lianxh.cn

借助 AI 应对内生性问题

连玉君 (中山大学)
arlionn@163.com

- Empirical Research with AI
- AI-for-Endog GitHub

Outline

1. AI 时代我们需要做哪些改变
2. 内生性问题的识别框架
3. 解决方案工具箱
    - 控制变量法、工具变量法等
4. 案例：政府引导基金介入与供应链稳定性
5. 总结

AI 时代更需要深厚的方法论基础

AI的作用

- ✅ 放大你的专业能力
- ✅ 加速文献检索和代码生成
- ✅ 提供多角度思考框架

AI不能替代

- ❌ 你对研究问题的深刻理解
- ❌ 你对方法论假设的判断
- ❌ 你对结果合理性的评估

使用 AI 工具的四大基石

1. 研究问题与数据背景

- 充分了解研究问题 | 熟悉数据特征和限制

2. 因果推断方法论

- 熟悉常用方法 | 知晓假设条件 | 理解适用/不适用场景

3. 编程技能

- Stata / R / Python → 能够整合 AI 生成的代码 | 调试+优化
- 建议：VScode + Anaconda + Jupyter Notebook

4. 经典论文

- 理解研究设计细节 | 精读+复现 → 研究设计理念 (本地库)
- 判别能力 → 品味 → 直觉

迷糊 → 不懂方法论直接问 AI → 更迷糊

问题： "我想研究政府补贴对企业创新的影响，帮我设计研究方案"

