17. Missing Data

Patrick T. Bradshaw, Ph.D.

17 April 2017

PH 250C: Advanced Epidemiological Methods
School of Public Health
University of California, Berkeley

References

Required:

1. Greenland S, Finkle WD. A critical look at basic methods for handling missing covariates in epidemiologic regression analysis. *American Journal of Epidemiology*. 1995.
2. Ibrahim JG, Chu H, Chen M-H. Missing data in clinical studies: issues and methods. *Journal of Clinical Oncology*. 2012.

Optional:

1. Chapter 11: Missing data. Vittinghoff E, Glidden DV, Shiboski SC, McCullough CE. *Regression Methods in Biostatistics*, 2^{nd}. 2012.

References

Additional:

1. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*. 2007.

Outline

Intro to Missing Data

Missing data in epidemiology

- Missing data is impossible to avoid.
- Problems arise when subjects with incomplete data differ from those with complete data.
- Even when these groups are comparable, it is statistically not efficient to discard data.
- Methods for analysis with missing data:
  - Allow us to make most efficient use of all available data.
  - Can reduce bias in estimation.

Notation

- Outcome: Y, fully observed
  - Most common situation.
  - Issues somewhat different (generally less problematic) for missing outcome data.
- Covariates: $\mathbf{x}$, some observed: $\mathbf{x}^o$, and some missing: $\mathbf{x}^m$ for a particular subject.
- In a particular study:
  - $\mathbf{x}^o$ could be all of the sociodemographic factors on each individual you know (age, race, etc...).
  - $\mathbf{x}^m$ could be things people might not want to report (BMI, education, income).

Missing Data Classification

Missing data classification

- The process that led to the data becoming missing is important.*
- Consider r an indicator of observed/missing data.
  - $r = 1$ if data are observed, $r = 0$ if data are missing. (some authors use opposite coding)
  - r can be a vector (when you have multiple covariates that are potentially missing).
- The types of factors that determine the probability of observed/missing ($\Pr[r = 1]$) drives the analysis.

*Little and Rubin *Statistical Analysis with Missing Data*, 2^{nd}. 2002.

Missing Completely at Random

Missing completely at random (MCAR):

- Probability of being observed/missing *does not* depend on any data.

$$P(R = 1|y, \mathbf{x}) = P(\overset{R}{\cancel{r}} = 1)$$

  - *e.g.* Laboratory error, lost data, patient moves for no particular reason.
  - Observed data are a random sample.

- Effectively reduces sample size (loss of efficiency, but no bias).

Its not a big deal if the missing data is small (<10%).

Missing at Random

Missing at Random (MAR):

- Probability of being observed/missing depends only on *observed* data. You can predict the data that is missing based on something we know (i.e., age, other missing data...)

$$P(R = 1|y, \mathbf{x}) = P(R = 1|y, \mathbf{x}^o)$$

Given what I know I can predict why data is missing

  - *e.g.* Older individuals less likely to report certain behaviors; interviewers at certain centers less likely to press for answers.
  - Data are a random sample *conditional on observed variables*.
- In general, MAR $\implies$ bias* and loss of efficiency.

I can do something about variables if they are missing at random

*There are special cases where it doesn't.

Not Missing at Random

Not Missing at Random (NMAR)[*]:

- Probability of being observed/missing depends on the missing data (and possibly observed data).

$$P(R = 1|y, \mathbf{x}) = P(R = 1|y, \underline{\mathbf{x}^m}, \mathbf{x}^o)$$ Dependent on missing data

Problematic if data is missing due to competing risks, unmeasured/unobserved covariates, etc (non-random)

  - *e.g.* social desirability bias/sensitive measures; longitudinal studies where ability/willingness to report outcome depends on how sick you are.
  - Most problematic analytically (high potential for bias, definite loss of efficiency).
    - Proceed cautiously!

[*] Also referred to a Missing Not at Random.

Monotone vs. Non-monotone missing data

Monotone missing data: a hierarchy of missingness.

