

结构方程模型拟合评鉴: 整体拟合、内部拟合与复核效度检验*

郑文智 吴文毅

(华侨大学工商管理学院, 泉州 362021)

摘 要: 该文主要介绍了评价和鉴定结构方程中构建的模型与数据拟合程度的三个方面: 整体拟合、内部拟合及复核效度检验。整体拟合评鉴主要是结合拟合指数, 当拟合指数处于临界值时, 应同时参照其他检验结果并根据模型构建的理论依据进行综合判断; 内部拟合从项目质量检验、测验信度、平均变异萃取量和效度四个角度进行检验; 复核效度检验在多样本分析中可分为宽松复核、温和复核和严格复核, 但如果为研究样本所限, 可以通过 ECVI 来实现。

关键词: 结构方程模型; 整体拟合; 内部拟合; 复核效度检验

中图分类号: B841.2

文献标识码: A

文章编号: 1003-5184(2014)01-0057-05

随着越来越多的人使用结构方程模型(Structural Equation Modeling, SEM)作为社会科学实证研究的测量统计方法, 如何评价所构建的模型与数据拟合匹配程度这一问题显得越来越突出。然而, 到目前为止, 学术界没有一个可供参考的“金标准”。评鉴一个结构方程模型拟合的好坏可以从三个方面进行: 整体拟合、内部拟合及复核效度检验。从这三个方面可以非常系统地评鉴结构方程中所构建模型的拟合程度。

1 模型的整体拟合

拟合指数是对模型整体拟合好坏的度量, 是评判模型整体拟合优度的主要指标, 但从结构方程的运行来看, 假设模型的整体检验是在完成参数估计之后才进行的工作。因而, 评估模型的第一步, 应是检验结构方程的假设模型及参数能否被顺利识别, 收敛并估计。然后借由模型参数的质量与适当性检验, 对于假设模型的拟合进行初步评判。

具体来说, 在做整体拟合评判之前, 研究者应先检视其他四个方面: 一是看参数估计值, 如参数显著性是否达到要求; 二是看每个方程的 R^2 , 由此观察潜变量和各测量指标之间的聚敛能力; 三是检查残差矩阵, 进而可知再生矩阵和原始矩阵的差异程度; 四是看修正指数。通过从这四个方面的局部检验就能初步发现错误模型了。

该使用哪些拟合指数及选取多大的临界值来评价模型的整体拟合, 目前国内外还没有统一的标准。问题主要集中在两个方面, 第一, 哪些指数能更好的衡量模型拟合的好坏? 这个问题, 从不同的角度分析

会有不同的结果, 这也是指数家族越来越庞大的主要原因。一个好的指数应当具有以下特征: (1) 不受样本容量系统的影响或者影响非常小; (2) 惩罚复杂的模型; (3) 对误设模型敏感(温忠麟, 侯杰泰, 马什赫伯特, 2004)。第二, 作为判断标准的拟合指数的界值取多大合适? 在相当长一段时间里, 比较公认的标准是: 相对指数在 0.90 或以上, 模型的拟合结果可以接受; RMSEA 小于 0.05 表示模型拟合得好, 在 0.05 ~ 0.08 之间表示模型基本可以接受。

在这方面 Hu 和 Bentler(1998, 1999) 所做的研究最具影响力, 他们建议使用七个指数: TLI(NNFI)、BL89(IFI)、RNI(CFI)、Gamma Hat、Mc、SRMR 和 RMSEA; 并建议相应的界值为: TLI(NNFI)、BL89、RNI(CFI)、Gamma Hat 为 0.95, Mc 为 0.9, SRMR 为 0.08, RMSEA 为 0.06; 同时推荐用 2-指数准则检验模型的拟合(即同时检验 SRMR 和其他六个指数中的一个, 界值和单个指数的相同)。

