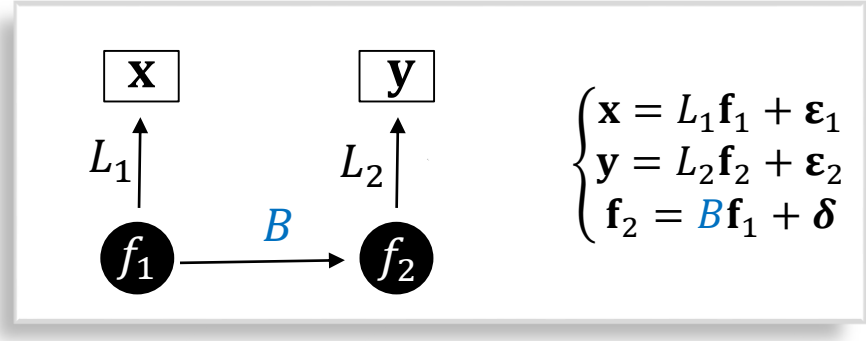
第十四讲 结构方程模型简介

2024.4.24

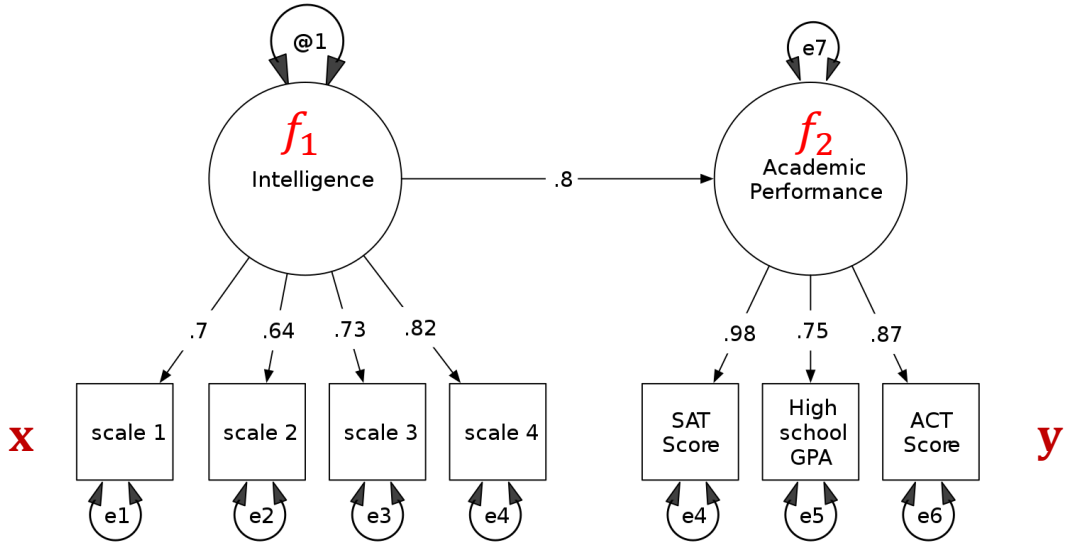


感兴趣的问题：变量f_1, f_2是否存在因果关系，但它们不可测，比如下图中的Intelligence, Academic performance.

我们只能测到它们的外显变量（或替代指标）**x, y**。

结构方程模型(Structural Equation Models，SEM)以可观测的显变量推断潜变量之间的关系，在社会学、经济学、政治学、行为科学、心理、教育学中应用广泛。

SEM模型以路径图(path diagram)表达：



SEM发展历史

- Charles Spearman （1863-1945，心理学家）：单因子分析
- Sewall Wright （1889-1988，遗传学家）：路径分析path analysis，因果研究先驱。
- Louis Thurstone (1887-1955，心理学家)，Herman Wold（1908-1992，经济学家，统计学家）：多因子分析
- Herbert Simon (1916-2001, 经济学家）：Under certain assumptions correlation is an index of causality.
- Hubert Blalock (1926-1991, 社会学家）：Simon-Blalock technique
- Otis Duncan （1921-2004，社会学家）：path analysis and causal models
- Karl Joreskog (统计学家）：协方差结构分析，SEM分析软件：LISREL (linear structural relations)

