内容回顾

- 因果推断问题可以用Rubin模型加以刻画, 关键挑战是我们不能同时观察到两个潜在结果
- 在社会科学中, 实验是研究因果关系的标尺和出发点
- 实验中的分配有完全随机和伯努利随机两种
- 实验效应可以用霍洛维茨-汤普森估计量加以估计
- 估计实验效应的不确定性可以用随机推断, 也可以用Neyman标准误

随机实验中的不顺从(Non-compliance)和扩散(intervention)

很多时候分配的处理(assignment)不等于接受的处理(exposure)

- 病人嫌苦, 没有吃分配的药物
- 控制组的病人从处理组得到了药物
- 给一个村子接种疫苗, 其他村子被感染的概率也下降了

此时数据中同时有Y, D和Z

随机实验中的不顺从(Non-compliance)

- 当Z只通过D影响Y的时候, 我们称之为不顺从
- 本质上分配Z是实际处理D的一个工具变量(instrumental variable)
- 我们可以只估计Z的处理效应, 这被称为意向处理效应(intention-to-treat effect)
- 对于政策制定者来说, ITT effect可能意义更大
- 但对于理论检验来说, D的效应比Z的效应更有意义

随机实验中的不顺从(Non-compliance)

估计D的效应需要定义"主分层(principal strata)", 即D的取值如何随着Z而变化:

永远接受者 (Always-takers):	D(0) = 1, D(1) = 1
从不接受者 (Never-takers):	D(0) = 0, D(1) = 0
顺从者 (Compliers):	D(0) = 0, D(1) = 1
违逆者 (Defiers):	D(0) = 1, D(1) = 0

我们一般假定不存在违逆者, Z独立于$(D_i(0), D_i(1), Y_i(0), Y_i(1))$, 以及Z只通过D影响Y (排他性假设, exclusive restriction)

局部处理效应的估计和推断

- 由于处理分配根本不影响永远接受者和从不接受者的处理状态, 我们只能估计顺从者的处理效应
- 这种效应被称为局部处理效应(local average treatment effect, LATE)
- 如果顺从者在样本中的比例只有20%, 那么就是20%的人造成了观察到的差异
- 所以LATE等于ITT effect除以顺从者的占比:

$$LATE = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}$$

- 这被称为沃德(Wald)估计量

随机实验中的不顺从(Non-compliance)

- 我们并不知道样本中谁属于哪个主分层
- 但由于Z的外生性, 我们可以计算每个主分层的占比 (π_a, π_n, 和π_c)
- 这些比例在$Z = 0$和$Z = 1$两组中是相等的
- $D = 0, Z = 1$的人只可能是从不接受者, 而$D = 1, Z = 0$的人只可能是永远接受者
- $D = 0, Z = 0$的人既可能是顺从者又可能是从不接受者; $D = 1$, $Z = 1$的人既可能是顺从者又可能是永远接受者

随机实验中的不顺从(Non-compliance)

一个例子: 总共200实验对象, 处理组100人, 控制组100人

	D = 0	D = 1
Z = 0	60	40
Z = 1	20	80

随机实验中的不顺从(Non-compliance)

一个例子: 总共200实验对象, 处理组100人, 控制组100人

	D = 0	D = 1
Z = 0	60	40
Z = 1	20	80

$\pi_a = 40/100 = 0.4$, $\pi_n = 20/100 = 0.2$. 两个组里各有永远接受者40人, 从不接受者20人, 因此顺从者在各组都是40人

因此, $\pi_c = E[D|Z=1] - E[D|Z=0]$, 即处理组中顺从者 + 永远接受者的比例, 减去控制组中永远接受者的比例

沃德估计量的性质

- 沃德估计量的分子分母都是随机变量

沃德估计量的性质

- 沃德估计量的分子分母都是随机变量
- 有偏但一致 (biased asymptotically but asymptotically unbiased)
- 其标准误可以由泰勒展开 (Delta method) 计算得到
- 如果存在控制变量, 只需要将沃德估计量中的期望变成条件期望
- 如果存在控制变量, 我们可以计算顺从得分 (compliance score), 并由此推算ATE

沃德估计量和两阶段最小二乘法 (2SLS)

