因果推断中的控制变量

- 好的控制变量
- 坏的控制变量
- 中性的控制变量
- 后门准则
- 应用

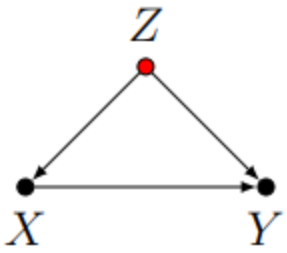
连玉君 (中山大学)
arlionn@163.com

北京友万信息科技有限公司
www.uone-tech.cn

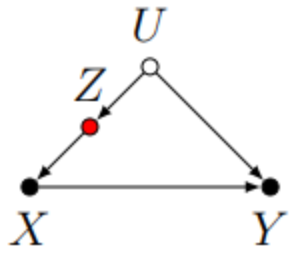
A. 好的控制变量

A1. 共同原因

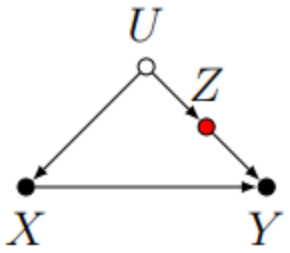
当 Z 作为共同原因或者共同原因的派生变量时，控制 Z 可以阻断虚假的因果路径。



(a) Model 1



(b) Model 2



(c) Model 3

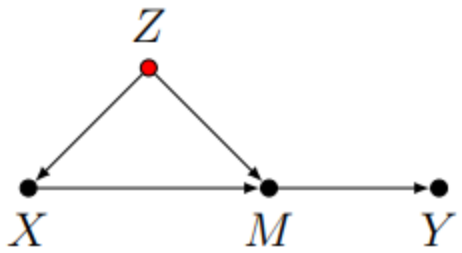
- 模型 1 中 Z 是共同原因，因此必须控制在模型中。
- 模型 2 或模型 3 中，Z 并不是传统意义上的混淆因素，但控制 Z 可以切断来自不可观测因素的混淆。

$$X = Z + u \quad \Rightarrow \quad Z = X - u$$
$$Y = Z + X + e \quad \Rightarrow \quad Y = 2X + e - u$$

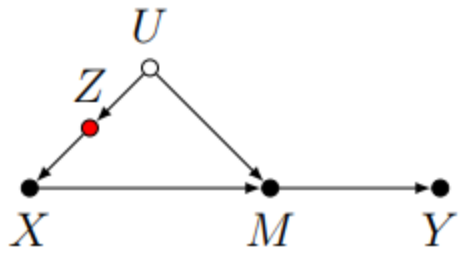
```
*-共同原因  X <-- Z --> Y

  clear
  set seed 135
  set obs 30
  gen Z = _n
  gen X = 1*Z + rnormal()
  gen Y = 1*X + 1*Z + 0.1*rnormal()

  eststo m1: reg Y X
  eststo m2: reg Y Z
  eststo m3: reg Y X Z

  esttab m1 m2 m3 , nogap compress
------------------------------------------------
                (1)         (2)         (3)
                  Y           Y           Y
------------------------------------------------
X             1.985***                  1.008***
             (79.98)                   (69.57)
Z                           1.995***    0.990***
                           (78.20)     (68.02)
_cons          0.309       0.0109      0.0245
              (0.70)       (0.02)      (0.71)
------------------------------------------------
N                30          30          30
------------------------------------------------
```

A2. 带有中介的共同原因

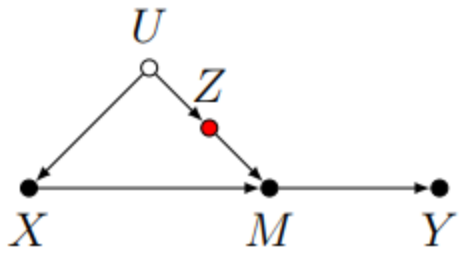
如果模型中同时存在共同原因和中介关系，那么同样必须阻断后门路径。



(a) Model 4



(b) Model 5



(c) Model 6

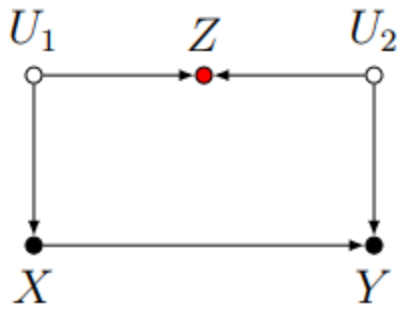
以上三个模型中同时包含了中介关系和共同原因

- 以模型 4 为例，其后门路径为 $X \leftarrow Z \rightarrow M \rightarrow Y$。
- 在模型 5 和模型 6 中，Z 是共同原因 U 的派生变量，因此同样可以阻断后门路径。

B. 坏的控制

B1. M 偏误 (共同结果)

该模型中，变量 Z 同时与处理变量和结果变量相关，因此其被称为 “预处理” 变量。尽管在传统的计量经济学中认为 Z 是一个好的控制，但实际上可能会打开一条后门路径 $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$，这种坏的控制称为 M **偏误。**



