因果推断基础

王也
纽约大学政治系

人民大学, 2019年7月3日

个人介绍

- 2007-2011 复旦大学数学系

个人介绍

- 2007-2011 复旦大学数学系
- 2011-2014 北京大学中国经济研究中心

个人介绍

- 2007-2011 复旦大学数学系
- 2011-2014 北京大学中国经济研究中心
- 2014-2015 威斯康星大学麦迪逊分校经济系博士

个人介绍

- 2007-2011 复旦大学数学系
- 2011-2014 北京大学中国经济研究中心
- 2014-2015 威斯康星大学麦迪逊分校经济系博士
  (太冷, 转学了)

个人介绍

- 2007-2011 复旦大学数学系
- 2011-2014 北京大学中国经济研究中心
- 2014-2015 威斯康星大学麦迪逊分校经济系博士 (太冷, 转学了)
- 2015至今 纽约大学政治系博士

个人介绍

- 2007-2011 复旦大学数学系
- 2011-2014 北京大学中国经济研究中心
- 2014-2015 威斯康星大学麦迪逊分校经济系博士 (太冷, 转学了)
- 2015至今 纽约大学政治系博士
- 领域: 政治学研究方法, 当代威权主义和政治转型

个人介绍

- 2007-2011 复旦大学数学系
- 2011-2014 北京大学中国经济研究中心
- 2014-2015 威斯康星大学麦迪逊分校经济系博士 (太冷, 转学了)
- 2015至今 纽约大学政治系博士
- 领域: 政治学研究方法, 当代威权主义和政治转型
- 2014年至今, 政见(CNPolitics)撰稿人

概览

- 基本概念
- 随机实验中的估计和推断
- 随机实验中的不顺从和干涉
- 因果推断中的回归
- 分块, 加权和匹配
- 双重稳健性
- 机器学习的应用
- 敏感性检验

另一条线索: 数据结构和假设

- 只有Y和D的情形(简单随机实验, 第一课)
- 有Y, D和Z的情形(简单随机实验中的不顺从和干涉, 第二课)
- 有Y, D和X且D外生的情形(实验的效率和异质性, 第二课)
- 有Y, D和X且D内生的情形(分块实验和观察性研究, 第三课)

社会科学和因果关系

因果关系在社会科学中无处不在：如果某个因素X改变了，结果Y会怎么变化？

- 征收房产税对购房需求有什么样的影响？
- 提供生活补贴是否能提高贫困学生在大学中的表现？
- 经济衰退是否会增加内战发生的可能性？

一般来说，我们关心的是某个原因的结果，而不是某个结果的原因（为什么？）

社会科学和因果关系

定义因果关系需要反事实(counterfactual)的概念

社会科学和因果关系

定义因果关系需要反事实(counterfactual)的概念

- 平行宇宙: 如果在那个时刻, 那个环境, 其他条件不变, 而X的取值从X1变成了X2, Y会如何相应改变?

社会科学和因果关系

定义因果关系需要反事实(counterfactual)的概念

- 平行宇宙: 如果在那个时刻, 那个环境, 其他条件不变, 而X的取值从X1变成了X2, Y会如何相应改变?
- 理想手段: 时光机器

社会科学和因果关系

定义因果关系需要反事实(counterfactual)的概念

- 平行宇宙: 如果在那个时刻, 那个环境, 其他条件不变, 而X的取值从X1变成了X2, Y会如何相应改变?
- 理想手段: 时光机器
- 现实情况: 只能观察到一种可能性

社会科学和因果关系

定义因果关系需要反事实(counterfactual)的概念

- 平行宇宙: 如果在那个时刻, 那个环境, 其他条件不变, 而X的取值从X1变成了X2, Y会如何相应改变?
- 理想手段: 时光机器
- 现实情况: 只能观察到一种可能性
- 注意: 有些时候反事实未必存在

Rubin模型

发明人: 清华大学的Donald Rubin教授

Rubin模型

发明人: 清华大学的Donald Rubin教授

Not!

Rubin模型

发明人: 清华大学的Donald Rubin教授

Not!
历史可以追溯到Neyman (1923)

Rubin模型

- 潜在结果 (potential outcome):

$$Y_i\ (\text{健康状况}) = \begin{cases} Y_i(1) \text{ if } D_i = 1\ (\text{吃了药}) \\ Y_i(0) \text{ if } D_i = 0\ (\text{没吃药}) \end{cases}$$

- 处理效应 (treatment effect):

$$\tau_i = Y_i(1) - Y_i(0)$$

- 平均处理效应 (ATE):

$$E[\tau_i] = E[Y_i(1)] - E[Y_i(0)]$$

Rubin模型

- 潜在结果 (potential outcome):

$$Y_i\ (\text{健康状况}) = \begin{cases} Y_i(1) \text{ if } D_i = 1\ (\text{吃了药}) \\ Y_i(0) \text{ if } D_i = 0\ (\text{没吃药}) \end{cases}$$

