对纵贯数据统计分析的认识*

任 强　　谢 宇

【内容摘要】在介绍了纵贯数据的设计思想及优缺点基础上，从统计方法的角度讨论了纵贯数据在社会科学中所能发挥的作用。纵贯数据的优点在于其可以帮助我们进行对总体异质性的识别、对因果机制干预的研究、对因果效应的研究和对“状态”变换的研究。以一些基于纵贯数据的研究设计为实例，文章阐述了在研究中假设与数据紧密衔接的重要性，以及利用统计方法分析纵贯数据时需要考虑的要点。但由于存在着由人类和人类行为内在变异性导致的这一无法避免的根本性缺陷，纵贯数据并不能解决所有问题。因而在纵贯数据的辅助下，研究者需要对社会现象有更深入的理解，将其进行更合理的概念化，并加以更精准的数据分析。

【关键词】纵贯数据；因果效应；变异性；异质性

【作者简介】任强，北京大学人口研究所副教授；谢宇，密歇根大学社会学系、北京大学中国社会科学调查中心教授。北京：100871

Statistical Analysis of Longitudinal Data

Ren Qiang　　Xie Yu

Abstract: The paper introduces the basic ideas of design for longitudinal survey data and its advantages and shortcomings, and discusses the rationales for collecting longitudinal data from the statistical perspectives. Longitudinal data are informative because they enable identification of population heterogeneity, study of intervening causal mechanisms, study of causal effects, and study of state transitions. Special considerations in longitudinal settings are addressed, as well as the importance of hypotheses, illustrated with examples of study designs using longitudinal data. Longitudinal data are not perfect, because the most serious shortcomings come from the intrinsic variability of humans and human behaviors. Given such severe limitations, what researchers of social phenomena can do is to develop better understanding, better conceptualization, and better data analysis, aided by longitudinal data.

Keywords: Longitudinal Data, Causal Effects, Variability, Heterogeneity

Authors: Ren Qiang is Associate Professor, Institute of Population Research, Peking University; Xie Yu is Professor, Department of Sociology at the University of Michigan and the Institute of Social Science Survey at Peking University. Beijing 100871. Email: renqiang@pku.edu.cn

---

* 此文曾在北京 2011 年 2 月 28 日～3 月 5 日举行的 1st International Conference on Challenges and Innovations in Longitudinal Surveys 会议上介绍，我们感谢参加会议的学者和於嘉、高丹雪所提供的建议。

1 使用纵贯数据的原因

在当前社会学、经济学、人口学领域，纵贯数据的使用已经成为主流，因为只有通过纵贯数据才有可能知道社会现象和个人行为的动态变化。目前国际上使用较多的纵贯数据有美国威斯康辛追踪调查（Wisconsin Longitudinal Survey，WLS）、美国收入动态追踪调查（Panel Study of Income Dynamics，PSID）、美国健康与养老研究（Health and Retirement Survey，HRS）等。使用较多的中国数据有中国健康与营养调查（China Health and Nutrition Survey，CHNS）和中国老人健康长寿影响因素研究（Chinese Longitudinal Healthy Longevity Survey，CLHLS）。目前，北京大学中国社会科学调查中心正在执行的中国家庭动态跟踪调查（China Family Panel Studies，CFPS）和中国健康与养老追踪调查（Chinese Health and Retirement Longitudinal Survey，CHARLS）也受到社会科学各界学者的欢迎和重视。

社会是非常复杂的系统，由于个体异质性（individual heterogeneity）、选择性偏误（selective bias）和忽略变量偏误（omitted - variable bias）的存在，完美无缺的数据是不存在的（谢宇，2006）。社会科学科研经费有限且较难申请，为什么要利用有限的经费进行如此昂贵的追踪调查？是否值得花费这么多钱以及我们能从花费巨大的纵贯数据中真正获得什么，是一个非常严肃的学术问题。面对如此复杂的社会，各类调查数据都往往存在某些缺陷，但我们在一定程度上可以通过改善研究设计、使用合适的统计方法来弥补数据的不足。而本文正是从统计方法角度阐述纵贯数据的重要性。

纵贯数据之所以在社会科学中如此重要，其核心原因有两点：（1）与截面数据相比较，纵贯数据在数据结构和提供的信息方面都更加丰富；（2）能够满足因果推论的需要。根据纵贯数据的属性，可以将其分为趋势数据（trend data）和追踪（面板）数据（panel data）两种类型。追踪数据是针对同一样本重复观测，如威斯康辛追踪调查、中国健康与营养调查、中国家庭动态跟踪调查和中国健康与养老追踪调查等等。趋势调查是针对同一总体在不同时期分别抽取不同样本进行重复观测，也被称为汇合的截面数据，如美国的综合社会调查（General Social Survey，GSS）和中国综合社会调查（China General Social Survey，CGSS），历次的人口普查和全国1%人口抽样调查等。

