Introduction to Structural Equation Modeling

Dr. Henrik Kenneth Andersen

2022-10-05

Introduction

Structural equation model (SEM) encompasses a variety of statistical techniques. It is a *unified framework* for empirical investigations using

- linear regression
- mediation analysis
- confirmatory factor analysis

and any combination of them.

In this course, we will look at linear regression and mediation analysis using SEM, confirmatory factor analysis which involves specifying *latent variables* as the cause for the observed responses, and the *general SEM* which combines linear regression and mediation together with confirmatory factor analysis to allow us to specify the relationships of interest between the *underlying latent variables*.

The general SEM gets us closer to the *constructs of interest* (attitudes, personality characteristics, etc.) by separating the *part of the response caused by the latent variable* and the *measurement error*.

Advantages of SEM

Compared to more traditional techniques (OLS, principle component analysis, etc.), SEM has the advantages of being able to

- *easily investigate mediation models* with multiple dependent variables
- account for measurement error in the observed variables and model relations at the *latent construct*-level
- evaluate the fit of *the entire model*, not just in terms of the error of a single dependent variable, and test the hypothesized structure
- *refine models* for future testing by allowing the data to "speak" (modification indices)

As noted, it is also a *unified framework* for just about any type of model. Practically *any model can be run in SEM*, meaning one does not have to switch between approaches/software/packages for different models. Many different types of models, from linear to logistic regression, to mediation, to multilevel, panel and experimental models can be run in `lavaan` or `Mplus`, etc.

Steps for running an SEM analysis

An investigation using SEM entails the following steps:

1. Specification
2. Identification
3. Estimation
4. (Evaluation)
5. (Respecification)
6. Interpretation

Most of these steps take place in a traditional type of model as well, even if they are not usually explicated. Steps 4 and 5, model evaluation and respecification are arguably specific to SEM.

Specification

Specification is the first step to any SEM. In fact, it is the first step in just about any statistical analysis, but the *flexibility of SEM makes it much more involved.*

Specifying a model means answering the following questions:

- what variables to include?
- how are the included variables distributed?
- in what way are the variables related to one another?
- what are the model assumptions?

In a normal OLS regression model, we go through these as well: we decide on the variables we want to investigate, with normally *one dependent variable* and *one or more independent variables.* Then, based on the *distribution of the dependent variable*, we run our OLS model or, in the case of, for example, a binary dependent variable, turn to a different type of model. We determine the *link function*, usually starting with a *linear model* and then *apply the model assumptions.* The most important assumptions being

- random sampling,
- linearity,
- variation in the independent variables, no excessive multicollinearity,
- zero conditional mean of the error,
- homoskedasticity of the error,
- normally distributed error.

We can relax a number of these assumptions and move to more general models (logistic/probit regression for binary outcomes, robust standard errors, etc.) but we will retain them in this workshop for the sake of simplicity.

In SEM, this process is more involved because we can have *multiple dependent variables in a single model.* This means we need to look at the distributions of each of the dependent variables and apply the assumptions to all of the error terms.

The zero conditional mean *assumption applies to all of the errors*, as well, and we need to determine how all of the independent variables are related to one another.

We need define not only how the substantively interesting variables are related to one another, but also how the *observed indicators are related to the underlying latent concepts* and each other.

The following code example shows how involved the specification of an SEM can be (though it can even be more complicated).

```
m1 <- '
# Define latent variables xeno: xenophobia, anom: anomia
  xeno =~ 1*px06 + l2*px07 + l3*px10 + l4*pa09 + l5*pa19
  anom =~ 1*lp03 + g2*lp04 + g3*lp05 + g4*lp06
# Mediation model
  xeno ~ 1 + b1*anomia + b2*classupper + b3*female + b4*yearseduc + b5*east + b6*age
  anom ~ 1 + t1*classupper + t2*female + t3*yearseduc + t4*east + t5*age
# Exogenous correlations
  clussupper ~~ female + yearseduc + east + age
  female ~~ yearseduc + east + age
  yearseduc ~~ east + age
  east ~~ age
# Exogenous variances
  classupper ~~ classupper
```

```
  female ~~ female
  yearseduc ~~ yearseduc
  east ~~ educ
  age ~~ age
# Disturbances or error variances
  xeno ~~ xeno
  anom ~~ anom
# Constrain disturbance correlation to zero
  xeno ~~ 0*anom
'
```

Identification

Identification broadly refers to whether enough data is available for unique estimates of all the model parameters.

```
...
## Model Test User Model:
##
##   Test statistic                               190.014
##   Degrees of freedom                               26
##   P-value (Chi-square)                          0.000
...
```

In a simple OLS regression, identification usually just means one has more observations than regression coefficients. In SEM, it is more complicated for two reasons:

1. SEMs can be complicated mediation models with latent and observed variables. We can specify *more than one regression coefficient per independent variable* and the SEM framework allows for various *other types of relations* that use degrees of freedom that are not possible in traditional analyses (e.g., correlated errors).
2. SEM does not look at individual observations but rather the *mean vector* and *covariance matrix* of the observed variables. We have as much data as there are non-redundant observed means and (co)variances. Therefore, we quantify the empirical information as the *pieces of non-redundant information* in the sample mean vector and covariance matrix. That is the number of observed means and the variances and covariances between the observed variables.