- 处理效应 (treatment effect):

$$\tau_i = Y_i(1) - Y_i(0)$$

- 平均处理效应 (ATE):

$$E[\tau_i] = E[Y_i(1)] - E[Y_i(0)]$$

写下这个模型的时候，其实我们已经假设了“稳定单位处理取值 (Stable Unit Treatment Value Assumption, or SUTVA)”

Rubin模型

- 潜在结果 (potential outcome):

$$Y_i\ (\text{健康状况}) = \begin{cases} Y_i(1) \text{ if } D_i = 1\ (\text{吃了药}) \\ Y_i(0) \text{ if } D_i = 0\ (\text{没吃药}) \end{cases}$$

- 处理效应 (treatment effect):

$$\tau_i = Y_i(1) - Y_i(0)$$

- 平均处理效应 (ATE):

$$E[\tau_i] = E[Y_i(1)] - E[Y_i(0)]$$

写下这个模型的时候, 其实我们已经假设了“稳定单位处理取值 (Stable Unit Treatment Value Assumption, or SUTVA)”
思考题: 一般来说, $Y_i(1)$和$Y_i(0)$这两个随机变量是独立的吗?

“因果推断的基本问题”

对于任意个体i, 我们不可能同时观察到$Y_i(1)$和$Y_i(0)$ (Holland, 1986)

“因果推断的基本问题”

对于任意个体i, 我们不可能同时观察到$Y_i(1)$和$Y_i(0)$ (Holland, 1986)
怎么办?

“因果推断的基本问题”

对于任意个体i, 我们不可能同时观察到$Y_i(1)$和$Y_i(0)$ (Holland, 1986)

怎么办?

- 科学方案 (施加假设)
- 统计学方案 (增大样本, 进行实验)

本质上来说, 因果推断是一个数据缺失问题: 对于处理组个体, 给定观察到的$Y_i(1)$, 如何推断其另一个潜在结果$Y_i(0)$?

科学方案

常见于中学课本: ”奥斯特发现, 导线通电之后, 旁边的小磁针会发生偏转”

科学方案

常见于中学课本: ”奥斯特发现, 导线通电之后, 旁边的小磁针会发生偏转”

- 这里的X和Y分别是什么?
- 做出因果论断依赖于怎样的假设?
- 为什么科学方案在社会科学里较少使用?

统计学方案

技能培训是否能帮助中国建筑工人在劳动力市场上获得更高的工资? 常见步骤:

- 从全国的建筑工人中抽取一个子样本
- 随机分配到培训组和控制组
- 比较培训之后两组的平均工资差异

思考题: 如何用潜在因果模型描述该研究设计?

统计学方案

假定处理D是“外生”的, 独立于潜在结果$Y_i(1)$和$Y_i(0)$:

$$D_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\}$$

统计学方案

假定处理D是“外生”的, 独立于潜在结果$Y_i(1)$和$Y_i(0)$:

$$D_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\}$$

那么

$$\begin{aligned} E[\tau_i] &= E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \end{aligned}$$

在实际操作中,

$$\widehat{ATE} = \frac{1}{N_1} \sum_{D_i=1} Y_i - \frac{1}{N_0} \sum_{D_i=0} Y_i$$

思考题1: 这里每个$Y_i(1)$的反事实是什么?
思考题2: 我们得到的”平均处理效应”是针对哪个群体而言的?

统计学方案

- 我们感兴趣的往往是对于整个人群的平均处理效应, 或者说总体平均处理效应(PATE), 但实验中能够得到的只是样本平均处理效应(SATE)的一个估计

统计学方案

- 我们感兴趣的往往是对于整个人群的平均处理效应, 或者说总体平均处理效应(PATE), 但实验中能够得到的只是样本平均处理效应(SATE)的一个估计
- $\widehat{ATE}$跟PATE的差异由什么造成?

