Structural Equation Modeling with R and lavaan

Day 1: Structural and Measurement Models

Constantin Manuel Bosancianu

WZB Berlin Social Science Center
Institutions and Political Inequality
bosancianu@icloud.com

September 20, 2021

Welcome! It's great to have you in the workshop!

Workshop structure (1)

First two days focus on 2 core components of SEM models:

1. a measurement model (akin to factor analysis)
2. a structural model (path models)

We also cover the intersection of these two: the full structural equation model (structural + measurement).

We go through: (1) specification; (2) identification (*basics*); (3) estimation (*basics*); (4) interpretation; and (5) assessment of fit.

Workshop structure (2)

Building on the Jan. 2021 workshop in MLM, the last day advances to multilevel SEM (MSEM).

The transition will be fast, so we focus only on 1–2 key specifications in ML-SEM (to ensure we can cover them well).

The majority of the lessons from the previous two days hold for ML-SEM specifications, so we will spend more time on the key differences and interpretation.

What gets obscured

Limited time means we have to give up on some topics:

- ✓ the connections between SEM and Judea Pearl's structural causal models approach (Pearl, 1998, 2012)
- ✓ non-recursive SEM (Finch & French, 2015; Kline, 2013)
- ✓ power analysis in SEM (Hancock & French, 2013)

We also won't be able to delve into more complex types of MSEM models, like random slopes MSEM specifications.

Logistics

Lecture based on .pdf. Labs based on .Rdata & .R, and carried out "live" (via Zoom).

Same as before, over the 3 days, time spent on lectures gradually decreases in favor of labs:

- D1: $\approx$ 100 min lecture & 100 min lab
- D2: $\approx$ 80 min lecture & 120 min lab
- D3: $\approx$ 60 min lecture & 140 min lab

Small part of the lab also devoted to questions about lectures.

SEM foundations

Value of SEM

It represents a *very* general statistical framework which can subsume many other approaches: ANOVA, standard regression, factor analysis, multilevel models.

Provides a mechanism to take into account *measurement error* in a set of indicators, which other methods (standard regression) have to assume away.

Can handle in the same specification (1) measurement issues, and (2) the estimation of causal relationships.

Smaller advantages: (1) can model multiple outcomes simultaneously; (2) can accommodate both direct and indirect effects; (3) can produce measures of both local and global fit; (4) can accommodate multiple data configurations.

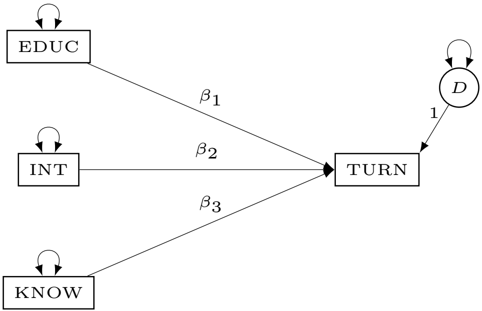
Questions probed with SEM

A marked gain in flexibility when considering the questions it can answer.

	Topic	
Political efficacy	Impact of attentiveness to political news on internal political efficacy	(Semetko & Valkenburg, 1998)
	Investigating the properties (reliability, validity) of a new scale	(Morrell, 2003)
Political trust	The structure of political trust across different educational groups	(Elsas, 2015)
	Association between education and generalized/political trust	(Hooghe, Marien, & de Vroome, 2012)
Justification of inequality	Link between "just-world" belief and justification of inequality	(Beierlein, Werner, Preiser, & Wermuth, 2011)
Political participation	The connection between personality traits and political participation	(Vecchione & Caprara, 2009)

Reasons for using SEMs (I)

Substantive: allows you to test more accurate depictions of social processes.

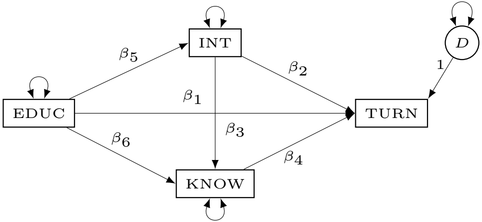


Standard regression

The standard regression approach will assume no measurement error and *independence of predictors*.

