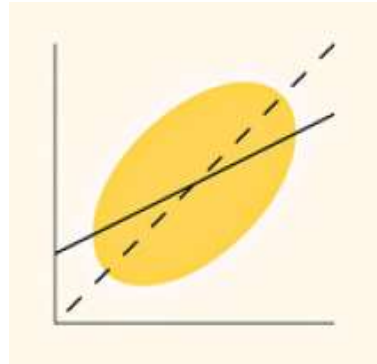


第一讲 回归分析简介

2022.9.2



内容

- 课程介绍
- 关联与因果
- 观察与试验
- 案例

课程简介

主讲 助教

杨亚宁 ynyang@ustc.edu.cn

周文斌 zhouwenbin@mail.ustc.edu.cn

李睿智 one_one@mail.ustc.edu.cn

先修

概率论数理统计; 线性代数

回归分析是两样
本t检验的推广

回归分析是线性
代数的应用

课程主页

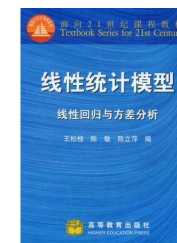
<http://staff.ustc.edu.cn/~ynyang/2022>

考察

期末考试（以课件为主）+ 作业（纸面、计算机）

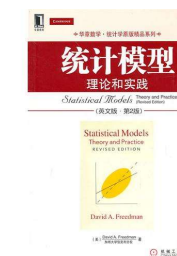
教材

1. 王松桂, 陈敏, 陈立苹编 (1999). **线性统计模型-线性回归与方差分析**. 高等教育出版社.



参考书

2. David A. Freedman (2009). **Statistical Models: Theory and Practice** (2nd ed). Cambridge University Press/机械工业出版社.



电子版下载地址:

<http://staff.ustc.edu.cn/~ynyang/2022/books/x.pdf> (x=1,2)

(将有x=3,4,...)

关于英文参考书:

D. A. Freedman (1938-2008): 伯克利大学统计学家, 所著《Statistical models: theory and practices》作为研究生教材主要介绍线性模型及统计思想, 非常简明易读。

第二版前言:

Some books are correct. Some are clear. Some are useful. Some are entertaining. Few are even two of these. This book is all four.

另外, Freedman撰写的《Statistics》是本科统计教材的典范。

课程概述

以线性回归模型为工具

- 研究变量之间的因果关系（生物统计、计量经济、科学研究）；
- 利用关联关系进行预测（人工智能、机器学习）。

线性模型

研究随机变量 y (响应, response) 与随机向量 \mathbf{x} (自变量、解释变量) 的关系。线性回归模型（或线性模型）假设

$$y = \alpha + \mathbf{x}^T \boldsymbol{\gamma} + \varepsilon$$

α , $\boldsymbol{\gamma}$: 回归系数, 未知参数

ε : 误差, 响应中不能被 \mathbf{x} 解释的部分。

线性

$\alpha + \mathbf{x}^\top \boldsymbol{\gamma}$ 是参数 $\boldsymbol{\gamma}$ 的线性函数。

注： $\mathbf{x}^\top \boldsymbol{\gamma}$ 是双线性函数，既是 $\boldsymbol{\gamma}$ 也是自变量 \mathbf{x} 的线性函数，

但有时允许关于 \mathbf{x} 不是线性，比如 $y = \alpha + e^{x_1} \beta_1 + x_2^2 \beta_2 + \varepsilon$

□ 所有样本 (y_i, \mathbf{x}_i) 满足 $y_i = \alpha + \mathbf{x}_i^\top \boldsymbol{\gamma} + \varepsilon_i, \quad i = 1, \dots, n$

