BIOSTATISTICS WORKSHOP:
MISSING DATA

Sub-Saharan Africa CFAR meeting
July 18, 2016
Durban, South Africa

Ideal World

- All datasets would be complete
  - Everyone will have filled in all the questions correctly
  - Everyone will have sent in all their questionnaires
  - All blood samples will make their way to the lab in time
  - All genotype data will have passed QC processes
  - No one will have a diagnosis date before their birth date
  - No men would be listed as having been pregnant

  - All researchers would have their own biostatistician to work with

2

Real World

- All datasets have issues (eh, no one's perfect)
  - People skip questions
  - Questionnaires are missing
  - We run out of blood samples
  - We have a QC process for a reason
  - Mistakes will happen

  - My inbox is overflowing

3

Missing data is a fact of life

- How you handle it matters
  - Need to consider the type of missingness
  - Different methods yield biased and/or inefficient estimates

"All Models are Wrong, but Some are Useful"

- There is no magic bullet
  - ...other than avoiding missing data at the design stage
  - Be aboveboard about limitations of your approach

George Box, PhD, 1919 - 2013

4

Missing data is a fact of life

- Ignore missing data: "Complete Case analysis"
  - Biased & Inefficient in all situations
  - Exception is for large samples sizes and very small amounts of missing data
    - Still biased and less efficient but not as noticeable because of sample size
- All alternative approaches have their own strengths and weaknesses
  - Dependent on type of missingness

5

Missing Data Definitions

Missing Completely At Random (MCAR)	$\Pr(\mathbf{M} \mid \mathbf{X}_{miss}, \mathbf{X}_{obs}) = \Pr(\mathbf{M})$
Missing At Random (MAR)	$\Pr(\mathbf{M} \mid X_{miss}, X_{obs}) = \Pr(\mathbf{M} \mid \mathbf{X}_{obs})$
Missing Not At Random (MNAR) a.k.a. "non-ignorable" or "informative"	$\Pr(\mathbf{M} \mid \mathbf{X}_{miss}, \mathbf{X}_{obs}) = \Pr(\mathbf{M} \mid \mathbf{X}_{miss}, \mathbf{X}_{obs})$

Where M = missing indicator (1=missing, 0=non-missing)
X_{miss} = missing values
X_{obs} = observed values

6

Missing Completely at Random (MCAR)

- P(M=1 | Xobs, Xmiss) = P(M=1)
  - Probability that X is missing is unrelated to the value of X or any other covariate

- Dropped lab sample
- Storm on day of clinic visit
- 2 pages of a questionnaire stuck together
- More?

7

Missing at Random (MAR)

- $P(M=1 \mid X_{obs}, X_{miss}) = P(M=1 \mid X_{obs})$
  - Probability that X_1 is missing is related to an OBSERVED value of another covariate X_2
  - After adjusting for the observed value X_2, X_1 is not associated with M

- Age/Income
  - Older age groups more likely to answer income question than younger age groups
  - Older age groups tend to make higher incomes
  - So overall average is inflated (if only look at non-missing)
  - Within age group. income level not related to missingness
    - So can control for age group to deal with missingness

8

Missing Not at Random (MNAR)

- $P(M=1 \mid X_{obs}, X_{miss}) = P(M=1 \mid X_{obs}, X_{miss})$
  - Probability that X is missing is related to an unknown/missing value

- Heavy drug users are less likely to report their drug use than light users
- So heavy users will have more missing values and
- Therefore overall average will be deflated
- So probability of missing drug use is related to higher frequencies of use

9

Missing Data

- Type of missing
  - MCAR - Missing Completely at Random
  - MAR – Missing at Random
  - MNAR – Missing Not at Random
- There may be different types of missingness in one dataset

- No one method is perfect
- There is no one method that fits every situation

- So now what?

Method	Advantages	Disadvantages
Complete case	Easy	Generally biased if data are not MCAR* Inefficient
Missing indicator	Easy for one variable A little more efficient	Biased Difficult for more than one variable
Weighted	Unbiased if data are MAR and missingness model correctly specified Point estimation easy Can be quite efficient**	Estimating standard errors can be difficult Can be inefficient**
Single imputation	Easy Can be unbiased in important situations (e.g. under the null) Can be quite efficient**	Generally biased Estimating standard errors can be difficult Can be inefficient**
Maximum likelihood	Unbiased if missingness model correctly specified (even for MNAR) Can be more efficient	Very difficult to implement

*Unbiased if missingness probability is "multiplicative" [Kleinbaum Morgenstern and Kupper (1981)]

**Loss of information depends on how accurately missing data can be predicted given observed data

11

Method	Advantages	Disadvantages
Complete case	Easy	Generally biased if data are not MCAR* Inefficient
Missing indicator	Easy for one variable A little more efficient	Biased Difficult for more than one variable
Weighted	Unbiased if data are MAR and missingness model correctly specified Point estimation easy Can be quite efficient**	Estimating standard errors can be difficult Can be inefficient**
Single imputation	Easy Can be unbiased in important situations (e.g. under the null) Can be quite efficient**	Generally biased Estimating standard errors can be difficult Can be inefficient**
Maximum likelihood	Unbiased if missingness model correctly specified (even for MNAR) Can be more efficient	Very difficult to implement

*Unbiased if missingness probability is "multiplicative" [Kleinbaum Morgenstern and Kupper (1981)]

**Loss of information depends on how accurately missing data can be predicted given observed data

12

Complete Case

- Limit dataset to only those subjects with **NO** missing data

- Issues with complete case analyses
  - Decrease sample size
  - Waste work, information, time
  - In most situations, this is biased

13

Complete Case

- "But we will only be dropping a few, what's the big deal?"

