sum_{i=1}^{n} U_i(\hat\theta)U_i(\hat\theta)^{\mathrm{T}}. \tag{1}$$

   General least-squares theory shows that if $y \sim N(X\beta, \sigma^2 W)$, then $\hat\beta_{LS} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ has expected value β and variance

   $$\sigma^2(X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}WX)(X^{\mathrm{T}}X)^{-1},$$

   where W is a diagonal matrix whose entries can be estimated using $\hat\epsilon$. This agrees with what I got using (1).

   So together (a) and (b) ask you to compare the simulation variance of $\hat\beta_1$ with its true variance under the model, the estimated variance using (1), and the estimated variance from the regression fit $(\sigma^2(X^{\mathrm{T}}X)^{-1})$.


n = 1 k = 5
n = 10 k = 5
n = 100 k = 5
n = 10 k = 20
n = 10 k = 50
n = 100 k = 50
Pvalue function
ψ
normal
exactFreq
exactBayes

Quasi-likelihood

- Recall: generalized linear model $y_1, \dots, y_n$ independent, with

$$f(y_i \mid x_i; \beta, \phi) = \exp[\{y_i\theta_i - c(\theta_i)\}/\phi + h(y_i, \phi)]$$

- ϕ a scale parameter in this exponential family
- $E(y_i) = \mu_i = c'(\theta_i)$ Exercises 1
- $\mathrm{var}(y_i) = \phi V(\mu_i) = \phi c''(\theta_i)$ variance function
- $g(\mu_i) = x_i^{\mathrm{T}}\beta$ link function
- link function converts $\theta_{n\times 1}$ to $\beta_{p\times 1}$
- Standard $V(\mu)$: Normal– 1; Gamma– μ^2; Poisson– μ; Bernoulli– $\mu(1-\mu)$
-

$$\ell(\beta, \phi; y) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - c(\theta_i)}{\phi} + h(y_i, \phi) \right\}$$

... quasi-likelihood

- log-likelihood

$$\ell(\beta, \phi; y) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - c(\theta_i)}{\phi} + h(y_i, \phi) \right\}$$

- score function

$$\frac{\partial \ell}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_r}$$

- MLE

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)} = 0$$

- Bartlett identity:

$$E\left\{ \frac{\partial \ell^2}{\partial \beta_r \partial \beta_s} + \frac{\partial \ell}{\partial \beta_r} \frac{\partial \ell}{\partial \beta_s} \right\} = 0$$

… quasi-likelihood

- Suppose instead of a generalized linear model, we had only a partially specified model:

$$\mathrm{E}(y_i) = \mu_i,\ \mathrm{var}(y_i) = \phi V(\mu_i),\ g(\mu_i) = x_i^{\mathrm{T}}\beta$$

- $g(\cdot), V(\cdot)$ known
- unbiased estimating equation

$$g(y;\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)}$$

- using result from Exercises, if $g(y;\tilde\beta) = 0$, then asymptotic variance of $\tilde\beta$ is

$$\mathrm{E}\left\{-\frac{\partial g(y;\beta)}{\partial\beta^{\mathrm{T}}}\right\}^{-1} \mathrm{var}\{g(y;\beta)\}\mathrm{E}\left\{-\frac{\partial g(y;\beta)}{\partial\beta}\right\}^{-1}$$

as with composite likelihood

- With $g(y;\beta) = \Sigma g_i(y_i;\beta) = \sum \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)}$
- $\mathrm{E}\left\{-\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}}\right\} = \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \frac{1}{g'(\mu_i)^2 \phi V(\mu_i)} = \phi^{-1} X^{\mathrm{T}} W X = \mathrm{var}\{g(y;\beta)\}$
- $W = \mathrm{diag}(w_j), \quad w_j = \{g'(\mu_j)^2 V(\mu_j)\}^{-1}, j = 1, \dots, n$

- quasi-likelihood function

$$Q(\beta; y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - u}{\phi V(u)} du$$

- this only works for models of this form
- called quasi-likelihood because $\partial Q / \partial \beta$ gives estimating equation with expected value 0, and 2nd Bartlett identity holds

Longitudinal data

- suppose now our observations come in groups: $y_{ij}, j = 1, \ldots, m_i; i = 1, \ldots, n$
- could be repeated measurements on subjects
- or measurements of members of the same cluster/family/group

- assume GLM-type structure $\mathrm{E}(y_{ij}) = \mu_{ij}$, $g(\mu_{ij}) = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i$
- random effects b_i induce correlation among observations in the same group; e.g. assume $b_i \sim N(0, \Omega_b)$

- GLM variance structure $\mathrm{var}(y_{ij}) = V_i(\beta, \alpha)$ for example
- α are extra parameters in the variance-covariance matrix
- QL-type estimating equations

