第12章： 断点回归

邱嘉平

大纲

1. 断点回归的直观理解
2. 断点回归的数据要求
3. RDD 的估计步骤
4. RRD运用实例 (详细过程见书）

断点回归的直观理解

例：政府针对低收入人群的医疗福利政策

- 政府根据病人的收入情况给每个人评分X_i，分数越高代表收入越高。同时病人的健康状况用健康指数Y_i衡量，指数越高代表越健康。

  - 接受治疗的平均潜在健康状况$Y_i(1)$和收入关系的函数为

$$E(Y_i(1)|X_i) = f(X_i)$$

  - 未接受治疗的平均潜在健康状况$Y_i(0)$和收入关系的函数为

$$E(Y_i(0)|X_i) = g(X_i)$$

  - 给定收入水平，病人平均治疗效果为

$$\tau(X_i) = E(Y_i(1)|X_i) - E(Y_i(0)|X_i) = f(X_i) - g(X_i)$$

  - 对于不同收入水平的平均治疗效果按不同收入的人数比率$p(x)$取平均值，总体平均治疗效果为

$$ATE = \sum_{x=20}^{80} p(x)\tau(x)$$

例：政府针对低收入人群的医疗福利政策

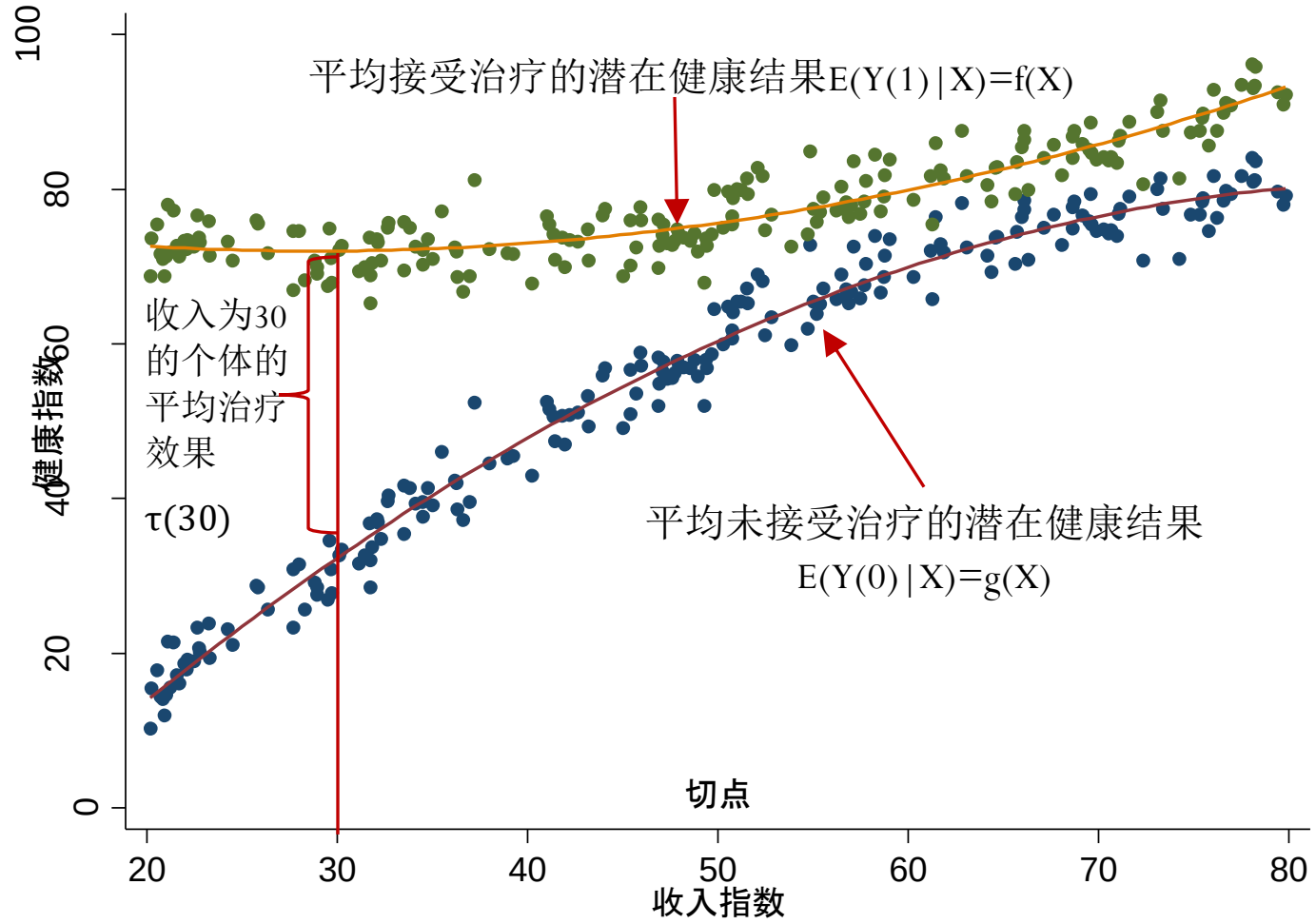

图12.1 同时观测到每个病人接受治疗和没接受治疗下的健康情况

例：政府针对低收入人群的医疗福利政策

- 问题：对于同一个病人，不可能同时观测到他接受治疗和没接受治疗两种情况下的健康状况。
- 假设：实际情况中，政府规定只对收入水平低于或等于50的病人提供这项治疗，而收入高于50的病人未能接受这项治疗，那么收入50就成为病人是否接受治疗的划分值，也称为**断点**c。
- 对于收入低于或等于50的人，我们观测到他们接受治疗的潜在健康指数，对于收入高于50的人，同样观测到他们没有接受治疗的潜在健康指数。那么平均观测值结果如下：

$$E(Y_i|X_i) = \begin{cases} E(Y_i(0)|X_i), X_i > 50 \\ E(Y_i(1)|X_i), X_i \leq 50 \end{cases}$$

例：政府针对低收入人群的医疗福利政策

- $\tau(x = 50) =$
  $E(Y_i(1)|X_i = 50) - (Y_i(0)|X_i = 50)$
  $\approx E(Y_i|X_i = 49.9) - (Y_i|X_i = 50.1)$

