因果中介效应分析的关键技术和多向分解方法

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍因果中介效应分析的5项关键技术和效应成分多向分解方法。5项关键技术的内容如下：①因果中介效应的识别；②因果中介效应分析的回归方法；③最大似然估计；④总效应与各种成分效应的估计；⑤自助法估计。多向分解方法包括3个双向分解、2个三向分解和1个四向分解。本文通过一个实例，借助SAS构建包含协变量和交互作用项的因果中介效应模型，对因果中介效应分析中的总效应进行双向分解、三向分解和四向分解，并对输出结果进行解释。

【关键词】 因果中介效应；效应识别；最大似然估计；自助法；多向分解
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20220911002

Key technology and multi-directional decomposition method of the causal mediation effect analysis

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
Corresponding author: Hu Liangping, E-mail: lphu927@163.com）

【Abstract】 The purpose of this paper was to introduce five key techniques and the multi-directional decomposition methods of effect components in the analysis of causal mediation effects. The contents of the five key technologies were as follows: ① identification of causal mediation effect; ② regression method of causal mediation effect analysis; ③ maximum likelihood estimation; ④ estimation of total effect and various component effects; ⑤ estimation by bootstrap method. The multi-directional decomposition methods included 3 bidirectional decompositions, 2 three-directional decompositions and 1 four-directional decomposition. Through an example, a causal mediation effect analysis model including covariates and interaction terms was constructed with the help of SAS, bidirectional decomposition, three-directional decomposition and four-directional decomposition were carried out for the total effect in the causal mediation effect analysis, and the output results were explained.

【Keywords】 Causal mediation effect; Effect identification; Maximum likelihood estimation; Bootstrap method; Multi-way decomposition

因果中介效应分析涉及多项关键技术，包括因果中介效应的识别、因果中介效应分析的回归方法、最大似然估计、总效应（total effect，TE）与各种成分效应的估计以及自助法估计。总效应可以分解成两种、三种或四种成分效应。本文先介绍前述提及的5项关键技术以及基于总效应分解出来的各种成分效应的含义，然后通过一个实例，采用SAS构建包含协变量和交互作用项的因果中介效应模型，对总效应进行双向、三向和四向分解。

1 因果中介效应分析的关键技术

1.1 因果中介效应的识别

为了便于识别因果中介效应，首先需要区分3类混淆协变量[1]。让C表示以下3类协变量：C_1代表混淆处理变量T与结果变量Y之间关系的通用协变量，它是一个处理变量与结果变量之间的混淆变量；C_2代表混淆中介变量M与结果变量Y之间关系的通用协变量，它是一个中介变量与结果变量之间的混淆变量；C_3代表一个混淆处理变量T与中介变量M之间关系的通用协变量，它是一个处理变量与中介变量之间的混淆变量。在回归分析中，控制C意味着所有类型的混淆协变量都被控制。

识别因果中介效应需要以下4个条件[2]：①在给定C的条件下，没有未测量的处理变量与结果变量之间的混淆变量；②没有未测量的处理变量与中介变量之间的混淆变量；③不存在受T直接或间接影响的中介变量与结果变量之间的混淆变量；④在

给定(C,T)的条件下,没有未测量的中介变量与结果变量之间的混淆变量。

受控直接效应(controlled direct effect,CDE)的识别需要上前述提及的前两个条件,自然直接效应(natural direct effect,NDE)和自然间接效应(natural indirect effect,NIE)的识别需要前述提及的所有4个条件,这4个条件统称为"无未测量混淆假设"[2-3]。在实践中,为了对因果中介效应进行有效的因果解释,研究者必须测量所有相关的混淆协变量C,并将其纳入因果中介效应分析中。

1.2 因果中介效应分析的回归方法

Proc causalmed过程采用回归方法来估计因果中介效应,该方法要求满足前文"1.1因果中介效应的识别"中关于识别条件的假设和以下两个模型的正确设定:①给定T、M和C的Y的结果模型;②给定T和C的M的中介模型。

对于一类广义线性模型,既往学者推导了用于计算不同变量类型的各种因果中介效应的分析公式[2,4-5],包括以下情况的组合:①结果变量Y,可以是二值、连续或计数的变量;②处理变量T,可以是二值或连续变量;③中介变量M,可以是二值或连续变量;④协变量C,可以是分类变量,也可以是连续变量。