- 在观察性研究中, 一般用两阶段最小二乘法来估计LATE:

$$Y_i = \alpha_1 + \gamma D_i + \beta_1 X_i + \varepsilon_i$$
$$D_i = \alpha_2 + \delta Z_i + \beta_2 X_i + \nu_i$$

沃德估计量和两阶段最小二乘法 (2SLS)

- 在观察性研究中, 一般用两阶段最小二乘法来估计LATE:

$$
\begin{aligned}
Y_i &= \alpha_1 + \gamma D_i + \beta_1 X_i + \varepsilon_i \\
D_i &= \alpha_2 + \delta Z_i + \beta_2 X_i + \nu_i
\end{aligned}
$$

- 我们认为D_i可能跟ε_i相关, 但Z_i跟ε_i无关
- 因此先估计一阶段, 得到D_i预测值$\hat{D}_i$, 再用$\hat{D}_i$取代二阶段的D_i, 估计其系数

沃德估计量和两阶段最小二乘法 (2SLS)

- 在观察性研究中, 一般用两阶段最小二乘法来估计LATE:

$$
\begin{aligned}
Y_i &= \alpha_1 + \gamma D_i + \beta_1 X_i + \varepsilon_i \\
D_i &= \alpha_2 + \delta Z_i + \beta_2 X_i + \nu_i
\end{aligned}
$$

- 我们认为D_i可能跟ε_i相关, 但Z_i跟ε_i无关
- 因此先估计一阶段, 得到D_i预测值$\hat{D}_i$, 再用$\hat{D}_i$取代二阶段的D_i, 估计其系数
- 如果不存在X_i, 两种方法的结果相同
- 两阶段最小二乘法是基于模型的分析, 模型中隐含了诸多假设, 在现实中未必成立

沃德估计量和两阶段最小二乘法 (2SLS)

- 在观察性研究中, 一般用两阶段最小二乘法来估计LATE:

$$Y_i = \alpha_1 + \gamma D_i + \beta_1 X_i + \varepsilon_i$$
$$D_i = \alpha_2 + \delta Z_i + \beta_2 X_i + \nu_i$$

- 我们认为D_i可能跟ε_i相关, 但Z_i跟ε_i无关
- 因此先估计一阶段, 得到D_i预测值$\hat{D}_i$, 再用$\hat{D}_i$取代二阶段的D_i, 估计其系数
- 如果不存在X_i, 两种方法的结果相同
- 两阶段最小二乘法是基于模型的分析, 模型中隐含了诸多假设, 在现实中未必成立

思考题: 为什么在两阶段最小二乘法中我们不需要排他性假设?

实践中的工具变量

- 工具变量是计量经济学家的发明

实践中的工具变量

- 工具变量是计量经济学家的发明
- 如何根据一组鱼获的价格和交易量估计供给曲线?

实践中的工具变量

- 工具变量是计量经济学家的发明
- 如何根据一组鱼获的价格和交易量估计供给曲线?
- 需要找到一个只影响供给, 不影响需求的变量 (海上的天气)

实践中的工具变量

- 工具变量是计量经济学家的发明
- 如何根据一组鱼获的价格和交易量估计供给曲线?
- 需要找到一个只影响供给, 不影响需求的变量 (海上的天气)
- 工具变量从90年代初开始变得愈发流行

实践中的工具变量

- 工具变量是计量经济学家的发明
- 如何根据一组鱼获的价格和交易量估计供给曲线?
- 需要找到一个只影响供给, 不影响需求的变量 (海上的天气)
- 工具变量从90年代初开始变得愈发流行
- 出生季度和入学年龄, 选举周期和警力部署, 殖民者死亡率和政治制度, 非洲降雨量和人均GDP....

实践中的工具变量

- 工具变量是计量经济学家的发明
- 如何根据一组鱼获的价格和交易量估计供给曲线?
- 需要找到一个只影响供给, 不影响需求的变量 (海上的天气)
- 工具变量从90年代初开始变得愈发流行
- 出生季度和入学年龄, 选举周期和警力部署, 殖民者死亡率和政治制度, 非洲降雨量和人均GDP....
- 工具变量: 实证研究中的黑魔法....

