

国立がん研究センター 生物統計セミナー

因果推論の応用： モデルを用いた交絡調整

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12月23日 18:30～20:00

国立がん研究センター 管理棟1階 第2会議室

前回と今回のアウトライン

- 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23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）

Take-home message (前回)

- 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

復習：記法

■ 観察変数

- ▶ X : 曝露(治療)
- ▶ Y : アウトカム
- ▶ C : 交絡変数

■ 潜在アウトカム変数

- ▶ $Y^{x=1}$: $X = 1$ を受けた場合に観察されるであろうアウトカム
- ▶ $Y^{x=0}$: $X = 0$ を受けた場合に観察されるであろうアウトカム
- ▶ X の値に応じて一方だけ観察される

■ $E(Y)$, $E(Y^x)$

■ $E(Y|X = x)$, $E(Y^x|C = c)$, $E(Y|C = c, X = x)$

復習：交絡と交絡変数

■ 交絡 (confounding)

- ▶ 曝露群：集団全員が曝露した場合と異なる

- $E(Y^{x=1} | X = 1) \neq E(Y^{x=1})$

and/or

- ▶ 非曝露群：集団全員が曝露しなかった場合と異なる

- $E(Y^{x=0} | X = 0) \neq E(Y^{x=0})$

■ 交絡変数 (confounders)

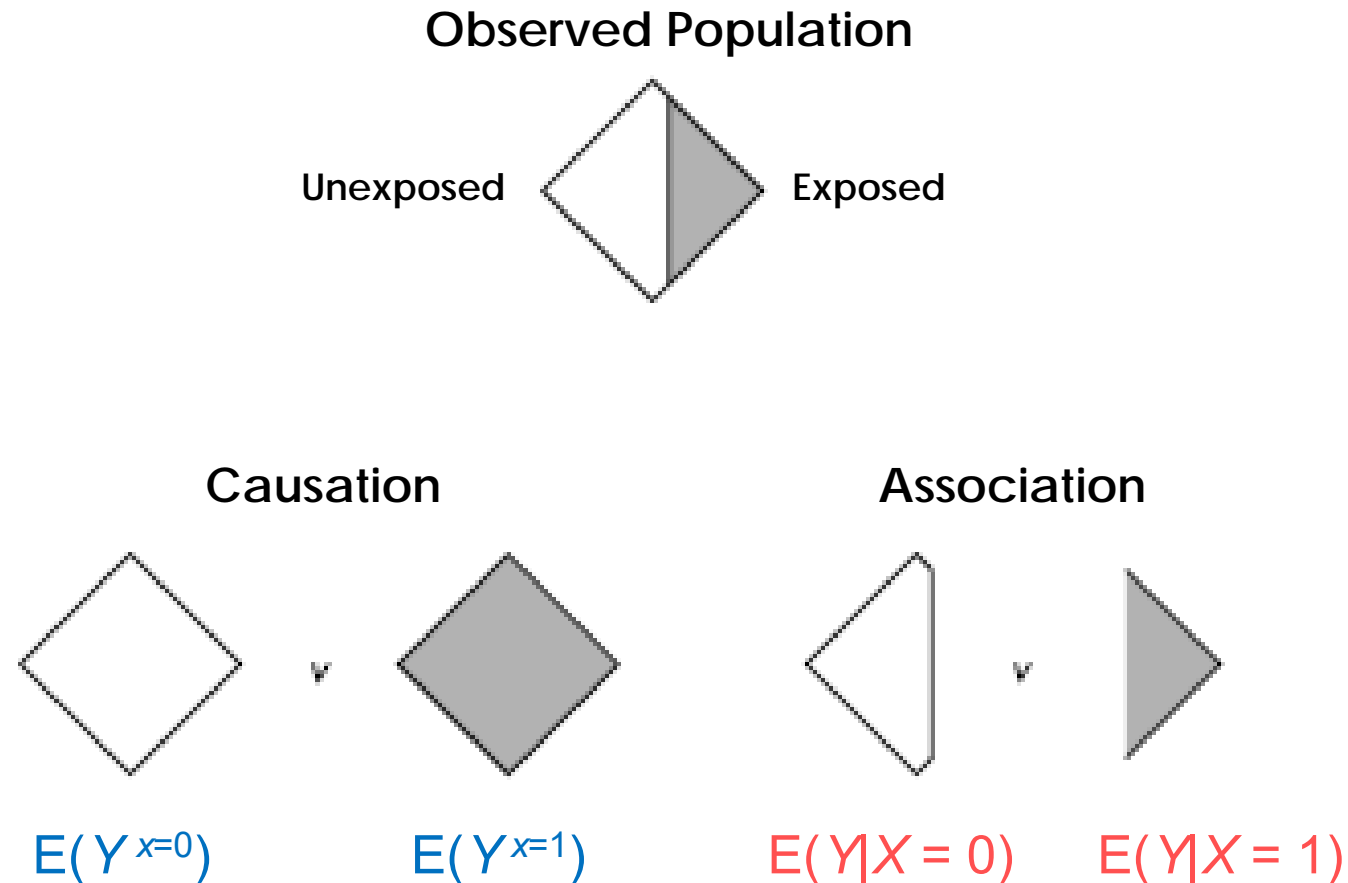
- ▶ その層内では「ランダム化」されていると見なせるような変数

- ▶ 交絡変数 C で層別すれば $Y^x (x = 0, 1)$ と曝露 X が独立

- $E(Y^{x=1} | X = 1, C = c) = E(Y^{x=1} | C = c)$

- $E(Y^{x=0} | X = 0, C = c) = E(Y^{x=0} | C = c)$

関心のある「平均因果効果」



Hernán. *J Epidemiol Community Health* 2004; 58: 265–71

Hernán, Robins. *Causal Inference: What If*. Forthcoming

層別できるなら

■ 層ごとの平均の重み付け平均

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

- ▶ $E(Y|X=x, C=c)$: 各層 ($C=c$) 内の曝露ごと ($X=x$) の回帰
- ▶ $P(C=c)$: 各層 ($C=c$) の大きさ

■ 交絡変数が C としてすべて測定されていれば

- ▶ = 交絡変数 C で層別すれば $Y^x(x=0, 1)$ と曝露 X が独立

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

層別解析

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	

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

層別解析

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	

$$\sum_c E(Y | X = 1, C = c) P(C = c) = 0.424(0.19) + \dots + 0.741(0.32) = 62.4\%$$

$$\sum_c E(Y | X = 0, C = c) P(C = c) = 0.258(0.19) + \dots + 0.737(0.32) = 58.3\%$$

回帰モデル

■ モデルで $E(Y|X, C)$ を近似

▶ 例: ロジスティック回帰モデル

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

$$\longrightarrow \hat{R}_{xc} = \hat{P}(Y=1|X=x, C=c) = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 X + \hat{\gamma}_2 C)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 X + \hat{\gamma}_2 C)}$$

層 k	C		X = 1			X = 0		
	Sex	Age	N	Y = 1	Risk	N	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	

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

回帰モデルの当てはめ

Sex	Age	X	Y	N
1	0	1	1	577
1	1	1	1	1709
0	0	1	1	1707
0	1	1	1	2978
1	0	0	1	423
1	1	0	1	711
0	0	0	1	713
0	1	0	1	722
1	0	1	0	783
1	1	1	0	1091
0	0	1	0	953
0	1	1	0	1042
1	0	0	0	1217
1	1	0	0	489
0	0	0	0	427
0	1	0	0	258

```
proc logistic desc;  
    model y = x sex age;  
    freq n;  
    output out = pred p = R;  
  
run;
```



変数	係数 γ	OR = $\exp(\gamma)$	95% CI	
1(切片)	0.2509			
X	0.2404	1.272	1.183	1.368
Sex	-0.8674	0.420	0.393	0.449
Age	0.7425	2.101	1.965	2.247

