断点回归模型原理、方法与应用①

刘维奇[1,2]，刘 唤[3]

（1. 山西大学 管理与决策研究所，太原 030006；2. 山西财经大学 财政金融学院，太原 030006；
3. 山西大学 经济与管理学院，太原 030006）

摘 要：近年来，断点回归（RDD）在实证研究中的应用越来越多。RDD 是能够有效利用实际约束条件分析变量之间因果关系的实证研究方法。与工具变量法和双重差分法相比，RDD 更接近于随机实验设计。本文首先对 RDD 的基本原理进行了阐述，并对 RDD 的估计方法和步骤进行了概述，指明各种估计方法适用的情形和具体操作步骤。RDD 可以分为精确断点回归（SRDD）方法和模糊断点回归（FRDD）方法，本文通过含有断点的模拟数据和国家退休政策对男性投资者风险偏好影响两个实例，展示了 SRDD 和 FRDD 在实证分析中的应用，并给出了 SRDD 和 FRDD 的 stata 命令实现。

关键词：断点回归；精确断点回归；模糊断点回归

中图分类号：C93，F8 **文献标识码**：A **文章编号**：（2019）01-0096-11

0 引言

近年来，断点回归（RDD）方法越来越受研究者的青睐，并且已经成为一种主流的实证研究方法。RDD 是能够有效利用实际约束条件分析变量之间的因果关系的实证研究方法。与工具变量法（IV）和双重差分法（DID）相比，RDD 更接近于随机实验，因此 RDD 方法得到的因果关系结论也更为可信。本文就 RDD 方法的基本原理和方法步骤进行了阐述。

在因果关系的实证研究中，最优的研究方法应当为随机实验，然而，随机实验需要花费较高的时间成本和经济成本。Thistlethwaite 和 Campbell[1]首次提出 RDD 是在非实验的情况下处理处置效应的一种有效的方法。但是在早期，RDD 在经济学研究领域并没有得到广泛应用，直到 20 世纪 80 年代，RDD 才被大量的应用到经济学研究领域，对经济变量之间的因果关系进行识别。

在实际中，新的政策或者制度的出台往往会引起实施对象跳跃性的变化。普通的回归方法只适用于分析研究对象的变化趋势，而不能精确的研究处置操作和研究对象之间的因果关系。RDD 方法是专门针对某一处置操作下，接受处置对象和处置操作之间因果关系的因果分析方法。RDD 方法也有一定的局限性，只适用于研究因变量的观测结果存在跳跃性变化的因果关系实证研究，对于本来不存在跳跃性变化的情况不可以用 RDD 方法研究，否则会导致错误的因果关系结论。

处置操作对个体的因果效应，可以通过对比个体在接受处置的情况下和没有接受处置的情况下结果变量的差异来研究。一般而言，个体在接受处置的情况下，无法观测到其没有接受处置的情况。RDD 方法下，假设表示个体某一特征的变量大于一个临界值时，个体接受处置，而表示该特征的变量小于临界值时，个体不接受处置。接受处置的个体进入实验组，未接受处置的个体进入控制组。这个决定个体接受或者不接受处置的变量称为分组变量。使用 RDD 方法进行因果效应研究之前，必须验证分组变量的不可操纵性，否则结果变量的跳跃可能是由分组变量引起的，而不完全是由处置变量引起的。RDD 方法要求控制组的潜在处置结果和实验组的实际处置结果有相同的基本特征分布，实验组的潜在未处置结果应该和控制组的实际观测结果有相同的基本特征分布。在 RDD 中，尤其是在变量连续的情况下，接受处置前临界

① 基金项目：山西省“1331 工程”重点创新团队建设计划资助项目（TD008）。
作者简介：刘维奇（1963—），男，山西忻州人，管理学博士，山西大学管理与决策研究所教授，博士生导师，研究方向：金融工程与风险管理，E-mail：liuwq@sxu.edu.cn；刘唤（1991—），女，山西临县人，山西大学经济与管理学院博士研究生，研究方向：金融工程与风险管理，E-mail：lhuaner@126.com。

值附近的样本是非常相似的，除了是否接受处置的区别外其他条件都可视为相同的，所以小于临界值的个体可以作为一个很好的控制组来反映个体没有接受处置时的情况，临界值两侧样本的差别可以很好地反映处置变量和结果变量之间的因果联系。门限模型中个体的行为也发生了彻底的改变，而断点回归模型中个体的行为并没有发生改变。

Lee 和 Card[2]认为在随机实验不可得的情况下，RDD 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。Lee 和 Lemieux[3] 对 RDD 在经济学研究中的应用规范进行了综述。Johnes 和 Tsionas[4]将 RDD 推广到用随机边界代替最优拟合的情形。

Trochim[5]在前人研究的基础上，将断点回归方法分为两类，一类是精确断点回归（SRDD），另一类是模糊断点回归（FRDD）。Hahn 等[6]证明，这两种 RDD 的方法都可以用断点两边的局部样本区间来研究处置效应这种因果关系。

本文对 RDD 的估计方法和步骤进行了综述，并对各种方法的使用情况做了全面分析，对 RDD 的合理有效应用具有指导性意义。本文将按照 RDD 的估计方法和步骤、RDD 的有效性和 RDD 案例分析及其 stata 命令实现，阐述 RDD 模型原理、方法及应用，旨在为相关研究者提供借鉴。