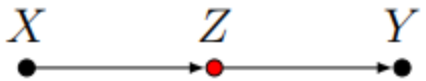
$$Z = 0.5X + Y + e,\ \ corr(X, Y) = 0$$

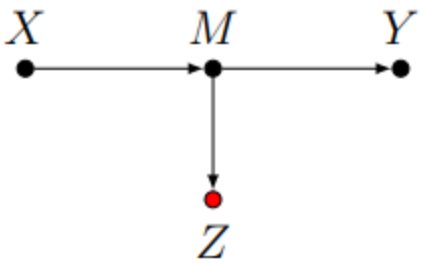
$$Y = -0.5X + Z - e$$

```
  clear
  set seed 135
  set obs 30
  gen X = rnormal()
  gen Y = rnormal()
  gen Z = -0.5*X + Y + 0.2*rnormal()

.   esttab m1 m2 m3 m4, nogap r2 compress

------------------------------------------------------------
                (1)          (2)          (3)          (4)
                  Y            Y            X            Y
------------------------------------------------------------
X             0.0840                                0.491***
              (0.58)                              (12.76)
Z                          0.722***    -0.572**      1.003***
                          (6.93)      (-2.91)      (21.95)
_cons         0.0234      -0.0495      -0.0223      -0.0386
              (0.14)      (-0.49)      (-0.12)      (-0.99)
------------------------------------------------------------
N                 30           30           30           30
R-sq           0.012        0.632        0.232        0.948
------------------------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

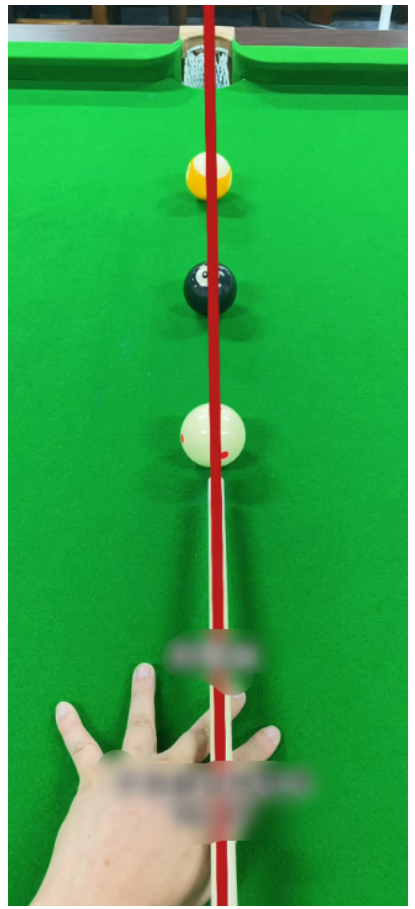
B2. 阻断正确路径

在因果推断中，一方面我们想要剔除所有可疑的路径，另一方面也要注意不能阻断正确的因果路径。下面两个模型显示了阻断因果路径的坏控制：



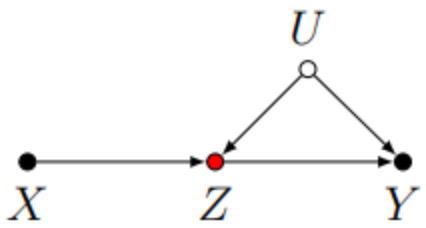


在这两个模型中，Z 分别作为中介变量和中介变量的派生变量，因此在模型中加入 X 之后，会完全和部分阻断正确的因果路径，导致不一致的估计。



B3. 打开混淆路径

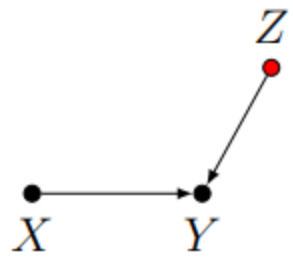
对具有中介变量的模型稍加改动。假设存在不可观测的因素 U 作为 Z 和 Y 的共同原因。此时路径 $X \to Z \leftarrow U \to Y$ 被 Z 这一共同结果阻断，加入 Z 之后反而会打开该路径。



C. 中性的控制

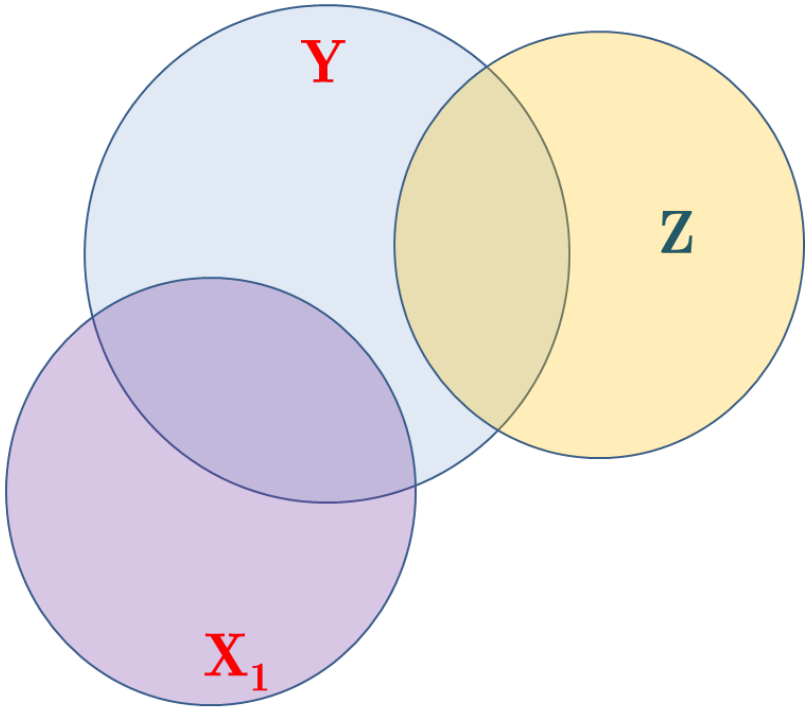
C1. 可能提高精度的情形

在很多情形下，加入某些控制变量是无害的，但也无法提供更多因果信息。例如在以下模型中，Z 并没有混淆因果关系，也没有阻断可疑的因果路径，因此 Z 是一个中性的控制。但加入 Z 之后，因果关系估计的标准误会下降，因此 Z 能够改善 ACE 的估计精度。



Q. 如果控制变量只与 Y 相关，而与 X 无关会怎样?

- **A.** 标准误变小 (t 值变大), 系数估计值不受影响
- Note: 这种情况其实很少发生

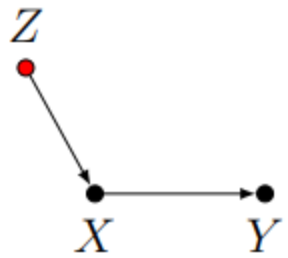


> 延伸阅读：Schjoedt, Bird, Carsrud, Brännback (2014, PDF)

C2. 可能降低精度的情形

与第一种情形相反，在下面的模型中虽然控制 Z 也不会影响从 X 到 Y 的因果关系，但是此时会放大 ACE 的估计方差，降低估计的精度。

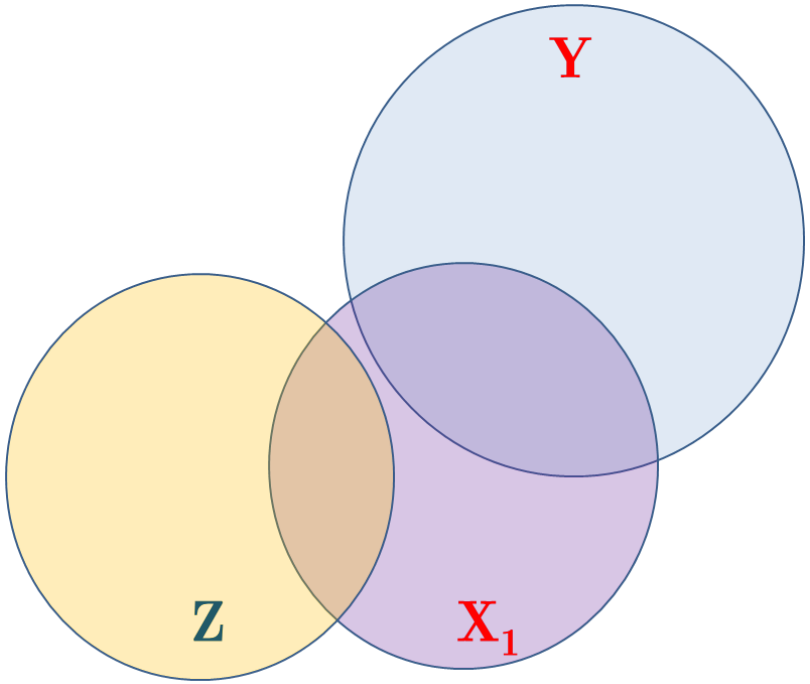
可见 X 的父变量会损害估计精度，而 Y 的父变量则会提高估计精度。



要注意的是，该模型与偏误放大情形非常类似，唯一的区别在于该模型中不存在与 X 和 Y 同时相关的不可观测因素。

Q. 如果控制变量只与 X 相关，但与 Y 无关会怎样?

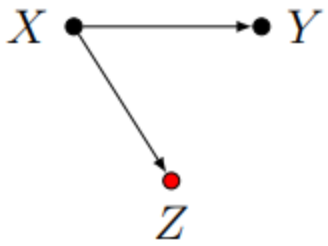
- A. 标准误变大 (t 值变小), 系数估计值不受影响 (Z 是冗余变量)

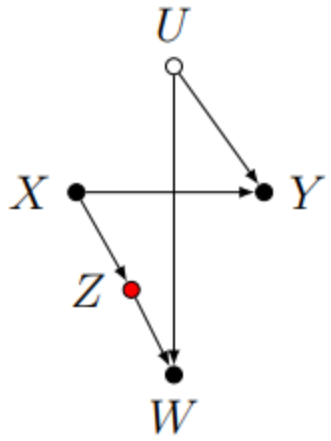


> 延伸阅读：Schjoedt, Bird, Carsrud, Brännback (2014, PDF)

C3. 可能缓解选择偏误的情形

与传统经济学不同，并非所有 “处理后” 变量都是坏的控制。在以下的两个模型中，Z 的加入并未打开混淆路径。

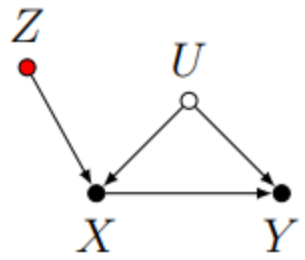




在这两个模型中，加入 Z 都会降低 X 的方差，因而损害估计的精度。但在右边的模型中，控制 Z 可以缓解关于 W 的选择偏误。

C4. 偏误放大

另一种关于 “预处理” 的控制是加入影响处理变量的因素。在这一情形下，不但无法分离出真实的因果效应，还会放大本身存在的偏误。



C. 理论基础

C1. 后门路径 (Back door path)

定义：**后门路径**：在连接处理变量 (X) 和结果变量 (Y) 的任何箭头序列 (无论其方向如何) 中，如果删除从 X 发出的箭头，某个序列依然存在，则称其为后门路径 (Pearl，2000)。

用途：切断所有后门路径。我们可以通过在回归模型中加入适当的控制变量来切断所有后门路径，以便识别 $X \rightarrow Y$ 的因果关系。

> **Source:** Hünermund P, Louw B. On the Nuisance of Control Variables in Regression Analysis[J]. arXiv preprint arXiv:2005.10314, **2022**. **R&R at Organizational Research Methods**, -PDF-

C2. 后门准则

因果图揭示了何种 $\mathbf{Z}$ 的设定会阻断正确的因果路径，我们需要做的是选择 $\mathbf{Z}$，以保证：

- 阻断所有虚假的路径；
- 避免阻断或部分阻断真实的因果路径；
- 避免打开其他虚假的路径。

C3. 三种因果关系

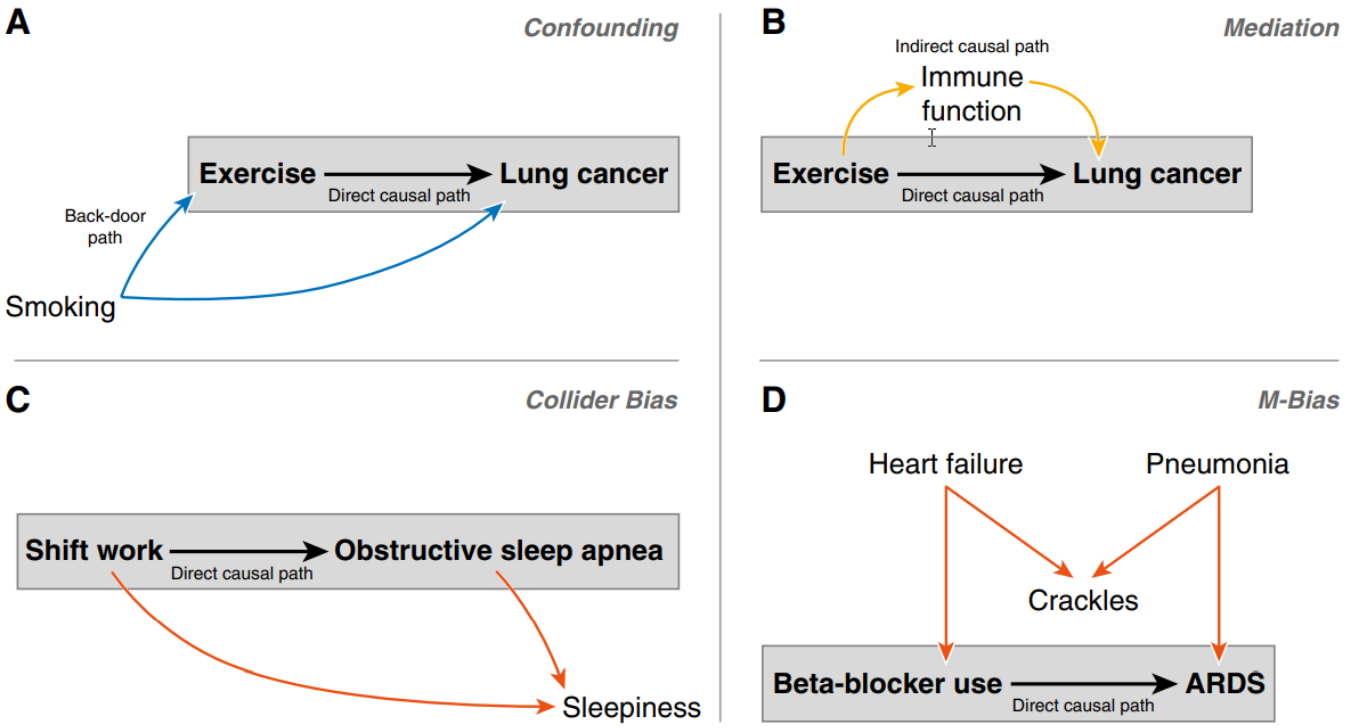
在一般的因果图中，需要理解三种重要的因果关系：

- 中介 (Chains)：中介指的是路径 $X \to Z \to Y$，即 X 对 Y 的因果影响是通过 Z 实现的。在方程中控制 Z 会阻断这一联系；
- 共同原因 (Forks)：共同原因指的是路径 $X \leftarrow Z \to Y$，即 Z 同时影响 X 和 Y。因此二者间存在非因果路径，在方程中控制 Z 会阻断这一联系；
- 共同结果 (Coliders)：共同结果指的是路径 $X \to Z \leftarrow Y$，这一路径本身是关闭的，但如果我们在方程中控制了 Z，则会打开这一非因果路径。

需要注意的是，控制某一变量的派生变量也视为部分控制了该因素。现在我们可以判断当以 $\mathbf{Z}$ 为条件时，路径 p 是否被阻断：

- 当路径是中介或共同原因时，$\mathbf{Z}$ 中会纳入中间节点能够阻断路径 p；
- 当路径是共同结果时，$\mathbf{Z}$ 中既不包含中间节点，也不包含其结果，则能够阻断路径 p。

C4. 简例

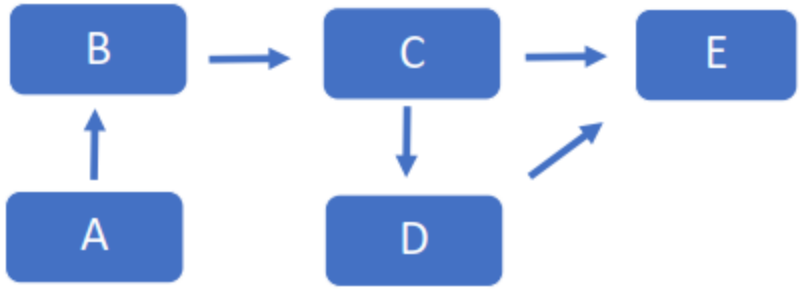


> Source: Control of Confounding and Reporting of Results in Causal Inference Studies: Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. -PDF-
> Kohler, Ulrich, Sawert, Tim, & Class, Fabian (2022). Replication material to "Control variable selection in applied quantitative sociology: A critical review". -Link-

D. 控制变量的选择与研究目标有关

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

构建因果模型时，可以借助因果图来识别遗漏变量的来源、理解因果关系，从而设定合理的模型。
我们用以下因果图来示例变量之间的因果关系：



- 如果关注 D → E 的影响，则 C 作为混杂因素，必须要控制。
- 如果关注 C → E 的影响，则 D 不应该被控制，因为 D 是 C 和 E 之间关系的中介。
  - Q：此时，A 和 B 是否是好的控制变量？

实际上，回归不能告诉我们是 C 导致 E 还是 E 导致 C，回归只是在给定条件下对相关性的估计，只有通过理论分析才能论证因果关系。

例子：工资的性别差异

假设我们想了解工资中性别差距的程度：

> $$wage_{it} = a + \beta Female_{it} + u_{it}$$

Q： 如何选择控制变量以克服遗漏变量偏误？

可能的控制变量：

- 工作经验 (Exp)、教育水平 (Edu)、职业、失业率 ($Unemp$)、住江景房、...

Q： 哪些是 Bad Controls？

例子：工资的性别差异 (Cont.)

> $wage_{it} = a + \beta Female_{it} + u_{it}$

- Exp, Edu
  - Good Controls：它们与性别相关，也会影响收入水平。
  - Bad Controls：如果我们认为性别导致教育或经验的变化，则二者便是 Bad Control。
- 控制您所在地区的失业率 ($Unemp$) 可能是 Bad control。
  - $Unemp$ 肯定会影响工资，但与性别无关，除非某些地区有更多/更少的女性。
  - 控制 λ_t (`i.year`) 是更好的选择，有助于控制宏观经济因素导致的年度差异
- 控制职业 (`i.occupation`) 是一种 Bad Control，因为它在很大程度上直击受性别影响，即 $Female \to Occupation \to Wage$

识别好和坏的控制变量

Angrist 和 Pischke (2009) 提供的**经验法则**：

- “好” 的控制变量是指在确定处理变量 X 时就已经固定的变量
- “坏” 的控制变量则是那些本身就是结果变量的变量。

几个例子

- 铲球次数 (Source: Hastie, 2021)
  - Y: 一名足球运动员在一个赛季中的铲球次数；W 和 H 是他的体重和身高。
  - 回归结果为 $\widehat{Y} = b_0 + 0.5W - 0.1H$。如何解释 $\hat{\beta}_2 = -0.1 < 0$？
- 若研究啤酒税 (Tax) 对交通死亡人数 ($Death$) 的影响，如下模型设定可行吗？ -Lk-

$$Death = \beta_0 + \beta_1 Tax + \beta_2 BeerSales + \dots$$

- 若研究气温 (Temp) 对暴力冲突 (Conflict) 的影响，如下模型是否可行?

$$Conflict = \beta_0 + \beta_1 Temp + \beta_2 Income + \dots$$

Bad control：模拟分析 A

- Literaturn on temprature, income and conflict
  - Dell et al. (2012, PDF): $Temp \xrightarrow{(-)} Income$
  - Miguel et al. (2004, PDF, Codes, cite) $Income \xrightarrow{(-)} Conflict$
  - $\underbrace{Temp}_{X} \xrightarrow{(-)} \underbrace{Income}_{W} \xrightarrow{(-)} \underbrace{Conflict}_{Y}$ or $X \longrightarrow W \longrightarrow Y$
- DGP1: $X \longrightarrow W \longrightarrow Y$
  - $temp \sim N(0,1)$, $e_1 \sim N(0,1)$, $e_2 \sim N(0,1)$
  - $income = \beta_1\, temp + e_1 \quad (\beta_1 = -0.5)$
  - $conflict = \beta_2\, income + e_2 \quad (\beta_2 = -1.0)$
  - Total effect: $\partial Y / \partial X = \beta_1 \times \beta_2 = -1.0 \times (-0.5) = +0.5$