国内有许多学者都围绕 Hu 和 Bentler 所做的研究进行了分析。温忠麟等人(2004)对 Hu 和 Bentler 的临界值标准提出了质疑, 建议使用传统的临界值(NNFI 和 CFI 界值为 0.9; RMSEA 界值为 0.08), 同时推荐他们的卡方准则。然而同一组作者后续发表的文章(Marsh, Hau, & Wen, 2004)又认为根据模型研究得到的卡方准则有欠妥当, 不再使用, 并继续坚持使用传统的临界值。而郭庆科等人(2008)仿照 Hu 和 Bentler 的研究做法, 利用模拟研究的方法得出新的拟合指数临界值标准, 推出 2 界值策略, 同时认为应当根据样本量的大小和是否正态分布采用不

* 基金项目: 华侨大学基本科研业务费专项基金(JB-SK1124)。

通讯作者: 郑文智, E-mail: zwz@hqu.edu.cn。

同的界值判断标准。温忠麟和侯杰泰(2008)对郭庆科等人(2008)的做法提出了质疑,他们用简单的Z检验说明了检验的临界值是不能通过模拟研究来

确定的。有趣的是,这两篇文章同时发表在《心理学报》上,而且是紧挨着的。

以下是用于评鉴模型整体拟合常用指数的说明表。

表1 常用拟合指数的说明

拟合指数	判定标准	数据非正态时能否很好估计	处理不同大小样本是否稳定	评估模型的简约性	说明
χ^2	$p > 0.05$	否	否	否	多组模型比较时特别有用
χ^2/df	一般要求介于1~2之间	否	否	否	多组模型比较时特别有用
RMSEA	< 0.08 < 0.05	不清楚	否	是	测量模型的绝对拟合,对参数多的模型加以惩罚
NFI	> 0.90 > 0.95	一般低估	是	否	不同模型评价时精确稳定,比较嵌套模型特别有用
NNFI	> 0.90 > 0.95	一般低估	不清楚	否	最大似然法估计时较好,最小二乘较差,可以比较嵌套模型
CFI	> 0.90	一般低估	不清楚	否	最大似然法估计时较好,最小二乘较差,可以比较嵌套模型

(资料来源:易丹辉,2006,p.127)

需要指出的是, χ^2 会非常敏感地拒绝参数过少的模型,而鼓励参数过多的模型,同时 χ^2 检验往往无法检测出错误的设定,倾向于接纳比较复杂的模型,加上 χ^2/df 容易受到样本量大小的影响,对于评价单个模型的拟合意义不大,更多的是用在模型间的比较(侯杰泰,温忠麟,成子娟,2008)。另外,有研究显示,在进行小样本分析时,RMSEA有被高估的现象,使模型拟合看起来不太理想(Bentler & Yuan,1999),而此时用CFI来评估模型却十分稳定(Bentler,1995),故建议在评价小样本数据的整体拟合时尽量选用CFI,避开RMSEA。

Yuan(2005)研究了不同条件(分布情况和样本量)对拟合指数的影响,发现所有的拟合统计量都或大或小地受到样本量变化的影响。因而Yuan认为拟合指数的临界值标准是难以确定的,但它们仍旧具有一定的意义,可以用来判断模型的相对拟合。实际上,结构方程模型拟合的判断确实是个非常复杂的问题,样本量不同、待估参数数量不同时拟合指数的界值往往也有所不同(郭庆科等,2007)。因此,在具体条件下要具体分析,并没有在所有条件下都普遍适用的“金标准”。就目前而言,学者们更倾向于使用传统的临界值。当拟合指数在临界值附近时,不能根据拟合指数一锤定音,应综合多种方法检验并修正模型,同时结合模型构建的理论背景和可解释程度,对模型的整体拟合进行综合评定。

