基于STATA模拟的内生性：本质、来源及应对

陈传波

chris@ruc.edu.cn

中国人民大学

1 内生性的本质

内生性导致OLS估计量的非一致性

$$y = \beta_1 x_1 + \ldots + \beta_k x_k + u$$

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

$$\mathbf{x}y = \mathbf{x}\mathbf{x}'\boldsymbol{\beta} + \mathbf{x}u$$

$$E(\mathbf{x}y) = E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + E(\mathbf{x}u) \xrightarrow{E(\mathbf{x}u)=0} E(\mathbf{x}y) = E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta}$$

$$E(\mathbf{x}u) = 0 \Rightarrow \boldsymbol{\beta} = \left[E(\mathbf{x}\mathbf{x}')\right]^{-1} E(\mathbf{x}y)$$

$$\hat{\boldsymbol{\beta}} = (\overline{\mathbf{x}\mathbf{x}'})^{-1}(\overline{\mathbf{x}\mathbf{y}})$$

p p p

内生性 $E(\mathbf{x}u) \neq 0$

只有不存在内生性时，估计量在样本取到无限大时才能“命中靶心”，即β的OLS估计为一致估计量。

给定期望和方差阵的虚拟数据回归

	lwage	educ	exper	urban	iq
1.	6.645091	12	11	1	93
2.	6.694562	18	11	1	119
3.	6.715384	14	11	1	108
4.	6.476973	12	13	1	96
5.	6.331502	11	14	1	74

	x	y	z	w	v
1.	6.004705	12.84209	10.09223	.3306129	93.23568
2.	7.020856	14.72017	5.910422	1.063821	121.9491
3.	7.721753	17.30684	7.045604	.4405274	115.508
4.	6.970091	13.44834	5.084117	1.541085	90.25694
5.	6.495118	14.11049	10.3113	.4325938	89.21008

lwage	Coef.	Std. Err.	t
educ	.0548946	.0072099	7.61
exper	.0198516	.00318	6.24
urban	.1728742	.0275652	6.27
iq	.0057722	.0009601	6.01
_cons	5.101422	.1201083	42.47

x	Coef.	Std. Err.	t
y	.0548946	.0072099	7.61
z	.0198516	.00318	6.24
w	.1728742	.0275652	6.27
v	.0057722	.0009601	6.01
_cons	5.101422	.1201083	42.47

```
use http://www.stata.com/data/jwooldridge/eacsap/nls80,clear
list   lwage educ exp urban  iq  in 1/5
reg   lwage educ exp urban  iq    //OLS
qui   corr  lwage educ exp urban iq,covar
mat    V=r(C)      //方差阵
qui   tabstat lwage educ exp urban iq,save
mat   M=(r(StatTotal))'    //期望
clear          //注意到原始数据已被消除
corr2data x y z w v, n(935) cov(V) means(M)
                                  //生成虚拟数据
 reg   *       //回归结果与原始数据相同
 list   in 1/5
```

2 内生性的主要来源

2.1 选择性偏误 Self-selections Bias

2.2 联立因果 Simultaneous Equations Bias

2.3 遗漏变量 OVB,Omitted-Variable Bias

2.4 测量误差 Measurement Error

2.1 选择偏误

不许脱鞋！如何估计鞋的增高效应？

- 研究问题：穿上5厘米的鞋能增高多少？
- 数据限制：不允许脱鞋
  - $y=\beta x+u$
- 其中y为身高，x=1表示穿鞋，x=0表示光脚
- 若测得光脚者身高为180厘米,而穿鞋者的身高为155厘米，则鞋的增高效应β的估计将是-25厘米
- WHY？

不同的人与同一个人的两种状态

$$b = y_i^1 - y_i^0 + y_i^0 - y_j^0 = (y_i^1 - y_i^0) + (y_i^0 - y_j^0)$$

cause effection　　select bias

-25　155　180　5　-30

下标表示第i个人和第j个人

上标1表示穿鞋，0表示光脚

若个子矮的人更倾向于穿鞋，即出现选择偏误

类似案例：吃药（同一个人不能既吃又不吃）

选择性偏误如何导致内生性？

$$y = \beta x + u \xrightarrow{E(xu)=0} P\lim \hat{\beta} = \beta$$

$$\begin{aligned} y &= y_0 + (y_1 - y_0)x \\ &= y_0 + \beta x \\ &= \alpha + \beta x + y_0 - E(y_0) \\ &= \alpha + \beta x + u \end{aligned}$$

$$E(xu) \neq 0$$

- 误差项u由光脚时的身高y_0决定，一旦越矮的人越倾向于穿鞋x=1，穿鞋x就与初始身高u负相关，即存在内生性

选择性偏误

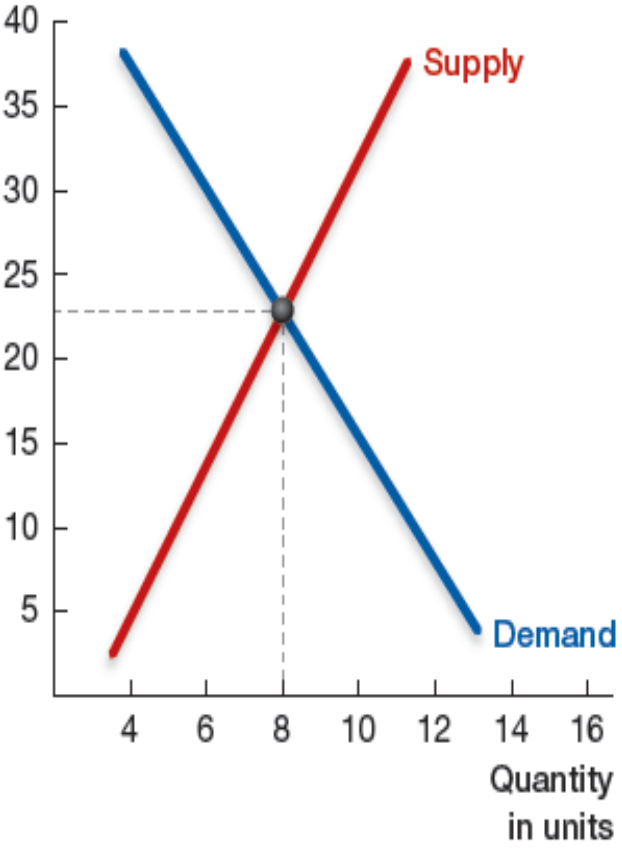
```
drawnorm u,n(100) seed(123) clear
sort   u
g  x=1 in 1/50
recode    x   .=0
scalar   b  =  1      //鞋高为1厘米
g    y= b* x+u
reg   y    x      //回归系数为-0.6165，穿鞋使人变矮 了0.6 厘米
```