1 断点回归原理

1935 年，统计学家 Fisher 通过对偶然因素的作用控制，完善了随机实验设。由于研究对象个体之间往往存在各种不可控差异，对研究造成了严重的干扰。Fisher 将实验对象随机地分配到控制组和实验组，根据大数定律，控制组和实验组之间的个体差异被随机分配过程中平均了，在平均意义上，这种随机分配方法得到的控制组和实验组可以被视为是无差异的。RDD 方法也引用了 Fisher 提出的随机分配的方法，RDD 方法假设断点两侧局部区域内个体是随机分配的，所以，平均而言，断点两侧个体是同质的。RDD 方法得到的因果关系为处置变量变化导致的因变量的平均变化。

Trochim[5]在已有的对 RDD 理论和方法研究之上，根据研究对象接受处置概率的特征，将 RDD 方法分精确断点回归（SRDD）和模糊断点回归（FRDD）。这两种 RDD 的方法都可以用断点两边的局部样本区间来研究处理效应这种因果关系。

使用 RDD 进行回归分析前，需保证协变量的条件密度函数在临界点处是连续的，这样才能保证因变量变化完全是由处置变量引起的。如果协变量的条件密度函数在临界点处是不连续的，因变量的变化也可能是由协变量引起的，这会影响处置变量和因变量之间的因果关系判断。将每个协变量分量作为因变量进行 RDD，可以判断协变量的条件密度函数是否是连续的。如果协变量回归结果在临界点也存在跳跃，则说明该协变量的条件密度函数在临界点处是不连续的，这说明因变量在临界点处的差异有一部分是由协变量引起的，而不只是处置变量的因果效应。McCrary[7]提出了通过检验某一点处左右极限是否相等来检验函数是否在该点连续。

1.1 精确断点回归

当处置变量是分组变量的确定性、不连续函数时，可以使用 SRDD。SRDD 中，临界值（断点）的一边都是未接受处置个体的观测值，临界值的另一边都是接受处置个体的观测值。在 SRDD 中，处置变量取值为 0 或者 1，表示个体接受处理的概率。SRDD 的潜在结果是连续的。

假设 D 为处置变量，$\boldsymbol{X}$ 为协变量（向量），$X=X_0$ 为临界值，Y 为结果变量。SRDD 中，处置变量 D 是协变量 $\boldsymbol{X}$ 的确定的、不连续函数，取值为 0 或 1［例如，图 1（左）］。$D=0$ 表示个体进入控制组，$D=1$ 表示个体进入处置组。结果变量 Y 是协变量 $\boldsymbol{X}$ 的函数。

由于在临界值两侧的局部区间内个体是随机分配的，所以可以通过估计临界值附近的局部平均处置效。

$$E(Y(1)-E(Y(0))\mid X_0)=E(Y(1)\mid X_0)-E(Y(0)\mid X_0) \tag{1}$$

假设在实验之前，结果变量 Y 与协变量 $\boldsymbol{X}$ 之间存在以下关系：

$$Y=a+f(X)+b*D+u_1 \quad (a,b \text{ 为常数}) \tag{2}$$

式中，$f(\cdot)$ 为 X 的多项式；u_1 为残差项，服从正态分布。当 $X<X_0$ 的个体不接受处置，当 $X\geqslant X_0$ 时，对个体进行处处置操作。如果处置操作对个体的结果产生了一定的效应，则在临界值 $X=X_0$ 处，结果变量 Y 会出现断点。$Y(1)-Y(0)$ 为处置效果。

由于决定个体是否进入处置组的是协变量 X 取值是否超过临界值 X_0，为了方便研究，我们可以将协变量和临界值的差 $X-X_0$ 作为自变量，则式（2）可以表示为

$$Y=a+f(X-X_0)+b*D+u_2 \tag{3}$$

式中，u_2 为残差项，服从正态分布。为了消除内生变量对结果的误差影响，模型中可以加入自变量和处置变量的交乘项 $f(X-X_0)D$，式（3）变形为

$$Y=a+f(X-X_0)+bD+f(X-X_0)*D+u \tag{4}$$

式中，u 为残差项，服从正态分布。如果在实验之前，结果变量 Y 和协变量 X 的关系式中包含高次项，那么可以引入高次项，并限定自变量的取值范围，保证样本数据的可靠性和估计结果的精确性。

$$Y=a+g(X-X_0)+b*D+g(X-X_0)*D+u, X_0-h\leqslant X\leqslant X_0+h \tag{5}$$

式中，$g(\cdot)$ 为 $(X-X_0)$ 的高次多项式，h 表示带宽。最优带宽可以预先给定，或者通过平均方差最小化来确定。

Thistlethwaite 和 Campbell[8]首次使用 SRDD 的方法研究了在升学考试中获得国家奖学金的大学生是不是更愿意读研究生。奖学金制度规定，只有升学考试分数超过规定分数线的时候才会获得国家奖学金。Thistlethwaite 和 Campbell[8]通过对比分数刚好低于和高于规定分数线的学生的研究生入学率来研究这一因果效应。文献用精确 RDD 的方法，拟合研究生入学率和入学成绩之间的关系，通过奖学金分数线附近入学率和入学成绩之关系是否存在断点来判断奖学金是否会影响高中生读研的决定。设发放奖学金这一行为视为处理操作，以奖学金分数线为临界值，获得奖学金的学生进入处理组，未获得奖学金的学生进入控制组。如果处理组和控制组之间的关系在临界值处形成了断点，说明存在因果关系。Dell[9]用精确 RDD 方法，以地理距离为变量，地理边界为断点，研究了 16～19 世纪西班牙殖民政府在秘鲁某些地区实行的米塔（Mita）劳役制度对经济发展的影响。Chen 等[10]用清晰 RDD 的方法，以秦岭和淮河界为断点，将中国地区分为南方和北方，研究了北方地区由于冬季供暖造成的环境污染对北方地区人均寿命的影响。

