因果推断在观察性研究中的应用 I：设计丨因果科学读书会第3期

【读书会相关信息】

“因果”并不是一个新概念，而是一个已经在多个学科中使用了数十年的分析技术。通过前两季的分享，我们主要梳理了因果科学在计算机领域的前沿进展。如要融会贯通，我们需要回顾数十年来在社会学、经济学、医学、生物学等多个领域中，都是使用了什么样的因果模型、以什么样的范式、解决了什么样的问题。我们还要尝试进行对比和创新，看能否以现在的眼光，用其他的模型，为这些研究提供新的解决思路。

如果大家对这个读书会感兴趣，欢迎报名：

https://pattern.swarma.org/mobile/study_group/10?from=wechat

【时间】2021年11月14日 21:00--23:00

【主讲人】

李昊轩，北京大学大数据科学研究中心博士研究生，导师为周晓华教授，专业为数据科学（统计学），研究兴趣为因果推断，推荐系统，强化学习。

【笔记小分队】

赵欣

刘曼霞

【讲座笔记】

记录人：赵欣

第一部分 实验性研究：线性回归

线性回归

- 这里我们关注因果模型的参数何时具有解释，而非线性回归本身。
  - 关注的自变量：接受治疗的指标变量 W_i 及一些治疗前变量 X_i
  - 视角：超总体
  - 回归方程参数估算方法：最小二乘法
  - 在条件均值模型中关心的系数为平均治疗效果（ATE）

- 这里不关注线性模型的有效性。不论真实模型是否被正确指定（即模型是否为线性），我们的estimate总是consistent

1. 不纳入协变量的情况

- 模型为：

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \epsilon_i$$

2. 纳入协变量的情况

- 考虑协变量的观测结果的回归模型为：

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \epsilon_i$$

当 X_i 是个体 i 的协变量向量

- 我们知道工业界应用均需要模型假设正确，但无论线性模型的线性假设是否正确（eg. 比如真实模型不是线性的，甚至是neural network），回归方法都具有无偏性和相合性线性。
  - 但当协变量能够解释潜在结果时，线性模型的最小二乘估计更加稳健。

3. 含交互作用情形

- 考虑协变量与治疗指标的交互作用，模型为：

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + W_i \cdot (X_i - \bar{X})\gamma + \epsilon_i$$

- 进一步提高了模型精度，同时提升了错误指定模型的鲁棒性。
  - 具体来说，模型在治疗和最对照组中的解释强度不同，也就是协变量系数不同。这个时候交互作用弥补了模型的bias。

检验因果效应

- 需要对线性模型进行假设检验来判断治疗效果的显著性。
- 在超总体观点下以上正确指定的模型有：

$$\mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} | X_i = x, W_i = w] = \alpha + \tau \cdot w + x\beta + w \cdot (x - \mu_x)\gamma'$$

1. 平均效果为零

- 在所有协变量相同的条件下，平均治疗效果为零的假设检验为：

$$H_0 : \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0) | X_i = x] = 0, \forall x$$

- 以及备择假设：

$$H_\alpha : \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0) | X_i = x] \neq 0, \text{对于某些} x$$

2. 平均效果为常数

- 对治疗效果为常数的假设检验为：

  H_0' : $\mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x] = 0, \forall x$

- 以及备择假设：

  H_α' : $\exists x_0, x_1$，使得
  $\mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X - i = x_0] \neq \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x_1]$

3. 渐进统计量

- 在超总体中进行有限抽样，并在该有限样本上进行完全随机检验。
  - 若存在常数 τ ，使得对于所有个体均有 $Y_i(1) - Y_i(0) = \tau$ ，则有 $\gamma^* = 0$ 且 Q_{const} 渐进于卡方分布 $\chi(\dim(X_i))$
- 若对于所有个体均有 $Y_i(1) - Y_i(0) = 0$ ，则有 Q_{zero} 渐进于 $\chi(\dim(X_i) + 1)$

例子：LRC-CPPT 胆固醇数据

- 案例：消胆胺对胆固醇水平的影响
  - 实验对象：$N_t = 165$ 接受消胆胺药物， $N_c = 172$ 接受安慰剂
  - 潜在结果：主要结果为 `cholf` 为随机分组后胆固醇读书的平均值，次要结果为 `comp` （依从性度量）
- 从描述性统计（ppt title有误）可以看到treatment的依从性在治疗比在控制组低。

Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	chol1	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	chol2	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	cholp	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	cholf	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	chold	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	comp	74.5	(21.0)	59.9	(24.4)	0	101.0

- 模型纳入协变量之后，在胆固醇含量上的鲁棒性提高了。

Table 7.2. ***Regression Estimates for Average Treatment Effects for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Effect of Assignment to Treatment on Post-Cholesterol Level		Effect of Assignment to Treatment on Compliance	
	Est	$\widehat{(\text{s. e.})}$	Est	$\widehat{(\text{s. e.})}$
No covariates	−26.22	(3.93)	−14.64	(3.51)
`cholp`	−25.01	(2.60)	−14.68	(3.51)
`chol1, chol2`	−25.02	(2.59)	−14.95	(3.50)
`chol1, chol2`, interacted with W	−25.04	(2.56)	−14.94	(3.49)

