SEM with observed variables: estimation
Estimation

- Tries to find a solution that best approximates ideally population covariance matrix \( \Sigma \), but in reality a sample estimate \( S \)

\[
F = f\left( \Sigma(\hat{\theta}) \mid S \right)
\]

- The “best” approximation is defined in various ways, leading to different fitting functions

- Our job is to find which is the best under what data conditions
Desirable fitting functions

- have the following properties:

\[ f(S, \Sigma(\hat{\theta})) \] is a scalar

\[ f(S, \Sigma(\hat{\theta})) \geq 0 \]

\[ f(S, \Sigma(\hat{\theta})) = 0 \text{ if and only if } \Sigma(\hat{\theta}) = S \]

\[ f'(S, \Sigma(\hat{\theta})) \] is continuous both in \( S \) and \( \Sigma(\hat{\theta}) \)

- Minimizing such fitting functions provides a consistent estimator of \( \theta \) (e.g., ML, ULS and GLS) --- true for all functions to be considered
Desirable asymptotic properties of estimators

• Unbiased

• Consistent

• Efficient

Note: asymptotic means “$N \to \infty$” by definition but its practical meaning is “as $N$ becomes sufficiently large” --- “how large is sufficient” will depend on many things such as complexity of the model, size of measurement errors, etc.
\( \theta \): parameters in the population

\( \hat{\theta}_N \): estimate of \( \theta \) from a sample of size \( N \)

- If \( E(\hat{\theta}_N) = \theta \), \( \hat{\theta}_N \) is unbiased

- If \( E(\hat{\theta}_N) = \theta \) as \( N \to \infty \), \( \hat{\theta}_N \) is asymptotically unbiased

- If \( P\left(|\hat{\theta}_N - \theta| < \alpha\right) = 1 \) as \( N \to \infty \) for any \( \alpha > 0 \), \( \hat{\theta}_N \) is consistent --- or called "\( \text{plim} \hat{\theta} = \theta \)"

- \( \hat{\theta}_N \) is efficient if its asymptotic variance is the minimum of all consistent estimator of \( \theta \) --- see Appendix B for more details on asymptotic properties of estimators
Maximum likelihood

• ML assumes:
  - Satisfactorily large sample
  - All observed variables distributed multivariate normal --- we will consider later a relaxed alternative to this for exogenous \( x \)
  - All observations independent and identically distributed

• Minimizing its fitting function \( F_{ML} \) maximizes joint log likelihood of the model parameters \( \theta \) given observed data \( S \)

\[
F_{ML} = \log|\hat{\Sigma}| + \text{tr}(S\hat{\Sigma}^{-1}) - \log|S| - (p + q)
\]

Obviously, both \( S \) and \( \hat{\Sigma} \) must be nonsingular for \( F_{ML} \) to be defined
How to optimize a fitting function?

• Find the partial derivatives w.r.t. all free model parameters \( \partial F(\theta)/\partial \theta \) and solve \( \partial F(\theta)/\partial \theta = 0 \) --- necessary for minimization

• The second-derivative matrix \( \partial^2 F(\theta)/\partial \theta \partial \theta' \) is positive definite (nonsingular) at the \( \hat{\theta} \) that minimizes \( f(\theta) \) --- sufficient for minimization

• Usually, there is no closed form solution to this problem; instead, the minimization attained by an iterative numerical method based on gradients \( g \) (i.e., the 1st derivatives), given

  1. starting values of \( \hat{\theta} \)
  2. step length (i.e., how much to change \( \hat{\theta} \) per iteration)
  3. when to stop iteration
• Parameters are initialized at some rational (instead of random) values --- e.g., reports from prior research, OLS estimates for loadings and causal paths, etc.

