当代 DID 方法：内生性问题

从理论到实践应用

许文立，wlxu@cityu.edu.mo

澳门城市大学金融学院

2025 年 3 月 10 日

因果推断回顾

- 回忆：线性回归方程的斜率系数 $\beta = \frac{\partial Y}{\partial T}$ 测度了处理变量 T 对结果变量 Y 的因果效应
- β 的无偏估计要求“外生性”：

$$cov(T, u) = 0$$

当随机扰动项 u 中包含遗落变量 W，外生性就不满足，“内生性”问题就出现了

- 计量经济学从以下几个方面来尝试解决内生性问题：
  - a 做实验，在实验中，我们可以控制所有其它因素为常数（即 $u = c$），从而使得 $cov(T, c) = 0$
  - b 在 RCT 中，随机配置处理 T（$T = rand$），从而使得 $cov(rand, c) = 0$
  - c 从 u 中把遗落遗漏变量 W “拿”出来，例如，回归调整或者匹配，控制住 W，比较相同 W 下的处理组与控制组结果差异
  - d 准自然实验方法，IV、RD、panel data、SC 和 DID 等

协变量控制方式回顾

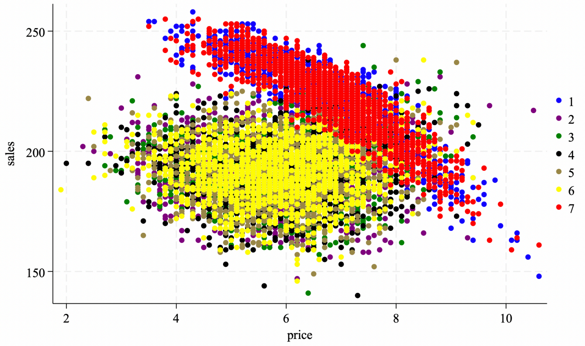
- 条件 PTA 下，最常用的 TWFE 方程包含时变协变量 X_{it}:

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + X_{it}'\beta + v_{it}$$

- 上述时变协变量给 TWFE 带来许多潜在问题
- 应用计量文献通常采用以下几种方式来应对：
  a 控制处理前协变量的值
  b 采用回归调整（RA）方式
  c 采用 IPW 方式
  d 采用双重稳健 DID 估计量
  e 其它一些方式：协变量平衡性检验、外生性检验、imputation DID 估计量、DID-INT 估计量、动态平衡协变量等等（但是注意，需要额外的假设）

Matheus Facure(2022): 冰淇淋销售

- 研究冰淇淋的价格对销售量的影响（测试数据集随机配置了价格，但是训练数据只有观测价格）



doflie:didml.do

Matheus Facure(2022): 冰淇淋销售

- 偏误的一个来源：周末（第一天和七天）的价格更高

Matheus Facure(2022): 冰淇淋销售

- 偏误的一个来源：周末（第一天和七天）的价格更高
- 也有其他的偏误来源：气温（temp）和成本（cost）

	temp	weekday	cost	price	sales
0	17.3	6	1.5	5.6	173
1	25.4	3	0.3	4.9	196
2	23.3	5	1.5	7.6	207
3	26.9	1	0.3	5.3	241
4	20.2	1	1.0	7.2	227

Matheus Facure(2022): 冰淇淋销售

- 偏误的一个来源：周末（第一天和七天）的价格更高
- 也有其他的偏误来源：气温（temp）和成本（cost）

	temp	weekday	cost	price	sales
0	17.3	6	1.5	5.6	173
1	25.4	3	0.3	4.9	196
2	23.3	5	1.5	7.6	207
3	26.9	1	0.3	5.3	241
4	20.2	1	1.0	7.2	227

- 因此，想要做因果推断，就必降低这些偏误

Matheus Facure(2022): 冰淇淋销售

- 控制变量法：利用线性回归模型来控制这些混淆因子，从而估计价格对销售量的效应：

$$Sales_i = \alpha + \tau price_i + \beta_1 temp_i + \beta_2 cost_i + \beta_3 Weekday_i + e_i$$

