

国立がん研究センター 生物統計セミナー

因果推論入門：交絡とその調整

東京理科大学 工学部情報工学科 篠崎 智大

shinozaki@rs.tus.ac.jp

2019年10月30日 18:00～19:30

国立がん研究センター 管理棟1階 第2会議室

今回と次回のアウトライン

- 因果推論Ⅰ（10月30日）：**因果推論の基礎**
 - ▶ 効果、交絡、交絡変数、ランダム化、ランダム化の仮定、層別解析、回帰、傾向スコア
- 因果推論Ⅱ（12月2日）：**モデルを用いた効果推定**
 - ▶ 回帰モデル（条件付き効果、周辺効果、回帰標準化）
 - ▶ 傾向スコアモデル（回帰調整、マッチング、逆確率重み付け IPW）
 - ▶ 二重ロバスト推定（augmented IPW）

因果効果

- ある人(Aさん)が、ある曝露(喫煙)を受けることであるイベント(10年後に死亡)を起こすかどうか
- データ(事実)
 - ▶ Aさんは喫煙し続けていたら10年後に亡くなった

喫煙はAさんの死亡に効果があった？

反事実アウトカムモデル

- Counterfactual-outcome model
- 事実データ (factual data)
 - ▶ Aさんが2009年4月から喫煙をやめることなく ($X_A = 1$)
10年後に死亡 ($Y_A = 1$)
- 反事実データ (counterfactual data)
 - ▶ Aさんが2009年4月から禁煙を開始したら ($X_A = 0$)
10年後に... ? ($Y_A = ? \text{ if } X_A = 0$)
 - If X_A had been 0, what Y_A would be?

潜在アウトカム

- Potential outcomes

- ▶ Y^x : 曝露 $X = x$ を受けた場合のアウトカム

- Aが曝露を受けた場合 ($X_A=1$) のアウトカム Y_A

- ▶ $Y_A^{x=1}$

- Aが曝露を受けなかった場合 ($X_A=0$) のアウトカム Y_A

- ▶ $Y_A^{x=0}$

潜在アウトカムによる因果効果

- Aが曝露を受けた場合 ($X_A=1$) のアウトカム Y_A
 - ▶ $Y_A^{x=1}$
- Aが曝露を受けなかった場合 ($X_A=0$) のアウトカム Y_A
 - ▶ $Y_A^{x=0}$
- Aさんに対する因果効果: $Y_A^{x=1} - Y_A^{x=0}$
- データとして得られるのは
 - ▶ 曝露ありだった場合 ($X_A=1$): $Y_A = Y_A^{x=1}$
 - ▶ 曝露なしだった場合 ($X_A=0$): $Y_A = Y_A^{x=0}$

必ず一方は反事実

潜在データと観察データ

対象者	X	$Y^{x=1}$	$Y^{x=0}$	Y	Causal effect ($Y^{x=1} - Y^{x=0}$)
A	1	1	1	1	0
B	1	1	0	1	1
C	0	0	0	0	0
D	0	1	0	0	1
E	1	0	1	0	-1
F	0	1	1	1	0
G	0	1	0	0	1