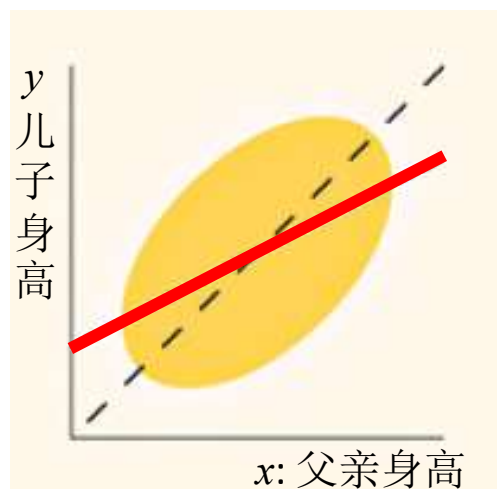
□ 以向量、矩阵表达： $\mathbf{y} = (y_1, \dots, y_n)^\top, X = \begin{pmatrix} 1 & \mathbf{x}_1^\top \\ \dots & \dots \\ 1 & \mathbf{x}_n^\top \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \boldsymbol{\gamma} \end{pmatrix}$ ，上述模型为 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

除了误差项，这是线性方程组的求解问题（求解最优 $\boldsymbol{\beta}$ ）。

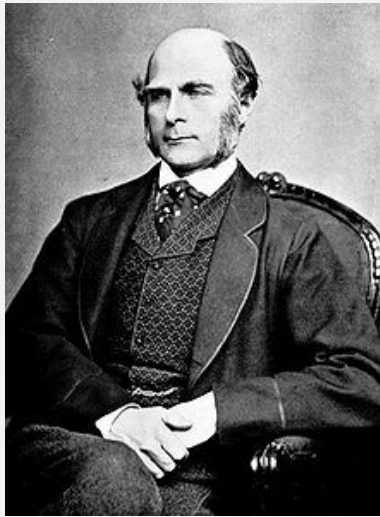
回归

归于平均现象 (Regression to the mean)

高尔顿研究父子身高关系的时候发现, 如果父亲很高,那么儿子也普遍较高(正相关), 但通常会比父亲矮一些; 如果父亲很矮,那么儿子也普遍较矮, 但会比父亲稍高一些。高尔顿把这种向平均靠拢的现象称为“回归regression”(图中**实线**)。



弗朗西斯·高尔顿
(Francis Galton)



弗朗西斯高尔顿 (1822 –1911), 爵士, 达尔文的表弟, 英国维多利亚时期的博物学家(polymath):

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

他自称为民间/个体学者 (private gentleman), 建立了众多学科或概念(包括统计、气象学、优生学、犯罪学...):

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) ...

关联与因果

宁可找到一个因果解释，而不愿获得一个波斯王位。
德谟克利特Democritus (460-370B. C.)

变量之间的关系包括关联(association)和因果(causation)。

因果关系是其它因素给定的条件下变量之间的（本质）关系；
关联是不考虑其它因素下得到的（表面）关系。

因果是一种特殊的关联，但关联并不蕴含着因果（Proverb:
Correlation does not imply causation）

药物治疗疾病

（因果，**试验**结果，困难）

伪劣保健品改善睡眠

（关联，**观察**结果，简单）

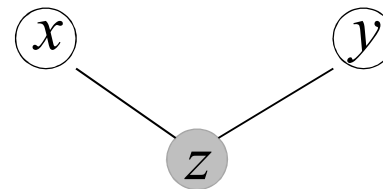
关联~不独立

若两个变量 x 与 y 不独立，称它们是关联或相关的。

相关在统计中表示线性关联。人们通常并不严格区分关联(association)与相关(correlation)

干扰因素/ 混杂因素

与 x, y 都有关的因素称为研究 x, y 关系的干扰或混杂因素 (confounding factor)。



上图中 z 是干扰因素。当 z 变化时， x, y 都变化，这使得 x 和 y 相关。但当 z 给定时， x, y 可能不再相关。参见下页例1。

因果~ 条件不独立

如果在所有干扰因素给定的条件下， x 与 y 相关， x 与 y 才可能存在因果关系。

例1(独立与条件独立). 同一家庭两个孩子的某个指标 x, y 相关, 它们都与父母以及共同的家庭环境有关 (以公共因子 z 表示), 此外它们分别与随机因素 $\varepsilon_1, \varepsilon_2$ 线性相关, 假设单因子模型:

$$\begin{cases} x = z + \varepsilon_1 \\ y = z + \varepsilon_2 \end{cases}, \text{ 其中 } z, \varepsilon_1, \varepsilon_2 \text{ 独立。}$$