实践中的工具变量

- 但是,

实践中的工具变量

- 但是,不论Wald还是2SLS都有偏, 且会放大已有偏误

实践中的工具变量

- 但是,不论Wald还是2SLS都有偏, 且会放大已有偏误
- 特别是在弱工具变量, 即D_i跟Z_i相关度低的情况下

实践中的工具变量

- 但是,不论Wald还是2SLS都有偏, 且会放大已有偏误
- 特别是在弱工具变量, 即D_i跟Z_i相关度低的情况下
- 直觉上来说, 如果分母非常小, 那么分子上的任何偏误都会被放大

实践中的工具变量

- 但是,不论Wald还是2SLS都有偏, 且会放大已有偏误
- 特别是在弱工具变量, 即D_i跟Z_i相关度低的情况下
- 直觉上来说, 如果分母非常小, 那么分子上的任何偏误都会被放大
- 反例: 出生季度和入学年龄

实践中的工具变量

- 但是,不论Wald还是2SLS都有偏, 且会放大已有偏误
- 特别是在弱工具变量, 即D_i跟Z_i相关度低的情况下
- 直觉上来说, 如果分母非常小, 那么分子上的任何偏误都会被放大
- 反例: 出生季度和入学年龄
- 排他性假设又无法检验, 大多数时候靠论证

实践中的工具变量

- 但是,不论Wald还是2SLS都有偏, 且会放大已有偏误
- 特别是在弱工具变量, 即D_i跟Z_i相关度低的情况下
- 直觉上来说, 如果分母非常小, 那么分子上的任何偏误都会被放大
- 反例: 出生季度和入学年龄
- 排他性假设又无法检验, 大多数时候靠论证
- Heather Sarsons, 降雨量和印度水坝

实践中的工具变量

- 现状: 工具变量很少被作为主要的识别策略, 多出现于稳健性检验中

实践中的工具变量

- 现状: 工具变量很少被作为主要的识别策略, 多出现于稳健性检验中
- 除非是实验中的不顺从, 或者对于分配机制有足够充分的了解

实践中的工具变量

- 现状: 工具变量很少被作为主要的识别策略, 多出现于稳健性检验中
- 除非是实验中的不顺从, 或者对于分配机制有足够充分的了解
- 越战抽签: 可能有人逃兵役, 可能有人志愿参军; 麦加朝圣: 同样的机制

实践中的工具变量

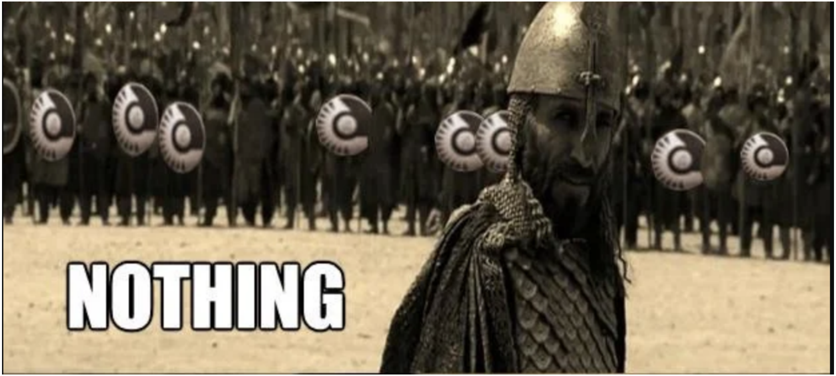
- 现状: 工具变量很少被作为主要的识别策略, 多出现于稳健性检验中
- 除非是实验中的不顺从, 或者对于分配机制有足够充分的了解
- 越战抽签: 可能有人逃兵役, 可能有人志愿参军; 麦加朝圣: 同样的机制
- 较糟的例子: 降雨, 到某地的距离, 邻居自变量的平均值...
- 目前出现了一些工具变量检验, 但通过不代表没有问题

随机实验中的干涉(Interference)

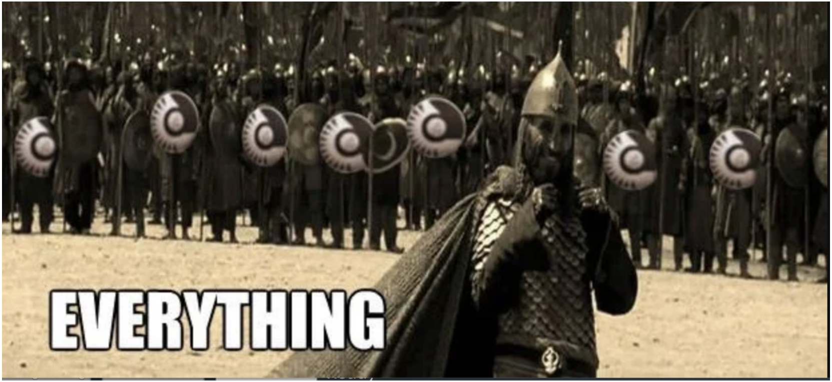
干涉改变了什么?