基于上述提及的各种情况下的分析公式,proc causalmed过程能实现全部计算。对于具有二值结果变量和连续中介变量的情况,分析公式假设结果变量Y是罕见事件。如果Y不是罕见的、并且Y是使用对数连接函数建模的,那么公式仍然有效。

让θ表示收集结果和中介模型中所有参数的向量。在回归模型的正确设定和识别的条件下,中介分析中的因果效应是θ的函数,条件是协变量值。也就是说,因果效应(ef)可以表示为θ的函数,给定C=c,此函数的形式见式(1)。

$$g_{ef}(\theta|C = c) \quad (1)$$

式(1)中,c代表协变量C的一些固定值。对于连续结果变量,中介效应$g_{ef}(\theta|C = c)$在原始尺度上定义;对于二值结果变量,中介效应$g_{ef}(\theta|C = c)$在优势比或多余相对风险尺度上定义。有关公式可参见文献[2,6]。

由于模型中可能存在非线性和交互作用项,对于不同的协变量值集,因果效应$g_{ef}(\theta|C = c)$通常是不同的。在默认情况下,proc causalmed过程使用$c=c_0$计算$g_{ef}(\theta|C = c)$,其中,c_0为C的样本平均值。此默认设置提供了各种因果中介效应的"总体"度量。这与文献[2]中对SAS宏的处理一致。对于分类协变量,此默认计算仍然适用。分类协变量的平均值由分类水平的虚拟编码0-1计算得出。然后将这些平均值放入计算整体因果中介效应的公式中。然而,这并不意味着proc causalmed过程要求用户对分类协变量进行虚拟编码加以分析。

1.3 最大似然估计

对于随机样本,proc causalmed过程通过最大似然法估计因果中介效应。对于结果模型和中介模型,首先估计θ的最大似然估计$\hat{\theta}$。各种因果中介效应最大似然估计的计算见式(2)。

$$g_{ef}\left(\hat{\theta}|C = c_0\right) \quad (2)$$

式(2)中,*ef*为效应指数,$C=c_0$是从样本计算的协变量值的平均值。对于分类协变量,c_0的定义假设水平被虚拟编码为0和1,这是由proc causalmed过程内部完成的。

给定$\hat{\theta}$的估计协方差矩阵,使用增量法估计因果效应$g_{ef}\left(\hat{\theta}|C = c_0\right)$的标准误差。在计算这些估计值时,协变量值$c_0$被视为固定值。在这种情况下,计算标准误差的增量的具体方法可参阅文献[3-4]。

一般来说,因果中介效应的评估取决于协变量的水平。除了在$C=c_0$时评估总体因果中介效应外,还可以使用bootstrap语句提供特定的协变量水平,例如$C=c_1$。最大似然估计值为$g_{ef}\left(\hat{\theta}|C = c_1\right)$,标准误差也可通过Delta方法进行类似计算[7]。

1.4 总效应与各种成分效应的估计

1.4.1 估计方法的概述

对于连续结果变量,四向分解的组成部分在原始尺度上进行计算;对于二值结果变量,四向分解的组成部分根据优势比和多余相对风险量表进行计算。具体计算公式,此处从略,可参见文献[6]。

除了四向分解之外,proc causalmed过程还使用与四向分解相同的分析技术来估计其他几种双向和三向分解的成分效应,可以在evaluate或proc causalmed语句中使用decomp选项进行估计。

为了计算这些成分效应及其百分比贡献的标准误差估计值,proc causalmed过程使用带有分析导数的δ方法[1,7];bootstrap(即自助法)方法也可用于

计算标准误差和置信区间，具体方法可参见后文“1.5 自助法估计”。

1.4.2 双向、三向和四向分解的含义

双向分解是将总效应TE分解成两种成分效应，有3种形式，见式(3)、式(4)、式(5)：

$$TE=NDE+NIE \tag{3}$$

$$TE=CDE+PE \tag{4}$$

$$TE=TDE+PIE \tag{5}$$

式(3)、式(4)和式(5)中各成分效应的含义如下。NDE+NIE：自然直接效应和自然间接效应；CDE+PE：受控直接效应和部分被剔除的效应；TDE+PIE：总直接效应和纯间接效应。

