小白学统计|面板数据分析与Stata应用笔记（十一（终））

本期内容： 断点回归设计

> 面板数据分析与Stata应用笔记整理自慕课上浙江大学方红生教授的面板数据分析与Stata应用课程，笔记中部分图片来自课程截图。
> 笔记内容还参考了陈强教授的《高级计量经济学及Stata应用（第二版）》

一、断点回归设计理论

断点回归设计(Regression Discontinuity Design)由Thistlewaite and Campbell(1960)首次使用，但直到1990年代末才引起经济学家的重视。Hahn et al(2001)提供了断点回归的计量经济学理论基础。目前，断点回归在教育经济学、劳动经济学、健康经济学、政治经济学以及区域经济学等领域的应用仍然方兴未艾。

我们以一个上大学的教育回报的案例来理解断点回归设计。

> 我们需要研究的是：上大学与不上大学对工资薪酬有没有影响？有多大的影响？

假设有一名叫Mike的同学：

- Mike不上大学的工资是 $Y_i(0)$；
- Mike上大学后能赚到的工资是 $Y_i(1)$。

那么 $Y_i(1)-Y_i(0)$ 就是读大学给Mike带来的薪酬改变。但是，对于研究人员来说，我们不能既让Mike上大学，又不让Mike上大学，即我们只能获得 $Y_i(0)$ 与 $Y_i(1)$ 其中的一个数据，这样我们就无法得知读大学对薪酬所带来的影响。此时，我们就需要引入断点回归设计。

我们假设： 上大学与否（ D_i ）完全取决于高考成绩 x_i 是否超过500分：

$$D_i=\begin{cases}1 & \text{若 } x_i \geqslant 500\\ 0 & \text{若 } x_i < 500\end{cases} \quad (1)$$

记不上大学与上大学的两种潜在的结果分别为 (y_{0i},y_{1i})。由于 D_i 是 x_i 的确定性函数，所以在给定 x_i 的情况下，可将 D_i 视为常数，即 D_i 不可能与任何变量有关系。

对于高考成绩为498、499、500或501的考生（500分附近的考生），可以认为他们在各方面（包括可观测变量与不可观测变量）都没有系统差异，他们高考成绩的细微差异只是由于“上帝之手”随机抽样的结果，从而导致成绩为500或501的考生上大学（进入处理组），而成绩为498或499的考生落榜（进入控制组）。因此，由于制度原因，仿佛对高考成绩在小邻域 $[500-\varepsilon,500+\varepsilon]$ 之间的考生进行了随机分组，所以可以视为准实验（quasi experiment）。

> 断点回归设计中处理组和对照组都不是随机的，但是存在一个分组规则，决定哪些个体属于处理组，哪些个体属于对照组。
> 考试成绩本身包含随机因素，如果考生能够事先知道分组规则，并通过自身的努力完全控制分组变量 x 的取值，从而可以自行选择进入处理组或对照组，那么断点回归将会失效。但在一般情况下考生无法精确地控制成绩，因此，在断点附近的考生，成绩大于或小于断点的概率大约都是二分之一，形成局部的随机分组。

由于存在随机分组，所以我们可以一致地估计在 $x=500$ 附近的局部平均处理效应（Local Average Treatment Effect，简记LATE），即

$$\begin{aligned}\text{LATE} &\equiv E(y_{1i}-y_{0i}|x=500)\\ &=E(y_{1i}|x=500)-E(y_{0i}|x=500)\\ &=\lim_{x\downarrow 500}E(y_{1i}|x)-\lim_{x\uparrow 500}E(y_{0i}|x)\end{aligned}\quad (2)$$

其中，$\lim_{x\downarrow 500}$ 与 $\lim_{x\uparrow 500}$ 分别表示从500的右侧与左侧取极限（即右极限与左极限）。在上式最后一步推导假设中，假设条件期望函数 $E(y_{1i}|x)$ 与 $E(y_{0i}|x)$ 为连续函数，所以其极限值等于函数取值。

更一般地，断点可以是某个常数c，而分组规则为：

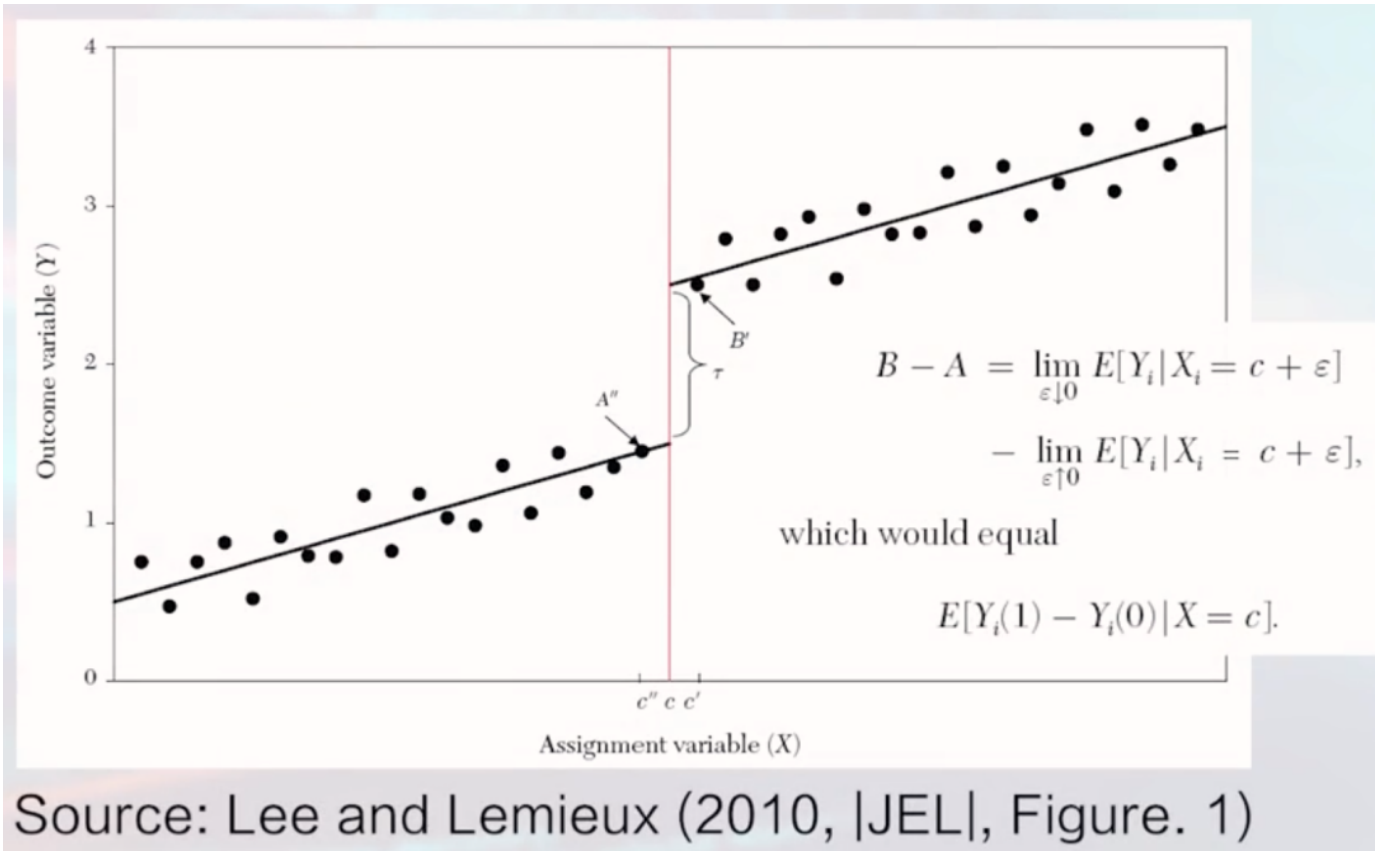
$$D_i=\begin{cases}1 & \text{若 } x_i\geqslant c\\ 0 & \text{若 } x_i<c\end{cases}\quad (3)$$

