Structural Equation Modelling

Mathieu Emily

12 mai 2021

State of The R in Rennes

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)
- Medicine and Genomics
  - Structural equation models for **pathway identification** (Xiong, 2001)
  - Application of Structural Equation Models to **GWAS** (Kim *et al.*, 2010)
  - The mediating effects of **public genomic knowledge** in precision medicine implementation: A structural equation model approach (Mogaka and Chimbari, 2020)
  - Bayesian structural equation modeling in **multiple omics data** (Maity, 2020)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)
- Medicine and Genomics
  - Structural equation models for **pathway identification** (Xiong, 2001)
  - Application of Structural Equation Models to **GWAS** (Kim *et al.*, 2010)
  - The mediating effects of **public genomic knowledge** in precision medicine implementation: A structural equation model approach (Mogaka and Chimbari, 2020)
  - Bayesian structural equation modeling in **multiple omics data** (Maity, 2020)
  - A comparison of methods for **inferring causal relationships between genotype and phenotype** using additional biological measurements (Ainsworth *et al.*, 2017)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)
- Medicine and Genomics
  - Structural equation models for **pathway identification** (Xiong, 2001)
  - Application of Structural Equation Models to **GWAS** (Kim *et al.*, 2010)
  - The mediating effects of **public genomic knowledge** in precision medicine implementation: A structural equation model approach (Mogaka and Chimbari, 2020)
  - Bayesian structural equation modeling in **multiple omics data** (Maity, 2020)
  - A comparison of methods for **inferring causal relationships between genotype and phenotype** using additional biological measurements (Ainsworth *et al.*, 2017)

B. Shipley, *Cause and correlation in Biology*, 2016

- SEM is a tool for modeling a **global system**
- SEM is one of the most **popular tool for investigating causality**

Outline

1. From Linear model to Path model
2. Model
3. Causality
4. Ending words

Outline

1. From Linear model to Path model
2. Model
3. Causality
4. Ending words

Outline

1 From Linear model to Path model
Modeling
First step with R

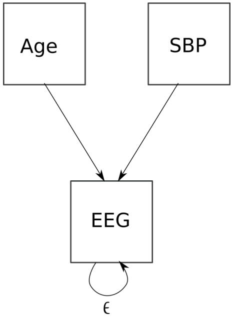
Introductive example : Electroencephalography for Alzheimer's patients
Multiple linear regression

- Three variables: z-scores for brain rate in the frontal region (=EEG), Age and Systolic Blood Pressure (SBP)
- Linear regression
  - $EEG = \beta_0 + \beta_1 Age + \beta_2 SBP + \varepsilon$
  - Coefficients (β_0, β_1 and β_2) are estimated by minimizing the residual variance $\sum(EEG - EEG_{Mod})^2$
- From a **system** point-of-view
  - Age and SBP values are determined outside the model and are imposed on the model (=**Exogeneous** variables)
  - EEG values are determined by the model (**=Endogeneous** variable)

Introductive example : Electroencephalography for Alzheimer's patients DAG visualisation

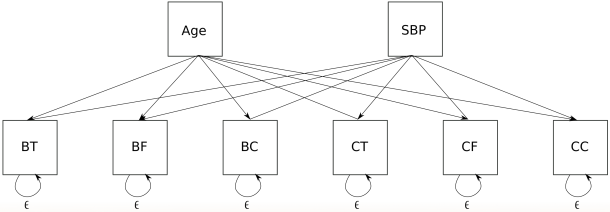
- Visualisation using a **Directed Acyclic Graph (DAG)**

$$EEG = \beta_0 + \beta_1 Age + \beta_2 SBP + \varepsilon$$



Introductive example : Electroencephalography for Alzheimer's patients
Multivariate regression

- **6 measures for EEG**: 3 regions (frontal, temporal, central) and 2 features (brain rate, complexity)
- **Multivariate** regression ($\sim$ Manova)
  - Basics for the estimation: minimizing the distance between the observed covariance for "response" variables and the model covariance
- DAG for a multivariate regression model

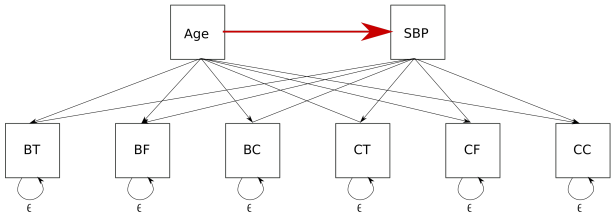


Introductive example : Electroencephalography for Alzheimer's patients Path modeling (1)

- *"An increase in (systolic)* **blood pressure** *has always been taken as an inevitable consequence of* **ageing**" (Pinto, 2007)
- How can we modify the modeling of the system?