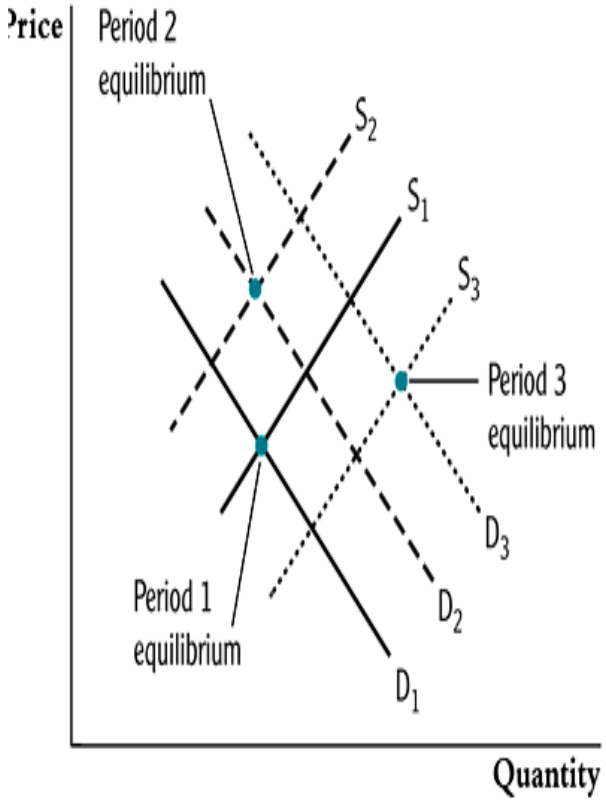
```
g    y=b* x+u

reg   y    x        //b=1，但回归系数为-0.6165

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =     23.23
       Model |  9.50118003         1  9.50118003   Prob > F        =    0.0000
    Residual |  40.0857134        98  .409037892   R-squared       =    0.1916
-------------+----------------------------------   Adj R-squared   =    0.1834
       Total |  49.5868935        99  .500877712   Root MSE        =    .63956

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.6164797   .1279121    -4.82   0.000    -.8703172   -.3626422
       _cons |   .7469744   .0904475     8.26   0.000     .5674842    .9264646
------------------------------------------------------------------------------
```

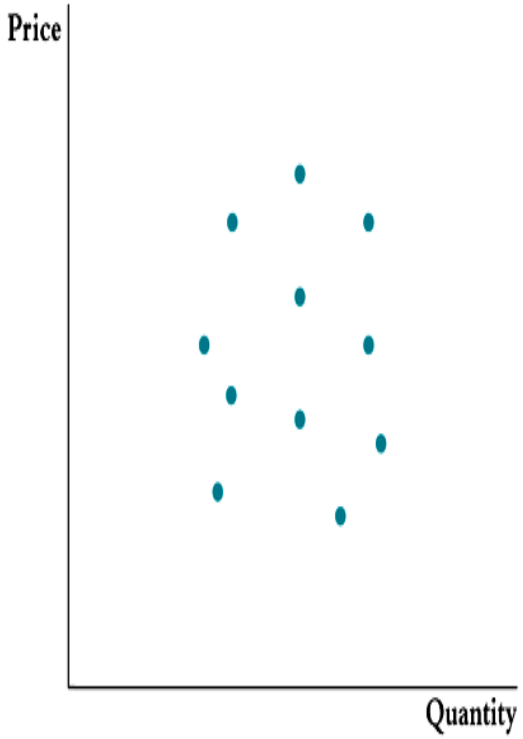
2.2 联立因果

$\ln Q = \beta \ln P + u$

β是需求弹性还是供给弹性？







联立因果为什么会导致内生性

$$y = \beta x + u \xrightarrow{E(xu)=0} P\lim \hat{\beta} = \beta$$

$$\left.\begin{aligned} y &= \beta x + u \\ x &= \alpha y + \varepsilon \end{aligned}\right\} \Rightarrow x = \alpha(\beta x + u) + \varepsilon = \alpha\beta x + \alpha u + \varepsilon \Rightarrow x = \frac{\alpha u + \varepsilon}{1 - \alpha\beta}$$

$$E(xu) = E\left(\frac{\alpha u + \varepsilon}{1 - \alpha\beta} u\right) = \frac{\alpha E(u^2) + E(u\varepsilon)}{1 - \alpha\beta} \neq 0$$

$$E(xu) \neq 0$$

联立因果模拟

```
drawnorm u v,n(1000) clear  seed(123)
scalar  b= -1   //b为需求弹性，应为负
scalar  a=2
g   x=(u-v)  /  (a-b)
g   y=b*x + u
reg   y   x          //b的估计结果显著不等于  -1
```

```
. g       y=b*x+u

. reg     y       x              //b的估计结果显著不等于1

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =     95.76
       Model |  46.2606009         1  46.2606009   Prob > F        =    0.0000
    Residual |  482.110581       998  .483076735   R-squared       =    0.0876
-------------+----------------------------------   Adj R-squared   =    0.0866
       Total |  528.371182       999  .528900082   Root MSE        =    .69504

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    .463179   .0473316     9.79   0.000     .3702981    .5560599
       _cons |   .0203864    .021979     0.93   0.354    -.022744    .0635168
------------------------------------------------------------------------------
```

2.3 遗漏变量

遗漏变量

$$y = x'\beta + z'\gamma + u, E(xu) = E(zu) = \mathbf{0}$$

$$y = x'\beta + \varepsilon, \qquad \varepsilon = z'\gamma + u$$

$$E(x\varepsilon) = E[x(z'\gamma + u)] = E(xz')\gamma$$

$$\left.\begin{array}{l} E(xz') \neq \mathbf{0} \\ \gamma \neq 0 \end{array}\right\} \Rightarrow E(x\varepsilon) \neq 0$$

遗漏变量模拟

```
use http://www.stata.com/data/jwooldridge/eacsap/nls80,clear
g  y=0.05*educ+0.02*exper+0.17*urban +.006*iq
reg  y  educ  exper  urban , r  //遗漏能力，教育回报被高估40%
```

```
g  y=0.05*educ+0.02*exper+0.17*urban +.006*iq

reg  y educ  exper  urban

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(3, 931)       =   1388.67
       Model |  25.0206656         3  8.34022186   Prob > F        =    0.0000
    Residual |  5.59148102       931  .006005887   R-squared       =    0.8173
-------------+----------------------------------   Adj R-squared   =    0.8168
       Total |  30.6121466       934  .032775318   Root MSE        =     .0775

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ |   .0714349   .0012989    55.00   0.000     .0688858    .073984
      exper |   .0202613   .0006513    31.11   0.000     .0189838     .02154
      urban |   .1703853   .0056458    30.18   0.000     .1593055   .1814652
      _cons |   .3156941   .0223184    14.15   0.000     .2718939   .3594943
------------------------------------------------------------------------------
```

“办一所名校的唯一要求是，招最优异的学生，然后让老师们远离他们。”

2.4 测量误差

因变量的测量误差

$$y = \beta x + u, E(xu) = 0$$

$$y^* = y + \varepsilon$$

$$y^* = y + \varepsilon = \beta x + u + \varepsilon, E(xu) = 0$$

$$E[x(u+\varepsilon)] \xrightarrow{E(xu)=0} = E(x\varepsilon) \begin{cases} \xrightarrow{=0} P\lim \hat{\beta} = \beta \\ \xrightarrow{\neq 0} P\lim \hat{\beta} \neq \beta \end{cases}$$

教育回报估计时，自报收入与真实收入之间存在测量误差，若测量误差与教育水平相关，如教育水平越低，越可能算不清楚他的收入，教育水平越高，越倾向低报收入。测量误差与自变量教育水平相关，导致内生性。

测量误差

```
drawnorm  x v,n(100) corr(1, .6 \ .6, 1) seed(123) clear     //因变量的测量误差与自变量相关
scalar   b= 1
g  y=b*x+invnormal(uniform())
g  ys=y+v                                  //因变量存在测量误差
reg  ys  x                                 //高估50%
```

ys	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	1.547198	.1230773	12.57	0.000	1.302955	1.791441
_cons	.0897636	.126588	0.71	0.480	-.1614462	.3409734

```
drawnorm  x v,n(100)   seed(123) clear     //因变量的测量误差与自变量不相关
scalar   b= 1
g  y=b*x+invnormal(uniform())
g  ys=y+v                                  //因变量存在测量误差
reg  ys  x                                 // 1落入置信区间【0.68，1.22】
```

ys	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.9531819	.1351688	7.05	0.000	.6849437	1.22142
_cons	.0943625	.1390244	0.68	0.499	-.1815269	.370252