```
clear
set seed 123
set obs 1000
gen temp = rnormal()  //temperature variable
gen  e_1 = rnormal()  //noise 1
gen  e_2 = rnormal()  //noise 2, uncorrelated with e_1
gen   income = -0.5*temp   + e_1
gen conflict = -1.0*income + e_2

eststo m1: reg conflict temp
eststo m2: reg conflict income
eststo m3: reg conflict income temp
esttab m1 m2 m3, nogap r2
```

```
conflict        (1)             (2)             (3)
------------------------------------------------------------
temp          0.522***                        0.0263
            (11.34)                         (0.71)
income                        -0.997***      -0.986***
                             (-34.33)       (-30.50)
_cons       -0.0613          -0.0406         -0.0413
            (-1.37)          (-1.26)         (-1.28)
------------------------------------------------------------
R-sq          0.114            0.542           0.542
```

- Dell, M., B. F. Jones, B. A. Olken, **2012**, Temperature shocks and economic growth: Evidence from the last half century, **American Economic Journal-Macroeconomics**, 4 (3): 66-95. -Link-, -PDF-, PDF2, Replication, -cited-
- Miguel, E., S. Satyanath, E. Sergenti, **2004**, Economic shocks and civil conflict: An instrumental variables approach, **Journal of Political Economy**, 112 (4): 725-753. -Link-, -PDF-, -Slides-, Replication, -cited-
    - Miguel, E., S. Satyanath, **2011**, Re-examining economic shocks and civil conflict, **American Economic Journal: Applied Economics**, 3 (4): 228-232. -Link-, -PDF-, Replication, -cited-

Bad control? 模拟分析 B

- DGP2: $Y \longleftarrow X \longrightarrow W \longrightarrow Y$
  - $conflict = -1.0\, income + 0.4\, temp + e_2$,
  - $income = -0.5\, temp + e_1$
- Note: 总效应 = $\underbrace{-0.5 \times (-1.0)}_{\text{直接效应}} + \underbrace{0.4}_{\text{中介效应}} = 0.9$