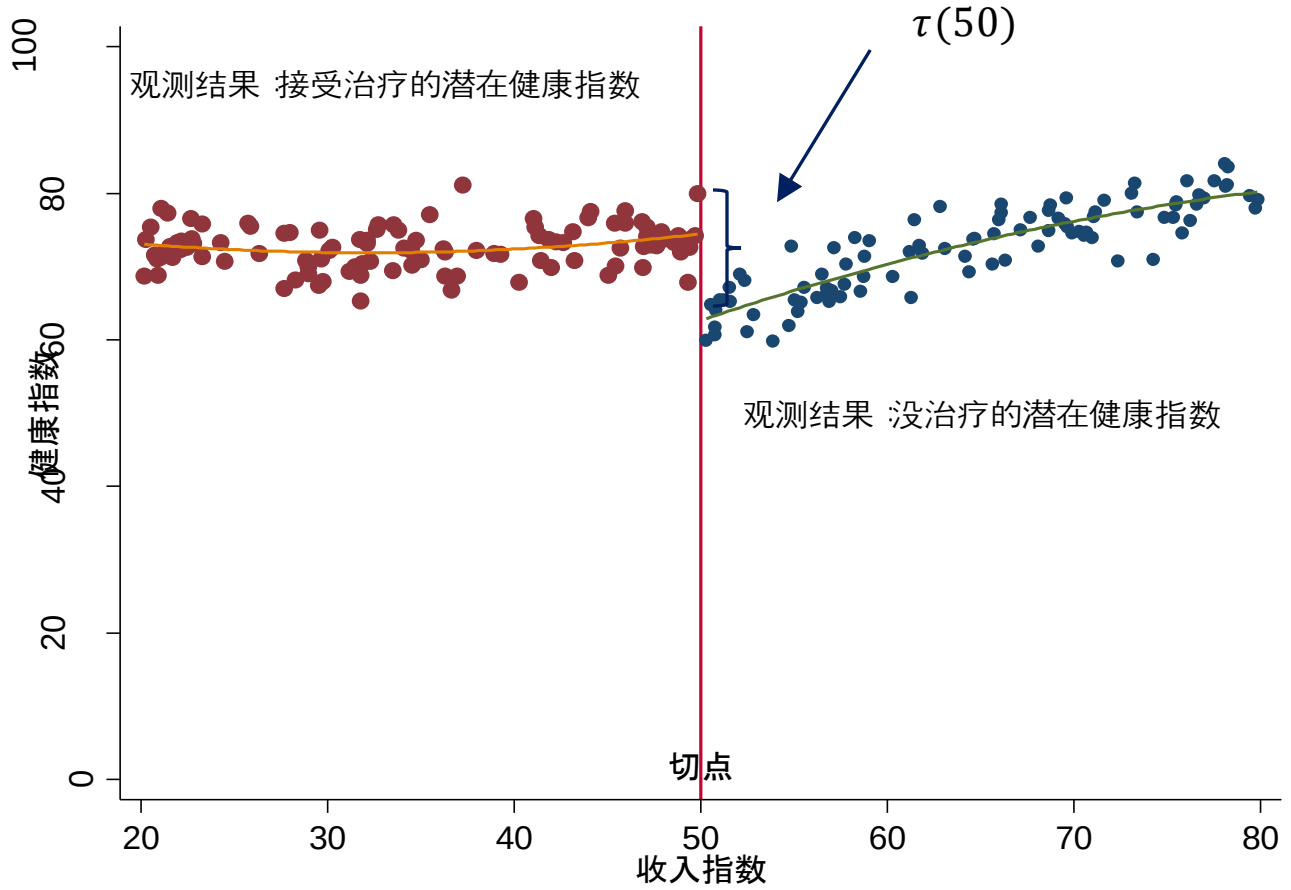

图12.2 收入低于（高于）50病人接受治疗（没接受治疗）情况下观测到的健康情况

断点回归要点一：连续性

- 如果潜在结果函数在断点C是连续的，我们可以用接近断点C的观测值去估计断点处的处置效应，因此处置效应可以表示为

$$\tau_c = E(Y_i(1)|X_i = c) - (Y_i(0)|X_i = c) = \lim_{x \uparrow c} E(Y_i|X_i = x) - \lim_{x \downarrow c} E(Y_i|X_i = x)$$

其中，τ_c是在$X_i = c$的处置效应，$x \uparrow c$ 表示x从小于c的地方无限趋近c，$x \downarrow c$表示x从大于c的地方无限趋近c。

- 潜在结果函数的连续性保证了通过断点左右附近的观测值去估计断点处的处置效应是合理的。

断点回归要点二：局部随机性

- **局部随机性**是指在断点附近的个体是否接受处置是随机的。
  - 本例中，如果某些随机发生的意外收入或支出使得部分病人收入刚好高于或低于50，意味着收入刚好高于50和收入刚好低于50是随机的，导致接受治疗与否也是随机的，从而保证了在断点50左右的病人除了接受治疗与否有差异外，其它的特征没有系统性差异。
- **由于在断点附近局部随机性的特点，断点回归也可被视作一种局部随机实验方法，虽然它使用的是观测数据而非真正的实验数据。**

