Statistical Methods for Handling Incomplete Data
Chapter 3: Computation

Outline

1. Introduction: Motivation

- Interested in finding the solution that

$$\hat{\theta} = \text{ arg max}_\theta L(\theta)$$

  Often the MLE can be computed from the score equation

$$S(\hat{\theta}) = 0$$

  which is generally a system of nonlinear equations.
- How to solve the score equation?

Methods for solving nonlinear equations: $g(\theta) = 0$

1. Bisection method: Use the intermediate value theorem.
   "If g is continuous for all θ in the interval $g(\theta_1)\, g(\theta_2) < 0$. A root of $g(\theta) = 0$ lie in the interval (θ_1, θ_2)"
2. Method of false positions (or Secant method): Use a linear approximation

   $$g(\theta) \cong g(a) + \frac{g(b) - g(a)}{b - a}(\theta - a)$$

   to get

   $$\theta = \frac{a g(b) - b g(a)}{g(b) - g(a)}.$$

   Thus, the method of false positions can be defined as

   $$\theta^{(t+2)} = \frac{\theta^{(t)} g\left(\theta^{(t+1)}\right) - \theta^{(t+1)} g\left(\theta^{(t)}\right)}{g\left(\theta^{(t+1)}\right) - g\left(\theta^{(t)}\right)}.$$

3. Newton's method (Or Newton-Raphson method): Use a linear approximation of $g(\theta)$ at $\theta^{(t)}$

$$g(\theta) \cong g\left(\theta^{(t)}\right) + \left[\frac{\partial g\left(\theta^{(t)}\right)}{\partial \theta'}\right]\left(\theta - \theta^{(t)}\right).$$

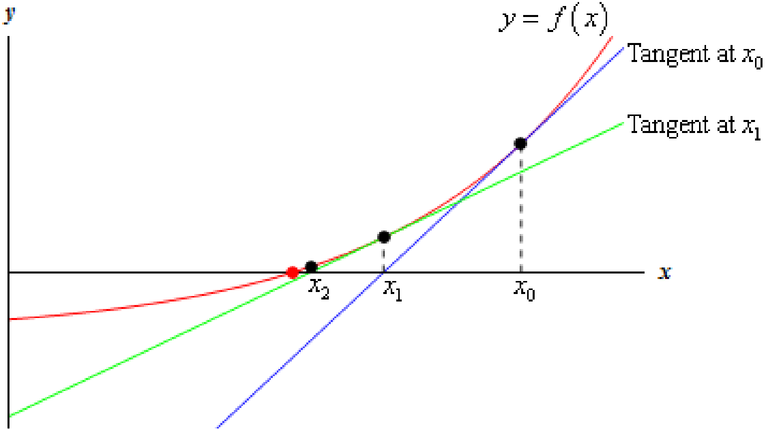
Thus,

$$\theta^{(t+1)} = \theta^{(t)} - \left[\frac{\partial g\left(\theta^{(t)}\right)}{\partial \theta'}\right]^{-1} g\left(\theta^{(t)}\right).$$

For score equation:

$$\theta^{(t+1)} = \theta^{(t)} + \left[I\left(\theta^{(t)}\right)\right]^{-1} S\left(\theta^{(t)}\right).$$

Newton's method



Other variants of Newton's method

1. Fisher scoring method: Use

$$\theta^{(t+1)} = \theta^{(t)} + \left[\mathcal{I}\left(\theta^{(t)}\right)\right]^{-1} S\left(\theta^{(t)}\right)$$

2. Ascent method:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha\left[\mathcal{I}\left(\theta^{(t)}\right)\right]^{-1} S\left(\theta^{(t)}\right)$$

   for $\alpha \in (0, 1]$. If $L(\hat{\theta}^{(t+1)}) < L(\hat{\theta}^{(t)})$, then use $\alpha = \alpha/2$ and compute $\theta^{(t+1)}$ again.

3. Quasi-Newton method:

$$\theta^{(t+1)} = \theta^{(t)} - \left[M^{(t)}\right]^{-1} S\left(\theta^{(t)}\right)$$

   where $M^{(t)}$ satisfies

$$S\left(\theta^{(t+1)}\right) - S\left(\theta^{(t)}\right) = M^{(t+1)}\left(\theta^{(t+1)} - \theta^{(t)}\right).$$

Example 3.1

Model

Logistic regression model

$$y_i \overset{i.i.d.}{\sim} Bernoulli\,(p_i)$$

with

$$\text{logit}\,(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}.$$

Log-likelihood

$$\begin{aligned} \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right] \\ &= \sum_{i=1}^{n}\left[y_i\left(\mathbf{x}_i'\boldsymbol{\beta}\right) - \ln\left(1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\right)\right] \end{aligned}$$

Example 3.1 (Cont'd)

Score function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - p_i(\boldsymbol{\beta})\} \mathbf{x}_i$$

$$I(\boldsymbol{\beta}) = -\frac{\partial}{\partial \boldsymbol{\beta}'} S(\boldsymbol{\beta}) = \sum_{i=1}^{n} p_i(\boldsymbol{\beta})\{1 - p_i(\boldsymbol{\beta})\}\mathbf{x}_i\mathbf{x}_i'$$