- A few here, a few there adds up fast.
- In studies with lots of covariates… lets think
  - If we were missing only 0.5% of each X (uncorrelated)

    - 1 outcome, 4 markers (X_1, X_2, X_3, X_4)
      - We would expect to be missing 1.9% of our data

    - 1 outcome, 100 markers (0.5% missing each)
      - We would expect to be missing 39% of our data

14

Complete Case

- MCAR – Missingness unrelated to any known or unknown variable
  - Unbiased
  - Loss of efficiency, especially in cases of large missingness
- MAR – Missing related to a measured variable
  - If related *only* to disease and/or exposure – as long as missingness is multiplicative then unbiased
  - If related to some measured covariate, adjusting for covariate should elevate any most bias
  - Lose efficiency in all cases
- MNAR – Missing related to some unmeasured/unknown or a measured but missing variable
  - Complete Case analysis will produce biased results!

Dementia and Memory Loss in HIV

- Ideal World: I created this dataset with n=1000 people (reality)
- Real World: I used this 'reality' dataset to make 3 'real' datasets with missingness
  - MCAR – missingness is not associated with anything
  - MAR – missingness is associated with age
  - MNAR – missingness is associated with an unknown variable

- Collect information on
  - **Score on memory test (continuous: higher is better)**
  - Age (continuous)
  - Clinic
  - Size of household (continuous)
- Model: Linear Regression
  - Memory Score = size_hh + age + clinic

Reality (n=1000)

```
> model1 <- lm(full$score ~ full$size_hh + full$age + full$clinicf)
> summary(model1)

Call:
lm(formula = full$score ~ full$size_hh + full$age + full$clinicf)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.6101  -5.4566  -0.1408   5.1860  22.4898 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          62.83261    3.57477  17.577  < 2e-16 ***
full$size_hh          0.47718    0.12041   3.963 7.93e-05 ***
full$age             -0.32426    0.04877  -6.649 4.87e-11 ***
full$clinicfclinic 2 -1.29629    0.69192  -1.873 0.061297 .  
full$clinicfclinic 3 -2.38112    0.63911  -3.726 0.000206 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.963 on 995 degrees of freedom
Multiple R-squared:  0.1217,	Adjusted R-squared:  0.1182 
F-statistic: 34.46 on 4 and 995 DF,  p-value: < 2.2e-16
```

Complete Case analysis

	Reality (n=1000)		MCAR		MAR		MNAR	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size_hh	0.48 (0.12)	10^{-5}						
Age	-0.32 (0.05)	10^{-11}						
Clinic 1	1.0							
Clinic 2	-1.29 (0.69)	0.06						
Clinic 3	-2.38 (0.64)	0.0002						

MCAR (n=553)

Missing
- size_hh (351)
- Age (148)
- Clinic (0)

missing at least 1 variable = 447 (45%)

with complete data = 553 (55%)

```
> modelMCAR <- lm(MCAR$score ~ MCAR$size_hh + MCAR$age + MCAR$clinicf)
> summary(modelMCAR)

Call:
lm(formula = MCAR$score ~ MCAR$size_hh + MCAR$age + MCAR$clinicf)

Residuals:
     Min       1Q   Median       3Q      Max
-26.5155  -5.4784  -0.2174   5.1320  22.4442

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           62.02601    4.88806  12.689  < 2e-16 ***
MCAR$size_hh           0.47424    0.16737   2.833  0.00477 **
MCAR$age              -0.30554    0.06603  -4.628 4.62e-06 ***
MCAR$clinicfclinic 2 -1.12189    0.95792  -1.171  0.24204
MCAR$clinicfclinic 3 -2.64161    0.85610  -3.086  0.00213 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.106 on 548 degrees of freedom
  (447 observations deleted due to missingness)
Multiple R-squared:  0.1199,    Adjusted R-squared:  0.1135
F-statistic: 18.67 on 4 and 548 DF,  p-value: 2.124e-14
```

Complete Case analysis

	Reality (n=1000)		MCAR (n=553)		MAR		MNAR	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size_hh	0.48 (0.12)	10^{-5}	0.47 (0.17)	0.005				
Age	-0.32 (0.05)	10^{-11}	-0.30 (0.07)	10^{-6}				
Clinic 1	1.0		1.0					
Clinic 2	-1.29 (0.69)	0.06	-1.12 (0.96)	0.24				
Clinic 3	-2.38 (0.64)	0.0002	-2.64 (0.86)	0.002				

