Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 4: Superpopulation models, and maximum likelihood

UNIVERSITY OF MICHIGAN

Superpopulation Modeling: Estimating parameters

- Various principles: least squares, method of moments, maximum likelihood
- Sketch main ideas of maximum likelihood, an important approach that underlies statistical inferences for many common models:
  - Linear and nonlinear regression
  - Generalized linear models (logistic, Poission regression)
  - Repeated measures models (SAS PROC MIXED, NLMIXED
  - Survival analysis – proportional hazards models

Finite population inference

- Modeling takes a predictive perspective on statistical inference – predict the non-sampled values
  - ML models for the sampling/nonresponse weights lie outside this perspective
- Inference about parameters is intermediate step in predictive superpopulation model inference about finite population parameters

Predict non-sampled values $\hat{y}_i = E(y_i \mid \hat{\theta}), \hat{\theta}$ ML estimate of θ

Estimate of total $T = \sum_{i \in s}^{n} y_i + \sum_{i \notin s}^{n} \hat{y}_i$, etc.

- Does not reflect uncertainty in ML estimate – Bayes incorporates this by intergrating over posterior distribution of parameters (as discussed later)

Definition of Likelihood

- Data Y
- Statistical model yields probability density $f(Y|\theta)$

  for Y with unknown parameters θ
- Likelihood function is then a function of θ

$$L(\theta|Y) = const \times f(Y|\theta)$$

- Loglikelihood is often easier to work with:

$$\ell(\theta|Y) = \log L(\theta|Y) = const + \log\{f(Y|\theta)\}$$

Constants can depend on data but not on parameter θ

Example: Normal sample

- $Y = (y_1, \ldots, y_n)$ univariate iid normal sample

$$\theta = (\mu, \sigma^2)$$

$$f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \right)$$

$$\ell(\mu, \sigma^2 \mid Y) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2$$

Example: Multinomial sample

- $Y = (y_1, \ldots, y_n)$ univariate K-category multinomial sample

  n_j = number of y_i equal to j (j=1,…,K)

$$\theta = (\pi_1, ..., \pi_{K-1}); \ \ \pi_K = 1 - \pi_1 - ... - \pi_{K-1}$$

$$f(Y \mid \pi_1, ..., \pi_{K-1}) = \frac{n!}{n_1! ... n_K!} \left(\prod_{j=1}^{K-1} \pi_j^{n_j} \right) (1 - \pi_1 - ... - \pi_{K-1})^{n_K}$$

$$\ell(\pi_1, ..., \pi_{K-1} \mid Y) = \left(\sum_{j=1}^{K-1} n_j \log \pi_j \right) + n_K \log(1 - \pi_1 - ... - \pi_{K-1})$$

Maximum Likelihood Estimate

- The maximum likelihood (ML) estimate $\hat{\theta}$ of θ maximizes the likelihood, or equivalently the log-likelihood

$$L(\hat{\theta} \mid Y) \geq L(\theta \mid Y) \text{ for all } \theta$$

- The ML estimate is the

  "value of the parameter that makes the data most likely"
- The ML estimate is not necessarily unique, but is for many regular problems given enough data

Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the likelihood equation

$$S(\theta \mid Y) = 0$$

where S is the score function, defined as the first derivative of the loglikelihood:

$$S(\theta \mid Y) \equiv \frac{\partial \log L(\theta \mid Y)}{\partial \theta}$$

For some models (e.g. multiple linear regression), likelihood equation has an explicit solution; for others (e.g. logistic regression) numerical optimization methods are needed

Normal Examples

- Univariate Normal sample $Y = (y_1, ..., y_n)$ $\theta = (\mu, \sigma^2)$

$$\hat{\mu} = \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

(Note the lack of a correction for degrees of freedom)

- Multivariate Normal sample

$$\hat{\mu} = \bar{y},\ \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$$

- Normal Linear Regression (possibly weighted)

$$(y_i \mid x_{i1}, ..., x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2 / u_i)$$

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) = \text{ weighted least squares estimates}$$

$$\hat{\sigma}^2 = \text{(weighted residual sum of squares)}/n$$

Multinomial Example

$$Y = (y_1, \ldots, y_n);\ y_i \sim \text{MNOM}(\pi_1, \ldots, \pi_K)$$

$$n_j = \text{ number of } y_i \text{ equal to } j\ (j = 1, \ldots, K)$$

Likelihood Equations:

$$\frac{\partial l}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_K}{1 - \pi_1 - \ldots - \pi_{K-1}} = 0,\ j = 1, \ldots, K-1$$

Hence ML estimate is

$$\hat{\pi}_j = n_j / n,\ j = 1, \ldots, K$$

Logistic regression

$$\Pr(y_i = 1 \mid x_{i1}, \ldots, x_{ip}) = \pi_i(\beta) = \frac{\exp\left(f_i(\beta)\right)}{1 + \exp\left(f_i(\beta)\right)}$$

$$f_i(\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i \pi_i(\beta) + (1 - y_i)(1 - \pi_i(\beta)) \right)$$