AI回复（过于笼统）：

- 建议使用 DID 或 PSM
- 列举了一堆控制变量
- 给出了标准代码模板

问题所在：

- 没有识别具体的内生性来源
- 没有评估方法假设是否满足
- 没有考虑数据的实际情况

后果： 研究设计有缺陷，审稿人一眼看出问题

设定合理预期：AI 能做什么？

- 快速梳理文献中的方法论
- 生成工具变量候选列表
- 提供代码框架和调试建议
- 设计稳健性检验方案
- 回应审稿意见的思路

设定合理预期：AI 无法做什么？

- 判断内生性的严重程度
- 评估工具变量的有效性
- 选择最合适的识别策略
- 解释经济学机制
- 做出最终的学术判断

第二部分：内生性问题基础知识

内生性的主要来源

1. 遗漏变量偏误
2. 反向因果
3. 样本选择偏误 / 自选择偏误
4. 测量误差
5. 模型设定偏误

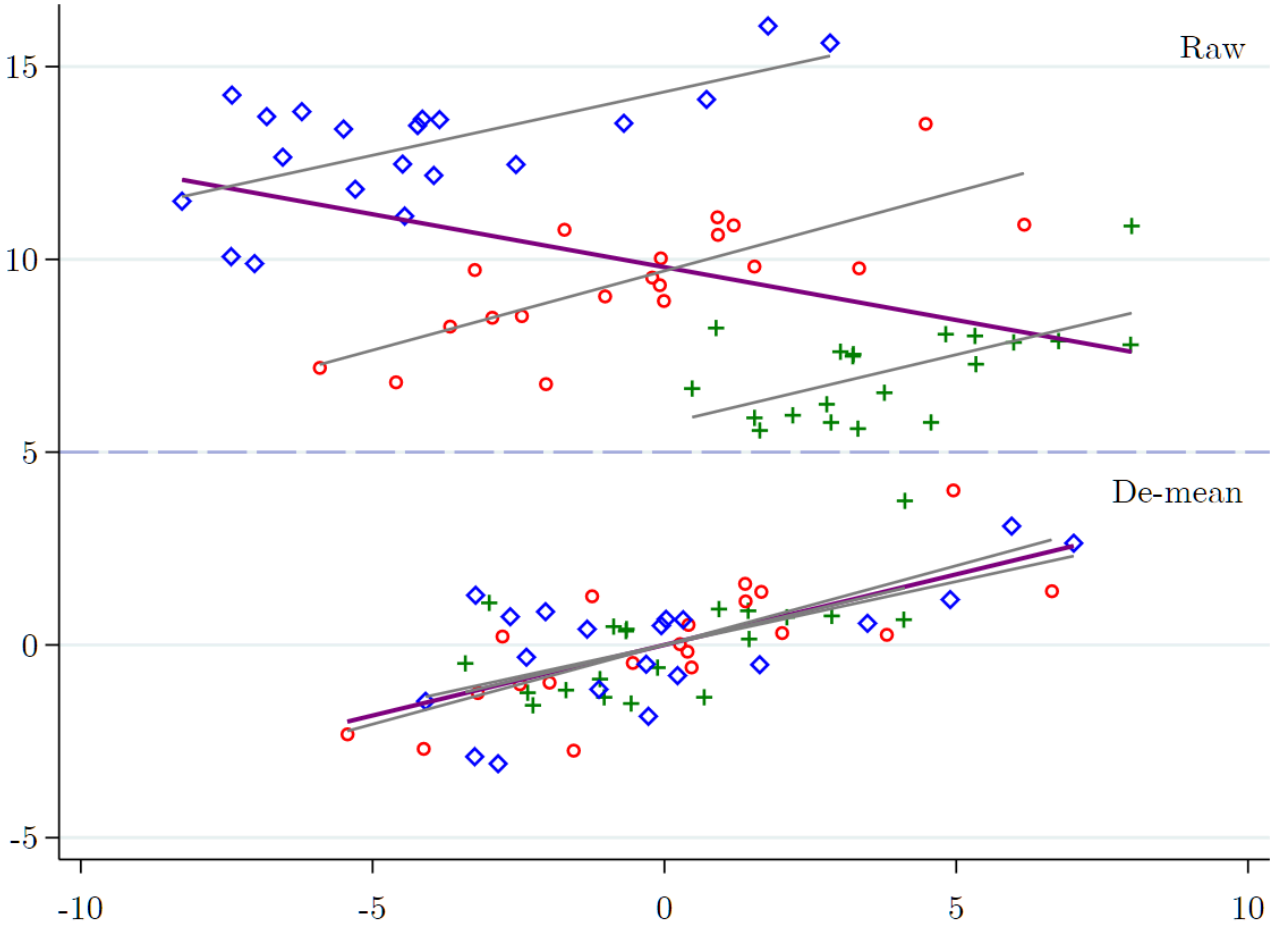
回顾 1：遗漏变量

$$True: \quad y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u_1$$

$$Estimate: \quad y = \alpha + \beta_1 x_1 + u_2$$

?

$$if\ Corr(x_2, x_1) \neq 0,\ then\ Corr(u_2, x_1) \neq 0,\ \text{Endog!}$$



例: 固定效应模型 (FE)

原始数据:

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}$$
$$= \sum_i^N \alpha_i D_i + x_{it}\beta + \varepsilon_{it}$$

组内去心 (De-meaned):

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

扩展：TWFE / HDFE

TWFE

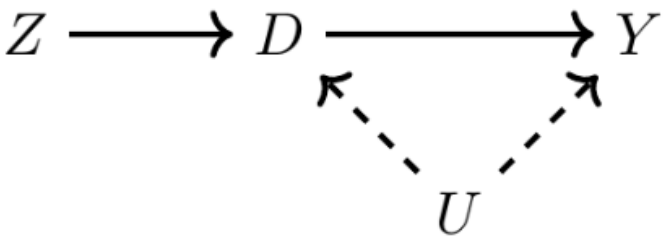
$$Y_{it} = c_0 + \alpha_i + \lambda_t + \mathbf{x}'_{it}\beta + v_{it}$$

HDFE

$$Y_{ijt} = c_0 + \alpha_i + \lambda_j + \delta_t + \mathbf{x}'_{ijt}\beta + v_{ijt}$$

