中级应用统计学

结构方程模型

张晨峰

华东理工大学商学院

2015年11月12日

9 结构方程模型

主要内容

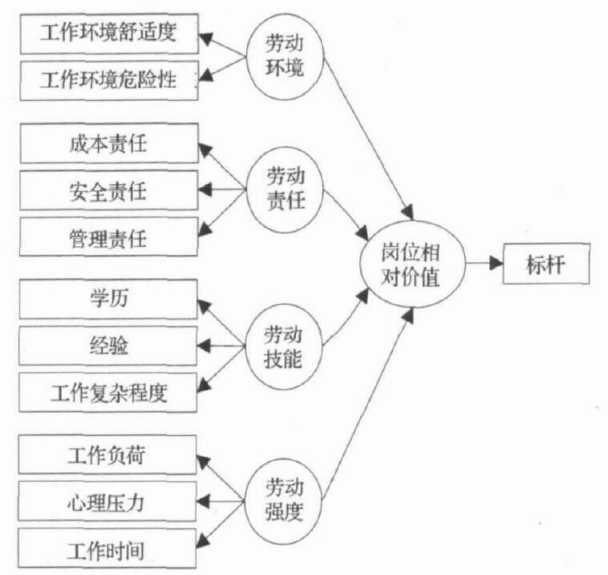
- 结构方程模型简介
- 结构方程模型的结构及模型设定
- 结构方程模型的识别和估计
- 结构方程模型的评价和修正
- 结构方程模型的应用

9.1 结构方程模型简介

表 1 岗位评价要素表

要素	子要素
劳动技能	学历 经验 工作复杂程度
劳动责任	成本责任 安全责任 管理责任
劳动强度	工作负荷 心理压力 工作时间
劳动环境	工作环境舒适度 工作环境危险性

9.1 结构方程模型简介



工作环境舒适度
工作环境危险性
劳动环境
成本责任
安全责任
管理责任
劳动责任
学历
经验
工作复杂程度
劳动技能
工作负荷
心理压力
工作时间
劳动强度
岗位相对价值
标杆

结构方程模型的概念

近年来，结构方程模型（structural equation modeling，SEM）作为统计分析的一般框架，被广泛地应用于社会科学的数据分析。**结构方程模型是基于变量的协方差矩阵来分析变量之间关系的一种统计方法，所以也称为协方差结构分析。**

9.1 结构方程模型简介

结构方程模型的基本思想

结构方程模型在估计一组**观察变量（observed variables）**与其代表的**潜变量（latent variables）**（或因子（factors））的关系的同时，分析各潜变量之间的关系。这样，潜变量之间的关系估计便不受**测量误差（measurement errors）**的影响。结构方程模型源于**因子分析（factor analysis）**和**路径分析（path analysis）**（或联立方程（simultaneous equations））。基于因子分析的测量模型（measurement model）与基于路径分析的结构模型（structural model）的整合，形成了一个数据分析的一般框架，叫做结构方程模型。

结构方程模型的技术特性

- **SEM具有理性先验性：**SEM被视为一种验证性（confirmatory）而非探索性（exploratory）的统计方法
- **SEM同时处理测量与分析问题：**主要的关键在于SEM将不可直接观察的概念，以潜变量的形式，利用观察变量的模型化分析加以估计
- **SEM以协方差的运用为核心：**协方差有两种功能：其一是描述性的功能；其二是验证性的功能，用以反映出理论模型所导出的协方差与实际观测得到的协方差的差异
- **SEM适用于大样本的分析**
- **SEM包含了许多不同的统计技术**

结构方程模型的优点

- **SEM能同时处理多个因变量**，而回归分析中，只能处理一个因变量
- **SEM容许自变量和因变量均包含测量误差**，而回归分析只允许因变量存在测量误差，假定自变量没有误差
- **SEM同时估计因子结构和因子关系**，因子间关系与因子的内部结构是可以互相影响
- **SEM容许更大弹性的测量模型**，相比较而言，传统因子分析难以处理一个指标从属多个因子，或则考虑高阶因子等有比较复杂的从属关系的模型
- **SEM估计整个模型的拟合程度**，相比较而言，传统的路径分析，只能估计每一条路径（变量间关系）的强弱

结构方程模型的建模步骤

- **模型设定**，根据理论研究或实践经验构建实证方程模型
- **模型识别**，识别设定模型的参数估计是否有唯一解
- **模型估计**，最常用的是最大似然估计法（MLE）
- **模型评估**，评估模型是否拟合数据
- **模型修正**，如果模型与数据拟合不好，则需要重新设定或修改模型

