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 21, 2024

Contents

开篇

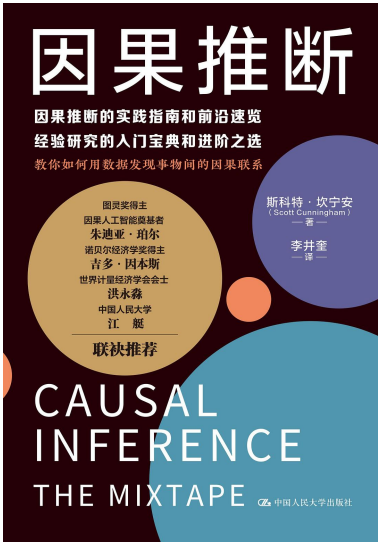
- 推荐的先修课教材
  - –《计量经济学导论：现代观点》(Introductory Econometrics: A Modern Approach), 伍德里奇 (Jeffrey M Wooldridge) 著。
  - –《计量经济学》(Introduction to Econometrics), 斯托克、沃森 (James H Stock and Mark W Watson) 著。
- 推荐阅读教材
  - –《因果推断》，斯科特·坎宁安，中国人民大学出版社，2023 年。
  - –《因果推断初步》，姚东旻，清华大学出版社，2022 年。
  - –《因果推断实用计量方法》，邱嘉平，上海财经大学出版社，2020 年。
  - –《基本无害的计量经济学》(Mostly Harmless Econometrics: An Empiricist's Companion), 安格里斯特 (Joshua D Angrist)、皮施克 (Jorn-Steffen Pischke) 著。格致出版社，2012 年。
  - –Mastering 'Metrics: The Path from Cause to Effect, by Joshua D Angrist and Jorn-Steffen Pischke, 2015.
  - –Microeconometrics Using Stata（用 STATA 学微观计量经济学）, Revised Edition, by A Colin Cameron and Pravin K Trivedi, 2010.

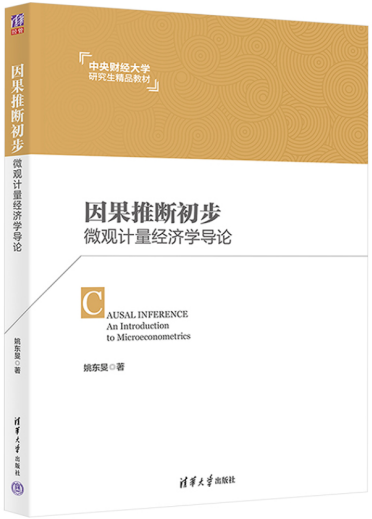
–《横截面与面板数据的计量经济分析》(Econometric Analysis of Cross Section and Panel Data), 伍德里奇 (Jeffrey M Wooldridge) 著。中文第二版，中国人民大学出版社，2016 年。

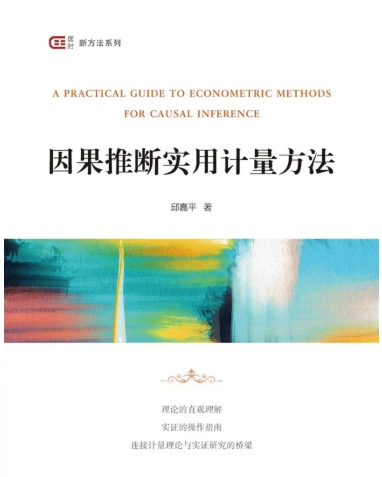
–Econometrics, Bruce Hansen. 2022 年。

–《计量经济分析》(Econometric Analysis), 格林 (William H Greene) 著。中文第六版，中国人民大学出版社，2011 年。英文第六版，中国人民大学出版社，2009 年。英文最新版，7th edition, 2011.

–《计量经济学》(Econometrics), 林文夫 (Fumio Hayashi) 著。中文版，上海财经大学出版社，2005 年。


因果推断
因果推断的实践指南和前沿速览
经验研究的入门宝典和进阶之选
教你如何用数据发现事物间的因果联系
图灵奖得主
因果人工智能奠基者
朱迪亚·珀尔
诺贝尔经济学奖得主
吉多·因本斯
世界计量经济学会会士
洪永淼
中国人民大学
江 艇
联袂推荐
斯科特·坎宁安
（Scott Cunningham）
著
李井奎
译
CAUSAL
INFERENCE
THE MIXTAPE
中国人民大学出版社


中央财经大学
研究生精品教材
因果推断初步
微观计量经济学导论
CAUSAL INFERENCE
An Introduction
to Microeconometrics
姚东旻 著
清华大学出版社


新方法系列
A PRACTICAL GUIDE TO ECONOMETRIC METHODS
FOR CAUSAL INFERENCE
因果推断实用计量方法
邱嘉平 著
理论的直观理解
实证的操作指南
连接计量理论与实证研究的桥梁

先讲 OLS

1 潜在结果分析框架与因果推断

1.1 Taking error terms seriously

- 一个例子：我们想研究 D 和 Y 之间的关系
  - D: some binary treatment (是否上大学)
  - Y: some outcome (40 岁时的工资水平)
- 假定数据 (D_i, Y_i) 独立同分布，于是我们写下一个线性模型：

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

- 那么问题来了：ε_i 是什么玩意儿?
- 显然，我们只有知道 ε_i 的含义，才能知道 β_0 和 β_1 的含义。

第一种理解：ε 表示期望残差 (expectational residual)

- 令

$$\beta_0 \equiv E[Y_i|D_i = 0]$$

$$\beta_1 \equiv E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

自然有

$$E[Y_i|D_i] = \beta_0 + \beta_1 D_i$$

(请注意，此处的线性并不是一个假设，所谓 saturated model [1])

- 然后定义 $\varepsilon_i \equiv Y_i - \beta_0 - \beta_1 D_i$. 根据以上构造 (By construction)

$$E[\varepsilon_i|D_i] = 0$$

于是 OLS 估计量无偏且一致。

[1] 饱和模型（saturated model）是指各观测变量之间均容许相关的最复杂模型。
Chatgpt: In statistics, a saturated model is a model that has the maximum number of parameters that can be estimated from the given data, such that the model perfectly fits the data. In other words, a saturated model has zero degrees of freedom, because there are no extra parameters that can be estimated to improve the fit of the model.
此例中，$E[Y_i|D_i] = \beta_0 + \beta_1 D_i$ 并不需要线性模型假设就可以得到。

- 即使 D 不是 binary，$E(Y|D)$ 通常也不是线性的，我们仍然可以类似地把 (β_0, β_1) 定义为

$$(\beta_0, \beta_1) = \arg\min_{\beta_0, \beta_1} E[(Y_i - \beta_0 - \beta_1 D_i)^2]$$

并将 $\beta_0 + \beta_1 D_i$ 理解为"best linear predictor"of Y_i given D_i. 在这种理解中，$\hat{\beta}$ 永远是 β 的一致估计。[2]

[2]这部分详见 Mostly Harmless Econometrics: An Empiricist's Companion。

第二种理解：ε 表示结构误差（structural error）

- 即不可观察的解释因素，例如能力。

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

那么，如果能力和 treatment 相关

$$Cov(D_i, \varepsilon_i) \neq 0$$

则 β_0,β_1 的 OLS 估计是有偏且不一致的。

- 我们把这种 β_0, β_1 称为 structual or causal parameters. 这时候，函数形式的假定就具有了实质性含义（即使对于 binary D 也是如此）。
- 接下来的问题是，**为什么第二种理解是重要的?**

1.2 条件期望与回归

- 我们花了很多精力学习回归理论。回归就是最小二乘，是一种估计一个变量在给定其它变量下的条件期望的工具。条件期望为什么重要?
  示例 1. 教育与收入（CFPS，2018）
- 条件期望函数 (conditional expectation function, CEF) $\mathbb{E}(y|\mathrm{x})$ 是给定 x 对 y 的最佳预测。

$$\mathbb{E}(y|\mathrm{x}) = \arg\min_{f(x)} \mathbb{E}(y - f(\mathrm{x}))^2$$

$$
\begin{aligned}
&\mathbb{E}[y - f(\mathbf{x})]^2 \\
=&\mathbb{E}[(y - E(y \mid \mathbf{x})) + (E(y \mid \mathbf{x}) - f(\mathbf{x}))]^2 \\
=&\mathbb{E}(y - E(y \mid \mathbf{x}))^2 + 2\mathbb{E}(y - \mathbb{E}(y \mid \mathbf{x}))(\mathbb{E}(y \mid \mathbf{x}) - f(\mathbf{x})) + \mathbb{E}(\mathbb{E}(y \mid \mathbf{x}) - f(\mathbf{x}))^2
\end{aligned}
$$

中间一项为：

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x},y}(y - \mathbb{E}(y \mid \mathbf{x}))(E(y \mid \mathbf{x}) - f(\mathbf{x})) \\
=&\mathbb{E}_{\mathbf{x}}\mathbb{E}_{y|\mathbf{x}}(y - \mathbb{E}(y \mid \mathbf{x}))(\mathbb{E}(y \mid \mathbf{x}) - f(\mathbf{x})) \\
=&\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}(y\mathbb{E}(y \mid \mathbf{x})) - \mathbb{E}(y \mid \mathbf{x})^2 - f(\mathbf{x})\mathbb{E}_{y|\mathbf{x}}y + f(\mathbf{x})\mathbb{E}(y \mid \mathbf{x})\right] \\
=&\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}(y \mid \mathbf{x})^2 - \mathbb{E}(y \mid \mathbf{x})^2 - f(\mathbf{x})\mathbb{E}(y \mid \mathbf{x}) + f(\mathbf{x})\mathbb{E}(y \mid \mathbf{x})\right] = 0
\end{aligned}
$$

因此，让 $E[y - f(\mathbf{x})]^2$ 最小的 $f(\mathbf{x})$ 为：

$$
f(\mathbf{x}) = \mathbb{E}(y|\mathbf{x})
$$

- 定义期望残差 $\tilde{\varepsilon} \triangleq y - \mathbb{E}(y|\mathbf{x})$，具有如下性质：
  - $\tilde{\varepsilon}$ 均值独立于 $\mathbf{x}$，即 $\mathbb{E}(\tilde{\varepsilon}|\mathbf{x}) = 0$.
  - $\tilde{\varepsilon}$ 期望为零，即 $\mathbb{E}(\tilde{\varepsilon}) = 0$.
  - $\tilde{\varepsilon}$ 与 $\mathbf{x}$ 不相关，即 $\mathbb{E}(\mathbf{x}\tilde{\varepsilon}) = 0$.
  - $\tilde{\varepsilon}$ 均值独立于 $\mathbf{x}$ 的任意函数，即 $\mathbb{E}(\tilde{\varepsilon}|f(\mathbf{x})) = 0$.
  - $\tilde{\varepsilon}$ 与 $\mathbf{x}$ 的任意函数不相关，即 $\mathbb{E}(f(\mathbf{x})\tilde{\varepsilon}) = 0$.
- 一种重要的特殊情形是线性条件期望函数：

$$\mathbb{E}(y|\mathbf{x}) = \mathbf{x}'\beta$$

- 定义总体最小二乘问题：

$$\min_{\beta} \mathbb{E}(y - \mathbf{x}'\beta)^2$$

- 显然，当 CEF 确为线性时，总体最小二乘问题的解即为 CEF。
- 但我们通常并不知道 CEF 的函数形式，因此线性只是对其的近似。可以证明，当 CEF 为非线性时，总体最小二乘问题的解是 CEF 的**最佳线性近似**。

$$\arg\min_{\beta} \mathbb{E}(y - \mathbf{x}'\beta)^2 = \arg\min_{\beta^*} \mathbb{E}(\mathbb{E}(y|\mathbf{x}) - \mathbf{x}'\beta^*)^2$$

- 我们把 $\mathbf{x}'\beta$ 称作总体回归函数。定义总体回归残差 $\tilde{\tilde{\varepsilon}} \triangleq y - \mathbf{x}'\beta$，得到如下线性回归模型：

$$y = \mathbf{x}'\beta + \tilde{\tilde{\varepsilon}}$$

求解总体最小二乘问题，可得[3]

$$\mathbb{E}(\mathbf{x}(y - \mathbf{x}'\beta)) = \mathbb{E}(\mathbf{x}\tilde{\tilde{\varepsilon}}) = 0$$

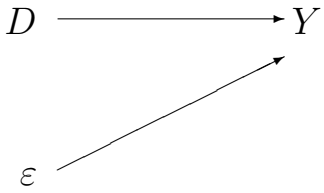
$$\beta = \left[\mathbb{E}(\mathbf{x}\mathbf{x}')\right]^{-1}\mathbb{E}(\mathbf{x}y)$$

[3]注意，$\mathbb{E}(\tilde{\tilde{\varepsilon}}|\mathbf{x}) = 0$ 未必成立，只有当 CEF 确实是线性时才成立。而 $\mathbb{E}(\tilde{\varepsilon}|\mathbf{x}) = 0$ 始终成立。

1.3 何为因果推断?

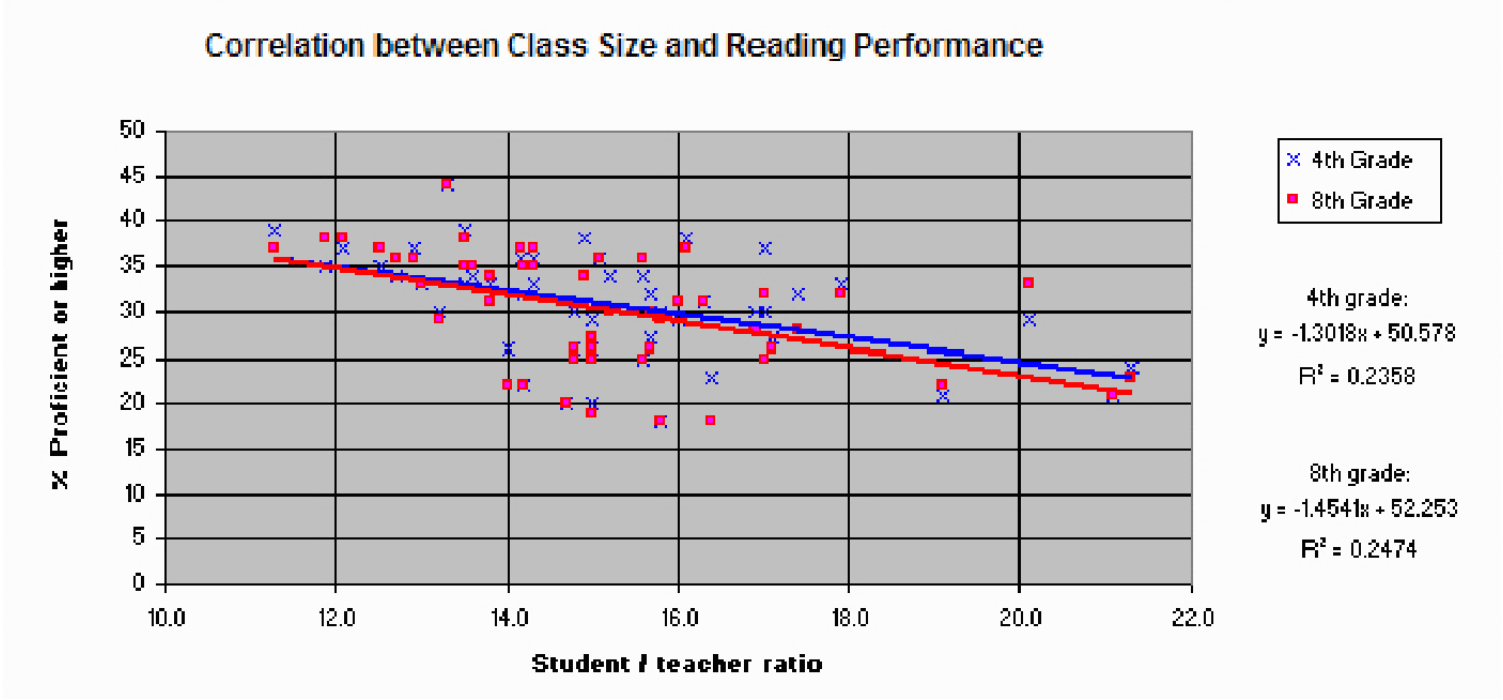
- 我们关注 CEF 的目的在于，它能帮助我们理解变量之间的因果关系。
- 我们真正关心的因果问题是：读书有没有用？一个人如果多上一年学，他的工资水平预期能增长多少?

- 用 Y 表示我们感兴趣的结果（outcome），或反应（response）。
- 用 D 表示我们感兴趣的原因（cause），或处理（treatment）、干预（intervention）。D 可以是离散的，也可以是连续的。
- 用 ε 表示影响结果的其他因素。
- 我们感兴趣的因果关系可以用如下的基本因果模型来刻画：

