计量经济分析

第一章　导论

徐秋华，xuqh@swufe.edu.cn

西南财经大学，金融学院

大纲

1. 什么是计量经济学
2. 计量经济学的两个公理
3. 经济学的数据类型
4. 因果关系

什么是计量经济学

定义（非正式）

- “econometrics”，“econo”即为经济，而“metric”的含义是测度
- 用定量的方式分析和检验经济学中定性理论的一门学科
- 任何有助于对经济问题进行分析和检验的定量方法都属于计量经济学的研究范畴
- 重要任务是结合经济学问题的特点改进已有定量方法（如数理统计）和发明新的定量方法
- “economic-tricks”，经济学家用来证明他们想要证明的结论时使用的“花招”
- “花招”也好，“方法”也罢，都说明了计量经济学的工具性质

Econometrica 创刊

国际计量经济学会（Econometric Society）的创立者欧文 · 费雪（Irving Fisher）指出“..... 经济学的定量方法有几个方面，这些方面中的任何一个都不应该与计量经济学相混淆。因此，计量经济学绝不等同于经济统计学。它也不等同于一般的经济理论，尽管这种理论的相当一部分具有明确的数量特征。计量经济学也不应被视为数学的经济学应用的同义词。经验表明，看待计量经济学的这三种观点，即统计学、经济理论和数学，其中的任何一种对于真正理解现代经济生活中的数量关系来说，都是必要的，但不是充分的。这三者的结合是功能强大的，也正是这三者的结合构成了计量经济学。”

应用举例

1. 通过一组个体的工资、受教育年限、年龄以及工龄等其他个体特征的数据，分析受教育年限对对数工资的影响；
2. 利用沪深 A 股市场的股票数据，分析我国的融资融券交易制度对股市波动率的影响；
3. 基于上证指数的历史数据预测其未来的波动率；
4. 基于某银行贷款申请以及违约情况的历史数据预测某贷款申请者在未来是否违约；
5. 检验资本资产定价模型（CAPM）在我国 A 股市场的有效性。

计量经济学的主要用途：对经济变量之间的因果关系进行推断、对经济变量进行预测以及对经济理论进行检验。

主要的计量方法：[Angrist and Pischke, 2008]

- 将可能影响因果推断的变量进行有效控制的多元回归模型；
- 分析真实实验和自然实验的工具变量方法；
- 利用重复观测控制遗漏变量的双重差分（differences-in-differences, DID）策略。

此外，还有以向量自回归（Vector Autoregression, VAR）模型为代表的时间序列方法。

“大数据”时代前所未有的挑战

- [López De Prado, 2018]：“计量经济学是经济学和金融学在过去 70 年没有取得有意义的进展的主要原因。”
- All models are wrong, but some are useful.
- 有监督 (supervised) 机器学习提供了解决高维预测问题的工具，主要关注的是以解释变量 $\boldsymbol{X}$ 为条件被解释变量 Y 的条件期望 $E[Y|\boldsymbol{X}]$ 或条件分布 $p(Y|\boldsymbol{X})$。
- 传统计量经济学也关心预测问题，但是 $\boldsymbol{X}$ 的维度通常较低。
- 因果关系：控制其他因素不变！