2 模型的内部拟合

模型的内部拟合指从模型的内在质量评价每一个潜变量的适切程度。具体操作上,研究者可以通过项目质量检验、测验信度、平均变异萃取量和效度来衡量模型的内部拟合。

2.1 项目质量检验

Bollen(1989)指出,建构一个有意义的潜变量最重要的前提,是组成能够反映潜变量的测量指标。

从项目来看,潜变量和测量指标之间必须有良好的聚敛关系。因子载荷一方面能反映测量误差的影响,同时也显示个别题目能够用来反映潜变量的程度。Hair等人(2006)认为一个足够大的因子载荷,能反映题项具有较好的聚敛效度,也即该题项的测量指标能够很好地聚敛起来测量潜变量。一般而言,当因子载荷(完全标准化解,也写为 λ)大于0.71时,即显示项目具有理想的质量,因为此时潜变量能够解释测量指标将近50%的变异量, λ 的大小可以看成是基本拟合中最为明确的一个判断。

不过,社会科学研究中的因子载荷都不会太高,此时建议采用Tabachnick和Fidell(2007)所建议的标准($\lambda \geq 0.55$ 便可宣称良好)。表2显示了因子载荷反映项目质量的判断标准。

表2 因子载荷的判断标准

λ	λ^2	状况
0.71	50%	优秀
0.63	40%	非常好
0.55	30%	好
0.45	20%	普通
0.32	10%	不好
0.32 以下	10% 以下	不及格

2.2 测验信度

测验信度(reliability)是衡量测验质量的重要指标。自从Cronbach(1951)提出 α 系数开始, α 系数几乎成了测验信度的化身。 α 系数被疯狂推崇的同时,文献上对 α 系数的质疑和评论也持续了很长的时间,目前有大量研究认为, α 系数不能很好地估计测验信度。Fornell和Larker(1981)提出了一个非常类似于内部一致性信度的潜变量合成信度(Composite Reliability,CR或 ρ_c),有时也被称为组合信度。

$$CR = \rho_c = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \Theta_{ii}} \quad \text{式1}$$

式中 λ_i 为各题项完全标准化因子载荷, Θ_{ii} 为测量误差。

在结构方程出现以前,由于缺乏对路径系数的有效计算和检验,许多学者使用 α 系数作为测验信度。然而随着结构方程的出现及合成信度的提出,研究者更倾向于使用克服了 α 系数诸多缺陷的合成信度。事实上,应用结构方程模型不仅可以算出潜变量的合成信度,根据 Lisrel 软件给出针对各个指标题项的个别项目信度,研究者还可将之作为选择量表题项的直接依据(徐万里,2008)。

合成信度是测验信度的下限,如果合成信度高到可以接受,报告合成信度并说明:测验信度不低于合成信度;如果合成信度过低,虽然不能说测验信度也一样低,但就目前所拥有的研究方法而言,只能认为测验的信度不能被接受,停止进一步的统计分析(温忠麟,叶宝娟,2011)。

那么,合成信度要多大才行呢?依据古典测验理论的观点,测验信度需达 0.70 才属于比较稳定的测量,但此标准的达成必须使各题项的完全标准化因子载荷(λ) 平均达 0.70 以上,而社会科学领域的量表难以达到这个要求,因此 Bagozzi 和 Yi(1988) 建议要求合成信度达到 0.60 即可。而 Raine - Eudy(2000) 的研究指出,当合成信度达到 0.50 时,所使用的测量工具可以基本稳定地反映真分数。目前相当多的文献也认为信度在 0.50 以上是可以接受的;低于 0.50 则表明信度不足,不应当接受。

由于合成信度是一种点估计,如果考虑模型估计有可能受随机误差的影响,信度估计应当采用区间估计的做法报告信度,以此说明该测量估计方法所得到的合成信度可以在 95% 或 99% 的置信区间上反映总体信度。

合成信度的区间估计(SE 为标准误):

$$95\% \text{ CI}(\rho_c) = \rho \pm 1.96\text{SE}(\rho_c) \quad \text{式 2}$$

$$99\% \text{ CI}(\rho_c) = \rho \pm 2.59\text{SE}(\rho_c) \quad \text{式 3}$$

鉴于 α 系数无法对单个题项的信度进行检验,Bollen(1989) 提出了使用个别题项信度(Squared Multiple Correlations, SMC, 也称为 R^2) 作为测量题项被潜变量所解释的程度,SMC 可以由结构方程得到,其计算公式如下:

$$\text{SMC} = \frac{\lambda_i^2}{\lambda_i^2 + \Theta_{ii}} \quad \text{式 4}$$

式中 λ_i 为各题项完全标准化因子载荷, Θ_{ii} 为测量误差。

因为 SMC 反映测量指标受潜变量的影响程度,因而 SMC 越大,表示真分数所占比重越大,测验信度也就越高,反之则表明越低,当 SMC 低到不可被接受时,需要修正或删除题项。需要注意的是,SMC