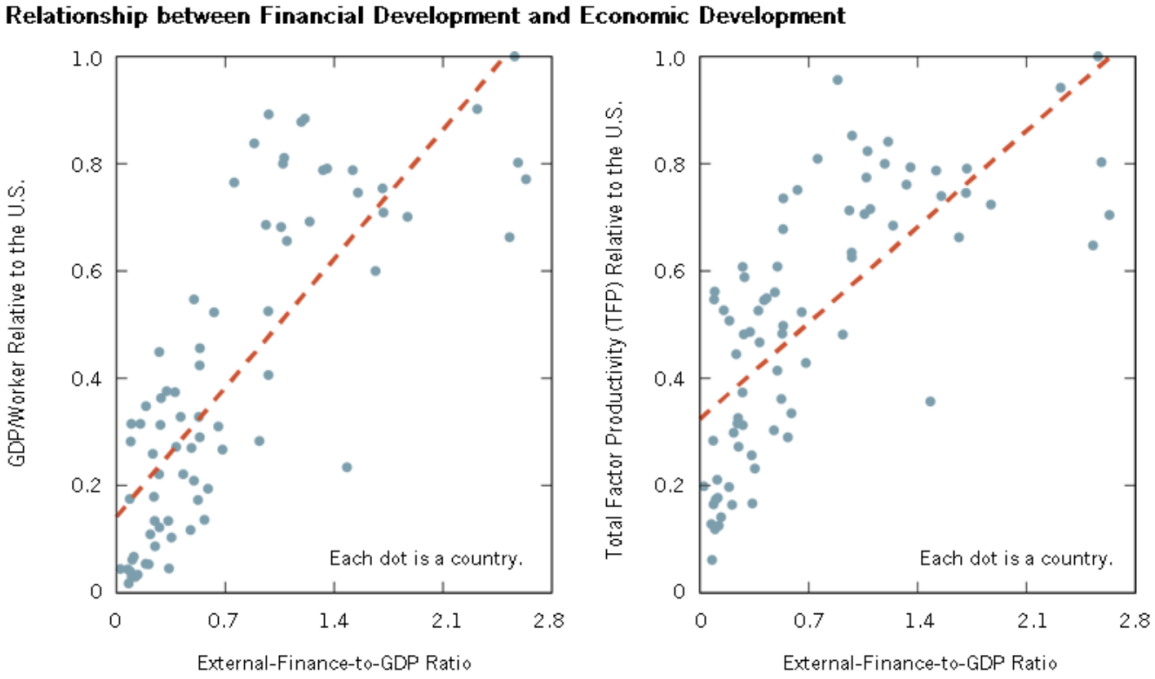


- 再来看三个例子，刻画了三组相关性事实，体会它们试图讲述什么因果故事？

示例. 班级规模与教育产出

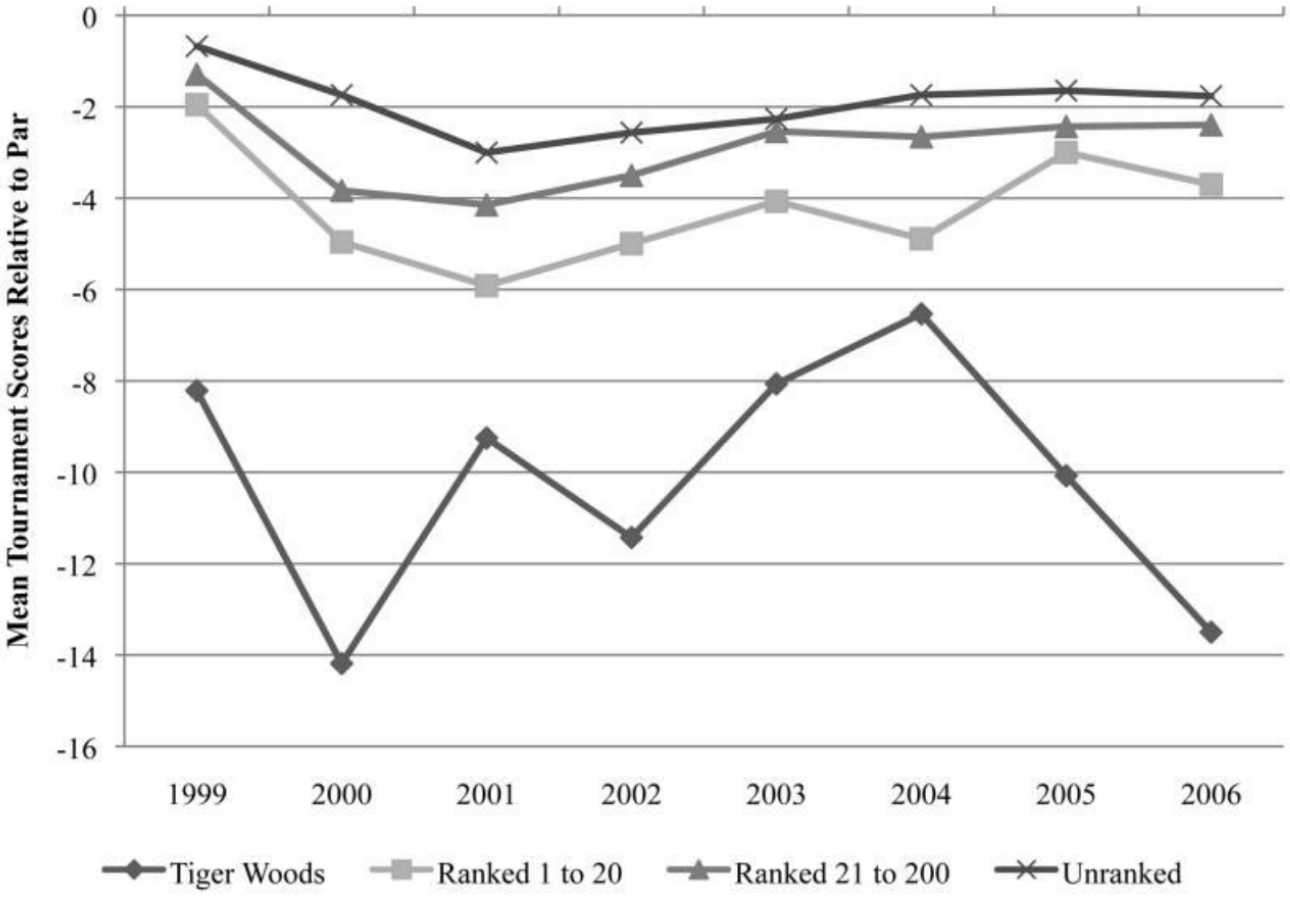


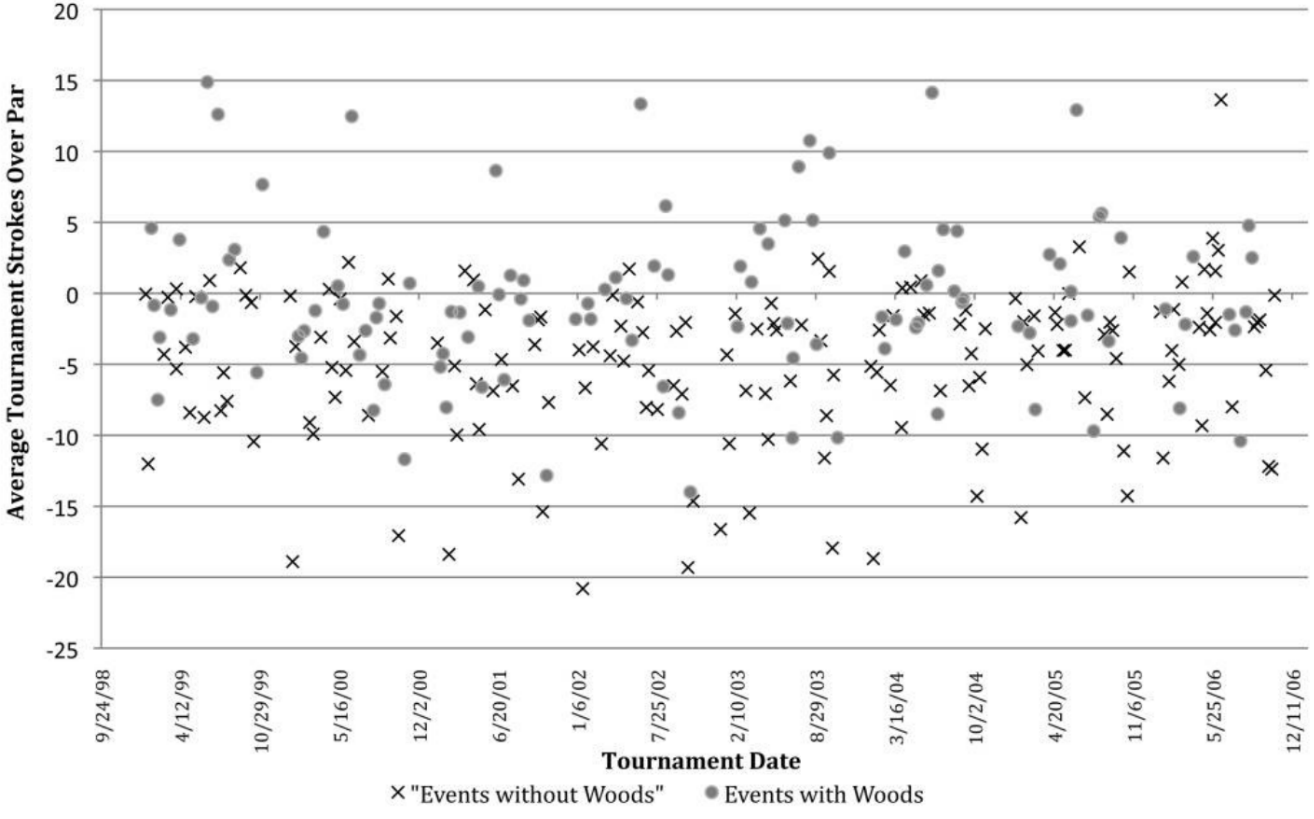
示例. 金融发展和经济增长



示例. 超级明星效应（Brown, 2011, JPE）

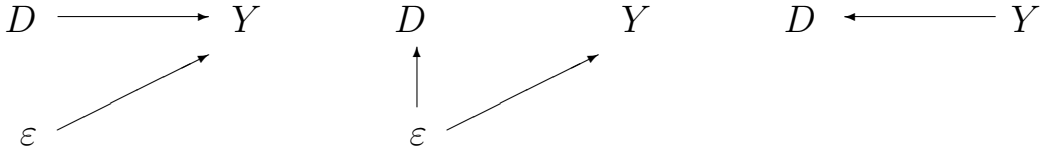
- 生活常识：竞争是一种重要的激励机制。考核相对绩效的锦标赛机制要想发挥作用，有一项重要前提——竞争者的能力必须相对均衡。存在“超级明星”时，锦标赛机制反而会产生负面效果。
- 研究情境：“老虎”伍兹，史上最伟大的高尔夫球手。1975 年出生，1996 年 20 岁时成为职业球手，职业生涯未满一年即跃居世界排名第一，在 1999 年 8 月至 2004 年 9 月以及 2005 年 6 月至 2010 年 10 月分别连续 264 周和 281 周保持世界排名第一。


Mean Tournament Scores Relative to Par
0
-2
-4
-6
-8
-10
-12
-14
-16
1999
2000
2001
2002
2003
2004
2005
2006
Tiger Woods
Ranked 1 to 20
Ranked 21 to 200
Unranked


Average Tournament Strokes Over Par
20
15
10
5
0
-5
-10
-15
-20
-25
9/24/98
4/12/99
10/29/99
5/16/00
12/2/00
6/20/01
1/6/02
7/25/02
2/10/03
8/29/03
3/16/04
10/2/04
4/20/05
11/6/05
5/25/06
12/11/06
Tournament Date
"Events without Woods"
Events with Woods

- 如果我们感兴趣的从 D 到 Y 的因果关系真的存在，那么 D 和 Y 之间的相关性必然存在，反之则不然。D 和 Y 相关这一事实可能被多个基本因果模型所合理化 (rationalize)：

模型 1　　模型 2　　模型 3

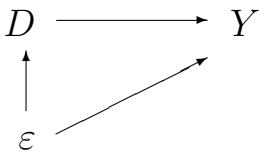


- 重新审视前述三个例子，除了最显然的因果故事以外，还有没有其它竞争性 (alternative/competitive) 的解释？
- 如果在特定的研究情境下，变量之间满足一定的假设条件，使得一个特定的因果模型没有与之竞争的观测上等价 (observationally equivalent) 的因果模型，我们就说这个特定的因果模型被识别 (identified)。
- 这样的假设被称作识别假设，我们马上就会看到，识别假设是永远无法严格证明的，只能根据社会科学理论加以论证。

- 因此任何因果推断问题都包含两部分：
  - **因果识别** (causal identification)：如果拥有整个总体，是否能够确定总体因果关系？这是社会科学理论的任务。因果识别的基本逻辑是：如果相关性不存在，则因果性不存在；如果相关性存在，且只有一种因果模型可以合理化这种相关性，则这种特定的因果性存在。
  - **统计推断** (statistical inference)：如何从样本数据获取关于总体因果关系的信息？这是统计学的任务。统计推断致力于发现 Y 和 D 在样本中的相关性，并由此评估其总体相关性。

- 真实世界的数据生成过程 (data generating process) 可能是多个基本因果模型同时作用

模型 4



的结果。

- 模型 1 和模型 4 都可以用如下的线性模型来表示：

$$Y_i = \beta_1 D_i + \varepsilon_i'$$

做一下技术处理：定义 $\beta_0 \triangleq \mathbb{E}(\varepsilon_i')$，定义 $\varepsilon_i \triangleq \varepsilon_i' - \beta_0$，则

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \mathbb{E}(\varepsilon_i) = 0$$

- β_1 是我们重点关注的未知的总体因果参数（population causal parameter），其含义是，保持 ε 不变，一项处理的实施（D 由 0 变到 1，或 D 变化一个单位），导致结果变化 β_1，称为因果效应（causal effect）或处理效应（treatment effect）。
- 这个模型看起来和线性回归模型长得很相似，但含义截然不同。这个模型被称为**结构模型**，因为它包含了我们关于因果关系的先验知识：ε 的含义，模型的线性性质，以及 D

和 ε 之间的关系，这些知识将帮助我们识别 β_1。因此 β_1 被称为结构参数，ε 被称为结构误差项或结构扰动项。

- 若 ε 均值独立于 D，即 $\mathbb{E}(\varepsilon|D)=0$，则 $\mathbb{E}(Y|D)=\beta_0+\beta_1 D$，因此线性回归能够识别 β_1；反之，若 $\mathbb{E}(\varepsilon|D)=h(D)\neq 0$,

$$\mathbb{E}(Y|D)=\beta_0+\beta_1 D+h(D)$$

线性回归仍然能够得到 $\mathbb{E}(Y|D)$ 或其最佳线性近似，但是无法识别 β_1。

- 注意，

$$Y=\beta_0+\beta_1 D+h(D)+\tilde{\varepsilon},\tilde{\varepsilon}=\varepsilon-h(D)$$

$\mathbb{E}(\tilde{\varepsilon}|D)=0$ 自动成立，但是 $\mathbb{E}(\varepsilon|D)=0$ 是否成立却需要借助社会科学理论加以判断，后者就是区分模型 1 和模型 4 的识别假设，它是无法从数学上或统计上加以证明的，因为 ε 是未加观测或无法观测的。

- 在教育回报率的例子中，用 D 表示是否上过大学，Y 表示工资水平，由于 D 为二元变量，因此 $\mathbb{E}(Y|D)$ 必然可以表示为 $\mathbb{E}(Y|D)=\gamma_0+\gamma_1 D$，事实上

$$\begin{aligned}\gamma_0 &= \mathbb{E}(Y|D=0)\\ \gamma_1 &= \mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)\end{aligned}$$

由此得到线性回归模型

$$Y = \gamma_0 + \gamma_1 D + \tilde{\varepsilon}$$
$$\mathbb{E}(\tilde{\varepsilon}|D) = 0$$

但 γ_1 并不具有因果含义，它只表示总体中上过大学人群和没上过大学人群的平均工资差异。
回忆 CEF，我们也需要“假设”$E(e|X) = 0$，此时，我们通过构造的方法实现了 $\mathbb{E}(\tilde{\varepsilon}|D) = 0$，也就是 $E(e|X) = 0$。
线性回归模型试图回答的是如下的**预测性问题：“如果我们观测到 D 的取值为 D_0，我们预期 Y 的取值为何？**

- 相反，当我们写下结构模型

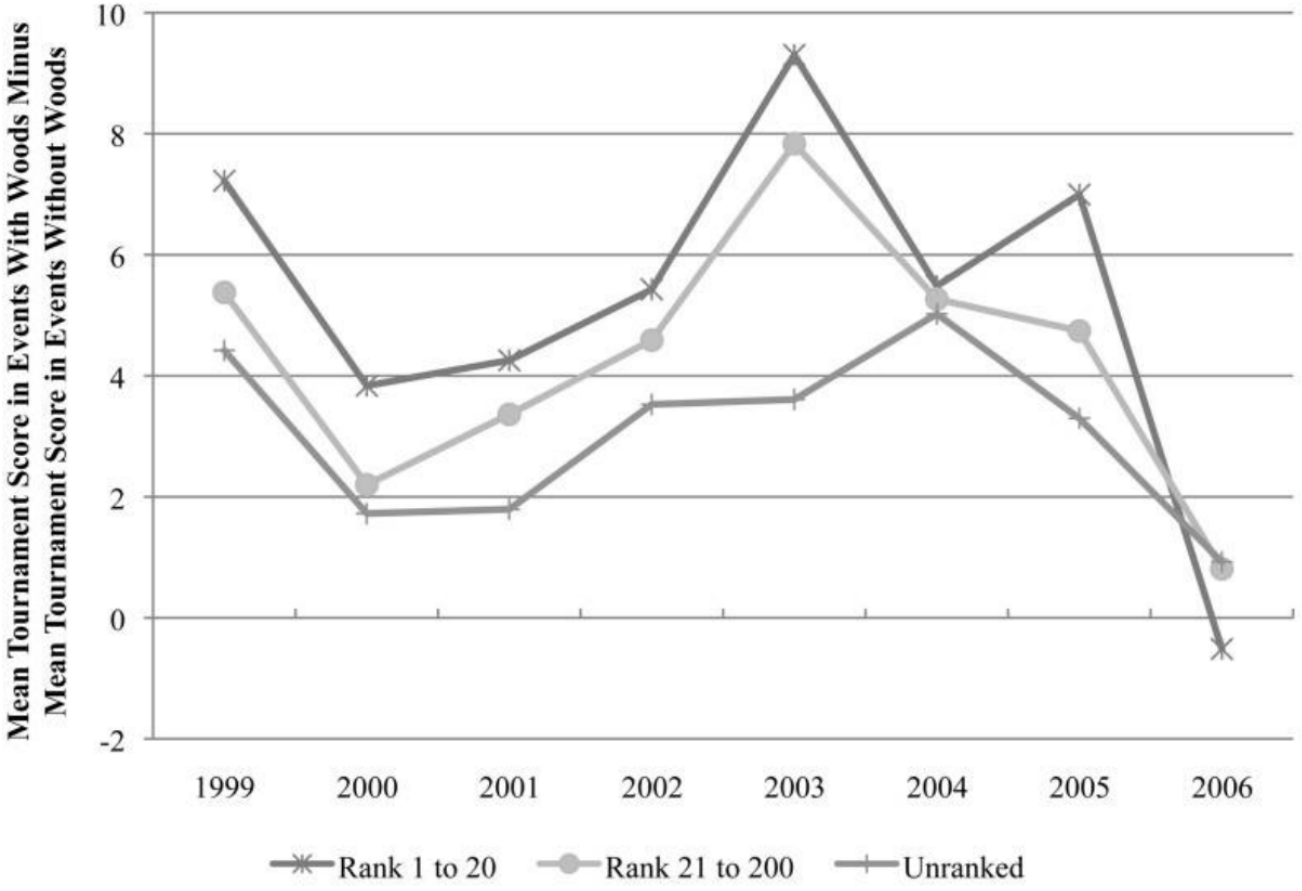
$$Y = \beta_0 + \beta_1 D + \varepsilon$$

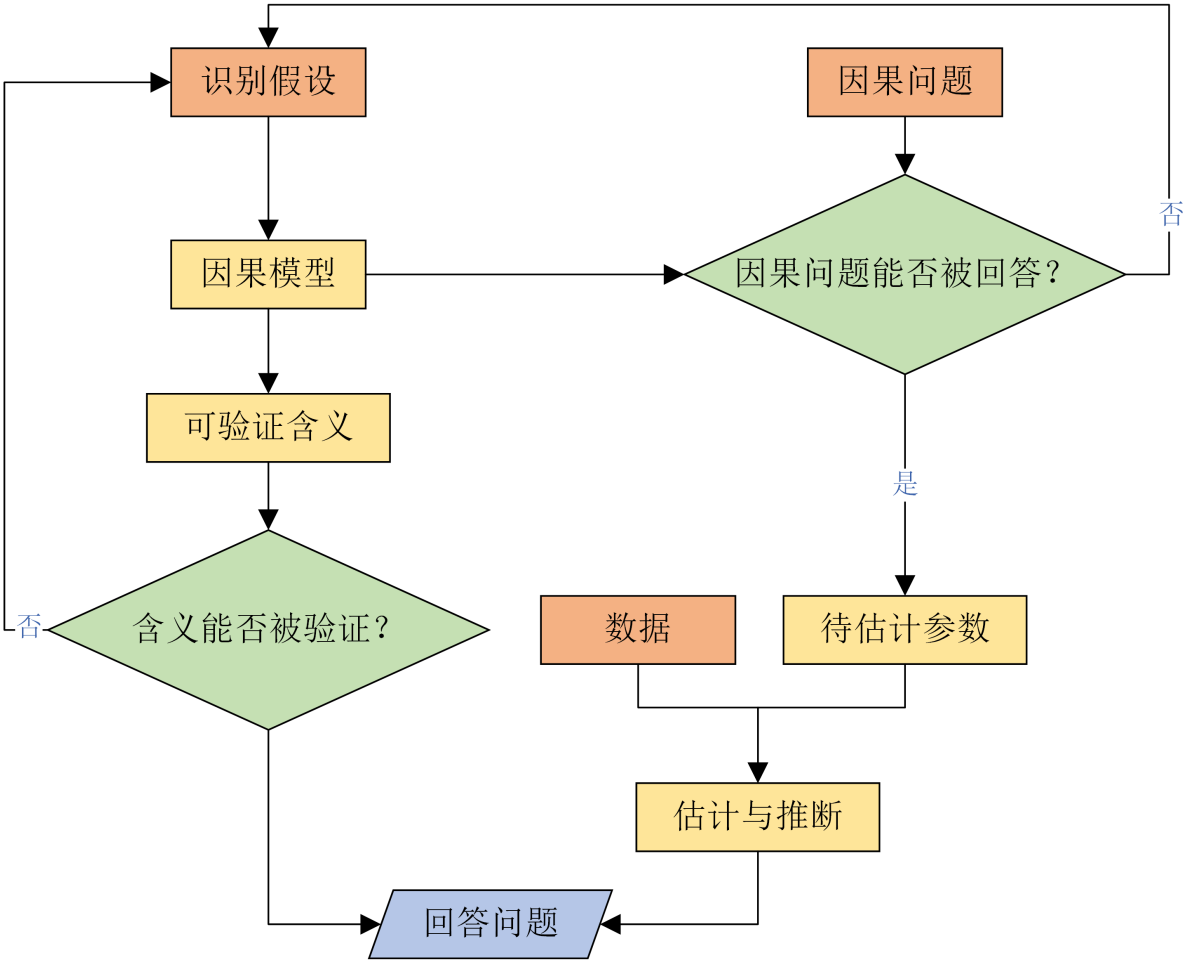
我们是把 ε 解释为“影响工资水平的不可观测的能力或积极性”，那么 ε 很有可能和 D 相关（能力越强的人越倾向于上过大学）。此时线性回归就无法识别因果效应 β_1。

$$\gamma_0 = \beta_0 + h(0)$$
$$\gamma_1 = \beta_1 + h(1) - h(0)$$

不管是 $Y = \beta_0 + \beta_1 D + \varepsilon$ 还是 $Y = \gamma_0 + \gamma_1 D + \tilde{\varepsilon}$，我们都只能用 OLS 去估计。得到的是 $\hat{\beta}_1$ 和 $\hat{\gamma}_1$，$\hat{\beta}_1 = \hat{\gamma}_1$。
但是因为对 ε 和 $\tilde{\varepsilon}$ 的理解不同，所以 $\mathbb{E}(\tilde{\varepsilon}|D) = 0$ 成立，而 $\mathbb{E}(\varepsilon|D) = 0$ 不成立。
结构模型试图回答的是如下的**因果性问题：如果我干预人们的上大学行为，将 D 的取值设定为 D_0**，我们预期 Y 的取值为何？

- 两种基本的识别策略：
  - 寻找特定的研究情境。不同的方法依赖于不同的识别假设，而不同的研究情境适用不同的识别假设 (make assumptions justifiable)。
  - 有时很难令人信服地论证识别假设的成立，此时尝试去挖掘因果模型更丰富的、可验证的相关性含义 (testable implications)，即提出这样的问题，“如果从 D 到 Y 的因果关系真的存在，那么我们还将观测到何种相关现象？”（异质性）


Mean Tournament Score in Events With Woods Minus Mean Tournament Score in Events Without Woods
10
8
6
4
2
0
-2
1999
2000
2001
2002
2003
2004
2005
2006
Rank 1 to 20
Rank 21 to 200
Unranked


识别假设
因果问题
因果模型
因果问题能否被回答？
否
可验证含义
是
含义能否被验证？
否
数据
待估计参数
估计与推断
回答问题

1.4 随机实验：因果推断的参照系

- 在一项随机实验中，研究者主动介入了数据生成过程，以确保只有模型 1 成为可能，因此随机实验是因果推断的理想情形和参照系，**所有的研究设计都致力于使得研究情境尽量接近于随机实验**。
- 研究者招募一批被试，将其随机划分为两组，对处理组个体实施处理，对控制组个体不实施处理。

$$D_i = \begin{cases} 1 & \text{对}i\text{实施处理（进入处理组、实验组）} \\ 0 & \text{不对}i\text{实施处理（进入控制组、对照组）} \end{cases}$$

- 尽管每位被试的 ε_i 各不相同，但随机分组保证了处理组个体和控制组个体的 ε 大体上保持平衡，因此两组个体结果的平均差异即反映因果效应。

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

$$\begin{aligned}\mathbb{E}(Y_i|D_i = 1) &= \beta_0 + \beta_1 + \mathbb{E}(\varepsilon_i|D_i = 1)\\ \mathbb{E}(Y_i|D_i = 0) &= \beta_0 + \mathbb{E}(\varepsilon_i|D_i = 0)\end{aligned}$$

若

$$\mathbb{E}(\varepsilon_i|D_i = 1) = \mathbb{E}(\varepsilon_i|D_i = 0) \tag{1}$$

则

$$\beta_1 = \mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0)$$

- (??) 式即为随机实验的识别假设，正式表述为
  Assumption LS 1: $\mathbb{E}(\varepsilon_i|D_i) = \mathbb{E}(\varepsilon_i)$
  称 ε_i 与 D_i **均值独立**，或
  Assumption LS 1': $\mathrm{Cov}(D_i, \varepsilon_i) = \mathbb{E}(D_i\varepsilon_i) = 0$
  称 ε_i 与 D_i 不相关。
- 一般而言，假设 1 比假设 1' 更强，当 D_i 为二元变量时，二者等价。但这一区别通常只具有数学上的意义，不具有社会科学上的意义。
- 回忆 β_1 的含义："保持 ε 不变，D 由 0 变到 1 所导致的变化"。这里存在一个因果识别的基本难题：一方面，必须在**干预性视角**下理解因果效应；但另一方面，理想的干预是不可行的。
- 假设 1 使得因果识别成为可能，此时蕴含着一种视角的转换——在干预性视角和**相关性视角**之间建立联系：控制组与处理组的比较，等价于对同一组个体施加干预与否的比较。

示例. 使用电脑有损学习成绩 (Carter et al, 2017, Economics of Education Review)

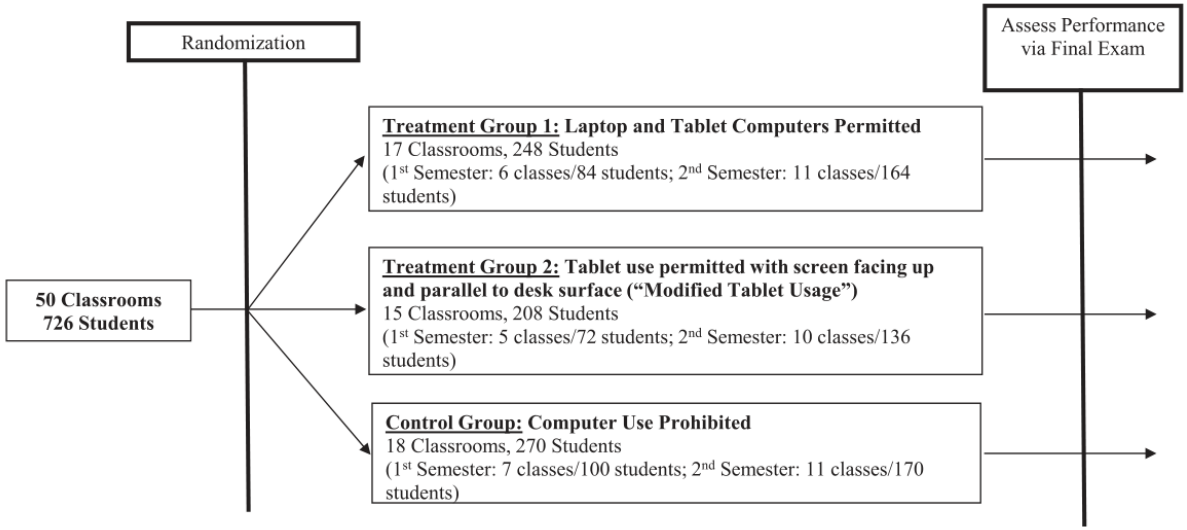


Fig. 1. Experimental design.