> 断点回归可以分为两种类型，一种是"精确断点回归"，另一种是"模糊断点回归。"精确断点回归"的特征是：在断点 $x=c$ 处，个体得到处理的概率从0跳跃为1；“模糊断点回归”的特征是：在断点 $x=c$ 处，个体得到处理的概率从a跳跃为b，其中0<a<b<1。笔记只整理了"精确断点回归"的内容，"模糊断点回归"的详细内容可以查看陈强教授的《高级计量经济学及Stata应用（第二版）》中的内容。

假设在实验前，结果变量 y_i 与 x_i 之间存在如下线性关系：

$$y_i=\alpha+\beta x_i+\varepsilon_i\quad (i=1,\cdots,n)\quad (4)$$

不失一般性，假设 $D_i=1(x_i\geqslant c)$ 的处理效应为正，则 y_i 与 x_i 之间的线性关系在 $x=c$ 处就存在一个向上跳跃的断点，如下图所示：



由于在 $x=c$ 附近，个体在各方面均无系统差别，所以造成条件期望函数 $E(y_i|x)$ 在此跳跃的唯一原因只可能是 D_i 的处理效应。

为了估计这一影响，我们构造断点回归设计RDD回归方程：

$$y_i=\alpha+\beta(x_i-c)+\delta D_i+\gamma(x_i-c)D_i+\varepsilon_i\quad (i=1,\cdots,n)\quad (5)$$

上式中，变量 $(x_i - c)$ 为 x_i 的标准化，这样可以使得 $(x_i - c)$ 的断点为0；引入交互项 $\gamma(x_i - c)D$ 是为了允许在断点两侧的回归线斜率可以不同。

> 在有交互项的情况下，如果在上述方程中使用 x_i 而不是标准化变量 $(x_i - c)$，则 $\hat{\delta}$ 虽然度量断点两侧回归线的截距之差，但并不等于这两条回归线在 $x = c$ 处的跳跃距离。这是因为，存在交互项意味着断点两侧的回归线斜率不同，并非平行线，所以在断点处的跳跃距离并不等二者的截距项之差。
> 上述方程等价于使用断点两侧的数据分别进行回归，然后计算两侧截距项之差。如果断点两侧的回归线斜率不同，但未包括交互项，即相当于强迫两侧的斜率相同。这会导致断点右（左）侧的观测值影响对左（右）侧截距项的估计（我们不希望有此影响），从而引起偏差。

当 $D = 0$ 时，得到的是上图中左边那段线，当 $D = 1$ 时，得到的则是右边那段线。

对上述方程进行OLS回归，得到的系数 $\hat{\delta}$ 就是在 $x = c$ 处局部平均处理效应（LATE）的估计量，也即政策效应。由于此回归存在一个断点，所以称为"断点回归"或"断点回归设计"。
对上述方程（5）进行断点回归的估计可能会出现两个问题。首先，如果回归函数包含高次项（比如二次项 $(x - c)^2$ ）则会导致遗漏变量偏差；其次，既然断点回归是局部的随机试验，那么原则上只应该使用断点附近的观测值，但方程（5）的估计是使用的整个样本。
为了解决以上两个问题，我们在方程（5）的基础上引入高次项（比如二次项），并限定 x 的取值范围为 $(c - h, c + h)$：

$$\begin{aligned} y_i =& \alpha + \beta(x_i - c) + \delta D_i + \gamma_1(x_i - c)D_i \\ &+ \beta_2(x_i - c)^2 + \gamma_2(x_i - c)^2 D_i + \varepsilon_i \qquad (6) \\ &(i = 1, \cdots, n; c - h < x < c + h) \end{aligned}$$

其中， δ 为对LATE的估计量，并可以使用稳健标准误来控制可能存在的异方差。

> 由于在断点附近仿佛存在随机分组，所以一般认为断点回归是内部有效性比较强的一种准实验。在某种意义上，断点回归可以视为"局部随机实验"；而且，可以通过考察协变量在断点两侧的分布是否具有显著差异来检验此随机性。另一方面，断点回归仅推断在断点处的因果关系，并不一定能推广到其他样本值，所以外部有效性受局限。

