心理统计

第二十五讲：降维分析、调节变量和中介变量

严超赣

Chao-Gan Yan, Ph.D.

yancg@psych.ac.cn
http://rfmri.org/yan

Institute of Psychology, Chinese Academy of Sciences

1

1

主要内容

2

多元线性回归

- 多元线性回归（ Multiple Linear Regression）

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p + \epsilon$$

Y – 因变量，Dependent variable
X_i – 自变量，Independent (explanatory) variable
b_0 –截距，Intercept
b_i –斜率，Slope
ϵ – 残差，Residual (error)

3

多重共线性(multicollinearity)

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p + \epsilon$$

- 其基本假设之一就是解释变量间相互独立。
- 如果某两个或多个解释变量之间出现了相关性，则称为多重共线性。

4

多重共线性

- 在设计矩阵中，至少有一个变量可以表示为其他变量的线性组合。则变量之间存在多重共线性。即：

$$c_1x_1 + c_2x_2 + c_3x_3 + c_ix_i + \epsilon = 0$$

- 若$\epsilon = 0$，则为完全共线性，但在现实情况并不常见，一般$\epsilon \neq 0$，即近似共线性。

5

多重共线性

- 多重共线性的产生是由于变量之间具有相关关系。
- 若解释变量间存在多重共线性，利用最小二乘法推导出的方程没有唯一解。

6

多重共线性的处理

- 如果完全共线性：直接删除不必要的变量
- 如果是近似共线性，处理依赖于经验
  - 整合解释变量（如降维的方法——主成分分析等）
  - 增加样本量

7

Part 1:
探索性因素分析(Exploratory Factor Analysis)

8

因素分析（因子分析）

- 以潜在较少数量的未观察变量（因子）用于描述观察到的、相关的变量之间的变异性。
- 用少数因子反映原始资料的大部分信息。
- 观测变量被建模为潜变量(latent variables)和“误差”项的线性组合。

9

因素分析

- 因素分析（factor analysis）：
  - 探索性因素分析（exploratory factor analysis, EFA）
  - 验证性因素分析法（confirmatory factor analysis, CFA）

10

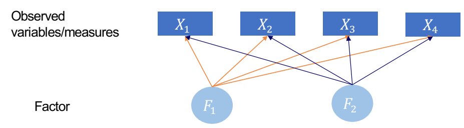
因素分析

- 探索性因素分析：探索多变量的内在结构并进行降维的处理技术，将多变量综合为少数几个核心因子。
  - 变量较多，根据相关性大小可将变量分组，组内间的变量间相关性较高，不同组的变量间相关性较低。
- 验证性因素分析：对因子结构及一系列原变量间的关系是否符合研究者的理论构想进行验证的技术。

11

探索性因素分析

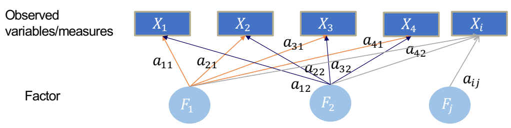
- 探索性因素分析的代数模型：



12

探索性因素分析

- 因素分析的代数模型：



a_{ij}：第i个变量和第j个因子之间的“相关系数”

13

探索性因素分析

- 探索性因素分析的代数模型：

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \cdots \\ X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \\ \cdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

X_1，X_2，... X_p分别表示某被试在第1个、第2个到第p个变量上的得分，以标准分表示

F_1，F_2，... F_m分别表示某被试在第m个公共因子上的得分，以标准分表示

a_{ij}：第i个变量对应回归方程中第j个因子的系数，也称因子载荷

14

探索性因素分析

- 因素分析的矩阵形式：
  - X = （$X_1, X_2, \ldots, X_p$）
  - F = （$F_1, F_2, \ldots, F_m$）
  - ε = （$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$）
  - $X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i$

X_1，X_2，... X_p分别表示某被试在第1个、第2个到第p个变量上的得分，以标准分表示

F_1，F_2，... F_m分别表示某被试在第m个公共因子上的得分，以标准分表示

a_{ij}：第i个变量对应回归方程中第j个因子的系数，也称因子载荷

15

探索性因素分析

- 变量的共同度：
  - $X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_1$
  - $h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2$
  - 共同度在数值上等于因子载荷矩阵中第i行因子载荷的平方和，反映“所有因子对这个变量共同制约的程度”，共同度越接近1越好

16

探索性因素分析

- 公共因子的方差贡献：
  - $X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1$

    $\cdots$
  - $X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p$
  - $v_j^2 = \sum_{i=1}^{p} a_{ij}^2 = a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2$
  - 反映公共因子对所有变量总的影响，以及其在所有公共因子中的相对重要性

