回归分析
杨亚宁 (ynyang@gmail.com)

第一讲. 关联与因果

纲要

- 课程基本信息
- 关联与因果
- 试验研究与观察研究

课程基本信息

■ 关键词：

关联/相关、因果、观察研究、回归、预测

■ 课程概述：

以线性回归模型为工具, 研究变量之间的关系:
(1) 对于观察研究数据，通过控制变量推断因果关系;
(2) 利用关联关系进行预测。

■ 先修：数理统计，线性代数

■ 教材：

1. David A. Freedman (2009). Statistical Models: Theory and Practice (2nd ed). Cambridge University Press/机械工业出版社. 统计思想
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. 技术细节,习题

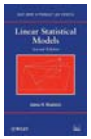
关于英文教材:

D. A. Freedman (1938-2008): 伯克利大学统计学家, 所著 “Statistics” 和“Statistical models：theory and practices” 是统计教材的典范。他尤其关注如何正确应用统计。

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. …

■ 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. M. Kutner, C. Nachtsheim, J. Neter, W. Li (2004) Applied Linear Statistical Models (5ed), McGraw Hill. 大全,备查



下载地址: http://staff.ustc.edu.cn/~ynyang/lm2014/books

■ 其它信息

- 目标/要求:
  掌握回归的思想和理论；使用软件分析解决实际问题。
- 上机实习（5-13周）：
  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
- 课程主页：
  http://staff.ustc.edu.cn/~ynyang/lm2014

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，是人类长期以来孜孜追求的探索目标；关联关系通常是表面、间接和相对容易获得的，是预测分析的主要工具。

■ 关联和因果

- 关联/相关(correlation)： 不独立/不相关。
- 因果： 所有其它因素相同/固定时，一个变量的变化导致另一个变量的变化。可以理解为条件不独立/条件不相关。

本课程学习使用线性回归模型，研究变量之间的关联关系和因果关系推断

Amos 3:3
Do two walk together unless they have agreed to do so ?

■ 关联不一定蕴含因果 (Correlation does not imply causation)

寻找和发现因果关系是一个古老的问题。因果通常隐藏在关联的背后, 透过关联发现因果是科学研究的重要目标。

很多时候，“在一起” 只是关联而不是因果，例如

“物体越重下落越快”
“饭后百步走，活到九十九”，

都是经验观察得到的关联。

例1：因果还是关联？

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。

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，所以打电话会导致癌症。

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. 比较肺癌病人和正常人的生活习惯，发现肺癌病人抽烟比率显著高于正常人，所以抽烟可导致肺癌。

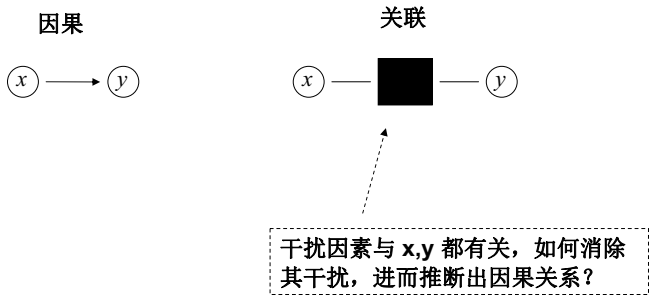
抽烟与肺癌有关联，但不一定是因果。存在其它因素(比如某种基因)导致抽烟，也导致癌症？即：两组人群有其它方面的差异？

4. 糖尿病人随机分配为两组，一组使用药物，一组使用安慰剂，双盲，发现药物组血糖有显著的降低，所以服用该药物可以有效降低血糖。

很可能是因果。由于随机化和双盲，两组人群基本没有其它系统性差异。

临床试验（随机化控制试验）

总结：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 (confounder), 则变量之间的关系可能只是关联而不是因果。



■ 推断因果的条件：其它条件均同
Ceteris paribus

欲推断独立变量(因)与响应变量(果)是否存在因果关系，其它变量必须保持相同，从而不干扰独立变量和响应变量关系的研究。

物理化学实验可以做到其它条件尽量相同，比如伽利略落体实验。但大多数情况下干扰因素太多以至于无法识别。Fisher 的随机化控制试验解决了这一难题。

统计意义上的“相同”为同分布，并容许5%的错误（I 型错误）

研究设计：试验与观察

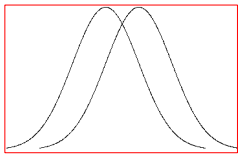
试验和观察研究是两种常见的数据采集方式(研究设计)，区分研究设计对于数据分析非常重要。

■ 随机化控制试验 (Fisher,1920’s)
使每个个体的因变量取值随机化，然后观察响应变量，因此因变量与其它干扰因素独立。

● 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。因而得到的因果关系是可靠的(除了5%的例外)。

例1中的糖尿病药物临床试验，双盲随机化保证了两组人除了用药与否不同之外在统计意义上相同。应用两样本t-检验得到显著性，则说明药物有治疗效果。

统计意义上的“相同”：同分布



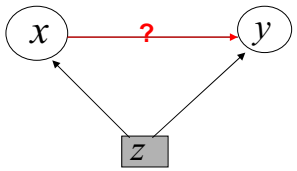
但是，大多数情况下只能被动观察，不能试验，特别是与人、社会有关的问题。

■ 观察研究(observational study)：
研究对象的数据只能被动观察，不能干预、改变其某些特性，因而变量之间的关系常常是关联而不是因果。

例如，药效研究中如果每个病人自己选择是否使用新药，则是观察研究。两组的病理指标差异不能反映药效作用。

例1中的1，2，3都是观察研究，所得到的结果都是关联关系。

一般地，“因变量” x 和 “果”变量/响应变量 y 作为研究对象个体的两个特性一起被观察到。

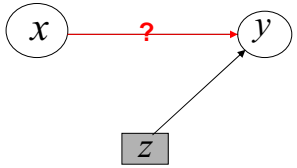


干扰因素 z 同时与 x, y 有关。

为了推断x，y之间的真实关系（因果），需要断开 x, z 之间的联系.

常用的两种策略：

- 分层：保持 z 是常数
- 回归：通过线性方程消除 z 的影响 (正交化)



需要强调的是，观察研究中，一般无法识别所有的干扰因素，所以分层与回归都不是完美的方法，但回归提供了一个比单纯的相关分析更接近因果的结论。