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta^{\mathrm{T}}}\right)^{\mathrm{T}} V_i^{-1}(\alpha, \beta)(y_i - \mu_i) = 0$$

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta^{\mathrm{T}}}\right)^{\mathrm{T}} V_i^{-1}(\alpha,\beta)(y_i-\mu_i)=0$$

- parameter α in variance function doesn't divide out, as in univariate case
- we will need an estimate $\hat{\alpha}$ from somewhere

many suggestions in the literature

- Liang & Zeger suggested using a "working covariance matrix" to get an estimate of β
- e.g. could assume independence, or $AR(1)$, or ...
- estimates of β will still be consistent, but the asymptotic variance will be of the sandwich form as the model is misspecified
- there is no integrated function that serves as a quasi-likelihood in this setting

- true model complex, but can be simulated as the system $y_t = G(y_{t-1}, x_t, u_t; \beta), t = 1, \ldots, T$
- G is known function, x_t is observed (‘exogenous’), u_t is noise (possibly i.i.d. F), $\beta \in \mathbb{R}^k$ to be estimated

- simpler working model has density $f(y_t \mid y_{t-1}, x_t; \theta)$; $\theta \in \mathbb{R}^p, p \geq k$
- maximum likelihood estimate $\hat{\theta} = \hat{\theta}(y)$ easy to obtain

- 1. simulate $\tilde{u}_1^m, \ldots \tilde{u}_t^m$ from F
  2. choose β and construct $\tilde{y}_1^m(\beta), \ldots, y_t^m(\beta)$
  3. use simulated data to estimate θ
- $\tilde{\theta}(\beta) = \text{argmax}_\theta \sum_{m=1}^M \sum_{t=1}^T \log f\{\tilde{y}_t^m(\beta) \mid \tilde{y}_{t-1}^m(\beta), x_t, \theta\}$
- estimate β by minimizing $d\{\hat{\theta}, \tilde{\theta}(\beta)\}$ for some distance measure

- $\tilde{\theta}(\beta) = \text{argmax}_\theta \sum_{m=1}^{M} \sum_{t=1}^{T} \log f\{\tilde{y}_t^m(\beta) \mid \tilde{y}_{t-1}^m(\beta), x_t, \theta\}$
- $\tilde{\beta} = \text{argmin}_\beta d\{\hat{\theta}, \tilde{\theta}(\beta)\}$
- as $T \to \infty$, $\tilde{\theta}(\beta) \to h(\beta)$ and $\hat{\theta} \to \theta_0$ pseudo-true value, θ^*
- if $p = k$ then $h(\cdot)$ is one-to-one and can be inverted: $h^{-1}(\theta_0) = \beta; h^{-1}(\hat{\theta}) = \hat{\beta}$
- when $p > k$ need to choose $d(\cdot, \cdot)$

1. Wald $\hat{\beta}_W = \text{argmin}_\beta \{\hat{\theta} - \tilde{\theta}(\beta)\}^{\mathrm{T}} W \{\hat{\theta} - \tilde{\theta}(\beta)\}$
2. score $\hat{\beta}_S = \text{argmin}_\beta S(\beta)^{\mathrm{T}} V S(\beta)$,
   $S(\beta) = \sum_{m=1}^{M} \sum_{t=1}^{T} \partial \log f\{\tilde{y}_t^m(\beta) \mid \tilde{y}_{t-1}^m(\beta), x_t, \hat{\theta}\} / \partial\theta$
3. $\hat{\beta}_{LR} = \text{argmin}_\beta \sum_{t=1}^{T} [\log f\{y_t \mid y_{t-1}, x_t, \hat{\theta}) - \log f\{y_t \mid y_{t-1}, x_t, \tilde{\theta}(\beta)\}]$

... indirect inference

- it is not necessary that $\hat{\theta}$ be the mle under the working model
- we could instead assume some statistic $\hat{s}(y)$ of dimension p
- perhaps obtained by solving $\sum_{t=1}^{T} h(y_t; s) = 0$ for an estimating equation
- we will probably have something like $\sqrt{T}\{\hat{s} - s(\beta)\} \to N_p(0, \nu)$

- $H(\beta; \hat{s}) = \{\hat{s} - s(\beta)\}^{\mathrm{T}} \nu^{-1} \{\hat{s} - s(\beta)\}$
- $\tilde{\beta} = \text{argmin}_{\beta} H(\beta; \hat{s})$
- $\exp\{-H(\beta; \hat{s})\}$ called 'indirect likelihood' Jiang & Turnbull '04