Pattern	Hypothetical Monotone Y	X_1	X_2	X_3	Hypothetical Non-monotone Y	X_1	X_2	X_3
1	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs
2	Obs	Obs	Obs	M	Obs	Obs	Obs	M
3	Obs	Obs	M	M	Obs	Obs	M	M
4	Obs	M	M	M	Obs	Obs	M	Obs
5					Obs	M	Obs	Obs
6					Obs	M	M	Obs

Monotone missing data pattern, there's a pattern in missing data (if x1 is missing, x2 and x3 are missing). So you can predict what data will be missing if x1 is missing

Non-monotone, there's no way to reorder patterns

Figure 1. Monotone and nonmonotone patterns of missingness (Obs = observed, M = missing)

From: Horton and Kleinman. Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models. *Am Stat*. 2007.

Monotone vs. Non-monotone missing data

- Monotone patterns of missingness can make for simpler models.
    - Rare in practice.

Analysis Methods for Missing Data

Ad-hoc methods

Generating a "missing" category will clump every observation (various levels of exposure) into one category. It's like clumping everyone who did not report BMI (all potential BMI categories) into one (unadjusted) category to then compare to the other (adjusted) BMI categories which makes the inference all bad!!

***Ad-hoc* missing data methods:**

- Add indicator variable for missing category.
- Replace missing value with that variable's mean/predicted value (e.g. regression-based).
- Replace missing value with observation randomly chosen from the data.
- Last observation carried forward (for longitudinal studies).

These appear intuitive, but no theory to justify their use.*

---

*Greenland S. and Finkle WD. A critical look at methods for handling missing data in epidemiologic regression analyses. *American Journal of Epidemiology*. 1995. 142(12):1255-64.

Ad-hoc methods

- Possible to find some cases where they are unbiased, but not in general. This method doesn't perform well
- Can actually induce bias and *reduce* precision.
- Greenland and Finkle (1995): Examples simulated in Greenland paper
  - Ordinary missing indicator:
    - CI coverage abysmal! ($\sim$ 40% vs 95%).
    - Bias in OR ($\sim$ 1.5 vs. 2).
    - Higher RMSE than other methods.
  - Regression imputation:
    - Potential for suboptimal CI coverage ($\sim$ 70%) and large bias (mean OR $>$ 3 vs. 2)! (see esp. Table 4).

Ad-hoc methods

Great resources for methods to correct for missing data

- For discussion/details:
  - Greenland and Finkle (1995).
  - Vach W and Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad-hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol* 1991. 134: 895-907.

Principled Methods

If you're going to predict missing values, it's important to describe the uncertainty of the values

- Acknowledge inherent uncertainty in this process.
- Model the missing data mechanism (R) and/or missing covariates ($\mathbf{X}^m$) (allows us to understand the properties of these methods).
- Methods today will assume data is MAR (or MCAR).
  - Many can be extended to data that is NMAR.
  - If you suspect NMAR, statistical consultation good idea.

Missing At Random (MAR)

Not Missing At Random (NMAR)

Principled Methods

We will consider:

1. Complete case analysis.
2. Inverse-probability of missing weighting.
3. Maximum likelihood. (briefly)
4. Multiple imputation.
5. Bayesian.

Missing data methods are really a sensitivity analysis because you never really know the mechanism of how data are missing (MAR or NMAR?)

Complete case analysis

Complete case (CC) analysis: The whole observation is removed

- Omits each record if any variable is missing.
- Reasonable if small percentage of data is missing ($< 10\%$) and sample is large. Small sample of 200, losing 20 may be significant
- Automatic in most software packages. Always check your # of observations in the output!
- Estimators unbiased if data is MCAR, or MAR but $P[R = 1]$ not dependent on outcome.
    - *But* efficiency suffers.

