Analysis of Longitudinal Data Subject to Drop-out

Xi Chen

xchen595@163.com

One issue in the analysis of longitudinal data that requires attention is the nature of any missing data, which can bias results.

Outline

- Example of missing data
- The "Taxonomy" for missing data
- Impact of missing data
- Analysis approaches
- Caveats

Example of Missing Data

Schizophrenia Treatment Trial

1. A randomized longidudinal study of haloperidol and risperidone
2. Primary outcome: PANSS score at 8 weeks
3. Intermediate outcomes at 1, 2, 3, and 6 weeks
4. 8 week completion by arm:

Placebo 27/88 = 31

Haloperidol 36/87=41

Risperidone(6mg) 52/86=60

Reasons for dropout

name	times
Abnormal lab result	4
Adverse experience	26
Inadequate response	183
Inter-current illness	3
Lost to follow-up	3
Uncooperative	25
Withdrew consent	19
Other	7

Taxonomy of Missing Data

- Missing Completely at Random (MCAR) This assumes that the probability of missing an observation does not depend on any variables. No selection bias
- Missing at Random (MAR) This assumes that missing an observation is predicted by variables that you have measured, but not further dependent on variables you have not measured. Possibly fixable selection bias.
- Nonignorable (NI) This assumes missing an observation is predicted by variables that you have not measured such as the outcome of interest. Not fixable selection bias!!!

Missing Data Issues&Implications

we have $R_{ij} = 1$ if subject i is observed at time j, $R_{ij} = 0$ if subject i is not observed at time j

- MCAR: if $p(R_i|y_i, X_i) = p(R_i|X + i)$, this implies that $E(Y_{ij}|R_{ij} = 1, X_i) = E(Y_{ij}|X_i)$;

Standard analysis using the available cases is valid

- MAR: if $P(R_i|Y_i^O, Y_i^M, X_i) = p(R_i|Y_i^O, X_i)$;

here the probability of missing data only depends on the observed values and not hte missing values. Trouble starts here since this implies $E(Y_{ij}|R_{ij} = 1, X_i) \neq E(Y_{ij}|X_i)$, Standard analysis of the available cases is potentially biased. However, there are methods that can provide valid analysis, but these require additional (correct) statistical modelling.

- NI: if $P(R_i|Y_i^O, Y_i^M, X_i)$ depends on Y_i^M;

Standard analysis of the available cases is likely biased. The bias can not be corrected since missingness depends on unobserved data and therefore can not be empirically modelled. The recommended approach is to conduct sensitivity analyses: "It is bad, but how bad could it be?"

Analysis Approaches

- MAR: Maximum likelihood (ML)
- A model is needed that links the missing outcomes to the factors that predict the missing outcomes.
- The factors that predict the missing outcomes need not be part of the "regression" model of interest (such as intermediate outcomes in a longitudinal study)
- For longitudinal data linear mixed models can help (SAS PROC MIXED)
- This general approach can also be taken for missing covariates

Multiple Imputation(MI)

- A model is needed that links the missing variables to the factors that predict the missing outcomes
- The factors that predict the missing variables need not be part of the "regression" model of interest (such as intermediate outcomes in a longitudinal study)
- Fill in the missing data
- Fill in and analyze multiple times
- Average the multiple summaries and combine within and between sample variances.

Multiple Imputation: Combine Results

- For each of M imputed data sets yhou get an estimate of the regression coefficient and the variance of the estimate.
- Average the M estimates.
- Average the M variance estimates and combine with the observed variance of the regression estimates across the imputed data sets to get a final variance (sd) estimate