Introductive example : Electroencephalography for Alzheimer's patients Path modeling (1)

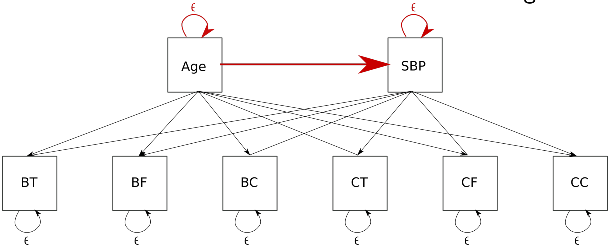
- *"An increase in (systolic)* **blood pressure** *has always been taken as an inevitable consequence of* **ageing**" (Pinto, 2007)
- How can we modify the modeling of the system?



- SBP is now an endogeneous variable

Introductive example : Electroencephalography for Alzheimer's patients Path modeling (2)

- **Measurement error** is also accounted for SBP and Age



Paradigm shift

- In **path modeling**, all observed variables in the system are considered in the estimation of the model
- The aim is to model the covariance matrix

Outline

1. From Linear model to Path model
   - Modeling
   - First step with R

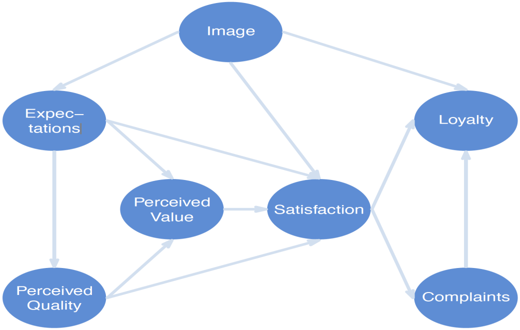
Illustrative example: European Customer Satisfaction Index (ECSI)

- The ECSI model is designed to measure the cause-effect relationships from the antecedents of Customer Satisfaction to its consequences.

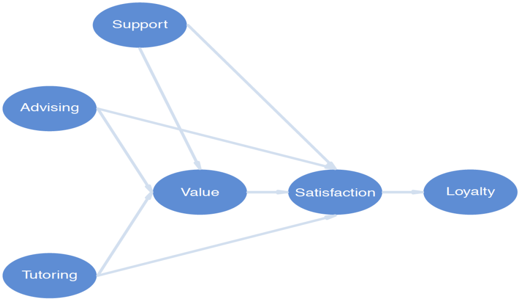
Overview of constructs in the ECSI model:

- **Image** refers to the brand name and the kind of associations customers get from the product/brand/company. It is expected that image will have a positive effect on customer satisfaction and loyalty. In addition, image is also expected to have a direct effect on expectations.
- **Expectations** is the information based on, not actual consumption experienced but, accumulated information about quality from outside sources, such as advertising, word of mouth, and general media.
- **Perceived Quality** comprises product quality (hardware) and service quality (software/humanware). Perceived product quality is the evaluation of recent consumption experience of products. Perceived service quality is the evaluation of recent consumption experience of associated services like customer service, conditions of product display, range of services and products, etc. Perceived quality is expected to affect satisfaction.
- **Perceived Value** is the perceived level of product quality relative to the price paid of the "value for the money" aspect of the customer experience.
- **Satisfaction** is defined as an overall evaluation of a firm's post-purchase performance or utilization of a service.
- **Complaints** implies the complaints of customers
- **Loyalty** refers to the intention repurchase and price tolerance of customers. It is the ultimate dependent variable in the model and it is expected that the better image and higher customer satisfaction should increase customer loyalty.

