《线性回归》
— 统计模型的作用

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.02.24

主要内容：统计模型的作用

1. 统计模型
   - 统计模型
   - 回归模型
2. 总体和样本
   - 目标: 了解总体的参数
   - 解决方案: 利用样本
   - 理想的抽样方法
   - 现实中抽样的例子
   - 方便样本和推广
3. 混杂
   - 监狱数据
   - 这意味着什么？
   - 观察研究与实验研究
   - 观测性研究中的解决方案
   - 实验研究中的解决方案
   - 混杂因素

统计模型

- **模型**是（复杂的）现实的一种简化。数学模型，物理模型，统计模型，···
- 统计模型是关于随机变量或者随机向量的一系列假定。
- 统计模型的可能用途（首先要明确欲研究的问题）：
  - 描述. 例如：描述收入如何依赖于受教育年龄，种族，性别，居住地区。
  - 预测. 例如：根据年龄、性别、前科、入狱犯罪类型，预测获释罪犯再次被捕的可能性。
  - 因果分析（控制）. 例如：囚犯参加教育项目会降低再次被捕的几率吗?
- 对于上面所有的问题，我们还想知道估计的精度。

回归模型：

- 回归模型是关于随机变量(向量)$\mathbf{X}$和$\mathbf{Y}$的的条件分布$(Y|X)$ 的一系列假定。(后面将进一步明确这些假定！)
- 回归分析法是用来发现变量之间关系【非因果关系】的使用最为广泛的统计工具
- 线性回归是研究响应变量Y与一个或者多个解释变量之间关系的一种统计方法.
- 常见的线性回归。一个解释变量的情形是简单线性回归(simple linear regression), 多个解释变量的情形是多重线性回归(multiple linear regression), 注意，这个词组与多元线性回归(multivariate linear regression) 是不同的，多元线性回归中的响应变量是有多个，而且它们可能是相关的。在统计中有很多的回归模型，我们将从最为简单的简单线性回归模型入手学习回归方法。

回归模型

- 回归模型研究变量与变量之间的关系。
  【注意变量的类型！不同类型的变量会对应不同的模型！】
- 在所有情况下，我们要考察单个因变量Y与一个或多个自变量$X_1, \cdots, X_k$之间的关系。
- 确认(前面例子中的)响应变量和协变量。
- 响应变量的其他名称：响应，因变量，结果
- 协变量的其他名称：预测变量，解释变量，回归变量，变量，协变量，独立变量
- 不同角度去看变量之间的关系时会产生不同的回归模型。均值回归，中位数回归，线性回归，非线性回归，参数回归，非参数回归，半参数回归，……
- 新的数据类型不断涌现，例如，函数型数据，线性模型的内涵在不断的开展。

不同领域对变量额称谓有不同的习惯，但本质上是一样的。

目标:了解总体的参数

- 我们经常想知道总体的一个参数。例子:
  - ♠ 中国居民的平均收入
  - ♠ 中国人的平均收入与受教育年限之间的关系，受教育时间越长，平均收入增加多少？
- 联系到每个人并询问他们的收入是不可行的.
- 所以我们永远不会确切知道总体参数。
- 测量和参数的历史变迁。

解决方案: 利用样本

- 解决方案:利用样本
  - 我们从人群中收集随机样本的数据。
  - 我们用样本中的平均收入来**估计**总体的平均收入。
  - 估计是随机的: 取一个新的样本会得到一个不同的估计。
  - 估计=总体参数+随机误差。
  - 为了从估计得到关于总体参数的结论，我们需要知道估计量的性质:
    - 误差有多大?
    - 误差与样本容量有什么关系?
  - 因此，我们在这门课程中花费很多时间来研究回归估计的分布。《统计推断》中的许多理论和方法在这里有用武之地。

理想的抽样方法

- 识别和确定总体
- 列出总体中所有的个体
- 用概率法随机抽取样本(这意味着你知道总体的每个个体被抽到的概率[可能相等也可能不等])
- 然后将样本结论推广到总体

