Latent Variable Analysis with R

张曜

课程内容

探索性因子分析

验证性因子分析

结构方程模型

潜在增长模型

课程目标

- 理解潜变量分析的用途/回答什么样的问题
- 思考线性回归、因子分析、结构方程模型的主要区别和关联
- 学会用R写探索性因子分析、验证性因子分析、结构方程模型的范式
- 学会对结果的解读&可视化
- 了解潜变量模型的一些变式

拓宽1种思路=学会100种方法

什么是潜变量模型?

使用外显变量对潜变量进行操作化，同时使用统计模型来估计外显变量与潜变量之间的关系，进而使用可观测的外显变量来间接估计不可直接观测的潜变量，或探索潜变量与潜变量的关系。这些统计模型被统称为潜变量模型(Latent Variable Modeling, LVM)。凡是涉及测量指标与潜变量之间关系的模型都是LVM。

太复杂了！ 不读了！ 往下看！

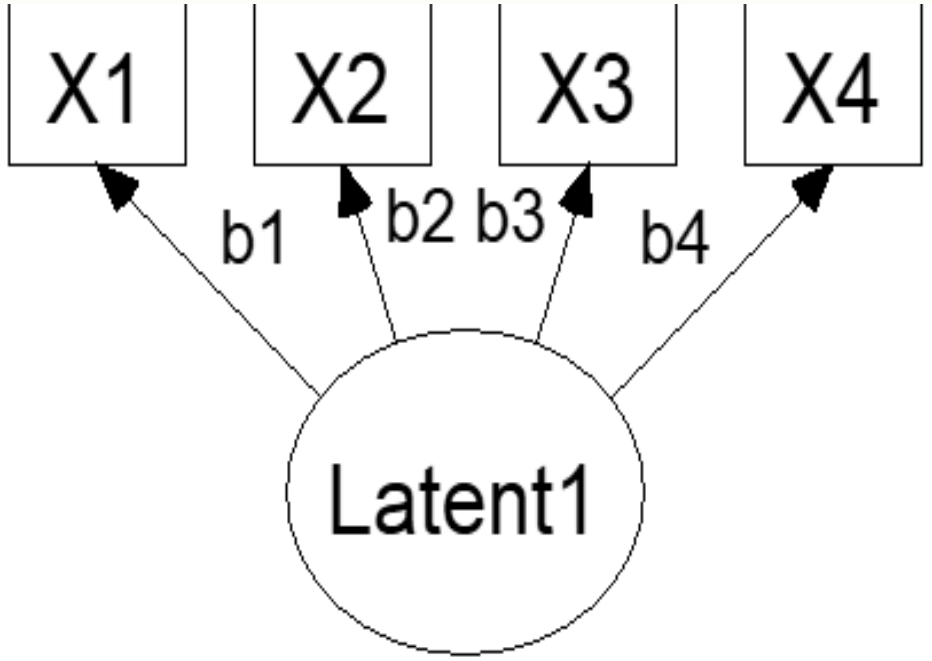
相关概念

- 什么是**爱**?
- 你对你的生活**满意**吗?
- 你觉得旁边的小伙子**帅**不帅?
- 你感到**焦虑**不?
- 你是什么**人格**?
- 你的**领导力**如何?
- 什么是**国家脆弱性**?
- 啥叫**现代化**? 啥叫**小康社会**?

相关概念

- **爱、满意、帅**这些都是我们脑中的构念(constructs)，无法直接测量或难以直接测量，又叫做**潜变量(latent variables)**
- “每天想他**100**次”、“**每天**都要和他说晚安”、“我们在一起**5年**了”等等这些都是可以直接测量(观测)的**显变量（Manifest variables)**

相关概念



例如左图

Latent1: 他**爱**我

X1-X4: 他爱我的各种**细节/迹象**

（e.g., 每天想我100次；每天都要拥抱我3次）

Y = b1*X1 + b2*X2 + b3*X3 + b4*X4 + …

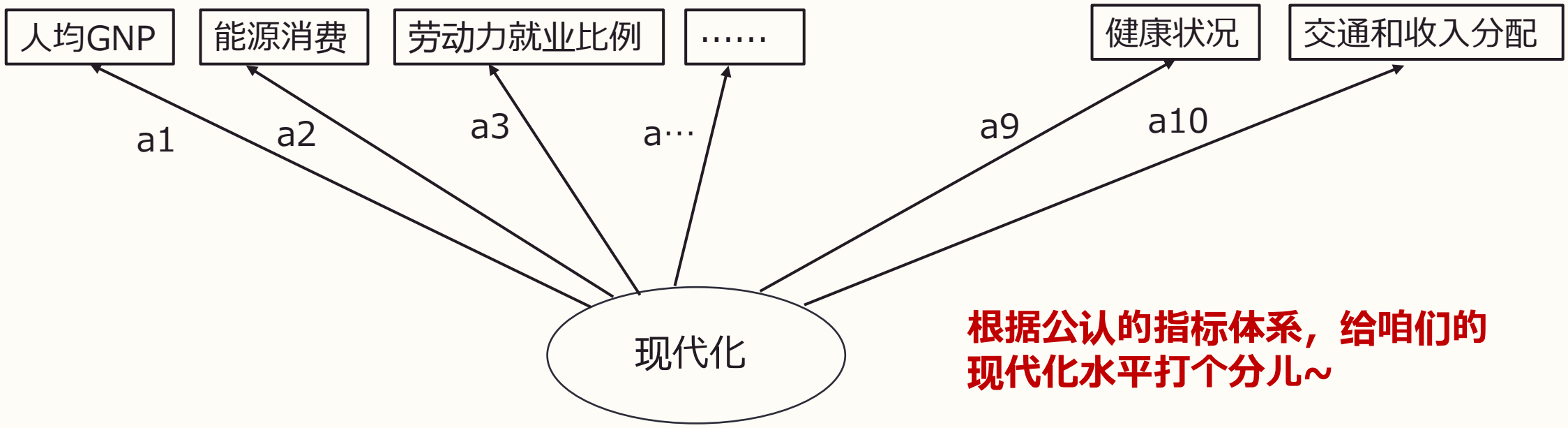
哪些指标可以表示他爱我？

Y ——→ 他到底有多爱我？？？

相关概念

例如：如何评估一个国家是否进入现代化？

普林斯顿大学国际研究中心的布莱克教授提出的**现代化10项指标**：人均GNP、能源消费、劳动力就业比例、各部门占GNP比例、最终产品占GNP比例、城市化、教育状况、健康状况、交通和收入分配。



