Improving Imputation With Auxiliary Variables

Missing At Random (MAR) Revisited

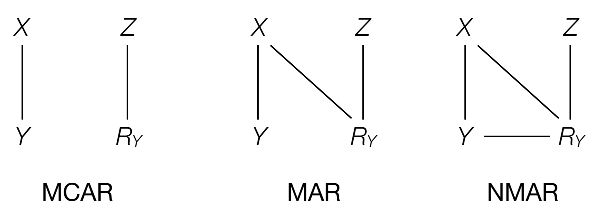
The probability of missing data on a variable *Y* is related to observed responses of other variables but is unrelated to the would-be values of *Y* itself

$$P\left(R|Y_{obs}, Y_{mis}\right) = P\left(R|Y_{obs}\right)$$

The probability of nonresponse varies across different observed score profiles

Diagram of Mechanisms

X represents a set of observed variables correlated with *Y*, *Z* represents a set of observed variables uncorrelated with *X* and *Y*, and R_Y is the missing data indicator for *Y*



MAR = observed and missing scores are the same, on average, after conditioning on (controlling for) other variables

MAR is satisfied only when we condition on all correlates of missingness.

Inclusive Analysis Strategy and Auxiliary Variables

The literature recommends an inclusive strategy that incorporates auxiliary variables into missing data handling

An auxiliary variable is not of substantive interest but is used to improve power or reduce nonresponse bias

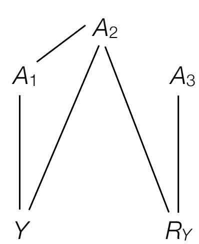
The benefit of an auxiliary variable depends on the pattern and magnitude of its correlations with the analysis variables and missing data indicators

Diagram of Auxiliary Variable Correlations

Conditioning on A_1 improves power but ignoring this variable does not introduce bias

Ignoring A_2 induces an NMAR mechanism and nonresponse bias

A_3 cannot introduce bias nor can it increase power



Motivating Example

Data from a sample of 250 chronic pain patients

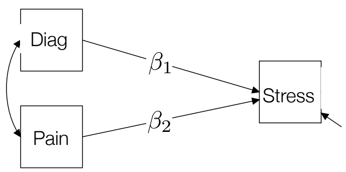
Variables include gender, the number of diagnosed physical ailments, sleep quality, pain ratings, positive and negative affect, and stress

Gender and number of diagnoses are complete, the remaining variables have up to 14% missing data

Analysis Model

The analysis is a multiple regression model that examines the influence of pain on stress, controlling for the number of diagnosed ailments

$$Stress = \beta_0 + \beta_1(Diagnose) + \beta_2(Pain) + e$$



Identifying Correlates of Missingness

Create a missing data indicator *R* for each incomplete variable (e.g., 0 = complete, 1 = missing) and examine correlations with variables not in the analysis

Correlations are the same as *t* tests with indicators as grouping variables but are easier to implement

Indicator correlations with non-zero effect sizes (e.g., greater than ± .10) identify potential auxiliary variables

Bivariate Correlations

	Female	Diagnose	Sleep	Pain	PosAff	NegAff	Stress
Female	1.00						
Diagnose	0.43	1.00					
Sleep	0.04	-0.21	1.00				
Pain	0.45	0.44	-0.32	1.00			
PosAff	0.11	-0.21	0.45	-0.19	1.00		
NegAff	-0.02	0.02	-0.02	-0.03	-0.24	1.00	
Stress	0.08	0.12	-0.23	0.29	-0.30	0.45	1.00

= variables in the analysis model, cannot be auxiliary variables

= potential auxiliary variables

Indicator Variable Correlations

Missing data handling automatically conditions on analysis variables

Indicators correlate with gender and to a lesser degree positive affect

	R_{Pain}	R_{Stress}
Female	0.11	-0.32
Diagnose	0.27	-0.04
Sleep	-0.01	0.08
Pain	NA	-0.16
PosAff	-0.14	0.10
NegAff	0.07	-0.03
Stress	0.09	NA

Conclusions

Gender and positive affect are potentially useful auxiliary variables because they correlate with missingness

Conditioning on gender may reduce non-response bias because it also correlates with analysis variables, and conditioning on positive affect may improve power

Include the extra variables in the imputation phase then ignore them when analyzing the data