```
conflict2              (1)              (2)              (3)
-------------------------------------------------------------
temp                0.922***                          0.426***
                   (20.03)                           (11.55)
income                                -1.151***       -0.986***
                                     (-37.25)        (-30.50)
-------------------------------------------------------------
R-sq                0.287            0.582            0.631

. reg income temp // regfit
                      income =  0.02 - 0.50*temp
                               (0.03) (0.03)
```

进一步讨论

- Income 能否作为控制变量?
  - *... variables such as income both affect and are affected by civil conflict, including them directly as regressorsis problematic* (Miguel et al., 2004, PDF; Burke et al., 2010)
- temperature 是否会直接影响 conflict ?
  - Hsiang, S., **2016**, Climate econometrics, **Annual Review of Resource Economics, Vol 8**, 8: 43-75. -Link-, -PDF-
  - Burke, M., S. M. Hsiang, E. Miguel, **2015**, Climate and conflict, **Annual Review of Economics**, 7 (1): 577-617. -Link-, -PDF-, PDF2, -cited-

后门准则应用：不同的控制方案

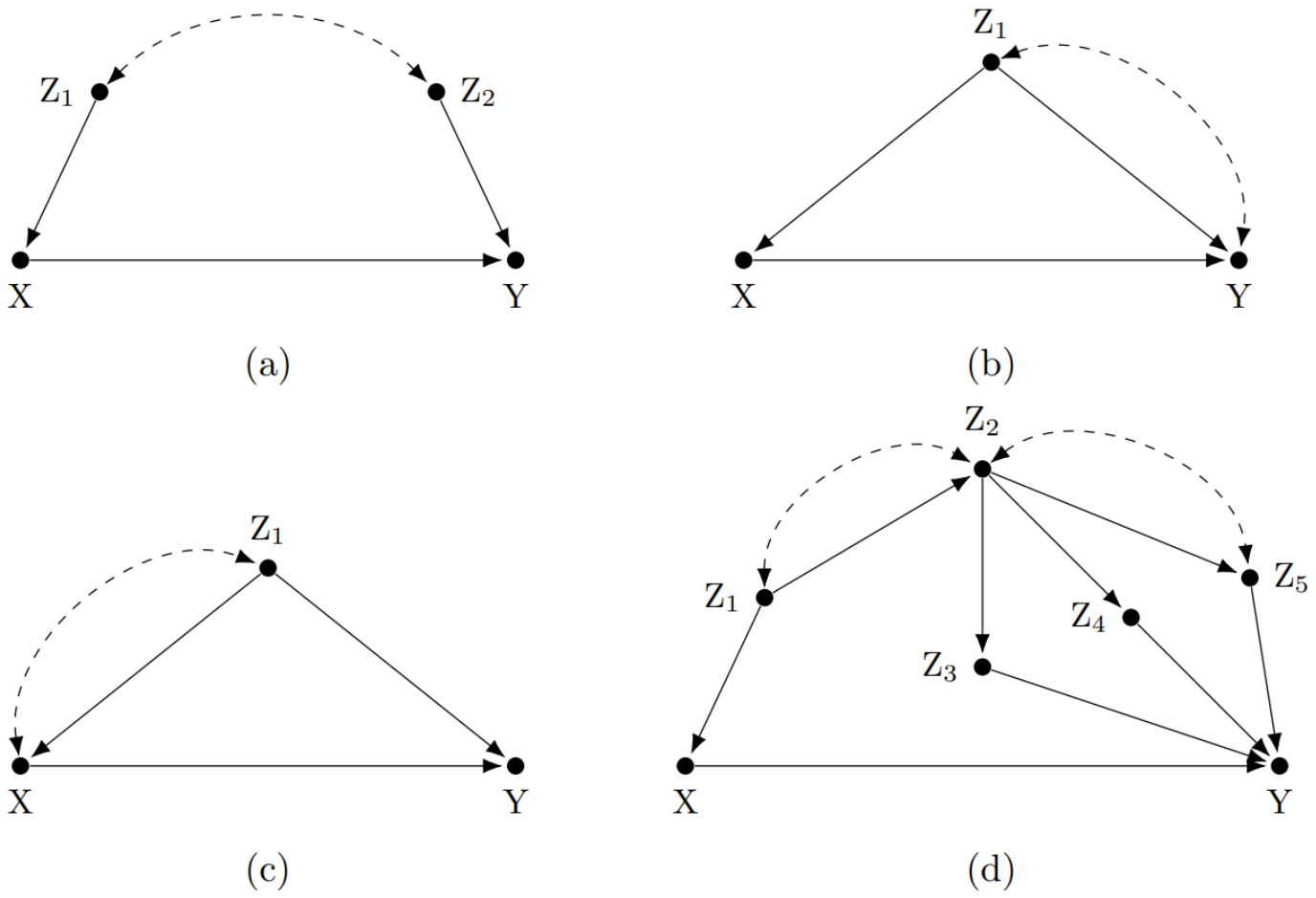


Figure 1: Examples of causal diagrams with valid control variable Z_1

Table 3: OLS regressions with varying adjustment sets

	Figure 1a			Figure 1b	Figure 1c	Figure 1d		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
X	1.017	1.004	1.015	0.993	1.001	0.991	1.006	1.003
	(0.015)	(0.006)	(0.010)	(0.012)	(0.008)	(0.057)	(0.007)	(0.010)
Z_1	0.499		-0.019	1.503	1.004	4.565		0.004
	(0.018)		(0.013)	(0.014)	(0.014)	(0.069)		(0.016)
Z_2		0.993	0.997					0.009
		(0.008)	(0.008)					(0.019)
Z_3							0.994	0.991
							(0.008)	(0.010)
Z_4							0.991	0.988
							(0.008)	(0.010)
Z_5							1.011	1.009
							(0.006)	(0.008)

FAQs

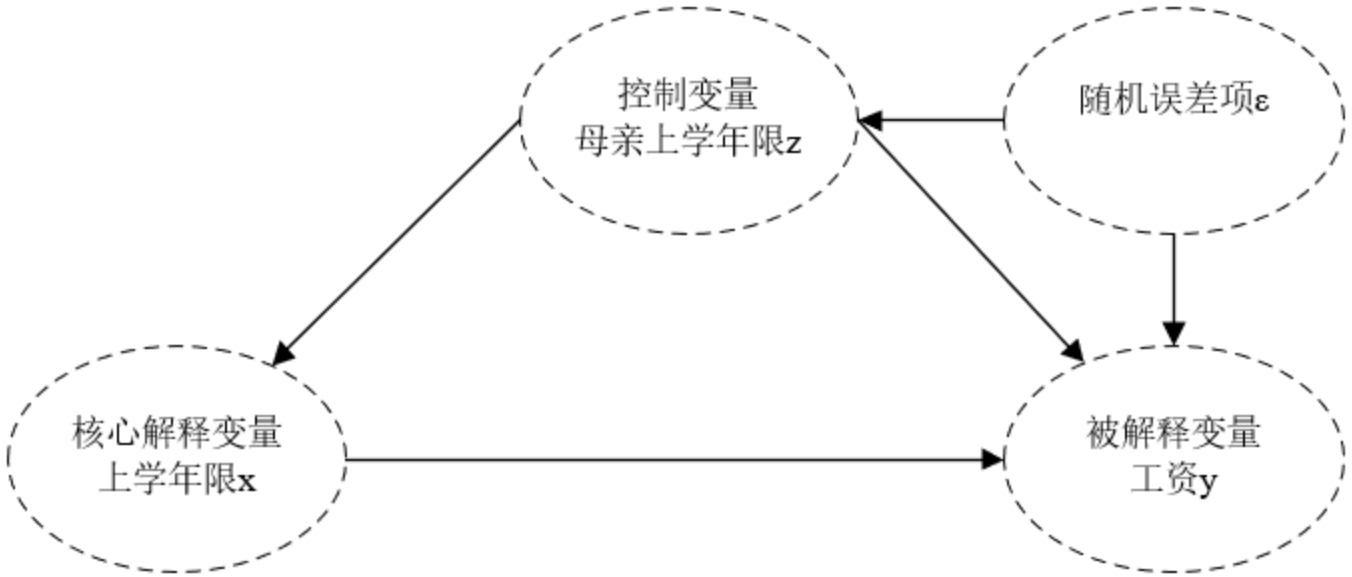
是否需要展示和讨论控制变量的经济含义?

- 不用太关心控制变量的系数!
- 虽然控制变量对于因果关系的识别至关重要，但其本身通常不具有结构性解释。即使是有效的控制变量，也常常会与其他未观察到 (或不能观测到) 的因素（unobserved factors）关联，从因果推断的角度来看，这使得它们的边际效应无法解释 (Westreich 和 Greenland，2013； Keele等，2020)。
- 因此，在因果识别和因果推断中，可以不必 (也很难) 讨论控制变量的系数含义。

控制变量可以是内生变量吗?

> Yes!

假设真实的因果模型如下图所示:



在如下回归模型中：

$$y = \beta_0 + \beta_x x + \beta_z z + \varepsilon$$

- β_x 可以作为因果解释：给定母亲的上学年限相同，平均每多上一年学能增加工资 β_x。这是因为母亲上学年限 z 作为控制变量之后，x 和 ε 就不再相关，$E(x\varepsilon|z) = 0$，β_x 无偏；
- β_z 只能作为相关解释：不能说母亲每多上一年学，就能增加子女的工资 β_z，只能说母亲的教育水平和子女的工资是正相关的。这是由于母亲上学年限在 ε 里面导致 z 与 ε 相关，$E(z\varepsilon|x) \neq 0$，β_z 有偏。
- **传统视角下的疑问**：控制变量 z 与扰动项 ε 相关。这显然不符合 OLS 的基本假设：所有的解释变量都与随机误差项不相关。否则，它们的估计值都不一致。

为何在因果框架下可以放松这一假定，而允许控制变量 z 和扰动项 ϵ 相关呢?

启示

- 在因果推断中，我们关注的是 X 的系数，此时可以「牺牲」控制变量 Z 系数的无偏性，这意味着：
  - 不必做过多讨论控制变量的系数，因为多数情况下，控制变量的系数估计值是有偏的。
  - 要借助理论分析和因果图 (DAG) 来辅助模型设定，说清楚上述设定思路
- 在因果框架下，我们通常只对回归方程中的一个核心解释变量感兴趣，特别希望得到对其系数的一致估计，并将其解释为核心变量对被解释变量的因果效应。
- 对于方程中的其他变量并无太大兴趣，之所以把它们放入回归方程，只是为了 “控制” 那些对被解释变量有影响的遗漏因素来避免 “遗漏变量偏差”。即使对控制变量系数估计不一致，我们也可以接受。

> 详见：Stata：控制变量与核心解释变量地位对等吗？。给出了上述处理方法的理论依据和证明过程。

IV 估计中的控制变量

工具变量需要满足的条件：

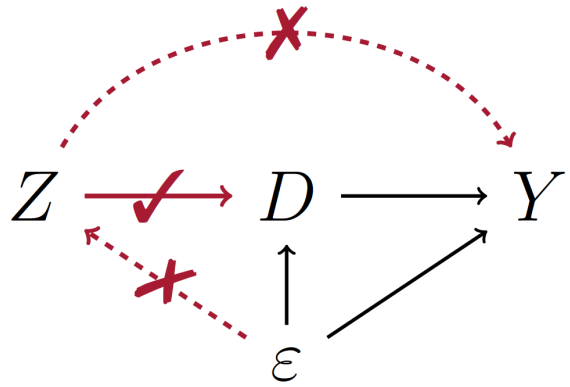
- 相关性：工具变量与内生解释变量相关。

$$\mathrm{Cov}(Z, D) \neq 0$$

- 外生性或独立性：工具变量与扰动项不相关。

$$\mathrm{Cov}(Z, \varepsilon) = 0$$

- 🍎 **排斥性约束**：工具变量只通过 X 或其他变量影响 Y，但不直接影响 Y。
  - 换言之，Z 不直接出现在结构方程右边。
- 教育回报率例子中的 Z 有哪些可能的选择?

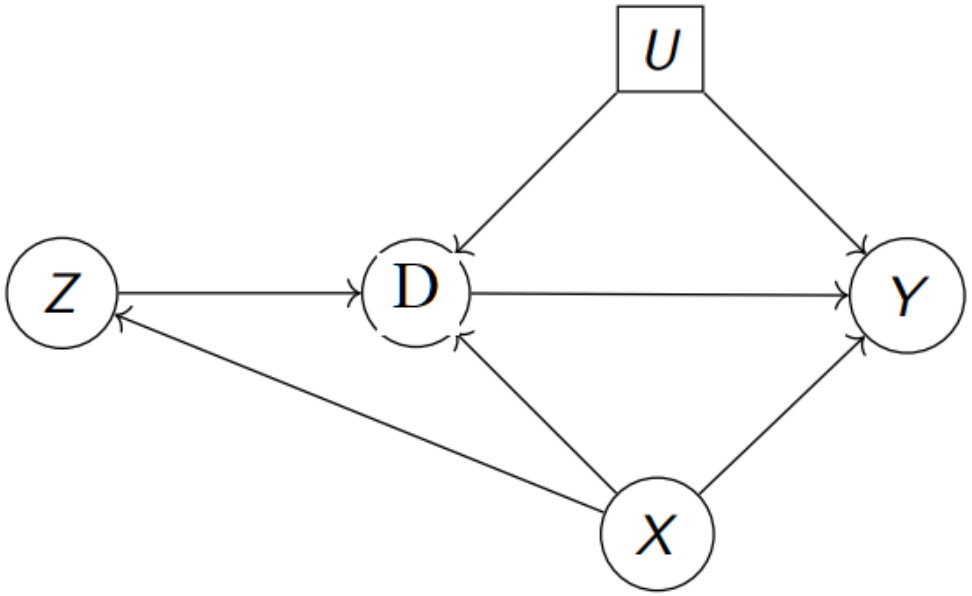


IV 中的控制变量 I：控制混淆因素

此时，X 是一个混淆变量，必须加以控制

例：

- Y: Income
- A：Education
- Z：Distance from College
- X: Family income, Father's Education, ...



IV 中的控制变量 II: 阻断其它路径

哪个说法正确？

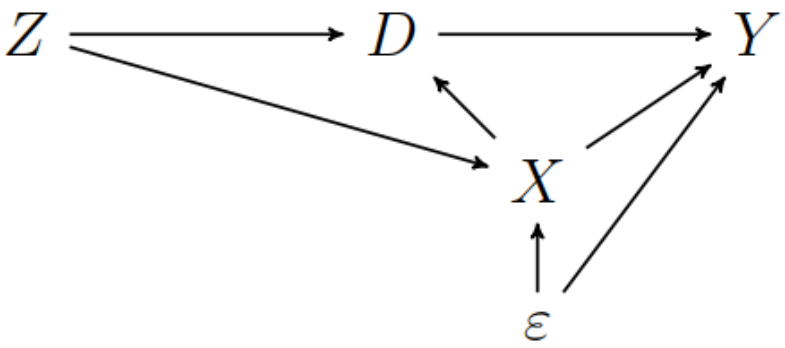
> A. Z 不能与 Y 相关，即 $\text{corr}(Z, Y) = 0$

> B. Z 可以与 Y 相关

- 但要满足 $Z \longrightarrow D \longrightarrow Y$，而非 $Y \longleftarrow Z \longrightarrow D$
- 甚至可以 $Z \longrightarrow D \longrightarrow Y$，且 $Z \longrightarrow W \longrightarrow Y$
  - 这时候就需要考虑控制变量的作用了 (next page)

IV 中控制变量的作用

- 排斥性约束要求工具变量 Z 只通过内生解释变量 D 影响 Y ，那么在 包含控制变量的工具变量回归中， Z 可不可以通过控制变量 X 影响 Y 呢?
  - **答：当然可以。**但这个问题应该反过来理解，正是因为担心工具变量有除了 D 之外影响 Y 的渠道，故而把这些潜在的渠道尽量控制起来，这些被控制起来的渠道就成了控制变量。这正是排斥性约束检验的主要思路。
  - 例子：Card (1995) 教育回报论文中，加入 South, ASMA 等控制变量就是这个目的。



Card (1995) 教育回报率

- Outcome Eq: $Income = \beta_1 Educ + \beta_2 X + e$
- First Eq: $Educ = \theta_0 + \theta_1 near4$

	(1) OLS1	(2) IV1	(3) OLS_argu1	(4) OLS2	(5) IV2	(6) OLS_argu2
ed76	0.074**	0.132**	0.074**	0.085**	0.221**	0.084**
	(20.32)	(2.73)	(20.12)	(23.05)	(5.44)	(22.55)
nearc4			0.020			0.063**
			(1.27)			(4.11)
black	-0.190**	-0.131**	-0.190**			
	(-10.88)	(-2.54)	(-10.88)			
reg76r	-0.125**	-0.105**	-0.122**	-0.196**	-0.100**	-0.184**
	(-8.13)	(-4.58)	(-7.87)	(-13.16)	(-2.98)	(-12.19)
smsa76r	0.161**	0.131**	0.155**			
	(10.64)	(4.41)	(9.73)			
exp	0.084**	0.107**	0.083**	0.085**	0.143**	0.085**
	(12.42)	(5.09)	(12.40)	(12.29)	(7.46)	(12.23)
exp2	-0.224**	-0.228**	-0.224**	-0.230**	-0.239**	-0.229**
	(-7.04)	(-6.59)	(-7.05)	(-6.96)	(-5.55)	(-6.95)
_cons	4.734**	3.753**	4.729**	4.677**	2.330**	4.653**
	(67.47)	(4.59)	(67.50)	(65.13)	(3.32)	(65.08)
N	3010	3010	3010	3010	3010	3010
r2	0.291	0.225	0.291	0.241	-0.134	0.245

t statistics in parentheses
* p<0.05

附：Stata codes

> Data Source: Hansen B E . 2021. **Econometrics**. Princeton University Press. Data and Contents, PDF

