基于因果图模型检验调整集和寻找共用调整集

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】 本文目的是介绍基于因果图模型检验调整集、寻找共用调整集以及用SAS软件实现统计计算。首先，介绍与因果图模型有关的基本概念；其次，介绍因果图理论的初级内容，包括因果图的组成和术语；最后，针对两个实例并借助SAS/STAT中的CAUSALGRAPH过程，完成以下两项任务：①检验调整集和枚举路径；②寻找多个因果图模型共用的调整集。

【关键词】 因果图模型；因果效应；结点和边；潜在结构；调整集

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20220710003

Checking adjustment sets and finding common 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 method of checking adjustment sets based on a causal graph model, finding common adjustment sets and implementing the statistical calculation with SAS software. Firstly, the basic concepts related to the causal graph model were introduced. Secondly, the primary contents of the causal graph theory were given, including the composition and terminology of the causality diagram. Finally, for the two instances and with the help of the CAUSALGRAPH procedure in SAS/STAT, the following two tasks were completed: the first task was to examine the adjustment set and enumerate paths; the second task was to find the adjustment set common to the multiple causal graph models.

【**Keywords**】 Causal graph model; Causal effect; Node and edge; Latent structure; Adjustment set

基于因果图中设定的变量之间的关系，不仅可以构建和搜索调整集，也可以检验调整集，还可以基于设定的多个因果图模型寻找共用的调整集。本文先介绍因果图的基础知识和理论，再结合两个实例并基于SAS软件检验调整集和寻找共用调整集。

1 与因果图有关的基本概念

1.1 非随机对照研究可能引起虚假关联

在非随机研究中，观察单位不会被随机分配到处理组和对照组。因此，一些协变量可以与组分配和结果变量相关联。在这种情况下，结果变量的值由处理的因果效应以及与协变量的虚假关联决定。由于协变量混淆了因果关系，如果没有某种形式的调整（例如匹配或分层）来消除虚假关联，就无法确定处理对结果的因果关系[1]。事实上，在不完善的随机对照研究中，由混杂协变量引起的虚假关联也有可能存在。

研究者可以使用因果图过程来确定在何时、何种情况下估计因果效应[2]。为此，需要以有向无环图（Directed acyclic graph，DAG）的形式定义因果图模型。因果图过程接受输入DAG，并输出对因果分析和效应估计有用的结果。

1.2 统计和因果概念

通常情况下，研究者可以运用统计模型来描述和分析变量之间的依赖关系，例如，分析某个特定事件发生的概率有多大，以及随着观察到不同的变量，该概率将如何变化[3]。然而，统计模型中的某些变量之间可能确实存在一定程度的关联，而另一些变量之间也可能存在虚假的关联，此时，仅凭统计方法不足以揭示变量之间的因果效应[4]。这是因为变量之间存在的虚假关联会在相当大的程度上产生误导[5]。为了通过非随机试验的数据来回答因果效应，必须用一组因果假设来补充联合分布函数，这些假设与专业上已知的信息共同构成因果图模型。

2 因果图理论的初级内容

2.1 因果图的组成

因果图模型由DAG的形式表示，DAG由三部分组成[4]，即节点、边和缺失边。

节点：DAG中的每个节点代表一个变量，假设该变量在研究的过程中起因或果的作用；每个变量可以有任何分布；DAG中的每个变量可以是已观测的，也可以是未观测的，但未观测的变量不宜过多；通常，误差随机变量（独立误差项）不在DAG中表示。

边：DAG中的所有边都是定向的，也就是说，边由从一个节点指向另一个节点的箭头组成，且一对节点之间最多只能有一条边；边是因果假设的图形表示，即DAG中的边表示一个变量对另一个变量可能产生的直接因果效应；假设这些因果效应是确定性的，但它们是完全非参数的，因为每条边都可以有任何函数形式；因为DAG中的每条边都给出了因果解释，所以每条边都与一对节点的时间顺序相关联，因此，DAG不能包含有向循环。CAUSALGRAPH过程对每个模型执行语义验证，以验证它不包含有向循环。PROC CAUSALGRAPH允许指定双向边。双向边被解释为两个变量之间未测量的混杂，图形仍然是DAG。例如，式（1）与式（2）中的符号“↔”与“←L→”被解释为一对边：

$$X \leftrightarrow Y \quad (1)$$

$$X \leftarrow L \rightarrow Y \quad (2)$$

其中，节点L代表一个或一些未测量的变量。

缺失边：DAG中缺失边（即两个节点不通过边直接连接）表示直接因果效应为零的假设。因此，DAG中缺失的边表示比边具有更强的假设。这是图形模型的强零假设，它在计量经济学文献中被称为排除限制。缺失边对因果模型所隐含的统计特性有影响。