17

探索性因素分析

- 因素分析的主要步骤：
  - 适合度检验
  - 构造因素模型并确定因子数
  - 因子旋转
  - 因子得分计算
  - 因子命名与解释

18

探索性因素分析

- 因素分析的适合度检验：
  - 相关矩阵：大部分相关系数低于0.3，且不显著就不适合因素分析
  - 巴特利特球形检验：H_0为“相关矩阵是一个单位阵”，根据行列式计算得到统计量，进行卡方检验，结果显著则可进行因素分析
  - 反像相关矩阵检验：从偏相关矩阵出发，每个元素都是负的偏相关系数，元素绝对值大，说明受其他变量更迭影响小，不适合因素分析

19

探索性因素分析

- 因素分析的适合度检验：
  - KMO取样适合度检验：结合偏相关矩阵和相关矩阵，比较某变量和其他变量的相关系数与偏相关系数的相对大小。如果相关矩阵中绝对值大，偏相关矩阵元素绝对值小，存在公共因子可能性高，适合因素分析
    - KMO > 0.9，非常合适；
    - 0.8 < KMO <0.9，合适；
    - 0.7 < KMO < 0.8，一般；
    - 0.6 < KMO <0.7，不太合适；
    - KMO <0.5，极不合适。

20

探索性因素分析

- 因子提取的方法：
  - 主成分分析法(principal axis factoring)
  - 最小二乘法
  - 极大似然法
  - 映像分析法

21

探索性因素分析

- 因子数的确定：
  - 抽取的m个因子对原变量方差的解释率要达到较高比例，一般建议达到80%，实际运用可根据具体情况调整，也可定在40%~60%。
  - 特征值大于1，低于1的特征值表明该因子可解释的变异信息比原变量的变异方差小。
  - 通过碎石图确定。选择碎石曲线从迅速下降到平缓的拐点确定因子数。
  - 结合定性分析。结合理论与经验分析。
  - Horn’s Parallel analysis: 与相同大小的随机矩阵的特征值进行比较

22

探索性因素分析

- 因子旋转：
  - 初始因子载荷矩阵的含义往往不明确，如果各列因子载荷值之间差异不明显，很难分化原变量与公共因子的对应关系。
  - 因子旋转计算将抽取的因子经过数学变换，分离个因子，凸显其意义，使得每个变量在某因子上有高负荷，在其他因子上负荷较低。

23

探索性因素分析

- 因子旋转的方法：
  - 正交旋转：假设各因子间没有相关关系，旋转过程，各因子轴间保持90°夹角。
  - 斜交因子旋转：不要求因子轴互相垂直，旋转后每条因子轴更靠近各自变量群，斜交因子轴夹角余弦值是两因子间的相关系数。
    - 难以解释各变量被公共因子解释的比例；
    - 斜交旋转后的每列因子载荷平方和只有偶然情况等于共同度；
    - 斜交旋转后因子可作为变量进行因素分析。

24

探索性因素分析

- 因子旋转的方法：

转轴方法	类型
1.最大变异数法(Varimax)	正交orthogonal
2.四方次最大值法(Quartimax)	正交
3.平衡最大值法(Equimax)	正交
4.标准正交法(Orthogonal with gamma)	正交
5.哈雷斯.凯斯法(Harris-Kaiser)	正交与斜交
6.最优斜交法(Promax)	斜交
7.斜交法(Procusces)	斜交

25

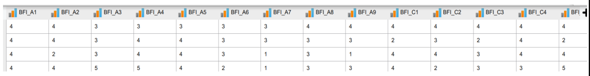
因子命名与因子分计算

- 因素分析反映的是因子与变量的相关关系，而非因果关系。
- 个体在各因子上得分(factor score)计算：
  - 因子分估计值：
  - $\hat{F}_j = W_{j1} \times ZX_1 + W_{j2} \times ZX_2 + \cdots + W_{jp} \times ZX_p$
    - W_{ij}为标准化后的数据矩阵X的加权系数，反映第i个变量和第j个因子的相关关系
    - 误差项为真因子分F与因子分估计值$\hat{F}$之间的差异，E = F- $\hat{F}$

26

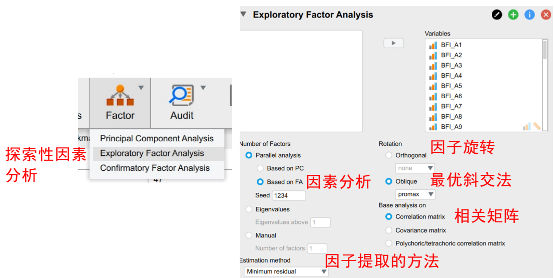
JASP操作

- 大五人格问卷题目分析，表中为每个被试在单个题目的得分
- 数据：example25.csv (Liu et al., 2020, J Open Psych Data)