根据公认的指标体系，给咱们的现代化水平打个分儿~

相关概念

潜变量：不可直接观测，是一种构念（construct）、一种因子（factor）；

显变量：可直接观测，是一个题项（item）、一个指标（indicator）

潜变量用圈圈，显变量用框框

探索性因子分析

从一堆指标当中进行筛选和赋予权重，探索出一个合理的简便的综合评估模型，降维思想。

比如：给同理心打个分/哪些因素影响你的同理心

步骤：

- 第一、确认待分析的原有若干指标是否适合做因子分析；
- 第二、确定公共因子个数、提取因子变量；
- 第三、利用旋转方法使因子变量更具有可解释性，推荐正交旋转；
- 第四、可视化结果；
- 第五、计算因子得分和总分。
- psych包

探索性因子分析

```
> fa(correlations, nfactors = 3, n.obs = 467, rotate = "none", scores = T, fm = "pa")
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 3, n.obs = 467, rotate = "none",
    scores = T, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1   PA2   PA3   h2   u2 com
ec1 0.48  0.13  0.15 0.27 0.73 1.3
ec2 0.66  0.02  0.34 0.55 0.45 1.5
ec3 0.73 -0.16  0.17 0.59 0.41 1.2
ec4 0.76 -0.33  0.06 0.70 0.30 1.4
ec5 0.77 -0.31 -0.06 0.68 0.32 1.3
ec6 0.68 -0.38 -0.02 0.61 0.39 1.6
pt1 0.34  0.39 -0.03 0.27 0.73 2.0
pt2 0.56  0.53  0.22 0.64 0.36 2.3
pt3 0.63  0.27 -0.01 0.47 0.53 1.4
pt4 0.62  0.08  0.11 0.41 0.59 1.1
pt5 0.69  0.09 -0.56 0.79 0.21 2.0
pt6 0.54  0.16 -0.37 0.45 0.55 2.0

                       PA1  PA2  PA3
SS loadings           4.81 0.94 0.68
Proportion Var        0.40 0.08 0.06
Cumulative Var        0.40 0.48 0.54
Proportion Explained  0.75 0.15 0.11
Cumulative Proportion 0.75 0.89 1.00

Mean item complexity =  1.6
Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the null model are  66  and the objective function was  5.4 wit
h Chi Square of  2489
The degrees of freedom for the model are 33  and the objective function was  0.43

The root mean square of the residuals (RMSR) is  0.04
The df corrected root mean square of the residuals is  0.06
```

PA表示成分载荷，即观测变量与因子的相关系数

h2表示公因子方差，即全部公共因子对每个指标的方差解释度

h2 = PA1^2 + PA2^2 + PA3^2

如果大部分变量的h2都高于0.8，则说明提取出的公共因子已经基本反映了各原始变量80%以上的信息，因子分析效果较好，是因子分析效果的重要指标，显然这个数据因子分析效果不好。

u2表示成分唯一性，即指标方差无法被因子解释的比例(1-h2)

Proportion Var：每个公共因子对整个数据集的解释程度，衡量因子重要程度。

Cumulative Var：累计解释率，纳入80%以上的因子，显然这里三个因子累计没达到80%，效果不好。

验证性因子分析

验证性因子分析(cfa)，经常用于检验问卷的效度和基于前人理论对现有指标进行**验证**。

- 建立模型
- 拟合模型
- 统计拟合情况
- 可视化
- lavaan包

SEM模型

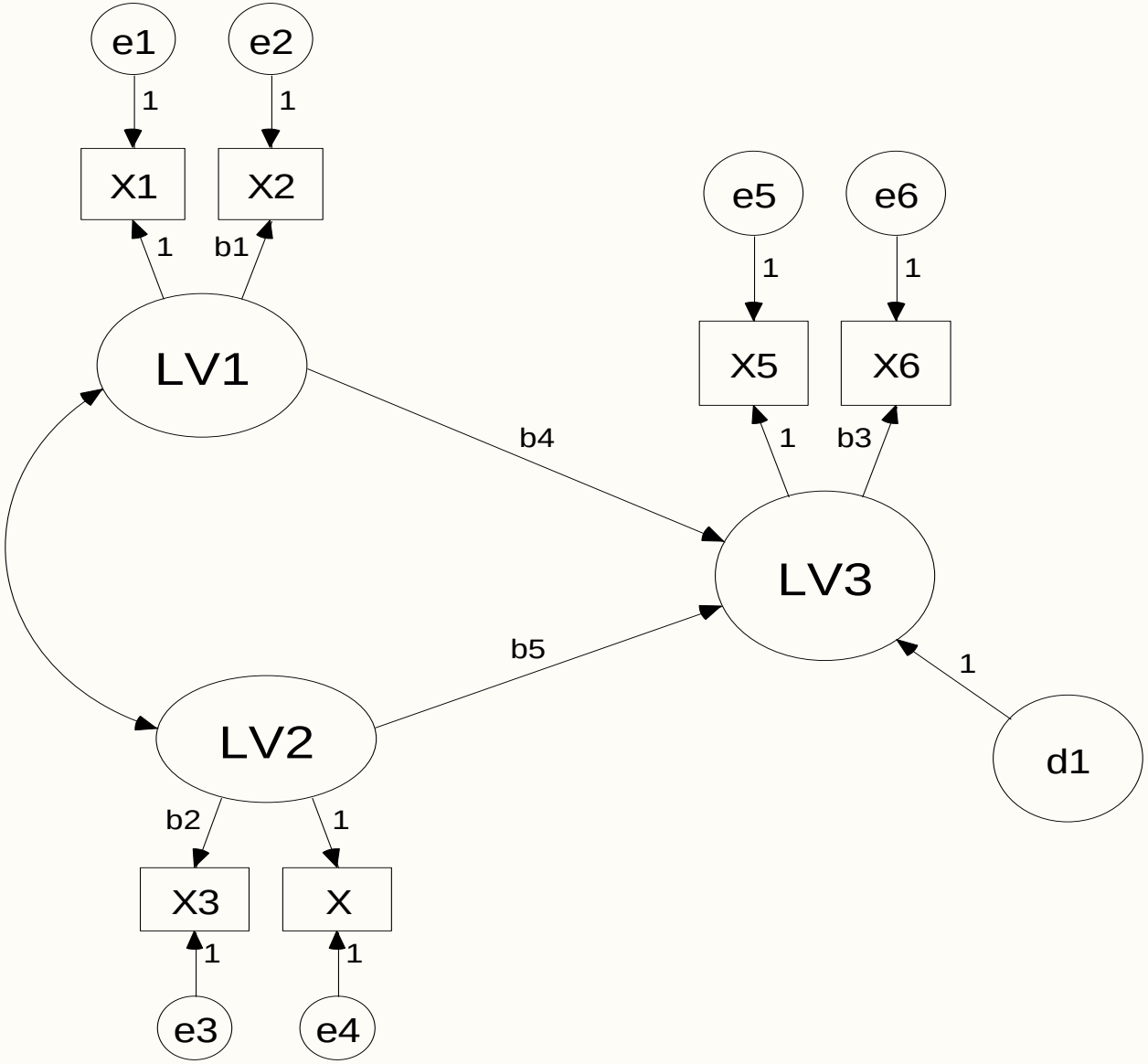
思考：

潜变量与潜变量之间可以进行分析吗？

外生潜变量(exogenous)?

