回归分析在线课堂

zoom会议号： 4741229681

时间：周三67，周五34

课程主页（课件、作业等）：
http://staff.ustc.edu.cn/~ynyang/lm2020
QQ讨论群：876416050

主讲: 杨亚宁 **ynyang@ustc.edu.cn**
助教: 赵明华 **zmh07@mail.ustc.edu.cn**
曾正浩 **zzh98052@mail.ustc.edu.cn**

课程简介

- 先修：数理统计，线性代数
- 课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020
- 课程概述：

以线性回归模型为工具, 研究变量之间的关系:

- 通过控制变量推断因果关系;
- 利用关联关系进行预测。

当前全社会关心疫情拐点何时出现（预测），以及有效药物的研发（因果）。

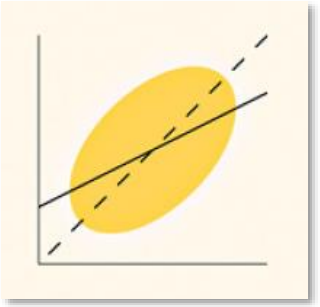
■ 线性回归模型 (Linear regression model)

线性回归模型以线性函数刻画响应变量 y 与自变量 $\mathbf{x}$ 之间的关系：

$$y = \alpha + \beta^{\mathsf{T}}\mathbf{x} + \varepsilon$$

■ 为什么叫做“回归 (regression)”？

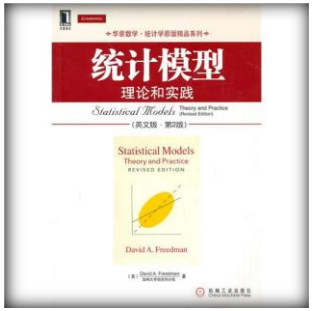
Regression to the mean:
如果自变量 (x轴)很极端，那么响应变量 (y轴)平均来看会趋于平庸（实线）。

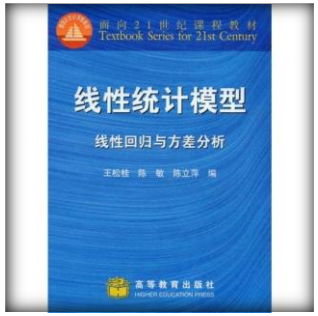


■ 教材：

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

2. **王松桂，陈敏，陈立萍编 (1999). 线性统计模型 – 线性回归与方差分析. 高等教育出版社.**





下载地址: http://staff.ustc.edu.cn/~ynyang/lm2020/books/1.pdf 2.pdf

关于英文教材:

D. A. Freedman (1938-2008): 伯克利大学统计学家, 所著《Statistics》是本科统计教材的典范，而 《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，非常简明清晰。都有中文版。

《Statistical models：theory and practices》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. Statistical Models: Theory and Practice is lucid, candid and insightful, a joy to read

■ 参考书：

1. **S. Weisberg (2005) Applied Linear Regression (3ed). Wiley.** 数据
2. **王松桂, 史建红, 尹素菊, 吴密霞 (2004) 线性模型引论. 科学出版社** 代数(2-3章)
3. **J.H. Stapleton (1995) Linear Statistical Models. Wiley** 投影
4. **R.Kabacoff (2015) R in action. Manning**

