微观计量经济学

阮 睿

中央财经大学
中国财政发展协同创新中心

March 27, 2025

Contents

1 匹配：截面数据的参数和非参数方法

1.1 第二类识别假设

- **第二类识别假设：分配机制不取决于潜在结果。**

$$\Pr(D=1|X,Y^0,Y^1)=\Pr(D=1|X)$$

- **第二类识别假设（重新表述）：给定可观测变量，潜在结果均值独立于处理状态。**

Assumption ID.2 $\mathbb{E}(Y^d|D,X)=\mathbb{E}(Y^d|X),\ \ d=0,1$

- **等价地，**

$$\mathbb{E}(Y^0|D=1,X)=\mathbb{E}(Y^0|D=0,X)$$

$$\mathbb{E}(Y^1|D=1,X)=\mathbb{E}(Y^1|D=0,X)$$

●**此时**可观测变量相同条件下的组间均值差异**能够识别条件平均处理效应** $\tau(X), \tau_1(x), \tau_0(x)$.

$$\begin{aligned} & \mathbb{E}(Y|X=x, D=1) - \mathbb{E}(Y|X=x, D=0) \\ = \ & \mathbb{E}(Y^1|X=x, D=1) - \mathbb{E}(Y^0|X=x, D=0) \\ = \ & \mathbb{E}(Y^1|X=x) - \mathbb{E}(Y^0|X=x) \\ = \ & \mathbb{E}(Y^1 - Y^0|X=x) \\ = \ & \tau(x) \end{aligned}$$

要想识别 $\tau_1(x)$**，需要用到** $\mathbb{E}(Y^0|D=1, X) = \mathbb{E}(Y^0|D=0, X)$.
要想识别 $\tau_0(x)$**，需要用到** $\mathbb{E}(Y^1|D=1, X) = \mathbb{E}(Y^1|D=0, X)$.
证明过程类似。因此

$$\tau(x) = \tau_1(x) = \tau_0(x)$$

•**进而能够识别平均处理效应**

$$\begin{aligned}\tau &= \mathbb{E}_X[\tau(x)] \\ \tau_1 &= \mathbb{E}_X[\tau_1(x)|D=1] \\ \tau_0 &= \mathbb{E}_X[\tau_0(x)|D=0]\end{aligned}$$

请注意，$\tau \neq \tau_1 \neq \tau_0$**，因为** $F_X(x) \neq F_{X|D=1}(x) \neq F_{X|D=0}(x)$.

$$\begin{aligned}\tau_1 &= \mathbb{E}_X[\mathbb{E}(Y|X,D=1) - \mathbb{E}(Y|X,D=0)|D=1] \\ &= \mathbb{E}(Y|D=1) - \mathbb{E}_X(Y|X,D=0)|D=1]\end{aligned}$$

$$\begin{aligned}\tau_0 &= \mathbb{E}_X[\mathbb{E}(Y|X,D=1) - \mathbb{E}(Y|X,D=0)|D=0] \\ &= \mathbb{E}_X[\mathbb{E}(Y|X,D=1)|D=0] - \mathbb{E}(Y|D=0)\end{aligned}$$

$$\begin{aligned}\tau &= \mathbb{E}_X[\mathbb{E}(Y|X,D=1) - \mathbb{E}(Y|X,D=0)] \\ &= \mathbb{E}_X[\mathbb{E}(Y|X,D=1)] - \mathbb{E}_X(Y|X,D=0)\end{aligned}$$

- **而简单的组间均值比较的表达式却是**

$$\begin{aligned}\tau &= \mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) \\ &= \mathbb{E}_X(\mathbb{E}(Y|X,D=1)|D=1) - \mathbb{E}_X(\mathbb{E}(Y|X,D=0)|D=0)\end{aligned}$$

- **现在的问题是：What is the best way to “condition” on X?**

1.2 非参数方法：匹配

- **一个准确匹配 (exact matching) 的假想例子**

id	treat	x	y
1	1	1	y_1
2	1	1	y_2
3	1	2	y_3
4	1	3	y_4
5	0	1	y_5
6	0	2	y_6
7	0	2	y_7
8	0	2	y_8

id	treat	control
$x = 1$	y_1, y_2	y_5
$x = 2$	y_3	y_6, y_7, y_8
$x = 3$	y_4	

•ATT 的估计

$$\begin{aligned}
&\hat{\tau}_1(x=1) = \left(\frac{y_1+y_2}{2}\right) - y_5 \\
&\hat{\tau}_1(x=2) = y_3 - \left(\frac{y_6+y_7+y_8}{3}\right) \\
&\hat{\tau}_1(x=3)\textbf{无法估计}
\end{aligned} \tag{1}$$

$$\begin{aligned}
\hat{\tau}_1 &= \hat{\tau}_1(x=1)\times\frac{2}{3} + \hat{\tau}_1(x=2)\times\frac{1}{3} \\
&= \left(\frac{y_1+y_2+y_3}{3}\right) - \left(y_5\times\frac{2}{3} + \frac{y_6+y_7+y_8}{3}\times\frac{1}{3}\right) \\
&= \frac{1}{3}\left((y_1-y_5)+(y_2-y_5)+\left(y_3-\frac{y_6+y_7+y_8}{3}\right)\right) \\
&= \frac{1}{n_T}\sum_{i\in T\bigcap C_T}(2D_i-1)w_i y_i
\end{aligned}$$

●ATU 的估计

$$\begin{aligned}
&\hat{\tau}_0(x=1) = \left(\frac{y_1+y_2}{2}\right) - y_5 \\
&\hat{\tau}_0(x=2) = y_3 - \left(\frac{y_6+y_7+y_8}{3}\right) \\
&\hat{\tau}_0(x=3)\text{无法估计}
\end{aligned} \tag{2}$$

$$\begin{aligned}
\hat{\tau}_0 &= \hat{\tau}_0(x=1)\times\frac{1}{4} + \hat{\tau}_0(x=2)\times\frac{3}{4} \\
&= \left(\left(\frac{y_1+y_2}{2}\right)\times\frac{1}{4} + y_3\times\frac{3}{4}\right) - \left(\frac{y_5+y_6+y_7+y_8}{4}\right) \\
&= \frac{1}{4}\left[\left(\frac{y_1+y_2}{2} - y_5\right) + (y_3-y_5) + (y_3-y_7) + (y_3-y_8)\right]
\end{aligned}$$

●ATE 的估计

$$\begin{aligned}
&\hat{\tau}(x=1)=\left(\frac{y_1+y_2}{2}\right)-y_5\\
&\hat{\tau}(x=2)=y_3-\left(\frac{y_6+y_7+y_8}{3}\right)\\
&\hat{\tau}(x=3)\text{无法估计}
\end{aligned}\tag{3}$$

$$\begin{aligned}
\hat{\tau}&=\hat{\tau}(x=1)\times\frac{3}{7}+\hat{\tau}(x=2)\times\frac{4}{7}\\
&=\left(\left(\frac{y_1+y_2}{2}\right)\times\frac{3}{7}+y_3\times\frac{4}{7}\right)-\left(y_5\times\frac{3}{7}+\left(\frac{y_6+y_7+y_8}{3}\right)\times\frac{4}{7}\right)\\
&=\frac{1}{7}\left[(y_1-y_5)+(y_2-y_5)+\left(y_3\frac{y_6+y_7+y_8}{3}\right)\right.\\
&\quad\left.+\left(\frac{y_1+y_2}{2}-y_5\right)+(y_3-y_6)+(y_3-y_7)+(y_3-y_8)\right]\\
&=\hat{\tau}_1\times\frac{3}{7}+\hat{\tau}_0\times\frac{4}{7}
\end{aligned}$$

1.3 参数方法：含控制变量的线性回归

•假定 $\mathbb{E}(Y|D=1,X)$ 和 $\mathbb{E}(Y|D=0,X)$ 的函数形式，然后使用线性回归分别对处理组和控制组进行估计。

$$\begin{aligned}\hat{\mu}_1(X) &= \mathbb{E}(Y|\widehat{D=1},X)\\ \hat{\mu}_0(X) &= \mathbb{E}(Y|\widehat{D=0},X)\\ \hat{\tau}(X) = \hat{\tau}_1(X) &= \hat{\tau}_0(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)\end{aligned}$$

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]\\ \hat{\tau}_1 &= \frac{1}{n_T}\sum_{i\in T}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]\\ \hat{\tau}_0 &= \frac{1}{n_C}\sum_{i\in C}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]\end{aligned}$$

- **常见情形一：**

$$\mathbb{E}(Y|D=1,X)=\alpha_1+\gamma X$$
$$\mathbb{E}(Y|D=0,X)=\alpha_0+\gamma X$$

即

$$\mathbb{E}(Y|X)=\alpha_0+(\alpha_1-\alpha_0)D+\gamma X$$

可以通过 Y **对** D **和** X **的线性回归进行估计**

$$Y_i=\hat{\alpha}_0+(\hat{\alpha}_1-\hat{\alpha}_0)D_i+\hat{\gamma}X_i+e_i$$

因此

$$\hat{\tau}(X)=(\hat{\alpha}_1+\hat{\gamma}X)-(\hat{\alpha}_0+\hat{\gamma}X)=\hat{\alpha}_1-\hat{\alpha}_0$$
$$\hat{\tau}=\hat{\tau}_1=\hat{\tau}_0=\hat{\alpha}_1-\hat{\alpha}_0$$

- **以** D **为核心解释变量，** X **为控制变量的多元线性回归，** D **的斜率系数估计即为对 ATE, ATT 和 ATU 的估计。此时隐含了同质处理效应的假定。**

•**常见情形二：当 X 为离散变量时，假定其有 $x_1, \cdots, x_R$ 种不同取值，构造一组虚拟变量 $W_r = \mathbb{1}(X = x_r)$**

$$\mathbb{E}(Y|D=1,X) = \alpha_1 + \sum_{r=2}^{R} \gamma_r W_r$$

$$\mathbb{E}(Y|D=0,X) = \alpha_0 + \sum_{r=2}^{R} \gamma_r W_r$$

即 $\mathbb{E}(Y|X) = \alpha_0 + (\alpha_1 - \alpha_0)D + \sum_{r=2}^{R} \gamma_r W_r$
可以通过 Y **对** D **和** $W_r (r = 2, \cdots, R)$ **的线性回归进行估计**

$$Y_i = \hat{\alpha}_0 + (\hat{\alpha}_1 - \hat{\alpha}_0)D_i + \sum_{r=2}^{R} \hat{\gamma}_r W_r + e_i$$

$$\hat{\tau} = \hat{\tau}_1 = \hat{\tau}_0 = \hat{\alpha}_1 - \hat{\alpha}_0$$

此时也隐含了同质处理效应的假定。

●常见情形三:

$$\mathbb{E}(Y|D=1,X)=\alpha_1+\gamma_1 X$$
$$\mathbb{E}(Y|D=0,X)=\alpha_0+\gamma_0 X$$

即 $\mathbb{E}(Y|X)=\alpha_0+(\alpha_1-\alpha_0)D+\gamma_0 X+(\gamma_1-\gamma_0)D\cdot X$
可以通过 Y 对 D, X 及其交互项的线性回归进行估计

$$Y_i=\hat{\alpha}_0+(\hat{\alpha}_1-\hat{\alpha}_0)D_i+\hat{\gamma}_0 X_i+(\hat{\gamma}_1-\hat{\gamma}_0)D_i\cdot X_i+e_i$$

因此 $\tau(X_i)=\tau_1(X_i)=\tau_0(X_i)=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)X_i$

$$\hat{\tau}=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\bar{X}$$
$$\hat{\tau}_1=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\bar{X}_{D=1}$$
$$\hat{\tau}_0=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\bar{X}_{D=0}$$

此时隐含了异质处理效应的特定形式（条件平均处理效应随 X 线性变化）假定。

- 当 X 为离散变量时，可以把常见情形二和常见情形三结合起来，例如

$$\mathbb{E}(Y|D=1,X)=\beta_1\mathbb{1}(X=1)+\beta_2\mathbb{1}(X=2)+\beta_3\mathbb{1}(X=3)$$
$$\mathbb{E}(Y|D=0,X)=\gamma_1\mathbb{1}(X=1)+\gamma_2\mathbb{1}(X=2)+\gamma_3\mathbb{1}(X=3)$$

即

$$\begin{aligned}\mathbb{E}(Y|D,X)=&\gamma_1\mathbb{1}(X=1)+\gamma_2\mathbb{1}(X=2)+\gamma_3\mathbb{1}(X=3)\\&+(\beta_1-\gamma_1)D\cdot\mathbb{1}(X=1)\\&+(\beta_2-\gamma_2)D\cdot\mathbb{1}(X=2)\\&+(\beta_3-\gamma_3)D\cdot\mathbb{1}(X=3)\end{aligned}$$

这一模型被称作饱和模型 (saturated model)，此时模型的线性形式不再具有限制性，它等价于匹配方法。

$$\tau(X = k) = \beta_k - \gamma_k$$

$$\tau_1 = \frac{1}{n_T} \sum_{i \in T} [(\beta_1 - \gamma_1) \cdot \mathbb{1}(X_i = 1) + (\beta_2 - \gamma_2) \cdot \mathbb{1}(X_i = 2) + (\beta_3 - \gamma_3) \cdot \mathbb{1}(X_i = 3)]$$

$$\tau_1 = \frac{1}{n_C} \sum_{i \in C} [(\beta_1 - \gamma_1) \cdot \mathbb{1}(X_i = 1) + (\beta_2 - \gamma_2) \cdot \mathbb{1}(X_i = 2) + (\beta_3 - \gamma_3) \cdot \mathbb{1}(X_i = 3)]$$

$$\tau_1 = \frac{1}{n} \sum_{i=1}^{n} [(\beta_1 - \gamma_1) \cdot \mathbb{1}(X_i = 1) + (\beta_2 - \gamma_2) \cdot \mathbb{1}(X_i = 2) + (\beta_3 - \gamma_3) \cdot \mathbb{1}(X_i = 3)]$$

这种等价性只有当 X 离散时才能实现；当 X 连续时，模糊匹配 (fuzzy matching) 和基于函数形式的外推 (functional extrapolation) 会得到不同的结果。

1.4 尤其重要的常见情形：控制固定效应

•考虑如下模型

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i$$

得到 $\hat{\beta}_1^{OLS}$ **的一种等价做法是，**Y_i **和** D_i **分别对** X_i **进行回归，保留残差，**

$$Y_i = c_1 + c_2 X_i + \tilde{Y}_i$$
$$D_i = d_1 + d_2 X_i + \tilde{D}_i$$

则

$$\hat{\beta}_1^{OLS} = \frac{\widehat{\text{Cov}(Y, D)}}{\widehat{\text{Var}(D)}} = \frac{\widehat{\text{Cov}(\tilde{Y}, \tilde{D})}}{\widehat{\text{Var}(\tilde{D})}}$$

其中后者还能给出正确的标准误。

- 要想准确估计一个解释变量的系数，这个解释变量必须具有足够的变动性 (variation)。例如，在教育回报率的例子中，

$$\text{wage}_i = b_0 + b_1 \cdot \text{edu}_i + e_i$$

为了估计 b_1，要求样本中存在各种不同受教育水平的个体。

- 现在方程右边加入性别控制变量（作用何在?），

$$\text{wage}_i = b_0 + b_1 \cdot \text{edu}_i + c \cdot \text{male}_i + e_i$$

这等价于

$$\widetilde{\text{wage}}_i = b_1 \cdot \widetilde{\text{edu}}_i + e_i$$

其中 $\widetilde{\text{wage}}_i$ 和 $\widetilde{\text{edu}}_i$ 分别是 wage_i 和 edu_i 对 male_i 回归得到的残差。