- 得出回归系数。

Table 7.3. ***Regression Estimates for Average Treatment Effects on Post-Cholesterol Levels for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Model for Levels		Model for Logs	
	Est	$\widehat{(\text{s. e.})}$	Est	$\widehat{(\text{s. e.})}$
Assignment	−25.04	(2.56)	−0.098	(0.010)
Intercept	−3.28	(12.05)	−0.133	(0.233)
`chol1`	0.98	(0.04)	−0.133	(0.233)
`chol2-chol1`	0.61	(0.08)	0.602	(0.073)
`chol1` × Assignment	−0.22	(0.09)	−0.154	(0.107)
(`chol2-chol1`) × Assignment	0.07	(0.14)	0.184	(0.159)
R-squared	0.63		0.57	

- 检验：服从 χ^2 的检验。在效果为常数的情况下，由于我们不知道causal effect τ ，此时只能做费希尔的精确P值。

Table 7.4. ***P-Values for Tests for Constant and Zero Treatment Effects, Using*** `chol1` ***and*** `chol2-chol1` ***as Covariates for the PRC-CPPT Cholesterol Data from Table 7.1***

		Post-Cholesterol Level	Compliance
Zero treatment effect	$\mathcal{X}^2(3)$ approximation	<0.001	<0.001
	Fisher exact p-value	<0.001	0.001
Constant treatment effect	$\mathcal{X}^2(2)$ approximation	0.029	0.270

第二部分 实验性研究：基于模型的推断方法

基于模型的推断方法

- 我们在有限样本中，把潜在结果视为随机变量并建立模型。这类随机模型通常依赖于一些未知参数，通过贝叶斯方法，使用假设模型来估算给定观察数据的确实潜在结果，并使用这些结果来对感兴趣的统计量进行推断。
  - 本质：对给定观察数据的缺失潜在结果进行估计
  - 对比：费希尔的精确P值方法、奈曼的重复抽样方法或回归方法相比，基于模型的方法非常灵活。

> 💡 贝叶斯方法是先验分布和观测数据的trade-off，以让预测结果更加稳定。

- 给定对因果效应估计的统计量 $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W})$ ，通过对缺失的潜在因果的估计，实验者可以推断感兴趣的统计量分布。
  - 对在没有纳入协变量的完全随机试验中基于贝叶斯模型的推理方法，主要目的是为缺失的潜在结果建立一个模型 $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W})$ ，由此导出感兴趣的估计值 $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$ 的分布，也可以用观察到的和缺失的潜在结果来表示该估计值 $\tau = \tau(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$ 。
- 需要两个输入：
  - 第一个输入是潜在结果的联合分布
  - 第二个输入是参数 θ 的先验分布 $p(\theta)$

例子：Lalonde NSW职业培训数据

- 案例：估计职业培训计划对收入的影响
  - 实验对象：在劳动力市场上处于实质性劣势的男性，大多数人的劳动力市场历史非常糟糕，很少有长期就业的情况。
  - 潜在结果：项目后劳动力市场指标，1978年的收入(`earn'78`)
- 描述性估计 （可下载），收入单位为1k

Table 8.1. *Summary Statistics: National Supported Work (NSW) Program Data*

Covariate	Mean	(S.D.)	Average Controls ($N_c = 260$)	Average Treated ($N_t = 185$)
age	25.37	(7.10)	25.05	25.82
education	10.20	(1.79)	10.09	10.35
married	0.17	(0.37)	0.15	0.19
nodegree	0.78	(0.41)	0.83	0.71
black	0.83	(0.37)	0.83	0.84
earn'74	2.10	(5.36)	2.11	2.10
earn'74=0	0.73	(0.44)	0.75	0.71
earn'75	1.38	(3.15)	1.27	1.53
earn'75=0	0.65	(0.48)	0.68	0.60
earn'78	5.30	(6.63)	4.56	6.35
earn'78=0	0.31	(0.46)	0.35	0.24

