12
断点回归

龙欣雨

CIRG 2020.12.17

框架

- 引言
- 断点回归的直观理解：一个例子
- 断点回归的数据要求
- RDD的估计步骤和相应的Stata命令
- RDD运用实例

引言

- 断点回归（Regression Discontinuity Design,RD）是一种研究非随机实验但接近随机实验数据的方法
- 适用：研究某些特定社会科学事件的后果
  - 事件特点：个体是否受到事件的影响取决于某个可观测特征的值是否大于给定的临界值
- 分类
  - 清晰断点回归（Sharp RDD）-本章集中讲解
  - 模糊断点回归（Fuzzy RDD）

断点回归的直观理解：一个例子

- 假设政府有一个针对低收入人群的医疗福利。政府根据病人收入为其评分，高评分代表高收入；同时病人健康指数越高代表越健康。
  - 平均接受治疗的潜在健康结果

$$E(Y(1)|X) = f(x)$$

  - 平均未接受治疗的潜在健康结果

$$E(Y(0)|X) = \mathrm{g}(x)$$

  - 两回归曲线间的距离代表给定收入水平的平均治疗效果。例如，对收入$X_i = 30$的病人，平均治疗效果为：

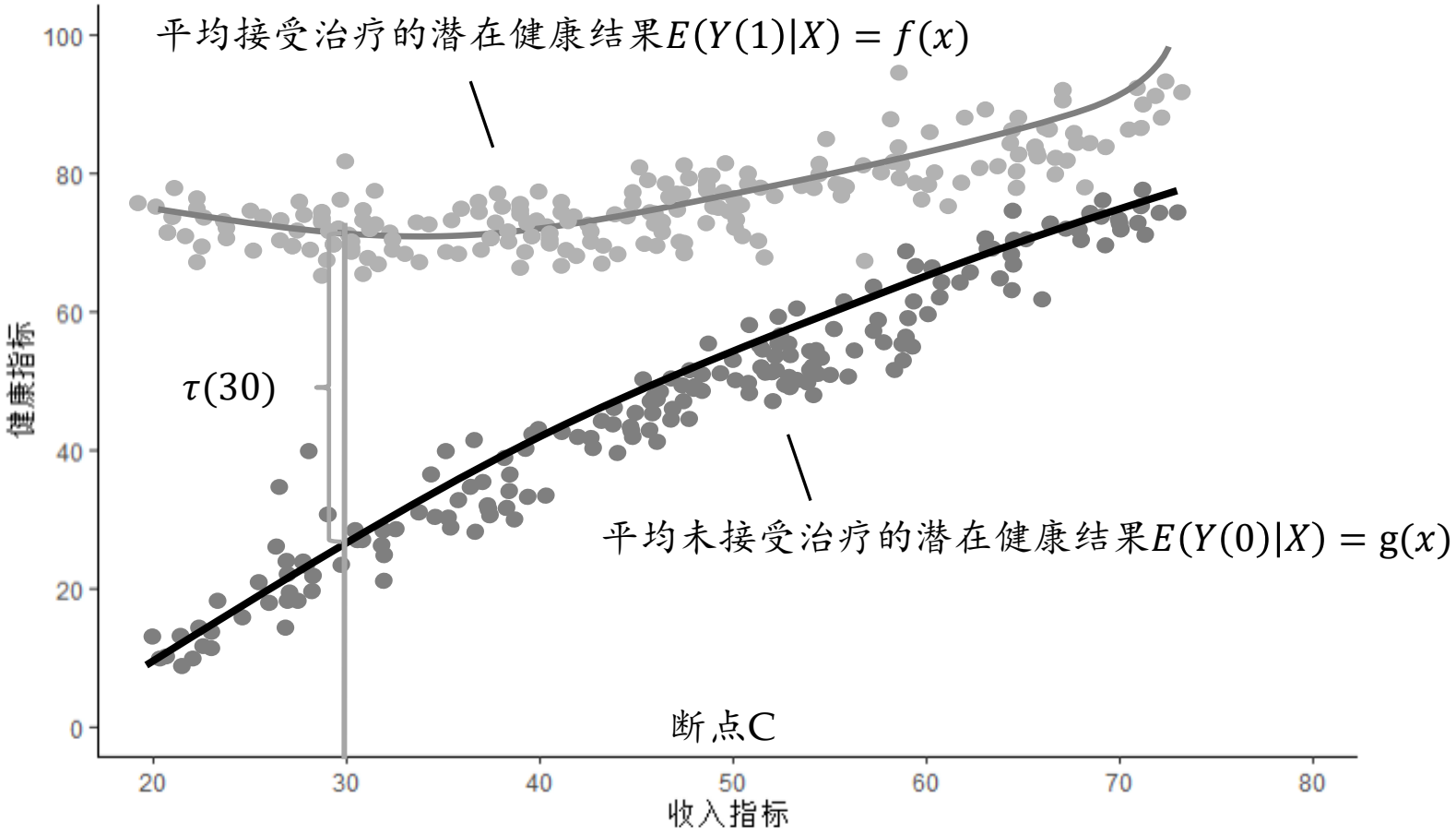
$$\tau(30) = E(Y_i(1)|X_i = 30) - E(Y_i(0)|X_i = 30) = f(30) - g(30)$$