•易知

$$\begin{aligned}\widetilde{\text{wage}}_i &= \begin{cases}\text{wage}_i - \bar{\text{wage}}_m & \text{if male}_i = 1\\ \text{wage}_i - \bar{\text{wage}}_f & \text{if male}_i = 0\end{cases}\\ \widetilde{\text{edu}}_i &= \begin{cases}\text{edu}_i - \bar{\text{edu}}_m & \text{if male}_i = 1\\ \text{edu}_i - \bar{\text{edu}}_f & \text{if male}_i = 0\end{cases}\end{aligned}$$

因此对于估计 b_1 而言真正有用的变动性不是受教育水平的整体变动性，而是其在性别组内的变动性。我们把控制 male_i 的操作称为控制性别固定效应。

- 类似地，如果在方程右边加入地区控制变量，

$$\text{wage}_i = b_0 + b_1 \cdot \text{edu}_i + \sum_{r=2}^{R} c^r \cdot \text{region}_i^r + e_i$$

其中

$$\text{region}_i^r = \begin{cases} 1 & \text{if } i \text{ is from region} r \\ 0 & \text{otherwise} \end{cases}$$

等价于

$$\widetilde{\text{wage}}_i = b_1 \widetilde{\text{edu}}_i e_i$$

其中 $\widetilde{\text{wage}}_i$ 和 $\widetilde{\text{edu}}_i$ 分别是 wage_i 和 edu_i 对所有地区虚拟变量回归得到的残差，也即去地区均值以后工资和受教育水平。控制了地区固定效应以后，真正有用的变动性是地区内部受教育水平的变动性

1.5 插曲：非参数回归

- 含控制变量的线性回归是回归调整法 (regression adjustment) 的特例。我们既可以用参数方法也可以用非参数方法估计 $\hat{\mu}_1(X)$ 和 $\hat{\mu}_0(X)$.
- 核回归 (kernel regression).
  - 用 $X = x$ 附近的 Y 的均值近似 $X = x$ 点处 Y 的均值。这里的 X 只是一个单变量。
  - $K(\cdot)$ 是核函数，本质就是权重函数，多表现为某个随机变量的密度函数形式。常见的核函数包括：
    - Uniform (rectangular) kernel:
    $$K(u) = \frac{1}{2}\mathbb{1}(|u| < 1)$$
    - Triangular kernel:
    $$K(u) = (1 - |u|)\mathbb{1}(|u| < 1)$$
    - Epanechnikov kernel:
    $$K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}(|u| < 1)$$
    - Gaussian kernel:
    $$K(u) = \phi(u)$$

–核估计量

$$\widehat{\mathbb{E}(Y|X=x)} = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$

当 $n \to \infty$ **时，** $h \to 0$**，称为带宽** (bandwidth)。

–核估计量是 Y **的加权平均，距离** $X = x$ **越近的观测值被赋予越高权重。**

- 局部线性回归 (local linear regression).
  - 考察如下 $X = x$ 附近处的回归

$$\min_{a,b} \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K\left(\frac{X_i - x}{h}\right)$$

$$\widehat{\mathbb{E}(Y|X = x)} = \hat{a}$$

  - 基本想法是，如果条件期望函数是平滑的，那么在 $X = x$ 附近就可以用线性函数近似。从这个视角，可以把核回归看作是局部常数项回归 (local constant regression)。局部线性回归通常比局部常数项回归偏误更小，尤其当 x 取边界值时。[1]

[1] 在回归断点设计中这种情形是重要的。

匹配就是良好的控制。

示例. 私立学校的经济回报 (Dale and Krueger, 2002, QJE).

The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

- **非参数估计结果：**
  - $\tau(A) = -5, \tau(B) = 30.$
  - $\tau = (-5) \times (3/5) + 30 \times (2/5) = 9.$
  - $\tau_1 = (-5) \times (2/3) + 30 \times (1/3) = 20/3.$
  - $\tau_0 = (-5) \times (1/2) + 30 \times (1/2) = 25/2.$

1.6 匹配与线性回归的比较

- **匹配作为控制 X 的非参数方法，无需假定同质处理效应或处理效应关于 X 的函数形式。**
- **但是，即使 $\tau(x)$ 各不相同并且高度非线性，是否有可能线性回归一致地估计了 $\tau = \mathbb{E}_X[\tau(x)]$ 呢?**
- **更重要的是，如前所述，匹配只是一种条件策略 (conditioning strategy)，而识别假设不会因为选择了一种特定的条件策略而变得更可信；换句话说，**
- 启示一：匹配的识别假设和线性回归的识别假设是一样的——**都是假设 ID.2. 事实上，当关于 $\mathbb{E}(Y^1|X)$ 和 $\mathbb{E}(Y^0|X)$ 的函数形式假设——即关于 $\mathbb{E}(Y|D=1,X)$ 和 $\mathbb{E}(Y|D=0,X)$ 的函数形式假设——正确时，可以证明，假设 $LS.2$ 和假设 $ID.2$ 是等价的。因此这两种方法的识别力度 (identification power) 是一样的，**认为匹配方法能够解决线性回归所无法解决的内生性问题，是大错特错。

- **所以我们需要深入探究**在异质处理效应前提下，**参数方法何时产生偏误。但必须牢记，**
- 启示二：对研究情境的充分认识和对协变量的妥善选择，以及在此基础上的因果识别工作，要优先于条件策略的选择。
- **要保证参数方法能够得到平均处理效应的一致估计，下面两个条件中必须至少满足一个** (Imbens, 2015, JHR)：
  - **关于条件期望的函数形式假设是正确的。**
  - **处理组和控制组的** X **分布相同。**

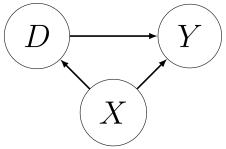
  否则，存在由模型误设 (**misspecification**) 引起的外推偏误 (**extrapolation bias**)。
- **通常而言，当处理组和控制组规模很不均衡时，其协变量分布往往也差异很大，此时可以考虑使用匹配方法。**

示例. 评估职业培训项目的效果 (Lalonde, 1986, AER; Dehejia and Wahba, 1999, JASA, 2002, REStat).

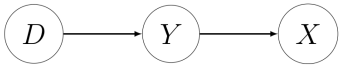
- 评估美国 1970 年代中期的某个职业培训项目。该项目是一项实地实验，招募一批劳动力市场上的弱势群体（戒毒瘾者、有犯案前科者、辍学者等），将其分为处理组和控制组，处理组个体可以获得 9-18 个月的工作机会。74、75 年为干预前，78 年为干预后。
- Lalonde (1986) 最先使用了这组数据，他先用处理组和控制组数据估计了这一职业培训项目的效果，然后用 PSID（收入动态面板调查）中的抽样数据代替控制组数据再次进行了估计，说明两者的差异。
- Dehejia and Wahba (1999, 2002) 发现，如果采用倾向得分匹配方法，那么即使用调查数据而非实际的控制组数据作为控制组，也能得到合意的结果。
- 实验组的干预前结果和调查数据控制组的干预前结果有很大差异。

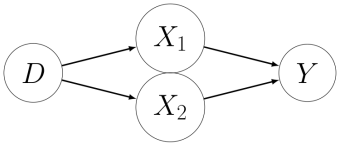
1.7 匹配方法实操

- **选择协变量 X**
  - **必须控制**

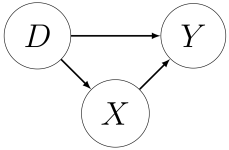


  - **必须不能控制**





–只有当关心直接效应时才控制（常常是“坏控制 (bad control)”）



- 匹配方法的分类
  - 以协变量 X 本身作为匹配对象的方法称为协变量匹配 (covariate matching)。当协变量为离散变量时，可以进行准确匹配。当用于匹配的协变量过多时，会遇到维度的诅咒。当匹配协变量为连续变量时，准确匹配无法实现。
  - 近似匹配。一种想法是构造处理组个体协变量向量和控制组个体协变量向量的距离指标，例如马氏距离 (Mahalanobis distance)，为处理组个体匹配到这一距离指标最小的控制组个体。

$$\min_{c \in C} (\mathbf{X}_c - \mathbf{X}_t)' \Omega (\mathbf{X}_t)^{-1} (\mathbf{X}_c - \mathbf{X}_t)$$

    其中 $\Omega(\cdot)$ 是样本方差-协方差矩阵。
  - 另一种想法是去估计倾向得分（注意，倾向得分可以模型化为匹配协变量的非线性函数），然后直接用倾向得分的估计值进行近似匹配，这就是第二种主要的匹配方法：倾向得分匹配 (propensity score matching, PSM)。

–是否放回 (replacement)：匹配过程中个体可否被重复使用，涉及 bias 和 variance 的权衡。

–一配多时可以选择固定数目的匹配个体——最近邻 k 匹配，也可以选择距离指标在一定范围内的匹配个体——半径 (radius)/卡钳 (caliper) 匹配。

–核 (kernel) 匹配：一配多计算匹配个体均值时，其权重是否以及如何随距离衰减。

$$\hat{E}(Y|X_t, D=0) = \sum_{c \in C} \frac{K\left(\frac{X_c - X_t)}{h}\right) Y_c}{K\left(\frac{X_c - X_t}{h}\right)}$$

–在大多数研究情境中，处理组规模往往小于控制组规模，能够为处理组个体找到匹配质量良好的控制组个体，因此 ATT 是主要的研究目标。

•协变量匹配时，因为匹配是不精确的，需要对估计量进行偏误修正。以 ATE 为例，

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^1 - \hat{Y}_i^0)$$

$$\hat{Y}_i^1 = D_i Y_i + (1 - D_i)\frac{1}{||T_i||}\sum_{t \in T_i}\{Y_t + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_t)\}$$
$$\hat{Y}_i^0 = (1 - D_i)Y_i + D_i\frac{1}{||C_i||}\sum_{c \in C_i}\{Y_c + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_c)\}$$

其中 $\hat{\mu}_d(X)$ 是 $\mu_d(X) = E(Y|X, D = d)$ 的 OLS 估计。

- 估计倾向得分
  - $h(\cdot)$ 形式的选择不是为了提供因果解释，而只是为了更好地近似条件期望函数。以 logit 为例，

$$P[D=1|x] \triangleq \pi(X) = \frac{\exp(h(X)'\gamma}{1+\exp(h(X)'\gamma)}$$

  - 基于逐步回归的方程设定搜索：
    1. 先根据经验决定有哪些变量必须加入（如果没有先验信息，就只加入截距项）。
    2. 然后逐一加入其余所有一次项，对新增项的系数显著性进行 likelihood ratio test, 统计量数值最大的一次项加入 $h(\cdot)$.
       似然函数和 AIC：

$$L = -(n/2)\cdot \ln(2\cdot\pi) - (n/2)\cdot\ln(SSE/n) - n/2$$

$$AIC = (2k + 2|L|)/n$$

    3. 对余下的一次项重复步骤 2，直到本轮新增项系数的检验统计量最大值低于临界值 $C_{lin} = 1$.
    4. 然后逐一加入所有二次项（包括平方项和交互项），进行类似于步骤 2-3 的操作，统计量临界值 $C_{qua} = 2.71$.
  - 根据最终确定的 $h(\cdot)$ 估计倾向得分。

●检查匹配样本的平衡性。

–计算组间的标准化差异 (normalized difference)。

$$\Delta = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{(s_C^2 + s_T^2)/2}}$$

$$s_C^2 = \frac{1}{N_C - 1}\sum_{c \in C}(X_c - \bar{X}_C)^2,\ s_T^2 = \frac{1}{N_T - 1}\sum_{t \in T}(X_t - \bar{X}_T)^2$$

–这个统计量和检验两个样本均值是否相等的 t 统计量长得很像，但不建议使用后者。(为什么?)

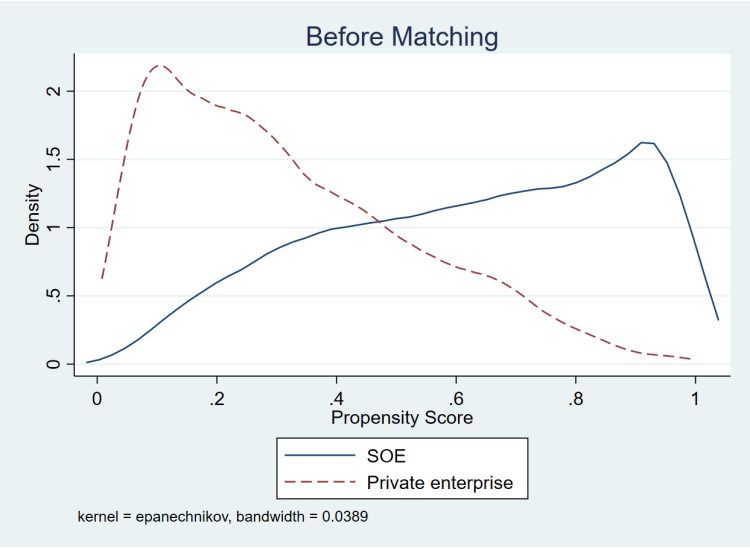
$$t_{stat} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\sum_{c \in C}(X_c - \bar{X}_C)^2/N_C + \sum_{t \in T}(X_t - \bar{X}_T)^2/N_T}}$$

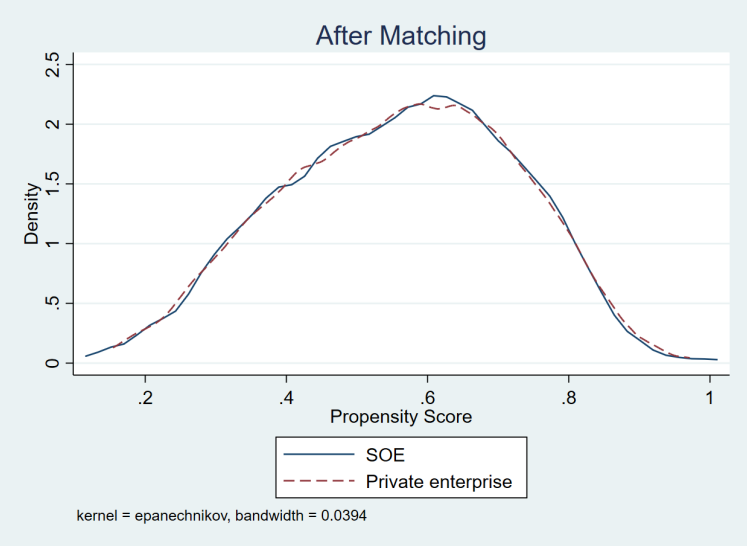
–如果标准化差异很大，则要考虑删截样本，可以根据倾向得分进行删节。

- 协变量匹配：teffects nnmatch 和 nnmatch 都可以实现，并且都提供了详尽的匹配细节，但后者报告的标准误估计是错误的。
- 倾向得分匹配：teffects psmatch 和 psmatch2 都可以实现，并且都提供了详尽的匹配细节，但后者报告的标准误估计是错误的。
- 不论是协变量匹配还是倾向得分匹配，可能都有必要预先进行一步准确匹配，例如性别、行业等。
- 尽管有研究认为协变量匹配优于倾向得分匹配 (King and Nielsen, 2019, Political Analysis)，但倾向得分匹配仍然被广泛使用。

```
foreach year of numlist 2008/2019{
cd "E:                                                    "
use "regdata20211020.dta",clear

**********************
* PSM
winsor2 investment roaa employee asset leverage roac roab cashflowd cashflowd operatingincome operatingprofit , replace cuts(1,99) by(year)
global controllist "leverage logasset tobinb grow roaa separation boardnum indboard_ratio largestholderrate  presidentisceo"


    display `year'

    keep if year==`year'
    psmatch2 soe_own $controllist ,noreplace neighbor(1) logit common

    gen pair = _id if _treated==0
    replace pair = _n1 if _treated==1
    bysort pair: egen paircount = count(pair)
    drop if paircount !=2
    keep stkcd
    gen year = `year'
    save `year'samples, replace

}
```

stata：kdensity




After Matching
Density
0
.5
1
1.5
2
2.5
.2
.4
.6
.8
1
Propensity Score
SOE
Private enterprise
kernel = epanechnikov, bandwidth = 0.0394

- 启示三：匹配的重点是找到“相像”的处理组和控制组个体，具体实现手段并不重要。(**言必称 PSM 只能反映对匹配方法的一知半解**。)

  示例. **信息不对称与融资决策**（Derrien and Kecskes, 2013, JF）

  these restrictions on both groups of firms. We require that candidate control firms have the same two-digit SIC code as our treatment firms. We also require candidate control firms to be in the same total assets quintile, Q quintile, and cash flow quintile as our treatment firms.6 We then retain candidate control firms that have the smallest difference in number of analysts compared to the corresponding treatment firms. We break any remaining ties based on the smallest differences in total assets, Q, and cash flow. To this end, we compute the difference between treatment firms and control firms for each of total assets, Q, and cash flow. We compute the rank of the difference for each of these three variables, and we compute the total rank across all three variables. We retain candidate control firms that have the lowest total rank.