Inverse Probability Weighting

Inverse probability weighting (IPW):

- *A.k.a.* Weighted estimating equation (WEE) method.
- Weight each observation by the inverse of the probability that it was observed (conditional on covariates).
- We already encountered this approach in the selection bias lecture. Correct for selection bias

* model the missing data mechanism as function of observed data and use to predict probability of missing; use predicted prob to calculate weight that is then applied to the observation
* think through who gets up-weighted and who gets down-weighted

Inverse Probability Weighting

Steps (one version):

1. Fit model for probability of being observed ($R = 1$) as a function of observed data (including outcome Y):

   Modeling the probability that is observed

   $$\text{logit}(\Pr[R = 1|\mathbf{x}^o, y]) = \mathbf{x}^o\gamma + y\delta$$

   Can only be used for MAR data (because this method doesn't depend on missing data)

2. Estimate predicted probability of being observed ($\hat{p}$) conditional on observed data and calculate weights*:

   $$w = 1/\hat{p}$$

3. Fit regression model for outcome among complete cases, weighted by w, using robust standard errors.

*Or stabilized weights by including marginal probability of R in numerator.

Inverse Probability Weighting

Pros

- Valid inference in large samples.
- Avoids specifying parametric model for missing covariates (unlike ML, MI, Bayes).
- Simple and can be executed in most any software package.

Cons

- Small sample properties probably not good.
- Not ideal for multiple missing variables (unless they are all missing simultaneously).
- Less statistically efficient than alternatives.

Not Patrick's go-to method because it lacks flexibility

Maximum Likelihood

Maximum Likelihood (ML):

- Recall: MLE maximizes the likelihood (based on the joint distribution of outcomes (Y) conditional on covariates ($\mathbf{x}$) and parameters (β)) *w.r.t.* β:

$$L(\beta|y,\mathbf{x}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i,\beta)$$

  (*a.k.a.* **complete data likelihood**). Likelihood assuming we had all the data
- Treats all $\mathbf{x}$ as fully observed and fixed (non-random).
  - When some $\mathbf{x}$'s are missing then we can treat them as random and incorporate their distribution into the model.

Maximum Likelihood

- With missing covariate data, **x** no longer fixed.
- Now require specification of *joint density* of outcome and missing covariate(s), which we often factor as:

Beta is the parameter on the outcome given the observed data

Now we have uncertainty around the models (joint probability of x and y)

$$p(y_i, \mathbf{x}_i^m | \beta, \alpha, \mathbf{x}^o) = \underbrace{p(y_i | \mathbf{x}_i^m, \mathbf{x}_i^o, \underline{\beta})}_{\text{density of } y \text{ conditional on (observed) covariate}} \overbrace{p(\mathbf{x}_i^m | \mathbf{x}_i^o, \underline{\alpha})}^{\text{from distribution of covariate}}$$

* we factor out y -- y can't show up in model for x

* selection model factorization

Alpha is the parameter on the distribution of covariates

where α are parameters that index the covariate distribution.

- Likelihood specified as before, plus a model for the distribution of the variables with missing data as a function of observed data: $p(\mathbf{x}^m | \mathbf{x}^o, \alpha)$.

* we model x to impute missing covariate values

Maximum Likelihood

- The **complete data likelihood** (likelihood if no missing data) would then be:

$$L(\beta, \alpha | y, \mathbf{x}^o, \mathbf{x}^m) = \prod_{i=1}^{n} p(y_i, \mathbf{x}_i^m | \mathbf{x}_i^o, \beta, \alpha)$$

(but there is missing data...)

Maximum Likelihood

- Integrate (sum) the individual contributions over the possible values of $\mathbf{x}^m$.
- The **observed data likelihood** is then[*]:

$$\widetilde{L}(\beta, \alpha | y, \mathbf{x}^o) = \prod_{i=1}^{n} \int p(Y_i | \mathbf{x}_i^m, \mathbf{x}_i^o, \beta) \, p(\mathbf{x}_i^m | \mathbf{x}_i^o, \alpha) \, d\mathbf{x}^m.$$

which we maximize with respect to α and β.

Like taking an expectation = your taking an average of alpha and beta

[*] If $\mathbf{X}^m$ are discrete then the integral is a sum.

Maximum Likelihood

- Very computationally demanding:
  - Integral/sum in the middle of the likelihood function.
    - Notoriously hard to code.
  - Very specialized: not usually found in commercial software except very specific/simplistic examples.
  - Particularly difficult to obtain standard errors for $\hat{\beta}$.
- Details of ML estimation in missing data outside scope of this course.
- These concepts are used in other approaches we will discuss (MI, Bayesian).