  - 总体平均治疗效果

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

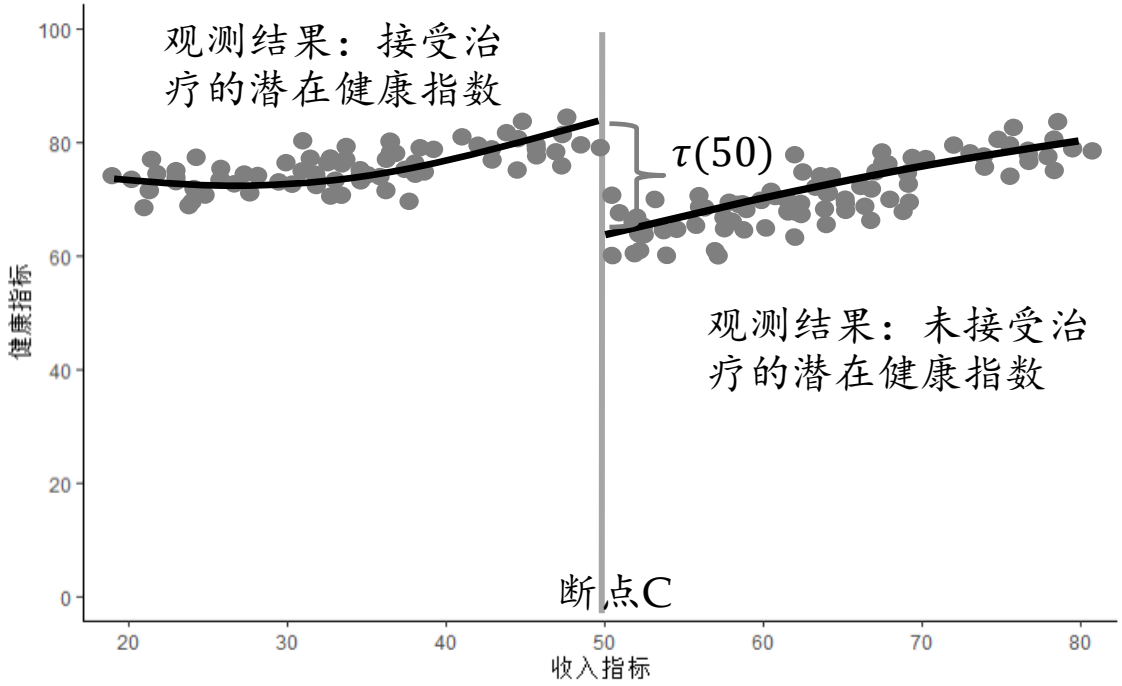
$P(x)$: 不同收入的人数比率

断点回归的直观理解：一个例子



断点回归的直观理解：一个例子

- 实际情况：政府只对收入水平低于或等于50的病人提供治疗，收入高于50的未接受治疗，因此收入50被称为断点C
- 没有匹配方法中要求的“共同支撑域”



断点回归的直观理解：一个例子

- 估计处置效应
  - 用$X_i = 50$的稍左侧接受治疗的平均观测值去近似$E(Y_i(1)|X_i = 50)$，用$X_i = 50$稍右侧未接受治疗的平均观测值去近似$E(Y_i(0)|X_i = 50)$
  - 治疗效果

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

  - 两个理解要点
    - 连续性

$$\tau_c = E(Y_i(1)|X_i = C) - E(Y_i(0)|X_i = C)$$
$$= \lim_{x \to c^-} E(Y_i(1)|X_i = x) - \lim_{x \to c^+} E(Y_i(0)|X_i = x)$$

    - 局部随机性

断点回归的直观理解：一个例子

- 局限

  仅适用于断点处的个体，不能推广到其他个体

断点回归的数据要求

- 数据需包含以下三个基本变量
  - 配置变量/驱动变量：连续特征变量
  - 断点
  - 观测结果
- 配置变量的值在断点附近无法被准确操纵
  - 偶然性，局部随机性
- 断点的选择不受配置变量影响
- 除处置状态在断点处发生跳跃式变化外，其他未处置的个体特征变量在断点处无显著变化

RDD的估计步骤和相应的Stata命令

- 常用Stata RDD命令
  - CCFT编写的非官方RDD命令组
  - 获取方式：http://sites.google.com/site/rdpackages

常用的Stata RDD命令和功能

命令	功能
rdplot	绘制结果变量和配置变量拟合图
rdbwselect	拟合图分区数量选择：可单独使用或包含在rdplot选项里
rdrobust	RDD局部多项式估计
rdwdensity	局部多项式估计带宽选择：可单独使用或包含在rdrobust选项里
rddensity	检验配置变量密度函数是否连续

RDD的估计步骤和相应的Stata命令

- RDD估计步骤
  - 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
  - 观察结果变量在断点处是否有明显跳跃
    - 散点图显示结果变量和配置变量的关系
    - 拟合图显示上述关系

    多项式回归拟合：选择多项式次数

    区间均值拟合：
    ①选择区间分割方式：按配置变量/观测数量
    ②选择区间数量：手动设置/*IMSE*最优区间数量
  - 检验数据是否符合使用RDD的前提条件
    - 配置变量密度函数在断点处的连续性（rddensity）
    - 非结果特征变量在断点处的连续性（rdplot,rdrobust）
  - 估计处置变量在断点处的跳跃程度和显著性-断点处处置效应的点估计
    - 全局多项式回归
    - 局部多项式回归