内生潜变量(endogenous)?

误差项(error item)?

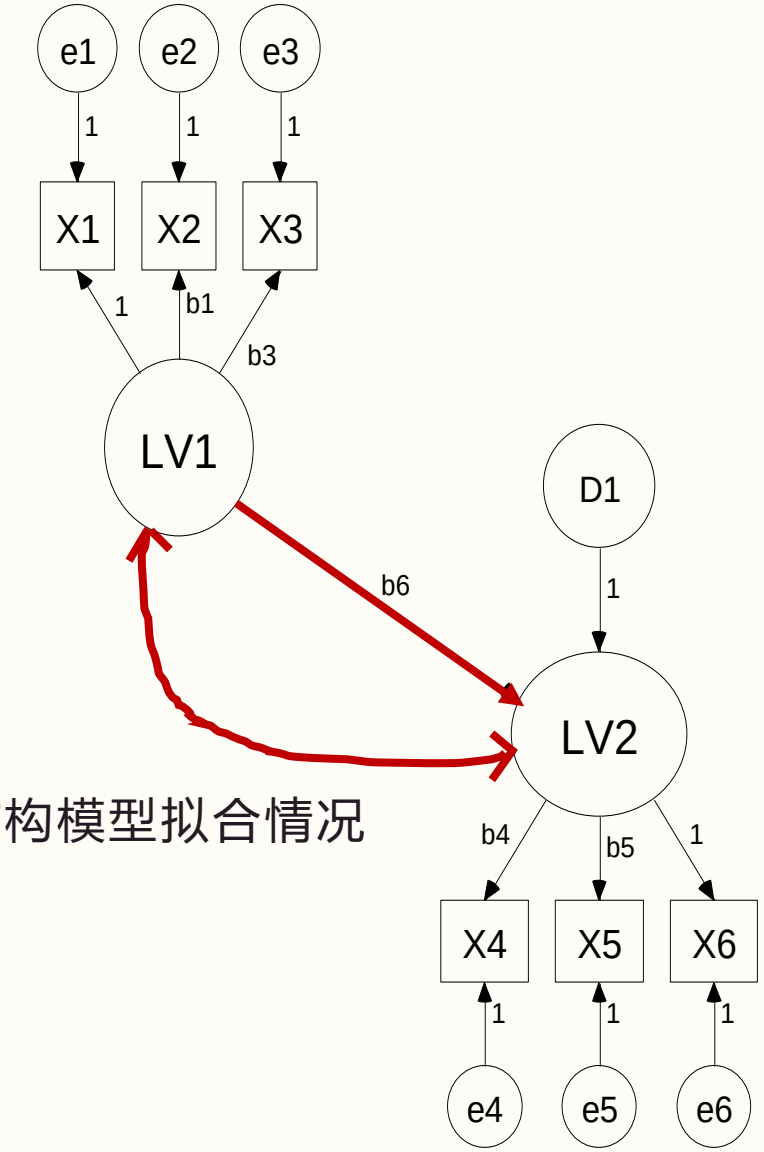


SEM模型

SEM VS. regression model?

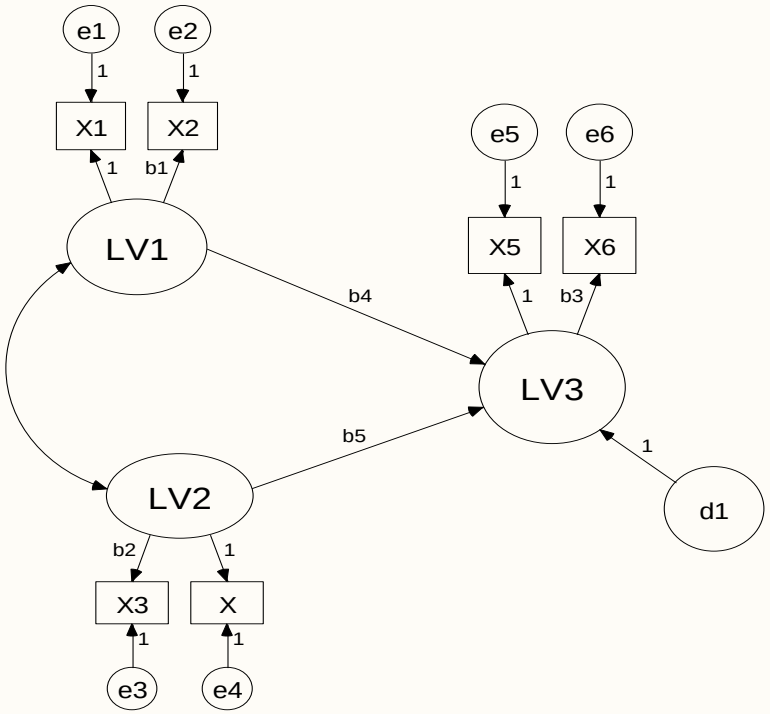
SEM VS. factor model?

- 线性回归模型假定自变量无测量误差
- Y=aX+b， x的测量真的无误差吗?



