基于工具变量识别因果效应以及用数据区分不同模型

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工具变量识别因果效应、用数据区分不同模型以及使用SAS软件实现计算的方法。首先，介绍因果图理论的4个主要内容，包括关联的来源、因果图模型的统计性质、识别和调整以及工具变量；其次，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用工具变量识别因果效应；②用数据区分不同模型。

【关键词】 因果图模型；因果效应；关联和偏差；识别和调整；工具变量

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20220710004

Identifying causal effects based on instrumental variables and distinguishing different models with data

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 methods of identifying causal effects based on instrumental variables, distinguishing different models with data, and using SAS software to realize calculation. Firstly, the four main contents of causal graph theory were introduced, including sources of association, statistical properties of causal models, identification and adjustment, and instrumental variables. Secondly, for two examples and with the help of the CAUSALGRAPH procedure in SAS/STAT, the following two tasks were completed: the first task was to identify causal effects using instrumental variables; the second task was to use data to distinguish different models.

【**Keywords**】 Causal graph model; Causal effect; Association and bias; Identification and adjustment; Instrumental variables

因果图理论的4个主要内容包括关联的来源、因果图模型的统计性质、识别和调整以及工具变量，本文在介绍这些理论的基础上，针对两个实例，并借助SAS软件完成用工具变量识别因果效应以及用数据区分不同模型的任务。

1 因果图理论的4个主要内容

1.1 关联的来源

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）。由DAG表示的因果图模型对信息在底层数据生成过程中的流动方式具有明确的定义。该信息流由三个图形结构封装，可用于在DAG中组装每条路径[1]。这三个结构对应因果图模型中关联的三个基本来源（即因果关系、混淆和内生选择），这些结构如图1所示。

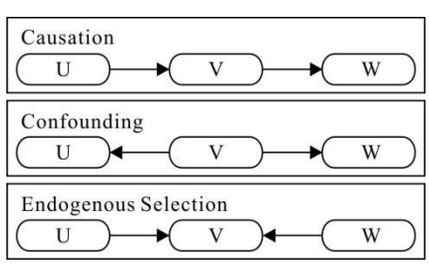

图1 关联的3个基本来源

Figure 1 Three fundamental sources of association

在因果关系U→V→W中，变量U和W是关联的，这种关联是因果链的结果。如果要对中介变量V进行调节，那么这将阻塞信息流，从而使变量U和W不再关联。

在混淆的关系U←V→W中，没有与变量U和W相关的因果路径，然而，U和W仍然是相关的，这种关联是由混杂变量V引起的，它是变量U和W的共同父项，如果研究者要以共同原因V为条件，那么这将阻塞信息流，从而使变量U和W不再关联。

在内生选择结构U→V←W中，变量U和W共同决定其共同子变量V的值，但变量U和W不相关，然而，如果研究者要以共同的结果变量V为条件，就可以创建一个信息流，变量U和W就会关联起来[2]。

粗略地说，如果在因果效应分析中有一个处理变量（如U）和一个结果变量（如W），那么目标是消除变量U和W之间的非因果关联，并保持因果关联不变。因此，这三个基本图形结构不仅对应关联的三个基本来源，也对应偏差的三个基本来源。一般来说，当研究者有一组处理变量和结果变量时，如果控制因果路径上的一个变量，就会阻塞流经该因果路径的信息流，这被称为过度控制偏差；同样，如果研究者无法控制一个令人困惑的共同原因，那么处理变量和结果变量之间的一些关联就是混淆的结果，被称为混淆偏差；而如果研究者控制变量的共同结果，就会在处理变量和结果变量之间建立非因果关系的关联，被称为内生选择偏差。

1.2 因果图模型的统计性质

1.2.1 局部马尔科夫性

有两种方法可以解释DAG中的假设：①DAG是“组织有关外部干预及其互动的一种结构”[3]；②DAG定义了一组变量之间的信息流。这两种解释在另外两种假设下是等效的[1]：①DAG中的变量满足局部马尔可夫性；②DAG满足弱忠实性属性。局部马尔可夫性质表明，DAG中的每个变量在统计上是独立的，取决于它的父变量和非退化变量集[4]。