Multiple imputation

"Proper imputation"

Multiple imputation (MI): estimating the missing values multiple times; reduces bias and characterizes variability.

Steps:

* alphas are the model parameters in the equation that models missingness?

1. Specify distributions for the missing covariates ($p(\mathbf{x}^m|\mathbf{x}^o, y, \alpha)$) and use them to predict the missing values in your dataset. Repeat M times.
2. Analyze each of these M datasets as you normally would.
3. Combine the results to get an overall parameter estimate and quantify its uncertainty.

Instead of predicting from regression "improper imputation", you use the values in your data set to predict the posterior predictive distribution (averages prediction models across a range of parameters) (Bayesian principle)

M=5 look like 5 full datasets with no missing data. Take the mean and then use that to make inferences

Multiple imputation

- Distributon of $\mathbf{X}^m$ a function of *all* observed data: $\mathbf{x}^o$ and y.[*]
- These aren't simply predictions from regression models ($\Rightarrow$ *improper imputation*).
    - *Regression imputation* method from Greenland and Finkle (1995).

[*] Moons et al. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*. 2006.

Multiple imputation

- For *proper imputation* sample from posterior predictive distribution:

$$p(\mathbf{x}^m|\mathbf{x}^o, y) = \int p(\mathbf{x}^m|\mathbf{x}^o, y, \alpha)\, p(\alpha)\, d\alpha$$

  (no longer conditional on α).
  - Requires prior on α. (Bayesian principles!)
- Draw sample of $\mathbf{x}^m$ from this distribution.

Multiple imputation

- Use these values to fill in missing data, creating M "complete" datasets.
- Size of M should be large enough to characterize the uncertainty (commonly $M = 5$ to 20).
    - Should increase with larger fraction of missing data, continuous $\mathbf{X}^m$.

Multiple imputation

- Pooled estimate of regression parameters:

$$\hat{\beta} = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}^{(j)}$$

- And its covariance matrix: $\mathbf{V}_{MI}$.*
- Likelihood and deviance don't translate into imputation framework.
  - No likelihood ratio tests!

    You've averaged M likelihood tests so you cant use likelihood ratio test after using MI, use a Wald test

*See Ibrahim et al. (2005) *JASA* for details.

Bayesian Methods

Bayesian methods for missing data: also referred to as "Fully Bayesian" approach. Best for NMAR variables

- Standard Bayesian analysis specifies distributions for outcome Y (sampling distribution), and priors on the model parameters (e.g. βs).
- Just one extra step when covariates are missing: specify distributions for each $\mathbf{X}^m$ (and corresponding priors).
- Very straightforward to implement once you know general Bayesian methods.

Bayesian Methods

- Very flexible–can specify very general missing data structures easily.
  - Can estimate models for NMAR data much easier than ML.
- Deep connections to MI and ML approaches.
  - MI was derived from Bayesian principles.
  - Bayesian methods use the observed data likelihood (sampling distribution) in the expression of the posterior.

Steps:

1. Specify model for sampling distribution of Y:

$$p(y|\mathbf{x}^m, \mathbf{x}^o, \beta)$$

Calculate beta, the probability of the outcome given the observed variables

2. Specify model for sampling distribution of missing covariates $\mathbf{X}^m$ as a function of observed covariates $\mathbf{x}^o$:

$$p(\mathbf{x}^m|\mathbf{x}^o, \alpha)$$

Calculate alpha, the probability of the missing values given the observed variables

3. Specify prior distributions on model parameters α and β: $p(\alpha, \beta)$.
4. Characterize posterior distribution of β and α by sampling from

$$p(\beta, \alpha|y, \mathbf{x}^o) \propto \underbrace{\int p(y|\mathbf{x}^m, \mathbf{x}^o, \beta)\, p(\mathbf{x}^m|\mathbf{x}^o, \alpha)\, d\mathbf{x}^m}_{\text{Observed data likelihood}}\; p(\alpha, \beta)$$

Example

Example

Simulated data (cohort study, $N = 1500$):