SEM软件

SEM商业软件

- LISREL (70's Karl Joreskog)： 应用最广泛的软件
- EQS (80's Peter Bentler)
- AMOS (90's James Arbuckle)
- Mplus (00's Bengt Muthen)
- SAS： proc Calis

R Package:

- lavaan (2010-, oll): latent variable analysis, 格式简单，仍在开发阶段，尚不能画path diagram。

- sem （2001-，John Fox）: 模型设定格式比较复杂，可画path diagram。
- Lava (2012-Klaus Holst)：

下面介绍，SEM的三种主要模型：

1. 确认因子分析（检验探索因子分析的结果）
2. 线性回归模型，路径分析（无潜变量）
3. 一般结构方程模型(包含上述两种)

1. 确认因子分析

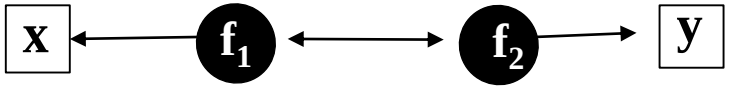
探索因子分析(EFA：exploratory FA)：即前面我们考虑的因子分析，用于探索发现潜在的因子，解释显变量（可以观测的变量）之间的相关性。

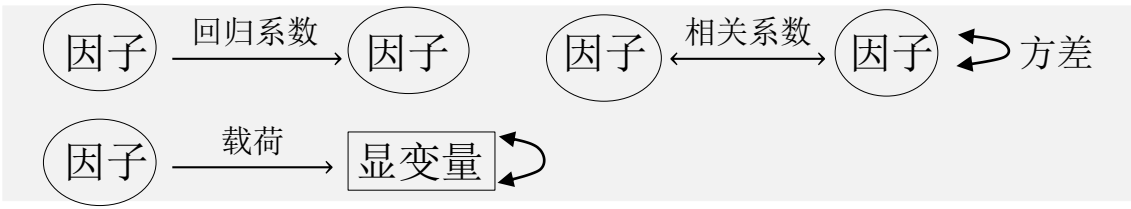
确认因子分析(CFA：confirmative FA)：基于EFA结果，CFA假设某些载荷为0，即明确假设哪些变量与哪个因子有关，但假设因子之间的相关性，然后检验假设的合理性。

简言之，CFA与EFA基本相同，它对EFA做了两点修正：

- 假设某些载荷为0，无需估计；
- 假设因子之间是相关的。

路径图
椭圆：潜变量
方形：显变量
单向箭头：因果
双向箭头：相关
弯曲双箭头：





SEM模型通常以路径图（path diagram)表示。路径图中，

- 椭圆代表因子(潜变量)，不可测量，但可推断。
- 长方形代表显变量(manifest variable,可以测量的变量）。
- 单箭头表示因果或载荷显示：因子之间的单箭头表示因果；因子到显变量的单箭头表示载荷显示（也是因果）。
- 双箭头表示相关：因子之间的双箭头表示相关。弯曲双箭头表示方差（有时为了简化，不显示）。
- 箭头旁边的数字为回归系数（因子→因子）、相关系数（因子↔ 因子）、载荷（因子→显变量） 。
- 特殊因子（误差）通常不显示。

例1.基于EFA结果，我们假设如下确认因子模型：

Gaelic, English, Histroy只和F2有关，

Arithmetic, Algebra, Geometry只和F1有关.

