LATENT VARIABLE MODELING

Missing data & Sample Size

Today's topics

- Missing data
- Sample size recommendations

Missing data

- Missing Completely At Random (MCAR)
  - As name implies, missingness is completely random (i.e., not associated with any variables in or outside of the model
  - Nothing systematic that makes some values more likely to be missing than others
- Missing At Random (MAR)
  - There is a systematic relationship between the missingness and the *observed* data, but *not* the missing data
  - Whether an observation is missing is independent from the missing value itself, but it may be dependent on other observed variables in the model
  - E.g., dataset with gender and weight. Gender has no missings, weight has missings
    - If women are more likely to have weight missing, missing is MAR
    - If people with higher weight are more likely to have missing weight, missing is MNAR
- Missing Not At Random (MNAR)
  - Missingness and missing values are dependent

Missing data

- Missing Completely At Random (MCAR)

  Unlikely in practice

  - As name implies, missingness is completely random ated with any variables in or outside of the model
  - Nothing systematic that makes some values more likely to be missing than others
- Missing At Random (MAR)

  Can be dealt with in analysis

  - There is a systematic relationship be gness and the *observed* data, but *not* the missing data
  - Whether an observation is missing is independent from the missing value itself, but it may be dependent on other observed variables in the model
  - E.g., dataset with gender and weight. Gender has no missings, weight has missings
    - If women are more likely to have weight missing, missing is MAR
    - If people with higher weight are more likely to have missing weight, missing is MNAR
- Missing Not At Random (MNAR)

  Problematic

  - Missingness and missing values are dependent

What is MNAR can be made MAR

- Example: Suppose, in a study on weight and depressive symptoms, women were more likely not to report their weight than men
- If gender is not included in the model, missing is MNAR
- If gender is included in the model as a predictor of weight, missingness in weight is made MAR
  - Assuming gendert is the only relevant predictor of missingness
- Thus, can make MNAR into MAR by adding variables to the model ('auxiliary variables' approach)

Missing data

- Listwise deletion
  - Best avoided, power much reduced
- Analyse covariance matrix based on pairwise complete observations
  - Unbiased parameter estimates when missing data are MAR
- Use full information maximum likelihood (FIML) estimation
  - Yields unbiased parameter estimates when missing data are MAR
  - Can add 'auxiliary variable(s)' (i.e., variable(s) that contain information about missing values, but were not part of the original model) to turn MNAR into MAR
- Multiple imputation (MI)
  - Outside scope of this course
  - Both FIML and MI perform similarly well when data are MCAR or MAR (Schafer & Graham, 2002)

Missing data

- If missing data < 5%, missingness is likely to be inconsequential
  - But listwise deletion is never a good idea (even though often the default in statistical software) as it yields more missing data
- Missing data *always* reduces power and increases standard errors, because less information has been observed from sample
- In lavaan:
  - For (robust) ML estimation: use FIML
    - Invoked by specifying missing="fiml" in call to functions lavaan(), cfa(), growth(), sem()
  - When using LS-type estimation (e.g., DWLS, with ordered categorical indicators): use pairwise complete observations
    - Invoked by specifying missing="pairwise" in call to functions lavaan(), cfa(), growth(), sem()
    - Correlation matrix (Pearson, tetra- or polychoric) is computed using pairwise complete observations and model is fitted on that matrix
    - Quite similar to what FIML does

Full information maximum likelihood

- Remember that with maximum likelihood estimation, value of the log-likelihood (LL) is maximized
  - The LL of the full dataset is the sum of the LLs of the individual observations
- When an observation has one or more missing values, its LL is computed based on only the variables that were observed
  - Thus, an observation contributes only to the parameter estimates involving observed variables it has non-missing values for
  - Note: Very similar to pairwise-complete approach

Sample size guidelines

- Many authors advice that *N:q* (the ratio of sample size to the number of estimated parameters) should equal at least 10 or 20 for SEM analyses to be adequately powered
- However, what is an adequately powered sample size depends not only on the *N:q* ratio, but also on:
  - Effect size: the size of (standardized) parameter estimates
    - Stronger effects are easier to recover
  - Number of indicators for a latent variable
    - More indicators per latent variable yields higher power, given same sample size
    - A 1-factor CFA with 6 items would actually require a ***lower*** number of observations than a 1-factor CFA with 3 items
      - May seem counterintuitive, because more parameters are estimated

Sample size guidelines

- Any SEM requires at least 200 observations (Kline, 2015)
- Sample size requirements range from 30 (for simple CFAs with four indicators and loadings around .80) up to 450 cases (mediation models) (Wolf et al., 2013)

References

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Press,.

Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*(2), 147-177.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models an evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934.