高级计量暑期班
(Seminar of Advanced Econometrics)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-06-28

RDD PART 02：断点效应评估

1.RDD原理是什么？(How Does It Work?)

2.RDD该如何实施？(How Is It Performed?)

3.RDD怎样高级进阶？(How the Pros Do It?)

1.RDD原理是什么？

(How Does It Work?)

1.1 RDD直观解释

1.2 RDD相关概念

1.3 RDD因果关系分析

1.4 RDD基本假设

1.5 RDD的基本过程

（引子）RDD典型分析：地理区隔与收入变化

示例：

- 圣地亚哥(San Diego)是美国南部一个大城市，占地面积超过300平方英里。它也很富裕，截至2019年，家庭平均年收入超过85000美元，比全国平均水平高出约50%。
- 当您向南进入城市的其他区域时，一些南部地区的收入会少一些。例如用，当你往南到达的圣伊西德罗(San Ysidro)地区时（靠近墨西哥边境），家庭收入已经下降到50000-55000美元左右。你越往南走，期望家庭收入就越低。
- 但是，当我们越过边境进入墨西哥的蒂华纳(Tijuana, Mexico)时会发生什么？一旦越过边境进入墨西哥的蒂华纳(Tijuana)。你会发现家庭收入，突然和急剧地下降到20000美元左右。

（引子）RDD典型分析：地理区隔与收入变化

思考：

- 我们从圣地亚哥(San Diego)市中心开车到南部区域圣伊西德罗(San Ysidro)，只有16英里距离，收入下降了25%。但是，只要继续往南步行几英尺越过边境进入墨西哥境内的蒂华纳(Tijuana, Mexico)，家庭收入则发生急剧下降。
- 当然，对于圣地亚哥南部的家庭，地理位置可能有所不同，这可以解释收入的一些差异。但是在边界线附近两端，家庭收入会出现显著跳跃，这是地理位置因素所难以解释的。

断点回归设计（Regression Discontinued Design, RDD）：

> RDD是一种用于检验**因果关系**（causal relationship）假设的分析方法 (Thistlethwaite and Campbell, 1960)

RDD主要用于如下情形(Cattaneo and Titiunik, 2021):

- 被研究对象（units）上可以观测到一个**运行变量**（running variable）
- 基于某些规则（rule）研究者可以给出运行变量上的一个（或若干个）**断点值**（cutoff），并据此对所有被研究对象设定**分配水平**（assignment level）：包括**处置条件**（treatment condition）和**控制条件**（control condition）。
- 在断点值以上的被研究对象将被分配**处置条件**（treatment condition），并被定义为**处置组**（treated group）；在断点值以下的被研究对象将被分配**控制条件**（control condition），并被定义为**控制组**（controlled group）
- 在满足某些**假设条件**下，断点附近处置条件分配概率的断点式变化，可以揭示出**处置条件**对**结果变量**（目标变量）的因果关系。

1.2 RDD相关概念：结果变量、运行变量、混淆变量

- **结果变量（output variable）**：研究的目标变量，一般记为 Y

> 例如，结果变量为观测到的病人是否猝死。

- **运行变量(Runing variable)**[a]：是一个可以观测得到的变量。一方面它将决定被研究对象（units）是否被处置（treated）；另一方面它本身也会影响到结果变量。一般记为 X

> 例如，医生测量病人的血压，如果收缩压高于135，医生会给病人开降压药，这里病人的血压就是运行变量。

- **混淆变量(Confound variable)**：是哪些不能被直接观测得到的变量，它们可能会同时影响到**运行变量**（进而干扰到马上要定义的**处置变量**）以及**结果变量**。一般记为 U

[a] 也被称为分派变量（assigning variable），或者强制变量（forcing variable）

- **断点（Cutoff）**：是运行变量中的一个具体取值，根据它的取值我们可以来决定对象是否需要处置。这一取值一般记为 $X = c_0$

> 以血压为例，假定断点值设置为收缩压135。如果你的血压高于135，就应该吃药。如果低于135，就无须吃药。

- **处置变量（Treatment variable）**：根据运行变量和断点值的关系，定义得到的关于是否要分配处置水平的虚拟变量。一般定义为：

$$D = \begin{cases} 0 & \text{if} \quad X < c_0 \\ 1 & \text{if} \quad X \geq c_0 \end{cases}$$

> 例如，给定运行变量 X为病人血压，断点值为 $c_0 = 135$，那么处置变量即为**是否用药**。具体地，所有血压值 $X \geq c_0$的病人都会进行**用药处置**，也即虚拟变量赋值 $D = 1 \quad (\text{if} \quad X \geq c_0)$；否则就**不用药**，虚拟变量赋值为0。

- **谱宽**(Band width)：是断点值附近的一个**邻域**的区间范围的长度，一般记为 h，此时这个领域的区间范围定义为 $b \equiv [c_0 - h, c_0 + h]$, and $h > 0$

示例：

- 研究者可以任意给定运行变量（血压）的一个**谱宽**为 $h = 10$，则**断点值**附近的一个**邻域**的区间范围为 $b \equiv [c_0 - h, c_0 + h] = [135 - 10, 135 + 10] = [125, 145]$

1.3 RDD因果关系分析：随机控制实验

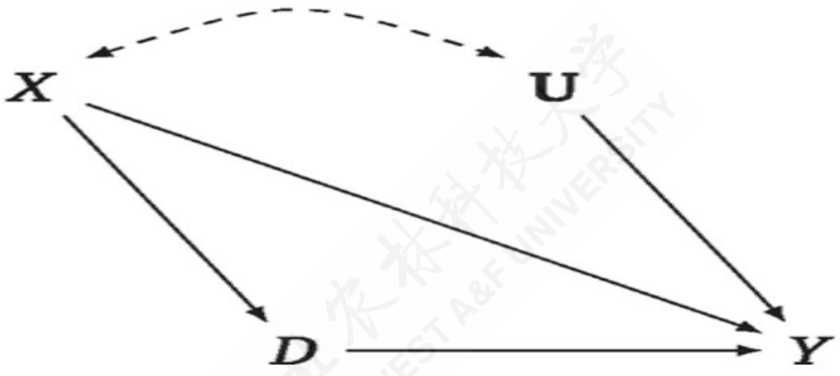
- **随机控制实验**（Randomized controlled experiments）：也称为随机对照实验，可以通过严格控制其他影响因素的变动，而准确分析特定一个影响因素对结果变量 Y 的作用。绝大部分自然科学研究都基于这一实验设计理念。
- **准自然实验**（Quasi-experiment or Natural experiment）：对于社会科学家而言，严格的随机控制实验往往无法获得或极难实施。但是在特定条件下，也还是可以得到某种“近似”（as if）随机性的数据生成机制（DGP）。
- **局部随机性实验**（Local randomized experiment）：在某些情形下，全局性（global）的随机对照实验难以满足或事实，但是却可以在局部范围内（local）进行近似随机的对照实验(Hausman and Rapson, 2018)。

（示例）局部随机控制实验

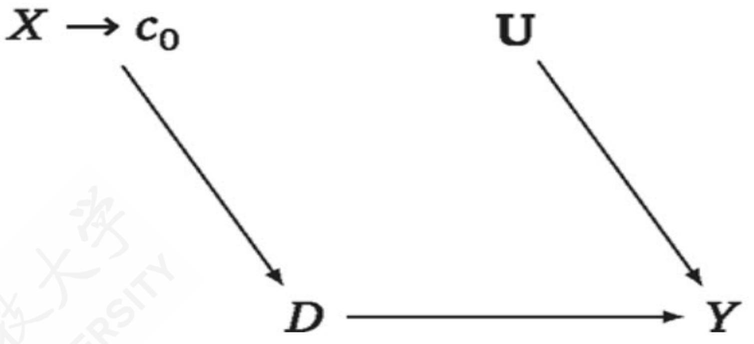
分数与录取案例：

- 高校根据高考成绩划定投档线和录取线，如果某省理工类一本录取最低控制分数线为450，该省内的一所重点高校N理工类最低录取分数线为520分。
- 那么该重点高校N最低录取线（520分）附近以下，例如516-519分之间未被录取的很多学生，与略高于最低录取线，例如520-522分之间被成功录取的很多学生，这两类学生群体理论上并无明显差异。
- 那么我们就可以基于这一局部观察，设计局部随机控制性实验分析。

1.3 RDD因果关系分析：断点与局部随机



a)数据生成机制DGP



b)RDD因果关系解析

- 图a)展示的是常见的数据生成机制（DGP）。因为**混淆变量** *U*的存在，使得难以有效分析出**处置变量** *D*对**结果变量** *Y*的作用关系（影响效应）。
- 图b)展示的是在RDD框架下，研究者能够很大程度上剥离**混淆变量** *U*的干扰，并有效分析出**处置变量** *D*对**结果变量** *Y*的作用关系（影响效应）。

1.3 RDD因果关系分析：可观测事实与反事实

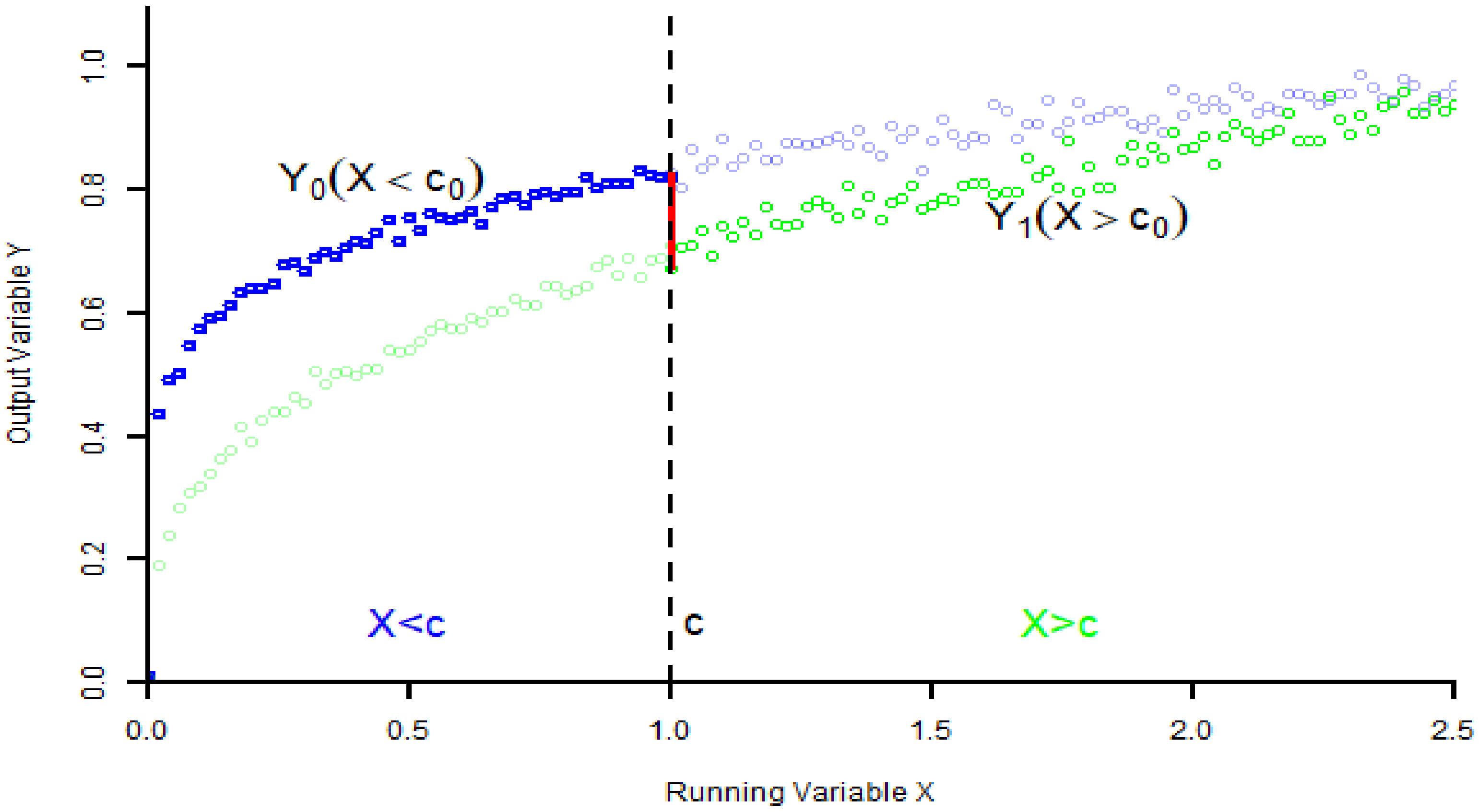
可观测事实（observed facts）：在给定研究对象某种分配条件下（例如处置条件或控制条件），可以分别得到处置组对象（treated group，**T**）和控制组对象（controlled group, **C**），就能分别观测到结果变量的表现，也即**可观测事实**。

> **可观测结果**（observed outcome）：此时，处置组和控制组的结果变量容易被观测得到，分别可记为 $[Y_i^1 \mid D=1]$以及 $[Y_i^0 \mid D=0]$

反事实（Counterfactual）：对于处置组的研究对象，如果不给它们分配处置条件，那么它们的结果变量会是如何呢？同理，对于控制组的研究对象，如果给它们分配处置条件，那么它们的结果变量又会是如何呢？显然，这些都是假想情形，实际并未发生的事实。

> **潜在结果**（Potential outcome）：此时，处置组和控制组的结果变量不能被直接观测得到，表现为**潜在结果**，我们分别可记为 $[Y_i^0 \mid D=1]$以及 $[Y_i^1 \mid D=0]$

（示例）图形演示：可观测事实与反事实



- 处置条件下结果变量（可观测的和潜在的）的期望：

$$
\begin{aligned}
&\equiv \mathbb{E}\left(Y_i^1 \mid X_i \geq c_0\right)+\mathbb{E}(Y_i^0 \mid X_i \geq c_0) \\
&\equiv \mathbb{E}\left(Y_i^1 \mid D=1\right)+\mathbb{E}(Y_i^0 \mid D=1) \\
&\equiv \mathbb{E}\left(Y^1 \mid c^+\right)+\mathbb{E}(Y^0 \mid c^+)
\end{aligned}
$$

- 控制条件下结果变量（可观测的和潜在的）的期望：

$$
\begin{aligned}
&\equiv \mathbb{E}\left(Y_i^1 \mid X_i < c_0\right)+\mathbb{E}(Y_i^0 \mid X_i < c_0) \\
&\equiv \mathbb{E}\left(Y_i^1 \mid D=0\right)+\mathbb{E}(Y_i^0 \mid D=0) \\
&\equiv \mathbb{E}\left(Y^1 \mid c^-\right)+\mathbb{E}(Y^0 \mid c^-)
\end{aligned}
$$

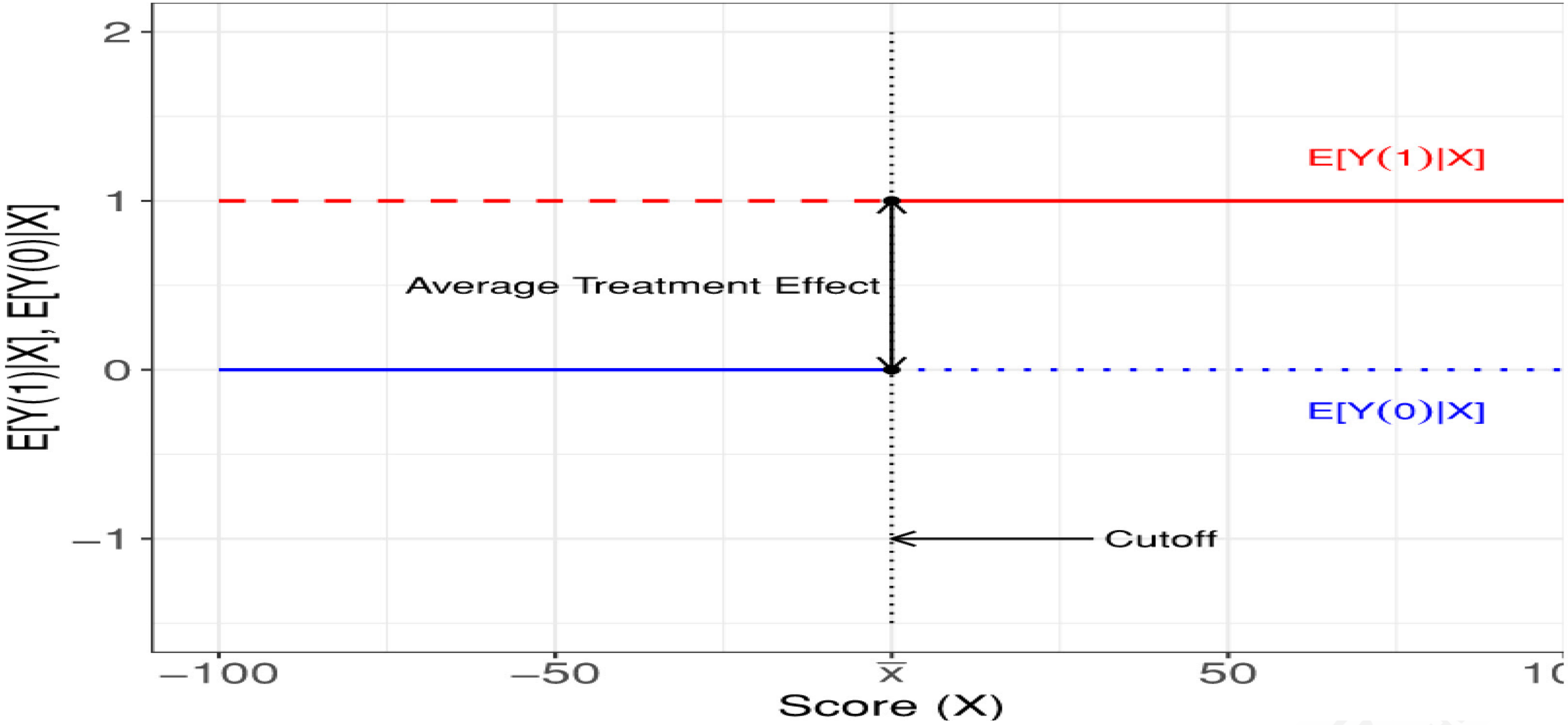
- 处置变量对结果变量的因果效应：

$$
\tau=\left[\mathbb{E}\left(Y^1 \mid c^+\right)+\mathbb{E}(Y^0 \mid c^+)\right]-\left[\mathbb{E}\left(Y^1 \mid c^-\right)+\mathbb{E}(Y^0 \mid c^-)\right]
$$

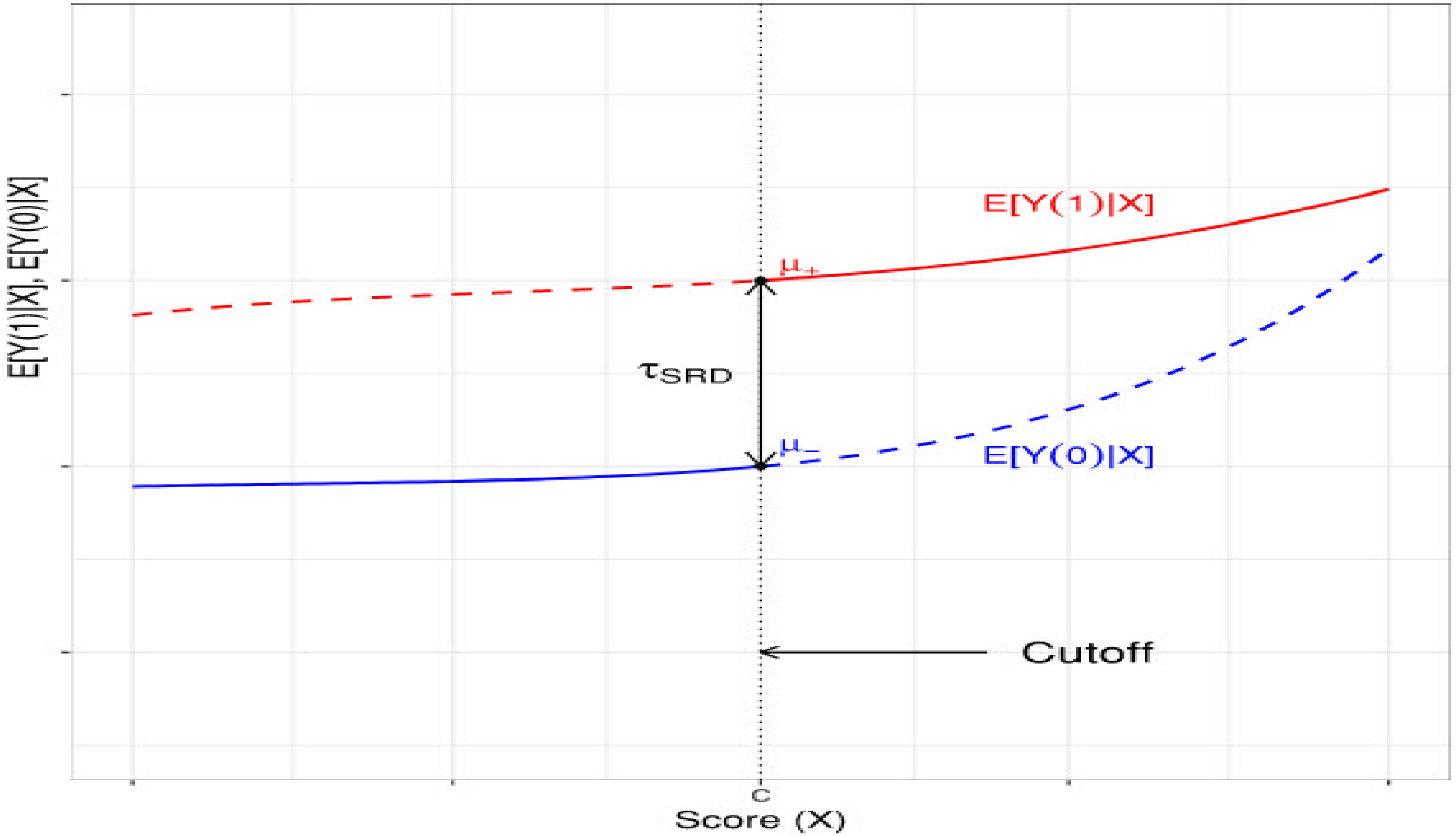
- 处置变量对结果变量的因果效应：

$$\begin{aligned}\tau &= \mathbb{E}\left(Y_i \mid X_i \geq c\right) - \mathbb{E}\left(Y_i \mid X_i < c\right) \\ &= \mathbb{E}\left(Y_i^1 \mid X_i \geq c\right) - \mathbb{E}\left(Y_i^0 \mid X_i < c\right) \\ &= \mathbb{E}\left(Y_i^1\right) - \mathbb{E}\left(Y_i^0\right)\end{aligned}$$

- 潜在结果变量的条件均值为常数的情形：

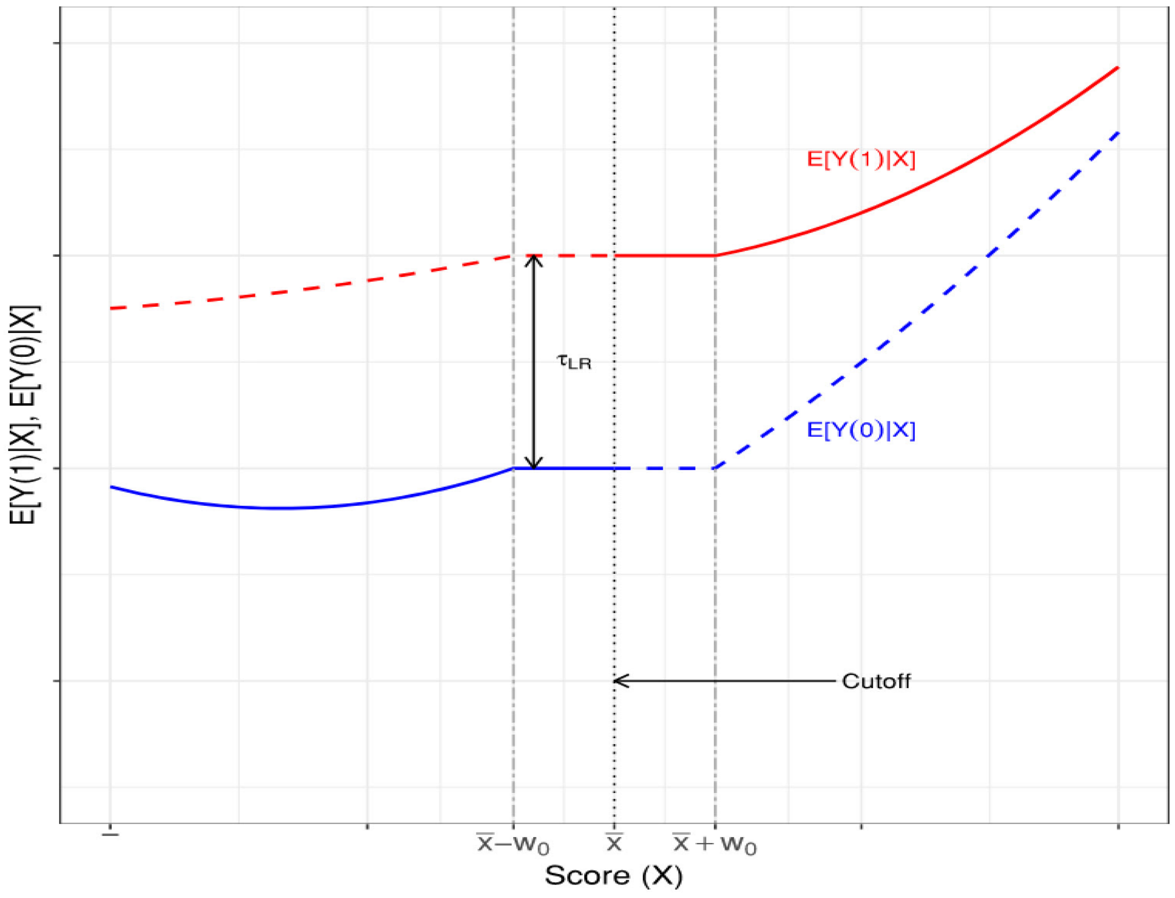


- 潜在结果变量的条件均值并**不是**常数的情形：



1.3 RDD因果关系分析：局部断点处置效应

- 断点处置效应具有局部性特征（the local nature of RD effect）。



1.4 RDD基本假设：连续性假设

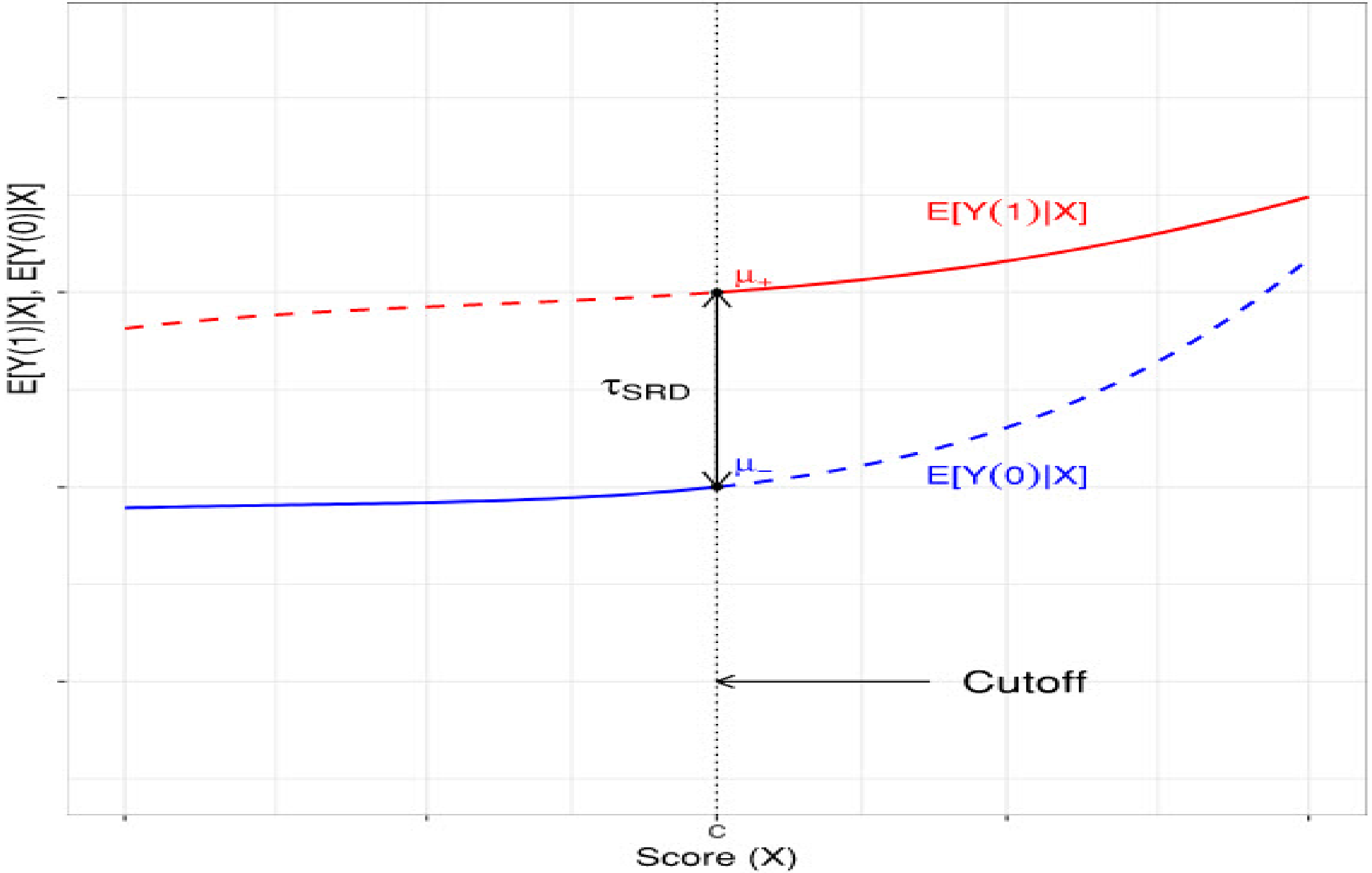
假设1：结果变量的期望值在**断点**处需要满足**连续性假设**（continuity assumption）：