Interactive FE

$$Y_{ijt} = c_0 + \alpha_i + \lambda_j + \delta_t + \phi_{it} + \psi_{jt} + \mathbf{x}'_{ijt}\beta + v_{ijt}$$



回顾 2：理解 IV 的原理

结构方程: $(1)\ y_i = \alpha + \beta x_i + \varepsilon_i$

第一阶段方程: $(2)\ x_i = \pi_0 + \pi_1 z_i + v_i$

将 (2) 带入 (1)，可得：

简约式: (3)

$$\begin{aligned} y_i &= (\alpha_0 + \beta\pi_0) + (\beta\pi_1) z_i + (\beta v_i + \varepsilon_i) \\ &= \gamma_0 + \gamma_1 z_i + (\beta v_i + \varepsilon_i) \end{aligned}$$

因此 (**LATE**, Local Average Treatment Effect)，

$$\beta = \frac{\gamma_1}{\pi_1} = \frac{\widehat{\mathrm{Cov}}(y, z)}{\widehat{\mathrm{Cov}}(x, z)}$$

工具变量需要满足的条件：

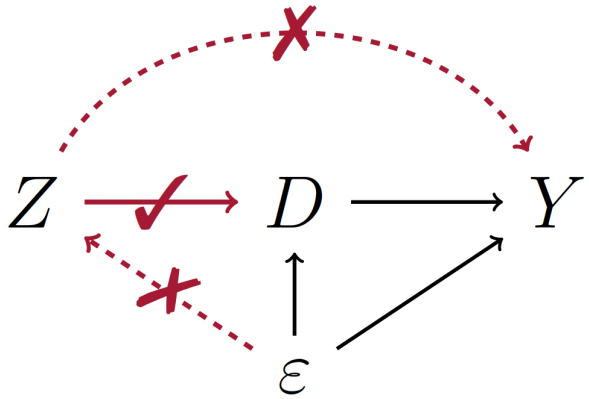
- 相关性：工具变量与内生解释变量相关。

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

- 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?



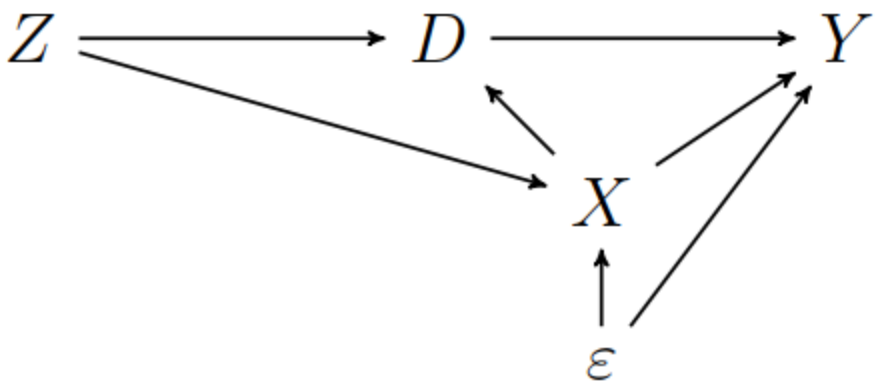
Which one is Right ?

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

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

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，把这个例子加进来。



回顾 3：不同的方法解决不同的内生性问题 (1)

- 控制变量法 (Control Variables)
  - 适用：遗漏变量 (可观测；或不可观测但可被代理)
  - 挑战：Good control variables? (基于因果路径图)
- RDD (Regression Discontinuity Design)
  - 适用：随机断点
  - 挑战：断点附近的外推性 / 局部效应

回顾 3：不同的方法解决不同的内生性问题 (2)

- Heckman 选择模型
  - 适用：样本选择偏误 (y 的非随机缺失)
  - 挑战：排他性限制 / 选择方程设定
- 匹配方法
  - 适用：不存在不可观测的遗漏变量
  - 应对：条件独立性假设 (CIA)

