Missing Data in Longitudinal Studies: To Impute or not to Impute?

Robert Platt, PhD
McGill University

McGill

Outline

- Missing data – definitions
- Longitudinal data – specific issues
- Methods
  - Simple methods
  - Multiple imputation
- Some broad recommendations

Missing Data

- Definition: If a measurement was intended to be taken and was not, it is missing
  - Observational study – what does “intended” mean?
- Compare to: Intentional lack of data (eg some subjects measured every hour, others every two hours)
- Intentional/structural unbalance can be handled straightforwardly
- Missing - must understand why…

McGill

Classifying Missing Data

- Completely Random (MCAR) if missingness independent of both observed and missing data
- Random (MAR) if missingness independent of missing data
- Informative if missingness depends on missing values
- Crucial distinction is between MAR and informative, because we observe the data to predict missingness.

McGill

MCAR– Missingness doesn't matter

- Complete cases are a random sample of the full dataset
  - A reduced dataset using only complete cases "looks like" the full dataset
  - Dropping cases with missing data gives unbiased estimates
  - Only issue is loss of power.

MAR– Can Model Missingness

- Missingness depends only on observed variables
  - – Overall estimates biased in complete cases
  - – BUT – within strata, estimates are unbiased
- Analogous to stratified sampling
- Can fix these problems in analysis

NMAR - Big Problem

- Missingness depends on the missing data
- No statistical approach can give unbiased estimates
- Best bet – try sensitivity analyses to determine extent of the missingness and what you can do about it

Key Result

- Crucial distinction is between MAR and informative, because the information about missingness and observed data can be separated.
- If data are MAR or MCAR, likelihood-based methods (eg mixed models) will work
- Methods like GEE for clustered data are not likelihood based.
  - Need extra care with missingness; weighted estimating equations

McGill

Problem

- MCAR/MAR/Informative?
  - How can you tell?
  - YOU CAN'T!
  - There is no test, and no guarantee whether it's one or the other…

Longitudinal data

Intermittent vs Dropout

- Dropout - from time T onward all obs missing.
- Intermittent - subjects miss individual values but return
- If intermittent mechanism is known, not really missing (see first slide)
- If unknown, must consider mechanism

Dropouts/Loss to follow up

- Problem - dropouts usually not ignorable
- Eg - dropout related to treatment side-effect?
- If dropouts are sicker, then the remaining subjects appear healthier than the population.
- Are the reasons for dropout measured?

Solutions?

Solutions

- Last observation carried forward
  - Fill in with the last completed value
  - Typical in pharma industry
  - Conservative if positive time trend in the outcome
- Complete cases only
  - OK if MCAR – rare
  - Biased if MAR or informative

McGill

Better solutions?

- Missing indicator/category
  - EG education:
    - <12 yrs
    - 12-16 yrs
    - 16+ yrs
    - Missing
- Problem – what does the "missing" category mean?
  - It's an average of all the other categories. Meaningful?

McGill

Better Solutions

- Single Imputation
  - Estimate a predicted value for the missing value
  - Use this in the analysis.
  - Unbiased if MCAR or MAR
  - Problem – uncertainty in that single prediction is not accounted for
  - Standard errors are too small

Best Solutions

- Multiple imputation
  - Impute several times
  - Use multiple values to estimate variability
  - Unbiased if MCAR/MAR
  - Variance estimates are valid.
- Inverse probability weighting
  - Inflate subjects by the inverse probability of being non-missing:
    - EG if 5 total subjects, 4 observed, reweight the 4 observed subjects by 5/4 (inverse probability of observed)
  - Unbiased if MCAR/MAR, variance estimates (via, e.g., bootstrap) valid

McGill

Multiple Imputation- Step 1

- A model for the missing data
  - Multivariate normal model
    - assume that the variables follow an MVN.
    - Estimate using Markov Chain Monte Carlo
    - Works well, even with binary or categorical variables

Multiple Imputation- Step 1

- A model for the missing data
  - Conditional model (e.g., multiple imputation using chain
    - Propose a model for the distribution of each variable conditional on the others
    - Estimate missing values for first variable
    - Use those predictions to estimate second variable
    - Repeat 10-20 times

Multiple Imputation

- Pros and cons
  - MVNorm:
    - Easy, theoretically grounded.
    - Non-continuous variables?
  - MICE:
    - Good results in practice
    - Very flexible
    - No formal theoretical grounds
    - Perfect predictions

How to build MI model

- Large model is good
  – but not too large…
- Always include the outcome in the MI process!
  – Omitting outcome causes parameter estimates to be biased towards zero
- Which approach?
  – MICE typically recommended, but not universally.

Generating Imputations

- Repeat imputation process several (m) times
  - (note – this is not the same as repeating the MICE steps 10-20 times)
- Each imputation generates a parameter estimate $\hat{\theta}_j$ and variance estimate $\hat{V}_j$ for j=1,…,m

Multiple Imputation – Step 2

- Compute within-imputation variance:

$$V = \frac{1}{m}\sum_{i=1}^{m} V_i$$

- And between-imputation variance:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\theta_i - \bar{\theta}\right)^2$$

- Final estimator is $\bar{\theta}$, variance $V + (1+1/m)B$

McGill

Inverse Weighting

- From survey sampling
- Idea –give higher weight to complete cases (proportional to the inverse of probability of being observed)
- Up-weight observed cases in rare strata
- Problem – very high weights for very rare cases?

Inverse Probability Weighting

- Very simple example:
  - 100 subjects
  - 40 report smoking status: 10 smokers, 30 nonsmokers. Estimated P(smoking) = 0.25
  - In 60, smoking status is missing.
  - Assuming MAR, what's the best estimate of the number of smokers out of 100 total?
    - 25 (100*0.25)

IPW

- P(observed) = 40/100 = 0.4
- Each observed subject "counts" for 100/40 = 1/(0.4) =2.5 subjects in total
- Best guess for number of smokers:
  - 10*(1/0.4) = 25
- Weight by the inverse of P(observed)
- Bootstrap to get variance estimates (or analytic approaches)

McGill

MI vs IPW

- MI is better if we can model the missing values
  - e.g., if SBP is the missing variable and using other characteristics we can predict SBP
- IPW may be better if we can model the missingness process
  - e.g., if we know that smokers are much less likely to respond to questions about drinking (but unable to estimate alcohol consumption

McGill

Is Imputation Necessary?

- Harrell (2001):
  - If <5% of cases have missing data, then complete case analysis usually fine
  - If >50% of cases have missing data, you should rethink the study!
  - In between, imputation is usually best option

What can you do?

- Recognize missingness as a problem
- Don't default to complete cases!
- Remember – MCAR vs MAR vs NMAR is untestable!
  - Could conduct sensitivity analyses
- If missingness is a significant problem, consult a statistician…

McGill

References

1. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health 2004;25:99-117.
2. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psycholog Methods 2002;7:147-77
3. Rubin DB. Inference and missing data. Biometrika 1976;63:581-92.
4. Greenland S. Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255-64
5. White et al. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med (2010) pp. XX
6. Moons et al. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology (2006) vol. 59 (10) pp. 1092-1101
7. Lee KJ, Carlin JB. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. Am J Epidemiol. 2010 Mar. 1;171(5):624–632.
8. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol. 2010;10:7.
9. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010 Sep. 13;:n/a–n/a.
10. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010 Dec. 31;10(1):112.

McGill