9.2 结构方程模型的结构及模型设定

潜变量

从统计的角度，潜变量（latent variables）是不能被直接观察，但可以由观察变量（observable variables）推断出来的变量，例如幸福感，道德感，生活质量等。

内生变量与外生变量

内生变量（endogenous variables）是指模型中，会受到其他变量影响的变量。而外生变量（exogenous variables）是指模型中不受任何其他变量影响，但会影响其他变量的变量。

9.2 结构方程模型的结构及模型设定

结构方程模型设定

模型设定最直接的方法是通过Wright（1934）提出的**路径图**（path diagram）来描述研究者感兴趣的模型。路径图是结构方程模型的基础，它可以清晰地表达研究人员对于变量间关系的方程，并可直接转换成建模所需要的方程。

结构方程模型结构

一般结构方程模型的组成：

- 测量模型（measurement model）
- 结构模型（structural model）

路径图的标示规则

- **长方形**：观测变量
- **椭圆**：潜变量（因子）
- **线条**：变量直接的关系用线条表示，如果两个变量之间没有线条相连，则表示两者间没有直接关系。**单向箭头**表示有方向性的关系（因果关系），**双向箭头**表示变量间的关联，即相互影响。

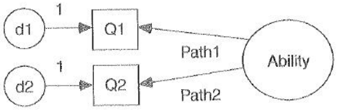
9.2 结构方程模型的结构及模型设定

Where: "Ability" is a latent variable measured by Q1 and Q2

Q1 and Q2 are observed variables (items on a questionnaire)
d1 and d2 are error terms

Given: Correlation *r* between Q1 and Q2 = 0.64

Influence of "ability" on Q1 and Q2 is the same



Results: Both path coefficients must be 0.80 (product of the paths = 0.64).

64% of Q1 is explained by "ability" (0.80^2).
64% of Q2 is explained by "ability" (0.80^2).
36% of Q1 must be explained by d1 (100% minus 64%).
36% of Q2 must be explained by d2 (100% minus 64%).

Figure 2.1 Calculating the Relationships of Observed Variables to a Latent Variable.

9.2 结构方程模型的结构及模型设定

Box 2-1 Proof of an SEM Property and a First Peek at SEM Notation

In Spearman's original work, he claimed that observed intercorrelations among scores on tests of different types of mental ability could be accounted for by a general underlying ability factor. Using our current example, we can imagine that the general ability factor affecting all test scores is the latent variable "ability." Scores on Q1 and Q2 in this example represent observed scores on two mental ability subtests. Variance in Q1 and Q2 that is not explained by "ability" is captured in d1 and d2, respectively. Denoting the two path coefficients (now called factor loadings) as λ_1 and λ_2 (lambda 1 and lambda 2), Spearman proved that the observed correlation between Q1 and Q2 (i.e., ρ_{12}) equals the product of the two factor loadings λ_1 and λ_2, or $\rho_{12} = \lambda_1\lambda_2$, or $0.64 = 0.80 * 0.80$. To prove this, we first express our model of Figure 2.1 in the following equations:

$$Q1 = \lambda_1 \text{Ability} + d1$$

$$Q2 = \lambda_2 \text{Ability} + d2.$$

Assuming we work with standardized scores for all variables, then the correlation ρ_{12} is simply the covariance of Q1 and Q2, or $\rho_{12} = \text{Cov}(Q1, Q2)$. Using the algebra of expectations, we can further write

$$\begin{aligned}\text{Cov}(Q1,Q2) &= E\left[(\lambda_1\text{Ability}+d1)(\lambda_2\text{Ability}+d2)\right]\\ &= E\left[\lambda_1\lambda_2\text{Ability}^2+\lambda_1\text{Abilityd2}+\lambda_2\text{Abilityd1}+d1d2\right]\\ &= \lambda_1\lambda_2E(\text{Ability}^2)+\lambda_1E(\text{Abilityd2})+\lambda_2E(\text{Abilityd1})+E(d1d2).\end{aligned}$$

Because E(Abilityd2) = 0 and E(Abilityd1) = 0 (because there is no correlation between the common factor Ability and each error), and E(d1d2 = 0) (because the two measurement errors are not correlated), then the equation becomes $\rho_{12} = \lambda_1\lambda_2\, E(\text{Ability}^2)$. Because $E(\text{Ability}^2)$ is Variance(Ability) and equals 1 (because Ability is a standardized score), then $\rho_{12} = \lambda_1\lambda_2$. That is, the observed correlation between two variables is a product of two path coefficients.