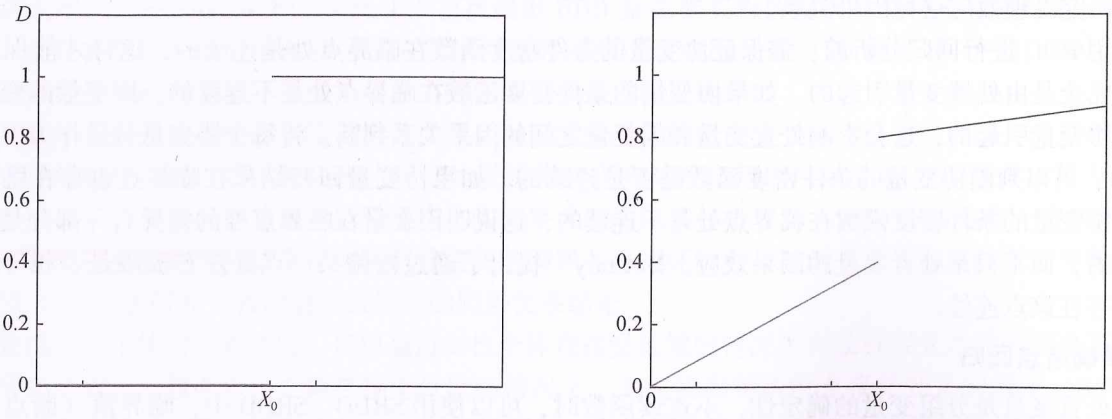

图 1　SRDD 的处置概率（图左），FRDD 的处置概了（图右）

Fig. 1　Disposal probability of SRDD (the left) and treatment probability of FRDD (the right)

1.2　模糊断点回归

FRDD 方法，处置变量不是分组变量的不确定、非连续函数，临界值两侧个体接受处置的概率不是从 0 跳到 1，而是从 a 跳到 b（例如，图 1（右）），其中 $0<a<b<1$。因为临界值两侧个体是否受到处置的情况是模糊的，所以称为 FRDD。由于处置变量不是分组变量的确定性函数，这会引起内生性问题。

同样假设 D 为处置变量，$\boldsymbol{X}$ 为协变量（向量），$X=X_0$ 为临界值，Y 为结果变量。FRDD 中，处置变量 D 不再是协变量 $\boldsymbol{X}$ 的确定性函数。假设 T 为个体接受处置的情况，取值为 0 或 1。$T=0$ 表示个体不接受处置，$T=1$ 表示个体接受处置。FRDD 可以用以下两阶段模型表示

$$D=a_0+a_1*T+g(X-X_0)+v \tag{6}$$

$$Y=\alpha_0+\beta*D+f(X-X_0)+u \tag{7}$$

式中，$f(\cdot)$ 和 $g(\cdot)$ 是关于分配变量（控制变量）和处置变量（决策变量）的函数，v，u 为残差项。如果 $T=1$，$D\leqslant a$，则表示处置个体被分配到控制组，$T=0$，$D\geqslant b$，表示未接受处置的个体被分配到了处置组。

用 FRDD 方法回归得到的处置因果效应为

$$Z=\frac{\lim\limits_{X\to X_0^-}E[Y|X-X_0]-\lim\limits_{X\to X_0^+}E[Y|X-X_0]}{\lim\limits_{X\to X_0^-}E[D|X-X_0]-\lim\limits_{X\to X_0^+}E[D|X-X_0]} \tag{8}$$

式（8）的分子就是精确断点回归的平均处置效应，分母是得到处置的概率在临界点处的跳跃。

FRDD 中，个体能够操纵参考变量，导致部分符合规定的个体进入了控制组，而部分不符合规定的个体进入了处置组。

Angrist 和 Lavy[11] 以以色列为例，用 FRDD 方法研究了班级规模对学生成绩的影响。在以色列，学校班级的规模不得超过 40 人。当学生总人数为 $K*40+P$ 时，所有学生将被分为 $K+1$ 个班。作者发现学校很多班级的人数未达到 40 人，但是被分到了两个班，说明参考变量是被操纵的。文章通过研究发现，班级规模对考试成绩具有显著影响。杨杨等[12] 用 FRDD 的方法研究了高质量内部控制对企业有效税率的影响。文献选用 2011～2016 年上证 50 指数和上证 180 指数的成分股为样本，企业进入上证 50 指数和上证 180 指数的概率为处理变量，研究企业内部控制质量与有效税率的因果关系。郭四维等[13] 以 1986 年义务教育法的实施时间为断点，利用 FRDD 的方法研究了教育对健康的影响。

2 断点回归（RDD）估计方法和步骤

使用 RDD 方法前提条件是存在断点，在使用 RDD 方法进行实证分析之前，必须先验证是否满足 RDD 的前提条件。例如，$y=3x$ 在（0，1）是连续函数，但是如果在 $x=a$，（$0<a<1$）左右邻域选取样本 $\{x_1\leqslant x_2\leqslant x_3\leqslant a\leqslant x_4\leqslant x_5\leqslant x_6\}$，并用 SRDD 方法进行回归，则会导致放大 $x=a$ 的右边邻域函数值，缩小 $x=a$ 的左边邻域函数值，导致 $x=a$ 处出现断点，得出 $y=3x$ 在 $x=a$ 处存在断点的结论。故，在使用断点回归方法之前，需先验证是否存在断点。