```
  use "Card1995_edu.dta", replace  // Hansen, 2021 Book

* New Variables
  gen exp  = age76 - ed76 - 6
  gen exp2 = (exp^2)/100
  gen age2 = (age^2)/100

* Drop observations with missing wage
  drop if lwage76==.

*-参考：Table 12.1 (Hansen 2021, book, modified)
  // SMSA: Standard Metropolitan Statistical Area
  //       (美国行政区划单位) 大城市及其郊区
   global cx1 "black reg76r smsa76r exp exp2"  // [a] common controls
   global cx2 "      reg76r         exp exp2"  // [c] black, smsa 是 cofounder ?
   global IV  "nearc4"

*-Note: 在 [c] 设定下，不再满足【排他性】假设
```

附：Stata codes (cont.)

```
*-(1) OLS
  reg lwage76 ed76 $cx1    , robust
  est store OLS1
  reg lwage76 ed76 $cx2    , robust
  est store OLS2

*-(4) 2SLS
  ivreg2 lwage76 $cx1 (ed76=$IV), robust
  est store IV1
  ivreg2 lwage76 $cx2 (ed76=$IV), robust
  est store IV2

*-(5) OLS + argument
  reg lwage76 ed76 $cx1 $IV, robust
  est store OLS_argu1
  reg lwage76 ed76 $cx2 $IV, robust
  est store OLS_argu2