这些假设体现在CFA中前三门课在F_1上载荷为0，后三门课在F_1上载荷为0。假设因子之间相关系数为ρ。

CFA模型：

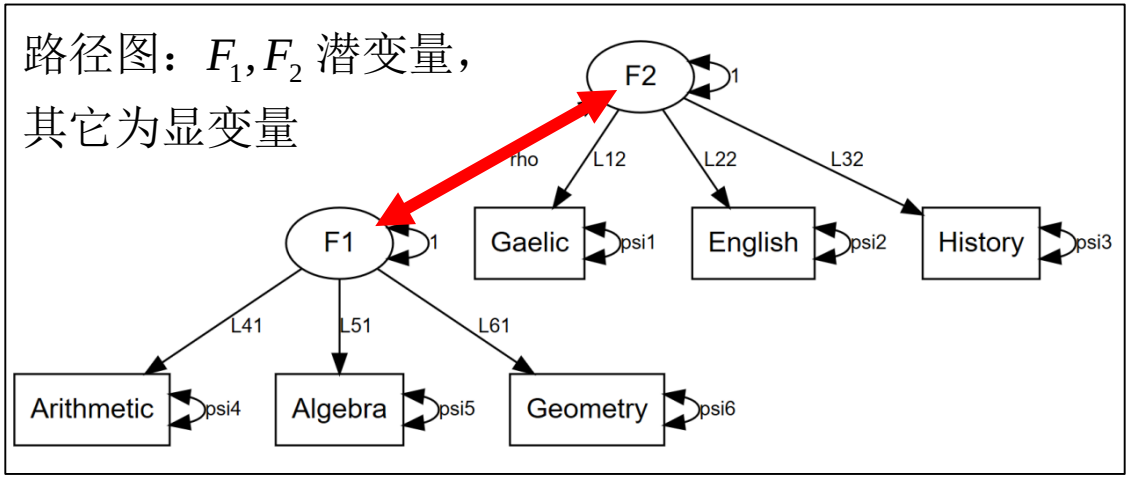
$$\text{Gaelic} = l_{12}F_2 + \varepsilon_1$$
$$\text{English} = l_{22}F_2 + \varepsilon_2$$
$$\text{History} = l_{32}F_2 + \varepsilon_3$$
$$\text{Arithmetic} = l_{41}F_1 + \varepsilon_4$$
$$\text{Algebra} = l_{51}F_1 + \varepsilon_5$$
$$\text{Geometry} = l_{61}F_1 + \varepsilon_6$$

$\varepsilon_i \sim N(0, \psi_i), i = 1,\ldots,6$独立.

$\varepsilon's$与F_1, F_2独立。

$$\begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

路径图：F_1, F_2 潜变量，其它为显变量



极大似然法

假设CFA模型：

$$\mathbf{x}_{p\times 1} = L_{p\times m}F_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1},\ \boldsymbol{\varepsilon} \sim N_p(0,\Psi),\ F \sim N_m(\mathbf{0},\Omega),\ \boldsymbol{\varepsilon} \perp\!\!\!\perp F$$

其中Ψ对角，载荷L有若干元素限制为0，Ω非对角，则

$$\mathbf{x} \sim N(0,\Sigma(\boldsymbol{\theta})),\Sigma(\boldsymbol{\theta}) = L\Omega L^{\mathsf{T}} + \Psi$$

其中$\boldsymbol{\theta}$代表所有参数L,Ψ,Ω,极大似然法等价于极小化

$$\log\det(\Sigma(\boldsymbol{\theta})) + \mathrm{tr}(S\Sigma(\boldsymbol{\theta})^{-1}),$$

其中S为$\mathbf{x}$的协方差矩阵或相关系数矩阵。

例1中，$\mathrm{L} = \begin{pmatrix} l_{12} & \\ l_{22} & \\ l_{32} & \\ & l_{41} \\ & l_{51} \\ & l_{61} \end{pmatrix}, F = \begin{pmatrix} F_2 \\ F_1 \end{pmatrix} \sim N_2(\mathbf{0},\Omega), \Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

拟合优度

拟合优度基于似然比统计量

$$W = n\log\left(|\Sigma(\hat{\boldsymbol{\theta}})|/|S|\right)$$

$\text{p-value} = \text{P}(\chi^2_{df} > W),\ df = v_1 - v_0$，其中

- $v_1 = p(p+1)/2$ (不假设结构情形时Σ参数个数);
- $v_0 = p + |L|_0 + m(m-1)/2$ (因子模型假设下参数个数).