因为最优带宽 h 通常是难以确定的，而且在设定函数的情况下还与函数的具体形式有关，所以，研究者们开始转向非参数回归，不再依赖于具体的函数形式，这种非参数回归的方法还可以通过最小化均方误差（MSE）来选择最优带宽 h，通常情况下使用的方法都是局部线性回归的方法，具体的模型设定是：

$$\min_{\{\alpha,\beta,\delta,\gamma\}} \sum_{i=1}^{n} K[(x_i - c)/h][y_i - \alpha - \beta(x_i - c) - \delta D_i - \gamma(x_i - c)D_i]^2 \qquad (7)$$

上式中， $[y_i - \alpha - \beta(x_i - c) - \delta D_i - \gamma(x_i - c)D_i]^2$ 为局部回归的残差平方和； $K[(x_i - c)/h]$ 为加权核函数，越接近断点，权重越大，一般默认为三角核函数，也可以使用矩形核（权重相等，相当于普通OLS回归），简而言之，三角核函数属于加权的OLS。

> 局部线性回归的实质是，在一个小邻域 $(c - h, c + h)$ 内进行加权最小二乘估计，此权重由核函数来计算，离 c 越近的点权重越大。

最优带宽 h 的确定通常有三种方法：

- Cross Validation(Ludwig & Miller,2007)，简记为CV;

- Imbensand Kalyanaraman(2012)，简记为IK;
- Calonico,Cattaneo,Titiunik(2012,2017)，简记为CCT。

> Stata的命令rd，默认使用IK法确定最优带宽；命令rdrobust提供CCT最新多种不同的最优带宽计算方法选项（mserd、cerrd，默认为mserd）。

断点回归设计也需要进行稳健性检验，由于断点回归在操作上存在不同选择，实践中一般建议同时汇报以下四种情形，以保证稳健性。

- 分别汇报三角核与矩形核的局部线性回归结果（后者等价于线性参数回归）；
- 分别汇报使用不同带宽的结果（比如，最优带宽及其二分之一或两倍带宽），即不同计算方式，需要自己计算出然后加入到回归命令中；
- 分别汇报包含协变量与不包含协变量的情形；
- 进行模型设定检验，包括检验分组变量（如高考分数）与协变量的条件密度是否在断点处连续，防止存在人为操作。

> 对于协变量的条件密度是否连续除了画图外还可以使用"跑回归"的方式，也即将协变量逐个作为被解释变量进行回归，看是否存在显著的政策效应，如果存在则表明协变量不连续。

二、断点回归设计的Stata命令

断点回归设计在Stata中的命令主要有：

- rdrobust：断点回归设计的命令
- rdbwselect：选择最优带宽的命令
- rdbinselect：处理箱体的命令
- rdplot：画图的命令，用来检验是否存在跳跃点
- McCrary Test：检验分组变量与协变量的条件密度是否在断点处连续

rdrobust命令的基本格式如下：

rdrobust depvar runvar [if] [in] [, c(cutoff) p(pvar) q(qvar) kernel(kernelfn) h(hvar) b(bvar) rho(rhovar) covs(covars) bwselect(bwmethod) delta(deltavar) vce(vcemethod) matches(nummatches) level(level) all]

其中，"depvar"是被解释变量；"runvar"是分组变量；"c(cuoff)"中填写分组变量的断点；"p(pvar)"是局部多项式点估计，默认为局部线性；"q(qvar)"指定多项式的阶数，默认为2阶；"kernel(kernelfn)"用来指定核函数，默认为三角核函数（另外两个选项分别为均匀核函数（uniform）和伊番捏切科夫核函数（epanechnikov），三种核函数的主要区别是误差项平方和前权重的大小是不同的）；"h(hvar)"为带宽选择，默认使用最优带宽（我们也可以手动计算带宽的0.5、1.5、2倍等值带入括号中，在稳健性检验时常常使用这种方法）；"b(bvar)"对箱体进行设定，比如括号中如果填写100，就是将箱体个数放大100倍；"covs(covars)"中放入协变量；"bwselect(bwmethod)"是估计带宽的方法，默认为CCT方法；“all”表示均使用默认值。

命令rdbwselect、rdbinselect与rdplot的语法格式与rdrobust相似。

McCrary Test命令的基本格式为：

DCdensity assign_var, breakpoint(#) generate(Xj Yj r0 fhat se_fhat) graphname (filename)

其中，"assign_var"为分组变量；必选项"breakpoint(#)"用来指定断点位置；必选项"generate(Xj Yj r0 fhat se_fhat)"用来指定输出变量名；选择项"graphname (filename)"用来指定密度函数图的文件名。

三、断点回归设计的案例操作

> 研究美国民主党候选人在本届的当选情况对自己在本辖区内下一届竞选得票率的影响，即考察当期在位者的优势。根据美国竞选法，当一个辖区内民主党得票率超过50%，即可获胜当选成为在位者。

显然，分组变量就是本届得票率，断点为0.5。如果将民主党与共和党的当期得票率作差，此时断点就变成0。

案例中具体包括如下变量：

- vote：民主党候选人在下次参议院选举中的得票比例。
- margin：民主党候选人在上次参议院选举中的得票比例减去共和党候选人在上次参议院选举中的得票比例（判断下次选举时是否连任，断点为0）。
- class：美国参议院有100个席位，分为三个组，每组有33-34人，每隔两年轮换一次，class指在哪一个组任职。
- termshouse：某议员在众议院任职任期数。
- termssenate：某议员在参议院任职任期数。
- population：某个州的人口总数。

