Statistical Analysis with Missing Data

Module 4

Propagating Imputation Uncertainty

UNIVERSITY OF MICHIGAN

Accounting for Imputation Uncertainty

- Imputation “makes up” the missing data
  - treats imputed values as the truth
- For statistical inference (standard errors, P-Values, confidence intervals) need methods that account for imputation error
  - (A) redo imputations using sample reuse methods – bootstrap, jackknife
  - (B) Multiple imputation (Rubin 1987)

Bootstrapping: with complete data

- A bootstrap sample of a complete data set S with n observations is a sample of size n drawn with replacement from S
  - Operationally, assign weight w_i to unit i equal to number of times it is included in the bootstrap sample

$$w_1, \ldots, w_n \sim \mathrm{MNOM}(n; \tfrac{1}{n}, \ldots, \tfrac{1}{n})$$

Bootstrap distribution

- Let $\hat{\theta}^{(b)}$ be a consistent parameter estimate from the *b*th bootstrap data set
- Inference can be based on the bootstrap distribution generated by values of $\hat{\theta}^{(b)}$
- In particular the bootstrap estimate is

$$\hat{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$$

with variance

$$\hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$

Bootstrapping with incomplete data

- For incomplete data:
  - bootstrap the complete and incomplete cases
  - impute bootstrapped data set
  - $\hat{\theta}^{(b)}$ = consistent estimate from *b*th data set, with values imputed; then as before:

$$\hat{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} \qquad \hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$

* Bootstrap then impute, <u>not</u>
* Impute then bootstrap

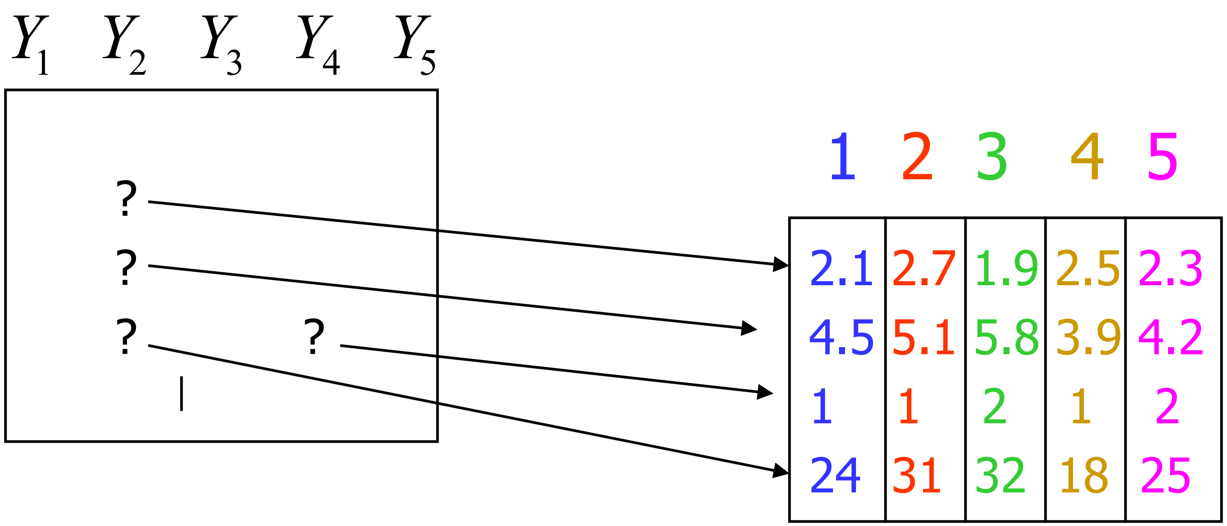
Imputing the bootstrap sample

- Impute so that the estimate $\hat{\theta}_b$ from imputed data is consistent. In particular:
  - conditional mean ok for linear statistics
  - conditional draw ok for linear or nonlinear statistics; more general, but loss of efficiency
- Computationally intensive: imputations created for each bootstrap data set

  B=200, 1000 are typical numbers

Multiple Imputation

- Create D sets of imputations, each set a draw from the predictive distribution of the missing values
  - e.g. D=5



Multiple Imputation Inference

- D completed data sets (e.g. D = 5)
- Analyze each completed data set
- Combine results in easy way to produce multiple imputation inference
- Particularly useful for public use datasets
  - data provider creates imputes for multiple users, who can analyze data with complete-data methods

