PSYC 650 APPLIED DATA ANALYSIS MISSING DATA

November 5, 2018

MISSING DATA MECHANISMS

Note. function of analysis (not the dataset)

Missing completely at random (MCAR)

- probability of missing data is completely unsystematic (where random = random effect, probabilistic)
- Ex. Research assistant drops random set of surveys in mud puddle

Missing at random (MAR)

- systematic missingness, where missing data is related to other measured variables in the analysis
- Ex. Dropout related to severity, which is a covariate

Missing not at random (MNAR) aka non-ignorable missingness (NIM)

- probability of missing data is related to missing values
- Ex. Dropout related to outcomes, not included in the model

WHAT CAN A RESEARCHER DO WITH MISSING DATA?

Mechanisms of missing data are rarely known

- What can you do?
  - Worst case…
    - Ignore missingness and risk biasing results
  - Best case…
    - Use theory to make educated guesses as to why data are missing
      - Include variables that predict missingness in the analyses
    - Use robust tools for handling missing data

APPROACHES TO HANDLING MISSING DATA

Deletion methods

Single imputation methods

- Mean substitution, regression substitution
- Stochastic regression imputation
- Longitudinal data
  - Last (or baseline) observation carried forward (LOCF, BOCF)
- Missing = failure

Model based methods

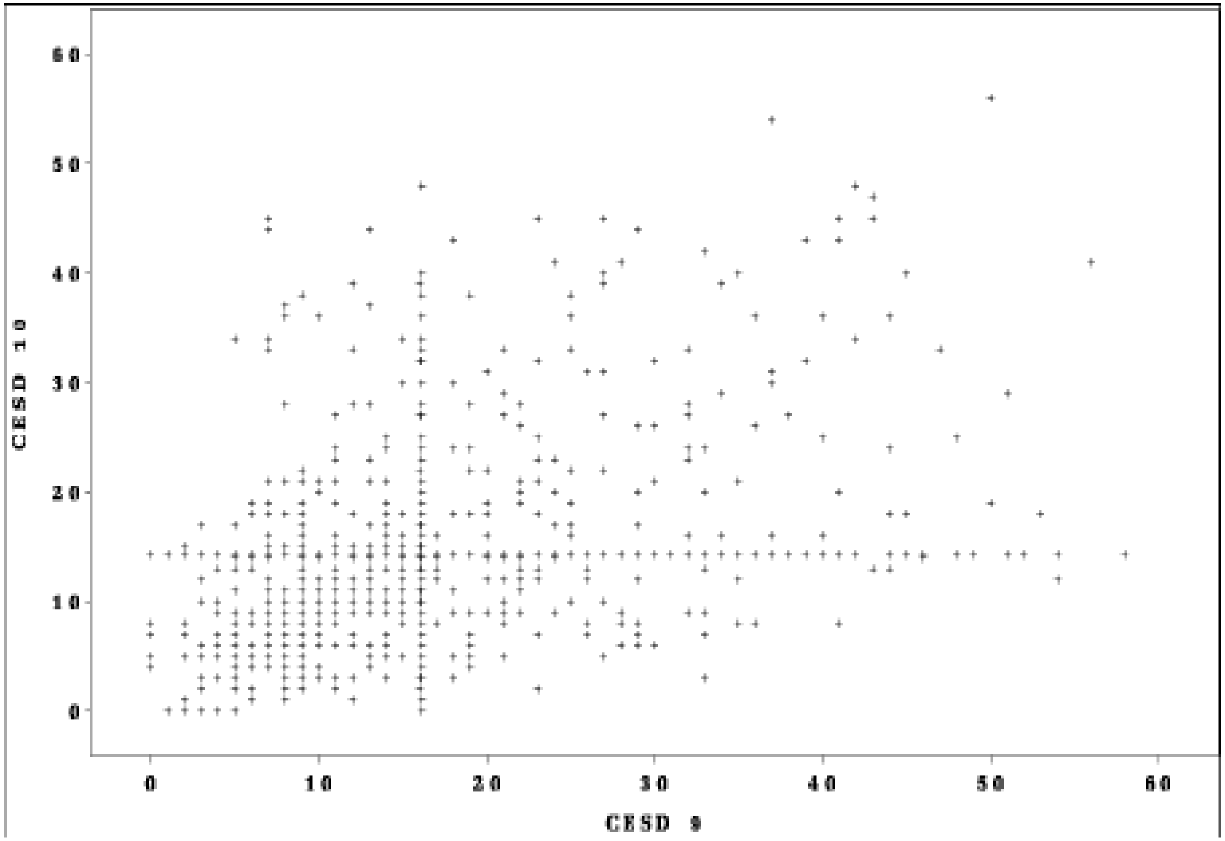
- Multiple imputation
- Maximum likelihood
- Pattern mixture models
- Selection models

