Applied Statistical
Causal Inference

应用统计因果推论

胡安宁 ◎编著

復旦大學出版社

应用统计因果推论

胡安宁 ◎编著

复旦大學出版社

图书在版编目(CIP)数据

应用统计因果推论/胡安宁编著. —上海：复旦大学出版社，2020.5
ISBN 978-7-309-14821-3

Ⅰ.①应… Ⅱ.①胡… Ⅲ.①统计分析-研究 Ⅳ.①C813

中国版本图书馆 CIP 数据核字(2020)第 020235 号

应用统计因果推论
胡安宁 编著
责任编辑/谢同君

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编：200433
网址：fupnet@fudanpress.com http://www.fudanpress.com
门市零售：86-21-65102580 团体订购：86-21-65104505
外埠邮购：86-21-65642846 出版部电话：86-21-65642845
上海丽佳制版印刷有限公司

开本 787×1092 1/16 印张 19 字数 282 千
2020 年 5 月第 1 版第 1 次印刷

ISBN 978-7-309-14821-3/C·389
定价：48.00 元

谨以此书献给我的爷爷　胡继海

幼时受教

受用终生

前　言

在过去几十年中，统计因果推论技术在社会科学不同领域中得到了长足发展。以传统分析相关关系为主的社会科学量化研究逐渐转向关注如何通过严格的研究设计或者统计操作来展示变量之间的因果联系。在此背景下，我们需要一本专门的书籍，来系统介绍社会科学统计因果推断的不同技术及其在现实环境中的实现过程。

与一般意义上的统计学教科书相比，本书的一个不同之处在于，使用了相当大的篇幅来讨论因果推论的基本原理。这样设计的一个考虑是，因果推断的统计实践过程，就操作上来讲并不复杂，在 STATA 或者 R 软件里面可能只需要一些简单的命令，分析结果就可以出来了。但对于社会科学的研究者而言，更为重要的问题或许是如何去理解、阐释分析的结果，让统计分析结果具有社会科学的实质意义。毋庸置疑，这个过程需要研究者准确把握相关分析背后的原理和机制，了解分析过程背后假设是什么以及得出这个结论的前提基础是什么。在一本介绍统计因果方法的教材中，笔者认为，这些非常重要的原理性问题需要特别的篇幅去澄清。基于这个考虑，本书大约三分之一的内容在介绍因果推断技术背后的基本统计原理。

当然，这些原理性的知识和后面具体的因果推断方法之间并不是彼此割裂的。当研究者掌握这些原理以后，对于后面的一些具体方法，不管是工具变量也好，还是回归断点设计也好，甚至是一些历时性数据分析也好，都会有一个更加深入的了解。因此，把原理性的东西向读者介绍清楚，读者在面对具体问题，使用相关方法的时候就会有的放矢，对研究工作的展开起到事半功倍的效果。此外，基于这些原理，我们甚至可以自己去开发新的方法。打个比方，这就好像是做菜，高级的食材往往需要用心去烹饪。社会科学研究亦如是。越来越多的研究主题需要研究者根据

“经验材料”开发独特的分析方法。此时，熟悉原理性的知识就变得非常有必要。当然，我们不是说每个研究者为了自己研究的问题都要亲自开发一套方法，但是真的碰到了这种情况，我们能够依靠的便是原理性的知识，而非别人封装(canned)好了的分析软件。

具体来讲，本书的内容分为三部分。第一部分是基础理论，主要是谈反事实的因果分析框架、随机实验的原理、观测性研究概念等主题。第二部分是基本的因果推断方法。所谓基本的方法，并不是说它们简单、低级，而是说这些方法已经比较成熟，成为学者研究过程中的“常规武器”。从一个使用者的角度来说，这些方法究竟怎么用，怎么判断用得好不好，已经基本上有定论。研究者通常按部就班使用这些方法就可以了，别人不会质疑具体的分析过程，毕竟程序上已经很常规化，不再是每一步的分析都需要争论。具体而言，基本方法部分涉及五个方法：匹配方法、倾向值方法、工具变量方法、回归断点设计方法和追踪数据的因果推断方法。第三部分是高级方法部分。这里所谓的高级方法，也不是说它们要比基本方法更加高深，而是说这些方法很多尚在开发过程中。因此，应该怎么使用，背后需要哪些条件，尚存在一定的争议。从使用者的角度来讲，高级方法大家还是要慎用。因为在分析具体问题的时候，别人有可能不会针对研究结论提出实质性的质疑，而是直接挑战方法论。如果研究者没有很好的应对方式，那么方法上的缺陷无疑是釜底抽薪，直接将经验研究根基动摇了。在高级方法部分，主要介绍广义倾向值方法、因果中介分析、敏感性检验和随时间变化的处理效应方法。在本书的最后，笔者附上两篇学术论文，分别处理的是统计模型的不确定问题以及主观变量解释主观变量的问题。这两个问题并非完全原理性的思辨，而是从实践的角度进行的讨论，这里也一并展示出来，供读者参考。

