Challenges for Imputing Missing Covariates in Meta-Regression

Jacob M. Schauer
jms@u.northwestern.edu
Institute for Policy Research
Northwestern University

NSF Grant #1841075

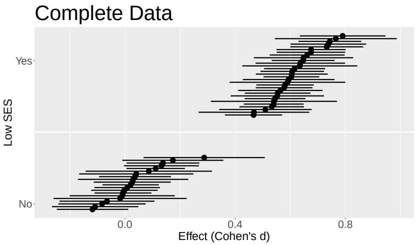
SRSM 2019

A Simple Meta-Regression

Is a reading intervention more/less effective for low-SES students?

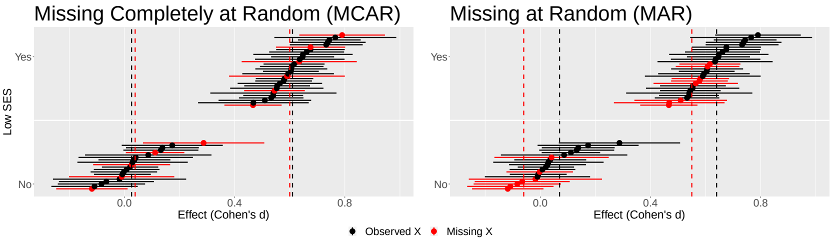
- Effect size esimates Y_i.
- Variance of effect estimates v_i.
- Did study i involve (mostly) students who are low-SES?
  - $X_i = 1$ for low-SES students.
  - $X_i = 0$ for not low-SES students.
  - X_i missing if information not given.
- Subgroup analysis/ANOVA (Hedges, 1982)
- Are the reasons the X_i are missing related to things you observe or not?

Missing (Completely) at Random



Missingness unrelated to everything

Depends on effect size/variance



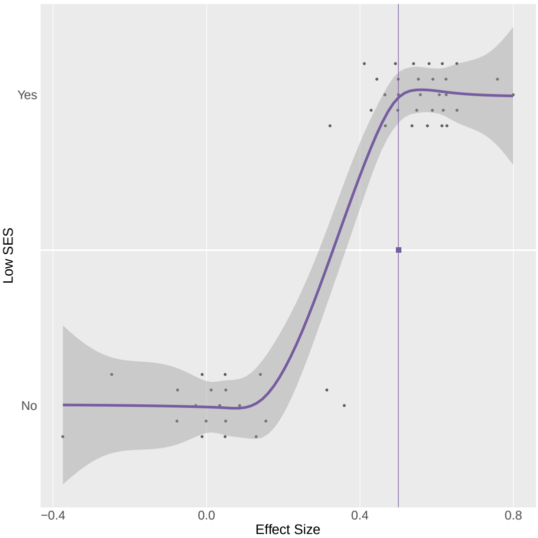
Handling Data Missing at Random

- **Complete case analysis:** Only analyze studies for which we observe effect sizes, variances, and predictors.
    - Common approach in practice (Tipton, Pustejovsky, & Ahmadi, 2019).
    - May only be a few complete cases, so **standard errors will be large**.
    - Unless the data are missing *completely* at random, **estimates can be biased** (Pigott, 2001; Rubin, 1976).
- **Multiple imputation:** Fill in values for missing predictors.
    - Strong theoretical/empirical justification (Little & Rubin, 1987; Rubin, 1996).
    - Commonly used in many other fields.
    - Tons of great software (e.g., `mice` in `R`)

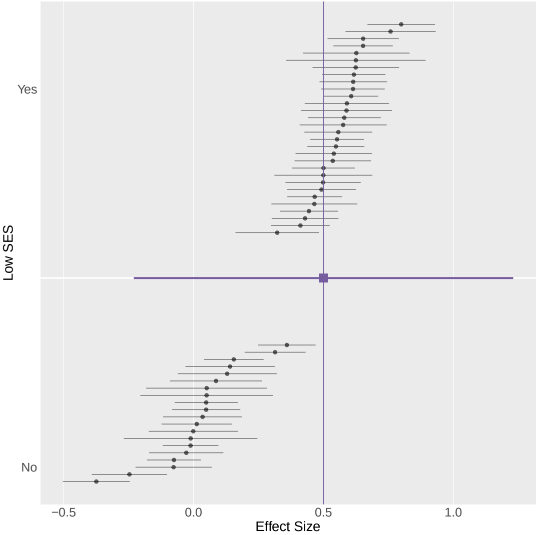
Multiple Imputation

- Fill in values for the missing X_i according to some predictive model.
  - Fill in multiple m values for each missing X_i, creating m complete datasets.
  - Run regression model on each dataset.
  - Pool regression models across datasets.
- Accuracy/validity can depend on the way in which we fill in the X_i (Little & Rubin, 1987).
  - Predict X_i given Y_i and/or v_i (and other observed data)
  - We want to use **appropriate** models to fill in X_i.