我们设定断点回归模型如下：

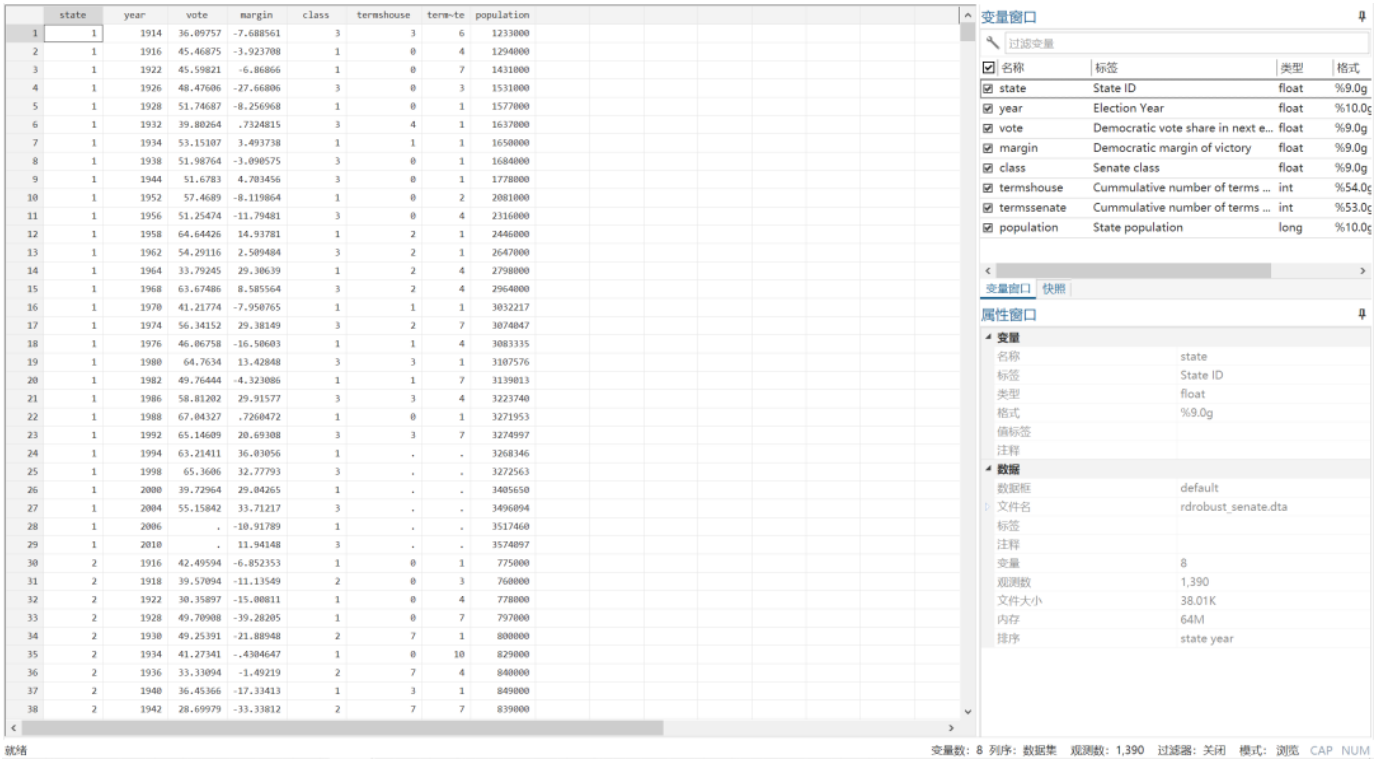
$$v_{i2} = \alpha w_{i1} + \beta v_{i1} + \gamma d_{i2} + \varepsilon_{i2}$$

$$d_{i2} = 1[v_{i1} \geqslant 0.5]$$

其中， v_{i2} 为民主党人下一期在选区 i 的得票率； v_{i1} 为本年度民主党人在选区 i 的得票率； w_{i1} 为协变量向量； d_{i2} 为第二年选举期间民主党是否为执政党的指标变量， v_{i1} 大于0.5就是在位者 ，如果是在位者 d_{i2} 就取值为1，否则为0。

在Stata中调用"rdrobust_senate.dta"数据集，并查看数据集信息。

```
use "D:\rdrobust_senate.dta"
des
sum vote margin class termshouse termssenate population,sep(2)
```



	state	year	vote	margin	class	termshouse	term~te	population
1	1	1914	36.09757	-7.688561	3	3	6	1233000
2	1	1916	45.46875	-3.923708	1	0	4	1294000
3	1	1922	45.59821	-6.86866	1	0	7	1431000
4	1	1926	48.47606	-27.66806	3	0	3	1531000
5	1	1928	51.74687	-8.256968	1	0	1	1577000
6	1	1932	39.80264	.7324815	3	4	1	1637000
7	1	1934	53.15107	3.493738	1	1	1	1650000
8	1	1938	51.98764	-3.090575	3	0	1	1684000
9	1	1944	51.6783	4.703456	3	0	1	1778000
10	1	1952	57.4689	-8.119864	1	0	2	2081000
11	1	1956	51.25474	-11.79481	3	0	4	2316000
12	1	1958	64.64426	14.93781	1	2	1	2446000
13	1	1962	54.29116	2.589484	3	2	1	2647000
14	1	1964	33.79245	29.30639	1	2	4	2798000
15	1	1968	63.67486	8.585564	3	2	4	2964000
16	1	1970	41.21774	-7.950765	1	1	1	3032217
17	1	1974	56.34152	29.38149	3	2	7	3074047
18	1	1976	46.06758	-16.50603	1	1	4	3083335
19	1	1980	64.7634	13.42848	3	3	1	3107576
20	1	1982	49.76444	-4.323086	1	1	7	3139013
21	1	1986	58.81202	29.91577	3	3	4	3223740
22	1	1988	67.04327	.7260472	1	0	1	3271953
23	1	1992	65.14609	20.69308	3	3	7	3274997
24	1	1994	63.21411	36.03056	1	.	.	3268346
25	1	1998	65.3606	32.77793	3	.	.	3272563
26	1	2000	39.72964	29.04265	1	.	.	3405650
27	1	2004	55.15842	33.71217	3	.	.	3496094
28	1	2006	.	-10.91789	1	.	.	3517460
29	1	2010	.	11.94148	3	.	.	3574097
30	2	1916	42.49594	-6.852353	1	0	1	775000
31	2	1918	39.57094	-11.13549	2	0	3	760000
32	2	1922	30.35897	-15.00811	1	0	4	778000
33	2	1928	49.70908	-39.28205	1	0	7	797000
34	2	1930	49.25391	-21.88948	2	7	1	800000
35	2	1934	41.27341	-.4304647	1	0	10	829000
36	2	1936	33.33094	-1.49219	2	7	4	840000
37	2	1940	36.45366	-17.33413	1	3	1	849000
38	2	1942	28.69979	-33.33812	2	7	7	839000


```
  obs:         1,390
 vars:             8                          7 Mar 2017 18:33
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
state           float   %9.0g                 State ID
year            float   %10.0g                Election Year
vote            float   %9.0g                 Democratic vote share in next election
margin          float   %9.0g                 Democratic margin of victory
class           float   %9.0g                 Senate class
termshouse      int     %54.0g                Cummulative number of terms served in U.S. House by
                                                congress of record
termssenate     int     %53.0g                Cummulative number of terms served in U.S. Senate by
                                                congress of record
population      long    %10.0g                State population
-------------------------------------------------------------------------------
Sorted by: state  year
```

```
. sum vote margin class termshouse termssenate population,sep(2)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        vote |      1,297    52.66627    18.12219         0        100
      margin |      1,390    7.171159    34.32488      -100        100
-------------+---------------------------------------------------------
       class |      1,390    2.023022    .8231983         1          3
  termshouse |      1,108    1.436823    2.357133         0         16
-------------+---------------------------------------------------------
 termssenate |      1,108    4.555957    3.720294         1         20
  population |      1,390     3827919     4436950     78000   3.73e+07
```

