第二讲. 案例及回归简介

2014.2.20

回顾

- 关系：关联、因果
  - 关联：表面关系，提供因果的线索；
  - 因果：深层次、本质的关系，科学探索的目标；
  - 推断因果的条件：***ceteris paribus***。
- 研究设计：随机化控制试验、观察研究
  - 随机化试验：推断因果的金方法，但常常无法实施；
  - 观察研究：更为常见，但存在其它因素干扰因果推断；
- 回归：提供了从观察研究中发现因果和预测的工具。

案例1：霍乱病源推断（教材第一章）

19世纪中期，人类对霍乱的传播几乎一无所知， 细菌学说只是众多理论中的一种。 **1855**年伦敦医生**John Snow** 利用通过精巧的分析发现霍乱**(cholera)** 是一种通过饮用水传染的疾病。

- **1848**年，伦敦爆发了霍乱。**Snow**找出了第一个病例**Harnold**：他是刚从霍乱流行的**Hamburg**乘船到伦敦的海员。并且发现第二例病人住过**Harnold**曾经住过的房间。

  这表明霍乱可能具有传染性

- 然后发现附近的两个公寓，一个公寓发生了霍乱，而另一个没有，而前者的饮水系统被污染了，而另外一个没有。

  这表明霍乱可能是通过饮用水传染的。

- 1854年伦敦又爆发了霍乱。Snow在地图上标识了疾病发生区域。很多病例集中在Broad Street的供水泵附近。

- 但该区域也有病例很少的地方：比如一个酿酒厂，该厂的工人习惯于喝麦芽酒，而且该厂有自己的供水系统；救济院(poor-house)也是如此。另外，Snow发现附近的零星病例，大多与Broad Street有关：例如，一个霍乱病例不住在Broad Street，但喜欢到Broad Street取水。

- 然后Snow使用统计方法做了生态学研究。他注意到伦敦有两大供水公司：
  - Southwark and Vauxhall (SV)公司：水源在泰晤士河下游，污染较重，
  - Lambeth 公司：水源在上游，污染不重。

Table 2. Death rate from cholera by source of water. Rate per 10,000 houses. London. Epidemic of 1854. Snow's table IX.

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

p=0, 显著不同。霍乱死亡率与供水公司有关联，而两家公司水质不同。是否水质差导致霍乱呢？

- Snow进一步研究发现：
  - 两个公司在伦敦的某些地区的供水并没分开，而是混在了一起，比如同一个房子的两侧两家可能选用不同的公司。
  - 而且两个公司的价格、服务各方面差异不大，用户一般不知道SV的水质被污染。各家选用供水公司几乎是随机的：不依赖于贫富、房子大小、房主的职业、所处位置等。

⇒

虽然是观察研究数据，但 Nature在该地区做了一个随机试验（称为天然试验）。比较该地区两种客户的发病率得到的关联只能归因于所属公司的不同。而公司的不同主要是水质。

所以，可以断言：饮用水传播霍乱

案例2：HIP trial (随机化试验)

- 乳腺癌在北美妇女中较为常见. Mammograph 是一种X光筛查方法,以期早期发现。 Mammograph 有用吗？第一个大型随机化控制试验在纽约进行(1960's)。

- HIP (Health Insurance Plan)是一种集体医疗保险，有700,000成员。其中62,000年龄在40-64的女成员被随机地分为处理组和对照组。
  - "处理"：邀请参加一年4次的Mammograph筛查，另外也参加一般临床检查。
  - "对照"：只参加一般临床检查，但不接受上述"处理"

- **1/3被邀请的人拒绝参加筛查**
- **试验开始后5年内的结果如下表：**

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

H0：Mammogrpah无效

成员自己决定是否接受筛查，这是嵌在随机试验中的一个观察研究！

如果接受筛查的(Screened)与拒绝的(Refused)有系统性差异, 那么比较 **Screened vs. Control**，结果会出现偏差。

作业：你能否从**Table1**数据中提供**Screened** 组与**Refused**组存在差异的证据？

事实上，富裕和教育程度高的人更倾向于接受邀请，但这些人乳腺癌发病率比其他人高 。

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

正确的分析方法是 *intention-to-treat analysis:*

比较Treatment组(被邀请的人,不论是否接受筛查!) 和 Control 组

$$u = \frac{0.0013 - 0.002}{\sqrt{\left(\frac{1}{31000} + \frac{1}{31000}\right) \times 0.00165 \times (1 - 0.00165)}} = -2.378$$

$$pvalue = P(|N(0,1)| \geq |-2.378|) = 0.017, \quad \text{其中} 0.00165 = (39 + 63)/62000$$

列联表分析得到同样结果:Pearson chi-square = 5.6563, p=0.017

所以, Mammograph有效果，而且是积极的效果。

为什么**intention-to-treat analysis**是正确的方法?