断点回归的数据要求

断点回归的数据要求一

- 断点回归的数据需要包含3个基本变量

  - ✓ **配置变量**（Assignment Variable），也称作**驱动变**量（Forcing Variables, Running Variables）：配置变量是个体的一个连续特征变量，匹配变量的值是否大于断点将决定个体是否接受处置。在上节例子中配置变量为收入。

  - ✓ **断点**（cut-point）：用于决定个体是否接受处置的阈值。上例中，收入50为断点。

  - ✓ **观测结果**：个体接受处置或未接受处置的观测结果。上例中，个体的观测健康指数为观测结果。

断点回归的数据要求二

- **配置变量的值在断点附近无法被准确操纵。**

  - ✓ 不能被准确操纵的意思是，存在一些随机因素，导致配置变量的值大于或小于断点存在偶然性。本例中，如果有病人为了接受治疗而能够将收入降低到50，或者有病人不愿意接受治疗而能够将收入增加超过50，那么收入稍微低于50的人和收入稍微高于50的人就不是局部随机分配的，可能存在系统性差异，也就不具备可比性。二者观测结果的差异就不能反映处置效应。

断点回归的数据要求三

- **断点的选择不受配置变量的影响。**
  - ✓ 在本例中，如果有个病人收入是52，这个某种原因，政府为了让这位病人接受治疗，将断点设为52，这时断点的选择就受到了配置变量（收入）的影响。这种情况同样造成在断点附近的个体不是局部随机形成的。