DELETION: COMPLETE CASE ANALYSIS

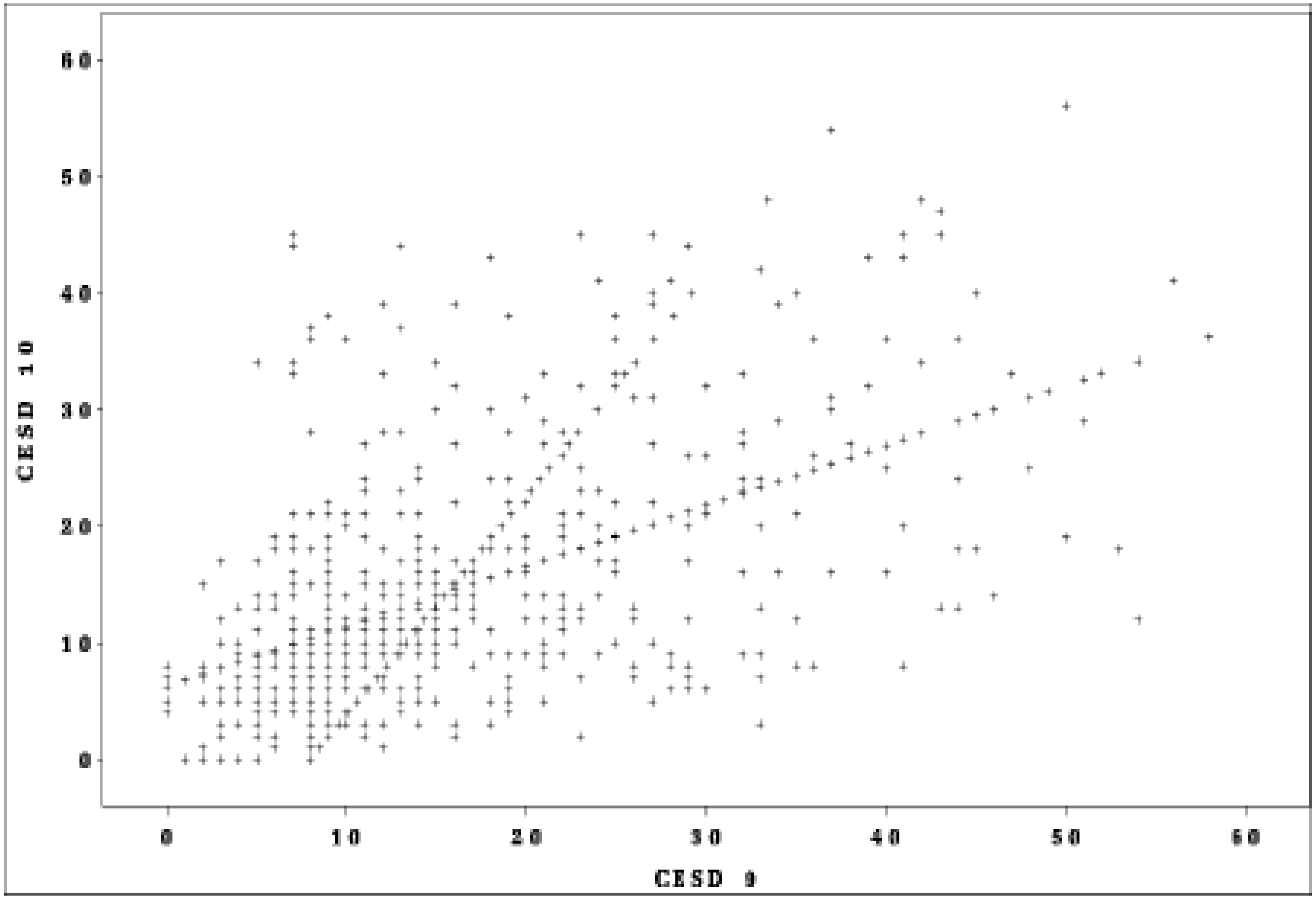
- Participants with incomplete data not included

