An overview of model based approaches to inference in the presence of missing data

M. Daniels

October 31, 2014

Notation

- y: full data response $(J \times 1)$
- y_{mis}: missing data
- y_{obs}: observed data
- r: missingness indicators
- x covariates (more later)

Focus

I will focus on model-based approaches

Standard classification

- Missing completely at random (MCAR):

$$p(r|y) = p(r)$$

- Missing at random (MAR):

$$p(r|y) = p(r|y_{\text{obs}})$$

- Missing not at random (MNAR):

$$p(r|y) \neq p(r|y_{\text{obs}})$$

Ignorability

- three conditions:
  - missingness is MAR
  - distinct parameters for $p(y|\boldsymbol{\theta})$ and $p(r|y,\gamma)$
  - a priori independence
- if any of these conditions are not satisfied, need to model $p(y,r|\boldsymbol{\omega})$ (full data model); but interest in

$$p(y|\boldsymbol{\omega}) = \int p(y|r;\boldsymbol{\omega})dF(r;\boldsymbol{\omega})$$

Data augmentation and the EM algorithm

- algorithms to sample from the posterior or maximize the likelihood
- not an assumption or default way to deal with missing data

Case study

- (Bayesian) Modeling of competing infectious pathogens with incomplete lab results

Data

- assume M different pathogens; here, $M = 3$, noninfluenza pathogen, influenza A, and influenza B
- observe sequential ILI (influenza like illness) onset times in infected subjects, $\tilde{t}_i$
- v_i: pathogen type responsible for ILI episode
- missingness in v_i since not all ILI episodes are tested (r_i set of missingness indicators)
- what do we observe?
  - $\{\tilde{t}_i, r_i\}$ completely observed
  - v_i partially observed
  - infection days $\hat{t}_i$ are not observable (but $\hat{t}_i < \tilde{t}_i$)
- so full data is $\{\hat{t}_i, \tilde{t}_i, v_i, r_i\}$

Components of models

- natural history of disease
  - incubation and latent period
    - assume duration of latent period is 0
    - assume infection occured at most H days before ILI onset times, $\tilde{t}$ (have no information on this)
  - infectious period (using a Beta distribution)
  - examine sensitivity to cross-immunity between pathogens
- transmission model (will examine various components here)
  - community to person transmission hazard
  - person to person transmission hazard
  - assume can only be infected by one pathogen on a given day (competing risks)

Missingness

- assume missing pathogen types are MAR (actually ignorable); so only need to specify model for the full data response
- recall full data is $\{\hat{t}_i, \tilde{t}_i, v_i, r_i\}$
- note: the observed data model is very hard to work with here

Computational highlights

- use data augmentation to fill in missing data (and create full data)
- here data augment $\hat{t}_i$ and v_i; sample all missing infection days and pathogen types simultaneously
- focus here is on missing infection types (but infection times are also missing)
- then MCMC algorithm proceeds as if full data

Study

- Pittsburgh Influenza Prevention Project (PIPP)
- randomized study of effectiveness of nonpharmaceutical interventions (NPI) conducted in 10 public elementary schools in PGH
- NPI: mostly hand hygiene (including sanitizer)
- ILI: fever (> 100) plus cough or sore throat
- 380 ILI episodes in 352 (out of 3959) students

Noingorable missingness

- extrapolation factorization:

$$p(y, r; \omega) = p(y_{mis}|y_{obs}, r; \omega)p(y_{obs}, r; \omega)$$

- unidentified parameters/sensitivity analysis/informative priors
  - not possible with fully parametric SM, $p(r|y)p(y)$ and SPM $\int p(r|b)p(y|b)p(b)db$
  - consistent with factoring (as pattern mixture model), $p(y|r)p(r)$

Simple example of sensitivity analysis

- consider a bivariate response, (y_1, y_2) with only missing in y_2
- let $r = I\{y_2 \text{ is observed}\}$
- consider a mixture model type factorization, $p(y|r)p(r)$
- now factor $p(y|r) = p(y_2|y_1, r)p(y_1|r)$
- extrapolation distribution is $p(y_2|y_1, R = 0)$
- can do sensitivity analysis on this

Simple example of sensitivity analysis (cont.)

- default: $p(y_2|y_1, R = 0) = p(y_2|y_1, R = 1)$
- try location shift or exponential tilt
- $E[Y_2|y_1, R = 0] = \delta + E[Y_2|y1, R = 1]$: put prior on δ
- $p(y_2|y_1, R = 0) = \frac{\exp(\delta y_2)p(y_2|y_1,R=0)}{\int \exp(\delta y_2)p(y_2|y_1,R=0)dy_2}$

What is a sensitivity parameter?

- if I vary it, the fit of the model to the observed data is unchanged
- if I fix it, all the parameters of the full data model are identified

How do i fit these models?