本书读者群的定位是希望了解统计因果推断的社会科学各学科的师生。为了很好地使用本书，读者需要一些基本的预备知识。例如，理解统计推断的基本逻辑，明白何为零假设、替代假设、p 值、置信区间等概念，知道 OLS 回归和 logistic 回归等。除此之外，本书的使用对于软件没有直接要求。考虑到 STATA 或者 R 软件在社会科学领域内的广泛应用，本书大部分的统计方法都附上两个软件的代码与分析结果，读者可以根

据自己的偏好，选择使用。

本书得以付梓，我要特别感谢复旦大学出版社谢同君老师对文稿细致的排版与编辑。此外，本书有幸由圣路易斯华盛顿大学郭申阳教授、香港科技大学吴晓刚教授以及南京大学陈云松教授撰写推荐语。三位皆为享誉海内外的研究方法专家，同时是社会科学研究的前辈先进。对于他们对于本书的肯定与支持，我由衷的表示感谢！

目录

第一部分 理论基础

第二部分 基本方法

第三部分　高 级 方 法

第一部分 理论基础

本部分有三章，第一章介绍了统计因果推断的基本原理、前提假设和一些研究设计上需要注意的理论点。第二章介绍了随机实验方法，除了实施过程之外，尤其着重展示了如何通过统计手段来分析随机实验收集的数据。第三章则从随机实验过渡到观测性研究，看在有限个潜在混淆变量的基础上，如何采用一些简单的分析手段达到对因果关系的估计。总体而言，本部分的三章可以看作后续章节的基础。虽然在这部分中，我们不会介绍太多具体的实务性分析技术，但是了解本部分的内容是所有因果关系的基础和前提，相关的内容在后面两个部分的多个章节中也会基于具体的研究情境反复出现。

第一章

反事实因果分析框架

本章的核心内容是反事实的因果推论框架。可以说，目前哲学也好，心理学也好，统计学也好，基本上绝大多数的经验导向的学科在做因果推断的时候，大家普遍采取的一个分析框架就是反事实的分析框架。因此，本章的内容可以说是非常基础性的理论。

在这一部分，主要讲四个问题。其一，什么是因果关系。这是一个基本的定义问题。做因果推断，我们首先就要确定什么样的关系可以称得上因果关系。其二，在定义好因果关系之后，下一个问题是如果要确定因果关系的话，需要满足什么条件。这里需要强调的是，任何统计模型都是在特定的条件下才成立的。比如，大家都知道最小二乘回归模型，需要满足线性关系假设。如果这个假设不满足，那么线性回归模型就不是那么可靠。比如，自变量和响应变量之间的关系有可能不是线性的，而是曲线关系，或者是波浪形关系，此时，我们的模型成立的假设前提就不满足了。后续即使做了工作，都将是无用功。同理，为了确定因果关系，我们需要满足一些基本假设。如果这些假设不成立，那么具体的分析技术，例如倾向值、回归断点设计等都会受到质疑。所以这一部分的内容是非常重要的。其三，我们简要讨论什么变量可以作为处理变量，或者英文表述为treatment variable。我们可以把处理变量理解成自变量，以此和响应变量进行区分。只是，我们通常所谈的自变量和响应变量更多强调的是相关关系情况下的变量联系，或者在控制了一些控制变量后的偏相关关系。这里将自变量命名为处理变量，更多的是和公共卫生领域的研究相呼应，

以求和响应变量(response variable)进行对应,强调的是因果关系。从定义上看,处理变量就是给个体一个干预,看结果是什么。在这部分的讨论中,我们所要回答的问题是,什么样的变量可以作为处理变量。其四,我们将讨论"结果的原因"和"原因的结果"两种研究范式的区分。这两个短语听起来很拗口,但彼此之间的差异很大。这里提到两者的区分,目的是希望推动读者去进一步的思考,自己在做的研究属于哪种性质的研究,多大程度是在科学性上讲是可靠的,多大程度上具有一般意义的科学研究价值。