- **结果变量**的期望值在**断点**处连续，也即 $E[Y_i(1)|X_i = x]$和 $E[Y_i(0)|X_i = x]$，可是作为 x的函数（$f(x)$），且在 $x = c_0$出连续。（见下图）
- **断点**值 c_0本身需要满足**外生性**（exogeneity）条件。也即，**断点**值 c_0在触发**处置变量**D的时候，不会有**其他变量**在同时期来干预这种“触发行为”。
- 在上述条件下，运行变量 X对结果变量 Y将**不再**具有直接影响（$X \rightarrow Y$），而是通过处置变量 D发生间接作用（$X \rightarrow D \rightarrow Y$）。
- **连续性假设**（continuity assumption）应该是RDD**最关键**的一个假设条件，而且这符合**经验事实**。

> **大自然不会跳跃！**[a] ---达尔文《物种起源》

[a] 事物的发展变化总是**渐进式**的，而不会陡然改变。常言道“量变引发质变”。

(示例)条件期望函数CEF的连续性假设



1.4 RDD基本假设：断点性假设

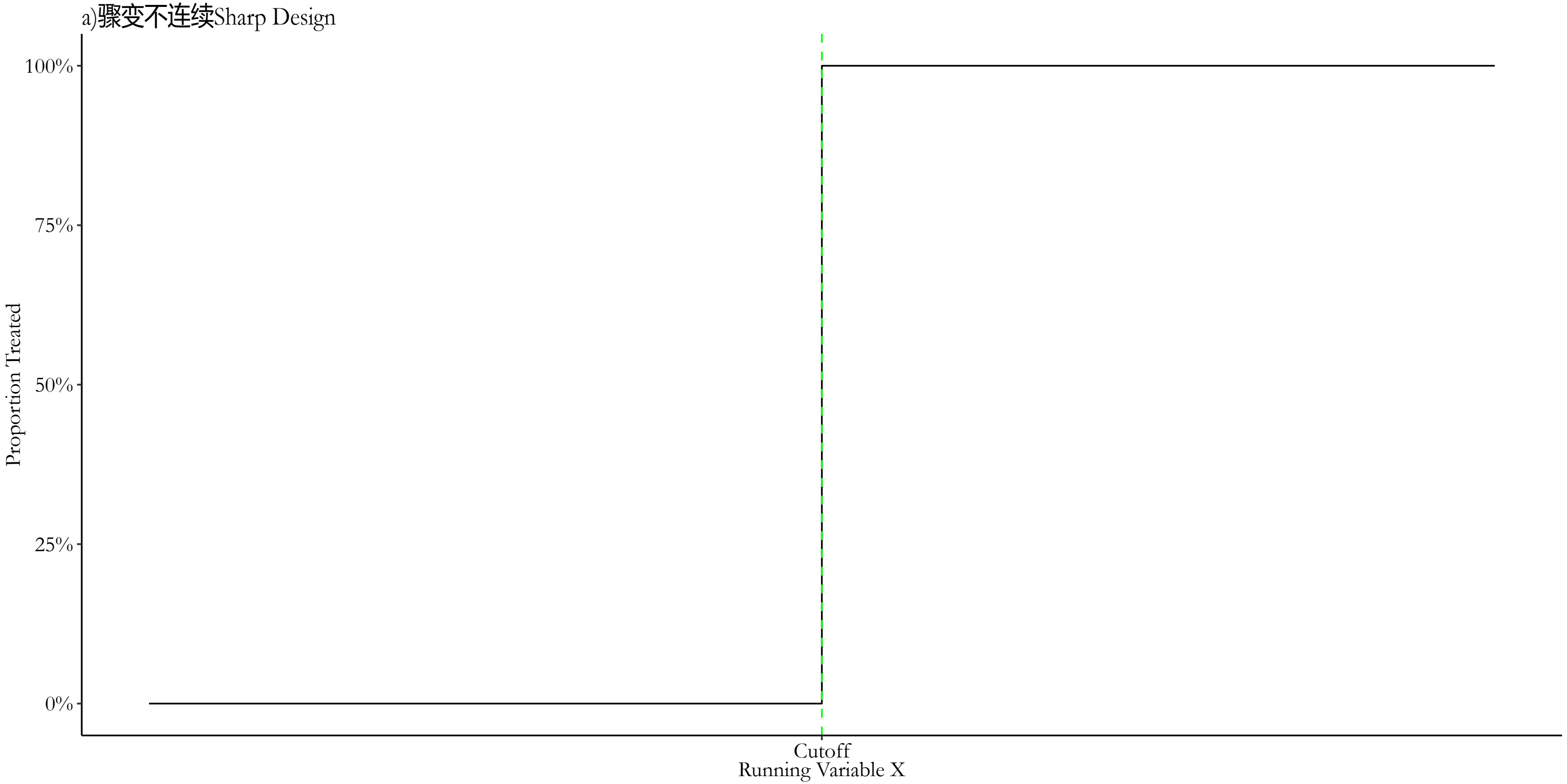
假设2：被研究对象被分配（assign）处置条件（treated condition）[1]的**条件概率**（Conditional Probability of Receiving Treatment）$P(D_i = 1 \mid X_i = c_0)$在断点处是**不连续的**（也即间断的）。

常见的处置分配概率不连续**模式**包括：

- **骤变不连续**（Sharp discontinuity）：处置条件分配的概率在断点处被完全决定。
- **模糊不连续**（Fuzzy discontinuity）：处置条件分配的概率在断点处**不能**被完全决定。

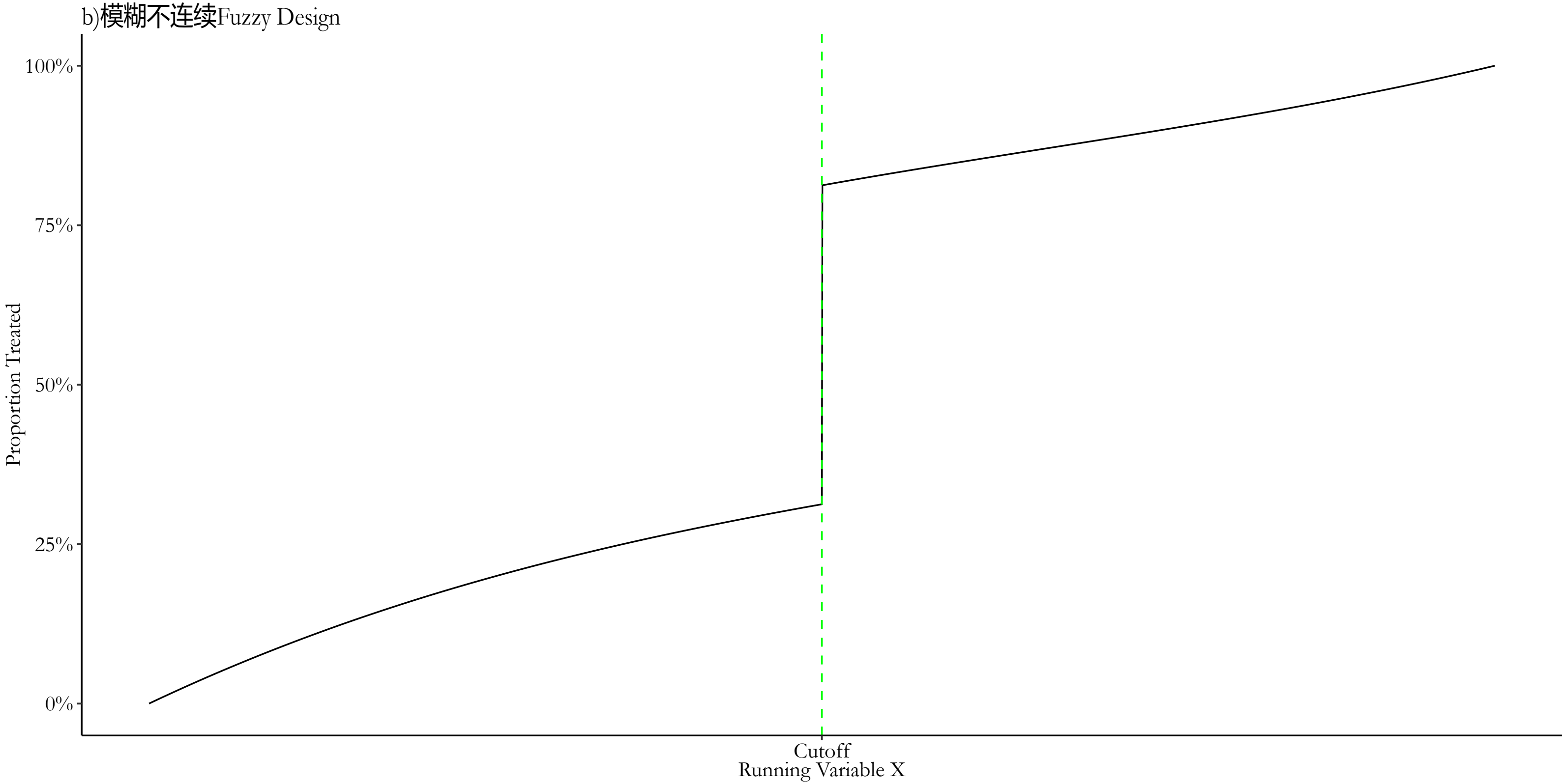
[1] 回顾**分配水平**（assign level）具有两个水平：处置条件（treated condition）和控制条件（controlled condition）

（示例）断点性假设：骤变（Sharp）不连续



- 处置条件的骤变（Sharp）不连续示例：小学入学年龄严格要求出生日期（X）在 $c_0 = 9\text{月}1\text{日}$之前。

（示例）断点性假设：模糊（Fuzzy）不连续



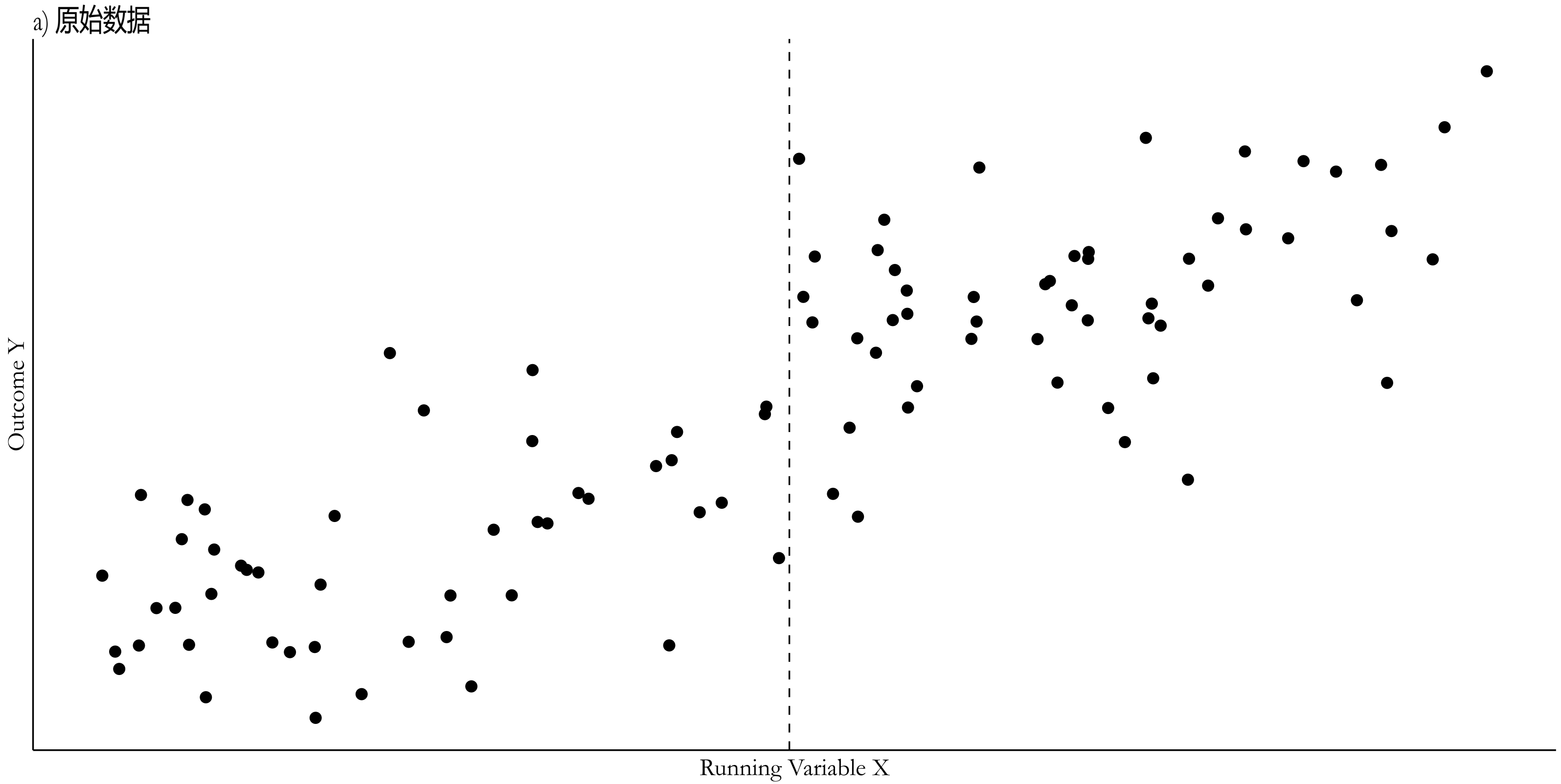
- 处置条件的模糊（Fuzzy）不连续：小学入学年龄要求出生日期（X）在 $X \in [8月1日, 9月30日]$期间，家长可以自己选择孩子是否上小学。

1.5 RDD的基本过程：概览

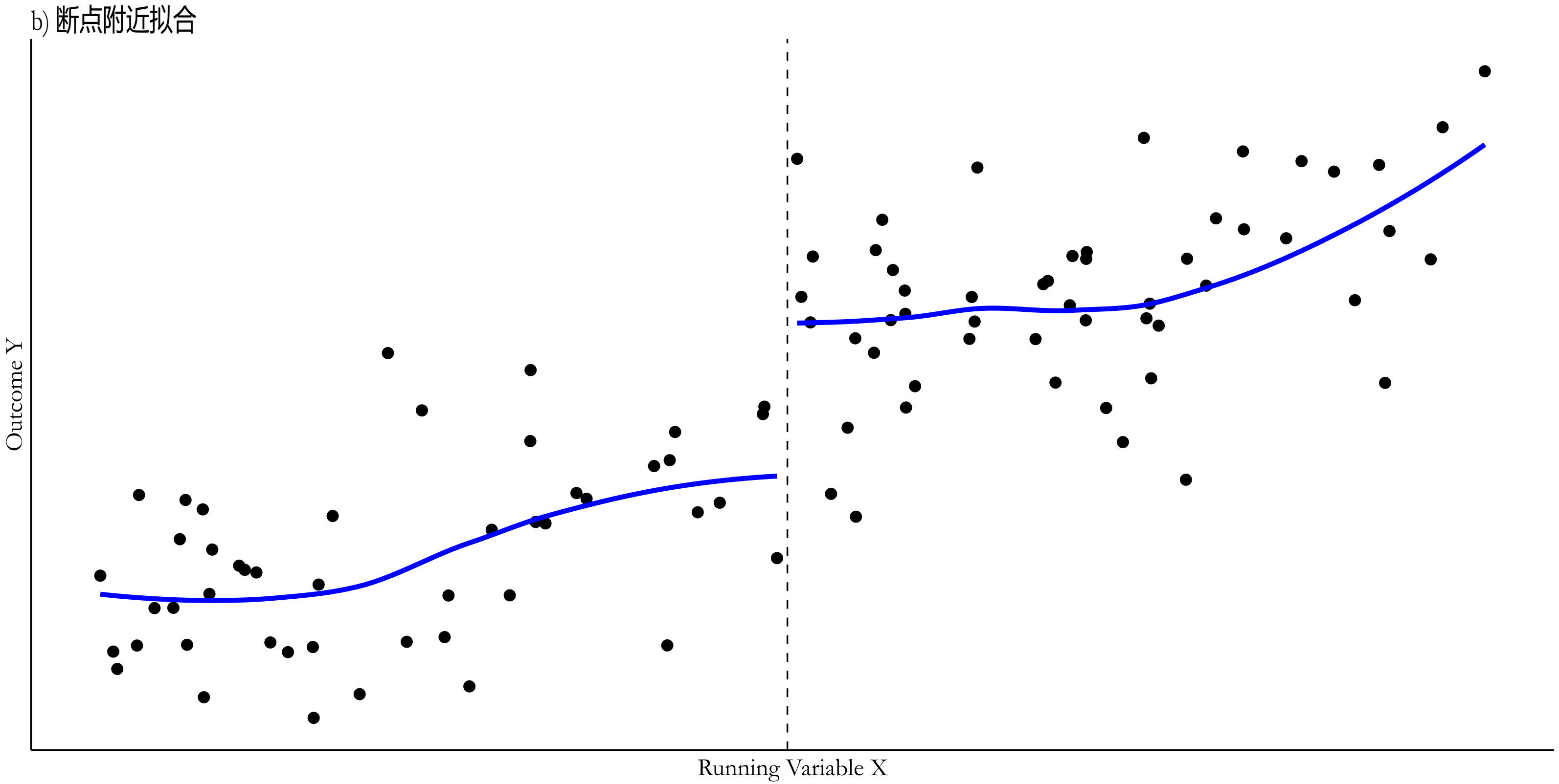
如果暂时忽略各种细节，一个最简化的RDD分析过程包括：

- 设定断点两边对结果变量的预测模型方法（predictive model）
- 选择局部谱宽（bandwidth）
- 估计并计算因果效应

（示例）RDD的基本过程1/4：原始数据

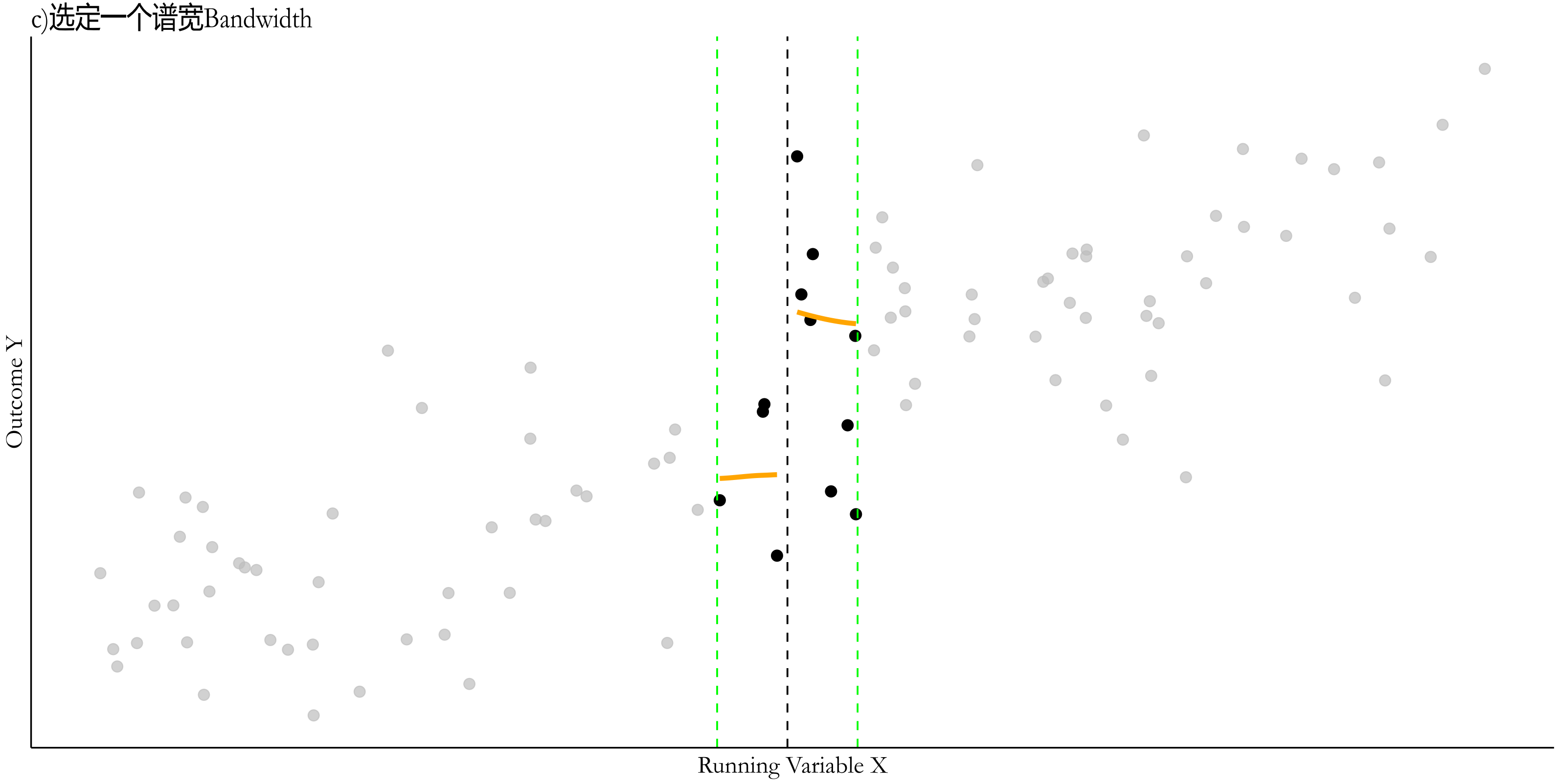


（示例）RDD的基本过程2/4：断点两边拟合



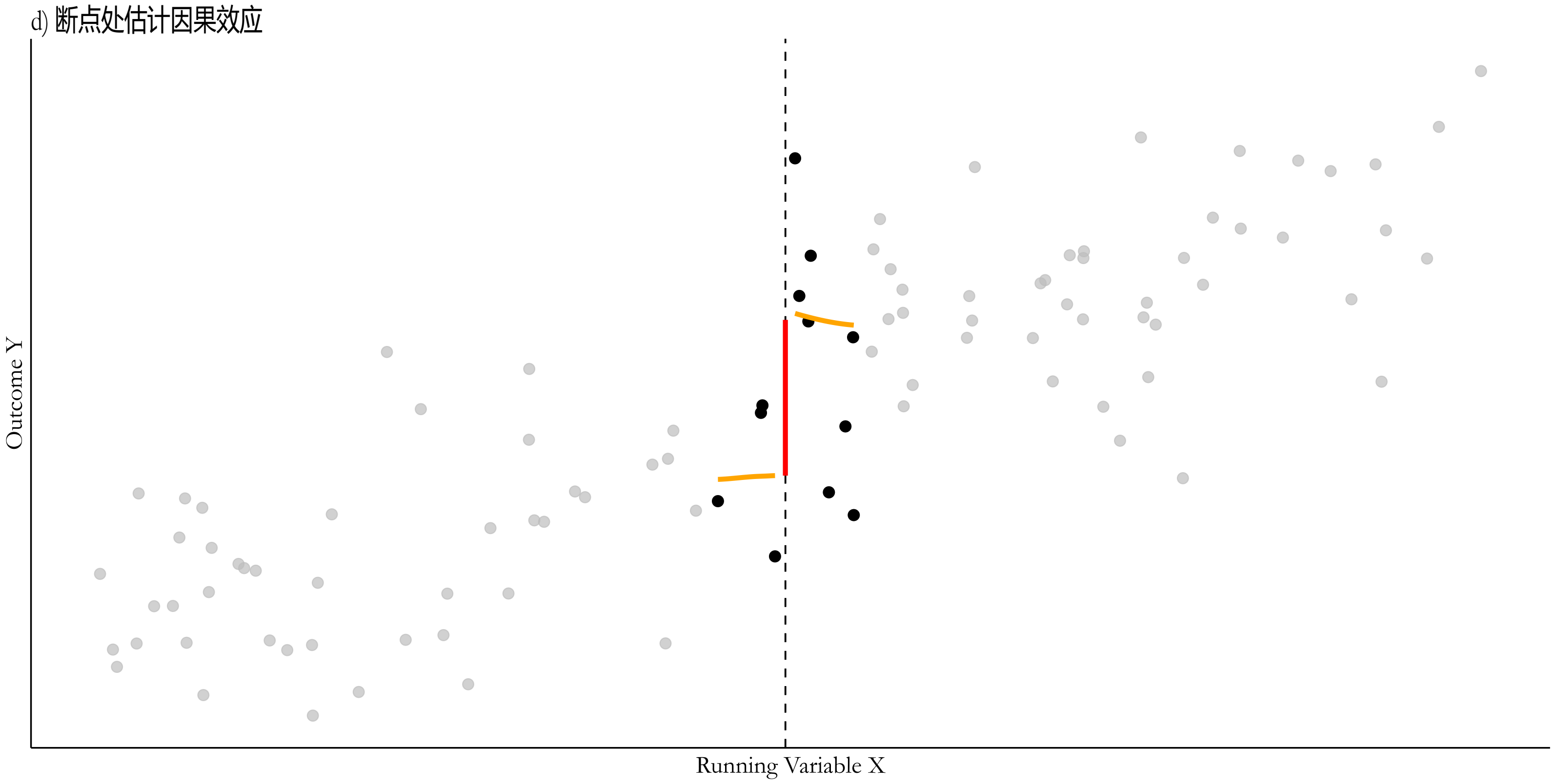
- 这里采用了**LL**方法拟合局部均值（local mean）

（示例）RDD的基本过程3/4：选定一个谱宽Bandwidth



- 我们暂时不关心远离断点处的观测值（因为混淆变量会产生作用）
- 最优化的谱宽选择可以基于某些准则，例如BIC等

（示例）RDD的基本过程4/4：断点处估计因果效应



- 谱宽范围内、断点两边的估计结果，表现出了“跳跃”效果（jumps）

2.RDD该如何实施？

(How Is It Performed?)