通过des命令，我们可以看到，这个数据集中有1390个观测值，8个变量。此外，我们也可以看到各变量的标签。

通过sum命令我们可以查看各变量的描述性统计情况，命令中"sep(2)"是指每两个变量隔一条线。

接下来，我们在Stata中实现断点回归设计。

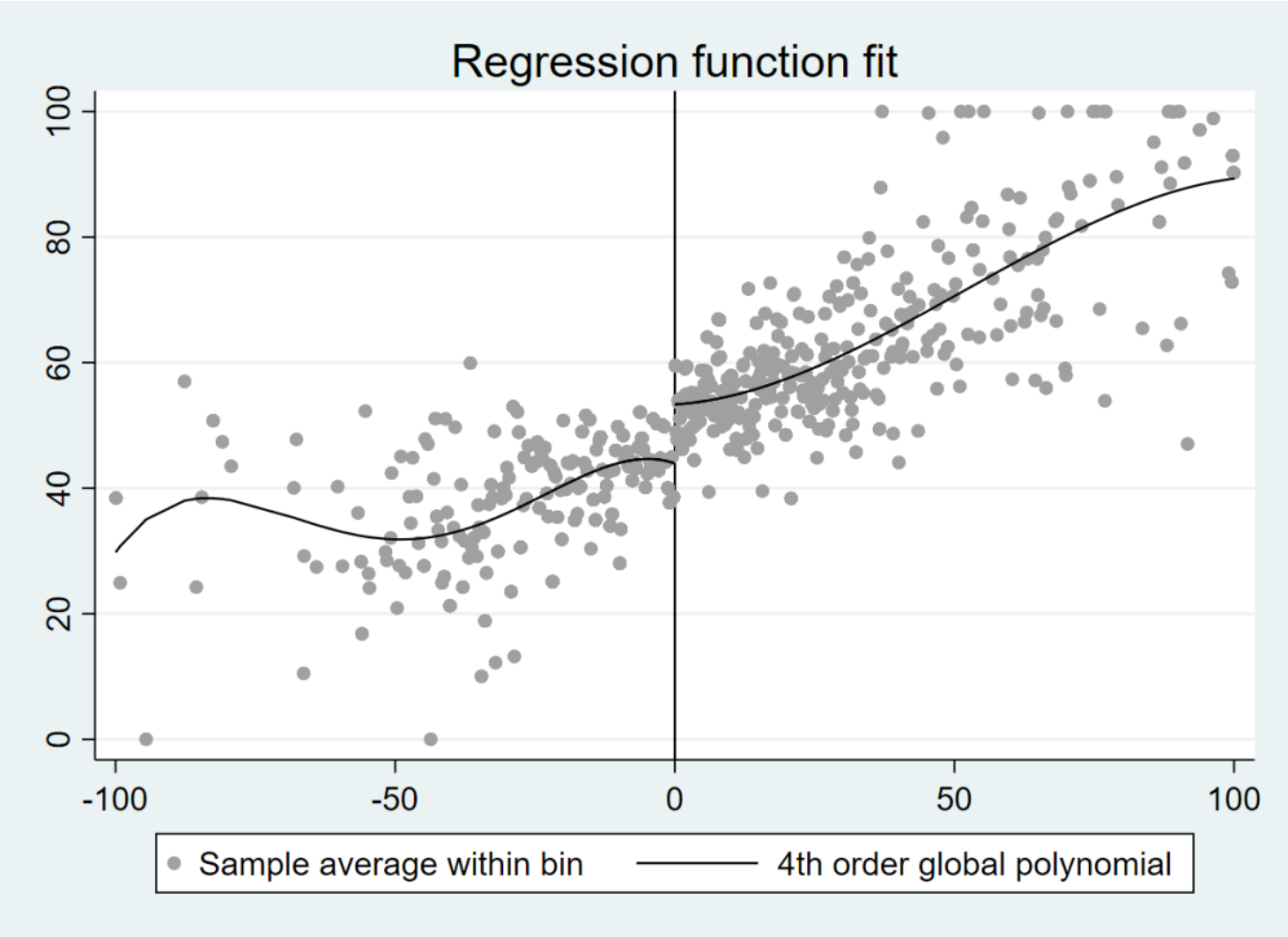
```
rdbinselect vote margin, scale(100)
```