27

JASP操作



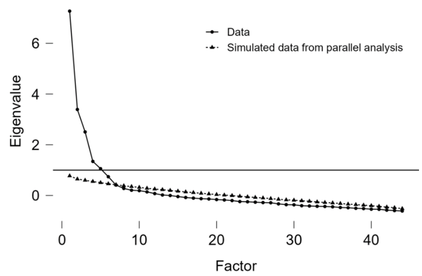
28

JASP操作

29

JASP操作

- 碎石图



30

JASP操作

因子载荷矩阵

Factor Loadings

残差

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Uniqueness
BFI_C6	0.707						0.410
BFI_N7	−0.706						0.461
BFI_N5	0.690						0.507
BFI_N3	0.650						0.441
BFI_N2	0.641						0.575
BFI_N1	0.637						0.520
BFI_N4	0.561						0.679
BFI_N8	−0.425						0.513
BFI_O2		0.771					0.447
BFI_O1		0.719					0.501
BFI_O4		0.662					0.561
BFI_O7		0.661					0.557
BFI_O8		0.419					0.717
BFI_N6		0.401					0.649
BFI_E6			0.729				0.449
BFI_E8			0.704				0.524
BFI_E1			−0.689				0.301
BFI_A8			0.551				0.671
BFI_E2			−0.507				0.456
BFI_C1				0.697			0.564
BFI_C8				−0.688			0.578
BFI_C5				0.576			0.619
BFI_C3				0.563			0.625
BFI_C9				−0.517			0.706
BFI_C2				0.496			0.672
BFI_C4				0.455			0.657
BFI_A1					0.608		0.610
BFI_A4					0.557		0.636
BFI_A5					0.495		0.795
BFI_A7					−0.469		0.638
BFI_A3					0.468		0.664
BFI_O10						0.645	0.575
BFI_O3						−0.590	0.501
BFI_O6						−0.477	0.638

31

JASP操作

Factor Characteristics

旋转前 旋转后

	Unrotated solution			Rotated solution		
	SumSq. Loadings	Proportion var.	Cumulative	SumSq. Loadings	Proportion var.	Cumulative
Factor 1	7.482	0.170	0.170	4.046	0.092	0.092
Factor 2	3.691	0.084	0.254	3.959	0.090	0.182
Factor 3	2.792	0.063	0.317	3.038	0.069	0.251
Factor 4	1.597	0.036	0.354	2.956	0.067	0.318
Factor 5	1.303	0.030	0.383	2.516	0.057	0.375
Factor 6	0.997	0.023	0.406	1.345	0.031	0.406

32

JASP操作

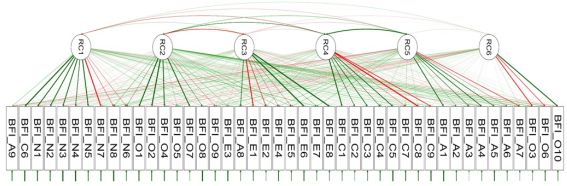
Parallel Analysis

	真实特征值 Real data factor eigenvalues	模拟特征值 Simulated data mean eigenvalues
Factor 1*	7.269	0.773
Factor 2*	3.391	0.649
Factor 3*	2.507	0.599
Factor 4*	1.344	0.552
Factor 5*	1.057	0.506
Factor 6*	0.740	0.461
Factor 7	0.411	0.424
Factor 8	0.278	0.382
Factor 9	0.208	0.351
Factor 10	0.191	0.320
Factor 11	0.143	0.284
Factor 12	0.075	0.253
Factor 13	0.020	0.230
Factor 14	−0.001	0.198
Factor 15	−0.042	0.170
Factor 16	−0.079	0.140
Factor 17	−0.104	0.115
Factor 18	−0.126	0.087
Factor 19	−0.137	0.062
Factor 20	−0.161	0.038
Factor 21	−0.176	0.016

33

JASP操作

- 路径图



34

Part 2:
主成分分析

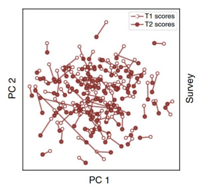
35

主成分分析

12维

Financial risk-taking
Eating control
Reward sensitivity
Emotional control
Mindfulness
Impulsivity
Goal-directedness
Ethical risk-taking
Risk perception
Sensation seeking
Social risk-taking
Agreeableness

2维



(Eisenberg et al. 2019)

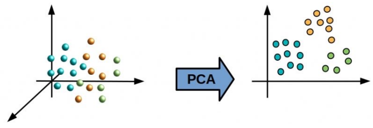
36

主成分分析

- PCA (Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。
- PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

37

主成分分析



https://starship-knowledge.com/pca-vs-linear-regression

38

标准化

- PCA对变量的方差较为敏感；
- 标准化可以将变量转化为标准正态分布，以使每个变量的量纲一致，保证PCA的准确性

39

协方差矩阵

- 数据能够进行降维的原因是因为不同变量之间的存在相关，包含冗余信息。
- 通过协方差矩阵，能够“识别”数据间的相关性。

40

协方差矩阵

- 协方差矩阵是一个p×p 对称矩阵（其中 p是维数），它是变量之间的所有相关联的协方差。
- 例如，对于具有 3 个变量x、y 和z 的三维数据集，协方差矩阵是以下 3×3 矩阵：