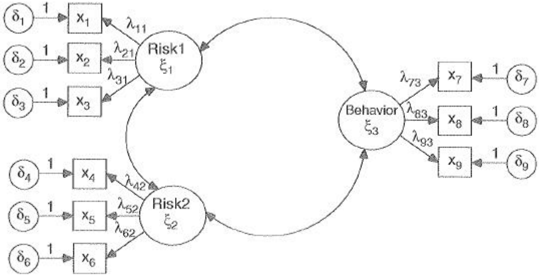


Figure 2.2 Measurement Model.

其中x是变量，δ是误差项，ξ是潜变量，λ是因子载荷。

测量模型

测量模型的基本目的是描述观察变量是否适合作为潜变量的测量手段。测量模型由验证性因子分析来完成和评估。测量模型的一部分：

$$x_1 = \lambda_{11}\xi_1 + \delta_1$$

$$x_2 = \lambda_{21}\xi_1 + \delta_2$$

$$x_3 = \lambda_{31}\xi_1 + \delta_3$$

结构模型

确定了测量模型中的潜变量之后，就可在结构模型中评估潜变量之间的相互关系。注意，如果结构模型中的变量都是观察变量而不是潜变量，那么结构方程就会变成一组观测变量之间结构关系的建模体系。这样，模型就简化为传统的社会学中的路径分析（path analysis）或计量经济学中的联立方程（simultaneous equations）。

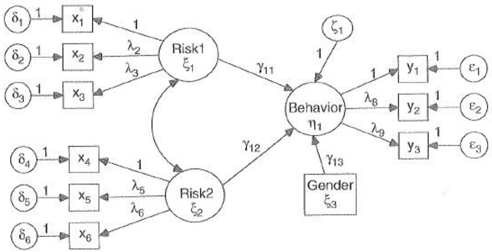


Figure 2.4 General Structural Model 1: Direct Effects Only.

上述的结构模型

$$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \gamma_{13}\xi_3 + \zeta_1$$

其中x是外生变量，y是内生变量，δ是误差项，ξ是外生潜变量，η是内生潜变量，λ是因子载荷，γ是路径系数，ζ是结构误差。

Box 2-3 Two Types of Error in SEM

In the discussion of measurement models starting on p. 20, we defined measurement error as "unique" and "residual" variation in scores of observed indicators that were not associated with the hypothesized factor model. An additional type of error is relevant to structural models and should not be confused with measurement error. SEM structural models, like other regression models, include *structural errors*. The *structural error* for any dependent variable in a structural model is the variance of the variable that is not explained by its predictor variables. Although the latent risk and behavior variables in Figure 2.4 are theoretically free of *measurement* error, we do not expect the risk and gender variables to predict Behavior perfectly. In other words, we do not expect 100% of the variance of Behavior to be explained by the two risk variables and gender. In a general structural model, any variable that is regressed on others in the model has an error term representing the *structural error* (this error can also be thought of as the "error of prediction"). The latent variable ζ_1 represents the error in our structural model—the variation in behavior scores that is not explained by Risk1, Risk2, and Gender.

结构方程模型的结构

测量模型：

$$x = \Lambda_x \xi + \delta$$

$$y = \Lambda_y \eta + \varepsilon$$

结构模型

$$\eta = B\eta + \Gamma\xi + \zeta$$

其中Λ_x和Λ_y是因子载荷矩阵，B和Γ是结构系数矩阵。

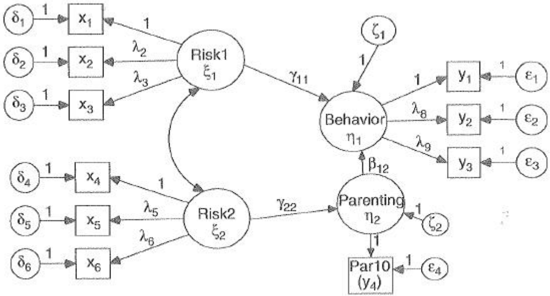


Figure 2.5 General Structural Equation Model 2: Direct and Indirect Effects.