EFA中$v_0 = p + |\text{L}| - m(m-1)/2$

例1中，确认因子模型的特殊方差个数为$p = 6$,

因子个数为$m = 2$（因子之间的相关系数个数为$m(m-1)/2 = 1$),

载荷矩阵L的非0个数记作$|L|_0 = 6$。

R程序包
lavaan

lavaan： 方便使用，仍在开发, 主要函数sem()
sem： 早期开发，可绘制路径图,主要函数sem()

```
> library(lavaan)
##1. 指定模型
>  mymodel = '
  # latent variable definitions  (潜变量/因子定义)
    factor1 =~ x1+x2..   # 显变量x1,x2 与因子factor1 有关
  # regressions （显变量之间或因子之间的线性回归关系）
    factor1 ~ factor2      # factor1=b*factor2+error
    x1 ~ x2                # x1 = b*x2+error
  # residual correlations （方差，协方差）
    factor1 ~~ factor2 #factor1与factor2相关 ‘

##2. 调用sem
> sem(model=mymodel, sample.cov=, sample.nobs=, std.lv= )
  #sample.cov: 样本协方差，另一种数据指定方式为 data=原始数据矩阵)
  #当不提供 原始数据矩阵时，需指定样本量; lv.std=TRUE: 因子的方差设定为1
```

=~和~对应于路径图中的单箭头

~~对应于路径图中的双箭头

例1 （代码）

```
> library(lavaan)

#指定模型：
> mymodel <- '
  # latent variable definitions
     F1 =~ Arithmetic + Algebra+Geometry   # “=~”read as “is manifested by”
     F2=~ Gaelic +English+History
 # residual correlations （方差，协方差）
    F1~~F2  ‘

#拟合模型：
> fit  <- sem(model=mymodel, sample.cov=r ,sample.nobs=220 ,  std.lv=TRUE )
   #std.lv=T 因子方差为1
> summary(fit)
```

> summary(fit) #输出结果（参数估计和方差协方差估计）

Model Test User Model:

Test statistic	7.990
Degrees of freedom	8
P-value (Chi-square)	0.434

Latent Variables:

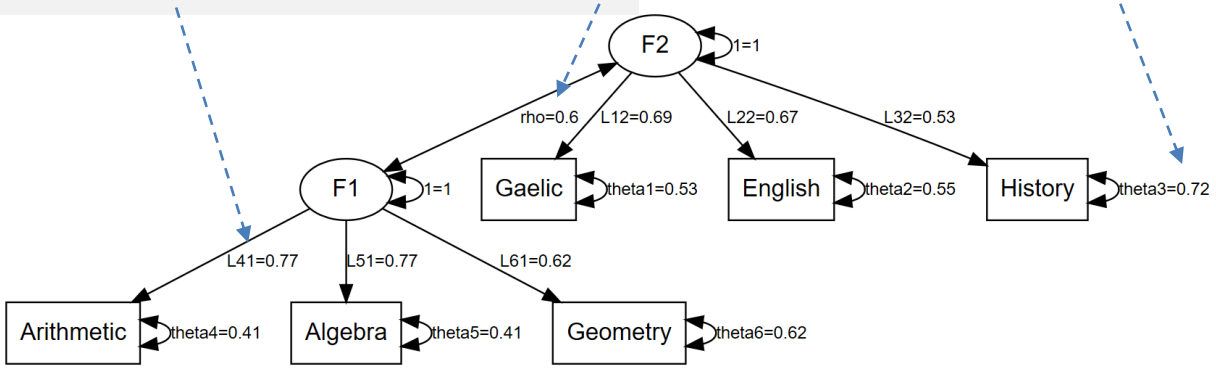
	Estimate	Std.Err	z-value	P(>|z|)
F1 =~				
Arithmetic	0.765	0.067	11.406	0.000
Algebra	0.767	0.067	11.436	0.000
Geometry	0.614	0.069	8.963	0.000
F2 =~				
Gaelic	0.685	0.075	9.096	0.000
English	0.671	0.075	8.920	0.000
History	0.531	0.075	7.060	0.000

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)
F1 ~~				
F2	0.597	0.072	8.332	0.000

Variances:

	Estimate	Std.Err	z-value	P(>|z|)
.Arithmetic	0.411	0.068	6.071	0.000
.Algebra	0.408	0.068	6.025	0.000
.Geometry	0.618	0.071	8.712	0.000
.Gaelic	0.526	0.082	6.437	0.000
.English	0.545	0.081	6.723	0.000
.History	0.713	0.081	8.754	0.000
F1	1.000			
F2	1.000			



