SQM

Week 5: Model Evaluation

FIT

- Fit of an SEM is defined as the differences between the observed covariance (correlation) matrix and the matrix as implied by the SEM model.
- These differences are the quantities that sem models try to minimize; the exact function is different between estimations methos, of which Stata offers two: ML and ADF. (Other SEM programs have more options, but this matters little in practice).
- Most commonly we use an overall statistics, which expresses the differences into a single quantity.
- The difference can be expressed in multiple ways, and a host of fit statistics has arisen around these model. Stata can show a modest amount.

The logic

- In testing an SEM model, we have to get used to a reversed H0/H1 situation. We want the (over-all) fit statistic to be as small as possible and preferably non-significant.
- This is so, because the research question is different. In common regression / anova models, we only ask about the significance of individual parameters, the model by itself always fits perfectly.
- Of course, SEN models also have test of significance of individual parameters. These tests have the usual interpretation.

LR-test

- The most commonly reported over-all fit statistic is the LR-test, which is connected with ML-estimation (most commonly used.)
- The test is relative to the DF (degrees of freedom) which is: # covariances.
- Notice that you should adjust the degrees of freedom if you analyze standardized data (correlations) and ignore (do not model) means and standard deviations – which I usually do.
- LR follows a chi-squared distribution, so the critical value is 3.84 in a 1 degree of freedom test.

LR and N

- If the specified model is the true population model and all distributional assumptions hold, the LR statistic will be non-significant at any N. (This applied in our simulated data.)
- However, neither condition is ever perfectly realized in real life.
- If the specified model is not exactly the population model *or* the distributional assumption do not exactly hold, LR will increase with N, and assessing fit becomes different from testing significance. The LR becomes very powerful: you reject H0 all the time, despite the misfit being trivial.
- One common approach to this problem is to test on an standard fictional N (e.g. 1000), effectively saying: the misfit would not be significant in a 1000 cases.
- Another approach has been to construct a fit index, which is a number that standardize LR on some maximum number. I do not find these numbers very informative.

Hierarchical testing

- A useful way to apply LR-statistics is for comparison between models. (This is similar to F-change in regression and anova models with dummy or polynomial predictors.)
- You make a table that starts at a model with many effects (possible saturated, at last at the latent level) and constrain them step-by-step to be equal of zero. You can also work the other way around.
- The question at each step: is the difference in LR statistically significant?
- These comparisons are useful, even if all the LR's are statistically significant!
- However, this approach remains vulnerable to large N.

Information Criteria

- Ideally, you want your model evaluation to be not too sensitive to large N – you do not want you power to be too large.
- After all, testing is all about the H0 – which may not interest you altogether.
- Information criteria discount the N. The most often used criterion is BIC (Bayesian Information Criterion). General rule: the lower the better.

Residuals and Modification

- One can also look at fit at a more detailed level, which are the residuals in the covariance (correlation) matrix. (We have seen these also in SPSS factor analysis.)
- However, SEM models try to minimize these residuals and close fit may therefore be misleading in individual numbers.
- Another way to look at details are the modification indices (which are displayed for each individual fixed parameter), that predict how much LR would change if a parameter is estimated in stead of fixed.
- Modification indices relate to residuals like influence statistics to residuals in OLS regression.
- ***Residuals can be misleading, but modification indices can (and often are!) nonsensical.*** *(Use a combination.)*

RMSEA

- RMSEA is a popular statistic in SEM evaluation that mixes LR, (mean) residuals and discounting N.
- It is like an average residual at large N, but makes adjustments when N become smaller.
- RMSEA also comes with an additional statistical test: it tests against a non-standard H0, usually taken at rmsea = .05.
- RMSEA may own it popularity to the fact that it is displayed as the standard statistic in the LISREL program; however, I think it works pretty well.

A more qualitative approach

- I tend to discard significance testing in SEM models with large N altogether. SEM models have too much statistical power and over-all fit tests address an uninteresting question: the H0.
- Think about an OLS situation in which you would have enormous N: what is the point of significance testing – are you ever going to make a type II error?
- In stead I favor a qualitative approach: how much do (other) model parameters (and their interpretation) change, when I introduce additional constraints in the model?
- Your argument then should be about specific parameters of interest: how do they change if you change the model specification?
- Notice that this does not mean that you should disregard significance testing for individual parameters.