Illustrative example: ECSI (Sanchez, 2013)



Application of the model: Satisfaction in Education
A simplified version of the ECSI



- Packages `dagitty`, `lavaan` and `semPlot`

Coding break 1: example of a confirmatory analysis

Outline

1 From Linear model to Path model

2 Model

3 Causality

4 Ending words

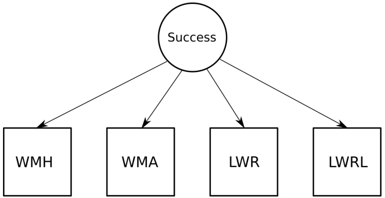
Latent variables by an example

- How to define a strategy of **success**?
- Data obtained from all teams in an entire season.

Variable	Description
GSH	total number of goals scored at home
GSA	total number of goals scored away
SSH	percentage of matches with scores goals at home
SSA	percentage of matches with scores goals away
GCH	total number of goals conceded at home
GCA	total number of goals conceded away
CSH	percentage of matches with no conceded goals at home
CSA	percentage of matches with no conceded goals away
WMH	total number of won matches at home
WMA	total number of won matches away
LWR	longest run of won matches
LRWL	longest run of matches without losing
YC	total number of yellow cards
RC	total number of red cards

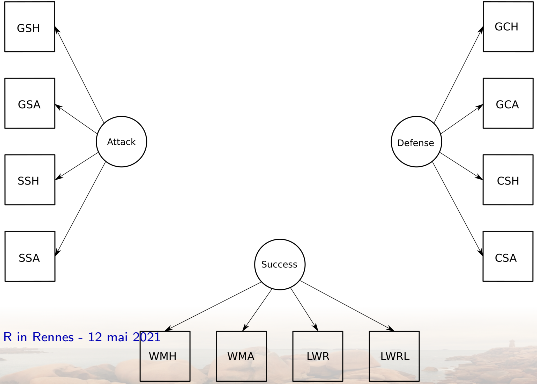
Football example
The concept of Success

- Success is easy to observe/measure but understanding how to achieve success is more complicated
  - Attack strategy
  - Defense strategy
  - Adapt to the opponent
- 4 variables are related to **concept the success**: WMH, WMA, LWR and LRWL



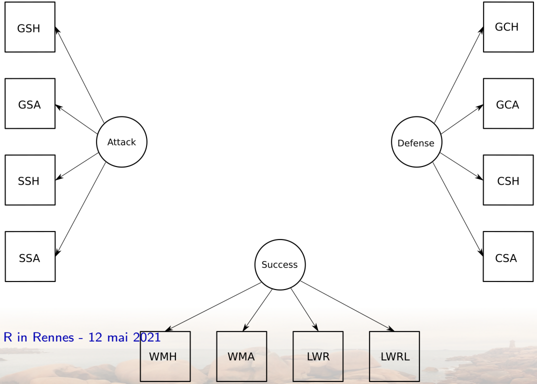
Football example
Latent modeling

- Similarly, the **concepts of Attack** and **Defense** can be modeled as:
  - Attack: GSH, GSA, SSH and SSA
  - Defense: GCH, GCA, CSH and CSA



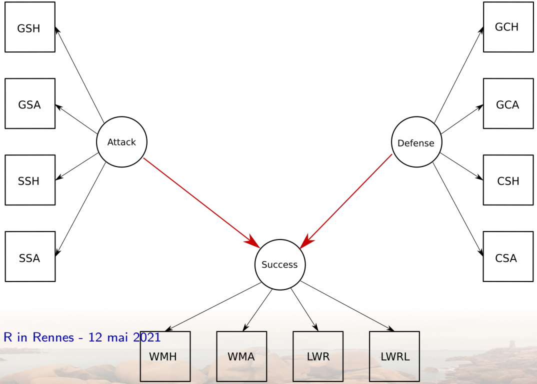
Football example
Latent modeling

- Similarly, the **concepts of Attack** and **Defense** can be modeled as:
  - Attack: GSH, GSA, SSH and SSA
  - Defense: GCH, GCA, CSH and CSA
- How to **link observed** and/or **latent** variables?



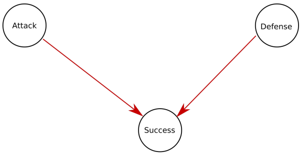
Football example
Latent modeling

- Similarly, the **concepts of Attack** and **Defense** can be modeled as:
  - Attack: GSH, GSA, SSH and SSA
  - Defense: GCH, GCA, CSH and CSA
- How to **link observed** and/or **latent** variables?