- 因子分析(efa)只反映测量变量的因子载荷，SEM还探索因子之间的结构模型拟合情况
- 因子分析(efa)是探索性分析，**SEM是验证性分析**

Taking diagrams → equations



Measurement equations:

X1 = 1.0 LV1 + e1

X2 = b1 LV1 + e2

X3 = b2 LV2 + e3

X4 = 1.0 LV2 + e4

X5 = 1.0 LV3 + e5

X6 = b3 LV3 + e6

Structural equations among latent variables:

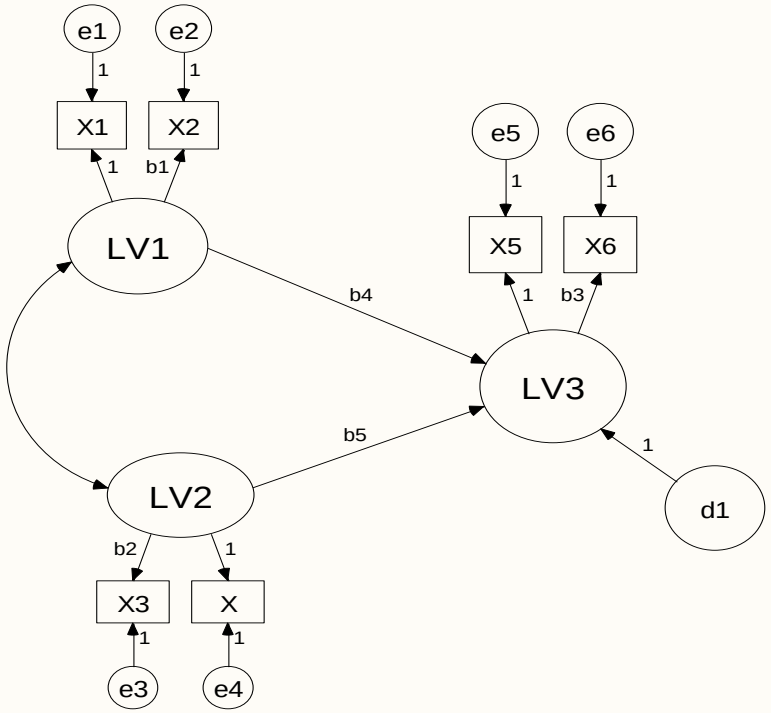
(“construct equations”)

LV3 = b4 LV1 + b5 LV2 + d1

Notes: It is common to have one indicator with a fixed measurement equation parameter value of 1.0.（固定载荷）

SEM模型设定

Model parameters includes:



Path coefficients:

b1, b2, b3, b4, b5

Error term variances:

Var (e1) var (e2) var (e3) var (e4)

Var (e5) var (e6) var (d1)

Exogenous latent variable variances:

Var (LV1) var (LV2)

Exogenous latent variable covariance:

Cov (LV1, LV2)

Error term covariances:

(NONE IN THIS MODEL)

Total number of model parameters: 15

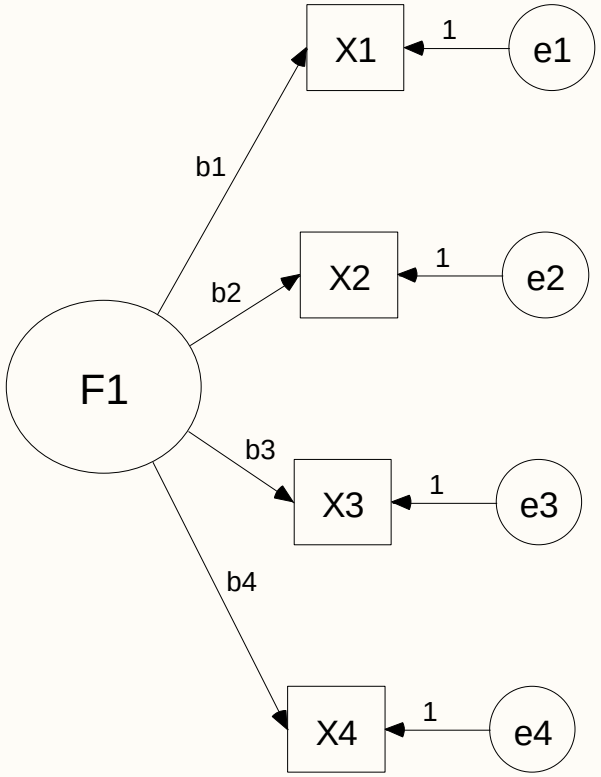
SEM模型识别

模型的识别依据：自由度（df）

Degrees of freedom = number of observed variances/covariances - number of parameters

(k+1)(k) / 2

df >=0, over-identified, necessary
df < 0, under-identified



With 4 manifest variables, we have 10 empirical variances and covariances

Parameters:

b1,b2,b3,b4 (1.0), var(F1)

VAR(e1) VAR(e2) VAR(e3) VAR(e4)

Total = 8

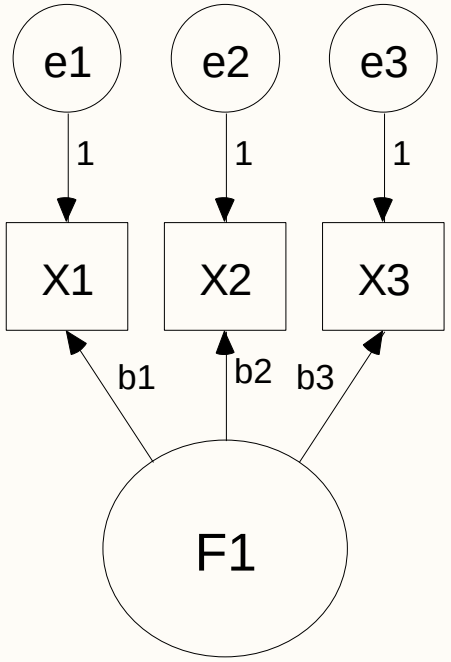
df = 2

Over-identified

SEM模型识别

Three indicator rule:

Latent variable model with at least 3 indicators will be identified



Parameters:

VAR(F1) *

VAR(E1)

VAR(E2)

VAR(E3)

b1 *

b2 *

b3 *

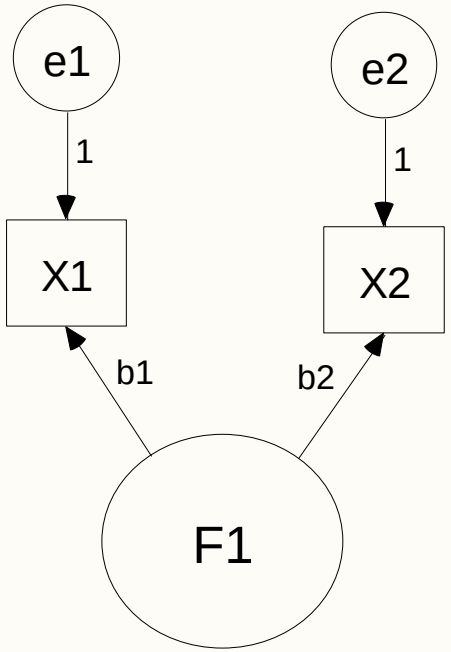
*one of these must be fixed to 1.0

Empirical covariances/variances: 6 Parameters: 6

df=0

Three-indicator rule

What if only 2 indicators?