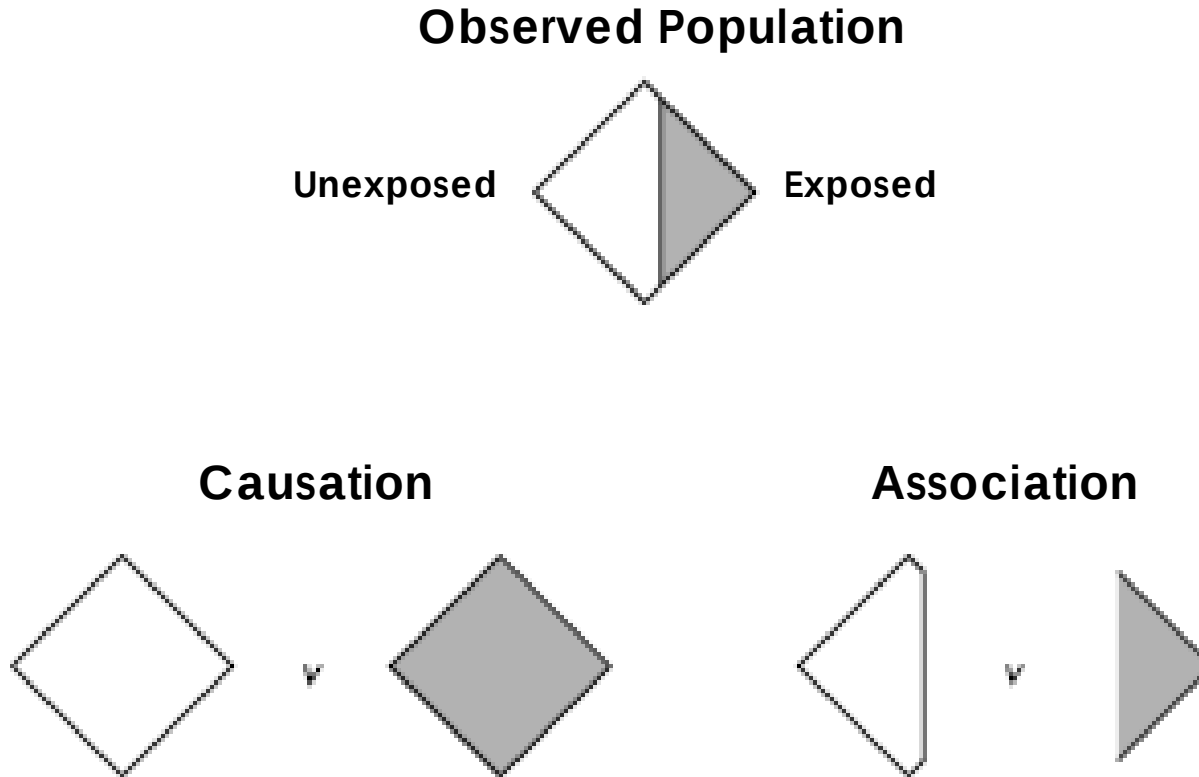
観察変数と反事実変数

- $Y^{x=1}, Y^{x=0}$
 - ▶ どちらか一方は実際に観察できない(反事実)
- 個人 i に対する因果効果 $Y_i^{x=1} - Y_i^{x=0}$ は調べられない
- 集団に対する効果
 - ▶ 平均因果効果 (average causal effect)

$$E(Y^{x=1}) - E(Y^{x=0})$$

- $E(Y^{x=1})$: 全員が曝露を受けた場合の期待値(平均)
- $E(Y^{x=0})$: 全員が曝露を受けなかった場合の期待値(平均)

Association is not causation



Hernán. *J Epidemiol Community Health* 2004; 58: 265–71
Hernán, Robins. *Causal Inference: What If*. Forthcoming

ランダム化

- 一方の曝露群に予後の悪い人が集まらない

- ▶ Y^x ($x = 0, 1$)と曝露 X が独立

- $Y^{x=1}, Y^{x=0}$

- ▶ $E(Y^{x=1} | X = 1) = E(Y^{x=1} | X = 0) = E(Y^{x=1})$

- ▶ $E(Y^{x=0} | X = 1) = E(Y^{x=0} | X = 0) = E(Y^{x=0})$

- X は交換可能 (exchangeable)

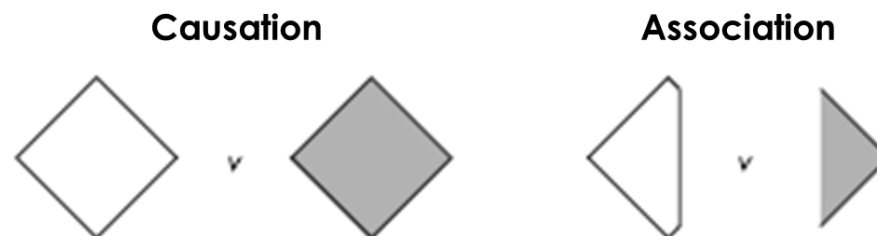
- ▶ $X = 1$ の集団

- ▶ $X = 0$ の集団

がおなじ潜在アウトカムの期待値をもつ

ランダム化試験では

- 平均因果効果: $E(Y^{x=1}) - E(Y^{x=0})$
 - ▶ $E(Y^{x=1})$
 - ランダム化の仮定が正しいので、 $E(Y^{x=1}) = E(Y^{x=1} | X = 1)$
 - データとして得られるから、 $E(Y^{x=1} | X = 1) = E(Y | X = 1)$
 - ▶ $E(Y^{x=0})$
 - ランダム化の仮定が正しいので、 $E(Y^{x=0}) = E(Y^{x=0} | X = 0)$
 - データとして得られるから、 $E(Y^{x=0} | X = 0) = E(Y | X = 0)$
- 平均因果効果 = $E(Y | X = 1) - E(Y | X = 0)$
 - ▶ Association is causation!



ランダム化の仮定

- ランダム化は実際にできていないが...
- 曝露 X が Y^x とは独立に選ばれているとみなす
 - ▶ ただし Y^x に関係しない変数にしたがって選ばれていてもよい
- 交換可能性の仮定とも
 - ▶ exchangeability assumption

観察研究

- ランダム化の仮定は正しくないと分かっている
 - ▶ Association is NOT causation
- 交絡 (confounding)
 - ▶ 曝露群: 集団全員が曝露した場合と異なる
 - $E(Y^{x=1} \mid X = 1) \neq E(Y^{x=1})$
 - and/or
 - ▶ 非曝露群: 集団全員が曝露しなかった場合と異なる
 - $E(Y^{x=0} \mid X = 0) \neq E(Y^{x=0})$

「条件付き」ランダム化の仮定

■ リスク因子 C で層別した中で

- ▶ $E(Y^{x=1} \mid X = 1, C) = E(Y^{x=1} \mid C)$
 - 曝露群の平均 = 全体(曝露 + 非曝露)が曝露した平均
- ▶ $E(Y^{x=0} \mid X = 0, C) = E(Y^{x=0} \mid C)$
 - 非曝露群の平均 = 全体(曝露 + 非曝露)が曝露しなかった平均

■ 交絡変数 (confounders)