Table 2
Summary statistics and covariate balance.

	Control (1)	Treatment 1 (laptops/tablets) (2)	Treatment 2 (tablets, face up) (3)	Both treatments vs. control (4)	Treatment 1 vs. control (5)	Treatment 2 vs. control (6)
A. Baseline characteristics						
Female	0.17	0.20	0.19	0.03	0.06	0.00
				(0.03)	(0.04)	(0.04)
White	0.64	0.67	0.66	0.02	0.02	0.02
				(0.04)	(0.04)	(0.05)
Black	0.11	0.10	0.11	−0.02	−0.02	−0.03
				(0.03)	(0.03)	(0.04)
Hispanic	0.13	0.13	0.09	0.00	0.02	−0.03
				(0.03)	(0.03)	(0.03)
Age	20.12	20.15	20.15	0.03	0.05	0.06
	[1.06]	[1.00]	[0.96]	(0.08)	(0.09)	(0.10)
Prior military service	0.19	0.19	0.16	−0.02	0.00	−0.01
				(0.03)	(0.04)	(0.04)
Division I athlete	0.29	0.40	0.35	0.05	0.07*	0.04
				(0.04)	(0.04)	(0.05)
GPA at baseline	2.87	2.82	2.89	−0.01	−0.05	0.03
	[0.52]	[0.54]	[0.51]	(0.04)	(0.05)	(0.05)
Composite ACT	28.78	28.30	28.30	−0.34	−0.37	−0.54
	[3.21]	[3.46]	[3.27]	(0.26)	(0.31)	(0.33)
P-Val (Joint χ^2 Test)				0.610	0.532	0.361
B. Observed computer (laptop or tablet) use						
any computer use	0.00	0.81	0.39	0.62***	0.79***	0.40***
				(0.02)	(0.03)	(0.04)
Average computer use	0.00	0.57	0.22	0.42***	0.56***	0.24***
				(0.02)	(0.02)	(0.03)
Observations	270	248	208	726	518	478

Notes: Columns 1–3 of this table report mean characteristics of student in the control group (classrooms where laptops and tablets are prohibited), treatment group 1 (laptops and tablets permitted without restriction), and treatment group 2 (tablets are permitted if they are face up). Standard deviations are reported in brackets. Columns 4–6 report coefficient estimates from a regression of the baseline characterisitics on an indicator variable that equals one if a student is assigned to a classroom in the indicated treatment group. The regressions used to construct estimates in columns 4–6 include (instructor) x (semester) fixed effects and (class hour) x (semester) fixed effects. The reported *p*-values in Panel A are from a joint test of the null hypothesis that all coefficients are equal to zero. Observed computer usage, reported in panel B, was recorded during three lessons each semester of the experiment. Any computer use is an indicator variable for ever using a laptop or tablet during one of these three lessons. For example, a student who uses a computer during one of these three lessons has a value of one for any computer use and has an average usage rate of one-third. Robust standard errors are reported in parentheses. ***,**, and * denote significance at the 1%, 5%, and 10% level, respectively.

Table 3
Laptop and modified-tablet classrooms vs. non-computer classrooms.

	(1)	(2)	(3)	(4)
A. Dependent variable: Final exam multiple choice and short answer score				
Laptop/tablet class	−0.21***	−0.20***	−0.19***	−0.18***
	(0.08)	(0.07)	(0.06)	(0.06)
GPA at start of course			1.13***	1.00***
			(0.06)	(0.06)
Composite ACT				0.06***
				(0.01)
Demographic controls		X	X	X
R^2	0.05	0.24	0.52	0.54
Robust SE *P*-Val	0.010	0.005	0.001	0.002
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
B. Dependent variable: Final exam multiple choice score				
Laptop/tablet class	−0.18**	−0.17**	−0.16***	−0.15**
	(0.08)	(0.07)	(0.06)	(0.06)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.06	0.24	0.46	0.48
Robust SE *P*-Val	0.027	0.019	0.009	0.016
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
C. Dependent variable: Final exam short answer score				
Laptop/tablet class	−0.22***	−0.22***	−0.21***	−0.19***
	(0.08)	(0.07)	(0.06)	(0.06)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.06	0.18	0.42	0.43
Robust SE *P*-Val	0.007	0.004	0.001	0.002
Wild Bootstrap *P*-Val	0.006	0.016	0.000	0.008
D. Dependent variable: Final exam essay questions score				
Laptop/tablet class	0.02	0.02	0.03	0.03
	(0.07)	(0.06)	(0.06)	(0.06)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.33	0.38	0.50	0.51
Robust SE *P*-Val	0.785	0.766	0.642	0.548
Wild Bootstrap *P*-Val	0.757	0.775	0.627	0.509

Table 4
Unrestricted laptop/tablet classrooms vs. non-computer classrooms.

	(1)	(2)	(3)	(4)
A. Dependent variable: Final exam multiple choice and short answer score				
Computer class	−0.28***	−0.23***	−0.19***	−0.18***
	(0.10)	(0.09)	(0.07)	(0.07)
GPA at start of course			1.09***	0.92***
			(0.07)	(0.07)
Composite ACT				0.07***
				(0.01)
Demographic controls		X	X	X
R^2	0.08	0.28	0.54	0.57
Robust SE *P*-Val	0.003	0.007	0.005	0.005
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
B. Dependent variable: Final exam multiple choice score				
Computer class	−0.25***	−0.20**	−0.16**	−0.15**
	(0.10)	(0.009)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.08	0.27	0.48	0.50
Robust SE *P*-Val	0.009	0.023	0.025	0.029
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
C. Dependent variable: Final exam short answer score				
Computer class	−0.25***	−0.21**	−0.18**	−0.17**
	(0.09)	(0.09)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.08	0.21	0.44	0.46
Robust SE *P*-Val	0.008	0.016	0.017	0.019
Wild Bootstrap *P*-Val	0.008	0.020	0.022	0.028
D. Dependent variable: Final exam essay questions score				
Computer class	−0.03	−0.01	0.02	0.02
	(0.08)	(0.08)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.32	0.37	0.50	0.51
Robust SE *P*-Val	0.705	0.912	0.801	0.755
Wild Bootstrap *P*-Val	0.549	0.811	0.721	0.641

1.5 控制变量的作用

- 在非实验研究中，研究者是数据生成过程的被动观测者，因此影响 Y 的因素很可能同时影响 D，意味着 ε 和 D 相关。此时若要研究“保持 ε 不变，D 由 0 变到 1”，有两种思路：
  - 将 ε 中与 D 相关的因素剥离出来，使得剩余的 ε 和 D 不相关。
  - 考察非 ε 所带来的 D 的变化。

  这里我们先讨论前一种思路
- 设想一个非实验情境：某学校允许学生在课堂上自由使用电脑，研究者记录下学生实际是否使用电脑及其考试成绩。仍将使用电脑的学生归作处理组，不使用电脑的学生归作控制组。
- 在这一研究情境中，研究者无法假设 ε 和 D 不相关。例如 ε 中可能包含一个因素叫“学习习惯”。一方面，学习习惯不佳的学生考试成绩较差；另一方面，学习习惯不佳的学生更倾向于在课堂上使用电脑。因此处理组和控制组考试成绩的差异既有可能反映使用电脑的效应，也有可能反映学习习惯的效应。

- 考虑到是否按时出勤一定程度上能够反映学习习惯，因此采用出勤率 (X) 作为学习习惯的代理变量 (proxy variable)，则线性模型可以改写作

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i'', \mathbb{E}(\varepsilon_i'') = 0$$

$$\mathbb{E}(Y_i|D_i = 1, X_i = x) = \beta_0 + \beta_1 + \beta_2 x + \mathbb{E}(\varepsilon_i''|D_i = 1, X_i = x)$$
$$\mathbb{E}(Y_i|D_i = 0, X_i = x) = \beta_0 + \beta_2 x + \mathbb{E}(\varepsilon_i''|D_i = 0, X_i = x)$$

若

$$\mathbb{E}(\varepsilon_i''|D_i = 1, X_i = x) = \mathbb{E}(\varepsilon_i''|D_i = 0, X = x) \tag{2}$$

则

$$\beta_1(x) = \mathbb{E}(Y_i|D_i = 1, X_i = x) - \mathbb{E}(Y_i|D_i = 0, X_i = x)$$

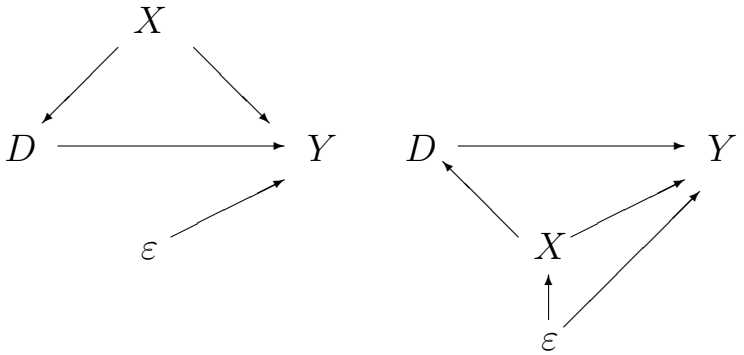
- (??) 式的直观含义是，原 ε 和 D 的相关性可以由它和 X 的相关性完全捕捉，在 X 相同的子总体内，新 ε'' 和 D 不再相关，新 ε'' 在处理组和控制组之间再次达到平衡——近似随机分组，因此可以将组间比较局限在 X 相同的子总体内以考察因果效应。[4]
- “把比较局限在 X 相同的子总体内”这个想法，我们经常简略地说成“给定 X”或“保持 X 不变”，英文的说法是”holding everything constant”或”other things being equal”，拉丁文

[4] 在不引起混淆的前提下，此后 ε'' 仍写作 ε.

的说法是"ceteris paribus"。称 X 为控制变量 (control variables) 或协变量 (covariates)。

- (??) 式可以正式表述为：
  Assumption LS 2: $\mathbb{E}(\varepsilon_i|D_i, X_i) = \mathbb{E}(\varepsilon_i|X_i)$
  称 ε_i 与 D_i **条件均值独立**，或
  Assumption LS 2': $\mathrm{Cov}(D_i, \varepsilon_i|X_i) = 0$
  称 ε_i 与 D_i **条件不相关**。
- 一般而言，假设 2 比假设 2' 更强，当 D_i 为二元变量时，两者等价。

- 注意，假设 2 并不要求 $\mathbb{E}(\varepsilon_i|X_i)=\mathbb{E}(\varepsilon_i)$ 或 $\mathrm{Cov}(X_i,\varepsilon_i)=0$，其区别见以下两图。



- 在左图中，$\mathrm{Cov}(D_i,\varepsilon_i)=\mathrm{Cov}(X_i,\varepsilon_i)=0$，这是传统计量教科书中的假设，但却是比假设 2 更强的假设，此时 D_i 和 X_i 都是外生的，$\hat{\beta}_1^{OLS}$ 和 $\hat{\beta}_2^{OLS}$ 能够分别反映使用电脑和出勤对考试成绩的因果效应，即 $\hat{\beta}_1^{OLS}\to_p\beta_1$ 且 $\hat{\beta}_2^{OLS}\to_p\beta_2$。但考察出勤变量的外生性给研究增加了额外的困难。因此，要么实证研究者在普遍地掩耳盗铃，要么这并不是实证研究者实际采用的假设。

- 在右图中，$\text{Cov}(D_i, \varepsilon_i) \neq 0$，但 $\text{Cov}(D_i, \varepsilon_i | X_i) = 0$，则 $\hat{\beta}_1^{OLS} \to \beta_1$ 成立，而 $\hat{\beta}_2^{OLS} \to_p \beta_2$ 不一定成立。

$$\varepsilon_i = \delta_0 + \delta_1 D_i + \delta_2 X_i + v_i$$
$$\text{Cov}(D_i, v_i) = \text{Cov}(X_i, v_i) = 0$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i \\ &= (\beta_0 + \delta_0) + (\beta_1 + \delta_1) D_i + (\beta_2 + \delta_2) X_i + v_i \\ \hat{\beta}_1^{OLS} &\to_p \beta_1 \Leftrightarrow \delta_1 = 0 \Leftrightarrow \text{Cov}(D_i, \varepsilon_i | X) = 0 \end{aligned}$$

而 δ_2 通常不为零，$\hat{\beta}_2^{OLS} \to_p \beta_2 + \delta_2 \neq \beta_2$

- 控制变量的双重作用：首先是控制 X 的直接效应 (β_2)，使得对回归系数的估计更准确（回归模型整体拟合程度更高，从而系数估计的标准误更小)；但更重要的是切断影响 Y 的其他因素（隐藏在扰动项中）与 D 的相关性，研究者希望，这个因素与 X 相关（δ_2）, 并且一旦控制 X 以后，这个因素不再与 D 相关。
- 因此 $\delta_2 \neq 0$ 是控制变量发挥作用的题中应有之义，其伴随的结果是无法准确识别控制变量的因果效应（还好我们并不关心)。
- 正确理解控制变量，要记住三句话：
  1. **一项因果推断研究待探究的原因往往只有一个，因此只能也只需处理某个特定 D 的内生性问题**。"XXX 的影响因素研究"不会是一项好的因果推断研究。
  2. 大部分影响 Y 的因素都被打包在 ε 中，**关键的控制变量一定是既影响 D 又影响 Y 的因素，只影响 Y 而不影响 D 的因素对于探究 D 对 Y 的因果关系往往并不重要**。
  3. **在研究中不要过度解读控制变量的系数估计结果**。

- 在示例（教育规模与教育产出）中，以学区为观测单位，Y 是学生的平均成绩，D 是班级的平均规模，X 是享受午餐补助的学生比例。ε 中包含学生的经济状况，它既影响 Y（反映课外学习机会的多寡），也影响 D（反映地区的财政实力）。如果控制 $\mathbf{X}$ 能够控制住经济状况与 D 的相关性，则可以一致地估计班级规模的因果效应。而 X 的系数很可能是负的，但这并不意味着取消午餐补助能够提高学生平均成绩，而是因为正的 β_2 被负的 δ_2 所抵消。

1.6 Potential outcomes framework

- 回答：因为我们关心 treatment(cause) 对 response(outcome) 的影响。在我们的心目中，(β_0, β_1) 应该反映的是 D 和 Y 之间的因果关系或结构性关系，而不仅仅是相关性。
- 我们需要一种定义 cause-and-effect relationships 的一般性框架：potential outcomes framework.
- 定义 treatment effect (causal effect)

$$
\begin{aligned}
Y_i^0 = &\ \text{Potential outcome when individual i} \\
&\ \text{receives control treatment 0 by intervention} \\
Y_i^1 = &\ \text{Potential outcome when individual i} \\
&\ \text{receives active treatment 1 by intervention}
\end{aligned}
$$

individual treatment effect: $Y_i^1 - Y_i^0$.

关于这一定义的几点评论：

- Causal effect 的定义只取决于 potential outcomes，而不取决于究竟哪个 potential outcome 被实际观察到（实际发生）。
- Causal effect 永远是在同一时点上的不同结果之间的比较，而且 treatment 必须发生在 outcome 之前，特别地，before-after comparison 不是 causal effect.
- 因果推断的基本难题就是数据缺失（missing data）：我们永远只能至多观察到一个 potential outcome. 没有观察到的那个 potential outcome 叫做 counterfactual.
- 因果分析的关键就是构造 counterfactual，也就是个 imputation 问题。而且 counterfactual 只能从 comparison group 个体中去寻找，也可以说因果分析的关键就是寻找好的对照组。
- 我们实际进行估计和推断所依赖的只能是 observed outcomes. 而且必须涉及多个个体或同一个体的不同时期，但无论是 cross-sectional comparison 还是 before-after comparison，都不是 causal effects 的定义，尽管它们对于估计和推断很重要。

- Observed treatment

$$D_i = \begin{cases} 1 & \text{treatment group} \\ 0 & \text{control group} \end{cases}$$

Observed outcome

$$Y_i = (1 - D_i) \cdot Y_i^0 + D_i \cdot Y_i^1$$

- 多个个体的存在本身并无助于解决因果推断的基本难题：treatment levels 和 potential outcomes 数量爆炸式增长。（考虑个体之间的相互影响。）
- 对于一个个体，其他个体是否受到处理，并不能影响处理对他的作用，以及他的行为。
- 想象一下：binary treatment, two individuals（A、B）, four treatment levels，four potential outcomes, only one observed outcome.
- four treatment levels：$(D_a = 0, D_b = 0), (D_a = 1, D_b = 0), (D_a = 0, D_b = 1), (D_a = 1, D_b = 1)$.
- four potential outcomes：$Y_a(D_a = 0, D_b = 0), Y_a(D_a = 1, D_b = 0), Y_a(D_a = 0, D_b = 1), Y_a(D_a = 1, D_b = 1)$. 所以我们才需要——

Stable unit treatment value assumption (SUTVA).

- 稳定个体处理假设。
- 每个个体的 potential outcome 不会因为其他个体接受的 treatment 的不同而不同；每个个体所接受到的每个 treatment 都只会产生一种 potenital outcome.
- 这代表两层含义：
  - No interference. 不存在同侪效应（peer effects），不存在一般均衡效应（general equilibrium effects）.
    - *一般均衡效应，例如：学校教育能够增加收入；但是如果大规模增加大学教育，会导致大学生的供给增加，可能会拉低大学生整体的工资水平。
  - No hidden variations in treatment. 实施 treatment 的方式不重要。(Consider Hawthorne effect [5]

).
  - 所有个体都以相同的剂量接受处理。
- 单有 SUVTA 还不够，我们还需要知道——
  识别假设

[5]霍桑效应（Hawthorne Effect）或称霍索恩效应，是指当人们知道自己成为观察对象，而会改变行为的倾向。起源于 1924 年至 1933 年间的一系列实验研究，由哈佛大学心理专家乔治·埃尔顿·梅奥（George Elton Mayo）教授为首的研究小组提出此概念，霍桑一词是美国西部电气公司座落在芝加哥的一间工厂的名称，是一座进行实验研究的工厂。实验最开始研究的是工作条件与生产效率之间的关系，包括外部环境影响条件（如照明强度、湿度）以及心理影响因素（如休息间隔、团队压力、工作时间、管理者的领导力）。(来自百度百科)

管理学发展：

- 古典管理理论的代表泰勒、法约尔等人着重强调管理的科学、合理、纪律；理论是基于这样一种假设，即社会是由一群群无组织的个人所组成的；他们在思想上、行动上力争获得个人利益，追求最大限度的经济收入，即“经济人”；管理部门面对的仅仅是单一的职工个体或个体的简单总和。工人被安排去从事固定的、枯燥的和过分简单的工作。
  - 泰勒制在使生产率大幅度提高的同时，也使工人的劳动变得异常紧张、单调和劳累，因而引起了工人们的强烈不满，并导致工人的怠工、罢工以及劳资关系日益紧张等事件的出现；另一方面，随着经济的发展和科学的进步，有着较高技术水平的工人逐渐占据了主导地位，体力劳动也逐渐让位于脑力劳动，也使得单纯使用古典管理理论已不合时宜。
- 照明实验：实验假设便是“提高照明度有助于减少疲劳，使生产效率提高”。可是经过两年多实验发现，照明度的改变对生产效率并无影响。具体结果是：当实验组照明度增大时，实验组和控制组都增产；当实验组照明度减弱时，两组依然都增产，甚至实验组的照明度减至 0.06 烛光时，其产量亦无明显下降；直至照明减至如月光一般、实在看不清时，产量才急剧降下来。
  - 实验初期，没有得到明确的结论。
- 梅奥接手实验以后：
  - 福利实验：查明福利待遇的变换与生产效率的关系。但经过两年多的实验发现，不管福利待遇如何改变（包括工资支付办法的改变、优惠措施的增减、休息时间的增减

等)，都不影响产量的持续上升。

*后经进一步的分析发现，导致生产效率上升的主要原因有：1、参加实验的光荣感。实验开始时 6 名参加实验的女工曾被召进部长办公室谈话，她们认为这是莫大的荣誉。这说明被重视的自豪感对人的积极性有明显的促进作用。2、成员间良好的相互关系。

–访谈实验: 研究者在工厂中开始了访谈计划。此计划的最初想法是要工人就管理当局的规划和政策、工头的态度和工作条件等问题作出回答，但这种规定好的访谈计划在进行过程中却大出意料之外，得到意想不到的效果。工人想就工作提纲以外的事情进行交谈，工人认为重要的事情并不是公司或调查者认为意义重大的那些事。访谈者了解到这一点，及时把访谈计划改为事先不规定内容，每次访谈的平均时间从三十分钟延长到 1-1.5 个小时，多听少说，详细记录工人的不满和意见。访谈计划持续了两年多。工人的产量大幅提高。

*工人们长期以来对工厂的各项管理制度和方法存在许多不满，无处发泄，访谈计划的实行恰恰为他们提供了发泄机会。发泄过后心情舒畅，士气提高，使产量得到提高。

–群体实验: 梅奥等人在这个试验中是选择 14 名男工人在单独的房间里从事绕线、焊接和检验工作。对这个班组实行特殊的工人计件工资制度。

实验者原来设想，实行这套奖励办法会使工人更加努力工作，以便得到更多的报酬。但观察的结果发现，产量只保持在中等水平上，每个工人的日产量平均都差不多，而

且工人并不如实地报告产量。深入的调查发现，这个班组为了维护他们群体的利益，自发地形成了一些规范。他们约定，谁也不能干的太多，突出自己；谁也不能干的太少，影响全组的产量，并且约法三章，不准向管理当局告密，如有人违反这些规定，轻则挖苦谩骂，重则拳打脚踢。进一步调查发现，工人们之所以维持中等水平的产量，是担心产量提高，管理当局会改变现行奖励制度，或裁减人员，使部分工人失业，或者会使干得慢的伙伴受到惩罚。

*这一试验表明，为了维护班组内部的团结，可以放弃物质利益的引诱。由此提出“非正式群体”的概念，认为在正式的组织中存在着自发形成的非正式群体，这种群体有自己的特殊的行为规范，对人的行为起著调节和控制作用。同时，加强了内部的协作关系。

第一类识别假设

- Assignment mechanism：每个个体如何接受 treatment.
- 显然，分配机制对于平均处理效应的识别至关重要。下面这个例子说明，不对分配机制进行正式建模，可能使得平均处理效应无法被正确识别。

考虑下面这个例子：

Unit	Potential Surgery	Outcomes Drug	Causal Effects S-D
Patient #1	7	1	6
Patient #2	5	6	-1
Patient #3	5	1	4
Patient #4	7	8	-1
Average Effect			2

Unit	Treatment	Observed Outcome
Patient #1	S	7
Patient #2	D	6
Patient #3	S	5
Patient #4	D	8
Average Difference		-1

- 因此，因果识别的关键假设往往就是关于分配机制的假设。
- 在潜在结果分析框架下，分配机制（倾向得分）可以表示为

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

比如在前面的例子中，

$$\pi = 1(Y^1 - Y^0 > 0)$$

- 第一类识别假设：分配机制不取决于潜在结果和可观测变量。

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

- 第一类识别假设（重新表述）：潜在结果均值独立于处理状态。

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

等价地

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

- 为什么主要研究 mean treatment effect? 因为 $E(\cdot)$ 是线性的。

$$E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

例如，$Median(Y^1 - Y^0) \neq Median(Y^1) - Median(Y^0)$，前者 > 0 表示至少一半人受益，后者 > 0 表示 median person 受益。目前关于 quantile treatment effect 的研究，研究对象基本都是后者，但前者 empirically more relevant.