则 x 与 y 不独立, 但给定 z 时 x, y (条件) 独立。

外人看来, 同一家庭的两个孩子很像:

x, y 不独立 (相关) 是因为它们共享 z 。

父母看来, 两个孩子差异较大:

给定 z 条件下, x, y 条件独立。

例2. 下述因果论断是否正确？识别干扰因素

1. 观察表明，常吃富含维生素食物的人的癌症发病率较低，所以维生素可预防癌症。

错误。生活习惯好的人常吃维生素，也不容易得癌症。生活习惯是干扰因素。

2. 一项研究收集了多个国家的人均电话装机量和女性乳腺癌死亡率数据，发现两者高度正相关，所以打电话会导致乳腺癌。

错误。发达国家人均电话量高而生育率低。女性生育是一个自我修复完善的过程，生育少则乳腺癌发病机会高。国家发达程度是一个干扰因素。

3. 数据表明借贷多的人健康状况较差，所以债务可导致疾病。

错误。因果颠倒，实际上可能是疾病导致借贷。

推断因果的前提: 条件均同 (Ceteris paribus)

条件均同

推断因果的前提是“条件均同”，即在推断因果关系时，需要控制所有干扰因素（变量、条件），使得它们不变（均同）。

控制实验条件

在同样的条件下进行研究和采集数据。

伽利略落体实验，保证球体阻力相同

以回归控制变量

研究对象所处“条件”不同时，在回归方程中添加“条件/变量”一项，以校正/对齐样本。

400m比赛的起点前置 $1.2 \times 2\pi = 7.54$ 米

研究设计：观察与试验

随机化控制试验（简称试验）和观察研究是两种常见的研究设计(数据采集方式)。试验与观察研究的区别在于自变量取值是由外界干预决定还是研究对象本身固有的。

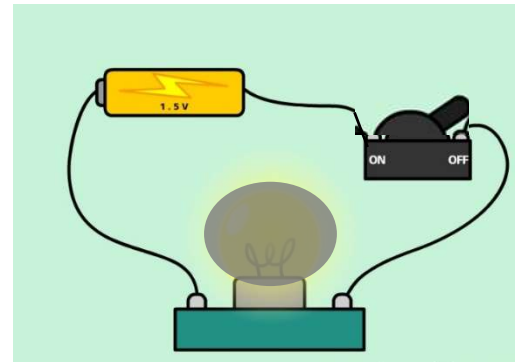
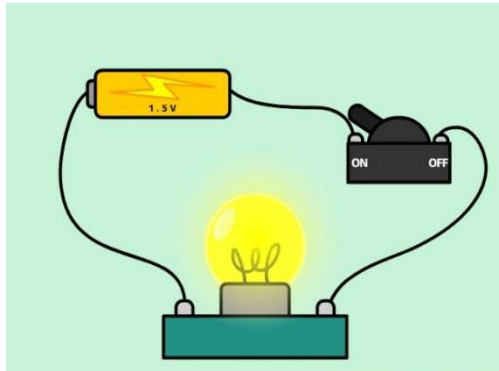
观察：研究对象不能被干预或改变，只能被动观察，干扰因素不可控或根本不可识别。比如例2.1，2.2。观察研究大多与人、社会有关。

试验：随机改变研究对象的条件/变量，比如临床试验。随机化控制试验是推断因果的最高原则。但多数情况下，比如社会现象，特别是与人有关的问题，不能试验，不能干预、改变研究对象。