三向分解是将总效应TE分解成3种成分效应，有2种形式，见式(6)、式(7)：

$$TE=NDE+PIE+IMD \tag{6}$$

$$TE=CDE+PIE+PAI \tag{7}$$

式(6)和式(7)中各成分效应的含义如下。NDE+PIE+IMD：自然直接效应、纯间接效应和中介交互作用项效应；CDE+PIE+PAI：受控直接效应、纯间接效应和归因于交互作用部分的效应。

四向分解是将总效应TE分解成4种成分效应，四向分解只有1种形式，见式(8)：

$$TE=CDE+IRF+IMD+PIE \tag{8}$$

式(8)中各成分效应的含义如下。CDE为受控直接效应，不是由交互作用项或中介变量产生的成分效应；IRF为参考交互作用效应，由交互作用项而非中介变量产生的成分效应；IMD为中介交互作用效应，由交互作用和中介变量引起的成分效应；PIE为纯间接效应，由中介变量而非交互作用项产生的成分效应。

1.5 自助法估计

如果指定bootstrap语句，proc causalmed过程将使用bootstrap重采样来计算因果中介效应和分解出的各成分效应的标准误差和置信区间。该过程根据用户在nboot=选项中指定的数量对自助样本数据集进行采样，然后进行计算。

Bootstrap置信区间仅针对效应及其相应的百分比进行计算，不是为结果或中介模型中的参数计算的。通过在bootstrap语句中使用bootci选项，可以指定以下一种或多种类型的自助置信区间。

Bootci(normal)选项要求系统基于正态近似法求自助置信区间。$(1-\alpha)100\%$ 正态自助置信区间见式(9)。

$$\hat{\mu}_j \pm \sigma_{\mu_j^*} z_{(1-\alpha/2)} \tag{9}$$

式(9)中，$\hat{\mu}_j$是原始样本中μ_j的估计值，$\sigma_{\mu_j^*}$是自助参数估计值的标准误差，$z_{(1-\alpha/2)}$是标准正态分布的第$100(1-\alpha)$百分位。

Bootci(perc)选项要求系统基于百分位数法求自助置信区间。置信区间是自助参数估计的第$100(\alpha/2)$和第$100(1-\alpha/2)$百分位，计算方法如下：设$\mu_{j,1}^*, \mu_{j,2}^*, \cdots, \mu_{j,B}^*$表示潜在结果平均值$\mu_j$的自助估计的有序值。设第k个加权平均百分位数为q，设$p=\frac{k}{100}$，并且，令：

$$Np=l+g \tag{10}$$

式(10)中，Np代表所求的百分位数，l是Np的整数部分，g是Np的小数部分。

第k个百分位q的计算见式(11)，它对应于proc univariate过程使用的默认百分位定义。

$$q=\begin{cases}\frac{1}{2}\left(\mu_{j,l}^*+\mu_{j,l+1}^*\right) & g=0\\ \mu_{j,l+1}^* & g>0\end{cases} \tag{11}$$

Bootci(bc)选项要求系统求偏差校正的bootstrap置信区间，该置信区间使用bootstrap参数估计的累积分布函数$G(\mu^*)$来确定置信区间的上下端点。偏差校正自助置信区间见式(12)。

$$G^{-1}\left[\Phi\left(2z_0 \pm z_{\alpha/2}\right)\right] \tag{12}$$

式(12)中，$\Phi(\cdot)$是标准正态累积分布函数，$z_{\alpha/2}=\Phi^{-1}(\alpha/2)$是标准正态分布横坐标轴上的一个分位数，其标准正态分布曲线下右侧尾端的概率为$\alpha/2$，z_0是偏移校正量，z_0的计算见式(13)。

$$z_0=\Phi^{-1}\left[\frac{N\left(\mu_j^* \leqslant \hat{\mu}_j\right)}{B}\right] \tag{13}$$

式(13)中，$\hat{\mu}_j$是输入数据集中μ_j的原始样本估计值，$N\left(\mu_j^* \leqslant \hat{\mu}_j\right)$是小于或等于$\hat{\mu}_j$的自助估计数$(\mu_j^*)$，B是获得处理效应估计数的自助重复数。