因果关系的定义

我们进入第一个问题,什么是因果关系。这里举个例子,大家都有感冒的经历。感冒了以后需要吃药,吃药以后感冒好了,这个时候我们就会说,多亏吃了药,不然的话,感冒还不知道什么时候能好。在这个日常生活的例子中,实际上一个人在吃药和感冒症状变化之间确定了一种因果联系。但这种生活化的语言并不严谨。例如,我们怎么就知道感冒好了就是因为吃药了呢?比如说小朋友们感冒,只要不是病毒性的,让他坚持一星期,不吃药感冒也有可能好了。这时候,如果当时给他吃药,他感冒自然会好,但是即使不吃药,感冒也会好。这时,我们如何判断吃药对减轻小朋友的感冒症状的因果性联系呢?如果不吃药感冒也会好,我们会认为,吃药和感冒痊愈之间没有什么因果关系。从这个角度看,当我们说吃药和症状痊愈两者之间有因果关系的时候,我们实际上有一个基本预设,即**"如果当时不吃药的话,感冒不可能好起来"**。只有吃药以后,症状才得到缓解。从这个例子可以发现,对于一个个体而言,有两个状态,这两个状态都是一开始吃药之前不知道的,一个状态是吃了药以后的状态,另一个状态是没有吃药的状态。只有这两个状态有所不同,我们才能说是因为吃了药,感冒才好的。

这里我们可以再举一个社会科学经常分析的例子。我们之所以上大学,一个很大的动力是,毕业以后收入会不错。也就是说,我们都觉得上了大学以后的社会经济地位要比不上大学的情况下更高。换句话说,我

们来上大学，一般会抱有一个预设，认为“上了大学拿到文凭后，到劳动力市场上去找工作，收入水平肯定要比那些不上大学、高中毕业后就直接进入社会的要高”。但是我们如何能够确证这一点呢？为什么上大学就一定让我们的收入提高呢？会不会出现相反的情况，上了大学反而比不上大学收入更低呢？为了回答这些问题，我们通常的做法是，通过列举一些身边人的例子来说明上大学的好处。比如，一个人可能会说，某某同学，他高中毕业进入社会，到现在还没有稳定的工作、收入微薄。而另一个某某同学大学毕业了以后工作稳定，收入可观。但问题在于，这些某某同学是“你”吗？当然不是。那么下一个问题是，如果“你”当年不上大学，会是什么样的境遇呢？一定就和这个无稳定工作的某某同学一模一样吗？这可能就要画一个问号了。很多时候你会说，我也不知道啊，因为我实际上已经上大学了，没上大学的情况谁知道呢。此时，你虽然可以想象，但是也确实无法确证，如果“你”当时没上大学，而是直接进入社会，有没有可能比如今上大学的“你”的境遇更好。也有可能，“你”大学毕业以后去公司上班，但如果当年没上大学的话，反而自己开公司，成为雇佣大学生的人了。如果是这样的话，你的结论就会有 180 度转弯，即上大学没有那么有价值，甚至“有损”个人利益。大学回报问题是很经典的因果推断问题，这在讲到倾向值和匹配方法的时候，会再拿出来讨论。这里提到这个例子，是希望说明，和感冒吃药的例子类似，我们在思考因果问题的时候，需要思考个体的多种状态，唯有如此，我们才能够了解某种处理变量（这里的是否上大学）对于某个响应变量（这里的社会经济地位）的影响。说到这里，读者都能理解，毕竟是大家都耳熟能详的事情，有什么稀奇的呢？为了回答这个问题，我们需要进行更加规范、严格的讨论，以此展现上述的思路的价值，这就涉及反事实（counterfactual）问题了。

首先来定义什么是反事实。从本质上讲，事实状态，就是我们自己可以看到、经历了的状态。在感冒吃药的例子中，事实状态就是你真的吃药以后的症状。在教育回报例子中，大学毕业后的收入情况就是一个事实状态。而反事实状态则指的是，处理变量取值为“现实生活中没有发生”的状态下，响应变量的取值。也就是，如果一个人没有吃药的情

况下会有什么症状，或者如果一个大学生当年如果没上大学的话，他或者她的收入情况。基于这些定义，我们很容易理解，事实状态和反事实状态的差异，即代表了某种因果关系，即某一个处理变量对个体的因果效果。

