专题三：坏控制变量

黄嘉平

2026.4.5

在经济学实证研究中，研究者们最为关心的问题之一是回归分析中潜在的遗漏变量带来的估计偏差。由于遗漏变量偏差广为人知，人们倾向于尽可能多地向回归方程中添加控制变量，甚至认为添加的控制变量越多，估计偏差越小。

但是，增加一个控制变量并不能保证让回归系数的估计值更接近真实值。事实是有些变量控制了反而会让估计偏差不降反增，学术界把这种变量称为“坏控制变量”（bad controls），而那些能够降低估计偏差的控制变量是“好控制变量”（good controls）。在 Angrist & Pischke (2009, 2014) 中，作者认为受处理变量影响的变量都是坏控制变量，而在处理发生时已经确定取值的变量是好控制变量。但是随着研究的深入，现在对坏控制变量的界定比过去更加清晰了（Cinelli et al., 2024)。接下来我们通过几个例子了解什么是坏控制变量，并尝试验证理论推断是否正确。

1. 变量间的因果关系和 DAG

区分控制变量好坏的第一步是正确描述一个因果模型。当我们要表达变量 X 的变化引起了变量 Y 的变化，即 X 是 Y 的因，Y 是 X 的果时，我们可以写成

$$X \to Y$$

这里符号 $\to$ 代表因果的方向。以此类推，$X \to Z \to Y$ 表达了一个因果链条，即 X 引起了 Z 的变化，Z 进一步引起了 Y 的变化。这里 Z 起到了中介的作用。$X \to Z \leftarrow Y$ 表达了 X 和 Y 都能引起 Z 的变化。$X \leftarrow Z \to Y$ 则表达了 Z 引起了 X 和 Y 的变化。用这个方法可以表达更长的因果链条，例如 $X \to M_1 \to M_2 \to \ldots \to M_k \to Y$，但无法表达单一链条以外的因果关系。

我们可以借助图论中的有向无环图（directed acyclic graph, DAG）来表达多个因果链条叠加在一起的因果网络。在图论中，图（graph）的含义等同于网络（network），包含节点和边两种图形概念。如果边有方向，则称为有向图。如果从一个节点出发，按照边的方向依次走到下一节点，最终能够走回出发点，那么就形成了一个环。有向无环图（DAG）就是指边有方向但是不包含环的图。图 1 中给出了两个 DAG 的例子。

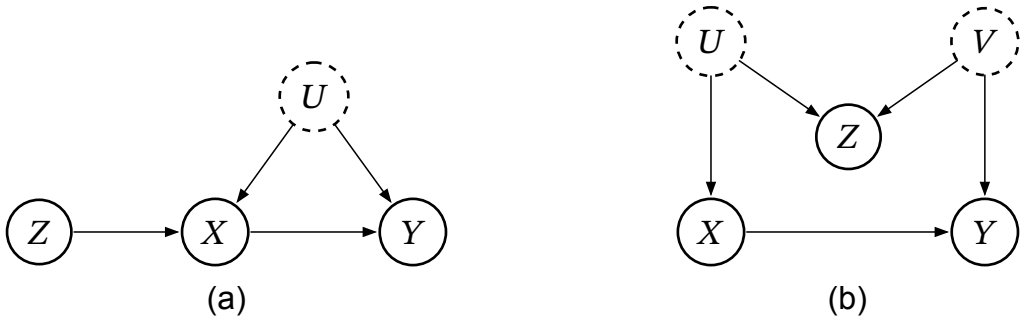

图 1　两个有向无环图

在结构方程模型的诞生初期，计量经济学家们曾经提倡用有向图描述变量间的关系，并将该方法称为 path diagram (Wright, 1920, 1921)，但并没有在实证研究中被广泛认知和采用。Pearl (1995) 将 DAG 和他的因果推断理论融合在一起。随着近年来因果推断在实证研究中的流行，用 DAG 表达因果关系的方法也逐渐被广泛接受。Pearl 的因果图中的核心概念是路径（path），即一串首尾相连的边，无论这些边的方向是否一致。例如 $X \to Z \to Y$，$X \to Z \leftarrow Y$，$X \leftarrow Z \to Y$ 都是连接 X 和 Y 的路径。这里面只有 $X \to Z \to Y$ 是因果路径（causal path）。

需要注意的是，用 DAG 表达的因果关系并不仅限于线性关系，这也是因果图和 path diagram 最根本的区别。因此，在因果分析中，所谓的“控制某一变量“意味着固定其取值，而不应该单纯的和在回归模型中添加控制变量划等号。但是，如果我们在因果图的基础上假设所有的因果关系都是线性的，那就可以写出对应的线性回归方程组，并利用最小二乘法对因果效应进行估计。