RDD运用实例

- 目的：考量美国众议院选举是否存在在职优势（一个地区现任政党对在该地区中获得选票的影响）

  Randomized Experiments from Non-random Selection in U.S. House Elections

  Lee, David S 2005-08-08

  - 配置变量：margin
  - 断点：margin=0 若margin>0,则选举胜出
  - 虚拟变量：win=1若margin> 0；win=0若margin ≤ 0
  - 结果变量：vote

RDD运用实例

Name	Label
state	State ID
year	Election Year
vote	Democratic vote share in next election
margin	Democratic margin of victory
class	Senate class
termshouse	Cummulative number of terms served in U.S. House by congress of record
termssenate	Cummulative number of terms served in U.S. Senate by congress of record
win(自己创建)	是否赢得选举

RDD运用实例：描述性统计及其总结

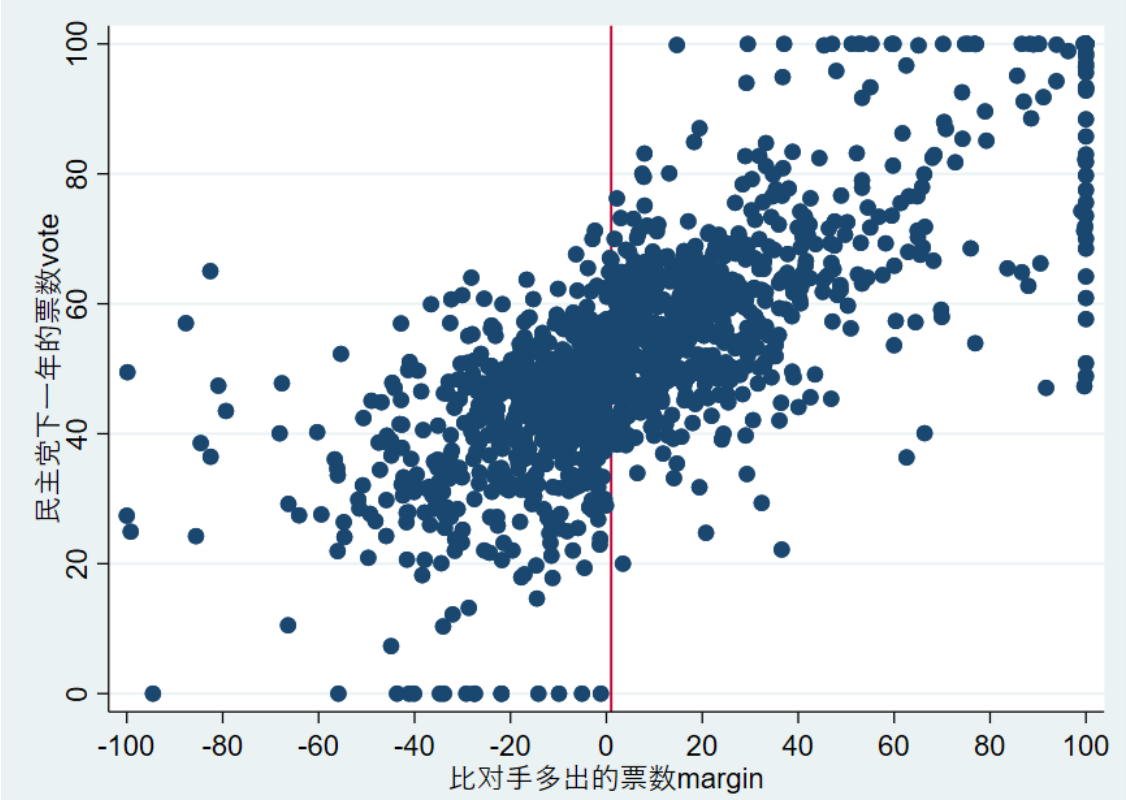
- 数据显示，平均民主党的票数为52.67；win的均值为0.5396，显示其赢得了53.96%的选举；样本包含了1914-2010年的选举数据

```
. sum vote win year
```

Variable	Obs	Mean	Std. Dev.	Min	Max
vote	1,297	52.66627	18.12219	0	100
win	1,390	.5395683	.4986113	0	1
year	1,390	1964.63	28.05466	1914	2010