Reasons for using SEMs (II)

A SEM allows the specification of a more complex causal structure.



Reasons for using SEMs (III)

Statistical:

- ✓ ability to model multiple variables simultaneously;
- ✓ ability to test for measurement invariance across groups;
- ✓ ability to leverage *latent* variables to account for measurement error.

It's a particularly strong tool for those with theories about the dimensionality of concepts, and who have multi-item measures in their data.

Development of SEMs

Draws on two separate traditions:

- ✓ path modeling, where a system of equations is used to describe functional relations among **observed** variables (Wright, 1918, 1921, 1934)
- ✓ factor analysis, where one tries to explain the (co-)variation in a set of observed measures with a reduced set of **latent** variables (Tucker, 1955)

By the 70s, theoretical work had joined these two into the full SEM framework, combining a measurement model with a structural one (Jöreskog, 1967, 1969, 1973; Jöreskog & Sörbom, 1979).

Foundational concepts in SEM

Observed vs. latent (I)

SEM operates with 2 types of variables:

- ✓ observed (manifest): the measures for which we have collected data, and which are part of the data file
- ✓ latent: the theoretical construct which is captured by the indicators we measure

Example: anxiety. Many measures, but a common one is the State-Trait Anxiety Inventory, asking for agreement with a series of factual questions: “I feel jittery”, “I feel content”, “I am worried”, “I am tense”, “I feel that difficulties are piling up” etc.

Other examples: intelligence, political efficacy, patriotism, homophobia, religiosity.

Observed vs. latent (II)

Observed variables can either be continuous or categorical, but in SEM latent variables are *only* continuous.

Differences in how they are represented in graphs:

$\boxed{O_1}$

Observed: square or rectangle

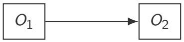
L_1

Latent: circle or ellipse

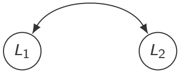
Error terms in the model are also depicted with circles or ellipses, as they are (in a sense) latent constructs.

Co-variation & causality

Two types of relationships are depicted in SEM:



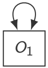
Directional causal effect



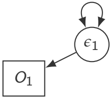
Unanalyzed association

Unanalyzed associations (*covariances*) are estimated by the software, but no theory explains them (we don't know *why* they covary).

A special type of this is unexplained variance in observed or latent indicators.



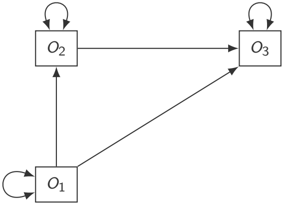
or



Exogenous & Endogenous

SEM no longer uses the DV/IV distinction, as a variable can be both outcome and predictor here. Instead:

- *exogenous*: the causes of which are not included in the model
- *endogenous*: which have at least one cause included in the model



Parameters estimated by the model:

- direct effects on endogenous variables
- variances and covariances of exogenous variables

Parameters: free, constrained, or fixed

SEM offers considerable flexibility in terms of specification; the user can "turn off" or "on" specific parameters in the model.

Parameters:

- *free*: the computer estimates this based on the data
- *constrained*: estimated by the computer within the bounds specified by the user (equality, proportionality)
- *fixed*: the computer accepts the parameter estimate provided by the user, irrespective of how the data looks like

Fixing or constraining parameters is a way to address problems with lack of identification based on theory—if $\beta_2 = 0.6 * \beta_1$, only β_1 needs to be estimated.

Observations & sample size

SEM works by analyzing a variance–covariance matrix of the variables included in the model.

The "ceiling" in terms of model identification is *not* the sample size, but the off-diagonal elements in this matrix (the *observations*).