9.2 结构方程模型的结构及模型设定

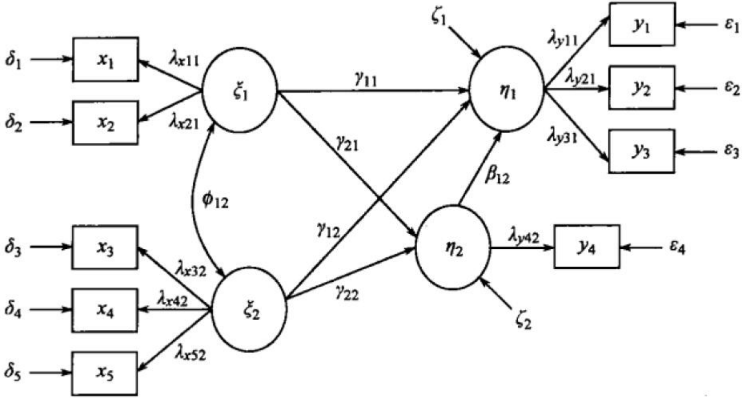


图 1.1-1 假设的结构方程模型路径图

9.2 结构方程模型的结构及模型设定

结构方程模型的基础矩阵

在结构方程模型中，共有8个基础矩阵。一个结构方程模型可由这8个参数矩阵的矩阵格式设定。虽然现在的SEM软件多不用矩阵格式设定模型，但在许多软件的输出结果中，仍然报告这8个参数矩阵的参数估计值信息。这些矩阵有助于研究者深入了解结构方程模型。

表 1.1.3-2　一般结构方程模型的 8 个基本参数矩阵

矩阵	定义	维度
系数矩阵		
Λ_y (lambda y)	y 与 η 之间的因子载荷	$p \times m$
Λ_x (lambda x)	x 与 ξ 之间的因子载荷	$q \times n$
B (beta)	η 与 η 之间的系数矩阵	$m \times m$
Γ (gamma)	ξ 与 η 之间的系数矩阵	$m \times n$
方差/协方差矩阵		
Φ(phi)	ξ 的方差/协方差矩阵	$n \times n$
Ψ(psi)	ζ 的方差/协方差矩阵	$m \times m$
Θ_ε(theta-epsilon)	ε 的方差/协方差矩阵	$p \times p$
Θ_δ(theta-delta)	δ 的方差/协方差矩阵	$q \times q$

注: p 是 y 的变量数; q 是 x 的变量数; m 是 ξ 的变量数; n 是 η 的变量数。

9.3 结构方程模型的识别和估计

识别问题

例如，对于一个二元一次方程式

$$X + Y = 10$$

符合该方程式条件的解有无限多种可能。此时，就是所谓无法识别或欠识别的状态。除非我们提出第二个方程式

$$X + 2Y = 20$$

此时，即能够求出唯一解$(0, 10)$。此时，就是所谓充分识别或恰好识别的状态。

识别问题

如果我们再提出第三个方程式

$$X + 3Y = 40$$

我们可以利用估计方法，求出符合这三个方程式的最优解。此时，就是所谓的过度识别的状态。

结构方程模型的识别

目前尚无公认的SEM模型识别的充要条件，通常需要检查以下两个必要条件：

- 数据点的数量（测量数据数，DP）一定不能少于自由参数的数量，即自由度不能为负数
- 模型中的每一个潜变量都必须设立一个测量尺度。一般有以下两种方式：其一，将一个观测变量的因子载荷固定为一个常数，通常为1；其二，将潜变量的方差固定为1（即将潜变量标准化）。

数据点的数量

数据点数即观察变量的方差/协方差矩阵的不同元素的数量。

$$DP = \frac{(p+q)(p+q+1)}{2}$$

假设模型中有10个测量变量，总计可以产生10个方差与45个协方差，共计为55个数据点数。

自由参数

模型要估计的参数

t法则

Bollen（1989）提出了一个衡量识别性的必要但非充分的识别条件计算法则——t法则，t值代表模型中的自由估计参数数量。t法则的判断原则如下：

- 当$t < DP$时，为过度识别
- 当$t = DP$时，为充分识别
- 当$t > DP$时，为识别不足，这种情况将导致无法进行任何参数估计

识别问题的解决方法

通常，我们可以增加一些潜变量的观察变量，从而获得**更多的数据点**。另外，预防模型识别问题发生的**重点在于参数设定**。模型识别取决于如何将参数设定为自由参数、固定参数或强制参数。通过固定或强制某些参数，可减少自由参数的数量。一般来说，在模型构建初始，自由参数数量应尽可能少，以构建**简约模型**。如果该模型能够被识别，则可随后逐步增加其他感兴趣的参数，最后通过比较所有的替代模型，以选出最恰当的模型。