断点回归的数据要求四

- **除了处置状态在断点处发生跳跃式变化外，其它未处置前的个体特征变量在断点处没有显著差异。**

  - ✓ 如果其它特征变量在断点处也有显著差异，则观测结果在断点处的变化不一定是由处置状态变化造成的。在上例中，如果收入略低于50的病人比收入略高于50病人在没接受治疗前的锻炼时间也明显多，那么前者在接受治疗后的健康水平比后者高可能是由于前者锻炼时间较多，而非治疗的效果。

RDD 的估计步骤和相应STATA命令

RDD 的估计步骤一

- **第一步的目的是理解使用RDD方法背后的经济机制，需要回答为什么要使用和为什么能使用RDD方法。**

  - **具体执行为：**

（1）讨论配置变量和断点的产生过程，确定配置变量和断点选择是独立的

RDD 的估计步骤二

- **第二步的目的是视觉上观察结果变量在断点处是否有明显的跳跃。**
  - **具体执行为：**

（2）用散点图显示结果变量和配置变量的关系（twoway scatter）

（3）用拟合图显示结果变量和配置变量的关系 (rdplot)，通常有多项式回归拟合和区间均值拟合两种方法。这两种方法都需要做一些选择。

（3.1）多项式回归拟合

- 选择多项式次数

（3.2）区间均值拟合

- 选择区间的分割方式
  - （a）按配置变量值平均分割
  - （b）按观测数量平均分割
- 选择区间的数量
  - （a）手动设置
  - （b）选择IMSE最优区间数量 （rdbwselect）

RDD 的估计步骤三

- **第三步的目的是通过统计方法具体检验数据是否符合使用RDD的前提条件。第三步和第一步是互补的，区别是前者着重从统计角度而后者着重从经济意义论证使用RDD的合理性。**

  - **具体执行为：**

  统计检验RDD的有效性

  (4.1) 检验配置变量密度函数在断点的连续性（rddensity，DC density）

  (4.2) 检验非结果特征变量在断点的连续性（rdplot, rdrobust）

RDD 的估计步骤四

- **第四步的目的是在确定使用RDD合理性后，估计处置变量在断点处的跳跃程度和显著性。**

具体执行如下：

（5）断点处处置效应的点估计。通常有以下两种方法：

(5.1) 全局多项式回归（regress）

多项式次数

(5.2) 局部多项式回归（rdrobust）

多项式次数

带宽选择

权重选择

RDD运用实例

文章背景

- **文章引用：** Murillo Campello, Janet Gao, Jiaping Qiu, and Yue Zhang, "Bankruptcy and the Cost of Organized Labor: Evidence from Union Elections," *Review of Financial Studies* 31, no. 3 (March 2018): 980–1013.

- **文章背景：** 本文的目的是研究工会对企业债券价格的影响。企业债券的价格取决于企业的破产概率，以及一旦企业破产，债权人能够拿到的企业剩余价值。理论上工会如何影响企业债券价格并没有一个简单的答案，工会如何影响债券价格是个实证问题。

为什么要使用RDD方法

- 要估计工会和企业债券价格的因果关系，一个简单的思路是比较有工会和没有工会企业的债券价格，但这个办法很有可能受到缺失变量问题的干扰，因为有工会和没有工会的企业除了工会差异之外，还有其它很多观测得到和观测不到的差异，因此很难识别工会和债券价格差异的因果关系。

- 本文利用美国企业成立工会投票事件来估计工会对债券价格的因果影响。当成立工会投票结果公布后，企业的债券价格会对投票结果做出反映，RDD的方法是通过比较工会得票率在50%左右企业的债券价格变化来估计工会的影响。

变量

- **配置变量**（vote_for_share）：支持成立工会的得票率；
- **断点**：得票率为50%。如果支持工会成立得票率大于或等于50%，企业必须成立工会，反之则不必成立工会。
- **虚拟变量**（win）：如果支持工会成立得票率大于或等于50%，win=1；如果得票率小于50%，win=0；
- **结果变量**（return）：工会选举结果后12个月的企业债券超额回报率，如果企业债券价格回报低于（超过）市场同期平均回报水平，其超额回报率为负（正），反之为正。