RDD运用实例：结果变量与配置变量关系-直观

- 讨论配置变量和断点的产生过程
  - 断点0由法律规定，不会受得票率的影响
- 用散点图显示结果变量和配置变量的关系



RDD运用实例：结果变量与配置变量关系-直观

```
. ttest vote, by(win)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	595	40.92053	.4989536	12.17079	39.9406	41.90045
1	702	62.6217	.6147484	16.28793	61.41474	63.82867
combined	1,297	52.66627	.5032002	18.12219	51.67909	53.65344
diff		-21.70118	.810499		-23.29121	-20.11114

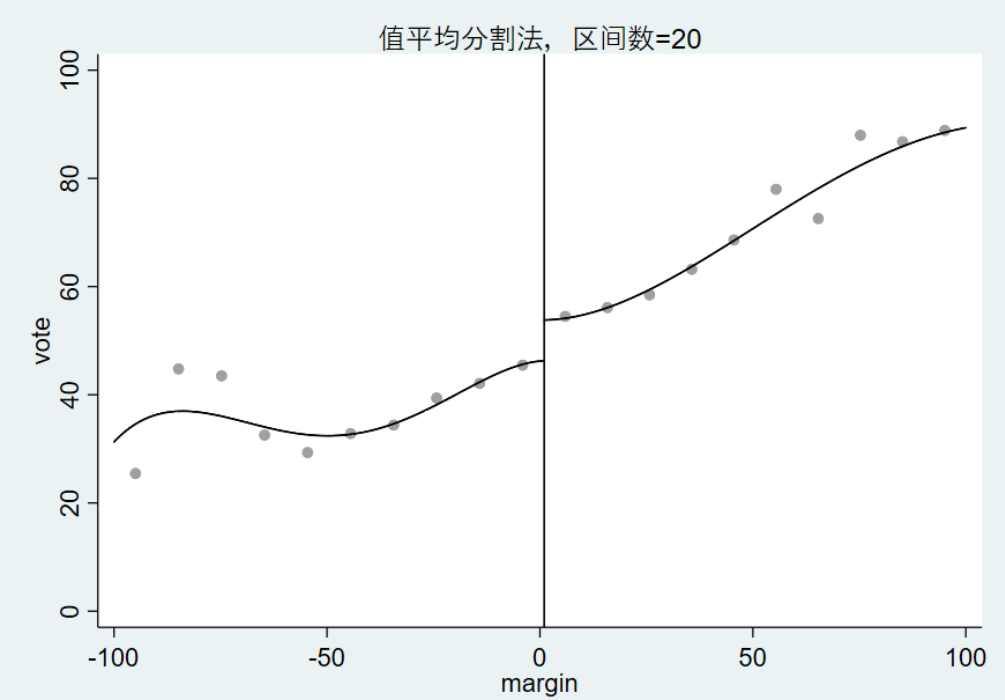
```
    diff = mean(0) - mean(1)                                  t = -26.7751
Ho: diff = 0                     degrees of freedom =     1295

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

RDD运用实例：结果变量与配置变量关系-直观

- 用拟合图显示结果变量和配置变量的关系
  - 多项式回归拟合：绘图过程中不必太严谨，多取2-4次
  - 分区均值拟合
    - 区间分割方法 $\begin{cases} \text{值平均分割}（ES） \\ \text{数量平均分割}（QS） \end{cases}$
    - 区间分割数量：IMSE最优区间数量-原理：如果区间增多，区间平均值估计量偏差（bias）降低，但由于区间内观测值较少，使得估计方差（variance）增加。IMSE最优区间数量指权衡这二者的因素，通过最小化估计量的IMSE得到的区间数量
  - 通过图形考察结果变量和配置变量的关系是RDD的关键第一步。如果平滑拟合后的图形没有明确显示结果变量在断点处有跳跃，即使统计方法发现了显著的处置效果，稳健性也是可疑的

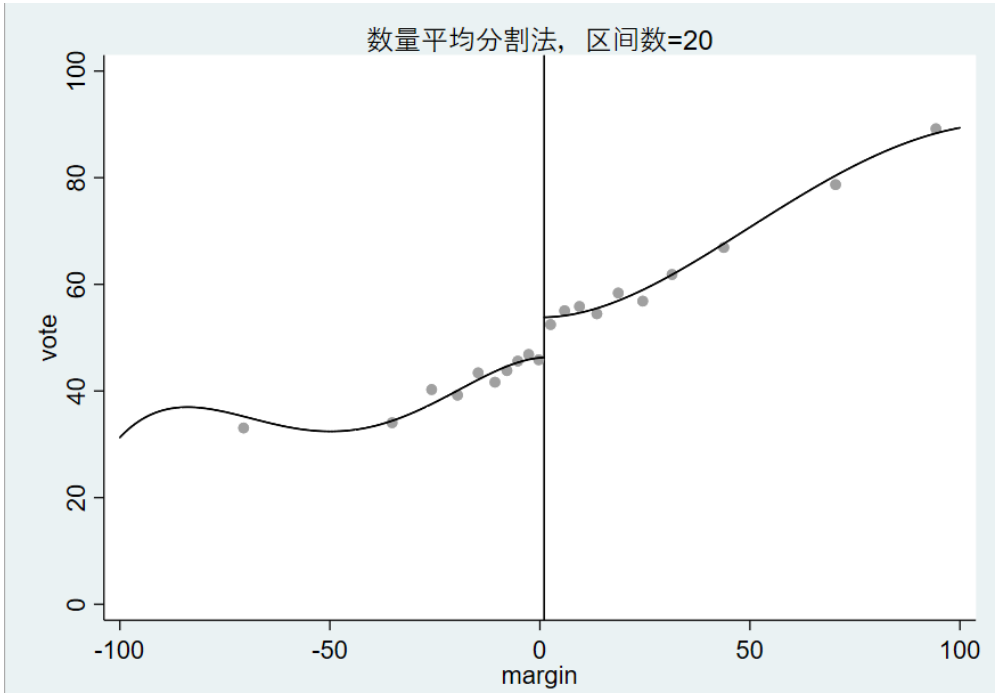
RDD运用实例：结果变量与配置变量关系-直观

- rdplot vote margin, **nbins(10 10) binselect(es)** c(1) p(4) graph_options(legend(off) xtitle("margin") ytitle("vote") ylabel(0(20)100,nogrid) subtitle("值平均分割法，区间数=20") legend(off))