回帰モデルの予測値

層 k	C		N	X = 1		N	X = 0	
	Sex	Age		Risk	R_{xc}		Risk	R_{xc}
1	Male	Young	1360	0.424	0.407	1640	0.258	0.351
2	Male	Old	2800	0.61	0.591	1200	0.593	0.531
3	Female	Young	2660	0.642	0.620	1140	0.625	0.562
4	Female	Old	4020	0.741	0.774	980	0.737	0.730
Total			10840			4960		

$$\hat{R}_{xc} = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 x + \hat{\gamma}_2 c)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 x + \hat{\gamma}_2 c)}$$

回帰モデルによる標準化

- $\sum_c P(Y = 1|X = x, C = c)P(C = c)$
 - ▶ ただし、 $P(Y = 1|X = x, C = c)$ を \hat{R}_{xc} で近似

$$\begin{aligned}\sum_c P(Y = 1|X = 1, C = c)P(C = c) \\ = 0.407(0.19) + \dots + 0.774(0.32) = 62.1\%\end{aligned}$$

$$\begin{aligned}\sum_c P(Y = 1|X = 0, C = c)P(C = c) \\ = 0.351(0.19) + \dots + 0.730(0.32) = 56.7\%\end{aligned}$$

$P(Y = 1|X = 1, C = C_i)$ と $P(Y = 1|X = 0, C = C_i)$
を個人ごとに予測して平均しているのとおなじ

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

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

$$\longrightarrow \hat{R}_{xi} = \hat{P}(Y = 1|X = x, C = C_i) = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 x + \hat{\gamma}_2 C_i)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 x + \hat{\gamma}_2 C_i)}$$

- 個人の値 C_i ごとに $E(Y|X = x, C_i)$ を予測
 - ▶ $x = 1$: $\hat{R}_{1i} = 1/\{1 + \exp(-\hat{\gamma}_0 - \hat{\gamma}_1 - \hat{\gamma}_2 C_i)\}$
 - ▶ $x = 0$: $\hat{R}_{0i} = 1/\{1 + \exp(-\hat{\gamma}_0 - \hat{\gamma}_2 C_i)\}$
- $\hat{R}_{1i} - \hat{R}_{0i}$ を全員で平均
 - ▶ 回帰モデルにもとづく標準化 (regression standardization)

周辺効果、条件付き効果(リスク差)

■ $P(Y = 1|X, C)$

- ▶ Risk : 割合で推定
- ▶ R_{xc} : ロジスティック回帰モデルで推定

条件付き効果

層 k	C		X = 1		X = 0		リスク差	
	Sex	Age	Risk	R_{xc}	Risk	R_{xc}	Risk	R_{xc}
1	Male	Young	0.424	0.407	0.258	0.351	0.166	0.056
2	Male	Old	0.61	0.591	0.593	0.531	0.017	0.059
3	Female	Young	0.642	0.620	0.625	0.562	0.017	0.058
4	Female	Old	0.741	0.774	0.737	0.730	0.040	0.045
標準化			0.624	0.621	0.583	0.567	0.041	0.054

周辺効果

周辺効果、条件付き効果(オッズ比)

■ $P(Y = 1|X, C)$

- ▶ Risk : 割合で推定
- ▶ R_{xc} : ロジスティック回帰モデルで推定

条件付き効果

層 k	C		X = 1		X = 0		オッズ比	
	Sex	Age	Risk	R_{xc}	Risk	R_{xc}	Risk	R_{xc}
1	Male	Young	0.424	0.407	0.258	0.351	2.12	1.27
2	Male	Old	0.61	0.591	0.593	0.531	1.08	1.27
3	Female	Young	0.642	0.620	0.625	0.562	1.07	1.27
4	Female	Old	0.741	0.774	0.737	0.730	1.02	1.27
標準化			0.624	0.621	0.583	0.567	1.19	1.25

周辺効果

層別解析の回帰モデルによる近似

- 回帰モデルが近似である以上、層別解析と結果は異なる
 - ▶ 一致するのは「飽和モデル」の場合のみ
- ロジスティック回帰モデル
 - ▶ 「条件付きオッズ比が一定」という制約下での近似
 - 交互作用項を入れることで緩めることも可能
- オッズ比
 - ▶ ロジスティックモデルの結果
 - 条件付きオッズ比 = 1.27 (各層で一定)
 - 周辺オッズ比 = 1.25 (!)
 - ▶ 併合不能性 non-collapsibility
 - 周辺効果指標が条件付き効果指標の凸包に含まれない
 - バイアスではなく、オッズ比(やハザード比)という指標の性質

層別できるなら Part 2

- 各層の曝露確率 $P(X = 1|C)$ を求める
 - ▶ 傾向スコア
 - ▶ propensity score (PS)
- 3つの方法(Cで層別できたらどれもおなじ)
 - ▶ 層別・回帰調整 (stratification/adjustment)
 - ▶ マッチング (matching)
 - ▶ 逆確率重み付け (inverse probability weighting)

逆確率重み付け IPW inverse probability weighting

■ 個人 i が実際に受けた曝露を受ける確率の逆数

- ▶ 曝露群では $W_i = 1/PS_i$
- ▶ 非曝露群では $W_i = 1/(1 - PS_i)$

$$W_i = \frac{1}{P(X = X_i | C = C_i)}$$

- ▶ 重み付けた集団で、そのまま差を計算

■ 交絡変数が C としてすべて測定されているならば Rosenbaum, JASA 1987

- ▶ = 交絡変数 C で層別すれば $Y^x (x = 0, 1)$ と曝露 X が独立

$$E_{IPW}(Y | X = x) = E(Y^x)$$

IPWの計算

1/PS

C		X = 1			X = 0			$P(X=1 C)$ = PS
Sex	Age	N	$N_{Y=1}$	IPW	N	$N_{Y=1}$	IPW	
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		

$$\text{IPW} = X/PS + (1 - X)/(1 - PS)$$

- $X = 1 \Rightarrow 1/PS$
- $X = 0 \Rightarrow 1/(1 - PS)$

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

傾向スコアモデル

- モデルで $P(X = 1|C)$ を近似
 - ▶ 例: ロジスティック回帰モデル

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

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

C		X = 1			X = 0			$P(X=1 C)$ = PS
Sex	Age	N	$N_{Y=1}$	IPW	N	$N_{Y=1}$	IPW	
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		

傾向スコアモデルの当てはめ

Sex	Age	X	Y	N
1	0	1	1	577
1	1	1	1	1709
0	0	1	1	1707
0	1	1	1	2978
1	0	0	1	423
1	1	0	1	711
0	0	0	1	713
0	1	0	1	722
1	0	1	0	783
1	1	1	0	1091
0	0	1	0	953
0	1	1	0	1042
1	0	0	0	1217
1	1	0	0	489
0	0	0	0	427
0	1	0	0	258

```
proc logistic desc;  
    model x = sex age;  
    freq n;  
    output out = pred p = ps;  
  
run;
```



変数	係数 α	OR = $\exp(\alpha)$	95% CI	
1(切片)	0.7331			
Sex	-0.7956	0.451	0.421	0.484
Age	0.8012	2.228	2.078	2.389

傾向スコアモデルの予測値

C		X = 1			X = 0			PS
Sex	Age	N	$N_{Y=1}$	IPW	N	$N_{Y=1}$	IPW	
Male	Young	1360	577		1640	423		0.484
Male	Old	2800	1709		1200	711		0.677
Female	Young	2660	1707		1140	713		0.675
Female	Old	4020	2978		980	722		0.823
Total		10840	6971		4960	2569		

$$\widehat{PS}_c = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 c)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 c)}$$

傾向スコアモデルによるIPW

C		X = 1			X = 0			PS
Sex	Age	N	$N_{Y=1}$	IPW	N	$N_{Y=1}$	IPW	
Male	Young	1360	577	2.06	1640	423	1.94	0.484
Male	Old	2800	1709	1.48	1200	711	3.09	0.677
Female	Young	2660	1707	1.48	1140	713	3.08	0.675
Female	Old	4020	2978	1.22	980	722	5.64	0.823
Total		10840	6971		4960	2569		

$$1/\widehat{PS}_c = \left\{ \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 c)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 c)} \right\}^{-1}$$

$$1/(1 - \widehat{PS}_c) = \left\{ \frac{1}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 c)} \right\}^{-1}$$