ML estimation requires iterative methods like method of scoring

ML for mixed-effects models

$y_i = (y_{\text{obs},i}, y_{\text{mis},i})$: k-dimensional vector of repeated measures

$(y_i \mid X_i, \beta_i) \sim N_k(X_{1i}\alpha + X_{2i}\beta, \Sigma)$

α are fixed effects; β are random effects: $\beta_i \sim N_q(0, \Gamma)$

Missing Data Mechanism: missing at random

ML requires iterative algorithms

e.g. Harville (1977), Laird and Ware (1982), SAS Proc Mixed

- Very flexible mean and covariance structures
- Normality not a major assumption if N large, and recent programs allow for non-normal outcomes

Properties of ML estimates

- Under assumed model, ML estimate is:
  - Consistent (not necessarily unbiased)
  - Efficient for large samples
  - not necessarily the best for small samples
- ML estimate is transformation invariant
  - If $\hat{\theta}$ is the ML estimate of θ

    Then $\phi(\hat{\theta})$ is the ML estimate of $\phi(\theta)$

Large-sample ML Inference

- Basic large-sample approximation:

  for regular problems,

$$\theta - \hat{\theta} \sim N(0, C)$$

  where C is a covariance matrix estimated from the sample

  - Frequentist treats $\hat{\theta}$ as random, θ as fixed; equation defines the sampling distribution of $\hat{\theta}$
  - Bayesian treats θ as random, $\hat{\theta}$ as fixed;

    equation defines posterior distribution of θ

Forms of precision matrix

- The precision of the ML estimate is measured by C^{-1} Some forms for this are:
  - Observed information (recommended)

$$C^{-1} = I(\hat{\theta}|Y) = -\left.\frac{\partial^2 \log L(\theta|Y)}{\partial\theta\,\partial\theta}\right|_{\theta=\hat{\theta}}$$

  - Expected information (not as good, may be simpler)

$$C^{-1} = J(\hat{\theta}) = E\left[I(\hat{\theta}|Y,\theta)\right]\Big|_{\theta=\hat{\theta}}$$

  - Sandwich estimator (robust properties

$$\hat{C}* = I^{-1}(\hat{\theta})\hat{K}(\hat{\theta})I^{-1}(\hat{\theta}), \text{ where } \hat{K}(\hat{\theta}) = D_\ell(\hat{\theta})D_\ell(\hat{\theta})^T$$

Bootstrap variance estimate

- A bootstrap sample of a complete data set *S* with *n* observations is a sample of size *n* drawn with replacement from *S*
  - Operationally, assign weight w_i to unit *i* equal to number of times it is included in the bootstrap sample

$$w_1, \ldots, w_n \sim \text{MNOM}(n; 1/n, \ldots, 1/n)$$

Bootstrap distribution

- Let $\hat{\theta}^{(b)}$ be ML estimate from the *b*th bootstrap data set
- Inference can be based on the bootstrap distribution generated by values of $\hat{\theta}^{(b)}$
- In particular the bootstrap estimate is

$$\hat{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$$

with variance

$$\hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$

Asymptotic properties similar to sandwich estimator

Interval estimation

- 95% (confidence, probability) interval for scalar θ is:

  $\hat{\theta} \pm 1.96\, C^{1/2}$, where 1.96 is 97.5 pctile of normal distribution
- Example: univariate normal sample

$$I = J = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4) \end{bmatrix} \Rightarrow C = \begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix}$$

Hence some 95% intervals are:

$$\bar{y} \pm 1.96\, s/\sqrt{n} \quad \text{for } \mu$$

$$s^2 \pm 1.96\, s^2/\sqrt{n/2} \quad \text{for } \sigma^2$$

$$\ln(s) \pm 1.96\sqrt{2/n} \quad \text{for } \ln(\sigma)$$

Significance Tests

Tests based on likelihood ratio (LR) or Wald (W) statistics:

$$\theta = (\theta_{(1)}, \theta_{(2)}); \theta_{(1)0} = \text{ null value of } \theta_{(1)};\ \theta_2 = \text{other parameters}$$

$$\hat{\theta} = \text{ unrestricted ML estimate}$$

$$\tilde{\theta} = (\theta_{(1)0}, \tilde{\theta}_{(2)}); \tilde{\theta}_{(2)} = \text{ ML estimate of } \theta_{(2)} \text{ given } \theta_{(1)} = \theta_{(1)0}$$

LR statistic: $\mathrm{LR}(\hat{\theta}, \tilde{\theta}) = 2\left[\ell(\hat{\theta} \mid Y) - \ell(\tilde{\theta} \mid Y) \right]$

Wald statistic: $W(\hat{\theta}, \tilde{\theta}) = (\theta_{(1)0} - \hat{\theta}_{(1)})^T C_{(11)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)})$

$$C_{(11)} = \text{ covariance matrix of } (\theta_{(1)} - \hat{\theta}_{(1)})$$

yield P-values $P = pr\left(\chi_q^2 > D(\hat{\theta}, \tilde{\theta}) \right)$

D = LR or Wald statistic; q = dimension of θ_0

χ_q^2 = Chi-squared distribution with q degrees of freedom