随机实验中的干涉(Interference)

干涉改变了什么?



随机实验中的干涉(Interference)



随机实验中的干涉(Interference)

此时SUTVA不再成立

换言之, 有可能$Y_i \neq Y_i(D_i)$, Rubin模型的基础不复存在

这既是挑战, 也是机会

随机实验中的干涉(Interference)

此时SUTVA不再成立
换言之, 有可能$Y_i \neq Y_i(D_i)$, Rubin模型的基础不复存在
这既是挑战, 也是机会
基本想法: 利用辅助信息, 根据分配的Z计算实际接受的处理D

随机实验中的干涉(Interference)

此时SUTVA不再成立
换言之, 有可能$Y_i \neq Y_i(D_i)$, Rubin模型的基础不复存在
这既是挑战, 也是机会
基本想法: 利用辅助信息, 根据分配的Z计算实际接受的处理D

- D等于同小组内接受处理者的比例
- D等于社交网络内邻居接受处理的比例
- D随着到处理个体的距离而递减

随机实验中的干涉(Interference)

此时SUTVA不再成立
换言之, 有可能$Y_i \neq Y_i(D_i)$, Rubin模型的基础不复存在
这既是挑战, 也是机会
基本想法: 利用辅助信息, 根据分配的Z计算实际接受的处理D

- D等于同小组内接受处理者的比例
- D等于社交网络内邻居接受处理的比例
- D随着到处理个体的距离而递减

我们可以将总的处理效应分解为直接效应和间接效应, 从而更好地理解处理的扩散作用

两步法设计

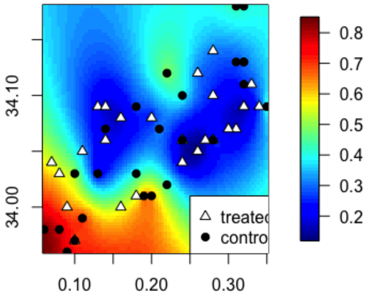
- 最简单的干涉实验设计
- 例子: Duflo and Saez, 2003
- 目的: 想提高某大学内职工的养老保险参保率
- 设计: 把全部系分成两组, 处理组中每个系有一半职工得到了宣传材料; 控制组中没有人得到材料
- 估计: 处理组中未得到材料者的参保率 - 控制组参保率 = 间接效应; 处理组中得到材料者的参保率 - 处理组中未得到材料者的参保率 = 直接效应
- 严格分析: Hudgens and Halloran (2008)

两步法设计

- 可以推广到多个组别, 每组中个体接受处理的概率不同
- 标准误估计可以用随机推断, 也可以用Tchetgen Tchetgen and VanderWeele (2012)中提供的保守估计
- 这里的关键是构造了从分配到实际处理的映射: $Z_i \to (Z_i, \bar{Z}_g)$
- 我们需要假设干涉只发生在各组之内
- 如果该假设不满足, 但我们知道个体之间的社会网络, 那么可以用Aronow and Samii (2017)中的方法, 基于网络构造映射

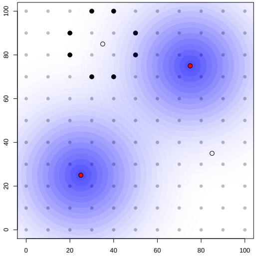
空间扩散效应

- Aronow, Samii, and Wang (2019)
- 在田野实验中，干涉未必局限在某个组别之内
- 在某个区域内的学校发放打虫药，该区域的住户可能同时受到多所学校的影响
- 我们考虑两层设计：处理分配发生在学校层面，但结果出现在家户层面