统计学方案

- 我们感兴趣的往往是对于整个人群的平均处理效应, 或者说总体平均处理效应(PATE), 但实验中能够得到的只是样本平均处理效应(SATE)的一个估计
- $\widehat{ATE}$跟PATE的差异由什么造成?
- 不确定性的两个来源: 抽样误差和设计误差

统计学方案

- 我们感兴趣的往往是对于整个人群的平均处理效应, 或者说总体平均处理效应(PATE), 但实验中能够得到的只是样本平均处理效应(SATE)的一个估计
- $\widehat{ATE}$跟PATE的差异由什么造成?
- 不确定性的两个来源: 抽样误差和设计误差
- 估计描述性统计量只需要考虑抽样误差, 估计SATE只需要考虑设计误差

统计学方案

- 我们感兴趣的往往是对于整个人群的平均处理效应, 或者说总体平均处理效应(PATE), 但实验中能够得到的只是样本平均处理效应(SATE)的一个估计
- $\widehat{ATE}$跟PATE的差异由什么造成?
- 不确定性的两个来源: 抽样误差和设计误差
- 估计描述性统计量只需要考虑抽样误差, 估计SATE只需要考虑设计误差
- 有时候我们想知道培训对于不同子群体产生的效应, 即条件平均处理效应(CATE):

$$\tau(x) = E[\tau_i|X_i = x] = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$$

统计学方案

- 我们感兴趣的往往是对于整个人群的平均处理效应, 或者说总体平均处理效应(PATE), 但实验中能够得到的只是样本平均处理效应(SATE)的一个估计
- $\widehat{ATE}$跟PATE的差异由什么造成?
- 不确定性的两个来源: 抽样误差和设计误差
- 估计描述性统计量只需要考虑抽样误差, 估计SATE只需要考虑设计误差
- 有时候我们想知道培训对于不同子群体产生的效应, 即条件平均处理效应(CATE):

$$\tau(x) = E[\tau_i|X_i = x] = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$$

思考题: 这里的X跟回归分析中常见的"控制变量"所扮演的角色相同吗?

评判统计学方案的标准

- 几个概念: estimand (待估计量), estimator (估计量), estimate (估计值)

评判统计学方案的标准

- 几个概念: estimand (待估计量), estimator (估计量), estimate (估计值)
- 我们希望重复多次实验, 由estimator得到的estimate平均值等于estimand: 无偏性 (unbiasedness)

评判统计学方案的标准

- 几个概念: estimand (待估计量), estimator (估计量), estimate (估计值)
- 我们希望重复多次实验, 由estimator得到的estimate平均值等于estimand: 无偏性 (unbiasedness)
- 我们希望随着样本量增加, 由estimator得到的estimate趋近于estimand: 一致性 (consistency)

评判统计学方案的标准

- 几个概念: estimand (待估计量), estimator (估计量), estimate (估计值)
- 我们希望重复多次实验, 由estimator得到的estimate平均值等于estimand: 无偏性 (unbiasedness)
- 我们希望随着样本量增加, 由estimator得到的estimate趋近于estimand: 一致性 (consistency)
- 我们希望estimate和estimand的差异尽可能小: 有效性 (efficiency)

评判统计学方案的标准

- 几个概念: estimand (待估计量), estimator (估计量), estimate (估计值)
- 我们希望重复多次实验, 由estimator得到的estimate平均值等于estimand: 无偏性 (unbiasedness)
- 我们希望随着样本量增加, 由estimator得到的estimate趋近于estimand: 一致性 (consistency)
- 我们希望estimate和estimand的差异尽可能小: 有效性 (efficiency)
- 有效性越高, 标准误越小, 结果越容易显著

评判统计学方案的标准

- 几个概念: estimand (待估计量), estimator (估计量), estimate (估计值)
- 我们希望重复多次实验, 由estimator得到的estimate平均值等于estimand: 无偏性 (unbiasedness)
- 我们希望随着样本量增加, 由estimator得到的estimate趋近于estimand: 一致性 (consistency)
- 我们希望estimate和estimand的差异尽可能小: 有效性 (efficiency)
- 有效性越高, 标准误越小, 结果越容易显著

思考题: 哪些estimator有偏但一致, 哪些无偏但不一致?

为什么要做实验?

- 实验是当今社会科学中的“黄金标准”
- 实验的好处: 内部效度(internal validity)非常高, 能够保证得到因果关系

为什么要做实验?

- 实验是当今社会科学中的“黄金标准”
- 实验的好处: 内部效度(internal validity)非常高, 能够保证得到因果关系
- 相比之下

为什么要做实验?

- 实验是当今社会科学中的“黄金标准”
- 实验的好处: 内部效度(internal validity)非常高, 能够保证得到因果关系
- 相比之下

  案例研究: 可以揭示内在机制, 但样本量小, 可信度低

为什么要做实验?

- 实验是当今社会科学中的“黄金标准”
- 实验的好处: 内部效度(internal validity)非常高, 能够保证得到因果关系
- 相比之下

  案例研究: 可以揭示内在机制, 但样本量小, 可信度低
  观察性研究: 代表性强, 但无法排除混淆变量

为什么要做实验?

- 实验是当今社会科学中的“黄金标准”
- 实验的好处: 内部效度(internal validity)非常高, 能够保证得到因果关系
- 相比之下

  案例研究: 可以揭示内在机制, 但样本量小, 可信度低
  观察性研究: 代表性强, 但无法排除混淆变量
  规范性研究: 需要坚实的因果关系作为基础

为什么要做实验?