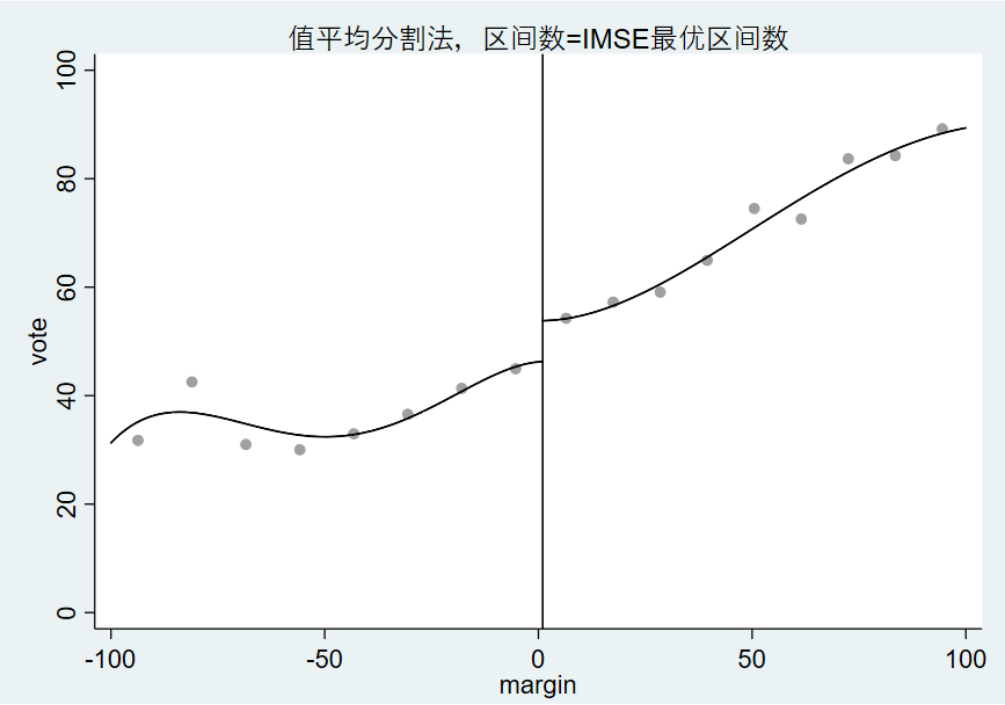
RDD运用实例：结果变量与配置变量关系-直观

- rdplot vote margin, **nbins(10 10) binselect(qs)** c(1) p(4) graph_options(legend(off) xtitle("margin") ytitle("vote") ylabel(0(20)100,nogrid) subtitle("数量平均分割法，区间数=20") legend(off))



RDD运用实例：结果变量与配置变量关系-直观

- rdplot vote margin, **[去掉nbins参数]** binselect(es) c(1) p(4) graph_options(legend(off) xtitle("margin") ytitle("vote") ylabel(0(20)100,nogrid) subtitle("值平均分割法，区间数=IMSE最优区间数") legend(off))



RDD运用实例：验证RDD的有效性

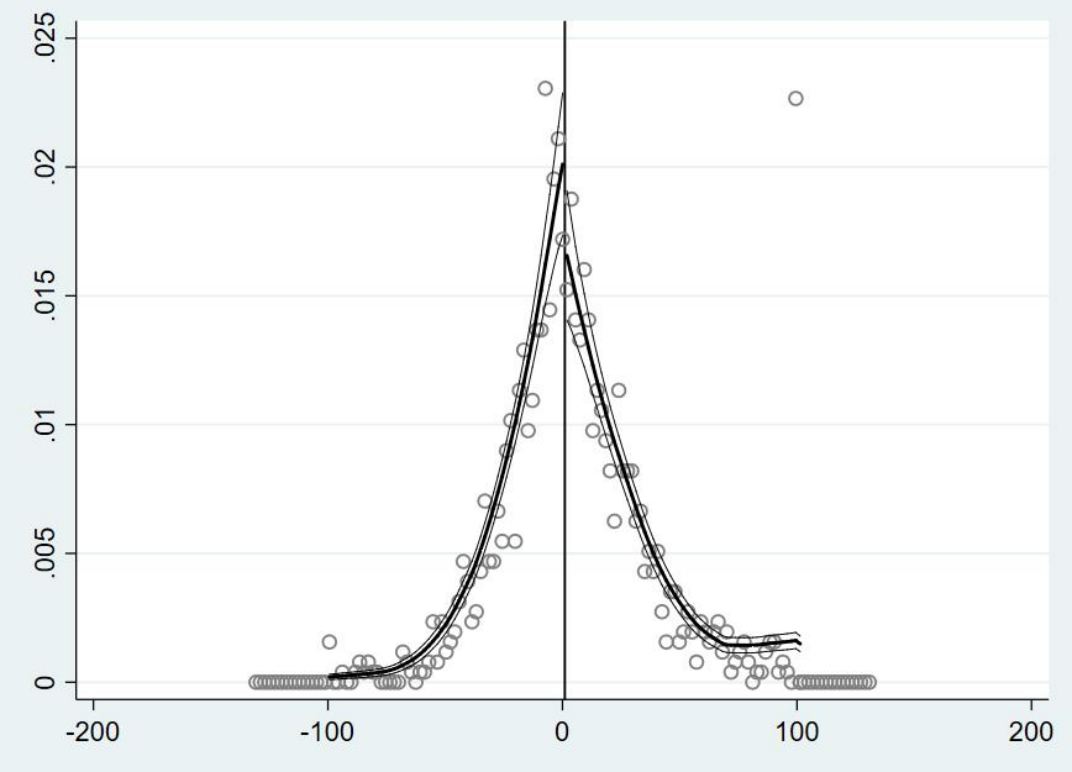
- 检验配置变量的概率分布连续性
  - 方法1

方法	命令	原理	缺陷
McCrary	DCdensity	在断点两边各选定一定的区间 将概率值对配置变量运行两个线性回归 分别得到回归直线在断点左边和右边的截距值 检验log(截距差)的显著性	需要选择局部线性回归的区间