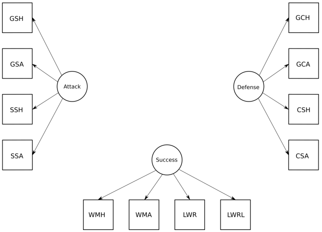


Structural model

- A structural model is made by 2 models:



Latent model



Measurement model

- Each arrow is a **linear** link between variables:
  - $Success = f(Attack, Defense) = \beta_1 Attack + \beta_2 Defense + \varepsilon$
  - $GSH = f(Attack) = \gamma_1 Attack + \varepsilon$
  - ...
- Remark: Success is an endogeneous latent variable while Attack and Defense are two exogeneous latent variables.

Outline

2 Model

General definition

Identification rules

Estimation and tests

Latent model

- Let consider a model with m endogeneous latent variables and n exogeneous variables

$$\eta = \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta$$

  - $\mathbf{B}$ is a $m \times m$ matrix of coefficients for latent endogeneous variables
  - $\mathbf{\Gamma}$ is a $m \times n$ matrix of coefficients for latent exogeneous variables
  - $\mathbf{\Phi} = \mathbb{E}[\xi\xi']$ is a $n \times n$ covariance matrix for ξ
  - $\mathbf{\Psi} = \mathbb{E}[\zeta\zeta']$ is a $m \times m$ covariance matrix for ζ
- Assumptions:
  - $\mathbb{E}[\eta] = 0$
  - $\mathbb{E}[\xi] = 0$
  - $\mathbb{E}[\zeta] = 0$
  - $Cov(\zeta, \xi) = 0$
  - $(I - B)$ nonsingular

Measurement model

- Let consider a model with p endogeneous observed variables and q exogeneous observed variables

$$\mathbf{x} = \mathbf{\Lambda_x}\xi + \delta$$

$$\mathbf{y} = \mathbf{\Lambda_y}\eta + \varepsilon$$

  - $\mathbf{\Lambda_x}$ is a $q \times n$ matrix of coefficients relating x to ξ
  - $\mathbf{\Lambda_y}$ is a $p \times m$ matrix of coefficients relating y to η
  - $\Theta_\delta = \mathbb{E}[\delta\delta']$ is a $q \times q$ covariance matrix for δ
  - $\Theta_\varepsilon = \mathbb{E}[\varepsilon\varepsilon']$ is a $p \times p$ covariance matrix for ε
- Assumptions:
  - $\mathbb{E}[\delta] = 0$
  - $\mathbb{E}[\varepsilon] = 0$
  - $Cov(\delta, \varepsilon) = 0$
  - $Cov(\delta, \zeta) = 0$ and $Cov(\delta, \xi) = 0$
  - $Cov(\varepsilon, \zeta) = 0$ and $Cov(\varepsilon, \xi) = 0$

Toy example of prostate cancer

Observed variables:

- Gleason score from biopsy
- PSA test from a blood sample
- HPC1 (hereditary prostate cancer 1) expression
- PcaP (predisposing for prostate cancer) expression
- PG1 (prostate cancer susceptibility gene 1) expression
- BMI
- Exposure to pollution
- Age

Toy example of prostate cancer

Observed variables:

- Gleason score from biopsy
- PSA test from a blood sample

Cancer measures

- HPC1 expression
- PcaP expression
- PG1 expression

Genetic measures

- BMI
- Exposure to pollution
- Age

Environnemental measures

Toy example of prostate cancer

$$\mathbf{B} = \begin{bmatrix} 0 \end{bmatrix}$$

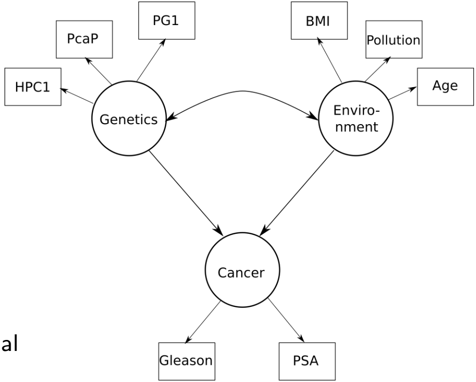
$$\mathbf{\Gamma} = \begin{bmatrix} \beta_{11} \\ \beta_{21} \end{bmatrix}$$

