Missing Data

Paul E. Johnson[1]

[1]Center for Research Methods and Data Analysis

2019

KU

Outline

1. Listwise Deletion is Evil
2. Tools
   - Multiple Imputation
   - 2 Methods of MI
3. SEM is a focal point
4. Conclusions

KU

Outline

Listwise Deletion

- Common "listwise deletion": cases with any "missing data" are dropped from the analysis.
- Will omit a case even if it is missing only on one variable, but has scores for 10 or 20 others.
- implications
  - Best Case: used sample size shrinks $\rightarrow$ larger standard errors, fewer "statistically significant" estimates
  - Worst Case: parameter estimates are biased, hypo-tests wrong
- Terms "MCAR", "MAR" , and "MNAR" are attributed to Rubin (1976, 1987)

Missing Completely at Random (MCAR)

- MCAR: the best case scenario
- Running example with 2 variables
  - x, Education (years)
  - y, Income (thousands of dollars)
- MCAR: static unpredicatbly destroys survey responses

Missing Completely at Random

- R is a "0" or "1" *missing indicator*"
- x or y represent two observed variables
- MCAR: the chance that an individual piece of data is missing does not depend on the value of either x or y

$$Pr(R = 1|x, y) = Pr(R = 1)$$

- If MCAR, listwise deletion not disastrous
  - still "unbiased/consistent" parameter estimates
- However, smaller N $\rightarrow$ higher standard errors, fewer "significant" tests

Worst Case Scenario: Missing Not at Random (MNAR)

- People who have a lot of money are more likely to tell us their income, no matter what the level of education is
- low earners are systematically NA. High-earning respondents will exaggerate the linkage between education and income.
- MNAR: chance that information will go missing depends on unobserved variables
- MNAR methods still not widely available.

Middle Scenario: Missing at Random (MAR)

- Common mistake: "missing at random" means "missing completely at random"
- Correct: "Missing at random": missings are predictably missing using observed variables.
  (In other words, the exact opposite of what most of us expect MAR to mean).

MAR

- The people who are poorly educated misunderstand questions, say "don't know".

Education	Prob. Missing
8	.3
9	.25
10	.2
11	.2
12	.15
13	.15

- The *probability that data goes missing is the same for each and every person within an education* level.
- MAR: Given the details on the respondent, the chance that information is missing is the same for all respondents in that group.

- MNAR danger: the chance of missing information depends on the questions that they are not asked, or refuse to answer

Outline

MI or FIML

- At the current time, the only credible methods are the following

1. Multiple Imputation
2. Full Information Maximum Likelihood

MI materials available in lit folder

- My lecture notes on MI: multipleImputation-1-lecture.pdf
- Notes from the CRMDA Saturday Seminar by Terry Jorgensen and Kyle Lang

MI Thumbnail sketch

1. Create m "completed data" sets: Make several educated guesses (trust me) about values of missings
2. Estimate same model with each completed data set
3. Combine estimates using Rubin's rules. With m sets of estimates
   - Parameter estimates: average the imputed, $\hat{\beta} = \sum_{j=1}^{m} \hat{\beta}_j$
   - Variance estimate for $\hat{\beta}$ is a sum of
     1. average of $\widehat{Var(\hat{\beta})}$, $\sum_{j=1}^{m} = \widehat{Var(\hat{\beta}_j)}$, plus
     2. a penalty for uncertainty between $\hat{\beta}_j$, $\frac{1}{1+m}\sum(\hat{\beta}_j - \hat{\beta})^2$.

MI Unsolved Problems

- Slow:
  1. Must create m sets
     1. Which variables to use?
     2. Diagnose quality of imputations
  2. Model must be estimated for each of the m data sets. TIME CONSUMING!
- Hypo testing nightmare.
  - Rubin's rules work for the "normally distributed" parameters–βs. Not R^2,F, χ^2, variances, etc
  - Undefined procedures for sequential "hypothesis testing". F test or likelihood ratio tests between two models have to be done for each of the m data sets. What if they disagree?
  - SEM users accustomed to "model chi-square test", RMSEA, CLI, etc. All are undefined in MI, or nearly so.