自变量的测量误差

$$y = \beta x + u, E(xu) = 0$$

$$x^* = x + \varepsilon, E(x\varepsilon) = 0, E(\varepsilon u) = 0$$

$$y = \beta x + u = \beta(x^* - \varepsilon) + u = \beta x^* + u - \beta\varepsilon$$

$$E[x^*(u - \beta\varepsilon)] = E[(x + \varepsilon)(u - \beta\varepsilon)] \xrightarrow{E(xu)=E(x\varepsilon)=E(\varepsilon u)=0} = -\beta E(\varepsilon\varepsilon) \neq 0$$

自变量测量误差必导致内生性。

例：y为学习成绩，$\boldsymbol{x}$为旷课次数，当一个人很少旷课时，他所报告的旷课次数更准确，相反，随着旷课次数的增多，他能够准确回忆并报告其次数的可能性也下降，因此$\boldsymbol{x}$存在测量误差，导致内生性。

自变量测量误差

```
drawnorm  x u v,  n(100)  seed(123)   clear     //即使自变量与其测量误差不相关
scalar   b=  1
g  y=b*x+u
g   xs=x+v                                      //自变量存在测量误差
reg  y  xs                                      //低估50%
```

```
. g  y=b*x+u

. g   xs=x+v                                    //自变量存在测量误差

. reg  y  xs                                    //低估50%

      Source |       SS       df       MS      Number of obs   =       100
-------------+------------------------------   F(1, 98)        =     31.27
       Model |  49.9069834     1  49.9069834   Prob > F        =    0.0000
    Residual |  156.422488    98  1.59614784   R-squared       =    0.2419
-------------+------------------------------   Adj R-squared   =    0.2341
       Total |  206.329472    99  2.08413608   Root MSE        =    1.2634

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          xs |   .5696856   .1018805     5.59   0.000      .367507    .7718641
       _cons |   .0509968   .1273849     0.40   0.690    -.2017944    .3037879
------------------------------------------------------------------------------
```

3 应对内生性的主要方法

3.1 加入控制变量

3.2 面板数据差分估计

3.3 匹配估计

3.4 工具变量估计

3.5 断点回归

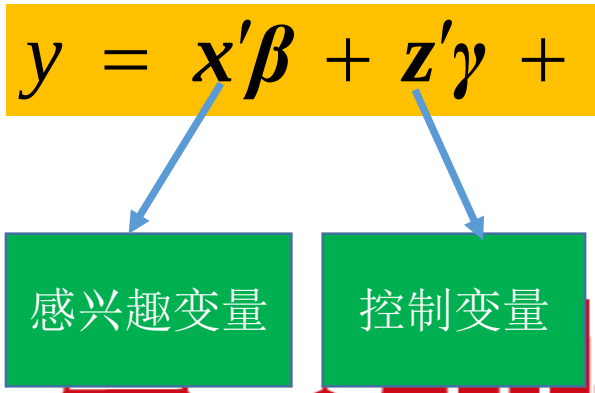
3.1 加入控制变量

关注核心自变量的斜率估计

真实的模型是不存在的,寻找“真实模型” 的目标注定是徒劳的。好的研究是由关键问题引导的，是由理论和假设所激发的，关注模型是否能回答研究问题就好。

聚焦某一个自变量，或者相关的几个自变量（核心变量/感兴趣的变量),尽最大努力，在数据收集有约束的情况下，去获得对核心变量而言“好的”斜率估计量。

$$y = x'\beta + z'\gamma + u$$



计量结果可信性吗？

- 多元回归非常流行，然而。。。。。。
- Leamer (1983)：
- “计量经济学的艺术就是，研究者在计算机终端中拟合许多(甚至上千个) 统计模型，从中选择一个或几个符合作者预期的估计结果在论文中进行报告。“
- “我们发现我们正处于一种令人沮丧和不科学的境地。没有人将数据分析看作严肃的事情; 或者，更准确地说，没有人把别人的数据分析当回事。”

敏感性分析

- 将计量经济学中的“谎言和欺骗”剔除出来——“敏感性分析”
- Sala—i—Martin (1997) 估计了200 万个包含62 个可能的解释变量的增长回归模型，其中3 个变量是主要变量(包括GDP、预期寿命和1960年的小学入学率) ，其余59 个其他解释变量的不同组合作为可能的模型设定，得到200 万个回归结果，以检验结果的稳健性。
- 但问题是，这些变量是否都是正确的控制变量? 为何选择这些变量作为控制变量? 为什么选择这样的函数形式而不是其他的函数形式? 为什么用这些观察值？

控制变量一般原则

在处理变量被决定之前就被测试到的变量，一般而言是好的控制变量，因为它们无法被处理所改变。相比之下，在处理变量被决定之后才被测度到的控制变量，可能会部分地被处理所决定，在这种情况下，这些变量其实并不是控制变量，而是处理产生的结果之一。

例如，上大学、成为白领与收入。

估计大学教育回报，不宜控制职业

```
reg income collage
```

Source	SS	df	MS	Number of obs =	6
				F(1, 4) =	0.38
Model	375000	1	375000	Prob > F =	0.5734
Residual	4000000	4	1000000	R-squared =	0.0857
				Adj R-squared =	-0.1429
Total	4375000	5	875000	Root MSE =	1000

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
collage	500	816.4966	0.61	0.573	-1766.958	2766.958
_cons	2000	577.3503	3.46	0.026	397.0187	3602.981

```
clear
input collage white income
1 0 1500
0 0 1000
0 0 2000
1 1 2500
1 1 3500
0 1 3000
end
reg income collage
reg income collage white
```

```
reg income collage white
```

Source	SS	df	MS	Number of obs =	6
				F(2, 3) =	5.06
Model	3375000	2	1687500	Prob > F =	0.1093
Residual	1000000	3	333333.333	R-squared =	0.7714
				Adj R-squared =	0.6190
Total	4375000	5	875000	Root MSE =	577.35

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
collage	9.28e-14	500	0.00	1.000	-1591.223	1591.223
white	1500	500	3.00	0.058	-91.22315	3091.223
_cons	1500	372.678	4.02	0.028	313.9723	2686.028

表 6.1 不好的控制变量如何造成选择偏误

工人类别	潜在职业		潜在收入		根据职业划分的平均收入	
	大学未毕业 (1)	大学毕业 (2)	大学未毕业 (3)	大学毕业 (4)	大学未毕业 (5)	大学毕业 (6)
始终是蓝领(AB)	蓝领	蓝领	1 000	1 500	蓝领 1 000	蓝领 1 500
可能是蓝领或白领(BW)	蓝领	白领	2 000	2 500	蓝领 1 500	白领 3 000
始终是白领(AW)	白领	白领	3 000	3 500	白领 3 000	

3.2 采用面板数据

差分估计

双重差分DID

采用面板数据

真实模型

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{\gamma} + u, E(\boldsymbol{x}u) = \boldsymbol{0}, E(\boldsymbol{z}\boldsymbol{x}') \neq \boldsymbol{0}$$

错误设定

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, E(\boldsymbol{x}\varepsilon) \neq 0$$