$$\mathbf{\Lambda_x} = \begin{bmatrix} \lambda^x_{11} & 0 \\ \lambda^x_{21} & 0 \\ \lambda^x_{31} & 0 \\ 0 & \lambda^x_{12} \\ 0 & \lambda^x_{22} \\ 0 & \lambda^x_{32} \end{bmatrix}$$

$$\mathbf{\Lambda_y} = \begin{bmatrix} \lambda^y_{11} \\ \lambda^y_{21} \end{bmatrix}$$

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

Ψ, Θ_δ and Θ_ε are diagonal



Covariance implied by the model

- Examples

$$
\begin{aligned}
Cov(HPC1, PSA) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^y_{21} Cancer + \varepsilon_2) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, Cancer) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, \beta_{11} Genetics + \beta_{21} Environ. + \zeta_1) \\
&= \lambda^x_{11}\lambda^y_{21}\beta_{11}\phi_{11} + \lambda^x_{11}\lambda^y_{21}\beta_{21}\phi_{12}
\end{aligned}
$$

$$
\begin{aligned}
Cov(HPC1, PG1) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^x_{31} Genetics + \delta_{31}) \\
&= \lambda^x_{11}\lambda^x_{31}\phi_{11}
\end{aligned}
$$

Covariance implied by the model

- Examples

$$
\begin{aligned}
Cov(HPC1, PSA) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^y_{21} Cancer + \varepsilon_2) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, Cancer) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, \beta_{11} Genetics + \beta_{21} Environ. + \zeta_1) \\
&= \lambda^x_{11}\lambda^y_{21}\beta_{11}\phi_{11} + \lambda^x_{11}\lambda^y_{21}\beta_{21}\phi_{12}
\end{aligned}
$$

$$
\begin{aligned}
Cov(HPC1, PG1) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^x_{31} Genetics + \delta_{31}) \\
&= \lambda^x_{11}\lambda^x_{31}\phi_{11}
\end{aligned}
$$

- Similarly, all covariances can be obtained thus leading to the **implied covariance** $\mathbf{\Sigma}(\theta)$ where θ is the set of unknown parameters of the model

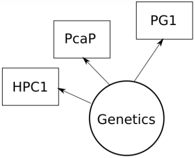
Estimation principle

- Choosing θ for $\mathbf{\Sigma}(\theta)$ to be as close to S as possible

Outline

2 Model
General definition
Identification rules
Estimation and tests

Issue with identification



- θ is **identified** if $\nexists$ θ_1 and θ_2 such as $\mathbf{\Sigma}(\theta_1) = \mathbf{\Sigma}(\theta_2)$
- Example:

	HPC1	PcaP	PG1
HPC1	$(\lambda_{11}^x)^2\phi_{11} + \Theta_{11}^\delta$		
PcaP	$\lambda_{11}^x\lambda_{21}^x\phi_{11}$	$(\lambda_{21}^x)^2\phi_{11} + \Theta_{22}^\delta$	
PG1	$\lambda_{11}^x\lambda_{31}^x\phi_{11}$	$\lambda_{21}^x\lambda_{31}^x\phi_{11}$	$(\lambda_{31}^x)^2\phi_{11} + \Theta_{33}^\delta$

- 7 parameters for only 6 observations: a **need for constraint**
  - Set the variance of the latent variable to 1 ($\phi_{11} = 1$)
  - Set $\lambda_{11}^x = 1$ to scale the *Genetics* to *HPC*1
  - Set $\lambda_{11}^x = \lambda_{21}^x = \lambda_{31}^x$ to balance the amount of variance/covariance in the latent space ($\tau-$equivalence)

Conditions for identification (Bollen, 1989)

- **The** *$t - rule$*

$$t \leq \frac{(p+q)(p+q+1)}{2}$$

where t is the number of free parameters in θ

  - A necessary but not sufficient condition ($t = 19$ in the general prostate model with $p + q = 8$ observed variables)