我们一般所讲的纵贯数据是指追踪（面板）数据。趋势数据实际上不是真正的纵贯数据，之所以将它与追踪数据共同提出，其目的在于强调二者的区别。追踪数据在结构上的特点为：（1）至少包含两个维度的信息——时间维度 t 和案例维度 i；（2）基本变量包含两类——时间独立或时间恒定变量（time - independent variable/time - invariant variable）与时间依赖或时变变量（time - dependent variable/time - varying variable）。一般来说，使用追踪数据进行研究的目的是控制未被观测到的异质性（unobserved heterogeneity）和对变化的趋势或过程进行描述和分析。Baltagi（2002）和 Hsiao（2003）认为，纵贯数据的优势在于：（1）控制个体异质性；（2）提供更加丰富的变异性信息，减少变量之间发生共线性的可能，增加自由度和提高估计的效率；（3）更好地对动态变化进行分析；（4）更好地识别和测量纯粹截面数据和时间序列数据中难以识别的效应；（5）建构和检验更加复杂、基于纯粹截面数据和时间序列数据无法实现的模型。当然，追踪数据具有一定的局限性，包括调查设计相对复杂，调查费用很高，以及由于很难长期追踪受访者，导致因无应答和样本规模的选择性缩减等问题而产生的偏差。

2 纵贯数据能做什么？

从社会科学研究的角度出发，研究者要清楚纵贯数据能满足或有助于我们回答什么样的问题。首先来回顾一下我们所主张的社会科学第一原理——变异性（variability）（谢宇，2006）。在社会科学研究中，所有的分析单元都是不一样的，而关于它们彼此之间是如何不一样的，则往往体现于数

据分析过程中所做的假设。我们之所以对追踪数据感兴趣，不仅在于关注总体变异（population variability）——这是要进行随机抽样的原因所在；更在于关注另一个维度的变异——时间维度的变异（temporal variability）。因此，在追踪数据中，我们将会面对更多的变异。这些变异不仅存在于分析单元层次上，而且存在于时间层次上，有时也存在于情景环境层次上。

关于纵贯数据在社会科学中的作用，我们认为主要体现在以下四个方面：（1）通过提供丰富的信息而有助于描述（describe）总体异质性（population heterogeneity）；（2）有助于揭示（reveal）干预的因果机制；①（3）有助于识别（identify）因外生性原因导致（exogenously imposed）的因果效应（这一点需要假设，我们后面将会着重讨论）；（4）它有助于描述/揭示/识别状态变换（state transition）。

2.1　对总体异质性的识别

正如社会科学研究强调的那样，由于所有的个体是不一样的，因而在总体水平存在大量的变异。当开展对总体异质性识别（identification of population heterogeneity）的研究时，我们经常使用固定效应模型（fixed – effects model）。有了追踪数据，就可能使用固定效应模型；如果没有追踪数据，则将没有足够的信息使用个体层次上的固定效应模型。此外，我们也使用增长曲线模型（growth – curve model）来描述总体异质性。在此我们将概述两种模型的主要差异以及各自包含的技术细节。

固定效应模型首先假设个体间是完全不一样的，有些特征是固有的、天生的，如智商（IQ）。但是，假设一个人的智商、能力或者个性是固定不变的，这实际上是不完全正确的。但出于需要，我们在数据分析时经常做这种假设。然而，由于所有个体差异在任何时候都不能被观测到，因此这种差异实际上会使个体随时间变化而表现出他们自己的特性。例如，有些人开始做事较晚，但努力赶超；有些人总是很早着手做事，而且很快完成。这样一来，个体之间因性格等未观察到的特征差异便导致他们体现出各自的工作风格。因此，当一些个体差异随时间变化而表现出来的时候，我们就可以用增长曲线模型来描述这些差异。