使用rdbinselect命令进行估计（也可以使用新命令rdplot进行这一估计）。其中，"vote"是被解释变量，"margin"是分组变量，"scale(100)"是将箱体个数扩大100倍。

```
. rdbinselect vote margin, scale(100)

Number of bins for RD estimates - Method: Evenly spaced

     Cutoff c = 0 |  Left of c     Right of c
------------------+----------------------------
     Number of obs |        595            702
      Poly. order  |          4              4
    Number of bins |          3              6
        Bin length |     33.307         16.661
------------------+----------------------------
            Scale  |        100            100
    Number of bins |        300            600
        Bin length |      0.333          0.167
```

从输出结果我们可以看到，在断点左侧有595个观测值，在断点右侧有702个观测值；“Poly.order”显示使用了四阶多项式进行拟合；断点左侧的箱体数目为3个，右侧为6个；断点左侧箱体的宽度为33.307，右侧箱体宽度为16.661。表格的下半部分是将箱体个数扩大了100倍后的结果，箱体个数扩大为原来的100倍，箱体宽度缩小为原来的100倍。

与此同时，通过rdbinselect命令我们还得到了回归函数的拟合图，如下图所示。



上图中的每个点对应着每个直方图的均值，300个箱体即为300个点，每个点就是箱体包含数据的均值。断点左侧表示在位者不是民主党人时，民主党下次当选的得票率；右侧表示在位者是民主党人时，民主党下一次还当选的得票率。观察上图，可以很明显的看到：在断点处有一个跳跃，这个跳跃一定程度上反映了在位者的优势。

接下来，我们对这一结果进行估计。

```
1  rdrobust vote margin,all
```

```
. rdrobust vote margin,all

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c           Number of obs =       1297
                   |                                 BW type       =      mserd
     Number of obs |       595         702           Kernel        = Triangular
Eff. Number of obs |       360         323           VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |    17.754      17.754
       BW bias (b) |    28.028      28.028
         rho (h/b) |     0.633       0.633

Outcome: vote. Running variable: margin.

------------------------------------------------------------------------------
      Method |     Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |     7.4141     1.4587     5.0826   0.000      4.5551     10.2732
Bias-corrected |     7.5065     1.4587     5.1460   0.000     4.64747     10.3655
      Robust |     7.5065     1.7413     4.3110   0.000      4.0937     10.9193
------------------------------------------------------------------------------
```

由结果可知，计算最优带宽的方法"BW type"是“mserd”，加权核函数"Kernel"为三角核函数"Triangular"。左侧的观测值是595个，右侧的观测值是702个，(p)=1说明是一次函数形式，最优带宽（h）为17.754，修正之后的最优带宽（最优带宽可能是有偏移的）为28.028。

我们可以看到显著的政策效应为7.4141，最优带宽修正之后的结果为7.5065，其所对应的p值均为0.000，小于0.01，说明在位者优势能够给下一期带来更多的选票。表格中最后一行的"Robust"是针对偏移修正的情况，提供了一个稳健标准误的结果。

最后，我们对上述的估计结果进行稳健性检验

依照第一部分的内容，有四种方法进行稳健性检验，我们应当同时汇报四种方法的结果。

第一种方法：使用另外一种计算最优带宽的方法检验。

```
rdrobust vote margin,all bwselect(cerrd) c(0) p(1) kernel(triangular)
```

```
. rdrobust vote margin,all bwselect(cerrd) c(0) p(1) kernel(triangular)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c           Number of obs =       1297
                   |                                 BW type       =      cerrd
     Number of obs |       595         702           Kernel        = Triangular
Eff. Number of obs |       284         248           VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |    12.407      12.407
       BW bias (b) |    28.028      28.028
         rho (h/b) |     0.443       0.443

Outcome: vote. Running variable: margin.

------------------------------------------------------------------------------
      Method |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |     7.6316     1.6801     4.5424   0.000      4.3387      10.9244
Bias-corrected |     7.6817     1.6801     4.5723   0.000     4.38884      10.9746
      Robust |     7.6817     1.8406     4.1735   0.000     4.07422      11.2892
------------------------------------------------------------------------------
```

可以看到，使用"cerrd"计算最优带宽得到的结果为7.6316与"mserd"得到的结果7.4141相差不大，且二者在1%的显著性水平下均显著。

第二种方法： 先选定一种计算最优带宽的方法，然后查看最优带宽的二分之一或者两倍的结果。

命令rdbwselect可以提供所有计算最优带宽的方法和结果

```
rdbwselect vote margin,all
```

```
. rdbwselect vote margin , all

Bandwidth estimators for sharp RD local polynomial regression.

      Cutoff c =  |  Left of c  Right of c          Number of obs =      1297
                  |                                 Kernel        = Triangular
     Number of obs|       595         702          VCE method    =        NN
     Min of margin|   -100.000       0.036
     Max of margin|     -0.079     100.000
   Order est. (p) |         1           1
   Order bias (q) |         2           2

Outcome: vote. Running variable: margin.
```

Method	BW est. (h) Left of c	BW est. (h) Right of c	BW bias (b) Left of c	BW bias (b) Right of c
mserd	17.754	17.754	28.028	28.028
msetwo	16.170	18.126	27.104	29.344
msesum	18.365	18.365	31.319	31.319
msecomb1	17.754	17.754	28.028	28.028
msecomb2	17.754	18.126	28.028	29.344
cerrd	12.407	12.407	28.028	28.028
certwo	11.299	12.667	27.104	29.344
cersum	12.834	12.834	31.319	31.319
cercomb1	12.407	12.407	28.028	28.028
cercomb2	12.407	12.667	28.028	29.344