计量经济学的两个公理

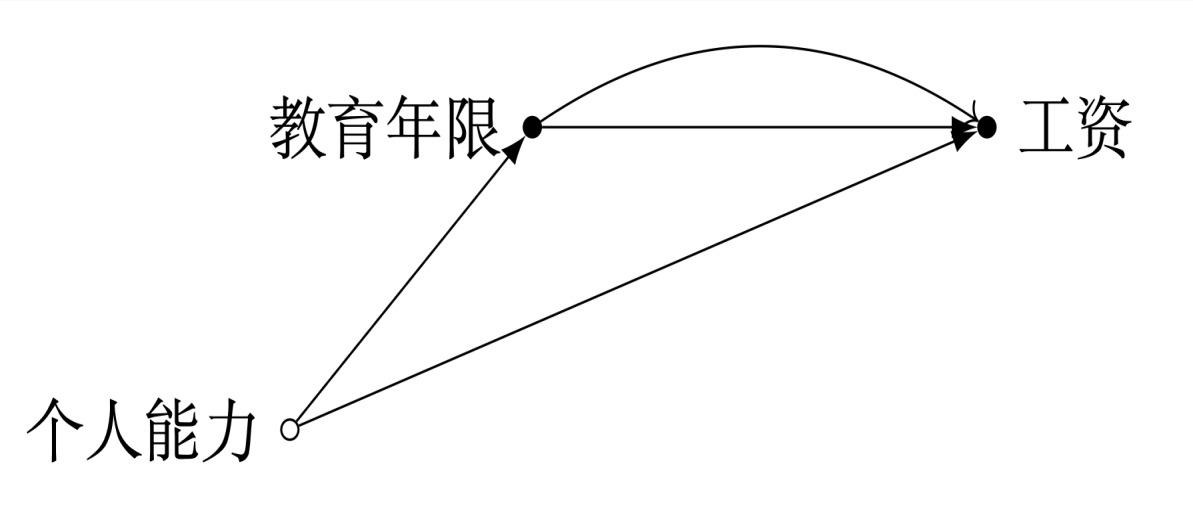
公理

计量经济学的两个公理

- 任何经济体都可以被视为服从某一概率法则的随机过程；
- 以数据形式概括的各种经济现象都可以被视为上述随机数据生成过程（data generating process, DGP）的实现。

对公理的解读

- 现代计量方法的重要特征是将解释变量和被解释变量的观测值理解为随机变量的实现
- 将解释变量视为随机变量为分析其和误差项的相关性提供了可能。
- 公理中的概率法则指解释变量和被解释变量的联合概率分布
- 计量经济学中所谓的模型是对此联合概率分布施加的一组假设



参数模型

若记 $\boldsymbol{W} = (Y, \boldsymbol{X})$，参数模型的完整设定是给出关于联合概率分布 $f_J(\boldsymbol{W} \mid \theta)$ 的具体描述，[1] 其中，θ 是 $f_J(\cdot)$ 中涉及的所有参数构成的向量。由于联合分布中涉及解释变量 $\boldsymbol{X}$ 的分布，因此，允许解释变量随机似乎增加了模型设定的难度。但是，如果 θ 可以被划分为彼此无关的两部分 θ_1 和 θ_2，并且满足：[2]

$$f_J(\boldsymbol{W} \mid \theta) = f_C(Y \mid \boldsymbol{X}, \theta_1) \times f_M(\boldsymbol{X} \mid \theta_2), \tag{1}$$

则对 θ_1 的统计推断不需要 $f_M(\boldsymbol{X} \mid \theta_2)$ 的相关信息，而 θ_1 恰恰是分析 $\boldsymbol{X}$ 与 Y 之间关系时要估计的参数。因此，计量经济学的模型假设中常常出现以 $\boldsymbol{X}$ 为条件。

---

[1] 下标“J”是英文单词 joint 的首字母，用来表示联合分布。类似地，公式 (1) 中的下标“C”和“M”分别表示条件（conditional）分布和边缘（marginal）分布。

[2] 这里用到了联合分布等于条件分布与边缘分布的乘积。

误差项

- 公理的另一个重要含义是用确定性的函数为经济变量之间的关系进行建模是不可行的。
- 没有任何模型可以涵盖复杂经济现象背后的随机性的方方面面。
- 研究者只能对经济变量之间的随机关系做一些假设，并在这些假设的约束下利用观测数据对模型涉及的参数做统计推断。
- 例如，线性模型只是对变量之间复杂随机关系的一种线性近似。为了体现变量之间关系的随机性，需要在线性关系基础上加入随机误差项 ε_i。