Subj ID	Treatment Condition	Baseline PHD	Within-treatment PHD	Post-treatment PHD
1	0	88	0	56
2	0	72	0	
3	0	100	24	100
4	1	100	0	0
5	1	100		
6	1	75	0	75

Type 1: Resulting data from Mean Substitution



Type 2: Data using Regression Imputation



SINGLE IMPUTATION: LAST OBSERVATION CARRIED FORWARD

- Missing values replaced with the last observed value of the same variable
- Must be done manually or through programming

Subj ID	Treatment Condition	Baseline PHD	Within-treatment PHD	Post-treatment PHD
1	0	88	0	56
2	0	72	0	0
3	0	100	24	100
4	1	100	0	0
5	1	100		100
6	1	75	0	75

SINGLE IMPUTATION: MISSING = FAILURE

- Missing values replaced with 100% heavy drinking days
- “Conservative” to assume dropout indicates relapse

Subj ID	Treatment Condition	Baseline PHD	Within-treatment PHD	Post-treatment PHD
1	0	88	0	56
2	0	72	0	100
3	0	100	24	100
4	1	100	0	0
5	1	100	100	100
6	1	75	0	75

PROBLEMS WITH SINGLE IMPUTATION METHODS

Mean substitution

- Unbiased means when data are MCAR
- Attenuated correlation, biased means when MAR, reduced variability

Regression imputation

- Unbiased means when data are MCAR or MAR
- Overestimate correlated, reduced variability

Stochastic regression imputation

- Unbiased when MCAR or MAR
- Underestimated standard errors - inflate Type I error

MULTIPLE IMPUTATION

Imputation step

- Multiple copies of a dataset with unique estimates of the missing values drawn at random
- Data augmentation with iterative algorithm
  - Imputation step – identical to stochastic regression
  - Posterior step – add random error to estimates

Analysis step

- Yields several estimates of each parameter and standard error (e.g., 20 datasets = 20 estimates for each parameter and each standard error)

Pooling step

- Pool parameter estimates adjusting for within imputation variance and between imputation variance

MULTIPLE IMPUTATION – PART 1

Generate plausible values for missing values

Residual error added to imputed values

			Imp. 1	Imp. 2	Imp. 3		Imp. *m*
Treatment	Baseline PHD	Within-treatment PHD	Post-treatment PHD	Post-treatment PHD	Post-treatment PHD	. . .	Post-treatment PHD
0	88	0	56	56	56		56
0	72	0	25	17	0		40
0	100	24	100	100	100		100
1	100	0	0	0	0		0
1	100		68	90	75		100
1	75	0	75	75	75		75

MULTIPLE IMPUTATION – PART 2

Generate plausible values for missing values, "*m*" times and then pool results to estimate effects

Imputation number	**Treatment effect (β)**	**Standard Error (SE)**
1	-4.31	3.11
2	-6.85	3.87
3	-2.88	2.99
…	…	…
m	-3.92	3.50
Pooled estimate	-4.49	3.57