上表呈现了10种计算最优带宽的方法和结果。

我们以上表中第一个结果为例，h=17.754，b=28.028。查看二分之一最优带宽和两倍最优带宽的结果。

```
rdrobust vote margin, all h(8.877) b(14.014)
rdrobust vote margin, all h(35.508) b(56.056)
```

```
. rdrobust vote margin , all h(8.877) b(14.014)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0  |  Left of c  Right of c          Number of obs =      1297
                    |                                 BW type       =    Manual
     Number of obs  |       595         702          Kernel        = Triangular
Eff. Number of obs  |       222         182          VCE method    =        NN
    Order est. (p)  |         1           1
    Order bias (q)  |         2           2
       BW est. (h)  |     8.877       8.877
       BW bias (b)  |    14.014      14.014
         rho (h/b)  |     0.633       0.633

Outcome: vote. Running variable: margin.

      Method  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+------------------------------------------------------------
Conventional  |      8.3587     1.9463     4.2946   0.000     4.54401      12.1734
Bias-corrected|      9.3753     1.9463     4.8169   0.000     5.56057       13.19
      Robust  |      9.3753     2.3638     3.9661   0.000     4.74221      14.0083
```

```
. rdrobust vote margin , all h(35.508) b(56.056)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0  |  Left of c  Right of c          Number of obs =      1297
                    |                                 BW type       =    Manual
     Number of obs  |       595         702          Kernel        = Triangular
Eff. Number of obs  |       514         499          VCE method    =        NN
    Order est. (p)  |         1           1
    Order bias (q)  |         2           2
       BW est. (h)  |    35.508      35.508
       BW bias (b)  |    56.056      56.056
         rho (h/b)  |     0.633       0.633

Outcome: vote. Running variable: margin.

      Method  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+------------------------------------------------------------
Conventional  |      7.1616     1.1018     6.4998   0.000     5.00205      9.32109
Bias-corrected|      8.0008     1.1018     7.2614   0.000     5.84127      10.1603
      Robust  |      8.0008     1.3482     5.9344   0.000     5.35836      10.6432
```

上述结果显示，在1%的显著性水平下结果均是显著的。

第三种方法：加入协变量查看结果是否稳健。

```
rdrobust vote margin, all covs(class termshouse termssenate)
```

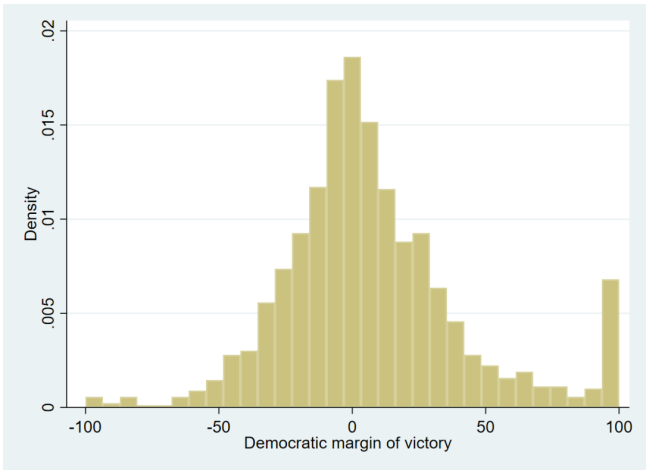
```
. rdrobust vote margin,all covs(class termshouse termssenate)

Covariate-adjusted sharp RD estimates using local polynomial regression.

      Cutoff c = 0 |  Left of c  Right of c          Number of obs =       1108
                   |                                BW type       =      mserd
     Number of obs |        491         617          Kernel        = Triangular
Eff. Number of obs |        315         283          VCE method    =         NN
    Order est. (p) |          1           1
    Order bias (q) |          2           2
       BW est. (h) |     18.033      18.033
       BW bias (b) |     28.988      28.988
         rho (h/b) |      0.622       0.622

Outcome: vote. Running variable: margin.

------------------------------------------------------------------------------
      Method |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |      6.8499     1.4067     4.8694   0.000     4.09275      9.607
Bias-corrected |      6.9884     1.4067     4.9679   0.000      4.2313     9.74556
      Robust |      6.9884     1.6636     4.2009   0.000      3.7279      10.249
------------------------------------------------------------------------------
Covariate-adjusted estimates. Additional covariates included: 3
```

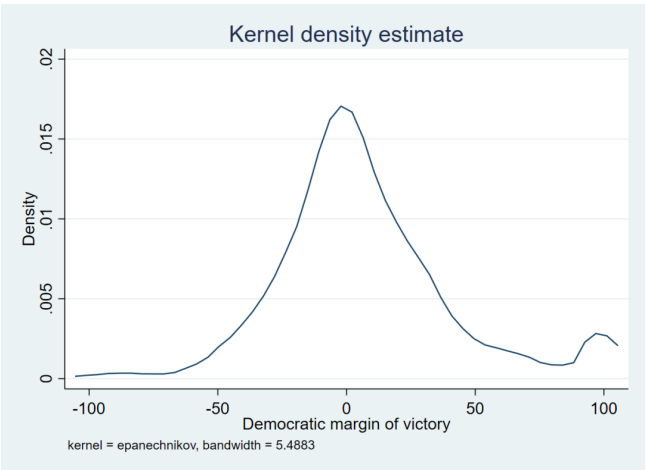
由得到的结果可知，政策效应为6.85，修正后的结果为6.99，两个结果在1%的显著性水平下均是显著的。