26

傾向スコアモデルによるIPW

C		X = 1			X = 0			PS
Sex	Age	N_{IPW}	$N_{Y=1,IPW}$	IPW	N_{IPW}	$N_{Y=1,IPW}$	IPW	
Male	Young	2807.7	1191.2	2.06	3180.7	820.4	1.94	0.484
Male	Old	4137.7	2525.5	1.48	3711.9	2199.3	3.09	0.677
Female	Young	3937.9	2527.1	1.48	3512.9	2197.1	3.08	0.675
Female	Old	4886.8	3620.1	1.22	5525.2	4070.6	5.64	0.823
Total		15770.0	9863.8		15930.7	9287.4		
▶ $E_{IPW}(Y X=1) = 9863.8/15770.0 = 62.5\%$								
▶ $E_{IPW}(Y X=0) = 9287.4/15930.7 = 58.3\%$								

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

$$\log \frac{P(X = 1|C = c)}{1 - P(X = 1|C = c)} = \alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \dots$$

$$\longrightarrow \widehat{PS}_i = \widehat{P}(X = 1|C = C_i) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 C_{1i} + \hat{\alpha}_2 C_{2i} + \hat{\alpha}_3 C_{3i} + \dots)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 C_{1i} + \hat{\alpha}_2 C_{2i} + \hat{\alpha}_3 C_{3i} + \dots)}$$

- 3つの方法

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

周辺効果、条件付き効果

■ 傾向スコアを割合で推定した場合

条件付き効果

C		X = 1			X = 0			リスク差
Sex	Age	N_{IPW}	$N_{Y=1,IPW}$	$Risk_{IPW}$	N_{IPW}	$N_{Y=1,IPW}$	$Risk_{IPW}$	
Male	Young	3000	1272.8	0.424	3000	773.8	0.258	0.166
Male	Old	4000	2441.4	0.610	4000	2370.0	0.593	0.017
Female	Young	3800	2438.6	0.642	3800	2376.6	0.625	0.017
Female	Old	5000	3704.0	0.741	5000	3683.7	0.737	0.004
Total		15800	9856.8	0.625	15800	9204.1	0.583	0.041

周辺効果

周辺効果、条件付き効果

■ 傾向スコアをモデルで推定した場合

条件付き効果

C		X = 1			X = 0			リスク差
Sex	Age	N_{IPW}	$N_{Y=1,IPW}$	$Risk_{IPW}$	N_{IPW}	$N_{Y=1,IPW}$	$Risk_{IPW}$	
Male	Young	2807.7	1191.2	0.424	3180.7	820.4	0.258	0.166
Male	Old	4137.7	2525.5	0.610	3711.9	2199.3	0.593	0.017
Female	Young	3937.9	2527.1	0.642	3512.9	2197.1	0.625	0.017
Female	Old	4886.8	3620.1	0.741	5525.2	4070.6	0.737	0.004
Total		15770.0	9863.8	0.625	15930.7	9287.4	0.583	0.042

周辺効果

傾向スコアのモデルによる近似

■ 周辺効果

- ▶ モデルが近似である以上、層別解析と結果は異なる
 - 一致するのは「飽和モデル」の場合のみ
 - 交互作用項、高次の項を入れることで近似精度を高められる
- ▶ 層別（調整）、マッチング、IPW いずれも同様
 - ただし、**標的集団をどこに置いているかを常に意識する必要あり**（☞ 前回）

■ 条件付き効果

- ▶ **Cで層別するなら傾向スコアモデルの近似精度によらずおなじ値（！）**
 - 傾向スコアは C の分布をバランスさせるだけ
 - ただ、Cで層別できるなら傾向スコア「モデル」は使わないわけで...
 - 「傾向スコアモデル ⇒ 回帰モデル」だと各モデルの近似精度により異なる値
- ▶ 層別（調整）、マッチング、IPW いずれも同様

層別しきれないほどのデータの解析例

傾向スコアによる「調整」

- 層別／回帰調整 (stratification/conditioning/adjustment)
 - ▶ PS(C) が同じ値のサブグループで効果推定、併合
 - 条件付きランダム化の仮定 + バランス性をそのまま利用
- マッチング (matching)
 - ▶ PS(C) が同じ値のサブグループからランダム抽出 ⇒ 周辺期待値の推定
 - 条件付きランダム化の仮定 + バランス性にもとづく確率サンプリング
- 逆確率重み付け (inverse-probability weighting)
 - ▶ $1/PS(C)$ or $1/(1 - PS(C))$ で重み付け ⇒ 周辺で効果推定
 - 重み付けた分布では X と C が独立 (statistically exogenous)
 - 条件付きランダム化の仮定と併せて、 Y^* と X が独立 (causally exogenous)
- いずれも曝露 X 間で共変量 C をバランスさせる方法

傾向スコア推定、実用上の注意点1

■ 傾向スコアに含めるべき共変量

▶ つまり、層別すべき共変量

1. 基本的には交絡変数 C

- ▶ 曝露ともアウトカムとも関連するもの
- ▶ 曝露の結果は含めない
- ▶ DAG (directed acyclic graph) でいうと「バックドア・パスを塞ぐ変数のセット」

2. 曝露に強く関連する変数は含めない

- ▶ アウトカムに関連しない: 操作変数 (instrumental variable)
- ▶ アウトカムに関連しても、含むと小標本バイアスを増幅することもある

3. アウトカムに影響する変数は含める

- ▶ 曝露と関連がなくても含む方がよい

VanderWeele, *European Journal of Epi* 2019

傾向スコア推定、実用上の注意点2

■ モデルの当てはまり ≠ モデルの判別力

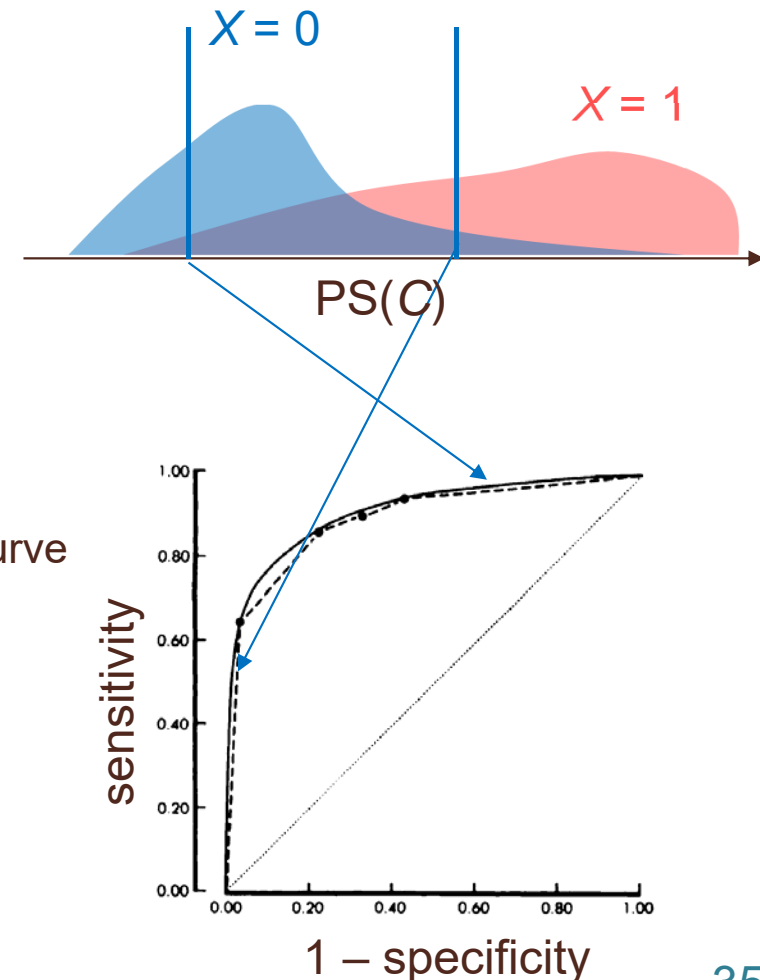
■ C-統計量

- ▶ $P(PS_i > PS_j | X_i = 1, X_j = 0)$
- ▶ 曝露あり($X_i = 1$ の i) の PS の方が
曝露なし($X_j = 0$ の j) の PS より
大きくなる割合