经济学的数据类型

数据类型

计量实证分析使用的数据集主要分为四种类型，它们分别是：

- 横截面（cross-sectional）数据
- 时间序列（time series）数据
- 面板（panel）数据
- 聚类（clustered）数据

这四种不同类型数据的主要分类依据是不同观测之间的相依结构。

横截面数据

- 横截面数据是指在给定的时间点上由不同个体的各个变量的样本观测值构成的数据。
- 例如，由 2020 年我国各个省份（直辖市、自治区）的第四季度 GDP 和地方财政一般预算收入构成的数据集便是横截面数据。
- 即使收集各省数据时存在时间上的差别（如，2020 年第四季度的不同星期），仍然可以将其看作是横截面数据。
- 横截面数据的特点是个体的排序不会影响计量分析的实证结果。
- 通常假设横截面数据是从总体中通过随机抽样得到的。
- 随机样本的重要性质是不同观测个体之间满足独立同分布(independent and identically distributed, IID)。
- 但并非所有的横截面数据都可以被视为随机样本。例如，表 1 包含了我国所有省份（直辖市、自治区）的相关数据，因此，它更像是一个总体而不是随机样本。

表 1: 2020 年我国各省份（直辖市、自治区）GDP 和地方财政一般预算收入（单位：亿元）

观测序号	年份	省（直辖市、自治区）	GDP	地方财政一般预算收入
1	2020	北京市	35943.25	5483.89
2	2020	天津市	14007.99	1923.11
3	2020	河北省	36013.84	3826.46
4	2020	山西省	17835.58	2296.57
5	2020	内蒙古自治区	17258.04	2051.20
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
30	2020	宁夏回族自治区	3956.34	419.44
31	2020	新疆维吾尔自治区	13800.74	1477.22

时间序列数据

- 时间序列数据是指由同一个体的各个变量在时间维度上的观测值构成的数据。
- 例如，由贵州茅台（600519.SH）2020 年所有交易日的开盘价、收盘价和成交量构成的数据集是时间序列数据。
- 因为时间具有一个天然的先后次序，所以，表 2 中的第一列不能随意的换序。
- 展示某变量的时间序列数据的最好方式是画时间序列图，其横轴为时间，纵轴为该变量在不同时间点的观测值。
- 时间序列数据的主要特征是存在时间维度上的相依关系。通常，某一时刻的数据会和邻近时刻的数据之间存在相依性。
- 若这种相依性体现为变量间的相关关系，则称时间序列存在序列相关（serial correlation）。

表 2: 2020 年贵州茅台的日开盘价（元）、收盘价（元）和成交量（万股）

观测序号	日期	开盘价	收盘价	成交量
1	2020-01-02	1128.00	1130.00	1480.9916
2	2020-01-03	1117.00	1078.56	1303.1878
3	2020-01-06	1070.86	1077.99	634.1478
4	2020-01-07	1077.50	1094.53	478.5359
5	2020-01-08	1085.05	1088.14	250.0825
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
242	2020-12-30	1870.00	1933.00	344.5210
243	2020-12-31	1941.00	1998.00	388.6007