观察研究

观察研究 (observational study) 只能被动观察，不能干预、改变研究对象。发现的变量关系常常是关联而不是因果。

观察两个线路。左图：开关on，灯亮；右图：开关off，灯不亮。



上述观察不能说明开关(x)能控制灯泡(y)。有可能电池、线路等干扰因素出现问题而不是开关导致右图不亮。

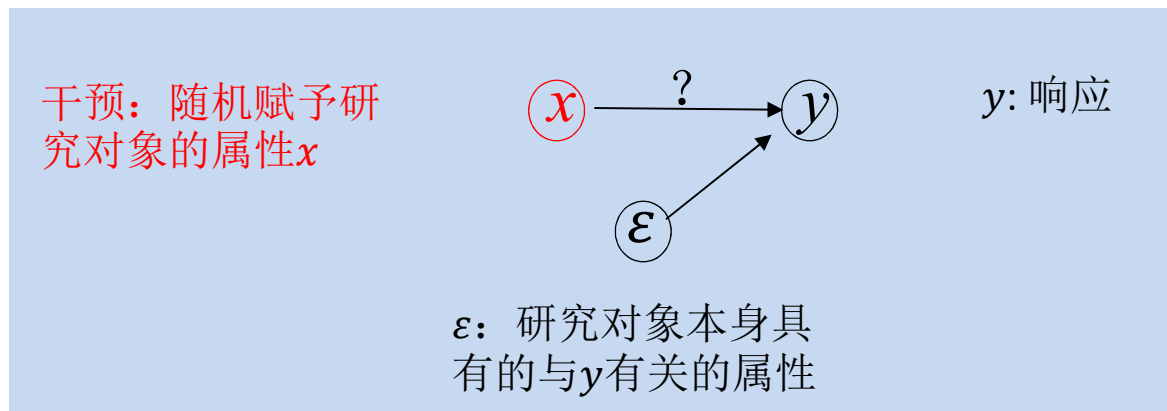
如何避免干扰？试验，在同一线路（条件）上改变开关状态。

随机化控制试验

随机化控制试验 (Randomized controlled experiment, Fisher 1920's): 研究对象的某个变量 (原因) 取值由外界随机分配 (干预), 因而该变量与研究对象本身具有的属性是独立的。

自变量 x 变化是否导致响应 y 变化 (因果) ?

随机化试验中, x 的取值由外界随机赋予, 而不是研究对象自身的属性, 所以 x 与研究对象独立, 与研究对象自身的影响 y 的所有因素 ε 独立 (图中 x 与 ε 无连线)。



临床试验

药物开发中的临床试验(clinical trial)是最常见的随机化控制试验。前期证明药物无害之后，三期临床试验将研究对象**随机**分配服用药物或安慰剂（自变量 x 等于0或1），双盲（患者和医生都不知道 x 值）。

x 的取值不由研究对象自身决定，也与响应的测量者（医生）无关，所以 x 与所有影响 y 的干扰因素 ε 独立（ ε 包括年龄、性别等和其它观测不到的因素，以及 y 的测量者）。

换言之，两组样本除了 x 不同之外，在其它因素上都相同（同分布）。比较两组的响应，如果有差异只能归因于 x 的不同（允许5%的错误率）。

注：如果一组随机变量是iid的，那么我们认为它们（在统计意义上）是相同的。

例3. 从服用某种降血压药物的患者中随机抽取若干人，另外从未服用该药物的患者中随机抽取若干人作为对照 (control)，测量血压指标。

假设两组样本分别 iid 来自于总体 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$ ，零假设 $H_0: \mu_1 = \mu_2$ 。

计算两样本 t -检验统计量，得 p 值 $p=0.002$ 。结论：在 0.05 水平下药物组和对照组的血压有显著性差异，检验结果能否说明药物有效？

不能。这是观察研究而不是临床试验。服用药物与否是病人或医生的选择，而这种选择可能与疾病程度有关，也可能与职业、年龄、性别等因素有关。所以结果显著不说明药物有效。实际上，我们没有充分理由假设“原假设下两组同分布”。

注

- Control: 在统计中，control variable 是“控制（给定变量）”的意思；在比较试验中 control 组译作“对照组”。
- “试验”还是“实验”？前者用于探索，后者用于验证。

