DOI:10.13546/j.cnki.tjyjc.2016.10.021

断点回归设计:基本逻辑、方法、应用述评

罗　胜

(中南财经政法大学 统计与数学学院,武汉 430073)

摘　要:断点回归设计是最接近于随机试验的拟随机实验方法,具有因果推断清晰、结果最为可信、易于检验的特点。文章从断点回归设计的基本逻辑、方法、应用和研究进展等几个方面介绍了这一方法。

关键词:断点回归设计;反事实;因果推断

中图分类号:C81　　**文献标识码:**A　　**文章编号:**1002-6487(2016)10-0078-03

0　引言

在过去二十多年来,社会科学领域中,通过严格的统计方法来进行因果推论受到越来越多的关注。随机实验方法是因果推论实证最优选择,这是学界已经达成的共识。但在社会科学研究中,随机实验方法的应用极为有限,并且随机实验的时间成本和经济成本都比较高,在随机实验不可得的情况下,一种近似于随机实验方法,即拟随机实验方法,受到更多的重视和研究。拟随机实验方法是以统计控制模拟实验控制,从而检验因果假设。断点回归设计就是仅次于随机实验的一种拟随机方法。Lee(2008)认为在随机实验不可得的情况下,断点回归能够避免参数估计的内生性问题,从而真实反映出变量之间的因果关系。随着越来越多的研究开始关注断点回归设计,这一拟随机实验统计方法有着极为广阔的应用前景,但在国内研究中鲜见该方法,基于此,本文将介绍断点回归设计基本逻辑、方法与应用,并探讨其最新发展趋势。

1　断点回归设计的基本逻辑

1.1　哲学逻辑

1848年J.S.密尔从方法论的角度上总结了因果归纳的逻辑,提出求同法和求异法。求同法是指一个群体中所有人都在两个变量上取值相同,而在其他变量上取值不同,那么这两个变量之间就有因果关系。求异法是指两个个体在因变量上的取值不同,存在某个自变量取值不同,而其他自变量上的取值相同,则那个取值不同的自变量和因变量之间存在着因果关系。Holland(1986)通过总结自然科学、社会科学的大量研究和讨论,提出科学的解决方案和统计的解决方案两种解决因果问题的方案,科学的解决方案主要包括重复实验和随机实验,研究者通过重复实验和随机分组实验来操纵和控制,进而研究二者之间的关系。断点回归的主要思想,控制研究的样本近似于随机分布在临界值附近,小于临界值的样本作为控制组,大于临界值的样本作为实验组,通过比较它们的差别来研究干预变量和结果变量之间的因果联系。

1.2　统计逻辑

1935年,统计学家Fisher通过对偶然因素的作用控制,完善了随机化实验设计,他将试验对象随机分配到控制组和实验组,由于个体的各种不可控差异,根据大数定理,在随机分配过程中被平均了,平均而言,两组实验对象可视作是同质总体。统计学对因果关系表述为:在相等条件下,X变化导致Y平均值的变化。Fisher随机实验设计的伟大贡献在于把这种机能联系和类似于硬币的随机翻转相连起来,以保证我们想要切断的联系确实被切断,因为我们可以假定这个随机硬币是不受任何我们可以测量到的因素所影响。Rubin(1984)通过仔细分析实验条件下的因果推论问题,认为这是一个反事实的问题。在统计学理论中,反事实指在相反情境下的某种状态。以新药物实验为例,一群病人在一个实验中被分到实验组接受新药物治疗。这一群病人接受治疗后效果是我们能够观察到的"事实"。而"反事实"则是指"假设"这同一群病人是被分到对照组,而不是实验组,即没有接受新药物治疗,那么他们的症状又如何。在统计学意义上,新药对于症状的因果性效果就是指这同一实验对象在实验组时和在对照组时的之间的差异 。换句话说,统计学上的因果关系可以表述为可观察到的"事实"与其"反事实"之间的差异 。在研究的非实验数据中,对实验组来说,研究者无法观测其在未接受实验时的表现;而对于控制组来说,研究者也无法预测其在接受实验时的表现。我们可以将实验定义为二分变量D,接受实验时,D=1,结果变量为Y1;未接受实验时,D=0,结果变量为Y0。根据反事实框架,因果关系可以表述为:

$$T=p*(E(Y_1|D=1)-E(Y_0|D=1))+(1-p)*(E(Y_1|D=0)-E(Y_0|D=0))$$

基金项目:中南财经政法大学研究生科研创新项目(2015B1902)

作者简介:罗　胜(1987—),男,湖北大冶人,博士研究生,研究方向:国民经济核算、经济统计分析。

其中，T表示因果关系，p表示研究对象接受实验的概率，$E(Y_1|D=1)$、$E(Y_0|D=0)$表示可观测的“事实”，而$E(Y_0|D=1)$、$E(Y_1|D=0)$表示不可观测的“事实”，也即反事实。在实际研究中，反事实是永远观察不到的，一个实验对象，要么在实验组，要么在对照组，只能二者选其一，不能同时出现在两个研究组中，Holland（1986）称之为“因果推论的基本问题”。为了简化因果推论公式，在统计学中做出了非混淆假设，即：

