NOTES TO SEM LECTURES (in yellow: not explicitly discussed)

Harry BG Ganzeboom

Melbourne, October 24 2017

Lecture 1

Arguments why you should be doing SEM for the rest of your life

1. SEM makes you think about the world in terms of causality, which happens to be what science is all about.
2. SEM makes you aware that all we are observing is covariation (correlations); it is the task of a scientist to invent a causal theory / model that explains the covariation. The world that we see is a correlation matrix.
3. SEM allows the estimation of a causal process by simultaneous (or: a system of) equations; this has certain statistical advantages (to come), but it is also conceptually relevant.
4. SEM allows for an easy method to include all available data in your estimation, when there are missing values. The method is called MLMV (Maximum Likelihood with Missing Values) of FIML (Full Information Maximum Likelihood).
5. SEM makes you distinguish between the facts and your data (observations) and can estimate the relationship between the two: measurement error.
6. SEM allows you not only to diagnose measurement error, but also so to correct for it.

Regression models

- SEM's are nothing but a bunch ('system') of regression equations = linear models of simple or multiple (partial) associations. So you will have to understand first how to interpret regression equation output.
    - Y = b0 + b1*X1 + b2*X2 … + residual
    - Intercept B0: **expectation** of Y when all X's are zero
    - b1, b2: (partial) slopes: how many units of Y you get for each unit of X
    - Z-standardization: recalculating X and Y to Z-scores: Z(X) = (Xi – M(X))/SD(X). Z(X) expresses each Xi relative to its mean M(X), so expresses Xi in units of SD(X). This sounds pretty abstract, but it is in fact quite useful, because it makes variables and effects comparable to one another.
    - Standardized effects (beta) are effects between standardized variables. Beta(YX) = B(YX) * SD(Y)/SD(X). Like correlations, standardized effects vary between -1 and +1.
    - Effect strength ('effect size') cannot be determined from unstandardized B's. Cohen (1988, 1992) about standardized effects (beta's or correlations):

        - Around 0.10 weak
        - Around 0.30 moderate
        - Around 0.50 strong

        Most of our research is about ***weak*** effects. Moderate and strong effects are mostly trivial and do not need further research or testing (but they should be present).

    - Strength and significance of effects are two different things. An effect is said to be (statistically) significant if we can reject the H0 (no effect) with P < .05. P denotes the probability that the observed effect would arise in an infinite number of samples of size N drawn from the same population if the H0 holds in this population. Given sample size N, effects are more likely to be significant when effects are stronger. Given an effect size beta, it is more likely that you can reject the H0 when N is larger.

    - The crucial statistical element in any estimation procedure is the Standard Error (SE). Most students know about the SE(mean), which is estimated as SD/sqrt(N). However, all statistics (== quantities that summarize data) have an SE, including regression coefficients B and Beta. A useful conceptual definition of an SE is the amount of variation (as measured as the SD of the ***sampling distribution***) that arises if you would draw a large amount of samples of size N from the same population (think bootstrapping). Formula's for many SE's exist and have different ingredients, but 1/sqrt(N) is always one of them: SE's become smaller if N becomes larger (but NOT in a linear way).

- SE's of regression B's are a function of (=become smaller when):
    - N is larger
    - Explained variance (R2) is higher
    - The range of X is wider
    - There is less correlation between the X of interest and the other X-vars in the model **(collinearity)**.
- Other quantities of interest in a regression output:
    - Total Sum of Squares: SS-total
    - Model Sum of Squares: SS-model
    - Residual Sum of Squares: SS-residual

    When divided by their degrees of freedom, SS's become **variances** [mean squares]. When you take a square root of a variance, you are back at **'average' deviation**:

    - Sqrt(Total Variance) = SD.

- Sqrt(Residual Variance) = Root Mean Squared Residual. SPSS calls this quantity the Standard Error of the Estimate. [I find this rather confusing, better would be: Standard Deviation of the Residuals.]

- R2 = SS-model / SS-total = 1 – (SS-residual / SS-total).

- Correlation R = sqrt(R2).