基本模型

假设我们有一个基本模型（basic model）

$$y_{it} = \alpha_{it} + \beta_{it}' x_{it} + \varepsilon_{it}$$

这里，$i = 1,2,\cdots,N$，$t = 1,2,\cdots,T$。α_{it}和β_{it}'都随个体 i 和时间 t 变化。由于一共有 K 个自变量，因此在此基本模型中，观测数量是 NT，常量参数数量是 NT，斜率参数数量是 NKT。在任一给定时间，我们在个体水平 i 和时间水平 t 观测到各类变量，在这个时间点每个个体有一个截距项、一组协变量系数和一个残差。因为自由度少于待估参数，如果没有对参数的约束，模型是不能够被识别的。但假如我们可以对参数进行适当的约束，就能够从总体上识别截距项的异质性、斜率的异质性和残差的异质性。此模型被称为随机效应模型（random effects model），它可以用来分析组间差异和组内差异，且此模型假定组间的差异是随机的。但随机效应模型无法完全解决忽略变量偏误或者生态学谬误的问题。

固定效应模型

当面对基于上述基本模型解决不了的问题时，应当如何处理呢？此时我们可以使用固定效应模型，即假设截距项不随时间变化，以此来控制未观测到的不变的异质性，也就是说假定个人的个性、智力或者某些生理特性保持不变。即它们是一个关于 i 的函数，而不是关于 t 的函数。换句话

① 我们使用“揭示”在于其中性词的属性，因为使用“识别”又过于明确，使用“描述”又过于模糊，所以使用了介于“识别”和“描述”之间的一个中性词汇。

说，我们不让截距项同时随 i 和 t 变化。另外，由于我们把识别协变量系数作为研究的主要目的，因此它们既不随 i 变化也不随 t 变化。此模型被称为固定效应模型。其表达为

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}$$

每个个体有自己的截距系数，以此来表达个体水平未观测到的异质性。此模型估计简单，因为我们假设异质性不随时间发生变化。我们可以使用追踪/面板数据识别这些不随时间变化的异质性。α 是一个完全关于 i 的函数，而与 t 无关。β' 是固定不变的，不随 i 和 t 改变。固定效应模型的优点是控制了不随时间变化的、个体上的异质性。但是，固定效应模型的缺点是浪费了过多的自由度，用来识别固定效应 α。

增长曲线模型

如果个体差异不随时间发生变化，该如何处理？此时我们可以用增长曲线模型（growth – curve models）（Raudenbush 和 Bryk，2002）。增长曲线模型也被称作多水平模型（multi – level models）、分层线性模型（hierarchical linear models）、随机系数回归模型（random – coefficient regression models）、混合效应模型或随机效应模型（mixed effect models or random – effect models）。在统计学中被称为混合模型（mixed models）。在心理学中常被称为增长曲线模型。

增长曲线模型是基本模型与固定效应模型的折中。增长曲线模型的基本思想是分解实际增长曲线（一般是线性的，但也可以是非线性的）。也就是说，将因变量的变异分解为两部分：个人本身的变化（即组内差异，随时间变化）（within – person，over time）和人与人之间的差异（即组间差异）（between – person）。如果 x 轴是时间 t，则每个人都有自己的截距项和自己的增长率。但研究者再把它们进一步分解为个人具体属性的函数。这是完全参数化的表达，因为它假设基础水平和增长率都是观测特征的函数——分别为下面多水平增长曲线模型中的第 2 个方程和第 3 个方程。当然，它也包含随机项成分。这是参数化与随机项结合的多水平增长曲线模型。模型表达为

$$y_{it} = \beta_{i0} + \beta_{i1} x_{it} + \varepsilon_{it}$$

$$\beta_{i0} = \gamma_0 + \lambda_0' W_i + \mu_{i0}$$

$$\beta_{i1} = \gamma_1 + \lambda_1' W_i + \mu_{i1}$$

在描述异质性增长率（heterogeneity growth rate）方面，此模型被经常使用。例如，小孩有时长得很快，有时长得相对较慢，如何解释这类现象？关键在于分解，即将因变量的变异分解为两部分：人与人之间的差异和个人本身的变化差异。人与人之间差异的模型是多水平模型部分，个人本身的变化是增长曲线模型。

2.2 对因果机制干预的研究

研究者往往对教育获得很感兴趣。教育获得会受到家庭环境的影响，但其同样会影响其它许多方面，如婚姻、收入、工作、政治参与等。因此，教育获得实际上可以被看作是一个干预结果，因为它一方面受到家庭环境、智商及其它属性的影响；另一方面它又影响其它方面。

