·科研方法专题·

基于因果图模型构造和搜索调整集

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图模型的基础知识、因果图过程的内容以及基于SAS/STAT中的CAUSALGRAPH过程构造和搜索调整集的方法。因果图模型是图论与概率论相结合的产物,它可以基于用户设定的变量之间的作用关系找到包含最小调整集在内的所有可能的调整集。因果图过程的内容主要包括三种识别标准、两种操作模式和一种验证检查方法。本文基于SAS中因果图过程对两个实例进行因果效应分析,并对输出结果做出解释。

【关键词】 因果图模型;因果效应;处理变量;工具变量;调整集

中图分类号:R195.1 **文献标识码**:A doi:10.11886/scjsws20220710002

Constructing and searching adjustment sets based on a causal graph model

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 the basic knowledge of the causal graph model, the contents of the CAUSALGRAPH procedure and the method of constructing and searching adjustment sets based on the CAUSALGRAPH procedure in SAS/STAT. The causal graph model was the product of the combination of graph theory and probability theory. It could find all possible adjustment sets including the minimum adjustment set based on the action relationship between the variables set by the user. The contents of the CAUSALGRAPH procedure mainly included three identification criteria, two operating modes and one verification checking method. This paper analyzed the causal effect of two instances based on the CAUSALGRAPH procedure in SAS, and explained the output results.

【**Keywords**】 Causal graph model; Causal effect; Treatment variable; Instrumental variable; Adjustment set

因果图模型是一种呈现多因素或多变量对结果变量影响情况的图形表达方式,它有别于以统计公式表达的多重回归模型,但从输出结果的实际效果来看,它与多重回归模型分析给出的自变量筛选结果非常相似。本文将介绍因果图模型的基础知识和基于SAS/STAT中因果图过程实现因果图模型分析的方法。

1 因果图过程简介

1.1 无需数据的因果图模型

无论是简单的假设检验和区间估计[1-2],还是复杂的多因素和多元统计分析[3-4],都必须基于统计数据进行统计分析。然而,SAS 9.4版本的SAS/STAT中[5]给出了一个“PROC CAUSALGRAPH”过程,它不需要任何统计数据就可按因果图中设定的变量之间的因果关系进行模拟和计算,并输出可能对因变量有影响的全部协变量(称为调整集)。构造和搜索调整集的过程和结果类似于多重回归分析中变量筛选的过程和结果。调整集是一组变量,可用于消除因果图模型中处理变量(即研究者着重考察的自变量)和结果变量之间的非因果关联。如果存在调整集,则可确定处理变量对结果变量存在因果效应。

1.2 因果图模型的理论基础

因果图模型就是用图形的形式呈现统计模型中变量之间依赖关系或因果关系的一种方法,它是图论与概率论相结合的产物。图论是一门古老的数学分支,主要研究用某种方式联系起来的若干事物之间的二元或多元关系。自20世纪40年代埃尔德什首次引入概率方法以来,特别是近些年

来，概率方法在图论中得到了深入的发展，并且日渐成为研究中的一个有力工具。由于研究方法和内容的不同，图论已产生了若干分支，如代数图论、极值图论、随机图论、因果图论、拓扑图论和应用图论等[6-7]。

1.3 因果图过程的内容

1.3.1 三种识别标准

因果图过程为确定因果处理效应提供了几个标准。用户可以使用因果图过程语句中的“METHOD=选项”指定以下任一识别标准：构造性后门标准[8]、后门标准[9]及工具变量[10]。

构造性后门标准（METHOD=adjustment）也称为调整标准，用于查找仅由观测变量组成的所有有效调整集。后门标准（METHOD=backdoor）同样可以找到由观测变量组成的调整集，但标准稍强一些。后门标准在计算上比调整标准效率更高，但它可能无法找到所有可能的有效调整集。后门标准的吸引力在于它具有直观的解释，并提供了一种快速构建有效调整集的方法[11]。工具变量法（METHOD=IV）寻找工具变量，以处置处理变量和结果变量之间存在的未测量的混淆变量。由于未测量的混淆变量可能会导致调整标准和后门标准失效，故需要采用工具变量予以调整。