将Refused 组包括进处理组，不会出现偏差(I型错误率控制在预定水平), 但可能降低显著性(功效).

如果筛查确实无效(原假设成立)，**Screened+Rufused** 组与**Control**应该无差异, 所以使用**Refused**数据可保证结果无偏(即I型错误控制在希望的范围之内).

如果筛查确实有效(原假设不成立), 比较**Screened+Rufused** 组与**Control** 可能会功效偏低, 这是因为**Refused**的那些人没有得到治疗.

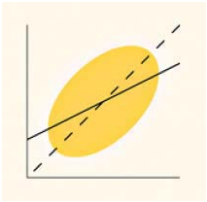
这反映了统计学/科学 宁愿降低功效也要控制好 I 型错误的基本思路，即宁可不发现 (不拒绝**H0**), 也不要做出错误的发现。

回归分析方法简介

■什么是回归 (regression)?

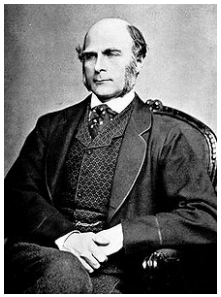
以线性方程描述随机变量(y,x)之间关系的方法称为线性回归。

$$y = a + bx_1 + \ldots + cx_p + \varepsilon$$



为什么称之为“回归”？
高尔顿(Galton)研究父子身高关系时发现儿子身高与父亲相比有趋中的趋势，称之为回归。

Francis Galton (弗朗西斯. 高尔顿)



Sir Francis Galton (1822 –1911), cousin of Charles Darwin, was an English Victorian polymath:

Anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician, and statistician.

He said he is a private gentleman. He was knighted in 1909. He is best known for:

Regression toward the mean (回归), Correlation (相关系数), Standard deviation (标准差), Galton board (高尔顿板), eugenics (优生学), Weather map (气象图), Fingerprint (指纹), nature vs nurture (先天与后天）…

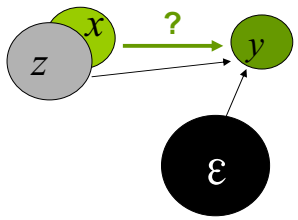
■ 为什么回归?

1. 因果推断：为了推断x，y是否有因果关系，如果z既与y有关，也与x 有关，在回归方程

$$y = a + bx + \varepsilon$$

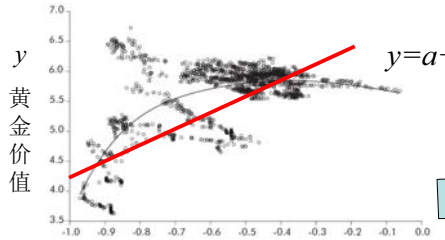
中添加一项，控制z：

$$y = a + bx + cz + \varepsilon$$

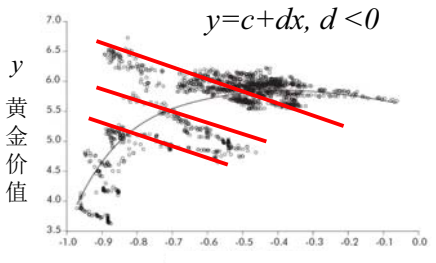


ε ： 所有与y有关，但与自变量无关的其它因素

Gold does not hedge against Dollar?



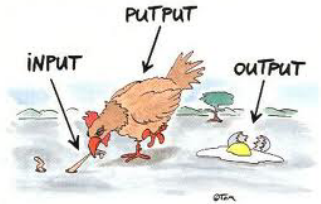
控制z (年代)



2. 预测：基于关联进行预测

莎士比亚: **凡是过去，皆为序曲。**

基于与目标变量(output), 关联的自变量(input), 使用线性回归模型或其它模型预测目标/响应变量。 预测模型可以有多种。

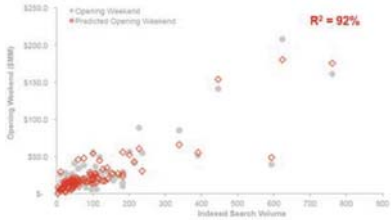


- 亚马逊购物推荐系统
- 谷歌利用搜索数据预测奥斯卡获奖者
- Neflix通过大数据分析深度挖掘了用户的喜好，捧红了《纸牌屋》。
- Detecting influenza epidemics using search engine query data, Nature 457, 1012-1014 (2009)
- Civil conflicts are associated with the global climate, Nature 476.438-441 (2011)

谷歌电影票房预测模型

使用如下4个变量预测一周后的票房

（1）电影放映前一周的电影的搜索量
（2）电影放映前一周的电影广告的点击量
（3）上映影院数量
（4）同系列电影前几部的票房表现



使用如下3个变量预测一月后的票房

（1）电影预告片的搜索量
（2）同系列电影前几部的票房表现
（3）档期的季节性特征