心理学报 2016, Vol. 48, No.11, 1489−1498
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2016.01489

缺失数据的结构方程建模：全息极大似然估计时辅助变量的作用*

王孟成 邓俏文

(广州大学心理系; 广州大学心理测量与潜变量建模研究中心, 广州 510006)

摘 要 本研究通过蒙特卡洛模拟考查了采用全息极大似然估计进行缺失数据建模时辅助变量的作用。具体考查了辅助变量与研究变量的共缺机制、共缺率、相关程度、辅助变量数目与样本量等因素对参数估计结果精确性的影响。结果表明, 当辅助与研究变量共缺时：(1) 对于完全随机缺失的辅助变量, 结果更容易出现偏差; (2) 对于 MAR-MAR 组合机制, 纳入单个辅助变量是有益的; 对于 MAR-MCAR 或 MAR-MNAR 组合机制, 纳入多于一个辅助变量的效果更好; (3) 纳入与研究变量低相关的辅助变量对结果也是有益的。

关键词 缺失数据; 缺失机制; 结构方程; 全息极大似然估计; 辅助变量; 蒙特卡洛模拟

分类号 B841

1 引言

在对心理学等社科调查的数据进行建模时, 常常遇到数据缺失的情况。例如, 研究参与者拒绝接受调查、不愿意回答或遗漏某些问题。毫不夸张的说, 数据缺失无法避免, 因此如何处理缺失数据就成了摆在研究者面前重要而又无法回避的问题。

全息极大似然估计(Full Information Maximum Likelihood, FIML)和多重插补(Multiple Imputation, MI)是目前缺失数据建模最为学者推崇的方法(Graham, 2009; Schafer & Graham, 2002)。这两种方法在特定条件下所得结果是等价的, 但鉴于建模软件的可获得性、统计处理的便捷性以及结果的稳健性(e.g., Yuan, Yang-Wallentin, & Bentler, 2012), 在方法学实践中 FIML 更加方便和灵活(Yuan et al., 2012; 王孟成, 叶浩生, 2014)。

在缺失数据建模实践中, 方法学者通常会建议纳入辅助变量(auxiliary variable)来提高结果的稳健性。采用 FIML 处理缺失数据时, 合理利用辅助变量可以使与 FIML 密切相关的缺失机制得到满足, 从而产生更可靠的参数估计, 增加统计功效(Collins, Schafer, & Kam, 2001; Graham, 2009)。然而, 当前方法学领域对纳入辅助变量的研究还有些重要的问题尚未探明。例如, 纳入自身就存在缺失的辅助变量是否有益？因此, 本研究拟采用蒙特卡洛模拟的方法对尚存的问题做进一步的探索, 希望为应用研究者合理使用辅助变量提供有益的指引。

1.1 数据的缺失机制与现代的处理方法

美国统计学家 Rubin (1976)最早将缺失机制分为 3 类：完全随机缺失、随机缺失和非随机缺失。完全随机缺失(Missing Completely at Random, MCAR)指变量缺失发生的可能性与变量自身及其他变量都无关, 即变量出现缺失这一事件是随机事件。随机缺失(Missing at Random, MAR)指变量缺失发生的可能性与模型中某些观测变量有关而与该变量自身无关, 即缺失发生的可能性与其他变量有关。非随机缺失(Missing Not at Random, MNAR)指变量缺失发生的可能性只与自身相关。

现代处理缺失数据的方法中, 最为研究者推崇的是全息极大似然估计和多重插补(Enders & Bandalos, 2001; Graham, 2009; Schafer & Graham, 2002)。随着统计软件的发展, 这两种方法得到了广

收稿日期: 2015−06−05

* 国家自然科学基金(31400904)和广州大学“创新强校工程” (2014WQNCX069)项目资助。

通讯作者: 王孟成, E-mail: wmcheng2006@126.com (由于篇幅限制, 本研究的完整结果可以和作者联系获取)

泛的应用(e.g., Kidger et al., 2015)。

在处理缺失数据时, FIML 使用所有观测变量的全部信息进行参数估计, 因而又称为全息极大似然估计。在满足 MCAR 和 MAR 的条件下, FIML 产生无偏和有效的参数估计。FIML 处理缺失值并没有使用替代值对缺失值进行替换, 而是根据未缺失数据的信息采用迭代的方式进行估计(Enders & Bandalos, 2001; Graham, 2009)。

MI 假设在数据随机缺失情况下, 用两个或更多能反映数据本身概率分布的值来插补缺失数据。一个完整的 MI 包含 3 步: 数据插补, 计算和汇总。理论上, 在插补次数无限的情况下, MI 与 FIML 结果一致。

两种方法都是目前处理缺失数据最有效的方法, 但两种方法又有着显著的不同。首先, FIML 是专门用于模型分析的参数估计方法, 严格意义上来说, MI 是基于统计模型的处理过程。其次, FIML 分析时并非填补缺失值, 而是根据已知信息采用迭代的方式进行估计, MI 需要填补数据再进行后续分析。最后, 运用 MI 处理缺失数据时辅助变量的效用还没有一致的结论, 尤其是在实际数据处理中(Mustillo, 2012)。基于以上 FIML 和 MI 特点的比较, 以及本文主要研究与辅助变量相关的内容, 本研究只关注 FIML 处理缺失数据时辅助变量的相关问题。

1.2 辅助变量

辅助变量是研究者不感兴趣, 但能为缺失数据建模提供有用信息的变量(Enders, 2008)。提供辅助信息的变量通常是造成数据缺失的原因变量或者与研究变量相关的变量。