Everything we estimate in the model uses *degrees of freedom*. We will look at an example later and see that in a simple bivariate regression model,

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

we have two observed means, $(\bar{x}, \bar{y})$, two observed variances, (σ_y^2, σ_x^2), and one observed covariance, $\sigma_{x,y}$. We must estimate one unconditional mean, $\hat{\bar{x}}$ and one intercept, $\hat{\beta}_0$, so the *mean structure is saturated* (exactly as many estimated parameters as pieces of empirical information, which is usually the case). We also need to estimate the slope coefficient, $\hat{\beta}_1$, the error variance, $\hat{\sigma}_u^2$, and the unconditional variance of x. We have three pieces of empirical information in the observed covariance matrix so we have $3 - 3 = 0$, i.e., *zero remaining degrees of freedom*.

Thus, or model is *just-identified*, i.e., there are exactly as many pieces of empirical information as there are parameters to estimated. This provides the rare opportunity to derive closed-form solutions for the parameters to be estimated, just like in traditional OLS, which we will see later. Normally, we will strive for *over-identified* models, i.e., where we have more empirical information than parameters to be estimated and the closed-form solution will not be available. This will come up again later when we discuss estimation.

We will look at this in more detail later and see how to use path models to help determine whether a model is identified or not.

Estimation

Once we have specified the model and ensured its identification, we need to *estimate the unknown parameters*. This is normally done using *maximum likelihood* but other estimators are available, like unweighted least squares, generalized least squares, as well as robust variants, etc.

```
m1.fit <- sem(model = m1,
              data = df,
              estimator = "ML")
```

Each of these come with their own benefits and sometimes drawbacks and can be chosen according the the requirements of the model at hand.

Estimation is a very complicated topic and requires a good grasp of algebra, calculus and statistics to fully understand. We can look at it in broad sense in this workshop; those interested can see Ferron and Hess (2007); Bollen (1989) for a much more detailed explanation.

Evaluation

After specifying and estimating the model, the *model fit* is normally evaluated. This step is restricted to over-identified SEMs and it is not something normally available to traditional methods. Namely, with an over-identified model, it is possible to test the model assumptions and restrictions empirically.

Over-identified models are those with positive degrees of freedom (df), where the df is calculated as the number of pieces of unique information in the observed mean vector, $\boldsymbol{o}$, and covariance matrix, $\boldsymbol{S}$, minus the number of parameters to be estimated. Namely, we will introduce *latent variables* as a way to *reduce complexity* and *decompose observed scores* into a *part that is explained by the underlying latent variable* and a *part that is due to measurement error*.

In other words, we write the observed means and covariances as *functions of the model parameters*. Usually, we are primarily interested in the *covariance structure* and we are essentially attempting to boil the observed covariance matrix down to functions of a smaller number of unknown parameters.

Once we have estimated the unknown parameters, we get our *model-implied* mean vector and covariance matrix. We can then compare the model-implied to the sample moments and compute a χ^2 statistic for the amount of discrepancy between the two. If the discrepancy is low, the model-implied moments closely resemble the sample counterparts and the χ^2 statistic is small. Then we say our model *fits the data well*. If the discrepancy is high, then the χ^2 statistic is large and we say the model does not fit the data well.

Notice that the χ^2 statistic gives a measure of fit for the entire model, not just the amount of explained variance in a single dependent variable. Also, it is well known that χ^2 has no upper bound, so interpreting its magnitude can be difficult (though a significance test is available), so a number of *comparative and absolute* fit indices (e.g., CFI, TLI, RMSEA, etc.) have been suggested, which we will discuss in detail later.

```
...
## Model Test User Model:
##
##   Test statistic                               190.014
##   Degrees of freedom                            26
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                              5699.550
##   Degrees of freedom                            36
```

```
##   P-value                                            0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                        0.971
##   Tucker-Lewis Index (TLI)                           0.960
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -25729.719
##   Loglikelihood unrestricted model (H1)      -25634.712
##
##   Akaike (AIC)                                51497.438
##   Bayesian (BIC)                              51609.802
##   Sample-size adjusted Bayesian (BIC)         51549.433
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                              0.048
##   90 Percent confidence interval - lower             0.042
##   90 Percent confidence interval - upper             0.055
##   P-value RMSEA <= 0.05                              0.681
...
```