案例1：乳腺癌研究（Freedman第一章）

Mammograph 是一种X光早期筛查乳腺癌的方法。为了检验Mammograph的有效性，1960's在纽约进行了一个大型随机化控制试验。

HIP (Health Insurance Plan) 医疗保险有700,000个成员。其中62,000个年龄在40-64的女成员被随机地分为处理组(treatment)和对照组(control):

- 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。1/3被邀请的人拒绝参加筛查。
- 对照：只参加一般临床检查，但不接受Mammograph筛查。

5年跟踪(followup)后的数据如下

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

我们希望检验 H_0 : Mammograph无效

错误的分析方法:

检验 Screened组和对照组的死亡率 p_1, p_2 是否相同, 即检验 $H_0': p_1 = p_2$

$$\hat{p}_1 = 23 / 20200 = 1.1 / 1000, \quad n_1 = 20200$$

$$\hat{p}_2 = 63 / 31000 = 2 / 2, \quad n_2 = 31000$$

$$\hat{p} = (23 + 63) / (20200 + 31000) = 0.00168$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}} = -2.431$$

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

p 值 = 0.015, 在0.05显著性水平下拒绝原假设。这是否说明筛查有效?

由于成员自己决定是否接受筛查，所以同意接受筛查的Screened组是处理组的一个有偏样本（Screened组数据是观察研究数据）。

事实上，Screened组与Refused组有系统性差异：
富裕和教育程度高的人更倾向于接受邀请。

即使筛查没有任何作用（原假设），Control组的死亡率 p_1 也可能比Screened组的死亡率 p_2 要高。换言之，

“ H_0 : Mammograph无效”与“ H_0' : $p_1 = p_2$ ”不等价

H_0 成立的条件下，两个概率 p_1 和 p_2 不一定相同，这导致针对 H_0' 构造的z检验的一型错误率偏大。因此上页比较Screened组和对照组的死亡率的做法是错误的。

正确的分析方法：
比较Treatment组(含拒绝的人)和Control组。

$$\hat{p}_1 = 1.3 / 1000, \hat{p}_2 = 2 / 1000,$$

$$\hat{p} = (39 + 63) / 62000 = 0.00165$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}} = -2.378$$

	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

$pvalue = 0.017$, 所以Mammograph筛查能显著地降低死亡率。

注：“ H_0 : Mammograph无效”成立时，由于随机化，Treatment组成员和Control组成员在统计意义上无差异（他们的死亡率相同），所以，如果使用所有62000个研究对象构建检验， $H_0 \Leftrightarrow H_0'$ ，此时，针对 H_0' 构造的z检验，也是 H_0 的有效检验（原假设下的分布正确，进而一型错误率能够正确地控制在给定的水平）。

回顾：案例1用到了两个二项分布概率相等性检验

$a \sim B(n_1, p_1)$, $c \sim B(n_0, p_0)$, 独立; $\hat{p}_1 = a/n_1$, $\hat{p}_0 = c/n_0$,

$H_0: p_1 = p_0$ 的两样本 z-检验:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_0)\hat{p}(1-\hat{p})}}, \quad \text{其中 } \hat{p} = (a+c)/(n_1+n_0),$$

H_0 下近似 $z \sim N(0,1)$, p 值 = $P(|Z^*| > |z| | z)$, $Z^* \sim N(0,1)$.

记 $b = n_1 - a$, $d = n_0 - c$, $m_1 = a + c$, $m_0 = b + d$, $n = n_1 + n_0 = m_1 + m_0$, 以列联表展示:

	1	0	总计
组1	a	b	n_1
组2	c	d	n_0
总计	m_1	m_0	n

Pearson卡方统计量(齐一性检验):

$$X^2 = \frac{n(ad-bc)^2}{n_1 n_0 m_1 m_0}$$

H_0 下近似地 $X^2 \sim \chi_1^2$.

验证: $X^2 = z^2$.