- ▶ その層内では「ランダム化」されていると見なせるような変数

■ 交絡変数が全て測定されている仮定

- ▶ No-unmeasured confounders
- ▶ 交絡変数 C で層別すれば $Y^x (x = 0, 1)$ と曝露 X が独立

平均因果効果

- $E(Y^{x=1}) - E(Y^{x=0})$
 - ▶ $E(Y^{x=1}) = \sum_c E(Y^{x=1}|C = c)P(C = c)$
 - C で層別
 - ▶ $E(Y^{x=1}|C = c) = E(Y^{x=1}|X = 1, C = c)$
 - C の中では曝露はランダム化
 - ▶ $E(Y^{x=1}|X = 1, C = c) = E(Y|X = 1, C = c)$
 - データとして観察
- $\sum_c \{E(Y|X = 1, C = c) - E(Y|X = 0, C = c)\}P(C = c)$
 - ▶ 層ごとの平均 (= 回帰、条件付き期待値) の重み付け平均
 - ▶ 層別解析
 - 重みを $P(C = c)$ としたとき、標準化 (standardization)

因果効果の識別可能条件

- 一般的な警句: Association is not causation
- ランダム化の下では
 - ▶ 単純平均の差が因果効果
 - ▶ No-confounders
- ランダム化と見なせない(交絡がある)状況では
 - ▶ 交絡変数が全て測定されているなら、単純平均の差の重み付け平均が因果効果
 - ▶ No-unmeasured confounders
 - ▶ 測定されていない交絡変数があると因果効果は求まらない

層別解析による平均因果効果推定

平均因果効果 $E(Y^{x=1}) - E(Y^{x=0})$

■ ランダム化試験では

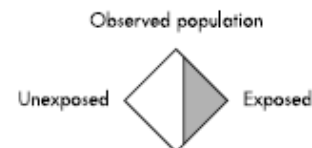
$$E(Y | X = 1) - E(Y | X = 0)$$

- ▶ $E(Y | X = x)$: 曝露群(x)ごとの平均値で推定

■ 交絡変数が全て測定されている観察研究では

$$\begin{aligned} & \sum_c E(Y | X = 1, C = c)P(C = c) \\ & - \sum_c E(Y | X = 0, C = c)P(C = c) \end{aligned}$$

- ▶ $E(Y | X = x, C = c)$: 層(c) × 曝露群(x)ごとに推定
- ▶ $P(C = c)$: 層(c)の大きさ



層別解析 stratified analysis

1. 交絡変数 C で層別

- ▶ C : 性、年齢、教育歴、...
- ▶ 組み合わせでサブグループに分ける
- ▶ サブグループ $k = 1, \dots, K$

2. 層ごとに推定

- ▶ 連続量: $\hat{E}(Y | X = 1, C)$ vs. $\hat{E}(Y | X = 0, C)$
- ▶ 二値アウトカム: Y_{1k}/N_{1k} vs. Y_{0k}/N_{0k}
 - Y_{1k}, Y_{0k} : サブグループ k の曝露群と非曝露群でのイベント数 (Y の合計)
 - N_{1k}, N_{0k} : サブグループ k の曝露群と非曝露群の人数

3. 層ごとの結果を併合

- ▶ サブグループの割合 $(n_k + m_k)/N$ を重み

仮想例

層 k	C		$X = 1$		$X = 0$	
	Sex	Age	N	$Y = 1$	N	$Y = 1$
1	Male	Young	1360	577	1640	423
2	Male	Old	2800	1709	1200	711
3	Female	Young	2660	1707	1140	713
4	Female	Old	4020	2978	980	722
Total			10840	6971	4960	2569

$$E(Y|X = 1) = 64.3\%$$

$$E(Y|X = 0) = 51.8\%$$

■ 未調整リスク差 = $64.3\% - 51.8\% = 12.5\%$

Step 1: 層ごとの平均値

層 k	C		N	X = 1		N	X = 0	
	Sex	Age		Y = 1	Risk		Y = 1	Risk
1	Male	Young	1360	577		1640	423	
2	Male	Old	2800	1709		1200	711	
3	Female	Young	2660	1707		1140	713	
4	Female	Old	4020	2978		980	722	
Total			10840	6971		4960	2569	

Step 1: 層ごとの平均値

層 k	C		N	$X = 1$		N	$X = 0$	
	Sex	Age		$Y = 1$	Risk		$Y = 1$	Risk
1	Male	Young	1360	577	0.424	1640	423	0.258
2	Male	Old	2800	1709	0.61	1200	711	0.593
3	Female	Young	2660	1707	0.642	1140	713	0.625
4	Female	Old	4020	2978	0.741	980	722	0.737
Total			10840	6971		4960	2569	

$$E(Y|X, C) = P(Y=1|X, C)$$

Step 2: 層ごとの重み

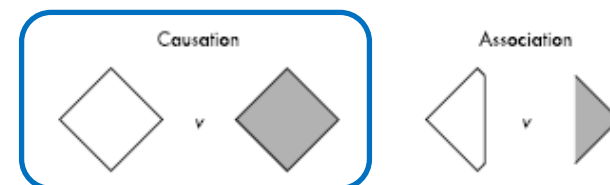
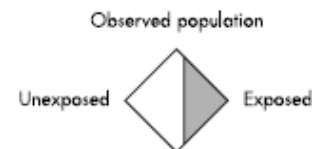
層 k	C		$X = 1$		$X = 0$		N_k	$P(C)$
	Sex	Age	N	Risk	N	Risk		
1	Male	Young	1360	0.424	1640	0.258		
2	Male	Old	2800	0.61	1200	0.593		
3	Female	Young	2660	0.642	1140	0.625		
4	Female	Old	4020	0.741	980	0.737		
Total			10840		4960		15800	