S. M. Iacus, G. King, and G. Porro. Causal inference without balance checking: Coarsened exact matching. Political analysis, 20(1):1–24, 2012.

”粗糙精确匹配”（Coarsened Exact Matching, CEM），因为无论是 1-k 匹配还是完全匹配都没有考虑外推区域，在另一个处理组中几乎没有或没有合理匹配存在的情况。

CEM 首先对所选的重要协变量进行粗化，即离散化，然后在粗化的协变量上执行精确匹配。例如，如果所选的协变量是年龄（年龄 >50 为 1，其他为 0）和性别（女性为 1，男性为 0）。在受治疗组中，年龄为 50 岁的女性患者被表示为粗糙化协变量值为（1，1）。

她只会与受治疗组中具有完全相同粗糙化协变量值的患者进行匹配。经过精确匹配后，整个数据被分成两个子集。在一个子集中，每个单位都有其精确匹配的邻居，而在另一个子集中，包含了外推区域的单位。外推区域中单位的结果是通过在匹配子集上训练的结果预测模型进行估计的。到目前为止，可以分别估计两个子集上的处理效应，最后一步是通过加权平均将两个子集上的处理效应结合起来。

STATA:

```
cem 变量1(分箱规则) 变量2(分箱规则), treatment(处理变量) [k2k autocuts(算法)]
```

例如：

```
cem age(10 20 30) education(scott), treatment(treated) k2k
```

表示对年龄按 10/20/30 分箱，教育按 Scott 规则分箱，并强制每层人数相等.

- 启示四：要从两个层次理解匹配，作为数据预处理手段的匹配和作为非参数估计方法的匹配，前者远比后者重要。

1.8 一些新的匹配方法

《A survey on causal inference》

1.8.1 Re-weighting methods

- **混淆因素导致了选择偏差问题，即观测数据是否接受处理的分配和协变量相关。样本重新加权（re-weighting）是克服选择偏差的有效方法。通过观测数据为每个单位分配适当的权重，可以创建一个** pseudo-population**，使得处理组和对照组的分布相似。**
- balancing score：Balancing score $b(x)$ is a general weighting score, which is the function of x satisfying:

$$W \perp\!\!\!\perp x \mid b(x)$$

  where W is the treatment assignment and x is the background variables.

- **构造** $b(x)$ **的方法有很多，最常用的是** propensity score**。**

$$e(x) = Pr(W = 1|X = x)$$

Propensity score based sample re-weighting Propensity scores can be used to reduce selection bias by equating groups based on these covariates. Inverse propensity weighting (IPW) , also named as inverse probability of treatment weighting (IPTW), assigns a weight r to each sample:

$$r = \frac{W}{e(x)} + \frac{1-W}{1-e(x)},$$

where W is the treatment assignment ($W = 1$ denotes being treated group; $W = 0$ denotes the control group), and $e(x)$ is the propensity score defined above.

After re-weighting, the IPW estimator of average treatment effect (ATE) is:

$$\widehat{\text{ATE}}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i^F}{\hat{e}\left(x_i\right)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\left(1-W_i\right)Y_i^F}{1-\hat{e}\left(x_i\right)},$$

and its normalized version, which is preferred especially when the propensity scores are obtained by estimation :

$$\widehat{\text{ATE}}_{IPW} = \sum_{i=1}^{n}\frac{W_i Y_i^F}{\hat{e}\left(x_i\right)} \Big/ \sum_{i=1}^{n}\frac{W_i}{\hat{e}\left(x_i\right)} - \sum_{i=1}^{n}\frac{\left(1-W_i\right)Y_i^F}{1-\hat{e}\left(x_i\right)} \Big/ \frac{\left(1-W_i\right)}{1-\hat{e}\left(x_i\right)}.$$

现实中，IPW 的准确性依赖于 propensity score 的估计是否准确，否则会导致 ATE 估计非常不准。为了解决这个问题，出现了 Doubly Robust estimator (DR), 又名 Augmented IPW (AIPW)。

DR estimator combines the propensity score weighting with the outcome regression, so that the esti-

mator is robust even when one of the propensity score or outcome regression is incorrect (but not both). In detail, the DR estimator is formalized as:

$$\begin{aligned}\text{ATE}_{DR} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{W_i Y_i^F}{\hat{e}\left(x_i\right)} - \frac{W_i - \hat{e}\left(x_i\right)}{\hat{e}\left(x_i\right)}\hat{m}\left(1, x_i\right)\right] - \left[\frac{\left(1-W_i\right)Y_i^F}{1-\hat{e}\left(x_i\right)} - \frac{W_i - \hat{e}\left(x_i\right)}{1-\hat{e}\left(x_i\right)}\hat{m}\left(0, x_i\right)\right]\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}\left(1, x_i\right) + \frac{W_i\left(Y_i^F - \hat{m}\left(1, x_i\right)\right)}{\hat{e}\left(x_i\right)} - \hat{m}\left(0, x_i\right) - \frac{\left(1-W_i\right)\left(Y_i^F - \hat{m}\left(0, x_i\right)\right)}{1-\hat{e}\left(x_i\right)}\right\},\end{aligned}$$

where $\hat{m}\left(1, x_i\right)$ and $\hat{m}\left(0, x_i\right)$ are the regression model estimations of treated and control outcomes. The DR estimator is consistent and therefore asymptotically unbiased, if either the propensity score is correct or the model correctly reflects the true relationship among exposure and confounders with the outcome [38]. In reality, one definitely cannot guarantee whether one model can accurately explain the relationship among variables. The combination of outcome regression with weighting by propensity score ensures that the estimators are robust to misspecification of one of these models.

The DR estimator consults outcomes to make the IPW estimator robust when propensity score estimation is not correct.

1.8.2 Stratification methods

把整个样本分成若干个同质的子样本。在理想情况下，每个同质子样本中，处理组和控制组在控制了协变量以后是相似的（即满足假设 2），可以视为随机控制实验，所以每个子样本中的 ATE（CATE）是可以估计出来的。

在得到 CATE 以后，可以使用这些子样本的 CATE 合成我们感兴趣的样本的 ATE。

In the following, we adopt the calculation of ATE as an example. In detail, if we separate the whole dataset into J blocks, the ATE is estimated as:

$$\text{ATE}_{\text{strat}} = \hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \left[\bar{Y}_t(j) - \bar{Y}_c(j)\right],$$

where $\bar{Y}_t(j)$ and $\bar{Y}_c(j)$ are the average of the treated outcome and control outcome in the j-th block, respectively. $q(j) = \frac{N(j)}{N}$ is the portion of the units in the j-th block to the whole units.

Stratification effectively decreases the bias of ATE estimation compared with the difference-estimator where ATE is estimated as: $\text{ATE}_{\text{diff}} = \hat{\tau}^{diff} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{CF} - \sum_{i:W_i=0} Y_i^{CF}$. In particular, if we assume the outcome is linear with the covariates, i.e., $\mathbb{E}\left[Y_i(w) \mid X_i = x\right] = \alpha + \tau * w + \beta * x$. The bias of the difference-estimator is:

$$\mathbb{E}\left[\hat{\tau}^{\text{diff}} - \tau \mid X, W\right] = \left(\bar{X}_t - \bar{X}_c\right)\beta.$$

While, the bias of the stratification estimator is the weighted average of the within-block bias:

$$\mathbb{E}\left[\hat{\tau}^{\text{strat}} - \tau \mid X, W\right] = \left(\sum_{j=1}^{J} q(j)\left(\bar{X}_t(j) - \bar{X}_c(j)\right)\right)\beta.$$

Compared with the difference estimator, the stratification estimator reduces the bias per covariate by the factor:

$$\gamma_k = \frac{\sum_j q(j)\left(\bar{X}_{t,k}(j) - \bar{X}_{c,k}(j)\right)}{\bar{X}_{t,k} - \bar{X}_{c,k}}$$

where $\bar{X}_{t,k}(j)$ $\left(\bar{X}_{c,k}(j)\right)$ is the average of k-th covariate of treated (control) group in j-th block, and $\bar{X}_{t,k}$ $\left(\bar{X}_{c,k}\right)$ is the average of k-th covariate in the whole treated (control) group.

The key component of stratification methods is how to create the blocks and how to combine the created blocks.

上述 stratification methods 都是根据处理前定变量分割样本。然而，在一些现实的应用中，需要根据一些处理后变量（post-treatment variables 表示为 S）来比较结果。

例如，疾病研究中的“替代”标志物（即中间结果），如艾滋病患者的 CD4 计数和病毒载量测量，都是处理后变量。

在比较艾滋病患者的药物的研究中，研究人员对 CD4 计数低于 200cell/mm^3 的群体上艾滋病药物的效果感兴趣。然而，直接比较观察到的 $S^{obs} < 200$ 群体的结果并不是真实效应，因为比较的两个子组: $\left\{i : W_i = 1, S^{obs} < 200\right\}$ 和 $\left\{j : W_j = 0, S^{obs} < 20\right\}$，其中 S^{obs} 是观察到的后处理值，如果治疗对中间结果有影响，则有很大的不一致性。

为了解决这个问题，主分层基于处理前定变量的潜在值构建子组。潜在预处理变量值，表示为 $S(W = w)$，是在值 w 的处理下 S 的潜在值。在潜在值的自然假设下，S 的潜在值与处理分配独立，子组的处理效应可以通过比较两组结果获得:

$$\left\{Y_i^{obs} : W_i = 1, S_i\left(W_i = 1\right) = v_1, S_i\left(W_i = 0\right) = v_2\right\}$$

和

$$\left\{Y_j^{obs} : W_j = 0, S_j\left(W_j = 1\right) = v_1, S_j\left(W_j = 0\right) = v_2\right\},$$

其中 v_1 和 v_2 是两个后处理值。基于后处理变量的潜在值进行比较确保了比较的两组相似，从而获得的处理效应是真实的效应。

1.8.3 Tree-based methods

使用分类或回归树等方法，把全样本分割成子样本，研究子样本中的因果关系。

Bayesian Regression Tree Models for Causal Inference 是一种统计模型，它结合了贝叶斯方法和回归树（如随机森林）来估计因果效应。这种模型特别适用于处理观察数据中的异质性治疗效果（heterogeneous treatment effects)、小效应大小（small effect sizes）以及由可观测变量引起的强烈混淆（strong confounding by observables）的情况。

这种模型的一个关键优势是它能够提供对处理效应的丰富推断，包括在不同水平上的平均处理效应（average treatment effects）和条件平均处理效应（conditional average treatment effects）的估计。

2 工具变量

- An explanatory variable is said to be endogenous if it is correlated with the error term. Endogeneity will cause OLS estimator to be inconsistent.

$$\begin{aligned}\mathbf{b} &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{'}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)\\ &\to_p \beta + (\mathrm{E}(\mathbf{x}_i^{'}\mathbf{x}_i))^{-1}(\mathrm{E}(\mathbf{x}_i\varepsilon_i)) \neq \beta\end{aligned}$$

if $\mathrm{E}(\mathbf{x}_i\varepsilon_i) \neq \mathbf{0}$.
In the simple regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,

$$b_1^{OLS} = \frac{\widehat{\mathrm{Cov}(y_i, x_i)}}{\widehat{\mathrm{Var}(x_i)}} \to_p \frac{\mathrm{Cov}(y_i, x_i)}{\mathrm{Var}(x_i)} = \beta_1 + \frac{\mathrm{Cov}(\varepsilon_i, x_i)}{\mathrm{Var}(x_i)} \neq \beta_1$$

if $\mathrm{E}(x_i\varepsilon_i) \neq 0$.

•Instrumental variable solution. A predetermined variable that is correlated with the endogenous regressor is called an **instrumental variable**, denoted z_i.

$$\begin{aligned}\text{Cov}(y_i, z_i) &= \text{Cov}(\beta_0 + \beta_1 x_i + \varepsilon_i, z_i) \\ &= \beta_1 \text{Cov}(x_i, z_i) + \text{Cov}(\varepsilon_i, z_i)\end{aligned}$$

$$b_1^{IV} \triangleq \frac{\widehat{\text{Cov}(y_i, z_i)}}{\widehat{\text{Cov}(x_i, z_i)}} \to_p \frac{\text{Cov}(y_i, z_i)}{\text{Cov}(x_i, z_i)} = \beta_1 + \frac{\text{Cov}(\varepsilon_i, z_i)}{\text{Cov}(x_i, z_i)} = \beta_1$$

if $\text{Cov}(\varepsilon_i, z_i) = 0$ and $\text{Cov}(x_i, z_i) \neq 0$.

- Two-stage least squares.
  - Stage 1: Regress x_i on z_i using OLS, and obtain $\hat{x}_i$.

$$x_i = \gamma_0 + \gamma_1 z_i + \omega_i = \hat{x}_i + \omega_i$$

  - Stage 2: Regress y_i on $\hat{x}_i$ using OLS.

$$y_i = \beta_0 + \beta_1 \hat{x}_i + [\varepsilon_i + \beta_1(x_i - \hat{x}_i)]$$

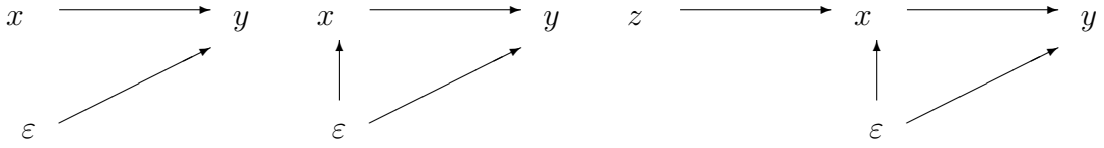
$$b_1^{2SLS} \triangleq \frac{\widehat{\text{Cov}(y_i, \hat{x}_i)}}{\widehat{\text{Var}(\hat{x}_i)}}$$

  - The second stage regressor satisfies orthogonality condition:

$$\text{Cov}(\hat{x}_i, \varepsilon_i + \beta_1(x_i - \hat{x}_i)) = \text{Cov}(\hat{x}_i, \varepsilon_i) + \beta_1 \text{Cov}(\hat{x}_i, \omega_i) = 0$$

- The equivalence between b_1^{IV} and b_1^{2SLS}.

•Summary of the basic idea.



–Relevance: IV should be correlated with the endogenous explanatory variable.

$$\text{Cov}(z, x) \neq 0$$

–Exclusion: IV should not appear on the right hand side of the structural equation.

–Exogeneity (independence): IV should be uncorrelated with the error term.

$$\text{Cov}(z, \varepsilon) = 0$$

2.1 内生性的种种来源

Simultaneity Bias

Simultaneity Bias arising from equilibrium conditions.

