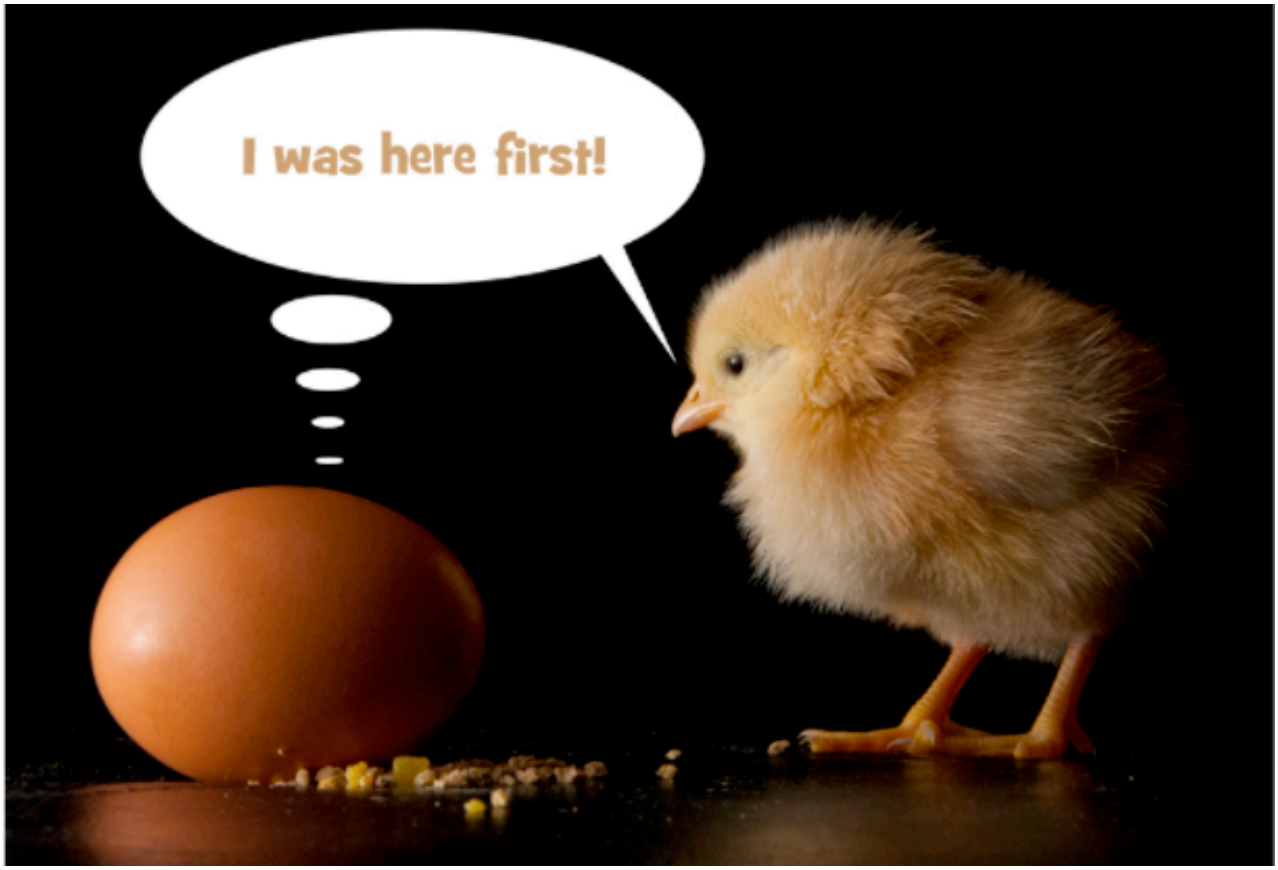
鸡生蛋还是蛋生鸡，数据分析之因果推断

Thursday, February 1, 2018
11:28 PM



先有鸡还是先有蛋这个因果困境想要表达的是一个“到底是先有蛋，还是先有鸡”的问题，谁是因，谁是果？

有人认为这个问题本身毫无意义，

不管谁生谁，我们有鸡又有蛋吃就对了。

这个问题不能当饭吃，鸡和蛋，倒是可以吃。

但是，这个问题引起我们去思考一些更重要更普遍的问题。
毕竟生活里除了吃喝，还有很多有重要又有趣的问题，等待我们去发现。

“蛋生鸡，鸡生蛋”的问题，指向数据分析里面一个重要的分支，

因果推断

对于一些我们观察到的现象，我们需要解释，数据分析里面的因果推断就是来帮助我们解释谁是因，谁是果。

因果推断，有啥用？
太有用了！
来来，给你举个例子：

举个例子来说

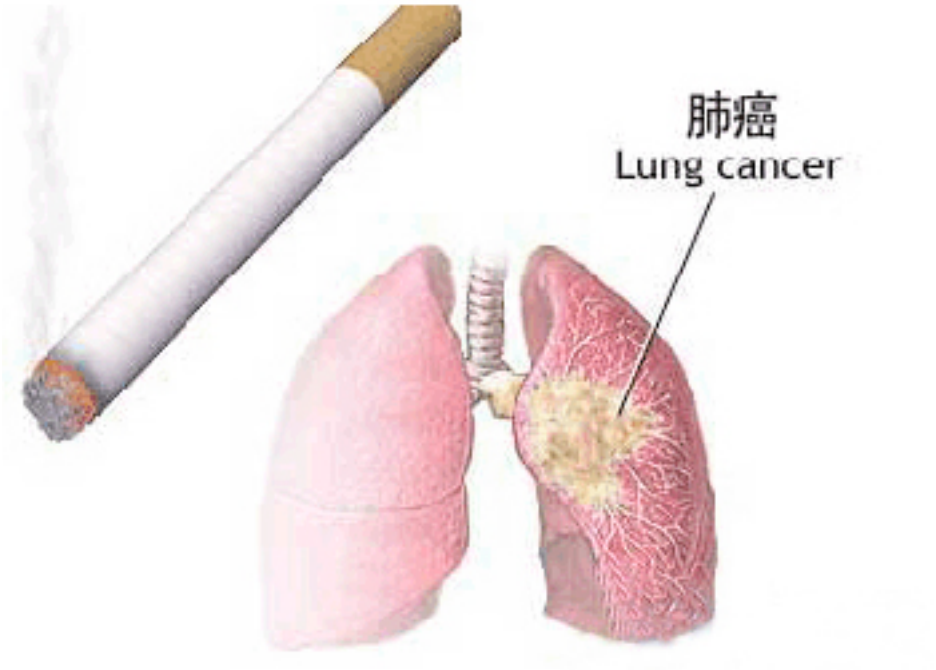
假设你是一个天外来客，一个胖子，造访地球，发现地球人崇尚苗条的身材，
你于是决定寻找一种方法变瘦，然后你碰到了她，认为瘦子都是吃出来的。
然后你照着做.....

很多场景下，我们不仅仅需要预测，更需要解释。

现在来点正经的：

一个很经典的问题：吸烟是否导致肺癌？

我们观察到很多得肺癌的人是吸烟的人，吸烟与肺癌正相关。



那我们能得出结论：“吸烟导致肺癌”吗？

毛，邓，周

这是因为可能存在一些未观测的因素，他既影响个体是否吸烟，同时影响个体是否得癌症。比如，某些基因可能使得人更容易吸烟，同时容易得肺癌；存在这样基因的人不吸烟，也同样得肺癌。此时，吸烟和肺癌之间相关，却没有因果作用。

相反的，我们知道放射性物质对人体的健康有很大的伤害，但是铀矿的工人平均寿命却不比常人短；这是流行病学中有名的 “健康工人效应”（healthy worker effect）。这样一来，似乎是说铀矿工作对健康没有影响。但是，事实上，铀矿的工人通常都是身强力壮的人，不在铀矿工作寿命会更长。此时，在铀矿工作与否与
寿命不相关，但是放射性物质对人的健康是有因果作用的。

好，我接受因果推断很重要，那有什么具体方法，才能排除干扰，确实得出因果关系呢？

推荐一本书，功夫计量（Mastering Metrics）。

随机试验 (randomized experiment)

介绍一篇经典：

Angrist, J. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review, 80*(3), 313-336. Retrieved from http://www.jstor.org/stable/2006669