• The iterative update can be written as:

\[ \hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - C^{(i)}g^{(i)} \]

  ➢ Steepest descent: \( C = I \)

  ➢ Newton-Raphson: \( C = H^{-1}, \ H = 2^{nd} \) partial derivatives (so called Hessian matrix)

• Iteration stops when \[ f(\hat{\theta})^{(i)} - f(\hat{\theta})^{(i+1)} < \alpha \] or \[ \hat{\theta}^{(i)} - \hat{\theta}^{(i+1)} < \alpha \] for all parameters, with \( \alpha \) at an arbitrarily small value (e.g., \( 10^{-7} \))
Properties of ML

- Asymptotically unbiased
- Consistent
- Efficient

- ML estimates are asymptotically normal --- while the 1\textsuperscript{st} PDs = 0 provides the estimates, square-roots of the diagonal entries of the 2\textsuperscript{nd} PDs give their standard errors, allowing for a z-test

- $F_{\text{ML}}$ is scale invariant:

$$F_{\text{ML}} \left( S, \hat{\Sigma} \right) = F_{\text{ML}} \left( DSD, D\hat{\Sigma}D \right)$$

with a diagonal matrix $D$, with diagonal entries all non-zero
• Scale freeness of ML estimates of $\theta$: $y = B y + \Gamma x + \zeta$ is functionally equivalent to $\tilde{y} = \tilde{B} \tilde{y} + \tilde{\Gamma} \tilde{x} + \tilde{\zeta}$ with $\tilde{y} = D_y y$, $\tilde{x} = D_x x$ and so $\tilde{B} = D_y B D_y^{-1}$, $\tilde{\Gamma} = D_y \Gamma D_x^{-1}$, $\tilde{\zeta} = D_y \zeta$

e.g., with $D_y = \text{diag}(s_{y_1}^{-1},...,s_{y_p}^{-1})$ and $D_x = \text{diag}(s_{x_1}^{-1},...,s_{x_q}^{-1})$, $\tilde{y}$ and $\tilde{x}$ become standardized variables, and $\tilde{B}$ and $\tilde{\Gamma}$ become standardized estimates of parameters

➢ What happens to the invariance and freeness if some parameters are subject to non-zero constant or equality constraints?
• Overall (badness of) fit of identifiable models:

\[(N - 1)F_{ML} \sim \chi^2\]

with \(H_0: S = \hat{\Sigma}, \) \(df = (p + q)(p + q + 1)/2 - t\)

- **Practical dilemma:** asymptotic theory requires sufficiently large \(N\) and the chi-square statistic proportionally increases with \(N\) while its model \(df\) doesn’t change

- **Relaxed condition for exogenous \(x\):** observed exogenous \(x\) don’t have to be multinormal; instead, if \(y\) conditional to \(x\) is multinormal and \(x\) is independent of \(\theta\), then all ML properties hold
Unweighted least squares

- ULS assumes nothing
- Consistent but not efficient
- Not scale invariant or scale free
- No statistical testing available; bootstrapping may be used for the overall fit

\[
F_{ULS} = \frac{1}{2} \text{tr} \left[ (S - \hat{\Sigma})^2 \right]
\]

*cf.* \[
\text{tr} \left[ (z - \hat{z})(z - \hat{z})' \right], \quad z = \begin{bmatrix} y \\ x \end{bmatrix}
\]
Generalized least squares

- GLS adjusts the residual matrix, $S - \hat{\Sigma}$ for unequal variances and covariances

$$F_{\text{GLS}} = \frac{1}{2} \text{tr} \left[ \left( S - \hat{\Sigma} \right) W^{-1} \right]^2$$

Analogous to weighted least-squares for OLS --- typically used to deal with heterogeneous variance over observations (vertically written):

$$\hat{b}_{\text{OLS}} = (X'X)^{-1} X'y$$

$$\hat{b}_{\text{WLS}} = (X'W^{-1}X)^{-1} X'W^{-1}y, \quad W = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$$
• Assuming that $W$ is positive-definite (asymptotically or given so), GLS estimates $\hat{\theta}$ are
  
  ➢ consistent
  
  ➢ distributed multinormal with known ACOV, and so a z-test available
  
  ➢ but not all $W^{-1}$ lead to efficient estimators

• Two more conditions needed for efficiency:
  