Example (现实中抽样的例子)

- 例子:
  - 我们想考察两种不同教学方法的效果。
  - 我们将某一高中班级的学生随机分为两种教学方法。
  - 我们发现A方法明显更好。
  - 你在另一所高中教书。你会改用A方法吗?
    - 从技术上讲，我们不能把方法A推广到那个特定高中特定班级以外的地方。
    - 但是如果你在另一所高中的班级是“相似的”，那么我们假设所得结果对新的班级也有效就是合理的。那么我们就会改变。

现实中抽样的例子

Example (现实中抽样的例子)

- 例子:
  - 一项医学研究想要检验一种药物的功效
  - 为此，要招募志愿者，然后随机化分为两组，一组接受药物，另外一组接受安慰剂[这里可能有双盲随机化实验的问题]
  - 研究发现两组之间有显著差异
  - 当局应该做出什么决定?
    - 志愿者可能不同于一般人群
    - 将研究组的几个特征与普通人群进行比较和验证
    - 如果它们看起来很相似，批准药物用于人群
    - 如果他们看起来非常不同，批准药物用于亚组(小的范围)，或进行进一步研究

【第一讲结束】

方便样本和推广

- 经常使用方便样本。
  例如，附近学校的学生，特定医院的病人。
- 我们经常想要推广到样本之外的人群。
  - 如果抽样的总体与要推广的总体相似，这是合理的。
- 在本课程中，我们总是假设我们有一个来自总体的代表性样本，每个个体都有相同的概率存在于总体中。

Example (监狱数据)

- 参加教育项目是否会降低再次被捕的几率?
- 因变量: 再次被捕(1=yes, 0=no)。
  自变量: 参与(1=yes, 0=no)。
-

	参与教育项目	未参与教育项目
再次被捕	10	50
未再次被捕	40	50
总数	50	100

- 在参与教育项目的人中，20%的人再次被捕。
- 在没有参与教育项目的人中，有50%的人再次被捕。

这意味着什么？

- 参加教育计划是否会降低再次被逮捕的几率？
- 这取决于具体的研究：
  - 如果囚犯决定是否参与研究—结论是：否。 观测研究
    - ✓ 差异可能是由于选择参与的人与选择不参与的人的系统原因造成的。
    - ✓ 想一想：犯罪类型，重新融入社会的动机，等等。
  - 如果囚犯被随机分配参与或不参与——结论可能：是。但也不是绝对确定：
    - ✓ 例如，可能是看管人员对这两个群体采取不同的行为。

实验研究

观察研究与实验研究

- 观察研究与实验研究的关键区别:
  - 观察性研究: 受试者决定治疗分配(例如: 吸烟者与非吸烟者，饮食选择)
  - 实验研究: 研究人员决定治疗分配(例如: 许多医学研究)
- 参见上面不同类型的研究
- 同一研究问题有可能是观测研究也可能是实验研究。二者得到的结论的可靠性是不一样的。
  例如研究儿童的身高发育随时间的变化。

观察性研究中的解决方案

- 比较除您感兴趣的因素之外相似的亚组。例子:
  - 比较积极参与的囚犯和不积极参与的囚犯
  - 比较无动机参与的囚犯和无动机不参与的囚犯
- 这叫做控制因素动机。
- 在回归中，我们可以通过将一个因子放入模型中来控制它。(后面我们还将回到这个问题上。)
- 问题: 我们永远无法确定我们是否控制了所有可能的相关因素。
- 尽管如此，这并不足以使每一项观察性研究丧失可信度。要质疑一项研究，你需要有说服力地辩称，可能是某个特定因素导致了这种模式。

