A new class of models for missing data

Donald B. Rubin
Department of Statistics, Harvard University

(joint work with AM Franks and EM Airoldi)

Outline

- Basic factorizations for data Y and response indicators R
- A recondite representation apparently due to JW Tukey
- Application to exponential family models
- Concluding remarks

Notation

- Complete-data vector is: $Y=(Y_1,\ldots,Y_n)^T$
- Response indicator vector: $R=(R_1,\ldots,R_n)^T$
- We treat (Y,R) as random variables (Rubin 1976) with joint i.i.d. distribution

$$P(Y,R|\theta) = \Pi_i f(Y_i,R_i|\theta)$$

- θ is the global parameter

Modeling missing data

- Two basic approaches
  - Selection model (Rubin 1974)

    Usual inferential objective is a complete-data quantity
  - Pattern mixtures (Rubin 1977; Little & Rubin 1987)

    Exposes sensitivity
- An approach apparently due to JW Tukey
  - Proposed in a discussion of Glynn, Laird & Rubin (1986) at an ETS conference on self-selected samples (Wainer ed. 1986; notes by Holland 1986)

Selection model

- Specify $P(Y_i,R_i|\theta)$ by using a distribution for the *complete* data, and a model for the missingness probability as a function of Y_i

$$f(Y_i, R_i=r\,|\theta) = f(Y_i|\theta)\, f(R_i=r|Y_i,\theta)$$

- Required specifications:
  - Complete-data distribution, $f(Y_i|\theta)$
  - Missingness mechanism, $f(R_i=r_i|Y_i,\theta)$
- $f(Y_i|\theta)$ can be highly speculative, depending on the pattern of missingness (e.g., extrapolation vs. interpolation)

Pattern-mixture model

- Specify $f(Y_i,R_i|\theta)$ as a mixture of observed data and missing data components:

$$f(Y_i, R_i=r|\theta) = f(Y_i|R_i=r,\theta)\, f(R_i=r|\theta)$$

- Required specifications:
  - Observed data distribution, $f(Y_i|R_i=1,\theta)$
  - Missing data distribution, $f(Y_i|R_i=0,\theta)$
  - Mixing weights, $f(R_i=1|\theta) = 1\text{-}f(R_i=0|\theta)$
- $f(Y_i|R_i=0,\theta)$ is highly speculative because there are no direct observations (income)

Tukey suggested model

- Specify $f(Y_i,R_i|\theta)$ by using a distribution for the *observed* data, and a model for the missingness probability as a function of Y_i

$$f(Y_i=y,R_i=r|\theta) = \frac{f(Y_i=y \mid R_i=1,\theta)}{f(R_i=1 \mid \theta)} \frac{f(R_i=r \mid Y_i,\theta)}{f(R_i=1 \mid Y_i,\theta)}$$

- Required specifications:
  - Observed-data distribution, $f(Y_i|R_i=1,\theta)$, easy
  - Missingness mechanism, $f(R_i|Y_i,\theta)$, sometimes logical, depending on the science
  - Normalizing constant, $f(R_i=1|\theta)$, easy

Remarks on Tukey's approach

- Improves on the pattern-mixture model by not requiring any specification for $f(Y_i|R_i=0,\theta)$
- Improves on the selection model by not requiring a difficult specification for $f(Y_i|\theta)$
- All distributions must be compatible

Exponential family / logistic models

- We consider models for the observed data in the following family

$$f(Y_i=y \mid R_i=1,\theta) = h(y)\, g(\theta)\, e^{T(y)'\theta}$$

with missingness mechanism

$$f(R_i=1 \mid Y_i=y,\theta) = \text{logit}\,(T(y)'\theta) = \frac{1}{1+e^{-T(y)'\theta}}$$

and implied odds of missingness

$$\frac{f(R_i=0 \mid Y_i=y,\theta)}{f(R_i=1 \mid Y_i=y,\theta)} = e^{-T(y)'\theta}$$

Tractable normalizing constant

- The normalizing constant for models in this family has a simple analytical form

$$f(R_i = 1 \mid \theta_{Y|R}, \theta_{R|Y}) = \frac{g(\theta_{Y|R} + \theta_{R|Y})}{g(\theta_{Y|R} + \theta_{R|Y}) + g(\theta_{Y|R})}$$

where $\theta=(\theta_{Y|R},\theta_{R|Y})^T$ in an obvious notation, and g(.) is the function that defines the exponential family density