多期数据

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{\gamma} + u_t$$

$$y_{t-1} = \boldsymbol{x}_{t-1}'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{\gamma} + u_{t-1}$$

$$\Delta y_t = \Delta \boldsymbol{x}_t'\boldsymbol{\beta} + \Delta u$$

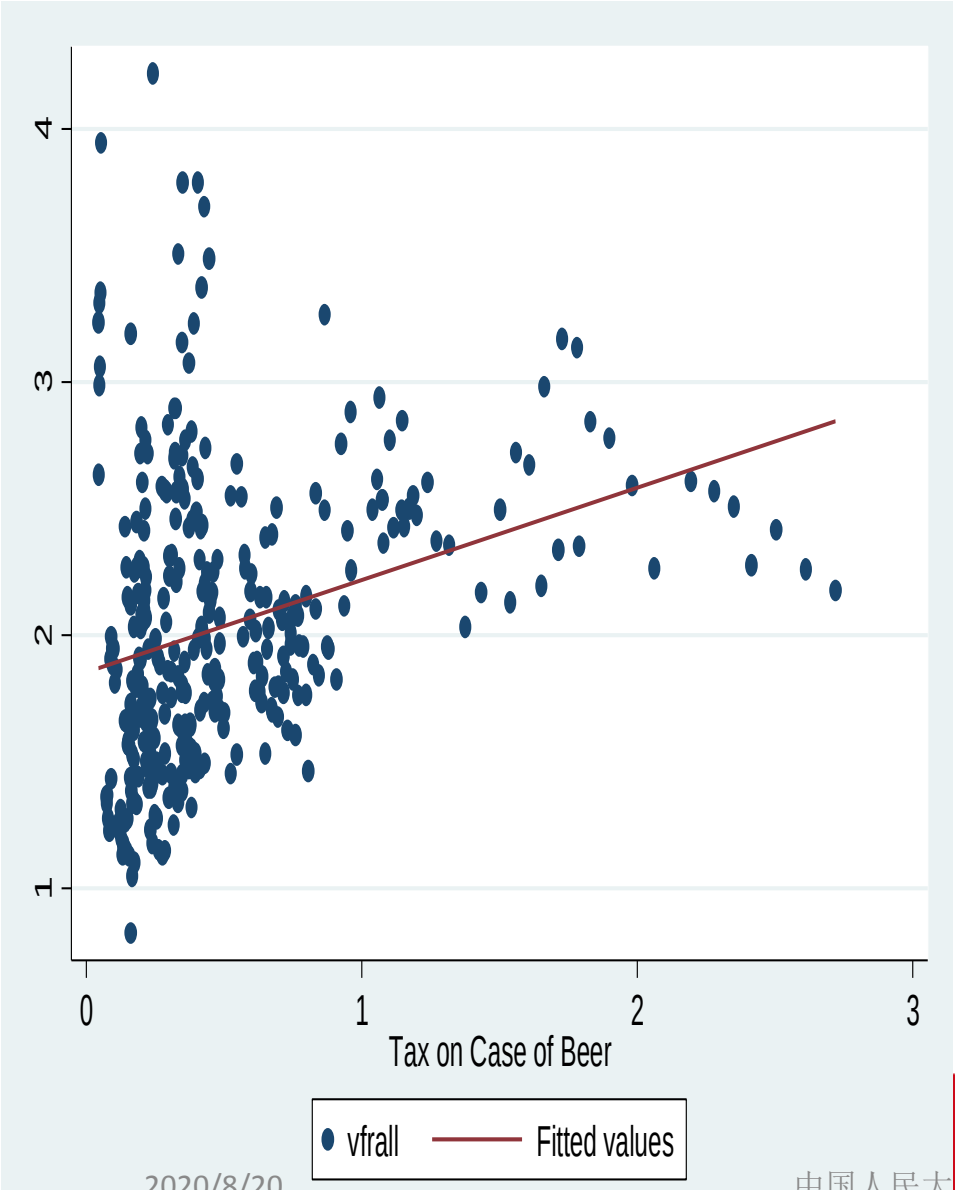
$$E(\boldsymbol{x}u) = \boldsymbol{0} \Rightarrow E(\Delta \boldsymbol{x}\Delta u) = \boldsymbol{0}$$