2. 三种基本因果机制

下面针对上一节中介绍的三种 X 和 Y 之间的机制，讨论是否应该对变量 Z 加以控制。我们的目标是正确估计 X 对 Y 的因果效应。以下假设线性模型成立，因此可以利用回归方法。

- $X \to Z \to Y$：这是一个因果中介的例子，Z 是一个**中介变量**（mediator）。Z 的存在不影响 X 和 Y 之间的因果关系，因此不需要进行控制。反之，如果控制了 Z 就会导致 X 的变化无法传递给 Y，使估计结果产生偏差。这一机制中的 Z 符合被广泛认知的坏控制变量的概念。

```
library(tidyverse)
library(broom) # 将回归结果转换为 tibble 形式的程序包，包含 tidy() 函数
n <- 1000
set.seed(111)

x <- runif(n,5,10)
z <- 2 + 3*x + rnorm(n,1,2)
y <- 5 + 2*z + rnorm(n,2,2)
# X 增加一个单位会导致 Y 增加 (3*2=6) 个单位

sim_data1 <- tibble(x, y, z)

lm(y ~ x, sim_data1) |> tidy()  # 不控制 Z （正确）
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    12.4     0.755       16.5 3.14e-54
2 x               6.05    0.0989      61.2 0
```

```r
lm(y ~ x + z, sim_data1) |> tidy()  # 控制 Z （抵消了 X 的变化）
```

```
# A tibble: 3 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   6.74      0.347     19.4   1.64e-71
2 x            -0.0126    0.105     -0.121 9.04e- 1
3 z             2.01      0.0316    63.7   0
```

- $X \leftarrow Z \rightarrow Y$：在这个机制中，$Z$ 是 X 和 Y 的共因，除此之外没有连接 X 和 Y 的路径。因此 X 和 Y 之间不存在因果关系，只存在由 Z 产生的相关关系（通常称为伪相关）。这种 Z 被称为**混杂因素**（confounder）。这时如果没有控制 Z，回归结果就会呈现出伪相关而无法正确估计因果效应，因此应该控制 Z。Pearl (1995, 2000) 将这种由指向 X 的边开始的路径称为**后门路径**，而此时控制 Z 就起到了阻断后门路径的作用。

```r
set.seed(222)
z <- runif(n,0,10)
x <- 2*z + rnorm(n)
y <- 5 + 3*z + rnorm(n) # X 和 Y 之间不存在因果关系

sim_data2 <- tibble(x, y, z)

lm(y ~ x, sim_data2) |> tidy() # 不控制 Z （伪相关）
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     5.37   0.112        47.7 6.80e-260
2 x               1.46   0.00979     149.  0
```

```r
lm(y ~ x + z, sim_data2) |> tidy() # 控制 Z （正确）
```

```
# A tibble: 3 × 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)  4.89       0.0631    77.5   0
2 x           -0.00559    0.0314    -0.178 8.59e-  1
3 z            3.02       0.0636    47.5   5.36e-258
```

- $X \rightarrow Z \leftarrow Y$：这是一个在过去的文献中常常被忽略的机制。这里 X 和 Y 同时影响 Z，而 Z 被称为**对撞变量**（collider）。和混杂因素相反，这里如果控制了 Z，就会制造出本不存在的伪相关，因此不去控制 Z 是正确的做法。需要注意的是对撞变量并不总是坏控制变量，关键要看它出现在什么位置。

```r
set.seed(333)
x <- runif(n,0,10)
y <- rnorm(n,2,5)
z <- x + 2*y + rnorm(n)  # Z 是对撞变量，X 和 Y 是相互独立的
```

```r
sim_data3 <- tibble(x, y, z)
```

```r
lm(y ~ x, sim_data3) |> tidy()  # 不控制 Z （正确）
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   2.25      0.313      7.18  1.32e-12
2 x            -0.0418    0.0543    -0.770 4.42e- 1
```

```r
lm(y ~ x + z, sim_data3) |> tidy()  # 控制 Z （制造了伪相关）
```

```
# A tibble: 3 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)   0.0244   0.0323      0.756   0.450
2 x            -0.492    0.00566   -86.9     0
3 z             0.491    0.00157   312.      0
```