2.1 RDD 的估计方法和步骤

RDD 模型常用的估计方法是参数估计法和非参数估计法。参数估计法主要是指局部多项式回归，当总体分布已知的时候可以选择用参数估计方法进行估计。RDD 多项式估计是关于协变量的多项式，多项式次数一般取一次到四次。非参数估计法主要指局部线性回归，当总体分布情况未知的时候选用非参数估计方法进行估计。由于 RDD 要求除了是否接受处置之外，控制组和处置组的其他情况几乎是一致的，所以影响 RDD 结果精确性的一个关键因素是样本带宽，过大或过小的带宽都会导致研究结果不准确。若带宽过大，样本所选个体的数据离临界点太远，个体间本身偏差较大，造成研究结果不准确；带宽过小，所选带宽内可用于研究的数据太少，数据有限导致研究结果方差过大。刘凯[14] 给出了 RDD 的非参数置信区间优化的方法。

1）使用 RDD 首要考虑的是分组变量选取问题，并且要考察分组变量是否被操纵。要保证所选分组变量没有被实验者所操纵，但是该分组变量仍然会对实验结果产生影响。如果分组变量也是被操纵过的，那么有可能最终因变量表现出来的因果效应是由分组变量引起的，而非处置变量引起的。例如，某公司规定，只要业绩达到 60 分就会有年终奖，而对于超过 60 分的部分没有任何奖励。在这种情况下，员工可以通过控制自身努力程度来操纵业绩。部分有能力达到 60 分以上的员工可能会控制自己的业绩，认为多余 60 分的付出是无意义的，这会导致 60 分以上的业绩几乎为零，在年终奖制度执行前后员工的平均业绩没有变化。在该例子中，除了年终奖激励这个处置变量在变化，员工自身努力也在变，如果用 RDD 方法分析年终奖制度对员工业绩的影响，回归结果中不会出现断点，会得出年终奖不会导致员工业绩更好的结论。事实上，年终奖制度导致员工努力程度发生了变化，年终奖制度已经对员工的业绩产生了影响，年终

奖制度和员工业绩之间是存在因果关系的。

缺乏实验数据的计量经济学识别方法往往需要建立在外生性假定基础上，处置变量对结果变量的影响与误差项无关。在处置变量直接导致结果变量变化的情况下，用回归的方法来识别因果效应才是充分有效的。在上述例子中，如果公司未公开年终奖的业绩达标线，那么员工会系统性地控制自己的业绩，但是由于年终奖业绩达标线是未知的，所以在业绩达标线局部区域内的员工可以被视为随机分配到业绩达标线两侧的，低于业绩达标线的员工进入控制组，高于业绩达标线的员工进入处置组。因此，在公司年终奖业绩达标线未知的情况下，可以使用 RDD 方法来分析年终奖制度对员工业绩之间的因果效应。

2）估计方法的选定。RDD 的估计方法有参数估计、半参数估计和非参数估计三种。首先画出结果变量关于分组变量的散点图，并将临界值左右两侧的分组变量取值划分成小区间，求出每个小区间内结果变量的平均值，并画出结果变量关于分配变量的曲线图。若结果变量的取值在临界值 X_0 处存在断点，则可以使用 RDD 方法分析因果效应。否则，有可能结果变量在临界值处本来是连续的，错误的应用 RDD 方法会导致错误的因果关系结论。对于 SRDD 的情形，参数估计法是将每个小区间内结果变量的均值作为因变量，协变量的高次多项式和处置变量作为自变量，在临界值两侧所选带宽范围内分别做回归，得到临界值两侧因变量的拟合值。用处置组因变量拟合值减去控制组因变量拟合值，得到因变量的处置效应。对于 FRDD 的情形，参数估计是将协变量及其与处置变量的乘积，或者分组变量的多项式和处置变量的乘积作为工具变量来进行 2SLS 估计。参数估计法的局限是要求总体分布形式已知，对于总体分布未知的情形是不能用参数估计法的。

对于总体分布形式未知的情形要用非参数估计法进行估计，利用所选样本来推断总体分布。RDD 非参数估计法是利用核密度函数局部线性回归来代替两步最小二乘估计（2SLS）中的一般线性回归（rdrobust 命令可以直接实现这种估计），非参数估计方法允许个体被分到处置组的概率在断点处不连续，并且此概率大于 0 小于 1。非参数估计法适用于总体分布不知道的情形。非参数估计的方法主要有核函数法、最近邻函数法、样条函数法、小波函数法，等。

SRDD 的非参数估计法，在一个小区间内进行加权最小二乘估计，权重由核函数来计算，离临界点越近的点权重越大，以核函数法为例，SRDD 的非参数估计法为最小化以下目标函数：

$$\min_{a,b,c,d}\{K((X-X_0)/h)[Y-(a+b*(X-X_0)+c*D+d*(X-X_0)*D)]\} \tag{9}$$

其中，$K(\cdot)$ 为核函数，一般为三角核或矩形核。

半参数估计法是将断点两侧的样本分别用非参数估计法对因变量进行拟合，再用右边因变量拟合值减去左边因变量拟合值得到处置变量对结果变量的因果效应。

3）检验间断点的起因。保证间断点处的跳跃只是由解释变量引起的，否则估计出的因果效应可能是由其他变量引起的，会导致因果效应分析错误。

Thistlethwaite 和 Campbell[8]研究了在升学考试中获得国家奖学金的大学生是不是更愿意读研究生的问题。在使用 RDD 方法研究国家奖学金制度和大学生上研究生决策之间的因果关系之前，必须先验证引起研究生升学率发生跳跃的原因。假如除了国家奖学金制度之外，同时有另外一个制度，规定所有毕业的研究生都保证百分之百就业，那么该研究生就业保证制度也会影响研究生升学率的上升，研究生升学率在间断点处的跳跃起因有国家奖学金制度和研究生就业保证制度两个变量引起的，而不是完全由奖学金制度导致的。这种情况下直接用 RDD 方法会放大国家奖学金制度对大学生是否上研究生的决策的影响。