2. 线性回归和路径分析（无潜变量）

线性回归和路径分析中所有变量（除了误差项）都是显变量。

路径分析包含多个线性回归方程。

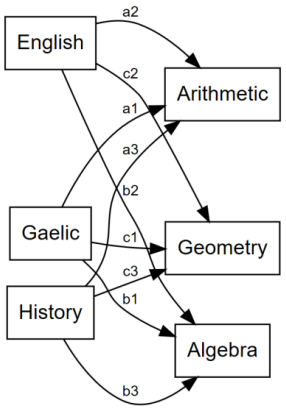
多元线性回归

$$\mathbf{y}_{m\times 1} = A_{m\times p}\mathbf{x}_{p\times 1} + \boldsymbol{\varepsilon}_{m\times 1}$$

$$\left\{\begin{array}{l} y_1 = a_1x_1 + ... + a_px_p + \varepsilon_1 \\ \vdots \\ y_m = c_1x_1 + ... + c_px_p + \varepsilon_m \end{array}\right\} \quad \varepsilon's与x's独立。$$

$$\text{lavaan}: y_1 \sim x_1 + ... + x_p,\ ...,\ y_m \sim x_1 + ... + x_p$$

```
model <- '
  Arithmetic ~ Gaelic +  English +History
  Algebra ~ Gaelic +  English +History
  Geometry   ~ Gaelic +  English +History
 # Arithmetic ~~Algebra + Geometry
   #具有相同自变量的响应之间自动计算协方差
 # Algebra ~~  Geometry  '
fit <- sem(model, sample.cov=r ,sample.nobs=220 )
```



路径分析

路径分析（path analysis）模型中没有潜变量，只有显变量，路径分析模型通过一系列线性回归方程表达因果关系。
注意，基本上只有基于时间次序建立的因果关系方程才是可靠的。

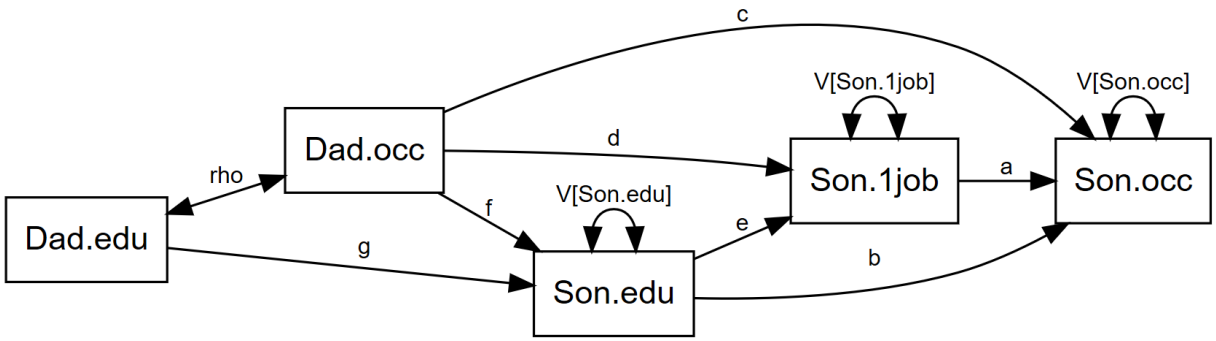
	Son.occ	Son.1job	Son.edu	Dad.occ	Dad.edu
Son.occ	1	0.541	0.596	0.405	0.322
Son.1job	0.541	1	0.538	0.417	0.332
Son.edu	0.596	0.538	1	0.438	0.453
Dad.occ	0.405	0.417	0.438	1	0.516
Dad.edu	0.322	0.332	0.453	0.516	1

按照时间次序，假设父子职业、教育程度等变量的关系：

$$\text{Son.edu} = f * \text{Dad.occ} + g * \text{Dad.edu} V + \varepsilon_{edu}, \varepsilon_{edu} \sim N(0, v_{Son.edu})$$

$$\text{Son.1job} = d * \text{Dad.occ} + \text{e} * \text{Son.edu} + \varepsilon_{job}, \varepsilon_{job} \sim N(0, v_{Son.1job})$$

$$\text{Son.occ} = a * \text{Son.1job} + b * \text{Son.edu} + c * \text{Dad.occ} + \varepsilon_{occ}, \varepsilon_{occ} \sim N(0, v_{Son.occ})$$