1.2.2 d分离

如果以下任一条件成立，则DAG中的路径称为由一组变量Z形成的d分离：①路径包含链U→V→W或分叉U←V→W，使得V∈Z；②路径包含一个碰撞器U→V←W，使得V∉Z以及V的后代不在Z中。

1.2.3 阻塞/非阻塞

一条d分离的路径被称为阻塞。如果X中的一个节点和Y中的一个节点之间的每条路径都被阻塞，则变量集X通过变量集Z与变量集Y进行d分离。阻塞/非阻塞术语反映了因果图模型中的信息流，如果路径被阻塞，则信息不会流经该路径；如果路径未被阻塞，则信息可能会流经该路径。d分离和信息流之间的联系体现在弱忠实性假设中。弱信度表示，如果一个DAG中的两个变量X和Y不是d分离的，那么这两个变量至少依赖于一个在DAG上分解的分布[5-6]。

1.2.4 全局马尔科夫性

通过将因果图模型解释为代表变量之间关联流的DAG，可以将DAG背后的因果假设转化为条件独立性。具体来说，如果两个变量在DAG中由集合Z进行d分离，那么这两个变量必须在统计上独立于Z。换言之，d分离是一个全局马尔可夫性质。如果条件独立性只包含观察到的变量，则可以使用观察到的数据执行统计检验，以查看独立性是否成立。因此，d分离标准确定了因果图模型具有可观测和可检验的含义[1]。

事实上，DAG的全局马尔可夫性质和局部马尔可夫性质在逻辑上是等价的[4]。如果研究者有一个局部或全局马尔可夫性质的完整列表，就可以使用semigraphoid公理[7-8]推导出另一个列表。在CAUSALGRAPH过程中，可以使用PROC CAUSALGRAPH语句中的IMAP选项来请求这些属性的列表。

1.3 识别和调整

一对变量之间的统计关联可以分为两个部分：因果部分和非因果部分（虚假部分）。如果所有虚假关联都能被消除，那么因果关系就会被识别出来。因此，一种可能的识别方法是调整识别，这是回归和匹配因果效应识别的基础[2]。

当研究者使用调整识别时，会寻找一个调整集，当在分析中进行控制时，它会阻塞DAG中的所有非因果路径，而不会阻塞任何因果路径。路径的因果属性是从模型中边的方向继承的。也就是说，因果属性是因果图模型的属性，在分析过程中不会改变。然而，路径是否被阻塞不仅取决于代表因果图模型的DAG结构，还取决于调整集中包含的变量集。因此，研究者必须谨慎地选择一个调整集，以便在不引入任何过度控制或内生选择偏差的情况下消除所有混淆偏差。

1.4 工具变量

在实际的医学研究中，一般来说，变量可分为自变量、中介变量和因变量（结果变量）。然而，统计学上还提出了一种“工具变量”[9]，其定义如下：某个变量Z与模型中某个自变量X高度相关，但却不与随机误差项相关，那么就可以用变量Z与模型中相应回归系数得到一个一致估计量，在模型的参数估计过程中，变量Z被作为一个工具使用，故称为工具变量。为了加深对工具变量的理解，举例如下[10]：

研究者将抽烟的孕妇随机分成两组，试验组接受“减少或停止抽烟”的劝告或鼓励，而对照组未接受劝告或鼓励。研究者记录每位受试者两个结果变量的取值，即孕妇在8个月的孕期内每天抽烟支数（记为S）及其婴儿的出生体重（记为B）。研究者关心的是S对B的因果效应（虽然S与B之间的关联可能存在混淆，如对S可能存在测量误差），为了解决S对B的因果效应的估计问题，研究者利用随机化方法（R），假设R可能与S高度相关，但R仅通过对S的影响进而对B产生影响（即以S为条件，随机化对B没有影响）。在这种情况下，变量R被称为工具变量。

2 因果图过程的应用

2.1 使用工具变量识别因果效应

2.1.1 实例与背景信息

【例1】沿用文献[9]中的“Example 34.2”，此例的结论为：“不可能使用调整集来确定持久性全氟烷基物质（PFAS）对母乳喂养持续时间（Duration）的因果效应”。试问：原因是什么？如何解决所面临的问题？