- Covariate: $Z \sim \text{Binomial}(1, 0.5)$.
- Exposure: $X \sim \text{Binomial}(1, 0.5)$.
- Outcome: $Y \sim \text{Binomial}(1, p_y)$ with

$$\text{logit}(p_y) = -1 + X - Z.$$

Example

```
require("blm")
require("geepack")
require("mi")
require("R2jags")
require("coda")

set.seed(111404)
N <- 1500 # Number of observations
Z <- rbinom(N,1,.5) # Binomial, p=0.5
px <- .5
X <- rbinom(N,1,px)
py <- expit(-1 + X - Z)
Y <- rbinom(N,1,py)
```

Sample Z from a binomial distribution

Expit is the inverse of the log function

Example

Imposed missingness:

Distributed binomially with the probability of something being observed (1) and the probability of something being missing (1-p)

- Indicator of being observed: $R \sim \text{Binomial}(1, 1 - p_m)$ with

$$\text{logit}(p_m) = -1 + .5Z - 5Y.$$

This is not dependent on the value of X

- Generates about 25% missing.

This gives the probability of something being missing

Example

Imposing missingness

```
p.miss <- expit(-1 + .5*Z - 5*Y) # Probability missing
R <- rbinom(N,1,1-p.miss) # Generate indicator of observed =1

X.miss <- X
X.miss[R==0] <- NA # If not observed set to missing (NA)
```

Wherever R is not observed, set it to NA (missing)

Example

- We assume data are MAR (probability of missing not dependent on unobserved data).
- Analyses:
  1. Complete case analysis.
  2. Inverse-probability weighting.
  3. Multiple imputation.
  4. Bayesian modeling.

Example

```
##### "True" analysis
summary(glm(Y~X+Z, family = binomial(link="logit")))

##### Complete-case analysis
summary(glm(Y~X.miss+Z, family = binomial(link="logit")))
```

Throws out any observation that is missing

Example: IP Weighting

1. Fit logistic regression of R on Z and Y.
2. Estimate predicted probability of $R = 1|Z, Y$: $\hat{p}_r$ and calculate IP weights $w = 1/\hat{p}_r$. 1/the probability of being observed
3. Estimate logistic regression of Y on X and Z, with robust standard errors.

Example

```
##### Inverse probability weighting
# Model for observed/missing
model.r <- glm(R~Z+Y, family=binomial)
# Predicted probability of observed
phat.r <- predict(model.r, type="response")
w <- 1/phat.r # Weight according to probability of being observed
w[R==0] <- 1 # Replace weight on unobserved obs.
id <- 1:length(Y)

data.cc <- na.omit(as.data.frame(cbind(Y,X.miss,Z,w,id)))

summary(geeglm(Y ~ X.miss + Z, family=binomial(link="logit"),
     weights = w, id=id,
     data=data.cc, std.err='san.se', corstr="exchangeable"))
```

Change NA to 0: If you leave R==0 as NA, the output is weird

Example: Multiple Imputation

Will use the `mi` package in R:*

1. Create $M = 20$ complete datasets filling in missing values for X^m from the posterior predictive distribution using a logistic regression of X^o on Z and Y.
   - Conduct quality checks for imputed data.
2. Analyze each of these datasets separately with a logistic regression of Y on X and Z:

$$\text{logit}\left(\Pr[Y = 1|X, Z]\right) = \beta_1 + \beta_2 X + \beta_3 Z$$

3. Calculate the pooled estimate of $\hat{\beta} = \frac{1}{M}\sum_m \hat{\beta}^{(m)}$ and calculate its standard error.

---

*See Su Y-S, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Software* 2011. 45(2).

Example: Multiple Imputation

Notes: In `mi` package:*

- Data frame should contain only variables in your model.
  - Extraneous variables (e.g. id's, variables from IPW, etc...) can cause problems.
- Always check the assumptions that the package makes on your variables:
  - Family (distribution), link, model.
  - Imputation method (allows proper and improper methods).
  - Transformations on dependent variables.
- Check for convergence (same mean for all variables across all M imputed datasets).

*See also `mi_vignette.pdf` (*An Example of mi Usage*) in optional readings folder.

