15. Missing Data

Patrick T. Bradshaw, Ph.D.

PBHLTH 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. *Am J Epidemiol*. 1995.
2. Ibrahim JG, Chu H, Chen M-H. Missing data in clinical studies: issues and methods. *J Clin Oncology*. 2012.

Optional:

1. Chapter 11: Missing data. Vittinghoff E, Glidden DV, Shiboski SC, McCullough CE. *Regression Methods in Biostatistics*, 2nd. 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. *Am Stat.* 2007.

Recent series in *AJE*

1. Perkins NJ, Cole SR, Harel O, *et al.* Principled approaches to missing data in epidemiologic studies. *Am J Epidemiol.* 2018; 187(3):568-575.
2. Harel O. Mitchell EM, Perkins NJ, *et al.* Multiple imputation for incomplete data in epidemiologic studies. *Am J Epidemiol.* 2018; 187(3):576-584.
3. Sun BL, Perkins NJ, Cole SR, *et al.* Inverse-probability-weighted estimation for monotone and non-monotone missing data. *Am J Epidemiol* 2018. 187(3):585-591.

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 (we use capital Y to demonstrate we 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.

standard missing data notation

- 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.* We often can't make a strong assertion about why something is missing or not but can make an educated guess.
- 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) ← check
  - 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*, 2nd. 2002.

Missing Completely at Random

"M-CAR"

Missing completely at random (MCAR):

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

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

  - *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).

Missing at Random

Missing at Random (MAR):

- Probability of being observed/missing depends only on *observed* data.

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

You could explain why it's missing based on what you know.

  - *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.

*There are special cases where it doesn't.

Not Missing at Random

("N-MAR", or sometimes "M-NAR")

Not Missing at Random (NMAR)*:

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

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

- *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.

Analysis Methods for Missing Data

Ad-hoc methods

***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.*

- Can induce bias and *reduce* precision (see ref).

*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.

Principled Methods

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

Really in the realm of sensitivity analysis at this point.

Principled Methods

We will consider:

1. Complete case analysis.
2. Maximum likelihood. (briefly)
3. Multiple imputation.
4. Bayesian.

Also *Inverse-probability of missing* weighted estimators (will not cover).

Complete case analysis

Rule of thumb:
10% you should really be doing something with your missing data.

Complete case (CC) analysis:

- Omits each record if <u>any</u> variable is missing.
- Reasonable if small percentage of data is missing ($< 10\%$) and sample is large.
- Automatic in most software packages.
- Estimators unbiased if data is MCAR, or MAR but $P[R = 1]$ not dependent on outcome.
  - *But* efficiency suffers.

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:

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

This is the likelihood you wish you had.

(*a.k.a.* **complete data likelihood**).

- 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, $\mathbf{x}$ no longer fixed.
- Now require specification of *joint density* of outcome and missing covariate(s), which we often factor as:

outcome

missing x

parameters for the y model

fully observed pts

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

per individual

parameters for the x^m model

probability of missing data given the observed data and $\propto$ (x^m model)

times

model for Y_i given x^m, x^o and β (Y model)

all our data as is

basically average this over all possible values. (max. likelihood)

where a are parameters that index the covariate distribution.

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

Maximum Likelihood

- So, the **complete data likelihood** (likelihood if you knew the values of the 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 you don't know the values of the missing data…)

$$X^m = (0,1) = P(Y | X^m = 0, X^o) P(X_m = 0 | X^o) + P(Y | X^m = 1, X^o) P(X_m = 1 | X^o)$$

P(Y if $X_m = 0$) times $P(X^m = 0)$ given that person's covariates

P(Y if $X_m = 1$) times $P(X_m = 1)$ given that person's covariates.

Maximum Likelihood

- With missing data, we 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.
- Computationally demanding and very difficult to code.
  - Details outside the scope of this course.
  - These concepts are used in other approaches we will cover (MI, Bayesian).

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

Multiple imputation

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

Steps:

1. Specify distributions for the missing covariates ($p(\mathbf{x}^m|\mathbf{x}^o, y, a)$) 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.

the trick is figuring out what M needs to be.

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* (which would be *improper imputation*).
  - *Regression imputation* method from Greenland and Finkle (1995).

or in a survival model, you could put the Nielsen-Aalen estimator of the cumulative hazard in the model in place of the outcome.

*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.
- 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 $\mathbf{X}^m$.

you're averaging over the distribution of the alphas. (smoothing out the influence of the α).

some say it should match the percentage of missing data.