- 实验的缺点:

为什么要做实验?

- 实验的缺点: 也很多!

为什么要做实验?

- 实验的缺点: 也很多!
- 贵, 而且越来越贵

为什么要做实验?

- 实验的缺点: 也很多!
- 贵, 而且越来越贵
- 能研究的问题十分有限

为什么要做实验?

- 实验的缺点: 也很多!
- 贵, 而且越来越贵
- 能研究的问题十分有限
- 结论的外部效度(external validity)低

为什么要做实验?

- 实验的缺点: 也很多!
- 贵, 而且越来越贵
- 能研究的问题十分有限
- 结论的外部效度(external validity)低
- 未必能反映真实的机制

为什么要做实验?

实验方法的风行改变了我们思考研究设计的方式

为什么要做实验?

实验方法的风行改变了我们思考研究设计的方式
设想如下的观察性研究:

- 从全国的建筑工人中抽取一个子样本做问卷调查
- 基于调查得到的数据进行回归分析
- 因变量和自变量分别是工资收入和是否接受过技能培训, 控制性别年龄工龄等协变量

为什么要做实验?

实验方法的风行改变了我们思考研究设计的方式
设想如下的观察性研究:

- 从全国的建筑工人中抽取一个子样本做问卷调查
- 基于调查得到的数据进行回归分析
- 因变量和自变量分别是工资收入和是否接受过技能培训, 控制性别年龄工龄等协变量

由此得到的回归系数跟平均处理效应有什么关系? 其标准误在多大程度上反映了真实的不确定性?

为什么要做实验?

实验方法的风行改变了我们思考研究设计的方式
设想如下的观察性研究:

- 从全国的建筑工人中抽取一个子样本做问卷调查
- 基于调查得到的数据进行回归分析
- 因变量和自变量分别是工资收入和是否接受过技能培训, 控制性别年龄工龄等协变量

由此得到的回归系数跟平均处理效应有什么关系? 其标准误在多大程度上反映了真实的不确定性?
我们可以将观察性研究想象成由"自然"设计并执行的随机实验, 并运用实验分析的工具去理解其结果

为什么要做实验?

实验方法的风行改变了我们思考研究设计的方式
设想如下的观察性研究:

- 从全国的建筑工人中抽取一个子样本做问卷调查
- 基于调查得到的数据进行回归分析
- 因变量和自变量分别是工资收入和是否接受过技能培训, 控制性别年龄工龄等协变量

由此得到的回归系数跟平均处理效应有什么关系? 其标准误在多大程度上反映了真实的不确定性?
我们可以将观察性研究想象成由"自然"设计并执行的随机实验, 并运用实验分析的工具去理解其结果
这被称为"基于设计的视角 (design-based perspective)"

为什么要做实验?

- 传统回归分析: 基于模型的视角 (model-based perspective)

为什么要做实验?

- 传统回归分析: 基于模型的视角 (model-based perspective)
- 假定线性模型是正确的, 培训的效应是恒定的, 不确定性来源于模型中的随机扰动项

为什么要做实验?

- 传统回归分析: 基于模型的视角 (model-based perspective)
- 假定线性模型是正确的, 培训的效应是恒定的, 不确定性来源于模型中的随机扰动项
- 但这些假设 1. 很难满足, 2. 无法验证, 3. 没有理论含义, 而且结果不太具有现实意义

为什么要做实验?

- 传统回归分析: 基于模型的视角 (model-based perspective)
- 假定线性模型是正确的, 培训的效应是恒定的, 不确定性来源于模型中的随机扰动项
- 但这些假设 1. 很难满足, 2. 无法验证, 3. 没有理论含义, 而且结果不太具有现实意义
- 更加有启发性的问题: 我们是如何得到这些样本的, 他们是否能代表我们感兴趣的总体 (抽样过程)? 为什么有些个体得到了处理有些没有, 背后的机制是什么 (分配过程)?

随机实验中的分配

最简单的实验: 只有Y和D

- 两种基本方式: 完全随机化(complete randomization)和伯努利随机化(bernoulli randomization)

随机实验中的分配

最简单的实验: 只有Y和D

- 两种基本方式: 完全随机化(complete randomization)和伯努利随机化(bernoulli randomization)
- 完全随机化: 从N个实验对象中随机抽取M个进入处理组(D=1), 余下N - M个进入控制组(D=0)

随机实验中的分配

最简单的实验: 只有Y和D

- 两种基本方式: 完全随机化(complete randomization)和伯努利随机化(bernoulli randomization)
- 完全随机化: 从N个实验对象中随机抽取M个进入处理组(D=1), 余下N - M个进入控制组(D=0)
- 伯努利随机化: 每一个实验对象有p的概率进入处理组, 1-p的概率进入控制组

随机实验中的分配

最简单的实验: 只有Y和D

- 两种基本方式: 完全随机化(complete randomization)和伯努利随机化(bernoulli randomization)
- 完全随机化: 从N个实验对象中随机抽取M个进入处理组(D=1), 余下N - M个进入控制组(D=0)
- 伯努利随机化: 每一个实验对象有p的概率进入处理组, 1-p的概率进入控制组
- 例子: 样本里有100位工人, 我们是随机抽取50位分配进处理组, 还是对每位工人扔一次骰子?