图 1: 贵州茅台 2020 年日收盘价的时间序列图

许多月度或季度的总量时间序列数据还可能表现出明显的季节(seasonal)趋势。在对时间序列数据进行分析之前，通常需要对数据中的季节趋势进行处理。

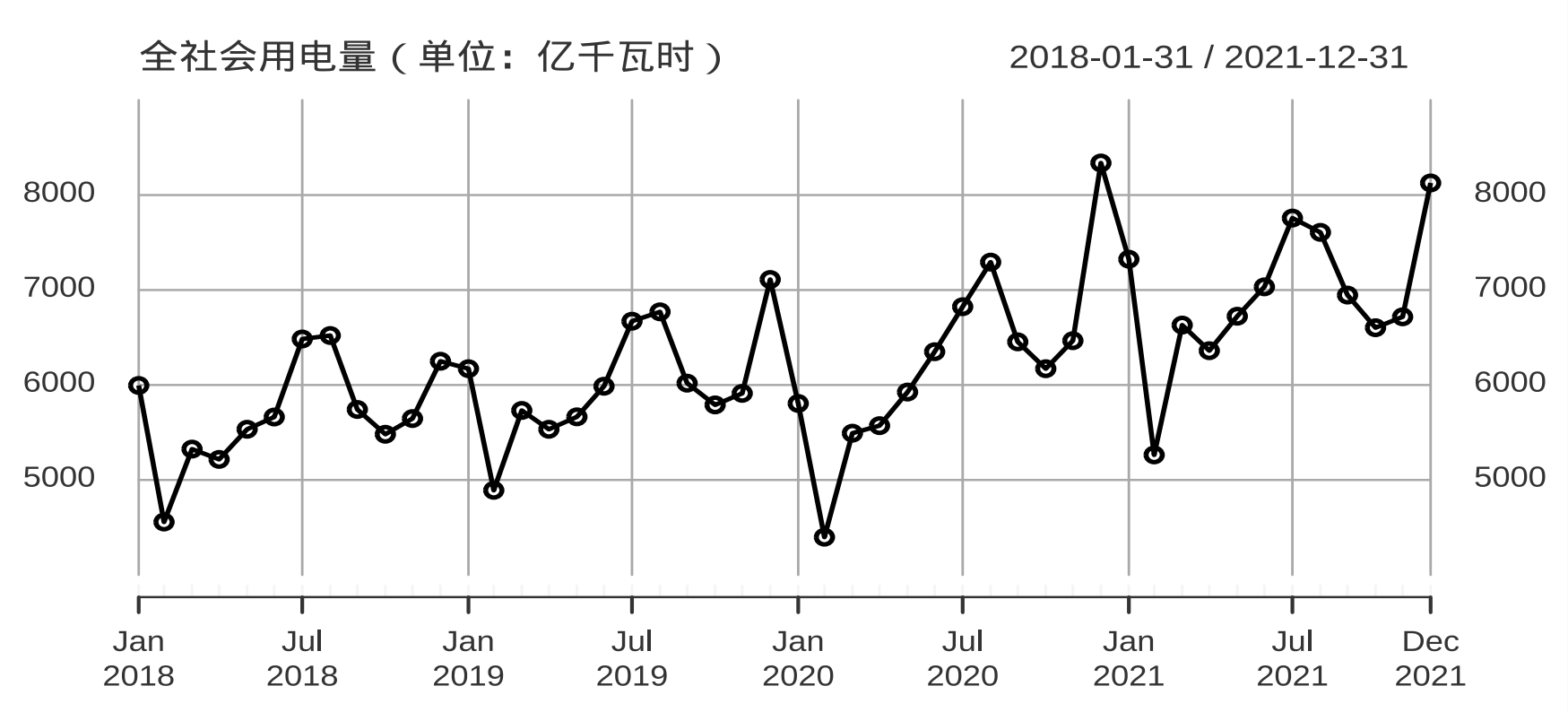


图 2: 我国 2018 年至 2021 年全社会用电量月度时间序列图

面板数据

- 面板数据，也称为纵向（longitudinal）数据，是由数据集中的每一个横截面个体的时间序列数据组合而成，并且满足不同个体的时间序列具有相同的时间点。
- 例如，由 2019 和 2020 年我国各个省份（直辖市、自治区）的第四季度 GDP 和地方财政一般预算收入构成的数据集是面板数据。
- 在存储面板数据时，需要用两列分别记录时间和个体名称。
- 与横截面数据类似，不同个体在数据集中的排序对数据分析没有影响，但是，同一个体的时间序列数据要根据时间顺序排列。
- 面板数据中每个个体的时间序列数据要集中排列。
- 在分析面板数据时通常假设横截面个体之间彼此独立，但是给定某一个体的各个时间点的时间序列数据是彼此相依的。
- 与横截面或时间序列数据相比，面板数据可以有效地控制个体异质性、提高自由度，并可以提取横截面或时间序列数据无法获得的动态信息。