- 对照组和治疗组收入：更多的对照组收入为零。

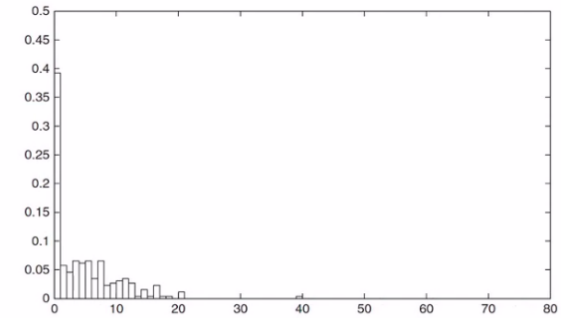


Figure 8.1. Histogram of earnings for control group – NSW job-training data

图 6: 对照组个体 1978 年收入直方图

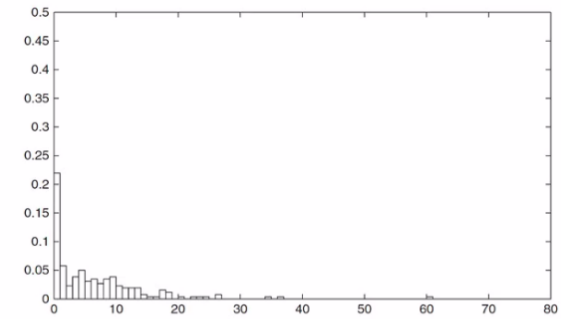


Figure 8.2. Histogram of earnings for trainee group – NSW job-training data

图 7: 治疗组个体 1978 年收入直方图

- 我们拿前六个样本进行分析， 需要对缺失的数据的分布进行估计

Table 8.2. *First Six Observations from NSW Program Data*

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
1	0	?	0	0
2	?	9.9	1	9.9
3	12.4	?	0	12.4
4	?	3.6	1	3.6
5	0	?	0	0
6	?	24.9	1	24.9

- 那么 causal estimand $\hat{\tau}$ 为

$$\hat{\tau} = \tilde{\tau}(\mathbf{Y}^{\text{obs}}, \hat{\mathbf{Y}}^{\text{mis}}, \mathbf{W}) = \frac{1}{6} \cdot \sum_{i=1}^{N} ((2 \cdot W_i - 1) \cdot (Y_i^{\text{obs}} - \hat{Y}_i^{\text{mis}}))$$

- 估计方法：
  - 均值插补：不理想

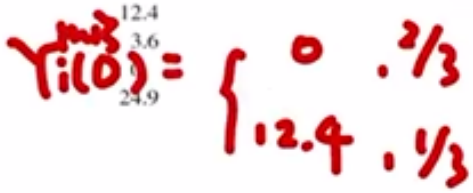
Table 8.3. *The Average Treatment Effect Using Imputation of Average Observed Outcome Values within Treatment and Control Groups for the NSW Program Data*

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
1	0	(12.8)	0	0
2	(4.13)	9.9	1	9.9
3	12.4	(12.8)	0	12.4
4	(4.13)	3.6	1	3.6
5	0	(12.8)	0	0
6	(4.13)	24.9	1	24.9
Average	4.13	12.8		
Diff (ATE):	8.67			

- 抽样法：对不同的treatment进行权重。这时候引入了随机性，得到ATE。

Table 8.4. *The Average Treatment Effect Using Imputed Draws from the Empirical Distributions within Treatment and Control Groups for the First Six Units from the NSW Program Data*

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
Panel A: First draw				
1	0	(3.6)	0	0
2	(12.4)	9.9	1	9.9
3	12.4	(9.9)	0	12.4
4	(12.4)	3.6	1	3.6
5	0	(9.9)	0	0
6	(0)	24.9	1	24.9
Average	6.2	10.3		
Diff (ATE):	4.1			
Panel B: Second draw				
1	0	(9.9)	0	0
2	(0)	9.9	1	9.9
3	12.4	(24.9)	0	12.4
4	(0)	3.6	1	3.6
5	0	(3.6)	0	0
6	(0)	24.9	1	24.9
Average	2.1	12.8		
Diff (ATE):	10.7			



具体步骤

- Step 1: 推导 $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W},\theta)$
- Step 2: 推导参数 θ 的后验分布 $p(\theta|\mathbf{Y}^{\text{obs}},\mathbf{W})$
- Step 3: 推导缺失潜在结果的后验分布 $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W})$
- Step 4: 推导估计量的后验分布 $f(\gamma|\mathbf{Y}^{\text{obs}},\mathbf{W})$

第三部分 分层随机试验 (Stratified Randomised Experiments)

- 为完全随机试验的直接推广。
- 设层数 $J, N(j), N_c(j)$和$N_t(j)$ 分别是层 j 内的总样本数，对照样本数和治疗样本数。对 $j = 1, ..., J$ ，设 $G_i \in \{1, ..., J\}$ 是个体 i 所在的层，记 $B_i(j) = \mathbf{1}_{G_i=j}$ 是示性函数，若个体 i 在层 j 则取值为1，否则为0。
- 分层/分配机制：需在试验之前指定好。

Fisher 精确P值方法

- 需要构造统计量，检验 Sharp's null hypothesis假设是否成立。
- 思想：先估计每个层内 j 的因果效应，然后进行加权得到总体的因果效应估计：

  $$T^{\text{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot (\bar{Y}_t^{\text{obs}}(j) - \bar{Y}_c^{\text{obs}}(j)) \right|$$

  权重为 $\lambda(j)$ 。
- 加权方法：
  - 使用每个层内的人数进行加权 (不建议）。
  - 使用每个层内的治疗效果的方差，用方差的倒数进行加权 （更鲁棒）
    - 具体来说，方差更大的样本有更低的权重。
  - 使用个体 i 在其所在层内的秩排序 R_i^{strat} 进行加权。

Neyman重复抽样法

- 由于不同层内个体的治疗分配独立，我们可以使用奈曼重复抽样方法。
- 总体的平均治疗效果为（用人数加权）：

  $$\tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m) = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

- 得到无偏估计量：

  $$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(m)$$

线性回归

- 在超总体的视角，考虑分层随机试验的回归方法。
  - 假设：层数为 j ，每个层内均有无限数量的个体。
- 层指标为：每层设置一个常数
- 模型为：

$$Y_i^{\text{obs}} = \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \epsilon_i$$