τ 是我们感兴趣的处理效应

Matheus Facure(2022): 冰淇淋销售

- 控制变量法：利用线性回归模型来控制这些混淆因子，从而估计价格对销售量的效应：

$$Sales_i = \alpha + \tau price_i + \beta_1 temp_i + \beta_2 cost_i + \beta_3 Weekday_i + e_i$$

τ 是我们感兴趣的处理效应

- 其它参数不是我们感兴趣的，但是我们也要关心它们

Matheus Facure(2022): 冰淇淋销售

- 控制变量法：利用线性回归模型来控制这些混淆因子，从而估计价格对销售量的效应：

$$Sales_i = \alpha + \tau price_i + \beta_1 temp_i + \beta_2 cost_i + \beta_3 Weekday_i + e_i$$

τ 是我们感兴趣的处理效应

- 其它参数不是我们感兴趣的，但是我们也要关心它们
- 例如，气候与销售量之间的关系可能是非线性的：

$$Sales_i = \alpha + \tau price_i + \beta_1 temp_i + \beta_2 temp_i^2 + \beta_3 cost_i + \beta_4 Weekday_i + e_i$$

Matheus Facure(2022): 冰淇淋销售

- 控制变量法：利用线性回归模型来控制这些混淆因子，从而估计价格对销售量的效应：

$$Sales_i = \alpha + \tau price_i + \beta_1 temp_i + \beta_2 cost_i + \beta_3 Weekday_i + e_i$$

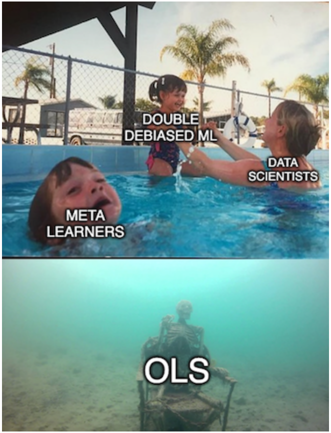
τ 是我们感兴趣的处理效应

- 其它参数不是我们感兴趣的，但是我们也要关心它们
- 例如，气候与销售量之间的关系可能是非线性的：

$$Sales_i = \alpha + \tau price_i + \beta_1 temp_i + \beta_2 temp_i^2 + \beta_3 cost_i + \beta_4 Weekday_i + e_i$$

- 成本和周几的参数呢？如果有几十个，上百个控制变量怎么办？

Matheus Facure(2022): 冰淇淋销售



图片来源于 Matheus Facure(2022)

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)
- 假设线性回归方程为：

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)
- 假设线性回归方程为：

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- 我们可以用下列步骤得到估计参数：

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)
- 假设线性回归方程为：

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- 我们可以用下列步骤得到估计参数：
  1. 用 y 对第二个自变量回归 $\hat{y}^* = \hat{\gamma}_1 X_2$

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)
- 假设线性回归方程为：

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- 我们可以用下列步骤得到估计参数：
  1. 用 y 对第二个自变量回归 $\hat{y}^* = \hat{\gamma}_1 X_2$
  2. 用 X_1 对 X_2 回归 $\hat{X}_1 = \hat{\gamma}_2 X_2$

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)
- 假设线性回归方程为：

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- 我们可以用下列步骤得到估计参数：
  1. 用 y 对第二个自变量回归 $\hat{y^*} = \hat{\gamma}_1 X_2$
  2. 用 X_1 对 X_2 回归 $\hat{X_1} = \hat{\gamma}_2 X_2$
  3. 获得余值 $\tilde{X}_1 = X_1 - \hat{X_1}$，$\tilde{y}_1 = y - \hat{y^*}$

FWL 定理

- FWL 定理是线性回归最有趣的定理之一，例如，Goodman-Bacon(2021),Goldsmith-Pinkham et al.(2024)
- 假设线性回归方程为：

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- 我们可以用下列步骤得到估计参数：
  1. 用 y 对第二个自变量回归 $\hat{y}^* = \hat{\gamma}_1 X_2$
  2. 用 X_1 对 X_2 回归 $\hat{X}_1 = \hat{\gamma}_2 X_2$
  3. 获得余值 $\tilde{X}_1 = X_1 - \hat{X}_1$，$\tilde{y}_1 = y - \hat{y}^*$
  4. 用结果余值对 X_1 的余值回归 $\tilde{y} = \hat{\beta}_1 \tilde{X}_1$