一些建议

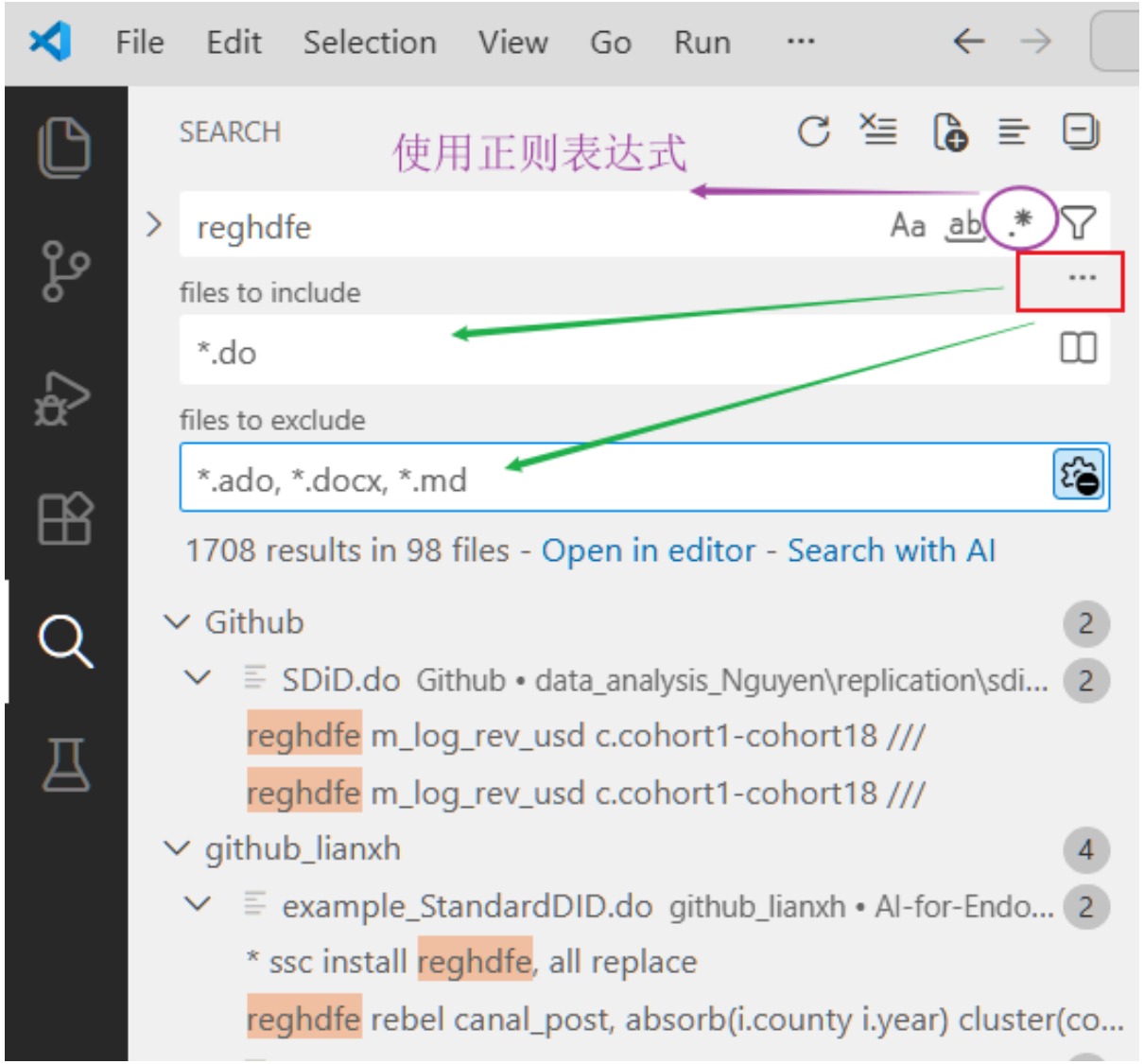
技巧 1：形成自己的知识库 (1)