Step 2: 層ごとの重み

層 k	C		$X = 1$		$X = 0$		N_k	$P(C)$
	Sex	Age	N	Risk	N	Risk		
1	Male	Young	1360	0.424	1640	0.258	3000	0.19
2	Male	Old	2800	0.61	1200	0.593	4000	0.25
3	Female	Young	2660	0.642	1140	0.625	3800	0.24
4	Female	Old	4020	0.741	980	0.737	5000	0.32
Total			10840		4960		15800	

$$\sum_c E(Y \mid X = x, C = c) P(C = c)$$

Step 3: 重み付け平均



■ ◆ : $E(Y^{X=1})$

- ▶ 集団全体が $X = 1$ だった場合
- ▶ $\sum_c E(Y | X = 1, C = c)P(C = c)$
 $= 0.424(0.19) + \dots + 0.741(0.32) = 62.4\%$

■ ◇ : $E(Y^{X=0})$

- ▶ 集団全体が $X = 0$ だった場合
- ▶ $\sum_c E(Y | X = 0, C = c)P(C = c)$
 $= 0.258(0.19) + \dots + 0.737(0.32) = 58.3\%$

■ 調整リスク差 = $62.4\% - 58.3\% = 4.1\%$

- ▶ 「標準化リスク差」とも

層別解析、Y から見るか？ X から見るか？



<https://www.amazon.co.jp>

因果効果の求め方(1)

Y から見る

■ 層別解析

1. 交絡変数 C の組み合わせで層に分ける
2. 層ごとにアウトカムの平均 $E(Y|X = x, C)$ を計算
3. 2 をさらに「層ごとの重み」で重み付け平均
 $E(Y|X = x, C)P(C)$

■ 集団全体の効果が求められる

$$E(Y^{x=1}) \text{ vs. } E(Y^{x=0})$$

因果効果の求め方(2)

Y から見る

- 層別しきれないほどの変数があるとき
- モデルで $E(Y|X, C)$ を近似

- ▶ 例: ロジスティック回帰モデル

$$\log \frac{P(Y = 1|X = x, C = c)}{1 - P(Y = 1|X = x, C = c)} = \gamma_0 + \gamma_1 X + \gamma_2 C$$

- ▶ $\exp(\gamma_1)$ は「調整オッズ比」

- 個人ごとに異なる C の組み合わせごとに $E(Y|X = x, C)$ を予測

- ▶ $x = 1$ を代入: $R_{1i} = 1/\{1 + \exp(-\hat{\gamma}_0 - \hat{\gamma}_1 - \hat{\gamma}_2 C_i)\}$
- ▶ $x = 0$ を代入: $R_{0i} = 1/\{1 + \exp(-\hat{\gamma}_0 - \hat{\gamma}_2 C_i)\}$

- $R_{1i} - R_{0i}$ を全員で平均

- ▶ 回帰モデルにもとづく標準化 (regression standardization)

因果効果の求め方(3)

X から見る

- 各層の曝露確率 $P(X = 1|C)$ を求める
 - ▶ 傾向スコア (propensity score: PS)

各層の傾向スコア

層 k	C		$X = 1$		$X = 0$		$P(X=1 C)$
	Sex	Age	N	$Y = 1$	N	$Y = 1$	
1	Male	Young	1360	577	1640	423	0.453
2	Male	Old	2800	1709	1200	711	0.700
3	Female	Young	2660	1707	1140	713	0.700
4	Female	Old	4020	2978	980	722	0.804
Total			10840	6971	4960	2569	

- 傾向スコアがおなじ層をまとめてみよう

傾向スコア PS で層別したデータ

層 k	$X = 1$			$X = 0$			$P(X=1 C)$ = PS	$P(\text{PS})$
	N	$Y = 1$	Risk	N	$Y = 1$	Risk		
1	1360	577	0.424	1640	423	0.258	0.453	0.19
2 & 3	5460	3416	0.626	2340	1424	0.609	0.700	0.49
4	4020	2978	0.741	980	722	0.737	0.804	0.32
Total	10840	6971		4960	2569			1

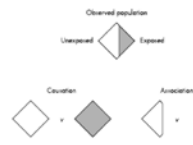
傾向スコアで「層別」

■ $P(X = 1|C)$ で層別解析 (標準化)

- ▶ $\sum_{PS} E(Y|X = 1, PS)P(PS)$
 $= 0.424(0.19) + 0.626(0.49) + 0.741(0.32) = 62.4\%$
- ▶ $\sum_{PS} E(Y|X = 0, PS)P(PS)$
 $= 0.258(0.19) + 0.609(0.49) + 0.737(0.32) = 58.3\%$

■ C そのもので層別解析した結果とおなじ

Step 3: 重み付け平均



- $E(Y^{X=1})$
 - ▶ 集団全体が $X = 1$ だった場合
 - ▶ $\sum_c E(Y|X = 1, C = c)P(C = c)$
 $= 0.424(0.19) + \dots + 0.741(0.32) = 62.4\%$
- $E(Y^{X=0})$
 - ▶ 集団全体が $X = 0$ だった場合
 - ▶ $\sum_c E(Y|X = 0, C = c)P(C = c)$
 $= 0.258(0.19) + \dots + 0.737(0.32) = 58.3\%$
- 調整リスク差 = $62.4\% - 58.3\% = 4.1\%$
 - ▶ 「標準化リスク差」とも

26

傾向スコア PS の推定

- C で層別できるなら、層 k の曝露割合

- ▶ 曝露 n_k 人、非曝露 m_k 人 $\longrightarrow PS_k = n_k / (n_k + m_k)$
- ▶ 求め方(1)の標準化と同じ推定量が導ける