4）稳健性检验。为了得到一个好的、精确的检验结果，要对检验结果做稳健性检验。当不断改变带宽和调整多项式次数，检验结果稳定，说明检验结果稳健。

5）为了进一步验证结果的稳健性，将前定变量加入回归模型中进行检验。如果加入前定变量，结果仍然稳定，说明检验结果稳定。

2.2 RDD 的有效性

RDD 的有效性可以通过检查临界值处结果变量是否存在不连续来进行测验。无效的或者低效的 RDD 可能会导致错误的因果效应结论。Guido 和 Thomas[15]认为 RDD 的有效性依赖于对协变量的外推，或者至少在协变量有不连续的那个领域内外推，所以条件函数的设定要精确。不精确的条件期望函数可能会导致

条件期望本身不连续的情况认为是个体处理的因果效应。由于条件期望函数的精确性是不可观测和预估的，所以我们只能通过缩小区间来降低这种误差。

3 RDD 应用案例及 stata 命令实现

为了简明清晰地展示 SRDD 方法和 FRDD 方法在因果关系实证研究中的应用，本文分别给出了 SRDD 方法应用和 FRDD 在具体案例中的应用，并给出了 stata 实现命令。

3.1 SRDD 应用的 stata 命令

假设结果变量为 y，分组变量 x 服从（0，1）均匀分布，引起结果变量 y 变化的其他变量为 z，z 服从均值为 0，方差为 0.25 的正太分布，D 为处置变量，取值为 0 或 1。我们假设结果变量是分组变量的二次多项式：

$$y=x+x^2+2D+z+e \tag{10}$$

3.1.1 数据与研究方法

为了研究方便，我们用 stata 生成一组数据，并取前 1000 个数据作为实验分析数据。设临界值为 $x_0=0.5$，当 $x>0.5$ 时个体接受处置，$D=1$，当 $x\leqslant 0.5$ 时，个体不接受处置，$D=0$。接受处置前结果变量 y 是连续的。在临界点 $x=0.5$ 两侧局部区间内，除了处置操作的差异外，在随机分布意义上个体是无差异的。因为 $\lim\limits_{x\to 0.5^-} y(D=0)\neq \lim\limits_{x\to 0.5^+} y(D=1)$，故结果变量 y 在临界点 $x=0.5$ 处出现跳跃。由于个体处置概率取值为分组变量 x 的确定性函数，所以我们用 SRDD 的方法来分析处置变量 D 和结果变量 y 之间的因果关系。

3.1.2 SRDD 的 stata 命令实现

在 stata 中新建一个 .do 文件，输入一下命令，并将该文件命名为 sharp RDD. do

```
* ————————————sharp RDD.do 命令开始————————————
clear
set obs 1000  //选取前 1000 个数组
set seed 123
gen x=runiform()  //生成均匀分布随机变量 x,作为分组变量
gen z=rnormal()*0.5  //生成正态分布随机变量 z
gen D=0  //处置变量 D 的初始值为 0
replace D=1 if x>0.5
//当分组变量 x 的取值大于 0.5 时,处置变量取值变为 1,个体接受处置
gen e=rnormal()/5  //误差项 e 服从正态分布
gen y1=D*2+x^2+x+0.5*z+e     //得到处置组的结果变量 y1
gen y0=x^2+x+0.5*z+e  //得到控制组的结果变量 y0
gen xc=x-0.5
label var x  "分组变量(Assignment variable)(x)"  //设置横坐标标签
label var xc "Centered Assignment variable(x-c)"
label var D  "D=1 for x>0.5,D=0 otherwise"
save "RDD_simu_data0.dta",replace
twoway(scatter y1 x,msymbol(+)msize(*0.4)mcolor(black*0.3))  ///
     (qfit y1 x if D==0,lcolor(red)  msize(*0.4))  ///
     (qfit y1 x if D==1,lcolor(red)msize(*0.4)),  ///
xline(0.5,lpattern(dash)lcolor(gray))  ///
//以 x 为横坐标,y 为纵坐标,画出 y 关于 x 的平面曲线图
text(3 0.2 "控制组(Control)")text(5 0.8 "实验组(Treat)")  ///
//设置标签位置
legend(off)xlabel(0 0.5 "Cut point" 1)  ///  //横坐标刻度设置
ytitle("结果变量(outcome variable)(y)")      //设置纵坐标标签
* ————————————命令结束————————————
```

3.1.3　SRDD 结果分析

在 stata 中运行上述命令，得到图 2。从图 2 结果显示可以看出，个体在接受处置后会有一个明显的跳跃，而除了处置状态的差异外，临界点两侧的观测值是无差异的，故该结果变量的跳跃是由处置变量引起的，所以该处置变量和结果变量之间存在因果关系，断点处的跳跃反映了这一因果效应。在具体的实证研究中，可以省去数据生成过程，直接将数据导入 stata 进行回归。