DAG中的节点、边和缺失边一起形成一个因果图模型，对研究者关于数据生成过程的假设进行编码。从这些生成数据的假设开始，研究者可以使用一组在DAG上操作的图形规则来导出统计关联关系。

2.2 因果图中的基本术语

因果图中的基本术语：①相邻，如果DAG中的两个变量由一条边直接连接，则它们是相邻的。②路径，路径是一个有序的变量列表，其中没有变量出现超过一次，列表中的连续变量在图中相邻，连接路径中连续节点的边可以指向任一方向。③因果关系与非因果关系，对于路径上的每一对连续变量，如果连接两个变量的箭头指向后一个变量，那么路径就是因果关系，否则，就称为非因果关系。④正确路径，如果路径以处理变量（即研究者关注的重要影响因素）开始，且不包含任何其他处理变量，则该路径是正确的[2]。⑤关系描述，DAG编码因果图模型中变量之间的特定关系的描述，对于一个边，如P→Q，其中，P是Q的父项，Q是P的子项；如果变量S到变量T之间存在因果路径，那么T是S的后代，S是T的祖先，因此，变量S的后代集是由S直接或间接引起的所有变量的集合，同样，变量T的祖先集是T的所有直接或间接原因的集合。⑥碰撞器，对于路径上的变量V，如果V有两个指向它的箭头（每侧一个），则V是路径上的碰撞器。否则，就是非碰撞器，碰撞器的定义是特定于路径的，变量可以是一条路径上的碰撞器，也可以是另一条路径上的非碰撞器。⑦因果与非因果关联，变量集之间的统计关联可分为因果部分和非因果部分或虚假部分，如果所有虚假的关联都能被消除，那么因果关系就被认为是确定的[2]。⑧调整或消除虚假关联，虚假关联通常通过某种形式的统计调整或条件作用来消除，例如，研究者可以通过在回归模型中包含一个变量作为回归自变量，或者通过按变量的水平对资料分层来进行调整计算，还可以使用因果图模型来确定调整集中必须包含哪些变量。

3 因果图过程的应用

3.1 检验调整集和枚举路径

3.1.1 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[6]）用于检验受试者血清尿酸盐（Urate）与心血管疾病（CVD）风险之间的关系。

图1模型中各变量的含义如下：Urate，尿酸盐（处理变量）；CVD，心血管疾病（结果变量）；AntiHypertensiveUse，抗高血压药物使用；Creatinine，肌酐（测量的血清肌酐水平）；Diabetes，糖尿病；Ethnicity，种族；Gender，性别；Gout，痛风；HbA1c，糖化血红蛋白；MedicationPropensity，药物倾向（反映个人服用处方药倾向的潜在变量）；Nutrition，营养（反映饮食或营养的潜在变量）；Obesity，肥胖（体重指数指标）；CurrentBP，当前血压；CurrentHDL，当前高密度

脂蛋白(高密度脂蛋白胆固醇指标);PreviousBP,先前血压(研究前血压指标);PreviousHDL,先前高密度脂蛋白(研究前高密度脂蛋白胆固醇指标);Smoking,吸烟(当前吸烟状态指标);StatinUse,他汀类药物使用。

变量 MedicationPropensity 和 Nutrition 对应潜在变量,故无法观察到。还假设未观察到变量 PreviousBP、PreviousHDL 和 Obesity。在考虑诸多协变量的前提下,试分析处理变量 Urate 对结果变量 CVD 是否具有因果效应。