由于 MAR 机制包含了未观测的数据, 因此在缺少这些未观测数据的条件下, 无法在统计上对数据是否满足 MAR 进行检验(Raykov, 2011)。为了克服这一不足, 方法学者提出使用纳入辅助变量的方式来提高满足 MAR 假设的可能性(Collins et al., 2001; Schafer & Graham, 2002; Yuan & Lu, 2008)。目前在 FIML 分析中, 最常用的纳入辅助变量的方式是通过 Graham (2003)提出的饱和关联模型(saturated correlates model, SCM)。在 SCM 中, 通常允许辅助变量间、辅助变量与外生观测指标以及内生观测指标的测量误差相关。在 SCM 提出之前, 运用 ML 处理缺失数据比较麻烦, 因此先前的研究多侧重通过专门运用 MI 的软件包 NORM (Schafer, 1999)进行缺失数据处理。采用 MI 处理缺失数据时, 纳入辅助变量很方便, 但由于插补次数所导致结果的不确定性, 人们开始寻找 ML 下如何处理缺失数据的方法(Graham, 2003)。采用 FIML 进行分析时, 若要纳入辅助变量, 某些潜变量建模软件(如 Mplus)会默认采用 SCM (Muthén & Muthén, 1998-2010)。SCM 的提出为 FIML/SEM 处理缺失数据提供了很大的便利, 也使得 Mplus 等自动采用 SCM 模型的软件成为运用 FIML/SEM 处理实际缺失数据或模拟研究常用的软件(Enders, 2008; 王孟成, 2014)。

2 问题提出

2.1 先前类似研究

先前的模拟研究多数只考虑了辅助变量不缺失的情况(Collins et al., 2001; Graham, 2003; Mustillo, 2012)。例如, Graham (2003)的研究通过 SCM 对单个不缺失的辅助变量进行 FIML/SEM 分析。这些模拟研究均发现不缺失的辅助变量与研究变量高度相关时, 纳入辅助变量能够改善模型参数估计。Mustillo (2012)的研究在回归模型中通过 MI 处理缺失值, 探究辅助变量类型、研究变量的缺失率与缺失机制的关系, 该研究发现纳入辅助变量对参数估计没有明显的改善。然而 Collins 等(2001)指出, 即使纳入与缺失变量无关的辅助变量, 得到最坏的结果也是中性的, 并不会恶化参数估计。而当与缺失变量相关的辅助变量被忽略时, 均值、方差、回归估计会产生实质性的偏差。

另外, 纳入辅助变量进行缺失数据分析时, 不仅研究变量存在缺失, 辅助变量也常常存在缺失。当辅助变量也存在缺失时, 情况又会如何? 为数不多的研究表明, 尽管纳入有缺失的辅助变量不如纳入完全的辅助变量那么有效, 但纳入总比忽略它更有益(Enders, 2008; Hardt, Herke, & Leonhart, 2012; Yoo, 2009)。例如, Enders (2008)采用 FIML/SEM 对研究变量的缺失机制(MAR)、单辅助变量的缺失机制(MCAR, MNAR)、辅助变量的缺失率(25%、50%)、相关程度(r = 0.54, r = 0.90)进行考查, 结果发现即使辅助变量缺失 50% (且辅助变量的缺失机制为 MNAR), 纳入它也有利于参数估计。虽然以上研究考虑到辅助变量有缺失的情况, 但它们主要研究辅助变量与研究变量各自的缺失率或设定辅助变量的缺失率后, 让研究变量的缺失率依辅助变量的缺失率而定(Enders, 2008), 而两变量各自的缺失并不等于研究样本中两变量共同的缺失(简称共缺)。

共缺指同一个体的数据在研究变量上有缺失, 在辅助变量上也有缺失, 共缺率则是共缺频数在样

本中的比例。因此, 研究变量与辅助变量共缺时, 参数估计的情况成了数据分析时的另一个问题。Von Hippel (2007)指出, 即使辅助变量与研究变量呈高相关, 当辅助变量与研究变量共缺时, 结果也得不到改善。Enders (2008)的研究无意中发现当辅助变量与研究变量的共缺率达到 15%时, 结果会产生明显的偏差。但这个问题目前并没有得到系统研究。因此更多的缺失机制组合、共缺率、辅助变量数等问题需要作进一步的探讨。

2.2　本研究的目的

通过文献回顾不难发现, 至少还有如下 4 个问题亟待解决：第一, 先前研究只探讨了单个辅助变量的情况。当有多个辅助变量且样本量足够大的时候, 以上问题会发生怎样的变化? 第二, 上述研究并没有对无意中发现的共缺率问题做进一步的探索; 第三, 以往研究仅局限在 MAR 机制的研究变量与 MCAR 和 MNAR 的辅助变量, 没有进一步探究研究变量与辅助变量其他的缺失机制组合(简称共缺机制); 第四, 先前的研究在参数设定时参考 Collins 等(2001)的模型, 该研究设定辅助变量与研究变量的相关程度过高(0.54, 0.90), 这在实际研究中并不多见。

针对以上 4 点, 本研究通过蒙特卡洛模拟, 采用 FIML 处理结构方程建模中的缺失数据, 主要目的是探究辅助变量与研究变量的共缺机制、共缺率、相关程度、辅助变量数与样本量这些因素对参数估计结果的影响。

3　研究设计

3.1　模拟研究设计

3.1.1　研究假设模型

本研究设定的模型参照 Enders (2008)的研究, 模型由两个因子(X 和 Y)构成结构模型, 每个因子有 3 个观测指标, 外生潜变量 X 对内生潜变量 Y 的回归系数设为 0.60。外生潜变量 X 的指标(x1, x2 和 x3)的数据不缺失, 内生潜变量 Y 的指标(y1, y2 和 y3)的数据存在缺失, 辅助变量 Z 是一个/组有缺失的观测变量(具体见下一小节)。