思考题: 这两种分配方式各自有什么好处?

随机实验中的估计

- 完全随机化: 组间均值之差(group mean difference)

$$
\begin{aligned}
\hat{\tau}_{gmd} = \widehat{ATE} &= \frac{1}{N_1} \sum_{D_i=1} Y_i - \frac{1}{N_0} \sum_{D_i=0} Y_i \\
&= \frac{1}{N_1} \sum_i Y_i D_i - \frac{1}{N_0} \sum_i Y_i (1 - D_i)
\end{aligned}
$$

随机实验中的估计

- 完全随机化: 组间均值之差(group mean difference)

$$\hat{\tau}_{gmd} = \widehat{ATE} = \frac{1}{N_1}\sum_{D_i=1} Y_i - \frac{1}{N_0}\sum_{D_i=0} Y_i$$
$$= \frac{1}{N_1}\sum_{i} Y_i D_i - \frac{1}{N_0}\sum_{i} Y_i(1-D_i)$$

霍洛维茨-汤普森估计量(HorvitzThompson estimator)

随机实验中的估计

- 完全随机化: 组间均值之差(group mean difference)

$$\hat{\tau}_{gmd} = \widehat{ATE} = \frac{1}{N_1}\sum_{D_i=1} Y_i - \frac{1}{N_0}\sum_{D_i=0} Y_i$$
$$= \frac{1}{N_1}\sum_i Y_i D_i - \frac{1}{N_0}\sum_i Y_i(1-D_i)$$

霍洛维茨-汤普森估计量(HorvitzThompson estimator)

- H-T估计量是无偏的

$$E\left[\hat{\tau}_{gmd}\right] = E\left[\frac{1}{N_1}\sum_i Y_i D_i - \frac{1}{N_0}\sum_i Y_i(1-D_i)\right]$$
$$= \frac{1}{N_1}\sum_i E\left[Y_i D_i\right] - \frac{1}{N_0}\sum_i E\left[Y_i(1-D_i)\right]$$
$$= E[Y_i|D_i=1] - E[Y_i|D_i=0] = E[\tau_i]$$

随机实验中的估计

- 伯努利随机化: N_1和N_0分别是多少?

随机实验中的估计

- 伯努利随机化: N_1和N_0分别是多少?
- 例子: 如果我们想让一半样本(50人)接受处理, 但实际上52人被分配进入处理组, 分母应该是50还是52?

随机实验中的估计

- 伯努利随机化: N_1和N_0分别是多少?
- 例子: 如果我们想让一半样本(50人)接受处理, 但实际上52人被分配进入处理组, 分母应该是50还是52?
- 答案: 都可以, 但后者效率更高

随机实验中的估计

- 更简单的方法: 回归估计
- $Y_i = \alpha + \tau D_i + \varepsilon_i$
- 问题: 组间均值之差得到的估计跟回归得到的估计是否相等?

随机实验中的估计

- 更简单的方法: 回归估计
- $Y_i = \alpha + \tau D_i + \varepsilon_i$
- 问题: 组间均值之差得到的估计跟回归得到的估计是否相等?
- 答案: 是的!
  注意到 $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$
  $\hat{\tau}_{ols} = \frac{Cov(Y_i,D_i)}{Var(D_i)} = \frac{Cov(Y_i(1)D_i+Y_i(0)(1-D_i),D_i)}{Var(D_i)} =$
  $E[Y_i(1)] - E[Y_i(0)] = ATE$

随机实验中的估计

- 更简单的方法: 回归估计
- $Y_i = \alpha + \tau D_i + \varepsilon_i$
- 问题: 组间均值之差得到的估计跟回归得到的估计是否相等?
- 答案: 是的!
  注意到 $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$
  $\hat{\tau}_{ols} = \frac{Cov(Y_i,D_i)}{Var(D_i)} = \frac{Cov(Y_i(1)D_i+Y_i(0)(1-D_i),D_i)}{Var(D_i)} =$
  $E[Y_i(1)] - E[Y_i(0)] = ATE$