■ ROC曲線下面積(AUC)

Area under the receiver operating characteristic curve

- ▶ すべてのカットオフ c に対して
 - 感度 = $P(PS > c | X = 1)$ を
 - $1 - \text{特異度} = P(PS > c | X = 0)$
- に対してプロット



モデルの当てはまり

- 傾向スコア $P(X = 1|C)$ の近似をする上で重要
 - ▶ モデル: C でもしも層別できた場合の確率を統一的に表現
 - ▶ 交絡変数 C は既に決まっている
 - 「変数の出し入れ」と「当てはまり(フィット)」は別次元
 - C に含まれる変数同士の交互作用や変数変換を検討すべき
 - 例: 尤度比検定やHosmer-Lemeshow統計量など
- C -統計量
 - ▶ グループ間の傾向スコアの分離度合 (discrimination)
 - モデルの当てはまりでなく、推定された傾向スコアの特性をあらわす
 - ▶ C -統計量が大きい \Rightarrow 傾向スコアの重なりが小さい
 - 傾向スコアによる調整解析の不安定さ(統計的効率の低さ)を示唆
 - 交絡変数分布が群間で大きく異なるので、交絡調整の必要性を示唆

実例 : Rotterdam Tumor Bank

■ rott

- ▶ 乳がん初発の女性 ($n = 2982$) の追跡データ
- ▶ R package **AF** より入手可能

■ アウトカム T

- ▶ 無再発生存期間

■ 曝露 X

- ▶ 化学療法**なし** ($= 1$; $n = 580$)、あり ($= 0$; $n = 2402$)
- ▶ なし: 2402名、あり: 580名

■ 共変量 C

- ▶ 年齢、閉経状況(あり/なし)、腫瘍径 ($\leq 20\text{mm}$, $>20-50\text{mm}$, $>50\text{ mm}$)、腫瘍グレード(2/3)、リンパ節転移数(0~34)、プロゲステロン受容体、エストロゲン受容体

変数名

- Time: RF
- Event: RFI
- Exposure: NO_CHEMO
- Covariates:
 - ▶ AGE
 - ▶ MENO (characteristic 0/1)
 - ▶ SIZE (characteristic 3-category)
 - ▶ GRADE (characteristic 2/3)
 - ▶ NODES (0-34) → EX_NODES: $\exp(-0.12 \cdot \text{NODES})$ に変換
 - ▶ PR
 - ▶ ER

傾向スコア推定

```
data rott;  
    set rott;  
    ex_NODES = exp(-0.12*NODES);  
  
run;  
proc logistic data = rott desc;  
    class MENO SIZE GRADE;  
    model NO_CHEMO = AGE MENO SIZE GRADE ex_NODES PR ER;  
    output out = rott2 p = PS;  
  
run;  
data rott2;  
    set rott2;  
    CHEMO = (1 - NO_CHEMO);  
    W = NO_CHEMO/PS + (1 - NO_CHEMO)/(1 - PS);  
  
run;  
proc univariate data=rott2;  
    var PS W; class NO_CHEMO; histogram;  
  
run;
```

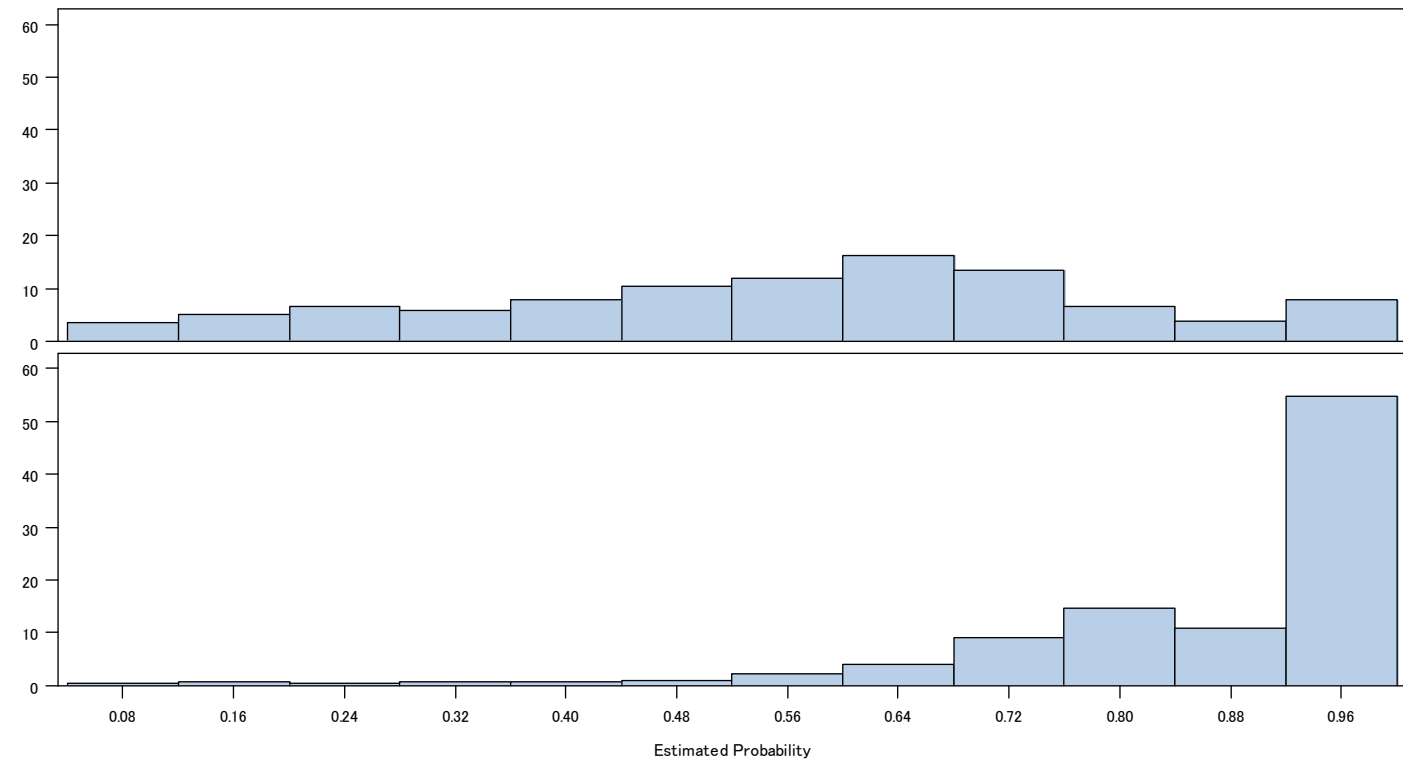
IPW用の重み

傾向スコア推定値の分布

■ $C = 0.875$

非曝露群
(Chemo)

曝露群
(No-chemo)



PSマッチング(1:1マッチング)

```
ods graphics on;  
proc psmatch data = rott region = treated;  
  class NO_CHEMO MENO SIZE GRADE;  
  psmodel NO_CHEMO(Treated='0') =  
    AGE MENO SIZE GRADE ex_NODES PR ER;  
  match method = greedy stat=lps caliper=0.2;  
  assess lps var = (AGE MENO GRADE ex_NODES PR ER)  
    / weight = none plots = all;  
  output out(obs = match) = rott3 matchid=_MatchID;  
run;
```