Inferential strategy

- Consider a simple model where $\theta_{R|Y} = (\beta_0, \beta_1)^T$

$$f(R_i = 1 \mid Y_i = y, \theta_{R|Y}) = \text{logit}\,(\beta_0 + \beta_1 y_i) = (1 + \exp\{-\beta_0 - \beta_1 y_i\})^{-1}$$

$$f(Y_i = y \mid R_i = 1) = \text{Normal}\,(0,1)$$

with normalizing constant: $f(R_i = 1 \mid \beta_0, \beta_1) = \dfrac{\beta_1^2}{\beta_1^2 - 2e^{\beta_0}}$

- Posit prior distributions on β_1 and the function $f(R_i=1|\theta)$ and compute the implied β_0
- More generally, one can work with a multivariate logistic model, where β_1 is a vector

Theory: Missing data distribution

Theorem 2. The missing-data density can be expressed a function of the observed-data density and the odds of missingness.

$$f(Y_i = y \mid R_i = 0, \theta) = \frac{f(R_i = 1 \mid \theta)}{f(R_i = 0 \mid \theta)} \frac{f(R_i = 0 \mid Y_i = y, \theta)}{f(R_i = 1 \mid Y_i = y, \theta)} f(Y_i = y \mid R_i = 1, \theta)$$

Theorem 3. If $f(Y_i \mid R_i=1,\theta)$ is exponential family, and the log odds of missingness are linear in the natural sufficient statistics of $f(Y_i \mid R_i=1,\theta)$, then $f(Y_i \mid R_i=0,\theta)$ is the same exponential family

Concluding remarks

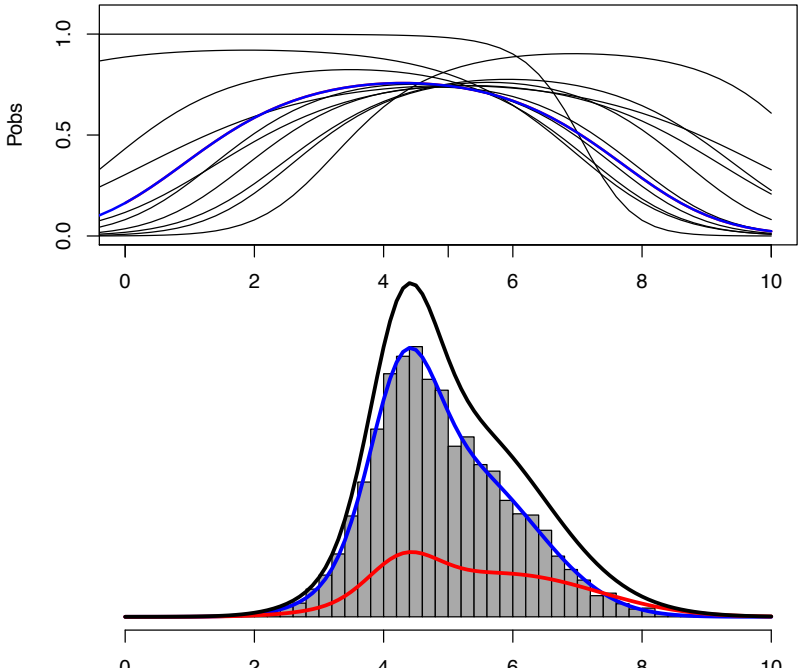
- The Tukey selection model is a fresh approach to modeling missing data
- Can be viewed as improving on the selection model by replacing a typically speculative complete-data model with an observed-data model
- Methods for exponential family hold more generally

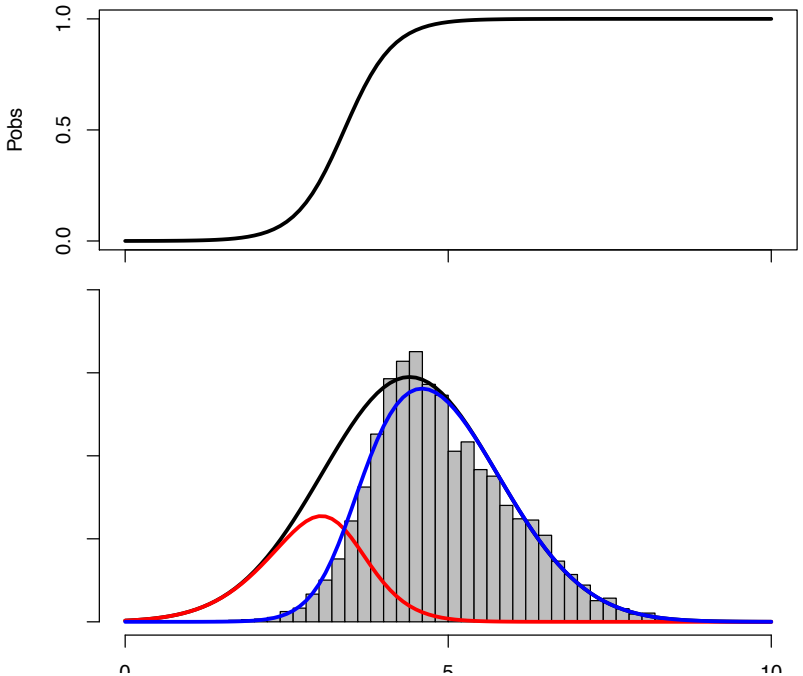
Read draft of the full paper:

http://arxiv.org/abs/1603.06045

BACK-UP

Example: Taxable income reporting



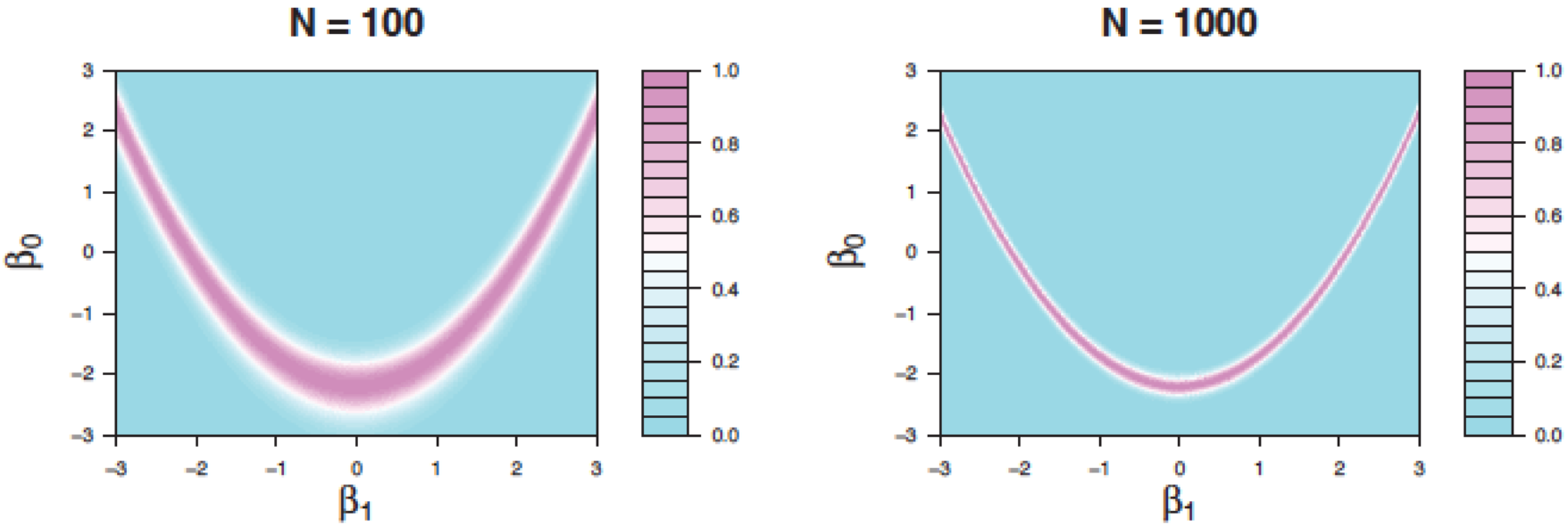


Tukey's approach

Selection factorization

Issue with naïve inference on $\theta_{R|Y}$

- As Q is estimated with higher precision, the region of positive support for the likelihood, in terms of θ_R, becomes increasingly constrained



Theory: Normalizing constant

- More generally, the normalizing constant is

$$f(R_i = 1 \mid \theta_{Y|R}, \theta_{R|Y}) = \left(1 + \int \frac{f(R_i = 0 \mid Y_i = y, \theta_{R|Y})}{f(R_i = 1 \mid Y_i = y, \theta_{R|Y})} f(Y_i = y \mid R_i = 1, \theta_{Y|R}) \, dy \right)^{-1}$$

Theorem 1. The normalizing constant Q is the population fraction of observed data.

$$\begin{aligned} \mathbb{E}[r_i \mid \theta_{Y|R}, \theta_{R|Y}] &= f(r_i = 1 \mid \theta_{Y|R}, \theta_{R|Y}) \\ &= \int f(y_i, r_i = 1 \mid \theta_{Y|R}, \theta_{R|Y}) \, dy_i \\ &= Q(\theta_{Y|R}, \theta_{R|Y}) \int f^{\text{obs}}(y_i \mid \theta_{Y|R}) \, dy_i \\ &= Q \end{aligned}$$