*---- 对比结果 ----
 local m "OLS1 IV1 OLS_argu1 OLS2 IV2 OLS_argu2"
 esttab `m' `s', mtitle(`m') nogap compress ///
        b(%6.3f) order(ed76 near*)          ///
                 s(N r2, fmt(0 3)) star(** 0.05)
```

动态面板中的 Controls

模型设定：

$$y_{it} = \gamma y_{i,t-1} + x'_{it}\beta + \alpha_i + \varepsilon_{it} \quad (1)$$

基本假设：

- H1: **无序列相关** $\mathrm{Corr}(\varepsilon_{it}, \varepsilon_{it-1}) = 0$
- H2: **外生性** $E(\varepsilon_{it} \mid y_{it-1}, x_{it}, \alpha_i) = 0$

估计：IV/GMM

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta x'_{it}\beta + \Delta \varepsilon_{it} \quad (2)$$

- $IV_1 = y_{it-2} \quad \rightarrow \quad E(y_{it-2} \cdot \Delta\varepsilon_{it}) = 0$
- $IV_2 = y_{it-3} \quad \rightarrow \quad E(y_{it-3} \cdot \Delta\varepsilon_{it}) = 0$

  $\cdots$ (好多！)

序列相关检验 [AB91]

- 水平方程: $y_{it} = \gamma y_{i,t-1} + x'_{it}\beta + \alpha_i + \varepsilon_{it} \quad (1)$
- 差分方程: $\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta x'_{it}\beta + \Delta \varepsilon_{it} \quad (2)$
- H1: **无序列相关** $\mathrm{Corr}(\varepsilon_{it}, \varepsilon_{it-1}) = 0$

$$\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$$
$$\Delta \varepsilon_{it-1} = \varepsilon_{it-1} - \varepsilon_{it-2}$$
$$\Delta \varepsilon_{it-2} = \varepsilon_{it-2} - \varepsilon_{it-3}$$

$$\rho_1 = \mathrm{Corr}(\Delta \varepsilon_{it}, \Delta \varepsilon_{it-1}) \to m_1$$

$$\rho_2 = \mathrm{Corr}(\Delta \varepsilon_{it}, \Delta \varepsilon_{it-2}) \to m_1$$