Collins 等(2001)的研究发现, 辅助变量与研究变量之间的相关系数最好能达到 0.4 以上, 在他们的研究中, 研究变量间还设定 ρ_{yz} = 0.90 的高度相关, 然而在实际研究中如此高的相关并不多见。因此, 在参考前人研究(Enders & Peugh, 2004; Hardt et al., 2012)的基础上, 本研究考虑如下两组相关水平：低相关设为 ρ_{xz} = 0.2, ρ_{yz} = 0.3 和中等偏高相关为 ρ_{xz} = 0.5, ρ_{yz} = 0.6。所有变量服从标准正态分布, 即均值为 0, 方差为 1, 固定因子方差为 1, 因子负荷为 0.70, 残差方差为 0.51。存在多个辅助变量时, 综合 Hardt 等(2012, r_{zs} = 0.1 或 r_{zs} = 0.5)与 Enders 和 Peugh (2004, r_{zs} = 0.3)的研究, 本研究设定辅助变量之间的相关系数为 0.4。

参照 Graham (2003)提出的饱和关联模型, 图 1 给出了单个辅助变量的饱和关联模型(辅助变量与研究变量呈低/高相关)。

3.1.2　缺失机制

当研究变量的缺失机制为 MNAR 时, 即变量缺失发生的可能性只与自身相关时, 纳入辅助变量无法改善参数估计结果(Enders, 2006; Yoo, 2009)。再者, 当研究变量与辅助变量的共缺率很大, 缺失机制都是 MNAR 时, 此时任何方法都难以得到无偏的估计结果, 因此本研究只考虑研究变量 Y 的 2 种缺失机制(MAR & MCAR)与辅助变量 Z 的 3 种缺失机制(MAR, MCAR, & MNAR), 共 6 种共缺机制组合。

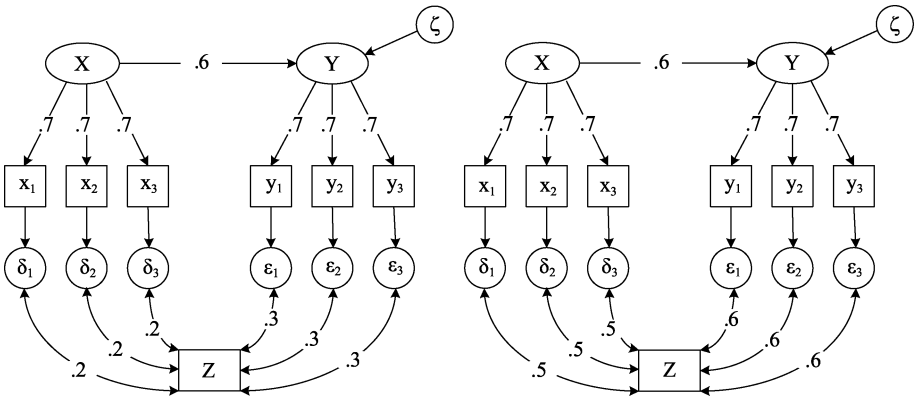

图 1　蒙特卡洛模拟所依据的模型路径图

与先前类似研究一致(Enders, 2008), 本研究采用逻辑回归生成缺失数据。具体来说, 当 Y 观测变量的缺失机制为 MAR 时, 设定 Y 的缺失与辅助变量 Z 有关, 斜率参数为正, 即 Z 的值越大, Y 的缺失率越大。当辅助变量 Z 的缺失机制为 MAR 时, 设定 Z 的缺失与 X 观测变量有关, 斜率参数设为负, 即 X 观测变量的值越大, Z 的缺失率越小。当辅助变量 Z 的缺失机制为 MNAR 时, Z 的缺失与自身相关, 设定斜率参数为负, 即 Z 的值越大, Z 的数据缺失率越小。这样, 当共缺组合形式为 MAR-MNAR 时, 辅助变量可为研究变量提供更多的信息, 同时可减少研究变量与辅助变量的缺失都是由辅助变量的缺失造成的可能性。

3.1.3 样本量和缺失率

表 1 汇总了先前相关模拟研究中, 样本量和缺失率的设置数据。大多数研究设置的样本量在 200~500 之间, 考虑到大多数心理学调查研究和 FIML 对样本量的要求, 本研究的样本量设为 100、200、500 和 1000。表中设定的缺失率指单纯研究变量的缺失率或辅助变量的缺失率。根据 Enders (2008)的研究, 当研究变量与单辅助变量的共缺率达到 8%时, 辅助变量的纳入能使结果得到改善, 所得参数估计结果偏差很小, 而当共缺率达到 15%时, 参数估计偏差显著增加。因此本研究设定辅助变量与研究变量的共缺率为 5%、10%、15%和 20%。

表 1 相关模拟研究设置的样本量与缺失率参数汇总

作者	分析方法	样本量	缺失率(%)
Arbuckle (1996)	ML	145, 500	0, 5, 10, 20, 30
Collins et al. (2001)	MI & ML	500	25, 50
Enders & Bandalos (2001)	FIML	100, 250, 500, 750	2, 5, 10, 15, 25
Newman (2003)	FIML & MI[a]	400	25, 50, 75
Enders & Peugh (2004)	ML & EM	200, 400,600	5, 15, 25
Enders (2008)	FIML	500	25, 50
Enders & Gottschall (2011)	MI	50, 250	5, 15, 25