2.1 平均和断点处置效应

2.2 骤变RDD的估计

2.3 骤变RDD谱宽选择

2.4 骤变RDD推断

2.5 RDD协变量分析

- 结果变量 Y
- 运行变量 X，断点值 c_0
- 处置变量 D:

$$D = \begin{cases} 0 & \text{if} \quad X < c_0 \\ 1 & \text{if} \quad X \geq c_0 \end{cases}$$

- 实验组对象**T** $(D = 1)$；控制组对象**C** $(D = 0)$

2.1 平均和断点处置效应：定义

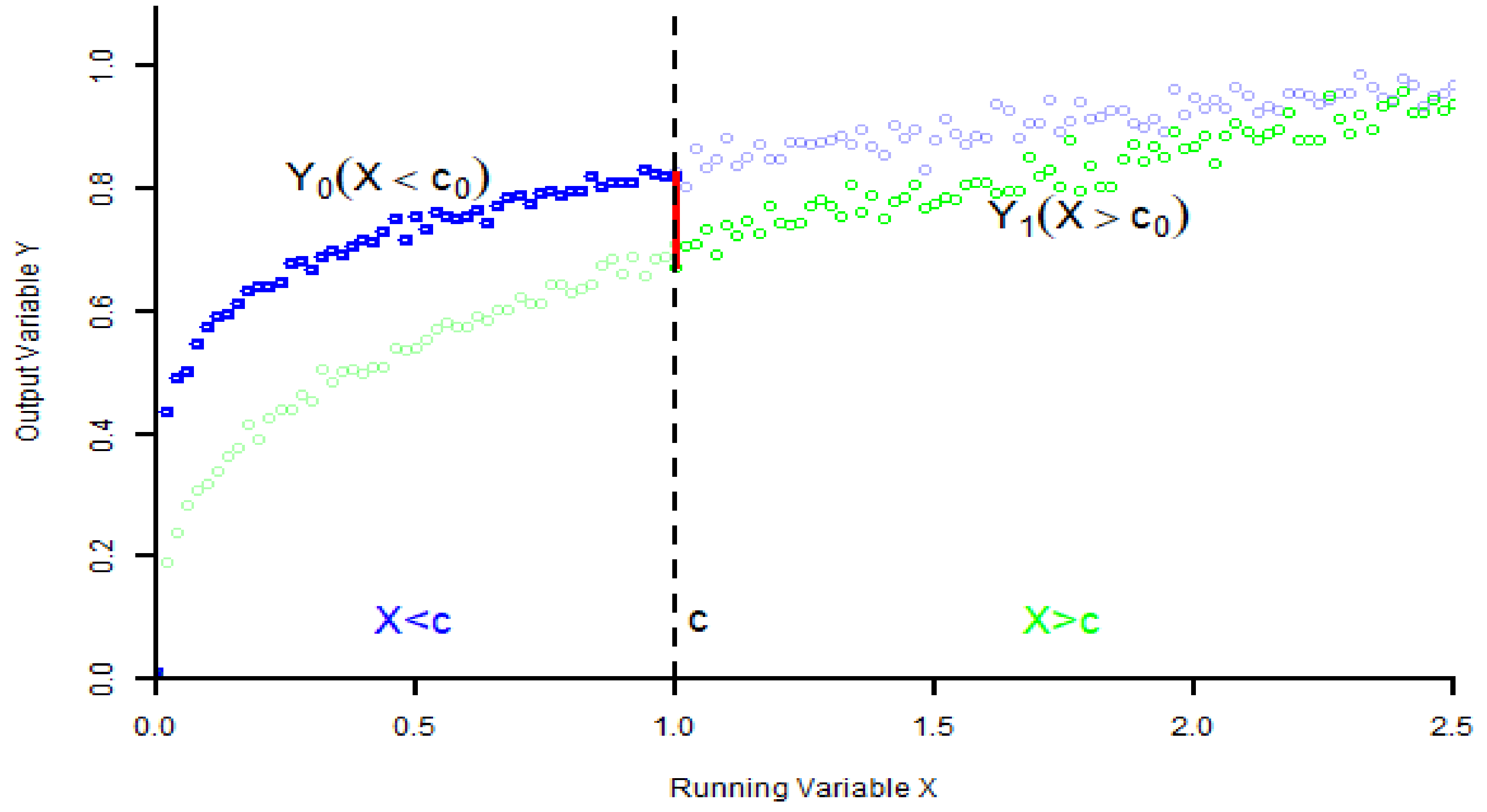
- 当个体 i被分配为“处置条件”时，其结果变量为 Y_1为；当个体 i被分配为“控制条件”时，其结果变量为 Y_0。
- 此时，个体 i的**处置效应**(treatment effect)记为 $\theta = Y_1 - Y_0$，因为其具有随机性，也被称为**随机处置效应**(random treatment effect)
- 给定一个可观测的协变量 X（运行变量），我们可以得到个体 i的**条件处置效应**(conditional treatment effect)，并记为：

$$\theta|(X = x) = (Y_1 - Y_0)|(X = x)$$

- 对于 $X = x$处的多个个体，我们可以得到它们的条件**平均处置效应**（average treatment effect, ATE），并记为：

$$\theta(x) \equiv \mathbb{E}(\theta \mid X = x)$$

（示例）个体和平均处置效应



给定结果变量的**条件期望函数**(conditional expect function, CEF)[a]如下：

$$m(x) \equiv \mathbb{E}(Y|X=x)$$

则可以分别得到**控制条件**和**处置条件**下的**条件期望函数**：

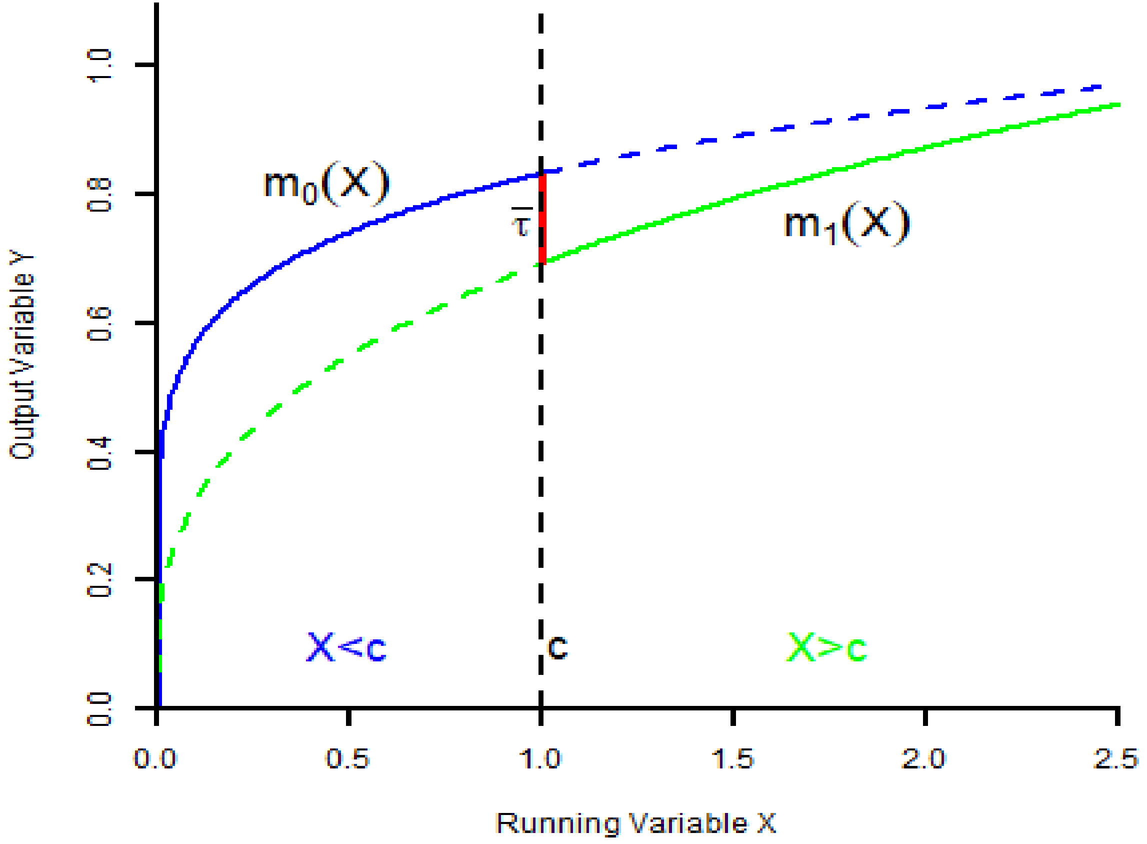
$$\begin{cases} m_0(x) = \mathbb{E}(Y_0|X=x) \\ m_1(x) = \mathbb{E}(Y_1|X=x) \end{cases}$$

进而，我们可以把**条件平均处置效应**(conditional ATE)表达为：

$$\begin{aligned} \theta(x) &\equiv \mathbb{E}(\theta \mid X=x) \\ &= \mathbb{E}[(Y_1 - Y_0) \mid X=x] \\ &= \mathbb{E}[(Y_1 \mid X=x) - (Y_0 \mid X=x)] \\ &= m_1(x) - m_0(x) \end{aligned}$$

[a] 这里先表达为**隐函数**形式，也即其具体函数表达式未知。

（示例）条件期望函数CEF与平均处置效应



结果变量的条件期望函数在断点处的**连续性**（continuity）假设：

> 给定断点值为 $x=c$，假设结果变量的条件期望函数 $m(x)$在断点处 $x=c$连续。
> 这也意味着在*控制条件***和***处置条件***下的***条件期望函数***也在断点处是连续的。也即 $m_0(x)$和 $m_1(x)$在断点处 $x=c$连续。

定义：我们把条件函数的 **极限**（ z从右边向 x值取极限，和 z从左边向 x值取极限）定义如下

$$
\begin{aligned}
m(x+) &= \lim_{z \downarrow x} m(z) \\
m(x-) &= \lim_{z \uparrow x} m(z)
\end{aligned}
$$

断点处置效应定理：给定处置分配规则为 $D = 1\{X \geq c\}$，而且假定结果变量满足断点处的连续性假设，也即结果变量的条件期望函数 $m(x)$在断点处 $x = c$连续，那么**断点处置效应**为：

$$\bar{\theta} = \lim_{z \downarrow c} m(z) - \lim_{z \uparrow c} m(z) = m(c+) - m(c-)$$

证明：首先，我们进一步定义结果变量：

$$Y \equiv Y_0 \cdot 1\{x < c\} + Y_1 \cdot 1\{x \geq c\}$$

两边对 $X = x$取期望，且根据结果变量的条件期望函数的定义，则有：

$$\begin{aligned}\mathbb{E}(Y|X = x) &= \mathbb{E}(Y_0|X = x) \cdot 1\{x < c\} + \mathbb{E}(Y_1|X = x) \cdot 1\{x \geq c\} \\ \Rightarrow m(x) &= m_0(x) \cdot 1\{x < c\} + m_1(x) \cdot 1\{x \geq c\}\end{aligned}$$

根据前面关于**条件处置效应**的定义及**连续性假设**，则有：

$$\begin{aligned}\theta(x) &\equiv \mathbb{E}(\theta \mid X = x) \\ &= \mathbb{E}[(Y_1 - Y_0) \mid X = x] \\ &= \mathbb{E}[(Y_1 \mid X = x) - (Y_0 \mid X = x)] \\ &= m_1(x) - m_0(x)\end{aligned}$$

$$\begin{aligned}\theta(c) &= m_1(c) - m_0(c) \\ &= \lim_{x \downarrow c} m(x) - \lim_{x \uparrow c} m(x) \qquad \leftarrow \text{(连续性假设)} \\ &= m(c+) - m(c-)\end{aligned}$$

断点回归设计（RDD）属于典型的**边界估计**（boundary estimation）问题，这里我们将优先采用**局部线性回归**（local linear regression, LLR）方法进行估计。

> 这里，我们将使用到非参数的**核函数**（kernel function）方法来除了回归的权重问题。

给定如下条件：

- 变量集

$$Z_i(x) = \begin{pmatrix} 1 \\ X_i - x \end{pmatrix}$$

- 核函数（kernel function） $K(u)$
- 谱宽（bandwidth） h

2.2 骤变RDD的估计：局部线性回归估计（CEF）

此时，可以证明**局部线性**方法下的系数估计为（证明略）：

- 对于断点左侧 $x<c$，系数估计为[a]：

$$\widehat{\beta}_0(x)=\left(\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) Z_i(x) Z_i(x)^{\prime} \cdot 1\left\{X_i<c\right\}\right)^{-1}\left(\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) Z_i(x) Y_i \cdot 1\left\{X_i<c\right\}\right)$$

- 对于断点左侧 $x \geq c$，系数估计为[b]：

$$\widehat{\beta}_1(x)=\left(\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) Z_i(x) Z_i(x)^{\prime} \cdot 1\left\{X_i \geq c\right\}\right)^{-1}\left(\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) Z_i(x) Y_i \cdot 1\left\{X_i \geq c\right\}\right)$$

[a] [b] 需要注意的是，这里我们得到的都是系数向量（vector）。

根据结果变量条件期望函数 $m(x)$的定义，我们可以使用上述系数估计 $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}(x), \widehat{\boldsymbol{\beta}}_{\mathbf{1}}(x)\}$，进一步得到结果变量条件期望函数的估计结果[a]：

$$\widehat{m}(x) = \left[\widehat{\boldsymbol{\beta}}_{\mathbf{0}}(x)\right]_1 \cdot 1\{x < c\} + \left[\widehat{\boldsymbol{\beta}}_{\mathbf{1}}(x)\right]_1 \cdot 1\{x \geq c\}$$

因此，根据**断点处置效应定理**，可以得到在断点 $x = c$处对总体平均处置效应 $\bar{\theta}$ 的样本估计结果 $\hat{\theta}$：

$$\hat{\theta} = \left[\widehat{\boldsymbol{\beta}}_{\mathbf{1}}(c)\right]_1 - \left[\widehat{\boldsymbol{\beta}}_{\mathbf{0}}(c)\right]_1 = \hat{m}(c+) - \widehat{m}(c-)$$

[a] 条件期望函数CEF只需要用到系数向量（vector）的第一个元素，因此用了下标$_1$表达。

- 容易证明**骤变RDD断点处置效应**也可以通过如下简单线性回归方法 等价地得到 $\hat{\theta}$的对应估计值：

$$Y = \beta_0 + \beta_1 X + \beta_3 (X - c) D + \theta D + e$$

需要注意的是：

- 上述等价模型，只是等价前面的基于**非正规化矩形核函数**（unnormalized Rectangular）谱宽下的**局部线性LL**断点处置估计效应值。
- 简单地，上述等价模型需要进行样本数据集的重新定义。具体地，运行变量的范围需要调整到 $X \in [c - h^*, c + h^*]$，其中 $h^* = \sqrt{3}h = \sqrt{3} \times 8$

2.3 骤变RDD谱宽选择：基本问题

- 基于**边界估计**的**局部线性回归**方法本质上需要进行**非参数估计**，这尤其体现在核函数的**谱宽**（bandwidth）估计。
- 目前还没有达成一致意见的最优谱宽选择方法。因此在进行LLR估计之前，研究者不得不多尝试多种**数据导向**（data based）的谱宽选择工具。
- 谱宽估计是一项具有挑战性的工作，有些具体估计方法会异常复杂。

当然，这里可以建议使用两种谱宽选择方法：

- **多项式**(polynomial, PN)谱宽选择法(Fan, Gijbels, Hu, et al., 1996)：这是一种经验方法。
- **交叉验证**（cross validation, CV）谱宽选择法

2.3 骤变RDD谱宽选择：多项式法

- 首先构造包含 q阶多项式和断点漂移项的模型：

$$m(x)=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_q x^q+\beta_{q+1} D$$

- 然后，通过估计得到的条件期望函数 $\widehat{m}(x)$计算二阶求导结果：

$$\widehat{m}''(x)=2\widehat{\beta}_2+6\widehat{\beta}_3 x+12\widehat{\beta}_4 x^2+\cdots+q(q-1)\widehat{\beta}_q x^{q-2}$$

- 再计算常量 $\overline{B}$，其中 $[\xi_1,\xi_2]$是运行变量 X内部的一个评价区间：

$$\widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\widehat{m}''\left(X_i\right)\right)^2 1\left\{\xi_1\leq X_i\leq\xi_2\right\}$$

- 最后，对于任意**正规化核**（normalized kernel），可以计算得到**谱宽**：

$$h_{\mathrm{FG}}=0.58\cdot\left(\frac{\widehat{\sigma}^2\left(\xi_2-\xi_1\right)}{\widehat{B}}\right)^{1/5} n^{-1/5}$$

根据核函数的不同，多项式法(polynomial)计算公式略有不同：

- 对于**非规化矩形核**（un-normalized rectangular kernel） $K(u)=1/2, \text{for}|u| \leq 1$ ：

$$h_{\text{pn}}=1 \cdot\left(\frac{\widehat{\sigma}^{2}\left(\xi_{2}-\xi_{1}\right)}{\widehat{B}}\right)^{1/5} n^{-1/5}$$

- 对于**非规化三角核**（un-normalized rectangular kernel） $K(u)=1-|u|, \text{for}|u| \leq 1$：

$$h_{\text{pn}}=1.42 \cdot\left(\frac{\widehat{\sigma}^{2}\left(\xi_{2}-\xi_{1}\right)}{\widehat{B}}\right)^{1/5} n^{-1/5}$$

2.3 骤变RDD谱宽选择：交叉验证法

- **交叉验证**（cross validation, CV）方法：主要形式是把训练集分成两部分，一部分用来训练模型，另一部分用来验证模型。
- **交叉验证**方法包括：**留出法**（holdout）、**留一法**（Leave-one-out, LOO）、K折法（K-fold）、**自助法**（Bootstrap）等。
- 这里介绍的**交叉验证**谱宽选择法主要采用**留一法**（Leave-one-out,LOO）。

留一法（Leave-one-out,LOO）选择谱宽的基本步骤：

- 初步选定一个临近断点的区间 $[\xi_1, \xi_2]$（去中心化后centered X的范围）
- 任意选择初始谱宽
- 通过**留一法**计算模型预测残差及其**残差平方和**
- 以**最小化**残差平方和为目标，分析谱宽的变化趋势[a]，并最终确定谱宽bandwidth。

[a] 可以绘制CV标准（如均方误差AMSE）与谱宽关系的图示法进行观察。

2.3 骤变RDD谱宽选择：方法评析

- 谱宽估计的噪点（noise）会进入到RDD估计进程中去，因此谱宽选择显得非常重要。
- 无论是**多项式法**还是**交叉验证法**，确定最终谱宽时，都考虑到了**全局性**准确度。

> 这意味着它们都用到了更多的样本数据，因此谱宽估计会比较稳定。

- 另一种**局部性**的谱宽选择方法，主要考察**断点**附近(near-by)的准确度。

> 因为局部性存在多种可能，所以这类方法得到的谱宽会更加不稳定。具体参看(Imbens and Kalyanaraman, 2012; Arai and Ichimura, 2018)。

- 通过**改变**谱宽值，来对RDD估计进行**稳健性检查**是很必要的。

> 更大的谱宽，一般会使得断点效应估计系数**方差减小**(reduce variance)，置信区间变窄，但同时也会**增加偏误**(increase bias)。

谱宽选择的**经验法则**：

- 实践操作中，我们往往需要同时结合**多项式法**和**交叉验证法**来确定一个谱宽 $\tilde{h}$。
- 在上述基础上，我们还需要适当调减谱宽值，例如 $h = 25\% \cdot \tilde{h}$，以减少估计偏误。

基于**局部线性回归**LLR估计结果，对断点处置效应参数 $\bar{\theta}$ 的推断陈述（inferential statement），都会受到其中**非参数估计**偏差的影响。

可以证明，**局部线性回归**（LLR）的估计量 $\hat{m}(x)$在**标准正则条件**（standard regularity conditions）下将服从**渐近正态分布**。

- 此时，RDD估计量 $\hat{\theta}$的渐近偏误（bias）和渐近方差分别为：

$$
\begin{aligned}
\text{bias}[\hat{\theta}] &= \frac{h^2\sigma^2_{K^*}}{2}(m''(c+) - m''(c-)) \\
\text{var}[\hat{\theta}] &= \frac{R^*_K}{nh}\left(\frac{\sigma^2(c+)}{f(c+)} + \frac{\sigma^2(c-)}{f(c-)}\right)
\end{aligned}
$$

上述理论方差，我们可以通过两个**边界回归**（断点两边）的系数估计量的**渐近方差**求和计算得到。我们首先给定如下条件：

- 变量集：

$$Z_i = \begin{pmatrix} 1 \\ X_i - c \end{pmatrix}$$

- **核函数**（kernel function） $K_i = k\left(\frac{X_i - c}{h}\right)$
- **谱宽**（bandwidth） h
- **留一法**[a]得到的模型预测**残差**（leave-one-out prediction error） $\tilde{e}_i$

[a] **留一法**(Leave One Out, LOO) 是一种 常见的交叉验证方法，其中每个观察集都被视为验证集test，其余的 $(n-1)$观测值被视为训练集training。此处原理类似，每次都去掉一个数据进行估计，然后根据估计结果进行预测，然后得到预测误差。

2.4 骤变RDD推断：样本方差

此时，我们可以得到**局部线性回归**LLR估计系数 $\hat{\theta}$ 的**方差协方差矩阵**分别为：

$$\widehat{\boldsymbol{V}}_0 = \left(\sum_{i=1}^{n} K_i Z_i Z_i' \cdot 1\left\{X_i < c\right\}\right)^{-1} \left(\sum_{i=1}^{n} K_i^2 Z_i Z_i' \tilde{e}_i^2 \cdot 1\left\{X_i < c\right\}\right) \left(\sum_{i=1}^{n} K_i Z_i Z_i' \cdot 1\left\{X_i < c\right\}\right)^{-1}$$

$$\widehat{\boldsymbol{V}}_1 = \left(\sum_{i=1}^{n} K_i Z_i Z_i' \cdot 1\left\{X_i \geq c\right\}\right)^{-1} \left(\sum_{i=1}^{n} K_i^2 Z_i Z_i' \tilde{\boldsymbol{e}}_i^2 \cdot 1\left\{X_i \geq c\right\}\right) \left(\sum_{i=1}^{n} K_i Z_i Z_i' \cdot 1\left\{X_i \geq c\right\}\right)^{-1}$$

进一步地，估计系数 $\hat{\theta}$ 的渐进方差为上述两个矩阵第一个对角元素之和：

$$\operatorname{Var}(\hat{\theta}) = \left[\widehat{\boldsymbol{V}}_0\right]_{11} + \left[\widehat{\boldsymbol{V}}_1\right]_{11}$$

最后，我们可以分别对断点两侧计算**逐点置信区间**（Pointwise Confidence Interval），并相应构建**置信带**。

$$\widehat{m}(x) \pm z_{1-\alpha/2}(n-1) \cdot \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

$$\widehat{m}(x) \pm 1.96\sqrt{\widehat{V}_{\widehat{m}(x)}}$$

（死亡率案例）背景说明1/2

援助项目与儿童死亡率：

- 案例基于(Ludwig and Miller, 2007)的研究，他们重点评估了美国联邦政府脱贫援助项目（Head Start）的**骤变RDD**政策效应。
- 该援助项目于1965年实施，为3-5岁贫困孩子及其家庭提供学前教育、健康和社会服务等方面的资金援助。对于该援助项目经费，联邦政府将决定通过公开竞标，分配给提交援助申请的中标县。
- 为了保障援助项目的针对性，联邦政府将重点考虑资助被认定的300个贫困县。其中**贫困县**是基于1960年美国统计测度得到的**贫困线水平**（poverty rate）予以划定。
- 最终，300个贫困县中，有80%的县获得了项目资助；而其他提交申请的县中（非贫困县），有43%的县也获得了项目资助。