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

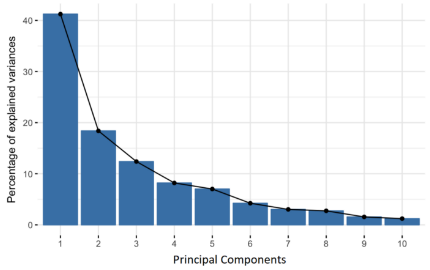
41

主成分

- 主成分是变量的线性组合的新变量。
- 这些组合的完成方式是，新变量（即主成分）不相关，并且变量中的大部分信息被压缩或压缩到第一个分量中。
- 10 维数据会得到 10 个主成分，但 PCA 试图令第一个主成分包含尽可能多的信息，然后令第二个主成分包含剩余信息中尽可能多的信息，依此类推，直到得到下面碎石图所示的东西。

42

主成分



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

43

主成分

- 主成分的可解释性较差，可能没有任何实际意义（它们仅是初始变量的线性组合）。
- 从几何上讲，主成分表示解释最大方差量的数据的方向，即捕获数据大部分信息的线。
- 一条线携带的方差越大，沿它的数据点的离散度就越大，沿一条线的方差越大，它拥有的信息就越多。
- 主成分可以被当作新的空间轴，帮助我们更好地看到观测值间的差异。

44

如何构造主成分

- 主成分的数量可与变量一样多；
- 第一个主成分是能够代表数据中最大可能的方差的成分
- 第二个主成分与第一个主成分不相关（即垂直于），能解释剩下方差中最大的比例。
- 不断地计算，直到得到p 个主成分。

45

沿主成分轴重铸数据

- 具体而言，在计算主成分过程中，使用协方差矩阵的特征向量形成的特征向量，将数据从原始轴重新定向到主分量表示的轴（因此称为主分量分析）。
- 通过将原始数据集的转置乘以特征向量的转置来完成。

46

PCA的本质

- PCA本质上是求出协方差矩阵的特征值和特征向量。
- 目前对协方差矩阵的特征值分解现在是基于奇异值分解进行。
- PCA对主成分进行选择其实就是选择特征值较大的几个特征向量。

47

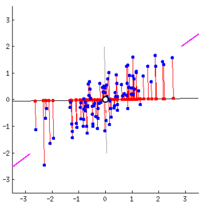
PCA的本质

- 数据降维度会导致信息的损失，所以降维的代价是牺牲精确性。
- PCA降维：舍弃方差较小的主成分，保留方差较大的主成分。

48

PCA的本质

- 紫色的线是最大化方差（从投影点（红点）到原点的平方距离的平均值）的线。



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

49

应用

- PCA可以用来进行图像压缩，通过舍弃信息较小的维度，保留信息较大的维度，从而实现降低图片的大小。

https://blog.csdn.net/ctyqy2015301200079/article/details/85325125

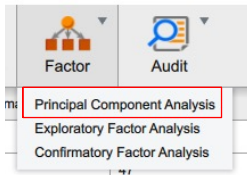
50

其它算法

- 在机器学习中，还有k-Neareast Neighbours(K-近邻)，随机森林等一系列降维算法。

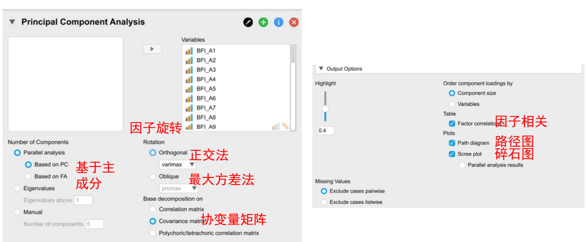
51

功能选择



52

模型设定



53

结果展示

Chi-squared Test ▼

	Value	df	p
Model	1420.108	697	< .001

- 卡方检验的结果表明，PCA是有效的。

54

结果展示

主成分
权重

残差

Component Loadings

	PC1	PC2	PC3	PC4	PC5	PC6	Uniqueness
BFI_N3	0.795						0.439
BFI_C6	0.771						0.363
BFI_N5	0.743						0.395
BFI_N7	−0.707						0.385
BFI_N1	0.647						0.446
BFI_N2	0.582						0.396
BFI_N4	0.426						0.390
BFI_N8	−0.422	0.516					0.415
BFI_O2		0.604					0.348
BFI_O4		0.575					0.458
BFI_O7		0.548					0.381
BFI_O1		0.547					0.371
BFI_E1		0.490	−0.726				0.305
BFI_E4		0.479					0.335
BFI_E2		0.471	−0.543				0.399
BFI_E5		0.468					0.287
BFI_E3		0.428					0.455
BFI_N6		0.403					0.471
BFI_E6			0.837				0.402
BFI_A8			0.758				0.557
BFI_E8			0.706				0.407
BFI_C8				−0.716			0.422
BFI_C9				−0.614			0.526
BFI_C3				0.567			0.469
BFI_C1				0.527			0.386
BFI_C2				0.484			0.469
BFI_C5				0.475			0.435
BFI_A7					−0.612		0.456
BFI_A9					−0.611		0.591
BFI_A6					−0.487		0.630
BFI_A4					0.415		0.386
BFI_O10						0.913	0.405
BFI_O6						−0.640	0.504
BFI_O3						−0.602	0.377
BFI_O9						0.438	0.653
BFI_A1							0.302
BFI_A2							0.500
BFI_A3							0.358
BFI_A5							0.429
BFI_C4							0.330
BFI_C7							0.532
BFI_O5							0.407
BFI_O8							0.429
BFI_E7							0.421

Note. Applied rotation method is varimax.