Example

Multiple imputation

```
data.all <- as.data.frame(cbind(Y,X.miss,Z,w,id))

# Keep only variables in analysis
to.drop <- names(data.all) %in% c("w","id")

# Convert to missing data frame
mdf <- missing_data.frame(data.all[!to.drop])
```

(MI wants all the data in the same place)

```
# Examine patterns of missing data
# and verify distributional assumptions
show(mdf)
summary(mdf)
```

"Show" tells you the number of missing data patterns. Make sure the variable type (e.g., binary), method for imputation is ppd (posterior predicted distribution), and model (e.g., logit for binary outcome) is what you want it to be

Summary will help confirm the number of missing data (just the one variable you are looking at)

Example

```
# Create 20 samples from posterior predictive distributions
# for missing variables:
imputations <- mi(mdf, n.iter=50, n.chains=20)
```

20 is the number of imputations

```
# Check convergence:
# Means should be same across chains for each variable
round(mipply(imputations, mean, to.matrix = TRUE), 3)
```

Give me the mean of all these variables rounded to 3 decimal places

```
# Rhats should be very close to 1:
Rhats(imputations)
```

The x variable should be approx the same across all chains (imputations)
NOTE: The output gives 1.00 values when the output should be binary, but just take the value and subtract 1

Example

* MI = multiple imputation; connection to bayesian methods

Pooling MI for logistic regression on outcome

```
# Individually analyze and pool data:
analysis <- pool(Y~X.miss+Z, data=imputations, family=binomial)
summary(analysis)
```

* imputed data set -- in imputed data frame -- you run your analysis

Example: Bayesian Modeling

Will use JAGS through R, as before.

1. Specify sampling distribution for Y and X as binomial random variables with 1 trial, with success probabilities defined as:

$$p_y = \text{logit}\left(\Pr[Y = 1|X, Z]\right) = \beta_1 + \beta_2 X + \beta_3 Z$$

   and

$$p_x = \text{logit}\left(\Pr[X = 1|Z]\right) = \alpha_0 + \alpha_2 Z.$$

* if multiple missing x's -- can layer in more models for those easily in Bayesian context

* want the Z to be as complex as it is in the model for Y -- it usually looks the same as in the Y model

2. Specify vague prior distributions on parameters β and α.

* we could also specify informative priors -- vague priors is essentially like maximum likelihood

3. Sample from the posterior distribution of $[\beta, \alpha]$ and calculate summary statistics (mean, median, credible intervals).

Example: Bayesian Modeling

Notes:

- Missing data models are more richly parameterized–will probably need to run for a lot of iterations. (price for flexibility)
- Like ML, working with joint distribution of Y and $\mathbf{X}$: beware trying to specify non-identified conditional distributions:

* have to factor it out in sensible way:

$$p(Y,X) \neq p(Y|X)\,p(X|Y)$$

(in ML and FB you can't include outcome in model for missing covariate.)

Example

* same as how we did it in bayesian lectures, but now we're just estimating two models (one for y and another for x)

```
logistic.model <- function() {
  # SAMPLING DISTRIBUTION
  for (i in 1:N) {
    logit(p[i]) <- b[1] + b[2]*X[i] + b[3]*Z[i];
    Y[i] ~ dbin(p[i],1);
  }
```

* binomially distributed with success prob p[i]

```
    # DISTRIBUTION ON COVARIATE WITH MISSING DATA:
    logit(p.x[i]) <- a[1] + a[2]*Z[i];
    X[i] ~ dbin(p.x[i],1);
  }
```

* logistic regression bc x is binary

* probability of x here (not y as above)

```
  # PRIORS ON BETAS
  b[1:N.y] ~ dmnorm(mu.b[1:N.y],tau.b[1:N.y,1:N.y])
```

* predictors in model for y = 3

```
  # PRIORS ON ALPHAS
  a[1:N.x] ~ dmnorm(mu.a[1:N.x],tau.a[1:N.x,1:N.x])
}
```

* predictors in model for x = 2

Example

```
N <- length(Y) # Number of observations
N.y <- 3 # Number of slope parameters in model for Y
N.x <- 2 # Number of slope parameters in model for X^m

# Data, parameter list and starting values
mu.b <- rep(0,N.y)
tau.b <- diag(0.001,N.y)

mu.a <- rep(0,N.x)
tau.a <- diag(0.001, N.x)

data.logistic <- list("N", "N.y", "N.x", "Y", "X", "Z",
          "mu.b", "tau.b", "mu.a", "tau.a")
parameters.logistic <-c("b","a") # Parameters to keep track of
inits.logistic <- function() list (b= rep(0,N.y, a=rep(0,N.x)))
```