的计算是以单维假设为前提的信度估计,如果一个测量指标受到两个及两个以上潜变量的影响,则该式不适用。

2.3 平均变异萃取量

测量题项的因子载荷越高,表示测量指标能够反映潜变量的能力越高,因子也就越能够解释各测量指标的变异程度,因而可以计算出一个平均变异萃取量(Average Variance Extracted, AVE 或 ρ_v),来反映一个潜变量被一组测量指标有效估计的聚敛程度指标。从计算过程来看, ρ_v 实际上反映的是当各个测量指标提供一个单位的变异时,潜变量的变异量占总变异的百分比。当 ρ_v 大于 0.50 时,表示潜变量的聚敛能力理想。

$$\text{AVE} = \rho_v = \frac{\sum \lambda_i^2}{\sum \lambda_i^2 + \sum \Theta_{ii}} \quad \text{式 5}$$

式中 λ_i 为各题项完全标准化因子载荷, Θ_{ii} 为测量误差。

2.4 效度

效度是指测量工具或方法能够准确测出所需测量的事物的程度。效度分为三种类型:效标效度、内容效度和构念效度。

效标效度指测验分数与测验所欲测量之特质之外的指标之间的相关程度。这种程度以相关系数来表示,称为效标系数。由于效标效度受到测量本身随机误差的影响,也受到效标误差的影响,不同的效标会产生不同的效标效度,使得效标系数难以确定,再者,许多社会科学的测量根本没有效标可用,效标系数的测量自然就无从谈起了。

内容效度是指一个测验本身所能包含的概念意义范围或程度,即测验内容是否针对欲测的目的,是否具有代表性和适当性。内容效度相当依赖逻辑的处理而非统计的分析。因此,理论的思维就显得非常重要,但统计结果也可以从一个侧面上反映内容效度。在结构方程中,如果因子载荷大于 0.5,即可从一定程度上说明该测验具有内容效度。

构念效度对社会科学的测量而言是相当重要的效度指标,因为许多社会科学方面的理论都是依据假设构想而来的。构念效度是指测量到理论上的特质或构念的程度,简单的说,就是指构念被其测量结果反映的程度。在操作技术上,可以使用竞争模式法来进行构念效度的区分,同时用相关系数区间估计法和平均变异萃取量比较法来验证。

竞争模式法是利用两个验证性因子模型(confirmatory factor analysis; CFA) 进行竞争比较。其中一个 CFA 模型令两个构念之间的相关自由估计(效度模型),而另一个 CFA 模型则是将两个构念之间的相关设定为 1.00(完全线性相关模型)或将两个

构念合并,置于一个因子下(单因素模型),由于这样的模型少了一个有待估计的参数,自由度多1,模型的拟合度也会相应的有所降低。如果效度模型没有显著地比完全相关模型或单因素模型优越,则可认为两个构念之间缺乏区分度。

相关系数区间估计法是指两个潜变量的相关系数95%的置信区间不能涵盖1,即上述公式2所得出的置信区间结果中不能包括1,否则表示构念缺乏区分度。平均变异萃取量比较法是指比较两个潜变量的平均变异萃取量(ρ_v)的平均值是否大于两个潜变量相关系数的平方(Fornell & Larcker, 1981),如果得到的答案是肯定的,则显示构念之间具有理想的区分度。