- 卡方检验的结果表明，PCA是有效的。

55

主成分相关

旋转前　　　旋转后

Component Characteristics

	Unrotated solution			Rotated solution		
	Eigenvalue	Proportion var.	Cumulative	SumSq. Loadings	Proportion var.	Cumulative
Component 1	6.884	0.183	0.183	4.190	0.111	0.111
Component 2	3.877	0.103	0.286	3.985	0.106	0.217
Component 3	2.890	0.077	0.362	3.339	0.089	0.306
Component 4	2.043	0.054	0.417	2.872	0.076	0.382
Component 5	1.739	0.046	0.463	2.251	0.060	0.442
Component 6	1.417	0.038	0.500	2.215	0.059	0.500

56

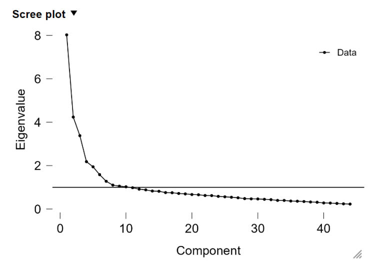
主成分相关

Component Correlations

	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
Component 1	1.000	0.000	0.000	0.000	0.000	0.000
Component 2	0.000	1.000	0.000	0.000	0.000	0.000
Component 3	0.000	0.000	1.000	0.000	0.000	0.000
Component 4	0.000	0.000	0.000	1.000	0.000	0.000
Component 5	0.000	0.000	0.000	0.000	1.000	0.000
Component 6	0.000	0.000	0.000	0.000	0.000	1.000

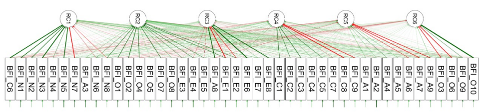
57

碎石图



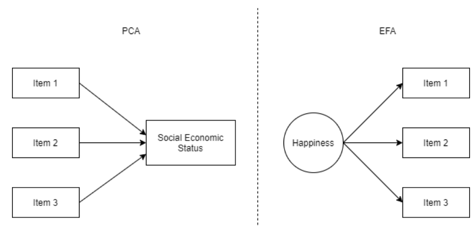
58

路径图



59

PCA与EFA



Formative model, the variation in the component is based on the variation in item responses.

Reflective model, the variation of the items is based on the variation of a construct.

https://towardsdatascience.com/exploratory-factor-analysis-vs-principal-components-from-concept-to-application-b67bbbdb82c4

60

Part 3:
聚类分析

61

聚类分析

- 将分类对象置于多维空间，按亲疏远近分类。
- 以事物间的定量差异为数学基础，定性差异为结果选择的依据。
- 聚类的目标是得到较高的簇内相似度和较低的簇间相似度，使得簇间的距离尽可能大，簇内样本与簇中心的距离尽可能小。

62

聚类分析

- 聚类得到的簇可以用聚类中心、簇大小、簇密度等来表示。
  - 聚类中心是一个簇中所有样本点的均值(质心)
  - 簇大小表示簇中所含样本的数量
  - 簇密度表示簇中样本点的紧密程度

https://blog.csdn.net/weixin_43584807/article/details/105539675

63

聚类分析的依据

- 用于对聚类结果进行评判，分为内部指标和外部指标两大类：
  - 外部指标指用事先指定的聚类模型作为参考来评判聚类结果的好坏
  - 内部指标是指不借助任何外部参考，只用参与聚类的样本评判聚类结果好坏

64

聚类分析的度量

- 内部指标
  - 在聚类分析中，对于两个m维样本$x_i = (x_{i1}, x_{i2}, ..., x_{im})$和$x_j = (x_{j1}, x_{j2}, ..., x_{jm})$。距离的测量包括距离和相似性系数两种类型。常用的距离度量有欧式距离、曼哈顿距离、切比雪夫距离和明可夫斯基距离等。