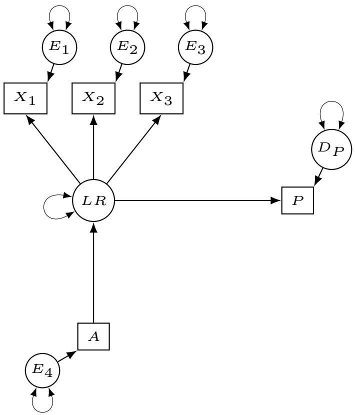
If v is the number of variables in the model, then we have $\frac{v(v+1)}{2}$ observations (q).

$$df_M = p - q \tag{1}$$

Importance of graphical models

Most of the choices discussed above will be presented graphically to your readers (matrix notation is also possible, though complex).

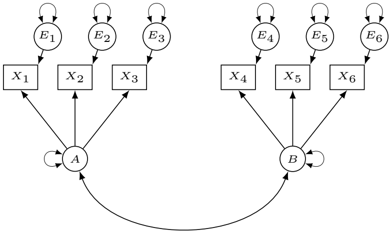
Graphics (1) organize knowledge and (2) represent your hypotheses. They also illustrate what parameters are fixed or free.



Measurement component: *factor analysis*

Factor analysis (I)

A technique to extract latent *factors* from a set of observed *indicators*—goal to explain variation with a reduced set of dimensions.

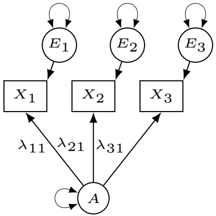


Confirmatory factor analysis (CFA) model

Factor analysis (II)

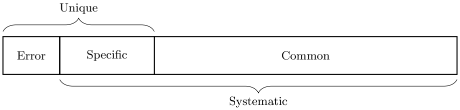
The association between indicators and factor is called a *factor loading*.

The loading represents a regression path from the factor to the corresponding indicator.



The method *decomposes* variance:

- *common* variance between indicators is assumed to be due to the factor
- *unique* variance: composed of specific variance and measurement error



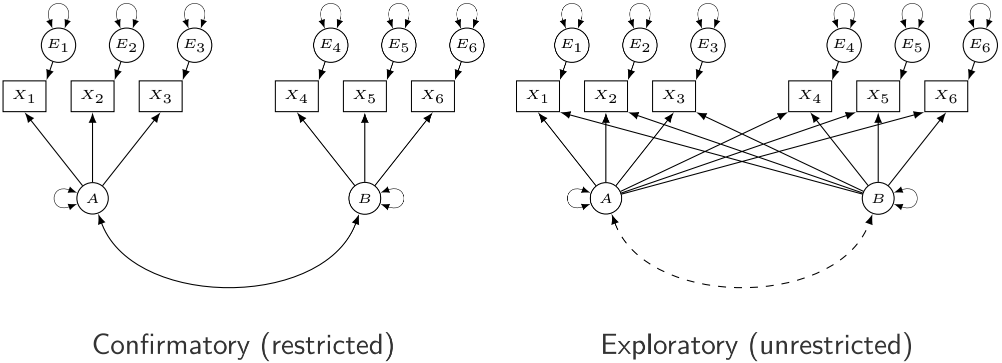
Exploratory vs. confirmatory (I)

Exploratory (EFA)	Confirmatory (CFA)
No need to pre-specify number of factors	Must always pre-specify number of factors
Cannot specify correspondence between indicators and factors	Can pre-specify correspondence between indicators and factors
Not identified by default; require *rotation*	Need to be identified before estimating; no *rotation* phase

Differences: EFA vs. CFA (Kline, 2015)

EFA is used to determine the dimensionality; CFA is used to confirm a theoretically-defined or hypothesized factorial structure.

Exploratory vs. confirmatory (II)



EFA does not require the estimation of factor covariances, though it is possible.

EFA: rotation types

EFA also includes a rotation stage before loadings are provided—it serves to increase interpretability.

It reweights the initial solution, so that each extracted factor explains as much variance as possible in a unique set of indicators (it increases *separation*).

Types:

- ✓ orthogonal: factors are uncorrelated (e.g. *varimax* rotation)
- ✓ oblique: factors will be correlated (e.g. *promax* rotation)

Keep in mind: CFA requires no rotation, as there you're already specifying the number of factors to extract.

Exploratory vs. confirmatory (III)

In practice, the border between EFA and CFA is fuzzy.