MI Inference for a Scalar Estimand

$\theta =$ estimand of interest

$\hat{\theta}_d =$ estimate from d th dataset $(d = 1,...,D)$

The MI estimate of θ is $\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$

$W_d =$ estimate of variance of $\hat{\theta}_d$ from d th dataset

The MI estimate of variance is $T_D = \bar{W}_D + (1 + 1/D)B_D$

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d = \text{Within-Imputation Variance}$$

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta}_D)^2 = \text{Between-Imputation Variance}$$

Example of Multiple Imputation

- First imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	2.1			
	4.5			
	24		1	

	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53\cdot 1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)

- Second imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	2.7			
	5.1			
	31		1	

Dataset (d)	Estimate (se^2) μ_1	$\beta_{53\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)

- Third imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	1.9			
	5.8			
	32		2	

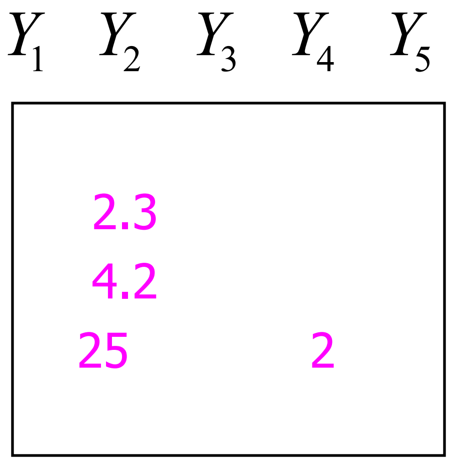
	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)

- Fourth imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
	2.5			
	3.9			
	18		1	

	Estimate (se^2)	
Dataset (d)	μ_1	$\beta_{53\cdot 1234}$
1	$12.6\ (3.6^2)$	$4.32\ (1.95^2)$
2	$12.6\ (3.6^2)$	$4.15\ (2.64^2)$
3	$12.6\ (3.6^2)$	$4.86\ (2.09^2)$
4	$12.6\ (3.6^2)$	$3.98\ (2.14^2)$

- Fifth imputed dataset



Dataset (d)	Estimate (se^2) μ_1	$\beta_{53\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)
4	12.6 (3.6^2)	3.98 (2.14^2)
5	12.6 (3.6^2)	4.50 (2.47^2)
Mean	12.6 (3.6^2)	4.36 (2.27^2)
Var	0	0.339

Summary of MI Inferences

	$\bar{\theta}_D$	$\bar{W}_D$	B_D	$\sqrt{T_D} = \sqrt{\bar{W}_D + 6/5\, B_D}$	$\hat{\gamma}_D = 1.2B_D / (1.2B_D + \bar{W}_D)$
μ_1	12.6	3.6^2	0	3.6	0
$\beta_{53\cdot 1234}$	4.36	2.27^2	0.339	2.36	0.073

$$\hat{\gamma}_D = \frac{(1+1/D)B_D}{(1+1/D)B_D + \bar{W}_D} = \text{ estimated fraction of missing information}$$

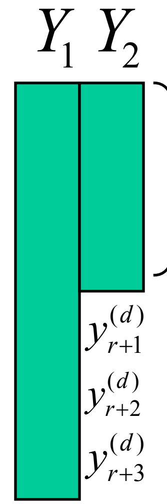
Creating Multiple Imputations

- Multiple Imputations created within a single model take into account within-model uncertainty
- Multiple Imputations can also be created under alternative models, to account for imputation model uncertainty
- Imputations can be based on implicit or explicit models, as for single imputation

Examples of draws for *d*th set of MI's

- Hot Deck: create D candidate donors that are close to incomplete case, and draw dth value from this set with replacement

- Regression: add normal draws $r_i^{(d)}$ to regression predictions



$$y_i^{(d)} = \hat{E}(y_{i2} \mid y_{i1}) + r_i^{(d)}$$

$$r_i^{(d)} \sim N(0, \hat{\sigma}^2)$$

These methods are simple but *improper* – do not account for parameter uncertainty
Later consider *proper* methods that take into account uncertainty in regression coefficients

Improper MI

(1) Estimate parameters (e.g. using complete cases)