$E(Y_1|D=1)=E(Y_1|D=0)$以及$E(Y_0|D=0)=E(Y_0|D=1)$

非混淆假设要求研究对象是随机地分配到实验组和对照组，即二分量D本身和最后的实验结果Y_1、Y_0没有关系，换句话说，Y_1、Y_0独立于D。传统方法中个体异质性和混杂因素的问题，在断点回归设计中都得到很好地解决。通过统计控制，使得非实验的调查数据尽可能地随机分布在临界值附近，同时，满足非混淆假设，就是要求结果变量独立于干预变量。

2 断点回归设计方法

断点回归最早是由美国西北大学心理学家Campbell在1958年设计出来的。Thistlethwaite & Campbell在1960年正式发表关于断点回归分析的文章，并提出在非实验条件下断点回归是处理处置效应的一种有效办法。此后也有很多学者对该方法进行研究和发展，1984年，Trochim综合前人对断点回归的理论和方法，将断点回归分为两类：一类是确定型的（Sharp RD）（如图1（a）），即个体在临界值X一边接受处置效应的概率为1，另一边则为0；一类是模糊型的（Fussy RD）（如图1（b）），即在临界值X附近，接受处理效应的概率是单调变化的。Hahn等（2001）从模型识别和模型估计上对断点回归进行了严格意义上的理论证明。断点回归的主要原理是：存在一个变量，如果该变量大于这个临界值时，接受处置效应，小于临界值时，不接受处置效应，可以视作是对照组。在确定型断点回归中，临界值是确定的，一边是完全接受处置效应，另外一边是完全不接受处置，而在模糊型断点回归中，临界值附近的观测值接受处置效应概率是单调随机的。

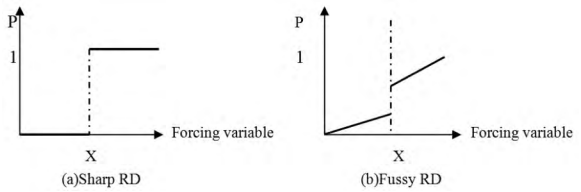

图1 干预效应概率

2.1 假设条件

清晰型断点回归：

（1）将研究对象随机分为对照组和实验组的干预变量是一种已知、确定性的决策规则。

（2）在断点X处，结果变量是连续的，即E(y(0)|x), E(y(1)|x)在断点处是连续的。

模糊型断点回归：

（1）将研究对象随机分为对照组和实验组的干预变量是一种随机的决策规则。

（2）在断点X处，倾向值函数$Pr(d_i=1|x)$存在着跳跃。

2.2 模型

由于清晰型断点回归设计是模糊型的一种特例，本文主要介绍模糊型断点回归模型，可由如下两阶段方程表示：

$$d_i=\phi+\varphi T_i+g(x_i-\tilde{x})+\lambda_i \quad (1)$$

$$y_i=\alpha+\beta d_i+f(x_i-\tilde{x})+\varepsilon_i \quad (2)$$

其中，y_i是研究个体的结果变量；T_i=1表示个体接受处置，反之，不接受；d_i=1表示基于决策规则被分配到实验组，反之，则分配到对照组；$f(x_i-\tilde{x})$和$g(x_i-\tilde{x})$是关于决策变量和控制变量的函数。

由于在模糊型断点回归中，d_i是随机的，T_i类似于d_i的工具变量，通过2SLS回归方法可估计出处置效应τ_{frd}。

$$\tau_{frd}=\frac{\lim\limits_{x\downarrow\tilde{x}}E(Y|X=x)-\lim\limits_{x\uparrow\tilde{x}}E(Y|X=x)}{\lim\limits_{x\downarrow\tilde{x}}E(T|X=x)-\lim\limits_{x\uparrow\tilde{x}}E(T|X=x)}$$

2.3 断点回归设计步骤

（1）检验在临界值X点处是否存在断点。构造箱体$(b_k, b_{k+1}]$，$b_k=c-(K_0-k+1)*h$，其中c为临界值，h为箱体的范围；计算每一个箱体中的样本数量，$N_k=\sum_{i=1}^{N}1\{b_k<X_i\le b_{k+1}\}$，其中$X_i$为决策规则的关键变量；求出每个箱体的平均值$Y_k=\frac{1}{N_k}\sum_{i=1}^{N}Y_i\cdot 1\{b_k<X_i\le b_{k+1}\}$，最后，画出$Y_k$和$X_i$二者之间的曲线图。若在临界值处，存在断点，则可使用断点回归设计。