3. 结构方程模型

一般的结构方程模型是CFA、路径分析的拓展，特别地，可以**用于研究潜变量的因果关系**。观察研究的显变量通常不满足线性模型“误差与自变量独立”的假设，一般难以研究因果关系。

例2. 工业化与政治民主 (数据集：PoliticalDemocracy， in lavaan package)．为了研究工业化程度与民主化程度的关系，收集了1960年和1965年75个国家的4个政治民主度量 (两年共8个) 以及3个工业化度量：

- 60年民主化指标（出版，政见，选举，立法）：
  Press60， *Oppo60*，*Elect60*， *Legis60*
- 65年民主化指标:
  Press65， *Oppo65*， *Elect65*，*Legis65*
- 60年工业化指标:
  GNP60 (人均国民生产总值), *Energy60* (人均能量消耗), *Labor60* (产业工人比例)

数据参见下页。

数据

	Press60	Oppo60	Elect60	Legis60	Press65	Oppo65	Elect65	Legis65	GNP60	Energy60	Labor60
1	2.5	0	3.33	0	1.25	0	3.73	3.33	4.44	3.64	2.56
2	1.25	0	3.33	0	6.25	1.1	6.67	0.74	5.38	5.06	3.57
3	7.5	8.8	10	9.2	8.75	8.09	10	8.21	5.96	6.26	5.22
4	8.9	8.8	10	9.2	8.91	8.13	10	4.62	6.29	7.57	6.27
5	10	3.33	10	6.67	7.5	3.33	10	6.67	5.86	6.82	4.57
6	7.5	3.33	6.67	6.67	6.25	1.1	6.67	0.37	5.53	5.14	3.89
7	7.5	3.33	6.67	6.67	5	2.23	8.27	1.49	5.31	5.08	3.32
8	7.5	2.23	10	1.5	6.25	3.33	10	6.67	5.35	4.85	4.26
…											

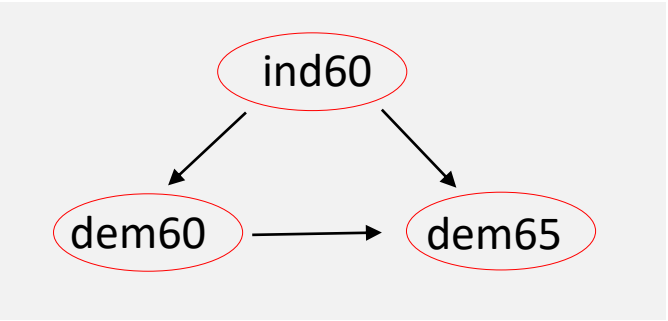
相关系数矩阵

	Press60	Oppo60	Elect60	Legis60	Press65	Oppo65	Elect65	Legisl65	GNP60	Energy60	Labor60
Press60	1	0.6	0.68	0.69	0.74	0.65	0.67	0.67	0.38	0.32	0.25
Oppo60	0.6	1	0.45	0.72	0.54	0.71	0.58	0.61	0.21	0.25	0.21
Elect60	0.68	0.45	1	0.61	0.58	0.43	0.65	0.53	0.33	0.31	0.23
Legis60	0.69	0.72	0.61	1	0.65	0.66	0.68	0.74	0.47	0.44	0.39
Press65	0.74	0.54	0.58	0.65	1	0.56	0.68	0.63	0.56	0.52	0.43
Oppo65	0.65	0.71	0.43	0.66	0.56	1	0.61	0.75	0.34	0.35	0.33
Elect65	0.67	0.58	0.65	0.68	0.68	0.61	1	0.71	0.39	0.4	0.35
Legis65	0.67	0.61	0.53	0.74	0.63	0.75	0.71	1	0.46	0.46	0.37
GNP60	0.38	0.21	0.33	0.47	0.56	0.34	0.39	0.46	1	0.89	0.8
Energy60	0.32	0.25	0.31	0.44	0.52	0.35	0.4	0.46	0.89	1	0.85
Labor60	0.25	0.21	0.23	0.39	0.43	0.33	0.35	0.37	0.8	0.85	1