1.3.2 两种操作模式

为了识别调整集或工具变量集，因果图过程有两种主要操作模式：其一，因果图过程语句中的列表选项使用户能够列举可用于估计因果效应的标准；其二，TESTID 语句允许检验用户指定的标准是否适用于估计因果效应。用户可以在一次运行过程中同时使用这两种模式；可以使用各种选项来微调所请求标准的输出列表；可以使用这些选项来限制列出的条件的数量，对列出的条件进行排序，提高搜索和列出的效率等。

在因果图过程中，每个因果图模型都必须是有向无环图（Directed acyclic graph，DAG）。用户可以使用 MODEL 语句输入因果图。MODEL 语句支持类似路径的语法来输入变量之间的因果关系，例如，要指定因果路径 X→Y，可以在 MODEL 语句中使用 X==>Y 或 Y<==X 语法。还可以将多个因果关系指定为因果路径链，例如，X==>Y==>Z，Z<==X==>Y<==W 等。因果路径中的每条边（指两变量之间的连线或箭头）表示一个变量对另一个变量的直接因果效应。

1.3.3 验证检查方法

因果图过程对指定的每个模型执行以下语义验证检查：其一，模型应弱连接，也就是说，当因果路径中的所有边都被视为无向时，任何一对变量之间都应该有一条路径；其二，模型不能包含任何定向循环。

因果图过程还支持指定双向边（或路径）。双向边语法，例如 X<=>Y（对于 X 和 Y），被解释为两个变量之间的未测量混淆，故图形仍然是 DAG。也就是说，X<=>Y 相当于 X<=L=>Y（对于 X、L、Y），其中节点 L 代表一些未测量的变量，用户可在 UNMEASURED 语句中指定这些变量。

在因果图模型分析中，区分测量变量和未测量变量很重要。在 UNMEASURED 语句中，列出的变量将被视为未测量或未观测的变量，所有其他变量均被视为已测量或观测到的变量。为了使因果效应评估有意义，必须始终测量用户指定的处理变量和结果变量。因果图模型中的未测量变量不能包含在统计分析中，因此，用户不能在因果处理效应的任何识别标准中使用它们。

2 因果图过程的应用

2.1 构造调整集

2.1.1 实例及其背景信息

【例 1】图 1 所示的因果图模型（改编自文献[12]）呈现了法罗群岛居民中母亲接触持久性全氟烷基物质（PFAS）与母乳喂养持续时间（Duration）之间的关系。

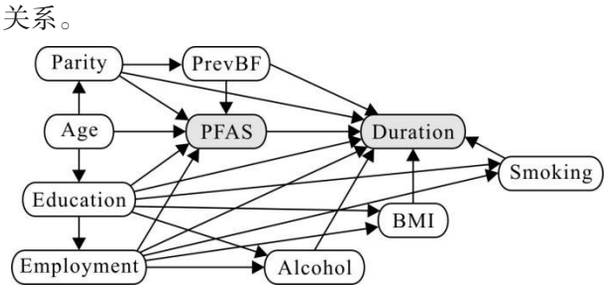

图 1 持久性全氟烷基物质对母乳喂养持续时间影响的因果图模型

Figure 1 Causal model of the effect of persistent perfluoroalkyl substances on breastfeeding duration

图 1 模型中各变量的含义如下：PFAS，持久性全氟烷基物质（危险因素或处理变量）；Duration，持续时间（结果变量）；Age，孩子出生时母亲的年

龄；Education，教育（母亲接受初等教育的时间）；Employment，就业（描述母亲就业状况的分类变量）；Parity，胎次（是否为母亲的第一次分娩的指标）；Alcohol，饮酒（母亲在怀孕期间是否饮酒的指标）；Smoking，吸烟（母亲在怀孕期间是否吸烟的指标）；BMI，体重指数（母亲孕前体重指数）；PrevBF，先前是否有母乳喂养经验。

假设在此例中，未观测到饮酒变量和吸烟变量，其他变量都被观测到了。试通过图 1 中设定的变量之间的关系，使用因果图过程来确定因果图模型中必须控制的协变量，以便估计具备有效因果解释的因果效应。