  - 方法2（Cattaneo, Jansson and Ma） rddensity 统计性质更优越

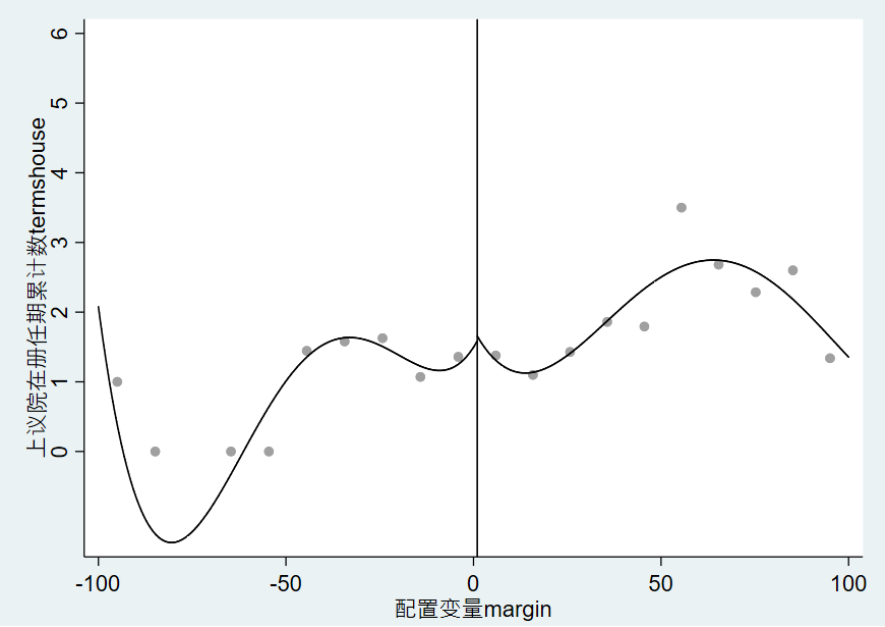
RDD运用实例：验证RDD的有效性

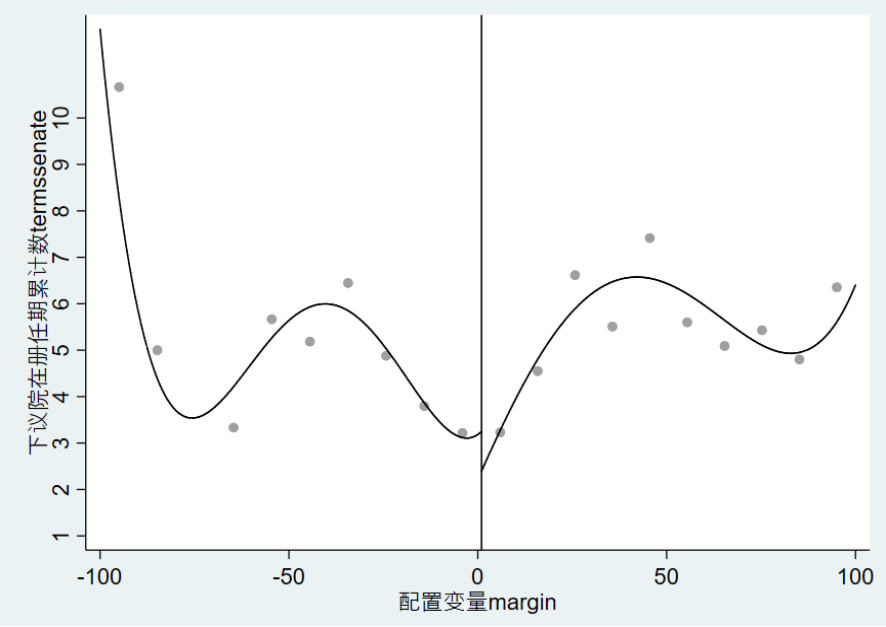
```
. DCdensity margin, breakpoint(1) generate(Xj Yj r0 fhat se_fhat)
Using default bin size calculation, bin size = 1.84133021
Using default bandwidth calculation, bandwidth = 30.3968964

Discontinuity estimate (log difference in height): -.194383955
                                                   (.110478306)
```



RDD运用实例：验证RDD的有效性

- 检验个体特征变量连续性
  - 使用rdplot绘制显示特征变量和配置变量关系的平滑拟合图，并观测在断点处是否连续（注：这些变量必须是在事件发生前的值）





RDD运用实例：断点处置效应估计

- 全局多项式回归
  - 将断点左边与右边数据分别回归

$$Y_i = \alpha_{l0} + \alpha_{l1}(x_i - C) + \alpha_{l2}(x_i - C)^2 + \cdots + \alpha_{lp_{left}}(x_i - C)^{p_{left}} + u_i$$

$$Y_i = \alpha_{r0} + \alpha_{r1}(x_i - C) + \alpha_{r2}(x_i - C)^2 + \cdots + \alpha_{rp_{right}}(x_i - C)^{p_{right}} + v_i$$