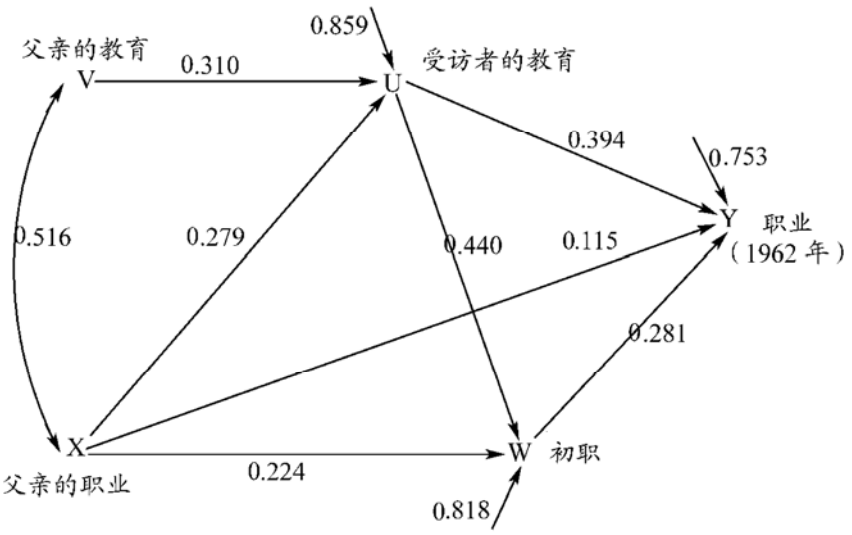
很多影响教育获得的因素是外部引入的，即外生的，但另外一些因素可能是内生的，随时间而表现出来。这也是为什么说，我们所揭示出的因果机制不是必然的原因，在某种程度上讲它是解释性的。

因此，这里的重点实际上是随时间变化的协变量，即变化中的自变量。在使用追踪数据时，要重视随时间变化的协变量。随时间变化的协变量是内生性的还是外生性的，研究者并不一定有明确答案。所以，对于“因果效应”（causal effects）和“因果机制”（causal mechanism），研究者往往并不能够有明确的解释，因为所发现的因果机制的真实原因并非必然是由外在效应引起的。

现在，让我们简单回顾一下图 1 的 Blau – Duncan 模型（Blau 和 Duncan，1967），这是社会学中身份获得的一个经典模型。模型告诉我们，家庭环境主要通过个人的教育获得来影响他的职业。因此，大部分影响来自家庭环境，通过教育间接起作用，即需要由开放的劳动力市场作中介。在现代社会，教育是很重要的，因为雇主并不知道雇员家庭背景是否会影响劳动者的工作效率，而他们往往根据劳动者的教育对其做出的评价。

图 1　Blau – Duncan 的身份获得模型

Figure 1　Blau – Duncan Model of Status Attainment



来源：Blau 和 Duncan（1967）。

另外一个例子是经典威斯康辛模型（Wisconsin Model）（Hauser、Tsai 和 Sewell，1983；Sewell、Haller 和 Portes，1969）。经典威斯康辛模型告诉我们，家庭环境因素影响孩子的成就，但是这种影响主要是通过孩子的非认知能力发生作用的，具体来讲就是孩子求学的愿望和工作的愿望。家庭环境的作用在现代社会是非常大的，其关键在于影响到一个人的愿望，即抱负。如果一个人很有抱负，那么他/她就有可能获得更高的教育水平。类似的，智商在家庭对个人成就的影响中起到了中介机制的作用。

最近，受到威斯康辛模型和 James Heckman 等人（2006）的影响，Hsin 和 Xie（2011）使用早期儿童追踪数据对一个类似模型进行了研究（见图 2）。他们利用追踪数据，试图分解家庭环境是如何通过影响孩子的认知和非认知因素从而对孩子的成就产生作用的。这正是追踪数据能够回答的问题，因为我们知道时变协变量是如何变化的，知道过去的情况是如何通过中间阶段发生的事件来影响未来的。