It happens that a confirmatory specification doesn't fit data well, and needs adjustment.

In EFA we frequently decide to take only the first k factors based on how much "sense" they make (which is where theory plays a role).

Data interlude (I)

Political efficacy

Let's take a short example: ANES 1987 pilot study. We're interested in people's sense of political efficacy ($N = 351$).

Multiple indicators available:

- A38a: I consider myself well-qualified to participate in politics
- A38b: I have good understanding of important political issues
- A38c: Other people seem to understand complicated political issues easier than I do
- A38e: I often don't feel sure of myself when talking with other people about politics
- ...: ...
- A15c: Most officials can be trusted to do what is right without having to constantly check on them
- A15d: Most officials are truly interested in what the people think
- A15e: Candidates for office are only interested in people's votes, not in their opinions

Observations: covariance matrix

	a38a	a38b	a38c	a38d	a38e	a38f	a15a	a15b	a15c	a15d	a15e	a15f	a15g
a38a	2.05	0.93	0.46	1.12	0.76	0.75	0.08	-0.04	0.05	0.12	0.16	0.09	0.14
a38b	0.93	1.42	0.30	0.74	0.53	0.76	0.02	-0.10	0.16	0.09	0.06	0.05	0.07
a38c	0.46	0.30	1.75	0.57	0.84	0.43	-0.18	0.15	-0.12	0.02	0.24	0.31	0.27
a38d	1.12	0.74	0.57	2.00	0.77	0.75	-0.09	-0.02	-0.14	-0.10	0.03	0.06	-0.04
a38e	0.76	0.53	0.84	0.77	2.00	0.66	-0.11	0.11	-0.16	0.03	0.24	0.23	0.18
a38f	0.75	0.76	0.43	0.75	0.66	1.81	-0.03	-0.03	0.02	0.04	0.14	0.17	0.13
a15a	0.08	0.02	-0.18	-0.09	-0.11	-0.03	1.07	0.15	0.43	0.45	0.14	0.15	0.24
a15b	-0.04	-0.10	0.15	-0.02	0.11	-0.03	0.15	0.79	0.20	0.22	0.34	0.49	0.37
a15c	0.05	0.16	-0.12	-0.14	-0.16	0.02	0.43	0.20	1.51	0.53	0.22	0.40	0.22
a15d	0.12	0.09	0.02	-0.10	0.03	0.04	0.45	0.22	0.53	1.13	0.40	0.36	0.40
a15e	0.16	0.06	0.24	0.03	0.24	0.14	0.14	0.34	0.22	0.40	1.60	0.65	0.61
a15f	0.09	0.05	0.31	0.06	0.23	0.17	0.15	0.49	0.40	0.36	0.65	1.29	0.76
a15g	0.14	0.07	0.27	-0.04	0.18	0.13	0.24	0.37	0.22	0.40	0.61	0.76	1.30

EFA: Factor loadings

	MR2	MR1	MR3	MR4	MR5
a38a	-0.04	0.39	0.12	0.29	0.17
a38b	0.00	0.88	0.00	0.00	-0.01
a38c	0.15	-0.02	-0.11	0.05	0.53
a38d	0.01	0.00	-0.01	1.00	0.00
a38e	-0.04	0.05	0.01	0.02	0.75
a38f	0.06	0.43	-0.04	0.09	0.18
a15a	-0.07	-0.04	0.64	0.06	-0.07
a15b	0.49	-0.20	0.12	0.05	0.07
a15c	0.12	0.14	0.47	-0.02	-0.18
a15d	0.08	0.00	0.68	-0.06	0.09
a15e	0.48	0.00	0.10	-0.04	0.11
a15f	0.90	0.02	-0.04	0.02	-0.04
a15g	0.62	0.03	0.10	-0.07	0.07

Pattern matrix with standardized loadings

Every possible factor loading (*pattern coefficient*) is estimated in this setup.

EFA: continuation

Some software takes a *stepwise* approach, and sequentially tests the fit with the data for a solution with 1, 2, 3, . . . factors.

It's up to the user, then, to select which of the solutions make more sense based on theory.