Newton-Raphson Method= Scoring method

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left[\sum_{i=1}^{n} p_i^{(t)}(1 - p_i^{(t)})\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \sum_{i=1}^{n}(y_i - p_i^{(t)})\mathbf{x}_i$$

where

$$p_i^{(t)} = p_i(\boldsymbol{\beta}^{(t)}).$$

Order of convergence

Definition

Let θ^* be the unique solution to $g\left(\theta\right) = 0$. A sequence $\left\{\theta^{(t)}\right\}$ is converges to θ^* of order β if

$$\lim_{t\to\infty} \|\theta^{(t)} - \theta^*\| = 0$$

and

$$\lim_{t\to\infty} \frac{\|\theta^{(t+1)} - \theta^*\|}{\|\theta^{(t)} - \theta^*\|^\beta} = c$$

for some constants $c \neq 0$.

Remark

Result

Under the regularity conditions, the sequence obtained from Newton's method converges at a second order rate.

Sketched Proof:

By the second order Taylor expansion,

$$
\begin{aligned}
0 &= g\left(\theta^{*}\right) \\
&\cong g\left(\theta^{(t)}\right)+\left\{\partial g\left(\theta^{(t)}\right)/\partial\theta\right\}\left(\theta^{*}-\theta^{(t)}\right)+\left\{\partial^{2} g\left(q\right)/\partial\theta^{2}\right\}\left(\theta^{*}-\theta^{(t)}\right)^{2}/2
\end{aligned}
$$

where q is between θ^{*} and $\theta^{(t)}$. Multiplying both sides of the above equation by $\left\{\partial g\left(\theta^{(t)}\right)/\partial\theta\right\}^{-1}$ and using the definition of the Newton method, we have

$$
\frac{\theta^{(t+1)}-\theta^{*}}{\left(\theta^{(t)}-\theta^{*}\right)^{2}}=\frac{\partial^{2} g\left(q\right)/\partial\theta^{2}}{2\partial g\left(\theta^{(t)}\right)/\partial\theta}
$$

Thus, the Lipschitz condition holds and

$$
\lim_{t\to\infty}\frac{\|\theta^{(t+1)}-\theta^{*}\|}{\|\theta^{(t)}-\theta^{*}\|^{2}}=\left|\frac{g^{''}\left(\theta^{*}\right)}{2g'\left(\theta^{*}\right)}\right|\neq 0.
$$

Statistical Methods for Handling Incomplete Data (Chapter 3.2: Factoring Likelihood Approach)

Example 3.4 (Bivariate Normal distribution)

- Model

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right]$$

- <u>Observation</u>

  r complete observations $\{(x_i, y_i)\,; i = 1, 2, \cdots, r\}$

  $n - r$ partial observations $\{x_i; i = r+1, r+2, \cdots, n\}$

  assume missing at random.

- The observed likelihood is

$$L_{obs}(\theta) = \prod_{i=1}^{r} f(x_i, y_i; \mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy}) \times \prod_{i=r+1}^{n} f(x_i; \mu_x, \sigma_{xx})$$

  Finding the MLE using direct maximization of the observed likelihood is computationally challenging.

Factoring likelihood approach (Anderson, 1957)

Idea: Use
"Joint pdf of (x, y) = (marginal pdf of x) $\times$ (conditional pdf of y given x)"
Alternative parametrization

$$\begin{aligned} X_i &\sim N(\mu_x, \sigma_{xx}) \\ Y_i \mid X_i = x &\sim N(\beta_0 + \beta_1 x, \sigma_{ee}) \end{aligned}$$

where

$$\begin{aligned} \beta_1 &= \sigma_{xy}/\sigma_{xx} \\ \beta_0 &= \mu_y - \beta_1 \mu_x \\ \sigma_{ee} &= \sigma_{yy} - \sigma_{xy}^2/\sigma_{xx}. \end{aligned}$$

Under the new parametrization,

$$\begin{aligned} L_{obs}(\theta) &= \prod_{i=1}^{n} f(x_i; \mu_x, \sigma_{xx}) \times \prod_{i=1}^{r} f(y_i \mid x_i; \beta_0, \beta_1, \sigma_{ee}) \\ &= L_1(\mu_x, \sigma_{xx}) \times L_2(\beta_0, \beta_1, \sigma_{ee}). \end{aligned}$$

Example 3.4 (Cont'd)

- The MLEs under the new parametrization are

$$\begin{aligned}\hat{\mu}_x &= \bar{x}_n \\ \hat{\sigma}_{xx} &= S_{xxn}\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}_1 &= S_{xyr}/S_{xxr} \\ \hat{\beta}_0 &= \bar{y}_r - \hat{\beta}_1\bar{x}_r \\ \hat{\sigma}_{ee} &= S_{yyr} - S_{xyr}^2/S_{xxr},\end{aligned}$$

where the subscript r denotes that the statistics are computed from the r respondents only and subscript n denotes that the statistics are computed from the whole sample of size n.