65

聚类分析的度量

- 欧式距离（Euclidean Distance）是计算欧式空间中两点之间的直线距离

$$dict_{ed} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

66

聚类分析的度量

- 曼哈顿距离（Manhattan Distance）也称绝对值距离，类似于在城市中两个地点之间的实际距离，即沿道路行驶的距离。
- 计算空间中两点各维度指标间差值的绝对值之和。

$$dict_{mand} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$

67

聚类分析的度量

- 切比雪夫距离（Chebyshev Distance），向量空间中的距离度量。
- 将空间坐标中两个点的距离定义为其各坐标数值差绝对值的最大值。

$$dict_{cd} = \max_{k} |x_{ik} - x_{jk}|$$

68

聚类分析的度量

- 测量指标的量纲统一
  - 数据中心化变换
  - 数据标准化变换
  - 极差正规化变换
  - 对数变换

69

聚类的分类

- 基于层次的聚类
- 基于划分的聚类(k-mean算法、k-medoids算法、k-prototype算法)
- 基于密度的聚类(DBSCAN算法、OPTICS算法、DENCLUE算法)
- 基于网格的聚类
- 基于模型的聚类(模糊聚类、Kohonen神经网络聚类)

70

基于层次的聚类

- 层次聚类分析就是逐次聚类分析，将观测样本的每一个个体或每一个变量看作小类，计算两两间距离，比较距离，距离最小的归为一类，不断进行此过程。
- 对个案的层次聚类分析叫Q聚类分析，对变量的层次聚类分析叫R聚类分析。

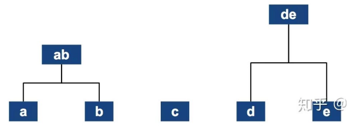
71

基于层次的聚类

- Iteration: Merge the two nearest clusters
- Clusters {d} and {e} are merged since d(d,e) is the lowest distance

Distance matrix

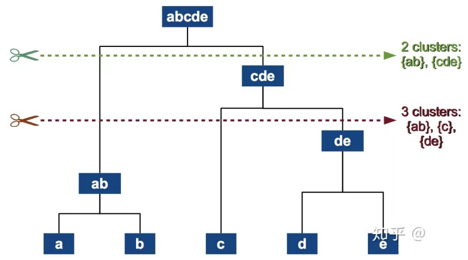
	a	b	c	d	e
a	0.00				
b	0.18	0.00			
c	0.39	0.32	0.00		
d	0.43	0.34	0.25	0.00	
e	0.39	0.41	0.27	0.21	0.00



https://zhuanlan.zhihu.com/p/37856153

72

基于层次的聚类



https://zhuanlan.zhihu.com/p/37856153

73

层次聚类分析

- 步骤：
  - 数据获取
  - 距离计算与逐步凝聚
  - 绘制凝聚状态表，树状图，冰柱图
  - 确定类别数和个体的类属关系

74

层次聚类分析

- 距离计算与逐步凝聚：
- 个案距离计算：
  - 连续变量：欧氏距离，切比雪夫距离，绝对值距离
  - 顺序，等级变量：卡方分析
  - 二分变量：二元欧氏距离平方

75

层次聚类分析

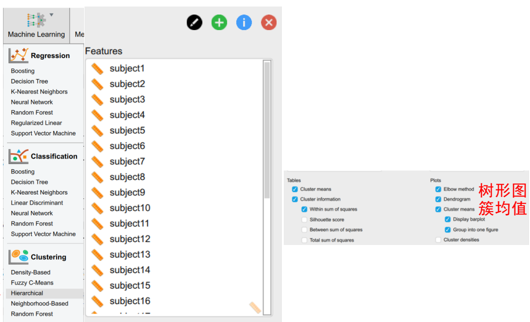
- 距离计算与逐步凝聚：
- 个案与小类，小类与小类距离计算：
  - 最短距离法
  - 最长距离法
  - 类间平均连锁法
  - 重心法

76

JASP操作

- 数据：example25_2.csv

层次聚类分析



77

JASP操作

Hierarchical Clustering ▼

Hierarchical Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
2	44	0.167	1195.190	1302.240	0.200

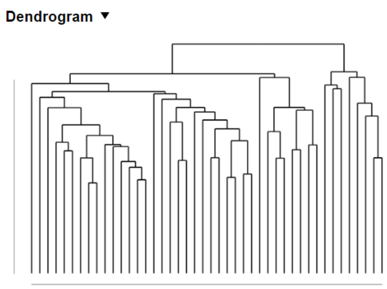
Note. The model is optimized with respect to the *BIC* value.