应用指南：Sargan 和序列相关检验为何无法通过？

Sargan 检验的目的：在于检验工具变量的合理性，简言之，是工具变量与差分后的干扰项是否存在统计上显著的相关性。相当于做如下回归：

$$\Delta\varepsilon_{it} = \beta_1 + y_{it-2}\beta_2 + y_{it-3}\beta_3 + \cdots + v_{it}$$

然后看 y_{it-s} $(s \geq 2)$ 的系数是否联合显著，如果显著，那就悲剧了。

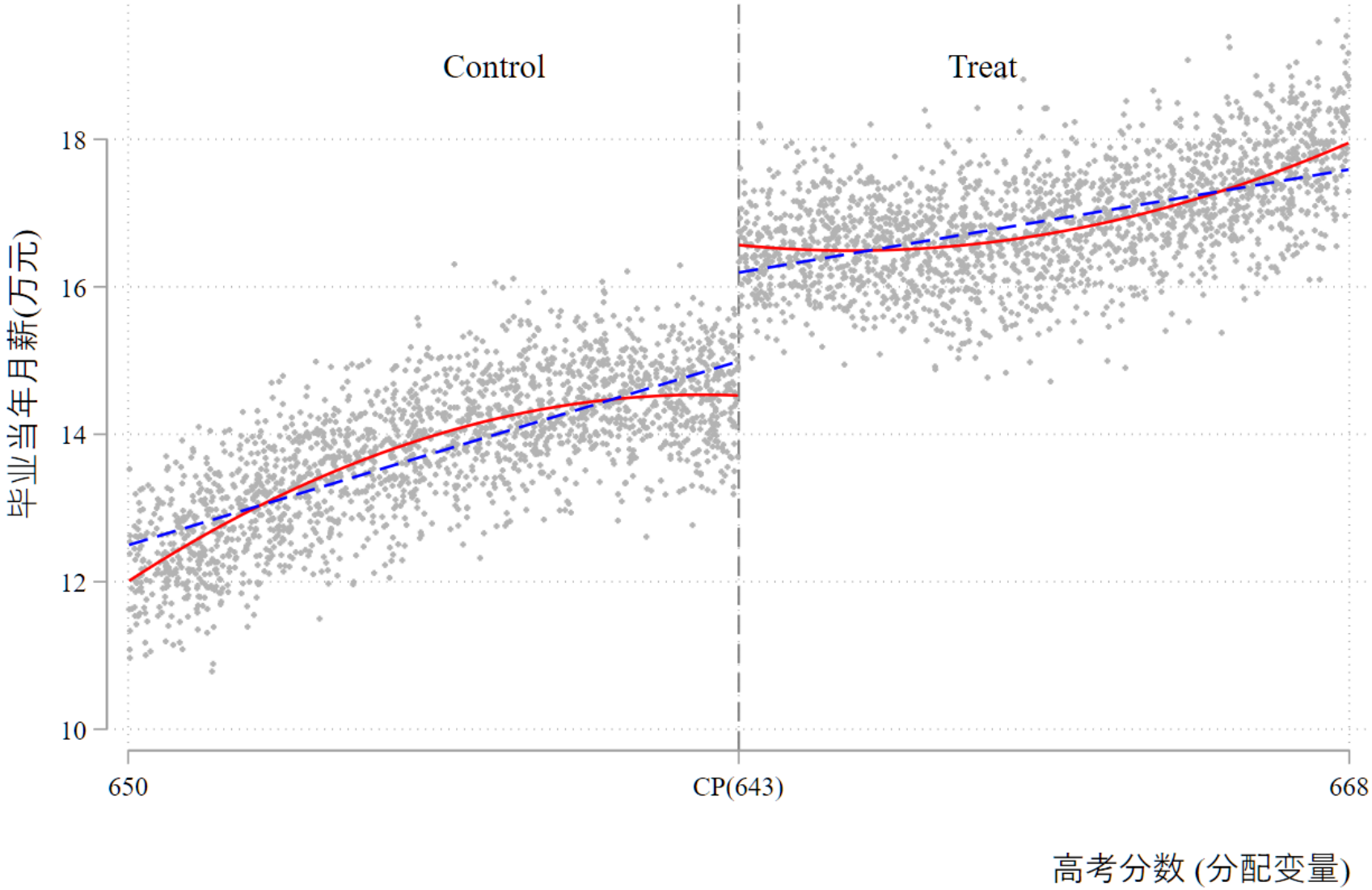
可以看出，我们可以从 $\Delta\varepsilon_{it}$ 和 y_{it-s} 两个角度入手。

- 让 $\Delta\varepsilon_{it}$ 干净一点
- 选择谁做 IVs？—— s 的选择
- **建议：** 🍎
  - 尽量加入 `i.year` (λ_t), 甚至 `i.industry#i.year` (λ_{jt})
  - 使用 `maxldep(#)` 选项控制冗余 IV 问题 (Weak IVs)

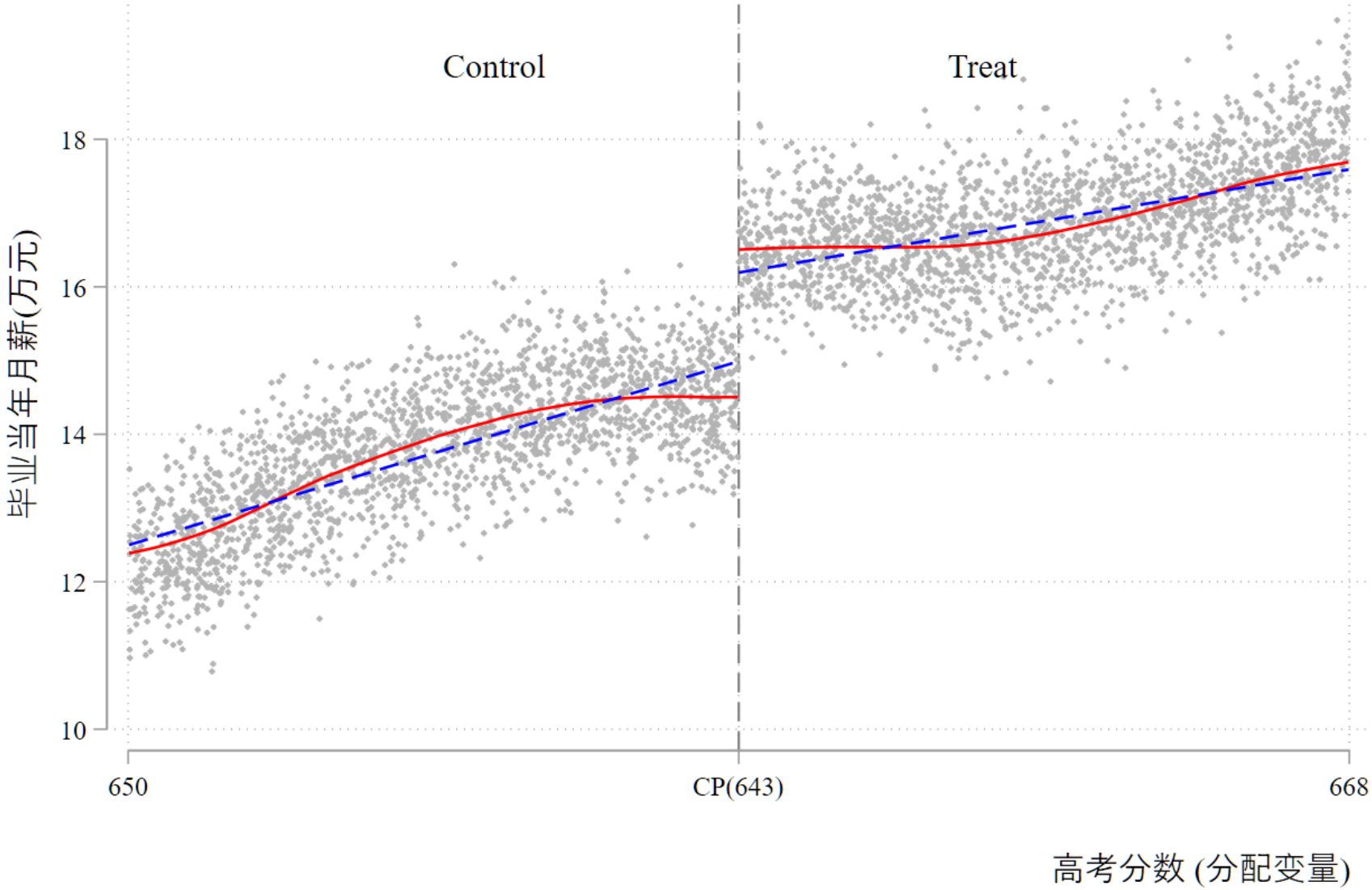
断点回归中的 Controls

- 若样本量很大，可以设定一个很窄的窗口，此时，无需加控制变量
- 若样本量较小，就需要设定一个较大的窗口，此时，需要加入控制变量 (通常是驱动变量的高阶项或交互项)，以捕捉非线性特征，从而降低模型误设偏误
- Calonico S, Cattaneo M D, Farrell M H, et al. Regression discontinuity designs using covariates[J]. Review of Economics and Statistics, 2019, 101(3): 442-451. -PDF-

参数法：三次函数



非参数方法：局部多项式回归，Stata: `rdrobust`



SCM 中的 Controls

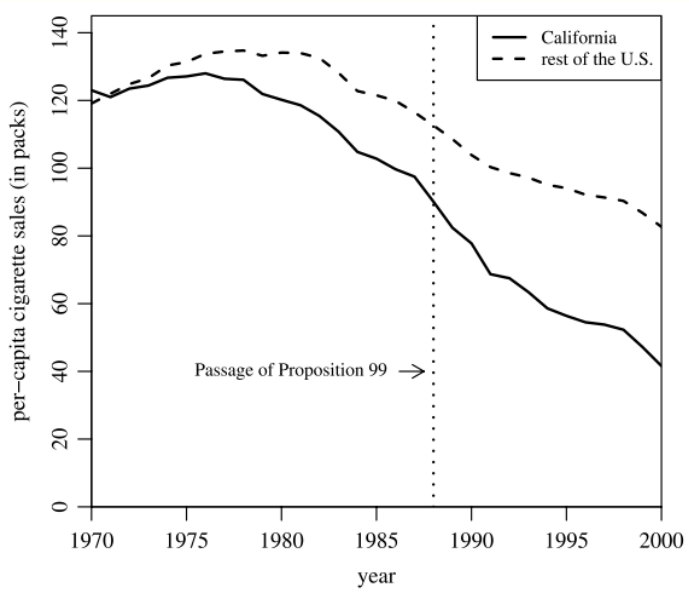
- 是否需要？ **YES**
- 如何设定？ 要兼顾 **样本外预测能力**，配合时间安慰剂检验
  - 可以使用 **Lasso** 筛选变量
  - 也可以使用回归控制法
  - 核心思想：采用交叉验证法筛选变量

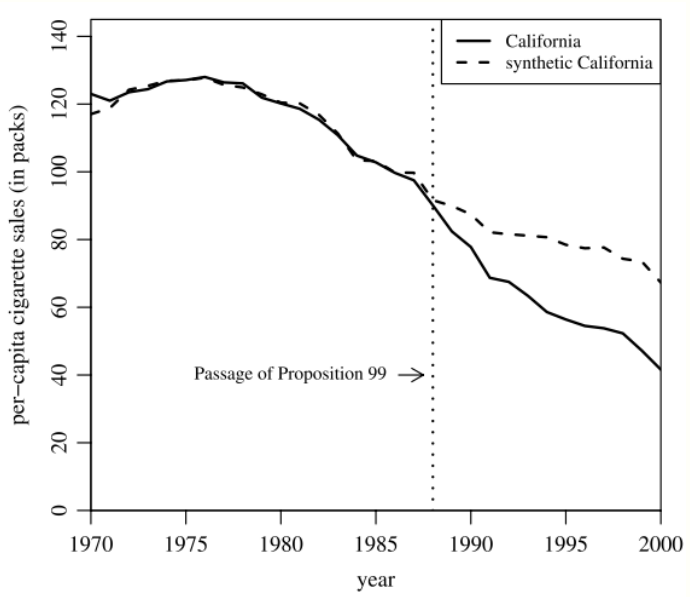
合成控制法：基本思想

Abadie et al. (2010, JASA) 假设结果变量由因子模型所确定:

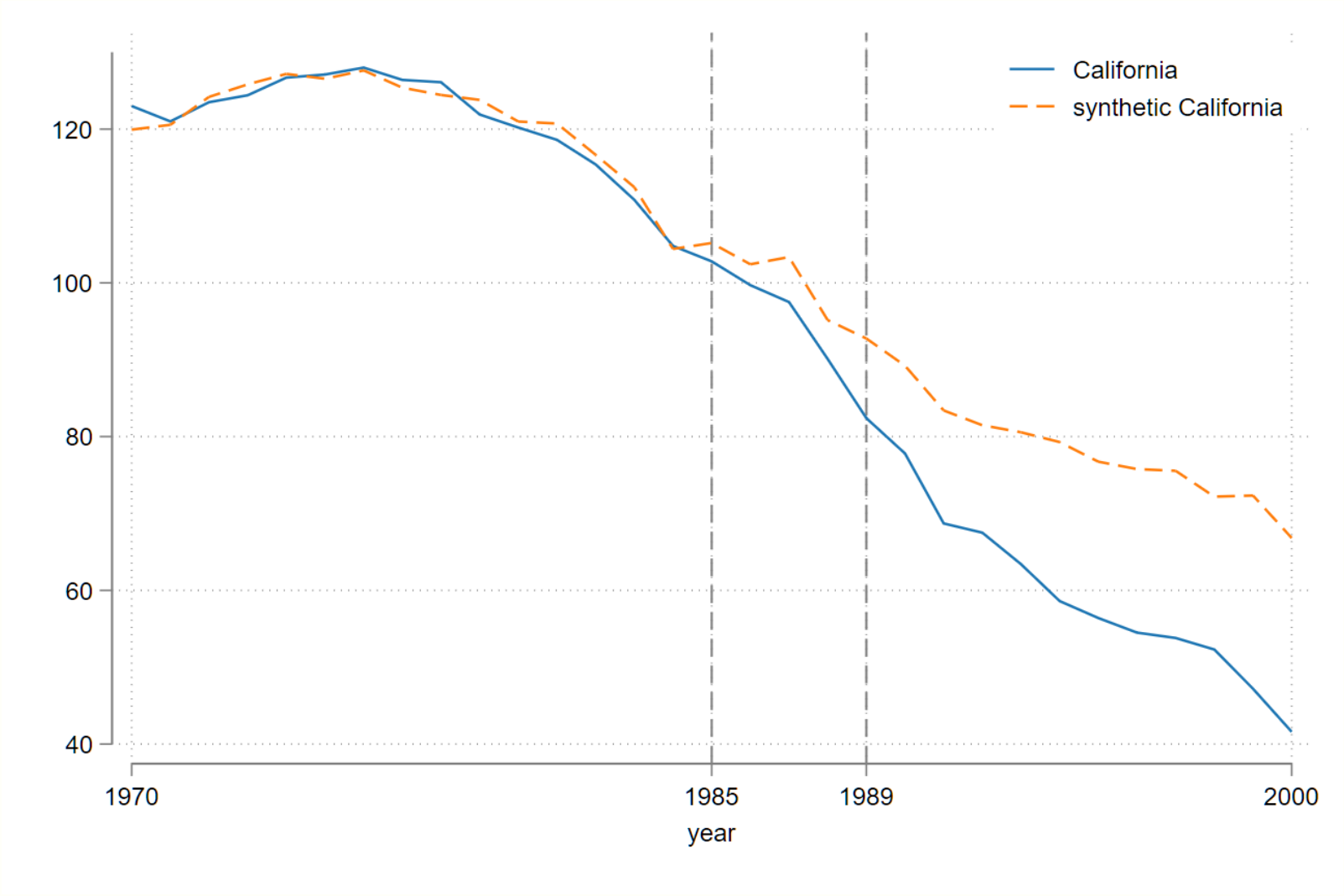
$$y_{jt} = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \varepsilon_{ij}$$

- Z_i 是一个 $(r \times 1)$ 阶可观测的预测变量：人均收入，啤酒消费量，人口结构等
  - 太多：过拟合
  - 太少：欠拟合

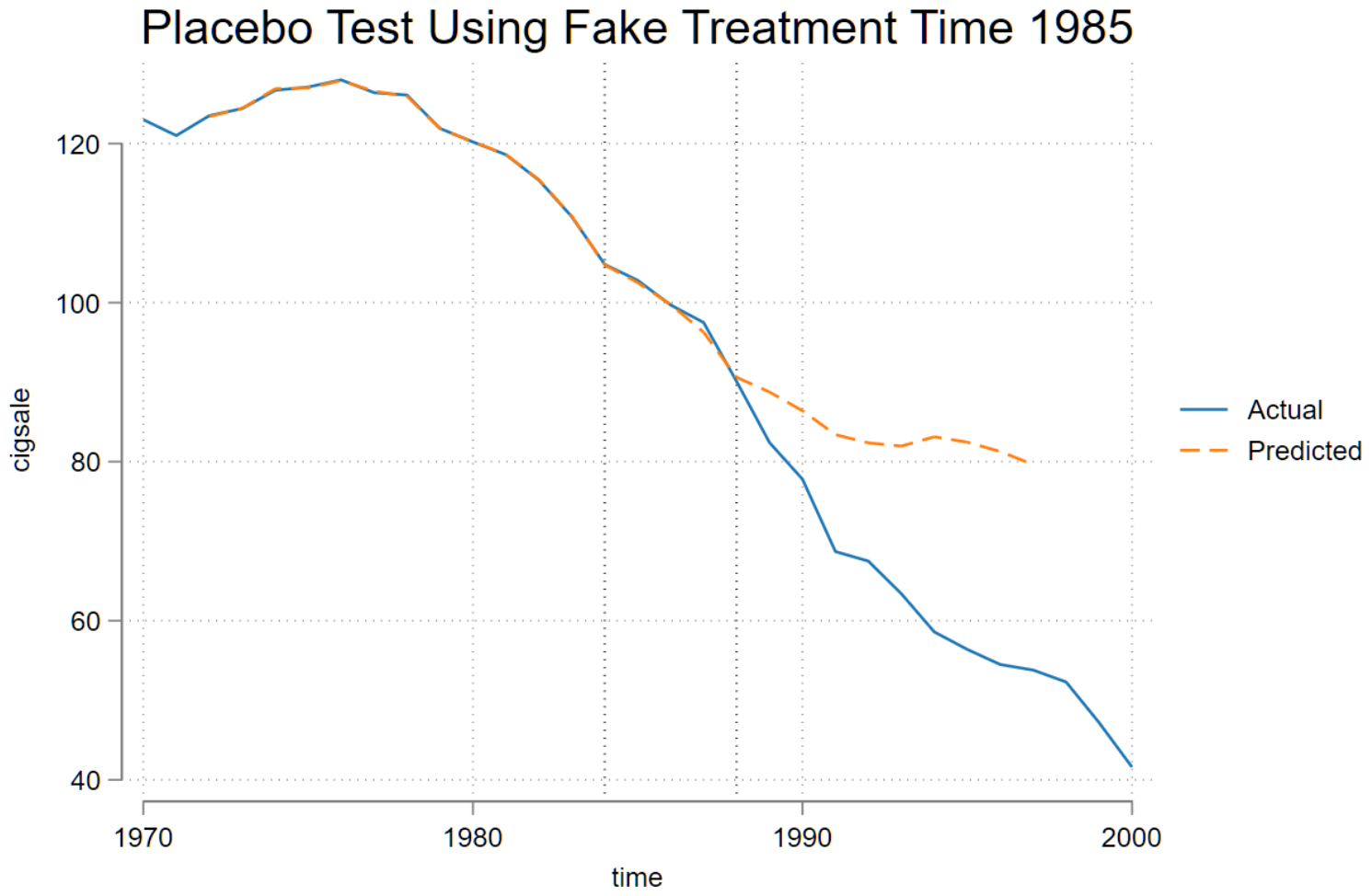




过拟合问题图示


California
synthetic California
120
100
80
60
40
1970
1985
1989
2000
year

使用 Lasso / RCM 的效果