实验研究中的解决方案

- 确保治疗分配是随机的
- 尽可能使用盲法:
  - 参与者致盲
  - 评价者/研究人员致盲

混杂因素

- 像监狱例子中的动机这样的因素被称为混淆因素(confounding factor)。
- 定义:
  - 这个因素影响因变量/结果
  - 这一因素与本研究关注的协变量有关
- 如果满足这两个条件，那么混淆因子的效应和感兴趣的协变量的影响就会混淆(混淆)。我们就无法确定导致这种影响的原因是什么。【在囚犯的例子中，到底是接受教育还是囚犯的动机导致了再次被捕难以区分】
- 参见上面的示例
- 关于混杂更准确的定义和判断北京大学的耿直教授有深入的研究。

回到监狱数据的例子

- 监狱的例子:
  - 动机影响再次被捕的机会
  - 动机与参与教育项目有关(参与的人更有动机)。
- 所以:
  - 那些积极参与该计划的囚犯群体很少再次被捕。
  - 这群没有动机和没有参与该计划的囚犯经常被再次逮捕。
- 我们不知道再次逮捕率的差异是由动机引起的，还是由参与该项目引起的。这些效果混杂了(混合在一起了)。

Example (加拿大难民上诉裁决问题)

- 对于加拿大难民的上诉申请，不同的法官是否做出类似的裁决?
- 加拿大难民数据(Fox，2016，3rd ed.，Table1.1，第5页，原表中有12个法官判决的结果)

法官	许可上诉	不许可上诉
Pratte	9%	91%
Desjardins	49%	51%

- 这些数据成为法院判决加拿大难民确定程序公平性的案例的基础.
- 因变量: 是否许可上诉(是/否)。
  自变量: judge (Pratte/Desjardin)。

John Fox. Applied Regression Analysis and Generalized Linear Models, 3rd Ed. SAGE Publications, Inc.

性别是一个混杂的因素吗?

- 情景1: 法官更倾向于给女性上诉，而Desjardins的女性申请者比例更高。
- 情景2: 法官更有可能给女性上诉，而且两名法官的女性申请者比例大致相同。
- 情景3: 申请人的性别不影响法官的决定，Desjardins的女性申请人比例较高。

混杂

随机化实验

- 混杂因素在随机试验中不是问题。
- 为什么?
- 自变量表示治疗组。通过随机化，治疗组在所有方面都将变得大致相同。因此，混杂因素的定义的第二个条件永远不满足。
- 所以我们总是愿意做随机实验。
- 但随机实验在特定情景中不总是可能的，也不总是合乎道德的。例如: 吸烟、气候变化。为了研究吸烟对疾病的影响，我们不可以将病人做随机化，一组吸烟，另外一组不吸烟。这是道德和伦理所不允许的。与人有关的医药实验在开始之前要得到伦理委员会的审核和批准。

参考书[数学较为浅显]

- ♡ Norman R. Draper and Harry Smith (1998). Applied Regression Analysis, 3rd Ed. Wiley.
- John Fox (1997), "Applied Regression Analysis, Linear Models, and Related Methods", Sage Publications.
- Sanford Weisberg (2005), "Applied Linear Regression", 3rd edition, Wiley.
- Paul D. Allison (1999), "Multiple linear regression, a primer", Thousand Oaks.
- Peter Dalgaard (2002), "Introductory Statistics with R", Springer.
- Cook, R. D. and Weisberg, S. (1999). Applied Regression including Computing and Graphics. New York, NY Wiley.

这些书中的数学都比较浅显，但有非常直观的例子，阅读这些书对于培养统计直觉非常有益。

参考书[数学相对严谨]

- ♡ G. A. F. Seber and A. J. Lee (2003). Linear regression analysis. Wiley.
- David J. Olive (2017). Linear Regression. Springer.
- C. R. Rao (1973). Linear Statistical Inference and its Application. 2nd ed. New York: Wiley. (中文版：C. R. 劳(1987). 线性统计推断及其应用，科学出版社，北京. 以严谨形式论述统计推断的最新理论与技巧（70年代以前）)
- 陈希孺，陈桂景，吴启光，赵林城(2010). 线性模型参数的估计理论。北京: 科学出版社(是作者在数理统计线性模型参数估计理论方面所做研究工作的总结。非常的数学！)