- often $\hat{s}$ will be a set of moments of the observed data

- a similar use of simulation to avoid computation of complex likelihood functions
- model $f(y \mid \theta)$, prior $\pi(\theta)$ posterior $\pi(\theta \mid y) \propto f(y \mid \theta)\pi(\theta)$
- Algorithm 1: assume y takes values in a finite set $\mathcal{D}$
    1. generate $\theta' \sim \pi(\theta)$ prior
    2. simulate $z_i \sim f(\cdot \mid \theta')$
    3. if $z_i = y$, set $\theta_i = \theta'$, repeat
- after N steps, $\theta_1, \dots, \theta_N$ is a sample from $\pi(\theta \mid y)$
- because $\sum_{z_i \in \mathcal{D}} \pi(\theta_i) f(z_i \mid \theta_i) 1\{y = z_i\} \propto \pi(\theta_i) f(y \mid \theta_i)$

- Algorithm 2: sample space not finite
    1. generate $\theta' \sim \pi(\theta)$
    2. simulate $z_i \sim f(\cdot \mid \theta')$
    3. if $d\{s(z_i), s(y)\} < \epsilon$, set $\theta_i = \theta'$, repeat
- need to choose $s(\cdot)$, $d(\cdot, \cdot)$

... ABC

- Algorithm 2: sample space not finite
  1. generate $\theta' \sim \pi(\theta)$
  2. simulate $z_i \sim f(\cdot \mid \theta')$
  3. if $d\{s(z_i), s(y)\} < \epsilon$, set $\theta_i = \theta'$, repeat
- after N steps, $\theta_1, \ldots, \theta_N$ is a sample from

$$\pi_\epsilon(\theta \mid y) = \int \pi_\epsilon(\theta, z \mid y) dz \propto \int \pi(\theta) f(z \mid \theta) 1\{z \in A_{\epsilon,y}\} dz$$

- $\pi(\theta \mid y) \simeq \pi_\epsilon(\theta \mid y)$ if ϵ 'small enough' and $s(z)$ a 'good' summary statistic

- many improvements possible, using ideas from MCMC
- which generates samples from the posterior by sampling from a Markov chain with that stationary distribution
- many techniques for trying to ensure that sampling is from regions of Θ where $\pi(\theta \mid y)$ is large, without knowing $\pi(\theta \mid y)$

High-dimensional inference

- $f(y;\theta), y \in \mathbb{R}^n; \theta \in \mathbb{R}^p$, p large relative to n, or $p > n$
    - Aside: empirical likelihood has $p = n - 1$, yet usual asymptotic theory applies
    - Partial likelihood has $p = n - 1 + d$, yet usual asymptotic theory applies
    - "Neyman-Scott problems" have $y_{ij} \sim f(\cdot;\psi,\lambda_i), j = 1,\dots,m; i = 1,\dots,k$, so $n = km$ and $p = 1 + k$ i.e. $p/n = O(1)$ if $m \to \infty$, k fixed; ususal theory does not apply

- we concentrate on Bühlmann paper — there is a very large literature
- $y = X\beta + \epsilon, \quad \mathrm{E}(\epsilon) = 0,\ \mathrm{cov}(\epsilon) = \sigma^2 I$ — running example, $n = 71, p = 4088$

$$\hat\beta_{ridge} = \arg\min_\beta\{\frac{1}{n}(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\sum_{j=1}^p \beta_j^2,$$

$$\hat\beta_{lasso} = \arg\min_\beta\{\frac{1}{n}(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\sum_{j=1}^p |\beta_j|$$

... high-dimensional inference

$$\hat{\beta}_{ridge} = \arg\min_{\beta}\{\frac{1}{n}(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\sum_{j=1}^{p}\beta_j^2,$$

$$\hat{\beta}_{lasso} = \arg\min_{\beta}\{\frac{1}{n}(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\sum_{j=1}^{p}|\beta_j|$$

- usual to assume $\Sigma_{i=1}^{n}x_{ij} = 0, \Sigma_{i=1}^{n}x_{ij}^2 = 1$ — so units are comparable

  $\hat{\beta}_0 = \bar{y}$ is not 'shrunk'
- Lasso penalty leads to several $\hat{\beta}_k = 0$ — sparse solutions
- there are many variations on the penalty term
- λ is a tuning parameter — often selected by cross-validation

... high-dimensional inference

- Inferential goals (§2.2)

  (a) prediction of surface $X\beta$ or $y_{new} = x_{new}^{\mathrm{T}}\beta$
  (b) estimation of β
  (c) estimation of $S = \{j : \beta_j \neq 0)$ ‘support set’

- re (c): define $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$; to have $\Pr(\hat{S} \to S\} > 1 - \epsilon$, say need very strong conditions: $\min|\beta_j| > c$, $c \sim \sqrt{\log p/n}$ $\simeq 0.34$
- plus condition on design
- often replaced by (c'): ‘screening’ $\hat{S} \supset S$ with high probability

  also needs conditions on X

- can solve (b) and (c'), if $|S| << n/\log p$ and $\log p << n$
- re (a) prediction accuracy can be assessed by cross-validation
- note that (a) can be estimable even if $p > n$, as long as $X\beta$ of small enough dimension