下载地址: http://staff.ustc.edu.cn/~ynyang/lm2020/books/3.pdf 4.pdf 5.pdf 6.pd

■ 其它信息

- 上机实习（时间待定）：
  9 次上机实习（**R**语言）

- 考核方式：
  总评 **= 15%**作业 **+ 15%**上机实习 **+ 70%**期末考试

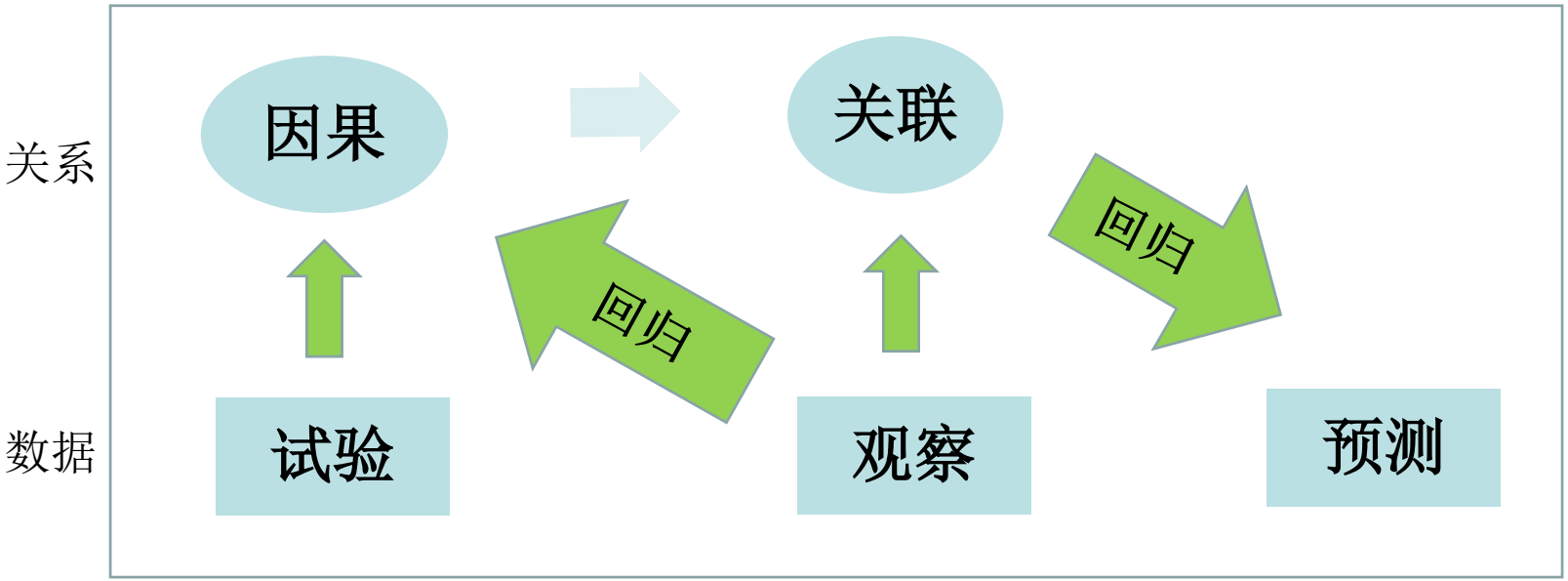
回归分析 **(01714601)**
主讲: 杨亚宁 **ynyang@ustc.edu.cn**
助教: 赵明华 **zmh07@mail.ustc.edu.cn**
曾正浩 **zzh98052@mail.ustc.edu.cn**

第一讲 关联与因果

2020.2.19

纲要

- 关联与因果
- 观察研究与试验研究
- 案例：霍乱传播



关联与因果

宁可找到一个因果解释，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

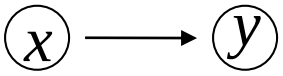
关联(association)和因果(causation)

- 因果关系是直接的、本质的关系。
- 关联是表面、间接和相对容易获得的。
- 关联未必蕴含因果。
- 关联可用于预测，为因果探索提供了线索。

■ 关联和因果 （概率含义）

- 关联： 不独立。
- 因果： 所有其它因素给定的情况下一个变量的变化导致另一个变量的变化。其必要条件为条件不独立。

因果



关联

*x*和*y*关联是因为它们都与z有关。

z称为干扰因素(confounder)

■ 因果蕴含关联，但关联不一定蕴含因果
Correlation does not imply causation

人们通常可能并不严格区分关联(association)与相关(correlation)。

如果“因”变量与“果”变量一起被观察到，并且存在其它干扰因素，则变量之间的关系可能是关联而不是因果。

排除干扰因素、透过关联发现因果是科学研究的重要目标。

■ 关联关系的表述不应解读为因果

人们总是倾向于用因果关系表达观察到的现象，例如

“饭后百步走，活到九十九”

表述的可能只是关联而不是因果。

而一些关联表述让人容易联想到因果关系，例如

“雄鸡一唱天下白”

半夜鸡叫的故事证明了两者没有因果关系。

例1. 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。

生活习惯好的人常吃维生素，也不容易得癌症。生活习惯可能是一个干扰因素。

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。

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。生育率是一个干扰因素。

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。

因果颠倒，实际上可能是疾病导致借贷。

推断因果的前提

其它条件均同(Ceteris paribus)

欲推断两个变量是否存在因果关系, 其它干扰因素必须保持固定不变（控制 **control**）。

物理化学试验可以做到其它条件尽量相同（伽利略落体实验)。但多数情况下，干扰因素太多以至于无法识别和人为控制。

试验

（随机化控制）试验通过外界干预解决了这一难题：人为（随机）地改变研究对象的因变量，然后观测结果。

观察

但绝大多情况下，我们不能随机地改变对象， 我们只能被动观察，回溯推断。

试验与观察：从t检验谈起

假设两组独立样本：

$$y_{11},\ldots,y_{1n_1} \text{ iid} \sim N(\mu_1,\sigma^2);$$

$$y_{21},\ldots,y_{2n_2} \text{ iid} \sim N(\mu_2,\sigma^2);$$

零假设$H_0:\mu_1=\mu_2$

$$\bar{y}_k=\sum_{i=1}^{n_k} y_{ki}/n_k, k=1,2; n=n_1+n_2.$$

$$s^2=\frac{1}{n-2}\left(\sum_{k=1}^{2}\sum_{i=1}^{n_2}(y_{ki}-\bar{y}_k)^2\right).$$

两样本 t-检验统计量：$t=\dfrac{\bar{y}_1-\bar{y}_2}{\sqrt{(1/n_1+1/n_2)s^2}} \overset{H_0}{\sim} t_{n_1+n_2-2},$

- 当$|t|\geq t_{n_1+n_2-2}(\alpha/2)$时 $\Leftrightarrow$ $p\leq\alpha$时，在α水平下拒绝原假设；
- 其中p值：$p=P(|T^*|\geq|t|\,|t)$其中$T^*\sim t_{n_1+n_2-2}$.