- 分层随机试验中，最小二乘估计 $\hat{\tau}^{\text{ols}}$ 不一定是超总体平均治疗效果 τ_{sp} 的相合估计，而是层内平均效应 $\tau_{\text{sp}}(j)$ 的加权平均值的相合估计。权重依赖于1）层内人数百分比及2）层内倾向性得分。
  - 当层内的倾向性得分倾向于0或者1时，层的权重就很小。
- 这时候我们需要另外一种建模方法去保证 $\hat{\tau}^{\text{ols}}$ 和 τ_{sp} 是相合估计。
  - 但并不是协变量越多，样本方差越小。

基于模型的推断方法

- 我们有两种情形基于模型的推断方法：
  - 当层数少且每层有很多个体：可以将参数先验地设置为独立，最终得到参数向量

    $\theta = (\mu_c(j), \mu_t(j), \sigma_c^2(j), \sigma_t^2(j), w = 0, 1, j = 1, ..., J)$

    此时 $\mu_c(j)$ 和 $\mu_t(j)$ 的先验服从正态分布，而 $\sigma_c^2(j)$ 和 $\sigma_t^2(j)$ 的先验服从逆卡方分布。
  - 当层数多且每层个体数较少：需要施加限制条件，最终得到参数向量

    $\theta = (\sigma_c^2, \sigma_c^2, \gamma_c, \gamma_t, \eta_c^2, \eta_t^2)$

第四部分 配对随机实验 (Pairwise Randomised Experiments)

- 配对随机试验是分层随机试验的一个特殊情况。
- 分配机制：其中每层正好包含两个个体我们随机分配一个（A，对照0）给治疗组，一个（B，对照1）给对照组。
- 问题：由于每一个阶层中治疗组和对照组均只有一个个体的事实，而奈曼对平均因果效应的方差估计量要求存在至少两个单位分配给内阁层内的治疗组与对照组，也就是说我们无法使用奈曼抽样方差的估计量。
- 此外，由于每个层内具有相同比例的处理单元，我们可以对称地分析层内估计，即平均治疗效果的自然估计量将对每一层以等权重进行加权。
- 我们有：

$$Y_{j,A}^{\text{obs}} = \begin{cases} Y_{j,A}(0) & \text{if } W_{j,A} = 0 \\ Y_{j,A}(1) & \text{if } W_{j,A} = 1 \end{cases} \quad Y_{j,B}^{\text{obs}} = \begin{cases} Y_{j,B}(0) & \text{if } W_{j,A} = 0 \\ Y_{j,B}(1) & \text{if } W_{j,A} = 1 \end{cases}$$

- 总体的平均治疗效果为：

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = \frac{2}{N}\sum_{j=1}^{N/2}\tau_{\text{pair}}(j)$$

Fisher精确P值方法

- 与分层随机试验中相同。通过费希尔假的尖锐零假设，可以评估因果效应的显著性，其中零假设依然为个体无治疗效果。

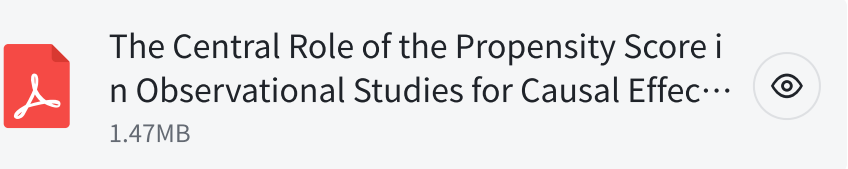
Neyman重复抽样法

- 问题：由于每层只有一个个体接受治疗和对照，无法直接估计方差，因此我们无法直接应用奈曼的重复抽样方法。
- 解决方法：假设层内及层间的治疗效果均为相同的常数且具有可加性
  - 可推出估计量的方差是真实方差无偏的估计量（统计保守）。

第五部分 因果推断的潜在结果框架在观察性研究的作用

因果推断在观察性研究中的非混淆性

- 非混淆性假设(unconfoundedness)：可以理解是协变量包含了充分的信息，即在两个个体具有同样的充分信息（协变量）的条件下，可以将总体的分配机制是为分层随机试验。
  - 本质：通过协变量匹配来推断个体的缺失潜在结果。
- 但由于协变量1）通常为高维且2）通常在样本中呈现出许多不同的值，因此可能会有相当多的个体无法精确匹配协变量。
- 在超总体的视角中，由于高维的协变量很难实现精确的匹配，在实践中我们引入平衡分数从而找到低维的协变量函数。
  - 平衡分数的相等足以消除与治疗前变量差异相关的偏差，因此接受积极治疗的概率不依赖于协变量。
  - 此时我们关心的不是协变量 X_i ，而是均衡得分（即协变量的函数)。
    - 注意均衡得分不唯一
    - 我们最感兴趣的是低维均衡得分，包括倾向性得分(propensity score)
      - 倾向性得分，见pdf：Rosenbaum and Rubin (1983)