Imputation Details

- the part where I said "*trust me*"
  1. Multivariate Normal approximation model.
  2. MICE: Multiple Imputation with Chained Equations

NORM was first

- Rubin (1976) and Little and Rubin (1987) brought missingness to forefront in statistics
- Schaffer's implementation, which was known as NORM, was the first widely available software based on the idea that all of the variables are multivariate normal.
- Political scientists learned from King et al. (2001). By a considerable margin, King was "faster to the market" than SAS and Stata with his program Amelia, an MVN-based imputer. Amelia II is the version for R (2011)
- Until 2009 (at least), these MVN-based imputers were the only workable software

What's Norm?

Here's my best effort

- Assume all variables are drawn from one Multivariate Normal Distribution, $MVN(\mu, \Sigma)$
- Use algorithms to estimate μ and Σ
- After estimating μ and Σ, then draw random samples from the $MVN(\hat{\mu}, \hat{\Sigma})$ to fill in missing values
- Then re-estimate μ and Σ, then re-draw samples, repeat until this "converges"
- Practical Dilemma: which variables should be included?
  - **auxiliary variable** not included in your regression model, but possibly relevant to missingness.

Reluctance about NORM

- Most people say "but my variables are not Normal." (gender, survey scales, etc)
  - imputation of 0.75 (or 1.5, or -0.5) not meaningful for Sex coded 0 or 1
- Not tolerant of highly inter-correlated variables.
  - Norm-based imputers fail, thinking those variables are redundant

MICE: Alternative MI implementation

MICE: Multiple Imputation via Chained Equations (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012).

- R package "mice"
- Separately process each column, predicting it from all the others. "The algorithm imputes an incomplete column (the target column) by generating 'plausible' synthetic values given other columns in the data."(van Buuren, 2012)

Competing R software packages

- mi (Su, Gelman, Yajima, 2011)
- rms (Harrell, 2017).

MICE: Alternative MI implementation ...

- customized regressions, one for each column
  - linear regression for numeric variables
  - logistic for dichotomies, ordinal
  - poisson for count data
  - multinomial for multicategory.
- Cycle through columns over and over, until model converges (in MCMC sense)

MICE: Alternative MI implementation ...

PMM "predictive mean matching" to select imputed values

- Find cases with similar predicted values to the case in question
- Draw imputations randomly from that subset of actual scores
- Solves the problem that imputations might have impossible values
  - Imputations for categorical variables always match the original scale (sex is always 0 or 1, never 0.64)
  - When a variable is badly skewed, the PMM always selects a realistic value.

Technical Issues

- MICE can be very slow

Maximum Likelihood Analysis

- In a nutshell, ML without missing data says
  1. you estimate the parameters (maybe they are β and σ_e^2, lets don't get bogged down in details).
  2. Calculate the chance of finding that data!
  3. Adjust estimates of parameters to maximize the chance of finding that data.

Casewise Maximum Likelihood Analysis

This is the way many statisticians think of ML.

1. The chance of observing the outcome y_i for case i is p_i.
   1. p_i depends on the model coefficients (say, β, σ, or θ) as well as the probability model you assumed.
   2. e.g. assume $y_i = X_i\beta + \epsilon$. Given X_i and β and σ_ε estimates, calculate probability of observed y_i.
2. The likelihood (L) of N cases is the product of individual likelihoods

$$L = p_1 \cdot p_2 \cdot p_3 \ldots p_N$$

For convenience, convert that to a sum by taking the natural logarithm ($\ln = \log_e$)

$$lnL = ln(p_1) + ln(p_2) + ln(p_3) \ldots ln(p_N)$$

$$lnL = \sum_{i=1}^{N} ln(p_i)$$

Casewise Maximum Likelihood Analysis ...

3. Maximize lnL by adjusting parameters (say, β, σ, or θ). Those are the maximum likelihood estimates.
4. Diagnostic statistics (variance estimates of parameters, etc) follow.