Sato & Matsuyama, *Epidemiology* 2003

- できないなら、モデル化

- ▶ 例: ロジスティック回帰

$$\log \frac{P(X = 1 | C = c)}{1 - P(X = 1 | C = c)} = \alpha_0 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3 + \dots$$

$$\longrightarrow \hat{P}(X = 1 | C = c) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 c_1 + \hat{\alpha}_2 c_2 + \hat{\alpha}_3 c_3 + \dots)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 c_1 + \hat{\alpha}_2 c_2 + \hat{\alpha}_3 c_3 + \dots)}$$

因果効果の求め方(4)

Y から見る

- 層別しきれないほどの変数があるとき
- モデルで $P(X = 1|C)$ を近似

$$\log \frac{P(X = 1|C = c)}{1 - P(X = 1|C = c)} = \alpha_0 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3 + \dots$$

$$\longrightarrow \hat{P}(X = 1|C = c) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 c_1 + \hat{\alpha}_2 c_2 + \hat{\alpha}_3 c_3 + \dots)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 c_1 + \hat{\alpha}_2 c_2 + \hat{\alpha}_3 c_3 + \dots)}$$

- 3つの方法
 - ▶ 層別・回帰調整 (stratification/adjustment)
 - ▶ マッチング (matching)
 - ▶ 重み付け (weighting)

傾向スコア

- 交絡変数 C で層別した中での条件付き曝露確率
 - ▶ $PS(c) = P(X = 1|C = c)$
- 単なる「曝露確率」ではない
 - ▶ 交絡変数を特定する必要
 - 交絡変数の定義が必要
 - 交絡の定義が必要
- 交絡変数の組み合わせごとに傾向スコアも異なる
 - ▶ 「そのひと特有の曝露確率」の推定が目的ではない
 - 目的は交絡調整、傾向スコアは手段

balancing・スコア

- Balancing score: $b(Z)$
 - ▶ 何らかの観察変数 Z の実数値関数で、
 - ▶ 同じスコア $b(Z)$ の集団の中では曝露 X と Z が独立
= 曝露群間で Z がバランス
- 最も細かい (finest) balancing・スコア
 - ▶ $b(Z)$: Z の全組み合わせ水準
 - $b(z) = (Z = z \text{ となる組み合わせ水準})$
- 最も粗い (coarsest) balancing・スコア
 - ▶ $b(z) = P(X = 1 \mid Z = z)$
 - 粗いスコアは細かいスコアの関数で表せる

Rosenbaum & Rubin, *Biometrika* 1983

傾向スコアはバランシング・スコア

- 傾向スコア: $PS(C)$
 - ▶ 交絡変数 C の最も粗いバランシング・スコア
 - ▶ $PS(c) = P(X = 1 \mid C = c)$
- 同じ傾向スコア $PS(C)$ の集団の中では
 - ▶ 曝露 X と交絡変数 C が独立
 - ▶ 曝露群間で交絡変数 C のバランスがとれる

条件付きランダム化の仮定の下で

- 条件付きランダム化の仮定
 - ▶ 交絡変数 C で層別すれば $Y^x(x=0, 1)$ と曝露 X が独立
- 傾向スコア $PS(C)$ のバランス性
 - ▶ $PS(C)$ で層別すれば曝露 X と交絡変数 C が独立



$PS(C)$ で層別すれば
 $Y^x(x=0, 1)$ と曝露 X が独立

Rosenbaum and Rubin, *Biometrika* 1983

傾向スコア、3通りの使い方

因果効果の求め方(4)

再掲

- 層別しきれないほどの変数があるとき
- モデルで $P(X = 1|C)$ を近似

$$\log \frac{P(X = 1|C = c)}{1 - P(X = 1|C = c)} = \alpha_0 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3 + \dots$$

$$\longrightarrow \hat{P}(X = 1|C = c) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 c_1 + \hat{\alpha}_2 c_2 + \hat{\alpha}_3 c_3 + \dots)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 c_1 + \hat{\alpha}_2 c_2 + \hat{\alpha}_3 c_3 + \dots)}$$

- 3つの方法
 - ▶ 層別・回帰調整 (stratification/adjustment)
 - ▶ マッチング (matching)
 - ▶ 重み付け (weighting)

傾向スコア利用法(1): 層別・調整

- adjustment (= conditioning)
- C で層別できるなら

傾向スコア PS で層別したデータ								
層 k	$X = 1$			$X = 0$			$P(X=1 C)$ = PS	$P(PS)$
	N	$Y = 1$	Risk	N	$Y = 1$	Risk		
1	1360	577	0.424	1640	423	0.258	0.453	0.19
2 & 3	5460	3416	0.626	2340	1424	0.609	0.700	0.49
4	4020	2978	0.741	980	722	0.737	0.804	0.32
Total	10840	6971		4960	2569			1

- ▶ C で層別しても $PS(C) = P(X = 1|C)$ で層別してもおなじ

Cで層別できないときのPS層別・調整

- モデルから $\widehat{PS}(c) = \widehat{P}(X = 1|C = c)$
- $\widehat{PS}(C)$ で層別
 - ▶ 例) 5分位でサブグループ化 (例: 0~0.3, 0.3~0.5, ..., 0.8~1)
 - ▶ 各グループ内で比較: $Y_{1k}/n_k - Y_{0k}/m_k$
 - $k = 1, \dots, 5$
 - ▶ 各グループの結果を併合
- 共変量調整
 - ▶ 例) ロジスティックモデル

$$\log \frac{\Pr\{Y = 1|X = x, \widehat{PS}(C)\}}{1 - \Pr\{Y = 1|X = x, \widehat{PS}(C)\}} = \gamma_0 + \gamma_1 X + \gamma_2 \widehat{PS}(C)$$