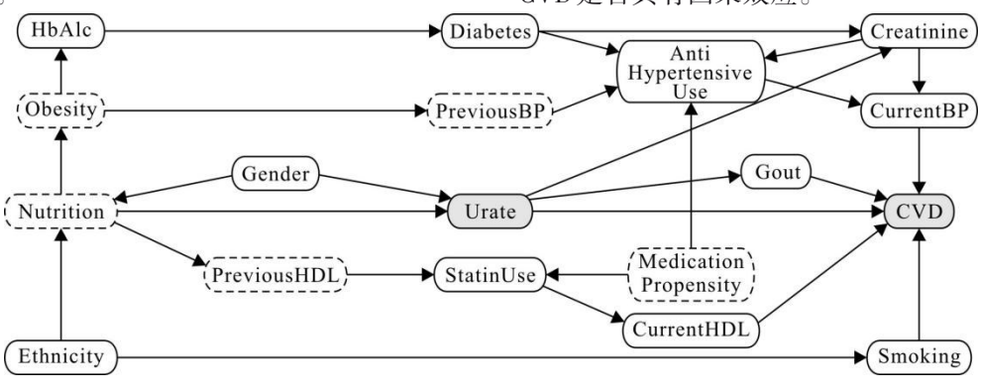

图 1 血清尿酸盐对心血管疾病风险影响的因果图模型

Figure 1 Causal model of the effect of serum urate on risk of cardiovascular disease

3.1.2 用 SAS 实现因果效应分析

3.1.2.1 检验调整集

根据图 1 所显示的因果图模型,变量 Urate 和 CVD 之间的统计关联反映了真实因果关联和附加虚假或非因果关联的组合。为了揭示这两个变量之间的真正因果联系,Thornley 等[6]考虑对变量 CurrentHDL、Ethnicity、Gender、HbA1c 和 Smoking 进行调整。以下程序调用因果图过程,以检验该调整集是否可用于根据因果图模型估计变量 Urate 对 CVD 的因果效应。设所需要的 SAS 程序如下:

```
proc causalgraph compact;
model "Thor12"
AntiHypertensiveUse ==> CurrentBP,
Creatinine ==> AntiHypertensiveUse CurrentBP,
CurrentBP ==> CVD, CurrentHDL ==> CVD,
Diabetes ==> AntiHypertensiveUse Creatinine,
Ethnicity ==> Nutrition Smoking, Gender ==> Nutrition Urate,
Gout ==> CVD, HbA1c ==> Diabetes,
MedicationPropensity ==> AntiHypertensiveUse StatinUse,
Nutrition ==> PreviousHDL Urate Obesity, Obesity ==> PreviousBP HbA1c,
PreviousBP ==> AntiHypertensiveUse, PreviousHDL ==> StatinUse,
Smoking ==> CVD, StatinUse ==> CurrentHDL,
Urate ==> PreviousBP Creatinine CVD Gout;
identify Urate ==> CVD;
unmeasured Nutrition Obesity PreviousBP MedicationPropensity PreviousHDL;
testid CurrentHDL Ethnicity Gender HbA1c Smoking;
run;
```

【SAS输出结果及解释】

检验调整集是否为有效的因果效应的输出结果,见表 1。

表 1 检验调整集是否为有效的因果效应的输出结果

Table 1 Output results for testing whether the adjustment set is a valid causal effect

模型名称	大小	有效	最小	C1	C2	C3	C4	C5
Thor12	5	否	否	*	*	*	*	*

注:C1~C5 分别代表协变量 CurrentHDL、Ethnicity、Gender、HbA1c、Smoking;*协变量对结果变量的影响具有统计学意义

【表 1 中有关内容的说明】第 2 列的“大小”指协变量的个数(本例有 5 个);第 3 列的“有效”指协变量的集合对研究变量 Urate 对 CVD 因果效应是否有效(本例经检验,其结果为“无效”,输出中用“否”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 5 个协变量,故它不是最小的调整集);表 1 中 5 个协变量对结果变量的影响具有统计学意义,即在研究变量 Urate 对 CVD 因果效应中是不可忽视的,但由第 3 列上的“否”可知,包含 5 个协变量的调整集不足以估计变量 Urate 对 CVD 的因果效应。

3.1.2.2 枚举路径和搜索有效调整集

要了解建议的调整集无效的原因,可以请求枚举将处理变量与模型中的结果变量联系起来的正

确路径，也可以使用该过程搜索有效的调整集。以下程序调用 CAUSALGRAPH 过程来执行这两项任务，SAS 程序与上文“3.1.2.1 节”的程序基本相同，仅以下两句稍有改变：

```
proc causalgraph compact list;
testid Gender HbA1c Ethnicity Smoking CurrentHDL / paths=(noncausal nonblocked);
```