Here, though, the full solution is given, leaving perhaps the confirmatory step for another data set (or the second split half of data).

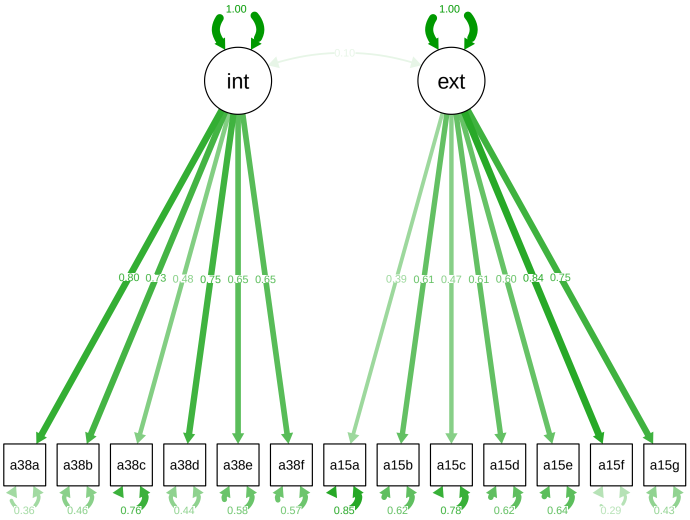
CFA: theory testing

If we have a pretty good idea of which factors we can extract, based on theory, we can go directly to *confirmatory* factor analysis.

Internal	External
Believe well-qualified to participate	Politicians well-qualified
Good understanding of political issues	Politicians not as honest as voters deserve
Others comprehend politics easier than I do	Public officials can be trusted to do right
I could do a good job in office	Public officials truly interested in what people think
Not sure of myself when talking politics	Candidates interested only in votes, not opinions
I'm equally well-informed about politics as most	Many politicians think they are masters, not servants
	Elected officials lose touch pretty quickly

Internal vs. external political efficacy

CFA: two factors



Deciding among solutions

How to choose whether a 1-factor or 2-factor solution is better? Is it all up to the user?

Not entirely: can use *model fit criteria* to decide among solutions.

We cover these in tomorrow's "full" SEM session.

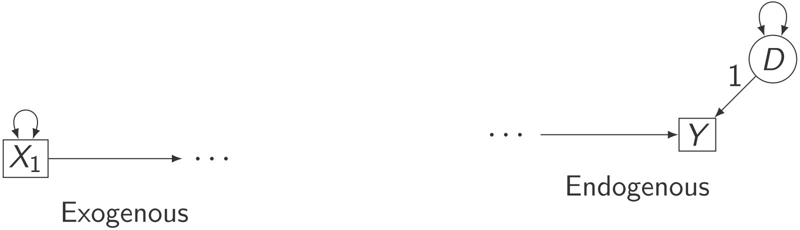
	χ^2 scaled	DF scaled	*p* scaled	RMSEA scaled	CFI scaled	TLI scaled	SRMR
2F	344.11	64	.000	.112	.883	.858	.093
1F	1254.27	65	.000	.229	.504	.404	.200

Structural component: *path analysis*

Basic features

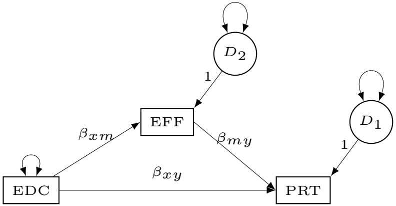
A system of equations *using only observed variables* that together specify a complex set of causal linkages.

This system has at least 2 *endogenous* variables and at least 1 *exogenous* one.



Every endogenous variable has a disturbance (denoted here with D)—unexplained variation.

Pared down example



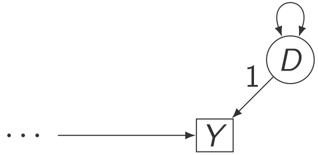
2 endogenous variables: *EFF* and *PRT*; 1 exogenous one: *EDU*.

Needn't use SEM to get estimates of βs; could use the Baron & Kenny approach (`http://davidakenny.net/cm/mediate.htm`).