注：a 指该研究中还涉及传统的分析方法; EM = 期望−最大算法(expectation maximization)。

3.1.4 辅助变量数

不少研究建议纳入辅助变量进行缺失数据分析, 但对于纳入多少个辅助变量的问题在以往的研究中并没有专门探讨。Enders (2008)主要研究一个辅助变量的情况, 本研究的模拟方法参考 Enders (2008)基于 FIML/SEM 的研究, 并与该研究结果进行比较。另外, 综合模拟设置的最小样本量、最大缺失率及结构方程模型的要求, 本研究主要考察 1 个、3 个和 5 个辅助变量的情况。

3.1.5 数据生成

本研究所有的数据生成与分析均采用 Mplus 7.0 (Muthén & Muthén, 1998-2010)完成。本研究共模拟 3 种情况：(1) 研究变量缺失, 辅助变量不缺失, 且在建模时纳入辅助变量; (2) 研究变量和辅助变量共缺, 且在建模时纳入辅助变量; (3) 研究变量与辅助变量共缺, 但在建模时不纳入辅助变量。第一种模拟条件由第二种模拟条件设定辅助变量不缺失得到, 其中的缺失机制组合形式实质上只是研究变量的缺失机制, 但这两种条件下研究变量的缺失率相同。第三种模拟条件由第二种模拟条件在模型分析时, 不纳入辅助变量进行分析而得到。

每种模拟共有 576 种组合, 每种组合均重复 5000 次。后两种模拟条件中控制的因素有：4 种样本量(100、200、500、1000)、4 种共缺率(0.05、0.10、0.15、0.20)、6 种共缺机制组合(MAR-MAR, MAR-MCAR, MAR-MNAR, MCAR-MAR, MCAR- MCAR, MCAR-MNAR)、3 种辅助变量数目(1、3、5)、两种相关程度(低相关$\rho_{xz}=0.2, \rho_{yz}=0.3$和中等偏高相关$\rho_{xz}=0.5, \rho_{yz}=0.6$)。

3.2 结果评价标准

采用模拟研究中常用的两个评估标准：参数估计的偏差(e.g., Enders & Bandalos, 2001; Yoo, 2009)和覆盖率(Coverage; Yoo, 2009)。

比较常用的估计偏差的指标是标准偏差(Standardized Bias; Collins et al., 2001; Enders & Gottschall, 2011), 标准偏差 = (平均估计值−理论值)/平均标准误(以示区分, 记为偏差 $_{S.E.}$)。如果偏差 $_{S.E.}$等于–0.5, 意味着该平均估计值处于理论值–0.5 个标准误的位置。Collins 等(2001)指出偏差 $_{S.E}$ 小于 0.4 为无偏估计, 后来的研究也采用此标准(Enders & Gottschall, 2011)。但有研究质疑以 0.4 作为判断标准的适切性(Graham, 2009), 因此当偏差出现的情况较多时, 本研究同时采用偏差 $_{100}$ ($Bias_{100}$)作为结果评价指标。

蒙特卡洛模拟中, 偏差可通过以下公式得到：

$$\bar{\hat{\phi}} = \sum_{k=1}^{K} \hat{\phi}_k / K \qquad \text{(公式 3)}$$

$$Bias_{100} = (\bar{\hat{\phi}} - \phi) / \phi \times 100 \qquad \text{(公式 4)}$$

其中, K 表示重复总次数, k 为重复次数($k=1$,

$2,\cdots,K$)。$\hat{\phi}_k$ 为第 k 次运算参数的估计值，ϕ 为参数的理论值，$\hat{\phi}$ 为 K 次运算的平均估计值。公式 3 和公式 4 可简化为：偏差=(平均估计值−理论值)/理论值 × 100 (以示区分，记为偏差 $_{100}$)。根据 Muthén, Kaplan 和 Hollis (1987)的建议，本研究认为偏差大于 10 为有偏估计。

另外，在模拟研究中，覆盖率表示每次重复模拟计算所得结果等于/接近真值的比例，类似于参数区间估计(频率论)的置信区间：区间包含的真值。但这里的置信区间所允许的犯错误的概率(即显著性水平)不是固定的，显著性水平=1−覆盖率。所以，当覆盖率为 0.95 时，意味着在抽样 1000 次(模拟计算)得到的结果所组成的区间中，有 950 次得到的估计值在总的区间中包含了真值，此时犯错误的概率为 0.05。前人的研究认为覆盖率小于 0.90 是不可接受的(Collins et al., 2001; Enders & Peugh, 2004)，本研究也采用 0.90 的标准。

4　结果

综合前人研究考虑的需要模型估计的参数(Collins et al., 2001; Enders, 2008)，本研究主要考虑因子负荷和回归系数的估计值。其中因子负荷的结果(偏差与覆盖率)由对应结果求平均数得到(即条目因子负荷之和除以条目个数)。由于版面限制本研究只呈现了部分结果(n = 500)，更多的结果(其他样本量与偏差 $_{100}$ 的结果)可与作者联系获得。

4.1　辅助变量不缺失

建模时纳入不缺失的辅助变量，估计结果的偏差和覆盖率都在可接受范围。总的来说，样本量越大，因子负荷和回归系数的偏差值越小。对于小样本(n = 100)，辅助变量越多，回归系数的偏差越大。对于回归系数的参数估计而言，随着样本量的增大，纳入单个辅助变量依然是有益的。另外，辅助变量越多(n = 100 除外)，X 因子负荷的偏差越小。在此模拟的各种条件下，覆盖率均达标且变化不大。

4.2　辅助与研究变量共缺：建模时纳入辅助变量

在此条件下，只有回归系数的参数估计产生偏差，其他参数均无偏。对于回归系数的结果，“MCAR-”组合形式的参数估计结果都无偏。偏差多出现在“MAR-”组合形式中，且样本量越大，出现偏差的情况越多。相关越高，越容易出现偏差。