- Demand curve

$$q_i^d = \alpha_0 + \alpha_1 p_i + u_i$$

- Supply curve

$$q_i^s = \beta_0 + \beta_1 p_i + v_i$$

- Market equilibrium

$$q_i^d = q_i^s$$

- Simplifying assumptions.

$$\mathrm{E}(u_i) = \mathrm{E}(v_i) = \mathrm{Cov}(u_i, v_i) = 0$$

•Price is endogenous in both demand and supply equations.

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}$$
$$q_i = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1}$$

$$\text{Cov}(p_i, u_i) = -\frac{\text{Var}(u_i)}{\alpha_1 - \beta_1}$$
$$\text{Cov}(p_i, v_i) = \frac{\text{Var}(v_i)}{\alpha_1 - \beta_1}$$

•Regressing q_i on p_i (and a constant) does not estimate either the demand curve or the supply curve consistently.

$$
\begin{aligned}
b_{OLS} &= \frac{\widehat{\text{Cov}(p_i, q_i)}}{\widehat{\text{Var}(p_i)}} \rightarrow_p \frac{\text{Cov}(p_i, q_i)}{\text{Var}(p_i)} \\
&= \frac{\alpha_i \text{Var}(p_i) + \text{Cov}(p_i, u_i)}{\text{Var}(p_i)} \text{ or } \frac{\beta_1 \text{Var}(p_i) + \text{Cov}(p_i, v_i)}{\text{Var}(p_i)} \\
&= \alpha_1 + \frac{\text{Cov}(p_i, u_i)}{\text{Var}(p_i)} \text{ or } \beta_1 + \frac{\text{Cov}(p_i, v_i)}{\text{Var}(p_i)}
\end{aligned}
$$

- In order to consistently estimate (for example) the demand slope, we need to find an IV for p_i in the demand equation, an observable factor that is predetermined w.r.t. the demand curve, but correlated with p_i. A natural candidate is a pure supply shifter.

- Suppose the supply equation can be written as

$$q_i^s = \beta_0 + \beta_1 p_i + \beta_2 z_i + \zeta_i, \text{Cov}(z_i, \zeta_i) = 0$$

and z_i does not affect demand,

$$\text{Cov}(z_i, u_i) = 0$$

then

$$\alpha_1^{IV} = \frac{\widehat{\text{Cov}(q_i, z_i)}}{\widehat{\text{Cov}(p_i, z_i)}}$$

To see why the IV estimator works in this specific case,

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} z_i + \frac{\zeta_i - u_i}{\alpha_1 - \beta_1}$$
$$q_i = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1} z_i + \frac{\alpha_1\zeta_i - \beta_1 u_i}{\alpha_1 - \beta_1}$$

$$\text{Cov}(p_i, z_i) = \frac{\beta_2}{\alpha_1 - \beta_1}\text{Var}(z_i) \neq 0$$

$$\text{Cov}(q_i, z_i) = \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1}\text{Var}(z_i)$$

$$\alpha_1^{IV} \to_p \frac{\text{Cov}(q_i, z_i)}{\text{Cov}(p_i, z_i)} = \alpha_1$$

Simultaneity Bias arising from reverse causalities.

•"Institutions affect economic performance" .

$$g_i = \alpha_0 + \alpha_1 d_i + u_i$$

•"Rich economies choose or can afford better institutions."

$$d_i = \beta_0 + \beta_1 g_i + v_i$$

Omitted Variables Bias

- The true model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \gamma q + \varepsilon$$

- q is unobservable. What we estimate is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u, u \triangleq \gamma q + \varepsilon$$

- Write the linear projection of q onto the observable regressors,

$$q = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_K x_K + v$$

$$\mathrm{E}(v) = 0, \mathrm{Cov}(x_k, v) = 0, k = 1, 2, \cdots, K$$

$$y = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1)x_1 + (\beta_2 + \gamma\delta_2)x_2 + \cdots + (\beta_K + \gamma\delta_K)x_K + (\gamma v + \varepsilon)$$

- It is easy to see

$$\mathrm{plim}_{n\to\infty} b_k = \beta_k + \gamma\delta_k$$

•Suppose all δ_k except δ_K are zero, then

$$\text{plim}_{n\to\infty} b_k = \beta_k, k = 1, 2, \cdots, K-1$$

$$\text{plim}_{n\to\infty} b_K = \beta_K + \gamma \frac{\text{Cov}(x_K, q)}{\text{Var}(x_K)}$$

•For example, x_K denotes years of schooling and q denotes unobserved ability. A more able person tends to have higher wage ($\gamma > 0$), and is also likely to receive more education ($\text{Cov}(x_K, q) > 0$), therefore OLS will overestimate the return on schooling.

OLS Solution: Find a proxy.

•Use IQ, denoted by z, as a proxy variable for unobserved ability.

$$q = \theta_0 + \theta_1 z + \omega, \text{Cov}(z, \omega) = 0$$

•Qualifications for a valid proxy.

1.Redundancy: z is irrelevant for explaining y once **x** and q have been controlled for. (For example, we don't need to control for IQ if ability were observable and hence controlled.)

$$\text{Cov}(z, \varepsilon) = 0$$

2.The correlation between q and **x** is zero once z is partialled out.

$$\text{Cov}(\mathbf{x}, \omega) = 0$$

•Consistent OLS estimator with proxy.

$$y = (\beta_0 + \gamma\theta_0) + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma\theta_1 z + (\gamma\omega + \varepsilon)$$

$$\text{Cov}(\mathbf{x}, \gamma\omega + \varepsilon) = 0, \text{Cov}(z, \gamma\omega + \varepsilon) = 0$$

- Including a proxy variable can actually reduce asymptotic variances as well as mitigate bias because $\text{Var}(\gamma\omega + \varepsilon) < \text{Var}(\gamma q + \varepsilon)$.
- The second qualification does not sound very intuitive. A necessary but not sufficient condition is that the proxy variable has to be correlated with the omitted variable. Including an irrelevant variable as proxy will actually exacerbate the inconsistency of OLS estimator. *Sketch of proof:* Assume a simple structural model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where an omitted variable is hidden in ε.

$$b_{\text{ov}} \to_p \beta_1 + \frac{\text{Cov}(x, \varepsilon)}{\text{Var}(x)}$$

•Suppose we choose z as a proxy for the omitted variable, and regress y on x and z. In light of FWL, the coefficient estimate on x is equal to the coefficient estimate on v in the regression of y on v, where v is the residual of projecting x on z.

$$x = \pi_0 + \pi_1 z + v$$

$$y = (\beta_0 + \beta_1 \pi_0) + \beta_1 v + (\varepsilon + \beta_1 \pi_1 z)$$

$$b_{\text{proxy}} \to_{\text{p}} \beta_1 + \frac{\text{Cov}(v, \varepsilon + \beta_1 \pi_1 z)}{\text{Var}(v)} = \beta_1 + \frac{\text{Cov}(v, \varepsilon)}{\text{Var}(v)}$$

If z is irrelevant, i.e., z is uncorrelated with the omitted variable and hence with ε.

$$\text{Cov}(z, \varepsilon) = 0$$

$$\text{Cov}(x, \varepsilon) = \text{Cov}(v, \varepsilon)$$

$$\text{Var}(x) = \pi_1^2 \text{Var}(z) + \text{Var}(v) > \text{Var}(v)$$

$$\left| \frac{\text{Cov}(x, \varepsilon)}{\text{Var}(x)} \right| < \left| \frac{\text{Cov}(v, \varepsilon)}{\text{Var}(v)} \right|$$

Proxy vs. IV. If we are unable to find a valid proxy (for ability), then we try to find a valid instrument (for years of schooling). Both a proxy variable and an instrument variable must be redundant (do not appear in the true model that explicitly contains the omitted variable). However, a proxy is with regard to the omitted variable, while an IV is with regard to the endogenous explanatory variable. In other words, a proxy should be highly correlated with the omitted variable, while an IV should be uncorrelated with the omitted variable. Therefore, a proxy makes a poor IV, and an IV makes a poor proxy.

Measurement Error

- The true model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{K-1} x_{K-1} + \beta_K x_K^* + \varepsilon$$

- Measurement error.

$$x_K = x_K^* + e_K$$

$$\mathrm{E}(e_K) = 0, \mathrm{Cov}(x_k, e_K) = 0 \forall k \neq K, \mathrm{Cov}(\varepsilon, e_K) = 0$$

- The classical errors-in-variables (CEV) assumption.

$$\mathrm{Cov}(x_K^*, e_K) = 0$$

- In some cases it is clear that the CEV assumption cannot be true.

- What we estimate is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u, u \triangleq \varepsilon - \beta_K e_K$$

$$\mathrm{Cov}(x_K, u) = -\beta_K \mathrm{Cov}(x_K, e_K) = -\beta_K \sigma_{e_K}^2 \neq 0$$

- OLS estimators of all β_k are inconsistent.
- The biases of b_k $(k \neq K)$ are difficult to characterize in general, but $\text{plim}_{n\to\infty} b_K$ can be characterized in any case. Attenuation bias (bias toward zero).
- Write linear projections of x_K^* and x_K onto the other regressors.

$$x_K^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_{K-1} x_{K-1} + r_K^*$$
$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_{K-1} x_{K-1} + r_K$$

$$r_K = r_K^* + e_K$$

- In cases where there is a single regressor $(K = 1)$ measured with error or $x_K^*(x_K)$ is uncorrelated with all other $\mathbf{x}_{-K}$,

$$b_K \to_p \beta_K \left(1 - \frac{\text{Var}(e_K)}{\text{Var}(x_K)}\right)$$

- Since $\frac{\text{Var}(e_K)}{\text{Var}(x_K)} < 1$, the projection coefficient shrinks the structural parameter ¯ towards zero. This is called measurement error bias or attenuation bias.
- Again, we shall try to find an instrument for the mismeasured variable.

Measurement error vs. proxy. The measurement error problem has a statistical structure similar to the proxy variable problem, but they are conceptually very different. In the proxy variable case, we are looking for a variable that is **somehow associated with the omitted variable (unobservable and usually not well-defined) in order to cope with the endogeneity of other explanatory variables**. We cannot estimate the effect of the omitted variable per se. In the measurement error case, the variable that we do not observe **has a well-defined quantitative meaning but our measure of it may contain error**. The mismeasured explanatory variable is the very one whose effect is of primary interest and its own endogeneity is what we are addressing.

Suppose we are estimating the effect of peer group behavior on individual learning output, where the behavior of one's peer group is self-reported. Self-reporting may be a mismeasure of actual peer group behavior. Does it cause a problem?

Repeated measurement solution.

•The true model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{K-1} x_{K-1} + \beta_K x_K^* + \varepsilon$$

•Measurement error.

$$x_K = x_K^* + e_1$$

$$\text{Cov}(x_K^*, e_1) = \text{Cov}(\varepsilon, e_1) = \text{Cov}(x_k, e_1) = 0, \forall k \neq K$$

•OLS estimator is inconsistent.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + (\varepsilon - \beta_K e_1)$$

•Assume there exists a second mismeasured variable $x_{K,2}$ that satisfies the same assumptions as $x_{K,1}$,

$$x_{K,2} = x_K^* + e_2$$

If $\text{Cov}(e_1, e_2) = 0$ then $x_{K,2}$ can be used as an IV for $x_{K,1}$.

–验证 $x_{K,2}$ 为什么可以是 IV：

$$\text{Cov}(x_{K,2}, \varepsilon - \beta_K e_1) = 0$$

2.2 两阶段最小二乘法

•**回顾 OLS 估计量的推导:**

$$S(\beta) = \mathrm{E}\left[(Y - X'\beta)^2\right]$$

$$\beta = \arg\min_b S(b)$$

$$\mathbb{E}[X\varepsilon] = 0$$

、

$$\begin{aligned}\mathbb{E}[Xe] &= \mathbb{E}\left[X\left(Y - X'\beta\right)\right] \\ &= \mathbb{E}[XY] - \mathbb{E}\left[XX'\right]\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY] \\ &= 0\end{aligned}$$

•Key assumption 1 (moment condition): $\mathbf{z}_i$ is predetermined.

$$\mathrm{E}(\mathbf{z}_i\varepsilon_i) = \mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)] = 0$$

•Key assumption 2: $\mathbf{z}_i$ and $\mathbf{x}_i$ are sufficiently linearly correlated.

•Method of moments: A method of estimating population moments (expectation of function of random vectors) with sample analogue (sample mean). OLS estimation is a method of moments.

•IV estimation as a method of moments.

$$\mathrm{E}(\mathbf{z}_i\varepsilon_i) = \mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)] = 0$$
$$\mathrm{E}(\mathbf{z}_i\mathbf{x}_i')\beta = \mathrm{E}(\mathbf{z}_i y_i)$$
$$\beta = (\mathrm{E}(\mathbf{z}_i\mathbf{x}_i'))^{-1}\mathrm{E}(\mathbf{z}_i y_i)$$
$$\mathbf{b}_{\mathrm{IV}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i\right) = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

- $\mathbf{b}_{\text{IV}}$ is a consistent estimator of β and is asymptotically normally distributed.
- IV estimation applies to just-identified cases only.
- Linear combinations of IVs are still valid IVs. In overidentified cases we can construct K combinations of all L available IVs to make it just identified. 2SLS offers such a way of construction.

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})\mathbf{Z}'\mathbf{X}$$

$$\begin{aligned}\mathbf{b}_{\text{2SLS}} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\end{aligned}$$

- $\mathbf{b}_{\text{2SLS}}$ is a consistent estimator of β and is asymptotically normally distributed.
- What goes wrong when we do the two stages manually?

2.3 Forbidden Regression

- Definition: improper use of instruments leads to inconsistency.
  - Case 1: Linear projection is incomplete.
  - Case 2: Expectation operator does not pass through nonlinear functions (of either the first or the second stage).
- Consider a model,

$$y_1 = z_1\delta_1 + \alpha y_2 + \alpha_2 y_2^2 + u_1$$

- The model is nonlinear in endogenous variables.
  - Step 1: $y_2 = z_1\pi_{21} + z_2\pi_{22} + v_2$. Let the predicted value be $\hat{y}_2$.
  - Step 2: run regression $y_1 = z_1\delta + \alpha_1\hat{y}_2 + \alpha_2(\hat{y}_2)^2 + e$
- This regression is sometimes called forbidden regression. It is wrong. The reason for this is easy:

$$E(y^2) \neq (E(y))^2$$

- The correct method should be:
  - Step 1: run two regressions:
    $y_2 = z_1\pi_{21} + z_2\pi_{22} + v_2$. Let the predicted value be $\hat{y}_2$.

$y_2^2 = g(z_1, z_2) + e$. Let the predicted value be $\hat{y}_2^2$, and $g(z_1, z_2)$ could be a nonlinear function of z_1 and z_2. For example:

$$y_2^2 = z_1\eta_1 + z_2\eta_2 + z_1^2\eta_3 + z_2^2\eta_4 + z_1z_2\eta_5 + v$$

And use the predicted value from this regression.

–Step 2:run 2SLS as in :

$$y_1 = z_1\delta + \alpha_1\hat{y}_2 + \alpha_2\hat{y}_2^2 + e$$

•Define a modified 2SLS estimator as $\hat{\beta}^{M2SLS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y}$ where $\tilde{\mathbf{X}}$ is an estimator of $E(\mathbf{X}|\mathbf{Z})$. Define an indirect least squares (ILS) estimator as $\hat{\beta}^{ILS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y}$.

2.4 相关检验

- Key reference: Baum et al. (2007, SJ7-4). "Enhanced Routines for Instrumental Variables/Generalized Method of Moments Estimation and Testing."

Testing the relevance of instruments.

- Test the joint significance of the excluded instruments in the first-stage regression. Rule of thumb when there is only one endogenous regressor: $Fstat > 10$. F -statistic might be misleading when there are multiple endogenous regressors.
- Stock-Yogo test of weak instruments.
  - H0: The set of instruments is weak. （第一阶段相关性不足，导致估计偏误超过可接受阈值）
  - Two characterizations of weak instruments:
    1. 2SLS bias can be large, and is in the same direction of OLS bias.
    2. 2SLS leads to size distortion of the joint significance test of the endogenous regressors.