For more complex configurations, though, a SEM approach will save considerable time.

Latent scales (I)

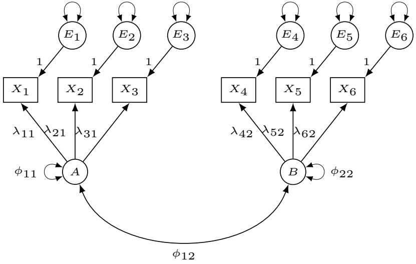
Just as in the case of a factor, disturbances are also considered latent variables (have to be estimated).



The 1 is a *scaling constant*; needed because latent variables have no inherent scale, so one needs to be imposed before estimating things about D.

This was also the case for EFA/CFA earlier, though it wasn't depicted yet on plots for the sake of simplicity.

Latent scales (II)

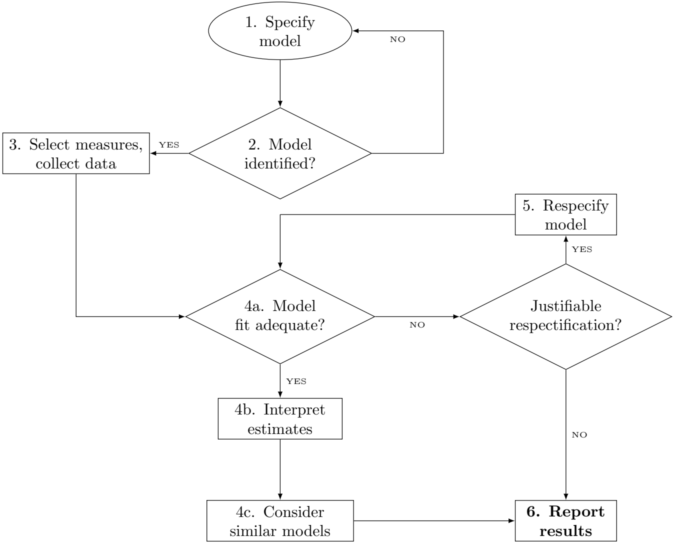


The scale of Es has to be fixed to 1. What about A and B?

Two strategies:

- ✓ set constraint on a unstandardized factor loading, e.g. $\lambda_{11} = \lambda_{42} = 1$ (what we did for Es; also called a **unit loading identification** (ULI) **constraint**)
- ✓ set constraint on their variances, e.g. $\phi_{11} = \phi_{22} = 1$ (also called a **unit variance identification** (UVI) **constraint**)

Stages in the analysis (Kline, 2015)



Identification of path models (I)

To very general rules:

- ✓ $df_M > 0$ (*counting rule*)
- ✓ Every disturbance must be assigned a scale (see UVI above)

If $df_M <= 0$, the model is **underidentified**.

Keep in mind: what matters is *not* sample size, but number of observations in the variance–covariance matrix.

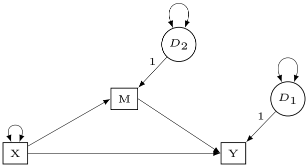
$a + b = 6$ is insufficient, but the following works (**just-identified**):

$$\begin{cases} a + b = 6 \\ 2a + b = 10 \end{cases} \tag{2}$$

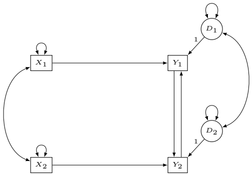
Identification of path models (II)

Two broad categories:

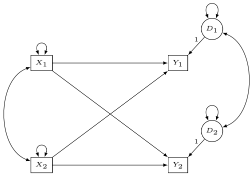
- **recursive**: (1) disturbances are uncorrelated; (2) causal effects are unidirectional
- **non-recursive**: either (1) or (2) (or both) no longer apply



Recursive



Non-recursive



(Partially) recursive

Identification of path models (III)

Good news: recursive path models are identified.

Much harder to say for non-recursive path models—see Kline (2015, pp.152–155) for a set of rules.