Notice:
- Betas are pretty close to reality
- SEs are larger
- p-values less significant

MAR (n=638)

Missing
- size_hh (362)

with complete data = 638 (64%)

```
> modelMAR <- lm(MAR$score ~ MAR$size_hh + MAR$age + MAR$clinicf)
> summary(modelMAR)

Call:
lm(formula = MAR$score ~ MAR$size_hh + MAR$age + MAR$clinicf)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.273  -5.691  -0.031   5.340  22.202 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           63.64856    4.58087  13.894  < 2e-16 ***
MAR$size_hh            0.52447    0.15479   3.388 0.000747 ***
MAR$age               -0.35122    0.06327  -5.551 4.18e-08 ***
MAR$clinicfclinic 2 -1.51119    0.89527  -1.688 0.091907 .  
MAR$clinicfclinic 3 -1.86610    0.83169  -2.244 0.025193 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.2 on 633 degrees of freedom
  (362 observations deleted due to missingness)
Multiple R-squared:  0.1247,	Adjusted R-squared:  0.1192 
F-statistic: 22.55 on 4 and 633 DF,  p-value: < 2.2e-16
```

Complete Case analysis

	Reality (n=1000)		MCAR (n=553)		MAR (n=638)		MNAR	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size_hh	0.48 (0.12)	10^{-5}	0.47 (0.17)	0.005	0.53 (0.15)	0.0007		
Age	-0.32 (0.05)	10^{-11}	-0.30 (0.07)	10^{-6}	-0.35 (0.06)	10^{-8}		
Clinic 1	1.0		1.0		1.0			
Clinic 2	-1.29 (0.69)	0.06	-1.12 (0.96)	0.24	-1.51 (0.90)	0.09		
Clinic 3	-2.38 (0.64)	0.0002	-2.64 (0.86)	0.002	-1.87 (0.83)	0.03		

Notice:
- Betas are pretty close-ish to reality*
- *missingness is associated with age, so by controlling for age we help alleviate the bias introduced by missingness
- SEs are larger
- p-values less significant

MNAR (n=890)

Missing
- size_hh (110)

with complete data = 890(89%)

```
> modelMNAR <- lm(MCAR$score ~ MNAR$size_hh + MNAR$age + MNAR$clinicf)
> summary(modelMNAR)

Call:
lm(formula = MCAR$score ~ MNAR$size_hh + MNAR$age + MNAR$clinicf)

Residuals:
     Min       1Q   Median       3Q      Max
-26.4587  -5.5548  -0.0569   5.2132  23.1123

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           66.56135    3.78372  17.591  < 2e-16 ***
MNAR$size_hh           0.31682    0.12764   2.482 0.013241 *
MNAR$age              -0.36247    0.05193  -6.979 5.81e-12 ***
MNAR$clinicfclinic 2 -1.50011    0.72933  -2.057 0.039994 *
MNAR$clinicfclinic 3 -2.29868    0.67288  -3.416 0.000664 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.922 on 885 degrees of freedom
  (110 observations deleted due to missingness)
Multiple R-squared:  0.1193,    Adjusted R-squared:  0.1153
F-statistic: 29.97 on 4 and 885 DF,  p-value: < 2.2e-16
```

Complete Case analysis

	Reality (n=1000)		MCAR (n=553)		MAR (n=638)		MNAR (n=890)	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size_hh	0.48 (0.12)	10^{-5}	0.47 (0.17)	0.005	0.53 (0.15)	0.0007	0.31 (0.13)	0.01
Age	-0.32 (0.05)	10^{-11}	-0.30 (0.07)	10^{-6}	-0.35 (0.06)	10^{-8}	-0.36 (0.05)	10^{-12}
Clinic 1	1.0		1.0		1.0		1.0	
Clinic 2	-1.29 (0.69)	0.06	-1.12 (0.96)	0.24	-1.51 (0.90)	0.09	-1.50 (0.73)	0.04
Clinic 3	-2.38 (0.64)	0.0002	-2.64 (0.86)	0.002	-1.87 (0.83)	0.03	-2.30 (0.67)	0.0007

Notice: even with the lease amount of missingness
- Betas are biased for size_hh
- SEs are similar because we are only missing ~ 10% of the data
- p-values less significant for biased estimates

Summary

Argumentum ad antiquitatem?
(proof from tradition)

"But Mom, everyone is doing it!"

- Ok, we get it – Complete Case is bad!
- Complete Case:
  - Only good when little missingness AND
  - Missingness is MCAR or MAR (correctly modeled)
- So what can we do?