曝露群 $n = 580$ 、非曝露群 $n = 2402$
⇒ 495ペア

Kaplan–Meier推定値

```
proc lifetest data = rott outsurv = s_crude;  
    time rf * rfi (0);  
    strata NO_CHEMO;
```

未調整の生存曲線

```
run;
```

```
proc lifetest data = rott2 outsurv = s;  
    time rf * rfi (0);  
    strata NO_CHEMO;  
    weight W;
```

IPW生存曲線

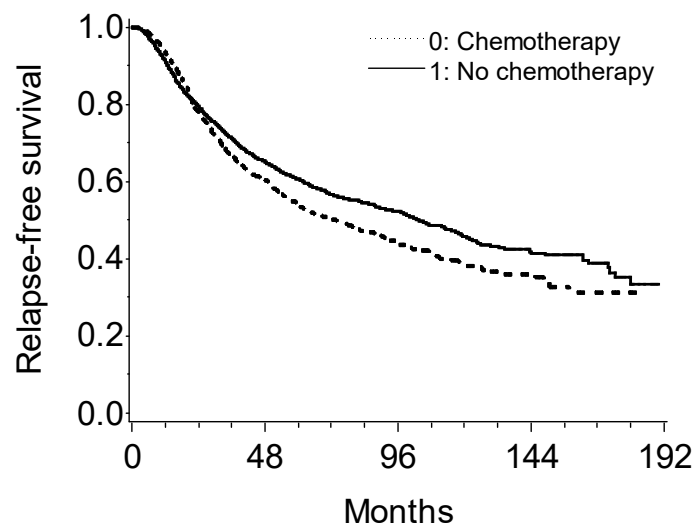
```
run;
```

```
proc lifetest data = rott3 outsurv = s_matched;  
    time rf * rfi (0);  
    strata NO_CHEMO;
```

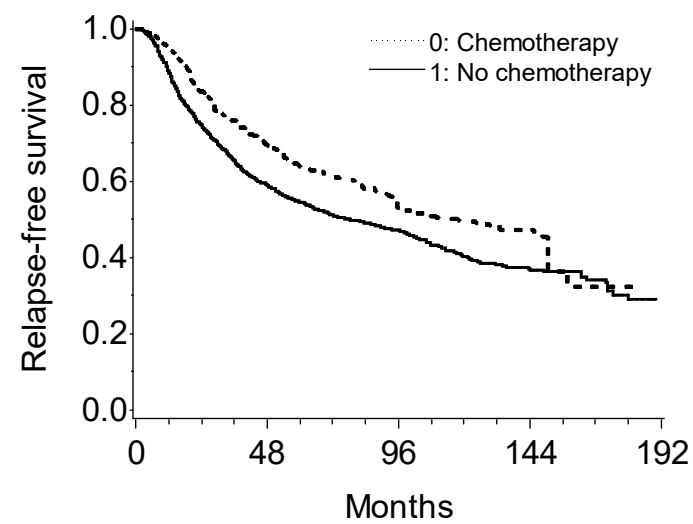
PSマッチド生存曲線

```
run;
```

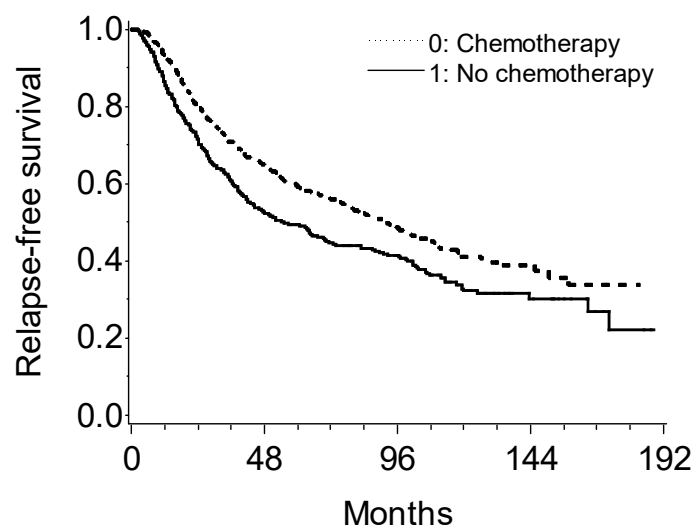
A) Original cohort



B) IPW cohort



C) Propensity-matched (1 to 1) cohort



傾向スコアによる
交絡調整により
効果推定値が逆転

Cox回帰のオプション

- オリジナル・コホート
 - ▶ 未調整モデル(バイアスあり)
 - ▶ 多変数調整モデル(C 条件付きモデル)
 - ▶ PS調整モデル(PS 条件付きモデル)
- PSマッチド・コホート
 - ▶ 未調整モデル(周辺モデル)
 - ▶ マッチド・ペアの層別モデル(PS 条件付きモデル)
 - ▶ 多変数調整モデル(C 条件付きモデル)
- IPWコホート
 - ▶ 未調整モデル(周辺モデル)
 - ▶ 多変数調整モデル(C 条件付きモデル)

オリジナル・コホート

* Unadjusted (biased) HR;

```
proc phreg data = rott;  
    model rf * rfi(0) = NO_CHEMO / r1;  
run;
```

* PS-adjusted (conditional) HR;

```
proc phreg data = rott;  
    model rf * rfi(0) = NO_CHEMO PS / r1;  
run;
```

* Multivariable-adjusted (conditional) HR;

```
proc phreg data = rott;  
    class MENO SIZE GRADE;  
    model rf * rfi(0) = NO_CHEMO AGE MENO SIZE GRADE  
        ex_NODES PR ER AGE*AGE ex_NODES*ex_NODES / r1;  
run;
```

マッチド・コホート

* Conditional (stratified) HR;

```
proc phreg data = rott3;  
    model rf * rfi(0) = NO_CHEMO/ r1 ties=exact;  
    strata pair; マッチド・ペアで層別  
run;
```

* Conditional HR;

```
proc phreg data = rott3;  
    class MENO SIZE GRADE;  
    model rf * rfi(0) = NO_CHEMO AGE MENO SIZE GRADE ex_NODES PR  
        ER AGE*AGE ex_NODES*ex_NODES / r1 ties=exact;  
run;
```

* Marginal HR in unexposed population;

```
proc phreg data = rott3 covs(aggregate); マッチド・ペア内の相関を考慮  
    model rf * rfi(0) = NO_CHEMO/ r1 ties=exact;  
    id pair;  
run;
```

IPWコホート

* IPW (marginal HR);

```
proc phreg data = rott2 covs(aggregate);  
    model rf * rfi(0) = NO_CHEMO / r1;  
    weight W;  
    id pid;
```

run;

* IPW + Multivariable-adjusted (conditional HR);

```
proc phreg data = rott2 covs(aggregate);  
    class MENO SIZE GRADE;  
    model rf * rfi(0) = NO_CHEMO AGE MENO SIZE GRADE  
        ex_NODES PR ER AGE*AGE ex_NODES*ex_NODES / r1;  
    weight W;  
    id pid;
```

run;

重み付けた場合は必ず
ロバスト分散

重み付けた場合は必ず
ロバスト分散

結果：ハザード比の推定値

Model	Analysis set	Target HR	HR	95% CI	
Unadjusted Cox model	Original cohort	Biased	0.84	0.75	0.95
Multivariable Cox model ^d	Original cohort	C-conditional	1.48	1.27	1.71
PS-adjusted Cox model	Original cohort	PS-conditional	1.34	1.16	1.55
Stratified Cox model ^a	Matched cohort	PS-conditional	1.47	1.18	1.83
Unstratified Cox model ^b	Matched cohort	Marginal ^e	1.33	1.13	1.57
Multivariable Cox model	Matched cohort	C-conditional	1.59	1.31	1.93
IPW Cox model ^c	IPW cohort	Marginal ^f	1.32	1.07	1.62
IPW multivariable Cox model ^{c,d}	IPW cohort	C-conditional	1.58	1.28	1.96

CI, confidence interval; IPW, inverse-probability weighted; HR, hazard ratio.