Respecification

An over-identified model has degrees of freedom to spare, so to speak. Respecification is a step available only in SEM and mainly involves looking at what would happen to our model fit if we loosened specific assumptions and used up those extra degrees of freedom. This information is found in the *modification indices*, a list of parameters that are currently fixed to zero with information about how much the model fit would improve (in terms of χ^2) if we relaxed the assumption and allowed the parameters to be freely estimated.

```
##     lhs op  rhs     mi    epc sepc.lv sepc.all sepc.nox
## 39 px07 ~~ px10 86.255  0.158   0.158    0.192    0.192
## 64 lp04 ~~ lp05 43.590 -0.032  -0.032   -0.251   -0.251
## 23 xeno =~ lp04 41.955  0.090   0.104    0.219    0.219
## 46 px10 ~~ pa09 29.877  0.081   0.081    0.114    0.114
## 62 lp03 ~~ lp05 16.882  0.016   0.016    0.132    0.132
## 22 xeno =~ lp03 13.904 -0.045  -0.052   -0.121   -0.121
## 41 px07 ~~ pa19 12.058  0.076   0.076    0.092    0.092
## 66 lp05 ~~ lp06 11.788  0.014   0.014    0.109    0.109
## 58 pa19 ~~ lp04  9.810 -0.025  -0.025   -0.075   -0.075
## 25 xeno =~ lp06  9.628 -0.040  -0.046   -0.101   -0.101
```

For example, in a typical CFA, it is assumed that the underlying latent variable is the sole cause of the observed covariances between indicators. The measurement error portion of the indicators are assumed independent and thus uncorrelated with one another. Sometimes, however, subsets of indicators may be closely related to one another (because of common wording, or scaling, for example) and there may exist a correlation between indicators above and beyond the common cause of the latent variable. In this case, the modification indices would tell us that the χ^2 statistic would improve if we allowed an *error correlation* to be estimated rather than fixed to zero.

Modification indices are *dangerous*, however, because they are entirely *data-driven*. I.e., typically we should specify our model a priori, before we have seen the data, in a *theory-driven* fashion based on the literature and our hypotheses. The modification indices have no regard for theory, they are dispassionate measures

of partial correlations. They need not make substantive sense, nor do they necessarily represent anything beyond sampling error or peculiarities of the specific model.

Also, it should be clear that chasing model fit is not the goal of the analysis. Rather, we should be focused primarily on capturing ceteris paribus (ideally "causal") relationships between variables, rather than trying to fully explain the entire world. And it should be clear that a *saturated model*, i.e., one with zero degrees of freedom (just as many estimated parameters as pieces of empirical information), tells us nothing about model fit. That is, a model with zero degrees of freedom fits perfectly every time, since without placing any restrictions on the model we can of course fully reconstruct the observed covariance matrix and mean vector. So, *just because we have degrees of freedom to spare, does not mean we should necessarily use them.*

For this reason, any respecification should be theoretically justifiable, based on common sense or a previously neglected aspect of the literature. Ideally, one could split the data beforehand and explore the modification indices with a smaller portion of the data before testing these new implications on the larger set. Or, take the opportunity to investigate plausible modifications to the model first theoretically before collecting new data to test in a confirmatory fashion.

Interpretation

Interpretation can mean different things depending on the type of model one is looking at.

For a simple linear regression model in SEM, we will be interpreting the magnitude and direction of the estimated coefficient, along with its statistical significance.

For a confirmatory factor analysis, we are primarily interested in the psychometric properties, i.e., the validity and reliability of the measurement model. That is, how well do the observed indicators measure the underlying construct? We look at the factor loadings — regressions of the observed indicators on the underlying latent variables — as well as model fit to establish validity and reliability, for example.

In a full SEM, we want to ensure our factor models (CFAs) are valid and reliable in order to be able to safely interpret the regression coefficients between (latent) variables.

Recommended literature

Comprehensive

Bollen, K. (1989). Structural Equations with Latent Variables. New York, NY: John Wiley & Sons.[1]

Gentle introduction

Muthén, B.; Muthén, L.; Asparouhov, T. (2016). Regression And Mediation Analysis Using Mplus. Los Angeles, CA: Muthén & Muthén.[2]

Kline, R. (2016). Principles and Practice of Structural Equation Modeling. Fourth Edition. New York, NY: The Guilford Press.

Hoyle, R. (2015). Handbook of Structural Equation Modeling. New York, NY: The Guilford Press.

SEM in `lavaan`

The `lavaan` project webpage.

Rosseel, Y. (2012). `lavaan`: An `R` Package for Structural Equation Modeling. Journal of Statistical Software, 48 (2).

[1] This is not antiquated, even if it is over 30 years old. It is the best book on SEM, hands down.
[2] Good basic introduction to regression and mediation analysis in SEM, even if you don't use Mplus.

Steinmetz, H. (2015). Lineare Strukturgleichungsmodelle: Eine Einführung mit R. München, Mering: Rainer Hampp Verlag.[3]

Special topics

Hoyle, R. (1995). Structural Equation Modeling: Concepts, Issues, and Applications. Thousand Oaks, CA: Sage Publications.[4]

References

Bollen, Kenneth. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Ferron, John, and Melinda Hess. 2007. "Estimation in SEM: A Concrete Example." *Journal of Educational and Behavioral Statistics* 32 (1): 110–20.

[3] I can't speak to the quality of this book but it is one of the few books on SEM in R using lavaan I know of. Plus it is in German. I would suggest reading the software-agnostic literature above along with the lavaan tutorial website.

[4] This is somewhat outdated, but there are some good easy-to-follow chapters on relevant topics, like nonnormal data.