    仍然是有偏的，同时导致系数膨胀问题。
  - The critical values (and hence the decisions) depend on the largest relative bias or the largest rejection rate we are willing to tolerate.
  - Cragg-Donald 统计量：适用于独立同分布（iid）误差假设，通过计算工具变量对内生变量的联合显著

性（F 统计量的推广形式）衡量相关性强度。Kleibergen-Paap 统计量：适用于异方差或自相关的误差结构，是 Cragg-Donald 统计量的稳健版本。

- IV 的系数膨胀问题。一般应控制在 3 倍以内。超过 5 倍的，很有可能是弱工具变量问题。

Testing overidentifying restrictions.

- When we have more instruments than needed to identify an equation, we can test whether the instruments are valid in the sense that they are uncorrelated with the error term.
- These are tests of the joint hypotheses of correct model specification and the orthogonality conditions. Rejection may be either because instruments are not truly exogenous, or because they are incorrectly excluded from the regression. Moreover, it may be either because the excluded instruments are not good, or because the predetermined regressors are actually endogenous.
- **假定 l 为工具变量个数，k 为内生变量个数，如果 $l > k$，就意味着矩（moments）的数量比待估计系数的数量多。因此需要做过度识别检验。**
- The instrumental variables model specifies $\mathbb{E}[Ze] = 0$. Equivalently, since $e = Y - X'\beta$ this is

$$\mathbb{E}[ZY] - \mathbb{E}\left[ZX'\right]\beta = 0.$$

This is an $\ell \times 1$ vector of restrictions on the moment matrices $\mathbb{E}[ZY]$ and $\mathbb{E}\left[ZX'\right]$. Yet since β is of dimension k which is less than ℓ it is not certain if indeed such a β exists.

To make things a bit more concrete, suppose there is a single endogenous regressor X_2, no X_1, and two instruments Z_1 and Z_2. Then the model specifies that

$$\mathbb{E}\left(\left[Z_1 Y\right] = \mathbb{E}\left[Z_1 X_2\right]\beta\right.$$

and

$$\mathbb{E}\left[Z_2 Y\right] = \mathbb{E}\left[Z_2 X_2\right] \beta.$$

Thus β solves both equations. This is rather special.

•For a general overidentification test the null and alternative hypotheses are $\mathbb{H}_0 : \mathbb{E}[Ze] = 0$ against $\mathbb{H}_1 : \mathbb{E}[Ze] \neq 0$. We will also add the conditional homoskedasticity assumption

$$\mathbb{E}\left[e^2 \mid Z\right] = \sigma^2.$$

it is best to take a GMM approach。To implement a test of $\mathbb{H}_0$ consider a linear regression of the error e on the instruments Z

$$e = Z'\alpha + v$$

with $\alpha = \left(\mathbb{E}\left[ZZ'\right]\right)^{-1}\mathbb{E}[Ze]$. We can rewrite $\mathbb{H}_0$ as $\alpha = 0$. While e is not observed we can replace it with the 2SLS residual $\widehat{e}_i$ and estimate α by least squares regression, e.g. $\widehat{\alpha} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}}$. Sargan (1958) proposed testing $\mathbb{H}_0$ via a score test, which equals

$$S = \widehat{\alpha}'(\widehat{\text{var}}[\hat{\alpha}])^{-}\widehat{\alpha} = \frac{\widehat{\boldsymbol{e}}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2}.$$

where $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$. Basmann (1960) independently proposed a Wald statistic for $\mathbb{H}_0$, which is S with $\widehat{\sigma}^2$ replaced with $\widetilde{\sigma}^2 = n^{-1}\widehat{\boldsymbol{v}}'\widehat{\boldsymbol{v}}$ where $\hat{\boldsymbol{v}} = \widehat{\boldsymbol{e}} - \boldsymbol{Z}\widehat{\alpha}$. By the equivalence of homoskedastic score and Wald tests (see Section 9.16) Basmann's statistic is a monotonic function of Sargan's statistic and hence they yield equivalent tests. Sargan's version is more typically reported.

The Sargan test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $S > c$ for some critical value c. An asymptotic test sets c as the $1-\alpha$ quantile of the $\chi^2_{\ell-k}$ distribution. This is justified by the asymptotic null distribution of S which we now derive.

•Hansen’ s J (Sargan) test.

$$S_n \rightarrow_d \chi^2(l-k)$$

- Testing a subset of overidentifying restrictions: Hayashi's C (difference- in-Sargan) test.
- L_1 个变量可以确信与 ε 不相关，而另外 L_2 个变量可能相关。如果加入 L_2 个变量能够显著提高 S_n，那么就有理由怀疑这些变量和 ε 是相关的。
- Suppose we can divide the L instruments into two groups: L_1 variables that are known to satisfy the moment conditions, and L_2 variables that are suspect. The moment conditions regarding L_2 are testable if $L_1 \geq K$. The idea is to compare two S_n from two separate GMM estimators of the same regression, one using only L_1 instruments, and the other using a full set of L instruments. If the inclusion of L_2 suspect instruments significantly increases S_n, that is a good reason for doubting the predeterminedness of the L_2 instruments.

$$C \triangleq S - S_1 \rightarrow_d \chi^2(L - L_1)$$

- If the unrestricted equation is exactly identified, the J statistic for the unrestricted equation will be zero and the C statistic will coincide with the J statistic for the original (restricted) equation, and this will be true irrespective of the instruments used to identify the unrestricted estimation. Therefore the overidentification test is an test for the failure of any of the instruments to satisfy the orthogonality conditions, but at the same time requires that the investigator believe that at least some of the instruments are valid.
- Even if partially testable, the exogeneity of instruments has to be justified mainly from a theoretical ground.

Testing for endogeneity of the regressors.

- H0: The regressors of interest are exogenous.
- Hayashi's C test.

Indirect test of the exclusion restriction. In samples where the first stage is zero, the reduced form should be zero as well. On the other hand, a statistically significant reduced-form estimate with no evidence of a corresponding first stage is cause for worry, because this suggests some channel other than the treatment variable links instruments with outcomes. We can construct "no-first-stage samples" and check whether they generate no evidence of significant reduced-form effects (Angrist and Pischke, 2014).

TABLE 7—REDUCED FORM RELATIONSHIP BETWEEN THE DISTANCE FROM THE COAST AND TRUST WITHIN AFRICA AND ASIA

	Trust of local government council			
	Afrobarometer sample		Asiabarometer sample	
	(1)	(2)	(3)	(4)
Distance from the coast	0.00039***	0.00031***	−0.00001	0.00001
	(0.00009)	(0.00008)	(0.00010)	(0.00009)
Country fixed effects	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes
Number of observations	19,913	19,913	5,409	5,409
Number of clusters	185	185	62	62
R^2	0.16	0.18	0.19	0.22

Notes: The table reports OLS estimates. The unit of observation is an individual. The dependent variable in the Asiabarometer sample is the respondent's answer to the question: "How much do you trust your local government?" The categories for the answers are the same in the Asiabarometer as in the Afrobarometer. Standard errors are clustered at the ethnicity level in the Afrobarometer regressions and at the location (city) level in the Asiabarometer and the WVS samples. The individual controls are for age, age squared, a gender indicator, education fixed effects, and religion fixed effects.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

TABLE 8—REDUCED FORM RELATIONSHIP BETWEEN THE DISTANCE FROM THE COAST AND TRUST WITHIN AND OUTSIDE OF AFRICA

	Intergroup trust				
	Afrobarometer sample		WVS non-Africa sample		WVS Nigeria
	(1)	(2)	(3)	(4)	(5)
Distance from the coast	0.00039***	0.00037***	−0.00020	−0.00019	0.00054***
	(0.00013)	(0.00012)	(0.00014)	(0.00012)	(0.00010)
Country fixed effects	Yes	Yes	Yes	Yes	n/a
Individual controls	No	Yes	No	Yes	Yes
Number of observations	19,970	19,970	10,308	10,308	974
Number of clusters	185	185	107	107	16
R^2	0.09	0.10	0.09	0.11	0.06

Notes : The table reports OLS estimates. The unit of observation is an individual. The dependent variable in the WVS sample is the respondent's answer to the question: "How much do you trust <nationality> people in general?" The categories for the respondent's answers are: "not at all," "not very much," "neither trust nor distrust," "a little," and "completely." The responses take on the values 0, 1, 1.5, 2, and 3. Standard errors are clustered at the ethnicity level in the Afrobarometer regressions and at the location (city) level in the Asiabarometer and the WVS samples. The individual controls are for age, age squared, a gender indicator, an indicator for living in an urban location, and occupation fixed effects.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Falsification test of plausibly endogenous IV.

- Key reference: Conley et al. (2012, REStat).

$$y = X\beta + Z\gamma + \varepsilon$$
$$X = Z\Pi + V$$

$$\hat{\beta} = (Z'X)^{-1}Z'Y \rightarrow_p \beta + \gamma/\Pi$$

There is typically a trade-off between **instrument strength** and **degree of violation of the exclusion restriction**.

- 具体做法：
  - 假设我们知道 γ 的支持集 G，即 γ 可能取值的范围。例如，我们可能认为 γ 在一个区间 $[-\delta, \delta]$ 内。
  - 对于 G 中的每一个可能的 γ 值 γ_0，我们都可以基于模型（1）进行估计，并得到一个关于 β 的置信区间。

$$(Y - Z\gamma_0) = X\beta + \varepsilon \tag{1}$$

Estimate β and construct a symmetric $(1-\alpha)$ confidence interval in the usual way:

$$CI(1-\alpha, \gamma_0) = [\hat{\beta}(\gamma_0) \pm z_{1-\alpha/2}\widehat{SE(\hat{\beta}(\gamma_0))}]$$

–通过改变 γ_0，我们可以为 G 中的每一个可能值都得到一个这样的置信区间。

$$CI(1-\alpha) = \cup_{\gamma_0 \in \Gamma} CI(1-\alpha, \gamma_0)$$

- 这种方法的优点是它不需要对 γ 的分布做出完全的假设，只需要知道 γ 的支持集。然而，这种方法的一个缺点是得到的置信区间可能会很宽，因为它在所有可能的情况下都要求正确的覆盖率，包括那些研究者可能认为不太可能的情况。
- 改进做法（B)：研究者需要指定一个关于 γ 的先验概率分布。这个分布反映了研究者对 · 可能取值的信念。例如，如果研究者认为 γ 很可能接近 0，那么在 0 附近的值应该有更高的先验概率。然后计算每个 γ_0 对应的 β 的置信区间，再将这些置信区间根据其 γ_0 先验概率进行加权，并取并集。
- 进一步改进（C)：

- γ local-to-zero approximation (局部到零).
- γ 被视为一个随机变量，并且与样本大小 N 成比例地缩小: $\gamma = \eta/\sqrt{N}$，η 是服从某个分布的随机变量。这种设定允许当 N 趋向于无穷大时，γ 趋向于 0。
- 利用上述设定，可以推导 2SLS 的 β 在 γ 局部到零的近似分布。这个分布包含了两部分：一部分是传统的 2SLS 渐近分布，另一部分是反映 γ 不确定性的影响。
- 为了简化计算，文章中通常假设 γ 服从正态分布。这样，β 的近似分布可以表示为正态分布 Suppose

$$\eta \sim N(\mu, \Omega)$$

$$\begin{aligned}\hat{\beta} &= (X'P_Z X)^{-1}X'P_Z y \\ &= (X'P_Z X)^{-1}X'P_Z(X\beta + Z\gamma + \varepsilon)\end{aligned}$$

$$\begin{aligned}\hat{\beta} &= (X'P_Z X)^{-1}X'P_Z\gamma + (X'P_Z X)^{-1}X'P_Z\varepsilon \\ \hat{\beta} &\sim N(\beta + A\mu, \widehat{Var(\hat{\beta})} + A\Omega A')\end{aligned} \tag{4}$$

where $A = (X'P_Z X)^{-1}X'Z$.

- stata 命令：plausexog

- **示例：国家能力与经济绩效**（Dincecco and Prado, 2012, JEG）.
  - $\log(Y_i/L_i) = \alpha + \beta F_i + \gamma' \mathbf{X_i} + \epsilon_i$
  - $F_i = \lambda + \zeta W_i + \delta' \mathbf{X_i} + v_i$
  - $\log(Y_i/L_i)$：**人均产出**；F_i：**财政能力**；W_i: **每平方公里领土在前现代战争中的伤亡人数**.
  - **问题**：IV **的排除性假设存疑**。

long live Keju:

探讨明清时期科举成功率（进士密度）对当代人力资本（平均受教育年限）的影响。 内生性问题：进士密度可能与未观测因素（如地方文化传统）相关，导致 OLS 估计偏误。

工具变量选择：各府到最近松竹产地的河流距离（bprvdist）。

6.2 Sensitivity analysis

To test the plausibility of our exogeneity assumption even further, we now perform a sensitivity analysis. This analysis evaluates the extent to which our key results regarding the

[34] As an alternative, we used the binary variable for historical colonial status from Comin et al. (2010). The key results did not change. Although it would also be worthwhile to use settler mortality rates from Acemoglu et al. (2001) to instrument for current property rights institutions, severe data limitations prevented this exercise (there are only 36 observations that overlap between Acemoglu et al.'s dataset and ours).

[35] We re-ran our benchmark IV specification using the same 44 observations as for the trust variable. The 2SLS estimate was 3.39 and is significant. As an alternative, we also constructed another control that measured the level of trust that individuals have for people of other nationalities (only 41 observations were available for this variable). The key results were unchanged.

positive performance effect of greater fiscal capacity can withstand violations of the exclusion restriction. We find that these results are indeed robust to moderate violations.

Recall from Sect. 5 that our exclusion restriction is that W_i in Eq. 2 does not appear in Eq 1. Following Conley et al. (2012), the sensitivity analysis instead assumes that W_i does in fact appear in Eq. 1:

$$\log(Y_i/L_i) = \alpha + \beta F_i + \boldsymbol{\gamma}'\mathbf{X_i} + \eta W_i + \epsilon_i. \tag{3}$$

To test how much of a violation of the exclusion restriction could exist before the positive effect of greater fiscal capacity on performance is no longer significant, we consider the following equation

$$\log(Y_i/L_i) - \eta W_i = \alpha + \beta F_i + \boldsymbol{\gamma}'\mathbf{X_i} + \epsilon_i, \tag{4}$$

where we allow η to take values other than zero.

Our inference strategy, which we base on Beber (2009) and Conley et al. (2012), makes the a priori assumption that η is normally distributed with mean zero and variance σ_η^2. We test many values of the standard deviation σ_η, which we vary systematically from 0 to 1 at 0.1 intervals. For each σ_η, we draw 10,000 η values from $\mathcal{N}(0, \sigma_\eta^2)$. The final output generates the percentage of 95% confidence intervals for our coefficient of interest β which are always positive.

Table 8 shows the results of the sensitivity analysis. Panel A reproduces the key findings from the eight 2SLS specifications in Table 5. The first row displays the coefficients for the second-stage relationships of fiscal capacity on economic performance, and the second row the coefficients for the first-stage relationships of past war casualties on fiscal capacity. Panel B shows the percentage of confidence intervals for β which are always positive. The first row of this panel replicates the benchmark 2SLS specification from column 1 of Table 5, where we assume that the exclusion restriction is met exactly and $\eta = 0$. In this case, all confidence intervals for β will always be positive. The second row increases σ_η to 0.1. The percentage of confidence intervals for β which are always positive remains 100 % of the time for most specifications. Moreover, for the three specifications for which this percentage falls below 100 (i.e., columns, 3, 6, and 8), it is still very high, ranging from 77 to 98 % of the time. Although the percentage of confidence intervals for β which are always positive gradually falls as we further increase σ_η, it always includes the majority (and sometimes, the vast majority) of specifications. For the case of $\sigma_\eta = 1$, the percentage of confidence intervals for β which are always positive always exceeds 60 % of the time, excluding the three specifications described above. For these cases, this percentage still occurs the majority of the time.