- 在给定个体的均衡得分 $b(X_i)$ 的条件下，个体的治疗指标 W_i 与潜在结果独立。若分配机制具有非混淆性，则在给定倾向性得分时，分配机制满足：

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1) \mid b(X_i)$$

- 倾向性得分是最粗的均衡得分，即倾向性的分时每个均衡得分的函数。

倾向性得分估计

- 协变量的选择：逐步回归思想
- Step 1: 基础协变量的选择
  - 基础协变量的选择： 1）对分配机制有解释作用的协变量，2）对潜在结果高度相关的协变量。
- Step 2: 增加协变量的线性组合项
  - 剩余的协变量被分别纳入已有的逻辑回归模型并计算似然比统计量（likelihood ratio statistic）。
  - 若所有协变量被分别纳入原回归模型中后：
    - 似然比统计量均低于一个预先设置的临界值 C_L ，则不纳入剩余的协变量。
    - 似然比统计量超过了 C_L ，则将似然比统计量最大的一个协变量纳入模型中。
- Step 3: 引入二次项和交互项
  - 仍旧用似然比统计量来评估新纳入的项在回归模型中系数为零的假设。

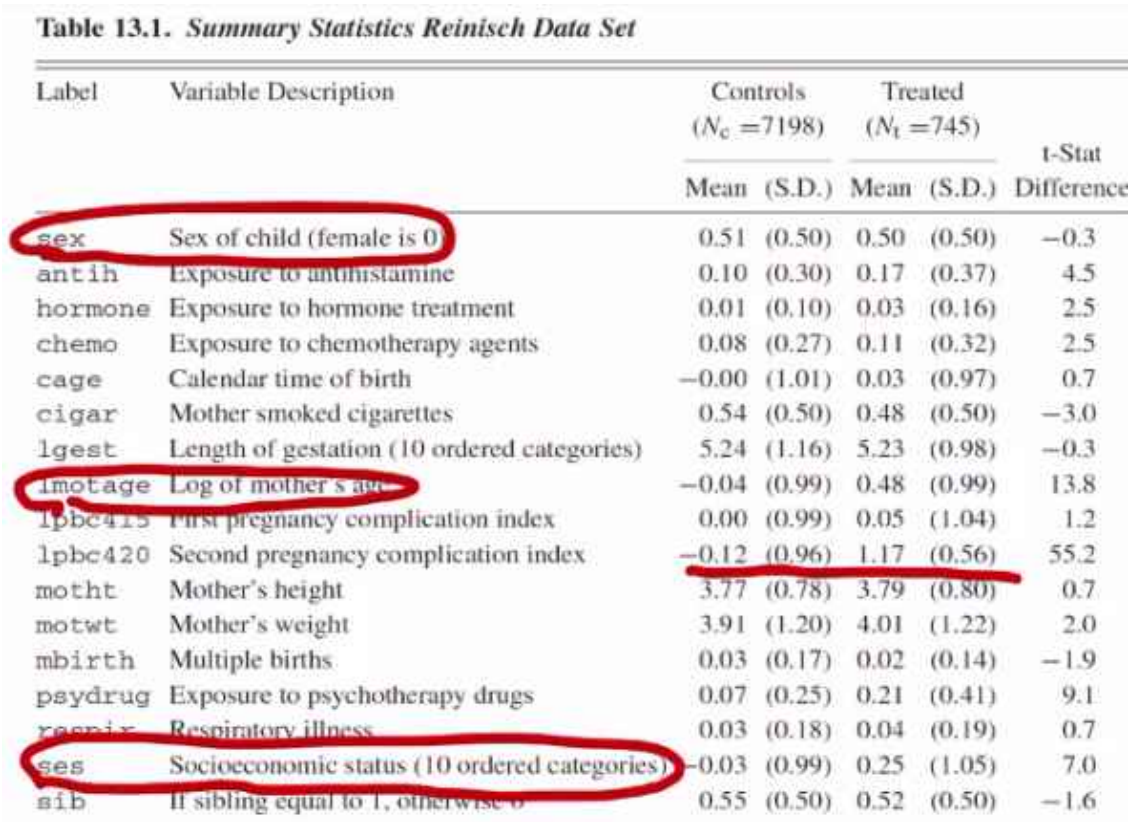
例子：The Reinisch et al. 巴比妥暴露数据

- 案例：产前暴露数据，感兴趣暴露对于多年后认知发展的影响。
- 描述性总结：共17个协变量。

Table 13.1. ***Summary Statistics Reinisch Data Set***

Label	Variable Description	Controls (N_c =7198)		Treated (N_t =745)		t-Stat
		Mean	(S.D.)	Mean	(S.D.)	Difference
sex	Sex of child (female is 0)	0.51	(0.50)	0.50	(0.50)	−0.3
antih	Exposure to antihistamine	0.10	(0.30)	0.17	(0.37)	4.5
hormone	Exposure to hormone treatment	0.01	(0.10)	0.03	(0.16)	2.5
chemo	Exposure to chemotherapy agents	0.08	(0.27)	0.11	(0.32)	2.5
cage	Calendar time of birth	−0.00	(1.01)	0.03	(0.97)	0.7
cigar	Mother smoked cigarettes	0.54	(0.50)	0.48	(0.50)	−3.0
lgest	Length of gestation (10 ordered categories)	5.24	(1.16)	5.23	(0.98)	−0.3
lmotage	Log of mother's age	−0.04	(0.99)	0.48	(0.99)	13.8
lpbc415	First pregnancy complication index	0.00	(0.99)	0.05	(1.04)	1.2
lpbc420	Second pregnancy complication index	−0.12	(0.96)	1.17	(0.56)	55.2
motht	Mother's height	3.77	(0.78)	3.79	(0.80)	0.7
motwt	Mother's weight	3.91	(1.20)	4.01	(1.22)	2.0
mbirth	Multiple births	0.03	(0.17)	0.02	(0.14)	−1.9
psydrug	Exposure to psychotherapy drugs	0.07	(0.25)	0.21	(0.41)	9.1
respir	Respiratory illness	0.03	(0.18)	0.04	(0.19)	0.7
ses	Socioeconomic status (10 ordered categories)	−0.03	(0.99)	0.25	(1.05)	7.0
sib	If sibling equal to 1, otherwise 0	0.55	(0.50)	0.52	(0.50)	−1.6

- 注意 `lpbc420` 第二个壬辰并发症指数在治疗组和对照组中的差异非常大，需要对此做出调整（下一期会讲到具体方法）。