【SAS 输出结果及解释】

输出了两条路径，第 1 条路径为：

Urate <== Nutrition ==> PreviousHDL ==> StatinUse <== MedicationPropensity ==> AntiHypertensiveUse ==> CurrentBP ==> CVD

第 2 条路径为：

Urate <== Nutrition ==> Obesity ==> PreviousBP ==> AntiHypertensiveUse ==> CurrentBP ==> CVD

注：没有满足“Thor12”的指定准则的调整集。

【关于上述两条路径的说明】第一条路径没有被阻塞，因为变量 StatinUse 是路径上的碰撞器，并且它的一个子变量 CurrentHDL 出现在调整集中。第二条路径不包含任何碰撞器，但不会被阻塞，因为它不包含建议调整集中的任何变量。这两条路径都是非因果关系的路径。如果使用建议调整集，变量 Urate 和 CVD 之间的某些关联可能归因于这两条非因果路径，并且因果效应无法正确估计。

3.1.2.3　减少不可测量变量

为了获得可识别性，研究者可以考虑收集附加数据。例如，如果研究者要收集变量 Obesity 和 PreviousHDL 的数据，以便这两个变量成为可测量的变量，这等价于在上面的输出结果中阻塞两条非因果路径。下面的程序可实现这一目标：

SAS 程序与前面“3.1.2.1 节”的程序基本相同，仅以下两句稍有改变：

```
unmeasured Nutrition PreviousBP Medication Propensity;
testid Gender HbA1c Ethnicity Smoking Current HDL PreviousHDL Obesity;
```

【SAS 输出结果及解释】

TESTID 语句中包含 7 个变量时的输出结果，见表 2。

表 2　TESTID 语句中包含 7 个变量时输出的结果

Table 2　Output results when the testid statement contains 7 variables

模型名称	大小	有效	最小	C1	C2	C3	C4	C5	C6	C7
Thor12	7	是	否	*	*	*	*	*	*	*

注：C1~C7 分别代表协变量 CurrentHDL、Ethnicity、Gender、HbA1c、Smoking、Obesity、PreviousHDL；*协变量对结果变量的影响具有统计学意义

由表 2 可知，建议的调整集（含 7 个协变量）在“有效”列中标记为“是”，说明该调整集足以估计变量 Urate 对 CVD 的因果效应。

3.2　寻找多个因果图模型共用的调整集

3.2.1　因果图模型确切结构未知时的解决方案

当研究者不确定因果图模型的确切结构或数据可能存在其他因果图模型时，如何使用因果图过程找到可用于估计因果效应的调整集？例如，研究者可能不确定是否要在因果图模型中包含哪些边，边具有什么方向，或者模型包含哪些协变量，则可以在过程中使用多个 MODEL 语句来指定每个合理的因果图模型，然后使用 PROC CAUSALGRAPH 语句中的公共选项来搜索对所有模型有效的调整集。

对于单个因果图模型，如果存在调整集，则可以使用该调整集来估计因果效应。如果多个因果图模型共享一个调整集，那么无论哪个因果图模型反映了真实的数据生成过程，都可以使用该调整集来估计因果效应。

3.2.2　实例及其背景信息

【例 2】沿用例 1 的实例和背景信息，在图 1 的基础上，添加一些新的信息。见图 2。

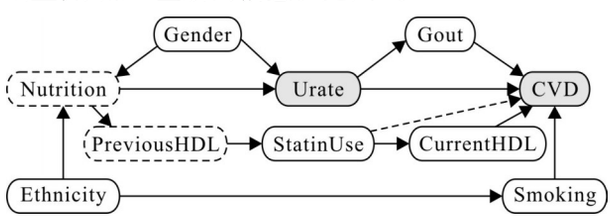

图 2　血清尿酸盐对心血管疾病风险影响的两种可能因果模型

Figure 2　Two possible causal models of the effect of serum urateon risk of cardiovascular disease