为了便利后续的讨论，这里有必要引入一些符号。假如说我们关心的响应变量标注为 Y，处理变量为 D。那么，Y 就是 D 的一个函数。如果 D 有两个状态，即上不上大学（$D=1$ 表示上大学；$D=0$ 表示没有上大学），Y 是收入，那么每个个体有两种状态，一种状态是他上大学以后的收入 $Y(1)$，一种状态是他不上大学的收入 $Y(0)$。针对个体 i，我们可以分别表示为 $Y_i(1)$ 和 $Y_i(0)$。需要提醒大家的是，为了更好地理解 $Y_i(1)$ 和 $Y_i(0)$ 的含义，这里要把我们手里的调查资料或者数据先放在一边，考虑的不是经验事实如何，而只是在理念上进行想象。这里之所以在 Y 后面采用括号，表示的是 Y“不是”一个真实的事实状态，而是我们假想的状态：如果 D 取值为某个值的话，Y 的取值如何。与 $Y_i(1)$ 和 $Y_i(0)$ 相比，Y_i 没有括号，因此这个符号表示我们真实观测到的个体 i 的 Y 的取值。比如，大家从某个调查数据中找到一个人，看到他的收入情况，就可以用 Y_i 来表示。对于个体 i 而言，因果效果 τ_i 可以表示为

$$\tau_i = Y_i(1) - Y_i(0)$$

那么 Y_i，$Y_i(1)$ 和 $Y_i(0)$ 三者之间是什么关系呢？这里，我们可以将 Y_i 写成 $Y_i(1)$ 和 $Y_i(0)$ 的函数：

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

如上文所述，Y_i 表示实际上的观测值，这个 D 是 0 或 1，1 的话是上大学，0 的话表示不上大学。基于上面的公式，如果一个人上大学了，则 $D=1$，后面不上大学的情况就不存在了（$1-1=0$），Y_i 就等于 $Y_i(1)$。同理，那些没有上大学的人，其观测值就是 $Y_i(0)$。通过这个函数关系，我们可以认为，当处理变量 D 为二分变量的时候，每个被研究对象都有两个状态，如果上大学其收入怎么样，如果不上大学其收入怎么样。基于这个基本的设定，研究者收集数据，找了一些上大学的人，一个人的收入情况 Y_i 就应该等于他或者她“如果”上学的话的 $Y_i(1)$ 收入情况。对于

那些实际上没有上大学的人，Y_i 的观测值就是如果该个体没上大学时候的 $Y_i(0)$ 收入情况。对于上大学的人而言，$Y_i(0)$ 即为其反事实状态，因为这一状态不可观测。同理，对于那些没上大学的人而言，$Y_i(1)$ 即为其反事实状态。

这里不难发现，$Y_i(1)$ 和 $Y_i(0)$ 是不可能同时观测到的，即反事实的两种状态并不能直接观察到。这被称为“因果推断的根本性问题”(Holland，1986)。这个根本性问题带来的最大困境，在于我们无法直接计算个体性的因果关系，即针对个体 i 而言，上不上大学对于他或者她的收入的影响是无法直接计算的。这里先简单提一下，面对这个困境，一个替代性的方案是，找一些和被研究对象“特别像”的人，他们不知道什么原因，比如说高考的时候因为突然发烧或者其他一些意外情况，没能够上大学，但他高考之前和被研究对象很像，例如成绩接近，也是班级前几名的，性别都是男(女)的，都是城市户籍，等等。但实际情况是，被研究对象上了大学，而与之相似的个体没能上大学。经过四年，我们把他们进行对比，如果发现收入差不多，此时我们会说，好像上大学没什么用。如果上大学的个体收入更高，那么我们会得出结论，上大学的确能够提升个体的社会经济地位。但无论结论是什么，之所以能够得出一个因果性的结论，是因为那个和被研究对象相似的高中同学可以近似地代表那个如果当年没有上大学的被研究对象的情况。在本书后面章节介绍的很多具体的因果推论方法都是遵循了这一思路，即找到和被研究对象近似，但是处理变量 D 取值又不同的人来作为探究反事实状态的凭借。

在确定了因果关系的基本定义之后，下面就要对因果关系进行进一步的分类。具体而言，有三种估计量。一个估计量叫做 ATT，一个叫做 ATU，一个叫做 ATE。ATT 全称是实验组平均因果处理效应(average treatment effect for the treated)，是针对那些接受处理变量影响的个体的因果效果。举个例子，研究者手里有一些数据信息，包括一些已经上大学的人，同时研究者知道他们的实际收入情况。此时，研究者想知道“这些大学毕业生”当年如果没上大学的话他们收入情况是怎样的。用他们的实际收入减去如果没上大学的情况下的收入情况，此时就能够知道上大学对于“大学生们”的影响，这就是所谓的 ATT。用公式表示