假设存在3个因子ind60, dem60, dem65, 含义如下：

- ind60：60年的工业化程度，表现为GNP60, Energy60, Labor60,
- dem60：60年的民主化程度，表现为Press60, Oppo60, Elect60, Legis60,
- dem65：65年的民主化程度，表现为Press65, Oppo65, Elect65, Legis65,

假设因子按照年代次序满足如下因果关系：



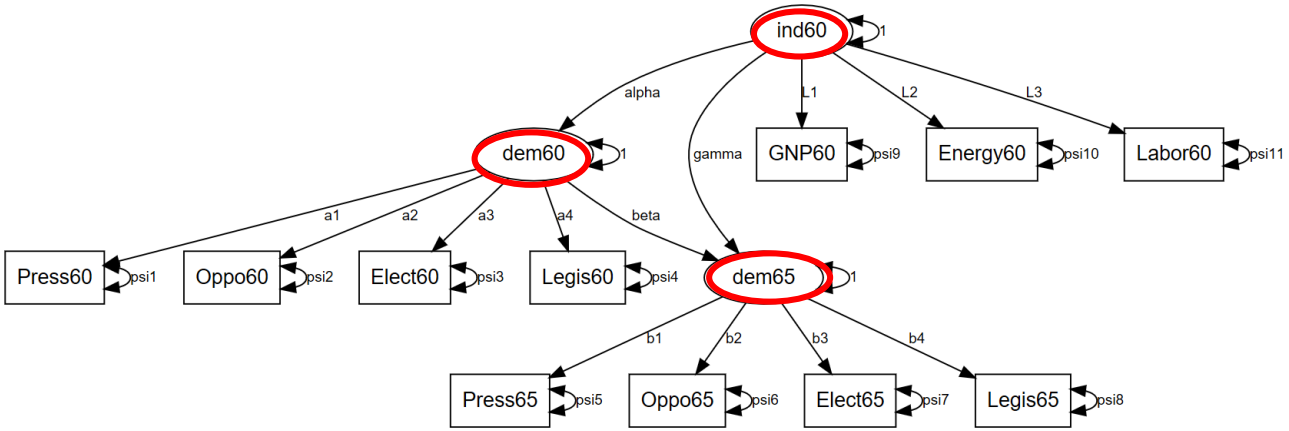
- ind60影响dem60, 也影响dem65
- dem60影响dem65

因子之间的回归模型：

$$\begin{cases} \text{dem60} = \alpha \times \text{ind60} + \text{e}_1, \\ \text{dem65} = \beta \times \text{dem60} + \gamma \times \text{ind60} + \text{e}_2 \end{cases}$$

另外，假设因子模型

$$\begin{cases} \text{GNP60} = L_1 \times \text{ind60} + \varepsilon_1, \ldots \\ \text{Press60} = a_1 \times \text{demo60} + \tilde{\delta}_1, \ldots \\ \text{Press65} = b_1 \times \text{demo65} + \tilde{\delta}_5, \ldots \end{cases}$$




```
#指定模型：
model <- '
  # latent variable definitions
     ind60 =~   GNP60 + Energy60 + Labor60
     dem60 =~ Press60+Oppo60+Elect60+Legis60
     dem65 =~ Press65+Oppo65+Elect65+Legis65
  # regressions
    dem60 ~ ind60
    dem65 ~ ind60 + dem60
  # residual correlations
    Press60 ~~ Press65
    Oppo60 ~~Legis60 +Oppo65
    Elect60 ~~ Elect65
    Legis60 ~~ Legis65
    Oppo65 ~~ Legis65‘
R=cor(PoliticalDemocracy)
fit <- sem(model, sample.cov=R ,sample.nobs=75,std.lv=T)
summary(fit)
```

#输出结果

Regressions:

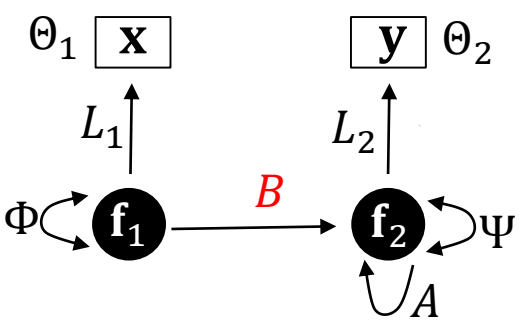
	Estimate	Std.Err	z-value	P(>|z|)
dem60 ~				
ind60	0.499	0.144	3.460	0.001
dem65 ~				
ind60	0.923	0.628	1.469	0.142
dem60	4.010	2.617	1.533	0.125

结论：dem60~ind60的p值=0.001，显著，说明同年的工业化程度对民主化有显著的影响。 60年的工业化和民主化对65年的民主化程度影响不显著。

结构方程模型的一般形式

假设观察研究有显变量**x**和**y**，它们分别是潜变量$\mathbf{f}_1$和$\mathbf{f}_2$的某种测量或外在展示(有时**x**, **y**称为是$\mathbf{f}_1$, $\mathbf{f}_2$的替代指标或带误差的测量)。

假定潜变量$\mathbf{f}_1$与$\mathbf{f}_2$存在因果关系，比如线性模型，我们以SEM研究这种假定的因果关系是否合理。



$$\begin{cases} \mathbf{x} = L_1\mathbf{f}_1 + \boldsymbol{\varepsilon}_1 \\ \mathbf{y} = L_2\mathbf{f}_2 + \boldsymbol{\varepsilon}_2 \\ \mathbf{f}_2 = B\mathbf{f}_1 + A\mathbf{f}_2 + \boldsymbol{\delta} \end{cases}$$

$$\mathrm{var}(\boldsymbol{\delta}) = \Psi;\ \mathrm{var}(\mathbf{f}_1) = \Phi$$

$$\mathrm{var}(\boldsymbol{\varepsilon}_1) = \Theta_1, \mathrm{var}(\boldsymbol{\varepsilon}_2) = \Theta_2$$

潜变量$\mathbf{f}_1$的弯曲双箭头表示$\mathbf{f}_1$内部的方差Φ，$\mathbf{f}_2$的单向弯曲箭头表示$\mathbf{f}_2$内部之间的回归关系:$\mathbf{f}_2 = A\mathbf{f}_2 + B\mathbf{f}_1 + \boldsymbol{\delta}$，$A$对角为0，代表$\mathbf{f}_2$的分量之间的回归。内生变量(单箭头指向的变量)旁边的参数Θ_1，Θ_2是特殊方差。

假设各个误差和因子服从多元正态分布，则显变量$\begin{pmatrix}\mathbf{x}_{q\times 1}\\ \mathbf{y}_{r\times 1}\end{pmatrix}\sim N_p(0,\Sigma(\boldsymbol{\theta})),$

$p=q+r, S$ 为$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$的协方差矩阵或相关系数矩阵，极大似然法极小化

$$\log\det(\Sigma(\boldsymbol{\theta}))+\operatorname{tr}(S\Sigma(\boldsymbol{\theta})^{-1})$$

即参数的MLE仅与样本协方差矩阵S有关。

拟合优度基于似然比统计量

$$W=-2\log\left(\frac{\max\limits_{\boldsymbol{\theta}} L(\Sigma(\boldsymbol{\theta}))}{\max\limits_{\Sigma} L(\Sigma)}\right)=-2\log\left(\frac{L(\Sigma(\hat{\boldsymbol{\theta}}))}{L(\hat{\Sigma})}\right)=n\log\left(|\Sigma(\hat{\boldsymbol{\theta}})|/|S|\right)$$

W近似服从χ^2_{df}，$df=v_1-v_0$，$v_1=p(p+1)/2$，v_0为结构方程模型的参数总数。

总结

- SEM本质上是对样本的协方差矩阵结构给予限制，是一种协方差矩阵结构分析方法。
- 结构方程模型所表达的因果关系都是假定的(putative)模型（注意：通常只能对有时间次序的变量建立因果模型）。