FWL 定理

- FWL 定理意味着用下列模型估计因果系数 τ:

$$Y_i - E[Y_i|X_i] = \tau \cdot (T_i - E[T_i|X_i]) + \epsilon$$

FWL 定理

- FWL 定理意味着用下列模型估计因果系数 τ:

$$Y_i - E[Y_i|X_i] = \tau \cdot (T_i - E[T_i|X_i]) + \epsilon$$

- 我们如何能处理大量的控制变量？如何能处理协变量效应的函数形式？

FWL 定理

- FWL 定理意味着用下列模型估计因果系数 τ:

$$Y_i - E[Y_i|X_i] = \tau \cdot (T_i - E[T_i|X_i]) + \epsilon$$

- 我们如何能处理大量的控制变量？如何能处理协变量效应的函数形式？
- 可以将 ML 应用于 FWL 来构造结果和处理变量的余值:

$$Y_i - \hat{M}_y(X_i) = \tau \cdot (T_i - \hat{M}_t(X_i)) + \epsilon$$

其中，$\hat{M}_y(X_i)$ 是 $E[Y|X]$ 的估计量，$\hat{M}_{(}(X_i)$ 是 $E[T|X]$ 的估计量

1. 因果推断中的内生性
2. DID 中的协变量
3. DID 中的机器学习
   - 一个例子：天气热了，买冰淇淋吗？
   - FWL 定理
   - 参数 DDML
4. 内生的处理配置
5. DID-IV 与 IV 型 DID
6. 参考文献

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：
    1. 用灵活的机器学习回归模型 M_y 来估计 Y 对 X 的回归

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：
    1. 用灵活的机器学习回归模型 M_y 来估计 Y 对 X 的回归
    2. 用灵活的机器学习回归模型 M_T 来估计 T 对 X 的回归

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：
  1. 用灵活的机器学习回归模型 M_y 来估计 Y 对 X 的回归
  2. 用灵活的机器学习回归模型 M_T 来估计 T 对 X 的回归
  3. 获得余值 $\tilde{Y} = Y - M_Y(X)$，$\tilde{T} = T - M_T(X)$

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：
  1. 用灵活的机器学习回归模型 M_y 来估计 Y 对 X 的回归
  2. 用灵活的机器学习回归模型 M_T 来估计 T 对 X 的回归
  3. 获得余值 $\tilde{Y} = Y - M_Y(X)$，$\tilde{T} = T - M_T(X)$
  4. 用结果余值对处理变量余值回归 $\tilde{Y} = \alpha + \tau\tilde{T}$

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：
  1. 用灵活的机器学习回归模型 M_y 来估计 Y 对 X 的回归
  2. 用灵活的机器学习回归模型 M_T 来估计 T 对 X 的回归
  3. 获得余值 $\tilde{Y} = Y - M_Y(X)$，$\tilde{T} = T - M_T(X)$
  4. 用结果余值对处理变量余值回归 $\tilde{Y} = \alpha + \tau\tilde{T}$
- "茅台虽好，可不要贪杯哦"

机器学习的因果推断

- 机器学习可以灵活地刻画交互和非线性
- 这意味着，用机器学习不用对协变量 X 和 Y(T) 之间的关系做任何参数假设
- 假设不存在不可观测混淆因子，那么用下列程序可以获得 ATE：
  1. 用灵活的机器学习回归模型 M_y 来估计 Y 对 X 的回归
  2. 用灵活的机器学习回归模型 M_T 来估计 T 对 X 的回归
  3. 获得余值 $\tilde{Y} = Y - M_Y(X)$，$\tilde{T} = T - M_T(X)$
  4. 用结果余值对处理变量余值回归 $\tilde{Y} = \alpha + \tau\tilde{T}$
- "茅台虽好，可不要贪杯哦"
- ML 的灵活性可能会导致过度拟合的问题