- 三种“no effect”.

$$Y_i^1 = Y_i^0 \ \ \forall i \Longrightarrow F_{Y^1} = F_{Y^0} \Longrightarrow E(Y^1) = E(Y^0)$$

- Average treatment effect (ATE)

$$\tau \equiv E(Y_i^1 - Y_i^0)$$

- 处理组平均处理效应 (Average treatment effect on the treated, ATT).

$$\tau_1 \equiv E(Y_i^1 - Y_i^0 | D_i = 1)$$

- 控制组平均处理效应 (Average treatment effect on the untreated,ATU).

$$\tau_0 \equiv E(Y_i^1 - Y_i^0 | D_i = 0)$$

- 除非是随机分组，一般来说以上三者不相等
- 三者存在如下关系：

$$\tau = \tau_1 \cdot \Pr(D_i = 1) + \tau_0 \cdot \Pr(D_i = 0)$$

- 识别：此时**组间均值差异**能够直接识别 $ATE(\tau), ATT(\tau_1), ATU(\tau_0)$.

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

想要识别 τ_1，需要用到 $\mathbb{E}(Y^0|D) = \mathbb{E}(Y^0)$.
想要识别 τ_0，需要用到 $\mathbb{E}(Y^1|D) = \mathbb{E}(Y^1)$.
证明过程类似。因此有

$$\tau = \tau_1 = \tau_0$$

- 估计：样本组间均值差异是总体组间均值差异的一致估计。

$$\bar{Y}_{D=1} - \bar{Y}_{D=0} \to_p \mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0)$$

- 第一类识别假设就是随机实验（也称随机控制试验，randomized controlled trial, RCT）的识别假设。

观测性研究的挑战：选择性

- 绝大多数社会科学数据都是观测数据，D 部分地由 ε 决定，因此 ε 的分布因 D 而异，原因在于，绝大多数人类行动都是选择的结果而不是分配的结果，即**人们自选择**(self-select)**接受某项处理**。D 的选择性或内生性，是观测性研究的根本挑战。选择性分为两种：
  - **基于可观测变量的选择性**（selection on observables）。这是指，个体是否接受某项处理，只受到可观测变量的影响，给定这些可观测变量，接受处理与否可视作近似随机。这等价于是说，给定这些可观测变量，假设 2 成立。此时因果推断的关键，就是寻找这些造成选择性的可观测变量。
  - **基于不可观测变量的选择性**（selection on unobservables）。这是指，至少有一部分造成选择性的变量是不可观测的，因此无法"给定"这些变量，即假设 2 不成立，无法将组间比较局限在这些变量相同的子总体内以考察因果效应。这相当于我们常说的遗漏变量问题。

- 选择性就是分配机制 (assignment mechanism)：每个个体如何接受处理，可以用倾向得分 $\pi \triangleq Pr(D=1|X,\varepsilon)$ 来表示。
  - 基于可观测变量的选择性：倾向得分是可观测变量的未知函数。

$$Pr(D=1|X,\varepsilon)=\pi(X)$$

  - 基于不可观测变量的选择性：倾向得分是不可观测变量的未知函数。

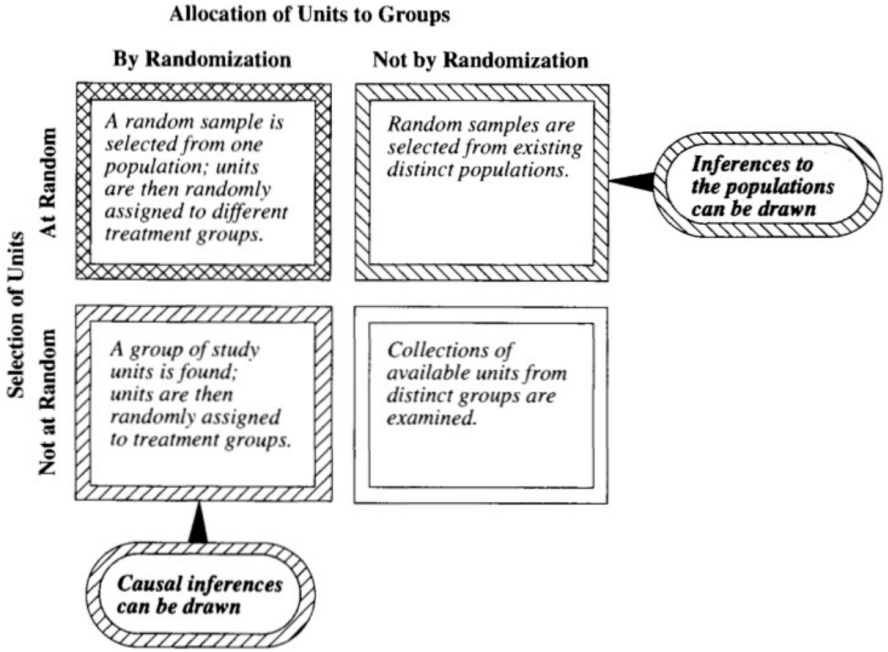
$$Pr(D=1|X,\varepsilon)=\pi(X,\varepsilon)$$

  - 若不存在选择性，则意味着倾向得分是常数，也即随机分配 (random assignment)。

$$Pr(D=1|X,\varepsilon)=\pi(\text{const.})$$

自选择 vs. 样本选择

- 我们通常假定所采用的样本来自随机抽样 (random sampling)，即总体中的每个个体都以相同概率进入样本，且抽取一个观测值不影响抽取其它观测值的概率，此时称样本中的每个观测值满足独立同分布 (independently and identically distributed)。
- 随机抽样意味着不存在样本选择 (sample selection)，例如收入调查中富人的应答率较低，或项目评估中处理组个体的非随机流失 (attrition)。
- 有时样本选择是由自选择引起的。例如在估计工资方程时，尽管我们所感兴趣的总体是所有工作年龄的劳动力，但只有实际参加工作的劳动力其工资才能被观测到，因此存在样本选择，其产生的原因正是劳动力自选择决定是否参加工作。这个问题应该被称作样本选择问题还是自选择问题？这并不重要。重点在于，我们在估计工资方程时必须正式处理样本的非随机特性。



在财经大学招募学生做试验是那种情形？

D 为连续变量的一般情形

- 此时 D 被称作连续处理 (continuous treatment)。而 β_1 的含义是，保持 ε 不变，当处理强度 (treatment intensity) 变化一个单位时，结果会变化 β_1 个单位。

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$
$$\mathbb{E}(Y_i|D_i) = \beta_0 + \beta_1 D_i + \mathbb{E}(\varepsilon_i|D_i)$$

若

$$\mathbb{E}(\varepsilon_i|D_i) = \mathbb{E}(\varepsilon_i) \equiv 0 \tag{3}$$

则

$$\beta_1 = \frac{d\mathbb{E}(Y_i|D_i)}{dD_i}$$

- 此时我们说，因果效应可以用“D_i 变化一个单位，Y_i 平均变化多少个单位”来衡量。请注意，这是一种衡量手段，而不是 β_1 的定义，因为这种衡量手段本质上依赖的是相关性，而当假设 (??) 式成立时，相关性可以揭示因果性。
- 此时的线性模型新增了一个限制性假设：边际效应不随着 D 的水平而变化，称这一假设为函数形式 (functional form) 假设或模型设定 (model specification) 假设。
- 对于含有控制变量的情形，β_1 的含义是，保持 X 和 ε 不变，当处理强度变化一个单位

时，结果会变化 β_1 个单位。

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i$$
$$\mathbb{E}(Y_i|D_i, X_i) = \beta_0 + \beta_1 D_i + \beta_2 X_i + \mathbb{E}(\varepsilon_i|D_i, X_i)$$

若

$$\mathbb{E}(\varepsilon_i|D_i, X_i) = \mathbb{E}(\varepsilon_i|X_i) = f(X_i) \tag{4}$$

其中 $f(X_i)$ 是（仅）关于 X_i 的未知函数。则

$$\beta_1 = \frac{\partial \mathbb{E}(Y_i|D_i, X_i)}{\partial D_i}$$

- 此时我们说，因果效应可以用“保持 X_i 不变，D_i 变化一个单位，Y_i 平均变化多少个单位”来衡量。
- 此时的线性模型施加了更强的函数形式假设：边际效应不随着 D 或 X 的水平而变化。

Causal estimand 被估计量

- estimand：待估计的量；
- estimate：估计结果；
- estimator：估计量。
- 例如：对于一个正态分布，有两个参数，均值 μ 和方差 σ^2。

$$s^2 = \sum_{i=1}^{n}(x_i - \hat{x})^2/(n-1)$$

是 σ^2 的 estimator，σ^2 是 estimand. 对于数据集 $x = 2, 3, 7$, 计算得到 $s^2 = 7$, 7 就是 estimate.

重新理解线性结构模型

线性结构模型：

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \mathbb{E}(\varepsilon_i) = 0$$

- 在假设 ID.1 下，样本组间均值差异是 τ 的一致估计；在假设 LS.1 下，样本组间均值差异是线性结构模型的结构参数 β_1 的一致估计。一个自然的问题是：τ 和 β_1 是什么关系？以及，假设 ID.1 和假设 LS.1 是什么关系？
- 定义潜在结果：

$$\begin{aligned} Y_i^1 &= \mathbb{E}(Y_i^1) + U_i^1 \triangleq \alpha_1 + U_i^1 \\ Y_i^0 &= \mathbb{E}(Y_i^0) + U_i^0 \triangleq \alpha_0 + U_i^0 \end{aligned} \tag{5}$$

- 线性结构模型的结构参数 β_1 等价于潜在结果框架下定义的平均处理效应 τ。

$$\begin{aligned} Y_i &= (1-D_i)Y_i^0 + D_i Y_i^1 \\ &= (1-D_i)(\alpha_0 + U_i^0) + D_i(\alpha_1 + U_i^1) \\ &= \alpha_0 + \underbrace{(\alpha_1 - \alpha_0)}_{ATE} D_i + [D_i U_i^1 + (1-D_i)U_i^0] \end{aligned}$$

- 并且线性结构模型并没有 $\beta_{1i} \equiv \beta_1$ 这样的限制性假设，允许处理效应存在异质性，β_1 衡量的是其平均值。
- 我们还能方便地证明，假设 ID.1 和假设 LS.1 是等价的。

证明：假设 ID.1 与假设 LS.1 等价。

假设 LS.1 要求

$$\text{Cov}(D, DU^1 + (1-D)U^0) = 0$$

$$\begin{aligned}
&LHS \\
=&\mathbb{E}(DU^1) - \mathbb{E}(D)\mathbb{E}(DU^1 + (1-D)U^0) \\
=&(1-\mathbb{E}(D))\mathbb{E}(DU^1) - \mathbb{E}(D)\mathbb{E}((1-D)U^0) \\
=&(1-\mathbb{E}(D))\mathbb{E}(U^1|D=1)P(D=1) - \mathbb{E}(D)\mathbb{E}(U^0|D=0)P(D=0) \\
=&\mathbb{E}(D)(1-\mathbb{E}(D))[\mathbb{E}(U^1|D=1) - \mathbb{E}(U^0|D=0)]
\end{aligned}$$

若假设 ID.1 成立，

$$\mathbb{E}(Y^d|D) = \mathbb{E}(Y^d)$$

则

$$\mathbb{E}(U^d|D) = 0$$

$$\mathbb{E}(U^1|D=1) - \mathbb{E}(U^0|D=0) = 0$$

Identifying treatment effects.

- 若 potential outcomes 均值独立于 treatment, 则 group-mean difference 能够识别 average treatment effect.

$$
\begin{aligned}
& E(Y^d|D) = E(Y^d) \Longrightarrow \\
& E(Y|D=1) - E(Y|D=0) \\
= \; & E(Y^1|D=1) - E(Y^0|D=0) \\
= \; & E(Y^1) - E(Y^0) \\
= \; & E(Y^1 - Y^0)
\end{aligned}
$$

- 对于 binary treatment，均值独立等价于不相关。

$$
\begin{aligned}
Cov(Y^d, D) = 0 \; & \Longleftrightarrow E(Y^d D) = E(Y^d)E(D) \\
& \Longleftrightarrow E(Y^d|D=1)P(D=1) = E(Y^d)E(D) \\
& \Longleftrightarrow E(Y^d|D=1) = E(Y^d)
\end{aligned}
$$

- 类似地，若想要识别 conditional treatment effect $E(Y^1-Y^0|X)$，关键的假设是 selection on observables:

$$E(Y^d|D,X) = E(Y^d|X)$$

进一步地，

$$\begin{aligned} E(Y^1 - Y^0) &= E\{E(Y^1 - Y^0|X)\} \\ &= \int E(Y^1 - Y^0|X = x)dF_X(x) \end{aligned}$$

selection on unobservables 则是指

$$E(Y^d|D,X,\varepsilon) = E(Y^d|X,\varepsilon)$$

- 定义在协变量上的条件平均处理效应 (Conditional ATE, CATE)

$$\tau(x) \triangleq \mathbb{E}(Y_i^1 - Y_i^0 | X_i = x)$$
$$\tau_1(x) \triangleq \mathbb{E}(Y_i^1 - Y_i^0 | X_i = x, D_i = 1)$$
$$\tau_0(x) \triangleq \mathbb{E}(Y_i^1 - Y_i^0 | X_i = x, D_0 = 0)$$

根据期望迭代定律，

$$\tau = \mathbb{E}\{\mathbb{E}(Y_i^1 - Y_i^0 | X_i = x)\} = \int \tau(x) dF_X$$
$$\tau_1 = \mathbb{E}\{\mathbb{E}(Y_i^1 - Y_i^0 | X_i = x, D_i = 1)\} = \int \tau_1(x) dF_{X|D=1}$$
$$\tau_0 = \mathbb{E}\{\mathbb{E}(Y_i^1 - Y_i^0 | X_i = x, D_i = 0)\} = \int \tau_0(x) dF_{X|D=0}$$

- randomization 确保了均值独立。若均值独立假设不成立，则 group-mean difference 存在 selction bias.

$$\begin{aligned} & E(Y|D=1) - E(Y|D=0) \\ = & \underbrace{E(Y^1|D=1) - E(Y^0|D=1)}_{\text{average treatment effect on the treated}} + \underbrace{E(Y^0|D=1) - E(Y^0|D=0)}_{\text{selection bias}} \\ = & E(Y^1) - E(Y^0) \end{aligned}$$

由此也可看出，想要识别 ATT，需要假定 $E(Y^0|d) = E(Y^0)$. 想要识别 ATU，则需要假定 $E(Y^1|D) = E(Y^1)$.

-

$$\begin{aligned} E(Y^1|D=1) = E(Y^1|D=0) & \qquad = E(Y^1) \\ E(Y^0|D=1) = E(Y^0|D=0) & \qquad = E(Y^0) \end{aligned}$$

• 假定 potential outcomes 是线性的，可以帮助我们更好地理解识别条件。

$$Y_i^d = \alpha_d + X_i'\beta_d + U_i^d, d = 0, 1$$

$$E(Y^1 - Y^0) = \alpha_1 - \alpha_0 + E(X)'(\beta_1 - \beta_0)$$

$$\begin{aligned}
& E(Y|D=1) - E(Y|D=0) \\
= \; & \alpha_1 - \alpha_0 + E(X|D=1)'\beta_1 - E(X|D=0)'\beta_0 \\
& + E(U^1|D=1) - E(U^0|D=0) \\
= \; & \underbrace{\alpha_1 - \alpha_0 + E(X)'(\beta_1 - \beta_0)}_{\text{average treatment effect}} \\
& + \underbrace{[E(X|D=1) - E(X)]'\beta_1 - [E(X|D=0) - E(X)]'\beta_0}_{\text{selection bias due to observables}} \\
& + \underbrace{E(U^1|D=1) - E(U^0|D=0)}_{\text{selection bias due to unobservables}}
\end{aligned}$$

• 类似地，

$$\begin{aligned} & E(Y|D=1,X)-E(Y|D=0,X) \\ = \ & \alpha_1-\alpha_0+X'(\beta_1-\beta_0) \\ & +E(U^1|D=1,X)-E(U^0|D=0,X) \end{aligned}$$

请注意，在此模型中，

$$E(U^d|D,X)=E(U^d|X) \Longleftrightarrow E(Y^d|D,X)=E(Y^d|X)$$

- 反过来，当我们写下一个线性模型时，从 potential outcomes framework 的角度看，意味着什么?

$$Y_i = \alpha + \beta D_i + U_i$$

令

$$Y_i^d = E(Y_i^d) + U_i^d \equiv \alpha_d + U_i^d, d = 0, 1$$

ATE

$$E(Y_i^1 - Y_i^0) = \alpha_1 - \alpha_0$$

$$Y_i = \alpha_0 + (\alpha_1 - \alpha_0)D_i + [D_i U_i^1 + (1 - D_i)U_i^0]$$

- OLS 估计的一致性要求

$$Cov(D, DU^1 + (1-D)U^0) = 0$$

即

$$E(D)(1-E(D))\left[E(U^1|D=1) - E(U^0|D=0)\right] = 0$$

$$\hat{\beta} \to_p (\alpha_1 - \alpha_0) + \left[E(U^1|D=1) - E(U^0|D=0)\right]$$

no selection bias **和线性回归模型的扰动项外生性假设是一回事。**

- 劳德悖论 (Lord's Paradox): 识别假设的重要性。
  - 研究儿童助学项目 (Head Start Program) 的效果。处理组和控制组都进行了项目开始前的预试和项目结束后的终试。处理组多是贫困 (low socioeconomic status, low-SES) 儿童，因此预试（X）和终试（Y）的成绩都低于控制组。
  - 在劳德的例子中，对于处理组和控制组，都有 $\bar{X} = \bar{Y}$. 两个研究者就此得出了截然不同的结论。
  - 一号研究者:

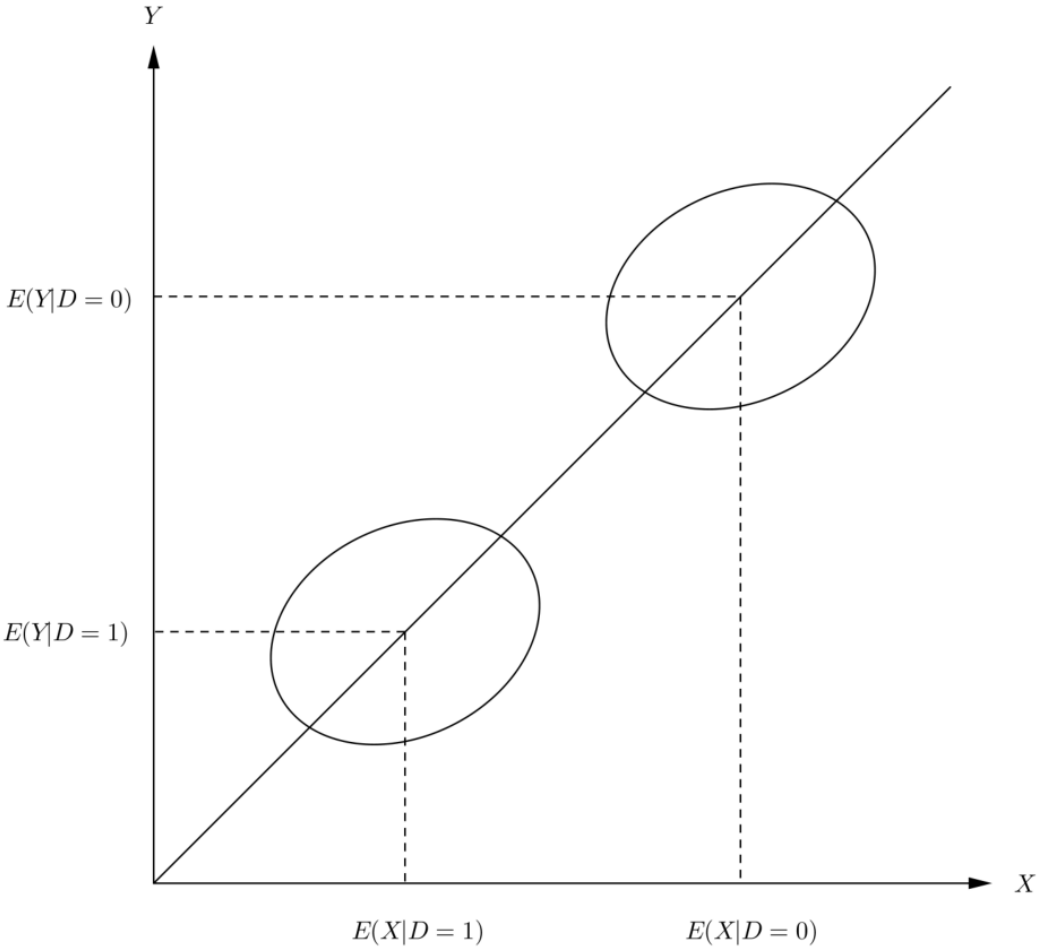
$$Y - X = \alpha + \beta D + \varepsilon$$

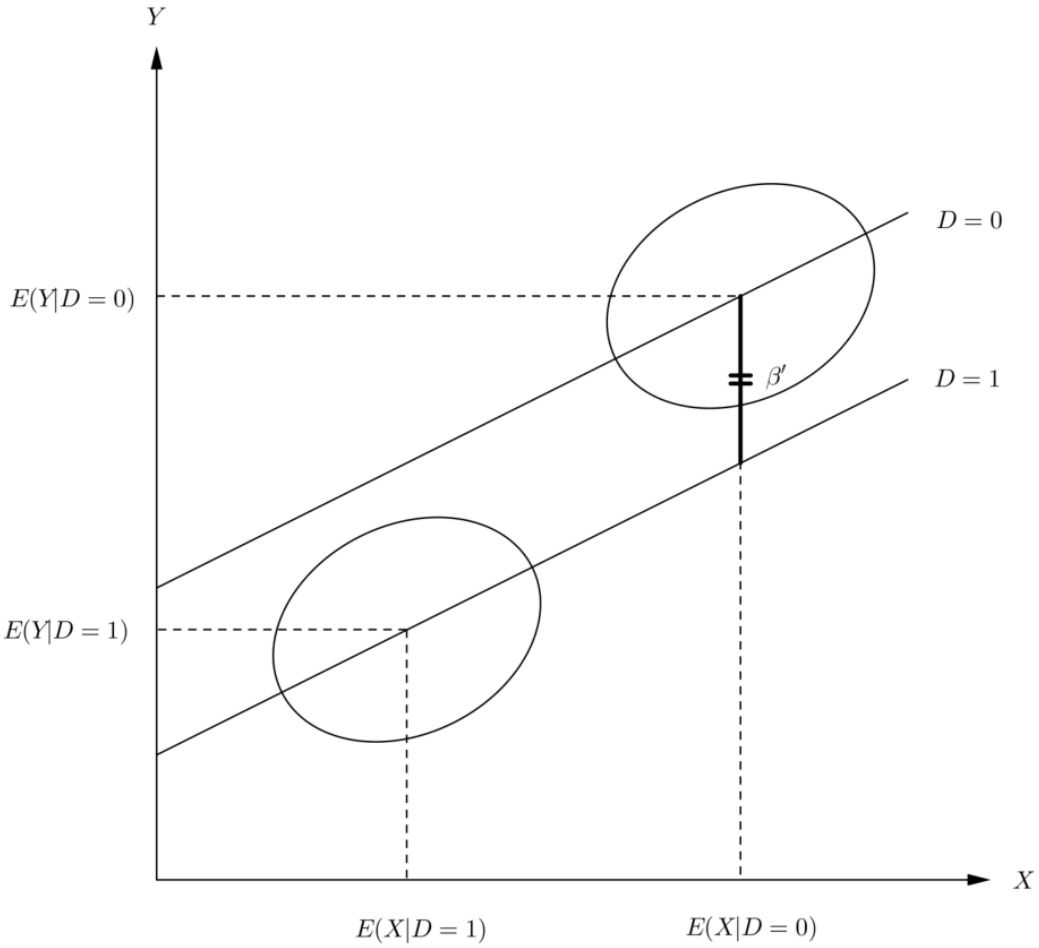
$$\beta = (E(Y|D=1) - E(X|D=1)) - (E(Y|D=0) - E(X|D=0))$$

  - 二号研究者:

$$Y = \alpha' + \gamma X + \beta' D + \varepsilon'$$

$$\beta' = (E(Y|D=1) - E(Y|D=0)) - \gamma(E(X|D=1) - E(X|D=0))$$


Y
E(Y|D = 0)
E(Y|D = 1)
X
E(X|D = 1)
E(X|D = 0)


Y
D = 0
E(Y|D = 0)
β′
D = 1
E(Y|D = 1)
X
E(X|D = 1)
E(X|D = 0)

- 真正的 ATT:

$$E(y = Y^1 - Y^0|D = 1)$$

- 一号研究者的隐含假设:

$$E(Y^0|D = 1) - E(X|D = 1) = E(Y^0|D = 0) - E(X|D = 0)$$

假若处理组儿童未接受处理，则其成绩的变化与控制组儿童相同。

- 二号研究者的隐含假设:

$$E(Y^0|D = 1) - \gamma E(X|D = 1) = E(Y^0|D = 0) - \gamma E(X|D = 0)$$

也即

$$E(Y^0|D = 1, X = x) = E(Y^0|D = 0, X = x) = \alpha' + \gamma x$$

(1) 假如未接受处理效应，则处理组儿童与预试成绩相同的控制组儿童的终试成绩相同。

(2) 处理组儿童未接受处理时的（反事实）终试成绩和预试成绩之间的线性关系，与控制组儿童的终试成绩和预试成绩之间的线性关系（由 γ 表示）相同。

- 评价一号研究者的假设：处理组多是 low-SES 儿童，他们在没有帮助的情况下所取得的进步会少于 high-SES 儿童，因此会低估 ATT.
- 评价二号研究者的假设：预试成绩相同的处理组和控制组儿童并不是在所有其他方面也都是可比的。

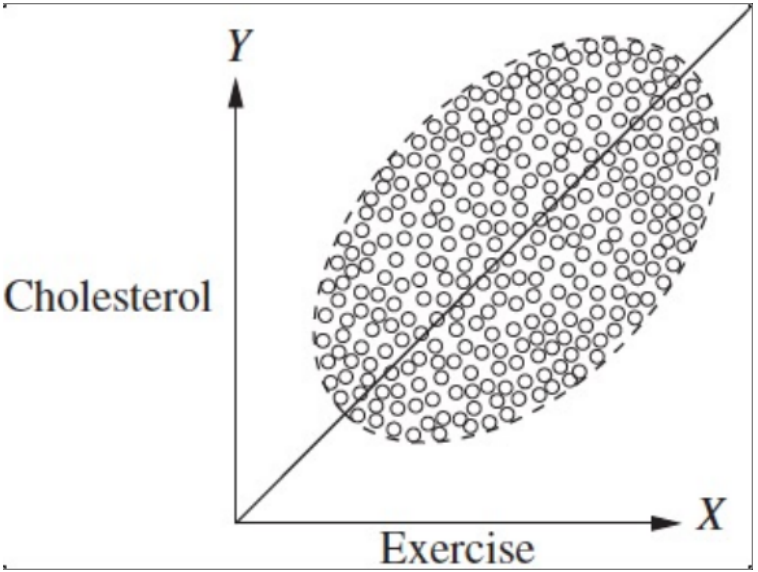
（1）正由于处理组儿童家庭背景较差，他们相较于预试成绩相同的控制组儿童可能拥有更好的学习能力和态度。因此会高估 ATT.

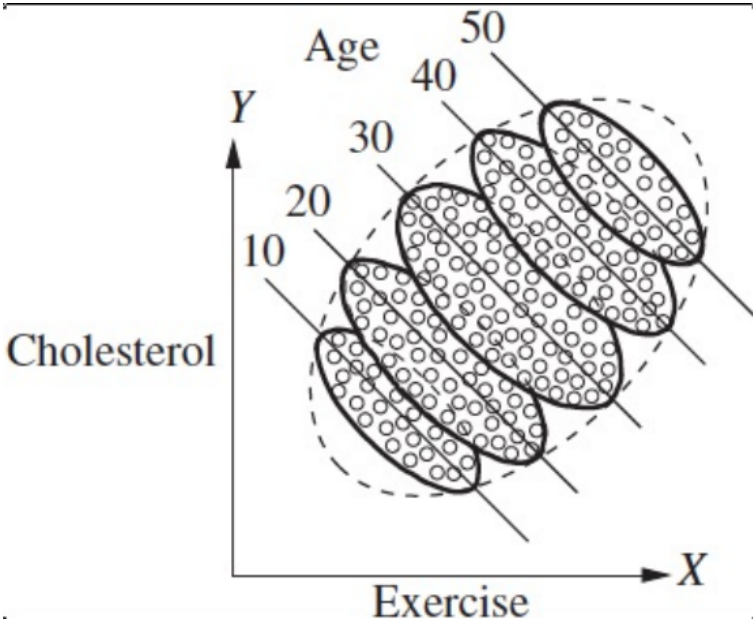
（2）处理组儿童可能会持续遇到的各种生活窘迫，会压低其在未接受处理时的成长曲线。因此会低估 ATT。

- 显然，对假设要求最弱的做法就是随机挑选儿童进入处理组。（此时仍然可以控制预试成绩，既为提高估计精度，又为检验可能的异质性效应，例如是否预试成绩更高的儿童从项目中受益也更多。）

• 辛普森悖论 (Simpson's Paradox): 部分的效应与总体的效应可能出现背离。

$$Sign(E(Y|D=1) - E(Y|D=0)) \neq Sign(E(Y|X, D=1) - E(Y|X, D=0))$$




Age
50
40
30
20
10
Y
Cholesterol
X
Exercise

	Drug($D=1$)	No drug($D=0$)
Men($X=0$)	81/87 (93%)	234/270(87%)
Women($X=1$)	192/263 (73%)	55/80(69%)
Combined data	273/350 (78%)	289/350(83%)

$$
\begin{aligned}
P(X=1) &= \tfrac{263+80}{700} = 49\% \quad \text{女性比例} \\
P(X=0) &= \tfrac{87+270}{700} = 51\% \quad \text{男性比例} \\
P(X=1|D=1) &= \tfrac{263}{350} = 75\% \quad \text{用药的病人中女性比例} \\
P(X=1|D=0) &= \tfrac{80}{350} = 23\% \quad \text{没用药的病人中女性比例}
\end{aligned}
$$

$$
\begin{aligned}
\tau(X=1) &= E(Y|X=1,D=1) - E(Y|X=1,D=0) \\
&= 73\% - 69\% = 4\% \\
\tau(X=0) &= E(Y|X=0,D=1) - E(Y|X=0,D=0) \\
&93\% - 87\% = 6\%
\end{aligned}
$$

$$
\begin{aligned}
& E(Y|D=1) \\
= \; & E(Y|X=1,D=1)P(X=1|D=1) \\
& +E(Y|X=0,D=1)P(X=0|D=1) \\
= \; & 73\% \times 75\% + 93\% \times 25\% \\
& E(Y|D=0) \\
= \; & E(Y|X=1,D=0)P(X=1|D=0) \\
& +E(Y|X=0,D=0)P(X=0|D=0) \\
= \; & 69\% \times 23\% + 87\% \times 77\% \\
\Longrightarrow \; & E(Y|D=1) - E(Y|D=0) = -5\%
\end{aligned}
$$

- 如果知道患者性别，应该开药；如果不知道患者性别，则不应该开药?！这个结论显然是荒唐的——如果该要对男性和女性都有效，则应该对任何人都有效。
- 直观解释：无论是否使用药物，女性的治愈率都较低；女性使用药物的几率较高。因此，药物在总体上似乎无效的原因在于：当我们随机挑选一位药物使用者时，该对象为女性的几率较高，因此平均而言治愈率较低。换言之，负向的女性效应抵消了正向的用药效应。

•ATE:

$$\begin{aligned}\tau &= E(E(Y|X,D=1)-E(Y|X,D=0))\\ &= \tau(X=1)P(X=1)+\tau(X=0)P(X=0)\\ &= 4\%\times 49\%+6\%\times 51\%\end{aligned}$$

可以看出，如果 $P(X|D)=P(X)$ 就不会有悖论了！

•ATT:

$$\begin{aligned}\tau_1 &= E(E(Y|X,D=1)-E(Y|X,D=0)|D=1)\\ &= \tau(X=1)P(X=1|D=1)+\tau(X=0)P(X=0|D=1)\\ &= 4\%\times 75\%+6\%\times 25\%=4.5\%\end{aligned}$$

•ATU:

$$\begin{aligned}\tau_0 &= E(E(Y|X,D=1)-E(Y|X,D=0)|D=0)\\ &= \tau(X=1)P(X=1|D=0)+\tau(X=0)P(X=0|D=0)\\ &= 4\%\times 23\%+6\%\times 75\%=5.5\%\end{aligned}$$

•三者的关系：

$$\tau=\tau_1\cdot P(D=1)+\tau_0\cdot P(D=0)$$

• 从线性模型角度来思考这一问题：

$$E(Y|X, D = d) = \alpha_d + \beta_d X$$

可以分组回归或使用交互项模型

$$E(Y|D, X) = \alpha_0 + (\alpha_1 - \alpha_0)D + \beta_0 X + (\beta_1 - \beta_0)D \cdot X$$

$$\begin{cases} E(Y|X = 1, D = 1) & = \alpha_1 + \beta_1 & = 73\% \\ E(Y|X = 0, D = 1) & = \alpha_1 & = 93\% \\ E(Y|X = 1, D = 0) & = \alpha_0 + \beta_0 & = 69\% \\ E(Y|X = 0, D = 0) & = \alpha_0 & = 87\% \end{cases}$$

$$
\begin{aligned}
\text{处理组的前后差异 } \tau_1 &= \frac{1}{N_T}\sum_{t\in T}[(\alpha_1-\alpha_0)+(\beta_1-\beta_0)X_t] \\
\text{控制组的前后差异 } \tau_0 &= \frac{1}{N_C}\sum_{c\in C}[(\alpha_1-\alpha_0)+(\beta_1-\beta_0)X_c] \\
\text{处理效应 } \tau &= \frac{1}{N}\sum_{i=1}^{N}[(\alpha_1-\alpha_0)+(\beta_1-\beta_0)X_i] \\
&= \frac{1}{N}\sum_{i=1}^{N}[\alpha_1+\beta_1 X_i]-\frac{1}{N}\sum_{i=1}^{N}[\alpha_0+\beta_0 X_i]
\end{aligned}
$$

- 辛普森悖论相当于犯了一个什么错误?

$$
\tau \neq \frac{1}{N_T}\sum_{t\in T}[\alpha_1+\beta_1 X_t]-\frac{1}{N_C}\sum_{c\in C}[\alpha_0+\beta_0 X_c]
$$

这相当于没有控制 X:

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

这也就是为什么在跑回归的时候要尽可能多地放入控制变量。

- Covariates (attributes, pretreatment variables) X 的最主要特征是不受 treatment assignment 影响。其重要性体现在:

–通过 control 一部分 outcome variation 使得估计更加精确。
–我们可能对由 covariate 定义的 subpopulation 感兴趣。
–会影响 assignment mechanism.

- 由倾向得分（propensity score）来定义 assignment mechanism.

$$\pi(D_i = 1|Y_i^1, Y_i^0, X_i)$$

- Overlap assumption

$$0 < \pi(D_i = 1|Y_i^1, Y_i^0, X) < 1$$

- Unconfoundedness (strong ignorability) assumption.

$$\pi(D_i = 1|Y_i^1, Y_i^0, X_i) = \pi(D_i = 1|X_i)$$

等价地，

$$D_i \perp (Y_i^1, Y_i^0)|X_i$$

- 随机试验与 observational studies 的区别在于前者的 assignment mechanism 的函数形式已知。
- 经典随机试验满足以上两个假设。Imbens and Rubin (2015) 把满足这两个假设的 observational studies 称作 regular observational studies.

- 什么是 irregular observational studies? 三种主要情形:
  - Unconfounded given latent variables that are only partially observed.
    $\Longrightarrow$ instrumental variables.
  - Unconfounded given pre data when pre and post data for both the treatment and control groups are available.
    $\Longrightarrow$ difference in difference
  - Violation of overlap when assignment to treatment is determined by a threshold eligibility rule.
    $\Longrightarrow$ regression discontinuity

这里是三个情形需要再理解一下。上次没讲清楚。

- Randomized experiments: 研究对象的 assignment mechanism 受研究者控制并已知。
- Regular observational mechanisms: 研究对象的 assignment mechanism 未知但满足 observational study。
- Irregular observational mechanisms.

1.7 随机推断（randomization inference）

- 关于随机实验的统计推断，Athey and Imbens (2017, Handbook) 建议采用基于随机化的方法，而非基于抽样的传统方法。
  - 基于抽样的方法假定处理的分配是固定的，而结果是随机的，统计推断基于这样的想法：被试样本是来自于总体的随机样本。
  - 基于随机化的方法假定被试的潜在结果是固定的，而处理的分配是随机的。在这种视角下，样本就是感兴趣的总体本身。此时平均处理效应的定义是所有被试接受处理的平均结果和所有被试不接受处理的平均结果的差异。但我们无法确定性地得到这个平均处理效应，此时不确定性的来源是，对于每个被试，只观测到两个潜在结果中的一个。

费雪的准确检验（Fisher's exact test）.

- 检验精确的原假说（sharp null hypothesis）

$$H_0 : Y_i^1 = Y_i^0, \forall i \quad \text{（注意：对于每一个个体。）}$$

- 当原假说为真时，我们能够知晓所有的潜在结果。

$$Y_i^0 = Y_i^1 = Y_i, \forall i$$

从而可以知道任何统计量 $S(D, Y^{obs}, X)$ 的准确分布（其中 D 为处理状态向量）。

- 计算准确 p 值 (exact p-value).

$$
\begin{aligned}
p &= \Pr\left(\left|S\left(\mathrm{D}, \mathrm{Y}^{obs}, \mathrm{X}\right)\right| > \left|S\left(\mathrm{D}^{obs}, \mathrm{Y}^{obs}, \mathrm{X}\right)\right|\right) \\
&= \frac{\sum_{\tilde{\mathrm{D}} \in \Omega_{\mathrm{D}}} 1\left\{\left|S\left(\tilde{\mathrm{D}}, \mathrm{Y}^{obs}(\tilde{\mathrm{D}}), \mathrm{X}\right)\right| > \left|S\left(\mathrm{D}^{obs}, \mathrm{Y}^{obs}, \mathrm{X}\right)\right|\right\}}{|\Omega_{\mathrm{D}}|}
\end{aligned}
$$

- 检验统计量举例

  –均值差异

$$
S_{\text{avg}} = \bar{Y}_t - \bar{Y}_c
$$

  –分位数差异

$$
\begin{aligned}
S_{\text{med}} &= \text{med}\left(Y_t\right) - \text{med}\left(Y_c\right) \\
S_q &= q_\delta\left(Y_t\right) - q_\delta\left(Y_c\right)
\end{aligned}
$$

  ––t 统计量 (help ttest)

$$
S_{t-\text{ stat}} = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}
$$

其中

$$s_t^2 = \frac{1}{n_t - 1} \sum_{i:D_i=1} \left(Y_i - \bar{Y}_t\right)^2, s_c^2 = \frac{1}{n_c - 1} \sum_{i:D_i=0} \left(Y_i - \bar{Y}_c\right)^2$$

–秩统计量 (rank statistic), 将所有观测到的结果按升序排列。

$$S_{\text{rank}} = \bar{R}_t - \bar{R}_c$$

–基于回归的统计量

$$\left(\hat{\beta}_0, \hat{\beta}_X, \hat{\beta}_D\right) = \arg \min_{\{\beta_0, \beta_X, \beta_D\}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_X X_i - \beta_D D_i\right)^2$$

$$S_{\text{coef}} = \hat{\beta}_D$$

回归模型是否正确设定不影响检验的有效性，但影响统计功效。

- 实现：可以穷举处理状态的所有分配方式，或采用模拟方法近似。(help permute)

  1.基于对分配机制的理解抽取 $\tilde{\mathrm{D}}$.

  2.根据 $\tilde{\mathrm{D}}$ 构造 $\mathrm{Y}^{obs}(\tilde{\mathrm{D}})$.

  3.计算 $S\left(\tilde{\mathrm{D}}, \mathrm{Y}^{obs}(\tilde{\mathrm{D}}), \mathrm{X}\right)$.

  4.重复以上步骤。

一个例子：

- 20 世纪 20 年代末一个夏日的午后，在英国剑桥，一群大学教员、他们的妻子以及一些客人围坐在室外的一张桌子周围喝下午茶。其中一位女士坚持认为，将牛奶倒入茶中和将茶倒入牛奶中的味道不同。在座的科学家均认为这种观点很可笑，两种液体混合，化学成分一样，能有什么区别。
- Fisher 设计的实验：准备 8 杯一样的奶茶，其中 4 杯是先奶后茶，4 杯先茶后奶，并随机打乱顺序。此时，请这位女士品尝这 8 杯奶茶，分辨其中先奶后茶的 4 杯。
- 这位女士至少需要分辨出几杯咖啡，才能证明她有这个能力？
- 处理发生方式：先加奶后加茶，$D = 1$。
- 原假设：两种奶茶没有差别：$Y_i(1) = Y_i(0) \ \forall i$。
- 女士完全正确的概率是 $Pr_0 = \frac{C_8^8}{C_8^4} \approx 0.0143$;
- 认错一杯的概率是 $Pr_1 = \frac{C_4^1 C_4^3}{C_8^4} \approx 0.2286$.