在默认情况下，使用偏差校正自助置信区间。Proc causalmed过程要求至少应有50个自助样本，如果产生的可用估计值少于40个样本，则不计算正态自助置信区间。Proc causalmed过程要求至少应有1 000个自助样本，如果产生的可用估计值少于900个样本，则不计算百分位数和偏差校正自助置信区间。如果在nboot=N选项中指定的样本数$N \leqslant 1\,000$，并且请求百分位或偏差校正自助置信区间，则忽略N的值。

2 因果中介效应分析的实例与SAS实现

2.1 实例与数据结构

2.1.1 资料来源与背景信息

【例1】文献[1]中的一个例子：仿照Marjoribanks讨论的理论教育模式[8]，模拟了一组包含6个变量、300个观测的数据集，旨在了解父母提供的鼓励性环境（Encourage）是否会影响儿童的认知发展（CogPerform）。关于数据集中6个变量及其含义参见文献[1]。试基于数据集中的数据，进行具有交互效应和四向分解的因果中介效应分析。

2.1.2 创建用于因果中介效应分析的数据集

因篇幅所限，创建数据集Cognitive的SAS程序见文献[1]。下面直接调用已创建的SAS数据集Cognitive。

2.2 用SAS实现因果中介效应分析

2.2.1 考虑协变量但不考虑交互作用项的因果中介效应分析

设所需要的SAS程序如下：

```
proc causalmed data=Cognitive;
model CogPerform=Encourage Motivation;
mediator Motivation=Encourage;
covar FamSize SocStatus;
run;
```

【SAS主要输出结果及解释】主要输出结果见表1。各效应项均有统计学意义。

表1 考虑协变量但不考虑交互作用项的总效应、直接效应和中介效应汇总

Table 1 Summary of total, direct and mediated effects considering covariates but excluding interaction term

各效应项	估　计	标准误差	Wald 95%置信区间	Wald χ^2	$Pr>\chi^2$
总效应	6.844	0.153	6.545~7.142	44.880	<0.010
受控直接效应（CDE）	4.296	0.110	4.081~4.511	39.140	<0.010
自然直接效应（NDE）	4.296	0.110	4.081~4.511	39.140	<0.010
自然间接效应（NIE）	2.547	0.156	2.241~2.854	16.300	<0.010
中介变量所占百分比	37.222%	1.752	33.787~40.656	21.240	<0.010
交互作用的百分比	0	–	–	–	–
剔除的百分比	37.222%	1.752	33.787~40.656	21.240	<0.010

2.2.2 考虑协变量和交互作用项的因果中介效应分析

通过在结果模型中包含Encourage变量与Motivation变量之间的交互作用项来扩展分析。设所需要的SAS程序如下：

```
proc causalmed data=Cognitive;
model CogPerform=Encourage | Motivation;
mediator Motivation=Encourage;
covar FamSize SocStatus;
run;
```

【SAS程序说明】model语句中设定Encourage | Motivation包括鼓励和动机的主效应及其交互作用效应，该语句的等价写法为：model CogPerform= Encourage Motivation Encourage * Motivation。

【SAS主要输出结果及解释】主要输出结果见表2。由表2可知，当包含交互作用项时，中介变量所占百分比略有变化，从37.222%（对于没有此项的模型，见表1）变为39.333%。虽然表2中交互作用的百分比具有统计学意义，为0.420%，但小于1.000%。因此，对结果的解释与没有交互作用项的分析结果并无实质性差异。

表2 考虑协变量和交互作用项的总效应、直接效应和中介效应汇总

Table 2 Summary of total, direct and mediated effects considering covariates and interaction term

各效应项	估　计	标准误差	Wald 95%置信区间	Wald χ^2	$Pr>\chi^2$
总效应	6.842	0.143	6.562~7.122	47.840	<0.010
受控直接效应（CDE）	4.180	0.047	4.088~4.272	89.000	<0.010
自然直接效应（NDE）	4.151	0.047	4.059~4.243	88.210	<0.010
自然间接效应（NIE）	2.691	0.145	2.407~2.976	18.530	<0.010
中介变量所占百分比	39.333%	1.370	36.647~42.018	28.700	<0.010
交互作用的百分比	0.420%	0.024	0.373~0.466	17.730	<0.010
剔除的百分比	38.913%	1.357	36.252~41.573	28.670	<0.010