机器学习的过度拟合问题

- Chernozhukov et al (2018) 详细说明了过度拟合问题

机器学习的过度拟合问题

- Chernozhukov et al (2018) 详细说明了过度拟合问题
- 假设是结果与协变量 X 的模型 $M_y(X)$ 存在过度拟合，那么，结果余值 $\tilde{Y} = Y - M_y(X)$ 可能比真实值要小

机器学习的过度拟合问题

- Chernozhukov et al (2018) 详细说明了过度拟合问题
- 假设是结果与协变量 X 的模型 $M_y(X)$ 存在过度拟合，那么，结果余值 $\tilde{Y} = Y - M_y(X)$ 可能比真实值要小
- 这意味着 $M_y(X)$ 刻画了过度 X 和 Y 之间过度的信息

机器学习的过度拟合问题

- Chernozhukov et al (2018) 详细说明了过度拟合问题
- 假设是结果与协变量 X 的模型 $M_y(X)$ 存在过度拟合，那么，结果余值 $\tilde{Y} = Y - M_y(X)$ 可能比真实值要小
- 这意味着 $M_y(X)$ 刻画了过度 X 和 Y 之间过度的信息
- 有一部分信息可能来自于 T 和 Y 之间的关系

机器学习的过度拟合问题

- Chernozhukov et al (2018) 详细说明了过度拟合问题
- 假设是结果与协变量 X 的模型 $M_y(X)$ 存在过度拟合，那么，结果余值 $\tilde{Y} = Y - M_y(X)$ 可能比真实值要小
- 这意味着 $M_y(X)$ 刻画了过度 X 和 Y 之间过度的信息
- 有一部分信息可能来自于 T 和 Y 之间的关系
- 这时，余值回归就会趋向于 0，即 $M_y(X)$ 包含了部分因果效应，从而使得余值回归的效应缩小

机器学习的过度拟合问题

- Chernozhukov et al (2018) 详细说明了过度拟合问题
- 假设是结果与协变量 X 的模型 $M_y(X)$ 存在过度拟合，那么，结果余值 $\tilde{Y} = Y - M_y(X)$ 可能比真实值要小
- 这意味着 $M_y(X)$ 刻画了过度 X 和 Y 之间过度的信息
- 有一部分信息可能来自于 T 和 Y 之间的关系
- 这时，余值回归就会趋向于 0，即 $M_y(X)$ 包含了部分因果效应，从而使得余值回归的效应缩小
- 同理，对于 M_T 的过度拟合问题，它会包含部分 T 的方差，从而使得处理变量余值方差过小，那么，最终估计量的方差会非常大

交叉验证法

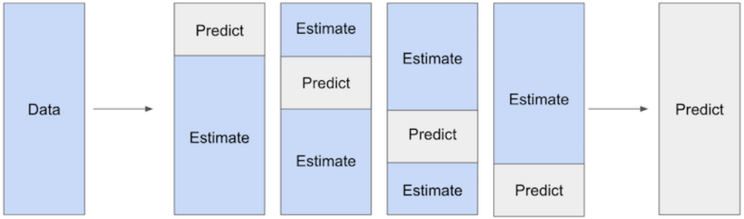
- 将数据划分成 K 个部分，用 K-1 个部分的数据估计 M，用剩余的部分数据来进行余值回归

交叉验证法

- 将数据划分成 K 个部分，用 K-1 个部分的数据估计 M，用剩余的部分数据来进行余值回归
- 结合所有 K 部分的预测值来估计最终的因果模型 $\tilde{Y} = \alpha + \tau \tilde{T}$

交叉验证法

- 将数据划分成 K 个部分，用 K-1 个部分的数据估计 M，用剩余的部分数据来进行余值回归
- 结合所有 K 部分的预测值来估计最终的因果模型
  $\tilde{Y} = \alpha + \tau \tilde{T}$
- 如下图