- **Two-Step rules**
  - Step 1 : Consider y and η as exogeneous variables (CFA)
    - Three-indicator rule
    - **Two-indicator rule**
  - Step 2 : Consider the identification as the latent model (as a measurement model)
  - A sufficient condition

- **MIMIC rule** (for Multiple Indicators and MultIple Causes model)

Outline

2 Model

General definition

Identification rules

Estimation and tests

Estimation

The **closeness** of $\mathbf{\Sigma}(\theta)$ to S is measured by fitting functions $F(S, \mathbf{\Sigma}(\theta))$ (with $F \geq 0$ and $F = 0$ iif $\mathbf{\Sigma}(\theta) = S$)

Estimation

The **closeness** of $\mathbf{\Sigma}(\theta)$ to S is measured by fitting functions $F(S, \mathbf{\Sigma}(\theta))$ (with $F \geq 0$ and $F = 0$ iif $\mathbf{\Sigma}(\theta) = S$)

- **ML (Maximum Likelihood)**

$$F_{ML} = log|\mathbf{\Sigma}(\theta)| + tr(S\mathbf{\Sigma}^{-1}(\theta)) - \log|S| - (p+q)$$

  - Asymptotically unbiased
  - Consistent
  - Asymptotically efficient
  - Scale freeness
  - Availibity of a Confidence Interval

Estimation

The **closeness** of $\mathbf{\Sigma}(\theta)$ to S is measured by fitting functions $F(S, \mathbf{\Sigma}(\theta))$ (with $F \geq 0$ and $F = 0$ iif $\mathbf{\Sigma}(\theta) = S$)

- **ML (Maximum Likelihood)**

$$F_{ML} = log|\mathbf{\Sigma}(\theta)| + tr(S\mathbf{\Sigma}^{-1}(\theta)) - \log|S| - (p+q)$$

  - Asymptotically unbiased
  - Consistent
  - Asymptotically efficient
  - Scale freeness
  - Availibity of a Confidence Interval

- **ULS (Unweighted Least Squares)**

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

- **GLS (Generalized Least Squares)**

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

Global Fit Measures

- Principle: **comparaison with the saturated model**
  - $\mathcal{M}_s$: Saturated model: no latent variable and one parameter for each variance/covariance for manifest variables
  - $\mathcal{D} = -2(\ell(\mathcal{M}) - \ell(\mathcal{M}_s)) \sim_{\mathcal{H}_0} \chi^2(df)$
  - $p = 0.010$: the model is rejected

- **Other measures** are proposed but *"their purpose is to determine the degree to which the rejected model is approximately correct"* (Shipley, 2016):
  - RMSEA (Root Mean Square Error of Approximation)
  - CFI (Bentler's comparative fit index)

Interpretation

The proposed model is rejected: game over?

- Yes in Confirmatory Factor Analysis (**CFA**)
  - The model is not confirmed by observed data
- No in Explanatory Factor Analysis (**EFA**)
  - How can we propose a more likely model?

Interpretation

The proposed model is rejected: game over?

- Yes in Confirmatory Factor Analysis (**CFA**)
  - The model is not confirmed by observed data
- No in Explanatory Factor Analysis (**EFA**)
  - How can we propose a more likely model?

Coding break 2: example of an explanatory analysis

SEM and Explanatory Factor Analysis
To sum up

- Relaxing constraints (=freeing parameters)
  - Modification index
- Adding constraints
  - Remove manifest variables
  - Residual checking
  - Coefficients testing (depends on constraints)
  - Regularized SEM (package `regsem`)
- Model comparison using AIC, BIC and other criteria

Outline

1. From Linear model to Path model
2. Model
3. Causality
4. Ending words

Eight myths about causality and SEM (Bollen and Pearl, 2013)

- Although SEM aims at incorporating causal assumptions, their ability to infer causality is still a matter of debate

Eight myths about causality and SEM (Bollen and Pearl, 2013)

- Although SEM aims at incorporating causal assumptions, their ability to infer causality is still a matter of debate
- Here 8 myths :
    1. SEMs aim to establish causal relations from associations alone
    2. SEMs and regression are essentially equivalent
    3. No causation without manipulation
    4. SEMs are not equipped to handle nonlinear causal relationships
    5. A potential outcome framework is more principled than SEMs
    6. SEMs are not applicable to experiments with randomized treatments
    7. Mediation analysis in SEMs is inherently non causal
    8. SEMs do not test any major part of the theory against the data.