表 3: 2019-2020 我国各省份（直辖市、自治区）GDP 和地方财政一般预算收入（单位：亿元）

观测序号	年份	省（直辖市、自治区）	GDP	地方财政一般预算收入
1	2019	北京市	35445.13	5817.10
2	2020	北京市	35943.25	5483.89
3	2019	天津市	14055.46	2410.41
4	2020	天津市	14007.99	1923.11
5	2019	河北省	34978.55	3738.99
6	2020	河北省	36013.84	3826.46
7	2019	山西省	16961.61	2347.75
8	2020	山西省	17835.58	2296.57
9	2019	内蒙古自治区	17212.53	2059.69
10	2020	内蒙古自治区	17258.04	2051.20
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
59	2019	宁夏回族自治区	3748.48	423.58
60	2020	宁夏回族自治区	3956.34	419.44
61	2019	新疆维吾尔自治区	13597.11	1577.63
62	2020	新疆维吾尔自治区	13800.74	1477.22

混合（pooled）横截面数据

- 容易与面板数据混淆的一种数据类型被称为混合（pooled）横截面数据，它将变量在不同时间点的横截面数据按时间顺序混合在一起。
- 不同时间点的横截面个体可以不同。例如，将西南财经大学金融科技专业 2019 年和 2020 年的学生金融计量学期末考试成绩混合在一起构成的数据集是混合横截面数据，除了少数重修的学生之外，两年参加考试的学生是不同的。
- 在存储混合横截面数据时，同一时间点对应的横截面数据要集中排列。

面板数据的分类

- 横截面个体数量 N 较大，时间序列时间点数量 T 较小的面板数据被称为短面板数据；
- N 较小，T 较大的面板数据被称为长面板数据；
- 如果所有 N 个个体在 T 个时间点的全部数据都是可以观测的，则称这样的面板数据为平衡面板（balanced panel）数据；
- 如果至少一个个体的某一时间点的数据缺失，则称这样的面板数据为非平衡面板（unbalanced panel）数据。

聚类数据

- 对于横截面数据，通常假设不同个体之间是满足 IID 的。但是，如果个体之间出现了某种关联性，则 IID 假设不再成立。
- 例如，同一个学校内的不同学生的学习成绩，或者同一个行业内的不同企业的盈利状况，都会由于一些共同因素导致个体之间出现某种程度的关联性。
- 刻画横截面个体之间的相关性有它固有的难点。这与时间序列数据是不同的，时间序列数据具有时间这一天然的顺序，时间间隔较长的数据彼此之间可以认为是渐近不相关的。
- 允许横截面个体存在相关性的常见数据类型是聚类数据。此类数据将不同的横截面个体划分到不同的聚类中，不同的聚类彼此之间相互独立，但是同一聚类内部的个体可以具有任意形式的相依关系。
- 分析聚类数据的计量方法通常不直接为横截面个体的相关关系建模，而是在考虑了这种相关性的情况下设计稳健（robust）的统计推断方法。

因果关系

因果关系

- 在大多数经济学的实证分析中，研究者的目标是推断一个变量对另一个变量是否具有因果意义上的影响。
- 虽然简单地发现两个变量之间的相关关系为进一步分析它们之间的因果关系提供了可能性
- 但是，并非所有的相关关系都正确地揭示了变量之间的因果关系