空间扩散效应

- 估计方法: 画圈圈



- 在各个距离d上计算组间均值之差, 得到的曲线是真实效应的无偏和一致估计
- 推断可以用随机推断或者Conley标准误

随机实验中的协变量

现在假定我们除了Y和D之外, 还拥有协变量X的信息

随机实验中的协变量

现在假定我们除了Y和D之外, 还拥有协变量X的信息
X跟Y, D的关系有两种可能性

随机实验中的协变量

现在假定我们除了Y和D之外, 还拥有协变量X的信息
X跟Y, D的关系有两种可能性

- X只跟Y相关, 不跟D相关; D仍然“外生”
- X跟D相关, 此时只有控制了X之后, D才独立于潜在结果:

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

随机实验中的协变量

现在假定我们除了Y和D之外, 还拥有协变量X的信息
X跟Y, D的关系有两种可能性

- X只跟Y相关, 不跟D相关; D仍然“外生”
- X跟D相关, 此时只有控制了X之后, D才独立于潜在结果:

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

- 我们先讨论第一种情况, 此时X也被称为调节变量 (moderator)

随机实验中的协变量

考虑经典的回归分析, 如果不控制X, D的系数会不会有偏误?

随机实验中的协变量

考虑经典的回归分析, 如果不控制X, D的系数会不会有偏误?
不会, 但我们大多数时候仍然应该控制X, 因为:

- X中包含的信息可以提高估计的效率 (有效性), 降低D系数估计的标准误
- 我们有时候想知道处理效应如何随着X变化(处理效应的异质性, heterogeneous treatment effect)
- 换言之, 我们想估计条件平均处理效应 (CATE):
  $\tau(x) = E[\tau_i | X_i = x]$

随机实验中的协变量

考虑经典的回归分析, 如果不控制X, D的系数会不会有偏误?
不会, 但我们大多数时候仍然应该控制X, 因为:

- X中包含的信息可以提高估计的效率 (有效性), 降低D系数估计的标准误
- 我们有时候想知道处理效应如何随着X变化(处理效应的异质性, heterogeneous treatment effect)
- 换言之, 我们想估计条件平均处理效应 (CATE): $\tau(x) = E[\tau_i|X_i = x]$
- 得到了CATE之后, 很容易就可以得到ATE: $\tau = \int_x E[\tau_i|X_i = x] * Pr(X_i = x)$

随机实验中的协变量

考虑经典的回归分析, 如果不控制X, D的系数会不会有偏误?
不会, 但我们大多数时候仍然应该控制X, 因为:

- X中包含的信息可以提高估计的效率 (有效性), 降低D系数估计的标准误
- 我们有时候想知道处理效应如何随着X变化(处理效应的异质性, heterogeneous treatment effect)
- 换言之, 我们想估计条件平均处理效应 (CATE):
  $\tau(x) = E[\tau_i|X_i = x]$
- 得到了CATE之后, 很容易就可以得到ATE:
  $\tau = \int_x E[\tau_i|X_i = x] * Pr(X_i = x)$
- 为了能够估计CATE, 我们需要在任何由x决定的子群体里都同时有处理组个体和控制组个体
- 这被称为“交叠性假设 (overlapping)”

随机实验中的协变量

- 例子: 在建筑工人的培训实验中, 我们认为工人的性别和学历这两个因素会影响其在劳动力市场上的收入
- 性别和学历不影响工人接受处理的概率
- 但我们想知道培训对于不同性别, 不同学历的工人有怎样的影响

	高中毕业	高中未毕业	总数
女性	30	20	50
男性	10	40	50
总数	40	60	100

随机实验中的协变量

- 例子: 在建筑工人的培训实验中, 我们认为工人的性别和学历这两个因素会影响其在劳动力市场上的收入
- 性别和学历不影响工人接受处理的概率
- 但我们想知道培训对于不同性别, 不同学历的工人有怎样的影响

	高中毕业	高中未毕业	总数
女性	30	20	50
男性	10	40	50
总数	40	60	100

- 我们可以在(女性, 高中毕业), (女性, 高中未毕业), (男性, 高中毕业), (男性, 高中未毕业)这四个子群体中分别计算处理效应
- 很显然这要求每个子群体中都有处理组和控制组个体

利用协变量: 分块和匹配

- 基本想法: 把协变量取值相同的观测放在一组, 分别计算组内的处理效应, 再进行加总
- 一种做法是在实验之前就将组分好, 在组内进行随机分配 (分块, blocking)
- 另一种做法是在实验之后根据协变量分组, 在各组内进行估计 (匹配, matching)
- 在完全随机实验中, 二者没有差别
- 在观察性研究中, 只能使用匹配

回归: 基于设计的视角

问题: 为什么不利用回归估计处理效应, 并在右手端控制协变量?