Method	Advantages	Disadvantages
Complete case	Easy	Generally biased if data are not MCAR* Inefficient
Missing indicator	Easy for one categorical variable A little more efficient	Biased Difficult for more than one variable
Weighted	Unbiased if data are MAR and missingness model correctly specified Point estimation easy Can be quite efficient**	Estimating standard errors can be difficult Can be inefficient**
Single imputation	Easy Can be unbiased in important situations (e.g. under the null) Can be quite efficient**	Generally biased Estimating standard errors can be difficult Can be inefficient**
Maximum likelihood	Unbiased if missingness model correctly specified (even for MNAR) Can be more efficient	Very difficult to implement

*Unbiased if missingness probability is "multiplicative" [Kleinbaum Morgenstern and Kupper (1981)]

**Loss of information depends on how accurately missing data can be predicted given observed data

26

Indicator Method – Simple Example

- Outcome: Memory Score
- Exposure: Size of household
- Confounders
  - Age (continuous)
  - Clinic (categorical)

- In this case only clinic has missing values

- Define clinic as 1/2/3/missing using dummy variables
- Model: score = size + age + c2 +c3 +cm
  - Those missing clinic value will be included as their own 'clinic'

	C2	C3	CM
Clinic 1	0	0	0
Clinic 2	1	0	0
Clinic 3	0	1	0
missing	0	0	1

Indicator Method

	Reality (n=1000)		MAR (n=818) Complete Case		MAR (n=1000)	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size_hh	0.48 (0.12)	10^{-5}	0.53 (0.13)	0.00007	0.46 (0.12)	0.0001
Age	-0.32 (0.05)	10^{-11}	-0.30 (0.06)	10^{-6}	-0.37 (0.05)	10^{-14}
Clinic 1	1.0		1.0		1.0	
Clinic 2	-1.29 (0.69)	0.06	-1.53 (0.84)	0.07	-1.71 (0.83)	0.04
Clinic 3	-2.38 (0.64)	0.0002	-2.33 (0.68)	0.001	-2.08 (0.67)	0.002
					-0.45 (0.80)	0.576

Notice:
- Beta for size_hh is biased when complete case is used
- Including all n=1000 with indicator for missing clinic helps alleviate the bias, but only because it is MAR associated with age (observed)
- MNAR would be biased even with indicator

Indicator Method - Issues

- For multivariate models
  - Indicator is created for every covariate, X, with any missing
  - Best used with only categorical Xs, but can make a continuous into categorical and then make a group for missing X
  - Need to be wary
    - Look for variation in the outcome in the missing levels for each covariate
      - Need at least 1 case and 1 control for every level
      - If not, subjects missing this value must be deleted
    - Look for 'perfect' missingness
      - groups of variables missing (pregnant men)
      - i.e. food frequency questionnaire
      - Can use 1 missing indicator variable

Method	Advantages	Disadvantages
Complete case	Easy	Generally biased if data are not MCAR* Inefficient
Missing indicator	Easy for one variable A little more efficient	Biased Difficult for more than one variable
Weighted	Unbiased if data are MAR and missingness model correctly specified Point estimation easy Can be quite efficient**	Estimating standard errors can be difficult Can be inefficient**
Single imputation	Easy Can be unbiased in important situations (e.g. under the null) Can be quite efficient**	Generally biased Estimating standard errors can be difficult Can be inefficient**
Maximum likelihood	Unbiased if missingness model correctly specified (even for MNAR) Can be more efficient	Very difficult to implement

*Unbiased if missingness probability is "multiplicative" [Kleinbaum Morgenstern and Kupper (1981)]

**Loss of information depends on how accurately missing data can be predicted given observed data

Inverse Probability Weighting (IPW)

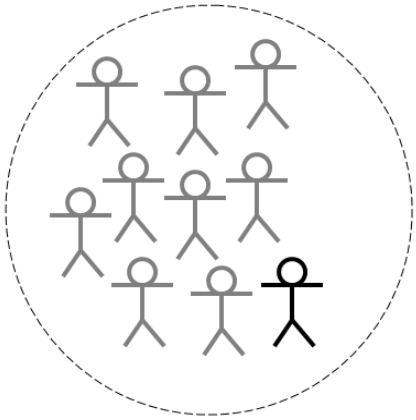
- Basic premise
  - Given the complete observed dataset
  - The sample is re-weighted to recreate the best estimate of the unobserved full & complete data
- Simple example
  - Y = Outcome (diagnosis of dementia)
  - X = Exposure (clinic)
  - Z = Confounder/covariate (age)

31

Why does IPW work?

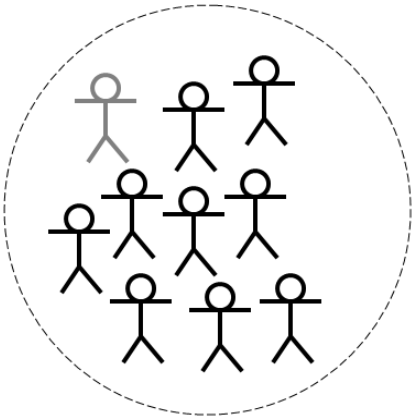
M=1 (Z unobserved)

M=0 (Z observed)



Group 1: X,Y identical = X_1,Y_1

One subject with M=0 has to "stand in" for all 10 subjects, hence weight = 1/.1 = 10



Group 2: X,Y identical = X_2,Y_2

Nine subjects with M=0 have to only "stand in" for 10 subjects, hence weight = 1/.9 = 10/9

KEY ASSUMPTION: Within each group the distribution of Z is identical for subjects with M=1 and those with M=0. I.e. conditional on X and Y, Z is independent of M. I.e. Z is "missing at random."