（2）在临界值附近选择合适的样本进行回归。在临界值附近选择样本，涉及到带宽h的选择，目前主要有交叉验证法和局部多项式回归方法，读者可具体参考相关文献。

3 断点回归设计的应用

断点回归被视为仅次于随机实验，能够有效利用现实约束条件分析变量之间的因果关系的实证方法，因而得到广泛的运用，特别是在教育学、政治学、经济学以及一些政策效应评估上。Angrist & Lavy(1999)利用以色列教育制度，用断点回归方法来研究班级大小对学生成绩的影响，在它们的制度中，班级大小必须小于等于40人，超过40人则必须分为两个班，通过研究发现班级人数越少，其班级成绩越好。Lee等（2004）利用50%的得票率是获选的关键决定因素来研究选举对政策的影响，结果发现选民会选择制定了对自己最有利政策的竞选者。Lee（2008）在同样的背景下，研究美国众议院当选者会不会利用本次当选所得到的权利来影响下次再次当选，通过断点回归结果表明众议院的当选者在下一届选举中获胜的可能性要大些。Cunat（2012）通过分析市场对企业公司治理年度会议中决策的通过与否的反应对股东价值的影响，发现政策建议的

通过会给股东带来正向回报。在国内学者研究中，相关实证文献较为缺乏。雷晓燕等（2010）利用政府对退休年龄的规定，男性65周岁，女性60岁退休，来研究退休对健康的影响，结果发现正常退休对男性健康有显著的负面影响，对女性健康影响不大。曹静等（2014）运用断点回归方法，对2008年北京奥运会后采取的限行政策对空气质量的影响进行评估，发现限行政策尤其是“尾号限行”对空气质量的影响甚微。

4 断点回归设计局限与研究进展

断点回归是拟随机实验方法中揭示因果效应最有效的一种方法，可以视作是一种特殊的倾向值匹配，它不需要对多个混淆变量控制，而是考虑一个个体是否接受某个自变量的影响，不用考虑太过复杂。但断点回归方法也存在着局限性：

（1）在使用断点回归时，如果其他协变量也存在着“中断”的情况，则不清楚是由于其他变量还是我们所关心的强制变量所导致的。

（2）非混淆假设条件严格。断点回归方法假设研究对象是同质的或近似同质的，即被放置对照的个体若放在实验组与放置在实验组的个体产生的效应是一样的，但在实际中很难保证，如若产生异质性反应，则估计结果是有偏的。

（3）断点回归衡量的是在临界值附近的局部平均效应，不是一个整体的平均效应，很难推广到整体研究中。

研究进展：（1）在进行局部线性回归时，选择一个合适的带宽使得估计量无偏且具有效率仍在研究中，Imbens & Lemieux（2007）认为在局部线性回归时矩形核估计最合适，lee（2008）认为核估计在局部线性回归中存在者偏误，夸大平均因果效应。Imbens &Kalyanaraman（2012）认为最优带宽选择是一个开放性的问题。因而最优带宽选择还有待进一步研究。（2）在模糊断点回归中，强制变量对Di的影响不是决定性的，而是随机的，即在临界值左侧也存在进入到实验组的个体，此时Di是一个内生变量，可以采用断点回归的工具变量法解决，Angrist（2009）认为模糊断点回归就是一个工具变量。因而可以采用工具变量的一些衡量估计量的方法来衡量。（3）Papay等（2011）将单个分配变量扩展到多个分配变量的断点回归模型，Reardon & Robinson（2012）根据教育政策的特点，提出了多评分维度的断点回归模型（multiple rating score regression discontinuity ，MRSRD），并讨论了多评分维度断点回归五种模型。Wong等（2013）具体介绍了多元断点回归模型（multivariate regression-discontinuity design ，MRDD），并针对出现两个分配变量时，提出边界方法、中心方法，单变量方法和工具变量法四种方法来估计干预效应。

5 简评

断点回归设计是最接近于随机实验的方法的拟随机实验方法，在微观政策评价方面进行因果推断具有较大的优势，因果推论清晰且易于检验。本文从断点回归设计的基本逻辑、方法、应用和研究进展等几个方面介绍该方法。对于中国这样一个处于转型和发展的国家，政策和规则的改变给实施断点回归设计提供了绝佳背景，这也是本文的着眼点。同时，在使用过程中还应当注意断点回归设计的假设条件和适用条件，避免一味选择该方法造成评价不当。

参考文献：

[1]Hahn J, Todd P, Van der Klaauw W. Identification and Estimation of Treatment Effects With A Regression-Discontinuity Design[J]. Econometrica, 2001, 69(1).

[2]Lee D S. Randomized Experiments From Non-Random Selection in US House Elections[J]. Journal of Econometrics, 2008, 142(2).

[3]Lee D S, Lemieux T. Regression Discontinuity Designs in Economics [J]. Journal of Economic Literature, 2010, (48).

[4]Thistlethwaite D L, Campbell D T. Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment[J]. Journal of Educational psychology, 1960, 51(6).

[5]Trochim W M K.Research Design for Program Evaluation: The Regression-Discontinuity Approach[M]. Newbury Park, CA: Sage, 1984.

（责任编辑/易永生）