在此模拟条件下，X、Y 因子负荷的覆盖率都在可接受范围内。在高相关、MAR-MAR 组合条件下，纳入 5 个辅助变量时，X 因子负荷的覆盖率比其他辅助变量数目条件下的稍高(差异在 0.01~0.02 之间)，然而 Y 因子负荷没有呈现此特点。同样条件下，回归系数的覆盖率更容易出现不可接受的结果(低相关条件下的结果几乎全无偏)，有偏的结果全部出现在“MAR-”组合上(n = 100 无偏)。与 Enders (2008)结果一致的是，当 n = 500 且研究变量的缺失机制为 MAR 时，MAR-MCAR 组合似乎比 MAR-MNAR 组合出现偏差的情况更多(见表 2)。而且在 MAR-MCAR 组合上，当共缺率达到 15%时，单辅助变量的情况下出现明显的偏差，这也是为什么 Enders (2008)的研究发现影响参数估计结果的因素主要是辅助与研究变量的共同缺失模式(共缺率)，

表 2　辅助变量与研究变量共缺且纳入辅助变量时回归系数的估计偏差 $_{S.E}$ (n = 500，中等偏高相关)

MR	AUX	MAR-MAR	MAR-MCAR	MAR-MNAR	MCAR-MAR	MCAR-MCAR	MCAR-MNAR
0.05	1	0.0469	–0.3646	**–1.4419**	–0.0058	–0.0058	–0.0058
	3	0.0418	–0.0526	0.1175	–0.0131	–0.0146	–0.0087
	5	0.1334	–0.0480	0.2910	–0.0029	–0.0029	0.0044
0.10	1	0.0374	–0.0230	**–1.7935**	–0.0043	–0.0058	–0.0044
	3	0.1449	**–0.8075**	0.1105	–0.0130	–0.0116	–0.0072
	5	0.2615	–0.1553	0.3471	–0.0029	–0.0029	–0.0014
0.15	1	0.1432	**–0.7913**	0.0261	–0.0014	–0.0028	–0.0029
	3	0.1958	–0.1301	0.1107	–0.0157	–0.0144	–0.0071
	5	**0.4418**	–0.1693	0.3388	–0.0014	–0.0014	0.0216
0.20	1	0.1207	**–0.9355**	0.0266	–0.0014	–0.0042	–0.0014
	3	**–0.4426**	–0.1681	–0.0505	–0.0156	–0.0172	0.0000
	5	–0.2518	–0.2358	0.1852	–0.0014	–0.0014	0.0301

注：粗体表示结果出现偏差; MR=缺失率, AUX=辅助变量数, MAR-MAR=横线左边代表研究变量的缺失机制，右边代表辅助变量的缺失机制, MAR=随机缺失, MCAR=完全随机缺失, MNAR=非随机缺失。下同。

而不是缺失机制。

4.3 辅助与研究变量共缺：建模时不纳入辅助变量

在此模拟条件下, 大多数共缺机制组合下的结果都出现严重的偏差, 且高相关或辅助变量数多的条件下更容易出现偏差, 偏差与覆盖率的结果呈现一致的规律。但是 Y 因子负荷的偏差都在接受范围内, 且随着样本量的增大, 可接受的偏差值减小。另外, 即使是 MCAR 的研究变量, 估计结果也会出现偏差。与前一种模拟情况类似, 偏差多出现在“-MAR”、“-MCAR”组合上, “-MNAR”组合较少出现偏差(见表 3)。

4.4 不同模拟条件间的比较

由于第一种模拟条件由第二种模拟条件设定辅助变量不缺失得到, 即两种条件下的研究变量的缺失率、缺失机制等是相同的, 通过比较这两种模拟条件下的结果发现, 辅助变量有缺失比不缺失时产生严重偏差的可能性更大。

相对于第二种模拟的条件, 在第三种模拟条件下, 偏差主要出现在 X 因子负荷和回归系数上。通过相关程度、辅助变量数、样本量的比较, 发现偏差主要出现在中等偏高相关条件下; 且辅助变量越多, 出现偏差的情况越多; 样本量越大, 出现偏差的情况也越多。通过比较高相关条件下 X 因子负荷偏差的结果, 发现在不纳入辅助变量的条件下, MAR-MCAR, MCAR-MAR 和 MCAR-MCAR 这 3 种组合的结果都出现偏差。纳入辅助变量之后, 结果得到明显的改善, 尤其是“MCAR-”组合, 估计结果的偏差全部都在可接受范围内。

通过比较第二种模拟和第三种模拟回归系数的结果(见表 4), 发现在不纳入辅助变量时, 所有缺失机制的组合结果都出现偏差。其中, 在“MAR-”组合形式的 MAR-MAR 出现偏差数量最少; 在“MCAR-”组合形式中的 MCAR-MNAR 出现偏差数量最少。随着样本量的增大, 出现偏差的数量增多。纳入辅助变量之后, “MCAR-”组合结果全部在可接受范围内。“MAR-”组合的结果在小样本的条件下无偏, 大样本时依然出现偏差, 但有偏的结果得到明显的改善。纳入辅助变量之后, 即使辅助变量与研究变量呈低相关, 因子负荷和回归系数的估计同样得到了改善, 可接受的偏差值变小。

5 讨论与结论

Enders (2008)的研究主要考虑辅助变量的缺失率、缺失机制对参数估计的影响。本研究在 Enders (2008)的基础上, 进一步考查辅助变量与研究变量的共缺率、共缺机制、相关程度、辅助变量数目及样本量对参数估计结果的影响。