2.1.2 用 SAS 实现因果效应分析

2.1.2.1 输出包含变量数尽可能多的调整集

为了确定变量 PFAS 对变量 Duration 的因果效应，用户可能会考虑一个调整集，包括所有观察到的协变量。以下语句调用 CAUSALGRAPH 过程来检验此调整集是否有效。设所需要的 SAS 程序如下：

```
proc causalgraph;
model "Timm17AllObs"
Age ==> Parity PFAS Education,
Parity ==> PrevBF Duration PFAS,
PrevBF ==> PFAS Duration,
PFAS ==> Duration,
Education ==> Duration Employment PFAS BMI Alcohol Smoking,
Employment ==> Duration PFAS BMI Alcohol Smoking,
BMI Alcohol Smoking ==> Duration;
Identify PFAS ==> Duration;
testid "All Covariates" Age Education Employment Parity
AlcoholSmoking BMI PrevBF;
run;
```

【SAS 程序说明】在 MODEL 语句中，指定要分析的因果图模型。语句中带引号的字符串标记模型的名称；MODEL 语句的其余部分指定了模型中的所有变量和边。这些变量和边反映了图 1 所示的假设数据生成过程。在 IDENTIFY 语句中，用户指定了感兴趣的因果效应。用户可以使用此语句指定一个或多个处理变量以及结果变量。处理变量与结果变量之间由一个“==>”符号隔开。在本例中，用户感兴趣的是检验处理变量 PFAS 对结果变量 Duration 的因果效应的识别。由于 PROC CAUSALGRAPH 语句中未指定 METHOD=选项，故该过程默认使用构造性后门标准（METHOD=adjustment），以检验用户在 TESTID 语句中指定的调整集对因果效应的识别。

【SAS 输出结果及解释】

第 1 部分输出结果：因果图模型中设定的 10 个测量变量，包括处理变量（PFAS）、结果变量（Duration）和协变量（Age、Alcohol、BMI、Education、Employment、Parity、PrevBF、Smoking）。没有未测量的变量。

第 2 部分输出结果：图形模型汇总结果，见表 1。

表 1 图形模型汇总

Table 1 Graphical model summary

模型名称	节点	边界	处理	结果	测量的	未测量的
Timm17AllObs	10	23	1	1	10	0

【表 1 中有关内容的说明】第 2 列的“节点”指因果图模型中包含的全部变量的个数（本例为 10 个）；第 3 列的“边界”指因果图模型中带箭头的线条数（本例为 23 条）；第 4 列的“处理”指因果图模型中处理变量的个数（本例为 1 个）；第 5 列的“结果”指因果图模型中结果变量的个数（本例为 1 个）；第 6 列的“测量的”指因果图模型中测量变量的个数（本例为 10 个）；第 7 列的“未测量的”指因果图模型中未测量变量的个数（本例为 0 个）。用户可以基于这些输出内容用作模型设定的定性检查。

第 3 部分输出结果：协变量调整检验的结果，见表 2。

表 2 基于 8 个协变量调整检验呈现 PFAS 对 Duration 的因果效应

Table 2 Causal effect of PFAS on Duration presented based on the adjustment test of 8 covariates

模型名称	大小	有效	最小	C1	C2	C3	C4	C5	C6	C7	C8
Timm17AllObs	8	是	否	*	*	*	*	*	*	*	*

注：C1~C8 分别代表协变量 Age、Alcohol、BMI、Education、Employment、Parity、PrevBF、Smoking；*协变量对结果变量的影响具有统计学意义

【表 2 中有关内容的说明】第 2 列的“大小”指协变量的个数（本例有 8 个）；第 3 列的“有效”指协变量的集合对检验 PFAS 对 Duration 的因果效应是否有效（本例经检验，其结果为“有效”，输出中用“是”

表示）；第4列的“最小”指所找到的调整集是否为最小的调整集（本例的调整集包含8个协变量，故它不是最小的调整集）。

根据计算的结果可知，基于由8个协变量组成的调整集足以确定PFAS对Duration的因果关系，但它不是一个最小的调整集。如果使用此调整集，因果效应的估计可能在计算上效率较低。此外，用户必需收集所有这些变量的数据，以估计因果效应。