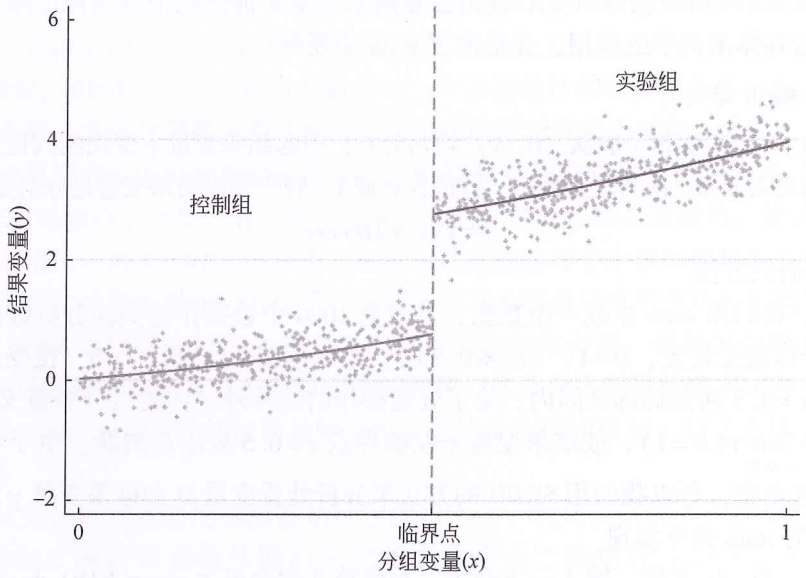

图 2　精确断点回归（SRDD）运行结果

Fig. 2　Results of SRDD

3.2　FRDD 应用案例分析

本文从退休的角度研究了退休对男性居民投资风险偏好的影响。国家退休政策规定男性正常退休年龄为 60 周岁。由于退休政策并不是完全强制实施的，部分居民即便达到了退休年龄仍未做到真正意义上的退休，而有的员工会由于一些特殊的原因而选择提前退休。所以本文采用 FRDD 的方法研究退休和男性居民投资风险偏好之间的因果关系。退休会对居民带来两方面的变化：一方面，居民在退休后没有了工作的压力，会有更多的时间来做风险投资，同时退休会使退休者的角色发生变化，导致其心理上也发生一定的变化，退休行为会导致退休者对投资的风险厌恶程度会有所降低；另一方面，由于高龄导致的身体素质问题，退休后人们也会花费更多的时间成本和经济成本去养生，退休行为会导致退休者对投资的风险厌恶程度增加。对上述两种影响下，退休行为对投资风险厌恶程度的影响，到底哪一种影响起主导作用，本文用 FRDD 方法做了实证分析。

3.2.1　数据与研究方法

本文采用问卷调查的数据收集方法，对某公司的 60 名在职和已退休同级男性员工发放问卷来收集数据。其中包括 10 名 57 岁，10 名 58 岁，10 名 59 岁的在职男性员工和 10 名 60 岁，10 名 61 岁，10 名 62 岁的已退休员工。问卷内容包括年龄（周岁），是否退休（是，否），近五年收入水平（金额），近五年投资水平（金额），近五年家庭中是否有意外发生（交通事故，重大疾病），儿女是否都已成婚并有稳定工作。调查内容不包括留薪停职和退休返聘信息，所以样本中男性员工真实的退休情况是模糊的。本文将信息完整，且近五年家中无意外发生，儿女已成婚并有稳定工作的 57 个调查对象作为最终样本。

男性员工的投资风险偏好（I）用投资净增长比来度量，具体表示为

$$\text{男性员工投资风险偏好}(I)=\frac{\text{上一年投资金额}-\text{本年投资金额}}{\text{上一年投资金额}}$$

表 1 为未达到国家规定退休年龄的男性员工变量的描述性统计结果，表 2 为已达到国家规定退休年龄的男性员工变量的描述性统计结果，表 3 为全样本变量的描述性统计结果。

表1 低于法定退休年龄子样本变量的描述性统计结果

Table 1 The descriptive statistics for ages below the statutory retirement

变量	Obs	Mean	Std.	Min	Max
age	28	58	0.861	57	59
T	28	0	0	0	0
D	28	0.1	0.43	0.05	0.15
I	28	0.0001	0.0003	-0.0007	0.0008

表2 达到法定退休年龄子样本变量的描述性统计结果

Table 2 The descriptive statistics for ages reach the statutory retirement

变量	Obs	Mean	Std.	Min	Max
age	29	60.966	0.823	60	62
T	29	1	0	1	1
D	29	0.898	0.412	0.85	0.95
I	29	0.0312	0.0178	-0.0073	0.0769

表3 全样本变量的描述性统计结果

Table 3 The descriptive statistics for ages reaching the statutory retirement

变量	Obs	Mean	Std.	Min	Max
age	57	59.509	1.713	57	62
T	57	0.509	0.504	0	1
D	57	0.506	0.405	0.05	0.95
I	57	0.0159	0.0201	-0.0073	0.0769

对比表1和表2，达到国家退休年龄的男性员工的平均风险投资偏好要比未达到国家退休年龄男性员工的平均风险投资偏好更大，风险偏好的波动程度也更大，也就是说，达到退休年龄的男性员工风险厌恶程度普遍会降低，风险投资的波动程度增加。这说明男性员工在退休后会有更多的时间去关注投资，花费更多的经济在风险资产的投资上。

对比表1、表2和表3，发现表3中男性员工投资风险偏好的波动程度要比表1和表2都大，而该波动的差异是由退休年龄临界点处的波动引起的，说明在退休年龄临界点处出现了一个更大的波动。上述分析表明，退休会引起男性员工风险偏好增加。

为了更清晰地观测退休前后男性员工投资风险偏好在法定退休年龄（60周岁）临界点处的变化，以及退休制度对投资风险偏好的因果效应，本文用FRDD方法对退休年龄限制和投资风险偏好进行了因果关系分析。以法定退休年龄60周岁为临界点，低于60周岁的样本为控制组，60周岁及60周岁以上的样本为处置组，用FRDD方法进行回归。员工年龄用age表示，假设男性员工的实际退休概率 D 为一线性函。员工是否达到法定退休年龄用 T 表示，当员工达到法定退休年龄时 $T=1$，否则 $T=0$。