Causal models

- Causal models combine two or more regression models. A useful way to start is the **elementary causal model** that shows the effects of a causal variable X on an outcome Y, controlled for a **confounding** variable Z. The same model can also be interpreted as how X influences Y via a **direct effect** and an **indirect** effect via M. In this reading of the model M is a **mediating** variable. Distinguishing between mediators and confounders is crucial step in causal analysis.

- While we can calculate the effects in a causal model using regression equations, we can also find the effects from the correlations (or: covariances) and the path-analytic theorem:

  Correlation = direct effect + indirect effects + confounding effects

  In which indirect effects and confounding effects are defined as a multiplication of their constituting direct effects: a ***chain*** and a ***fork.***

- Between three variables X M Y in the elementary causal model this works out as the following system of equations:

  - r(MX) = a
  - r(YX) = c + a*b
  - r(YM) = b + a*c

  We have three equations (one for each correlation) with three unknowns (a b c), which can be solved with some high school algebra. The system is ***exactly identified***.

- Notice that this operation leads us to the standardized regression coefficients without actually doing regression analysis!! It is important to understand the operation because it is the way SEM programs work. Also note that we do not need the individual data to do this – we only need the correlations!

- (We could set up the system of equations also with covariances rather than correlations. This would lead us to the unstandardized regression coefficients – as well as the variances of X M Y.)

- The path-analytic theorem allows us to calculate direct, indirect and confounding effects and to express these as percentages (of total effects). In the given example father's occupation influences ***(explains)*** occupation for about 65% ***via*** education, the rest is direct ***(=unexplained)***. This operation is called **mediation** analysis.

Mediation analysis

- In mediation analysis we are interested in how much of a total effect (X → Y) runs via M. In practical analyses (using plain regressions), the focus is often on how the total effect changes into a direct (= 'remaining') effect, in a stepwise ('forward') analysis. In a sense this is the wrong focus. Rather, we should be looking at the size of the indirect effect and its (two) constituents. Notice that in the usual mode of operation (popularized by Baron & Kenny (1986)) does not allow for the calculation of the percentage explained, when the indirect effect has a different sign than the direct effect ('suppression').

- The SE and the significance of the indirect effect are not provided by the usual regression approach. For this we need the sobel-test, inspired by Sobel (1982). It is available online (http://quantpsy.org/sobel/sobel.htm), but now also as a Stata **ado** procedure: **sgmediation**. In the online test, we would need to specify either one of:

    - Effects a and b, and their associated SE's. In this case you obtain the SE and significance of the indirect effect (but not its size!!).

    - The T-values for a and b. In this case you obtain the significance of the indirect effect.

    NB1: The size of the indirect effect is a*b.

    NB2: Stata's **sgmediation** produces all the right quantities.

- Mediation (or: indirect effect) analysis is often done with standardized coefficients, which has indeed an advantage. You do not only want to make a comparison between c and a*b (or: between c' and a*b), but also between a and b. This is only possible when X M Y are all expressed in the same unit.

    NB: this is the reason why some (economists) would say that path analysis is nonsensical…

- The SE and the significance of the indirect effect can also be obtained from the SEM estimation, using **estat teffects**.

Measurement models

- A measurement model is just a special case of a causal model – and the same basic path-analytic theorem applies. We assume the presence of latent variables that ***cause*** the (cor)relations between observed indicators. Only confounding effects ('forks') are in the model. In the case of three indicators and one latent variable, we obtain:

    - r12 = a*b

    - r23 = b*c

    - r13 = a*c

- A measurement model is exactly identified with three indicators. Again we can work out the coefficients with some high school algebra. This model (also called: ***common factor analysis***) leads us to the untrue but useful rule that for perfect measurement you need three indicators.

- If we have more than three indicators in a measurement procedure, two phenomena arise:

- When there still is only one latent factor at work, we can estimate the measurement coefficients (factor loadings) with more precision (smaller standard errors).
- We can test / examine whether there is more than one latent variable at play (***multiple*** factor analysis).