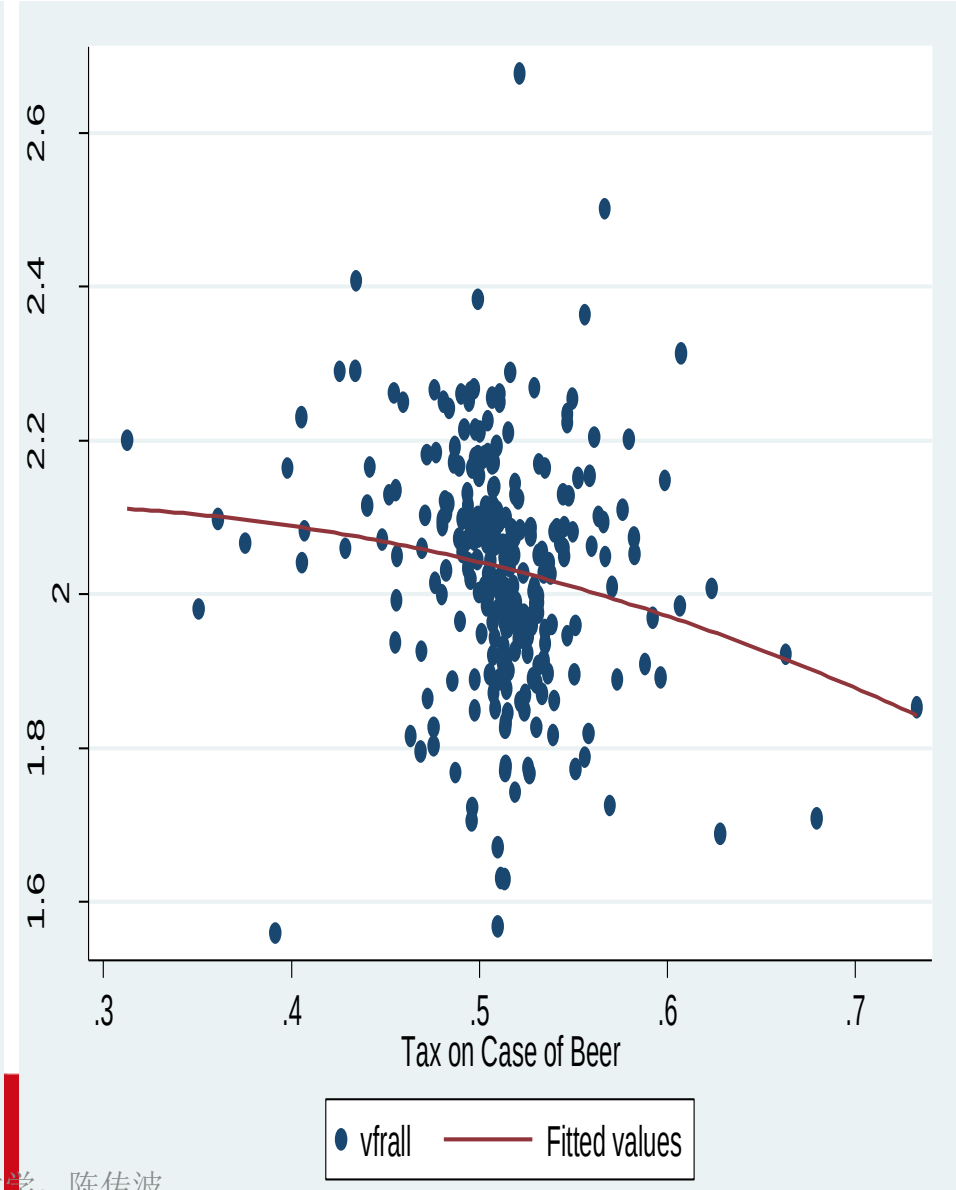
关键条件：

z的时不变性

$$E(\mathbf{x}_1 u_0) = E(\mathbf{x}_0 u_1) = 0$$

酒后驾车与交通事故死亡率





DID：双重差分

$$y_{it} = \gamma_i + \lambda_t + \beta x_{it} + u_{it}$$

$E(y \mid \gamma, \lambda)$	$\lambda_{t=0}$	$\lambda_{t=1}$	一重差分 Δ
$\gamma_{i=0}$	$\gamma_0 + \lambda_0$	$\gamma_0 + \lambda_1$	$\lambda = \lambda_1 - \lambda_0$
$\gamma_{i=1}$	$\gamma_1 + \lambda_0$	$\gamma_1 + \lambda_1 + \beta$	$\lambda + \beta$
一重差分 Δ	$\gamma = \gamma_1 - \gamma_0$	$\gamma + \beta$	双重差分 β

3.3 匹配

处理前的异质性

$$y_i^1 - y_j^0 = (y_i^1 - y_i^0) + (y_i^0 - y_j^0)$$

y	d	y1	y0	dy=y1-y0
162	1	162	160	2
170	0	172	170	2

y是我们观察到的身高，d是干预（=1表示穿鞋，0表示光脚），y1和y0是两种可能的结果，光脚时身高为y0，穿鞋时身高为y1，则dy是鞋的高度。其中有底纹的数据是无法观察到的，我们手头上只有头两行的数据

处理效应异质性

不同性别的人所穿鞋的高度不一样，女生穿高跟鞋（4cm），男生穿平底鞋（2cm）。
因个体不同处理效应不同，称为**处理效应的异质性偏差**

y	d	w	y1	y0	dy=y1-y0
164	1	60	164	160	4
170	0	40	172	170	2

平均处理效应与协变量分布相关：

当男女生各占一半时，平均增高3cm=2×0.5+4×0.5
若男生只点40%，则增高效应为3.2cm=2×0.4+4×0.6

偏误的分解

y	d	w	y1	y0	dy=y1-y0
164	1	60	164	160	4
170	0	40	172	170	2

若只能观察到y,d和w这前三列数据，彩底中的潜结果观察不到。

观察到的穿鞋者和光脚者身高差异为-6=164-170

只有观察到另外一种可能性（彩底格），才能得正确的ATE=3.2

$$y_i^1 - y_j^0 = [p(y_i^1 - y_i^0) + (1-p)(y_j^1 - y_j^0)] + [y_i^0 - y_j^0] + (1-p)[(y_i^1 - y_i^0) - (y_j^1 - y_j^0)]$$
$$= [p\beta_i + (1-p)\beta_j] + [y_i^0 - y_j^0] + [(1-p)(\beta_i - \beta_j)]$$

平均处理效应为3.2，处理前异质性偏差为-10=160-170，处理效应异质性偏差为0.8=（1-0.6）*(4-2)，三者之和恰好为-6 =3.2-10+0.8

分组随机实验仍然可能导致内生性

基于一些可观察到的**协变量x**（比如x=0或1），先将样本分到不同组。在不同组内区分处理对象和参照对象时采用的概率可以不同，即按x分成的不同组中被处理的个体所占比例各不相同。

y	d	x	w	y1	y0	dy
172	1	1	8	172	170	2
164	1	0	45	164	160	4
170	0	1	32	172	170	2
160	0	0	15	164	160	4

$$y = \beta d + u, E(du) \neq 0$$

$$E(y^0 \mid d = 0) \neq E(y^0 \mid d = 1)$$

$$166.8 = \frac{170 \times 32 + 160 \times 15}{32 + 15} \neq \frac{170 \times 8 + 160 \times 45}{8 + 45} = 161.5$$

处理d与u不独立，因更多女生（75%）穿鞋，仅20%男生穿鞋，女生平均比男生矮

条件均值独立假设（CI）成立

y	d	x	w	y1	y0	dy
172	1	1	8	172	170	2
164	1	0	45	164	160	4
170	0	1	32	172	170	2
160	0	0	15	164	160	4

$$166.8 = E(y^0 \mid d = 0) \neq E(y^0 \mid d = 1) = 161.5$$

$$E(y^0 \mid d = 1, x = 1) = 170 = E(y^0 \mid d = 0, x = 1)$$
$$E(y^0 \mid d = 1, x = 0) = 160 = E(y^0 \mid d = 0, x = 0)$$

但如果按性别先分组（相当以x为条件)，只要男生组和女生组内，穿鞋与否随机选取，则男生组内抽出组与对照组在光脚时平均身高必然相等；对女生组来说亦然。

条件均值独立性：给定的x（如固定为男生），处理与否（穿鞋否）与身高无关（即组内是随机分配的）。

匹配估计

要求得id=1的样本光脚时的身高，我们在参照组（d=0）中寻找和他的协变量（x_1=1）最为接近的样本，这个样本应该是id=3,因为x_3=1，第三个样本对应的y=170,于是我们将这个值作为id=1这个样本光脚时的身高

id	y	d	x	w	y^1	y^0
1	172	1	1	8	172	y_3=170
2	164	1	0	45	164	y_4=160
3	170	0	1	32	y_1=172	170
4	160	0	0	15	y_2=164	160

$$\beta_{POM} = E(y_x) = E[E(y_x \mid x)] = \sum_x p_x y_x = 160 \times 0.6 + 170 \times 0.4 = 164$$

$$\beta_{ATE} = E(\beta_x) = E[E(\beta_x \mid x)] = \sum_x p_x \beta_x = 4 \times 0.6 + 2 \times 0.4 = 3.2$$

$$\beta_{ATET} = E(\beta_x \mid d = 1) = \sum_x p_{x,d=1} \beta_x = 4 \times \frac{45}{53} + 2 \times \frac{8}{53} = 3.6981132$$