Example 3.4 (Cont'd)

- Thus, the MLE's for the original parametrization are

$$
\begin{aligned}
\hat{\mu}_y &= \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_x = \bar{y}_r + \hat{\beta}_1 \left(\hat{\mu}_x - \bar{x}_r \right) \\
\hat{\sigma}_{yy} &= S_{yyr} + \hat{\beta}_1^2 \left(\hat{\sigma}_{xx} - S_{xxr} \right) \\
\hat{\sigma}_{xy} &= S_{xyr} \frac{\hat{\sigma}_{xx}}{S_{xxr}}.
\end{aligned}
$$

- The MLE of μ_y is called the regression estimator.

Remark

- The regression estimator can be expressed as the sample mean of the best predictors of y_i:

$$\hat{\mu}_y = \frac{1}{n}\left\{\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i\right\} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$$

  where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- The asymptotic variance of the regression estimator can be shown to be

$$V(\hat{\mu}_y) \doteq \frac{1}{n}\sigma_x^2\beta_1^2 + \frac{1}{r}\sigma_e^2 = \frac{1}{n}\sigma_y^2\rho^2 + \frac{1}{r}\sigma_y^2(1-\rho^2)$$

  where $\rho = Corr(X, Y)$. (Recall Example 2.7)

Example 3.5 (Bivariate categorical distribution)

$$(Y_1, Y_2) = \begin{cases} (1,1) & \text{with prob. } \pi_{11} \\ (1,0) & \text{with prob. } \pi_{10} \\ (0,1) & \text{with prob. } \pi_{01} \\ (0,0) & \text{with prob. } \pi_{00} \end{cases}$$

<u>Observation</u>
r complete observations $\{(y_{1i}, y_{2i})\,; i = 1, 2, \cdots, r\}$
$n - r$ partial observations $\{y_{1i}; i = r+1, r+2, \cdots, n\}$

<u>Observed likelihood</u> for $\theta_1 = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$

Example 3.5 (Cont'd)

Alternative parametrization: $\theta_2 = (\pi_{1+}, \pi_{1|1}, \pi_{1|0})$ where

$$\begin{aligned} \pi_{1+} &= Pr(Y_1 = 1) \\ \pi_{1|1} &= Pr(Y_2 = 1 \mid Y_1 = 1) \\ \pi_{1|0} &= Pr(Y_2 = 1 \mid Y_1 = 0) \end{aligned}$$

Observed likelihood for θ_2

Example 3.5 (Cont'd)

<u>MLE</u>

Because we can write

$$L_{\text{obs}}(\pi_{1+}, \pi_{1|1}, \pi_{1|0}) = L_1(\pi_{1+})L_2(\pi_{1|1})L_3(\pi_{1|0})$$

for some $L_1(\cdot)$, $L_2(\cdot)$, and $L_3(\cdot)$, we can obtain the MLE by separately maximizing each likelihood component. Thus, we have

$$\begin{aligned} \hat{\pi}_{1+} &= \frac{1}{n}\sum_{i=1}^{n} y_{1i} \\ \hat{\pi}_{1|1} &= \frac{\sum_{i=1}^{r} y_{1i}y_{2i}}{\sum_{i=1}^{r} y_{1i}} \\ \hat{\pi}_{1|0} &= \frac{\sum_{i=1}^{r}(1-y_{1i})y_{2i}}{\sum_{i=1}^{r}(1-y_{1i})}. \end{aligned}$$

The MLE for π_{ij} can then be obtained by $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{j|i}$ for $i = 0, 1$ and $j = 0, 1$.

Remark

1. The factoring likelihood approach is particularly useful for *monotone missing patterns*, where we can relabel the variable in such a way that the set of respondents for each variable is monotonely nested:

$$R_1 \supset R_2 \supset \cdots \supset R_p$$

where R_i denotes the set of respondents for Y_i after relabeling. In this case, under MAR, the observed likelihood can be written as

$$L_{\text{obs}}(\theta) = \prod_{i \in R_1} f(y_{1i}; \theta_1) \times \prod_{i \in R_2} f(y_{2i} \mid y_{1i}; \theta_2) \times \cdots \times \prod_{i \in R_p} f(y_{pi} \mid y_{p-1,i}; \theta_p)$$

and the MLE for each component of the parameters can be obtained by maximizing each component of the observed likelihood (Rubin, 1974).

2. For non-monotone missing data, we cannot directly apply the factoring likelihood method.

Missingness Patterns (✓ indicates "observed")

Data	Study Variable			Monotone Missing (?)
	Y_1	Y_2	Y_3	
A	✓	✓	✓	Yes
	✓	✓		
	✓			
B	✓	✓	✓	Yes
	✓	✓		
		✓		
C	✓	✓	✓	No
	✓	✓		
		✓	✓	
D	✓	✓	✓	Yes
		✓		
		✓	✓	

REFERENCES

Anderson, R. L. (1957), 'Maximum likelihood estimates for the multivariate normal distribution when some observations are missing', *Journal of the American Statistical Association* **52**, 200–203.

Rubin, D. B. (1974), 'Characterizing the estimation of parameters in incomplete data problems', *Journal of the American Statistical Association* **69**, 467–474.