（死亡率案例）背景说明2/2

援助项目与儿童死亡率(续):

- (Ludwig and Miller, 2007)重点关注**援助项目**对中长期**儿童死亡率**影响。其中儿童死亡率定义为：1973-1983年间、儿童年龄范围在8-18岁、儿童死亡原因为Head Start定义的相关原因（如结核病等）。因而而援助项目希望努力消减这些儿童死亡情形的发生。
- 我们关注的问题：**脱贫援助项目**（Head Start）对**儿童死亡率**（`Y=mortality rate`）的因果效应。我们将采用骤变RDD非参数回归估计，**运行变量**为县贫困率（`X=poverty rate`），**断点值**(cut-off)设定为$c = 59.1984$。将使用**子样本数据**的样本数为n=2783。

（死亡率案例）样本数据集

援助项目数据集(n=2783)

obs	X	Y	D
1	15.2085	0.6846	0
2	15.2118	2.0734	0
3	15.2175	3.3101	0
4	15.2254	0.0000	0
5	15.2411	0.0000	0
6	15.2583	1.0910	0
7	15.2761	0.0000	0
8	15.2817	0.0000	0

Showing 1 to 8 of 2,783 entries

Previous 1 2 3 4 5 … 348 Next

- 样本数据的描述性统计如下：

```
      X              Y                D
Min.   :15    Min.   :  0    Min.   :0.0
1st Qu.:24    1st Qu.:  0    1st Qu.:0.0
Median :34    Median :  0    Median :0.0
Mean   :37    Mean   :  2    Mean   :0.1
3rd Qu.:47    3rd Qu.:  3    3rd Qu.:0.0
Max.   :82    Max.   :136    Max.   :1.0
```

（死亡率案例）样本数据集：分组描述性统计

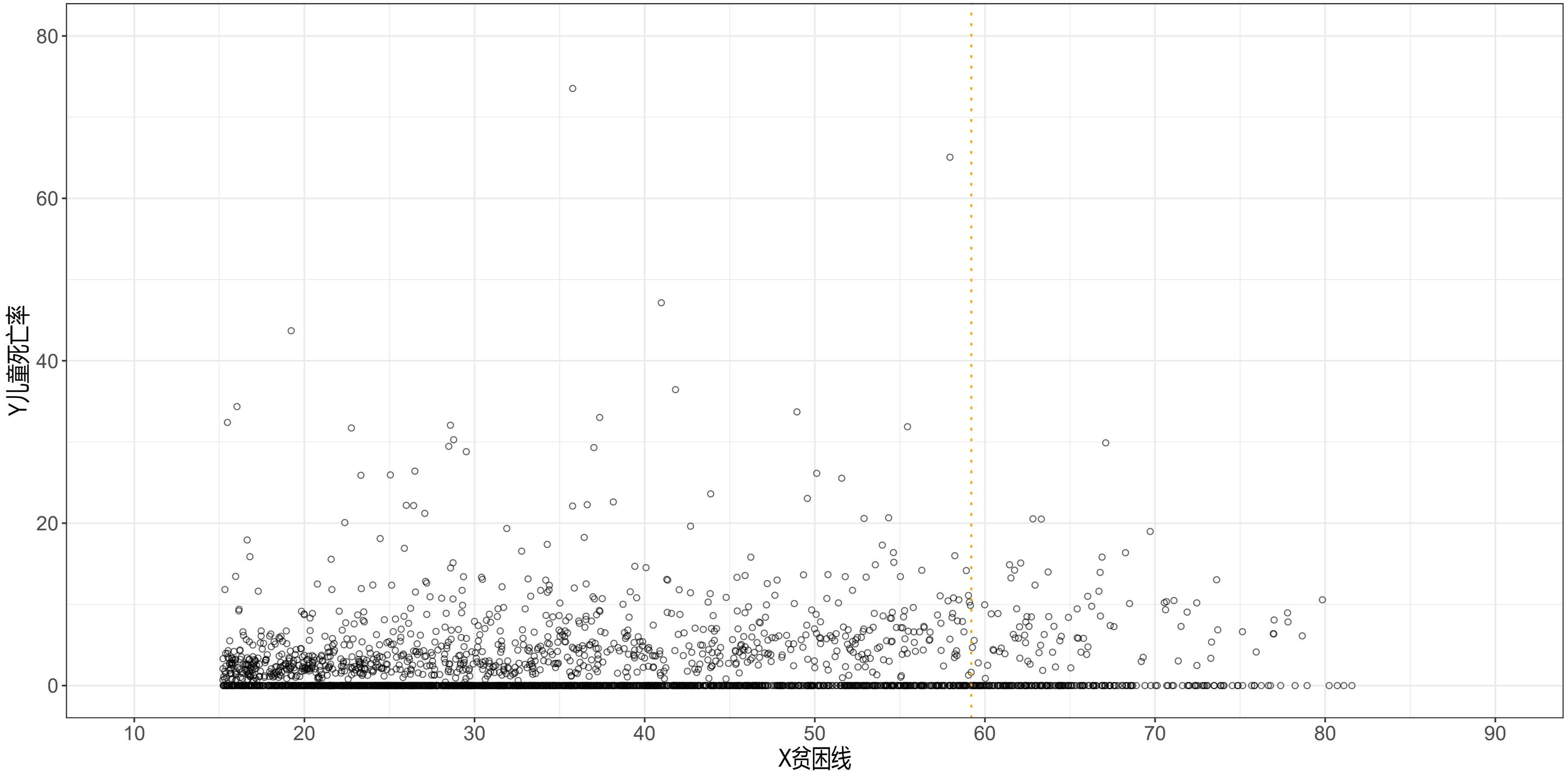
处置组和控制组描述性统计(q=2783)

stats	D0	D1
n	2,489.00	294.00
x_mean	33.29	65.87
x_min	15.21	59.20
x_max	59.19	81.57
x_sd	12.05	5.26
y_mean	2.23	2.42
y_min	0.00	0.00
y_max	136.05	29.90
y_sd	5.85	4.51

Showing 1 to 9 of 9 entries

Previous 1 Next

（死亡率案例）样本数据散点图



（死亡率案例）谱宽选择及CEF估计的规则策略

- 规则1：我们设定**先验谱宽**为 $h = 8$，断点值设定为 $c = 59.1984\%$。
- 规则2：分别设定断点两边**箱组中心点**序列值（center of bins）。我们将采用非对称箱组设置方法：

> - 控制组（断点左边）的评估范围为 $[15, 59.2]$，序列间隔为`0.2`。评估总箱组数为 $g1 = 222$，待评估序列值为 $15.0, 15.2, 15.4, 15.6, 15.8, \cdots, 58.6, 58.8, 59.0, 59.2$。
> - 处置组（断点右边）的评估范围为 $[59.2, 82]$，序列间隔为`0.2`。评估总箱组数为 $g2 = 115$，待评估序列值为 $59.2, 59.4, 59.6, 59.8, 60.0, \cdots, 81.4, 81.6, 81.8, 82.0$。

- 规则3：基于**三角核函数**（triangle kenerl）采用局部线性估计法，分别对断点两侧进行条件期望函数CEF $m(x)$进行估计，并得到估计值 $\widehat{m}(x)$（见下面附表和附图）。

（死亡率案例）CEF m(x)估计：计算附表

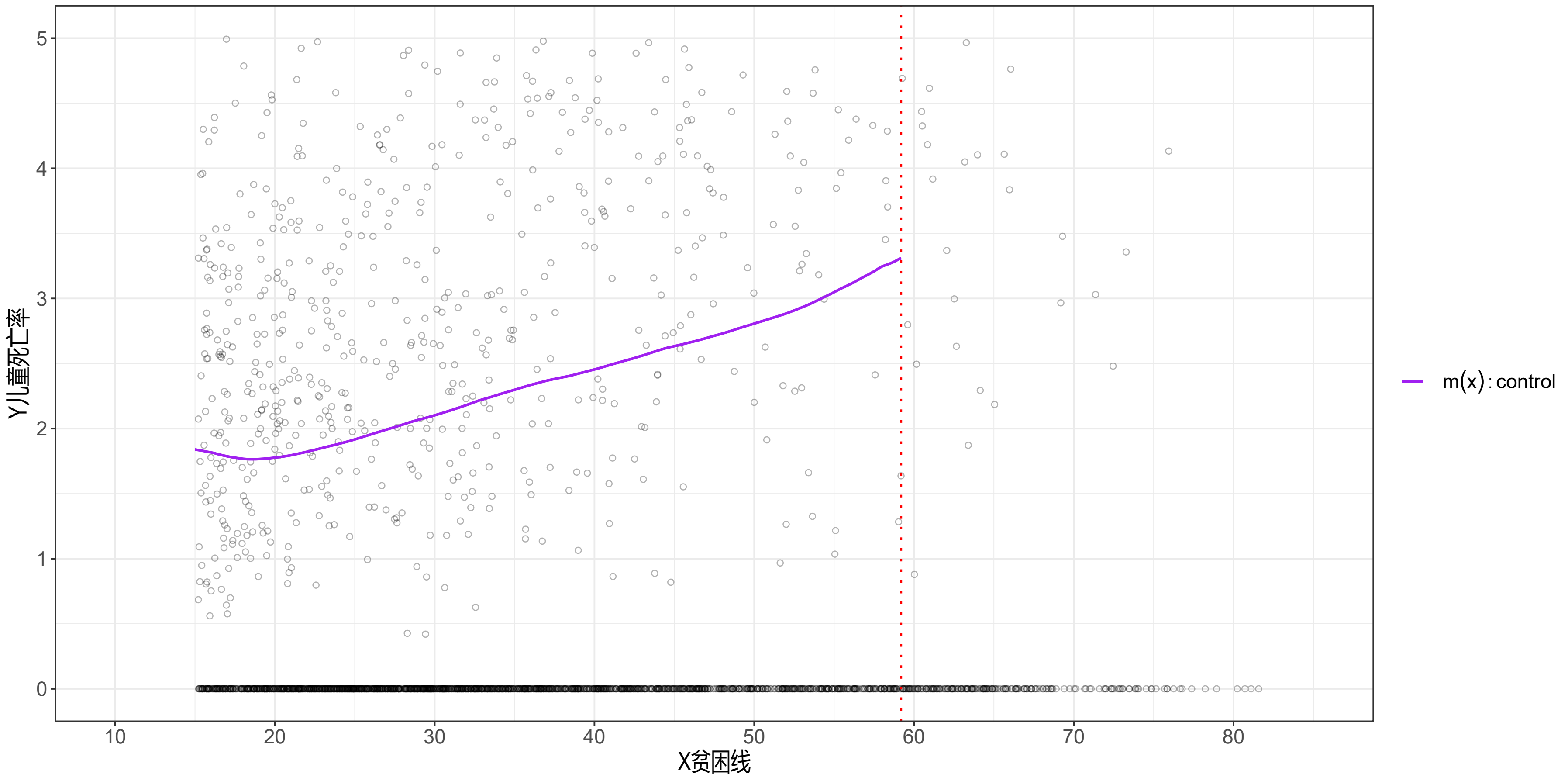
局部线性估计LL方法对m(x)的估计结果

index	group	xg	mx
1	control	15.0	1.8395
2	control	15.2	1.8347
3	control	15.4	1.8310
4	control	15.6	1.8260
5	control	15.8	1.8210
6	control	16.0	1.8169
7	control	16.2	1.8116
8	control	16.4	1.8042

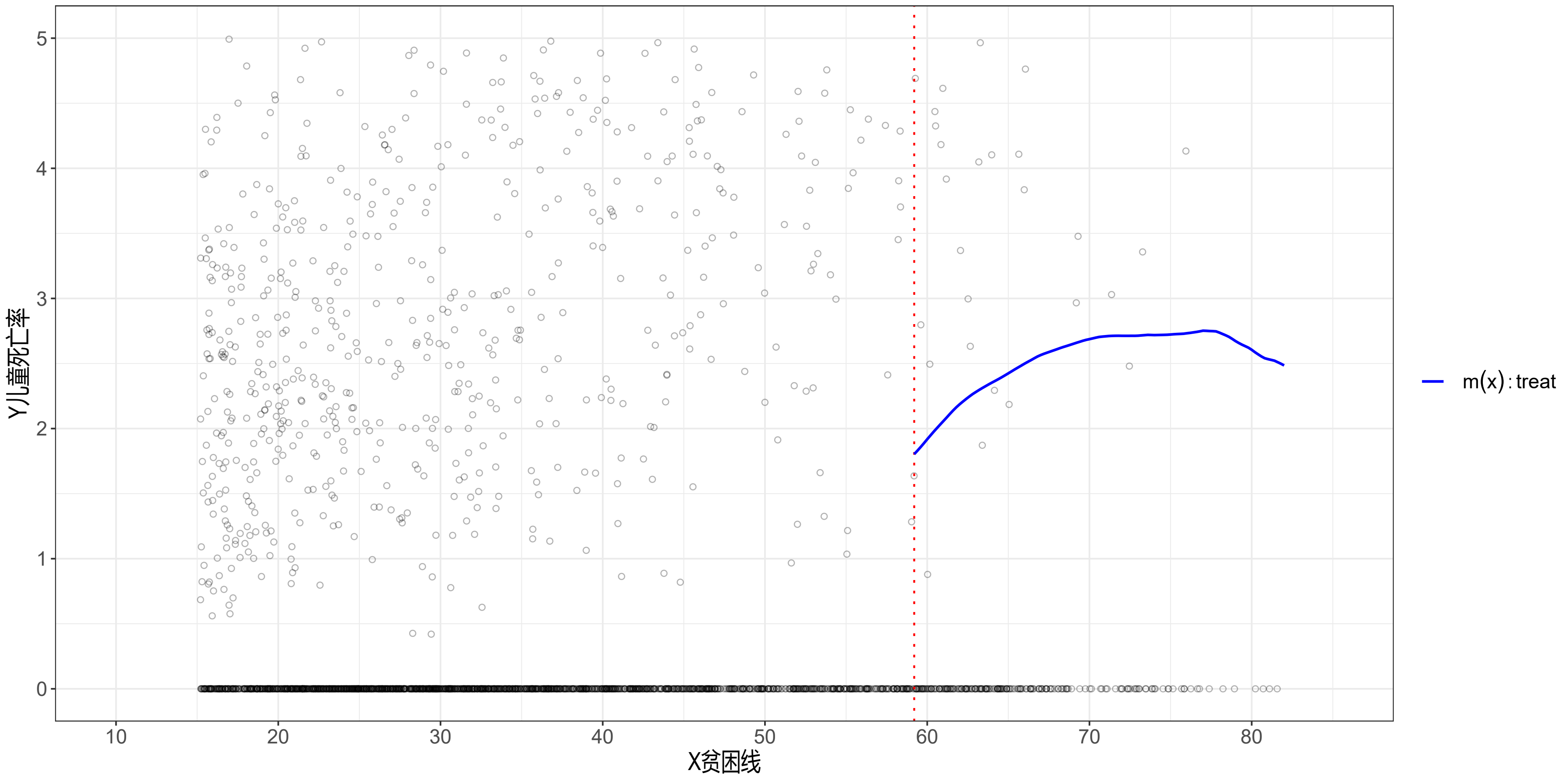
Showing 1 to 8 of 337 entries

Previous 1 2 3 4 5 … 43 Next

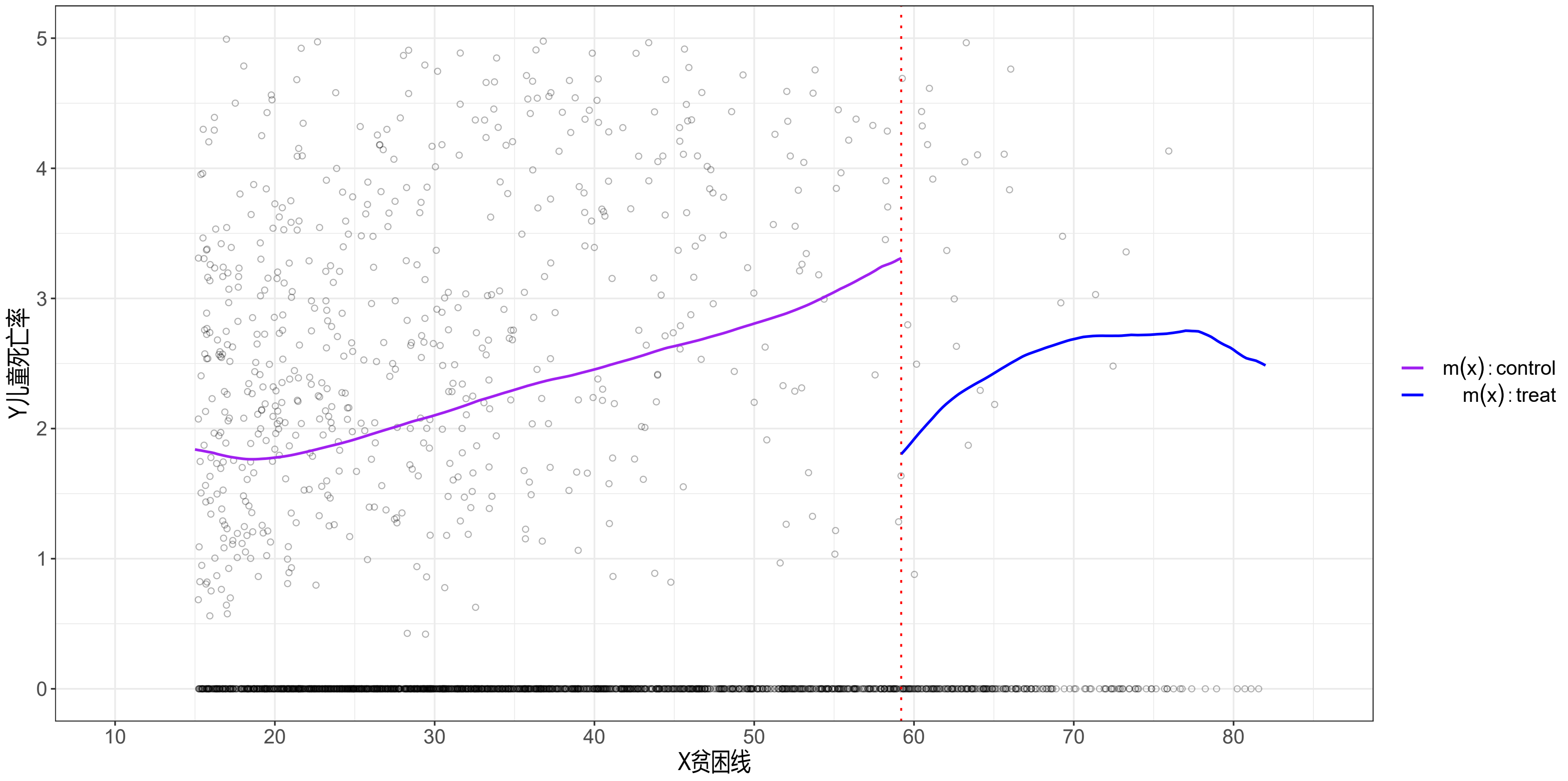
（死亡率案例）CEF m(x)估计图示：断点左侧（控制组）



(死亡率案例)CEF m(x)估计图示：断点右侧(处置组)



（死亡率案例）CEF m(x)估计图示：断点两侧（对比）



（死亡率案例）CEF方差估计：计算方差、标准差

- 直接使用谱宽[a] $h=8$进行局部线性LL估计，并利用**留一法**法计算得到**预测误差** $\tilde{e}$，并最终分别得断点两侧的协方差矩阵（见下式），从而进一步计算得到CEF估计值的方差和标准差（见后面附表）。

$$
\widehat{\boldsymbol{V}}_0=\left(\sum_{i=1}^{n}K_iZ_iZ_i'\cdot 1\left\{X_i<c\right\}\right)^{-1}\left(\sum_{i=1}^{n}K_i^2Z_iZ_i'\tilde{e}_i^2\cdot 1\left\{X_i<c\right\}\right)\left(\sum_{i=1}^{n}K_iZ_iZ_i'\cdot 1\left\{X_i<c\right\}\right)^{-1}
$$

$$
\widehat{\boldsymbol{V}}_1=\left(\sum_{i=1}^{n}K_iZ_iZ_i'\cdot 1\left\{X_i\geq c\right\}\right)^{-1}\left(\sum_{i=1}^{n}K_i^2Z_iZ_i'\tilde{e}_i^2\cdot 1\left\{X_i\geq c\right\}\right)\left(\sum_{i=1}^{n}K_iZ_iZ_i'\cdot 1\left\{X_i\geq c\right\}\right)^{-1}
$$

[a] 这里我们没有再次评估条件方差估计中的最优谱宽，而是简单直接地使用了CEF估计时的谱宽。但是我们还是要注意，二者的最优谱宽可以完全不相同！

（死亡率案例）CEJ方差估计：计算方差估计值（附表）

m(x)的样本方差和标准差估计结果

index	group	xg	mx	s	s2
1	control	15.0	1.8395	0.2396	0.0574
2	control	15.2	1.8347	0.2339	0.0547
3	control	15.4	1.8310	0.2284	0.0522
4	control	15.6	1.8260	0.2225	0.0495
5	control	15.8	1.8210	0.2166	0.0469
6	control	16.0	1.8169	0.2111	0.0446
7	control	16.2	1.8116	0.2050	0.0420
8	control	16.4	1.8042	0.1990	0.0396

Showing 1 to 8 of 337 entries

Previous 1 2 3 4 5 … 43 Next

（死亡率案例）CEF的置信区间和置信带

- 进一步计算**局部线性估计**下的**逐点置信区间**（Pointwise Confidence Interval）（见后面附表），并得到置信带（见后面附图）。

$$\widehat{m}(x) \pm z_{1-\alpha/2}(n-1) \cdot \sqrt{\widehat{V}_{\widehat{m}(x)}}$$
$$\widehat{m}(x) \pm 1.96\sqrt{\widehat{V}_{\widehat{m}(x)}}$$

（死亡率案例）CLM的置信区间和置信带（附表）

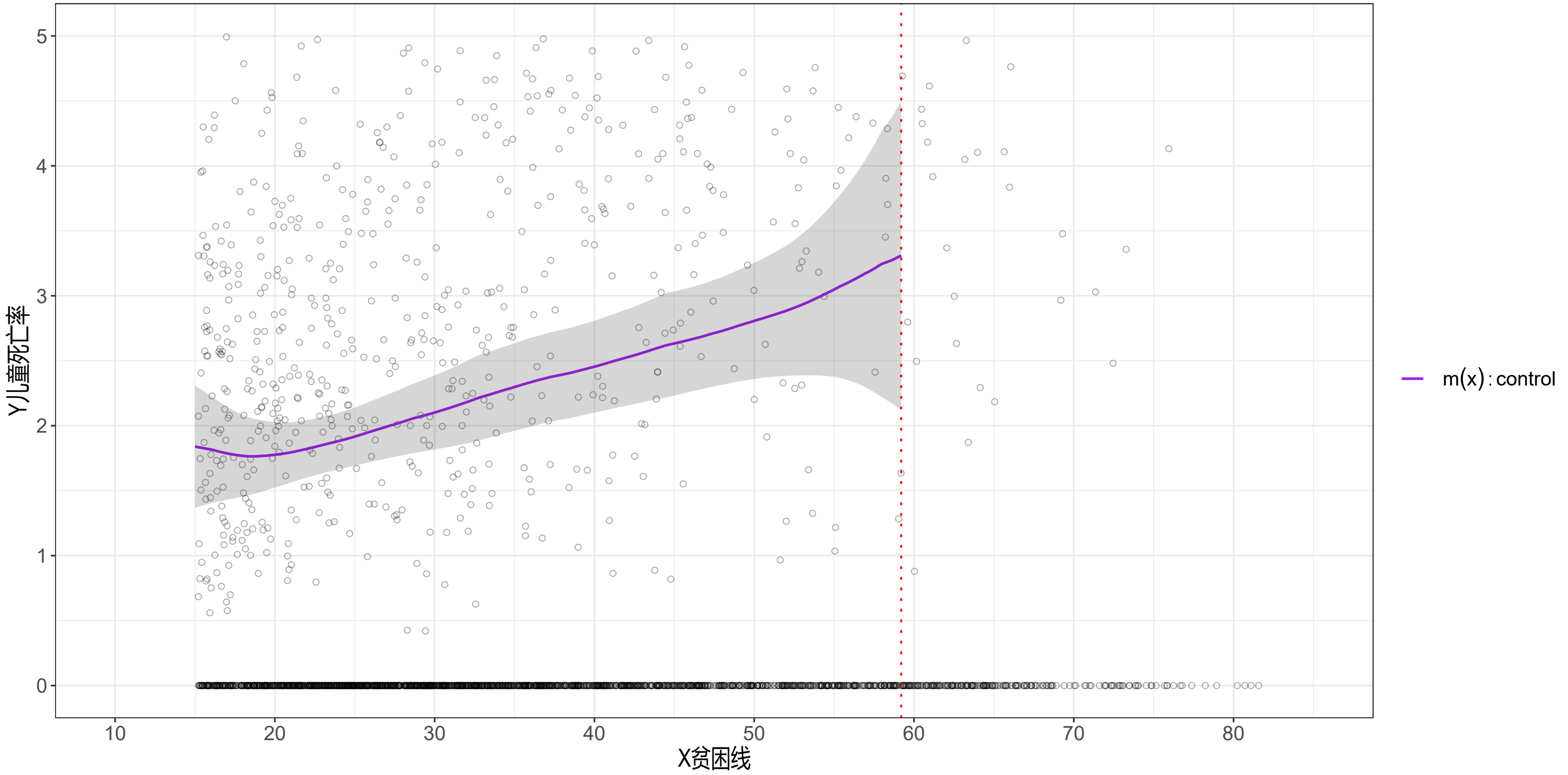
m(x)的逐点置信区间估计结果

group	index	xg	mx	s	lwr	upr
control	1	15.0	1.8395	0.2396	1.3699	2.3092
control	2	15.2	1.8347	0.2339	1.3762	2.2931
control	3	15.4	1.8310	0.2284	1.3833	2.2787
control	4	15.6	1.8260	0.2225	1.3899	2.2622
control	5	15.8	1.8210	0.2166	1.3964	2.2456
control	6	16.0	1.8169	0.2111	1.4032	2.2307
control	7	16.2	1.8116	0.2050	1.4098	2.2134
control	8	16.4	1.8042	0.1990	1.4142	2.1942