5.1 缺失机制的影响

当辅助变量与研究变量共缺时, 相对于 MNAR 的辅助变量, MCAR 的辅助变量更容易出现参数估计偏差。这说明, 即使辅助变量的缺失机制为 MNAR, 纳入模型进行分析也有利于改善参数估计(Enders, 2008)。Enders (2008)指出由于 MCAR 辅助变量的无方向性或不确定性, 增大了辅助变量的缺失机制与研究变量的缺失机制是由共同因素造成

表 3 辅助变量与研究变量共缺但不纳入辅助变量时 X 因子负荷的估计偏差 $_{S.E}$ (中等偏高相关)

MR	AUX	MAR-MAR	MAR-MCAR	MAR-MNAR	MCAR-MAR	MCAR-MCAR	MCAR-MNAR
0.05	a1	0.1011	**–0.5050**	–0.0071	–0.3565	**–0.5150**	–0.2463
	a3	**1.0821**	**–0.6143**	–0.3699	**1.1631**	**–0.6272**	–0.1963
	a5	–0.1801	**8.6938**	–0.2664	**1.6054**	**8.9708**	–0.2685
0.10	a1	**0.4234**	**–0.5153**	0.0021	**0.4412**	**–0.5150**	–0.0889
	a3	**0.6445**	–0.1684	–0.2803	**1.1298**	**–0.6198**	–0.1977
	a5	–0.2134	**4.3897**	–0.2358	**1.5734**	**8.4461**	**–0.4087**
0.15	a1	–0.1038	**–0.4218**	–0.2568	**0.4120**	**–0.5168**	–0.0896
	a3	**0.5556**	**–0.5610**	–0.2629	**1.0989**	**–0.6095**	–0.1991
	a5	–0.1607	**3.9751**	–0.1288	**1.5368**	**7.8865**	–0.2956
0.20	a1	0.3036	–0.3732	–0.2561	0.3744	**–0.5168**	–0.0896
	a3	–0.1244	**–0.5357**	–0.1745	**1.0712**	**–0.6019**	–0.2006
	a5	–0.0566	**2.4203**	–0.2840	**1.5028**	**7.1880**	–0.2970

表 4　辅助变量与研究变量共缺时纳入与不纳入辅助变量时回归系数的估计偏差 $_{S.E}$ 比较
(n = 500，中等偏高相关)

MR	AUX	MAR-MAR		MAR-MCAR		MAR-MNAR		MCAR-MAR		MCAR-MCAR		MCAR-MNAR	
		Z included	No Z	Z included	No Z	Z included	No Z	Z included	No Z	Z included	No Z	Z included	No Z
0.05	1	0.0469	**–0.4588**	–0.3646	**–2.1466**	**1.4419**	**–1.2277**	–0.0058	**–0.8048**	–0.0058	**–1.2625**	–0.0058	**–0.4986**
	3	0.0418	**1.4034**	–0.0526	**–4.2416**	0.1175	**–1.4980**	–0.0131	**1.9365**	–0.0146	**–4.1561**	–0.0087	**–0.4017**
	5	0.1334	**–3.7225**	–0.0480	**–26.2865**	0.2910	**–2.7701**	–0.0029	**2.4776**	–0.0029	**–26.6406**	0.0044	**–0.6177**
0.10	1	0.0374	**0.7844**	–0.0230	**–1.3159**	**–1.7935**	**–1.4928**	–0.0043	**0.8393**	–0.0058	**–1.2797**	–0.0044	–0.1164
	3	0.1449	0.3156	**–0.8075**	**–7.3373**	0.1105	**–1.5320**	–0.0130	**1.8892**	–0.0116	**–4.1764**	–0.0072	**–0.4102**
	5	0.2615	**–4.9350**	–0.1553	**–19.7723**	0.3471	**–3.1737**	–0.0029	**2.4364**	–0.0029	**–26.2012**	–0.0014	**–1.0154**
0.15	1	0.1432	**–1.8333**	**–0.7913**	**–3.2520**	0.0261	**–0.6556**	–0.0014	**0.7922**	–0.0028	**–1.2980**	–0.0029	–0.1154
	3	0.1958	0.2564	–0.1301	**–4.5317**	0.1107	**–1.6986**	–0.0157	**1.8429**	–0.0144	**–4.2121**	–0.0071	**–0.4178**
	5	**0.4418**	**–7.1851**	–0.1693	**–19.0169**	0.3388	**–3.3226**	–0.0014	**2.3860**	–0.0014	**–25.6260**	0.0216	**–0.6969**
0.20	1	0.1207	**0.5294**	**–0.9355**	**–3.5947**	0.0266	**–0.7276**	–0.0014	**0.7260**	–0.0042	**–1.3167**	–0.0014	–0.1160
	3	**–0.4426**	**–6.9937**	–0.1681	**–4.6797**	–0.0505	**–1.4407**	–0.0156	**1.8009**	–0.0172	**–4.2336**	0.0000	**–0.4206**
	5	–0.2518	**–9.8203**	–0.2358	**–17.3550**	**0.1852**	**–1.1406**	–0.0014	**2.3386**	–0.0014	**–24.7607**	0.0301	**–0.7033**

注：粗体表示结果出现偏差; MR=缺失率, AUX=辅助变量数, MAR-MAR=横线左边代表研究变量的缺失机制，右边代表辅助变量的缺失机制, MAR=随机缺失, MCAR=完全随机缺失, MNAR=非随机缺失, Z included=纳入辅助变量进行数据分析, No Z=不纳入辅助变量进行数据分析。