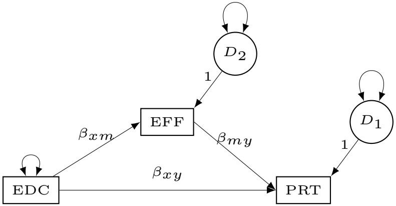
Strategies for non-identified non-recursive models:

1. Drop paths (increases df_M)
2. Enforce parameter constraints, e.g. equality on bi-directional pathways
3. Add instruments for each endogenous variable in a non-recursive link

Mediation (I)

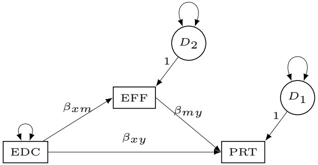
If you remember from our January workshop, *moderation* simply refers to *effect heterogeneity*: β_{XY} varies depending on a third variable.

Mediation is a causal concept, with deeper implications: effect *passes through* another variable on its way to the outcome.



Mediation (II)

Main challenge: without longitudinal data, or an experimental design, it's hard to show the $X \rightarrow M \rightarrow Y$ sequence.



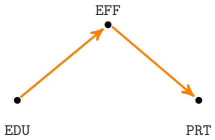
Direct effect: β_{xy}

Indirect effect: $\beta_{xm} * \beta_{my}$

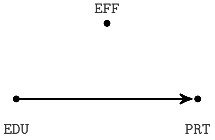
Total effect = direct + indirect

Configurations

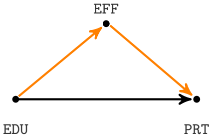
If $\beta_{xy} \approx 0$ and $\beta_{xm}\beta_{my} \neq 0$: full mediation.



If $\beta_{xy} \neq 0$ and $\beta_{xm}\beta_{my} \approx 0$: no mediation.



If $\beta_{xy} \neq 0$ and $\beta_{xm}\beta_{my} \neq 0$: partial mediation.



Data interlude (II)

JOBS II program

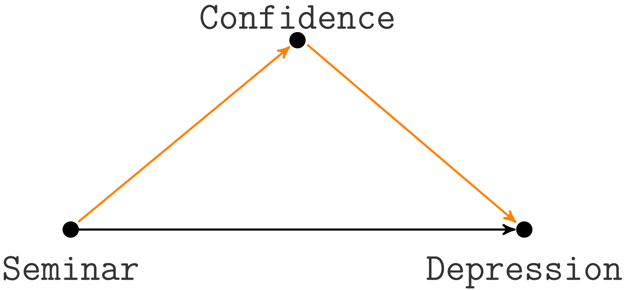
The JOBS II was a field experiment in Southeast Michigan in early 1990s.

Treatment was a job-search skills seminar administered to unemployed persons. Control was a booklet with search tips.

Expected effects:

- ✓ increased employment
- ✓ increased self-confidence $\Rightarrow$ decreased depression

JOBS II program



How much of the effect of the program on depression is *transmitted* through improved self-confidence?

JOBS II data

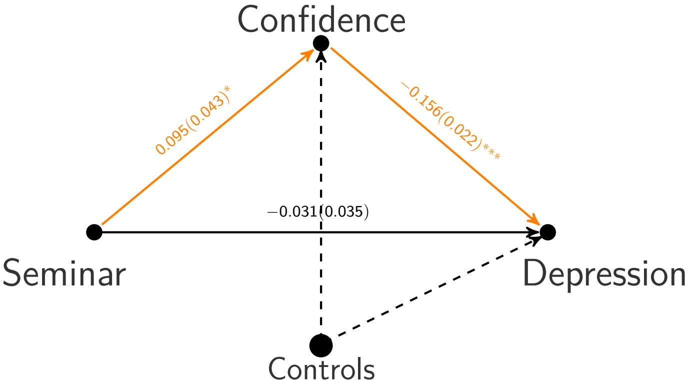
Outcome: continuous, score on Hopkins Symptom Checklist, measured 6 months after treatment.

Mediator: job search self-efficacy, measured 2 months after treatment.