倾向值匹配

倾向值匹配将协变量综合成一个倾向值得分，再用该得分进行近邻匹配。

用d对协变量做probit或logit估计并预测出倾向值后，又有三种处理办法：

第一种是用倾向值作为新的协变量，用近邻匹配进行估计；

第二种是用倾向值作为权重，进行逆概率加权；

第三种是将没有匹配的样本删除，或者仅保留倾向值在某一区间内如（0.1，0.9）的样本进行简单回归估计

鞋的平均增高效应估计

```
clear
input y d x w
170 0 1 32
160 0 0 15
172 1 1 8
164 1 0 45
end
expand w     //按照权重扩展成100个观察值

*近邻匹配
teffects nnmatch (y x) (d)  //匹配估计：基于协变量的直接匹配
teffects nnmatch (y x) (d),atet  //ATET

*倾向值匹配
logit d x     //先估计一个处理d对协变量的logit函数
predict ps  //得到处理的条件概率预测Pi1=0.2, Pj0=0.75
teffects nnmatch (y ps) (d)  //以倾向值为协变量的匹配
teffects psmatch (y) (d x)  //psm匹配估计，与上述三条命令等价
teffects psmatch (y) (d x),atet  //ATET

*逆概率加权
replace ps=1-ps if d==0   //对参照组的概率是1-p; pi0=0.8,pj1=0.25
reg y d [w=1/ps]        //用逆处理或未处理概率加权回归得到
ATE=3.2
teffects ipw (y) (d x)    //逆概率加权：结果与上述三条件命令等价
```

```
*分组回归
reg y x if d
predict y0
reg y x if !d
predict y1
su y0 y1
teffects ra (y) (d)      //结果与上述五行命令相同
*回归+逆概率加权
reg y x [w=1/ps] if d
predict wy0
reg y x [w=1/ps] if !d
predict wy1
su wy0 wy1
teffects ipwra (y) (d x)      //结果与上述五行命令等价

*回归
reg  y  d  //观察到的穿鞋与光脚者的身高差异=E[y|d=1]-
E[y|d=0]=(172*8+164*45)/(8+45)-(170*32+160*15)/(32+15)=
-1.6009635
bys x:reg y d  //分层回归得到条件协处理效应，再加权
table x d,c(m y n y)  //按协变量分层估计，平均因果效应再
要用差值加权
g dx=d*x
reg y d x dx
reg y d x dx  //带协变量x的回归:相当于对bx加权，权重为
方差Pdx*（1-Pdx）
*Px,b=(2*0.2*0.8*0.4+4*0.75*0.25*0.6)/(0.2*0.8*0.4+0.75*0
.25*.06)=3.2747875,结果与真实的ATE=3.2不同，除非
pdx=0.5
```

培训对收入的影响

设　　定	对样本的完全比较			用 p 得分值筛选后的样本	
	NSW (1)	CPS-1 (2)	CPS-2 (3)	CPS-1 (4)	CPS-3 (5)
普通的比较	1 794	−8 498	−635		
	(633)	(712)	(657)		
用人口特征做控制变量	1 670	−3 437	771	−3 361	890
	(639)	(710)	(837)	(811)	(884)
				[139/497]	[154/154]
1975 年的收入	1 750	−78	−91	无观测值	166
	(632)	(537)	(641)	[0/0]	(644)
					[183/427]
用人口以及 1975 年的收入做控制变量	1 636	623	1 010	1 201	1 050
	(638)	(558)	(822)	(722)	(861)
				[149/357]	[157/162]
用人口以及 1974 和 1975 年的收入做控制变量	1 676	794	1 369	1 362	649
	(639)	(548)	(809)	(708)	(853)
				[151/352]	[147/157]

注：表格引自Angrist等《其中无害的计量经济学》

3.4 寻找工具变量

外生性

相关性

GMM

工具变量获得一致估计的原理和前提

$$y = \mathbf{x}'\boldsymbol{\beta} + u, E(\mathbf{x}u) \neq \mathbf{0} \Rightarrow \mathbf{Plim}\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$$

$$\mathbf{z}y = \mathbf{z}\mathbf{x}'\boldsymbol{\beta} + \mathbf{z}u$$

$$E(\mathbf{z}y) = E(\mathbf{z}\mathbf{x}')\boldsymbol{\beta} + E(\mathbf{z}u)$$

$E(\mathbf{z}u)=\mathbf{0}$ $\rightarrow$ $E(\mathbf{z}y) = E(\mathbf{z}\mathbf{x}')\boldsymbol{\beta}$

$E(\mathbf{z}\mathbf{x}')\neq\mathbf{0}$ $\rightarrow$ $\boldsymbol{\beta} = [E(\mathbf{z}\mathbf{x}')]^{-1}E(\mathbf{z}y)$

$$\hat{\boldsymbol{\beta}}_{IV} = (\overline{\mathbf{z}\mathbf{x}'})^{-1}\overline{\mathbf{z}y}$$

供求价格弹性估计

```
use http://people.brandeis.edu/~kgraddy/datasets/fishdata.dta,clear
*供给弹性的OLS估计，得到-0.4
reg qty price stormy mixed , r
```

qty	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
price	-.40211	.1849944	-2.17	0.032	-.7688399	-.0353802
stormy	-.2737641	.1868029	-1.47	0.146	-.6440791	.0965508
mixed	-.1061517	.1678509	-0.63	0.528	-.4388966	.2265932
_cons	8.556986	.1267244	67.52	0.000	8.30577	8.808203

```
*供给弹性的工具变量估计2SLS
ivreg qty (price= day1 day2 day3 /// day4 rainy cold)  stormy  mixed, r
```

qty	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
price	1.072254	1.507905	0.71	0.479	-1.916991	4.061499
stormy	-.9177926	.6934222	-1.32	0.188	-2.292421	.4568359
mixed	-.4540534	.4016877	-1.13	0.261	-1.250352	.3422457
_cons	9.134773	.5910338	15.46	0.000	7.963118	10.30643

Instrumented: price
Instruments: stormy mixed day1 day2 day3 day4 rainy cold

供求弹性的GMM与3SLS估计

GMM

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
price	-1.296153	.401392	-3.23	0.001	-2.082866	-.5094388
day1	-.1553275	.2098246	-0.74	0.459	-.5665761	.2559211
day2	-.3755973	.1856015	-2.02	0.043	-.7393695	-.011825
day3	-.3721443	.1932593	-1.93	0.054	-.7509256	.006637
day4	.0868227	.1742969	0.50	0.618	-.254793	.4284384
rainy	.0562628	.1424187	0.40	0.693	-.2228727	.3353984
cold	.035882	.1364588	0.26	0.793	-.2315724	.3033364
_cons	8.424108	.1817484	46.35	0.000	8.067887	8.780328
xc						
price	1.029111	1.429805	0.72	0.472	-1.773256	3.831477
stormy	-1.085186	.6486846	-1.67	0.094	-2.356584	.1862126
mixed	-.651776	.3847183	-1.69	0.090	-1.40581	.102258
_cons	9.235452	.5602337	16.48	0.000	8.137414	10.33349