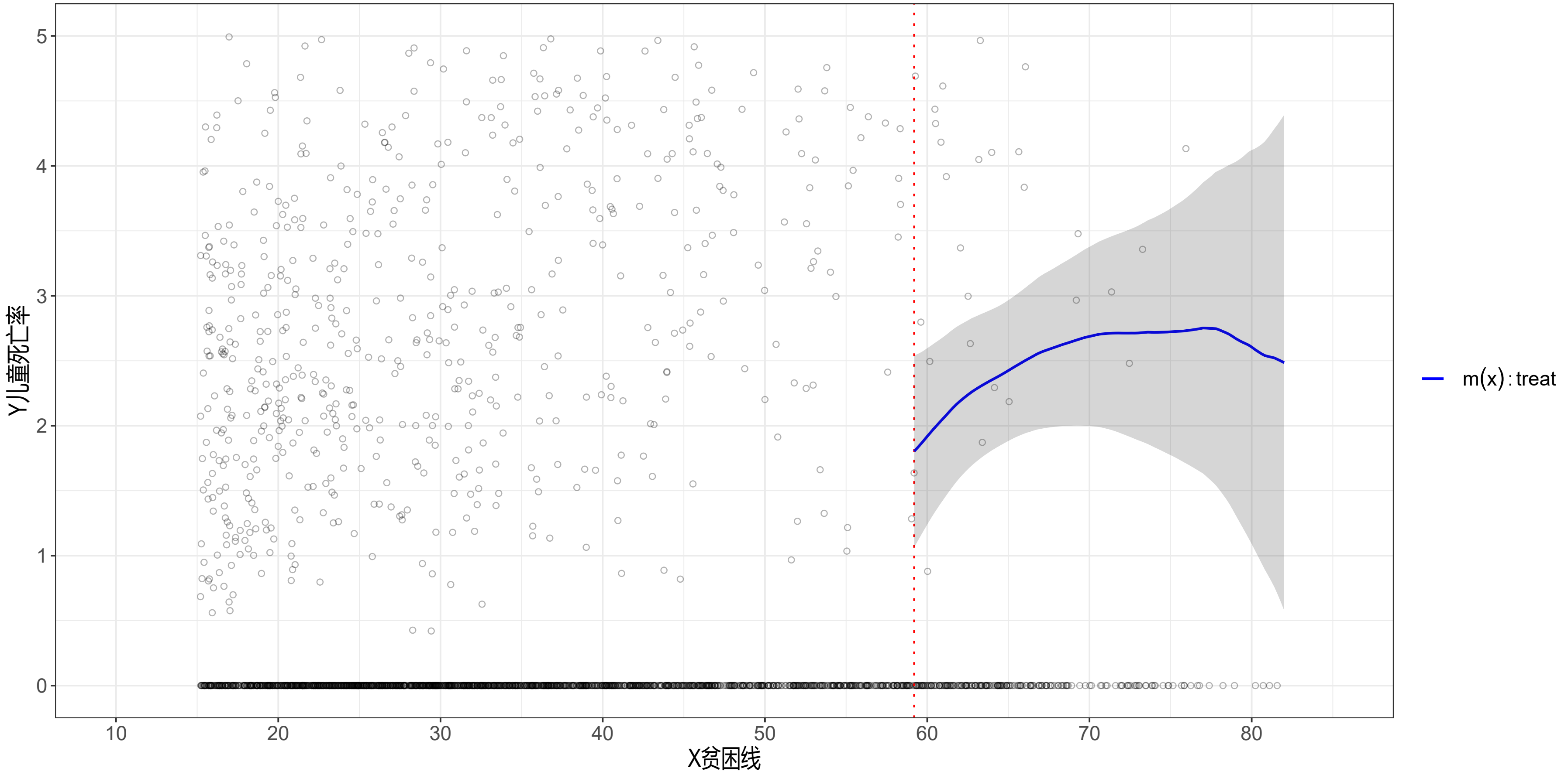
Showing 1 to 8 of 337 entries

Previous 1 2 3 4 5 … 43 Next

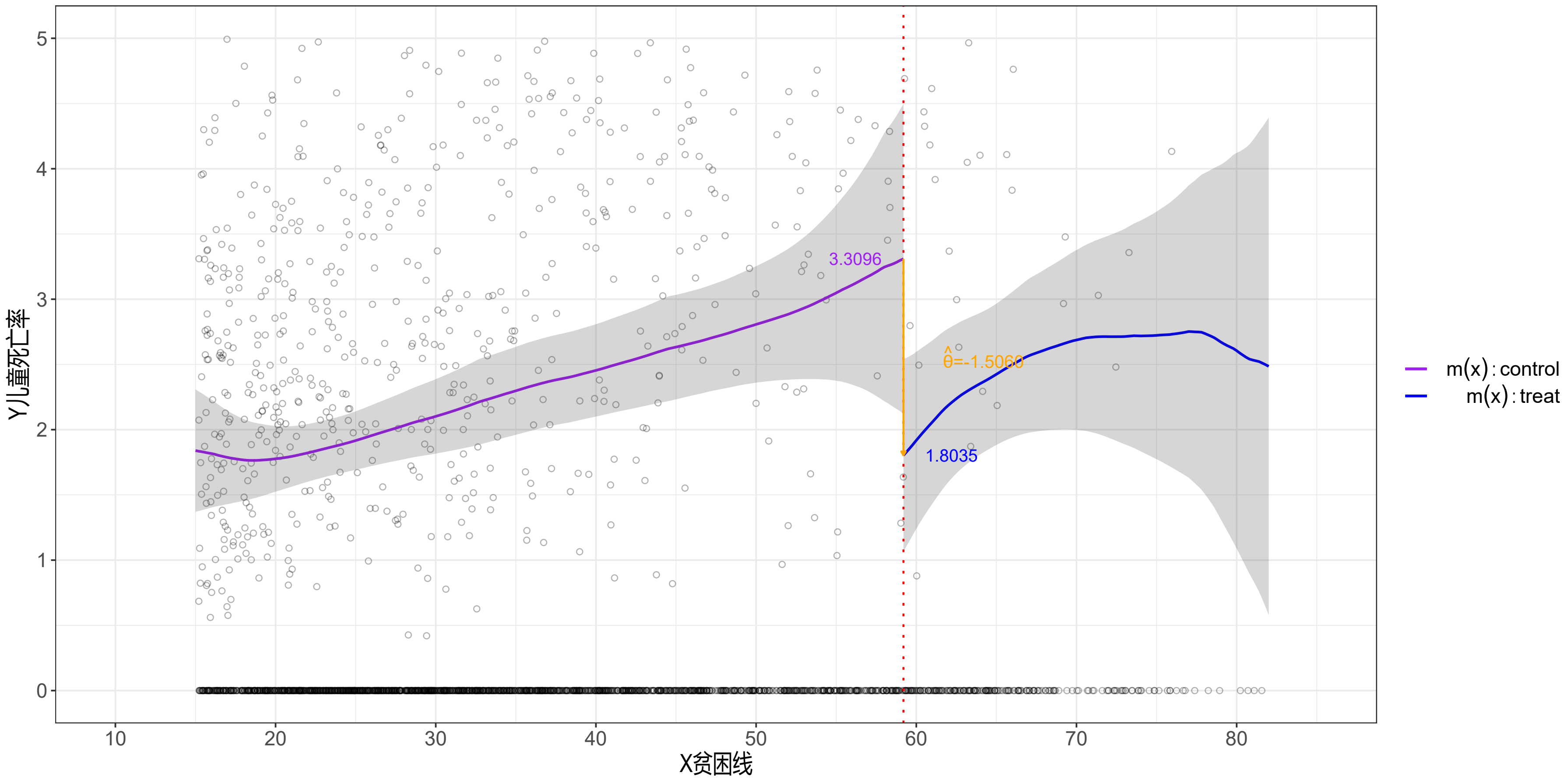
(死亡率案例)CEF的置信区间和置信带：断点左侧(控制组)



（死亡率案例）CEF的置信区间和置信带：断点右侧（处置组）



(死亡率案例)CSY的置信区间和置信带：断点两侧侧(对比)



（死亡率案例）RDD断点处置效应：计算结果

- 根据**断点处置效应定理**，可以得到在断点 $x = c = 59.1984$处对总体平均处置效应 $\bar{\theta}$的样本估计结果 $\hat{\theta}$：

$$
\begin{aligned}
\hat{\theta} &= \left[\widehat{\boldsymbol{\beta}}_{\mathbf{1}}(c)\right]_1 - \left[\widehat{\boldsymbol{\beta}}_{\mathbf{0}}(c)\right]_1 \\
&= \hat{m}(c+) - \widehat{m}(c-) \\
&= 3.3096 - 1.8035 = -1.5060
\end{aligned}
$$

- 断点处置效应估计值为 $\hat{\theta} = -1.5060$。

> - 断点左边的条件期望(CEF)的估计值 $\widehat{m}(c-) = 3.31$;
> - 断点右边的条件期望(CEF)的估计值 $\widehat{m}(c+) = 1.8$;

- **结论**：援助项目的实施，减低了儿童死亡率，使得10万个孩子中约1.51个小孩免于遭受死亡。相比不实施项目援助，儿童死亡率由3.3096，下降到1.8035，降幅接近50%。

（死亡率案例）RDD断点处置效应：估计误差及显著性检验

- 进一步地，估计系数 $\hat{\theta}$ 的**渐进方差**为两个方差协方差矩阵第一个对角元素之和：

$$
\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \left[\widehat{\boldsymbol{V}}_0\right]_{11} + \left[\widehat{\boldsymbol{V}}_1\right]_{11} \\
&= 0.3673 + 0.1417 = 0.5090 \\
se((\hat{\theta})) &= \sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{0.5090} = 0.7134
\end{aligned}
$$

> - 断点左边的条件期望(CEF)的估计值 $\widehat{m}(c-) = 3.3096$；
> - 断点右边的条件期望(CEF)的估计值 $\widehat{m}(c+) = 1.8035$；

- **结论**：援助项目的实施，减低了儿童死亡率，使得10万个孩子中约-1.5060个小孩免于遭受死亡。相比不实施项目援助，儿童死亡率由3.3096，下降到1.8035，降幅接近50%。

（死亡率案例）等价线性回归：调整运行变量范围

- 如前所述，**骤变RDD断点处置效应**也可以通过如下简单线性回归方法等价地得到 $\hat{\theta}$的对应估计值：

$$Y = \beta_0 + \beta_1 X + \beta_3 (X - c) D + \theta D + e$$

- 简单地，上述等价模型需要进行样本数据集的重新定义。具体地，**运行变量** X的范围需要调整到 $X \in [c - h^*, c + h^*]$，其中 $h^* = \sqrt{3}h = \sqrt{3} \times 8 = 13.86$

（死亡率案例）等价线性回归：调整后的数据集

调整过后的数据集(n=757)

obs	X	Y	D	XcD
1	45.3427	4.2082	0	0.0000
2	45.3510	0.0000	0	0.0000
3	45.3609	2.6114	0	0.0000
4	45.3783	0.0000	0	0.0000
5	45.3821	2.7903	0	0.0000
6	45.4081	0.0000	0	0.0000
7	45.4197	0.0000	0	0.0000
8	45.4256	13.3280	0	0.0000

Showing 1 to 8 of 757 entries

Previous 1 2 3 4 5 … 95 Next

- 样本数据的描述性统计如下：

```
       X               Y              D
 Min.   :45    Min.   : 0    Min.   :0.00
 1st Qu.:50    1st Qu.: 0    1st Qu.:0.00
 Median :55    Median : 0    Median :0.00
 Mean   :56    Mean   : 3    Mean   :0.34
 3rd Qu.:62    3rd Qu.: 4    3rd Qu.:1.00
 Max.   :73    Max.   :65    Max.   :1.00
```

（死亡率案例）等价线性回归：分组描述性统计

处置组和控制组描述性统计(n=757)

stats	D0	D1
n	500.00	257.00
x_mean	51.89	64.38
x_min	45.34	59.20
x_max	59.19	73.04
x_sd	4.11	3.61
y_mean	2.83	2.41
y_min	0.00	0.00
y_max	65.06	29.90
y_sd	5.46	4.61

Showing 1 to 9 of 9 entries

Previous 1 Next

（死亡率案例）等价线性回归：OLS估计结果

$$
\begin{array}{llllll}
\widehat{Y} = & -1.0987 & +0.0758X_i & +0.0331XcD_i & -1.5454D_i \\
(s) & (2.9382) & (0.0564) & (0.1060) & (0.7375) \\
(t) & (-0.37) & (+1.34) & (+0.31) & (-2.10) \\
(over) & n = 757 & \hat{\sigma} = 5.1830 & & \\
(fit) & R^2 = 0.0059 & \bar{R}^2 = 0.0019 & & \\
(Ftest) & F^* = 1.48 & p = 0.2191 & &
\end{array}
$$

- 用上述等价回归法估计得到的断点处置效应估计值为 $\hat{\theta} = -1.5454$，样本t统计量为 $t^* = -2.10$，对应的概率值为 $p = 0.0180$，表明是统计显著的。

- 回顾**断点处置效应**定理：

> 给定处置分配规则为 $D = 1\{X \geq c\}$，而且假定结果变量满足断点处的连续性假设，也即结果变量的条件期望函数 $m(x)$在断点处 $x = c$连续，那么**断点处置效应**为：

$$\bar{\theta} = m(c+) - m(c-)$$

- 根据前面的讨论，就效应**估计**和**推断**而言，RDD分析中完全没有必要引入其他协变量（Z）进入模型。
- 当然，为了提高模型预测**准确度**，我们可以引入一些额外的、有价值的协变量。

2.5 协变量RDD：符号表达

- 给定变量集为：(Y, X, Z)，其中 Z为含有 k个元素的协变量向量（covariates vector）
- 同前，Y_0和 Y_1分别为**控制条件**和**处置条件**下的**结果变量**（观测的或反事实的）
- 并进一步假定条件期望函数CEF是如下的**线性形式**，且断点两边的方程中协变量系数是相同的 β'：

$$
\begin{aligned}
\mathbb{E}\left[Y_0 \mid X = x, Z = z\right] = m_0(x) + \beta' z \\
\mathbb{E}\left[Y_1 \mid X = x, Z = z\right] = m_1(x) + \beta' z
\end{aligned}
$$

- 那么，结果变量 Y的条件期望函数CEF将可以表达为：

$$
m(x, z) = m_0(x) \cdot 1\{x < c\} + m_1(x) \cdot 1\{x \geq c\} + \beta' z
$$

- 此时，可以证明**断点处置效应**结果为：

$$
\bar{\theta} = m(c+, z) - m(c-, z)
$$

2.5 协变量RDD：估计方法

RDD协变量估计方法有很多种，这里重点讨论(Robinson, 1988)提出了一种**半参数效率**估计方法，主要步骤如下：

- **步骤1**：直接采用前面的RDD局部线性回归方法（RDD LLR），用 Y_i对 X_i进行回归，并得到第1阶段的**结果变量**的拟合值 $\widehat{m}_i = \widehat{m}_i(X_i)$
- **步骤2**：依次做 Z_{i1}对 X_i、Z_{i2}对 X_i、...的局部线性回归Z（LL），并分别得到**协变量**的拟合值 $\hat{g}_{1i}, \hat{g}_{2i}, \ldots, \hat{g}_{ki}$
- **步骤3**：做 $Y_i - m_i$对 $Z_{i1} - \hat{g}_{1i}, Z_{i2} - \hat{g}_{2i}, \ldots, Z_{ik} - \hat{g}_{ki}$的回归，并得到**估计系数** $\hat{\beta}$及**其标准误**
- **步骤4**：构造残差 $\hat{e}_i = Y_i - Z_i'\hat{\beta}$
- **步骤5**：再次采用RDD局部线性回归方法（LLR），进行 $\hat{e}_i$对 X_i的回归，并计算得到非参数估计量 $\widehat{m}(x)$，断点效应估计值 $\hat{\theta}$及其标准误。

（死亡率案例）背景说明

案例说明：

我们继续使用前面(Ludwig and Miller, 2007)的研究案例，来评估美国联邦政府脱贫援助项目（Head Start）对儿童死亡率的**骤变RDD**政策效应。现在我们考虑使用两个协变量（covariates）：

- 县级**黑人人口占比**（`black pop percentage`） Z_a
- 县级**城镇人口占比**（`urban pop percentage`） Z_a

- 上述两个协变量，本质上可以视作为**收入**变量（`income`）的**代理变量**（proxy）。
- 下面我们将使用(Robinson, 1988)的**半参数效率**估计方法来评估项目援助的断点处置效应（RDD ATE）。

（死亡率案例）样本数据集

增加协变量的援助项目数据集(n=2783)

obs	X	Y	Za	Zb	D
1	15.2085	0.6846	0.3	70.2	0
2	15.2118	2.0734	8.4	67.0	0
3	15.2175	3.3101	1.4	51.2	0
4	15.2254	0.0000	0.5	26.9	0
5	15.2411	0.0000	0.0	26.5	0
6	15.2583	1.0910	11.8	92.2	0
7	15.2761	0.0000	1.4	54.4	0
8	15.2817	0.0000	0.1	43.2	0

Showing 1 to 8 of 2,783 entries

Previous 1 2 3 4 5 … 348 Next

- 样本数据的描述性统计如下：

```
       X                Y                 Za
 Min.   :15     Min.   :  0     Min.   : 0
 1st Qu.:24     1st Qu.:  0     1st Qu.: 0
 Median :34     Median :  0     Median : 2
 Mean   :37     Mean   :  2     Mean   :11
 3rd Qu.:47     3rd Qu.:  3     3rd Qu.:15
 Max.   :82     Max.   :136     Max.   :83
```

（死亡率案例）样本数据集：分组描述性统计

处置组和控制组描述性统计(q=2783)

stats	D0	D1
zb_sd	26.99	18.09
zb_min	0.00	0.00
zb_mean	31.01	13.39
zb_max	100.00	93.70
za_sd	12.81	26.37
za_min	0.00	0.00
za_mean	7.87	33.91
za_max	75.50	83.40
y_sd	5.85	4.51
y_min	0.00	0.00

Showing 1 to 10 of 17 entries

Previous 1 2 Next

（死亡率案例）协变量RDD：规则策略

在进行协变量RDD分析之前，我们设定如下的规则策略：

- 规则1：我们设定**先验谱宽**为 $h = 8$，断点值设定为 $c = 59.1984\%$。
- 规则2：分别设定断点两边**箱组中心点**序列值（center of bins）。我们将采用非对称箱组设置方法：

> - 控制组（断点左边）的评估范围为 $[15, 59.2]$，序列间隔为`0.2`。评估总箱组数为 $g1 = 222$，待评估序列值为 $15.0, 15.2, 15.4, 15.6, 15.8, \cdots, 58.6, 58.8, 59.0, 59.2$。
> - 处置组（断点右边）的评估范围为 $[59.2, 82]$，序列间隔为`0.2`。评估总箱组数为 $g2 = 115$，待评估序列值为 $59.2, 59.4, 59.6, 59.8, 60.0, \cdots, 81.4, 81.6, 81.8, 82.0$。

- 规则3：如果使用局部线性估计法（LL），则采用**三角核函数**（triangle kenerl）。
- 规则4：我们将使用(Robinson, 1988)的**半参数效率**估计方法来评估断点处置效应（RDD ATE）。

（死亡率案例）协变量RDD：第1阶段LLR估计残差

- **步骤1**：直接采用前面的局部线性回归方法（LLR），用 Y_i对 X_i进行LL回归，得到第1阶段的**结果变量**的拟合值 $\widehat{m}_i = \widehat{m}_i(X_i)$，并进一步构造**留一法**残差[a] $Y_i - \widehat{m}_i$

RDD LL估计得到的残差(n=2783)

obs	D	X	Y	Za	Zb	e
1	0	15.2085	0.6846	0.3	70.2	-1.1544
2	0	15.2118	2.0734	8.4	67.0	0.2399
3	0	15.2175	3.3101	1.4	51.2	1.4815
4	0	15.2254	0.0000	0.5	26.9	-1.8413
5	0	15.2411	0.0000	0.0	26.5	-1.8409
6	0	15.2583	1.0910	11.8	92.2	-0.7454

Showing 1 to 6 of 2,783 entries

Previous 1 2 3 4 5 … 464 Next

[a] 这个阶段的残差序列用e命名。

（死亡率案例）协变量RDD：第2阶段LLR估计残差

- **步骤2**：同上，依次做 Z_a对 X_i、 Z_b对 X_i的**局部线性回归**（LLR），并分别得到**协变量**的拟合值 $\hat{g}_{1i},\hat{g}_{2i}$，及其对应残差[a] $(Z_a-\hat{g}_{1i}),(Z_b-\hat{g}_{2i})$

RDD LL估计得到的残差(n=2783)

obs	D	X	Y	Za	Zb	e	Ra	Rb
1	0	15.2085	0.6846	0.3	70.2	-1.1544	-1.2525	20.8178
2	0	15.2118	2.0734	8.4	67.0	0.2399	6.8787	17.6103
3	0	15.2175	3.3101	1.4	51.2	1.4815	-0.1493	1.7575
4	0	15.2254	0.0000	0.5	26.9	-1.8413	-1.0537	-22.6246
5	0	15.2411	0.0000	0.0	26.5	-1.8409	-1.5575	-23.0015
6	0	15.2583	1.0910	11.8	92.2	-0.7454	10.2859	42.9783

Showing 1 to 6 of 2,783 entries

Previous 1 2 3 4 5 … 464 Next

[a] 这个阶段的两个残差序列分别用Ra和Rb命名。

- **步骤3**：利用前面两个阶段的残差，做 $Y_i - m_i$对 $Z_{i1} - \hat{g}_{1i}, Z_{i2} - \hat{g}_{2i}, \ldots, Z_{ik} - \hat{g}_{ki}$的**无截距**的普通最小二乘回归（OLS），并得到**估计系数** $\hat{\beta}$及其**标准误**

$$
(Y_i - m_i) = \hat{\beta}_1(Z_{ia} - \hat{g}_{1i}) + \hat{\beta}_2(Z_{ib} - \hat{g}_{2i})
$$

$$
e = \hat{\beta}_1 R_a + \hat{\beta}_2 R_a
$$

（死亡率案例）协变量RDD：第3阶段OLS估计（结果）

- 上述模型，未矫正标准误下OLS估计的结果如下[a]：

$$
\begin{aligned}
\hat{e} = &\quad +0.0265Ra_i &-0.0094Rb_i \\
(s) &\quad (0.0083) & (0.0045) \\
(t) &\quad (+3.19) & (-2.08) \\
(p) &\quad (0.0014) & (0.0377) \\
(over)n = 2783 && \hat{\sigma} = 5.7091
\end{aligned}
$$

- 上述模型，进行稳健标准误矫正OLS估计的结果如下[b]：

稳健标准误OLS估计(n=2783)

term	estimate	std.error	statistic	p.value
Ra	0.0265	0.0073	3.62	0.0003
Rb	-0.0094	0.0046	-2.04	0.0412

[a b] 两种OLS估计程序下，回归系数都相同，只是系数对应的标准误不一样。这里我们仅需要用到回归系数，因此不影响后续步骤。

（死亡率案例）协变量RDD：构造残差

- **步骤4：** 利用前面的OLS估计系数，我们就可以构造得到残差 $\hat{e}_i = Y_i - Z_i'\hat{\beta}$

RDD LL估计得到的残差(n=2783)

obs	D	X	Y	Za	Zb	e	Ra	Rb	RZ
1	0	15.2085	0.6846	0.3	70.2	-1.1544	-1.2525	20.8178	1.3390
2	0	15.2118	2.0734	8.4	67.0	0.2399	6.8787	17.6103	2.4826
3	0	15.2175	3.3101	1.4	51.2	1.4815	-0.1493	1.7575	3.7560
4	0	15.2254	0.0000	0.5	26.9	-1.8413	-1.0537	-22.6246	0.2405
5	0	15.2411	0.0000	0.0	26.5	-1.8409	-1.5575	-23.0015	0.2500
6	0	15.2583	1.0910	11.8	92.2	-0.7454	10.2859	42.9783	1.6477

Showing 1 to 6 of 2,783 entries   Previous 1 2 3 4 5 … 464 Next

[a] 这个步骤构造出来的残差序列RZ。

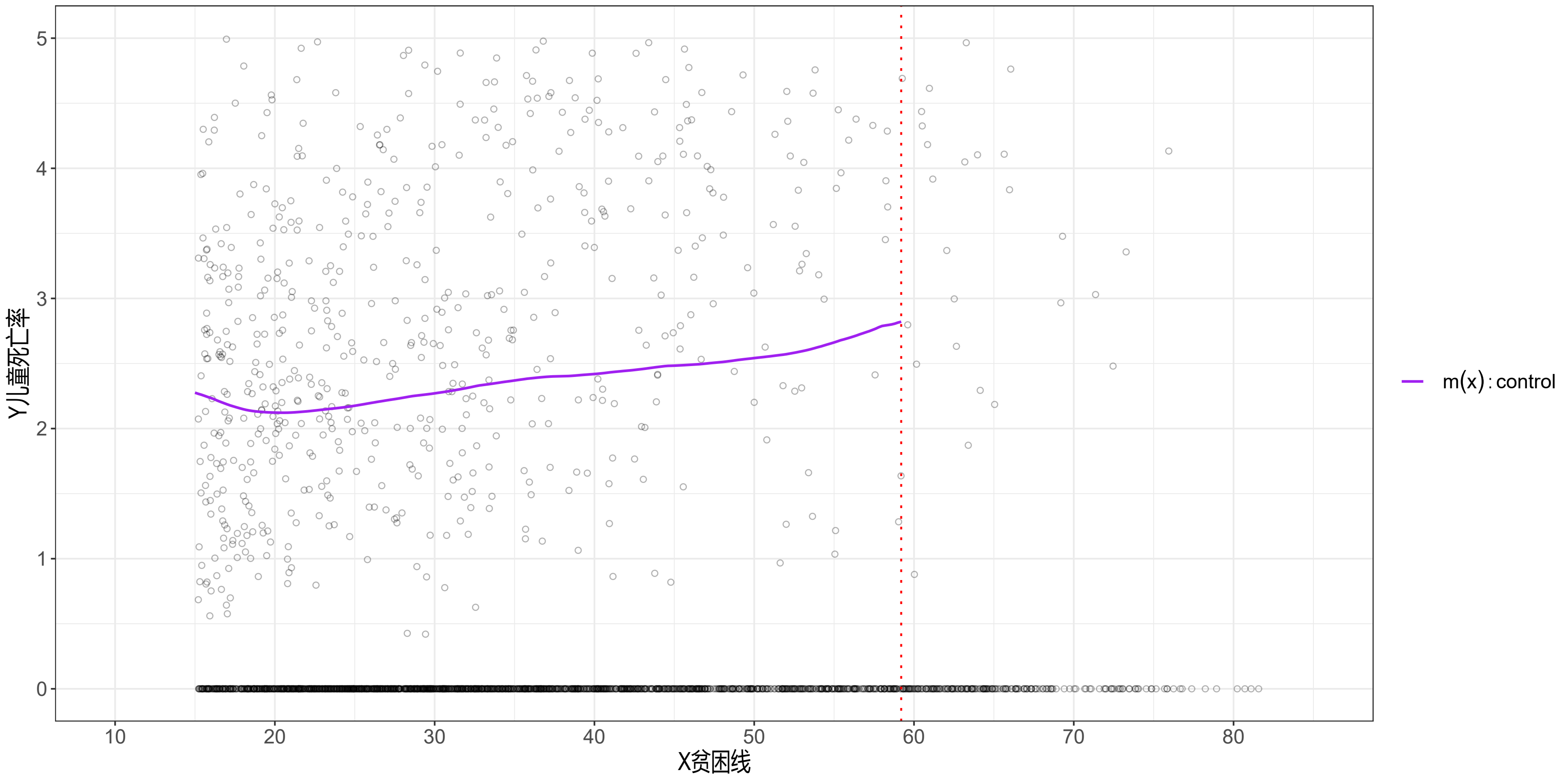
- **步骤5：** 再次采用RDD局部线性回归方法（LLR），进行 $\hat{e}_i$对 X_i的回归，并计算得到非参数估计量 $\widehat{m}(x)$，断点效应估计值 $\hat{\theta}$及其标准误。