案例2：霍乱传播

19世纪中期，人类对霍乱(cholera)传染几乎一无所知，细菌学说只是众多理论中的一种。1855年伦敦医生John Snow 利用通过流行病学调查（流调）和精巧的分析发现霍乱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，他是刚从霍乱流行的汉堡乘船到伦敦的海员。并且发现第二例病人住过第一个病例住过的房间。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，另一个没有。前者的饮水系统被污染了，而后者没有。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。

□ Snow发现其它地区的零星病例，大多与Broad Street有关

□ 另外，Broad Street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

所有证据都表明，疫情可能与供水系统有关。

■ Snow注意到伦敦有两大供水公司:

- Southwark & Vauxhall公司: 水源在泰晤士河下游, 污染较重;
- Lambeth公司: 水源在上游, 污染不严重。

■ Snow比较了这两个公司用户的霍乱死亡率:

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值=0, 两家公司用户死亡率有显著差异。但我们不能断言这是水质不同导致的, 因为这是一个观察研究, 我们不知道是否由其它因素导致两组死亡率不同 (比如, Lambeth的客户是否具有好的医疗条件)。

■ 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，比如同一个房子两侧的两家可能选用不同的公司。

两个公司的价格、服务各方面差异不大，用户一般不知道两个公司水源差异。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。

所以，虽然这是观察研究数据，但实际上这是一个**天然的随机化试验**！经过随机化，两个公司的用户除了水源即水质不同之外，在其它任何方面都在统计意义上相同（同分布），所以死亡率的不同只能归因于水质的不同。

基本可以断言：饮用水传播霍乱。

案例3：贫困的原因

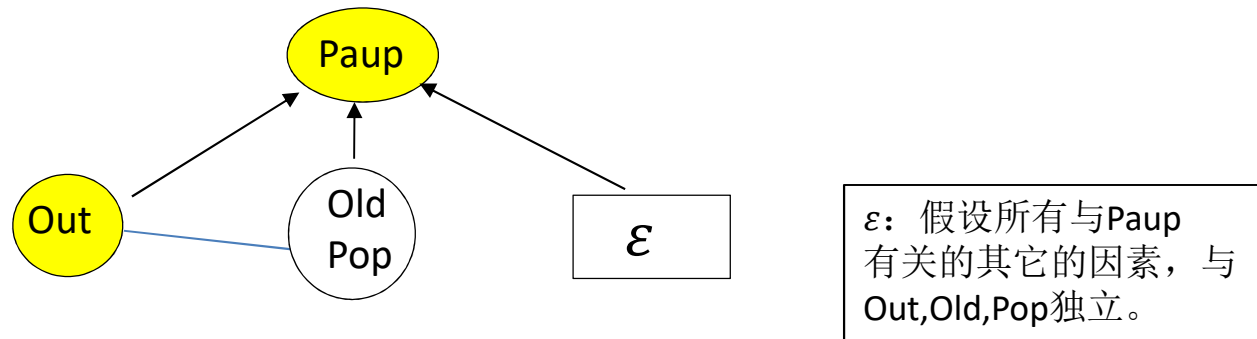
在十九世纪的英国，穷人（pauper）的生存主要依赖救济院（poor-houses）救济（但被要求必须在救济院工作），或救济院外的政府救济(out-relief)。英国统计学家 Yule (1899) 研究了院外救济是否会增加贫困人口比例。

Yule收集了1871年若干区会的数据：

- Paup: 穷人占总人口的比例
- $Out=N/D$: 救济院外和救济院内救济人数之比，其中 $N = \# \text{ out-relief}$;
 $D = \# \text{ inside poor-houses}$

Yule注意到老年人比例（变量Old）和人口总数（Pop）可能既与Paup有关，也与Out有关，是干扰因素，需要加以控制

假设没有其它干扰因素，变量之间的关系如下图表示：



建立如下线性模型（将Old,Pop加入模型加以控制）：

$$\text{Paup} = a + b \times \text{Out} + c \times \text{Old} + d \times \text{Pop} + \varepsilon$$