傾向スコア調整の注意点

■ 層別解析

- ▶ 層内でも傾向スコアは連続的に変化
- ▶ 交絡変数のバランスが完全にはとれない(= 残差交絡)
 - できるだけ細かく層別した方がよい？

■ 共変量調整

- ▶ モデルの仮定
 - PSの単位上昇あたりのオッズ比が単調(= γ_2)
 - PSの値によらず曝露効果一定という仮定
 - (ただしN大で線形モデルでは g-推定量に一致、モデル誤特定下でも解釈可)
- ▶ 求めているのは**集団全体**の効果ではなく **PS値ごと**の効果
 - ロジスティック回帰(オッズ比)、Cox回帰(ハザード比)で問題
 - Non-collapsibility(後述)
 - さらに**回帰標準化**(次回)を行うことも

傾向スコア利用法(2): マッチング

■ Matching

1. 傾向スコア $PS(C_i)$ を個人 i ごとに求める
2. 曝露群と非曝露群で $PS(C_i)$ の同じ or 近い人をペアにする
3. ペアに選ばれた人だけ解析に利用する

■ 傾向スコアでマッチされた集団

- ▶ Propensity-matched cohort
- ▶ 各ペアは「同じ交絡変数分布」からのランダムサンプル
 - 個々のペア内で交絡変数が似ている保証はないが...

■ 全体として「交絡がない」とみなして解析

Cで層別できるとき

- 傾向スコア PS の値がおなじ層で人数をそろえる

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS
	N	$Y = 1$	N	$Y = 1$	
1	1360	577	1640	423	0.453
			$\Rightarrow 1360/1640$		
2 & 3	5460	3416	2340	1424	0.700
	$\Rightarrow 2340/5460$				
4	4020	2978	980	722	0.804
	$\Rightarrow 980/4020$				
Total	10840		4960	2569	

$P(X = 1|C)$ でマッチしたコホート

■ PS matched cohort

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS
	N	$Y = 1$	N	$Y = 1$	
1	1360	577	1360	351	0.453
2 & 3	2340	1464	2340	1424	0.700
4	980	726	980	722	0.804
Total	4680	2767	4680	2497	

▶ $E_{\text{PSM}}(Y|X = 1) = 2767/4680 = 59.1\%$

▶ $E_{\text{PSM}}(Y|X = 0) = 2497/4680 = 53.4\%$

標的集団

- $E(Y^{x=1}) - E(Y^{x=0})$
 - ▶ もとの集団全体が $X = 1$ だった場合 vs. $X = 0$ だった場合
- 層別解析(標準化)
 - ▶ $\sum_c E(Y | X = x, C = c)P(C = c)$
 - ▶ $\sum_p E(Y | X = x, PS = p)P(PS = p)$
 - ▶ 標的集団: もとの集団全体
- マッチング
 - ▶ $E_{PSM}(Y|X = 1)$
 - ▶ 標的集団: もとの集団からマッチされた集団
 - $E(Y^{x=1} | PS\text{-matched}) - E(Y^{x=0} | PS\text{-matched})$

交絡変数の分布 $P(C)$ の変化

■ マッチング前

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS	$P(PS)$
	N	$Y = 1$	N	$Y = 1$		
1	1360	577	1640	423	0.453	0.19
2 & 3	5460	3416	2340	1424	0.700	0.49
4	4020	2978	980	722	0.804	0.32
Total	10840	6971	4960	2569		1

■ PSマッチング後

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS	$P(PS)$
	N	$Y = 1$	N	$Y = 1$		
1	1360	577	1360	351	0.453	0.29
2 & 3	2340	1464	2340	1424	0.700	0.5
4	980	726	980	722	0.804	0.21
Total	4680	2767	4680	2497		1

PSの分布を変えないマッチング

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS	$P(PS)$
	N	$Y = 1$	N	$Y = 1$		
1	1360	577	1640	423	0.453	0.19
			$\Rightarrow (980/1640) * (0.19/0.32)$			
2 & 3	5460	3416	2340	1424	0.700	0.49
			$\Rightarrow (980/2340) * (0.49/0.32)$			
4	4020	2978	980	722	0.804	0.32
Total	10840	6971	4960	2569		1

PSの分布を変えないマッチング

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS	$P(PS)$
	N	$Y = 1$	N	$Y = 1$		
1	1360	577 $\Rightarrow 588/1360$	588	152	0.453	0.19
2 & 3	5460	3416 $\Rightarrow 1529/5460$	1529	930	0.700	0.49
4	4020	2978 $\Rightarrow 980/4020$	980	722	0.804	0.32
Total	10840	6971	3097	1804		1

PSの分布を変えないマッチング

層 k	$X = 1$		$X = 0$		$P(X=1 C)$ = PS	$P(PS)$
	N	$Y = 1$	N	$Y = 1$		
1	588	249	588	152	0.453	0.19
2 & 3	1529	957	1529	930	0.700	0.49
4	980	726	980	722	0.804	0.32
Total	3097	1932	3097	1804		1