Summarizing, the sensitivity analysis indicates that our key results are robust to moderate violations of the exclusion restriction. While we still cannot completely exclude endogeneity concerns, this analysis thus provides further support for the plausibility of our IV approach.

虽然 IV 可能一定程度上违反了 exclusion restriction，但是 F 对 Y 直接产生作用的影响仍然很可能是存在的。

Table 8 Sensitivity analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 2SLS regressions from Table 5								
Fiscal capacity on	4.98***	4.44***	4.24***	4.73***	5.30***	4.12**	7.31***	1.81***
performance	(1.16)	(1.19)	(1.36)	(1.31)	(1.32)	(1.70)	(1.85)	(0.62)
War casualties on	0.16***	0.15***	0.14***	0.15***	0.17***	0.12***	0.11***	0.16***
Fiscal capacity	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.03)	(0.05)
Latitude		Yes				Yes		
Democracy			Yes			Yes		
Government size				Yes		Yes		
Trade openness					Yes	Yes		
Area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Continents	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: percentage of confidence intervals over which β is always positive								
σ_η								
0.0	100	100	100	100	100	100	100	100
0.1	100	100	98	100	100	77	100	81
0.2	99	93	84	93	99	64	97	68
0.3	93	84	74	84	93	61	90	62
0.4	86	78	68	77	86	57	83	59
0.5	81	72	65	72	81	56	78	58
0.6	77	68	63	68	77	54	75	56
0.7	74	67	62	67	73	54	71	56
0.8	70	64	60	64	71	54	69	55
0.9	69	63	58	64	69	53	66	54
1.0	67	62	58	63	67	53	65	53

See Table 5 for details
*** Significant at 1 %; ** significant at 5 %; * significant at 10 %

Two important options in ivreg2.

- Two-way clustering: cluster(*varname1 varname2*).
- Partialling out some exogenous regressors: partial, especially useful when using cluster() and the number of clusters is less than L, or when requesting a robust covariance matrix and the regressors include dummies.

stata help:

Partialling-out exogenous regressors

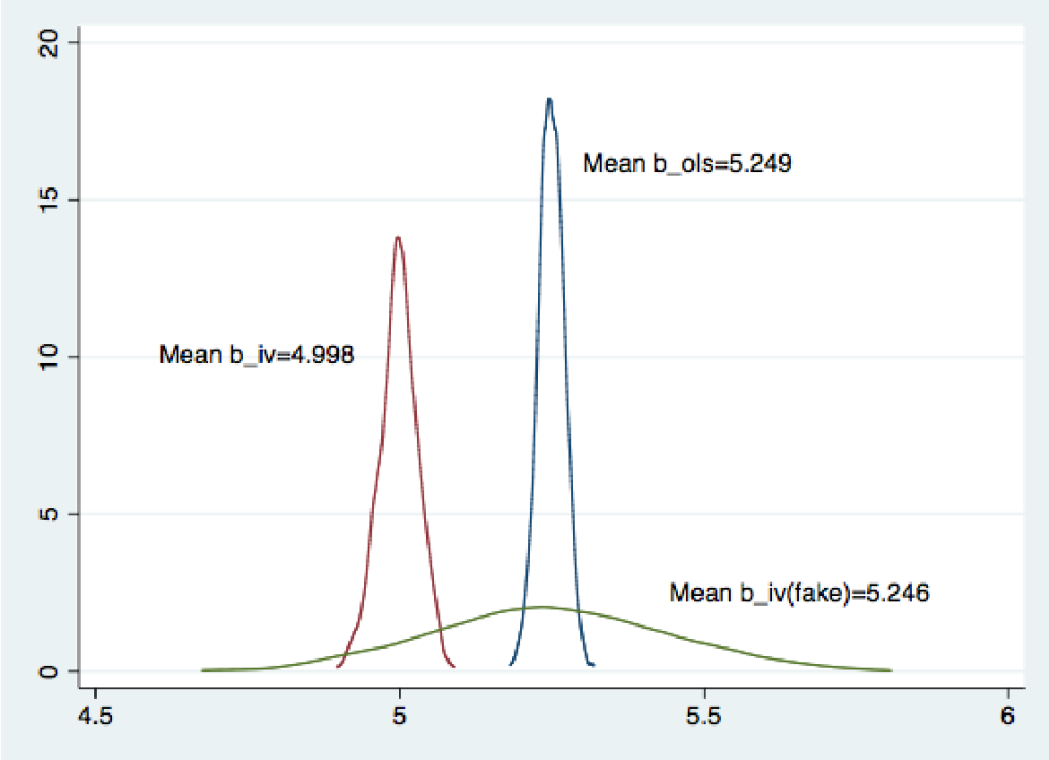
The partial(varlist) option requests that the exogenous regressors in varlist are partialled-out from all other variables (other regressors and excluded instruments) in the estimation. If the equation includes a constant, it is also automatically partialled out as well. The coefficients corresponding to the regressors in varlist are not calculated. By the Frisch-Waugh-Lovell (FWL) theorem in IV, two-step GMM and LIML estimation the coefficients for the remaining regressors are the same as those that would be obtained if the variables were not partialled out. (NB: this does not hold for CUE or GMM iterated more than two steps.) **The partial() option is most useful when using cluster() and ♯clusters < (♯exogenous regressors + ♯ excluded instruments). In these circumstances, the covariance matrix of orthogonality conditions S is not of full rank, and efficient GMM and overidentification tests are infeasible since the optimal weighting matrix $W = S^{-1}$ cannot be calculated.** The problem can be addressed by using partial() to partial out enough exogenous regressors for **S to have full rank**.

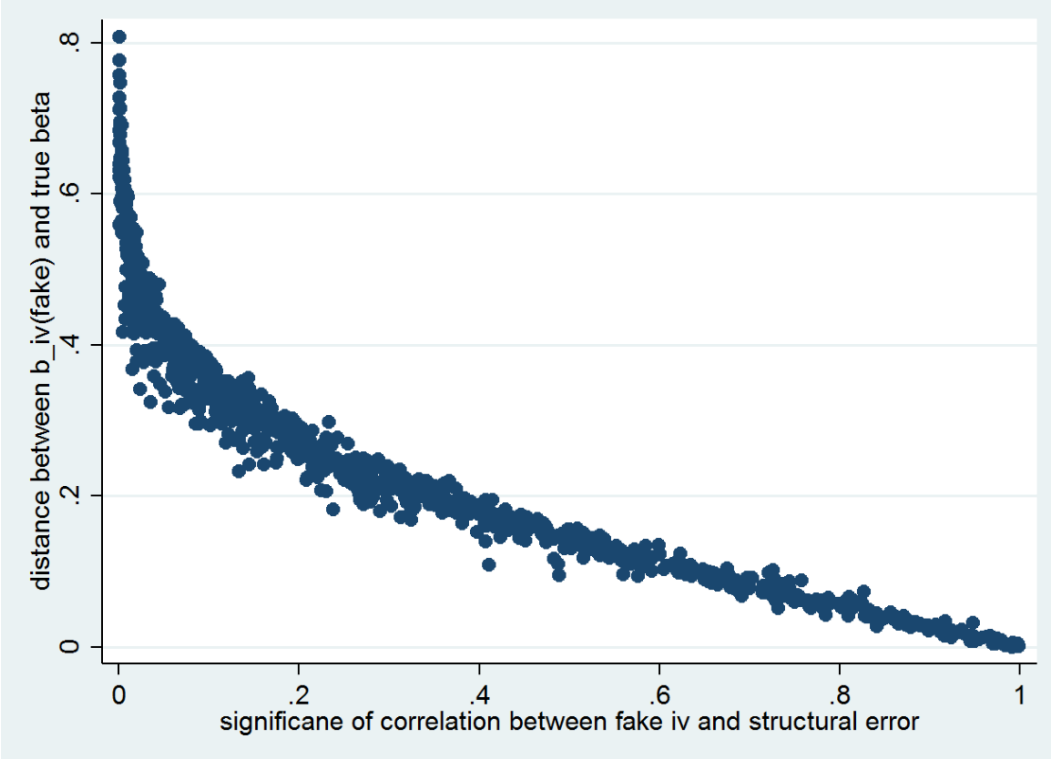
A similar problem arises when **the regressors include a variable that is a singleton dummy, i.e., a**

variable with one 1 and N-1 zeros or vice versa, if a robust-covariance matrix is requested. The singleton dummy causes the robust covariance matrix estimator to be less than full rank. Here partialling-out the variable with the singleton dummy solves the problem. Specifying partial(_cons) will cause just the constant to be partialled-out, i.e., the equation will be estimated in deviations-from-means form. When ivreg2 is invoked with partial(), it reports test statistics with the same small-sample adjustments as if estimating without partial(). After estimation using the partial() option, the postestimation predict can be used only to generate residuals, and that in the current implementation, partial() is not compatible with endogenous variables or instruments (included or excluded) that use time-series operators.

为什么不能自己制造 **IV**？

- **内生性来自 u 和 v 的相关性。**
- **问题：生成一个和 x_1 足够相关的随机变量。既然是任意生成的，就肯定不会影响 y（满足** exclusion restriction**）。这样的变量会是好的** IV **么？**
- Monte-Carlo experiment: Draw u and v from multivariate standard normal distribution with correlation .5. $\beta_1 = 5, \beta_2 = \gamma_1 = \gamma_2 = 1$.


20
15
10
5
0
Mean b_ols=5.249
Mean b_iv=4.998
Mean b_iv(fake)=5.246
4.5
5
5.5
6


distance between b_iv(fake) and true beta
0
.2
.4
.6
.8
0
.2
.4
.6
.8
1
significane of correlation between fake iv and structural error

What should we do in practice?

- Report the first stage result and think about whether it (magnitude and sign) makes sense.
- Report the reduced-form regression of the dependent variable on instruments.
- Pick your best single instrument and report just-identified estimates.
- Check over-identified 2SLS and GMM estimates. Worry if they are very different.
- Carry out specification tests.

●示例

1.Quarter of birth and schooling（Angrist and Krueger, 1991, QJE）.

2.Economic shocks and civil conflict（Miguel et al., 2004, JPE）.

3.Family business succession（Bennedsen et al., 2007, QJE）.

4.Colonial origins of comparative development（Acemoglu et al., 2001, AER）.

5.Tiebout choice in public education（Hoxby, 2000, AER）.

2.5 LATE

- Wald estimator: IV estimator with a binary instrument Z for a binary regressor D.
- $Y = D\beta + \alpha + e$, Z 为工具变量，$\mathbb{E}[e|Z] = 0$。
- Take expectations of the structural equation given $Z = 1$ and $Z = 0$, respectively. We obtain

$$\mathbb{E}[Y \mid Z = 1] = \mathbb{E}[D \mid Z = 1]\beta + \alpha$$
$$\mathbb{E}[Y \mid Z = 0] = \mathbb{E}[D \mid Z = 0]\beta + \alpha.$$

- Subtracting and dividing we obtain an expression for the slope coefficient:

$$\beta = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]} \tag{5}$$

It shows that the structural slope coefficient is the expected change in Y due to changing the instrument divided by the expected change in X due to changing the instrument. Informally, it is the change in Y (due to Z) over the change in X (due to Z).

- The natural moment estimator replaces the expectations by the averages within the "grouped data" where

$Z_i = 1$ and $Z_i = 0$, respectively. That is, define the group means

$$\bar{Y}_1 = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i}, \quad \bar{Y}_0 = \frac{\sum_{i=1}^n (1-Z_i) Y_i}{\sum_{i=1}^n (1-Z_i)}$$
$$\bar{D}_1 = \frac{\sum_{i=1}^n Z_i D_i}{\sum_{i=1}^n Z_i}, \quad \bar{D}_0 = \frac{\sum_{i=1}^n (1-Z_i) D_i}{\sum_{i=1}^n (1-Z_i)}$$

The moment estimator:

$$\widehat{\beta} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$$

It shows that the slope coefficient can be estimated by a the ratio of a difference in means. 下面证明上面这个式子就是 IV 估计量。

$$Y_i = X_i'\widehat{\beta}_{\text{iv}} + \widehat{\alpha}_{\text{iv}} + \widehat{e}_i$$

orthogonality:

$$\sum_{i=1}^{n}\left(Y_i - D_i'\widehat{\beta}_{\text{iv}} - \widehat{\alpha}_{\text{iv}}\right) = 0$$
$$\sum_{i=1}^{n} Z_i\left(Y_i - D_i'\widehat{\beta}_{\text{iv}} - \widehat{\alpha}_{\text{iv}}\right) = 0$$

上式代入下式:

$$\sum_{i=1}^{n} Z_i\left(\left(Y_i - \overline{Y}\right) - \left(D_i - \bar{D}\right)'\widehat{\beta}_{\text{iv}}\right) = 0$$

$$\begin{aligned}\widehat{\beta}_{\text{iv}} &= \left(\sum_{i=1}^{n} Z_i\left(D_i - \bar{D}\right)'\right)^{-1}\left(\sum_{i=1}^{n} Z_i\left(Y_i - \bar{Y}\right)\right) \\ &= \left(\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)\left(D_i - \bar{D}\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)\left(Y_i - \bar{Y}\right)\right)\end{aligned}$$

$$\widehat{\beta}_{\text{iv}} = \frac{\sum_{i=1}^{n} Z_i\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} Z_i\left(D_i - \bar{D}\right)} = \frac{\bar{Y}_1 - \bar{Y}}{\bar{D}_1 - \bar{D}}$$

Notice:

$$\bar{Y}_1 - \bar{Y} = \bar{Y}_1 - \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \bar{Y}_1 + \frac{1}{n} \sum_{i=1}^{n} (1 - Z_i) \bar{Y}_0 \right) = (1 - \bar{Z}) \left(\bar{Y}_1 - \bar{Y}_0 \right)$$

and similarly

$$\bar{D}_1 - \bar{D} = (1 - \bar{Z}) \left(\bar{D}_1 - \bar{D}_0 \right)$$

and hence

$$\widehat{\beta}_{\text{iv}} = \frac{(1 - \bar{Z}) \left(\bar{Y}_1 - \bar{Y}_0 \right)}{(1 - \bar{Z}) \left(\bar{D}_1 - \bar{D}_0 \right)} = \widehat{\beta}$$

●Heterogeneous treatment effect 会怎样?

$$Y_i^1 - Y_i^0 = \underbrace{E(Y_i^1 - Y_i^0)}_{\equiv\beta} + \underbrace{(Y_i^1 - Y_i^0 - E(Y_i^1 - Y_i^0))}_{\equiv U_i}$$

$$\begin{aligned} Y &= Y^0 + (Y_i^1 - Y_i^0)D \\ &= Y^0 + (\beta + U)D \\ &= \beta D + (Y^0 + UD) \end{aligned}$$

–一般来说，$Cov(Z, UD) \neq 0$，因为 $Cov(Z, D) \neq 0$, 除非 U 均值独　于 (Z, D)。

–但这一假设不合理，因为 U 是 treatment effect 的一部分。

–因此 β 的 IV 估计量不是 $E(Y_i^1 - Y_i^0)$ 的一致估计。

•那么，IV 估计到底估计的是什么?