MAXIMUM LIKELIHOOD

Uses all available data to identify parameter values that have the highest probability of producing the sample data

- Same concept as OLS regression
- Identify parameter estimates that maximize the sum of the log-likelihood values; repeated until estimates that minimize the distance to the observed data

FULL INFORMATION MAXIMUM LIKELIHOOD

Use all available data:

ID	Treatment	Baseline PHD	Within-TX PHD	Post-TX PHD
1	0	88	0	56
2	0	72	0	
3	0	100	24	100
4	1	100	0	0
5	1	100		
6	1	75	0	75

To create variance-covariance matrix:

	Treatment	Baseline PHD	Within TX PHD	Post Tx PHD
Treatment	0.250	0.003	0.014	0.001
Baseline PHD	0.003	0.082	0.014	0.015
Within Tx PHD	0.014	0.014	0.082	0.067
Post Tx PHD	0.001	0.015	0.067	0.099

Identify parameter values that have highest probability (i.e., maximize the likelihood) of producing the sample data

AUXILIARY VARIABLES

Improve estimation (reduces SE) without directly influencing parameter estimates

Best to use auxiliary variables that are highly correlated with incomplete analysis model variables

MISSING DATA MODELS IN MPLUS

FIML is default

Listwise deletion by adding to DATA: command

```
DATA:
FILE is filename.csv;
LISTWISE IS ON;
```

MI is two step process

- Create imputation datasets
- Pool estimates across imputed datasets

MULTIPLE IMPUTATION: IMPUTATION STEP

```
VARIABLE:
USEVARIABLES ARE  opioids asitot0-asitot12;

DATA IMPUTATION:
IMPUTE opioids (c) asitot0-asitot12;

NDATASETS 50;

SAVE = asimi*.dat;

ANALYSIS:
  TYPE = BASIC;
OUTPUT:
  TECH8;
```

AFTER IMPUTATION FILE

asimi1.dat
asimi2.dat
asimi3.dat
asimi4.dat
asimi5.dat
asimi6.dat
asimi7.dat
asimi8.dat
asimi9.dat
asimi10.dat
asimi11.dat
asimi12.dat
asimi13.dat
asimi14.dat
asimi15.dat
asimi16.dat
asimi17.dat
asimi18.dat
asimi19.dat
asimi20.dat
…
asimilist.dat

MULTIPLE IMPUTATION: ANALYSIS STEP

```
DATA:

FILE is asimilist.dat;

TYPE = IMPUTATION;


VARIABLE:

NAMES ARE opioids asitot0-asitot12;

MISSING ARE * ;

USEVARIABLES ARE opioids asitot12;


MODEL:

asitot12 on opioids;
```

MULTIPLE IMPUTATION RESULTS

```
MODEL RESULTS

                        Estimate       S.E.  Est./S.E.    P-Value

 ASITOT12 ON

   OPIOIDS               8.209       4.099      2.003      0.045

 Intercepts

   ASITOT12             11.006       2.027      5.430      0.000

 Residual Variances

   ASITOT12            682.786      74.273      9.193      0.000
```

MAXIMUM LIKELIHOOD

```
VARIABLE:
MISSING ARE ALL (999)
USEVARIABLES ARE opioids asitot12;
AUXILIARY ARE (m) asitot0-asitot11;

MODEL:
asitot12 on opioids;

OUTPUT:
SAMPSTAT CINTERVAL STANDARDIZED;
```

EXAMPLE OF MAXIMUM LIKELIHOOD WITH AUXILIARY VARIABLES

```
MODEL RESULTS
                      Estimate      S.E.  Est./S.E.    P-Value
 ASITOT12 ON
    OPIOIDS              8.975     4.286      2.094      0.036
 Intercepts
   ASITOT12             11.295     2.146      5.262      0.000
 Residual Variances
   ASITOT12            690.321    80.537      8.571      0.000
```