第四中方法：设定检验，检验协变量和分组变量的条件密度在断点处是否连续。

命令hist margin可以绘制频数直方图；命令kdensity margin可以绘制密度直方图。

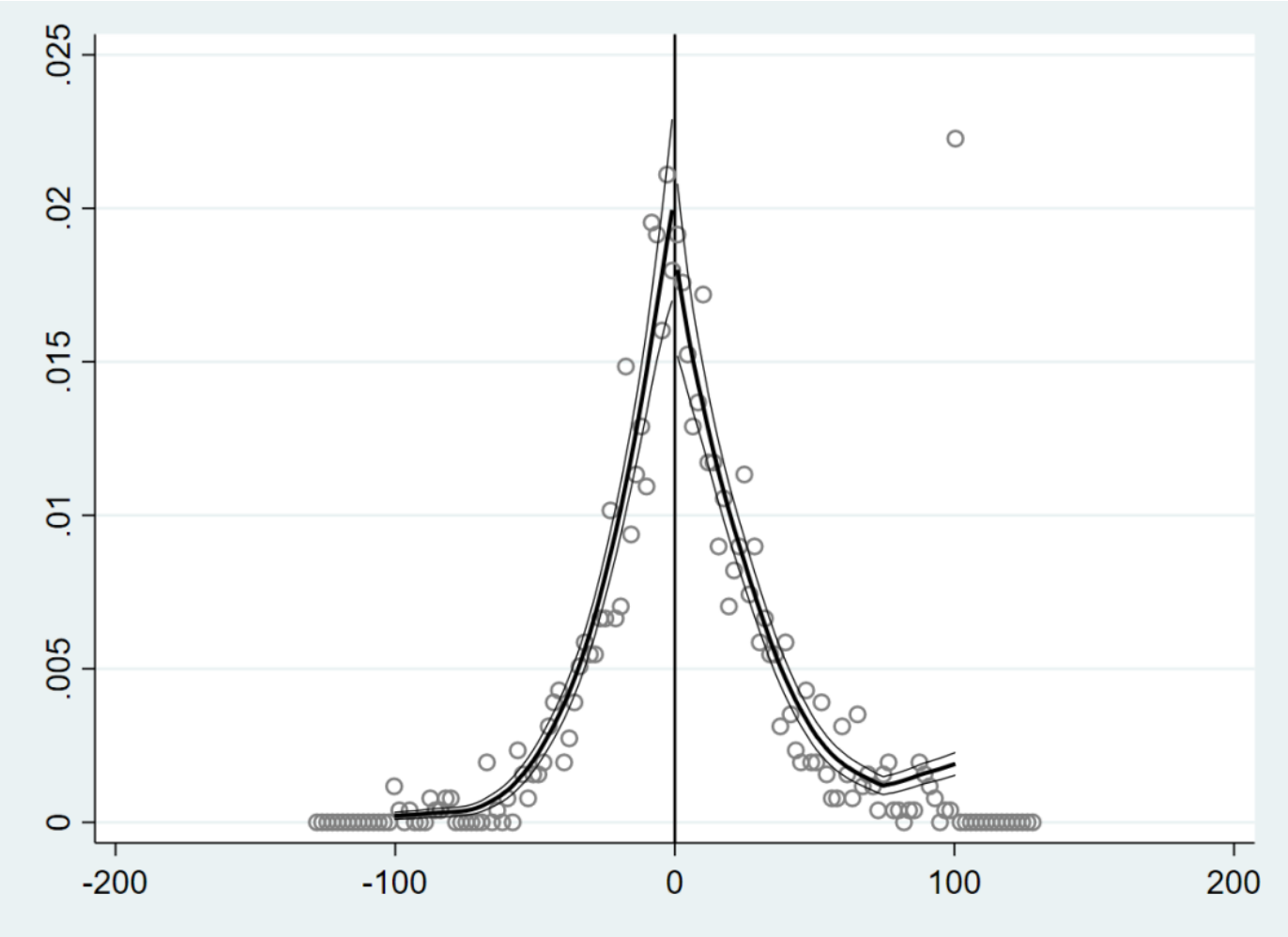
```
hist margin
kdensity margin
```





上图中的频数直方图与密度直方图在断点的左右两侧分布相对均匀。我们通过DCdensity命令绘制分组变量的条件密度函数图。

```
DCdensity margin, breakpoint(0) generate(Xj Yj r0 fhat se_fhat) graphname(rd.ep
```



上图中，黑色的线即为条件密度，黑线旁边的两条线表示一个置信区间。可以看到在断点处两个区间是有交集的，所以可以判断分组变量在断点处是连续的。

另外一种协变量条件密度检验的方法为"跑回归"，即分别将协变量作为被解释变量进行断点回归，预期的结果应当是不显著的。

```
rdrobust class margin,all
rdrobust termshouse margin,all
rdrobust termssenate margin,all
```

```
. rdrobust class margin,all

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c           Number of obs =       1390
                   |                                 BW type       =      mserd
     Number of obs |       640         750           Kernel        = Triangular
Eff. Number of obs |       412         381           VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |    20.924      20.924
       BW bias (b) |    32.813      32.813
          rho (h/b) |     0.638       0.638

Outcome: class. Running variable: margin.

------------------------------------------------------------------------------
      Method |     Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |   -.02135    .12054    -0.1771  0.859    -.25761      .214915
Bias-corrected |   -.01863    .12054    -0.1546  0.877    -.254893     .217632
      Robust |   -.01863    .14311    -0.1302  0.896    -.299114     .261854
------------------------------------------------------------------------------
```

```
. rdrobust termshouse margin,all

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c           Number of obs =       1108
                   |                                 BW type       =      mserd
     Number of obs |       491         617           Kernel        = Triangular
Eff. Number of obs |       282         257           VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |    15.657      15.657
       BW bias (b) |    25.431      25.431
          rho (h/b) |     0.616       0.616

Outcome: termshouse. Running variable: margin.

------------------------------------------------------------------------------
      Method |     Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |   -.17257    .42515    -0.4059  0.685    -1.00585     .660715
Bias-corrected |    -.2902    .42515    -0.6826  0.495    -1.12349      .54308
      Robust |    -.2902     .4996    -0.5809  0.561    -1.26941     .689003
------------------------------------------------------------------------------
```

```
. rdrobust termssenate margin,all

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 |  Left of c  Right of c           Number of obs =       1108
                   |                                 BW type       =      mserd
     Number of obs |        491         617           Kernel        = Triangular
Eff. Number of obs |        291         267           VCE method    =         NN
    Order est. (p) |          1           1
    Order bias (q) |          2           2
       BW est. (h) |     16.443      16.443
       BW bias (b) |     25.956      25.956
         rho (h/b) |      0.634       0.634

Outcome: termssenate. Running variable: margin.

------------------------------------------------------------------------------
      Method |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |   -.19177     .61552    -0.3116   0.755    -1.39816      1.01463
Bias-corrected |   -.09363     .61552    -0.1521   0.879    -1.30003      1.11276
      Robust |   -.09363     .74571    -0.1256   0.900     -1.5552      1.36794
------------------------------------------------------------------------------
```

观察上述三个命令的拟合结果可知，与预期相符所有协变量断点回归的结果均不显著。