- ▶ $E_{PSM}(Y|X = 1) = 1932/3097 = 62.4\%$
- ▶ $E_{PSM}(Y|X = 0) = 1804/3097 = 58.3\%$
 - 集団全体を標的とした標準化推定量に一致

傾向スコア利用法(3): 重み付け

■ Inverse probability weighting (IPW)

- ▶ 個人 i が実際に受けた曝露を受ける確率の逆数で重み付け
 - 曝露群では $W_i = 1/PS_i$
 - 非曝露群では $W_i = 1/(1 - PS_i)$
- ▶ 重み付けた集団で、そのまま差を計算

$$W_i = \frac{1}{P(X = x_i | C = c_i)}$$

Rosenbaum, JASA 1987

■ W_i で重み付けた集団では

- ▶ 曝露 X と交絡変数 C が独立 (statistically exogenous)

■ 条件付きランダム化の仮定と合わせることで

- ▶ $Y^x(x = 0, 1)$ と曝露 X が独立 (causally exogenous)

Hernan et al., JASA 2001

直観的な理解

■ 欠測データ解析の重み付け推定

▶ Horvitz–Thompson推定量

- 観測確率の逆数で重み付け
- 観察確率 p_i の i さんは、たまたま確率 p_i で観測された
- 全体では、同じ観察確率 p_i の人が $1/p_i$ 人があるはず

対象者	X	$Y^{X=1}$	$Y^{X=0}$	Y	Causal effect ($Y^{X=1} - Y^{X=0}$)
A	1	1	1	1	0
B	1	1	0	1	1
C	0	0	0	0	0
D	0	1	0	0	1
E	1	0	1	0	-1
F	0	1	1	1	0
G	0	1	0	0	1

■ 曝露・非曝露を $Y^{X=1}$, $Y^{X=0}$ の欠測問題とみなせば...

▶ IPW推定量

- 曝露群の i ($Y^{X=1}$ が観察) と同じ曝露確率 p_i の人は全体に $1/p_i$ 人
- 非曝露群の j ($Y^{X=0}$ が観察) と同じ曝露確率 p_j の人は全体に $1/(1-p_j)$ 人

IPWの計算

C		X = 1			X = 0			$P(X=1 C)$
Sex	Age	N	$N_{Y=1}$	IPW	N	$N_{Y=1}$	IPW	= PS
Male	Young	1360	577	2.21	1640	423	1.83	0.453
Male	Old	2800	1709	1.43	1200	711	3.33	0.700
Female	Young	2660	1707	1.43	1140	713	3.33	0.700
Female	Old	4020	2978	1.24	980	722	5.10	0.804
Total		10840	6971		4960	2569		

▶ $IPW = X/PS + (1 - X)/(1 - PS)$

■ $X = 1 \Rightarrow 1/PS$

■ $X = 0 \Rightarrow 1/(1 - PS)$

IPWで重み付け

C		X = 1			X = 0			$P(X=1 C)$ = PS
Sex	Age	N_{IPW}	$N_{Y=1,IPW}$	IPW	N_{IPW}	$N_{Y=1,IPW}$	IPW	
Male	Young	3000	1272.8	2.21	3000	773.8	1.83	0.453
Male	Old	4000	2441.4	1.43	4000	2370.0	3.33	0.700
Female	Young	3800	2438.6	1.43	3800	2376.6	3.33	0.700
Female	Old	5000	3704.0	1.24	5000	3683.7	5.10	0.804
Total		15800	9856.8		15800	9204.1		

▶ $N_{IPW} = N * IPW$

▶ $N_{Y=1,IPW} = N_{Y=1} * IPW$

IPWで重み付け

C		X = 1			X = 0			$P(X=1 C)$ = PS
Sex	Age	N_{IPW}	$N_{Y=1}$	IPW	N_{IPW}	$N_{Y=1}$	IPW	
Male	Young	3000	1272.8	2.21	3000	773.8	1.83	0.453
Male	Old	4000	2441.4	1.43	4000	2370.0	3.33	0.700
Female	Young	3800	2438.6	1.43	3800	2376.6	3.33	0.700
Female	Old	5000	3704.0	1.24	5000	3683.7	5.10	0.804
Total		15800	9856.8		15800	9204.1		

- ▶ $E_{IPW}(Y|X = 1) = 9856.8/15800 = 62.4\%$
- ▶ $E_{IPW}(Y|X = 0) = 9204.1/15800 = 58.3\%$

統計モデル

- アウトカム回帰モデル
 - ▶ $E(Y|X = x, C = c)$ を近似
- 傾向スコアモデル
 - ▶ $P(X = 1|C = c)$ を近似

回帰モデルの係数

$$\log \frac{\Pr(Y = 1|X = x, C = c)}{1 - \Pr(Y = 1|X = x, C = c)} = \gamma_0 + \gamma_1 X + \gamma_2 C$$

■ 「調整」オッズ比: $\exp(\gamma_1)$

- ▶ $P(Y^{x=1}|C = c)$ vs. $P(Y^{x=0}|C = c)$
- ▶ $C = c$ という層内での因果効果 = 条件付き効果
- ▶ 層別解析の「各層内の効果」

- ▶ $P(Y = 1|X = x, C = c) = P(Y^{x=1} = 1|X = x, C = c)$
 - 曝露 $X = x$ を受けているので
- ▶ $P(Y^{x=1} = 1|X = x, C = c) = P(Y^{x=1} = 1|C = c)$
 - C の層内では交絡がないので