回归: 基于设计的视角

问题: 为什么不利用回归估计处理效应, 并在右手端控制协变量?
我们来看一下回归模型背后隐含的假设:

$$Y_i = \alpha + \tau D_i + \delta X_i + \varepsilon_i$$

回归: 基于设计的视角

问题: 为什么不利用回归估计处理效应, 并在右手端控制协变量?
我们来看一下回归模型背后隐含的假设:

$$Y_i = \alpha + \tau D_i + \delta X_i + \varepsilon_i$$

当我们写下这个方程的时候, 已经假设了:

- 处理效应恒定 ($\tau_i = \tau$ for any i)
- 协变量对结果的影响是线性的 (没有高阶项和交叉项)

回归: 基于设计的视角

问题: 为什么不利用回归估计处理效应, 并在右手端控制协变量?
我们来看一下回归模型背后隐含的假设:

$$Y_i = \alpha + \tau D_i + \delta X_i + \varepsilon_i$$

当我们写下这个方程的时候, 已经假设了:

- 处理效应恒定 ($\tau_i = \tau$ for any i)
- 协变量对结果的影响是线性的 (没有高阶项和交叉项)

思考题: 为什么在没有协变量的时候, 使用回归分析没有这些问题?

回归: 基于设计的视角

- 如果处理效应并非恒定, 那么回归估计将是有偏的, 其期望不等于平均处理效应 (Aronow and Samii, 2014)
- 事实上此时其期望仍然等于各组CATE的加权平均, 但使用的权重是错误的 (Athey et al., 2017)
- 在面板数据中同样有这样的问题 (Imai and Kim, 2019)
- 这意味着, 很多时候基于回归的观察性研究未必就更有代表性 (Samii, 2017)

回归: 基于设计的视角

- 我们想要做的, 其实是用某个函数$f(\mathbf{X}, \mathbf{Y})$来近似无法观测的潜在结果$Y_i(0)$
- 在组间均值之差里, $f(X_i, Y_i) = \frac{1}{N_0} \sum_{D_i=0} Y_i$
- Lin (2013)建议用以下的线性方程来近似Y_i: $Y_i = \bar{Y}_i + \beta(X_i - \bar{X}_i) + \varepsilon_i$
- 如果我们想预测一堆树叶的平均面积, 怎么做更有效率?
- 可以用树叶的重量提高预测的准确性

林回归

- $Y_i(0) = \bar{Y}_i(0) + \beta_0(X_i - \bar{X}_i) + \varepsilon_{0i}$
  $Y_i(1) = \bar{Y}_i(1) + \beta_1(X_i - \bar{X}_i) + \varepsilon_{1i}$
- 因此,

$$\begin{aligned} Y_i =& D_i Y_i(1) + (1 - D_i) Y_i(0) \\ =& \bar{Y}_i(0) + (\bar{Y}_i(1) - \bar{Y}_i(0)) D_i + \\ & \beta_0(X_i - \bar{X}_i) + (\beta_1 - \beta_0) D_i (X_i - \bar{X}_i) + \varepsilon_i^{'} \\ =& \alpha + \tau D_i + \beta(X_i - \bar{X}_i) + \delta D_i (X_i - \bar{X}_i) + \varepsilon_i^{'} \end{aligned}$$

- 结论: 回归中控制处理变量, 减掉均值的协变量, 以及二者的交叉项
- 这样得到的系数估计是渐进无偏的, 而且能知道系数如何随着协变量取值而变化

林回归背后的假设

- 林回归仍然依赖于线性假设

林回归背后的假设

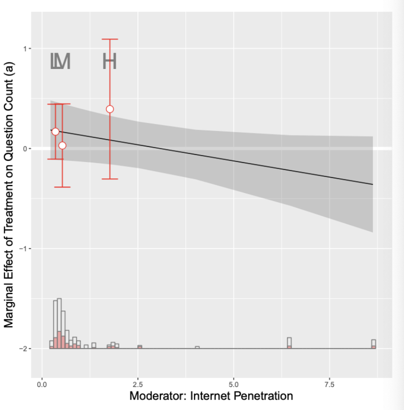
- 林回归仍然依赖于线性假设
- 如果假设不对会发生什么?