2.3　对因果效应的研究

2.3.1　因果关系是社会科学的中心问题

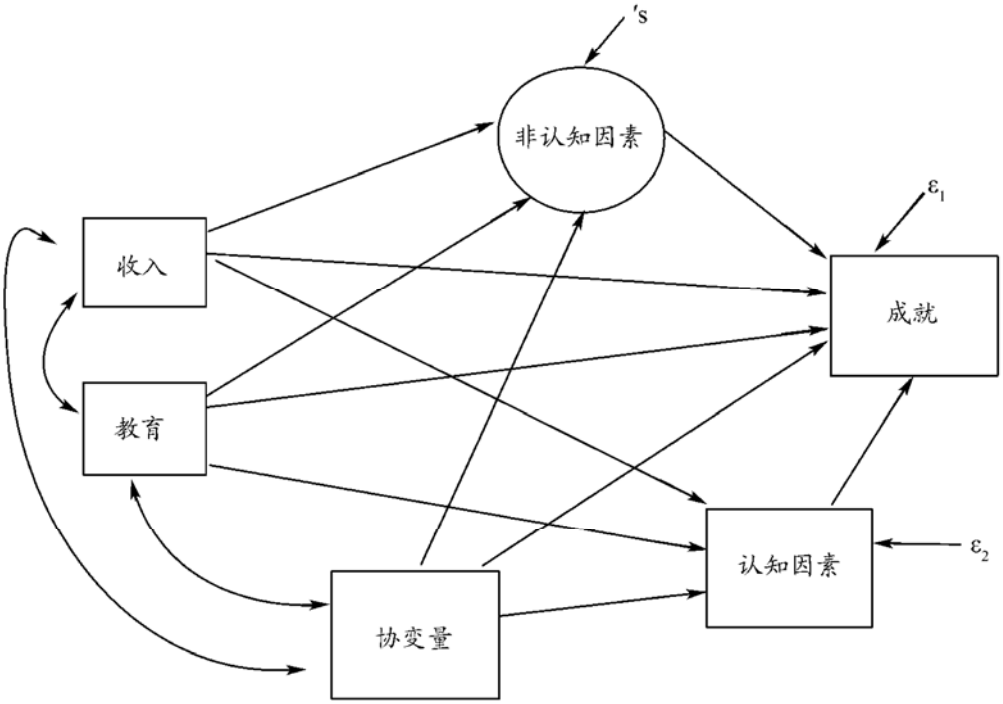
因果推论是目前社会科学研究中一个非常重要的主题。事实上，揭示因果关系是所有科学试图达到的最终目标。纵观科学史，从亚里士多德（Aristotle）到现代基因科学都始终如此。正确认识因果关系可以帮助对未来做出准确预测。对因果推断的重要性，许多科学家，包括许多经济学家，都持有非常强的观点——只有关于因果关系的研究才是真正的社会科学研究。虽然并没有那么极端，但我们认为因果关系是政策干预的科学基础，因果关系也丰富了我们的理论知识。

我们都知道，在时间上真正的原因应该发生在事件之前。原因和效果应该具有时间上的顺序。

但也有例外，有时因果关系只是表面上的时间顺序。关于这一点可以举一个例子，本文作者谢宇曾有幸与社会学家 Duncan 进行过一次交谈，讨论是否可以通过事件发生的前后顺序判断因果关系，在前的是原因，在后的是结果。Duncan 立刻表示了不同意意见。他说圣诞节前一般都会有一个购物高潮，是圣诞节导致购物高潮，还是购物高潮导致圣诞节？细想此问题确实有道理，因为人是有理性的，人们能够按照期望、预先的目的做事。人不像动物，人们实际上是被目的驱使的。这就导致有些事情结果在前，原因在后。目的也是一种原因。所以，不能简单地相信时间顺序能够解决对原因和结果的识别。

图 2 影响子女成就的、含时变变量的结构方程模型

Figure 2 Structural Model with Time – varying Covariates on Children' s Achievement



来源：Hsin 和 Xie（2011）。

截面数据在做因果推断的研究中作用是有限的，因为它们实际上不能告诉我们事件发生时间上的顺序。面对多个有关态度选项的时候，我们不会知道哪个是原因，哪个是效应，因为它们是在同一时间观测到的。而有了追踪数据我们能知道事件发生的时间。这些时间信息对因果推断的研究是非常有帮助的。

2.3.2 因果效应通常是一个反事实问题

出于道德和伦理的原因，许多诸如政策、培训等干预措施不可能对人类进行实验。比如说，在评估学前教育对个人成就的影响时，即使我们能够进行长期追踪，但很难为了设定控制组，人为地让部分幼儿不进行学前教育。因此，在人类社会中极为缺乏实验数据。对于因果推论，我们只能够询问反事实问题，对于同一个接受过“干预”的人，我们想问，如果他/她没有被“干预”过，那将会发生什么。而对于没有接受过“干预”的人，我们也会问同样的问题，如果他/她接受了“干预”，那又会发生什么。因此，对于同一个人，我们会面临多种反事实的可能性。当然，这些反事实的问题在个体上都不可能得到回答，因为在社会科学领域，变异是普遍存在的，个体和个体之间是不可比的，而这也正是我们需要在分组水平上进行因果分析的原因。面对这种情况，社会科学家在一定的假设前提下，充分利用追踪数据的信息，在分组水平上还是能够有所作为的。下面我们用一些具体实