先对处置变量 D_i 进行回归，处置变量回归模型如式（11）所示：

$$D_i=a_0+a_1*T_i+a_2*\text{age}_i+a_3*T_i*\text{age}_i+\mu_i \quad (i=1,2,\cdots,57) \tag{11}$$

第二阶段对结果变量 I_i 回归，结果变量回归模型如式（12）：

$$I_i=\alpha_0+\beta*D_i+\alpha*\text{age}_i+\gamma*\text{age}_i*D_i+\varepsilon_i \quad (i=1,2,\cdots,57) \tag{12}$$

3.2.2 FRDD的stata命令实现

在stata中新建一个.do文件，输入一下命令，并将该文件命名为fuzzy RDD.do

```
* ————————————fuzzy RDD.do 命令开始————————————
clear
*use "C:\Users\A\Desktop\retirement12-a.dta"
  //利用未达到法定退休年龄的男性员工子样本数据
*use "C:\Users\A\Desktop\retirement12-b.dta"
  //利用已达到法定退休年龄的男性员工子样本数据
use "C:\Users\A\Desktop\retirement1219.dta"
  //利用全样本数据
globalsizebin 1  //样本划分小区间的区间长度
gen bin=floor(age/ $sizebin)  //确定每个小区间内的样本
genmidbin=bin* $sizebin+0.5* $sizebin  //对每个小区间,取区间中点
bys bin:egen mean=mean(I)
  //求每个小区间内结果变量 I1(男性员工投资风险偏好)的均值
gen T=0
replace T=1 if age>=60  //年龄达到法定退休年龄时 T=1,否则 T=0
reg D T age m,robust  //处置变量回归,(10)式
gen n=age*D  //生成分组变 age 和处置变量 D 的交乘项
reg I D age n,robust  //结果变量回归,(11)式
twoway(scatter I age)  ///
    (qfit I age if T==0,sort lcolor(red)lwidth(thick))
    (qfit I age if T==1,sort lcolor(blue)lwidth(thick))
xtitle("age")  ///
    //以 age 为横坐标,I 为纵坐标,画出 I 关于 age 的平面曲线图
legend(off)xlabel(56(1)63)  //设置横坐标刻度
restore
*————————————命令结束————————————
```

3.2.3　FRDD 结果分析

在 stata 中运行上述命令，得到图 3。从图 3 结果显示可以给看出，该公司男性员工在达到退休年龄之后，投资风险偏好会有一个明显的向上的跳跃，而在样本中的所有个体，除了是否达到退休年龄这个处置状态的差异外，临界点两侧的观测值是无差异的，故该男性员工投资风险偏好的跳跃是由退休制度导致的，所以该处置变量和结果变量之间存在因果关系，断点处的跳跃反映了这一因果效应。

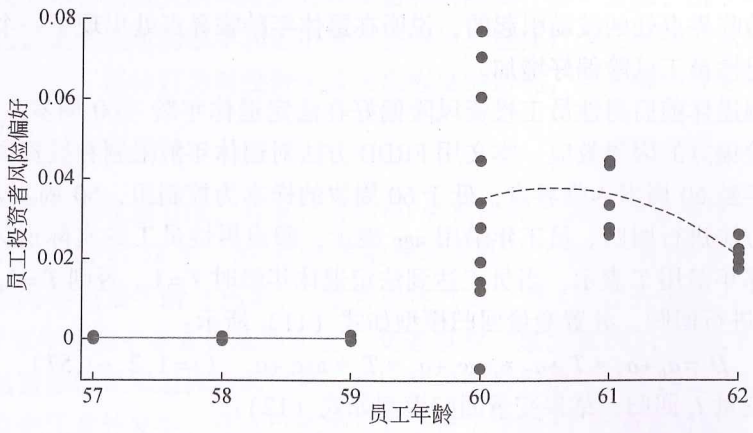

图 3　国家退休年龄制度对男性员工投资风险偏好的影响

Fig. 3　Retirement age system and the investment risk preference of male employees

图 4 为退休前（左图）后（右图）该公司男性员工风险投资偏好和年龄的关系。结果显示，退休前（左图）男性员工的风险投资偏好比较稳定，退休后（右图）男性员工风险投资偏好不稳定，并且在退休

后风险投资偏好会随着年龄的增长有所降低。结果表明，由于退休后工作压力变小，员工会有更多的时间花费在投资问题的考虑上，会使得男性员工的投资风险偏好增加。但是随着年龄的增长，员工会将时间和经济的中心向养身问题偏移，故男性员工的投资风险偏好会随着年龄增长而降低。