^a Stratified on matched pairs.

^b Using a robust variance estimator aggregating residuals within pairs.

^c Using a robust variance estimator aggregating residuals within an individual woman.

^d Adjusted for age at surgery, menopausal status, tumor size, tumor grade, progesterone receptors, oestrogen receptors, and $\exp(-0.12 \times \text{the number of positive lymph nodes})$. Age and the transformed number of nodes were included by linear and quadratic terms.

^e Target population is the matched part of unexposed population (treated with chemotherapy).

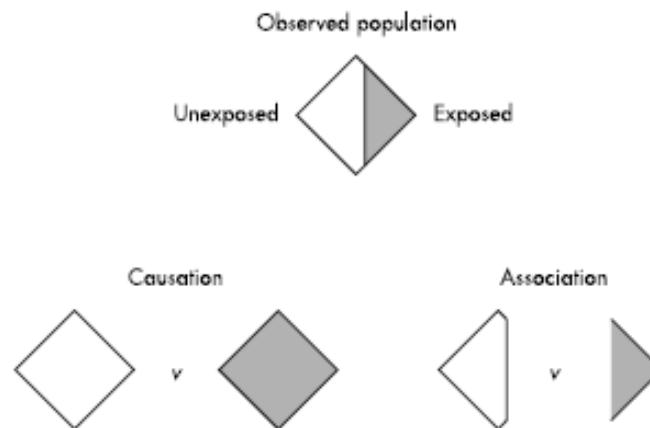
^f Target population is total (unexposed and exposed) population.

解析結果のまとめ

- 傾向スコアの交絡調整は生存時間アウトカムでも有効
 - ▶ マッチング、IPW、調整
- 周辺ハザード比 vs. 条件付きハザード比
 - ▶ 調整方法によって異なる指標(estimands)が標的
 - 一般に、周辺ハザード比は条件付きハザード比より1に近い
 - オッズ比でも同様の現象([non-collapsibility](#))
 - ▶ 注: 周辺モデルと条件付きモデルは同時に成り立たない
 - 比例ハザード性の成り立たない状況でのハザード比の解釈?

生存時間アウトカム

■ 平均因果効果？



■ 特定時点の生存確率 $S^{x=1}(t)$ vs. $S^{x=0}(t)$

- ▶ IPW 生存関数推定量
- ▶ Cox回帰モデルによる標準化推定量
- ▶ 二重ロバスト生存関数推定量

Bai et al., *Biometrics* 2013
Shinozaki & Nojima, *Epidemiology* 2019

集団全体(曝露群 + 非曝露群)を標的とした 二重ロバスト推定

平均因果効果の推定方法

- モデルの要らない交絡調整
 - ▶ 層別解析
 - ▶ 層別解析の2つの見方
- 交絡調整にモデルが必要な状況
 - ▶ アウトカム回帰モデル
 - ▶ 傾向スコアモデル
- モデルを使わざるを得ないなら、どちらをつかうか？
 - ▶ 両方つかう二重ロバスト推定
 - ▶ モデルに必要な仮定は？

AIPW推定量

- *Augmented* IPW (+ reweighted)
- IPW (+ reweight) 推定量にアウトカム回帰を組み込んだ局所セミパラメトリック有効推定量
 - ▶ $E(Y | X = x, C = c)$ のモデル
 - ▶ $P(X = 1 | C = c)$ のモデル
- いずれかのモデルが正しければ平均因果効果の一致推定量
 - ▶ 二重ロバスト性 (double robustness)

IPW推定量

■ 傾向スコアモデル

$$\log \frac{P(X=1|C=c)}{1-P(X=1|C=c)} = \alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \dots$$

▶ 個人 i に対して予測値 $\hat{\pi}_i$ を得る

- $\hat{\pi}_i = \{1 + \exp(-\hat{\alpha}_0 - \hat{\alpha}_1 C_{1i} - \hat{\alpha}_2 C_{2i} - \hat{\alpha}_3 C_{3i} - \dots)\}^{-1}$

■ 重み付け平均

$$\hat{\mu}_{0,IPW} = \frac{\sum_i \frac{1}{1-\hat{\pi}_i} (1-X_i) Y_i}{\sum_i \frac{1}{1-\hat{\pi}_i} (1-X_i)}$$

$$\hat{\mu}_{1,IPW} = \frac{\sum_i \frac{1}{\hat{\pi}_i} X_i Y_i}{\sum_i \frac{1}{\hat{\pi}_i} X_i}$$

回帰標準化推定量

■ アウトカム回帰モデル

$$\log \frac{E(Y|X=1, C=c)}{1 - E(Y|X=1, C=c)} = \beta_{10} + \beta_{11}c_1 + \beta_{12}c_2 + \beta_{13}c_3 + \dots$$

$$\log \frac{E(Y|X=0, C=c)}{1 - E(Y|X=0, C=c)} = \beta_{00} + \beta_{01}c_1 + \beta_{02}c_2 + \beta_{03}c_3 + \dots$$

▶ 個人 i に対して予測値 $(\hat{m}_{0,i}, \hat{m}_{1,i})$ を得る

- $\hat{m}_{0,i} = \{1 + \exp(-\hat{\beta}_{00} - \hat{\beta}_{01}C_{1i} - \hat{\beta}_{02}C_{2i} - \hat{\beta}_{03}C_{3i} - \dots)\}^{-1}$
- $\hat{m}_{1,i} = \{1 + \exp(-\hat{\beta}_{10} - \hat{\beta}_{11}C_{1i} - \hat{\beta}_{12}C_{2i} - \hat{\beta}_{13}C_{3i} - \dots)\}^{-1}$

■ 平均化

$$\hat{\mu}_{0,\text{reg}} = \hat{E}(\hat{m}_0)$$

$$\hat{\mu}_{1,\text{reg}} = \hat{E}(\hat{m}_1)$$

AIPW(二重ロバスト)推定量

■ 傾向スコアモデル

$$\log \frac{P(X=1|C=c)}{1-P(X=1|C=c)} = \alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \dots$$

$$\hat{\pi}_i = \{1 + \exp(-\hat{\alpha}_0 - \hat{\alpha}_1 C_{1i} - \hat{\alpha}_2 C_{2i} - \hat{\alpha}_3 C_{3i} - \dots)\}^{-1}$$

■ アウトカム回帰モデル

$$\log \frac{E(Y|X=1, C=c)}{1-E(Y|X=1, C=c)} = \beta_{10} + \beta_{11} C_1 + \beta_{12} C_2 + \beta_{13} C_3 + \dots$$

$$\log \frac{E(Y|X=0, C=c)}{1-E(Y|X=0, C=c)} = \beta_{00} + \beta_{01} C_1 + \beta_{02} C_2 + \beta_{03} C_3 + \dots$$

$$\hat{m}_{0,i} = \{1 + \exp(-\hat{\beta}_{00} - \hat{\beta}_{01} C_{1i} - \hat{\beta}_{02} C_{2i} - \hat{\beta}_{03} C_{3i} - \dots)\}^{-1}$$

$$\hat{m}_{1,i} = \{1 + \exp(-\hat{\beta}_{10} - \hat{\beta}_{11} C_{1i} - \hat{\beta}_{12} C_{2i} - \hat{\beta}_{13} C_{3i} - \dots)\}^{-1}$$

AIPW(二重ロバスト)推定量

■ IPW平均 + 回帰標準化 – IPW回帰標準化

$$\hat{\mu}_{1,AIPW} = n^{-1} \sum_i \left[\frac{X_i}{\hat{\pi}_i} Y_i + \hat{m}_{1,i} - \frac{X_i}{\hat{\pi}_i} \hat{m}_{1,i} \right]$$

$$\hat{\mu}_{0,AIPW} = n^{-1} \sum_i \left[\frac{1-X_i}{1-\hat{\pi}_i} Y_i + \hat{m}_{0,i} - \frac{1-X_i}{1-\hat{\pi}_i} \hat{m}_{0,i} \right]$$