的可能性，而 MNAR 辅助变量与研究变量的缺失机制能够重合的机会较少。所以当辅助与研究变量共缺时，如果采用纳入辅助变量的方法进行缺失数据分析，不能因为辅助变量的缺失机制为 MNAR 而有过多的顾虑，因为共缺组合机制为 MAR-MNAR 或 MCAR-MNAR 所得到的结果比 MAR-MCAR 或 MCAR-MCAR 要好。另外，尽管 MAR 与 MCAR 都是可忽略缺失，但 MCAR 的假设更加严格(Rubin, 1976)。因此，在辅助变量与研究变量呈中等偏高相关，纳入辅助变量时，研究变量的缺失机制为 MAR 较 MCAR 更容易出现偏差。

最后，需要考虑的一个重要问题是如何判断实际数据的缺失机制是否满足模拟设计下的共缺机制情况，这涉及到缺失数据机制的检验问题。关于这个问题一直都是这个领域研究的难点，目前对其的研究也不多(孙婕，金勇进，戴明锋, 2013)。尽管我们在本研究中设计了几种缺失值机制的组合，但是并未涉及如何判断实际研究中如何检验其机制的问题，我们也没有打算这么做，因为这个问题超出了本研究的范围。但是，了解数据缺失的可能原因是必要的，可以根据经验猜想缺失的可能性，并通过事后调查或根据已收集到的基本信息进行判断。

5.2　相关程度的影响

本研究发现，当辅助变量与研究变量的相关只有 0.2~0.3 时，纳入辅助变量也有利于得到无偏估计。这一发现与先前的研究结果不同, Hardt 等(2012)发现相关太低时(r = 0.1 与 r = 0.5)辅助变量作用不大。Enders 和 Peugh (2004; r = 0.1 与 r = 0.3)也得到类似的结论。因为辅助变量与研究变量的相关越高，辅助变量能为研究变量提供的信息越多(Collins et al., 2001; Yoo, 2009)。然而，这可能是由于他们设定的相关太低(r = 0.1)，导致辅助变量的改善情况不明显。而且, Enders (2008)指出当辅助变量也存在缺失时，相对于辅助变量与研究变量高相关条件下的估计结果，中等相关条件下的估计结果更接近于参数或辅助变量完全时的结果。因此，根据本研究结果，当辅助变量与研究变量的相关达到 0.2~0.3 时，即可考虑纳入该辅助变量，尤其是当共缺组合机制为 MAR-MCAR 时。

本模拟结果还发现，在辅助变量不缺失的情况下，相关程度对研究结果影响不大，这可能与研究变量的缺失率较低、相关较低有关(Collins et al., 2001)。

5.3　辅助变量的数目

过往的研究发现纳入辅助变量对缺失数据建模是有益的(e.g., Collins et al., 2001; Enders, 2008)，本研究也支持这一结论。但是，很少有研究探讨纳入辅助变量的数目对参数估计的影响，本研究对此问题做了有益的尝试。本研究发现，当辅助变量与研究变量存在共缺时，对于 MAR-MAR 组合机制，纳入单个辅助变量是有益的；对于 MAR-MCAR 或 MAR-MNAR 组合机制，纳入多于一个辅助变量的

效果更好。

5.4 样本量的影响

在辅助变量不缺失的情况下, 样本量越大, 结果越好。当不纳入辅助变量进行分析时, 样本量越大, 出现偏差的情况越多。另外, 根据 Muthén 和 Muthén (2002)的观点, 在辅助变量对参数估计的影响中, 样本量并非独自起作用, 它还受到变量间的缺失率、缺失机制等因素的影响。结合本研究的结果, 对于回归系数偏差的参数估计(偏差 $_{S.E.}$、偏差 $_{100}$), 当辅助与研究变量呈低相关, 共缺率为 0.20, 辅助变量数为 3 个的时候, 如果样本量为 200 或 500, 在“MCAR-”组合机制条件下得到的偏差值最大; 而当样本量为 1000 时, 在“MCAR-”组合机制条件下得到的偏差值最小。控制相关程度、共缺率、辅助变量数不变的情况下, 结果表明“MCAR-”组合机制条件下得到的偏差值随着样本量的增大而减小。对于 Y 因子负荷偏差的参数估计(偏差 $_{S.E.}$、偏差 $_{100}$), 相同条件下, 如果样本量为 100、500 或 1000, 在“MCAR-”组合机制条件下得到的偏差值最小。控制相关程度、共缺率、辅助变量数不变的情况下, 结果表明“MCAR-”组合机制条件下得到的偏差值随着样本量的增大而减小。

5.5 共缺率的影响

本模拟结果仅发现纳入不缺失的辅助变量时, 相同样本量的情况下, 共缺率越大, Y 因子负荷的偏差越大, 其他条件下共缺率的影响并不明显。本研究通过观察所有变量的缺失模式, 计算辅助变量与研究变量共同缺失的比例, 从而得到共缺率。因此, 可能出现如下的情况: 辅助变量越多, 每个辅助变量与研究变量的平均共缺率越低, 以至于共缺率对参数估计的影响差别不大。奇怪的是, 在大样本量(n = 500 或 1000)、MAR-MAR 组合机制条件下, 随着共缺率的增大, 单个辅助变量时的参数估计的偏差总体呈增大的趋势, 但并非与共缺率同步增大。因此, 缺失数据研究中, 缺失率的影响有待进一步的研究。

5.6 不足与展望

本研究也存在一些不足: 第一, 本研究设定的共缺率较低, 这影响了共缺率对辅助变量作用的研究。以后的研究可以考虑模拟更高的共缺率, 以考察共缺率与共缺机制对辅助变量作用的影响。当然共缺率很高的情况在实践中并不常见, 因此本研究设置的共缺率更具有实践指导意义。第二, 虽然本模拟研究表明辅助变量与研究变量存在共缺时, 样本量并非越大越好, 但对于多大的样本量是合适的, 本研究并不能提供明确的参考。第三, 本研究主要模拟结构方程模型下辅助变量的效用, 对于结果能否推广到其他模型仍有待进一步的研究。另外, 本研究模拟的数据服从正态分布, 而实际研究中数据满足正态性的情况相对较少, 以后的研究可以考虑数据非正态的情况。总之, 本研究在前人研究的基础上对缺失值建模进行了更深入的分析, 当然缺失值建模领域尚存很多问题需要探索。

致谢: 作者非常感谢美国亚利桑那州立大学的 Craig Enders 博士在研究设计和数据模拟过程中给予的指导和帮助。作者同时感谢审稿专家在本文审稿过程中给予的指导和建议。

参 考 文 献

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced Structural Equation Modeling: Issues and Techniques*, *3*, 243–277.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330–351.