【分析与解答】此例采用的是Timmermann等[11]的因果图模型，研究了法罗群岛居民中母亲接触PFAS与Duration之间的关系。该例表明，研究者无法构建调整集来估计处理变量PFAS对结果变量Duration的因果效应。这是因为处理变量和结果变量之间存在混淆偏差，这些混淆来自未观察到的变量“行为被认为是健康的程度”（HealthBehavior）和“先前是否有母乳喂养经验”（PrevBF）。

在许多存在未测量混淆的情况下，如果愿意假设因果图模型中的某些边具有特定的参数形式，通过使用工具变量[12-13]，仍然可以估计因果效应。

2.1.2 用SAS实现因果效应分析

2.1.2.1 初步分析

以下语句调用PROC CAUSALGRAPH过程列出可用于估计因果效应的工具变量。设所需要的SAS程序如下；

```
proc causalgraph method=iv; /*iv 为工具变量的缩写*/
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】

输出结果为两类变量：第一类为“工具变量”，即孩子出生时母亲的年龄（Age）；第二类为“条件变量”，包括饮酒（Alcohol）、母亲孕前体重指数（BMI）、母亲接受初等教育的时间（Education）、胎次（Parity）和吸烟（Smoking）。

输出结果表明：变量Age可用于确定处理变量PFAS对结果变量Duration的因果效应。

2.1.2.2 改变条件变量再分析

构造工具变量时产生的条件集可能不是最小的。例如，以下检验表明，如果仅调整Education和Parity两个变量，也可以使用变量Age作为工具变量。

所需要的SAS程序与“2.1.2.1节”的SAS程序相同，只需要在UNMEASURED语句之后增加以下语句：

```
testid "Minimal CIV" Age / conditional = (Education Parity);
```

【SAS程序说明】这里的新功能是使用TESTID语句。研究者想调查当条件变量为Education和Parity时，Age是否可以作为因果效应分析的工具变量。这组条件变量是之前分析中提出的条件变量的适当子集。

【SAS输出结果及解释】

当条件变量为Education和Parity时，以Age作为工具变量，研究处理变量PFAS对结果变量Duration的因果效应是有效的。

2.2 用数据区分模型

2.2.1 实例与背景信息

【例2】沿用文献[9]中的“Example 34.3”，研究者设定了两个模型，见图2、图3。

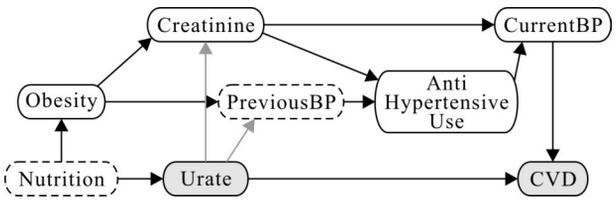

图2 血清尿酸盐对心血管疾病风险影响的第一个可能因果图模型

Figure 2 The first possible causal models of the effect of serum urate on risk of cardiovascular disease

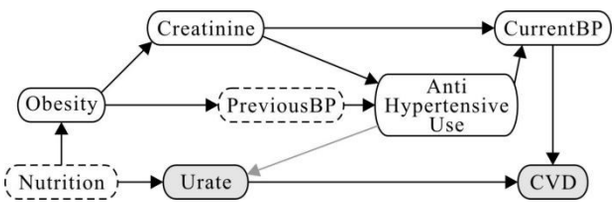

图3 血清尿酸盐对心血管疾病风险影响的第二个可能因果图模型

Figure 3 The second possible causal models of the effect of serum urate on risk of cardiovascular disease

在图2中，假设血压（PreviousBP）和使用抗高血压药物（AntiHypertensiveUse）介导了变量尿酸盐（Urate）对心血管疾病（CVD）的影响。在图3中，因果方向相反，并且假设使用AntiHypertensiveUse对Urate直接产生因果效应（对CVD而言，Urate就成了中介变量）。

两个模型之间的差异以灰色（特指变量AntiHypertensiveUse到Urate之间的灰色箭头）突出显示。两个模型中的处理变量Urate和结果变量CVD均位于阴影框内。注意，变量营养（Nutrition）对应于潜在结构（在图中以虚线框表示），故不进行测量或观察。还假设未测量变量先前血压（PreviousBP）。当研究者有多个可能的因果图模型时，若能找到一个对所有模型都有效的公共调整集，便可使用调整技术来估计数据的因果效应。

试基于图2和图3中设定的因果图模型，构建它们的公共调整集。

2.2.2 用SAS实现因果效应分析

以下语句调用PROC CAUSALGRAPH来构造公共调整集。设所需要的SAS程序如下：