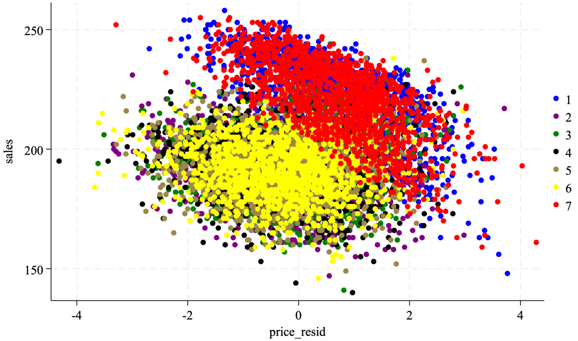
图片来源于 Matheus Facure(2022)

例子：夏天到了，冰淇淋畅销吗？

- 我们用 stata 命令 ddml 来估计处理模型 M_t，协变量是 temp，weekday，cost

例子：夏天到了，冰淇淋畅销吗？

- 我们用 stata 命令 ddml 来估计处理模型 M_t，协变量是 temp，weekday，cost
- 余值 $\tilde{T} = T - M_t(X)$ 可以视为所有来自 X 的混淆偏误都被移除了



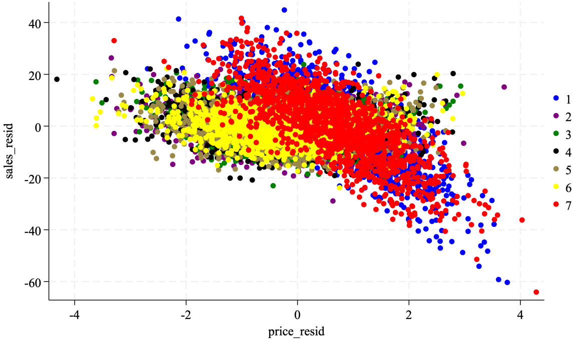
dofile:didml

例子：夏天到了，冰淇淋畅销吗？

- 处理模型 M_t 的作用是纠偏，那么，M_y 呢？它的作用是移除 Y 的方差

例子：夏天到了，冰淇淋畅销吗？

- 处理模型 M_t 的作用是纠偏，那么，M_y 呢？它的作用是移除 Y 的方差
- 直觉上，M_y 创建了一个结果变量，移除了所有 X 引起的方差



dofile:didml

例子：夏天到了，冰淇淋畅销吗？

```
. *手动余值回归
.
. gen double Y1_resid = $Y-Y2_pystacked_1

. gen double D1_resid = $T-D1_reg_1

. reg Y1_resid D1_resid, robust

Linear regression                               Number of obs     =     10,000
                                                F(1, 9998)        =    1054.24
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1441
                                                Root MSE          =     8.7592

------------------------------------------------------------------------------
             |               Robust
    Y1_resid | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    D1_resid |  -3.268805   .1006746   -32.47   0.000    -3.466148   -3.071462
       _cons |   .0056613   .0875904     0.06   0.948    -.1660336    .1773562
------------------------------------------------------------------------------
```

dofile:didml

例子：夏天到了，冰淇淋畅销吗？

```
DDML estimation results:
spec  r     Y learner     D learner         b        SE
   1  1        Y1_reg        D1_reg     1.619   (0.148)
   2  1        Y1_reg   D2_pystacked   -2.998   (0.154)
*  3  1   Y2_pystacked       D1_reg    -3.269   (0.101)
   4  1   Y2_pystacked  D2_pystacked   -3.938   (0.099)
* = minimum MSE specification for that resample.

Min MSE DDML model
y-E[y|X]  = y-Y2_pystacked_1                    Number of obs   =     10000
D-E[D|X]  = D-D1_reg_1
```

sales	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
price	-3.268805	.1006746	-32.47	0.000	-3.466124	-3.071486
_cons	.0056613	.0875904	0.06	0.948	-.1660128	.1773355

dofile:didml

1 Sant' Anna, DiD Resources, 2025

2 Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68

3 Alves, Matheus Facure. "Causal inference for the brave and true." São Paulo, Brazil. Available at https://matheusfacure.github.io/python-causality-handbook/01-Introduction-To-Causality.html (2022).

4 Goldsmith-Pinkham P, Hull P, Kolesár M. Contamination bias in linear regressions[J]. American Economic Review, 2024, 114(12): 4015-4051.

Thanks!