定义 potential treatment

Storage type	Bytes
Compliers (C)	$D = 1$ when $z = 1$ and $D = 0$ when $z = 0$
Defiers (D)	$D = 0$ when $z = 1$ and $D = 1$ when $z = 0$
Always-takers (A)	$D = 1$ regardless of the value of z
Never-takers (N)	$D = 0$ regardless of the value of z

Imbens and Angrist (1994) 证明，若

1.$P(D = 1|Z = 1) \neq P(D = 1|Z = 0)$: 存在 compliers

2.$D_i^1 \geq D_i^0 \ \forall i$: 不存在 defiers

3.$(Y^0, Y^1, D^0, D^1) \perp Z$

则

$$
\begin{aligned}
&E(Y|Z=1)-E(Y|Z=0)\\
=&E(DY^1+(1-D)Y^0|Z=1)-E(DY^1+(1-D)Y^0|Z=0)\\
=&E(D^1Y^1+(1-D^1)Y^0|Z=1)-E(D^0Y^1+(1-D^0)Y^0|Z=0)\\
=&E(D^1Y^1+(1-D^1)Y^0)-E(D^0Y^1+(1-D^0)Y^0)\\
=&E((D^1-D^0)(Y^1-Y^0))\\
=&E(Y^1-Y^0|D^1-D^0=1)P(D^1-D^0=1)\\
=&E(Y^1-Y^0|\text{compliers})P(\text{compliers})
\end{aligned}
$$

因为

$$D^1-D^0=1\Longleftrightarrow D^1=1,D^0=0(\text{compliers})$$

$$
\begin{aligned}
&E(D|Z=1)-E(D|Z=0)\\
=&P(D=1|Z=1)-P(D=1|Z=0)\\
=&P(\text{always-takers or compliers})-P(\text{always-takers})\\
=&P(\text{compliers})
\end{aligned}
$$

因此

$$E(Y^1 - Y^0|\text{compliers}) = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} = \hat{\beta}^{iv}$$

Intuition of LATE

Observable:

	$D=1$	$D=0$
$Z=1$	n_{11}	n_{10}
$Z=0$	n_{01}	n_{00}

Unobservable (假定不存在 defiers):

		$Z=0$	
		$D=0$	$D=1$
$Z=1$	$D=0$	Never-taker ($n_n^o,\ n_n^u$)	Defier
	$D=1$	Complier ($n_c^t,\ n_c^c$)	Always-taker ($n_a^o,\ n_a^u$)

	$D=1$	$D=0$
$Z=1$	Complier (n_c^t) Always-taker (n_a^u)	Never-taker (n_n^o)
$Z=0$	Always-taker (n_a^o)	Complier (n_c^c) Never-taker (n_n^u)

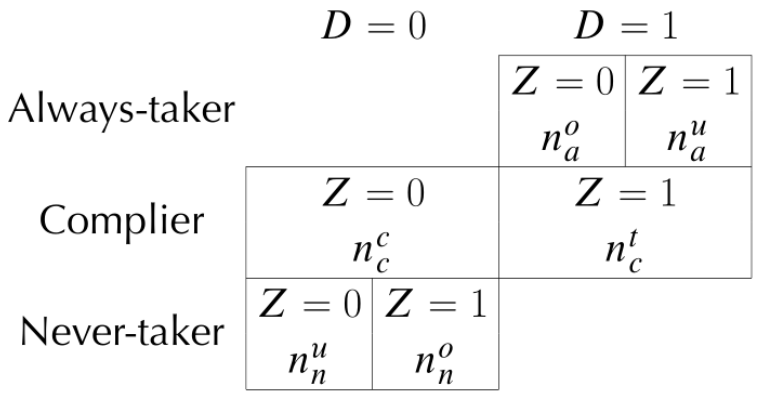
因为 Z 是随机的，因此 $Z=0/1$ subset 中的　例代表总体　例。

$$n_a = \frac{n_{01}}{n_{01}+n_{00}}, n_n = \frac{n_{10}}{n_{11}+n_{10}}$$

$$n_c = 1 - n_a - n_n = \frac{n_{11}n_{00} - n_{10}n_{01}}{(n_{11}+n_{10})(n_{01}+n_{00})}$$

$$n_a^u = \frac{n_{01}}{n_{01}+n_{00}} \cdot (n_{11}+n_{10}), n_n^u = \frac{n_{10}}{n_{11}+n_{10}} \cdot (n_{01}+n_{00})$$

$$n_c^t = n_{11} - n_a^u = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{01}+n_{00}}$$
$$n_c^c = n_{00} - n_n^u = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}+n_{10}}$$



- LATE 是关于 compliers: $n_c^c + n_c^t$.
- ATT 是关于 always-takers 和 a subset of compliers: $n_a^o + n_a^u + n_c^t$.
- 若 n_a^o 较小，则认为不存在 always-takers, 此时 ATT 是关于 n_c^t, 又因为 Z random assignment, 因此 LATE 等于 ATT.
- First stage 恰是 complier 的比例，全部为 complier 时，first-stage=1.

- 示例 Adams et al. (2009, Journal of Empirical Finance).
  - Y : 公司绩效 (Tobin’s Q and ROA)
  - D: CEO 是否为公司创始人
  - ”Eligibility” IV: 创始人死亡比例；创始人数量
  - Method: Probit and ILS

$$E(Y|Z=1,D=1) = \frac{n_c}{n_c+n_a} E(Y^1|\text{complier}) + \frac{n_a}{n_c+n_a} E(Y^1|\text{always-taker})$$

$$E(Y|Z=0,D=1) = E(Y^1|\text{always-taker})$$

$$E(Y|Z=0,D=0) = \frac{n_c}{n_c+n_n} E(Y^0|\text{complier}) + \frac{n_n}{n_c+n_n} E(Y^0|\text{never-taker})$$

$$E(Y|Z=1,D=0) = E(Y^0|\text{never-taker})$$

We can compare

$$E(Y^1|\text{complier}) \text{vs. } E(Y^1|\text{always-taker})$$

and

$$E(Y^0|\text{complier}) \text{vs. } E(Y^0|\text{never-taker})$$

If there is little difference, it is plausible that the average effect for compliers is indicative of average effects for other compliance types.

2.6 一些构造 IV 的“套路”

2.6.1 地理等不随时间变化的变量

LONG LIVE KEJU! THE PERSISTENT EFFECTS OF CHINA’S CIVIL EXAMINATION SYSTEM，EJ2020

中国历史上的科举制度（keju）对当代人力资本成果的持久影响。历史上在科举考试中表现优异的地区（以明清时期进士（jinshi）的密度为代表）与今天当地人口的平均受教育年限有显著的正相关关系。每增加一个进士（每 10,000 人）与 2010 年平均受教育年限增加 8.5% 有关。

工具变量：每个州府到最近的松树和竹子栖息地的河流距离。

如果一个州府在科举考试中表现更好，这可能与其能够接触到书籍和印刷资源的能力密切相关；

- 只有 19 个印刷中心分布在中国的 278 个州府中，这些中心在明代和清代期间出版的书籍占到了总出版书籍的 80%。因此，一个州府能够获得印刷书籍的难易程度，可能与其科举考试的成功有关联。

而主要印刷中心位于松树和竹子栖息地附近，且印刷所需的原料主要通过主要可航行的河流运输。因此，一个州府到最近的竹子和松树栖息地的河流距离，可以作为一个合理的工具变量来代理进士密度。

最小生成树：

The telegraph and modern banking development, 1881–1936，JFE2021

19 世纪末，电报技术被引入中国, 探讨了信息技术在银行发展中的重要性。构建了一个包含 1881 年至 1936 年间 287 个府级单位的电报局和银行分布的数据集。使用 OLS（普通最小二乘法）和 IV（工具变量法）方法来估计电报对银行发展的影响。主要发现：电报显著扩大了银行分支机构的数量和地理范围。

工具变量：早期清朝政府以军事目的建设电报系统干线。一个地区距离这个假设的军事电报干线 (HMTN）的远近，可以作为一个与当地银行发展无直接关联的工具变量。同时，作者通过连接清朝时期的 21 个军事中心，并沿着最小建设成本的路径来构建假设的军事电报干线（HMTN)。这个假设的干线考虑了地形和水文特征对建设成本的影响。

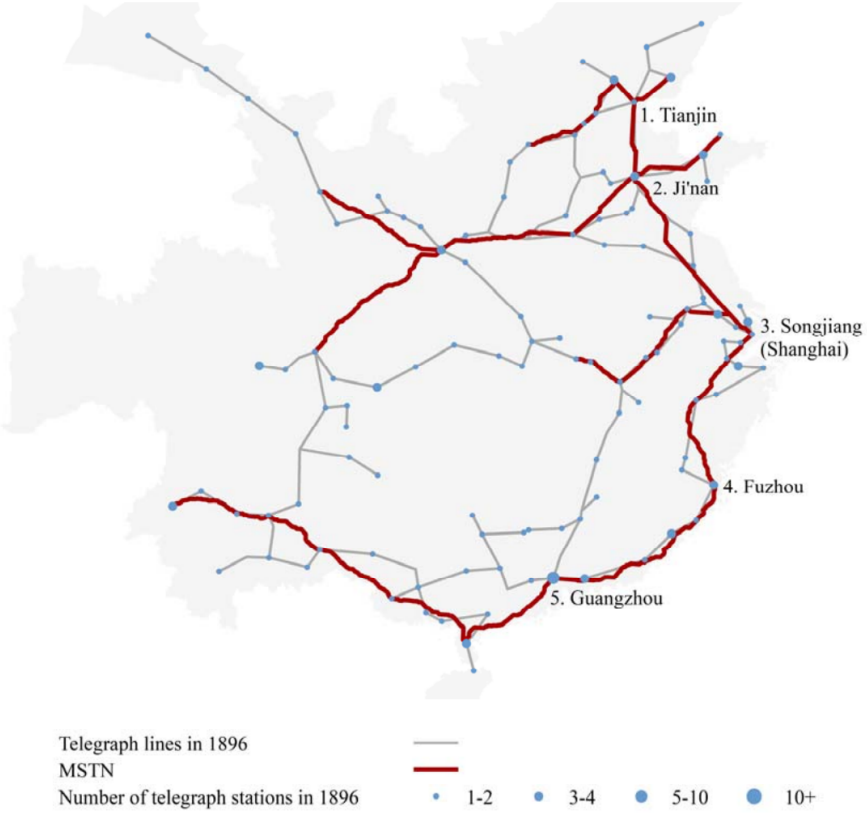


Fig. A.2. Alternative hypothetical telegraph trunk: minimum spanning tree network.
The minimum spanning tree network (MSTN) is constructed based on the minimum construction cost of the entire telegraph trunk system. We first calculate the least cost bilateral path between each pair of military centers, and then calculate the MSTN of the entire trunk based on the greedy algorithm method (Kruskal, 1956). The nodes (military centers) are the same as those in Fig. 2. The map covers 287 sample prefectures in China proper.

清朝时期假想的军事电报干线网络（Hypothetical Military Telegraph Network, HMTN）和最小生成树网络（Minimum Spanning Tree Network, MSTN）的构建过程

HMTN 的构建：

- HMTN 是基于连接清朝 21 个军事中心的电报线路的最小建设成本构建的。
- 为了计算电报线路的最小建设成本，研究者使用了来自 Shuttle Radar Topography Mission (SRTM) 的地形（坡度和起伏）和水文信息。
- 利用 ArcGIS 软件，数据被组织在 1km × 1km 的网格单元级别，每个网格单元的建设成本基于中国道路建设成本函数计算得出。
- 成本距离算法被用来计算两个军事中心之间的最小成本路径，生成累积成本栅格和方向反向链接栅格。
- 对于给定的军事中心，将累积成本栅格、方向反向链接栅格和目标节点（其他 20 个军事中心）输入到成本路径算法中，得到 20 条最小成本路径。
- 通过重复此过程，为所有军事中心生成共 210 条最小成本路径（21 × 20 / 2），然后选择与 1896 年实际电报线路最接近的路径作为 HMTN。

MSTN 的构建：

- MSTN 假设清朝政府试图最小化整个电报网络的建设成本，使用最小生成树方法生成全国最经济的电报网络。
- 首先，计算连接任意两个军事中心的每条假设电报线的最小建设成本路径，方法与 HMTN 相同，但提取所有可能的双边最小成本路径的总建设成本。
- 使用贪心算法（Kruskal, 1956）生成连接 21 个军事中心（节点）的整体最小成本网络。该方法从一个节点开始，计算从该节点到所有其他节点的最小成本路径，并为所有节点重复此过程。
- 基于局部最优解计算全局最优解。在本研究的背景下，天津-松江线和松江-广州线作为最早的电报线路被视为给定的，即不由生成树重塑。
- 在计算 MSTN 时，保持这两条沿海线路的最小建设成本路径，同时计算其他线路的 MSTN。

2.6.2 基于预测的构造方法

Lipscomb, Molly, A. Mushfiq Mobarak, and Tania Barham. 2013. "Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil." American Economic Journal: Applied Economics, 5 (2): 200-231.

文章探讨了 1960 年至 2000 年期间，巴西电气化对发展的影响。研究通过模拟巴西的电力网络发展，分析了基于地理成本考虑的基础设施投资对发展的影响。使用固定效应工具变量（IV）方法来检验 1960 年至 2000 年期间电气化对县级发展的影响。

主要发现：

- 电气化能够提高劳动生产率、就业和教育投资，
- 电气化对发展的影响在 OLS 中被低估
- 电气化对住房价值和联合国人类发展指数有显著的正面一下。

内生性问题：电气化投资决策可能与地区发展水平之间的相互关联。

- 政府目标和电气化投资：政府可能会基于政治或社会目标，将电气化项目投资于特定的地区，尤其是那些经济落后或需要支持的地区。这种选择性支持可能和地区发展相关联，电气化看起来对发展有正面影响，但可能只是政府想要重点发展这些地方。
- 电气化投资的需求侧因素应能够促进社会发展：电气化投资可能和人口密度、经济活动水平和资源可用性等因素相关。
- 逆向因果：地区发展可能促进电气化。

工具变量的思路：通过模拟基于地理成本因素的电力网络扩张来预测电气化，从而尝试区分电气化对发展的单向因果效应。这种方法假设地理成本因素是外生的，不受地区发展水平的影响，因此可以作为有效的工具变量。

构建过程： Our instrument (Z) is a prediction on electricity availability at each grid point in each decade, based on a model that simulates the evolution of generation plants and transmission lines in a way that minimizes construction cost. The model takes as inputs data on the geographic characteristics of each location and the national bud- get for each decade, and produces predictions for whether each of the 33,342 evenly spaced grid points has electricity access in each of the 5 time periods of data between 1960 and 2000. The geographic data are matched to existing hydropower dam data by 12 km buffer zones around each grid point.

如果基础设施投资仅基于地理成本考虑，巴西的电力网络将如何演变：

1. 确定每个十年的国家预算，决定将要建设的水电站数量。
2. 根据地理位置的地形特征（如水流、河流梯度和亚马逊地区的位置）对所有潜在的大坝建设点进行排名，以确定建设的优先顺序。
   - 在第一个十年中，地形适宜性排名最高的网格点将首先接收水坝，直到预算（在第一步中计算）用完。在下一个十年中，排名次高的未预测到有电力的网格点将接收水坝，直到达到该十年的国家预算。
3. 使用成本最小化算法来确定输电线路的布局，连接新建的水电站和现有的电网。

回归模型：

$$Y_{ct} = \alpha_c^1 + \gamma_t^1 + \beta \hat{E}_{c,(t-1)} + \epsilon_{ct},$$

where $\hat{E}$ is instrumented electricity provision, predicted on the basis of our model forecasting the expansion of electricity in the first stage:

$$E_{c,(t-1)} = \alpha_c^2 + \gamma_t^1 + \theta Z_{c,(t-1)} + \eta_{ct}.$$

E_{ct} 是时期 t 时县 c 中电网点的电气化比例。Z_{ct} 是一个预测值，表示根据作者的模型，县内的电网点预计在给定十年内将被电气化的比例。

电力供应滞后十年，因为配电网络的发展可能需要在水电站建设完成后的几年内完成。

作者论述工具变量的有效性：

In order for the instrument to be valid in a time and county fixed effects IV model, the demand side (people or firms) must not move independently over time along the same spatial lines as the forecasted placement of the electricity network within a county——from the lowest cost locations (robust water flow with a steep river gradient) in the early decades to slightly more expensive (flatter and less water-rich) locations in later years. We build confidence in the validity of the instrument by presenting evidence that the expansion of settlements followed a different spatial pattern than modeled electricity provision, and that the results remain robust to limiting the source of identification of the instrument. At the extreme, we rely solely on the nonlinearities and discontinuities built in to the forecasting model through decade budgets, and exclude the direct effects of the geographic variables.