```
proc causalgraph common(only);
model "Thor12SimpleBP" /*(与图2对应)*/
AntiHypertensiveUse ==> CurrentBP,
Creatinine ==> AntiHypertensiveUse CurrentBP,
Nutrition ==> Urate Obesity, Obesity ==> PreviousBP Creatinine,
CurrentBP ==> CVD, PreviousBP ==> AntiHypertensiveUse,
Urate ==> PreviousBP Creatinine CVD;
model "Thor12AltBP" /*(与图3对应)*/
AntiHypertensiveUse ==> CurrentBP Urate,
Creatinine ==> AntiHypertensiveUse CurrentBP,
Nutrition ==> Urate Obesity, Obesity ==> PreviousBP Creatinine,
CurrentBP ==> CVD, PreviousBP ==> AntiHypertensiveUse,
Urate ==> CVD;
identify Urate ==> CVD;
unmeasured Nutrition PreviousBP;
run;
```

【SAS输出结果及解释】

NOTE：没有满足指定准则的适用于所有模型的调整设置。

输出结果中的注释表明，在本例中，未能找到共用调整集。因此，研究者必须找到单独的调整集，然后使用每个模型分别估计因果效应，或者必须确定最能代表数据生成过程的模型。PROC CAUSALGRAPH语句中的IMAP（条件独立性假设）选项可对模型属性进行分析。

2.3 枚举和检验隐含的统计特性

2.3.1 隐含的统计特性

研究者可以根据可用数据检验因果图模型中隐含的统计特性。如果隐含的统计特性在数据中不存在，则应考虑修改或放弃模型。如果有多个模型，则可以比较这些模型的统计含义，以找到一个模型中具有的隐含统计特性，而不是其他模型中隐含的统计特性。然后可以在数据中检验此特性，并使用相应的检验结果来确定哪个模型最能代表真实的数据生成过程。

下面的语句调用PROC CAUSALGRAPH过程来枚举本例中两个模型具有的统计特性。对于每个模型，过程生成的条件独立属性表由ODS OUTPUT语句保存到一个数据集。与“用数据区分模型”的SAS程序基本相同，下面仅列出不同之处：

```
proc causalgraph imap=global;
ods output imap=SimpleBPIndep;
proc causalgraph imap=global;
ods output imap=AltBPIndep;
```

【SAS程序说明】PROC CAUSALGRAPH过程中的IMAP=GLOBAL选项都会为使用MODEL语句指

定的每个模型生成一个全局马尔可夫属性表。每个全局马尔可夫属性由两个变量组成,这两个变量在统计上独立于另一个变量集(可能为空,在这种情况下,独立性是无条件的)。如果观察到马尔可夫属性中的每个变量,则可以使用数据执行统计检验(例如可以检验零偏相关),以查看该属性是否可以被修改。而涉及一个或多个未测量变量的独立性属性无法检验。

【SAS输出结果及解释】

与图2对应的输出结果显示,可以找到仅包含变量肥胖(Obesity)的调整集,而且,它是最小的调整集。基于变量 Obesity,可有效地进行 Urate 对 CVD 的因果效应分析。

与图3对应的输出结果显示,可以找到11个调整集,因篇幅所限,具体输出结果从略。

2.3.2 输出数据集中的观测

以下程序输出每个模型的前10个观察到的条件独立性属性。为了简洁起见,本例重点介绍前10个结果。设所需要的SAS程序如下:

```
proc print data=SimpleBPIndep(obs=10);
var Set1 Set2 CondSet; where Observable = 1;
run;
proc print data=AltBPIndep(obs=10);
var Set1 Set2 CondSet; where Observable = 1;
run;
```

【SAS输出结果及解释】

因篇幅所限,与模型1(见图2)和模型2(见图3)对应的"条件独立性属性"的输出结果从略,现概要解释如下。

对这两个模型进行比较,两个模型有4个条件集,使得变量 AntiHypertensiveUse 的使用在条件上独立于变量 CVD,并且这4个条件集对于两个模型都是相同的。因此,研究者无法通过检验变量 AntiHypertensiveUse 和 CVD 之间的条件独立属性来区分这两个模型。接下来,比较变量肌酐(Creatinine)和 CVD 的条件独立性属性。Thor12SimpleBP 模型(模型1)有4个这样的属性,但 Thor12AltBP 模型(模型2)有5个这样的属性。Thor12AltBP 模型给出了 Creatinine 和 CVD 在集合上独立的统计含义(AntiHypertensiveUse、CurrentBP 及 Obesity),但 Thor12SimpleBP 模型中未给出此含义。

如果研究者要找到变量 Creatinine 和 CVD 之间的非零偏相关(在排除 AntiHypertensiveUse、CurrentBP 及 Obesity 后),研究者将有证据拒绝 Thor12AltBP 模型。研究者可以继续对两个模型中唯一的独立属性进行类似分析。最好的模型是其条件独立性属性与可用数据中的零偏相关最为匹配的模型。

3 讨论与小结

3.1 讨论

工具变量在因果效应分析中起着重要作用,如果在某个实际问题中确实存在一个或多个工具变量,它们必将对其他变量起混杂效应。因此,找出全部工具变量,并在其后的统计分析中充分发挥其作用(例如工具变量回归分析[9-10],它有别于通常的回归分析[14-15]),将有助于获得正确的统计结果和结论。

3.2 小结

本文介绍了因果图理论的4个主要内容,包括关联的来源、因果图模型的统计性质、识别和调整以及工具变量;针对两个实例并基于SAS软件[9],完成了使用工具变量识别因果效应以及用数据区分不同模型的因果效应分析任务。

参考文献

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

[2] Elwert F, Winship C. Endogenous selection bias: the problem of conditioning on a collider variable[J]. Annu Rev Sociol, 2014, 40: 31-53.

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

[4] Koller D, Friedman N. Probabilistic graphical models: principles and techniques[M]. Cambridge, MA: MIT Press, 2009: 24-79.

[5] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search [M]. 2nd edition. Cambridge, MA: MIT Press, 2001: 103-168.

[6] Pearl J. Causality: models, reasoning, and inference [M]. 2nd edition. Cambridge: Cambridge University Press, 2009: 52-136 .

[7] Pearl J, Verma T. The Logic of representing dependencies by directed graphs[A]. Proceedings of the sixth national conference on artificial intelligence [C]. Menlo Park, CA: AAAI Press, 1987: 374-379.

[8] Geiger D, Pearl J. On the Logic of causal models[A]. Proceedings of the fourth annual conference on uncertainty in artificial intelligence[C]. Amsterdam: North-Holland, 1988: 136-147.

[9] SAS Institute Inc. SAS/STAT®15. 1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2243-2300.

[10] Armitage P, Colton T. Encyclopedia of biostatistics [M].

2nd edition. Hoboken, NJ: John Wiley & Sons, 2005: 2788-2792.
[11] 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.
[12] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables[J]. J Am Stat Assoc, 1996, 91(434): 444-455.
[13] Imbens GW. Instrumental variables: an econometrician's perspective[J]. Stat Sci, 2014, 29(3): 323-358.
[14] 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.
[15] 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 yearafter discharge[J]. Sichuan Mental Health, 2020, 34(1): 69-74.

（收稿日期:2022-07-10）
（本文编辑:戴浩然）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955年8月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事、中国生物医学统计学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等10余种杂志编委；现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和3种医学杂志编委；主编统计学专著48部、参编统计学专著10部；发表第一作者和通信作者学术论文300余篇、发表合作论文130余篇；获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的30年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告100余场，举办数十期全国统计学培训班，培养20多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加100多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和SAS与R软件实现、各种层次的统计学教学培训和咨询工作。