图 2 中的模型源自 Thornley 等[6]开发的更大模型。这些模型研究了受试者血清 Urate 与 CVD 风险之间的关系。有关模型中变量的含义，请参见本文例 1。图 2 描述了变量 StatinUse 对 CVD 效应的两个合理因果假设。一个模型假设变量 StatinUse 对 CVD 的所有因果效应都是由变量 CurrentHDL 介导的；另一个模型，除介导效应外，还包括变量 StatinUse 对 CVD 的直接影响。这两个模型的不同之处在于一条边，如图 2 中的虚线箭头所示。在本例的两个模型中，假设变量 Nutrition 和 PreviousHDL 是不可测量的，故由虚线轮廓的节点表示。

3.2.3　用 SAS 实现因果效应分析

调用 PROC CAUSALGRAPH 为图 2 中的两个模

型构造调整集。所需要的SAS程序与例1中的程序基本相同，此处从略。

【SAS输出结果及解释】

因篇幅所限，具体输出结果从略。现将输出结果分3部分概述如下。

第1部分输出结果：图2中定义的第1个因果图模型（Thor12SimpleHDL）的计算结果，由6个可观测的协变量（CurrentHDL、Ethnicity、Gender、Gout、Smoking及StatinUse）构造出18个调整集，其中，有4个调整集属于最小调整集（每个最小调整集仅包含2个协变量）。第2部分输出结果：图2中定义的第2个因果图模型（Thor12AltHDL）的计算结果，由6个可观测的协变量（CurrentHDL、Ethnicity、Gender、Gout、Smoking及StatinUse）构造出12个调整集，其中，有2个调整集属于最小调整集（每个最小调整集仅包含2个协变量）。第3部分输出结果：2种因果图模型共用的调整集。巧合的是，在两个模型中有效的调整集与在具有额外边的模型中有效的调整集相同。

由于至少存在一个共用调整集，该分析表明，通过“第3部分输出结果”中的任何一个调整集，无论变量StatinUse是否对CVD有直接效应，都可以估计变量Urate对CVD的因果效应。

4 讨论与小结

4.1 讨论

基于因果图模型所做的统计分析，其结论正确与否，主要取决于以下关键点：图中所设定的变量之间的关系（包括变量的先后顺序，发出箭头和接收箭头的变量，变量是否为观测变量，哪些变量是处理变量、协变量和结果变量）；是否有专业知识为依据；是否符合基本常识和常规逻辑。这几个方面在多重线性回归分析[7]、多重Logistic回归分析[8]、多重Cox模型回归分析[9]和结构方程模型分析中[10]，也都是极为重要的。

4.2 小结

本文介绍了因果图模型的基础知识和理论，涉及因果图的组成和术语。针对两个实例并借助SAS软件，实现了两项因果图分析任务，即检验调整集和寻找多个模型的共用调整集。

参考文献

[1] 方积乾，陆盈. 现代医学统计学[M]. 北京：人民卫生出版社，2002：512-534.
Fang J，Lu Y. Advanced medical statistics[M]. Beijing：People's Medical Publishing House，2002：512-534.

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

[3] Pearl J. Comment：graphical models，causality and intervention[J]. Stat Sci，1993，8(3)：266-269.

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

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

[6] Thornley S，Marshall RJ，Jackson R，et al. Is serum urate causally associated with incident cardiovascular disease?[J]. Rheumatology，2013，52(1)：135-142.

[7] 孙振晓，于相芬. 新冠肺炎疫情期间封闭管理精神科医护人员焦虑抑郁症状及相关因素调查[J]. 四川精神卫生，2020，33(2)：102-106.
Sun Z，Yu X. Anxiety and depression symptoms and related factors of medical staff in closed-door psychiatric ward during COVID-19outbreak[J]. Sichuan Mental Health，2020，33(2)：102-106.

[8] 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 strategies in untreated depressive disorder patients[J]. Sichuan Mental Health，2020，33(1)：44-48.

[9] 李欣洁，何红波，张杰. 精神障碍住院患者出院后1年内再住院的危险因素[J]. 四川精神卫生，2021，34(1)：69-74.
Li X，He H，Zhang J. Risk factors of rehospitalization in psychiatric inpatients within one year after discharge[J]. Sichuan Mental Health，2020，34(1)：69-74.

[10] 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.

（收稿日期：2022-07-10）
（本文编辑：戴浩然）