2.2.3　考虑协变量和交互作用项并进行效应成分的多向分解

在第2.2.2节的SAS过程步程序中的第一句之后添加 decomp 选项(proc causalmed data=Cognitive decomp;)时,proc causalmed 过程将生成一个表,如表3所示,显示总体效应的各种分解,包括3个双向分解、2个三向分解和1个四向分解[1]。表3中,各种分解的成分均有统计学意义。

表3　考虑协变量和交互作用项的总效应的分解

Table 3　Decompositions of the total effect considering covariates and interaction term

分　解	效　应	估　计	标准误差	Wald 95% 置信区间	*Z*	*Pr*>|Z|
NDE+NIE	自然直接	4.151	0.047	4.059~4.243	88.210	<0.010
	自然间接	2.691	0.145	2.407~2.976	18.530	<0.010
CDE+PE	直接控制	4.180	0.047	4.088~4.272	89.000	<0.010
	剔除部分	2.663	0.144	2.381~2.944	18.520	<0.010
TDE+PIE	直接总计	4.208	0.047	4.116~4.300	89.630	<0.010
	纯间接	2.634	0.142	2.355~2.913	18.510	<0.010
NDE+PIE+IMD	自然直接	4.151	0.047	4.059~4.243	88.210	<0.010
	纯间接	2.634	0.142	2.355~2.913	18.510	<0.010
	中介交互	0.057	0.004	0.051~0.064	16.930	<0.010
CDE+PIE+PAI	直接控制	4.180	0.047	4.088~4.272	89.000	<0.010
	纯间接	2.634	0.142	2.355~2.913	18.510	<0.010
	交互作用部分	0.029	0.002	0.025~0.033	14.300	<0.010
四因子	直接控制	4.180	0.047	4.088~4.272	89.000	<0.010
	引用交互	−0.029	0.002	−0.033~−0.025	−14.300	<0.010
	中介交互	0.057	0.003	0.051~0.064	16.930	<0.010
	纯间接	2.634	0.142	2.355~2.913	18.510	<0.010
合计	总效应	6.842	0.143	6.562~7.122	47.840	<0.010

注:NDE=CDE+IRF,NIE=PIE+IMD,PAI=IRF+IMD,PE=PAI+PIE,TDE=CDE+PAI

3　讨论与小结

3.1　讨论

在进行因果中介效应分析时,除了要考虑协变量,还应将处理变量与中介变量之间的交互作用项一并纳入统计模型进行分析,以更精准地估计因果中介效应模型中的各效应项。对于总效应的分解,有3种双向分解,即每种双向分解将总效应分解成2项,例如,第一个双向分解“NDE+NIE”就是将总效应分解成自然直接效应(NDE)与自然间接效应(NIE)之和;同理,可理解2个三向分解以及1个四向分解的含义。

3.2　小结

本文介绍了因果中介效应分析的5项关键技术,包括因果中介效应的识别、因果中介效应分析的回归方法、最大似然估计、总效应与各种分解的成分效应的估计以及自助法估计。基于一个实例,展示了将交互作用项纳入因果中介效应模型进行分析的方法,还借助 decomp 选项实现了对总效应进行3个双向分解、2个三向分解和1个四向分解。

参考文献

[1] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2301-2364.

[2] Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros[J]. Psychol Methods, 2013, 18(2): 137-150.

[3] VanderWeele TJ. Explanation in causal inference: methods for mediation and interaction[M]. New York: Oxford University Press, 2015: 138-210.

[4] VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition[J]. Stat Interface, 2009, 2(4): 457-468.

[5] VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome[J]. Am J Epidemiol, 2010, 172(12): 1339-1348.

[6] VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition[J]. Epidemiology, 2014, 25(5): 749-761.

[7] 周勇. 广义估计方程估计方法[M]. 北京: 科学出版社, 2013: 67-71.
Zhou Y. Estimation method of generalized estimation equation[M]. Beijing: Science Press, 2013: 67-71.

[8] Marjoribanks K. Environments for learning[M]. London: National Foundation for Educational Research Publications, 1974: 68-127.

(收稿日期:2022-09-11)

(本文编辑:陈　霞)