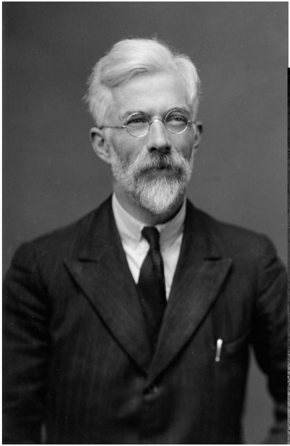
在只有Y和D的随机实验中, 回归可以得到对因果效应的无偏估计

随机实验中的推断(Inference)

- 余下的问题: 估计的标准误是多少? 是否统计显著?
- 统计推断的基本思想: 证伪
- 假设某个零假设成立(效应为零), 我们得到的估计是否跟这个假设矛盾?

随机实验中的推断(Inference)

Fisher vs. Neyman: 延续百年的争论



随机实验中的推断(Inference)

Fisher的方法: 随机推断(randomization inference)

随机实验中的推断(Inference)

Fisher的方法: 随机推断(randomization inference)

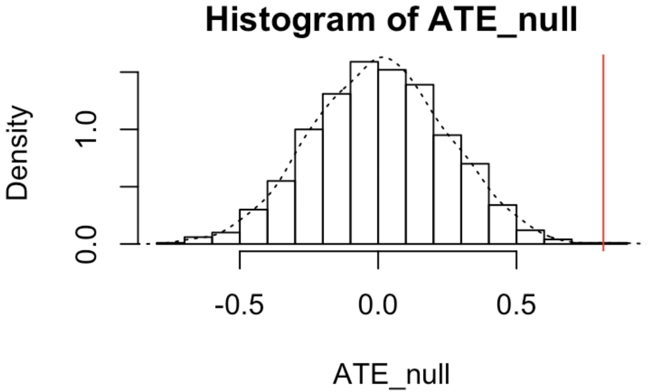
- 零假设(null hypothesis): $\tau_i = 0$ for any i
- 处理对于任何个体的效应都是零(严格零假设, sharp null)

随机实验中的推断(Inference)

Fisher的方法: 随机推断(randomization inference)

- 零假设(null hypothesis): $\tau_i = 0$ for any i
- 处理对于任何个体的效应都是零(严格零假设, sharp null)
- 如果严格零假设成立, 那么$Y_i(1) = Y_i(0) + \tau_i = Y_i(0)$, 我们同时观察到了每个个体的全部潜在结果!
- 想知道分配过程带来的不确定性, 只需要重复分配过程即可
- 一般来说只需要重复多次(1000次)
- Fisher随机推断可以用于更一般的研究, 要求我们对分配过程有足够的了解

随机实验中的推断(Inference)



随机实验中的推断(Inference)

Neyman的方法: 先分析方差, 再推导渐进分布, 以得到p值

- 方差推导: $\hat{\tau}$的不确定性来自哪里?

随机实验中的推断(Inference)

Neyman的方法: 先分析方差, 再推导渐进分布, 以得到p值

- 方差推导: $\hat{\tau}$的不确定性来自哪里?
- 在随机抽样和随机分配的情况下, 假定总体包含n个个体: $Var(\hat{\tau}) = Var(\bar{Y}_i(1)) + Var(\bar{Y}_i(0)) - 2Cov(\bar{Y}_i(1), \bar{Y}_i(0)) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{n}$
- 方差表达式由三部分组成: $Y_i(1)$的样本方差, $Y_i(0)$的样本方差, 以及参数的总体方差

随机实验中的推断(Inference)

Neyman的方法: 先分析方差, 再推导渐进分布, 以得到p值

- 方差推导: $\hat{\tau}$的不确定性来自哪里?
- 在随机抽样和随机分配的情况下, 假定总体包含n个个体: $Var(\hat{\tau}) = Var(\bar{Y}_i(1)) + Var(\bar{Y}_i(0)) - 2Cov(\bar{Y}_i(1), \bar{Y}_i(0)) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{n}$
- 方差表达式由三部分组成: $Y_i(1)$的样本方差, $Y_i(0)$的样本方差, 以及参数的总体方差
- Abadie et al. (2017): $Var(\hat{\tau}) = Var_{design|sampling} + Var_{sampling} = Var_{sampling|design} + Var_{design}$
- 比如说, $Var_{design} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}$, $Var_{sampling|design} = \frac{S_\tau^2}{N}(1 - \frac{N}{n})$

随机实验中的推断(Inference)

- 在方差公式中, S_1^2和S_0^2都可以基于数据估计, 但S_τ^2中包含了$Y_i(0)$和$Y_i(1)$的协方差, 无法估计

随机实验中的推断(Inference)