局部线性估计LL方法对m(x)的估计结果

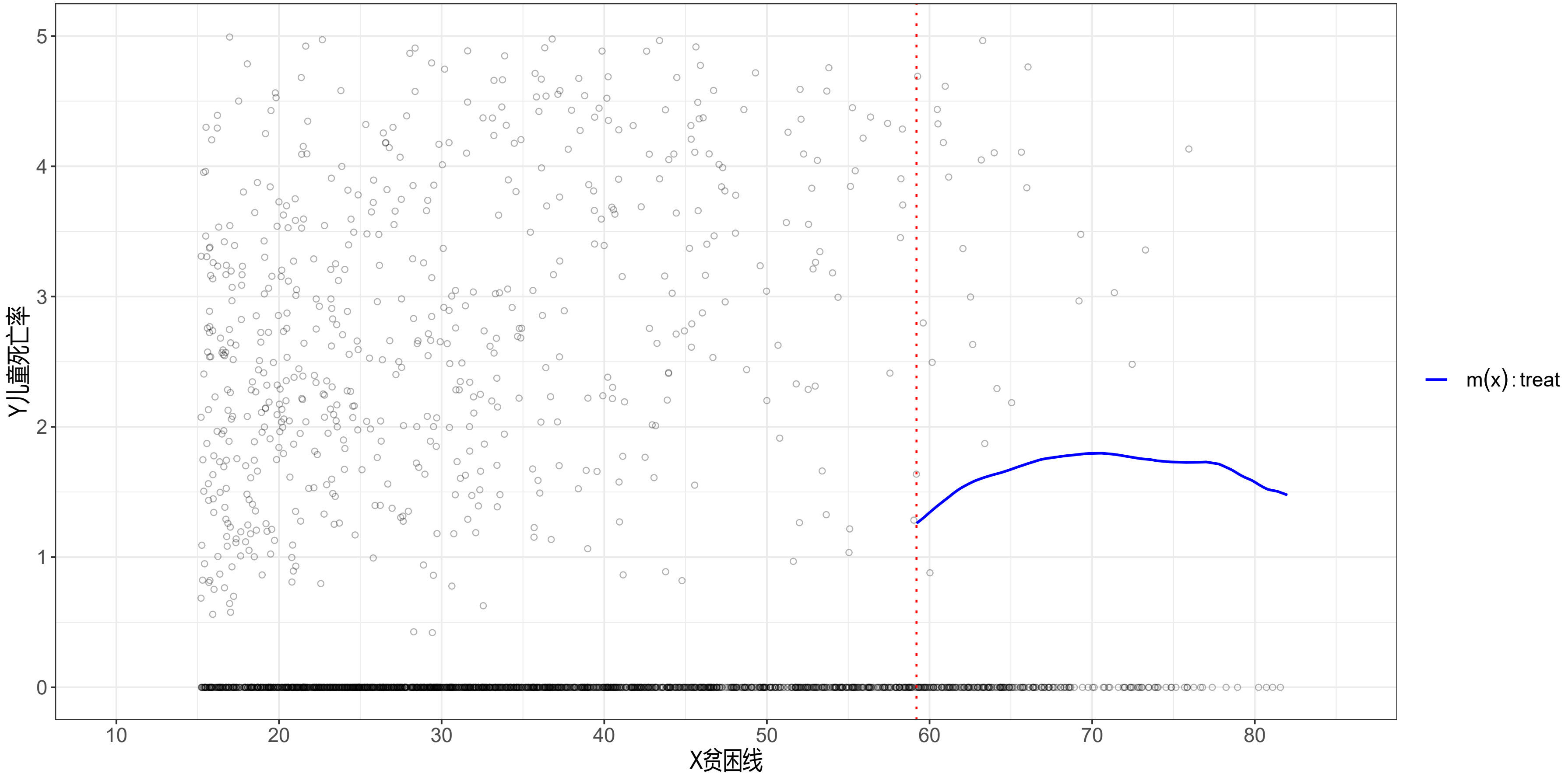
index	group	xg	mx
1	control	15.0	2.2757
2	control	15.2	2.2674
3	control	15.4	2.2601
4	control	15.6	2.2516
5	control	15.8	2.2428
6	control	16.0	2.2350

Showing 1 to 6 of 337 entries　　Previous 1 2 3 4 5 … 57 Next

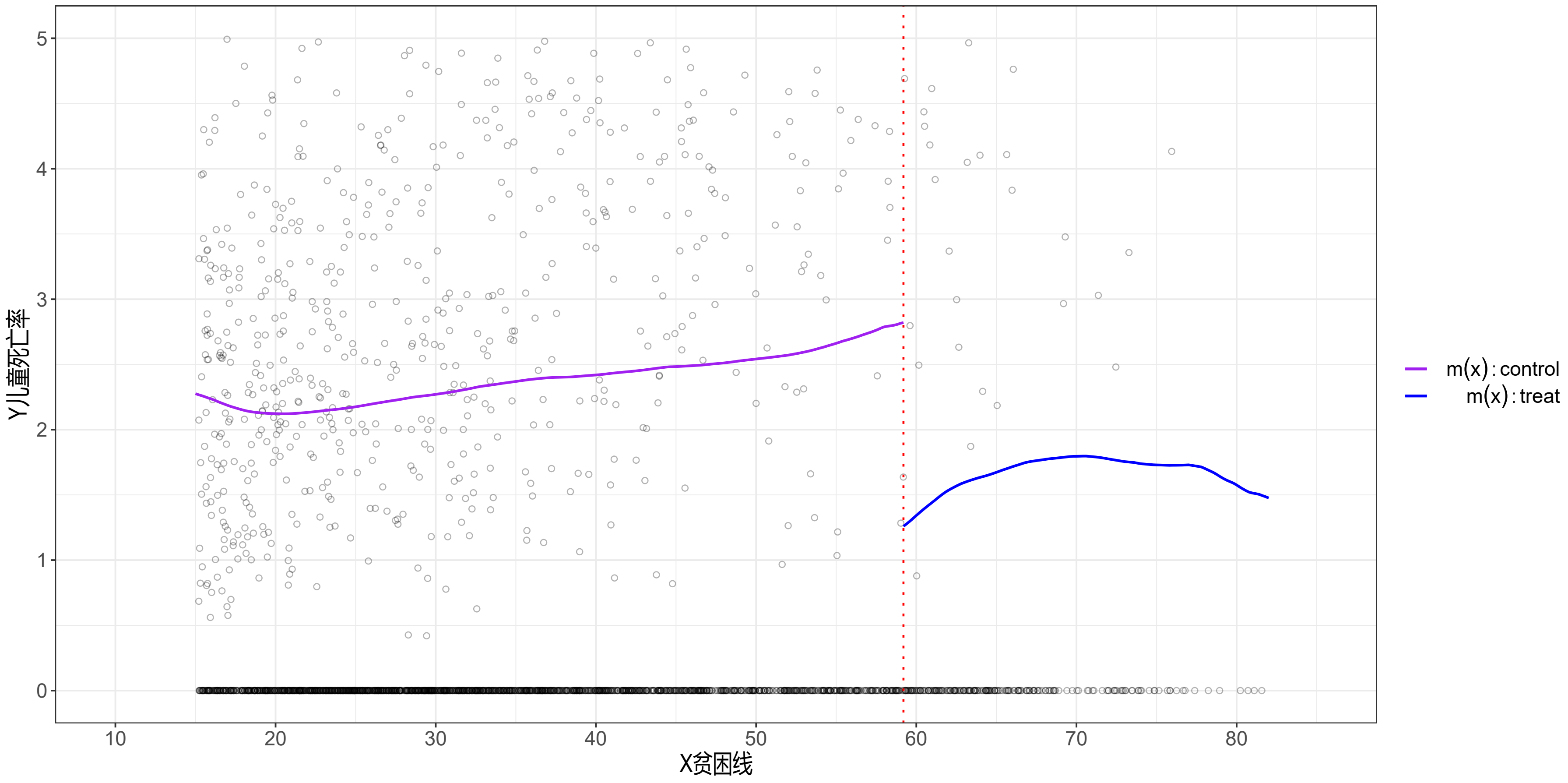
(死亡率案例)协变量RDD：断点左侧CEF(控制组)



（死亡率案例）协变量RDD：断点右侧CEF（处置组）



（死亡率案例）协变量RDD：断点两侧（对比）



（死亡率案例）协变量RDD：标准差和置信区间（附表）

- 同前，进一步计算得到CEF估计值的方差和标准差以及95%置信区间

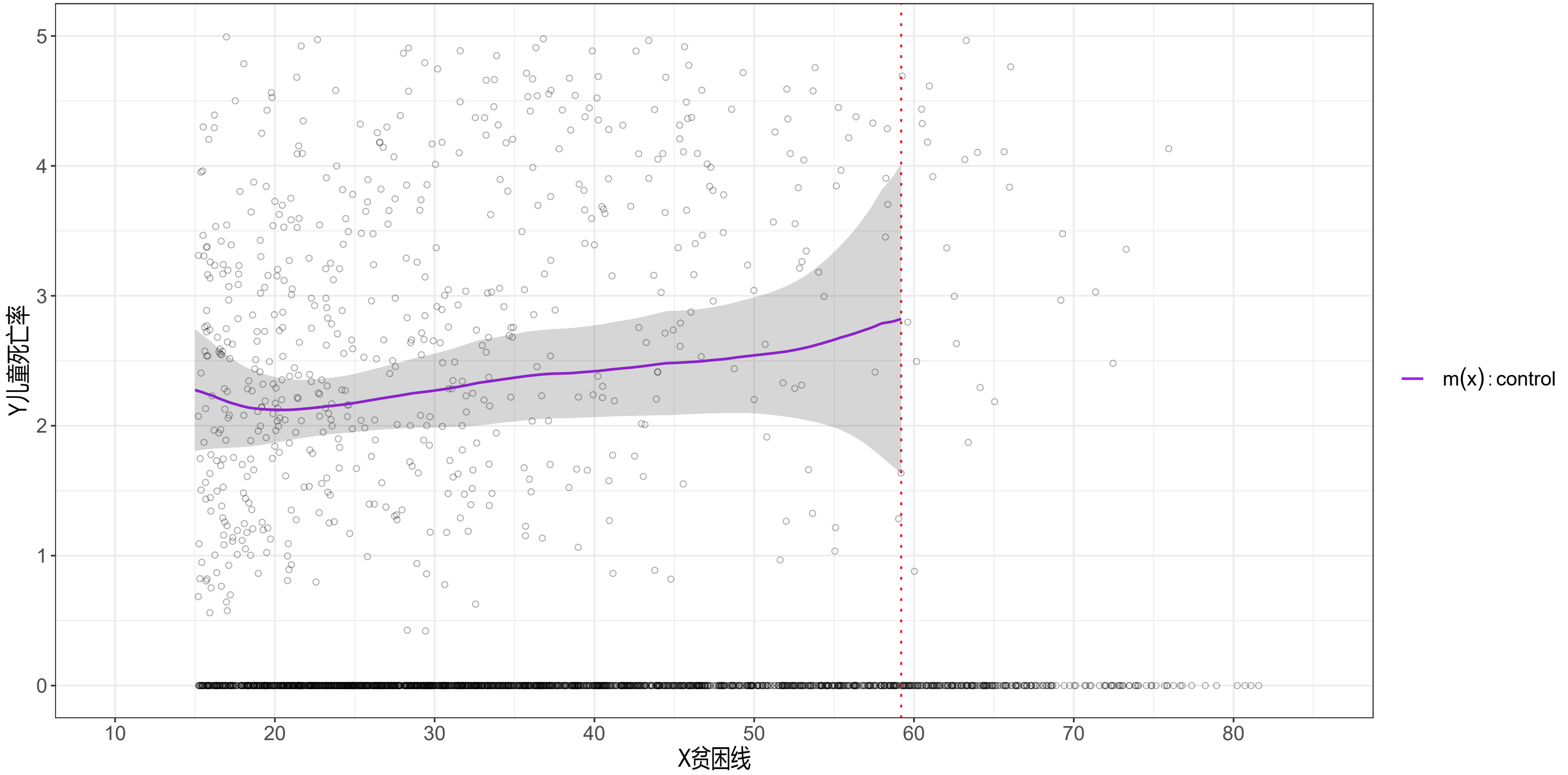
m(x)的样本方差和标准差估计结果

index	group	xg	mx	s	lwr	upr
1	control	15.0	2.2757	0.2392	1.8068	2.7445
2	control	15.2	2.2674	0.2335	1.8096	2.7251
3	control	15.4	2.2601	0.2281	1.8131	2.7071
4	control	15.6	2.2516	0.2222	1.8161	2.6871
5	control	15.8	2.2428	0.2163	1.8187	2.6668
6	control	16.0	2.2350	0.2109	1.8217	2.6482
7	control	16.2	2.2259	0.2048	1.8245	2.6273
8	control	16.4	2.2148	0.1989	1.8251	2.6046

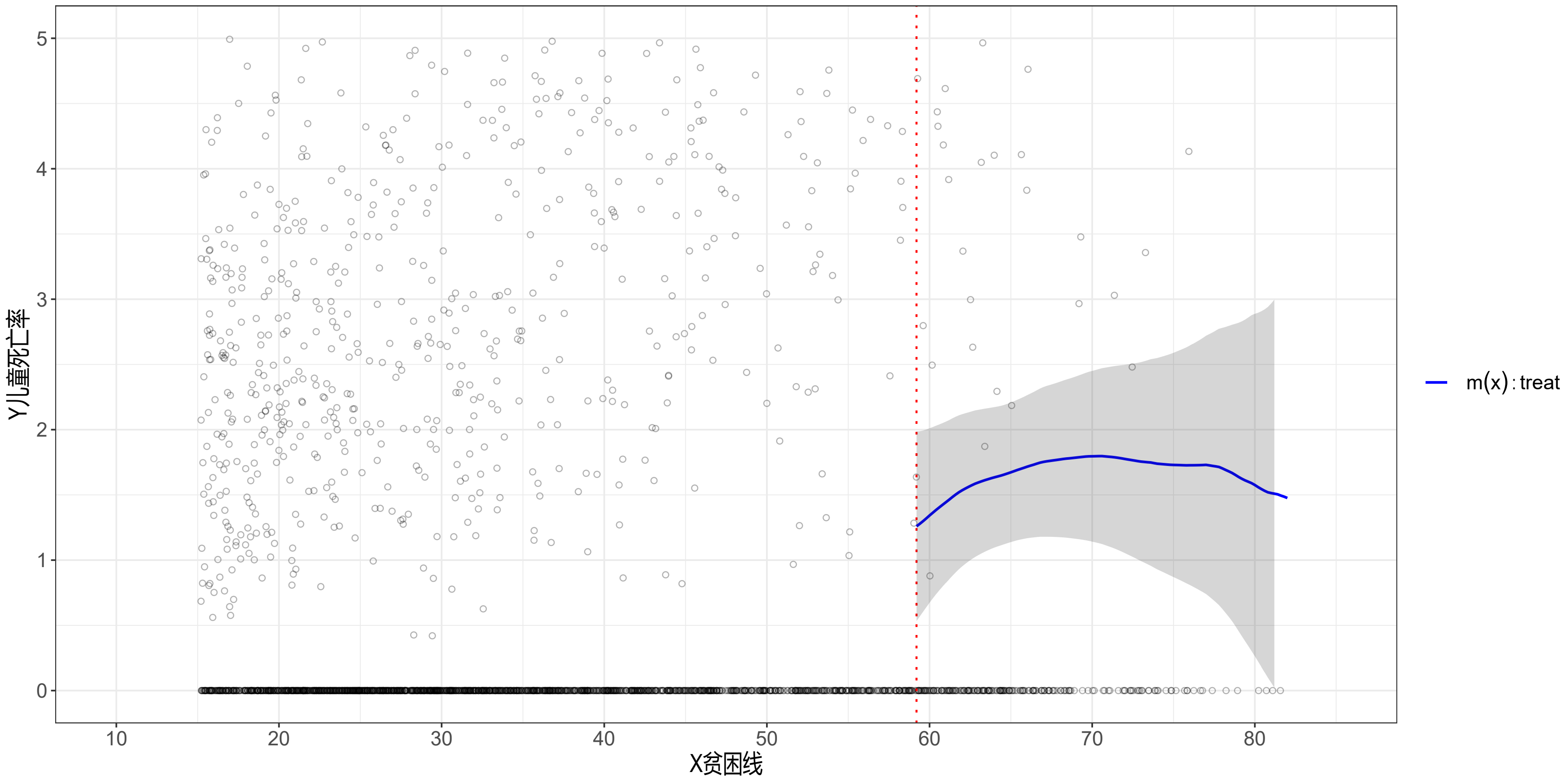
Showing 1 to 8 of 337 entries

Previous 1 2 3 4 5 … 43 Next

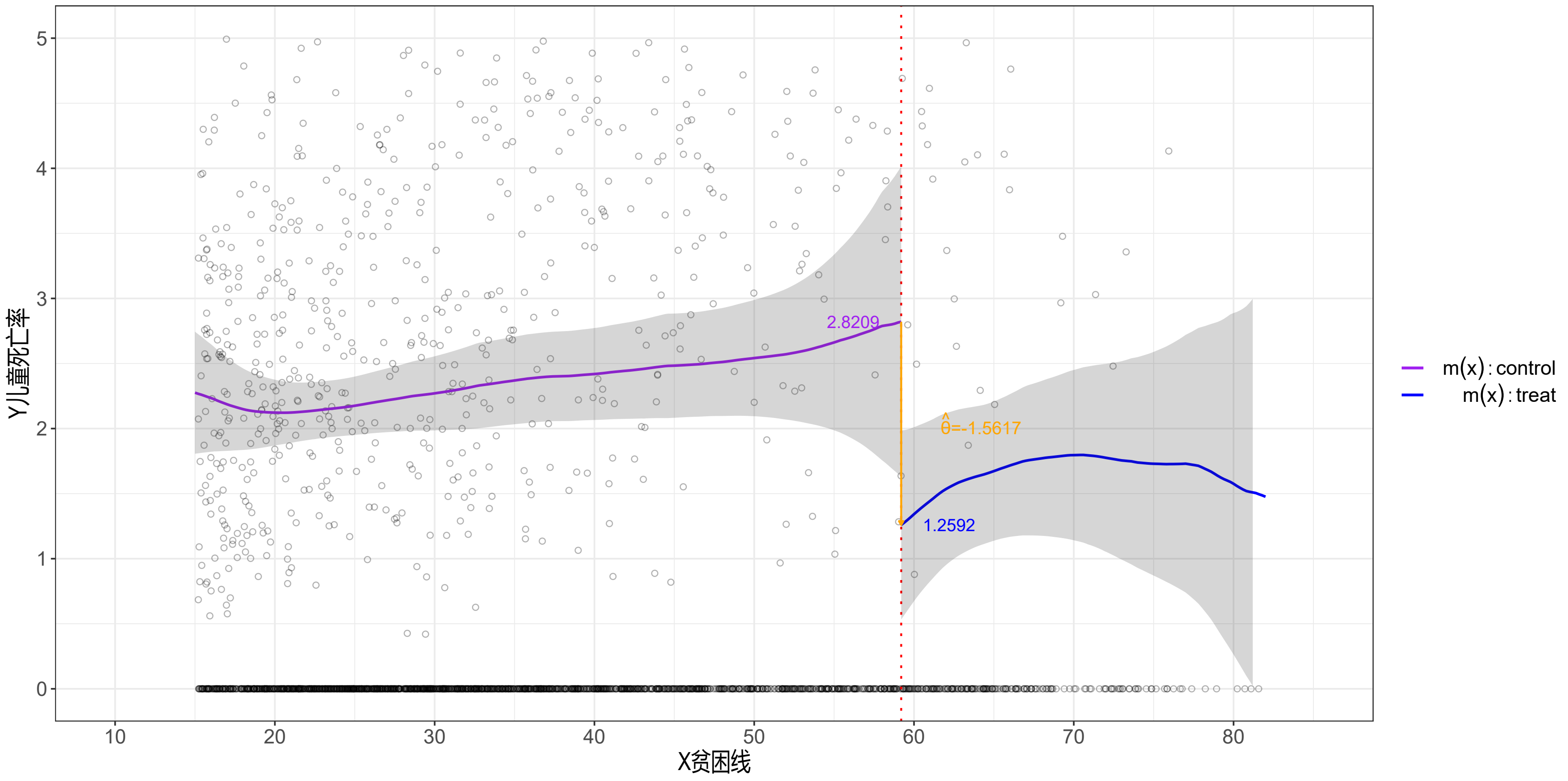
（死亡率案例）协变量RDD：断点左侧置信带（控制组）



(死亡率案例)协变量RDD：断点右侧置信带(处置组)



（死亡率案例）协变量RDD：断点两侧



（死亡率案例）协变量RDD：断点处置效应

- 根据**断点处置效应定理**，可以得到在断点 $x = c = 59.1984$处对总体平均处置效应 $\bar{\theta}$的样本估计结果 $\hat{\theta}$：

$$\begin{aligned}\hat{\theta} &= \left[\widehat{\boldsymbol{\beta}}_{\mathbf{1}}(c)\right]_1 - \left[\widehat{\boldsymbol{\beta}}_{\mathbf{0}}(c)\right]_1 \\ &= \hat{m}(c+) - \widehat{m}(c-) \\ &= 2.8209 - 1.2592 = -1.5617\end{aligned}$$

- 断点处置效应估计值为 $\hat{\theta} = -1.5617$。
  - 断点左边的条件期望(CEF)的估计值 $\widehat{m}(c-) = 2.8209$；
  - 断点右边的条件期望(CEF)的估计值 $\widehat{m}(c+) = 1.2592$；
- **结论**：援助项目的实施，减低了儿童死亡率，使得10万个孩子中约-1.5617个小孩免于遭受死亡。相比不实施项目援助，儿童死亡率由2.8209，下降到1.2592，降幅接近50%。

（死亡率案例）协变量RDD：估计误差及显著性检验

- 进一步地，估计系数 $\hat{\theta}$ 的**渐进方差**为两个方差协方差矩阵第一个对角元素之和：

$$
\begin{aligned}
\operatorname{Var}(\hat{\theta}) &= \left[\widehat{\boldsymbol{V}}_0\right]_{11} + \left[\widehat{\boldsymbol{V}}_1\right]_{11} \\
&= 0.3673 + 0.1417 = 0.5090 \\
se((\hat{\theta})) &= \sqrt{\operatorname{Var}(\hat{\theta})} = \sqrt{0.5090} = 0.7122
\end{aligned}
$$

- 因此RDD断点处置效应估计值 $\hat{\theta}$ 的**标准误**为 $se(\hat{\theta}) = 0.7122$；最后我们可以计算得到**RDD断点处置效应**对应的t统计量：$t^* = \frac{\hat{\theta}}{se(\hat{\theta})} = -2.19$，其概率值为 $p = 0.0283$.
- 综上，RDD结果表明援助项目降低了儿童死亡率，使得10万个孩子中约1.51个小孩免于遭受死亡。并且t统计量检验表明，援助项目在降低了儿童死亡率上具有统计显著性（置信度超过95%）。