(2) Impute missing values given estimated parameters

(3) Repeat (2) for MI data sets

(4) Use MI formula for variance

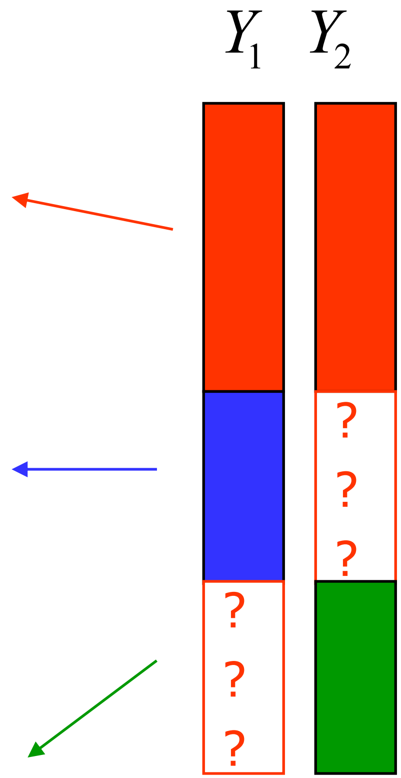
Note: only works for small amounts of missing data

Example: 2x2 Table

	$Y_2 = 1$	$Y_2 = 2$
$Y_1 = 1$	100	50
$Y_1 = 2$	75	75

Y_1	
1	30
2	60

$Y_2 = 1$	$Y_2 = 2$
28	60



Estimands:

Cell (1,1) proportion

Odds ratio

Multiple Imputation (D=5): Draw 5 sets of independent Binomial random variables

A~Bin(30,100/150)

B~Bin(60,75/150)

C~Bin(28,100/175)

D~Bin(60,50/125)

Improper!

	$Y_2 = 1$	$Y_2 = 2$
$Y_1 = 1$	100+A+C	50+(30-A)+D
$Y_1 = 2$	75 +(28-C)+B	75+(60-D)+(60-B)

Y_1 \ Y_2	1	2
1	138	83
2	107	150

Y_1 \ Y_2	1	2
1	137	84
2	112	145

Y_1 \ Y_2	1	2
1	136	81
2	114	147

Y_1 \ Y_2	1	2
1	133	86
2	115	144

Y_1 \ Y_2	1	2
1	135	82
2	128	133

$\hat{\theta}_{11}$	$\hat{\theta}_{11}(1-\hat{\theta}_{11})/478$	$\ln(OR)$	$Var(\ln(OR)) = (1/\hat{\theta}_{11}+1/\hat{\theta}_{12}+1/\hat{\theta}_{21}+1/\hat{\theta}_{22})/478$
0.29	0.00043	0.85	0.0353
0.29	0.000428	0.75	0.0350
0.28	0.000426	0.77	0.0353
0.28	0.00042	0.66	0.0348
0.28	0.000424	0.54	0.0349
0.28	0.000425	0.71	0.0351
1.62E-05		0.014	

MI estimates

Within-variance

Between-variance

$$\mathrm{var}_{MI}(\hat{\theta}_{11}) = 0.000425 + \frac{5+1}{5}1.62\times10^{-5} = 0.000445$$

$$r_m = \frac{\frac{5+1}{5}1.62\times10^{-5}}{0.000425 + \frac{5+1}{5}1.62\times10^{-5}} = 0.044$$

Fraction of Missing information

$$df = (5-1)/(0.044)^2 \approx 2066$$

$$95\%\ \mathit{confidence\ interval}:$$

$$0.28 \pm 1.96\times\sqrt{0.000445} = (0.24, 0.32)$$

$$\mathit{Complete-case}:$$

$$0.33 \pm 1.96\times\sqrt{\frac{0.33\times0.67}{300}} = (0.28, 0.38)$$

$$\text{var}_{MI}(\log(OR)) = 0.0351 + \frac{5+1}{5}0.014 = 0.0519$$

$$r_M = \frac{\frac{5+1}{5}0.014}{0.0351 + \frac{5+1}{5}0.014} = 0.325$$

$$df = (5-1)/(0.325)^2 \approx 38$$

$$95\%\ Confidence\ interval:$$

$$0.71 \pm 2.024 \times \sqrt{0.0519} = (0.25, 1.17)$$

$$Complete - case:$$

$$0.69 \pm 1.96 \times \sqrt{\frac{1}{100} + \frac{1}{75} + \frac{1}{50} + \frac{1}{75}} = (0.22, 1.16)$$

- The proper imputation approach should reflect uncertainty in the estimated proportions used in the binomial distribution.
- Using software that creates proper multiple imputation (CAT [discussed later]) on the same data set, we get

$$\hat{\theta}_{11} = 0.2812, SE = 0.02088$$

$$\log(OR) = 0.7364, SE = 0.2276$$

Creating proper MI's via bootstrap

(1) Take Bootstrap sample
(2) Estimate parameters (e.g. using complete cases) on BS sample
(3) Impute missing values given estimated parameters
(4) Repeat (1)-(3) for MI data sets
(5) Use MI formula for variance