例加以说明。

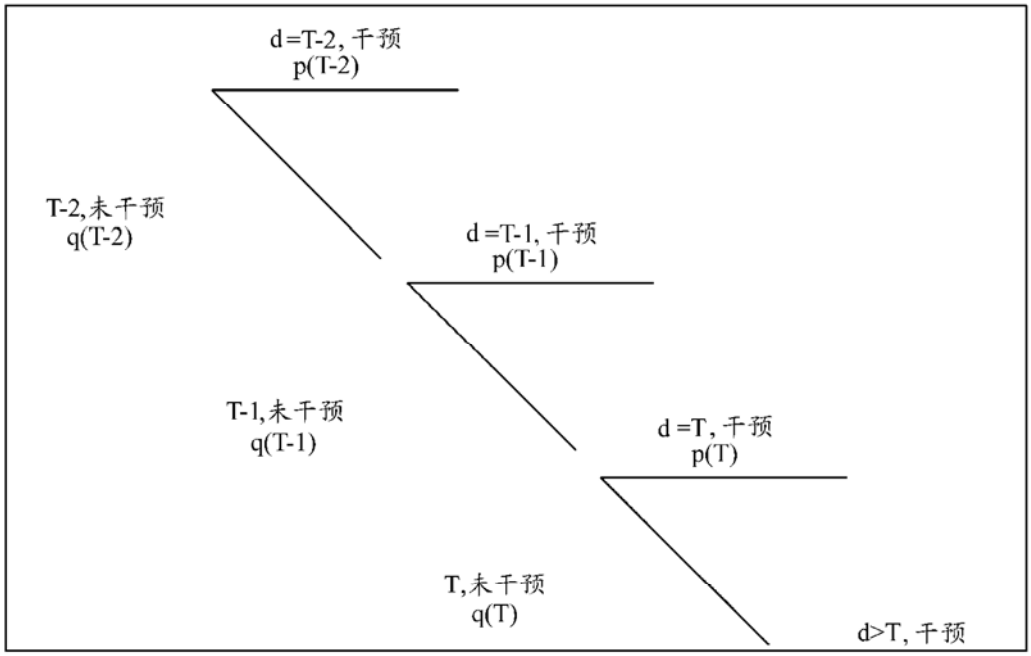
2.3.3　一些利用追踪数据进行研究设计的例子

如果某一干预变量是外生性的，那么可以将它看作是准试验（quasi－experiment）。这是因果关系研究中一个很重要的分支，现在在劳动经济学领域正变得越来越重要，即自然实验（natural experiment）。从方法论来讲，这可以看作是某种革命，要求我们利用识别策略，如运用工具变量方法（instrumental variable），回归间断点（regression discontinuity）和差分法（difference in difference）。这些策略都是某种自然实验，我们试图从这种实验中获得外部的干预，即某些随机发生的事件。而且，我们想知道是否能够用这种外部的干预来识别真正的因果效应。在这个领域的文献中（如，Angrist 和 Krueger，1999；Angrist 和 Pischke，2009），有许多利用管理数据进行的研究。而且这种情况下，信息量往往非常有限。事实上，这样的数据并不多，我们不能只指望利用这种外部的干预作为工具变量。

我们同样可以运用聚类的设计方案（clustering design）研究因果关系问题。例如，同卵双胞胎具有 100% 相同基因，异卵双胞胎具有 50% 相同基因。我们可以用这一差别来进行对于基因影响的因果效应研究。当然，所有这些都需要假设。如果没有假设，那么就不存在因果效应推论的研究。世上没有一个不加任何假设的、非常完美的设计。不同设计之间的差别只是在于使用了不同的假设罢了。

图 3　未来不同时点干预结果树状图（起始为 d = T－2）

Figure 3　Forward Tree of Treatment at Various Time Points（from d = T－2）



注：p(.) + q(.) = 1

来源：Brand 和 Xie（2007）。

在没有其它更好的信息时，研究者还会引入“可忽略性假设”（ignorability assumption）。在此假设下，所有与干预相关联的系统性差异，都可以通过一系列观察到的协变量来概括（Rosenbaum 和 Rubin，1984）。可忽略性假设可以帮助研究者估计因果效应。

但是，在追踪设计框架下需要特殊考虑哪些因素呢？传统因果关系的反事实模型只考虑两个数据时间点（Rubin，1974，1978）：干预时间和结果时间。然而，数据往往形成随时间变化的轨迹，