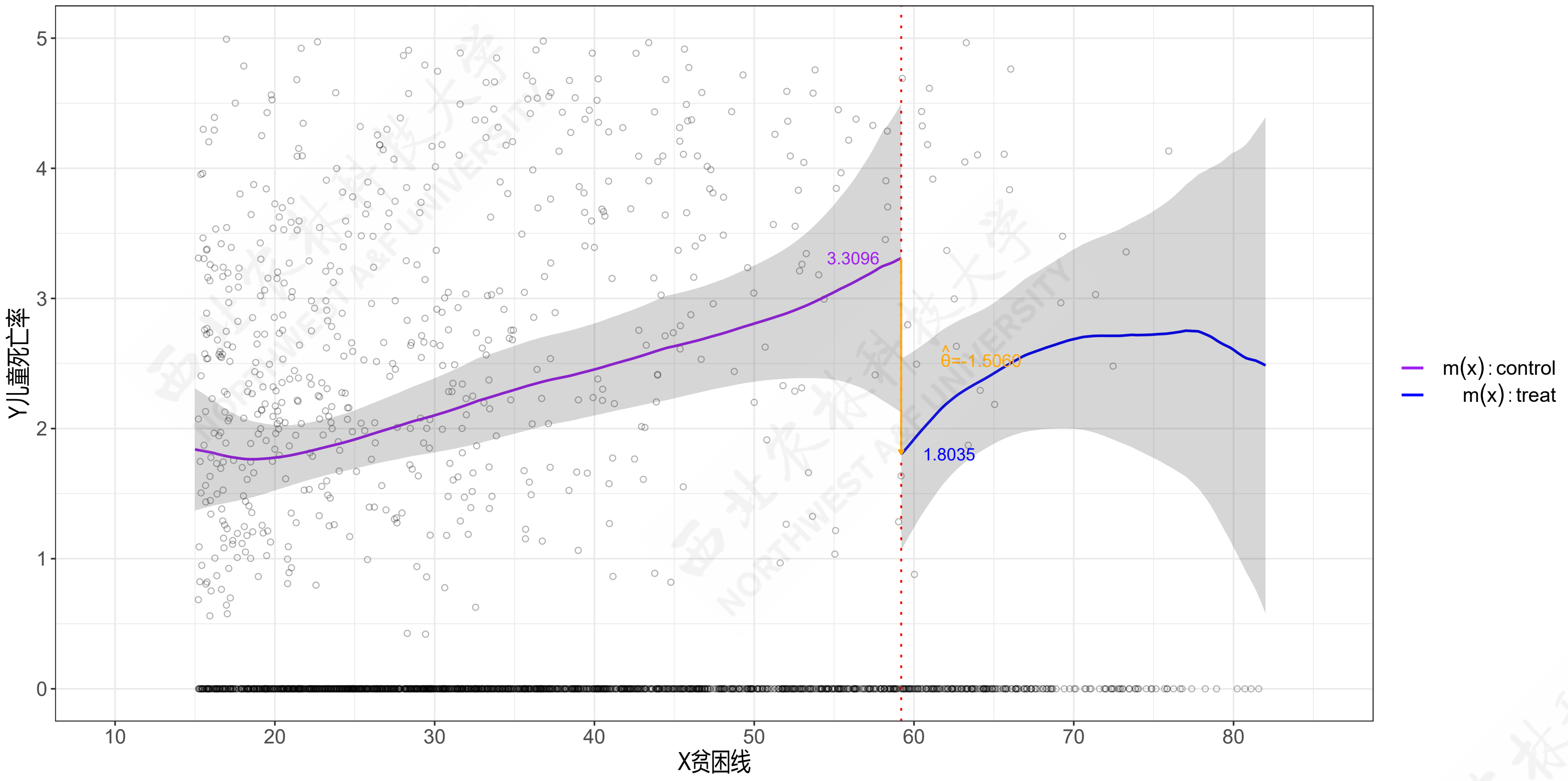
(死亡率案例)总结：系数和标准误比较

基准RDD和协变量RDD估计结果对比

pars	stats	baseline	covariate
theta	est	-1.5060	-1.5617
theta	se	0.7134	0.7122
black	est		0.0265
black	se		0.0073
urban	est		-0.0094
urban	se		0.0046

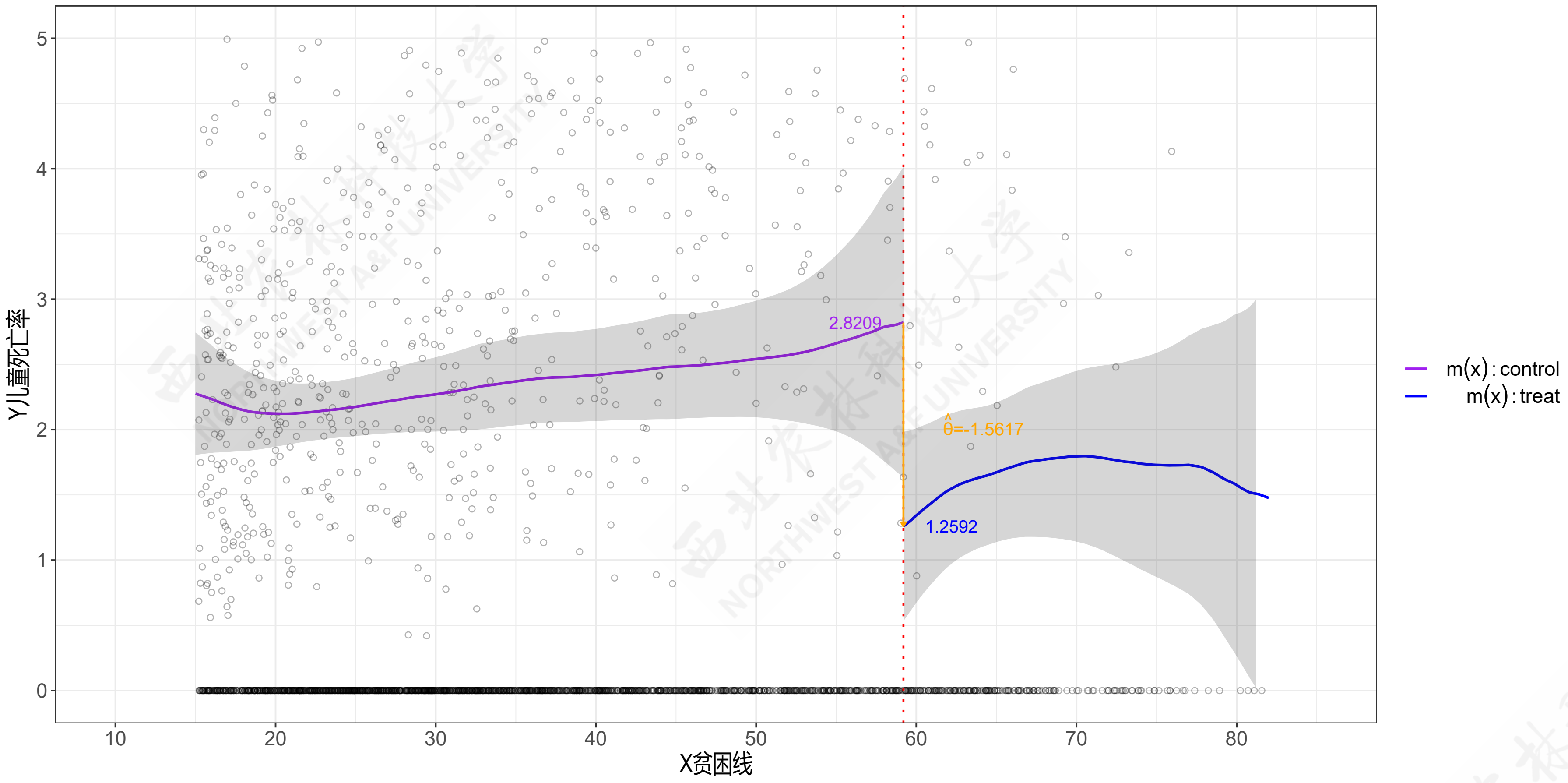
- 结论1：与基准RDD相比，两个协变量的引入没有明显改变断点处置效应估计值大小。
- 结论2：但是是否引入协变量，对CEF估计值 $\widehat{m}(x)$的影响较大。可以看到基准RDD更陡峭，而协变量RDD更平缓。（见后面附图对比）
- 结论3：考虑到两个协变量可以视作收入的**代理变量**，可以看到黑人人口比重负向影响儿童死亡率，而城镇人口比重正向影响儿童死亡率。

（死亡率案例）总结：CEF估计值图形比较1/2



基准RDD：局部线性回归及断点效应估计

（死亡率案例）总结：CEF估计值图形比较2/2



协变量RDD：局部线性回归及断点效应估计

3.RDD怎样高级进阶？

(How the Pros Do It?)

3.1 模糊RDD

3.2 拐点RDD (RKD)

3.3 多断点RDD

3.4 安慰剂检验

3.1 模糊RDD分析

模糊RDD（fuzzy regression discontinuity design,FRDD）：是指处置条件的条件分配概率在断点处是不连续的（跳跃的），但又不是从0直接跳跃到1的一种RDD分析情形。

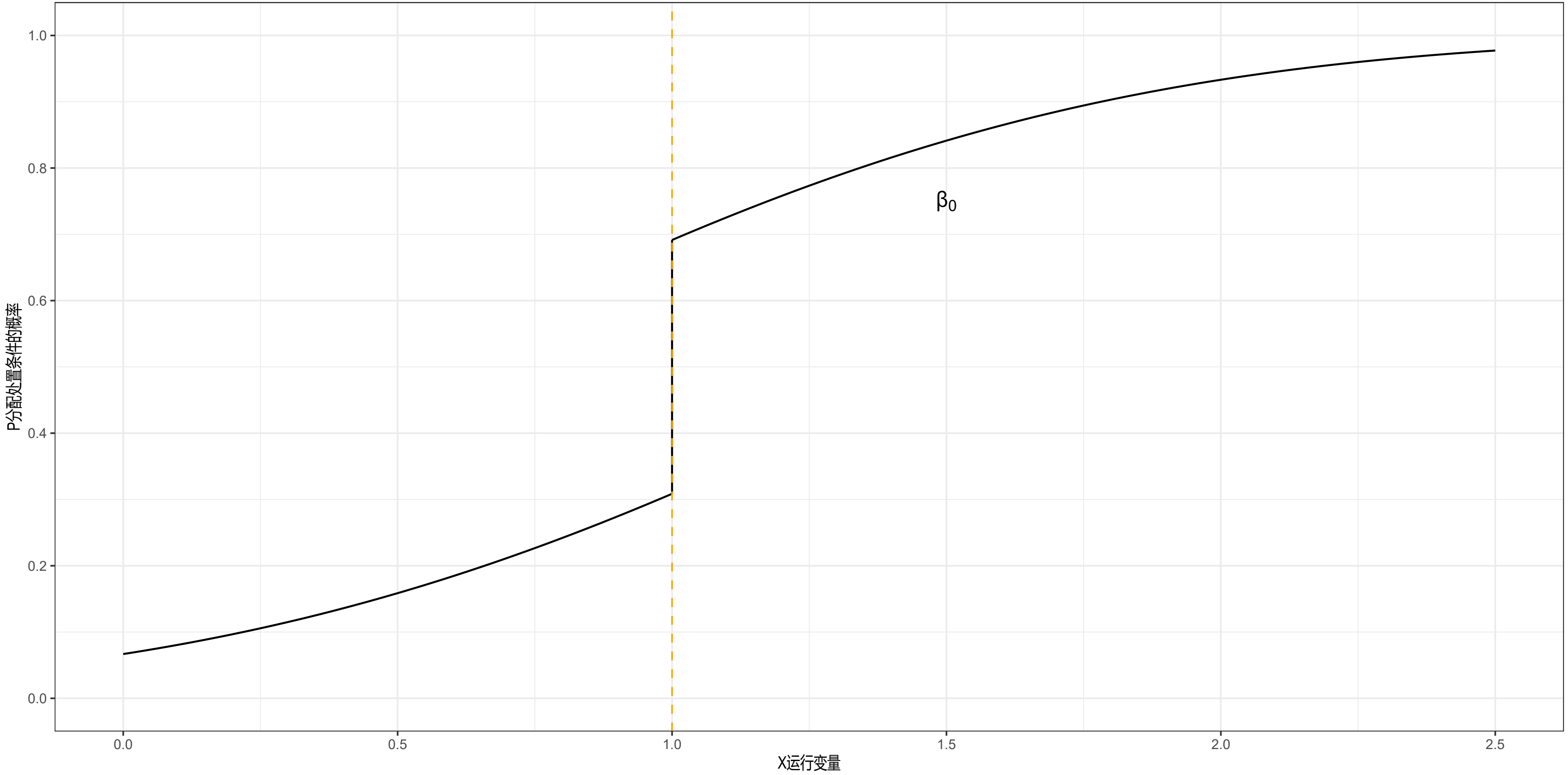
> **骤变RDD**中，在断点两边，处置条件的条件分配概率在断点处是跳跃的，而且是直接从0跳跃到1。

- 我们定义处置条件的**条件分配概率**为：

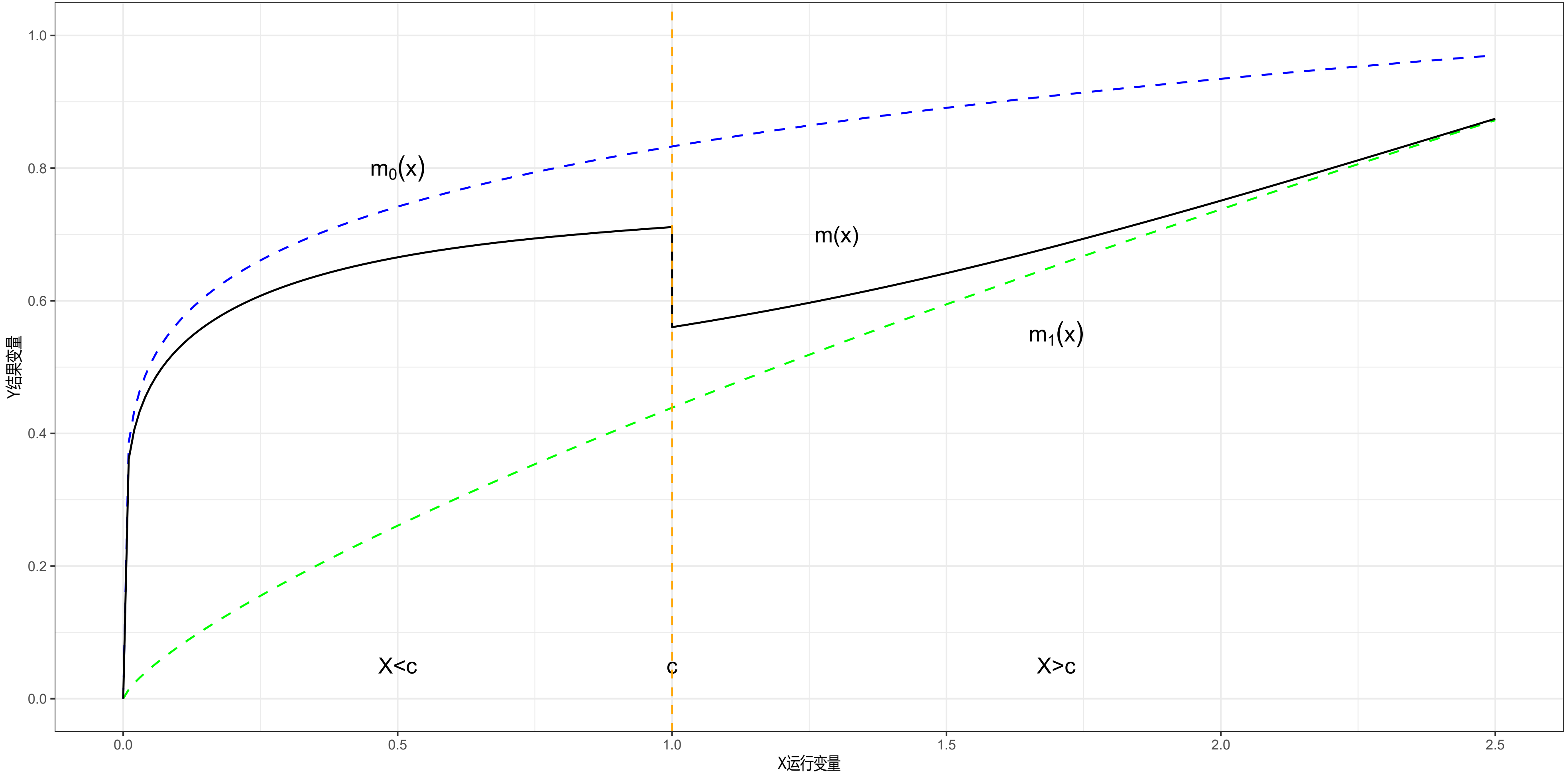
$$p(x) = \mathbb{P}[D = 1|X = x]$$

- 那么在断点 $x = c$左右两边的极限**条件分配概率**则分别定义为：$p(c-), p(c+)$。
- 因此，对于**模糊RDD**而言，则意味着：$p(c-) \neq p(c+)$

（示例）模糊RDD分析：处置水平的分配概率



（示例）模糊RDD分析：反事实与条件期望函数



定理： 模糊RDD下的断点处置效应ATE。

- 假定 $m_0(x)$和 $m_1(x)$在点 $x=c$处连续， $p(x)$在点 $x=c$处不连续，且 在断点附近处置变量 D与真实参数 $\theta|X$相互独立，则断点处置效应ATE为：

$$\bar{\theta} = \frac{m(c+) - m(c-)}{p(c+) - p(c-)}$$

此时，我们考虑如下的模型：

$$
\begin{aligned}
Y &= Y_0 1\{D=0\} + Y_1 1\{D=1\} \\
&= Y_0 + \theta 1\{D=1\}
\end{aligned}
$$

- 对两边在 $x=c$附近同时取期望：

$$m(x) = m_0(x) + \mathbb{E}[\theta 1\{D=1\} \mid X=x] = m_0(x) + \theta(x)p(x)$$

- 在 $x=c$处取极限，则有：

$$m(c+) = m_0(c) + \bar{\theta}p(c+); \qquad m(c-) = m_0(c) + \bar{\theta}p(c-)$$

- 最后可以证明：

$$m(c+) - m(c-) = \bar{\theta}p(c+) - \bar{\theta}p(c-) \Rightarrow \qquad \bar{\theta} = \frac{m(c+) - m(c-)}{p(c+) - p(c-)}$$

3.1 模糊RDD分析：断点ATE的估计（精简表达法）

- 对于分母部分，我们可以使用前面介绍的局部线性回归（LLR）对总体参数 $m(c+)-m(c-)$进行估计，得到断点 $x=c$附近的两边的估计值：

$$\widehat{m}(c+)-\widehat{m}(c-)\equiv[\widehat{\beta}_1(c)]_1-[\widehat{\beta}_0(c)]_1$$

- 同理，分子部分我们也同样使用局部线性回归（LLR）对总体参数 $p(c+)-p(c-)$进行估计，得到断点 $x=c$附近的估计值：

$$\hat{p}(c+)-\hat{p}(c-)\equiv[\widehat{\alpha}_1(c)]_1-[\widehat{\alpha}_0(c)]_1$$

- 最终，我们可以得到断点ATE的估计值为：

$$\hat{\theta}=\frac{\widehat{m}(c+)-\widehat{m}(c-)}{\hat{p}(c+)-\hat{p}(c-)}$$

对于模糊RDD断点ATE的估计：

$$\hat{\theta} = \frac{\widehat{m}(c+) - \widehat{m}(c-)}{\hat{p}(c+) - \hat{p}(c-)}$$

- 上式实际上是两类断点估计的比率值。而且当 $\hat{p}(c+) - \hat{p}(c-) = 1$时，以上估计式即为**骤变RDD**的估计情形！
- 上述估计值的计算总共会需要进行4次**局部线性回归**，是不是需要都使用相同的**谱宽**（bandwidth），或者断点两侧是否要采用不同数量的**箱组**（bins），可以进行多次尝试！

事实上，上述模糊RDD断点ATE的估计可以使用**工具变量法**（IV）等价得到。

- 简单地，可以把 D视作为 X的工具变量，然后把 Y对它们二者进行**局部加权工具变量估计**（Locally weighted IV estimation），从而得到断点ATE估计值 $\hat{\theta}$。
- 断点处置效应能否被识别，有赖于在断点附近处概率 $p(x)$的跳跃性程度。如果跳跃不大，那么就会带来**弱工具变量**问题（Weak Instruments Problem）。
- ATE估计的标准误，其计算过程类似于IV回归法。我们先把估计量 $\widehat{m}(c+) - \widehat{m}(c-)$的标准误定义为 $s(\hat{\theta})$，那么就可以使用下式计算得到ATE估计 $\hat{\theta}$的标准误：

$$s(ate) = \frac{s(\hat{\theta})}{|\hat{p}(c+) - \hat{p}(c-)|}$$

回顾与思考：

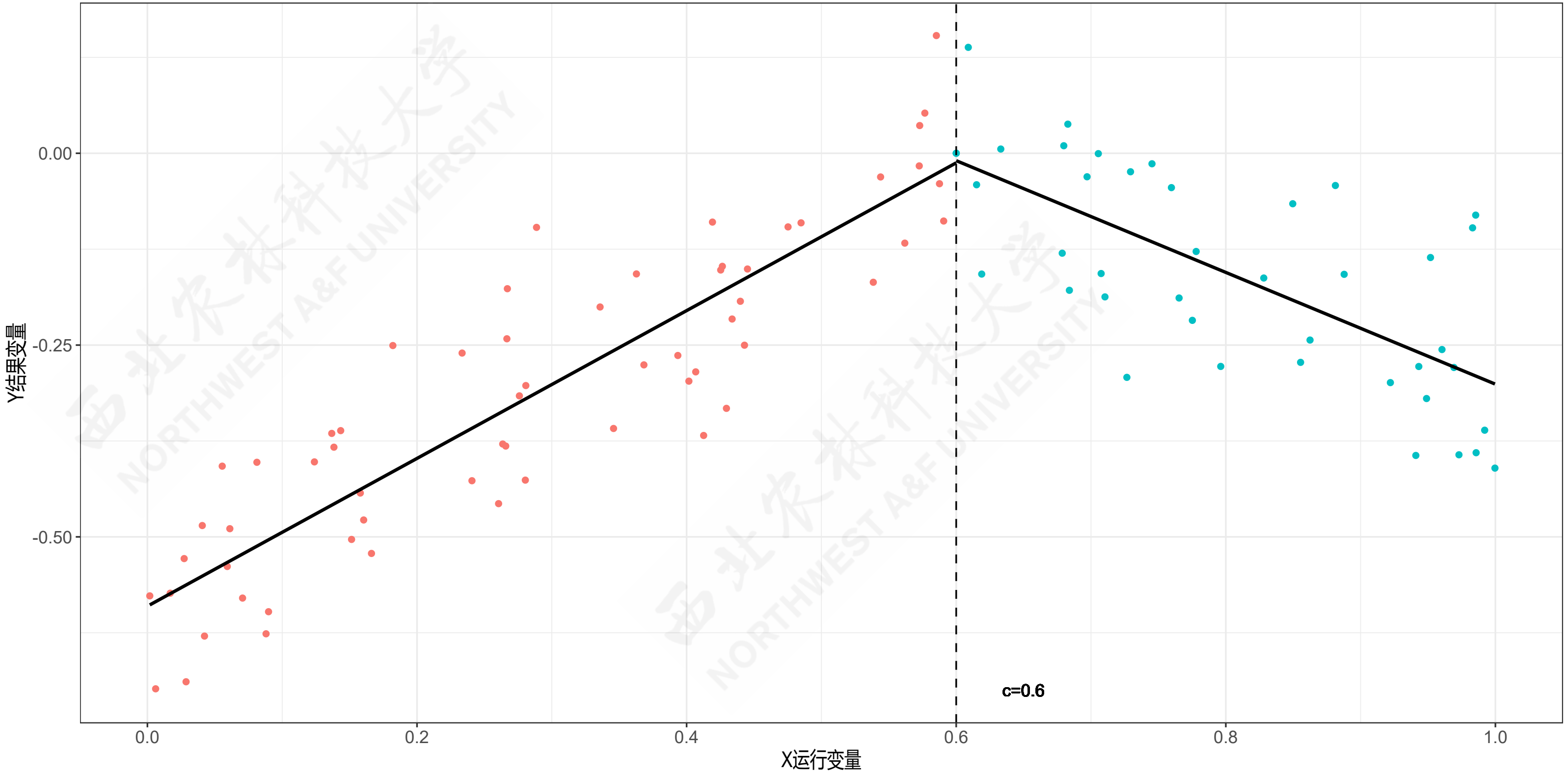
- 断点回归设计（RDD）探讨的是**结果变量** Y的条件期望值（均值）$\mathbb{E}(Y|X=x) \equiv m(x)$在**断点**附近是否存在跳跃性（jump）的不连续。
- 那么，我们能不能分析除此之外，其他对象的跳跃性或不对称性呢？

> 例如结果变量的**标准差**是否跳跃？中位数是否跳跃？或者**箱组内**（bins）局部回归的判定系数 R^2是否跳跃？

拐点回归设计（Regression Kink Design, RKD）：是探讨**结果变量** Y 对**运行变量** X 的**斜率**（slope）是否存在显著改变（change）的一种处置效应回归分析设计框架。

- 处置条件只是改变了**斜率**，但是并没有引起结果变量的跳跃，也即结果变量在**拐点处**（kink point）还是连续的！
- 在有些情形下，研究者还可以关注**处置水平**D对**运行变量**X的变化率的**拐点效应**。

（示例）拐点回归RKD分析：基于模拟数据



结果变量对运行变量的变化率（斜率）具有拐点效应

- 讨**结果变量** Y对**运行变量** X的变化率（**斜率**）具有**拐点效应**：

案例1： 政府择机扶持科技公司。

- 政府在特定时间点开始，决定大量关注并投资科技公司。此时公司雇员人数为**结果变量**Y，时间观测为**运行变量**X，政府是否决定大量投资则为**处置变量**D。
- 这种情况下，公司雇员总人数Y可能并不会在**拐点** $x=c$处立刻跳跃（不连续），但是我们预期在**拐点后**的公司的雇员增长率（Y对X的斜率）会比之前会有一个明显变化！

3.2 拐点回归RKD分析：应用案例2

- **结果变量** Y**对运行变量** X的变化率（**斜率**）具有**拐点效应**：

案例2：失业保险政策(Card, Lee, Pei, et al., 2015)。

- 案例背景为澳大利亚。公民的失业保险补贴水平大概为其正常工作收入的55%，并且有一个补贴最高上限值。因此，失业保险政策设计下，公民的正常工作收入会正向地影响失业保险补贴水平。工作收入越高，补贴会越多，直到达到一个补贴上限值。
- 此时，我们定义：公民的正常工作收入为**运行变量**X，公民是否能获得失业补贴为**处置变量**D。一个公民如果失业，把他从失业那一刻算起，直到他找到一份新工作，期间他所愿意的**等待时长**定义为**结果变量**Y，

- **结果变量** Y 对**运行变量** X 的变化率（**斜率**）具有**拐点效应**：

案例2(续)：失业保险政策(Card, Lee, Pei, et al., 2015)。

- 这种情况下，政策如果给予更高的补贴水平，那么我们可以预期公民的就业等待时长Y可能会更长！因此，我们也可以预期，在达到最高补贴水平（拐点c）之前，公民正常的工作收入X越高，那么他的就业等待时长Y也会更长！
- 显然，在拐点之后（最高补贴之后），**等待时长**Y对正常**工作收入**X的比率应该会变得比拐点之前更加平缓（斜率更小）！——也即出现了Y对X的斜率具有**拐点效应**！同时，我们还可以预期到就业**等待时长**Y并不会在**拐点** $x = c$ 处立刻跳跃（不连续）！