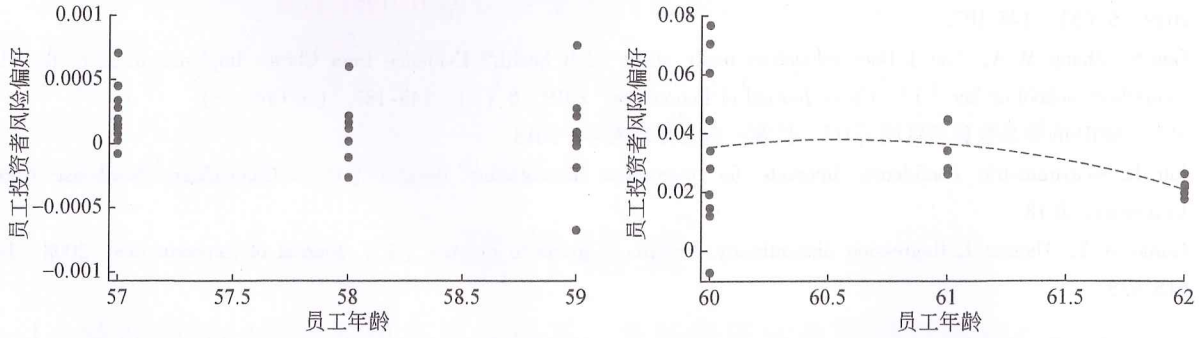

图4 男性员工退休前投资风险偏好（图左）及男性员工退休后风险投资偏好（图右）

Fig. 4 Investment risk preference of male employees before (the left) and after (the right) retirement

4 结语

近年来RDD方法越来越受学者的青睐，被广泛应用于各个学术领域的实证研究中。RDD的使用具有一定的规范，必须在满足条件的情况下才能准确进行估计，故在利用RDD方法进行研究之前，首先要对实验对象的条件进行检验，只有当实验对象满足RDD的使用条件时，RDD才能给出准确的因果关系估计。本文对RDD的方法原理、具体操作步骤和估计方法进行了阐述，并给出了RDD方法的应用案例，为RDD的准确应用提供了参考。

参考文献：

[1] Thistlthwaite D L, Campbell D T. Regression-discontinuity analysis: an alternative to the ex-post factor experiment [J]. Journal of Educational Psychology, 1960, 51 (6): 309-317.

[2] Lee D S, Card D. Regression discontinuity inference with specification error [J]. Journal of Econometrics, 2008, 142: 655-674.

[3] Lee D S, Lemieux T. Regression discontinuity designs in economics [J]. Journal of Economic Literature, 2010, 48 (2): 281-355.

[4] Johnes G, Tsionas M G. A regression discontinuity stochastic frontier model with an application to educational attainment [J]. Stat, 2019, 8 (1): e242: 1-7.

[5] Trochim W M K. Research design for program evaluation: the regression discontinuity approach [J]. Journal of the American Statistical Association, 1985, 8: 377-378.

[6] Hahn J, Todd P, Van der Klaauw W. Identification and estimation of treatment effects with a regression-discontinuity design [J]. Econometrica, 2001, 69 (1): 201-209.

[7] McCrary J. Manipulation of the running variable in the regression discontinuity design: a density test [J]. Journal of Econometrics, 2008, 142 (2): 698-714.

[8] Thistlethwaite D L, Campbell D T. Regression-discontinuity analysis: an alternative to the ex-post factor experiment [J]. Journal of Educational Psychology, 1960, 51 (6): 309-317.

[9] Dell M. The persistent effects of Peru'S miningmita [J]. Econometrica, 2010, (6): 1863-1903.

[10] Chen Y, Ebenstein A, Greenstone M, et al. Evidenc on the impact of sustained exposure to air pollution on life expectancy from China' S Hnai River policy [J]. Proceedings of the National Academy of Sciences, 2013 (32): 12936-12941.

[11] Angrist J D, Lavy V. Using Maimonides' rule to estimate the effect of class size on scholastic achievement [J]. Quarterly Journal of Economics, 1999, 114 (2): 533-575.

[12] 杨杨，于芝麦，杜剑. 高质量的内部控制能否降低企业有效税率？——基于模糊RDD的数据实证检验 [J]. 财经论

丛（浙江财经学院学报），2019（8）.
Yang Y，Yu Z M，Du J. Can high-quality internal control reduce the effective tax rate of enterprises? —An empirical test based on fuzzy discontinuity regression design [J]. Collected Essays on Finance and Economics，2019（8）.（in Chinese）
[13] 郭四维，张明昂，曹静. 教育真的可以影响健康吗？——来自中国 1986 年义务教育法实施的证据 [J]. 经济学报，2019，6（3）：148-187.
Guo S，Zhang M A，Cao J. Does education really affect adult health? Evidence from Chinas implementation of the 1986 compulsory schooling law [J]. China Journal of Economics，2019，6（3）：148-187.（in Chinese）
[14] 刘凯. RDD 的非参数置信区间 [D]. 长春：东北师范大学，2018.
Liu K. Nonparametric confidence intervals for regression discontinuity designs [D]. Changchun：Northeast Normal University，2018.
[15] Guido W I，Thomas L. Regression discontinuity designs：a guide to practice [J]. Journal of Econometrics，2008，142：615-635.

Theory Analysis，Operating Steps and Applications of Regression Discontinuity Design

Liu Weiqi[1,2]，*Liu Huan*[3]

1. Institute of Management and Decision，Shanxi University，Taiyuan 030006，China；
2. School of Finance Shanxi University of Finance and Economics，Taiyuan 030006，China；
3. School of Economic and Management，Shanxi University，Taiyuan 030006，China

Abstract：There is a growing application of regression discontinuity design（RDD）to empirical research in recent years. RDD can effectively use the actual constraint conditions to analyze causal relationship between variables. RDD is closer to randomized experimental design than instrumental variable method and difference to difference method. The paper described the theory of RDD，summarized the estimation methods and operating steps. Due to the different treatment probabilities，RDD can be divided into sharp regression discontinuity design（SRDD）and fuzzy regression discontinuity design（FRDD）. This paper demonstrated the application of SRDD and FRDD in empirical studies by two cases，and showed the stata command for SRDD and FRDD.

Key words：Regression Discontinuity Design；Sharp Regression Discontinuity Design；Fuzzyregression Discontinuity Design