Empirical covariances/variances: 3 Parameters: 4

df = -1

Under-identified

One possible solution: impose constraints

e1 = 0.48

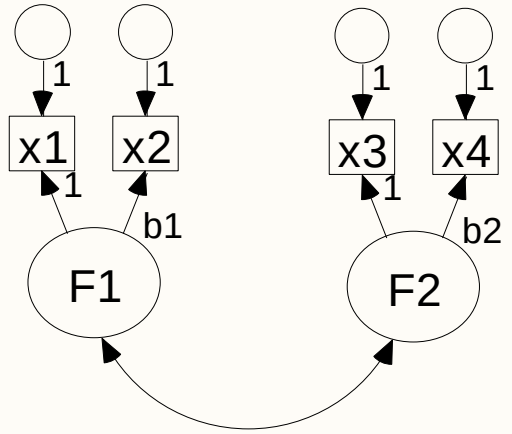
e2 = 4e1

b2 = 2b1

Another possible solution: add 2-indicator models

Cov(F1,F2) ≠ 0

df = 10 – 9 = 1



SEM模型拟合与检验

Chi-square test, $X^2 = F*(N-1)$, **sample-size dependent**

Other commonly-used fit measures:

- GFI (Goodness of fit index, 拟合优度指数, affected by N though)
- AGFI(Adjusted GFI)
- NFI (Normed Fit Index, 标准拟合指数)
- CFI (Incremental, Comparative fit indices)
- RMSEA (Root Mean Square Error of Approximation, 均方根残差)
- AIC/BIC (信息指数)
- ……