军队服役与终身收入的关系。

负相关，但是不能推断当兵导致经济能力变弱。

俗话说，好男不当兵。

美国警察，很多都是退役士兵，高中文化。

不要惹警察，一言不合就拔枪。

没有直接证据，直到这篇文章。

- 近似随机试验，越战，美国军队抓壮丁的事儿
- 类似汽车尾号限行

随机试验的关键本质:

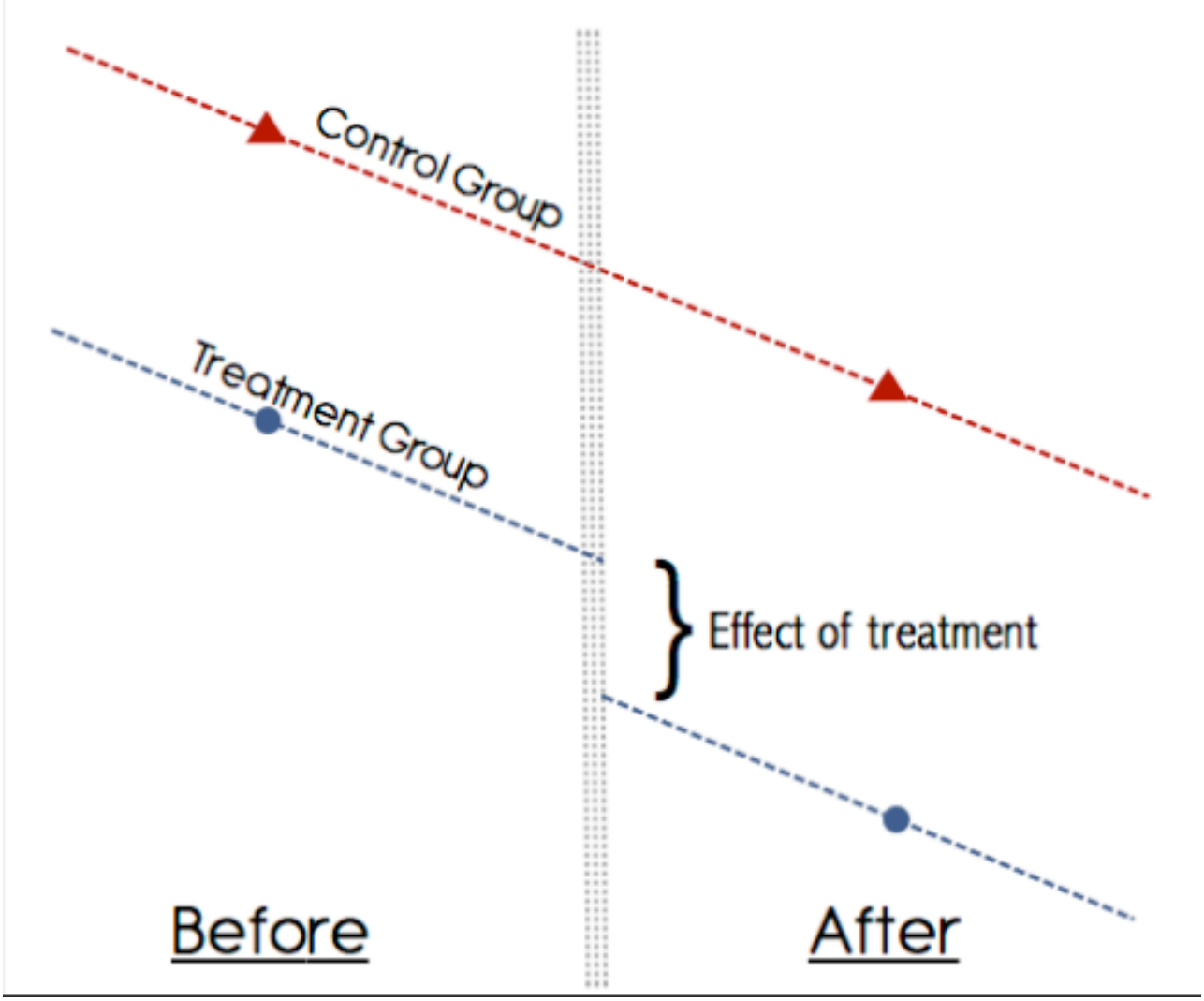
对照组，其他条件不变（Ceteris paribus）。

那另外一个问题来了，很多时候，我们没有办法做随机试验，成本太高，或者有违道德。此时我们要做因果推断，

经济学家有办法，

双重差分 (Difference in Difference)：

不必其他条件不变，只要干预之前，两组的差异差不多。例如上补习班，或者是再就业培训的效果评估，等等。



双重差分吗，就是差分两次。

需要两期数据：

首先

- 算出施加政策前，二者的差距
- 算出施加政策后，二者的差距

然后

- 将两个差距相见，得出政策效果的评估结果。

公式在这里：

$$Y_{it} = \alpha D_i + \beta T + \gamma\left(D_i \times T\right) + u_{it}$$