虽然许多学者强调理论内容对构念效度评鉴的重要性,但不可否认的是,基于数理统计上的逻辑推理所得出的结论,是对构念之间具有区分效度的极好说明。

3 复核效度检验

所谓复核效度(cross-validation)就是指一个模型在许多不同总体样本下,而非只是在其所衍生的样本下能够复制的程度。复核效度检验通过以后,可以说明模型具有跨样本的稳定性和有效性。复核效度的主要目的是检验模型的预测效度,根据 Diamantopoulos 和 Siguaw(2000)的看法,复核效度有四种类型:

表3 复核效度的类型

模型数目	单一模型 模型比较	复核样本	
		相同总体	不同总体
		模型稳定 模型选择	模型效度延展 模型普遍化

复核效度检验的典型做法就是把所收集到的样本分为两个样本,然后用这两个不同的样本对同一假设模型进行估计,其中的一个样本称为校正样本,另一个样本则称为效度样本。首先用校正样本对需检验的结构方程模型进行估计后,将所估计得到的参数套用在效度样本上,再进行模型拟合检验,分析所得结果的差异程度。

复核效度也可以将样本的角色互换,即将校正样本改为效标样本,效标样本改为校正样本。照此再做一次复核效度,那就可以称为双复核效度。

3.1 多总体大样本中的复核效度

在多总体大样本的操作中,结构方程的复核效度检验可以区分为三种方式:宽松复核、温和复核以及严格复核。宽松复核的做法是将效度样本上的模式套用校正样本上的模式进行界定,但效度样本的参数仍采用自由估计方式。相比之下,严格复核要求两个样本的假设模型具有完全相等条件,因而不将效度样本的模式套用校正样本的模式进行界

定,而且参数也要套用校正样本的估计值。如果效度样本的参数不是完全套用校正样本的估计值,只是部分套用,则称为温和复核。

复核效度检验主要是检查卡方统计量和卡方贡献百分比,这两个数据代表效度样本的加入对模型整体拟合的影响程度。检查三种策略下模型拟合度的卡方增加量是否达到显著水平,如果卡方增量没有达到显著水平,则表明复核效度通过检验,反之,则检验不通过。由于进行复核效度检验时,越严格的程序对稳定性检验的条件要求越苛刻,但在要求越高的条件下,效度样本的卡方增量未达显著水平,越能反映样本复核效度的能力。因此,建议研究者先采用严格复核检验,若卡方增量达到显著水平,再改为温和复核(邱皓政,林碧芳,2009)。

3.2 样本不够大时的复核效度检验

多总体复合效度检验的处理方法,无疑是需要相当大的样本数。但很多时候,研究者所使用的样本都不足以分为两个样本,即将收集到的样本分为校正样本与效度样本的同时又不失样本特性。而复核效度的适当性在评鉴模型又是必须的,此时可以借用 Lisrel 软件产生的 ECVI 指数来评判: ECVI 的值越小,表示该模型的复核效度越好。复核效度指数(Cross-Validation Index, CVI)是 Cudeck 和 Browne(1983)发展起来用于评鉴模型复核效度适当性问题的指数,为了避免样本太小而无法一分为二, Cudeck 和 Browne(1993)建议采用复合效度指数的期望值(Expected value of Cross-Validation Index, ECVI)作为复核效度检验的标准,当 ECVI 值越小时,表明模型的复核效度越好。由于所获得的 ECVI 是点估计,故 Cudeck 和 Browne(1993)提出了区间估计的方法, Lisrel 会产生 90% 置信区间的 ECVI 范围。同时,在实践中,所选择模型的 ECVI 值应小于独立模型和饱和模型的 ECVI 值。

4 总结

在评价模型的整体拟合时,主要是通过拟合指数来实现的,目前大多采用传统的临界值(NNFI 和 CFI 界值为 0.9; Mc 界值 0.85; RMSEA 界值为 0.08),但当拟合指数处于临界值附近时,不能凭临界值一锤定音,应结合其他方面检验结果的同时,根据模型构建的理论基础来做出最终的评判;在评价模型的内部拟合时,可以从项目的质量检验、测验信度、平均变异萃取量和效度四个方面综合评价;多样本分析中的复核效度检验分为宽松复核、温和复核和严格复核,主要是检查加入效度样本后的卡方增量是否达到显著水平,但如果为研究样本所限,可以通过 ECVI 来实现, ECVI 越小,复核效度越好。