■ 二重ロバスト性(証明は略記)

- ▶ $E(X | C) = \pi(C)$ と正しく特定
 - $E(Xm_1/\pi) = E\{E(Xm_1/\pi | C)\} = E\{(m_1/\pi)E(X|C)\} = E(m_1)$
- ▶ $E(Y | X = x, C) = m_x(C)$ と正しく特定
 - $E(XY/\pi) = E\{E(XY|C)/\pi\} = E[E\{X E(Y|X, C) | C\}/\pi]$
 $= E[\{1 \cdot m_1 \cdot \Pr(X=1|C) + 0 \cdot m_0 \cdot \Pr(X=1|C)\}/\pi] = E\{m_1 \cdot E(X|C)/\pi\}$

標的集団を変えた二重ロバスト推定

■ IPW回帰推定量

1. アウトカム回帰モデル $m_x(\beta)$ を 重み W で重み付け推定

- 標的集団 = 集団全体 $\Rightarrow W = \text{IPW}$
- 標的集団 = 曝露群 $\Rightarrow W = \text{IPW} \times P(X = 1|C)$
- 標的集団 = 非曝露群 $\Rightarrow W = \text{IPW} \times P(X = 0|C)$

2. 重み付けずに回帰標準化

- $W = \text{IPW}$
 \Rightarrow 集団全体 $(1, \dots, N)$ で予測値 $\hat{E}(Y|X = x, C_i)$ を平均
- $W = \text{IPW} \times P(X = 1|C)$
 \Rightarrow 曝露群 $(1, \dots, n_1)$ で予測値 $\hat{E}(Y|X = x, C_i)$ を平均
- $W = \text{IPW} \times P(X = 0|C)$
 \Rightarrow 非曝露群 $(1, \dots, n_0)$ で予測値 $\hat{E}(Y|X = x, C_i)$ を平均

Disc. in Kang & Schafer, *Stat Sci* 2007

いくつかある二重ロバスト推定量

■ 効果の種類、標的集団に応じた推定量

▶ 周辺効果、条件付き効果

- 集団全体、曝露群($X = 1$)、非曝露群($X = 0$)
- マッチングされた集団
- $C = c$ のサブグループ

▶ 推定手法

- AIPW推定
- G-推定
- TMLE (targeted minimum loss-based/maximum likelihood estimation)

■ データの特性に応じた推定量

- ▶ 打ち切りを伴う生存時間データ
- ▶ 繰り返し／継続される治療データ

実例：NHEFSデータ

- NHANES I Epidemiologic Follow-Up Study
 - ▶ National Health and Nutrition Examination Survey
 - www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm
 - Hernan & Robins, *Causal Inference*, 2018
- ベースラインデータ(1971年)
- 10年後のフォローアップデータ(1982年)
 - ▶ “For the study, medical and behavioral information were collected during an initial physical examination, and again at follow-up interviews approximately one decade later”

<https://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp>[FIND DATA](#) [SEARCH/COMPARE VARIABLES](#) [DATA-RELATED PUBLICATIONS](#) [RESOURCES FOR STUDENTS](#) [HELP](#)

Filters

Subject ▾

[demographic characteristics \(43\)](#)
[health status \(42\)](#)
[social indicators \(42\)](#)

Search Results

45 results.

NHEFS

GO

VIEW ALL

[search tips ▾](#)[INFORMATION FOR:](#) [Prospective Students](#) [Current Students](#) [Alumni](#) [Faculty & Staff](#) [Friends & Supporters](#)

Miguel Hernan

<https://www.hsph.harvard.edu/miguel-hernan/>[Home](#) > [Miguel Hernan](#) > Causal Inference Book

MIGUEL HERNAN

Search this section

[Home](#)[Research ▾](#)[Teaching ▾](#)[Causal Inference Book](#)

Causal Inference Book

My colleague Jamie Robins and I are working on a book that provides a cohesive introduction to causal inference. Much of this material is currently scattered across journals in several different fields. The book will be of interest to anyone interested in causal inference, e.g., epidemiologists, sociologists, political scientists, computer scientists... The book is divided into three parts: causal inference with models, causal inference from complex longitudinal data, and causal inference from experimental data.

主な変数コード

変数名	コーディング
seqn	UNIQUE PERSONAL IDENTIFIER
sex	0: MALE 1: FEMALE
race	0: WHITE 1: BLACK OR OTHER IN 1971
age	AGE IN 1971
education	AMOUNT OF EDUCATION BY 1971: 1: 8TH GRADE OR LESS, 2: HS DROPOUT, 3: HS, 4:COLLEGE DROPOUT, 5: COLLEGE OR MORE
smokeintensity	NUMBER OF CIGARETTES SMOKED PER DAY IN 1971
smokeyrs	YEARS OF SMOKING
qsmk	QUIT SMOKING BETWEEN 1ST QUESTIONNAIRE AND 1982, 1:YES, 0:NO
wt82_71	INCREASE IN NUMBER OF CIGARETTES/DAY BETWEEN 1971 and 1982
death	DEATH BY 1992, 1:YES, 0:NO
exercise	IN RECREATION, HOW MUCH EXERCISE? IN 1971, 0:much exercise,1:moderate exercise,2:little or no exercise
active	IN YOUR USUAL DAY, HOW ACTIVE ARE YOU? IN 1971, 0:very active, 1:moderately active, 2:inactive
wt71	WEIGHT IN KILOGRAMS IN 1971

解析方針

- 曝露 X
 - ▶ 禁煙 `qsmk` (1:止めた、0:喫煙)
- アウトカム Y
 - ▶ 10年間の死亡 `death` (1:死亡、0:生存)
- 交絡変数
 - ▶ `sex`、`race`、`age` (以上は連続値:2乗項まで調整)、`education`、`smokeintensity`、`smokeyrs`、`exercise`、`active`、`wt71`
- 集団全体での効果 $E(Y^{x=1}) - E(Y^{x=0})$ を推定
 - ▶ IPW
 - ▶ アウトカム回帰
 - ▶ 二重ロバスト

} SAS/PROC CAUSALTRT

IPW推定

```
ods graphics on;
proc causaltrt data = nhefs_nmv desc;
  class exercise active education /desc;
  psmodel qsmk(ref="0") = sex race age age*age education
                        smokeintensity mokeintensity*smokeintensity
                        smokeyrs smokeyrs*smokeyrs
                        exercise active wt71 wt71*wt71
                        / plots = (psdist weightdist
                                   pscovden(effects(age smokeyrs)) );
  model death / dist = bin;
run;
```

Analysis of Causal Effect							
Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > Z
POM	1	0.1884	0.0189	0.1514	0.2254	9.98	<.0001
POM	0	0.1839	0.0112	0.1619	0.2058	16.41	<.0001
ATE		0.004563	0.0207	-0.03603	0.04515	0.22	0.8256

従来のプログラムでは...

```
proc logistic data = nhefs_nmv desc;  
  class exercise active education;  
  model qsmk = sex race age age*age education  
             smokeintensity smokeintensity*smokeintensity  
             smokeyrs smokeyrs*smokeyrs  
             exercise active wt71 wt71*wt71/ lackfit;  
  output out = nhefs_ps p = ps;  
run;  
  
data nhefs_ps;  
  set nhefs_ps;  
  w = qsmk/ps + (1 - qsmk)/(1 - ps);  
run;  
  
proc genmod data = nhefs_ps;  
  class seqn;  
  model death = qsmk /dist = normal;  
  weight w;  
  repeated sub = seqn;  
run;
```

差を得るため DIST = NORMAL
(推定値はIPWRと一致、分布の誤特定下でも分散は妥当)
比が欲しければ DIST = POISSON

ここでのロバスト分散は傾向スコアモデルの推定を考慮していないため、「近似的」な漸近分散推定量
ただし、バイアスは保守的な(過大評価する)方向

比が欲しかったら

■ デルタ法

- ▶ 現段階の CAUSALTRT ではデフォルトで出力されない

田栗, 2017 “SASによる因果推論: CAUSALTRTプロシージャの紹介”