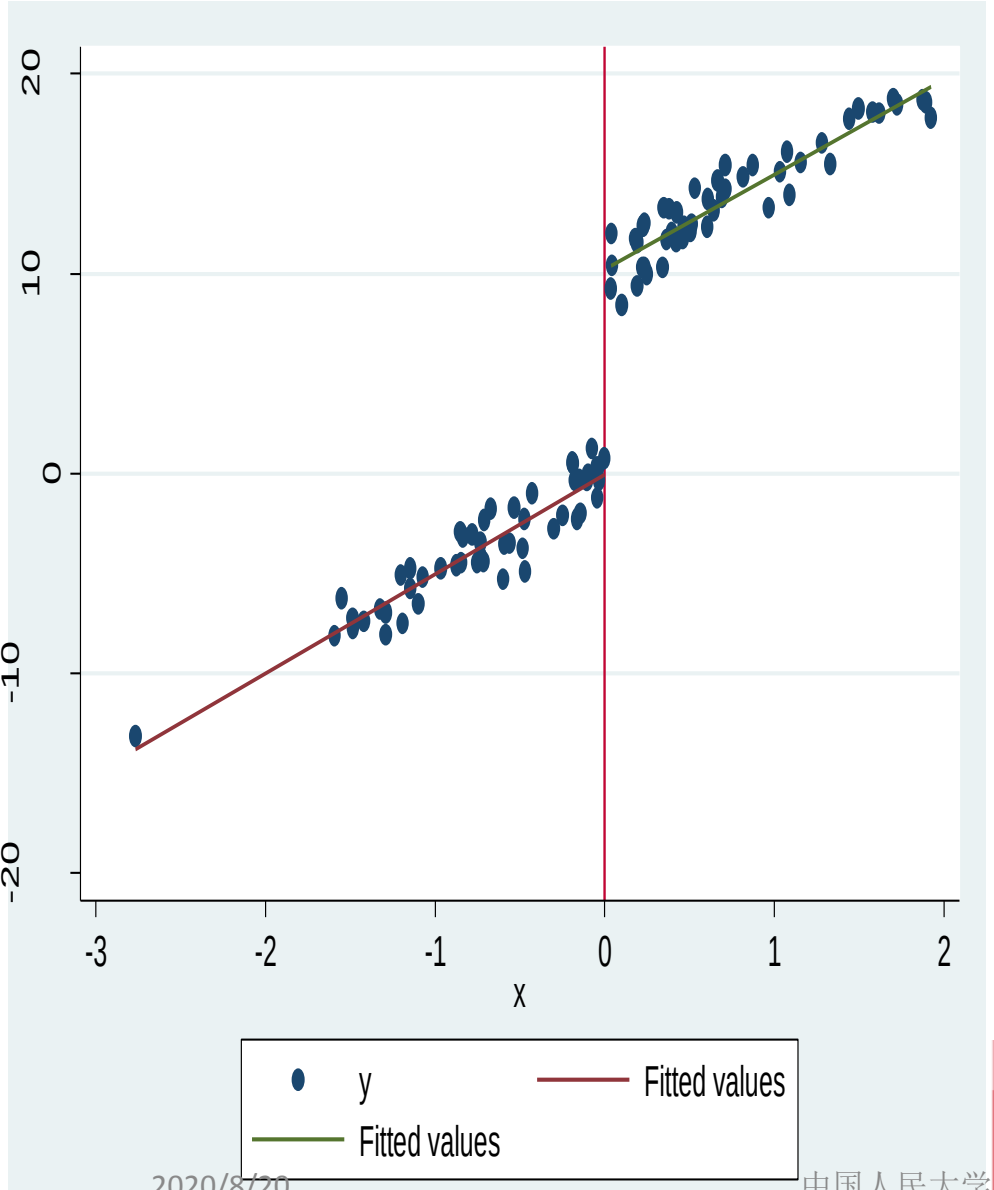
3SLS

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
qty						
price	-1.020842	.3867494	-2.64	0.008	-1.778857	-.2628274
day1	-.1150445	.1855372	-0.62	0.535	-.4786907	.2486018
day2	-.356868	.1721766	-2.07	0.038	-.694328	-.0194081
day3	-.366436	.171311	-2.14	0.032	-.7021993	-.0306726
day4	.091598	.1630694	0.56	0.574	-.228012	.4112081
rainy	.0339369	.1423345	0.24	0.812	-.2450336	.3129074
cold	.0718278	.1290774	0.56	0.578	-.1811592	.3248148
_cons	8.430028	.1588551	53.07	0.000	8.118678	8.741378
2qty						
price	1.045412	1.383568	0.76	0.450	-1.666331	3.757156
stormy	-.893848	.6353879	-1.41	0.159	-2.139185	.3514895
mixed	-.5123643	.3710589	-1.38	0.167	-1.239626	.2148978
_cons	9.140533	.5572253	16.40	0.000	8.048391	10.23267

```
use
http://people.brandeis.edu/~kgraddy/datasets/fishdata.dta

*GMM估计
gmm                                                    ///
 (qty-{xb:price day1 day2 day3 day4 rainy cold _cons})  ///
 (qty-{xc:price stormy mixed _cons}),                   ///
instr(day1 day2 day3 day4 rainy cold stormy mixed)     ///
 winitial(unadjusted, independent)

*3SLS估计
reg3 (qty price day1 day2 day3 day4 rainy cold)        ///
       (qty price stormy mixed),endog(price)
```