辛普森悖论（Simpson's Paradox）

考虑如下某种新研制的药物的存活率实验数据（见 [Pearl, 2009]）：

表 4: 某药物存活率的实验数据

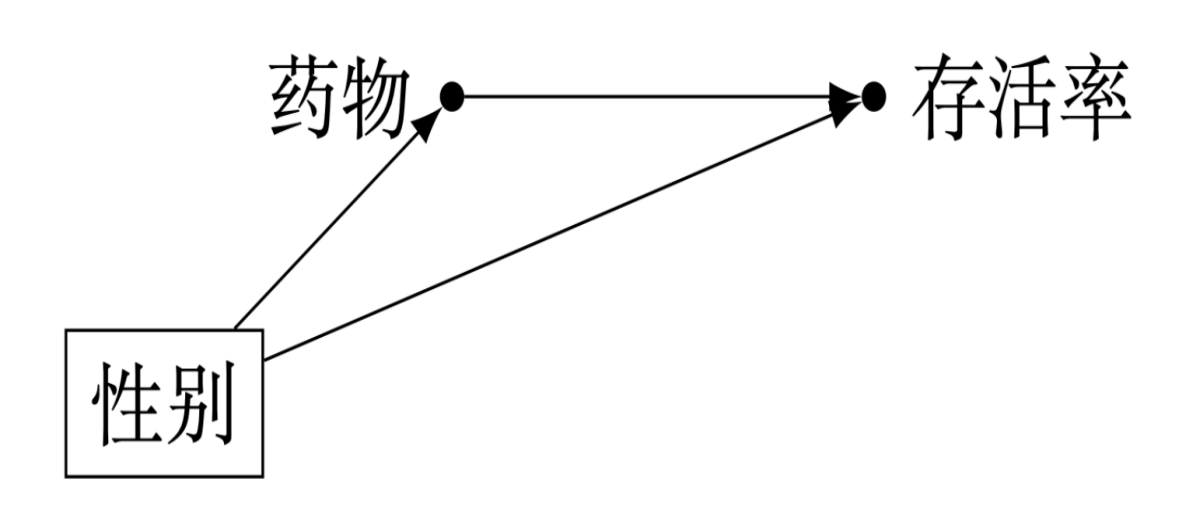
全体			
	存活	死亡	存活率
处理组	20	20	50%
控制组	16	24	40%
男性			
	存活	死亡	存活率
处理组	18	12	60%
控制组	7	3	70%
女性			
	存活	死亡	存活率
处理组	2	8	20%
控制组	9	21	30%

表 4 中的数据显示，在整个处理组加控制组的人群中，处理组具有较高的存活率（50%），说明新研制的药物具有“正向”作用。然而，无论对于男性群体还是女性群体，新研制的药物都表现出了“负向”的作用，即降低了存活率。这种现象被称为辛普森悖论（Simpson's Paradox）。

对例子的分析

- 如果只关注男女合并的群体数据，就会得到服用新药物可以提高存活率的错误结论。
- 前文中提到建立因果关系的核心是“控制其他因素不变”，服用新药物与存活率之间的正相关关系只有在有效控制了所有其他可能影响药物服用和存活率的因素之后才可以揭示两者之间的真正因果关系。
- 在例子中，更多的男性在处理组，而更多的女性在控制组，因此，性别与谁服用新药物的分配机制有关。
- 同时，女性的存活率明显低于男性的存活率，因此，性别也与存活率有关。

性别变化可以改变个体进入处理组（服用新药物）的概率，也会影响存活率。于是，性别的变化会导致服用新药物和存活率之间出现除因果关系之外的相关性。



性别被称为是推断服用新药物和存活率之间因果关系的“混杂(confounding)因素”。“控制其他因素不变”就是要控制住这种混杂因素，只有这样才可以发现变量间真正的因果关系。

政策评估

- 推断因果关系在经济学中的重要应用是对经济政策实施之后的政策效果进行评估。政策评估通常关心的是二值政策处置变量 D 对结果变量 Y 的因果效应，[3] 被称为处置效应（treatment effect）。

- 例如，评估融资融券交易制度对我国股市波动率的影响，D 表示是否加入融资融券标的股，Y 表示资产的波动率。正确地识别政策的处置效应依赖于真实结果与反事实（counterfactual）结果的比较。对于 $D = 1$ 的个体，反事实结果是其在 $D = 0$ 的情况下出现的结果；类似地，对于 $D = 0$ 的个体，反事实结果是其在 $D = 1$ 的情况下出现的结果。