■ SASではブートストラップ標本を出力可

- ▶ デフォルトでは変数 `_TRTPOM_`、`_CNTPOM_`、`_ATE_` に

VIEWTABLE: Work.Bootipws			
	Bootstrap sample estimate for Control Potential Outcome Mean	Bootstrap sample estimate for Treatment Potential Outcome Mean	Bootstrap sample estimate for Average Treatment Effect Treated
1	0.1504778876	0.1106642159	-0.079054076
2	0.1508252221	0.1367215302	-0.048844701
3	0.1534402572	0.1375840047	-0.046660066
4	0.1538615485	0.1393440369	-0.04487216
5	0.1551464369	0.1404315419	-0.044761185
6	0.1556760707	0.1406643752	-0.044039853
7	0.156204791	0.1431206381	-0.043185577
8	0.1564566971	0.1449251657	-0.042891978
9	0.1575005064	0.1461100067	-0.04195715

回帰標準化

```
ods graphics on;  
proc causaltrt data = nhfs_nmv poutcomemod desc;  
  class exercise active education /desc;  
  psmodel qsmk(ref="0") ;  
  model death = sex race age age*age education  
               smokeintensity  
               smokeintensity*smokeintensity  
               smokeyrs smokeyrs*smokeyrs  
               exercise active wt71 wt71*wt71  
               / dist = bin;  
  bootstrap bootci(all) plot seed=1234;  
run;
```

回帰標準化

Analysis of Causal Effect								
Parameter	Treatment Level	Estimate	Robust Std Err	Bootstrap Std Err	Wald 95% Confidence Limits		Bootstrap Wald 95% Confidence Limits	
POM	1	0.1799	0.0171	0.0179	0.1465	0.2134	0.1448	0.2151
POM	0	0.1851	0.0112	0.0114	0.1631	0.2071	0.1628	0.2075
ATE		-0.00516	0.0191	0.0199	-0.04258	0.03226	-0.04418	0.03385

Bootstrap Percentile 95% Confidence Limits		Bootstrap Bias Corrected 95% Confidence Limits		Z	Pr > Z
0.1468	0.2177	0.1454	0.2143	10.55	<.0001
0.1638	0.2093	0.1629	0.2080	16.46	<.0001
-0.04216	0.03642	-0.04095	0.03758	-0.27	0.7868

従来のプログラムでは...

```
data nhefs_nmv;
    set nhefs_nmv;
    wt1 = qsmk;
    wt0 = 1 - qsmk;

run;
proc genmod data = nhefs_nmv desc;
    class exercise active education;
    model death = sex race age age*age education
                smokeintensity smokeintensity*smokeintensity smokeyrs
                smokeyrs*smokeyrs exercise active wt71 wt71*wt71/ dist = bin covb;
    weight wt1; qsmk = 1 のみ当てはめ  $\Rightarrow E(Y^{x=1} | C)$  の予測値を全員に得る
    output out = pred3 p = pdeath1;

run;
proc genmod data = pred3 desc;
    class exercise active education;
    model death = sex race age age*age education
                smokeintensity smokeintensity*smokeintensity smokeyrs
                smokeyrs*smokeyrs exercise active wt71 wt71*wt71/ dist = bin covb;
    weight wt0; qsmk = 0 のみ当てはめ  $\Rightarrow E(Y^{x=0} | C)$  の予測値を全員に得る
    output out = pred3 p = pdeath0;

run;
proc means data = pred3;
    var pdeath1 pdeath0;

run;
```

分散共分散行列によるデルタ法
またはブートストラップ法で分散推定

AIPW(二重ロバスト)推定量

```
ods graphics on;
proc causaltrt data = nhfs_nmv ppsmodel poutcomemod desc;
  class exercise active education /desc;
  psmodel qsmk(ref="0") = sex race age age*age education
                        smokeintensity smokeintensity*smokeintensity
                        smokeyrs smokeyrs*smokeyrs
                        exercise active wt71 wt71*wt71;
  model death = sex race age age*age education
              smokeintensity smokeintensity*smokeintensity
              smokeyrs smokeyrs*smokeyrs
              exercise active wt71 wt71*wt71/ dist = bin;
  bootstrap bootci(all) plot seed=1234;
run;
```

AIPW(二重ロバスト)推定量

Analysis of Causal Effect						
Parameter	Treatment Level	Estimate	Robust Std Err	Bootstrap Std Err	Wald 95% Confidence Limits	
POM	1	0.1856	0.0189	0.0197	0.1486	0.2226
POM	0	0.1857	0.0112	0.0114	0.1637	0.2078
ATE		-0.00014	0.0208	0.0215	-0.04081	0.04053

Analysis of Causal Effect							
Bootstrap Wald 95% Confidence Limits		Bootstrap Percentile 95% Confidence Limits		Bootstrap Bias Corrected 95% Confidence Limits		Z	Pr > Z
0.1470	0.2242	0.1492	0.2261	0.1477	0.2236	9.83	<.0001
0.1634	0.2081	0.1640	0.2094	0.1632	0.2084	16.54	<.0001
-0.04232	0.04204	-0.03903	0.04570	-0.03900	0.04574	-0.01	0.9946

モデルに含める変数が異なると

- 「二重ロバスト性」は担保されない

- ▶ 十分な C を条件づけた回帰
- ▶ 十分な C を条件づけた傾向スコア

のモデル特定に対する性質

- モデルに含める変数に対する「ロバスト性」ではない

Keil et al., *Am J Epidemiol* 2018 (Letter)

- アドホックな組み合わせ (やらないほうがいいです)

- ▶ 交絡変数の一部を傾向スコアモデルでバランスさせておいて
- ▶ バランスされていない一部を回帰モデルで調整

Shinozaki and Nojima, *Epidemiology* 2019

まとめ：交絡調整方法のチョイス

- 交絡変数 C があるなら、層別解析
 - ▶ 交絡変数が C として十分測定されている (= 条件付きランダム化) 仮定
- 層別解析ができない場合
 - ▶ 層内に十分な人数がない、個人ごとに C の値が異なる
- 統計モデル：層別解析の近似手段
 - ▶ 回帰モデル ☞ $E(Y|X, C)$
 - ▶ 傾向スコアモデル ☞ $P(X = 1|C)$
 - 個人ごとに異なる値が得られるが...
 - あくまで交絡変数 C の層ごとの値が欲しい
 - ▶ どちらかを使う
 - ▶ 両方つかう ⇒ 二重ロバスト推定

まとめ：各手法に必要な条件

■ 必要な仮定

1. C で層別できたら X がランダムに決まる ➡ 層別解析に共通

2a. $E(Y|X, C)$ を正しく特定 ➡ 回帰モデル

2b. $P(X = 1|C)$ を正しく特定 ➡ 傾向スコアモデル

2c. $E(Y|X, C)$ または $P(X = 1|C)$ (または両方) を正しく特定
➡ AIPW(二重ロバスト)推定量

Take-home message(再掲)

- 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
 - ▶ 傾向スコアモデル
- モデルをつかう目的
 - ▶ 層別できないほど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

Why propensity score?

- セミパラメトリックモデルの推測 semiparametric inference
 - ▶ 「興味のあるパラメータ」以外の仮定を極力排したモデル
 - ▶ 二重ロバスト推定 (doubly robust estimation)
- 時間依存性交絡の調整 adjustment for time-varying confounding
 - ▶ 治療が複数時点行われる状況 (X_1, X_2, \dots, X_k)
 - ▶ ある変数が
 - a. 後ろの治療の交絡変数であり、かつ
 - b. 前の治療の影響を受ける場合、その変数を調整してもしなくてもバイアス
- 直接効果・間接効果 mediation analysis
 - ▶ 治療 X 後の中間変数 M を介した効果・介さない効果
- 動的な治療レジメン dynamic treatment regimens
 - ▶ 治療 X_1 開始後の共変量 L_1 の値に応じて次の治療 X_2 を決定