2.1.2.2 输出所有可能的调整集

用户可以使用因果图过程查看是否有任何较小的调整集可用于识别图1所示的因果效应。以下语句列出了所有可能的调整集，可用于估计PFAS对Duration的因果效应。

在上文“2.1.2.1节”的SAS程序中，删除“TESTID语句”，输出结果如下：

【SAS输出结果及解释】

输出结果的形式与表1类似，所有的调整集共有16组。这些调整集包含变量的个数分别为4、5、6、7、8个，其中，含4个变量的调整集只有1组，这4个变量分别是Education、Employment、Parity、PrevBF，它是本例中最小的调整集。也就是说，在研究PFAS对Duration的因果效应时，必需观测的最少的协变量个数为4个。含5、6、7、8个协变量的调整集分别有4、6、4、1组。因篇幅所限，详细输出结果从略。

2.1.2.3 仅输出最小的调整集

可以在PROC CAUSALGRAPH语句中使用MAXLIST=、MAXSIZE=或MINIMAL=选项来减少输出调整集的数量。例如，将上文“2.1.2.2节”SAS程序的第1句修改为：

```
proc causalgraph minimal;
```

于是，就只输出一行仅包含Education、Employment、Parity、PrevBF这4个变量的最小调整集。具体输结果从略。

2.1.2.4 存在未观测变量时寻找调整集

若在上文图1中，Alcohol和Smoking两个变量未观测到，如何构建调整集？所需要的SAS程序如下：

在上文“2.1.2.1节”的SAS程序中，删除“TESTID语句”，再增加以下未测量语句：

```
unmeasured Alcohol Smoking;
```

【SAS输出结果及解释】

寻找全部有效的调整集结果，见表3。

表3 全部有效的调整集

Table 3 All valid adjustment sets

编 号	大小	最小	C1	C3	C4	C5	C6	C7
1	4	是			*	*	*	*
2	5	否	*		*	*	*	*
3	5	否		*	*	*	*	*
4	6	否	*	*	*	*	*	*

注：C1、C3~C7分别代表协变量Age、BMI、Education、Employment、Parity、PrevBF；*协变量对结果变量的影响具有统计学意义

表3中各列和各行内容的含义，参见上文中“表2中有关内容的说明”，此处从略。由表3可知，共有4个有效的调整集。每一行都包含一个调整集，第1行为最小调整集。假设因果图模型是准确的，用户可以使用这些调整集中的任何一个来估计PFAS对Duration的因果效应。

2.2 高效搜索调整集

2.2.1 实例及其背景信息

【例2】沿用例1的资料和背景信息，不同的是：假定因果图模型包括一个额外的变量HealthBehavior，它被认为是一个潜在的结构（简称潜变量或隐变量），代表一个人的行为被认为是健康的程度；同时假设变量HealthBehavior和PrevBF未被观测到。此时，对应的因果图模型见图2。

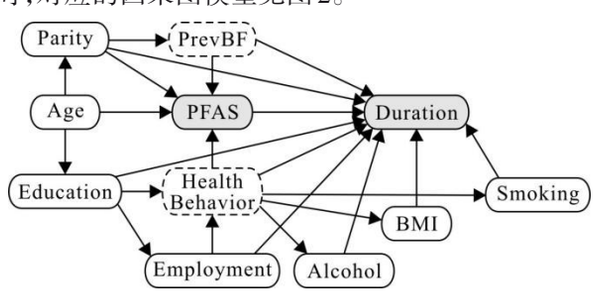

图2 持久性全氟烷基物质对母乳喂养持续时间影响的因果图模型

Figure 2 Causal model of the effect of persistent perfluoroalkyl substances on breastfeeding duration

图2中各变量的含义见前面的例1，此处从略。试通过图2中设定的变量之间的关系，使用因果图过程快速确定特定因果图模型中是否存在调整集。如果存在调整集，则确定处理变量对结果变量的因果效应。

2.2.2 用SAS实现因果效应分析

以下语句调用因果图过程，以确定是否有可能找到用于估计因果效应的调整集。设所需要的SAS程序如下：

```
proc causalgraph method=adjustment maxlist=1 nosort;
model "Timm17HealthBehavior"
Age ==> Parity PFAS Education,
Parity ==> PrevBF Duration PFAS,
PrevBF ==> PFAS Duration,
PFAS ==> Duration,
Education ==> Duration HealthBehavior Employment,
HealthBehavior ==> PFAS Duration BMI Alcohol Smoking,
Employment ==> HealthBehavior Duration,
BMI Alcohol Smoking ==> Duration;
identify PFAS ==> Duration;
unmeasured PrevBF HealthBehavior;
run;
```