Method	Advantages	Disadvantages
Complete case	Easy	Generally biased if data are not MCAR* Inefficient
Missing indicator	Easy for one variable A little more efficient	Biased Difficult for more than one variable
Weighted	Unbiased if data are MAR and missingness model correctly specified Point estimation easy Can be quite efficient**	Estimating standard errors can be difficult Can be inefficient**
Single imputation	Easy Can be unbiased in important situations (e.g. under the null) Can be quite efficient**	Generally biased Estimating standard errors can be difficult Can be inefficient**
Maximum likelihood	Unbiased if missingness model correctly specified (even for MNAR) Can be more efficient	Very difficult to implement

*Unbiased if missingness probability is "multiplicative" [Kleinbaum Morgenstern and Kupper (1981)]

**Loss of information depends on how accurately missing data can be predicted given observed data

33

Imputation and Likelihood

- The literature is HUGE!

- The goal of today is to give an overview
- Examples and terminology

- Little RJA and Rubin DB (2002) Statistical Analysis with Missing Data. Hoboken: Wiley Interscience. Chapters 1, 3-5.
- Harrell FE (2001) Regression Modeling Strategies. New York: Springer. Chapters 3 and 8.
- Steyerberg EW (2009) Clinical Prediction Models. New York: Springer. Chapters 7 and 8.
- Greenland S and Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol Dec 15;142(12):1255-64.
- SAS PROC MI manual or R "MI" package
- http://www.lshtm.ac.uk/msu/missingdata/biblio.html

Imputation

- Concept:
  - Replace missing values (covariates) with a value derived from the data
    - Select at random
    - Probability (Expected value based on complete data)
  - Single imputation
    - Impute once
    - Analyze as if completed data were observed
  - Multiple imputation
    - Impute multiple times
    - Analyze each imputed data set as if completed data were observed
    - Appropriately summarize results across data sets

35

Single Imputation

Observed Data

	d	x1	x2
1	0	1.147	NA
2	1	-0.101	0.108
3	1	0.308	NA
4	0	0.267	NA
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	NA
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

Impute once →

Completed Data

	d	x1	x2
1	0	1.147	0.073
2	1	-0.101	0.108
3	1	0.308	0.366
4	0	0.267	0.980
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	0.432
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

Analyze once →

Results

Analyze as if completed data were observed

Multiple Imputation

Observed Data

	d	x1	x2
1	0	1.147	NA
2	1	-0.101	0.108
3	1	0.308	NA
4	0	0.267	NA
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	NA
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

Impute Multiple times →

Multiple Complete Datasets

	d	x1	x2
1	0	1.147	1.052
2	1	-0.101	0.108
3	1	0.308	0.708
4	0	0.267	5.786
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	0.886
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

	d	x1	x2
1	0	1.147	0.073
2	1	-0.101	0.108
3	1	0.308	0.366
4	0	0.267	0.980
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	0.432
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

	d	x1	x2
1	0	1.147	2.171
2	1	-0.101	0.108
3	1	0.308	0.565
4	0	0.267	0.810
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	0.766
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

	d	x1	x2
1	0	1.147	0.171
2	1	-0.101	0.108
3	1	0.308	0.567
4	0	0.267	1.220
5	1	-1.290	1.800
6	1	0.662	1.091
7	1	0.686	3.002
8	0	-0.099	1.790
9	0	0.850	0.548
10	0	0.335	2.717

Analyse Multiple times →

Results

Analyze each imputed data set as if completed data were observed; appropriately summarize results across data sets

Imputation

- Both methods require user to specif ... ed data
- Lot's of assumptions
  - "close enough"
  - Fudging
- All component
  - Model fo
  - Mo
  - ... ng we like to do)

$$\frac{\ldots \mathbf{X}^{miss}, \mathbf{X}^{obs}, Y)\Pr(Y \mid \mathbf{X}^{miss}, \mathbf{X}^{obs}), \Pr(\mathbf{X}^{miss}, \mathbf{X}^{obs})}{\ldots \mathbf{X}^{miss} \Pr(\mathbf{M} \mid \mathbf{X}^{miss}, \mathbf{X}^{obs}, Y)\Pr(Y \mid \mathbf{X}^{miss}, \mathbf{X}^{obs}), \Pr(\mathbf{X}^{miss}, \mathbf{X}^{obs})}$$

This slide is only here to tell you that imputation isn't as simple as it appears

Caveat

"The idea of imputation is both **seductive** and **dangerous**. It is seductive because it can lull the user into the pleasurable state of believing the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."