EXAMPLE OF MAXIMUM LIKELIHOOD WITHOUT AUXILIARY VARIABLES

```
MODEL RESULTS

                      Estimate      S.E.  Est./S.E.    P-Value

 ASITOT12 ON
  OPIOIDS               9.138     4.332     2.109     0.035

 Intercepts
   ASITOT12            10.903     2.166     5.033     0.000

 Residual Variances
   ASITOT12           689.524    80.429     8.573     0.000
```

COMPARISON OF ESTIMATES

	n	Estimate	SE	Est/SE	p-value
Listwise deletion	146	9.139	4.362	2.095	0.038
Mean imputation	197	6.793	3.247	2.092	0.038
BOCF	193	8.605	6.811	1.263	0.208
Multiple imputation	200	8.209	4.099	2.003	0.045
Maximum likelihood	197	8.975	4.286	2.094	0.036

WHAT ABOUT MNAR DATA?

Pattern mixture models

- assume that the substantive data are conditional on the missing data mechanism
- Include missing data patterns as main effects with random effect means mixed across the different missing data patterns to yield single estimates of parameters

Selection models

- assume that the missing data mechanism is conditional on the substantive data
- Incorporate indicators of the probability of missing data, regressed on outcomes

PATTERN MIXTURE MODELS

- assume that the substantive data are conditional on the missing data mechanism

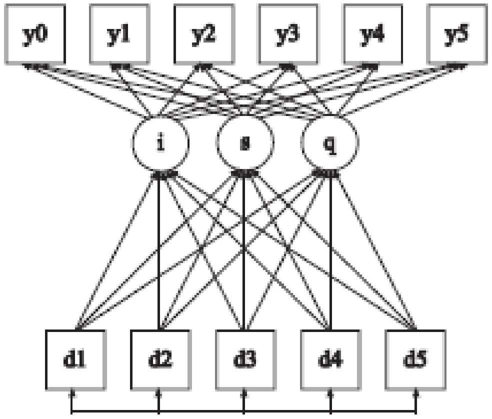


Figure 2. Pattern-mixture modeling (*d*s are dropout dummy variables).

SELECTION MODELS

- assume that the missing data mechanism is conditional on the substantive data

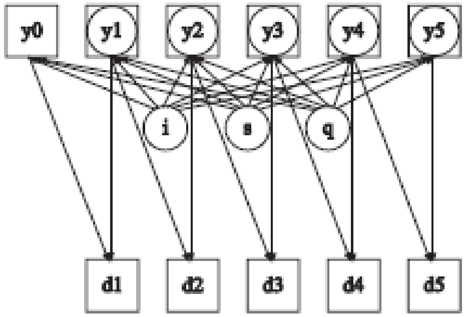


Figure 4. Diggle–Kenward selection modeling (*d*s are survival indicators).

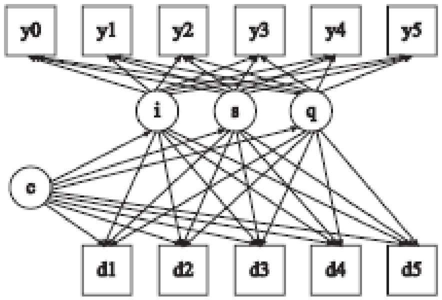


Figure 8. Beunckens mixture model (mixture Wu–Carroll model).

RESULTS SUGGEST THAT LGC RESULTS ARE ROBUST TO MISSING DATA MAR ASSUMPTION

Model	BIC	Intercept	Linear Slope	Quad Slope
LGC	39602.4	27.88 (0.62)	-5.60 (0.29)	0.33 (0.04)
Dropout PMM	39634.6	27.68 (0.80)	-5.91 (0.37)	0.36 (0.05)
Diggle Kenward	45452.2	28.23 (0.64)	-5.90 (0.32)	0.36 (0.04)
Wu Carroll	45639.3	28.23 (0.65)	-5.67 (0.34)	0.33 (0.05)