* make sure to count in the intercept term here

* vague priors for both mean and precision for mode parameters

* things that don't change -- considered "data"

* give starting values of 0s -- maybe also do for some positive and negative values

* we don't make inference on the alphas but we still want to make sure that they converge so we still keep track of them

Example

```
set.seed(114011)
logistic.sim<-jags(data=data.logistic,
          inits=inits.logistic,parameters.logistic,n.iter=50000,
          n.burn=25000, model.file=logistic.model, n.thin=5,
          n.chains = 3)

print(logistic.sim,2)

# Convergence diagnostics
logistic.mcmc <- as.mcmc(logistic.sim)

plot(logistic.mcmc)
autocorr.diag(logistic.mcmc)

geweke.diag(logistic.mcmc)
```

* convert simulation output into mcmc object for diagnostics

Example: Results

* stata has good ways to do multiple imputation; for fully bayes --> R is best

Table 1: β coefficients and standard errors with probability of X missing: $\Pr[R = 1|Z, Y] = -1 + .5 * Z - 5 * Y$

* introduce random missingness into full data using this model

* truth = how data was generated

Parameter (truth)	Full	Under 25% Missing X CC	IPW	MI	FB
β_1 (= -1)	-1.21 (0.11)	-0.83 (0.12)	-1.18 (0.12)	-1.20 (0.10)	-1.21 (0.11)
β_2 (=1)	0.95 (0.12)	0.88 (0.13)	0.88 (0.13)	0.87 (0.11)	0.95 (0.12)
β_3 (= -1)	-0.75 (0.12)	-0.65 (0.13)	-0.74 (0.13)	-0.73 (0.11)	-0.75 (0.12)

* full is just one sample, so not exactly truth

* Full = total sample -- no missingness -- ideally would want to get back to these estimates
* CC --> throw out observations with missing (only uses 75% of sample) -- not too diff. but not great
* IP weighting --> improvement in intercept and beta3
* MI is closer (and close to IPW), more precise; Bayesian handling of missingness gives the exact same answers

- All similar in estimate of association of X.
- Bayesian approach produced closest estimates for slope parameters.

* people usually do these repeatedly and then look at avg. performance

- Can't really tell the properties with just 1 simulation.
- These all rely on correct specification of each model.

* different fields deal with missing data in diff. ways

* IPW -- models probability of x missing, not the x variable itself (like MI and FB) -- if don't know much about x var itself but do about why it's missing, IPW may make more sense

General Guidelines

Guidelines

1. Try not to have missing data!
2. Assess extent of missingness.
3. Determine if MCAR, MAR or NMAR are reasonable.
   - Sensitivity analysis!
4. Select appropriate strategy (IPW, MI, Bayesian, others...).
5. Make imputation model (for $\mathbf{X}^m$) as richly parameterized as possible.
   - Should be at least as parameterized (ideally more) than outcome model.
6. Sensitivity analysis:
   - Always do a complete-case analysis for comparison.
   - Vary covariates in models.
   - If NMAR likely, consider more sophisticated methods (selection model)–consult statistician.

* missing at random
* not missing at random
* Bayes is best if significant missing issues
* in terms of predictors of x

Summary

Summary

- Principled approach to missing data can improve inference.
  - Avoid *ad-hoc* methods.
- Gains in efficiency (precision) and bias often realized.
- Analyses in presence of missing data are always subject to untestable assumptions.
  - Sensitivity analyses important.
- *Always* better to have complete data.

* missing at random -- still throw out some of your sample; so don't get bias, but lose some precision -- depends on how much missingness you have
* get sense of how sensitive results are to assumptions we're making

17. Missing Data

Patrick T. Bradshaw, Ph.D.

17 April 2017

PH 250C: Advanced Epidemiological Methods
School of Public Health
University of California, Berkeley