```
use "smoking_wide.dta", clear

*-Lasso
lasso linear cig0 cig1-cig39 lny1-lny39  ///
          if year<=1984, select(cv, fold(10))

*-回归控制法
rcm cig $xx, trunit(3) trperiod(1989) ///
             method(lasso)  ///
             criterion(cv) fold(10)   ///  // 10-fold CV
             placebo(period(1985))
```

总结

控制变量的选择 —— 好的和坏的

- 好的：
  - 控制混淆因素
  - 阻断后门路径
  - 可以有多种模型设定方式
- 坏的
  - X 的结果变量 (中介变量)
  - X 和 Y 的共同结果变量

控制变量的选择 —— 不要过度控制

> Source: Schjoedt, Bird, Carsrud, Brännback (2014, PDF)

- 在理论和实践上对 y 的影响不明确的变量：尽量不要加入
  - Carlson, K. D., J. Wu (2012, PDF): 'When in doubt, leave them out' (p. 413)
- 控制变量主要用于排除混杂因素
- 避免包含多个控制变量以涵盖所有可能性。这会导致 II 类错误（例如，在实际上有效果时得出没有效果的结论）
- 加入过度冗余变量的代价：降低统计检验力；降低研究结果的普适性

主要文献

- Cinelli, C., A. Forney, J. Pearl, **2022**, A crash course in good and bad controls, **Sociological Methods & Research**, forthcoming. -Link-, -PDF-, Replication-R-Codes, -cited-
  - 中文解读，A-理论部分：控制变量！控制变量！Good-Controls-Bad-Controls
  - Stata 模拟：B-Stata模拟：控制变量！控制变量！Good-Controls-Bad-Controls
- Whited, R. L., Q. T. Swanquist, J. E. Shipman, J. R. Moon, Jr., **2022**, Out of control: The (over) use of controls in accounting research, **The Accounting Review**, 97 (3): 395-413. -Link-, -PDF-, PDF2, 推文
- Wysocki, A. C., K. M. Lawson, M. Rhemtulla, **2022**, Statistical control requires causal justification, **Advances in Methods and Practices in Psychological Science**, 5 (2): 25152459221095823. -Link-, -PDF-, PDF2, -cited-

- Becker, T. E., G. Atinc, J. A. Breaugh, K. D. Carlson, J. R. Edwards,P. E. Spector, **2016**, Statistical control in correlational studies: 10 essential recommendations for organizational researchers, **Journal of Organizational Behavior**, 37 (2): 157-167. -Link-, -PDF-, -cited-
- Carlson, K. D., J. Wu, **2012**, The illusion of statistical control: Control variable practice in management research, **Organizational Research Methods**, 15 (3): 413-435. -Link-, -PDF-, -cited-
- Loh, Wen Wei, and Dongning Ren. **2021**. “Data-driven Covariate Selection for Confounding Adjustment by Focusing on the Stability of the Effect Estimator.” PsyArXiv. September 27. -PDF-
- Hünermund, Paul, Louw, Beyers and Caspi, Itamar. "Double machine learning and automated confounder selection: A cautionary tale" Journal of Causal Inference, vol. 11, no. 1, 2023, pp. 20220078. https://doi.org/10.1515/jci-2022-0078, -PDF-

相关推文

- 专题：内生性-因果推断
  - A-理论部分：控制变量！控制变量！Good-Controls-Bad-Controls
  - B-Stata模拟：控制变量！控制变量！Good-Controls-Bad-Controls
  - 敏感性分析B-Stata实操：控制变量内生时的系数敏感性分析-regsensitivity
  - 敏感性分析A-理论基础：控制变量内生时的系数敏感性分析-regsensitivity
- 专题：论文写作
  - 控制变量如何选？大牛们的10条建议

- 专题：回归分析
  - 控制变量越多越好吗?
  - Stata：控制变量与核心解释变量地位对等吗?
  - 调节效应是否需要考虑对控制变量交乘?
  - 控制变量！控制变量!
  - 不用太关心控制变量，真的!
  - 加入控制变量后结果悲催了!

- 专题：断点回归RDD
  - RDD：断点回归可以加入控制变量吗?
- 专题：内生性-因果推断
  - 因果推断：双重机器学习-ddml
  - Stata：控制变量组合的筛选-tuples
  - Lasso一下：再多的控制变量和工具变量我也不怕-T217
  - Stata：双重机器学习-多维聚类标准误的估计方法-crhdreg
  - DDML：从随机实验到双重机器学习
  - Stata新命令-pdslasso：众多控制变量和工具变量如何挑选?
  - DID偏误问题：多时期DID的双重稳健估计量(下)-csdid
  - DID偏误问题：两时期DID的双重稳健估计量(上)-drdid

谢　谢