- 在 SUTVA 假设下，在处理 $d=1$ 或 $d=0$ 时，每个个体的潜在处理结果是 $Y_i(d)$。个体的潜在处理效应为：

$$\tau_i = Y_i(1) - Y_i(0)$$

- 对于每个个体，我们仅能观测到其中的一个，也就是同一个体不能同一时间既接受处理又不接受处理，因此潜在处理效应 τ 不可观测。设 D_i 为处理分配机制，则观测结果为：

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

- 完全随机分配意味着每种处理发生的概率相同且独立，即

$$Pr(D = d) = \frac{1}{C_N^{N_1}} = \frac{N_1!}{N!}$$

- 例：下表是关于某种药物杀灭致病菌效果的小型随机试验中搜集的数据。共有 12 名被试参加了该项试验，随机选择其中 6 名服用该药物（处理组），另外 6 名服用安慰剂（控制组）。结果是关于被试体内致病菌寄生程度的指标。

控制组	处理组
8.62	0.06
1.48	1.72
8.93	2.19
9.57	7.32
2.65	7.53
7.30	7.62

我们希望检验该药物对这 12 名被试无效的原假设。

- H0：$Y_i^1 - Y_i^0 = 0$

控制/处理组	观测值	Y^0	Y^1
0	8.62	8.62	-
0	1.48	1.48	-
0	8.93	8.93	-
0	9.57	9.57	-
0	2.65	2.65	-
0	7.30	7.30	-
1	0.06	-	0.06
1	1.72	-	1.72
1	2.19	-	2.19
1	7.32	-	7.32
1	7.53	-	7.53
1	7.62	-	7.62

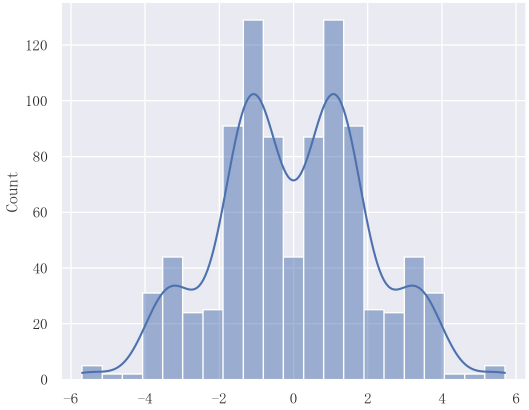
在原假设成立的情况下，我们其实知道控制组的 Y^1 和处理组的 Y^0:

控制/处理组	观测值	Y^0	Y^1
0	8.62	8.62	8.62
0	1.48	1.48	1.48
0	8.93	8.93	8.93
0	9.57	9.57	9.57
0	2.65	2.65	2.65
0	7.30	7.30	7.30
1	0.06	0.06	0.06
1	1.72	1.72	1.72
1	2.19	2.19	2.19
1	7.32	7.32	7.32
1	7.53	7.53	7.53
1	7.62	7.62	7.62

如果我们随机地抽取处理组，那么抽取到如此“小”的概率是多少？

$$\sum Y_i^1/6 - \sum Y_i^0/6 = -2.0183333333333326$$

- 计算 12 个个体中，任意取 6 个作为处理组的全部组合的处理效应，$C_{12}^{6} = 924$。
- 所有组合处理效应的分布：



- 有 125 个组合小于 -2.018333：
- $p = 125/924 = 0.13528138528138528$

- 如果样本比较大怎么办?
  - 想象一下，2000 个个体里面，取 1000 个个体，有多少种组合?

- 与回归分析的区别
  - Different sampling perspectives.
    - Fisher： finite population, potential outcomes as fixed, treatment assignments as the sole source of randomness.
    - Regression analysis: Infinite super-population. Potential outcomes in the sample are random. Properties of the estimators are assessed by resampling from that population, sometimes conditional on covariates as well as the treatment indicator.

- Key features of regression models for randomized experiments. They are models
  - for observed outcomes rather than for potential outcomes.
  - for the conditional mean rather than for the full distribution.
  - in which the estimand is always an average treatment effect and is a parameter of a statistical model.
  - in which inference can be viewed as inference for parameters of the statistical model.
  - in which whether the conditional mean is correctly specified is immaterial.
  - in which all results are asymptotic (large sample) results.

2 当我谈最小二乘法时我在谈些什么

2.1 OLS 的代数学和几何学

- Question: We have a sample with n observations ($i = 1, \cdots, n$) on individual wage y and $K - 1$ background characteristic $x_2, \cdots, x_K$. How in this sample are wages related to the other observables?
- Strategy: Which linear combination of $x_2, \cdots, x_K$ and a constant gives a good approximation of y?

$$\tilde{\beta}_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_K x_K$$

- Solution: OLS

$$\min_{\tilde{\beta}} S\left(\tilde{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\tilde{\beta}\right)^2$$

$$\mathbf{x}_i = (1\ x_{i2} \cdots x_{iK})', \tilde{\beta} = \left(\tilde{\beta}_1 \cdots \tilde{\beta}_K\right)'$$

$$\sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \mathbf{x}_i'\tilde{\beta}\right) = 0$$

$$\mathrm{b} = \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

- So far the linear approximation does not depend on any economic or statistical theory, and it holds irrespective of how the data are generated. Moreover, the linear approximation is an in-sample result. It does not give information about observations outside the sample, and there is no direct interpretation of the coefficients.

• 特例，$K = 2$

$$
\begin{aligned}
y_i &= b_0 + b_1 x_1 + e_i \\
b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{Cov}(x, y)}{\hat{Var}(x)} \\
b_0 &= \bar{y} - b_1 \bar{x}
\end{aligned} \tag{6}
$$

如果个体 i 为男性，则 $x_i = 1$; 如果不为男性，则 $x_i = 0$. 我们定义

$$
\bar{y}_m = \frac{\sum x_i y_i}{\sum x_i}, \bar{y}_f = \frac{\sum (1 - x_i) y_i}{(1 - x_i)}
$$

那么，

$$
b_0 = \bar{y}_f, b_1 = \bar{y}_m - \bar{y}_f
$$

• Matrix notation

$$\mathrm{X}_{n\times K} = \begin{pmatrix} \mathrm{x}_1' \\ \vdots \\ \mathrm{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nK} \end{pmatrix} = (\hat{\mathrm{x}}_1 \cdots \hat{\mathrm{x}}_K)$$

$$\mathrm{y}_{n\times 1} = (y_1 \cdots y_n)'$$

$$\min_{\hat{\beta}} S\left(\tilde{\beta}\right) = \left(\mathrm{y} - \mathrm{X}\tilde{\beta}\right)' \left(\mathrm{y} - \mathrm{X}\tilde{\beta}\right)$$

$$\mathrm{b} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

$$\mathrm{y} = \mathrm{X}\mathrm{b} + \mathrm{e}$$

$$\mathrm{X}'\mathrm{e} = 0$$

从代数的角度定义的最小二乘问题:

考虑 $A\boldsymbol{x}$ 作为 $\boldsymbol{b}$ 的一个近似. $\boldsymbol{b}$ 和 $A\boldsymbol{x}$ 之间的距离越小, $\|\boldsymbol{b}-A\boldsymbol{x}\|$ 近似程度越好. 一般的最小二乘问题就是找出使 $\|\boldsymbol{b}-A\boldsymbol{x}\|$ 尽量小的 $\boldsymbol{x}$, 术语"最小二乘"来源于这样的事实, 即 $\|\boldsymbol{b}-A\boldsymbol{x}\|$ 是平方和的平方根.

最小二乘问题最重要的特征是无论怎么选取 $\boldsymbol{x}$, 向量 Ax 必然属于列空间 $\operatorname{Col} A$. 因此我们寻求 $\boldsymbol{x}$, 使得 $A\boldsymbol{x}$ 是 $\operatorname{Col} A$ 中最接近 $\boldsymbol{b}$ 的点, 见下图（当然，如果 $\boldsymbol{b}$ 恰好在 $\operatorname{Col}A$ 中, 那么对于某个 $\boldsymbol{x}, \boldsymbol{b}$ 等于 $A\boldsymbol{x}$, 且这样的 $\boldsymbol{x}$ 是一个"最小二乘解"）.

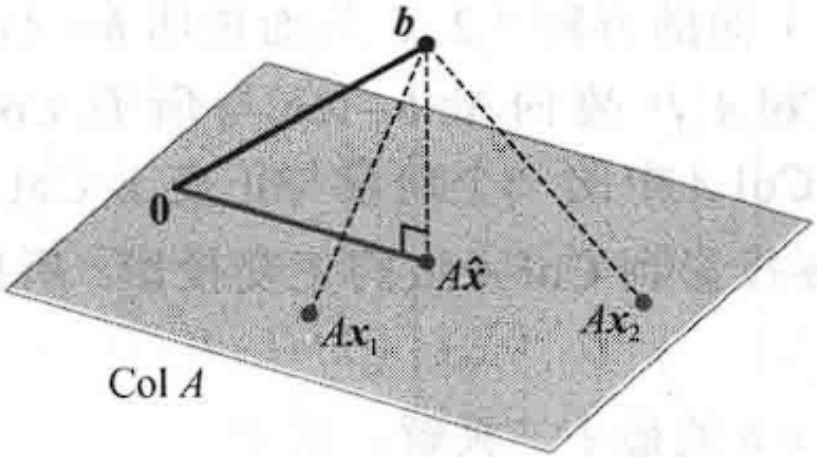
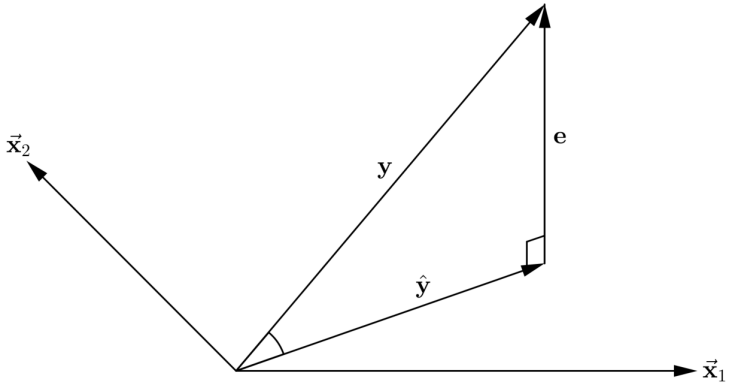


图 6-23　对 $\boldsymbol{x}$，$\boldsymbol{b}$ 与 $A\hat{\boldsymbol{x}}$ 的距离小于与 $A\boldsymbol{x}$ 的距离

- OLS 几何学解释
  - y,e 和 X 的每一列都可以视为空间 $\mathbb{E}^n$ 上的向量。$(\vec{\mathrm{x}}_1, \cdots, \vec{\mathrm{x}}_K)$ 是 K 维空间的基向量。
  - e 和 X 的线性张成（span）子空间正交（orthogonal）。
  - $\hat{\mathrm{y}}$ 可以视为 n 维向量 y 在 K 维子空间上的正交投影（orthogonal projection）。OLS 的思想就是最小化 y 到空间 X 上的距离。



- Frish-Waugh-Lovell theorem
  - Suppose we regress y on X_1 and X_2 and get

$$y = X_1 b_1 + X_2 b_2 + u$$

  - We can get the same b_1 and u in the following way: Regress y and X_1 on X_2 separately and store the corresponding residual e_y and e_1, and then regress e_y on e_1.

$$e_y = e_1 b_1 + u$$

  - We don't have to net out the effect of X_2 on y to get b_2 unless we also wish to have u.
  - What's so good about FWL? Once we purge the regression of less interesting variables, we can reduce it to the bivariate case and make use of the "covariance-to-variance"formula for subsequent analysis.
  - Especially useful in bivariate correlation plot.
    示例. 女人和犁（Alesina et al, 2013, QJE）

- 两种对女性家庭地位的看法：
  - 女性应该和男人一样自由地选择做什么工作。
  - 男主外，女主内。
- 两种农耕传统：
  - 迁移耕地：用锄头，需要的力气更小。
  - 用犁耕地：用犁，需要的力气更大，男人的优势更加明显。
- 黄色：不用犁。棕色：用犁，但犁技术从外部输入。深棕色：用犁，且犁技术是本土发明的。

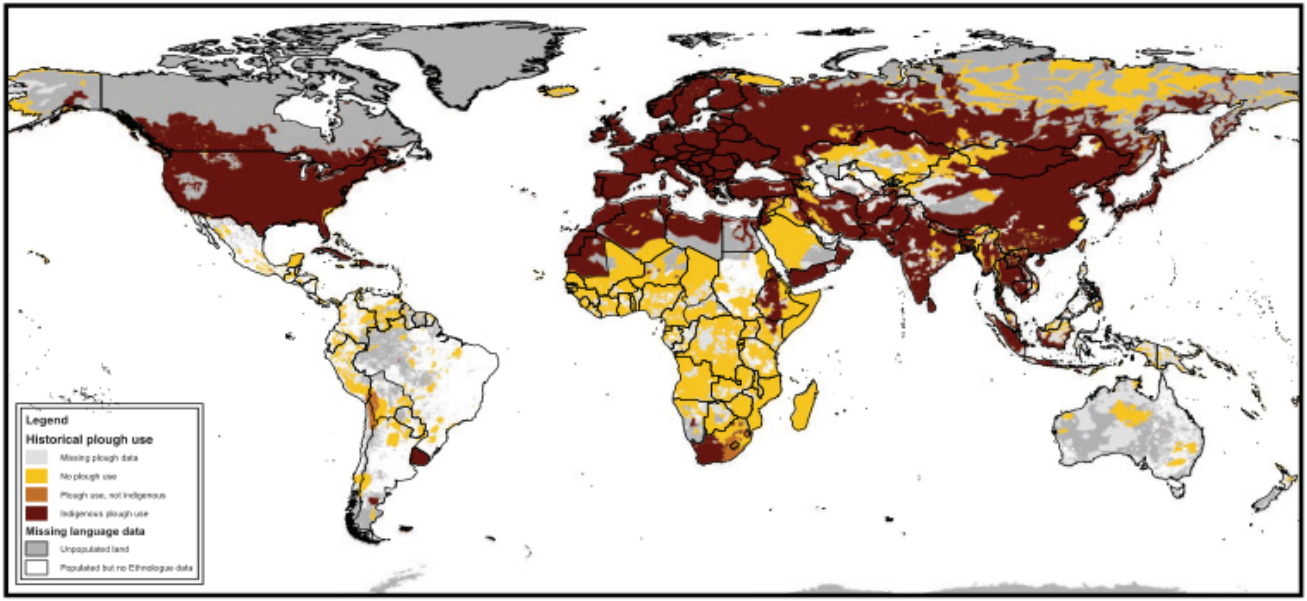
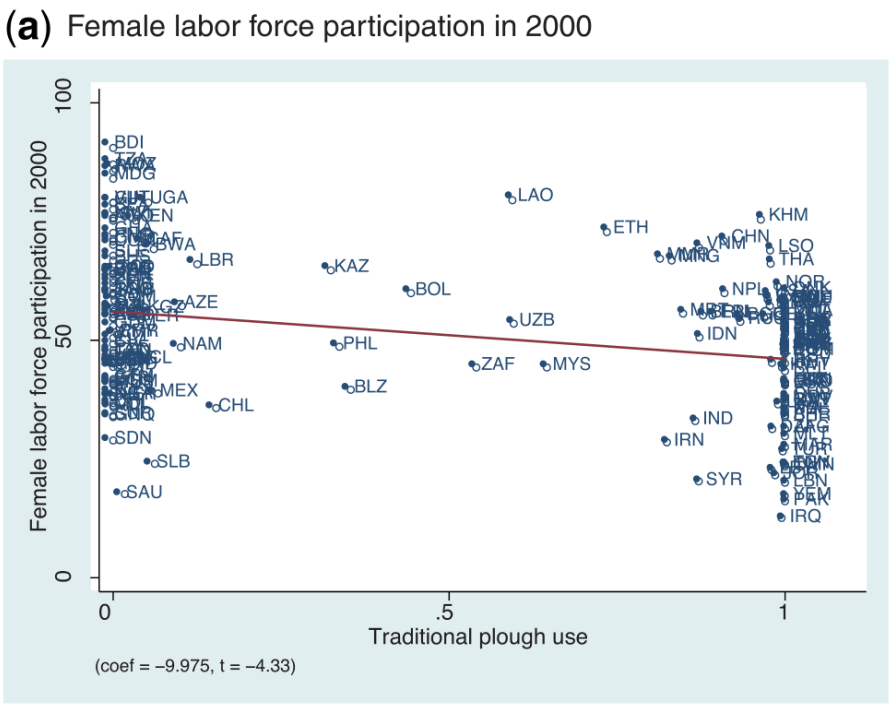


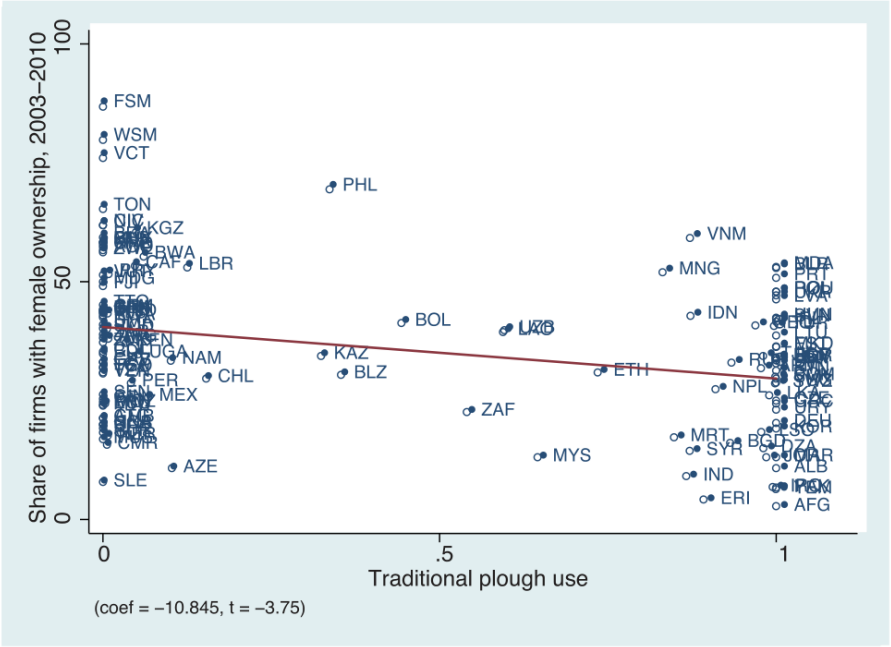
FIGURE II

Traditional Plough Use among the Ethnic/Language Groups Globally

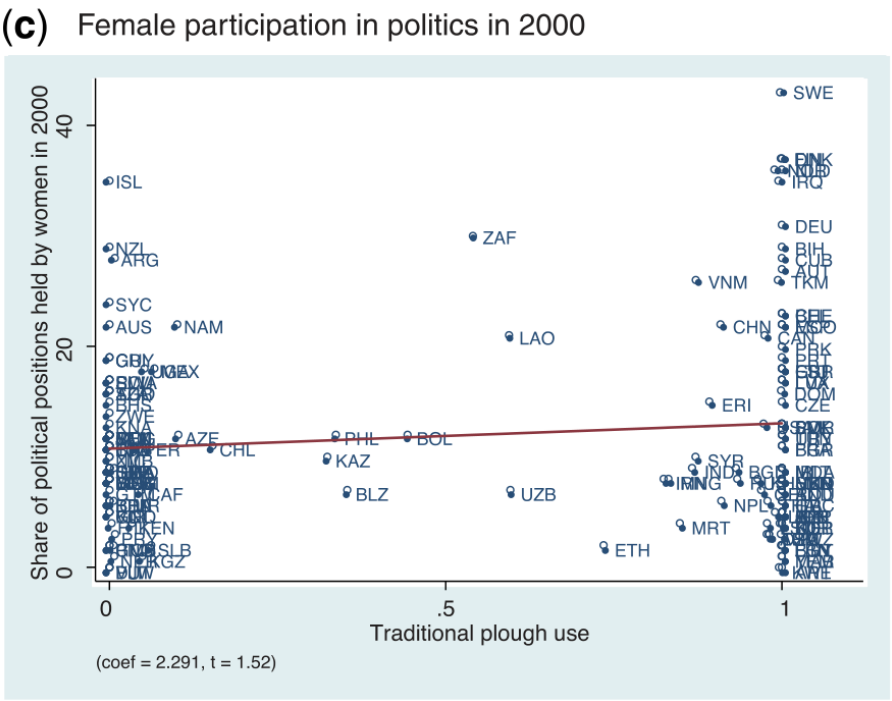


女性劳动参与率和使用犁耕地负相关，显著。

(b) Female firm ownership, 2003–2010



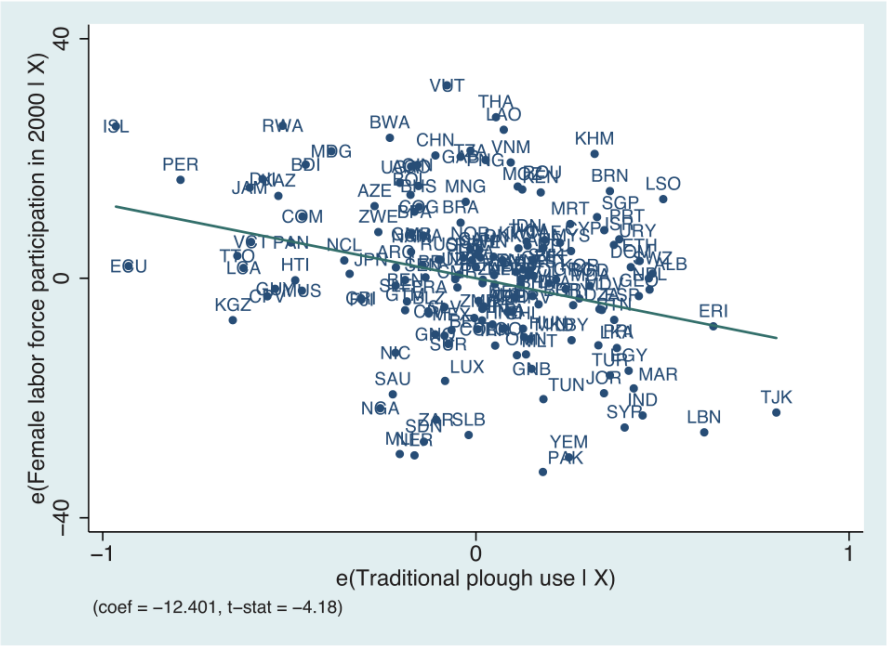
女性企业所有权和使用犁耕地负相关，显著。



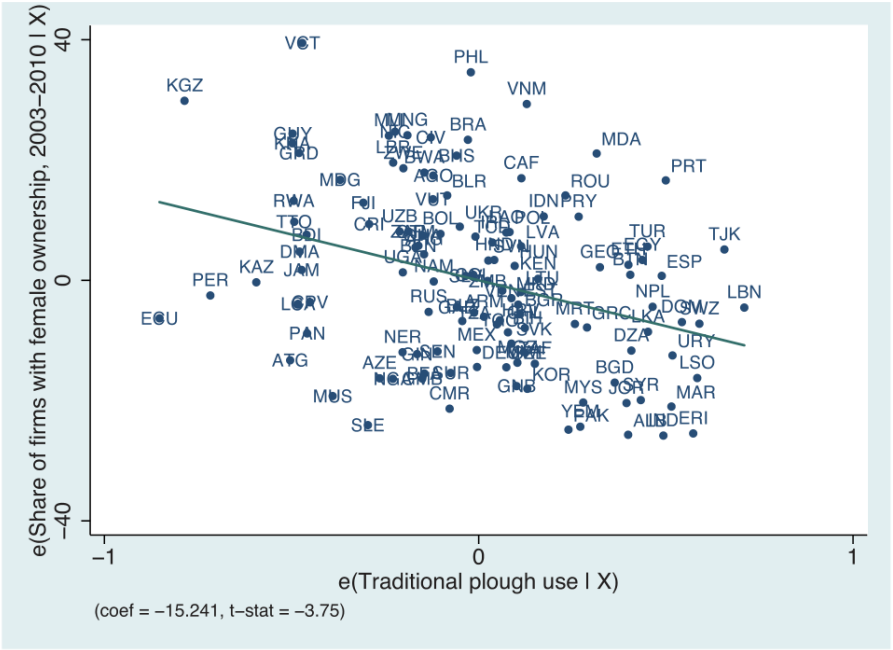
女性政治参与和使用犁耕地正相关，不显著。

- 原因是人均收入变量被遗漏。人均收入和女性政治参与正相关，同时和犁地正相关，并且与女性政治参与正相关程度高于与犁地正相关的程度。
- 因此，即使犁地和女性政治参与负相关，由于犁地和人均收入也正相关，当给定其他条件不变时，直接观察犁地和女性政治参与也可能得到正相关关系。