- 作者认为协变量 `sex`，`lmotage` 和 `ses` 会影响认知。因此根据先验信息，有三个基本斜变量。

Table 13.3. ***Estimated Parameters of Propensity Score: Baseline Case with*** `lpbc420` ***Added; Barbituate Data***

Variable	EST	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	−3.71	(0.10)	−36.3
`sex`	0.07	(0.09)	0.8
`lmotage`	0.22	(0.05)	4.7
`ses`	0.15	(0.05)	3.3
`lpbc420`	2.11	(0.08)	27.2
LR statistic	1308.0		

- 接下来我们用三个协变量和剩下的14个协变量分别做14个模型，并计算似然比统计量 (临界值为 $C_L = 1$) 。

Table 13.3. ***Estimated Parameters of Propensity Score: Baseline Case with*** `lpbc420` ***Added; Barbituate Data***

Variable	EST	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	−3.71	(0.10)	−36.3
`sex`	0.07	(0.09)	0.8
`lmotage`	0.22	(0.05)	4.7
`ses`	0.15	(0.05)	3.3
`lpbc420`	2.11	(0.08)	27.2
LR statistic	1308.0		

- 似然比最大的统计量为 `lpbc420` 并超过临界值 C_L ，接下来是 `mbirth` ，以此向右类推。

Table 13.4. ***Likelihood Ratio Statistics for Sequential Selection of Covariates to Enter Linearly; Barbituate Data***

Covariate	Step →										
`sex`	–	–	–	–	–	–	–	–	–	–	–
`antih`	17.5	0.5	1.6	1.3	2.1	1.8	1.6	1.6	1.7	1.3	–
`hormone`	3.9	0.3	0.7	0.7	0.4	0.8	0.7	0.7	0.7	0.8	0.9
`chemo`	10.0	36.6	41.9	–	–	–	–	–	–	–	–
`cage`	0.8	5.8	6.4	7.2	7.6	7.9	–	–	–	–	–
`cigar`	4.3	2.3	3.5	3.7	3.0	2.1	2.1	1.7	2.1	–	–
`lgest`	0.4	11.1	5.0	6.4	7.3	5.5	5.6	–	–	–	–
`lmotage`	–	–	–	–	–	–	–	–	–	–	–
`lpbc415`	0.6	0.0	0.2	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
`lpbc420`	1308.0	–	–	–	–	–	–	–	–	–	–
`motht`	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
`motwt`	6.1	1.5	0.6	1.2	2.5	2.7	2.4	3.4	–	–	–
`mbirth`	4.6	66.1	–	–	–	–	–	–	–	–	–
`psydrug`	93.1	29.8	38.9	46.8	–	–	–	–	–	–	–
`respir`	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
`ses`	–	–	–	–	–	–	–	–	–	–	–
`sib`	21.0	13.8	12.5	15.0	15.7	–	–	–	–	–	–

- 按照纳入顺序，所有线性一次及二次项的倾向性得分的估计：

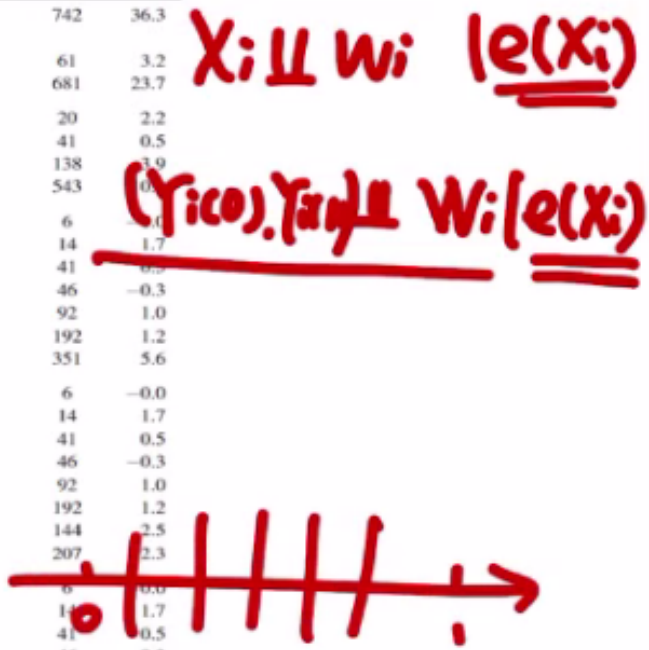
Table 13.6. ***Estimated Parameters of Propensity Score: Final Specification; Barbituate Data***