Cluster Information ▼

Cluster	1	2
Size	36	8
Explained proportion within-cluster heterogeneity	0.821	0.179
Within sum of squares	882.532	192.654

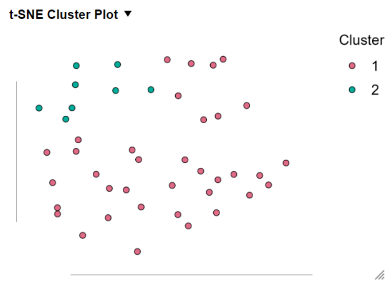
78

JASP操作



79

JASP操作



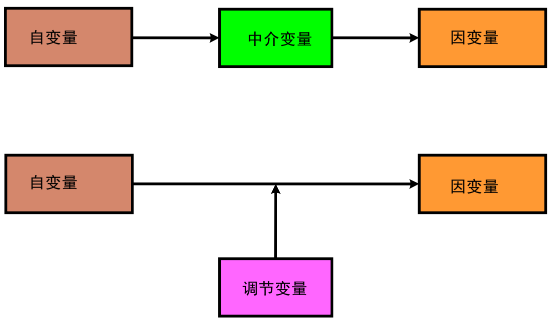
80

总结

·	EFA	PCA	Clusters
原理	变量是因子的线性组合	主成分是变量的线性组合	相似变量归为一类
应用	Reflective model	Formative model	分类
新变量	独立/相关	独立	/
数据预处理	/	标准化	/

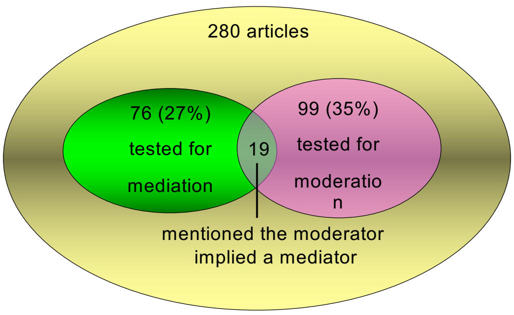
81

调节变量和中介变量



82

调节变量和中介变量



. Incidence of tests of mediation and moderation in the *Journal of Applied Psychology* (volumes 84-86).

83

调节变量的定义(moderating effect)

- 如果变量Y和变量X之间的关系是变量M的函数,则M称为调节变量
- 调节变量可以是定性的(如性别,种族等)
- 也可以是定量的(如年龄,刺激次数等)
- 它影响因变量和自变量关系的方向和强弱.

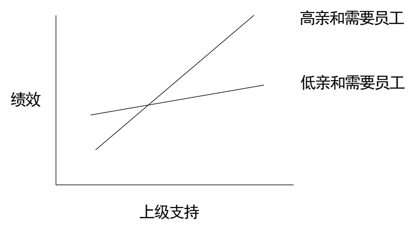
84

研究情境

- 也称为边际条件 “***boundary condition***.”
- 基于上述发现:
  - 不一致的发现.
  - 阻碍的条件.
  - 促进的条件.

85

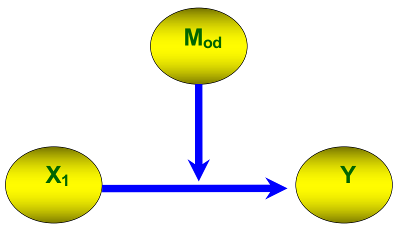
调节变量的例子（1）



86

调节变量亦即交互作用

- 在模型中有特殊的表达方法



87

如何检验调节效应

- 取决于变量的类型
- 如果自变量是二分变量，调节变量也是二分变量
  - 2x2 ANOVA

88

二分变量的调节效应

- 例：奖励食物数量影响动物的作业水平,内驱力是调节变量

89

自变量或调节变量是连续变量

- 比较复杂
- 必须假定没有测量误差
  - 效力较低
  - 必须有较大的样本容量
- 不要将连续变量分成二分变量
  - 有些做法从均值或中数分成2组，这样作降低了检验的效力

90

连续变量的调节效应

- 中心化或标准化自变量和调节变量
  - 中心化为了降低共线性
  - 标准化容易用SPSS、JASP、MATLAB或R程序计算

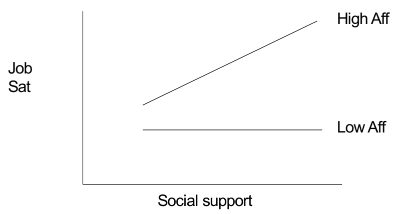
91

连续变量的调节效应

- 用两变量的乘积作为交互作用项
  - 在SPSS、JASP、MATLAB或R中建立一个新变量
- 进行层次化多元回归
  - 第一层：主效应
  - 第二层：加入交互作用项

92

一个例子



93

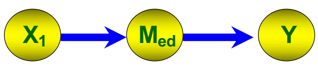
连续变量的调节效应的完整程序

1. 标准化自变量和调节变量
2. 建立交互作用变量
3. 层次化多元回归
4. 绘制和解释交互作用项

94

中介变量的定义(mediating effect)

- 考虑自变量*X* 对因变量*Y*的影响,如果*X* 通过影响变量*M* 来影响*Y*,则称*M* 为中介变量。
- 中介变量的意义是解释自变量对因变量的影响机制
- 在模型中有特殊的表达方法