It is possible that due to water scarcity- the population moved to new counties based on water availability during our period of analysis, leading settlement of counties in Brazil to independently follow the same pattern as electricity grid expansion. While this seems unlikely since Brazil has 13 percent of the world's freshwater resources, and all inhabited land is well covered by a dense network of small rivers and groundwater (Lipscomb and Mobarak 2011), we use census population data for 1910 onward to examine whether Brazil's counties were settled before the start of the analysis in 1960 (see online Appendix Figure A7). At a low-population density cutoff of 0.5/sq-km,15 all counties in Brazil were already settled by the starting period of our analysis, except for some counties in the Amazon. Even at a high-population density cutoff of 5/sq-km, only 23 out of 2,184 counties are settled for the first time during the analysis period of 1960-2000. Water scarcity is therefore unlikely to drive population movements and settlement patterns directly during the period of analysis.

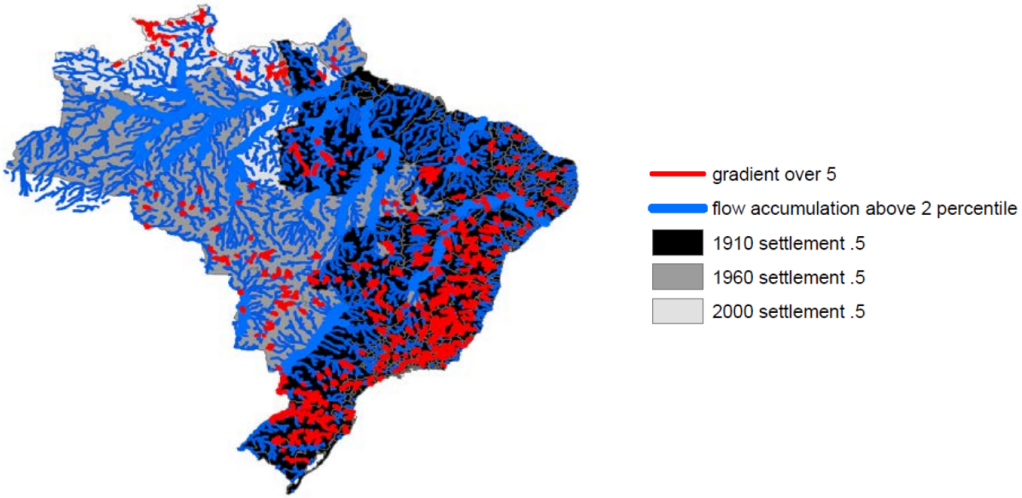


Figure A7: Evolution of population settlements across Brazil
Source: Authors' calculations

Another way to directly examine the question of whether people and/or firms move independently over time in the same spatial pattern as our forecast of electricity grid expansion, is to create a rank ordering of locations predicted to have the highest population density or highest GDP, using the same method we use to predict the most suitable locations for dam construction. To do so, we regress population density and GDP per capita on the same geographic characteristics as those used in the electricity forecasting model: water flow, river gradient, and Amazon. We then rank-order the points predicted to have the highest population and GDP by those regressions. We examine the Spearman rank order correlation between the suitability rank for hydropower generation and the suitability rank for population density in Table 4A and the correlation between hydropower suitability and GDP in Table 4B. We find that for each one of the five major regions of Brazil, the rank order correlation for population and hydropower suitability is low, and varies between -0.03 to +0.06. This is a conservative test of our identification assumptions, since this region fixed effects analysis is much less stringent than the county fixed effects we employ in all our regressions. The rank order correlations for GDP per capita rank and hydropower suitability rank for the typical region-decade is also very close to zero, and ranges from —0.06 to +0.1().

Another way to directly examine the validity of the instrument is to examine whether the placement of power plants simulated by the forecasting model can be predicted by development indicators in earlier years. Results in Table 5 show that the point estimates on decade-lagged values of development indicators that serve as our main outcome variables of interest (housing values and county HDI) are close to zero and statistically insignificant. This suggests that at least lagged development indicators do not predict the spatial allocation of hydropower dams and transmission lines, and provides some confidence that the model's simulation of cost-minimizing electrification is orthogonal to demand side factors.

TABLE 4A—SPEARMAN CORRELATIONS—HYDROPOWER SUITABILITY AND POPULATION DENSITY

Region	Year				
	1960s	1970s	1980s	1990s	2000s
Amazon	+0.0369	+0.0368	+0.0354	+0.0594	+0.0432
North East (including Bahia, Ceara, etc.)	−0.0298	−0.0290	−0.0341	−0.0321	−0.0349
Central West (including Pantanal)	+0.0217	+0.0172	+0.0141	+0.0442	+0.0375
South East (including Minas Gerais, Rio de Janeiro, Sao Paulo)	+0.0070	+0.0124	−0.0019	−0.0247	−0.0318
South (including Parana, Rio Grande do Sul)	+0.0486	+0.0447	+0.0506	+0.0532	+0.0631

Note: Each cell presents the Spearman rank order correlation between the suitability rank for hydropower generation and the rank for population density, by region and decade.

TABLE 4B—SPEARMAN CORRELATIONS—HYDROPOWER SUITABILITY AND GDP

Region	Year				
	1960s	1970s	1980s	1990s	2000s
Amazon	+0.0714	+0.0082	+0.0557	+0.0700	+0.0679
North East (including Bahia, Ceara, etc.)	+0.0098	+0.0127	+0.0259	+0.0968	+0.0689
Central West (including Pantanal)	+0.0067	−0.0155	+0.0141	−0.0078	−0.0138
South East (including Minas Gerais, Rio de Janeiro, Sao Paulo)	−0.0557	−0.0631	−0.0554	−0.0826	−0.0603
South (including Parana, Rio Grande do Sul)	+0.0031	−0.0043	−0.0237	−0.0104	+0.0238

Note: Each cell presents the Spearman rank order correlation between the suitability rank for hydropower generation and the rank for GDP, by region and decade.

TABLE 5—ROBUSTNESS CHECK FOR REVERSE CAUSALITY

	Instrument for electricity infrastructure	
Lagged housing value	0.000	
	(0.00)	
Lagged county HDI		−0.045
		(0.04)
R^2	0.984	0.984
Observations	6,549	6,549

Note: Standard errors clustered by county in parentheses. All regressions have county size weights and year dummies.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Lu, Fangwen, Weizeng Sun, and Jianfeng Wu. 2023. "Special Economic Zones and Human Capital Investment: 30 Years of Evidence from China." American Economic Journal: Economic Policy, 15 (3): 35-64.

经济特区提高了当地的高中入学率。以技术为导向的区域鼓励教育，而以出口导向的区域抑制教育。有以下几个机制：

1. 收入：经济特区提供了更好的就业机会，增加家庭收入，家长有能力给子女进行人力资本投资。
2. 就业机会渠道：如果经济特区提供的是低技能工作，会增加上学的机会成本，降低高中入学率。如果经济特区提供高技能工作机会，将增加高中教育的吸引力。

Following the forecasting method proposed by Lipscomb, Mobarak, and Barham (2013), we predicted the likelihood of the placement of SEZs based on predetermined local geographic attributes over each of the five-year periods (see Appendix Table 1).

To further alleviate the endogeneity concern, we adopted a jackknife method developed by Jackson, Johnson, and Persico (2016), and estimated the probability of having an SEZ by excluding all data from its host province. This "leave-out" estimate partly alleviates the concern about a weak instrument problem as it is less likely to violate the exclusion restriction criterion. We ranked the counties within each province based on the estimated probability, as was done by Duflo and Pande (2007) and Lipscomb, Mobarak, and Barham (2013), and generated a 0/1 variable on the predicted $\widehat{SEZ_{it}}$, for the top N_j counties if the province launched N_j SEZs during that period. The predicted $\widehat{SEZ}_{it}$ serves as an instrumental variable for the actual SEZ_{it}.

THE EFFECTS OF SCHOOL SPENDING ON EDUCATIONAL AND ECONOMIC OUTCOMES: EVIDENCE FROM SCHOOL FINANCE REFORMS

研究利用了美国 K-12 公立学校在 20 世纪 70 年代初至 80 年代中期经历的一系列学校财政改革作为自然实验，在 1970 年代以前，地方教育支出是通过地方财产税筹集的。所以富裕地区的学生支出更多。为了减少高收入地区和贫困地区人均教育支出的大幅差异，州法院在 1971 年-2010 年推翻了 28 个州的学校财政拨款系统。研究者通过将学校支出和学校财政改革数据与全国范围内代表性的儿童数据（出生于 1955 年至 1985 年，追踪至 2011 年，有学生成年以后的情况）相链接，使用事件研究和工具变量模型来分析学校支出对学生长期成果的影响。

研究发现，每增加 10% 的人均学校支出，可以使学生完成的教育年数增加 0.27 年，工资提高 7.25%，并减少 3.67 个百分点的成年贫困发生率。这些效应对于低收入家庭的儿童更为显著。此外，学校支出的增加与学校质量的显著改善相关，包括师生比例的降低、教师工资的增加和学年长度的延长。

内生性问题：

- 同时性偏误（Simultaneity Bias）：学校支出的变化可能与地区内部的其他因素相关联，这些因素也可能影响学生的成绩和成人后的经济成果。例如，一个地区的经济增长可能会导致更高的税收收入和学校支出的增加，同时也可能提高当地居民的收入和教育水平，从而影响学生成果。
- 遗漏变量偏误（Omitted Variable Bias）：如果模型没有控制所有影响学生成果的重要变量，那么估计的支出对成果的影响可能会受到遗漏变量的影响。例如，家庭背景、社区环境和教育质量等因素都可能对学生成果产生重要影响，如果这些因素没有被充分考虑，可能会导致对学校支出影响的估计存在偏误。
- 选择偏误（Selection Bias）：学校支出的增加可能是由于某些特定类型的地区或学校面临的特定需求或政

策变化引起的。如果这些地区或学校在选择增加支出时存在某种系统性差异，那么这些差异可能会影响学生成果，从而导致内生性问题。

Our goal is to identify the causal effect of per pupil public school spending during childhood on adult outcomes. Because the correlation between per pupil spending in an area and the adult outcomes of students who attended those schools is likely con- founded by other factors (due to residential segregation, Tiebout sorting, compensatory spending increases, etc.), we search for ex- ogenous variation in per pupil spending. To this aim, we use only variation in school spending during childhood that can be attrib- uted to the passage of court-ordered SFRs.

工具变量：研究者选择了由法院命令的学校财政改革（SFRs）作为工具变量，因为这些改革在不同时间和不同地区实施，且主要目的是改变学校的资金结构，从而提供了学校支出变化的外生性变化。

As pointed out in Hoxby (2001), the effect of a SFR on school spend- ing depends on (i) the type of school funding formula introduced by the reform and, (ii) how the funding formula interacts with the specific characteristics of a district. To capture some of this com- plexity, we follow the typology outlined in Jackson, Johnson, and Persico (2014) and categorize reforms into five main types. Foundation plans guarantee a base level of per pupil school spend- ing and are designed to increase per pupil spending for the lowest-spending districts. Spending limits prohibit per pupil spending levels above some predetermined amount. Such plans tend to reduce spending for high spending and more affluent districts and may reduce spending in the long run for all districts.

In predicting adult outcomes, our endogenous treatment variable is the natural log of average school spending over the previous 12 years. This measures average school spending across all school-age years (5 to 17) for expected high school graduates that year. Having predicted the spending change a district will experience with the passage of a court order, we now show how the interaction between this district-specific prediction, Spend $_d$, and the timing of court-ordered reforms isolates plausibly exogenous variation in school spending that is unrelated to potentially confounding district-level determinants of school spending. We estimate equation [3] where $\ln\left(\bar{\Phi}_{\text{dst}}\right)$ is the natural log of average school spending over the previous 12 years, Spend d_d is our scalar district-specific prediction of the reform-induced spending change, I_y^{court} are flexible event time indicators, θ_d is a district fixed effect, θ_t is a year fixed effect, and ε_{dt} is random error.

$$\ln\left(\bar{\Phi}_{dst}\right) = \alpha + \left(\text{ Spend }_d \sum I_y^{\text{court}}\right) \pi_{\text{spend },y}^{\text{court}} + Z_{dt} + \theta_d + \theta_t + \varepsilon_{dt}$$

The coefficients $\pi^{\text{cont}}_{\text{Spend},\,y}$ map out the spending change associated with a court-mandated reform for a district that is predicted (based on similar districts in other states) to double school spending by years 3 through 8 post reform. To show the changes in spending both by duration of exposure and by predicted treatment intensity, Figure 2 plots the estimated flexible event study coefficients for a 5 percent predicted change, a 10 percent change, and a 20 percent predicted change. If our instrument has identified exogenous variation, districts that saw larger versus smaller predicted spending increases due to reforms should be on very similar trajectories prior to reforms. Also, if the instrument is valid, after reforms districts with larger versus smaller predicted spending increases due to reforms should experience larger versus smaller actual spending increases.

2.6.3 Bartik IV

Shift-Share IV (SSIV)，是由地区行业层面的份额和国家行业层面增长率的交互项构成的工具变量。基于不同暴露情况下的地区或行业在遭受相同外生冲击时的反应进行设计。

问题设定：假设我们想要估计某个地区工资增长率（因变量 y_{lt} ）对于就业增长率（自变量 x_{lt} ）的因果关系。这里，下标 l 表示地区，t 表示时间。

内生性问题: 在实际中，就业增长率可能受到工资增长率的影响（反向因果），或者存在遗漏变量，使得普通最小二乘 (OLS) 估计产生偏误。

Bartik IV 构建:

- 就业增长率的分解: 首先，将就业增长率 x_{lt} 表示为各行业就业份额 z_{ltk} 与相应行业就业增长率 g_{ltk} 的内积，即 $x_{lt} = Z_{lt}G_{lt}$ ，其中 Z_{lt} 是一个 k 维向量，包含地区 l 在时间 t 的各行业就业份额，G_{lt} 是相应的行业就业增长率向量。
- 行业增长率的进一步分解: 将 g_{ltk} 分解为国家层面的行业增长率 g'_{tk} 和地区特有的行业增长率 g''_{ltk} ，即 $g_{ltk} = g'_{tk} + g''_{ltk}$ 。
- Bartik IV 的构造: 构造的 Bartik IV 是基于初始年份（第 0 期）各地区行业的就业份额 Z_{l0} 与国家层面行业的就业增长率 G_t 的内积，即 $B_{lt} = Z_{l0}G_t$ 。

2.6.4 其他

非常无脑的做法之一：用 X_{t-1} 作为 X_t 的工具变量。

非常无脑的做法之二：同省份/同行业/同省份同行业其他公司的 X 的平均值。

以上两种做法在前几年用的比较多，现在已经不推荐使用了。

对于第一种做法，原因是，很多时候变量有自相关性，$t-1$ 期和 t 期之间是相关的，所以 $E(X_{t-1}\varepsilon_t)=0$ 不成立。

对于第二种做法，原因是企业之间也可能有明显的模仿效应，同行业同地区的不同企业的行为也是相关的。所以 $E(X_{-i}\varepsilon_i)=0$ 也是不成立的。