Little and Rubin pg 59

39

Single Imputation (4 methods)

- Unconditional Mean
- Unconditional Draw
- Conditional Mean
- Conditional Draw

- Unconditional vs. Unconditional
  - Unconditional: Do not use other variables to 'help' imputation
  - Conditional: Use other variables to 'help' imputation
- Mean vs. Draw
  - Mean: Set missing X to the mean of non-missing
  - Draw: Set missing X to a random draw from non-missing distribution

40

Unconditional mean imputation

- How:
  - Find mean of all non-missing values
  - Replace all missing values with that mean

$$X_{ij}^{(observed)} \sim N(\overline{X}_j, s_j^2)$$

- Advantage:
  - easy
- Disadvantage:
  - underestimates the amount of variability in Xj, and
  - weakens any associations with the other Xs and the outcome Y.

- It's the missing indicator method without the missing indicator

Unconditional draw imputation

- How:
  - Find the mean and SD of all non-missing values
  - Take a random sample from a distribution with that mean and SD

$$X_{ij}^{(observed)} \sim N(\overline{X}_j, s_j^2)$$

- Advantage:
  - easy,
  - a little better at handling variability in Xj
- Disadvantage:
  - still underestimates the amount of variability in Xj, and
  - still weakens any associations with the other Xs and the outcome Y.

Conditional Mean Imputation

- How: Let's say X_1 has missing values
  - Using complete data model: $X_1 = X_2 + X_3 + \ldots + X_k$ (do NOT outcome!)
  - Using that model, 'predict' all the missing X_1s
  - Repeat for all possible combinations of missingness

$$X_{i3}^{(imputed)} = \hat{\alpha} + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$$

- Advantages:
  - Maintains efficiency (use all data)
  - Good for MCAR and MAR
- Disadvantages:
  - Not easy, especially when complicated patterns of missingness

Important note: this is the one imputation approach where one **CANNOT** use outcome to predict missing data values

It will create an association where none really exits

Conditional Draw Imputation

- How:
  - Same as Conditional Mean except include a variance term
  - This time you are drawing at random from a distribution, rather than selecting the 'predicted' value

$$X_{i3}^{(imputed)} \sim N(\hat{\alpha} + \hat{\beta} Y_i, \hat{s}^2)$$

- Advantages:
  - Reintroduces variability in the imputed Xs, so less likely to introduce to much bias
- Disadvantages:
  - Not easy, especially when complicated patterns of missingness

Multiple Imputation

- So basically:
  - Impute M datasets (impute missing values)
  - Yields M β estimates $\beta_1 \ldots \beta_M$
  - Final β estimate is mean of $\beta_1 \ldots \beta_M$

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

45

Multiple Imputation

- So basically:
  - And the variance is.....

$$V_\beta = \frac{1}{M}\sum_{j=1}^{M}\hat{\sigma}^{2(j)} + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\beta}^{(j)}-\hat{\beta}\right)^2\right) = A + \left(1+\frac{1}{M}\right)B$$

$$\boxed{A = \frac{1}{M}\sum_{j=1}^{M}\hat{\sigma}^{2(j)}} \qquad \boxed{B = \left(\frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\beta}^{(j)}-\hat{\beta}\right)^2\right)}$$

46

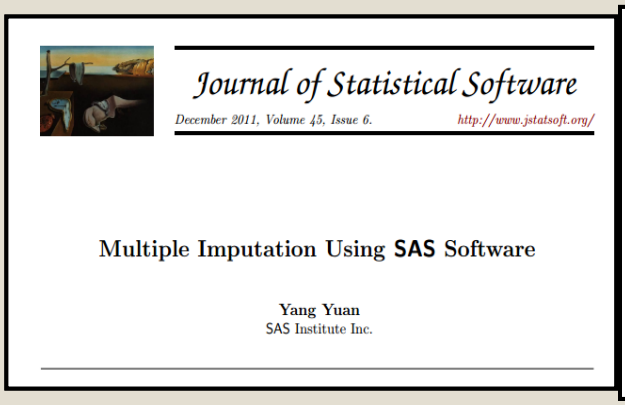
Multiple Imputation

- We want to impute the values for any variable missing in record i using all the observed data on i
- This gets difficult when different people have different missing data patterns—
  - e.g. you have to fit different models for X_3 on Y,X_1,X_2 and X_3 on Y,X_1 and X_3 on X_2 and X_3 on Y
- Ideally you'd want to fit one model for the joint distribution of all the variables, using all available data, even the incomplete records
  - This is what PROC MI (SAS) and 'mi' package (R) does, although at a price
    - it assumes the variables [or some simple transformations of the variables] are multivariate normally distributed
  - It does this via Markov Chain Monte Carlo methods