3. 工具变量是坏控制变量

工具变量法常用来解决回归中的遗漏变量偏差问题。图 1a 描绘了包函工具变量的常见因果模型。X 和 Y 之间存在一个共因 U，但是我们无法观测 U 的取值（不可观测变量在图中以虚线表达），因此无法直接进行控制。这导致了回归模型 $Y = \alpha + \beta X + \varepsilon$ 中系数 β 的最小二乘估计量有偏。而图中的 Z 能且仅能通过 X 影响 Y，因此是一个有效工具变量。由于 Z 不受 U 的影响，它的变化是外生的，由它引起的 X 的变化也是外生的，因此可以通过工具变量 Z 将 X 对 Y 的因果效应从总相关性中分离处来。

如果我们错误地理解了 Z 的作用而将它当作控制变量会发生什么呢? 结论是控制 Z 会放大估计偏差。这是因为由 Z 带来的外生变化被固定了，导致回归系数中内生部分的比重被放大。DiTraglia (2023) 给出了更加详细的解释。

```r
library(ivreg) # 进行工具变量估计的程序包
```

```r
set.seed(444)
u <- rnorm(n,0,2)
z <- runif(n,0,10)
x <- 1 + 2*z + 2*u + rnorm(n)
y <- 2 + x + 3*u + rnorm(n) # X 对 Y 的因果效应为 1
```

```r
sim_data4 <- tibble(x, y, z, u)
```

```r
lm(y ~ x, sim_data4) |> tidy() # 不控制 Z （遗漏变量偏差）
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    -3.13    0.297      -10.5 9.72e-25
2 x               1.48    0.0230      64.3 0
```

```
ivreg(y ~ x | z, data = sim_data4) |> summary() # 工具变量法
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)    2.24     0.422       5.31 1.34e-  7
2 x              0.980    0.0345     28.4  1.50e-130
```

```
lm(y ~ x + z, sim_data4) |> tidy() # 控制 Z （放大了偏差）
```

```
# A tibble: 3 × 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)    0.523    0.112       4.68 0.00000331
2 x              2.41     0.0136    178.   0
3 z             -2.80     0.0329    -85.1  0
```

4. M-偏差

图 1b 是一种典型的因果机制，其中 Z 是和 X-Y 间的因果关系无关的其他变量，如果它的取值在处理之前已经确定，则传统观点认为它是个好控制变量，至少可以提高估计结果的精确度。然而在这个因果机制中，Z 是一个典型的坏控制变量，因为它是一个对撞变量，控制 Z 会打开后门路径 $X \to U \to Z \leftarrow V \to Y$，使 X 的回归系数中包含伪相关。这种估计偏差被称为 M-偏差（因果链条的形状像英文字母 M）。

```
set.seed(555)
u <- rnorm(n,0,2)
v <- runif(n,0,10)
z <- u + 0.5*v + rnorm(n)
x <- 1 + 2*u + rnorm(n)
y <- 2 + 1.5*x + 3*v + rnorm(n)

sim_data5 <- tibble(x, y, z, u, v)

lm(y ~ x, sim_data5) |> tidy() # 不控制 Z （正确）
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    16.8     0.285       59.2 0
2 x               1.48    0.0689      21.4 4.36e-84
```

```r
lm(y ~ x + z, sim_data5) |> tidy() # 控制 Z （M-偏差）
```

```
# A tibble: 3 × 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)    9.53     0.247      38.6  2.55e-200
2 x             -0.257    0.0590     -4.36 1.46e-  5
3 z              3.75     0.0903     41.5  1.57e-219
```

Cinelli et al. (2024) 中包含了更多关于坏控制变量的讨论。如果想要全面了解 Pearl 的因果推断理论，推荐阅读 Pearl & Mackenzie (2018)。

5. 思考题

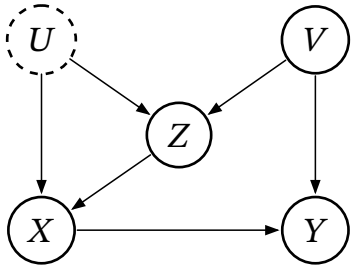

图 2　一个因果模型

针对图 2 中的因果模型，在假定线性关系的前提下，控制哪些变量才能正确估计出 X 对 Y 的因果效应？利用蒙特卡洛仿真确认你的推断。

参考文献

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Cinelli, C., Forney, A., & Pearl, J. (2024). A Crash Course in Good and Bad Controls. *Sociological Methods & Research*, 53(3), 1071-1104. https://doi.org/10.1177/00491241221099552

DiTraglia, F. (2023). A Good Instrument is a Bad Control. https://www.econometrics.blog/post/a-good-instrument-is-a-bad-control/

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4), 669-688.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Wright, S. (1920). The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs. *PNAS*, 6(6), 320-332.

Wright, S. (1921). Correlation and causation. *Journal of Agriculture Research*, 20(7), 557-585.