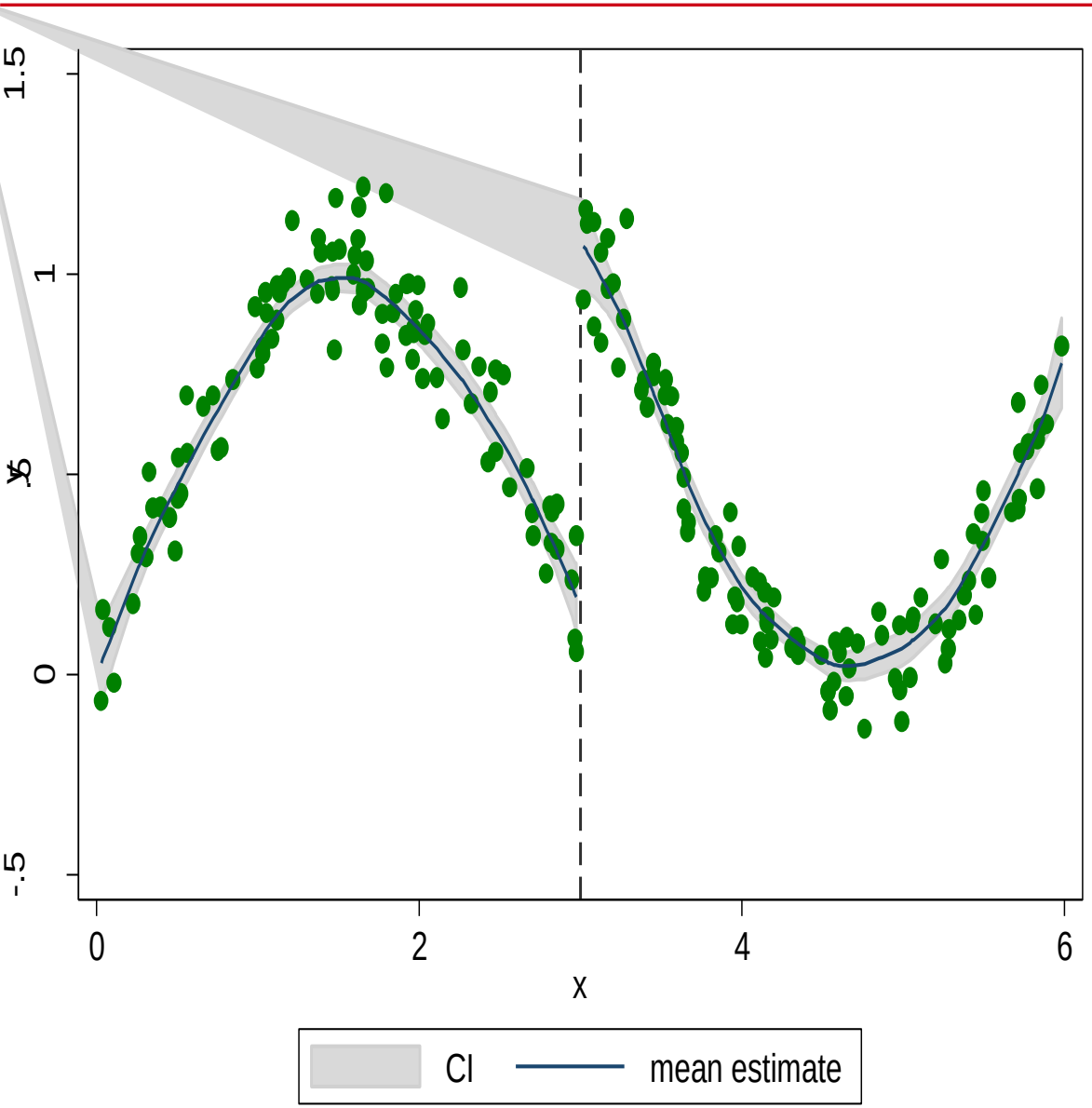
3.5 断点回归

清晰断点回归



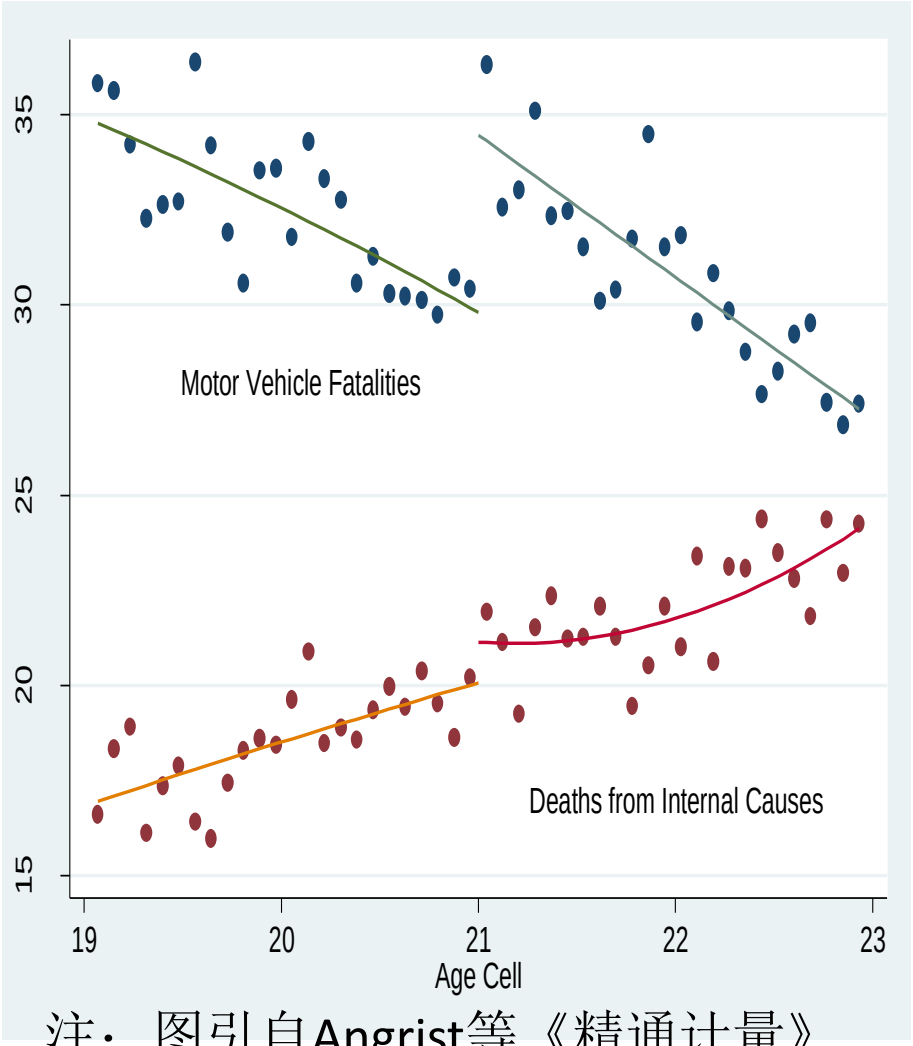

```
drawnorm x u,n(100) clear
g y=5*x+u                    //生成y
tw (sc y x) (lfit y x )
replace y=y+10 if x>0
tw (sc y x) (lfit y x if x<0)  (lfit y x if
x>0),xline(0)    //绘制图
g d=x>0              //生成虚拟变量d
reg y x d            //清晰断点回归
```

非线性断点回归




```
drawnorm u,n(200) clear
g double x = runiform() * 6
g y0= sin(x) + 0.1*u
g  d=x>3          假设断点为3
g y = d + sin(x) + 0.1*rnormal()
ssc install rdcv   //交错鉴定法
rdcv y x , thr(3) ci
tw (sc y x) (lpoly y x if x<3) (lpoly
y x if x>3)    //多项式拟合
tw (sc y x) (lowess y x if x<3)
(lowess y x if x>3)   //移动加权
```

生日与葬礼：法定饮酒年龄与死亡率

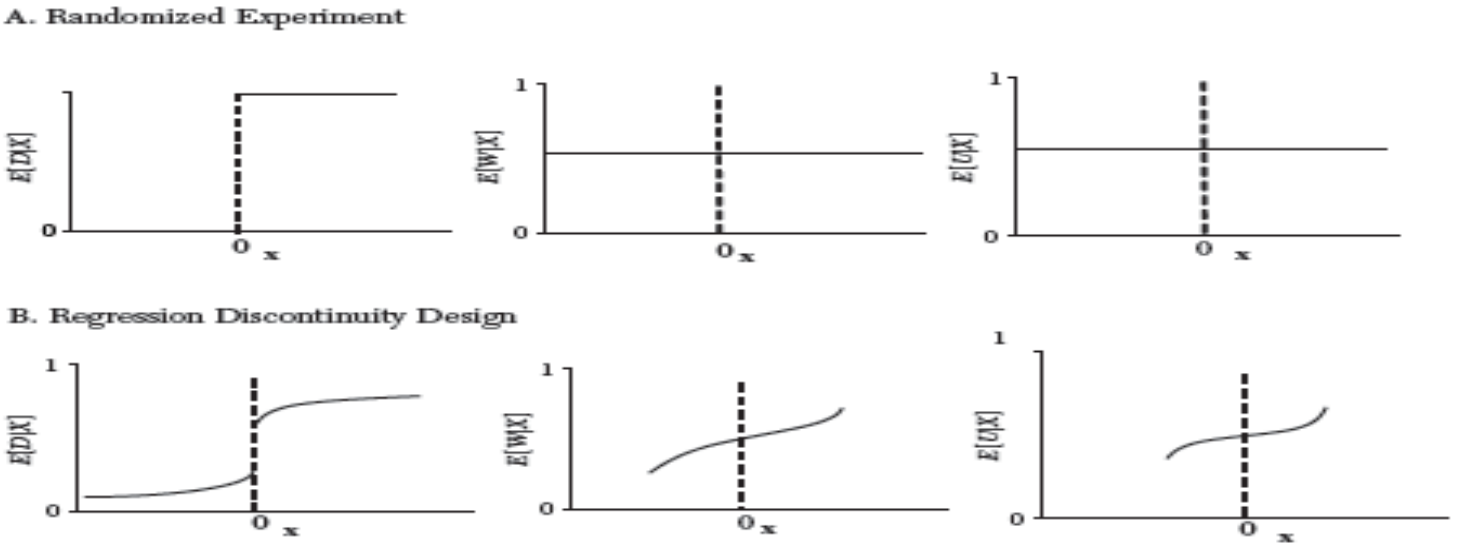


注：图引自Angrist等《精通计量》

```
u AEJfigs,clear
gen age = agecell - 21
gen over21 = agecell >= 21   //虚拟变量，相当于d
gen age2 = age^2             //二次项
gen over_age = over21*age    //交互项dx
gen over_age2 = over21*age2  //二次项与d的交互项
reg mva age age2 over21 over_age over_age2
predict exfitqi
reg internal age age2 over21 over_age over_age2
predict infitqi
twoway (scatter  mva internal agecell) ///
(line exfitqi infitqi agecell if agecell < 21) ///
 (line exfitqi infitqi agecell if agecell >= 21), ///
   legend(off) text(28 20.1 "Motor Vehicle Fatalities") ///
    text(17 22 "Deaths from Internal Causes")
```

四种主要方法对比

四种方法对比



Y：因变量

D：处理变量

X：分组变量

W：协变量

U：误差项

Z:工具变量