Variable	EST	(s. e.)	t-Stat
Intercept	−5.67	(0.23)	−24.4
Linear terms			
sex	0.12	(0.09)	1.3
lmotage	0.52	(0.11)	4.7
ses	0.06	(0.09)	0.6
lpbc420	2.37	(0.36)	6.6
mbirth	−2.11	(0.36)	−5.9
chemo	−3.51	(0.67)	−5.2
psydrug	−3.37	(0.55)	−6.1
sib	−0.24	(0.22)	−1.1
cage	−0.56	(0.26)	−2.2
lgest	0.57	(0.23)	2.5
motwt	0.49	(0.17)	2.9
cigar	−0.15	(0.10)	−1.5
antih	0.17	(0.13)	1.3
Second-order terms			
lpbc420 × sib	0.60	(0.19)	3.1
motwt× motwt	−0.10	(0.02)	−4.5
lpbc420 × psydrug	1.88	(0.39)	4.8
sesx sib	−0.22	(0.10)	−2.2
cagex antih	−0.39	(0.14)	−2.8
lpbc420× chemo	1.97	(0.49)	4.0
lpbc420 × lpbc420	−0.46	(0.14)	−3.3
cage × lgest	0.15	(0.05)	3.0
lmotage × lpbc420	−0.24	(0.10)	−2.5
mbirthx cage	−0.88	(0.39)	−2.3
lgest × lgest	−0.04	(0.02)	−2.0
sesx cigar	0.20	(0.09)	2.2
lpbc420× motwt	0.15	(0.07)	2.0
chemo × psydrug	−0.93	(0.46)	−2.0
lmotagex ses	0.10	(0.05)	1.9
cage × cage	−0.10	(0.05)	−1.8
mbirth × chemo	−∞	(0.00)	−∞

- 对倾向性得分从0到1进行分块：
  - 一旦t statistic超过临界值就进行分块。
  - 然后对每个块内协变量的均衡性进行评估。

Table 13.7. *Determination of the Number of Blocks and Their Boundaries; Barbituate Data*

Step	Block	Lower Bound	Upper Bound	Width	# Controls	# Treated	t-Stat
1	1	0.00	0.94	0.94	4462	742	36.3
2	1	0.00	0.06	0.06	2540	61	3.2
	2	0.06	0.94	0.88	1922	681	23.7
3	1	0.00	0.02	0.01	1280	20	2.2
	2	0.02	0.06	0.05	1260	41	0.5
	3	0.06	0.20	0.14	1163	138	3.9
	4	0.20	0.94	0.74	759	543	0
4	1	0.00	0.01	0.00	644	6	0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.94	0.57	301	351	5.6
5	1	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.50	0.13	181	144	2.5
	8	0.50	0.94	0.44	120	207	2.3
6	1	0.00	0.01	0.00	644	6	0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.42	0.05	101	61	0.3
	8	0.42	0.50	0.08	80	83	0.7
	9	0.50	0.61	0.11	73	90	0.8
	10	0.61	0.94	0.34	47	117	−0.3



总结

- 实验性研究：
  - 线性回归：估计量的相合性不依赖于模型的正确假设。
  - 基于模型的推断方法：为所有潜在结果建立一个随机模型，通过贝叶斯方法，使用假设模型来估算给定观察数据的缺失潜在结果。
- 分层随机试验：线性回归模型如何设定，使得估计量具有相合性？
- 配对随机试验：如何解决层内治疗组与对照组样本方差不可估？
- 因果推断在观察性研究中的非混淆性：均衡的分与倾向性得分。
- 倾向性得分估计：基于逐步回归的思想。

【精彩问答】

记录人：刘曼霞

1. treatment对效应是有影响，是不是就是所谓的条件效应？

A：是，可以这样去理解。

2. 错误指定是比如没有放入某些X，算错误指定？

答：错误指定是指真实的模型或者是真实的Y与X、W的关系是不是linear，而不是说哪些X拉进来、哪些不拉进来，当然这两个都是untestable，你永远不知道这个模型是否被正确指定的，哪怕是在机器学习里面建立一个神经网络，你不知道这个模型是否被正确指定。这里我们实质上想说的是，是否需要一个critic specify的一个假设，因为你会发现在W这些里面，要求模型是正确指定。如果你的模型不是正确指定，你会发现，那根本不是一个unbiased estimate。但是在这里结论里面，你会发现你的模型可以不是线性的，可以是二次，甚至可以是一个neural network，但是我们用线性模型去建模，那么我们关心的是一个统计量的相合性，相合性就是说你的样本量趋于无穷时，我这个量是否能够通过某些收敛，收敛到一个真实的参数，这是我们关心的。所以像刚才的问题，如果我们只把一部分的问题放进来，其实会不会相合，也是会相合的，只是说，我们建议可能用有解释性的、先验全部X，包含协变量筛选的过程，也是现在比较常见的practical方法。

3. 那是否可以说随机试验可以随便放X?