是时间的函数，而且它们的将来是不确定的。我们认为 Rubin 的模型过于简单化，因为在干预的时点，未来的轨迹并不是一定的。因此，在 Brand 和 Xie（2007）合作的一篇文章中，他们特别研究了此问题。如图 3 所示，人们可能在不同的时点接受干预。但什么是真正的因果关系问题呢？这是研究者们需要非常谨慎考虑的，主要包括两方面：(1) 在设定个体具有不同发展轨迹的条件下，进行未来反事实结果之间的比较；(2) 按照实际概率整合多种未来的（未知的）结果。

在追踪背景下的因果效应，可以通过下列公式表达

$$\Delta_i^t = y_i^{d=t} - y_i^{*\,d>t}$$

此公式表达的含义是，如果某一个体在 $d = t(t = 1, \cdots, T)$ 时被干预，则 $y_i^{d=t}$ 是可能被观测到的结果取值。$y_i^{*\,d>t}$ 是同一个体直到 t 时都没有受到干预的综合结果取值。当然，在个体层面上，我们无法做因果分析。但是，我们可以在总体上定义我们想要知道的统计量

$$\delta_T^{d=T} = E(y_T^{d=T}) - E(y_T^{d>T})$$

为了计算这一统计量，我们会用观测到干预的概率作为整合未来可能性的方法。这只是一种思路，但不是唯一的，其想法是在追踪情景下概念化因果关系。这样一来我们就可以将某一时间 t 的结果与未来的结果作比较。人们可能在时间 t 结婚，将来可能单身，但也可能明年结婚，或者后年结婚。这些在未来都是可能发生的。我们可以用将来实际发生的概率来计算在未来时间 T 时的轨迹。

在现实社会中，对于某些人类行为，人们经常猜想最终结果，在与最终结果比较的基础上推断“干预”对人们行为的影响。如在早婚与晚婚之间做比较，在早孕与晚孕之间作比较。可是这实际上是有悖于因果推论的。鉴于此，我们实际上需要在某一干预时点考虑一个反事实的问题，即将所有未来可能的轨迹看作不确定的。例如，我们可能结婚，但我们真的不知道将来是否离婚（有离婚的可能）。但是，结婚结果如何，可以通过比较没有结婚的人的结果与结婚以及结婚以后所有可能的结果而获得，而结婚后可能的结果包括离婚、因意外而失去配偶、生育子女、结婚后面临的许多事情。

从图 3 可以看到，正确的比较应该是在没有结婚的结果与未来不确定因素综合作用的平均结果之间的比较。这种正确的比较方法，我们称之为“前瞻性系列预期”方法（forward - looking sequential approach）（Brand 和 Xie，2007）。具体来讲，我们并不知道所有可能的结果，不能比较所有未来的结果，我们所能比较的只是各种未来结果在统计上的平均。在设定个体之间存在不同的发展轨迹的情形下，利用反事实假设在未来之间进行比较，要么是 A，要么是 B；如果是 B，将有许多种可能性；如果是 A，也同样存在许多种可能性。但是，所有我们能够考虑到的只是在某一时刻的干预。而反事实结果是整合多种未来结果的平均。

2.4 对“状态”变换的研究

最后，我们讨论有关状态变换（state transition）的研究。为什么追踪数据经常与事件史分析相联在一起？原因很简单。事件史分析（event history analysis）是研究状态变换的一种技术。何为“状态”？状态是一个相对同质性的情况，如在婚、退休。它们是相对稳定的。横截面数据根本不可能告诉我们状态的变化，因为我们只知道什么个体处于哪些状态，如被雇佣，在婚，或者活着。而状态变换是属性随时间的变化。

状态变换的研究需要追踪数据——具有回顾性信息的截面数据也可以。因此，状态变换研究经常与追踪数据联系在一起。其原因在于将个人作为分析单元，其从某一状态变换成另一状态需要时间，如从就业状态变为失业状态，从结婚状态变为离婚状态，从没有孩子到有孩子，等等。这些

状态变换随时间都会发生。这就是我们为什么做事件史分析的原因，因为我们关注状态变换的风险率，而研究从某一属性状态变为另一种属性状态以及在某一时间变成其它属性状态的风险率，就被称之为事件史分析。

状态变换的研究可以是“描述”、“揭示”，或“识别”异质性的。我们通常采用的方法是事件史分析，也被称为期间分析（duration analysis）、风险率模型（hazard rate models），以及变换/风险率模型（transition/hazard rate models）。我们一般将利用变换率或风险率的模型简单地称之为率模型（rate models）。使用了几个世纪的生命表（life table）也属于此类。生命表是一种统一技术，也是研究状态变换的一种非参数方法。我们在此不就风险率模型展开讨论，但它是我们在特定条件下研究状态变换概率的基础。