- 在方差公式中, S_1^2和S_0^2都可以基于数据估计, 但S_τ^2中包含了$Y_i(0)$和$Y_i(1)$的协方差, 无法估计
- 但很显然, $Var(\hat{\tau}) \leq \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$
- 因此一般就用$\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$作为实验估计的标准误
- 在有限总体中, 这是一个保守估计
- 在无限总体中, 第三项等于零, 保守估计等于真实估计

随机实验中的推断(Inference)

- 在方差公式中, S_1^2和S_0^2都可以基于数据估计, 但S_τ^2中包含了$Y_i(0)$和$Y_i(1)$的协方差, 无法估计
- 但很显然, $Var(\hat{\tau}) \leq \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$
- 因此一般就用$\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$作为实验估计的标准误
- 在有限总体中, 这是一个保守估计
- 在无限总体中, 第三项等于零, 保守估计等于真实估计
- Aronow et al. (2014) 给出了第三项的严格上下界, 由此可以得到更精确的标准误估计

随机实验中的推断(Inference)

回归估计也可以用于得到标准误, 但这个标准误是什么?

随机实验中的推断(Inference)

回归估计也可以用于得到标准误, 但这个标准误是什么?

- 回归中的HC2稳健 (HC2 robust) 标准误恰好等于$\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$
- 因此对于简单随机实验, 回归分析的结果跟Neyman框架下得到的估计和标准误是相等的
- 稳健标准误跟Neyman标准误也相差不大

随机实验中的推断(Inference)

但仅仅有标准误还不足以得到p值

- 我们还需要知道$\hat{\tau}$的分布
- 得到统计量分布的方法一般有两种: 渐进理论和自助法

随机实验中的推断(Inference)

但仅仅有标准误还不足以得到p值

- 我们还需要知道$\hat{\tau}$的分布
- 得到统计量分布的方法一般有两种: 渐进理论和自助法
- 渐进理论: 由中央极限定理可以知道, $\sqrt{N}(\hat{\tau}-\tau) \rightarrow N(0, Var(\hat{\tau}))$
- 在零假设下, $\tau = E[\tau_i] = 0$, 因此$\frac{\hat{\tau}}{Var(\hat{\tau})}$服从t分布
- 由t值可以得到p值和统计显著性

随机实验中的推断(Inference)

- 自助法 (bootstrap) 的核心思想是用样本分布函数近似总体分布函数
- 从样本中有放回地反复抽样, 并基于每个样本计算统计量的值, 由此可以得到任意统计量的分布

随机实验中的推断(Inference)

- 自助法 (bootstrap) 的核心思想是用样本分布函数近似总体分布函数
- 从样本中有放回地反复抽样, 并基于每个样本计算统计量的值, 由此可以得到任意统计量的分布
- 比如我们可以得到$\sqrt{N}(\hat{\tau} - \tau)$的分布, 并依此进行推断
- 自助法跟随机推断有相似之处, 但基于不同的理念 (零假设不同)

观察性研究中的不确定性

- 问题: 跨国回归中系数为什么会有不确定性?
- 我们已经观察到了所有感兴趣的个体, 此时系数标准误为什么不是零?

观察性研究中的不确定性

- 问题: 跨国回归中系数为什么会有不确定性?
- 我们已经观察到了所有感兴趣的个体, 此时系数标准误为什么不是零?
- 抽样误差是零, 但仍然存在设计误差
- 可以想象多个平行宇宙, 每个里面自变量的取值都有差异: 超总体 (super population)
- 每个设计都相当于是在超总体中进行了抽样

观察性研究中的不确定性

- 问题: 跨国回归中系数为什么会有不确定性?
- 我们已经观察到了所有感兴趣的个体, 此时系数标准误为什么不是零?
- 抽样误差是零, 但仍然存在设计误差
- 可以想象多个平行宇宙, 每个里面自变量的取值都有差异: 超总体 (super population)
- 每个设计都相当于是在超总体中进行了抽样

多值处理变量

- 目前我们一直假设处理D是一个二值变量

多值处理变量

- 目前我们一直假设处理D是一个二值变量
- 但实际上D也可以取多个值, 甚至可以是向量

多值处理变量

- 目前我们一直假设处理D是一个二值变量
- 但实际上D也可以取多个值, 甚至可以是向量
- 在D是标量的时候, 我们需要决定将其视作连续变量还是因子变量

多值处理变量

- 目前我们一直假设处理D是一个二值变量
- 但实际上D也可以取多个值, 甚至可以是向量
- 在D是标量的时候, 我们需要决定将其视作连续变量还是因子变量
- 如果是因子变量, 可以向回归中加入多个虚拟变量