Multiple Imputation

- "Monte Carlo" refers to estimating properties of distribution (mean, variance, etc.) using repeated draws from the distribution
  - Want to know if a coin is fair? Flip it 1,000 times and count the number of heads
- "Markov Chain" is a clever method for sampling from complicated distributions
  - e.g. instead of sampling all missing values at once, conditional on observed data, sample just one missing value
  - Start with a guess for parameters describing the joint distribution and the missing data values, then randomly update to move to the next link on the chain
  - Even though you start drawing values from a distribution that looks very different from the distribution you want, if you've done things right, "eventually" the Kth link will be a draw from the target distribution

Multiple Imputation

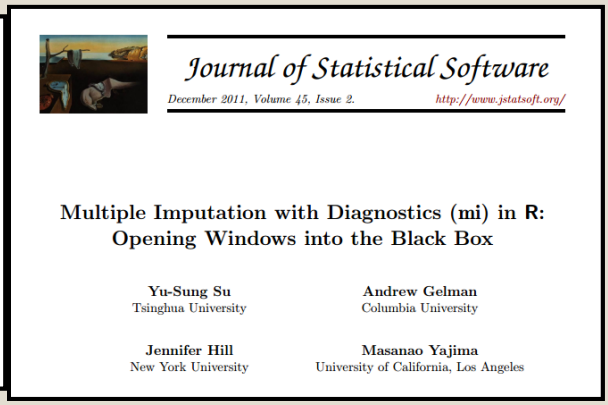


Journal of Statistical Software

December 2011, Volume 45, Issue 6. http://www.jstatsoft.org/

Multiple Imputation Using SAS Software

Yang Yuan
SAS Institute Inc.



Journal of Statistical Software

December 2011, Volume 45, Issue 2. http://www.jstatsoft.org/

Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box

Yu-Sung Su
Tsinghua University

Andrew Gelman
Columbia University

Jennifer Hill
New York University

Masanao Yajima
University of California, Los Angeles

So far so good

- Some analysis methods to deal with incomplete data
  - Weighted Regressions
  - Does not replace missing values, just tries to control for it in the analysis step
- Imputation Techniques
  - Replaces missing value with "best guess"
  - Continuous Measures
    - Mean & draw, conditional & unconditional
    - Single and multiple imputation
  - Categorical Variables
    - Multiple Imputation
    - HotDeck

50

Hot Deck Imputation

- Replaces missing value with the value from the most similar person in the dataset

- Recipient – subject with missing value
- Donor – similar subject with non-missing value
  - Donor pool – group of subjects similar to 'recipient'

*Andridge & Little, *Int Stat Rev.* 2010

51

Hot Deck Imputation

Pros

- No distribution assumptions
- Non-parametric
- Less sensitive to model specifications
- Only plausible values imputed
- Better coverage with skewed data

Cons

- More complicated
  - Many macros available

- Can be biased
  - especially with MNAR
  - Not enough donors – 1 donor over-represented

52

Hot Deck Imputation

- Replaces missing value with the value from the most similar person in the dataset

- A few options:
  - Replace with 1 donor that is most similar
  - Replace with a random donor from a donor pool of similar subjects
  - Replace with mean (or other summary measure) from donor pool of similar subjects
  - Create multiple Hot Deck imputed datasets and then summarize across datasets

53

Hot Deck Imputation

- Lots of SAS macros and R code available (google is our friend)
  - Less complicated (basically matching algorithms) to more complicated

- Differ based on
  - Methods (previous slide)
  - Definition of “similar”
  - Can it take into account multiple covariates
  - assumptions

54

Hot Deck Imputation

- Lots of SAS macros and R packages available
  - MIDAS: A SAS Macro for Multiple Imputation Using Distance-Aided Selection of Donors
  - R:
    - "hot.deck"
    - "HotDeckImputation"


Journal of Statistical Software
February 2009, Volume 29, Issue 9. http://www.jstatsoft.org/

MIDAS: A SAS Macro for Multiple Imputation Using Distance-Aided Selection of Donors

Juned Siddique, Northwestern University
Ofer Harel, University of Connecticut

Take Away

- It is easy to take care of missing data at the data *collection* stage than the data analysis stage
- How you deal with it will make a difference in the precision and accuracy of your results
- There are multiple different methods, each with pros and cons

- Analysis stage: Indicator method & Weighed regression
- Imputation: replace missing
  - "predicted value": conditional, unconditional, single, multiple
  - Someone similar: HotDeck

QUESTIONS?