Example: Imputing X_i



Example: Imputing X_i



(Statistical) Compatibility

- In order guarantee MI produces valid inferences, imputation models need to be **congenial** with the regression model (Meng, 1994; Liu et al., 2014).
  - What the Y_i and v_i say about the X_i should reflect what the X_i say about the Y_i.
- Explicitly using the analysis model as part of the imputation model can help ensure congenial imputations (Bartlett, et al., 2015).
- While common software are flexible, **none provide congenial imputations for meta-regression/sub-group analyses**.
  - Only on my laptop for now!

Implications

- Using standard MI software results in **uncongenial imputations** for meta-regression.
- Analyses using **uncongenial imputations** are not guaranteed to produce unbiased or appropriately precise results.
- Using out-of-the-box MI software (e.g., `mice`) **can result in bias**.
- **Compatible imputations are possible** and can provide unbiased and more precise analyses, but are not widely available (yet).

Forthcoming Findings

- If there is a very small amount of missing data, complete-case analysis and out-of-the-box software may work reasonably well.
- If there is a moderate to large amount of missing data:
  - **complete case analyses will be biased** (or impossible)
  - **out-of-the-box MI software will also be biased**, but can be less biased and more precise than complete-case analyses
  - **MI with compatible imputations will have very small or no bias**, and will often be more precise than than incompatible imputations.

Thank you!

jms@u.northwestern.edu

Works Cited

- Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4), 462—487.
- Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- Meng, Xiao-Li. (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558.
- Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation & the Health Professions*, 24(3), 277—307.
- Rubin, D. B. (1996). Multiple imputation after 18+ Years. *Journal of the American Statistical Association*, 91, 473–489.

(Statistical) Compatibility

The model we use to impute X_i must be **compatible** with the regression model.

- The imputation model $g(X_i|Y_i, v_i, \eta)$ and regression model $f(Y_i|X_i, v_i, \tau^2, \beta)$ must imply a joint distribution $p(X_i, Y_i|v_i, \eta, \beta, \tau^2)$ that exists.

In meta-regression, a compatible model can be written as (Bartlett et al., 2015):

$$g(X|Y, v, \beta, \tau^2) \propto f(Y|X, v, \beta, \tau^2)$$

Compatible Imputations for Predictors: Example

Imputing a single binary X:

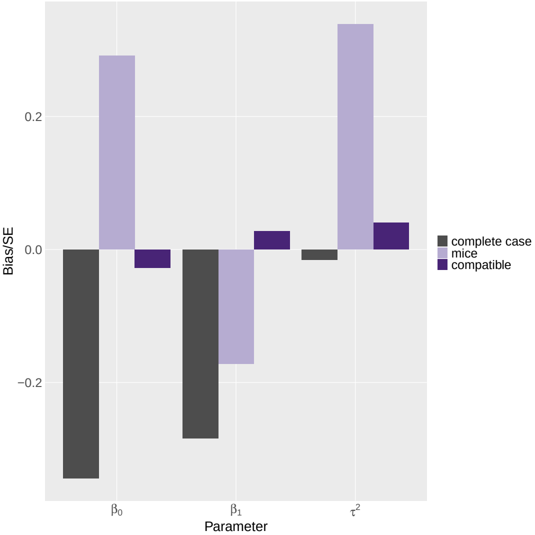
1. Fit meta-regression model on complete cases.
2. Regress Y on X for complete cases.
3. Draw β, τ^2 from their posterior distributions.
4. Draw $X \sim Bernoulli(\pi)$ where $\pi = \frac{f(Y|1,v,\beta,\tau^2)}{f(Y|1,v,\beta,\tau^2)+f(Y|0,v,\beta,\tau^2)}$
5. Repeat 3-4 m times.

Standard implementations of MI do not have options for imputations that are compatible for meta-regression.

Simulation Details

- $\beta_0 = 0.025$
- $\beta_1 = 0.575$
- $\tau^2 = 0.005$
- $v_i \in [0.003, 0.1]$
- $k = 50$
- $X_i = 1$ for 30 studies (truth)
- Missing about 50% of X_i
  - Generate 5,000 datasets $(X, Y)_j$ according the the model.
  - Randomly delete X_i with a probability that depends on
  - Run `mice` and to impute and estimate regression model.
  - Use compatible imputation model to estimate regression model.

Results: Bias



Results: Standard Errors