截距项为多项式在断点的截距

断点处的处置效应=处置边多项式截距-未处置边多项式截距

  - 将上述两个方程合并为一（引入虚拟变量win）

$$Y_i = \alpha_{l0} + \tau T + \alpha_{l1}(x_i - C) + \alpha_{l2}(x_i - C)^2 + \cdots + \alpha_{lk}(x_i - C)^k$$
$$+(\alpha_{r1} - \alpha_{l1})T(x_i - C) + (\alpha_{r2} - \alpha_{l2})T(x_i - C)^2 + \cdots + (\alpha_{rk} - \alpha_{lk})T(x_i - C)^k$$
$$+ \alpha_{r,k+1}T(x_i - C)^{k+1} + \cdots + \alpha_{r,k+w}T(x_i - C)^{k+w} + e_i$$

  - 方程次数的选择：AIC/尝试报告一系列的次数（1-6次）来观察结果是否对次数选择很敏感
  - 优点：数据多，方差小；缺点：结果对多项式方程设置敏感

RDD运用实例：断点处置效应估计

```
. gen X1 = margin-0

. gen X2 = X1^2

. gen X3 = X1^3

. gen X4 = X1^4

. gen win_X1 = X1*win

. gen win_X2 = X2*win

. gen win_X3 = X3*win

. gen win_X4 = X4*win
```

RDD运用实例：断点处置效应估计

```
. reg vote win X1 X2 X3 X4 win_X1 win_X2 win_X3 win_X4, cluster(class)

Linear regression                               Number of obs     =      1,297
                                                F(1, 2)           =          .
                                                Prob > F          =          .
                                                R-squared         =     0.5958
                                                Root MSE          =     11.562

                                    (Std. Err. adjusted for 3 clusters in class)
------------------------------------------------------------------------------
             |               Robust
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         win |   9.407075   1.556146     6.05   0.026      2.71152    16.10263
          X1 |    -.31181   .1971207    -1.58   0.255    -1.159952     .536332
          X2 |  -.0371986   .0072708    -5.12   0.036    -.0684824   -.0059147
          X3 |  -.0007185   .0001245    -5.77   0.029    -.0012541   -.0001829
          X4 |  -3.92e-06   7.32e-07    -5.35   0.033    -7.07e-06   -7.69e-07
      win_X1 |   .3682319    .114532     3.22   0.085    -.1245594    .8610231
      win_X2 |   .0456877   .0139022     3.29   0.081    -.0141285     .105504
      win_X3 |   .0006644    .000093     7.15   0.019     .0002644    .0010644
      win_X4 |   3.91e-06   1.37e-06     2.87   0.103    -1.96e-06    9.79e-06
       _cons |   43.93729    1.71457    25.63   0.002      36.5601    51.31449
------------------------------------------------------------------------------
```

RDD运用实例：断点处置效应估计

- 局部多项式回归
  - 原理：将回归区域限制在断点附近的区域，通常使用低次多项式（例如，一次/二次多项式）
  - 模型次数选择：局部线性回归

    $$Y_i = \alpha_{l0} + \tau T + \alpha_{l1}(x_i - C) + (\alpha_{r1} - \alpha_{l1})T(x_i - C) + \varepsilon_i$$
  - 核函数权重选择：建议三角核函数（triangular kernel function），给越接近断点处的值赋予更高权重

    与最优带宽一起使用时可以得到均值平方误差（MSE）最优的系数估计量$\hat{\tau}$
  - 带宽（“局部”的范围）选择
    - 带宽越大，受多项式方程设置和数据异常值影响越大，偏差越大；带宽越小，可用数据越少，估计方差越大
    - 常使用MSE最优带宽，原理：

      $$MSE(\tau) = Bias^2(\tau) + Variance(\tau)$$

RDD运用实例：断点处置效应估计

```
. rdrobust vote margin, c(1) kernel(triangular) p(1) bwselect(mserd) vce(cluster class)

Sharp RD estimates using local polynomial regression.
```

Cutoff c = 1	Left of c	Right of c
Number of obs	620	677
Eff. Number of obs	426	363
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	22.778	22.778
BW bias (b)	32.080	32.080
rho (h/b)	0.710	0.710
Number of clusters	3	3