多值处理变量

- 目前我们一直假设处理D是一个二值变量
- 但实际上D也可以取多个值, 甚至可以是向量
- 在D是标量的时候, 我们需要决定将其视作连续变量还是因子变量
- 如果是因子变量, 可以向回归中加入多个虚拟变量
- 如果是连续变量, 需要施加更强的假设: 处理效应如何随着变量取值变化: $TE(d) = f(d)$

多值处理变量

- 在D是向量的时候, 本质上是个析因设计 (factorial design)

多值处理变量

- 在D是向量的时候, 本质上是个析因设计 (factorial design)
- 我们不但可以分析每个分量的作用, 还能够检查它们的交互效应

多值处理变量

- 在D是向量的时候, 本质上是个析因设计 (factorial design)
- 我们不但可以分析每个分量的作用, 还能够检查它们的交互效应
- 例子

	政治极化	政治不极化
民主党主张	分支1	分支2
共和党主张	分支3	分支4

- 可以在回归中控制两个处理变量及其交叉项, 也可以逐组比较计算处理效应
- $Y_i = \alpha + \tau_1 D_{1i} + \tau_2 D_{2i} + + \tau_3 D_{1i} * D_{2i} + \varepsilon_i$ (饱和模型)

多值处理变量

- 常见的一种析因设计是联合分析法 (Conjoint Analysis)

多值处理变量

- 常见的一种析因设计是联合分析法 (Conjoint Analysis)
- 我们给被试提供两个在各个维度都有所不同的选项, 让他们不断进行二选一

多值处理变量

- 常见的一种析因设计是联合分析法 (Conjoint Analysis)
- 我们给被试提供两个在各个维度都有所不同的选项, 让他们不断进行二选一
- 例子

官员1	官员2
无博士学历	有博士学历
党龄长	党龄短
有基层经验	无基层经验
男性	男性
群众评价高	群众评价低

多值处理变量

- 常见的一种析因设计是联合分析法 (Conjoint Analysis)
- 我们给被试提供两个在各个维度都有所不同的选项, 让他们不断进行二选一
- 例子

官员1	官员2
无博士学历	有博士学历
党龄长	党龄短
有基层经验	无基层经验
男性	男性
群众评价高	群众评价低

- 由此我们可以识别各个维度对个体偏好的影响

多值处理变量

- 常见的一种析因设计是联合分析法 (Conjoint Analysis)
- 我们给被试提供两个在各个维度都有所不同的选项, 让他们不断进行二选一
- 例子

官员1	官员2
无博士学历	有博士学历
党龄长	党龄短
有基层经验	无基层经验
男性	男性
群众评价高	群众评价低

- 由此我们可以识别各个维度对个体偏好的影响
- 在没有交互项的情况下, 可以直接使用回归估计和标准误

多值处理变量

- 常见的一种析因设计是联合分析法 (Conjoint Analysis)
- 我们给被试提供两个在各个维度都有所不同的选项, 让他们不断进行二选一
- 例子

官员1	官员2
无博士学历	有博士学历
党龄长	党龄短
有基层经验	无基层经验
男性	男性
群众评价高	群众评价低

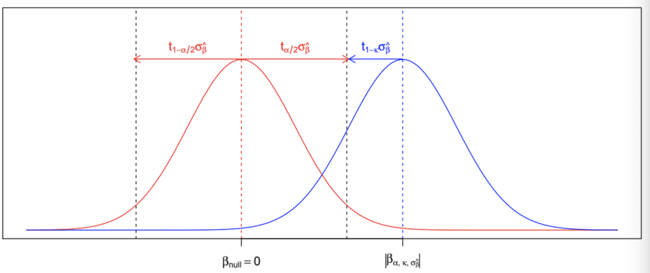
- 由此我们可以识别各个维度对个体偏好的影响
- 在没有交互项的情况下, 可以直接使用回归估计和标准误
- R包: cjoint

随机实验的效力 (Power)

- 随机实验在实践中面对的一大问题是: 到底需要多少样本?
- 因此目前的实验研究都要求事先对统计效力进行计算
- 统计效力能告诉我们: 如果处理效应真的存在, 那么我们检测不到它的概率有多大 (第二类错误)
- 一般来说, 我们要求这个概率小于20%
- 显然这个概率由效应的大小, 不确定性的大小, 和样本量决定

随机实验的效力 (Power)

- 实践中我们一般是给定效应和不确定性的大小, 计算所需的样本量
- 效应和不确定性的大小可以从pilot study或者先前的研究中得到
- 如图



一般来说, 我们需要效应是标准误的2.8倍来得到80%的效力

参考文献

- Angrist and Pischke: Mostly Harmless Econometrics
- Imbens and Rubin: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction
- Gerber and Green: Field Experiments: Design, Analysis, and Interpretation
- Aronow and Miller: Foundations of Agnostic Statistics
- Cochran: Sampling Technique

谢谢！