↑
果

↑
因

↑ ↑
干扰
因素

↑
其它与Paup有
关的因素

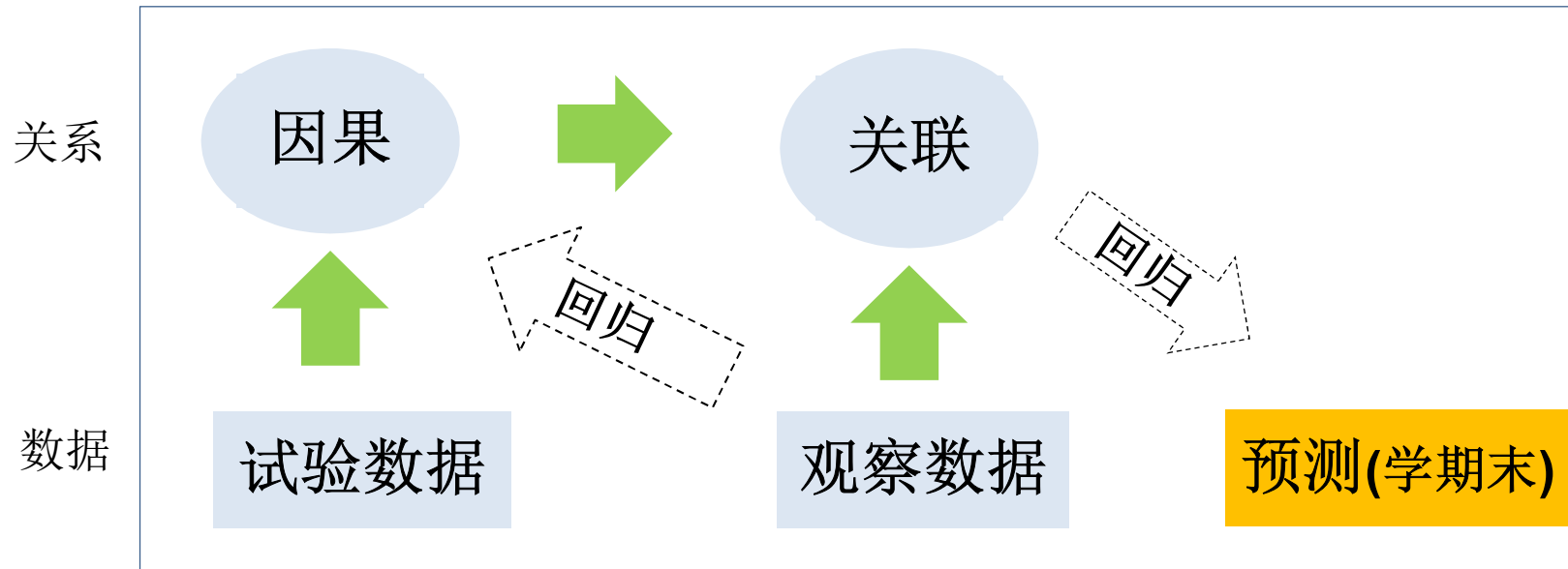
其中Old, Pop两项的作用在于控制干扰因素，你可以把方程理解为

$$\text{Corrected.Paup}^{\Delta} = \text{Paup} - c \times \text{Old} - d \times \text{Pop} = a + b \times \text{Out} + \text{error}$$

即先消除Old,Pop对Paup的影响，再研究校正后的Corrected.Paup与Out的关系。如果模型正确且 $b > 0$ ，那么我们有因果关系：

其它变量不变时，Out增加一个单位导致 Paup增加 b 个单位。

如果模型是不正确的，即 ε 与Out,Old,Pop不独立（比如还存在比如社会经济状况，政府管理的效率等其它干扰因素），那么上述结果就不是因果关系，而是关联。



回归分析:

- (1) 观察研究中控制干扰因素、推断因果;
- (2) 利用关联关系进行预测。