林回归背后的假设

- 林回归仍然依赖于线性假设
- 如果假设不对会发生什么?
- Hainmueller, Mummolo, and Xu (2018)
- 有两种常见的错误: 1. 在处理效应不是线性的时候使用线性模型拟合; 2. 在交叠性假设不满足的时候使用线性模型

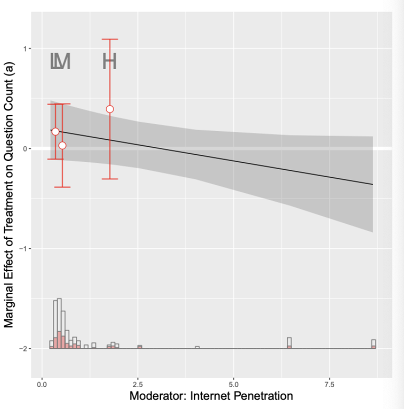
林回归背后的假设

- Malesky et al. (2012): 在互联网渗透率较高的地区，个人网站降低了越南议员的连任概率



林回归背后的假设

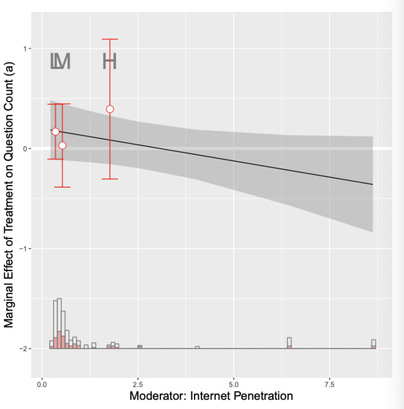
- Malesky et al. (2012): 在互联网渗透率较高的地区, 个人网站降低了越南议员的连任概率



- 但越南的互联网渗透率分布极不平衡, 这个结果几乎完全是由几个大城市驱动的

林回归背后的假设

- Malesky et al. (2012): 在互联网渗透率较高的地区, 个人网站降低了越南议员的连任概率



- 但越南的互联网渗透率分布极不平衡, 这个结果几乎完全是由几个大城市驱动的
- 建议: 1. 分段估计交互项; 2. 使用局部回归非参地估计交互效应

新趋势: 用机器学习估计异质性

- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?

新趋势: 用机器学习估计异质性

- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?
- 可以让机器来决定!

新趋势: 用机器学习估计异质性

- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?
- 可以让机器来决定!
- 我们希望模型可以精确捕捉到Y的变化

新趋势: 用机器学习估计异质性

- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?
- 可以让机器来决定!
- 我们希望模型可以精确捕捉到Y的变化
- 但实际观察到的结果Y里包含两部分, 信号和噪音

新趋势: 用机器学习估计异质性

- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?
- 可以让机器来决定!
- 我们希望模型可以精确捕捉到Y的变化
- 但实际观察到的结果Y里包含两部分, 信号和噪音
- 我们想要拟合前一部分, 否则得到的模型在新的数据集里会表现不佳

新趋势: 用机器学习估计异质性

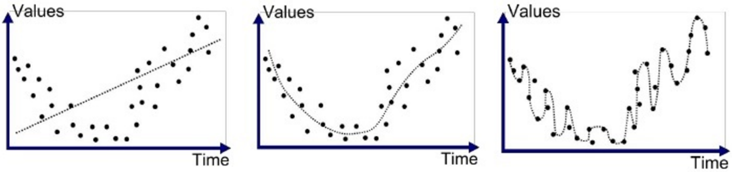
- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?
- 可以让机器来决定!
- 我们希望模型可以精确捕捉到Y的变化
- 但实际观察到的结果Y里包含两部分, 信号和噪音
- 我们想要拟合前一部分, 否则得到的模型在新的数据集里会表现不佳
- 如果一味提高拟合精度, 那估计的偏误会很小, 但对于新数据会非常敏感