Multiple imputation

- Pooled estimate of regression parameters:

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

Average of the $\hat{\beta}$s over the M samples.

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

But can do Wald tests.

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

Bayesian Methods

Bayesian methods for missing data: also referred to as "Fully Bayesian" approach.*

- 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.

Consider the covariates with missing values as parameters

When you have MNAR, this tends to be the best way to handle.

*Austin PC, Escobar MD. Bayesian modeling of clinical data in medical research. *Computational statistics & data analysis.* 2005. 49: 821-836.

Bayesian Methods

- Very flexible–can specify very general missing data structures easily.
- Fundamental 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.

Bayesian Methods

Steps:

1. Specify model for sampling distribution of Y:

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

i.e. a logistic regression - what we've been doing all along.

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

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

← the fully observed people inform αs.

use vague priors

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

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

The missing data got integrated over.

Example

Example

Simulated data ($N = 1000$):

- Covariate: $Z \sim \mathrm{N}(0, 0.5)$.
- Exposure: $X \sim \mathrm{Binomial}(1, p_x)$, with

$$\mathrm{logit}(p_x) = -2 + 3z.$$

- Outcome: $Y \sim \mathrm{Binomial}(1, p_y)$, with

$$\mathrm{logit}(p_y) = -1 + 1.5x - 3z.$$

← true model for outcome Y.

Example

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

########## DATA GENERATION
set.seed(111404)
N <- 1000 # Number of observations
Z <- rnorm(N, 0, .5)
px <- expit(-2 + 3*Z)
X <- rbinom(N, 1, px)
py <- expit(-1 + 1.5*X - 3*Z)
Y <- rbinom(N, 1, py)
```

Example

This is missing dependent on things that are fully observed (MAR).

Imposed missingness:

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

$$\text{logit}(p_m) = -1 - 2\,Y + 5\,Z.$$

- Generates about 1/3 missing.

[$R \sim \text{Binomial}(1, .5)$ would be 50% MCAR.]

Example

```r
p.miss <- expit(-1 - 2*Y + 5*Z)
R <- rbinom(N, 1, 1-p.miss) # Generate indicator of observed =1
table(R)
mean(R) # Proportion observed

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

Usually, life does this for you.

Example

- We assume data are MAR (probability of missing not dependent on unobserved data).
- Analyses:
  1. Complete case analysis.
  2. Multiple imputation.
  3. 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")))
```

by default
it throws out
any datas created
to have NAs in Patrick's prior step.

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 (using functions within `mi` package).

*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

```r
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])

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

Set up your df to only include the variables you need.

This is where you start telling R you're trying to address missing data.

This shows you the assumptions, which you might want to modify (check package documentation).

Example

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

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

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

→ M

This is like the $\hat{R}$ the coda package gives you for Bayesian analyses.

Example

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

You can give this the same sort of options you would give the glm command.

All 20 (M) datasets are now stored in this dataset.

Example: Bayesian Modeling

Will use JAGS through R, as before. ← except now we add a model for X. miss.

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

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

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

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

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

because of the way we're factoring the joint likelihood.

$(P \mid y, x) = P(y \mid x) P(x)$

can't have y in here.

Example

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

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

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

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

model for y, what you would have already done before

now adding a model for X

you would need to add a model for every X with missing data

Need priors for both model parameters.

Example

```r
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.miss", "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)))}
```

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)
```

Example: Results

Table: Regression coefficients (standard errors) with ~33% X missing.

(If data weren't missing)

simulation based on truth

Parameter (truth)	Full	CC	MI	FB
β_1 (= -1)	-1.05 (0.10)	-0.76 (0.11)	-1.06 (0.10)	-1.07 (0.10)
β_2 (=1.5)	1.42 (0.23)	1.73 (0.31)	1.63 (0.30)	1.68 (0.30)
β_3 (= -3)	-2.95 (0.22)	-2.45 (0.25)	-2.96 (0.23)	-2.99 (0.23)

- All similar in estimate of association of X.
- Bayesian approach produced closest estimates for slope parameters.
  - Can't really tell the properties with just 1 simulation.
  - These all rely on correct specification of each model.

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 (MI usually sufficient, Bayesian is flexible, IPW [not covered] may be useful).
5. Make imputation model (for $\mathbf{X}^m$) as richly parameterized as possible.
   (Same number of terms)
   - 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.

15. Missing Data

Patrick T. Bradshaw, Ph.D.

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