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. Hancock & R. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 313–342). Greenwich, CT: Information Age.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling, 15*, 434–448.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430–457.

Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling, 18*, 35–54.

Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling, 11*, 1–19.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10*, 80–100.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology, 12*, 184–196.

Kidger, J., Heron, J., Leon, D. A., Tilling, K., Lewis, G., & Gunnell, D. (2015). Self-reported school experience as a predictor of self-harm during adolescence: A prospective

cohort study in the South West of England (ALSPAC). *Journal of Affective Disorders, 173*, 163–169.

Mustillo, S. (2012). The effects of auxiliary variables on coefficient bias and efficiency in multiple imputation. *Sociological Methods & Research, 41*, 335–361.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431–462.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599–620.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*, 328–362.

Raykov, T. (2011). On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling, 18*, 419–429.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Schafer, J. L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park: Pennsylvania State University, Department of Statistics.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Sun, J., Jin, Y. J., & Dai, M. F. (2013). Discussion on testing the mechanism of missing data. *Mathematics in Practice and Theory, 43*, 166–173.

[孙婕, 金勇进, 戴明锋. (2013). 关于数据缺失机制的检验方法探讨. *数学的实践与认识,* 43, 166–173.]

Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology, 37*, 83–117.

Wang, M. C. (2014). *Latent variable modeling with Mplus*. Chongqing, China: Chongqing University Press.

[王孟成. (2014). *潜变量建模与 Mplus 应用.* 重庆: 重庆大学出版社.]

Wang, M. C., & Ye, H. S. (2014). Planned missing data design: Through intended missing data make research more effective. *Advances in Psychological Science, 22*, 1025– 1035.

[王孟成, 叶浩生. (2014). 计划缺失设计——通过有意缺失让研究更高效. *心理科学进展, 22*, 1025–1035.]

Yoo, J. E. (2009). The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educational and Psychological Measurement, 69*, 929–947.

Yuan, K. H., & Lu, L. (2008). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research, 43*, 621–652.

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research, 41*, 598–629.

The mechanism of auxiliary variables in full information maximum likelihood–based structural equation models with missing data

WANG Meng-Cheng; DENG Qiaowen

(*Department of Psychology*, *Guangzhou University*; *Center for Psychometric and Latent Variable Modeling*, *Guangzhou University*, *Guangzhou* 510006, *China*)

Abstract

In social and behavioral studies, missing data cannot be avoided in the process of data collection, especially in longitudinal studies. Because sample with missing data lose the balance characteristics of their complete counterparts, which may distort parameter estimates and degrade the performance of confidence intervals, special methods have to be developed for these analysis. Two modern missing data analysis techniques, maximum likelihood estimation and multiple imputation, have been widely studied in the methodological literature during the last decade. Since the maximum likelihood estimation and multiple imputation require the MAR (missing at random) assumption, including auxiliary variables can help fine-tune the missing data handling procedure, either by reducing bias or by increasing power. A useful auxiliary variable is a potential cause or a correlate of the incomplete variables in the analysis model. Notably, Graham (2003) proposed a "saturated correlates model", which allows us to include auxiliary variables in FIML-based structural equation models easily. However, some questions about the inclusion of auxiliary variables are needed to further study. The main research question was under what condition the auxiliary variables will be effective in the FIML-based structural equation modeling.

The current study investigates the effect of including auxiliary variables during estimation of structural equation modeling parameters with FIML estimation through Monte Carlo simulation. It focused on the missing values of the auxiliary variables and variables of interests simultaneously. The simulation repeated 5,000 times for each of 576 combinations: common missing rates (5 percent, 10 percent, 15 percent, and 20 percent), missing

mechanism combinations (MCAR-MCAR, MCAR-MAR, MCAR-MNAR, MAR-MCAR, MAR-MAR, and MAR-MNAR), correlations (low, moderate to high), number of auxiliary variables (1, 3, 5), and sample sizes (100, 200, 500, 1000). The evaluation criteria are bias and confidence intervals coverage of parameters. Data generates according to Enders (2008) model. All data generate and analyze by Mplus 7.0.

Auxiliary variables without missing values outperformed auxiliary variables with missing values. Including auxiliary variables which had missing values in the analysis procedure was found to improve parameter estimation efficiently in most cases. Results showed that the bias was more serious when the missing mechanism of the auxiliary variables was MCAR than MNAR. In the FIML-based structural equation modeling, the inclusion of more than a single auxiliary variable for MAR-MCAR or MAR-MNAR combined mechanisms is beneficial, while for MAR-MAR combined mechanism, a single auxiliary variable would be better. In addition, it is beneficial to include auxiliary variables which had low correlation with variables of interests in this model. However, simulation results indicated that the common missing rates had little impact on bias.

Overall, this study indicates that the inclusion of incomplete auxiliary variables is beneficial, even if the auxiliary variables and variables of interests have a relative proportion of missing data.

Key words missing data; missing mechanism; SEM; full information maximum likelihood; auxiliary variable; Monte Carlo simulation