例2. 从服用某种降血糖药物的糖尿病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，测量血糖指标. 假设两组样本分别来自于总体$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$, 零假设$H_0:\mu_1=\mu_2$.

假如计算得到 p=0.02，则我们可以拒绝零假设。 结论：在0.05水平下药物组和对照组的血糖有显著性差异.

检验结果能否说明药物有效？

不能！这是一个典型的观察研究。服用药物与否很可能是病人自己或医生的选择，而这种选择很可能与疾病程度有关，也可能与职业、年龄、性别等因素有关。结论是关联（血糖与分组有关）而不是因果。

应用 t-检验推断药物有效性的关键：

原假设下，即药物无效的情况下，两组分布是否可以假设相同？

其它类型的两组比较、多组检验、甚至回归分析与此类似。

例3(临床试验:随机化双盲试验). 随机抽取若干糖尿病患者，随机给病人服用药物或安慰剂(对照)，病人和医生都不知道每人服用的是药物还是安慰剂（双盲）.

随机分配药物和双盲使得两组人除了服用的药物不同之外，
其它因素都相同（*iid*,统计意义上相同）, 因而两组等同,具有可比性.

因此可以假设两组样本分别来自于总体$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$,
两个总体除了血糖均值不同外，其它相同（都是N，σ^2）.

如果t检验结果显著，比如用药组血糖指标显著的地低于对照组，
则说明药物有显著的治疗效果 (因果,一定的错误率下).

试验与观察

试验和观察研究是两种常见的研究设计(数据采集方式)，区分研究设计对于数据分析和结果解释非常重要。

■ **随机化控制试验**

(Randomized controlled experiment, Fisher 1920's)

研究对象的“因”变量取值由外界随机分配。一个完善的随机化试验基本上能消除掉干扰因素的影响，得到可靠的因果关系。

“试验”还是“实验”？前者用于探索，后者用于验证

随机化控制试验是推断因果的最高原则。随机化在统计意义上切断了因变量与所有其它干扰因素的关系。

特别地，在两组比较问题中，随机化使得被比较的两组几乎没有差异。因而得到的因果关系是可靠的(除了5%的例外)。

最主要和最成功的随机化控制试验是药物开发中的临床试验(clinical trial)。

■ 观察研究(observational study)：

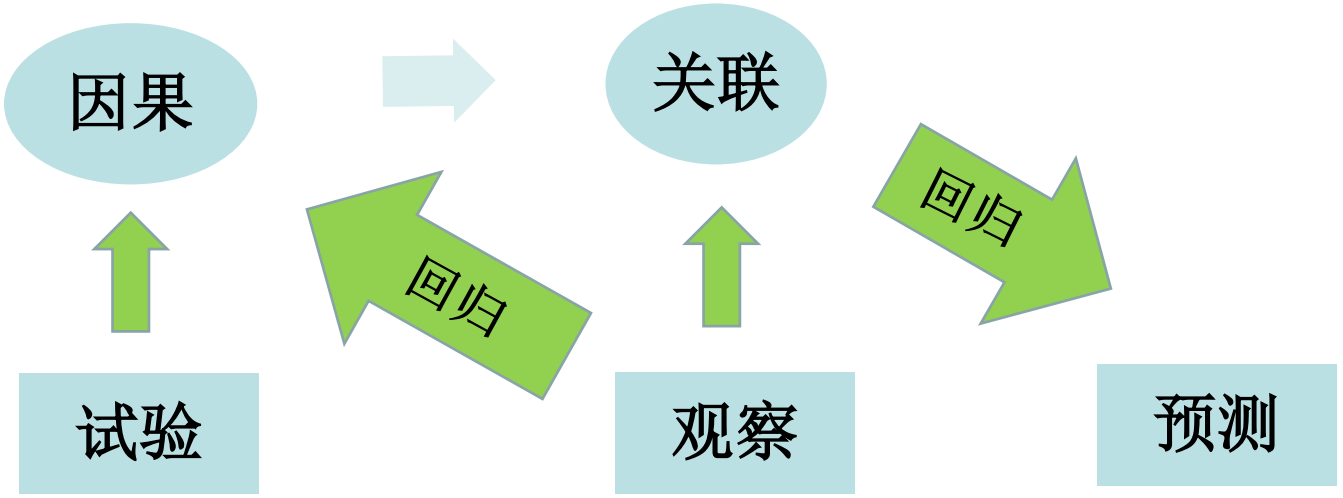
只能被动观察，不能干预、改变研究对象，因而变量之间的关系常常是关联而不是因果。

例1，例2都是观察研究

大多数情况下只能被动观察，不能试验，特别是与人、社会有关的问题。基于观察研究数据推断因果是很困难甚至是不可能的。

总之，试验与观察研究的区别在于因变量取值是由外界干预决定还是研究对象本身固有的。

如何从观察研究数据推断因果？回归分析



回归分析也用于通过建立因果或关联关系进行预测