评鉴模型拟合程度的好坏并没有一个基于数据上的临界值标准,模型的拟合评鉴只是评鉴所构建

的模型和原始数据之间的拟合或差异程度,并不能完全凭数据拟合的好坏去评价模型构建是否合理。无论是评价模型的整体拟合、内部拟合还是进行复核效度检验,大多是根据统计学结果得出相应的结论,而实际上,理论逻辑应当摆在第一位而且是最重要的一位。如果假设模型的构建缺乏理论逻辑基础,所有的模型评鉴都只能是“空中楼阁”,毫无意义可言。

参考文献

- 郭庆科,王炜丽,陈雪霞,韩丹. (2007). 验证性因素分析中模型拟合的判断. *心理学探新*, 27(4), 83-87.
- 郭庆科,李芳,王炜丽,陈雪霞,孟庆茂. (2008). 不同条件下拟合指数的表现及临界值的选择. *心理学报*, 40(1), 109-118.
- 侯杰泰,温忠麟,成子娟. (2001). *结构方程模型及其应用* (p. 156). 北京: 北京科学出版社.
- 邱皓政,林碧芳. (2009). *结构方程模型的原理与应用* (pp. 335-344). 北京: 中国轻工业出版社.
- 温忠麟,侯杰泰,马什赫伯特. (2004). 结构方程模型检验: 拟合指数与卡方准则. *心理学报*, 36(2), 186-194.
- 温忠麟,侯杰泰. (2008). 检验的临界值: 真伪差距多大才能辨别? ——评《不同条件下拟合指数的表现及临界值的选择》. *心理学报*, 40(1), 119-124.
- 温忠麟,叶宝娟. (2011). 测验信度估计: 从 α 系数到内部一致性信度. *心理学报*, 43(7), 821-829.
- 徐万里. (2008). 结构方程模式在信度检验中的应用. *统计与信息论坛*, 23(7), 9-13.
- 易丹辉. (2008). *结构方程模型方法与应用* (p. 127). 北京: 中国人民大学出版社.
- Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16(1), 74-94.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, 34, 181-197.
- Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research*, 17, 303-316.
- Cudeck, R., & Browne, M. W. (1993). *Alternative ways of assessing model fit. Testing structural equation models*. CA: Sage.
- Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LIS-REL: A guide for the uninitiated*. Thousand Oaks, CA: Sage.
- Hair, J. F. JR., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th Ed.). Upper Saddle River, NJ: Prentice-Hall.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in Overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling*, 11(3), 320-341.
- Raine-Eudy, R. (2000). Using structural equation modeling to test for different reliability and validity: An empirical demonstration. *Structural Equation Modeling*, 7(1), 124-141.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, A: Allyn & Bacon.
- Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, 40(1), 115-148.

Evaluation of Structural Equation Modeling Fitting: The Overall Fitting, the Internal Fitting and Cross - Validation Testing

Zheng Wenzhi Wu Wenyi

(College of Business Administration, Huaqiao University, Quanzhou 362021)

Abstract: The article introduces mainly the evaluation on modeling and three aspects of data fitting degree in structural equation: the overall fitting, the internal fitting and cross-validation testing. The overall fitting evaluation refers that the evaluation should consider the goodness of fit index. When the fit index is on the critical value, the evaluation should refer to other inspection results at the meantime and make a comprehensive judgment based on the theoretical basis for building model. The internal fitting is inspected from four angles including project quality inspection, reliability testing, average variance extracted and validity. Cross-validation testing can be divided into loose replication strategy, moderate replication strategy and strictly replication strategy in the analysis of diversity, but it can be achieved through ECVI if it is limited by research samples.

Key words: structural equation modeling; the overall fitting; the internal fitting; cross-validation testing