- WinBUGS, JAGS

Case Study: Estimation of (flu) vaccine efficacy using validation samples with selection bias

- estimate vaccine efficacy
  - outcome of interest only measured on a subset of study participants (validation sample)
  - typically, this sample is a convenience (not a random sample) so MAR unlikely to hold
  - Result: selection bias (MNAR)

Motivating Study I

- field study of trivalent, cold-adapted, influenza virus vaccine (CAIV-T) in Texas during 2000-2001 influenza season
- primary clinical outcome was a non-specific case definition, medically attended acute respiratory infection (MAARI)
- eligible healthy children and adolescents aged 18 months thru 18 years were offered the vaccine through specific clinics
- any individual presenting with a history of fever and any respiratory illness at the clinics were eligible to have a throat swab for the influenza virus culture

Motivating Study II

- specific case definition, *culture confirmed influenza*; that is the response of interest, y
- selection bias (for culture) concern: if physicians tend to select children whom they believe to have influenza for culturing (MNAR)

Age (years)	Vaccine Status	Children	MAARI cases	MAARI proportion	MAARI cases cultured	Num pos cult	Fract cult pos	Fract cult
1.5-4	CAIV-T	537	389	0.72	16	0	0	0.041
	None	1844	1665	0.90	86	24	0.28	0.052
5-9	CAIV-T	807	316	0.39	17	2	0.12	0.054
	None	2232	1156	0.52	118	53	0.45	0.102
10-18	CAIV-T	937	219	0.23	19	3	0.16	0.087
	None	5249	1421	0.27	123	56	0.46	0.087
Total	CAIV-T	2281	924	0.41	52	5	0.10	0.056
	None	9325	4242	0.45	327	133	0.41	0.077

Notation

- Z: vaccination indicator
- $A(z)$: indicator of MAARI for person with vaccination status z (only observe one of these)
- $Y(z)$: influenza status (biologically confirmed by culture for subset of participants)
- R: validation/missing data indicator ($R = 1$ if sampled for validation; only occurs for those with $A = 1$)
- X: age category (1.5-4 years; 5-9 years; 10-18 years)
- $P_{z,x}[A] = P[A|Z = z, X = x]$

Scientific question of interest

- causal effect of vaccination on the outcome Y
  - age-specific vaccine efficacy

$$\text{VE}_{S,x} = 1 - \frac{P_x[Y(1) = 1]}{P_x[Y(0) = 1]}$$

  - overall vaccine efficacy

$$\text{VE}_{S,x} = 1 - \frac{\sum_{x=0}^{2} P_x[Y(1) = 1]P[X = x]}{\sum_{x=0}^{2} P_x[Y(0) = 1]P[X = x]}$$

- problem: $P_x[Y(z) = 1]$ not identified from the observed data (more next)

Identifying $P_x[Y(z) = 1]$

- formulate as a selection model

$$\text{logit } P_{z,x}[R = 0|A = 1, Y = y] = h_{z,x} + \alpha_{z,x} y,$$

where

$$h_{z,x} = \log \left\{ \frac{1}{c_{z,x}} \frac{P_{z,x}[R = 0|A = 1]}{P_{z,x}[R = 1|A = 1]} \right\}.$$

- for subjects with $Z = z, X = x$ and MAARI, $\alpha_{z,x}$ is interpreted as the log odds ratio of being unvalidated for influenza vs. no influenza subjects (this is the sensitivity parameter)
- So, $\alpha_{z,x}$ positive or negative indicates that influenza subjects have lower or higher odds of being validated, respectively.

Elicitation of sensitivity parameters

- asked question: *If a physician were doing surveillance cultures during an influenza season, what is the probability that he would select the children who actually had true influenza over the children who just had nonspecific respiratory symptoms to culture?*
- for fully Bayesian analysis, construct a prior

Multiple imputation

- explicitly fill in missing responses using multiple (M) values
- why not with just one value?
- how do inference?
  - for each 'complete' dataset ($m = 1, \ldots, M$), fit the model with parameter θ to obtain an estimate $\theta^{(m)}$
  - $\bar{\theta} = \text{mean}(\theta^{(m)})$
  - then fix up the variance using

$$var(\bar{\theta}) = \frac{1}{M}\sum Var(\theta^{(m)}) + (1 + \frac{1}{M})\frac{1}{M-1}\sum(\theta^{(m)} - \bar{\theta})^2$$

- note that the limiting case of this is data augmentation

Multiple imputation

- how to impute?
- a lot of packages out there that do imputation based on a specific imputation model
- be careful especially with MICE (very popular)
- do based on a joint or sequential conditionals

SNP example

- multiple imputation of missing phenotype data for QTL mapping (Bobb et al. 2011, SAGMB)
- all done in WinBUGS

Missing covariates

- suppose interest is in a regression model, $p(y|x)$
- in what follows let r be an indicator that x is observed
- all that matters is

$$p(r|x, y) = p(r|x)$$

- if this holds, can just use complete cases
- for efficiency, need a model for $p(x)$ (only really need model for covariates that have missingness)