结构方程模型的估计

结构方程模型估计与多元回归不同，它不是极小化因变量拟合值与观察值之间的差异，而是**极小化样本方差/协方差与模型估计的方差/协方差之间的差异**。结构方程模型分析的基本假设是

$$\Sigma = \Sigma(\theta)$$

其中$\Sigma(\theta)$为模型估计的方差/协方差矩阵。模型估计的目标是找到一组模型参数θ，并使$\Sigma - \Sigma(\theta)$最小化。

结构方程模型的估计

因为Σ和$\Sigma(\theta)$均未知，因此实际上是最小化$S-\hat{\Sigma}$。模型估计过程须应用一种特殊拟合函数以尽可能地减少$S-\hat{\Sigma}$的差异。最常见的是似然函数

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

F_{ML}是对该差异函数的测量，称为最小差异函数。如果一个模型与数据完全拟合，其最小差异函数为零。除了ML方法，还有常见的估计方法包括：无加权最小二乘法（ULS），广义最小二乘法（GLS）等。

结构方程模型的评价

结构方程模型分析的一个重要特征是在零假设$\Sigma = \Sigma(\theta)$的基础上，对整体模型拟合进行检验。多种方法可用于评估S和$\hat{\Sigma}$之间的接近程度，因此有多种模型拟合指数。

- **绝对拟合指数**：评价模型估计的方差/协方差矩阵与观察方差/协方差矩阵之间的差异程度，例如模型卡方统计，拟合优度指数（GFI）等
- **相对拟合指数**：通过比较设定模型与基准模型（数据拟合最差的模型），检查模型拟合相对基准模型的改善比例，例如规范拟合指数（NFI）增值你和指数（IFI）等
- **信息标准指数**：从理论信息的角度评价模型拟合状况，例如AIC和BIC等。

模型卡方值χ^2

模型卡方是最早用于结构方程模型的拟合指数，定义为：

$$\chi^2 = F_{ML}(N-1)$$

此处我们希望卡方检验呈非显著性，即卡方值越小，模型拟合数据越好。卡方检验的明显不足在于：其一，卡方值定义为$N-1$倍的拟合函数，因此对样本量高度敏感；其二，如果样本量不是足够大，拟合函数可能不服从χ^2分布；其三，χ^2对变量是否为多元正态分布非常敏感；其四，χ^2随模型中变量数量的增加而增加。

模型比较

在结构方程模型分析中，不应仅考虑单一模型，而应考虑不同备选模型，以便通过模型比较来选出最好的模型。

- 嵌套模型：常用似然比（LR）检验
- 非嵌套模型：常用信息准则，例如AIC，BIC等

9.4 结构方程模型的评价和修正

模型设定探索

初始模型中常有设定不当之处，因而需要寻找模型拟合不良的可能原因，确定导致模型设定错误的因素，然后修正模型并用同一数据进行检验。该过程称为“模型设定探索”。

修正指数

为了改善数据拟合不良的初始模型，常使用**修正指数（modification indices，ML）**作为诊断指标来帮助修改模型设定。修正指数与模型的固定参数联系在一起。一个固定参数的MI值相当于自由度$df = 1$的模型卡方值，就是说，如果将模型中某个受限制的固定参数改为自由参数，则模型卡方值将减少，减少值相当于为该参数的MI估计值。

9.4 结构方程模型的评价和修正

模型残差

检查模型拟合质量的另一重要方法是**检查模型残差**。结构方程模型分析中的残差是残差矩阵$S-\hat{\Sigma}$中的元素。残差取决于观察变量的度量单位，因此常将残差标准化。一个较大的标准化残差表示某个具体的方差或协方差在S与$\hat{\Sigma}$间的差异较大。一般来说，如果一个标准化残差大于2.58，可以认为该标准化残差较大。

模型修正

必须强调，修正模型应是“数据驱动（data-driven）”和“理论驱动（theory-driven）”相结合。我们不能以改善模型拟合度为目的而单纯地增加或删除参数。虽然我们的目的是找到一个统计上对实际数据拟合良好的模型，但**更重要的是模型的所有参数估计都要有实际意义**。