注意：拟合指数良好不代表模型就是正确的，该模型在理论上不合理，那么拟合结果再好也没有意义。

比如：用我一天吃几碗饭、洗几次澡衡量我有多温柔体贴……

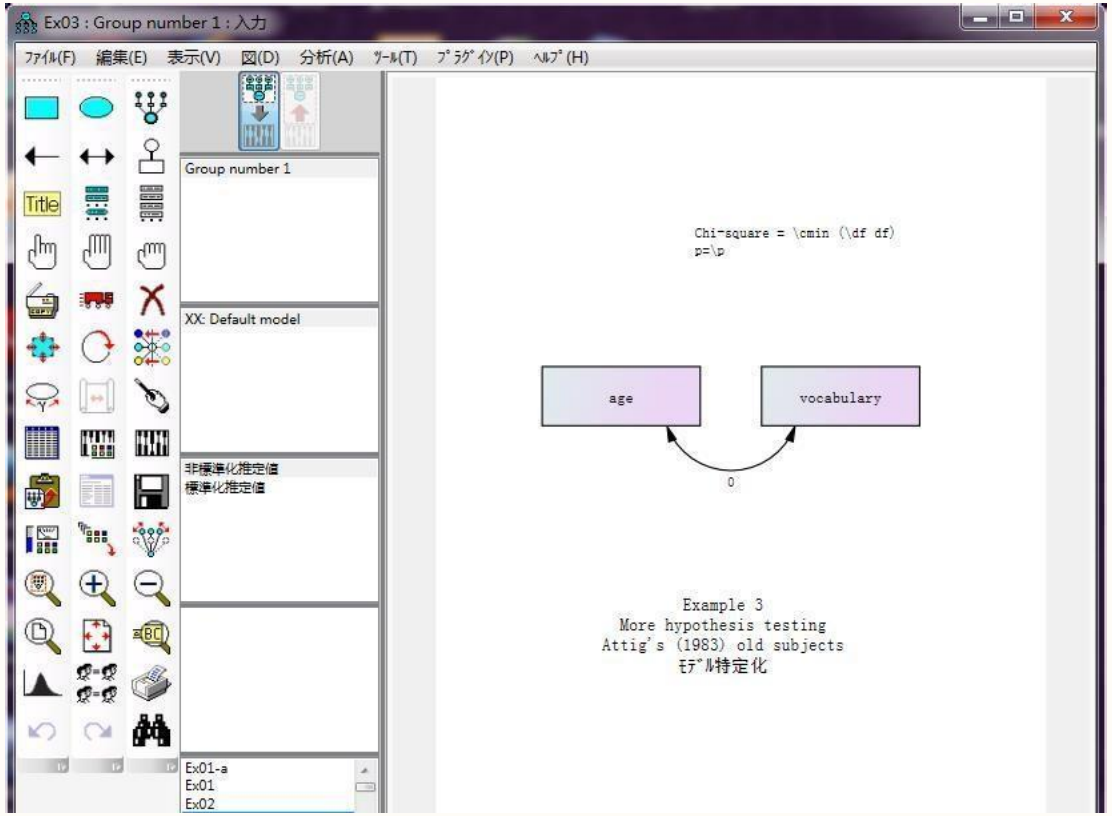
SEM模型拟合与检验

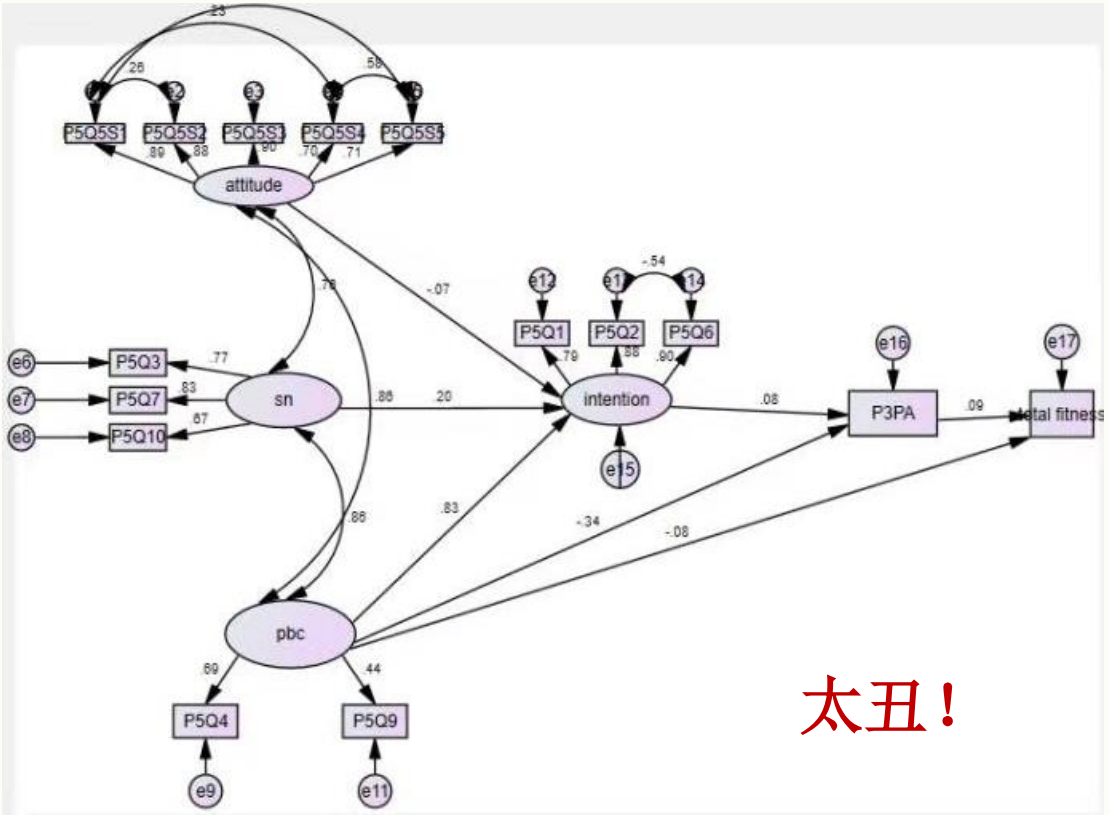
X^2 越小越好， $p < 0.05$

GFI, AGFI, NFI, CFI, TLI et al., > 0.9

RMSEA < 0.08

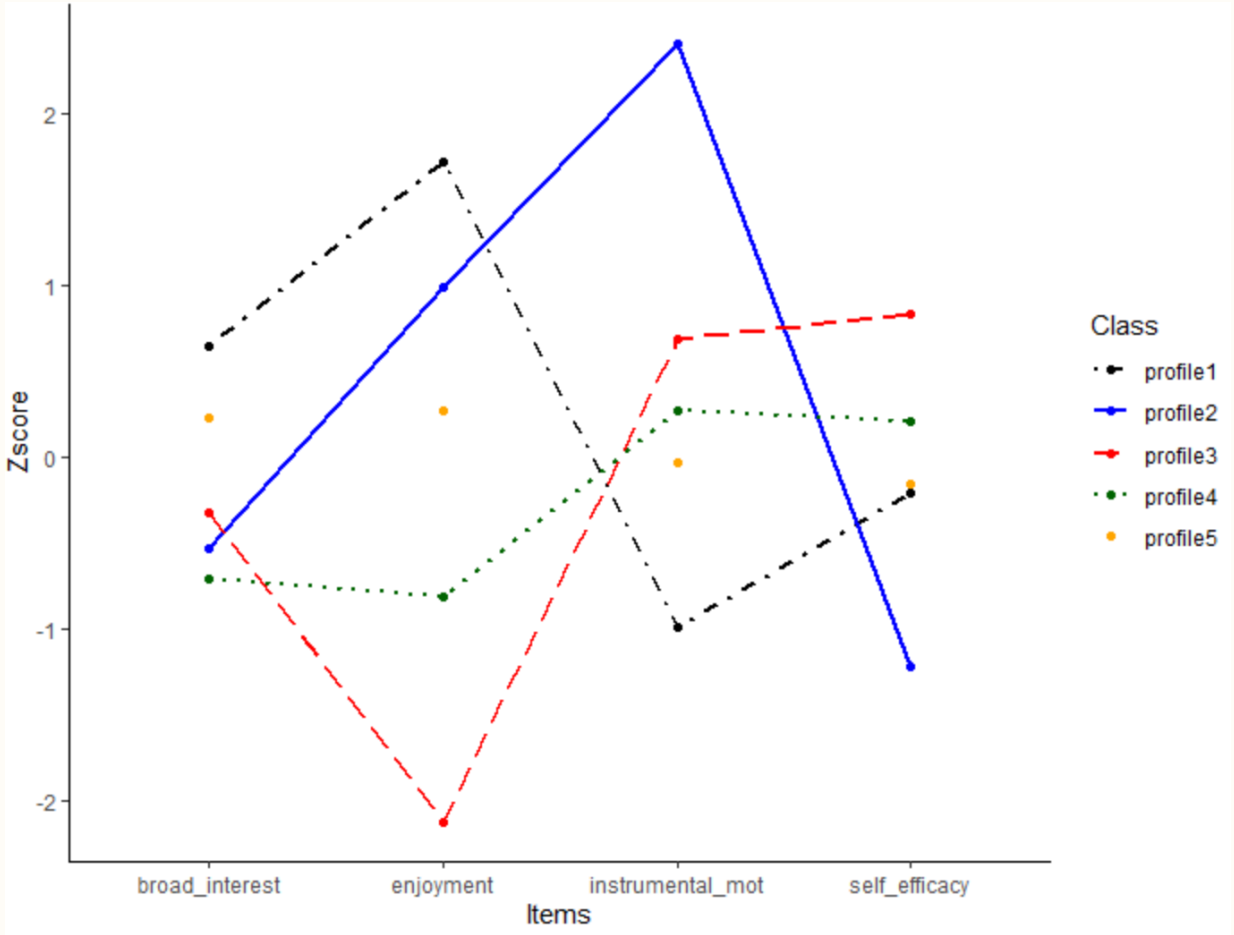
Why R?