新趋势: 用机器学习估计异质性

- 我们怎么知道用哪些协变量可以更好地解释处理效应的变化?
- 可以让机器来决定!
- 我们希望模型可以精确捕捉到Y的变化
- 但实际观察到的结果Y里包含两部分, 信号和噪音
- 我们想要拟合前一部分, 否则得到的模型在新的数据集里会表现不佳
- 如果一味提高拟合精度, 那估计的偏误会很小, 但对于新数据会非常敏感
- 这被称为误差-方差取舍 (bias-variance tradeoff)
- 误差太小是过拟合 (overfitting), 误差太大则是欠拟合 (underfitting)

新趋势: 用机器学习估计异质性

- 我们希望调整模型的参数, 使之有最优的预测表现, 在误差和方差之间取得平衡



Underfitted　　Good Fit/Robust　　Overfitted

新趋势: 用机器学习估计异质性

- 机器学习的基本想法: 用一个超参数 (hyperparameter) 来控制模型的简洁程度

新趋势: 用机器学习估计异质性

- 机器学习的基本想法: 用一个超参数 (hyperparameter) 来控制模型的简洁程度
- 例子: 回归模型里的变量个数

新趋势: 用机器学习估计异质性

- 机器学习的基本想法: 用一个超参数 (hyperparameter) 来控制模型的简洁程度
- 例子: 回归模型里的变量个数
- 为了估计超参数, 我们将数据随机分为两份: 训练集 (training set) 和测试集 (test set)

新趋势: 用机器学习估计异质性

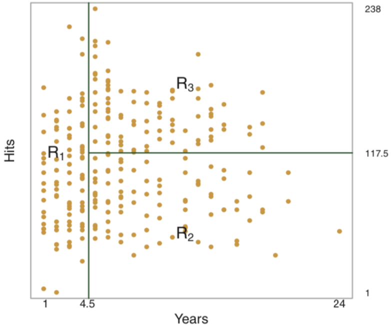
- 机器学习的基本想法: 用一个超参数 (hyperparameter) 来控制模型的简洁程度
- 例子: 回归模型里的变量个数
- 为了估计超参数, 我们将数据随机分为两份: 训练集 (training set) 和测试集 (test set)
- 给定一个超参数的值, 我们在训练集上估计模型, 然后检查其在测试集上的表现

新趋势: 用机器学习估计异质性

- 机器学习的基本想法: 用一个超参数 (hyperparameter) 来控制模型的简洁程度
- 例子: 回归模型里的变量个数
- 为了估计超参数, 我们将数据随机分为两份: 训练集 (training set) 和测试集 (test set)
- 给定一个超参数的值, 我们在训练集上估计模型, 然后检查其在测试集上的表现
- 我们不断改变超参数的值, 直到训练出来的模型在测试集上表现达到最佳

从因果树到因果森林

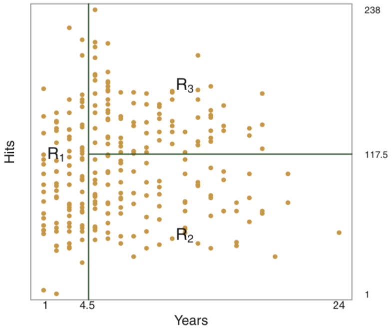
- 树模型在估计异质性方面有天然的优势



- 这里的超参数是“叶片”的数目
- 我们可以尝试不同的划分，使得协变量对处理效应的解释力尽可能强

从因果树到因果森林

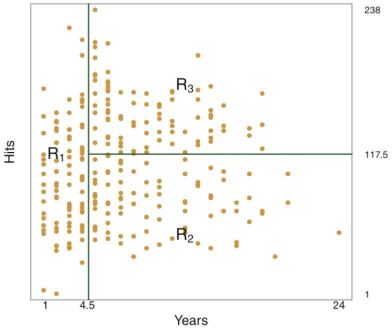
- 树模型在估计异质性方面有天然的优势



- 这里的超参数是“叶片”的数目
- 我们可以尝试不同的划分, 使得协变量对处理效应的解释力尽可能强
- Athey and Imbens (2018): 在训练集和测试集之外, 还应当专门有一部分数据用于效应的估计

从因果树到因果森林

- 树模型在估计异质性方面有天然的优势



- 这里的超参数是“叶片”的数目
- 我们可以尝试不同的划分, 使得协变量对处理效应的解释力尽可能强
- Athey and Imbens (2018): 在训练集和测试集之外, 还应当专门有一部分数据用于效应的估计
- 如果每次用一部分数据生成一棵树, 就得到了随机森林

谢谢!