答：不，我还是要说一下，就是说首先你随便放X的结论是成立的，因为刚才我们讲过这些结论，但是不是说越多越好，包括我们刚才讲的两个模型，一个是不纳入协变量去考虑，一个是纳入协变量去考虑，这个也不是说纳入协变量一定比不纳入协变量效果好，因为你可以去对比这两个variance。只有你这个X对Y有解释性的，也就是说你的conditional variance，大小关系不一定，只有当X对Y的解释性很强的时候，纳入X之后，你的parameter condtional variance会显著低于前面的量，这个就是说拉进来效果会更好。这里我们怎么去评价效果？看的还是variance，因为你永远是unbiased，在随机实验里面，所以我们会关注方差。

4. 在43、44页，

$$\mathcal{L}\left(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}\right)$$

是不是写反了?

答：不是，这个是log-likelihood的定义。

5. 在前面有六个样本，在采用这个方法的时候，推测mis值的时候，要不要考虑样本的问题？第二个问题是前面推导出来的miss值，是根据已知的theta和观测值来推导 ，最后一步利用推测出来的miss值和系数和theta之间是不是有一个检验的过程?

答：第一个问题，为什么会有样本的考虑？因为事实上前面这个方法，我们没有用到后面这个复杂的东西，这个引入是非常简单的，也是原书的一个引入。 这里面你会看到，其实根本没有参数，是直接给出来的。这个我认为是对所有的样本都是有这样一个分布， 所以对样本应该不会有太大的影响。关于检验过程，你想要去确认的是什么？是前端结果与theta是否真有联系还是什么?

6. 因为前面这个，我在想， 这个miss值是根据先验theta与已知的y1来推导出来的，最后需要用y0、y1来反推出真实的系数值， 但是miss值是根据已知的theta来推导， 这个结果是不是有偏呢?

答：是的，这个是所有的贝叶斯方法都会有你说的这个问题。因为你的逻辑是说你会有一个prior distribution，然后你会去修正，得到你的W之后，是你参数的一个后验，当然，我们还会有一个Y和

第一步得到的Ymiss，Yobs，theta，然后把他俩combine在一起，会得到y,f(Ymis|Yobs,W)。这个影响是肯定会的，因为所有贝叶斯方法都会有这样的一个问题。

7. 这个就会涉及到观察量、miss值数量可能会对结果有比较大的影响。

答：这个问题其实说的非常好，因为你会有两个input，一个Input就是theta,一个Input是f (Y0, Y1|theta)，其实你可以通过P(theta)一些方法来让它rely on prior distribution，最后这个结果是说对p(theta)敏感性到底有多少?因为其实贝叶斯方法可以理解为一个trade-off，你有两部分的信息来源，一个信息来源是prior distribution,另外一个信息来源是observable 数据，整个建模过程就是这两个东西的trade-off。在医学里面，我也不知道一个新的病，只能根据以往的流行性疫情的参数进行建模，在这种情况下，我可能会要求更加信任观察的数据，尤其是像现在Covid-19，数据量非常大。但是如果根据以往的这个病去建模，可能观察性不是很强。那么我可以通过手动去设置，让最后结果more stable。这个的一个技巧就是说，让Inverse prior distribution设成10000，这是个方差，设成一万，这个时候你会发现，最后就不是很stable，这也可以说我更加信任我的数据，这个可以根据建模时去调整。

8. 假设是什么？为什么说condtional on post-treatment，就不会是unbaised?

答：这个证明主要依赖，不是我们放了哪些协变量、没有放哪些协变量，我们主要关心的是在这几个假设下，进行理论性质的证明。不是去考虑这些X带着哪些noise，你的问题是说X带noise和不带noise，这个结论是否仍然成立。我的回答是，当你的unconfoundness这个假设，就没有问题。post-treatment也可以，只要unconfoundness假设是成立的。比如考虑极端情况，Xi就是Potential outcomes，当然这个肯定不合理，但是也违背了unconfoundness假设。

9. 关于协变量，我们会考虑他们是pre-treatment还是post-treatment，然后把它作为是否作为协变量的一个标准，我的理解就是是不是要把协变量放到模型里，协变量会不会受到分配机制的影响？为什么会看重pre-treatment和post-treatment?

答：这个不是我分享的一个想法，这个可能是大家比较关注，其实更多的还是回到刚才的假设上。还是需要分清楚协变量与结局变量的一个区别，这个重点不在pre还是post-treatment assignment mechanism，就是解释为什么这个人去接受treatment，另外一个是接收control，但是去看模型的显著性，这里面也会有一些问题。因为，显著不代表有因果机制，这其实是一个小问题，assignment问题还是在做一个相关性建模，就是用协变量去预测treatment indicator是1还是0.

【读书会PPT】