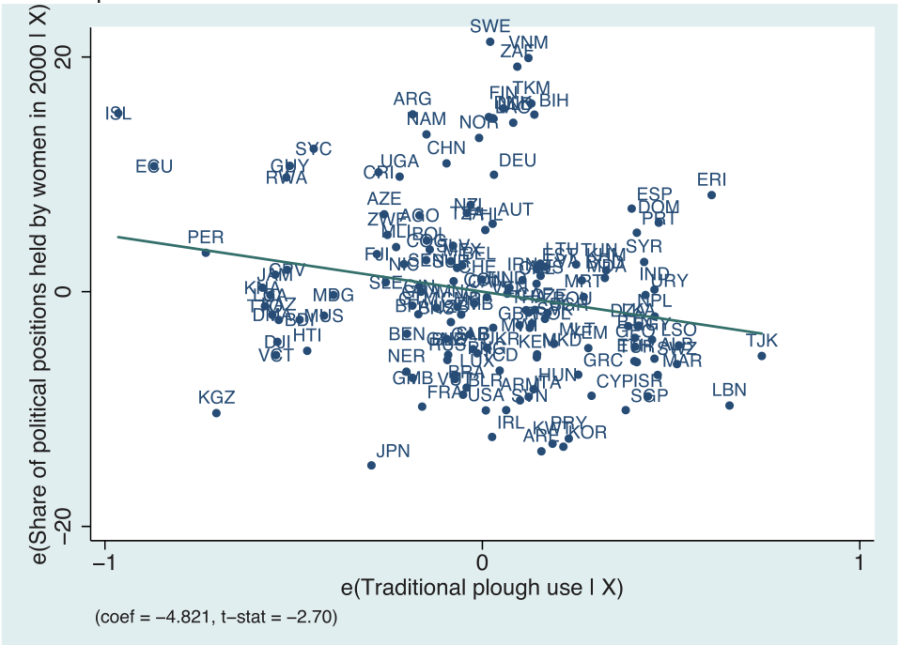
(**a**) Traditional plough use and current FLFP



(**b**) Traditional plough use and current female firm ownership



(c) Traditional plough use and current female participation in politics



2.2 小样本理论：假设、性质与推断

- 线性统计模型：

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

- 关键假设 1：严格外生性

$$E(\varepsilon|\mathrm{X}) = 0$$

$$E(\mathrm{y}|\mathrm{X}) = \mathrm{X}\beta$$

- OLS estimator 的性质，无偏性：

$$E(\mathrm{b}|\mathrm{X}) = \beta$$

- 在包含控制变量的回归中我们实际隐含的往往是条件均值独立性假定——如果我们并不关心控制变量的因果效应。（控制变量控制的到底是什么？回顾 1.5 的内容。）

- Key assumption 2: Spherical error variance.

$$\text{Var}(\varepsilon|X) = E(\varepsilon\varepsilon'|\text{X}) = \sigma^2\text{I}_n$$

1. Conditional homoskedasticity (条件同方差).

$$\text{Var}(\varepsilon_i|\text{X}) = \sigma^2$$

2. No correlation between observations.

$$\text{Cov}(\varepsilon_i, \varepsilon_j|\text{X}) = 0$$

- Conditional variance of b

$$\text{Var}(\text{b}|\text{X}) = \sigma^2(\text{X}'\text{X})^{-1}$$

In the simple linear regression,

$$\text{Var}(b_1|\text{X}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\text{Var}(b_k|\text{X})} = \frac{s^2}{(n-1)\left(1 - R_{c,k}^2\right)\widehat{\text{Var}(x_k)}}$$

where $R_{c,k}^2$ is the centered-R^2 in the regression of x_k on all other variables.

– The larger the sample size,

– The greater the variation in x_k,

– The less the correlation of x_k with the other variables,

– The better the overall fit of the regression,

the lower the variance of b_k will be.

- An important implication is that the presence of two highly collinear variables does not affect the precision of estimating a third variable in interest.

- Don't make too much of R^2.
  - Varies with different samples.
  - Sensitive to the definition of y.
  - No absolute benchmark to judge whether it is high or low.
  - One can easily manipulate a high R^2.
  - Measures the quality of linear approximation, while has nothing to say about whether the model is true, whether the assumptions are valid, or whether the estimator has good statistical properties.
- Hypothesis testing.

Decision/Hypothesis	TRUE	FALSE
Reject	Type I error (α)	No error ($1-\beta$)
Cannot reject	No error ($1-\alpha$)	Type II error (β)

- Consider the following example of testing the null regarding the mean of a normally distributed population against a particular alternative,

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1$$

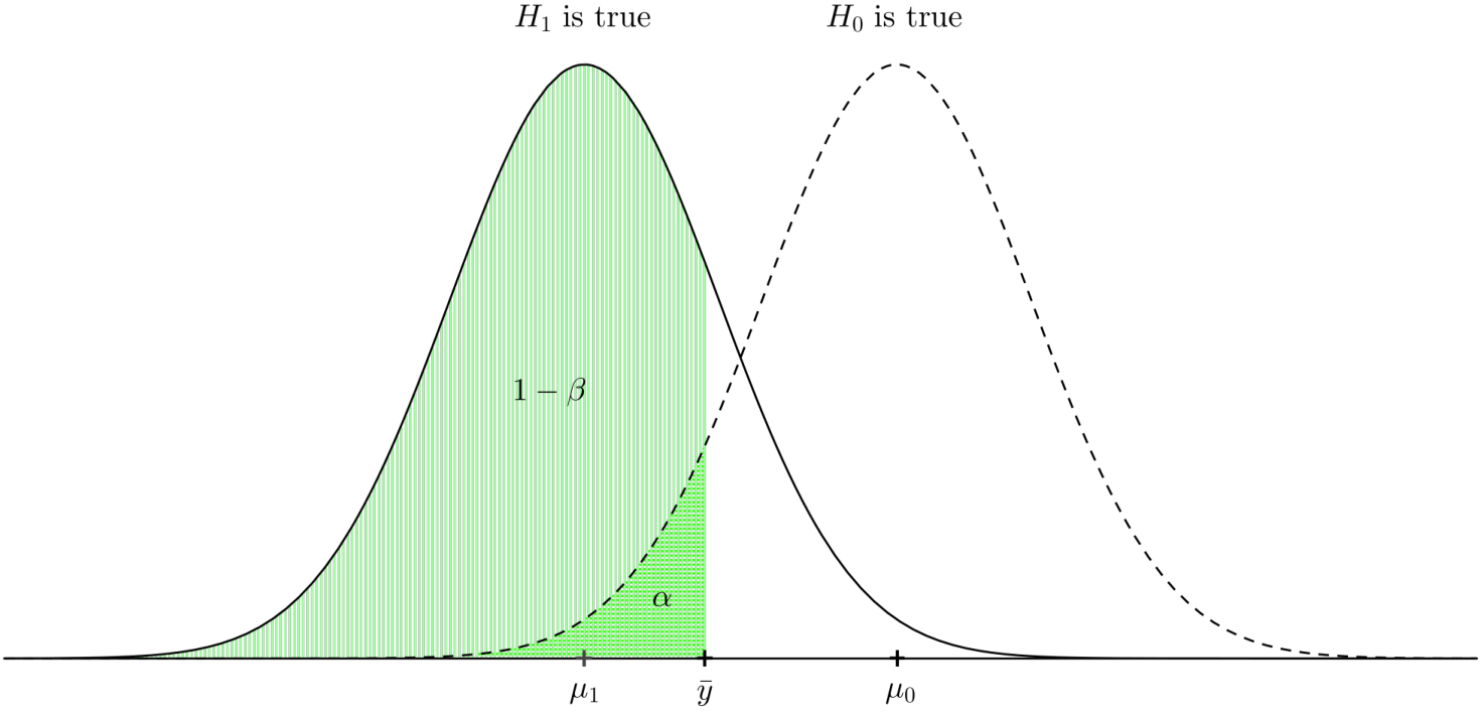
where without loss of generality $\mu_1 < \mu_0$.

$$\begin{aligned}
\text{Power} &= \Pr\left(z = \frac{\bar{y} - \mu_0}{\text{SE}(\bar{y})} < z_\alpha | \mu = \mu_1\right) \\
&= \Pr\left(\bar{y} < \mu_0 + z_\alpha \text{SE}(\bar{y}) | \mu = \mu_1\right) \\
&= \Pr\left(\frac{\bar{y} - \mu_1}{\text{SE}(\bar{y})} < z_\alpha + \frac{\mu_0 - \mu_1}{\text{SE}(\bar{y})} | \mu = \mu_1\right) \\
&= \Pr\left(z < z_\alpha + \frac{\mu_0 - \mu_1}{\text{SE}(\bar{y})}\right) \\
&= \Phi\left(z_\alpha + \frac{\mu_0 - \mu_1}{\text{SE}(\bar{y})}\right)
\end{aligned}$$

Power increases when

- the size α, increases,
- the magnitude of the difference we want to detect increases,
- the variability (of $\bar{y}$) decreases, and in particular, the sample size increases.

This implies that Type II errors become increasingly unlikely if we have large samples. To compensate for this, we typically reduce the probability of Type I errors by lowering the size α.


H_1 is true
H_0 is true
$1-\beta$
α
μ_1
$\bar{y}$
μ_0

• Key assumption 3: Normality of error distribution.

$$\varepsilon|\mathrm{X} \sim N(0, \sigma^2 \mathrm{I}_n)$$
$$(\mathrm{b} - \beta)|\mathrm{X} \sim N(0, \sigma^2(\mathrm{X}'\mathrm{X})^{-1})$$

• t-test of individual regression coefficients.

$$H_0 : \beta_k = \bar{\beta}_k$$

$$t_k \triangleq \frac{b_k - \bar{\beta}_k}{\widehat{\mathrm{SE}(b_k)}} = \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 \left((\mathrm{X}'\mathrm{X})^{-1}\right)_{kk}}} \sim t(n-K)$$

- Decision rules for t-test.
  - –Test based on critical value.

$$\text{Prob}(-t_{\alpha/2}(n-K) < t_k < t_{\alpha/2}(n-K)) = 1-\alpha)$$

    Accept H_0 if t_k falls within this range. Reject otherwise.
  - –Test based on p-value.

$$p = \text{Prob}(t > |t_k|) \times 2$$

    Accept H_0 if $p > \alpha$, reject otherwise.
  - –Test based on confidence interval.

$$b_k - \hat{\text{SE}}(b_k)t_{\alpha/2}(n-K) < \beta_k < b_k + \hat{\text{SE}}(b_k)t_{\alpha/2}(n-K)$$

    Accept H_0 if $\bar{\beta}_k$ falls within this range. Reject otherwise.

- t-test of one linear restriction.

$$H_0 : r_1\beta_1 + \cdots + r_K\beta_K = \mathbf{r}'\beta = q$$
$$t_{stat} \triangleq \frac{\mathbf{r}'\mathbf{b} - q}{\hat{\mathrm{SE}}(\mathbf{r}'\mathbf{b})} = \frac{\mathbf{r}'\mathbf{b} - q}{\sqrt{s^2\mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}r}} \sim t(n-K)$$

- F-test of general linear hypotheses.

$$H_0 : \mathrm{R}\beta = \mathbf{q}$$

$$\begin{aligned} F_{stat} &= (\mathrm{R}\mathbf{b} - \mathbf{q})' \left(s^2\mathrm{R}(\mathbf{X}'\mathbf{X})^{-1}\mathrm{R}'\right)^{-1} (\mathrm{R}\mathbf{b} - \mathbf{q})/\dim(\mathbf{q}) \\ &\sim F(\dim(\mathbf{q}), n-K) \end{aligned}$$

where $\dim(q)$ denotes the number of linear restrictions.

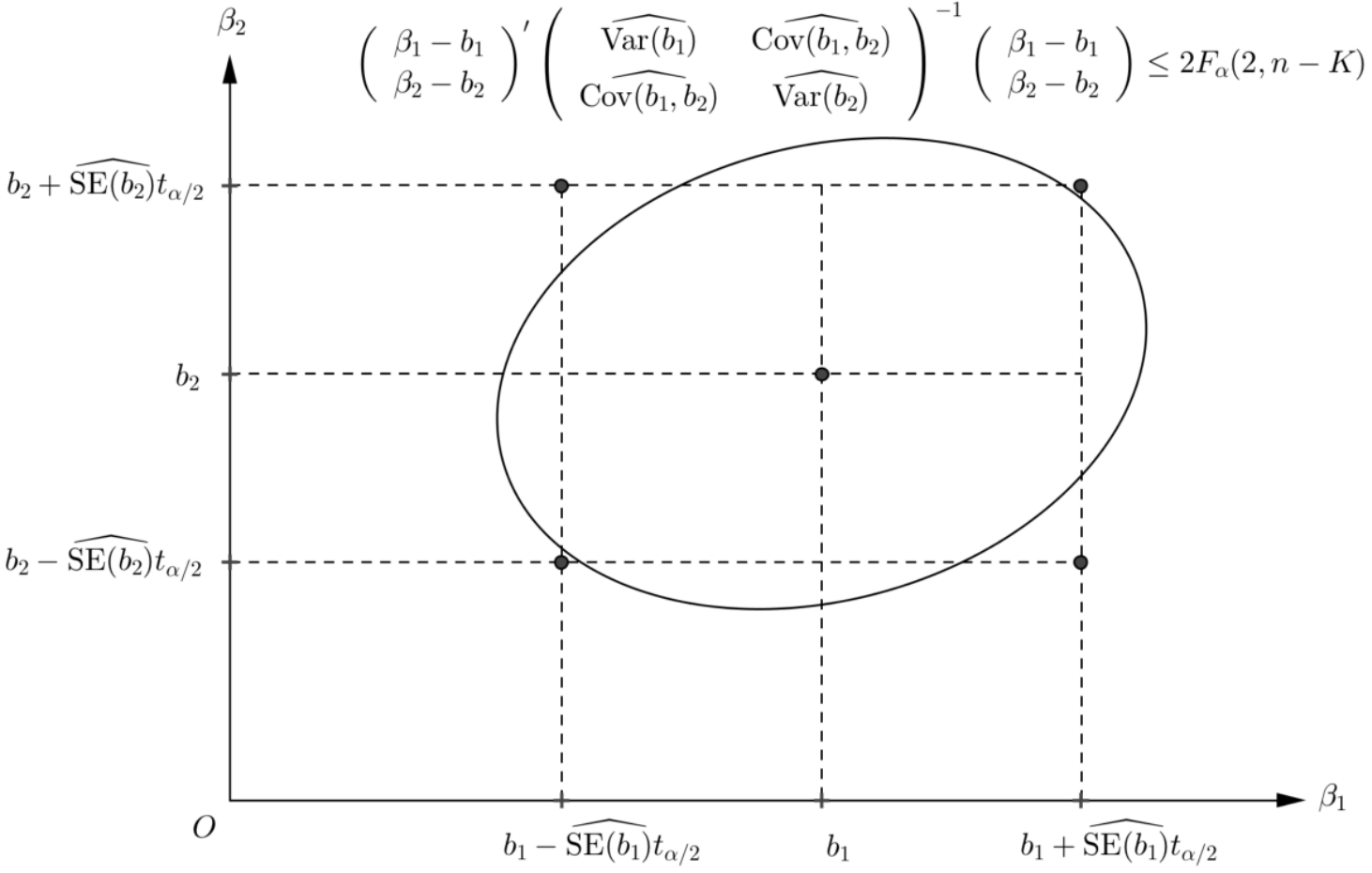
- Decision rules for F -test.
  - Test based on critical value.

$$\text{Prob}(F < F_\alpha(\text{dim}(\text{q}), n - K)) = 1 - \alpha$$

    Accept H_0 if $F_{stat} < F_\alpha(\text{dim}(\text{q}), n - K)$, reject otherwise.
  - Test based on p-value.

$$p = \text{Prob}(F > F_{stat})$$

    Accept H_0 if $p > \alpha$, reject otherwise.
- The previous t-tests are special cases of the F-test.
- Why don't we use multiple t-tests for multiple linear restrictions?


$\begin{pmatrix} \beta_1 - b_1 \\ \beta_2 - b_2 \end{pmatrix}' \begin{pmatrix} \widehat{\mathrm{Var}(b_1)} & \widehat{\mathrm{Cov}(b_1, b_2)} \\ \widehat{\mathrm{Cov}(b_1, b_2)} & \widehat{\mathrm{Var}(b_2)} \end{pmatrix}^{-1} \begin{pmatrix} \beta_1 - b_1 \\ \beta_2 - b_2 \end{pmatrix} \leq 2F_\alpha(2, n-K)$
β_2
$b_2 + \widehat{\mathrm{SE}(b_2)}t_{\alpha/2}$
b_2
$b_2 - \widehat{\mathrm{SE}(b_2)}t_{\alpha/2}$
O
$b_1 - \widehat{\mathrm{SE}(b_1)}t_{\alpha/2}$
b_1
$b_1 + \widehat{\mathrm{SE}(b_1)}t_{\alpha/2}$
β_1

2.3 大样本理论：假设、性质与推断

- The key assumption for finite-sample theory to hold is normal disturbances. Large-sample theory establishes that, when sample size is sufficiently large ($n \to \infty$), OLS estimators will have desirable properties even if this assumption is relaxed. It turns out that we can also relax the strict exogeneity assumption.
- Law of large number. $\{x_i\}$ i.i.d., $E(x_i) = \mu, \text{Var}(x_i) < \infty$, then

$$\bar{x}_n \longrightarrow_p \mu$$

A weaker version of LLN also holds when $\{x_i\}$ is identically distributed and having only weak persistence. That is to say, it allows for (serial or cross-sectional) correlation.

- Central limit theorem. $\{x_i\}$ i.i.d., $E(x_i) = \mu, \text{Var}(x_i) = \Sigma$, then

$$\bar{x}_n \to_d N(\mu, \text{AVar}(\bar{x}_n))$$

where the asymptotic variance of $\bar{x}_n$,

$$\text{AVar}(\bar{x}_n) = \frac{\Sigma}{n}$$

here are weaker versions of CLT when $\{x_i\}$ is non-independent but uncorrelated and when $\{x_i\}$ is correlated. In the latter case $\text{AVar}(\bar{x}_n)$ involves not only the variance of x_i but also its (auto)covariances.

- Key assumption: Predetermined (前定) regressors.

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

This condition is called moment condition (矩条件) or orthogonality condition (正交条件).

- Consistency of b.

$$\text{b} - \beta = (\text{X}'\text{X})^{-1}\text{X}'\varepsilon = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right) \to_p 0$$

- Asymptotic normality of b when $\mathbf{x}_i\varepsilon_i$ is uncorrelated.

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \to_d N\left(0, \frac{\text{Var}(\mathbf{x}_i, \varepsilon_i)}{n}\right) = N$$

$$\text{b} \to_d N(\beta, \frac{1}{n}(E(\mathbf{x}_i\mathbf{x}_i^{'}))^{-1}E(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i^{'})(E(\mathbf{x}_i\mathbf{x}_i^{'}))^{-1})$$

  - Importantly, the error term can be conditionally heteroskedastic.
  - Call $\widehat{\text{SE}(b_k)}$ heteroskedasticity-robust (Huber-White) standard error. STATA 实现：robust
  - Robust standard errors can be more misleading than homoskedasticity-only standard errors in situations where heteroskedasticity is modest. A conservative rule of thumb (but rarely adopted!) is to use the maximum of the conventional standard errors and the robust standard errors.