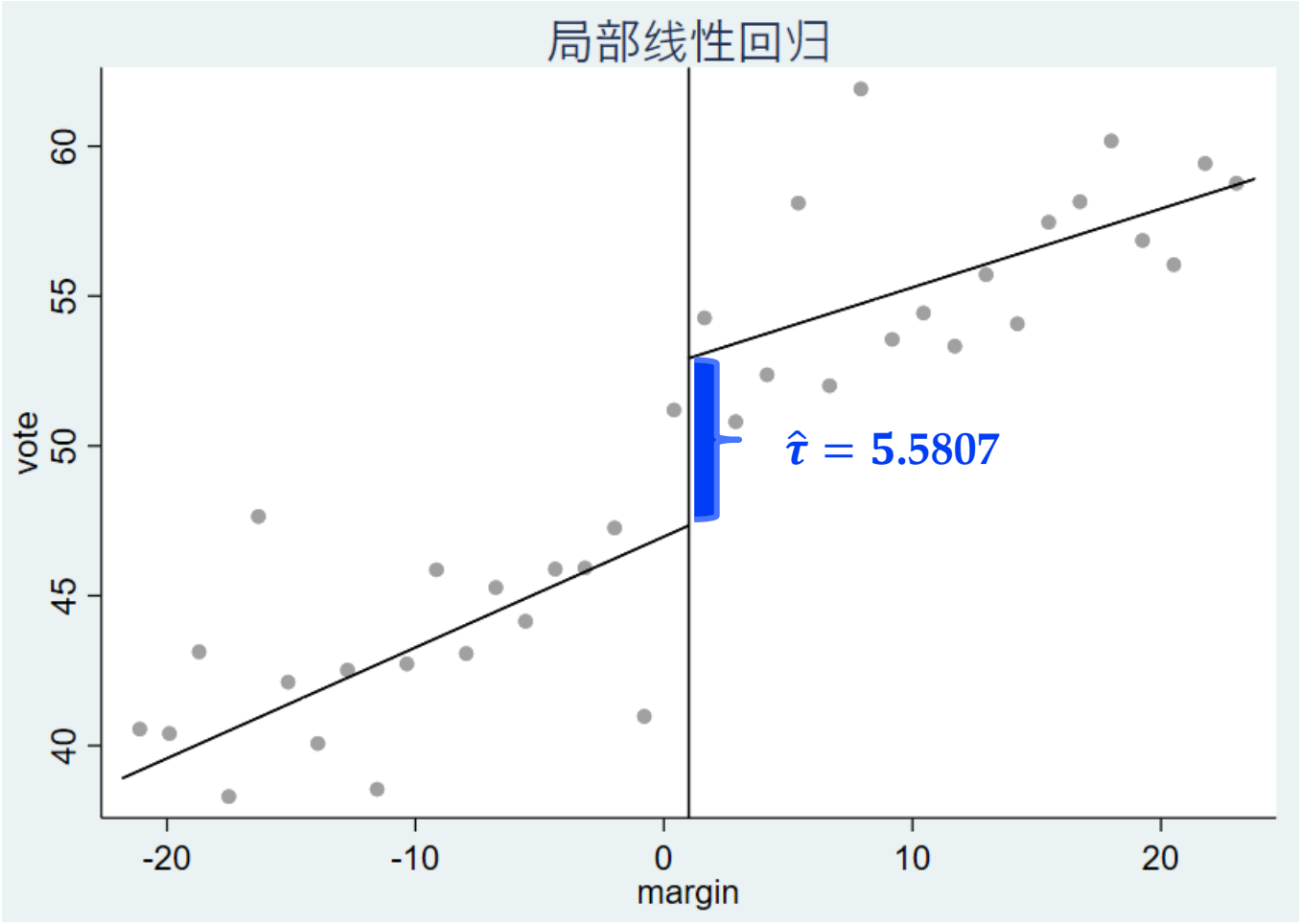
Number of obs =	1297
BW type =	mserd
Kernel =	Triangular
VCE method =	Cluster

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	5.5807	1.7028	3.2774	0.001	2.24328	8.91805
Robust	-	-	2.6198	0.009	1.25916	8.73903

Std. Err. adjusted for clusters in class

RDD运用实例：断点处置效应估计



RDD运用实例：断点处置效应估计

- 局部多项式回归
  - 点估计的置信区间
    - 全局回归

$$95\%Conf.Interval = (\hat{\tau} - 1.96\sqrt{Variance},\ \hat{\tau} + 1.96\sqrt{Variance})$$

    - 局部线性回归：偏差调整稳健置信区间

$$95\%Conf.Interval = [(\hat{\tau} - \widehat{Bias}) - 1.96\sqrt{Variance_{bc}},\ (\hat{\tau} - \widehat{Bias}) + 1.96\sqrt{Variance_{bc}}]$$

Bias: 与带宽选择相关的估计值偏差

$Variance_{bc}$: robust bias-corrected方差

RDD运用实例：断点处置效应估计

```
. rdrobust vote margin, c(1) kernel(triangular) p(1) bwselect(mserd) vce(cluster class)

Sharp RD estimates using local polynomial regression.
```

Cutoff c = 1	Left of c	Right of c
Number of obs	620	677
Eff. Number of obs	426	363
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	22.778	22.778
BW bias (b)	32.080	32.080
rho (h/b)	0.710	0.710
Number of clusters	3	3

Number of obs =	1297
BW type =	mserd
Kernel =	Triangular
VCE method =	Cluster

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	5.5807	1.7028	3.2774	0.001	2.24328	8.91805
Robust	-	-	2.6198	0.009	1.25916	8.73903

Std. Err. adjusted for clusters in class

RDD运用实例：断点处置效应估计

- 结果稳健性检验
  - 考虑在多项式中加入其他特征变量Z作为自变量

$$Y_i = \alpha_{l0} + \tau T + \alpha_{l1}(x_i - C) + \alpha_{l2}(x_i - C)^2 + \cdots + \alpha_{lk}(x_i - C)^k$$
$$+(\alpha_{r1} - \alpha_{l1})T(x_i - C) + (\alpha_{r2} - \alpha_{l2})T(x_i - C)^2 + \cdots$$
$$+ (\alpha_{rk} - \alpha_{lk})T(x_i - C)^k + \alpha_{r,k+1}T(x_i - C)^{k+1} + \cdots$$
$$+ \alpha_{r,k+w}T(x_i - C)^{k+w} + \boldsymbol{Z}'\boldsymbol{\beta} + e_i$$

加入控制变量后，标准误应该降低；若其增大，则说明其他特征变量也影响了结果变量在断点处的跳跃

  - 安慰剂检验，即检验一个不应该受处置事件影响的结果变量在断点处是否也有跳跃