Note: estimating parameters on BS sample propagates imputation uncertainty

Example -- Dose-Titration Study of Tacrine for Alzheimer's Disease

- Randomized, double-blind dose-escalation study (Knapp et al. 1994). Outcome - ADAS-COG

Treatment	Time (t)						
	1	2	3	4	5	[6]	[7]
Placebo	0	0	0	0	0	0	0
80mg	40	80	80	80	80	120	120
120mg	40	80	120	120	120	160	160

The Drop-Out Problem

- Titration to higher dosages to avoid side-effects on liver function
- Patients with side effects removed from double-blind study
- Other drop-outs from lack of compliance, dose-related adverse events
- Substantial differential drop-out rate at t=5:
  - Placebo 44/184 (24%)
  - 80mg 31/61 (51%)
  - 120mg 244/418 (57%)

MI model

- Missing values of ADAS-COG multiply imputed using a regression on dose, previous ADAS-COG values and baseline covariates. Two models:
  - Continuing dose model: assumes same dose after dropout as last dose before dropout
  - Zero-dose model: dose goes to zero after drop-out
- Contrast Intent-to-treat, where dose is based on original randomization

Ex. 1 contd. Tacrine Dataset

IT Analysis, Continuing Dose MI Model: 80mg vs Placebo

MI number	*Treat.diff (s.e.)*	*p-value*	*95 %C.I.*
1	-3.486 (0.951)	0.0003	(-5.35,-1.62)
2	-3.682 (0.876)	0.0000	(-5.40,-1.97)
3	-3.142 (0.944)	0.0009	(-4.99,-1.29)
4	-4.889 (0.908)	0.0000	(-6.67,-3.11)
5	-4.633 (0.910)	0.0000	(-6.42,-2.85)
6	-4.146 (0.920)	0.0000	(-5.95,-2.34)
7	-5.239 (0.925)	0.0000	(-7.05,-3.43)
8	-4.463 (0.933)	0.0000	(-6.29,-2.63)
9	-4.511 (0.953)	0.0000	(-6.38,-2.64)
10	-3.497 (0.899)	0.0001	(-5.26,-1.73)
MI Inference	**-4.169 (1.173)**	**0.0039**	**(-6.72,-1.62)**

Uncongeniality in Multiple Imputation

- Multiple imputation is designed to handle missing data once for multiple analysts
- Uncongeniality occurs when the assumptions made by the imputer and analysts are different
- Two broad categories:
  - Model assumptions are different
  - Model assumptions are the same but the estimation strategies are different

Examples

- Situation 1
- Imputer model:

$$y \mid x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

$$\Pr(\beta_0, \beta_1, \sigma) \propto \sigma^{-1}$$

- Analyst model:

$$y \sim N(\theta, \tau^2)$$

- Repeated analysis calculations under the analyst model
  - MI estimates less efficient
  - Wider confidence intervals (conservative)

- Situation 2
- Imputer model:

$$y \mid x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

$$\Pr(\beta_0, \beta_1, \sigma) \propto \sigma^{-1}$$

- Analyst model:

$$y \mid x \sim N(\alpha_0 + \alpha_1 x, \tau^2 x^2)$$

- Repeated analysis calculations under the analyst model
  - Bias

Examples (Contd.)

- Imputer and Analysts use the same model

$$y \sim N(\alpha, \tau^2)$$

- Analysts goal to estimate

$$\theta = \Pr(y \le 1)$$

- Analyst-1 estimate (uncongenial, conservative inferences)

$$\hat{\theta}_1 = \sum_{i=1}^{n} I_{\{y_i \le 1\}} / n$$

- Analyst-2 estimate (Congenial)

$$\hat{\theta}_2 = \Phi[(1-\hat{\alpha}) / \hat{\tau}]$$

General Conclusions

- Generally not an issue if the imputation models are carefully developed and capture important features in the data
- Large imputation model is preferred over a parsimonious model to accommodate multiple analysts
- Analyst should try to use the best method under his/her stated model assumption
- Using inefficient estimates may lead to conservative inferences

Summary of Multiple Imputation

- Retains advantages of single imputation
  - Consistent analyses
  - Data collectors knowledge
  - Rectangular data sets
- Corrects disadvantages of single imputation
  - Reflects uncertainty in imputed values
  - Corrects inefficiency from imputing draws
    - estimates have high efficiency for modest M, e.g. 10