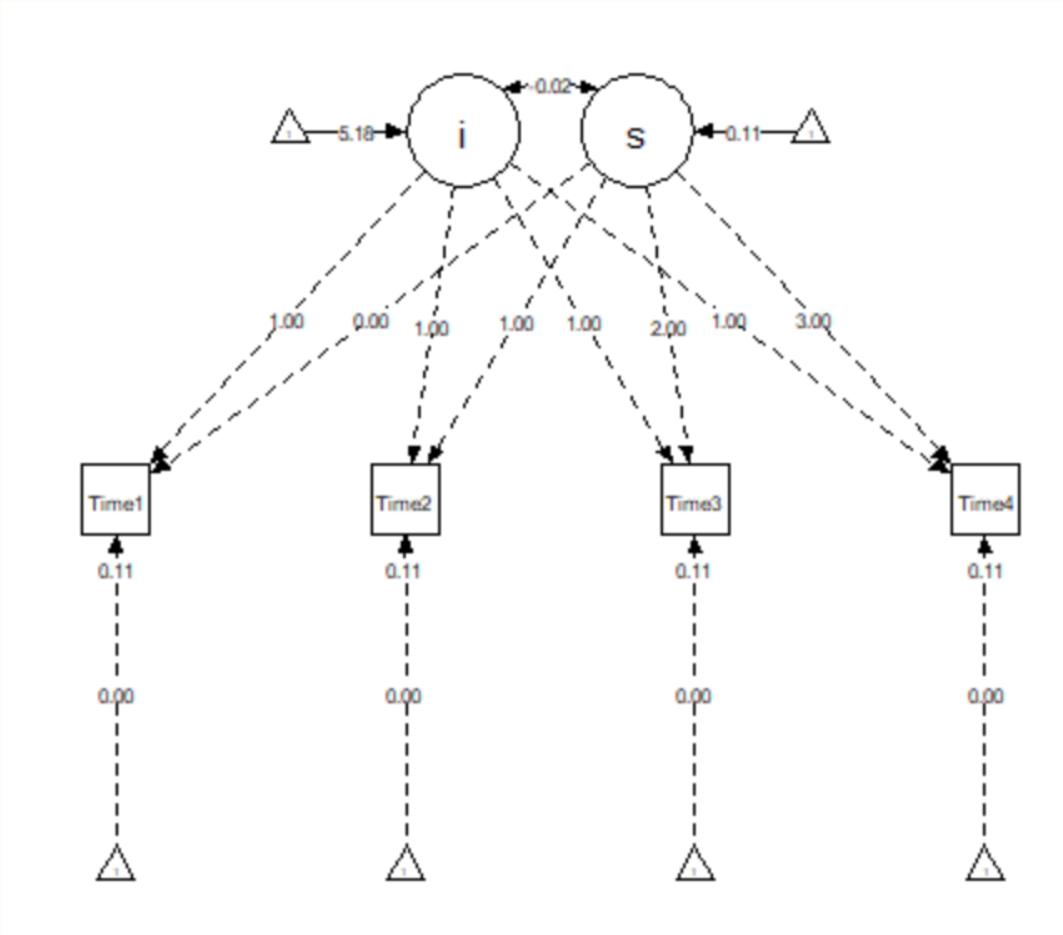




是AMOS点点鼠标不香？还是Mplus不够专业？

Why R?





R画图专业、好看，R的代码说人话、易理解

潜变量模型的分类-SEM的变式

根据潜变量和显变量是否连续：

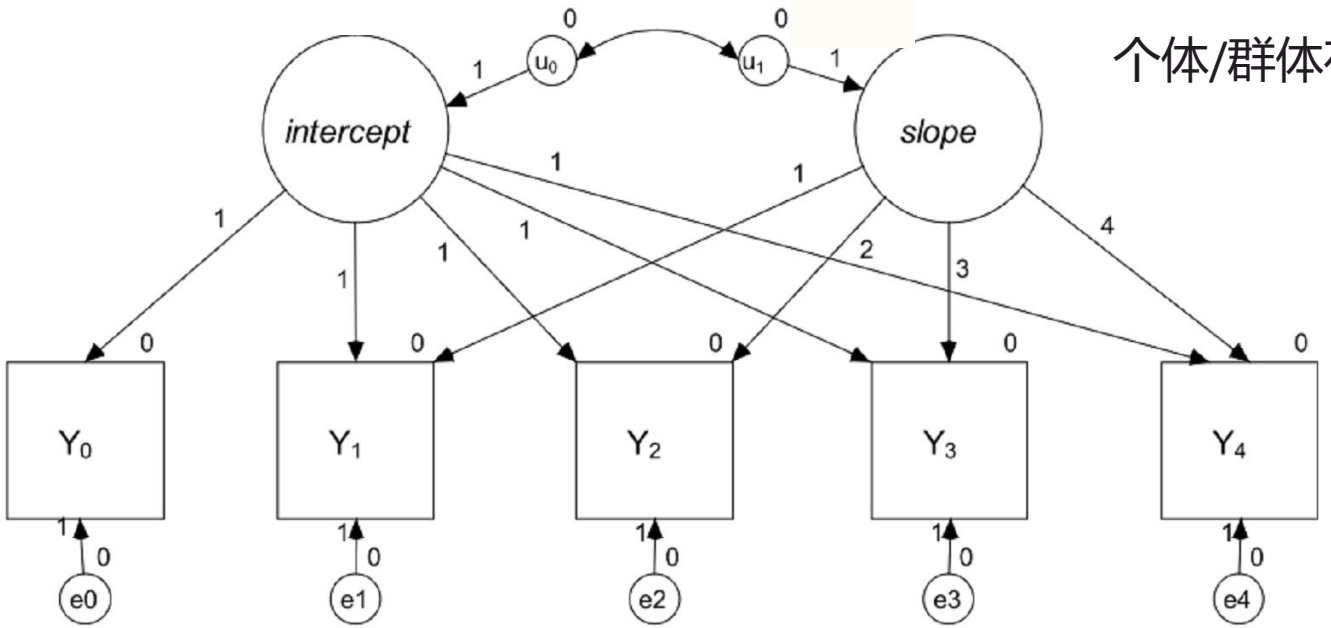
潜变量	外显变量	
	类别	连续
类别	潜在类别分析 Latent Class Analysis	潜在剖面分析 Latent Profile Analysis
连续	潜在特质分析/项目反应理论 Latent Trait Analysis Item Response Theory	因素分析 Factor Analysis

根据潜变量类型和数据类型：

	连续潜变量	类别潜变量	混合
横断面模型 Cross-section Models	因子分析(Factor Model，FA)，SEM	回归混合模型 (Regression Mixture Modeling, **RMM**) 潜类别分析(Latent Class Analysis, **LCA**)	因子混合模型(Factor Mixture Modeling, **FMM**)
纵向模型 Longitudinal Models	增长模型 (Growth Model)	潜在转换分析 (Latent Transition Analysis, **LTA**); 潜类别增长模型(latent class growth modeling, **LCGM**)	增长混合模型(Growth Mixture Model, **GMM**)