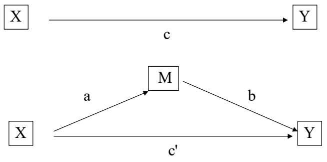
95

研究情境

- 当X 和Y 的关系给定后,目的是了解”为什么”的机制.
- 递进的因果关系

96

分析过程



c: X 和 Y的直接关系
a: X和中介变量的关系
b: 中介变量和Y的关系
c': 当中介变量在方程中时X 和Y的关系

97

要注意的问题

- 要注意的问题
  - 整个模型的效力
  - 中介变量的信度
- 这些问题都可以通过结构方程模型有效地解决

98

传统方法如何检验中介作用

- 考察自变量和因变量间的关系
  - 相关或回归
  - 求出"c"
- 将自变量(IV) 向中介变量(DV)回归
  - 标准化系数 B 就是 "a"

99

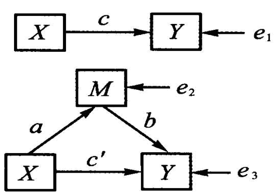
如何检验中介作用

- 进行层次化回归
  - 第一层: 放入 X
  - 第二层: 放入中介变量(得到 b)
- 变量X的β 降低的数量(c') 就是中介作用的大小
  - 完全中介, β 变为 0 (n.s.)
  - 部分中介, β显著减小

100

完全中介

- 如果变量X对Y的作用完全被中介变量M引起的a和b路径所分解,则称如果变量X对Y的作用被变量M完全中介.

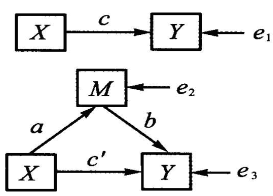


c≠0, c'=0

101

部分中介

- 如果变量X对Y的作用完全被中介变量M引起的a和b路径显著, 而X至Y的路径仍显著,则称如果变量X对Y的作用被变量M部分中介.



c≠0, c' ≠ 0

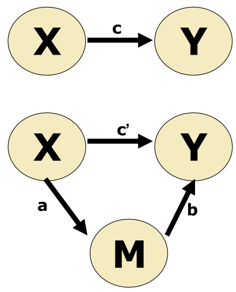
102

测量间接效应

- 间接效应即中介作用的大小
  - 完全中介或部分中介 间接效应=(c - c')
  - 理论上, c - c' = a*b
- 但是，这些对于证实中介作用是不够的，需要有一个统计检验.

103

测量间接效应



间接效应= a × b

如果 a × b = c,
完全中介模型

如果 a × b < c,
部分中介模型

如果 a × b > c,
抑制模型

104

基于自抽样程序的中介效应检验

- 所谓自抽样程序（bootsrap），是以样本（如，样本大小n=100）来代表总体，在此样本中进行放回抽样（已被抽取的个体允许被再度抽取），直至抽取n个（如100个），组成一个样本。
- 这样的程序反复进行多次（k次），亦即产生多个样本，每个样本都可以算出一个间接作用估计值，由此可以算出k个值，形成一个实际的分布。这个分布近似于从原始总体中取样的分布。
- 一般建议最少抽样1000次（亦即k=1000），推荐抽样5000次（Preacher & Hayes, 2008）。
- 中介作用真值的统计推论根据这k个估计值产生的ci%置信区间得到。
- 如果置信区间不包括0，那么中介作用显著，支持中介作用的假设；如果包括0，则不显著，不支持中介作用的假设

105

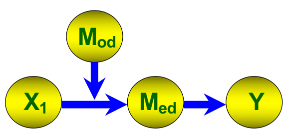
调节变量和中介变量的比较

	调节变量 M	中介变量 M
研究目的	X何时影响 Y或何时影响较大	X如何影响 Y
关联概念	调节效应、交互效应	中介效应、间接效应
什么情况下考虑	X对 Y的影响时强时弱	X对 Y的影响较强且稳定
典型模型	$Y = aM + bM + cXM + e$	$M = aX + e_2$ $Y = c'X + bM + e_3$
模型中 M 的位置	X, M在 Y前面, M可以在 X前面	M在 X之后、Y之前
M的功能	影响 Y和 X之间关系的方向 (正或负)和强弱	代表一种机制, X通过它影响 Y
M与 X、Y的关系	M与 X、Y的相关可以显著或不显著 (后者较理想)	M与 X、Y的相关都显著
效应	回归系数 c	回归系数乘积 ab
效应估计	$\hat{c}$	$\hat{a}\hat{b}$
效应检验	c是否等于零	ab是否等于零
检验策略	做层次回归分析, 检验偏回归系数 c的显著性 (t检验); 或者检验测定系数的变化 (F检验)	做依次检验, 必要时做 Sobel检验[18]

106

更复杂的模型(1)

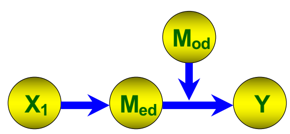
- 有中介的调节模型:
  - A moderating effect is transmitted through a mediator variable (Baron & Kenny, 1986)



107

更复杂的模型(2)

- 有调节的中介模型:
  - A mediating effect is through to be moderated by some variable (Baron & Kenny, 1986)



108