3.2 拐点回归RKD分析：应用案例3

- **处置变量**D对**运行变量** X的变化率（**斜率**）具有**拐点效应**：

案例3：妇女育儿支持政策(Bana, Bedard, and Rossin-Slater, 2020)。

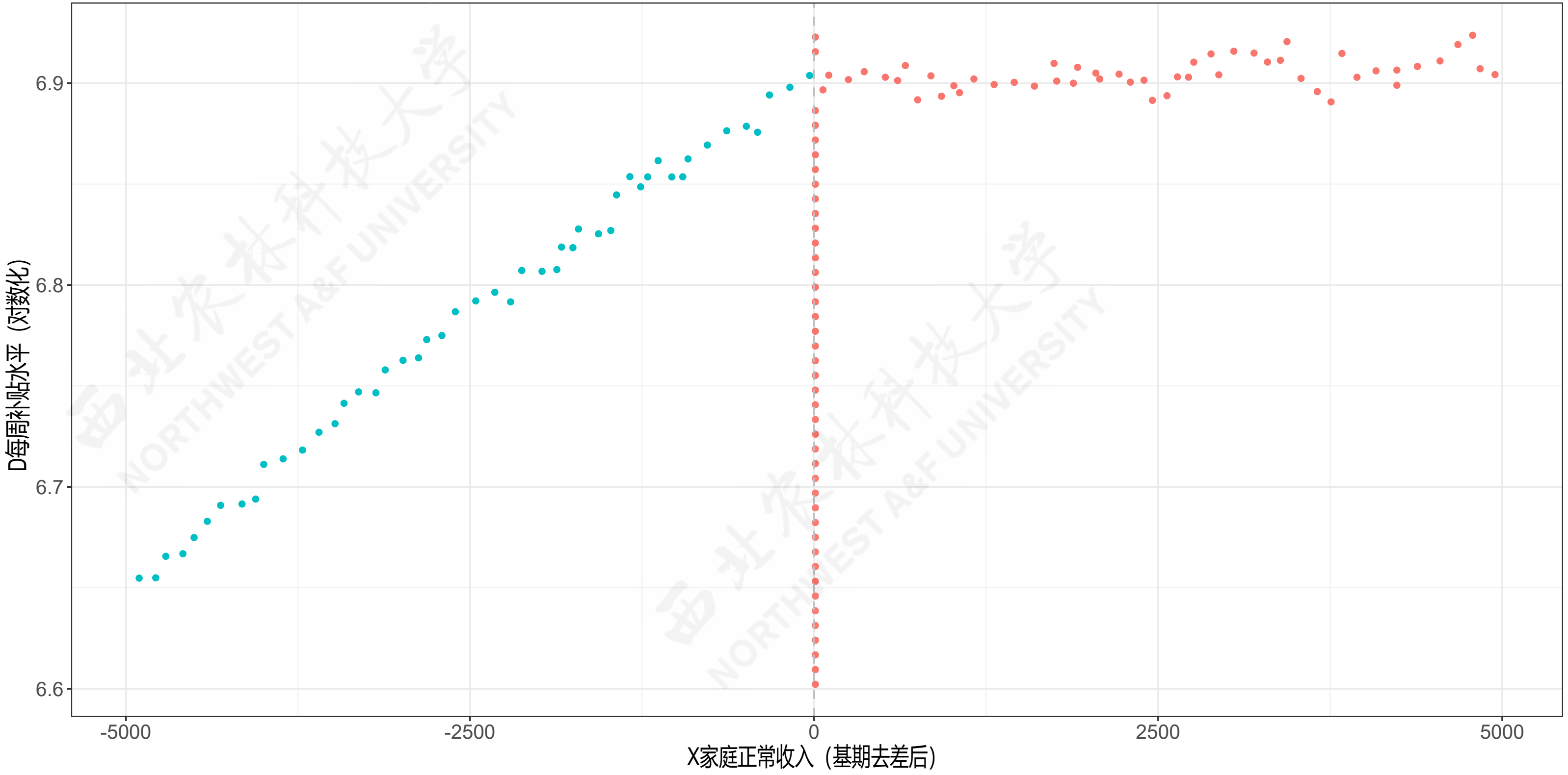
- 案例背景为美国加利福尼亚州。政府制定了一项**妇女育儿家庭支持**政策（paid family leave）。对于符合条件的家庭，加州政府根据家庭正常工作收入，将补贴其家庭收入的55%直至一个最高最高上限值。
- 因此，**妇女育儿家庭支持**政策设计下，家庭的正常工作收入会正向地影响补贴水平。工作收入越高，补贴会越多，直到达到一个补贴上限值。

- **处置变量**D对**运行变量** X的变化率（**斜率**）具有**拐点效应**：

案例3（续）：妇女育儿支持政策(Bana, Bedard, and Rossin-Slater, 2020)。

- 此时，我们定义：家庭的正常工作收入为**运行变量**X，家庭获得政策补贴水平为**处置变量**D。妇女获得的**育儿假时长**为结果变量Y。
- 显然，在拐点之前（最高补贴之前），家庭的**政策补贴水平**D越高，也意味着家庭正常**工作收入**X越高；而在拐点之后（最高补贴之后），家庭的**政策补贴水平**D会保持不变——也即意味着D对X的斜率为0！
- 因此，**处置变量**D对**运行变量** X的变化率（**斜率**）具有**拐点效应**：

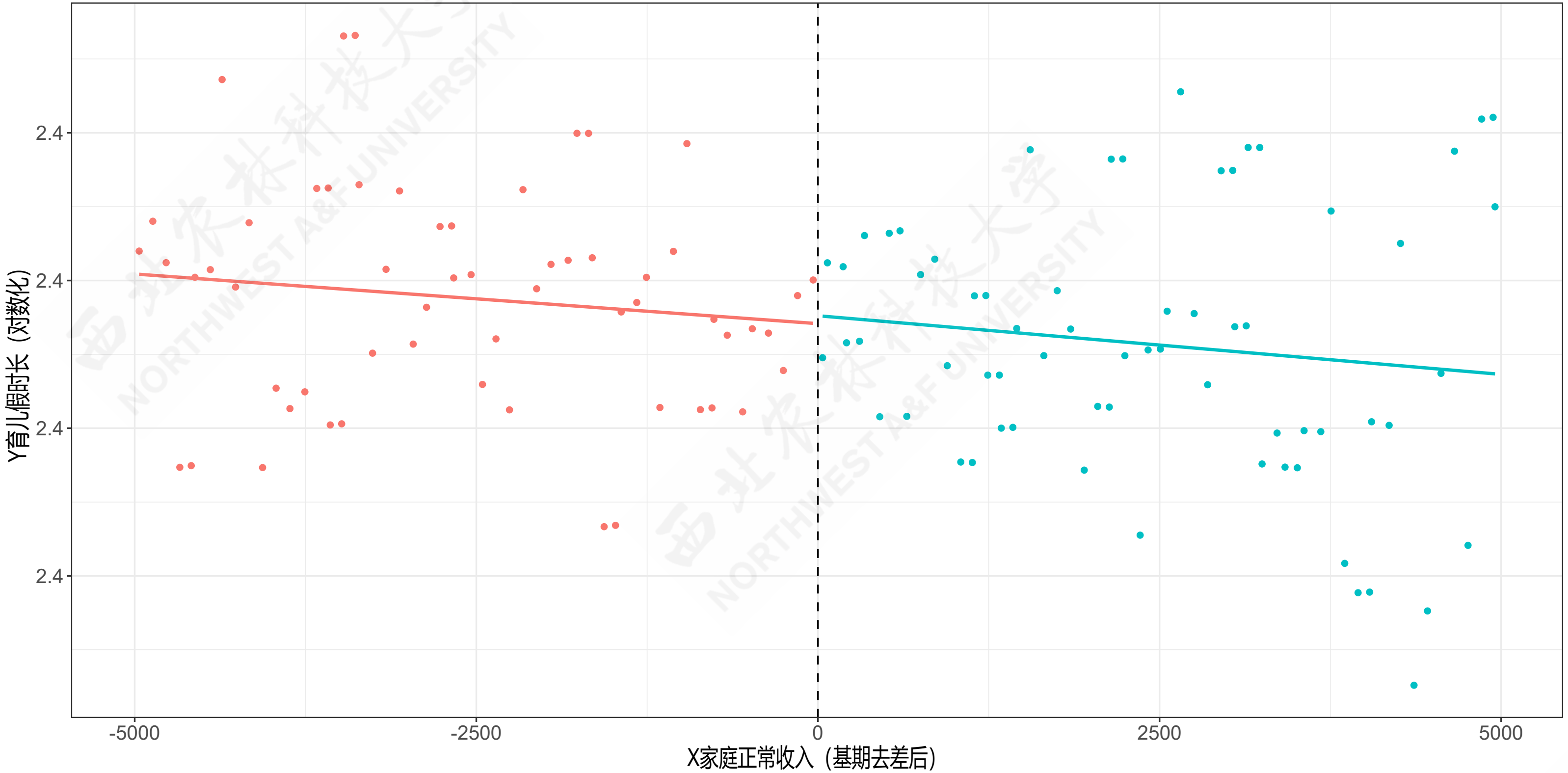
（育儿支持案例）处置变量对运行变量的拐点效应



处置变量对运行变量的变化率（斜率）具有拐点效应

（育儿支持案例）结果变量对运行变量的拐点效应1/2

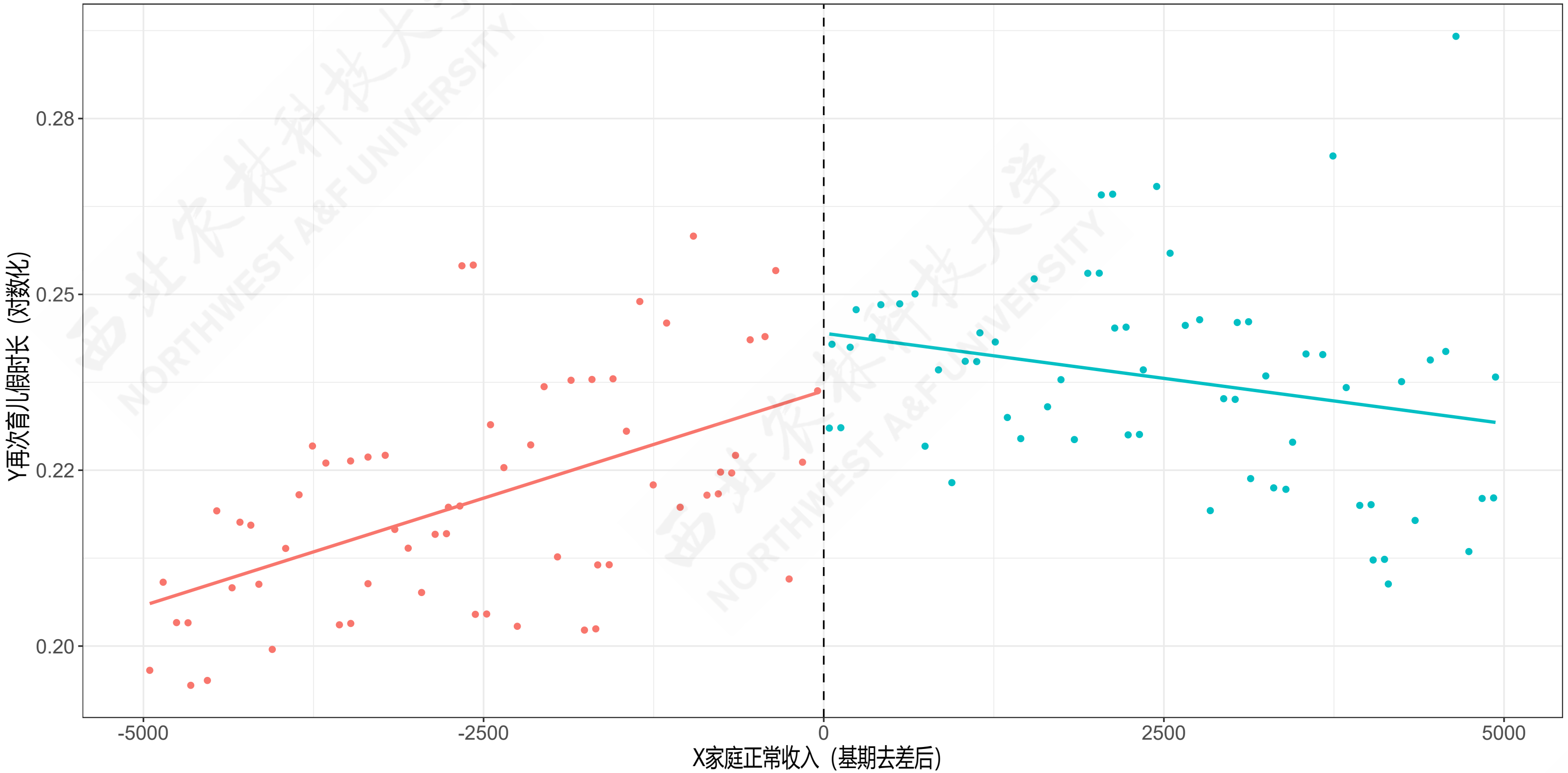
- 对于不打算再要孩子的家庭



结果变量对运行变量的变化率（斜率）没有拐点效应

（育儿支持案例）结果变量对运行变量的拐点效应2/2

- 对于打算再要孩子的家庭，不仅仅只是**拐点**效应，而是**断点**效应



结果变量对运行变量的变化率（斜率）具有拐点效应

3.2 拐点回归RKD分析：拐点效应ATE估计

总体而言，关于拐点效应ATE的估计方法，与之前的RDD估计过程基本类似：

- 确定核函数
- 选择谱宽
- 局部线性回归LLR或局部多项式回归LPR
- 安慰剂效应检验

编程提示：

- 对于`R`或`stata`用户而言，可以使用分析包`rdrobust`
- 拐点回归RKD估计时，仅需要设定参数`deriv=1`即可

3.3 多断点RDD：引子

回顾与思考：

- 截止目前为止，我们已经接触了骤变断点（SRDD）、模糊断点（FRDD）、斜率改变拐点（RKD）
- 那么，我们能不能考虑政策存在多个断点（或拐点）的情形呢？

（示例）多断点RDD应用案例

应用案例：

- 在**育儿支持案例**中，能拿到**最高**补贴支持的**家庭季度收入**（x=c），也是随着年度变化而进行调整的。例如在2004年这个收入水平划定为25000美元，而到了2005年则被划定在20000美元。
- 一国猪肉储备投放政策中，会考虑根据猪粮比价（X）变动，分别设定红色、橙色、蓝色和绿色预警窗口（多个断点），来决定如何干预市场（如生产收储或市场投放，以及数量多少等）。

（示例）多断点RDD应用案例

应用案例（续）：

- 高考招生政策中，会考虑根据不同招生类型（如普通招录生、师范特招生、体育特招生等），设定不同的高考成绩录取线（多个断点）。而且普通高考招录的录取线，对于不同的省份也是不同的（例如，某省的某个高校在全国各省的招录录取线就会各不相同——一般本省录取线会更低）。
- 在多党派多家的政党选举中，某个政党能否竞选胜出执政，有的年度可能需要50.1%的投票率，但是有的年份可能只需要42.7%就能胜选。

3.3 多断点RDD：问题描述

- 很多情况下，**断点**（或拐点）本身就是政策指定者最为关注的议题
- 一些情形下，**多个断点**的政策设计具有很强的现实意义或价值。

多断点分析（Cutoffs cut off Analysis）：在运行变量X上，存在多个**断点**，断点值的划定，往往基于不同群体、不同地区，或不同时间段上的运行变量取值。

回顾与启发：

- 在经典的RDD估计中，**断点平均处置效应**（ATE）是把断点附近的处置效应做了简单平均（正如其名！）。
- 但是，在**多断点**的RDD情形下，事情变得复杂（不同的断点针对不同的群体），因此不能再直接、粗暴地进行**简单平均**——我们必须考虑到不同的群体区块！

编程提示：

- 对于`R`或`stata`用户而言，可以使用分析包`rdmulti`
- 多断点RDD估计时，可以使用函数`rdmulti::rdmc()`进行分析

3.4 安慰剂检验：原理

安慰剂检验（Placebo Tests）：RDD分析的前提假设是，处置变量的作用是“干净的”、没有后门的（no back doors）。如果不使用结果变量Y，而是使用**协变量**作为“结果变量”进行正常的RDD估计，如果也表现出与之前同样显著的断点处置效应ATE，那么我们就要质疑我们的RDD设计框架了。

- 选择合理的协变量，将其视作为“结果变量”
- 进行常规的RDD分析流程
- 比较结果并得出检验结论。

> 理论上，上述操作不应该得到——**“显著存在断点处置效应”**——的结论！

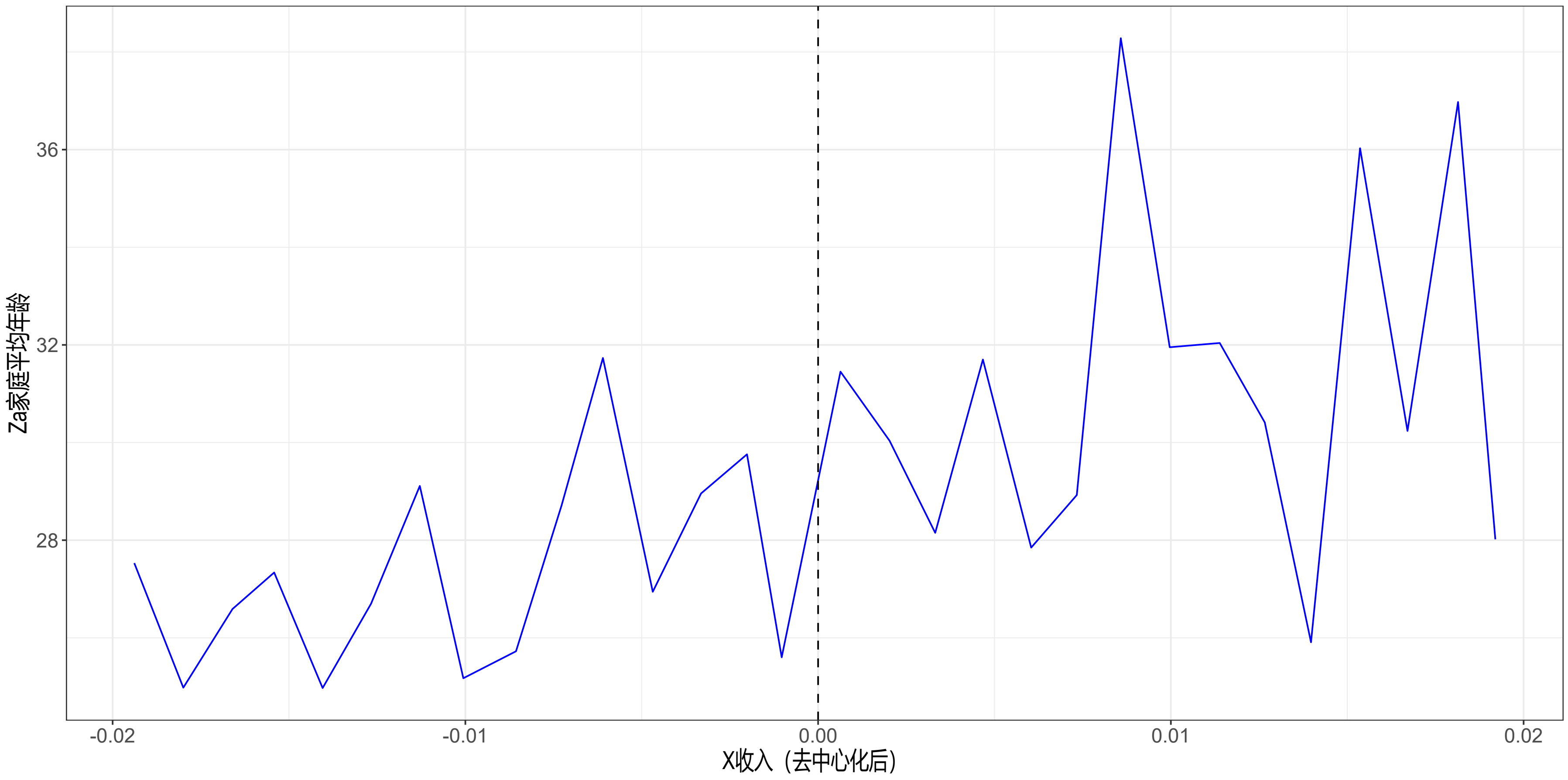
（政府转移支付案例）背景

政府转移支付案例：

- (Manacorda, Miguel, and Vigorito, 2011)分析了乌拉圭的一个大型扶贫项目，该项目削减了很大一部分贫困人口。论文关注的话题是：获得政府转移支付资金是否会让人们更有可能支持新成立的政府？
- 研究人员对一群接近收入临界值的人进行了调查，看看他们之后对政府的支持程度。收入低于临界值的人比收入高于临界值的人支持政府要更多吗？

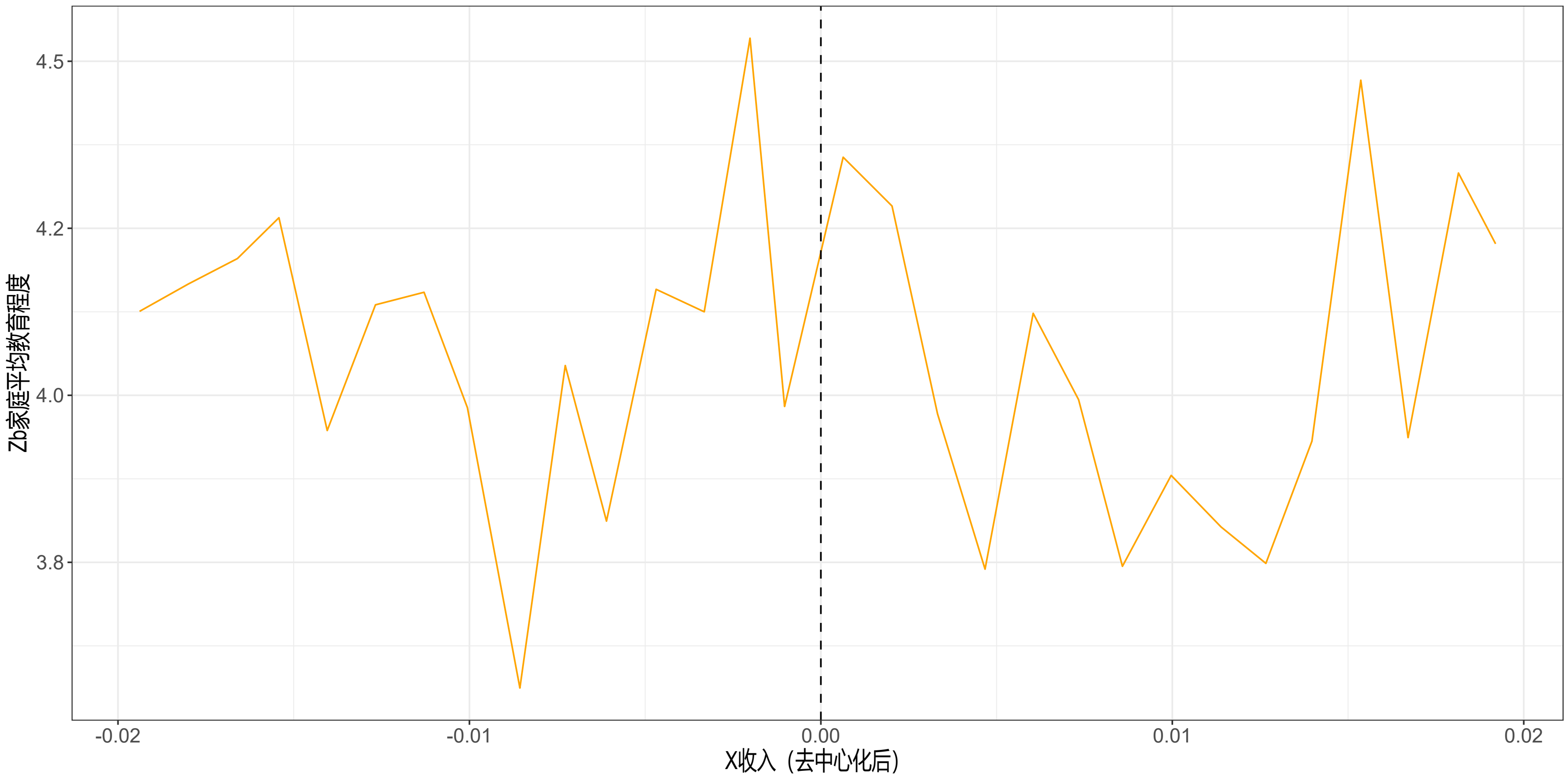
（政府转移支付案例）年龄协变量下的安慰剂检验

- 以协变量年龄作为结果变量进行安慰剂检验



（政府转移支付案例）教育协变量下的安慰剂检验

- 以协变量教育年数作为结果变量进行安慰剂检验



本章参考文献

Arai, Y. and H. Ichimura (2018). "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator". In: *Quantitative Economics* 9.1, pp. 441-482.

Bana, S. H., K. Bedard, and M. Rossin-Slater (2020). "The Impacts of Paid Family Leave Benefits: Regression Kink Evidence from California Administrative Data". In: *Journal of Policy Analysis and Management* 39.4, pp. 888-929.

Card, D., D. S. Lee, Z. Pei, et al. (2015). "Inference on Causal Effects in a Generalized Regression Kink Design". In: *Econometrica* 83.6, pp. 2453-2483.

Cattaneo, M. D. and R. Titiunik (2021). "Regression Discontinuity Designs" , p. 48.

Fan, J., I. Gijbels, T. Hu, et al. (1996). "A Study of Variable Bandwidth Selection for Local Polynomial Regression". In: *Statistica Sinica*, pp. 113-127.

参考文献（References）：2/3

Hausman, C. and D. S. Rapson (2018). "Regression Discontinuity in Time: Considerations for Empirical Applications". In: *Annual Review of Resource Economics* 10.1, pp. 533-552. DOI: 10.1146/annurev-resource-121517-033306.

Imbens, G. and K. Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". In: *The Review of economic studies* 79.3, pp. 933-959.

Ludwig, J. and D. L. Miller (2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design". In: *The Quarterly journal of economics* 122.1, pp. 159-208.

Manacorda, M., E. Miguel, and A. Vigorito (2011). "Government Transfers and Political Support". In: *American Economic Journal: Applied Economics* 3.3, pp. 1-28.

Robinson, P. M. (1988). "Root-N-consistent Semiparametric Regression". In: *Econometrica: Journal of the Econometric Society*, pp. 931-954.

参考文献（References）：3/3

Thistlethwaite, D. L. and D. T. Campbell (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." In: *Journal of Educational psychology* 51.6, p. 309.

本章结束