  ➢ $s_{ij}$ is unbiased estimator of $\sigma_{ij}$
  
  ➢ Entries of $S$ are asymptotically distributed multinormal, with mean of $\sigma_{ij}$ and asymptotic covariance of

  $$\text{acov} \left( s_{ij}, s_{gh} \right) = N^{-1} \left( \sigma_{ig} \sigma_{jh} + \sigma_{ih} \sigma_{jg} \right)$$
The ACOV condition is satisfied if $x$ and $y$ are multinormal, or if “tail” of the distribution is not excessively thick or thin

- Most common choice of $W$ is $S$, and so $\text{plim} W = \Sigma$

$$F_{GLS} = \frac{1}{2} \text{tr} \left[ \left( I - \Sigma(\theta)S^{-1} \right)^2 \right]$$

which is scale invariant and its estimates are scale free

- Like $F_{ML}$, $(N - 1)F_{GLS}$ is asymptotically chi-square distributed; consequently, with large $N$, $F_{ML}$ and $F_{GLS}$ should be close to each other if $H_0$ is true (i.e., the model correctly specified) --- see, e.g., Table 4.3, p. 121
What if parametric assumptions not met

• Give up statistical inferences

• Use goodness (or badness) of fit indices (e.g., GFI, CFI, TLI, RMSEA, AIC, BIC, etc.) --- these non-statistical measures assume nothing and, in consequence, parametric testing unavailable

• Use non-parametric testing (e.g., bootstrap, randomization, etc.) --- to be elaborated in the following

• If nonnormality is the only major issue, data might be transformed to approximate normality --- consequently, the modeled linear relationships are for the transformed variables, which should be substantively meaningful
The bootstrap

• The idea of bootstrapping evolved from jackknifing (due to Tukey) by introducing random selection of observational units from the sample, with replacement.

• The bootstrap is one way of “empirically” obtaining a sampling distribution of a statistic (e.g., $F_{ML}$, $F_{GLS}$, direct & indirect effects, etc.).

• The bootstrap distribution is asymptotically normal (as original $N \to \infty$).

• When only covariance or correlation data are provided, resampling is not possible --- pseudo raw data may be sampled from a parametric distribution, e.g., $N(0, S)$ --- called “parametric bootstrapping.”
"Bollen-Stine" bootstrap method

• Transforms the data $X$ so that the resulting covariance matrix equals to model-implied cov matrix $\hat{\Sigma}$ as (vertically written):

$$Z = XS^{-0.5} \hat{\Sigma}^{0.5}$$

which makes $H_0$ true in the bootstrap population $Z$, and sampling error of the statistic becomes smaller ("efficient") leading to a more confident statistical inference

• The idea of transforming $X$ to $Z$ with which $H_0$ is true was given by Beran and Srivastava (1985) more generally for the eigen structure of $\text{cov}(X)$


• Is the Bollen-Stine method always good? --- depends on whether the specified model is correct:

The transformation “forces” $H_0$ to be true in the bootstrap population $Z$ so as to produce (narrower) sampling distribution of statistics (overall fits) for more powerful testing against $H_A$ (typically a nested model, $df_0 < df_A$)

However, when $H_0$ is substantially false, a test for $H_A$ based on a “wrong” transformation may be misleading
Scale and standardization

\[ Y = -3 + 0.5X_1 + 2X_2 + 1.8X_3 + \zeta \]
\[ \tilde{Y} = 0.67\tilde{X}_1 + 0.83\tilde{X}_2 + 0.75\tilde{X}_3 + \tilde{\zeta} \]  (standardized)

- \( Y \): job satisfaction (1 = least, 7 = most satisfied; \( s_y = 1.2 \))
- \( X_1 \): annual salary in $10K (\( s_{x1} = 1.6 \))
- \( X_2 \): annual bonus in $10K (\( s_{x2} = 0.5 \))
- \( X_3 \): gender (1 = female, 0 = male, \( s_{x3} = 0.5 \))

• How should we compare effect of the salary and bonus?

• What is meant by 1.8 of gender effect? Would it be comparable to other effects if standardized?

• If alternative equations considered only with \( X_1 \) and \( X_2 \), separately for each gender, any concern for comparability?