Inference about β, $p > n$

- p-values for testing $H_{0,j} : \beta_j = 0$
- three methods suggested: multi-sample splitting, debiasing, stability selection
- multi-sample splitting: fit the model on random half, say, of observations, leads to $\hat{S}$
- use $X^{(\hat{S})}$ in fitting to the other half
- $P_j = p$-value for t-test of H_j if $j \in \hat{S}$, o.w. 1
- $P_{corr,j} = P_j \times |\hat{S}|, j \in \hat{S}$, o.w. 1
- Repeat B times and aggregate $P^b_{corr,j}$
- de-biasing $\hat{\beta}_{ridge/Lasso,corr,j} = \hat{b}_j - \text{bias}$ see paper
- Can show resulting estimate $\hat{\beta}_{ridge/Lasso,corr,j}j \sim N(\beta_j, \sigma^2_\epsilon w_j)$ w_j known
- $\hat{\beta}_{ridge/Lasso,corr,j} \neq 0$, any j, so need multiplicity correction $p = 4088$
- on their example; Lasso selects 30 features; multi-sample selects 1; bias-corrected Ridge selects 0; stability selection selects 3

Non-linear models

- example y_i independent, $\mathrm{E}(y_i) = \mu_i(\beta); g(\mu_i) = \beta_0 + x_i^{\mathrm{T}}\beta$

- Lasso-type 'mle': $\arg\min\{-\frac{1}{n}\ell(\beta, \beta_0; y) + \lambda\Sigma_{j=1}|\beta_j|\}$ $\quad \beta = (\beta_1, \ldots, \beta_p)$

- can use multi-sample splitting or stability selection

- a version of de-biasing applies to GLMs, based on weighted least squares

- a marginal approach would fit $y = \alpha_0 + \alpha_j x^{(j)}$, one feature at a time
- leading to 4088 p-values, and then need techniques for controlling FWER or FDR

- Model: $y_i = x_i^{\mathrm{T}}\beta + Z_i, \quad i = 1, \dots, n$ independent
- M-estimation:

$$\sum_{i=1}^{n} x_i \psi(y_i - x_i^{\mathrm{T}}\hat{\beta}) = 0 \tag{1}$$

- result: if ψ is monotone, and $p \log(p)/n \to 0$, and conditions on X, then

there is a solution of (1) satisfying $||\hat{\beta} - \beta||^2 = O(p/n)$

- "rows of X behave like a sample from a distribution in $\mathbb{R}^p$"

- if $p^{3/2} \log n/n \to 0$, then

$$\max |x_i^{\mathrm{T}}(\hat{\beta} - \beta)| \xrightarrow{p} 0$$

- and

$$a_n^{\mathrm{T}}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2)$$

$\sigma^2 = a_n^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}a_n \mathrm{E}\psi^2(Z)/\{E\psi'(Z)\}^2$

- Model: $y_i \sim \exp\{\theta^{\mathrm{T}} y - \psi(\theta)\}, i = 1, \dots, n$ independent; $p = p_n$
- maximum likelihood estimate $\psi'(\hat\theta_n) = \bar y_n$
- under conditions on the eigenvalues of $\psi''(\theta)$ and moment conditions on y, Fisher information matrix

$$||\hat\theta_n - \theta_n||^2 \le c\frac{p}{n}, \text{ in probability },$$

-

$$||\hat\theta - \theta - \bar y|| = O_p(p/n) \text{ if } p/n \to 0,$$

- $p^{3/2}/n \to 0$:

$$\sqrt{n} a_n^{\mathrm{T}}(\hat\theta - \theta) \xrightarrow{d} N(0,1),$$

  likelihood ratio test of simple hypothesis asymptotically χ^2_p
- "asymptotic approximations are trustworthy if $p^{3/2}/n$ is small, but may be very wrong if p^2/n is not small"
- MLE 'will tend to be' consistent if $p/n \to 0$

cf. also El Karoui et al., 2013, PNAS

Asymptotic theory, overview

- Sartori '03
  - Neyman Scott problems as above
  - profile likelihood inference okay if $p = o(n^{1/2})$
  - modified PL inference okay if $p = o(n^{3/4})$
- Portnoy '84
  - MLE "will tend to be consistent" if $p/n \to 0$
  - asymptotic approixmations okay if $p^{3/2}/n \to 0$
  - and fail if $p^2/n \to 0$
- classical $p/n \to 0$, p fixed, $n \to \infty$
- semi-classical $p_n/n \to 0$ or $p_n^{3/2}/n \to 0$ — Portnoy, Sartori
- moderate dimensions $p_n/n \to \beta$ — Sur & Candes '17
- high dimensions $p_n/n \to \infty$
- ultra-high dimensions $p_n \sim e^n$