Visualize a 2 parameter optimization problem

Convergence

- ML adapts estimates, continuing until a convergence point is obtained.
- Like climbing a mountain

Properties of ML Parameter Estimates

- small sample properties either unknown or sometimes poor (biased), but
- consistent (as $N \to \infty$, $\hat{\beta} \to \beta$), and
- asymptotically Normally distributed (central limit theorem).

Software versus Concept

- The software we usually have is based on the idea that p_i is calculated with the scores of all cases on all of the variables
- If a case is missing, the software "doesn't work" and listwise deletion happens
- *Stop right there*. We are letting a software limitation govern us.

Sketch of the FIML Concept

id	x1	x2	x3	x4	y
1	2	3	4	4	18
2	4	1	4	2	14
3	3	.	4	2	13
4	3	.	2	4	12
5	4	2	.	4	.
6	3	1	.	4	.
7	.	3	5	4	.

Notes

1. Group the rows into 'missing data patterns"
   1. {1,2}{3,4}{5,6}{7}
2. Develop a probability model for each missing data pattern (grouping)
3. Then maximize

$$lnL = \sum_{i=1}^{N} ln(p_i)$$

The singleton {7} may be unmanageable, but we've got good ideas what to do on the rest.

Outline

KU

SEM magnifies damage of listwise deletion

- Suppose we have a data set with 200 cases.
- 1 outcome and 4 predictors.
  - if the probability of non-missing data for each variable is 0.95,
  - the chance that data is available for all 5 variables on each case is $0.95^5 = 0.77$
  - In this "good case scenario", we end up using three-fourths of the rows of data .

SEM magnified damage of listwise deletion

- Suppose an SEM model uses 20 data indicators (5 latent variables, with 4 indicators on each one).
- With probability of non-missing at .95, the chance we observe all 20 variables is $0.95^{20} = 0.35$.
- Listwise deletion leaves us with one-third of our data, even in this comparatively optimistic data situation.
- Bad news scenario $0.90^{20} = 0.12$.

KU

Covariance structure analysis well suited to FIML

Remember this?

id	x1	x2	x3	x4	y
1	2	3	4	4	18
2	4	1	4	2	14
3	3	.	4	2	13
4	3	.	2	4	12
5	4	2	.	4	.
6	3	1	.	4	.
7	.	3	5	4	.

Analyze groups given missing data pattern
Except for $\{7\}$, a covariance matrix can be estimated for each sub-group
And a covar matrix is then used to calculate p_i for each person in a pattern group.

- It is still a covariance structure analysis, but it happens across the subgroups.
- Subgroups assumed to have same values for same parameters

Outline

Missing data in Social Science Culture

- For several decades, most social scientists have been happy enough to follow software defaults that omit cases on which there are some missing scores.
- It has been known, since the 1970s, that listwise deletion renders inferior parameter estimates
- Any type of MI, or FIML if it is available, is more desirable than listwise deletion

Situation is Changing 1: Imputation

- Missing data imputation does not seem so exotic as it once did.
- MI estimation is now integrated into many models in Stata (Cynic in me says "great, now people who run models they don't understand can now benefit from an imputation scheme that they also don't understand")

Situation is Changing 2: Covariance Structures Analysis

- SPSS's AMOS program for SEM was at the cutting edge of the FIML estimation wave (Arbuckle, 1996).
- Implementations of FIML for SEM now exist for numeric indicator variables in almost all SEM programs
- FIML for categorical indicators is still "iffy"
  - unavailable until very recently, anywhere
  - included in Mplus 7.2, but very slow and only cooperates with data structures and models of small/moderate size.

References

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *The American Political Science Review*, 95(1), 49–69.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics. New York: Wiley.

References ...

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3): 1-67.

Stef van Buuren (2012). Flexible Imputation of Missing Data. Boca Raton, FL: Chapman & Hall/CRC Press.

Yu-Sung Su, Andrew Gelman, Jennifer Hill, Masanao Yajima. 2011. "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box". Journal of Statistical Software. 45(2)

Session

```
sessionInfo()
```

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 19.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/atlas/libblas.so.3.10.3
LAPACK: /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
     LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8
     LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8
    LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] compiler_3.6.0 tools_3.6.0
```