層ごとの効果

層 k	C		$X = 1$		$X = 0$		Risk Difference
	Sex	Age	N	Risk	N	Risk	
1	Male	Young	1360	0.424	1640	0.258	0.166
2	Male	Old	2800	0.61	1200	0.593	0.017
3	Female	Young	2660	0.642	1140	0.625	0.017
4	Female	Old	4020	0.741	980	0.737	0.04
Total			10840		4960		

■ $E(Y^{x=1}|C = c) - E(Y^{x=0}|C = c)$

- ▶ 交絡変数 $C = c$ の層の全員が $X = 1$ の場合 vs. $X = 0$ の場合
- ▶ 条件付き効果 (⇔ 周辺効果 $E(Y^{x=1})$ vs. $E(Y^{x=0})$)

回帰モデルによる効果推定

- あくまで交絡変数 C で層別したい
 - ▶ 交絡変数が全て測定されている仮定
- できないので、 C で層別された層ごとの平均を近似
 - ▶ $E(Y|X, C)$ を十分近似できている仮定
- 回帰係数は条件付き効果
 - ▶ 交絡変数 C がおなじ値のサブグループでの効果
 - ▶ サブグループ効果が等しいと仮定(交互作用項いれないとき)
- 周辺効果(平均因果効果)を知るには予測値から標準化
 - ▶ 集団全体が $X = 1$ だった場合 vs. $X = 0$ だった場合

傾向スコアモデルによる効果推定

- あくまで交絡変数 C で層別したい
 - ▶ 交絡変数が全て測定されている仮定
- できないので、 C で層別された層ごとの曝露確率を近似
 - ▶ $P(X = 1|C)$ を十分近似できている仮定
- 求まる効果はその後の解析方法しだい
 - ▶ 層別・回帰調整(+ 標準化)
 - ▶ マッチング
 - ▶ IPW

傾向スコア PS の層ごとの効果

層 k	$P(X=1 C) = \text{PS}$	$X = 1$		$X = 0$		Risk Difference
		N	Risk	N	Risk	
1	0.453	1360	0.424	1640	0.258	0.166
2 & 3	0.700	5460	0.626	2340	0.609	0.017
4	0.804	4020	0.741	980	0.737	0.04
Total		10840		4960		

■ $E(Y^{x=1} | \text{PS} = p) - E(Y^{x=0} | \text{PS} = p)$

- ▶ 傾向スコア $\text{PS} = p$ の層の全員が $X = 1$ の場合 vs. $X = 0$ の場合
- ▶ 条件付き効果

傾向スコアにより推定される効果

1. 傾向スコアの層別、回帰調整

- ▶ そのままではPSの値が同じ層での条件付き効果
- ▶ (回帰)標準化をつかえば周辺効果 (= 平均因果効果)

2. 傾向スコアマッチング

- ▶ 標的集団の変化
- ▶ マッチング後の解析によって条件付き効果 or 周辺効果

3. 傾向スコアによる重み付け(IPW)

- ▶ 集団全体での周辺効果 (= 平均因果効果)
 - 他の集団での周辺効果の重みも
- ▶ 重み付け後の解析によって条件付き効果も

まとめ

- 観察研究データから因果的命題を導くのはたいへん
 - ▶ 因果命題は反事実命題
 - ▶ 反事実モデルにおける仮定が必要
- 仮定さえ置けば、やりようはある
 - ▶ 層別解析
 - ▶ 回帰
 - ▶ 傾向スコア

} 層別解析をどちらから見るか
- 仮定に対する感度はどの手法も共有
 - ▶ 条件付きランダム化の仮定
 - ▶ 交絡変数 C が高次元になるとモデル化の仮定(👉次回くわしく)



<https://www.amazon.co.jp>

傾向スコア

- 交絡変数 C で層別した中での条件付き曝露確率
 - ▶ $PS(c) = P(X = 1|C = c)$
- 「対象者の曝露確率」ではない
 - ▶ 「交絡変数により定義されるサブグループ」での確率
- 交絡変数の組み合わせごとに傾向スコアも異なる
 - ▶ 「その人特有の曝露確率」ではない

疑似ランダム化データ？

- 傾向スコアでマッチング・層別すれば
- 曝露確率の逆数で重み付ければ
 - ▶ 「バランス性」から傾向スコア推定に使った変数がそろうのは当然
 - ▶ 潜在アウトカム Y^x と曝露 X が独立？
- 「疑似ランダム化」とみなすための必要条件
 - ▶ 条件付きランダム化の仮定
 - ▶ 傾向スコア $P(X = 1|C = c)$ が正しく得られること
- 「傾向スコアを用いれば疑似ランダム化」
 - ▶ 「条件付きランダム化を仮定している」という主張に過ぎない

Take-home message

- 交絡調整 = 層別解析
 - ▶ 交絡変数でサブグループに分けて(= 層別)比較
- 交絡調整における2種類の統計モデル
 - ▶ アウトカム回帰モデル  $E(Y|X, C)$
 - ▶ 傾向スコアモデル  $P(X = 1|C)$
- モデルをつかう目的
 - ▶ 層別できないくらいの交絡変数があるとき
 - ▶ 「もし層別できた場合」の結果を近似
 - アウトカム回帰
 - 傾向スコア
 - ▶ 各モデルの推定自体が目的ではない