Sarah Ryan posted on Thursday, April 21, 2011 - 2:23 pm
Hello Again,

Following up here on the post above. I did go ahead and read Enders' book, which helped my understanding of MI immensely. I remain undecided. I have a rather complex model- several covariates (missingness on some, one reason making MI attractive), two latent exogenous constructs, several manifest exogenous measures, a latent mediator and continuous outcome. I will also use multiple group invariance testing. This is complex survey data and I'm using data from both parents and students. Student missingness ranges from 0% to 11%, but parent missingness ranges from 0% to 28% (16%-20% missingness on most parent-level varbs).

I know I would need to run MI for the full sample (includes all race/eth groups) as well as for each of the two subgroups of comparison (race-level comp). I also gather from reading on these boards that it would be best to fix parameter estimates for one of the MI sets (for full sample, and each subgroup) at the pooled imputation average AND THEN run the analysis model using FIML.

I'm wondering if, given the high levels of parent-missing and some missing on covariates, using MI would produce more accurate analysis model parameter estimates.

Do you have a stronger (convince me, please!) argument for why staying within the FIML framework (which would be simpler) would likely still produce just as accurate estimates even giving missingness issues?

Craig Enders posted on Thursday, April 21, 2011 - 4:42 pm
Sarah,

In my mind, the answer to your question largely depends on the scaling of your variables, particularly the covariates. My post is a bit long, so I have to split it into separate chunks.

First, suppose that all of your variables are continuously scaled, or approximately so. In this case, I think that the choice of FIML vs. MI is personal preference. Theoretically, there is no reason to expect noticeable performance differences (FIML might have slightly smaller SEs because MI uses a more complex saturated model to deal with missing data, but this difference is usually negligible). There are a couple nuances here. With FIML, you will need to take care of the missing data on the covariates (you seem to be aware of this issue). Suppose you have two covariates, X1 and X2. You would do this as follows:

X1 X2;
X1 with X2;

As I understand it, explicitly listing X1 and X2 effectively makes the covariates single indicators of a latent construct -- a programming trick that converts the Xs to Ys while still maintaining the exogenous status of the variables in the model. The missing data on your outcomes would be automatically handled by FIML.

Craig Enders posted on Thursday, April 21, 2011 - 4:43 pm
Same situation -- continuous variables. Turning to MI, the imputation process is a bit simpler than you describe. You would impute the data separately for your ethnic groups -- there is NO need to then impute the data for the whole sample. Separate-group imputation automatically fills in the missing values in a way that preserves mean differences among the ethnic groups and all group by variable interactions in the data. Said differently, you would have filled in the data using the most general model possible, and the set of imputations that you get from this procedure would be appropriate for all your analyses (the imputation routine would need to include all covariates, outcomes, etc.). I'm not sure what you are referring to when you talk about fixing the estimates at the pooled average, then running the model using FIML. I *think* you might be describing the method for pooling likelihood ratio tests, but that would not be necessary -- Mplus reports the pooled chi-square. On my website (appliedmissingdata.com)I have an example of separate-group imputation in Mplus 6.

Craig Enders posted on Thursday, April 21, 2011 - 4:45 pm
Next, suppose that one of your covariates is binary. I think that the situation becomes murkier here. Take that programming trick that lists the variances of the X variables and their covariances. Again, as I understand it, this would make the covariate a single indicator of a latent construct. However, it isn't so clear to me that this is appropriate with a binary covariate, because you would be assuming normality for the covariate, and employing a linear factor model to convert the X to a pseudo Y might cause problems. The first problem is more conceptual; the programming trick does not produce a model that is a one to one translation of the complete-data analysis. Whether this introduces bias, I don't know. Second, I could imagine that Mplus would issue a warning about the standard errors because the mean and the variance of the binary variable are linearly dependent when you use the linear factor model to handle the missing data. I'm not sure that either of the problems are substantial ...

MI would provide a useful alternative. Mplus allows you to specify variables as categorical or continuous in the imputation model. In the case of an incomplete categorical covariate, imputation would use a logistic regression to fill in plausible values. The MI procedure would be identical to what I describe above (impute separately for each ethnic group). You would simply use the (c) option to denote categorical variables.

Craig Enders posted on Thursday, April 21, 2011 - 4:48 pm
Continuing on ... Finally, if your outcomes/indicators are not continuous -- say Likert items -- then the choice isn't all that clear to me. FIML would assume multivariate normality, as you know. The missing data handling probably won't behave any differently than complete-data ML. With MI, you have two options: (a) use linear regression to impute the incomplete variables, or (b) use logistic regression to impute. The former assumes normality and would produce fractional imputed values (not a

problem beyond aesthetics, you would not want to round these). The latter does not assume normality and would produce discrete imputations. I know of no studies that have compared these two imputation approaches, but I suspect that the logistic imputation might lead to larger SEs because the multinomial model would have more parameters than a linear model. I would probably still go with FIML and MLR standard errors , but the choice isn't so clear cut.

The other thing to consider is your model testing procedures. It sounds like you plan to perform likelihood ratio tests. FIML is probably going to be easier to deal with here. Mplus computes pooled LR tests, but I'm not sure if there is a way to automate the computation of these tests when you are comparing fit between models from two separate analysis runs (e.g., your invariance tests). If you have to compute the LR tests by brute force, it would be time consuming.