- If we have only two indicators, the factor loadings are not identified, although we can still have an idea about their average value (take a=b), min and max.
- If we have only one indicator, there seems to be no way to learn about measurement error. Let alone we would have no indicator at all. ***Or so it seems***.
- A latent variable is NOT a column in your data matrix, it cannot even be constructed as such – it is principally unobserved. However, do not take this to mean that the latent score does not exist. RATHER: the latent score / variable is the real thing (the "facts"), the stuff you see in your data matrix are only numbers that represent these real things in an imperfect way (think Plato in his cave). To me, this is one of the principal reasons why SEM is so good: it makes you think about the difference between data and facts.

Combining structural and measurement models

- SEM's can be usefully thought off as combining structural (causal) models and measurement models in one system of equations. This system implies a complex bunch of equations, one for each observed correlation, still using the same path-analytic theorem.
- All SEM programs do is solve this system of equations, using slightly more complicated algebra than what we learned in high school.
- Unlike the two elementary models, more complicated systems of equations are often ***overidentified***: there are more equations than unknowns. In this case there generally is not an exactly fitting solution of the coefficients. SEM solves the coefficients is such a way that the covariance / correlation matrix implied by the model is closest to the observed matrix. The difference between the two ***(misfit)*** can be tested, using LR (chi-squared) statistics. Traditionally, SEM modelers care a lot about "fit". The Stata SEM program pays less attention to fit than other SEM programs.

Stata

- Stata files can be produced by SPSS. SPSS can read Stata(10) files, but Stata cannot read SPSS files.
- Stata is case sensitive. The most useful command to know before you start doing SEM in Stata therefore is: `rename _all, lower.`
- Stata is similar to, but in almost every detail different from SPSS.

Elements of SEM in Stata

- **`sem (effects) (effects) (effects) , var() covar() standardized iterate(k) method(mlmv).`** The presence of the comma is a standard Stata feature: before the comma there is the model, after the comma the options. The period (.) is not part of the statement.

- Effects are specified as **`(x→y) (x z→y).`** See manual. This is all very intuitive.

- Observed variables start in lower case, latent variables start in UPPER case. In SEM literature the difference is conventionally shown by using squares and circles, and/or by latin and γρεεκ symbols.

- I find it very odd that the **`var()`** and **`covar()`** elements are after the comma – but that is the way it is.

- Effects can be fixed at a certain value by using **`@value`**. E.g. **`(x→y@1)`** would fix the effect at 1, **`(x→y@a) (x→z@a)`** would fix the two effect at **a** (i.e. **constrain** the two effects at **a**).

- Observed variables have a unit of measurement, which can be a z-standardized metric. Latent variables do not have a unit of measurement of their own; we need to fix this somehow. Two options are:
    - Fix one of the measurement effects (factor loadings) at 1. We call this the **reference effect.** Notice that this does NOT imply that the latent variable is identical to the observed variable, only that their unit of measurement is the same.
    - Fix the variance of the latent variables at 1. This option is can only be used for ***exogenous*** latent variables (like you would have in a traditional factor analysis).

  In many models, Stata will solve the identification of the unit of measurement of the latent variables on its own, but occasionally it needs your help.
  Traditional SEM wisdom is to choose the strongest indicator as the reference effect. Stata seems to prefer the first one.

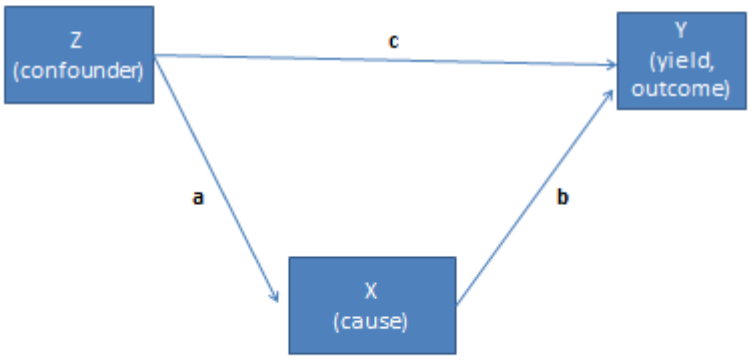
- When you make mistakes or try to estimate an unidentified model, Stata will often go into an infinite amount of iterations (***not converge***). This can be terminated by the **`iterate(k)`** option. You then get a not-converged solution that may lead you to clues about the location of the wrong specifications.