[3] $D = 0$ 表示未接受政策处置，$D = 1$ 表示接受了政策处置。

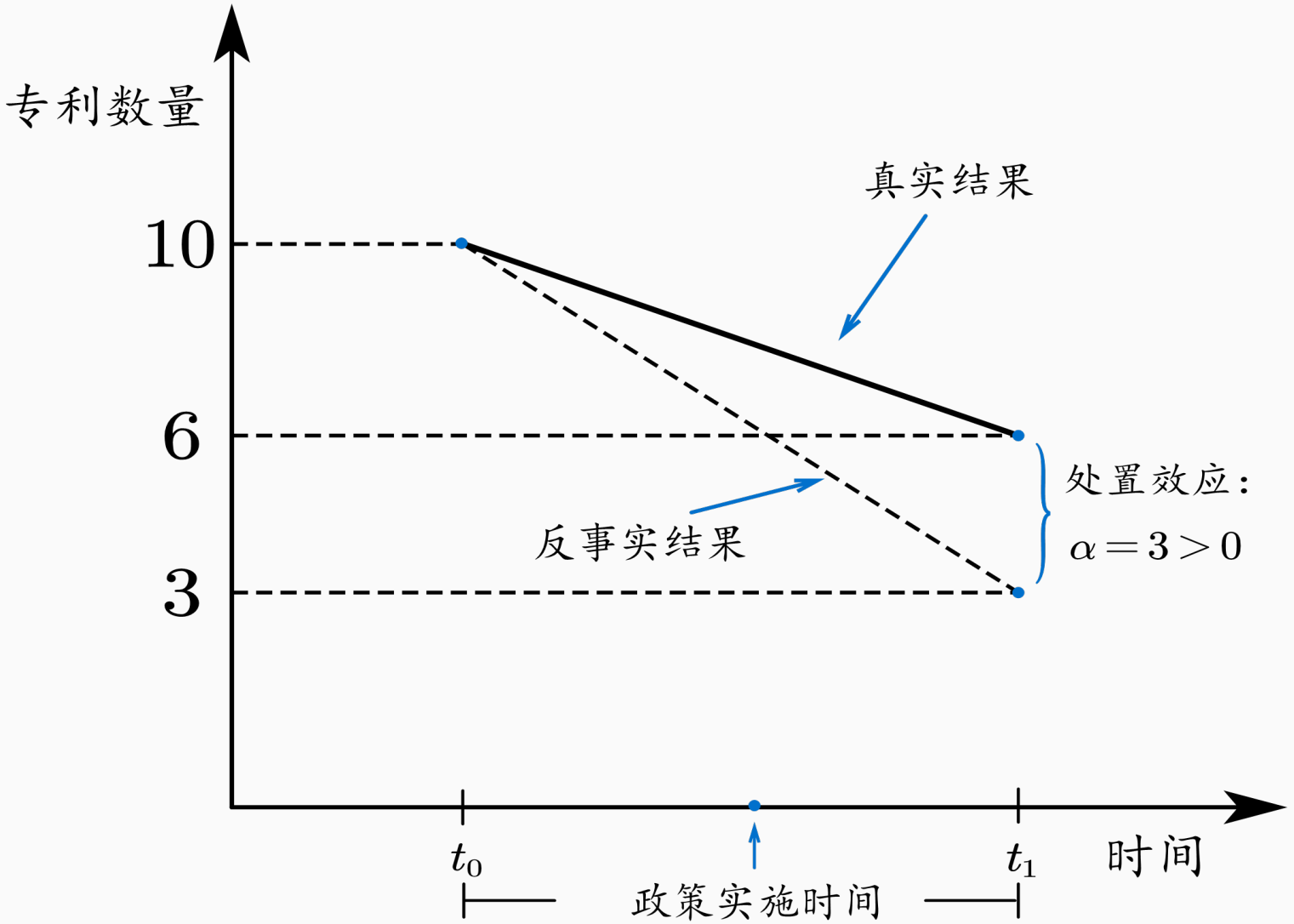


图 3: 激励企业增加专利数量的政策效果

“控制其他因素不变”与“反事实结果”

事实上，反事实比较体现了因果推断中“控制其他因素不变”的思想。真实结果与反事实结果是同一个体在两种不同的政策处置状态下呈现出的结果，因此，除了 $D = 1$ 和 $D = 0$ 的区别之外，所有其他因素都保持不变。因果推断的难点在于反事实结果是无法观测的，基于观测样本构造反事实结果是计量经济学的重要任务。

Questions?

References i


Angrist, J. D. and Pischke, J.-S. (2008).
***Mostly harmless econometrics: An empiricist's companion*.**
Princeton University Press.


López De Prado, M. (2018).
***Advances in financial machine learning*.**
John Wiley & Sons.


Pearl, J. (2009).
***Causality*.**
Cambridge University Press.