此方法需要的数据为具有回顾性信息的截面数据（cross - sectional data with retrospective information）和追踪（面板）数据（panel data）。其基本模型为

$$P_{ij} = P_r(T_{i=j} \mid T_{i \geqslant j})$$

即事件发生的条件概率，具体来说就是给定某事件尚未发生，该事件对于个体 i 而言在时间区间 j 内发生的概率。此条件概率只有在知道事件发生时间的情况下才能进行估计。根据时间变量的类型，可以分为离散时间方法（discrete - time methods）和连续时间方法（continuous - time methods）。选择何种方法取决于时间测量的精确程度。率模型要求具有人 - 期（person - period）或时段取向（episode - oriented）的数据结构。只有追踪数据，或具有关于个体从前不同时点状态回顾性信息的截面数据才能满足此条件。因此可以说，事件史分析方法无论从思维逻辑还是从数据结构上都与追踪数据的设计思路非常匹配。

3　结论

纵贯数据的建立都是非常伟大的系统工程，除了非常昂贵之外，收集过程也很困难。但是其回报是显而易见的。然而，纵贯数据并不是完美无缺的，研究者对纵贯数据不要有不切实际的期望。纵贯数据本身不会“点石成金”，并不能使质量较差的数据变得更好。而在正文中我们已经讨论了为什么真正因果关系分析的识别是非常困难的，原因在于它们通常需要将假设与数据良好地结合在一起。

纵贯数据很容易因为样本流失过多或无应答过多而失去它本来的功能。但导致其功能受限的根本原因并不在于流失或无应答本身，而是在于人类和人类行为的内在变异性。虽然有很多关于纵贯数据的缺陷和无应答方面的讨论，但这并不意味着这些缺陷一定会限制我们的研究。确切地讲，我们通过改善研究设计和合理运用统计方法能够弥补这些方面的不足，但这也需要较高的数据质量。所以，纵贯数据的数据质量很重要。当然，再好的数据也不能取代社会科学研究者对社会现象的理解和分析。

参考文献/References:

1　Angrist, Joshua D. and Alan B. Krueger. 1999. Empirical Strategies in Labor Economics:1277 - 1366 in Handbook of Labor Economics, vol. 3A, Edited by O. Ashenfelter and D. Card. Amsterdam: Elsevier.

2　Angrist, Joshua D. and J · rn - Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist' s Companion . Princeton, NJ: Princeton University Press.

3　Baltagi, Badi. H. 2002. Econometric Analysis of Panel Data. New York: Wiley.

4　Blau, Peter M. and O. D. Duncan. 1967. The American Occupational Structure. New York: Wiley.

5　Brand, Jennie and Yu Xie. 2007. Identification and Estimation of Causal Effects with Time - Varying Treatment and

Time – Varying Outcomes. Sociological Methodology 37:393 – 434.

6 Hauser, Robert M., Shu – Ling Tsai, and William H. Sewell. 1983. A Model of Stratification with Response Error in Social and Psychological Variables. Sociology of Education 56:20 – 46.

7 Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. Journal of Labor Economics 24: 411 – 482.

8 Hsiao, Cheng. 2003. Analysis of Panel Data (Second edition). Cambridge University Press.

9 Hsin, Amy and Yu Xie. 2011. Social Determinants and Consequences of Children' s Non – Cognitive Skills: An Exploratory Analysis. Paper Presented at the 2011 Spring Meeting of the ISA RC28, University of Essex, UK 13th – 16th April 2011.

10 Raudenbush, Stephen W. and Anthony S. Bryk. 2002. Hierarchical Linear Models: Applications and Data Analysis Methods (Second edition). Thousand Oaks: Sage Publications.

11 Rosenbaum, O. R. and Donald B. Rubin. 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of American Statistical Association 79: 516 – 524.

12 Rubin, Donald B. 1974. Estimating Causal Effects of Treatment in Randomized and Non – randomized Studies. Journal of Educational Psychology 66: 688 – 701.

13 Rubin, Donald B. 1978. Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics 6:34 – 58.

14 Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. The Educational and Early Occupational Attainment Process. American Sociological Review 34:82 – 92.

15 谢宇. 社会学方法与定量研究. 社会科学文献出版社,2006.
Xie Yu. 2006. Sociological Methodology and Quantitative Research. Social Sciences Academic Press (China).

(责任编辑:宋 严 收稿时间:2011 – 10)