断点回归（regression discontinuity）：

- 上好大学的回报

- “淮河 RD - 空气污染 - 北方人预期寿命少 5.5 年”

  Life expectancy and air pollution in China，

  Yuyu Chen, Avraham Ebenstein, Michael Greenstone, Hongbin Li

  Proceedings of the National Academy of Sciences Aug 2013, 110 (32) 12936-12941; DOI:10.1073/pnas.1300018110

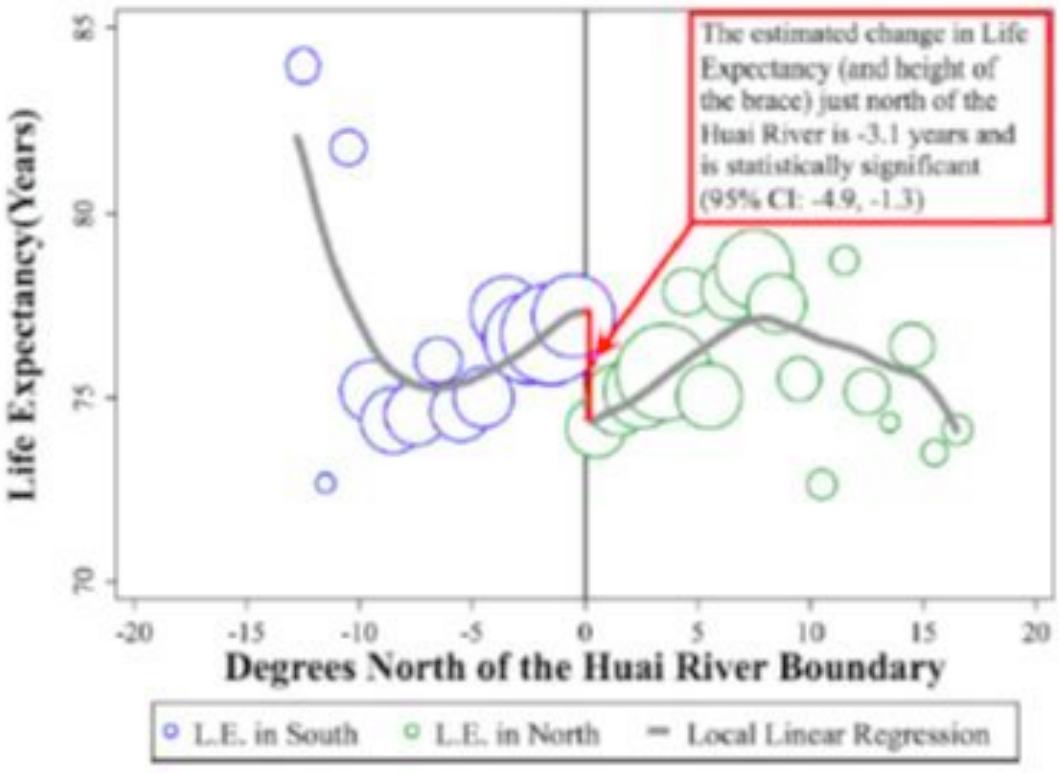


Fig. 3. Fitted values from a local linear regression of life expectancy (L.E.) on distance from the Huai River estimated in the same manner as in Fig. 2.