EXTRA SLIDES

Complete Case - MCAR

- Assume data are MCAR so
  - $P(X_1=\text{missing} \mid D, E, X_1 \ldots X_k) = P(X_1=\text{missing}) = f$

	E=1	E=0
Case (D=1)	*f*a*	*f*c*
Control (D=0)	*f*c*	*f*d*

$$so\ OR = \frac{fa * fd}{fc * fb} = \frac{a * d}{b * c}$$

- So OR is a valid estimate (unbiased)
- However,
  - Sample size is reduced by (1-*f*) x 100% and thus
  - Efficiency is reduced

59

Complete Case - MAR

- Probability that X_1 is missing is associated with an observed variable
- In this case missingness of X_1 is associated with disease status
  - So, probability of missing values in X_1 is different for cases and controls

$P(X_1=\text{missing} \mid D=1, E, X_1 \ldots X_k) = f_D=1$ ← Probability of missingness for cases

$P(X_1=\text{missing} \mid D=0, E, X_1 \ldots X_k) = f_D=0$ ← Probability of missingness for controls

-

60

Complete Case – MAR

- Assume data are MAR, related to disease status

	E=1	E=0
Case (D=1)	$f_{D=1} * a$	$f_{D=1} * c$
Control (D=0)	$f_{D=0} * c$	$f_{D=0} * d$

$$so\ OR = \frac{f_{D=1}a * f_{D=0}d}{f_{D=0}c * f_{D=1}b} = \frac{a*d}{b*c}$$

- Again OR is a valid estimate (unbiased)
- However,
  - Sample size is reduced
  - Efficiency is reduced

61

Complete Case – MAR

- Assume data are MAR, related to exposure status

	E=1	E=0
Case (D=1)	$g_{E=1} * a$	$g_{E=0} * c$
Control (D=0)	$g_{E=1} * c$	$g_{E=0} * d$

$$so\ OR = \frac{g_{E=1}a * g_{E=0}d}{g_{E=1}c * g_{E=0}b} = \frac{a*d}{b*c}$$

- Again OR is a valid estimate (unbiased)
- However,
  - Sample size is reduced
  - Efficiency is reduced

$P(X_1\text{=missing} \mid D, E=1, X_1 \ldots X_k) = g_{E=1}$

$P(X_1\text{=missing} \mid D, E=0, X_1 \ldots X_k) = g_{E=0}$

62

Complete Case - MAR

- What if missingness is related to another covariate, X_2
  - We can *control* for X_2 in our analysis and thus also control for missingness
  - This only works if the covariate,X_2 is not at all associated to the outcome or exposure
  - For continuous outcomes
    - Even if missingness is multiplicative the complete case method yields biased estimates

$X_2 = 0$	E=1	E=0
D=1	$f_{D=01} * a_0$	$f_{D=01} * c_0$
D=0	$f_{D=00} * c_0$	$f_{D=00} * d_0$

$X_2 = 1$	E=1	E=0
D=1	$f_{D=11} * a_1$	$f_{D=11} * c_1$
D=0	$f_{D=10} * c_1$	$f_{D=10} * d_1$

$X_2 = 2$	E=1	E=0
D=1	$f_{D=21} * a_2$	$f_{D=21} * c_1$
D=0	$f_{D=20} * c_2$	$f_{D=20} * d_1$

Take away message: If you can model your missingness you can control for it in your analysis. You will lose efficiency, but your estimates should be unbiased *if modeled correctly*

This means your missingness must be explained by an observed variable

63

Complete Case ... MNAR

- Probability of missingness is related to some unknown or unobserved value
  - Meaning missing depends on outcome, exposure, covariate, effect modifiers...
  - A different pattern of missingness that depends on something we do not have information on (we cannot model)

	E=1	E=0
Case (D=1)	$f_{11}*a$	$f_{10}*c$
Control (D=0)	$f_{01}*c$	$f_{00}*d$

$$so\ OR = \frac{f_{11}a * f_{00}d}{f_{10}c * f_{01}b} = \frac{ad}{bc} * \frac{f_{11}f_{00}}{f_{10}f_{01}}$$

This time OR is clearly a biased estimate

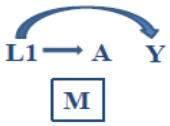
64

DAGs for missing data

In general, missing data is a form of censoring. In contrast to censoring that we discuss in other classes, the censoring we are discussing in 215 is of covariates (confounders), NOT the outcome or exposure.

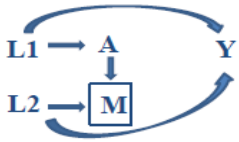
Note: these are examples of how missing data could appear in DAGs. These are NOT meant to represent "the MAR" or "the MNAR" DAGs, as there are many possible data structures. These are not part of the core course material, and are provided by the TA to help understand how missing data works. For a comprehensive discussion of this, BIO293 may be of interest (Statistical methods for incomplete data, taught by Dr. Tchetgen).

MCAR

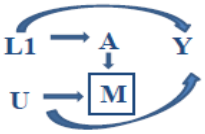


MAR

This is the smoking, silica, lung cancer IPW example from the first missing data lecture



MNAR



Y: Outcome, A: Primary Exposure, L1 : Confounder with missing data; L2: some other measured covariate; U: some unmeasured covariate; M: indicator variable for missing data on L1, where M=1 indicates missing and M=0 indicates non-missing

(M is implicitly conditioned on because you can only observe non-missing values!)

Remember: As with everything else in EPI, assumptions are key! The assumptions you make about why data are missing, or which variables are associated with missingness determine whether your data is MAR or MNAR, and what your DAG would look like.