【SAS输出结果及解释】

NOTE:没有满足“Timm17HealthBehavior”的指定准则的调整设置。

由以上输出结果可知,对于图2中的因果图模型,没有满足Timm17HealthBehavior指定标准的调整集。也就是说,不可能使用调整集来确定PFAS对Duration的因果效应。

尽管无法使用调整集来估计图2中的因果效应,但如果用户愿意在模型中做出额外的参数假设,仍然可以估计因果效应。因篇幅所限,此处从略。

3 讨论与小结

3.1 讨论

在常规的统计分析中,需要先给定统计数据,才能选择统计分析方法对数据进行分析。然而,将图论方法与概率论知识有机结合起来,只要能结合专业知识绘制出反映变量之间因果关系的因果图,就可计算出所有可能的调整集。这样就可以基于研究者的人力、物力、财力和时间,制定合适的研究方案,有针对性地收集资料,进而提高科研工作效率,节省科研经费。

3.2 小结

本文介绍了因果图模型的理论基础、因果图过程的内容(包括三种识别标准、两种操作模式和一种验证检查方法)以及基于SAS软件对两个实例进行了因果图模型的分析,输出所有可能的调整集和最小调整集。

参考文献

[1] 秦玄,王婷,王心羽,等.住院精神障碍患者家属照护过程的质性研究:基于生物生态学理论[J].四川精神卫生,2020,33(1):49-52.
Qin X, Wang T, Wang X, et al. Qualitative study on the care process of family members of inpatients with mental disorders: a perspective of bio-ecological theory[J]. Sichuan Mental Health, 2020, 33(1): 49-52.

[2] 张炳智,郑在江,田国娇,等.雅安市社区在管严重精神障碍患者现状研究[J].四川精神卫生,2020,33(1):53-56,60.
Zhang B, Zheng Z, Tian G, et al. Study on the status of patients with severe mental disorders under management in Ya'an[J]. Sichuan Mental Health, 2020, 33(1): 53-56, 60.

[3] 徐海婷,刘嫣然,吕婧,等.未治疗抑郁障碍患者自杀风险与认知情绪调节策略的关系[J].四川精神卫生,2020,33(1):44-48.
Xu H, Liu Y, Lyu J, et al. Association between suicide risk and cognitive emotion regulation strategiesin untreated depressive disorder patients[J]. Sichuan Mental Health, 2020, 33(1): 44-48.

[4] 孙彬.基于结构方程模型的精神疾病患者照料者抑郁焦虑的影响因素研究[J].四川精神卫生,2020,33(1):61-66.
Sun B. Analysis of influencing factors of depression and anxiety status among caregivers of patients with mental illness based on structural equation model[J]. Sichuan Mental Health, 2020, 33(1): 61-66.

[5] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2243-2300.

[6] 王元.数学大辞典[M].2版.北京:科学出版社,2017:658-688.
Wang Y. Dictionary of mathematics[M]. 2nd edition. Beijing: Science Press, 2017: 658-688.

[7] Béla Bollobás. Random graphs[M]. 2nd edition. Cambridge: Cambridge University Press, 2001: 34-159.

[8] van der Zander B, Liskiewicz M, Textor J. Constructing separators and adjustment sets in ancestral graphs[A]. Proceedings of the UAI 2014 conference on causal inference: learning and prediction[C]. Aachen: CEUR-WS.org, 2014: 11-24.

[9] Pearl J. Causality: models, reasoning, and inference[M]. 2nd edition. Cambridge: Cambridge University Press, 2009: 34-69.

[10] van der Zander B, Textor J, Liskiewicz M. Efficiently finding conditional instruments for causal inference[A]. Proceedings of the 24th international conference on artificial intelligence[C]. Menlo Park, CA: AAAI Press, 2015: 3243-3249.

[11] Morgan SL. Handbook of causal analysis for social research[M]. Dordrecht: Springer, 2013: 245-273.

[12] Timmermann CAG, Budtz-Jørgensen E, Petersen MS, et al. Shorter duration of breastfeeding at elevated exposures to perfluoroalkyl substances[J]. Reprod Toxicol, 2017, 68: 164-170.

(收稿日期:2022-07-10)
(本文编辑:戴浩然)