另外根据变量之间是否有双向因果关系：
只有单项因果——递归模型
有双向因果——非递归模型

潜在增长模型

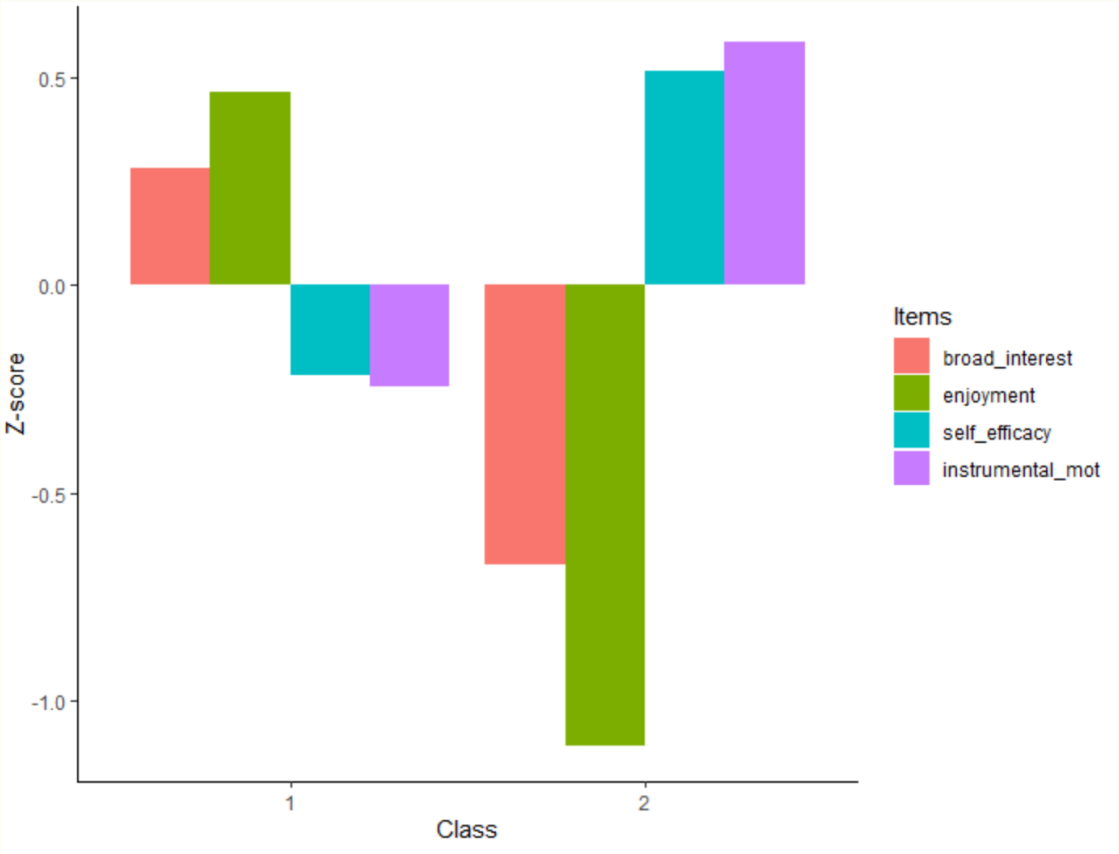


潜在增长模型(Latent Growth Model，LGM) 可以分析某一变量的变化趋势，用不可测量或难以测量的潜变量来描述总体的平均增长趋势，还可分析总体发展趋势和群组之间存在的差异，也可以分析个体之间的发展差异。

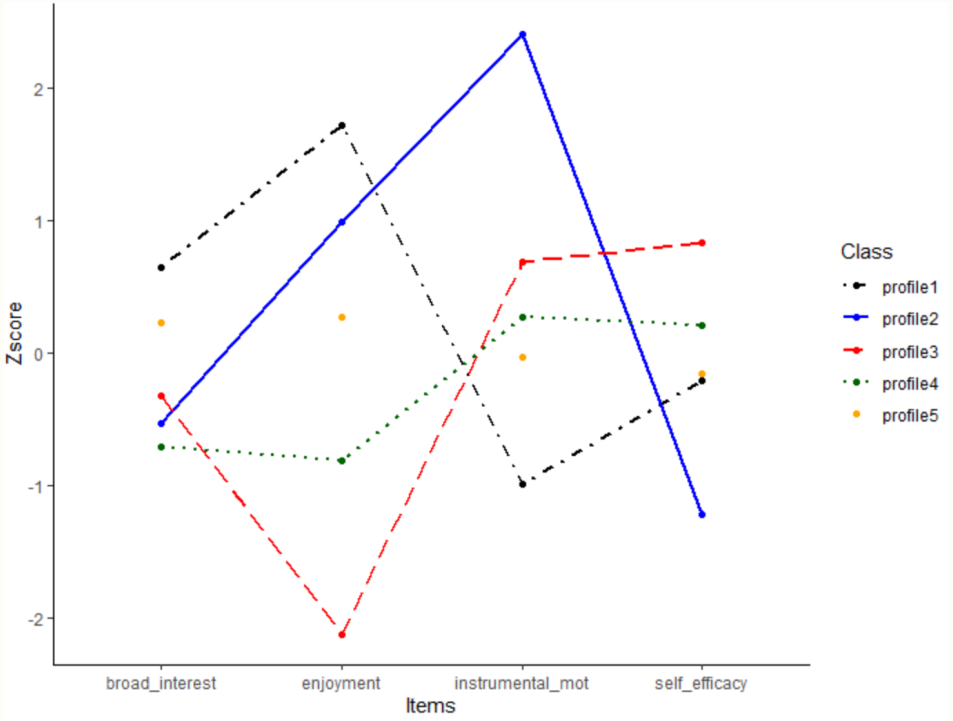
潜在增长模型是将截距（baseline）和斜率（changing rate）定义为潜变量，以描述纵向数据的变化特征

个体/群体在时间上的变化可以是线性的也可以是非线性的

潜在剖面模型



简单来说，潜在剖面模型就是用来探讨个体是否在某个潜变量上存在着差异或不同类别，以此来更有针对性的对个体进行教育、干预和治疗。比如：发现一群青少年的抑郁情绪可以分为低、中、高三类，而中等的那一类青少年表现出高BMI和家庭贫穷等特性。该方法在心理学、教育学、体育学、医学中都非常重要，一般用来分析横截面数据。可以理解为非监督分类的一种方法。



课后思考题：

1. multilevel模型 VS. 潜在增长模型？

2. 对tidyLPA包中的empathy数据构建SEM模型

```
install.packages(“tidyLPA”)
library(tidyLPA)
data(empathy)
```

3. 尝试用你们的数据写出路径分析/SEM/潜在增长模型/潜在剖面模型

Thanks for listening !