Myth $\sharp 1$: SEMs aim to establish causal relations from associations alone

- Inputs of SEM:
  - Qualitative causal assumptions
  - Empirical data
- Outputs of SEM
  - Failure to fit the data
    - Doubt on causal assumptions (*e.g.* zero coefficients or zero covariance)
    - Guides to repair structural misspecifications
  - Fitting the data
    - Not a proof of causal assumptions...but it makes more plausible

"Positive results need to be replicated and to withstand the criticisms of researchers who suggest other models for the same data"

Causality and DAG

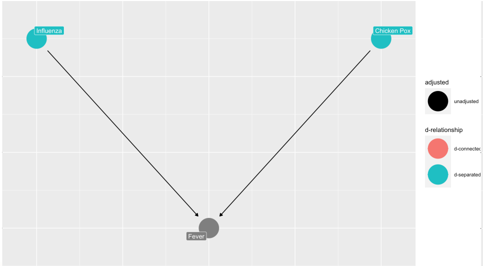
How graphical models help inferring causality?

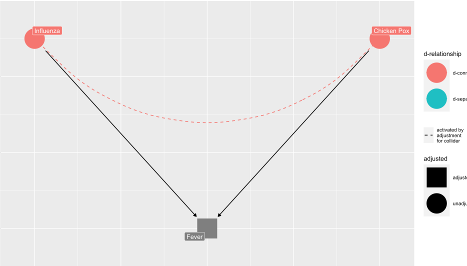
Given an acyclic causal graph, every d-separation relation is mirrored in an equivalent statistical independency in the observational data **if the causal model is correct**.

Collider

Definition

A collider vertex on an undirected path is a vertex with arrows pointing into if from both directions

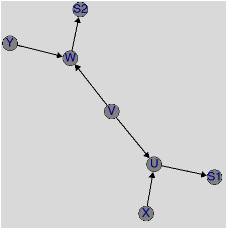




d-separation

Definition

- Given a causal graph G, if X and Y are two different vertices in G and W is a set of vertices in G that does not contain X or Y, then X and Y are d-separated given W in G if and only if there exists no undirected path U between X and Y such that
  - (a) every collider on U is either in W or else has a descendant in W
  - (b) no other vertex on U is in W



- X and V are d-separated
  - X and V are unconditionally independent
- U and W are d-separated given V
  - X and V are independent, conditioned on V
- X and Y are d-separated
  - X and Y are unconditionally independent
- X and Y are not d-separated given U and W
  - X is not independent of Y, conditioned simultaneously on U and W

D-sep tests

- Use the d-separation to predict a set of conditional probabilistic independencies
- Extract the minimum set of d-separation statements in a *basis set*
- Test for independence by using unconditional or conditional correlation coefficient (for example)
- Use Fisher's combined test to obtain a composite probability of the entire set of causal statements

Glossary

- `ggm`
  - DAG
  - dSep
  - basiSet
  - pcor.test
  - shipley.test
- `daggitty`
  - isCollider
  - dseparated
  - impliedConditionalIndependencies
  - localTests

Coding break 3: example of causal inference

Outline

1 From Linear model to Path model

2 Model

3 Causality

4 Ending words

Take-home messages

- SEM is a tool for **modeling (complex) systems via causal assumptions**
- Design of models should not be performed with a pure statistical point-of-view
- SEM can used for **CFA** and **EFA**
- SEM are **easy to use in R**

- Modeling specification and estimation can lead to **unusable models**
  - Convergence issues
  - Constraint sensitivity
  - Negative variance
  - ...
- SEM does **not solve causal inference** but **gives clues**

Extensions

- Multilevel SEM modeling
- Meta-Analysis in SEM
  - testing the consistency of the estimates and effect sizes in different studies
  - estimation of a polled effect size
  - identification of potential moderators that influence the model's structure
- Multi-group SEM
- Latent growth curve modeling (LGCM)
- Non-linear SEM
  - Package `piecewiseSEM`

Thank you for your attention!