- 使用 AI 做文献综述：
  - 找到相关的经典论文
  - 总结不同方法的优缺点
  - 或者直接找到几篇经典的综述论文，抽取其中的典型论文
- 获取经典论文的复现数据和代码
- 使用 VScode 的全文搜索能力，快速定位相关代码和数据文件
  - 举例：我收集了约 100 篇顶刊论文的复现代码和数据，放在 `PX_papers1`，`PX_papers2` 等文件夹中，形成了一个自己的代码库

技巧 1：形成自己的知识库 (2) - VScode 全文搜索

使用 VScode 的「在文件夹中搜索」功能，快速定位相关代码片段：

- **建立工作区**：File → Add Folder to Workspace... → 选择代码库所在的文件夹。按此操作，即可将多个文件夹添加到同一个工作区中
- **全文搜索**：Ctrl+Shift+F (Windows) 或 Cmd+Shift+F (Mac)
  - Note: 单击搜索框右下角的 `...` 图标，可以展开更多搜索选项，比如使用通配符或正则表达式



技巧 2：复现和解读期刊论文

- 适用：精读一些识别策略或研究设计比较经典的论文
- 在一个文件夹中放置该论文的 PDF, 复现代码和数据
- 使用 Claude code 或 notebooklm，逐步复现论文中的关键结果
  - 提示词：论文中文精要-识别策略
  - eg：Wang-2024-EE - AI 使用率与绿色创新效
    - Wang, et al. (2024). AI adoption rate and corporate green innovation efficiency. **EE**, 132, 107499. Link, PDF, Google, -cited-, Replication

技巧 3：多篇论文对比形成几种识别方案

- 适用：针对一个研究问题，找到多篇相关论文
- 形成一个文件夹，放置这些论文的 PDF, 复现代码和数据

技巧 4：积累自己的 Prompt 模板

- 连玉君的提示词：https://github.com/lianxhcn/myprompt

案例：政府引导基金介入与供应链稳定性

研究背景：

- 自2015年起，中国各级政府设立大量政府引导基金
- 通过股权投资支持本地企业，促进产业升级和技术创新
- 相比传统补贴，资金规模大、市场化程度高
- "投早、投小、投科技"原则

研究问题：
政府引导基金介入（D）对企业供应链稳定性（Y）的影响?

复杂性：

- 正面效应：资金支持 → 优化库存管理和采购 → 增强稳定性
- 负面效应：行政干预和资源错配 → 加剧不稳定性

与 AI 对话的思路 (1)

1. 说明研究背景和问题
2. 帮我分析 Y 可能受那些因素影响？(识别混杂因素)
3. 帮我分析 D 可能受那些因素影响？(识别内生性来源)
4. 帮我分析 D 和 Y 之间可能存在的内生性问题？(遗漏变量、反向因果等)
5. 详细解读每一种内生性问题的原因
6. 排序：哪些内生性问题最严重？为什么？

与 AI 对话的思路 (2)

7. 逐个讨论：针对每一种内生性问题，帮我设计解决方案
    - 控制变量法 (确保基准方程设定正确)
    - 工具变量法 (要明确说明内生性来源)
    - 双重差分法 (需要满足平行趋势假设)
    - 匹配方法 (条件独立性假设)
    - 其他方法
8. 帮我设计 10 个合理的工具变量 (IV)
9. 帮我评估这些 IV 的有效性 (相关性、外生性、排他性)
10. 推荐：最好的 3 个 IV，并说明理由
11. 其他可能的 IV：Shift-Share IV，Judge IV 等

与 AI 对话

> 具体过程参见 连玉君-2025-11 如何借助 AI 寻找工具变量?

- **ChatGPT 对话 01** - 政府引导基金介入与供应链稳定性