- The **`standardized`** option will show you the coefficients in a Z-standardized metric in which both the observed and latent variables are standardized to unit variance. This is similar to standardized regression and standardized factor analysis, but with some important differences:
    - There are SE's in standardized solutions (this is a unique feature of State, I have never seen it in other SEM programs). The resulting t-tests are different from the ones in the unstandardized solution.
    - The constants (intercepts) are not standardized.

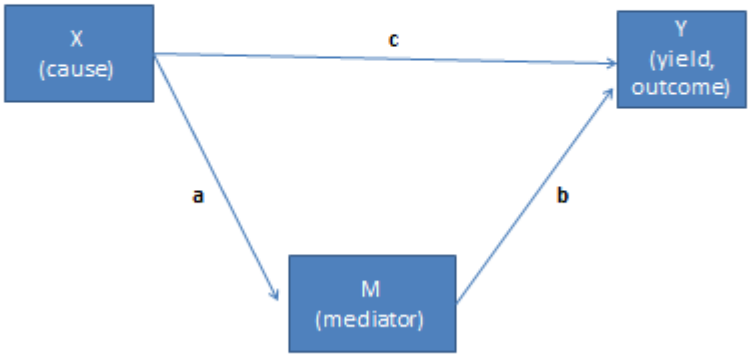
- The **`method(mlmv)`** instructs the program to use all available data to estimate the model, i.e. the correlations calculated on a pairwise basis. This increases the computing time considerably

and occasionally leads to non-convergence. The bonus is that you now have more data available and (usually) obtain smaller SE's (more power). More about missing values treatment is coming up.

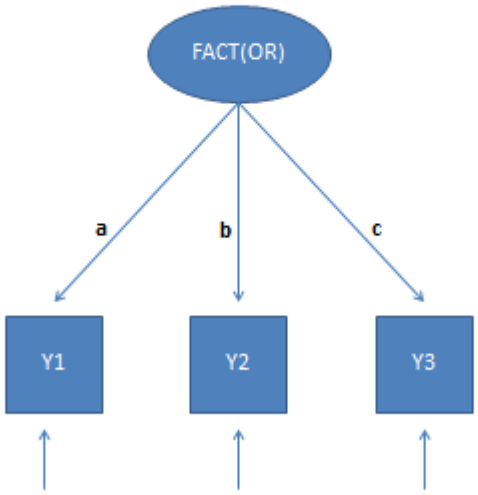
ELEMENTARY CAUSAL MODEL
Interpretation 1: confounding



ELEMENTARY CAUSAL MODEL
Interpretation 2: mediation



ELEMENTARY MEASUREMENT MODEL



```
use "U:\)Teaching\SEM\Melbourne\issp_2009_sem.dta", clear

rename _all, lower

!! elementary status attainment model !!

sum fisei degree isei

pwcorr fisei degre isei, obs
corr fisei degre isei

regr isei fisei
regr isei fisei, beta

regr isei fisei degree
regr isei fisei degree, beta

sem (fisei -> degree isei) (degree -> isei)
sem (fisei -> degree isei) (degree -> isei), standardized

estat teffects
sgmediation isei, mv(degree) iv(fisei)

!! elementary measurement model !!

summ att1 att2 att3
corr att1 att2 att3

sem (ATT -> att1 att2)
sem (ATT -> att1@a) (ATT -> att2@a)

sem (ATT -> att1 att2 att3)
sem (ATT -> att1 att2 att3) , standardized
sem (ATT -> att1 att2 att3) , var(ATT@1)

!! status attainment model with double indicators !!

sem (FOCC -> fisei fosei) (OCC -> isei osei) (FOCC -> OCC),
standardized

sem (FOCC -> fisei fosei) (OCC -> isei osei) (FOCC -> OCC),
standardized

sem (FOCC -> fisei fosei) (OCC -> isei osei) (FOCC -> OCC),
standardized method(mlmv)
```