- Asymptotic normality of b when $\mathrm{x}_i\varepsilon_i$ is correlated within clusters. Clustering arises when sampling is of independent groups but errors for individuals within the group are correlated.
  Let $\mathrm{X}_g(n_g \times K)$, $\varepsilon_g(n_g \times 1)$, and $\mathrm{e}_g(n_g \times 1)$ denote the regressor matrix, the error vector, and the residual vector for group g, respectively; n_g is the number of observations in group g; G is the number of groups.

$$\mathrm{b} - \beta = \left(\frac{1}{G}\sum_{g=1}^{G}\mathrm{X}_g'\mathrm{X}_g'\right)^{-1}\left(\frac{1}{G}\sum_{g=1}^{G}\mathrm{X}_g'\varepsilon_g\right)$$

$$\mathrm{b} \to_d N\left(\beta, \frac{1}{G}\left(E\left(\mathrm{X}_g'\mathrm{X}_g\right)\right)^{-1}\left(E\left(\mathrm{X}_g'\varepsilon_g\varepsilon_g'\mathrm{X}_g\right)\right)\left(E\left(\mathrm{X}_g'\mathrm{X}_g\right)\right)^{-1}\right)$$

$$\widehat{\mathrm{Var}(\mathrm{b})} = (\mathrm{X}'\mathrm{X})^{-1}\left(\sum_{g=1}^{G}\mathrm{X}_g'\mathrm{e}_g\mathrm{e}_g'\mathrm{X}_g\right)(\mathrm{X}'\mathrm{X})^{-1}$$

where

$$\sum_{g=1}^{G}\mathrm{X}_g'\mathrm{e}_g\mathrm{e}_g'\mathrm{X}_g = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}e_{gi}e_{gj}\mathrm{x}_{gi}\mathrm{x}_{gj}'$$

–STATA 实现: cluster(clustvar)

– It is especially necessary to use cluster-robust standard errors when an aggregated or macro variable is included while data are on an individual basis.
– Two high-profile papers advocate the use of cluster-robust standard errors. (Bertrand et al., 2004, QJE; Petersen, 2009, RFS).
– 问题：聚类到省级层面 vs. 聚类到市级层面，哪种做法更稳健（隐含的扰动项假设更弱）?

如何实现复杂的 cluster?

- 什么是复杂的 cluster?
  - 同时考虑同省份企业和同行业企业的相关性。
  - 此时只 cluster(province) 或 cluster(sector) 无法实现。
- 一种错误的做法

```
egen double_cluster=group(idcode year)
regress ln_wage age i.race, vce(cluster double_cluster)
```

- 正确的做法：使用 reghdfe 命令，并同时 cluster(group1 group2)

```
reghdfe ln_wage age i.race, absorb(temp) cluster(idcode year)
```

- 参考资料：
  http://www.tomzimmermann.net/2018/08/22/two-way-clustering-in-stata/

	firm1			firm2			firm3		
firm1	ε_{11}^2	$\varepsilon_{11}\varepsilon_{12}$	$\varepsilon_{11}\varepsilon_{13}$	0	0	0	0	0	0
	$\varepsilon_{12}\varepsilon_{12}$	ε_{12}^2	$\varepsilon_{11}\varepsilon_{12}$	0	0	0	0	0	0
	$\varepsilon_{13}\varepsilon_{11}$	$\varepsilon_{13}\varepsilon_{12}$	ε_{13}^2	0	0	0	0	0	0
firm2	0	0	0	ε_{21}^2	$\varepsilon_{21}\varepsilon_{22}$	$\varepsilon_{21}\varepsilon_{23}$	0	0	0
	0	0	0	$\varepsilon_{22}\varepsilon_{21}$	ε_{22}^2	$\varepsilon_{22}\varepsilon_{23}$	0	0	0
	0	0	0	$\varepsilon_{23}\varepsilon_{21}$	$\varepsilon_{23}\varepsilon_{22}$	ε_{23}^2	0	0	0
firm3	0	0	0	0	0	0	ε_{31}^2	$\varepsilon_{31}\varepsilon_{32}$	$\varepsilon_{31}\varepsilon_{33}$
	0	0	0	0	0	0	$\varepsilon_{32}\varepsilon_{31}$	ε_{32}^2	$\varepsilon_{32}\varepsilon_{33}$
	0	0	0	0	0	0	$\varepsilon_{33}\varepsilon_{31}$	$\varepsilon_{33}\varepsilon_{32}$	ε_{33}^2

Clustering issue

- Because our key regressor $g - investG_{rt}$ is a provincial-level variable, though we focus on its interaction effect with a firm-level variable, we still checked whether the firm observations should be clustered at the provincial level. It indicates that all the firms in the same province are correlated, which is equivalent to adjusting heteroskedasticity among provinces. From Figure 3 we can find that distribution of the residuals from our basic regression is quite similar among most of the provinces (except Tibet, because of fewer observations), which supports the claim that there is no obvious heteroskedasticity problem among provinces. Besides, when there are few clusters in the sample (in our case, 31 clusters with more than one million observations), there is no consensus in the literature on a truly satisfactory solution, and clustered standard errors can suffer from few-cluster bias and over-reject significant coefficients (Cameron and Miller 2013). Therefore, we assume correlation among firms'performance in different years and cluster our regression at the firm level.

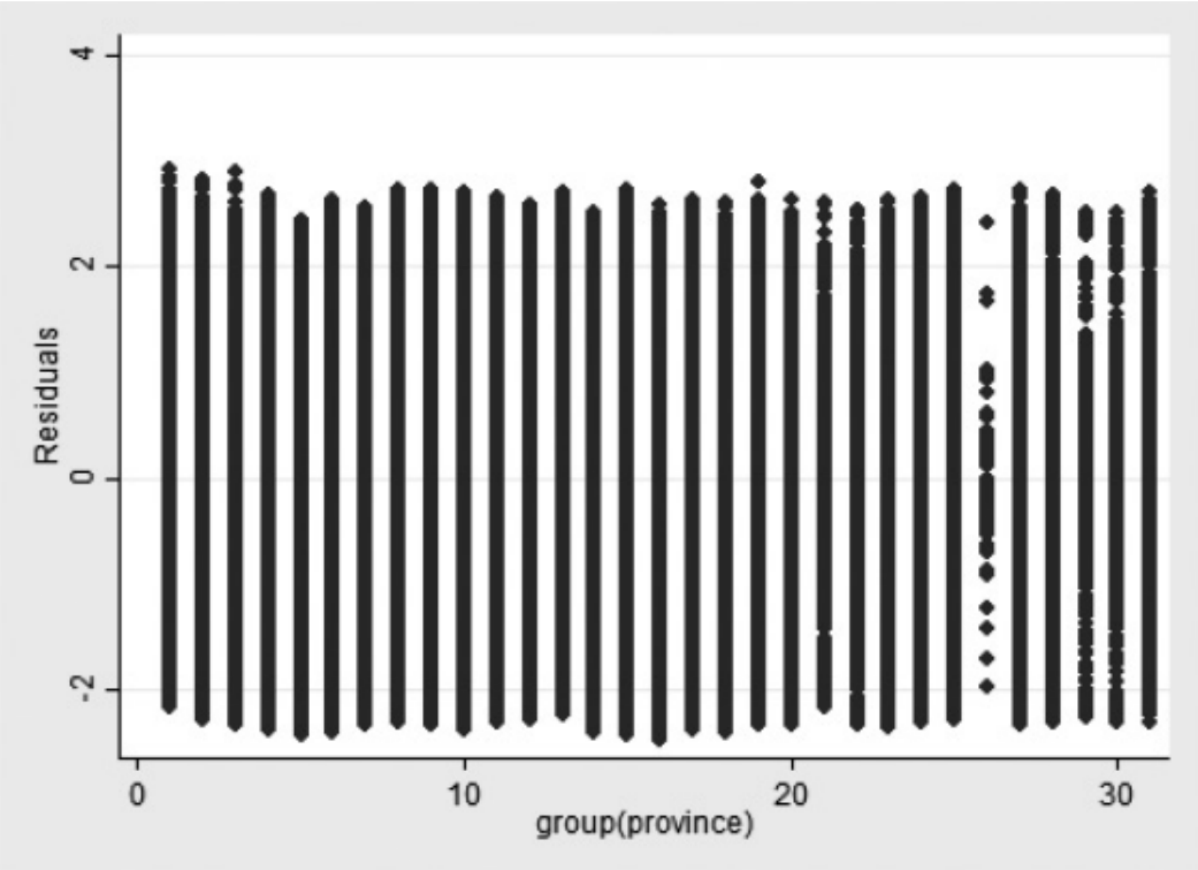


Figure 1: Distribution of residuals by province from basic regression

- z-test of individual regression coefficients.

$$H_0 : \beta_k = \bar{\beta}_k$$

Robust t-statistic

$$t_k \triangleq \frac{b_k - \bar{\beta}_k}{\widehat{\mathrm{SE}(b_k)}} \to_d N(0,1)$$

- χ^2-test of linear restrictions.

$$H_0 : \mathrm{R}\beta = \mathrm{q}$$

Wald statistic

$$W = (\mathrm{Rb} - \mathrm{q})' \left(\mathrm{R}\widehat{\mathrm{Var}(\mathrm{b})}\mathrm{R}' \right)^{-1} (\mathrm{Rb} - \mathrm{q}) \to_d \chi^2(\dim(\mathrm{q}))$$

- Stata still reports results of t-test and F -test, which are asymptotically valid and may work better for moderate sample sizes. (But there is no guarantee.)

2.4 那些“伪”非线性模型：多项式、对数、交互项

Polynomial models

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- Marginal effect

$$\frac{\mathrm{dE}(y|x)}{\mathrm{d}x} = \beta_1 + 2\beta_2 x$$

  STATA 实现：lincom; margins

- 计算定点

$$\frac{-\beta_1}{2\beta_2}$$

  STATA 实现：nlcom.

Logarithmic models.

1. Linear-log model.

$$y = \beta_0 + \beta_1 \log x + \varepsilon, \beta_1 = \frac{\mathrm{d}y}{\mathrm{d}\log x} = \frac{\mathrm{d}y}{\mathrm{d}x/x}$$

2. Log-linear model.

$$\log y = \beta_0 + \beta_1 x + \varepsilon, \beta_1 = \frac{\mathrm{d}\log y}{\mathrm{d}x} = \frac{\mathrm{d}y/y}{\mathrm{d}x}$$

3. Log-log model.

$$\log y = \beta_0 + \beta_1 \log x + \varepsilon, \beta_1 = \frac{\mathrm{d}\log y}{\mathrm{d}\log x} = \frac{\mathrm{d}y/y}{\mathrm{d}x/x}$$

- When is log transformation appropriate?
  - To better approximate normal distribution when the residuals are strongly positively skewed (long right tail).
  - To attenuate heteroskedasticity when the variation in the residuals is proportional to the fitted values.
  - When residuals are believed to reflect multiplicatively (rather than additively) accumulating random errors.
  - To narrow the range of the variable so as to make estimates less sensitive to outliers.
  - To linearize a relationship.
  - When economic theory indicates.
- Rules of thumb for taking logs.
  - For positive pecuniary amount and head counts, log is often taken.
  - Variables measured in years and variables representing a proportion or a percentage usually appear in level forms.

Interaction effects

1. Difference in differences in means.

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + \varepsilon$$

2. Heterogeneity in slopes.

$$y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 (x \times D) + \varepsilon$$

3. Substitutes vs. complements.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2) + \varepsilon$$

• 互补与替代

例一：

$$\begin{aligned} &\text{采取合作模式的概率} \\ =\ &\beta_1\text{制度质量} + \beta_2\text{契约复杂性} + \beta_3\text{制度质量} \times \text{契约复杂性} \end{aligned}$$

$$\frac{\partial\text{采取合作模式的概率}}{\partial\text{制度质量}} = \beta_1 + \beta_3\text{契约复杂性}$$

$\beta_1 > 0, \beta_3 > 0$ 意味着制度质量越高的地区，企业越有可能采取合作模式；在契约复杂性越高的行业，制度质量对采取合作模式的**正面效应越强**。

$$\frac{\partial\text{采取合作模式的概率}}{\partial\text{契约复杂性}} = \beta_2 + \beta_3\text{制度质量}$$

$\beta_2 < 0, \beta_3 > 0$ 意味着契约复杂性越高的行业，企业越不可能采取合作模式；在制度质量越高的地区，契约复杂性对采取合作模式的**负面效应越弱**。

例二

$$\begin{aligned} & \text{出人才的概率} \\ = \ & \beta_1 \text{素质教育程度} + \beta_2 \text{计划经济色彩} \\ & + \beta_3 \text{素质教育程度} \times \text{计划经济色彩} \end{aligned}$$

$$\frac{\partial \text{出人才的概率}}{\partial \text{素质教育程度}} = \beta_1 + \beta_3 \text{计划经济色彩}$$

$\beta_1 > 0, \beta_3 < 0$ 意味着素质教育程度越高的地区，出人才的概率越高；而计划经济色彩越浓的地区，越推行素质教育越容易有暗箱操作空间，出人才的概率反而会下降。因此，计划经济色彩越浓的地区，要想出人才，越**不能强调**素质教育。

$$\frac{\partial \text{出人才的概率}}{\partial \text{计划经济色彩}} = \beta_2 + \beta_3 \text{素质教育程度}$$

$\beta_2 < 0, \beta_3 < 0$ 意味着计划经济色彩越浓的地区，出人才的概率越低；而素质教育程度越低的地区，反正只有高考一条独木桥，市场经济较之计划经济的优势比较有限。因此，素质教育程度越高的地区，要想出人才，越需要**削弱**计划经济色彩。

- Calculate the marginal effect at mean.

$$\frac{\partial y}{\partial x_1}|_{x_2=\bar{x}_2} = \hat{\beta} + \hat{\beta}_3 \bar{x}_2$$

We can transform the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 - \bar{x}_1) \times (x_2 - \bar{x}_2) + \varepsilon$$

and the new $\hat{\beta}_1$ is what we want.

- 分组回归还是交互项?
- Insignificance may indicate not the end of the world, but the light of hope: countervailing mechanisms or group-specific heterogeneities are waiting for your call (LU Ming).

Selection on Observables vs. Unobservables

- Key reference: Altonji et al. (2005, JPE).
- 当控制了最重要的控制变量之后，如果关键解释变量的系数估计不再随着更多控制变量的加入而发生大幅变化，那么就说明潜在的遗漏变量偏误可能很小了。
- Omitted variable bias formula. Suppose the structural model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

If we regress y on x only, then

$$\hat{\beta}_1 \to_p \beta_1 + \beta_2 \cdot \frac{Cov(x,z)}{Var(x)}$$

- 原理：控制了最重要的关键控制变量之后，其它控制变量与关键解释变量的相关性比较小，有理由相信，倘若存在遗漏变量，其与关键解释变量的相关性也比较小。
  示例. 私立学校的经济回报 (Dale and Krueger, 2002, QJE).

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
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

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	

- 不可观测变量的选择性偏误强度的测量指标：要把关键解释变量的效应全部归因于不可观测变量的选择性偏误，这一偏误（相对于可观测变量的选择性偏误）必须达到多大。

$$\frac{\hat{\beta}^F}{\hat{\beta}^R - \hat{\beta}^F}$$

其中 $\hat{\beta}^F$ 为包含全部控制变量回归的系数估计，$\hat{\beta}^R$ 为包含部分控制变量回归的系数估计。

分母越小，说明估计值受可观测变量的选择性影响越小，则不可观测变量（相对于可观测变量）的选择性必须更大才能完全解释整个效应；分子越大，说明需要解释的效应越大。

示例. 奴隶贸易与信任 (Nunn and Wantchekon, 2011, AER).

TABLE 4—USING SELECTION ON OBSERVABLES TO ASSESS THE BIAS FROM UNOBSERVABLES

Controls in the restricted set	Controls in the full set	Trust of relatives (1)	Trust of neighbors (2)	Trust of local council (3)	Intragroup trust (4)	Intergroup trust (5)
None	Full set of controls from equation (1)	4.31	4.23	3.03	4.13	3.32
None	Full set of controls from equation (1), ethnicity-level colonial controls, and colonial population density	11.54	6.98	2.65	9.22	3.80
Age, age squared, gender	Full set of controls from equation (1)	4.17	3.99	2.89	3.91	3.12
Age, age squared, gender	Full set of controls from equation (1), ethnicity-level colonial controls, and colonial population density	10.93	6.52	2.57	8.44	3.59

Notes: Each cell of the table reports ratios based on the coefficient for ln (1 + exports/area) from two individual-level regressions. In one, the covariates include the "restricted set" of control variables. Call this coefficient β^R. In the other, the covariates include the "full set" of controls. Call this coefficient β^F. In both regressions, the sample sizes are the same, and country fixed effects are included. The reported ratio is calculated as: $\beta^F/(\beta^R - \beta^F)$. See Table 3 for the description of the full set of controls from equation (1), the ethnicity-level colonial controls, and colonial population density.

更新的方法：Oster(2019). Unobservable Selection and Coefficient Stability: Theory and Evidence

$$Y = \beta X + \Phi w^0 + W_2 + \epsilon$$

w^0 是可观测控制变量，W_2 是不可观察控制变量，$\mathring{\beta}$ 是加可观测控制变量的估计结果，$\tilde{\beta}$ 是不加可观测变量的估计结果，β^* 是正确的结果，

$$\beta^* = \tilde{\beta} - [\mathring{\beta} - \tilde{\beta}]\frac{R_{\max} - \tilde{R}}{\tilde{R} - \mathring{R}}$$

$$\beta^* \approx \tilde{\beta} - \delta[\mathring{\beta} - \tilde{\beta}]\frac{R_{\max} - \tilde{R}}{\tilde{R} - \mathring{R}}$$

```
eststo a11: xtreg leverage bepu_soe soe_own $controllist i.year#i.sector, fe cluster(stkcd)
disp e(r2_w)
disp e(r2_w)*1.3
*R2 = 0.1848
local a = e(r2_w)*1.3
psacalc delta bepu_soe, rmax(`a') mcontrol(separation boardnum indboard_ratio largestholderrate  presidentisceo)
```

2.5 面板数据模型

- Time-constant omitted variables.

$$y_{it} = \mathbf{x}_{it}'\beta + v_{it} = \mathbf{x}_{it}'\beta + u_i + \varepsilon_{it}, t = 1, 2, \cdots, T$$

- Pooled OLS. View the panel as a "cross section"with $n \times T$ observations.

$$\mathrm{b}_{POLS} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{it}\mathbf{x}_{it}' \right)^{-1} \left(\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{it} y_{it} \right)$$

When is pooled OLS valid?

$$\mathrm{E}(\mathbf{x}_{it}, v_{it}) = 0, \;\; i.e. \mathrm{E}(\mathbf{x}_{it}\varepsilon_{it}) = 0 \text{ and } \mathrm{E}(\mathbf{x}_{it}u_i) = 0$$

- Key assumption for fixed-effects models:

$$\mathrm{E}(\mathbf{x}_{is}\varepsilon_{it}) = 0, \; s, t = 1, \cdots, T$$

u_i and X_i can be correlated in any arbitrary form.

• Within transformation.

$$\bar{y}_i = \bar{\mathbf{x}}_i'\beta + u_i + \bar{\varepsilon}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$, $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$

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

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}'\beta + \ddot{\varepsilon}_{it}$$

Fixed-effects estimator.

$$\mathrm{b}_{FE} = \left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}\left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right) = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}\ddot{\mathbf{x}}_{it}'\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}\ddot{\mathrm{y}}_{it}\right)$$

$$\widehat{\mathrm{Var}(\mathrm{b}_{FE})} = \left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}\left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\hat{\ddot{\varepsilon}}_i\hat{\ddot{\varepsilon}}_i'\ddot{\mathbf{X}}_i\right)\left(\sum_{i=1}^{n} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}$$

- STATA's robust option for panel data is automatically cluster(id). However, we may want to cluster at more aggregate levels.
- Stock and Watson (2008, ECMA) suggest that the right t-statistic testing one element of β is $\sqrt{\frac{n}{n-1}}t_{n-1}$, and the F -statistic testing p elements of β is $\left(\sqrt{\frac{n}{n-p}}\right)F_{p,n-p}$. STATA does not seem to have adjusted it.

- Least squares dummy variable (LSDV) regression.
  Traditionally we view u_i as parameters rather than random variables.

$$y_{it} = \mathbf{x}_{it}'\beta + \sum_{j=1}^{n} \delta_j D_{it}^j + \varepsilon_{it}, D_{it}^j \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

  – One can easily show that $\mathrm{b}_{LSDV} = \mathrm{b}_{FE}$.
  – From the FOCs of LSDV regression, we have

$$\hat{u}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\mathrm{b}_{FE}$$

    The "estimates"of fixed effects are inconsistent (because consistency would rely on large-T asymptotics, and LSDV regression masks this fact).
  – Estimate the constant in the fixed effects model.

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i$$

  – How to make in and out of sample prediction?

• First-differencing transformation.

$$\Delta y_{it} = \Delta \mathrm{x}_{it}'\beta + \Delta\varepsilon_{it}, \quad t = 2, \cdots, T$$

$$\begin{aligned} \mathrm{b}_{FD} &= \left(\sum_{i=1}^{n} \Delta \mathrm{X}_i' \Delta \mathrm{X}_i \right)^{-1} (\Delta \mathrm{X}_i' \Delta \mathrm{y}_i) \\ &= \left(\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta \mathrm{x}_{it} \Delta \mathrm{x}_{it}' \right)^{-1} \left(\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta \mathrm{x}_{it} \Delta y_{it} \right) \end{aligned}$$

• Comments.

– In both FE and FD, parameters are identified off the variation within groups over time, not between groups. This variation may require justification.

– FE and FD are equivalent when $T = 2$.

– Both FE and FD require ε_{it} to be uncorrelated with lagged, con- temporaneous, and future x. If we maintain only contemporaneous exogeneity, i.e., $\mathrm{E}(\mathrm{x}_{it}\varepsilon_{it}) = 0$, FE is generally less inconsistent than FD when T is large. That's why FE is more popular in estimating static panel data models.

– It does not make sense to instrument an explanatory variable by its lags in fixed-

effects model.

– One shortcoming for fixed-effects models is that, if the variable of primary interest has little variation across time, FE or FD will be very imprecise. In this case, we may want to estimate time-specific slopes for this variable.

- In practice we usually include in the equation both individual-specific effects and time-specific effects, i.e., two-way models.
- How to construct rich controls of fixed effects?

reghdfe 和 xtreg 如何选择?

为什么更加推荐 reghdfe?

由于 reghdfe 背后应用的是 FWL 定理，先用多个固定效应组成的消灭矩阵（annihilate matrix）对 y 和 x 做第一步处理，如果 y 或者 x 和固定效应完全共线，那么会被直接 drop 掉。因此也就不可能估计出相应的系数。以本文估计的固定效应模型为例（见下面的代码），直接就被 drop 掉了。(《从一篇论文的错误开始吐槽》)

高维固定效应

```
reghdfe y x1 x2 , absorb(province # year)
```