Covariates: age, gender, ethnic minority, depression measured during treatment, economic hardship. Step-by-step approach (pre-SEM):

$$
\begin{aligned}
Depression_i &= \alpha_1 + \beta_1 \overbrace{Seminar_i}^{treatment} + \zeta_1 X_i + \epsilon_{i1} && (3)\\
Confidence_i &= \alpha_2 + \beta_2 Seminar_i + \zeta_2 X_i + \epsilon_{i2} && (4)\\
Depression_i &= \alpha_3 + \gamma Seminar_i + \beta_3 \underbrace{Confidence_i}_{mediator} + \zeta_3 X_i + \epsilon_{i3} && (5)
\end{aligned}
$$

Results



Standard path analysis through the sem() function in lavaan

Interpreting estimates

Coefficients interpreted as for a standard regression model: β quantifies change in Y resulting from 1-unit change in X.

For standardized coefficients, the interpretation changes to standard deviation units.

Indirect effect has similar interpretation: change in Y from 1-unit change in X transmitted through changes in the mediator.

Computing effects

	Direct	Indirect
β	-0.031	-0.015^{*}
SE	(0.035)	(0.007)

Total effect: $\beta_{total} = -0.046(0.035)$.

A case of pure mediation: a negative indirect effect of seminar attendance on depression (via confidence).

Thank you for the kind attention!

References I

Beierlein, C., Werner, C. S., Preiser, S., & Wermuth, S. (2011). Are Just-World Beliefs Compatible with Justifying Inequality? Collective Political Efficacy as a Moderator. *Social Justice Research*, *24*(3), 278.

Elsas, E. (2015). Political Trust as a Rational Attitude: A Comparison of the Nature of Political Trust across Different Levels of Education. *Political Studies*, *63*(5), 1158–1178.

Finch, W. H., & French, B. F. (2015). Modeling of Nonrecursive Structural Equation Models With Categorical Indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 416–428.

Hancock, G. R., & French, B. F. (2013). Power Analysis in Structural Equation Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 117–159). Charlotte, NC: Information Age Publishing.

Hooghe, M., Marien, S., & de Vroome, T. (2012). The Cognitive Basis of Trust: The Relation between Education, Cognitive Ability, and Generalized and Political Trust. *Intelligence*, *40*(6), 604–613.

Jöreskog, K. G. (1967). Some Contributions to Maximum Likelihood Factor Analysis. *Psychometrika*, *32*(4), 443–482.

References II

Jöreskog, K. G. (1969). A General Approach to Confirmatory Maximum Likelihood Factor Analysis. *Psychometrika*, *34*(2), 183–202.

Jöreskog, K. G. (1973). A General Method for Estimating a Linear Structural Equation System. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 83–112). New York: Seminar Press.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Cambridge, MA: Abt Books.

Kline, R. B. (2013). Reverse Arrow Dynamics: Feedback Loops and Formative Measurement. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 41–80). Charlotte, NC: Information Age Publishing.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York: The Guilford Press.

Morrell, M. E. (2003). Survey and Experimental Evidence for a Reliable and Valid Measure of Internal Political Efficacy. *The Public Opinion Quarterly*, *67*(4), 589–602.

Pearl, J. (1998). Graphs, Causality, and Structural Equation Models. *Sociological Methods & Research*, *27*(2), 226–284.

References III

Pearl, J. (2012). The Causal Foundations of Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York: Guilford Press.

Semetko, H. A., & Valkenburg, P. M. (1998). The Impact of Attentiveness on Political Efficacy: Evidence from a Three-Year German Panel Study. *International Journal of Public Opinion Research*, *10*(3), 195–210.

Tucker, L. R. (1955). The Objective Definition of Simple Structure in Linear Factor Analysis. *Psychometrika*, *20*(3), 209–225.

Vecchione, M., & Caprara, G. V. (2009). Personality Determinants of Political Participation: The Contribution of Traits and Self-Efficacy Beliefs. *Personality and Individual Differences*, *46*(4), 487–492.

Wright, S. (1918). On the Nature of Size Factors. *Genetics*, *3*(4), 367–374.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*, *20*, 557–595.

Wright, S. (1934). The Method of Path Coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215.