北大SPH・統計的因果推論と臨床疫学④

周辺構造モデルと標準化

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 周辺構造モデル
  - 平均因果効果をモデル化
  - IPW法による擬似集団でそのままあてはめ
- 標準化
  - 平均因果効果を求めるもう一つのアプローチ
- IPW法か標準化か
  - 曝露確率のモデルかアウトカムのモデルか
  - 二重ロバスト推定量

2

潜在アウトカム potential outcome

3

- $Y^{a=1}$
  - 曝露$a=1$を受けた場合のアウトカム
- $Y^{a=0}$
  - 曝露$a=0$を受けた場合のアウトカム
- アウトカムも2値(0,1)の場合

	$Y^{a=1}$	$Y^{a=0}$
Doomed	1	1
Helped	1	0
Hurt	0	1
Immune	0	0

3

個人での因果効果

4

	$Y^{a=1}$	$Y^{a=0}$	Causal effect $Y^{a=1}-Y^{a=0}$
Doomed	1	1	$1-1=0$
Helped	1	0	$1-0=1$
Hurt	0	1	$0-1=-1$
Immune	0	0	$0-0=0$

- データとして観察はできない
  - 反事実アウトカムとの比較で定義可能
- Sharp causal null hypothesis
  - Doomed, Immuneな人しかいない

4

平均因果効果 Average Causal Effects

5

- $E[Y^{a=1}]-E[Y^{a=0}]$
  - 集団全員が曝露を受けた場合と集団全員が曝露を受けなかった場合の差
- Null hypothesis of no average causal effect
  - $E[Y^{a=1}]=E[Y^{a=0}]$
  - Sharp causal null hypothesisに加え、Helpedな人とHurtな人が同数いる場合も成立

5

周辺構造モデル

6

- Marginal Structural Model
  - 潜在アウトカムの周辺平均に関するモデル
- $E[Y^a]=\Pr[Y^a=1]=\beta_0+\beta_1 a$
  - β_1が平均因果リスク差
- $\text{logit}\ \Pr[Y^a=1]=\beta_0+\beta_1 a$
  - β_1が平均因果オッズ比

6

効果修飾を表現

7

- $E[Y^a|V] = \beta_0 + \beta_1 a + \beta_2 Va + \beta_3 V$
  - 効果修飾因子Vと曝露との交互作用項をモデルに含める
  - Vごとに効果の大きさを求めているので“marginal”モデルではないものの・・
- $\beta_2 \neq 0$であれば効果修飾あり

7

曝露が連続変数にも対応

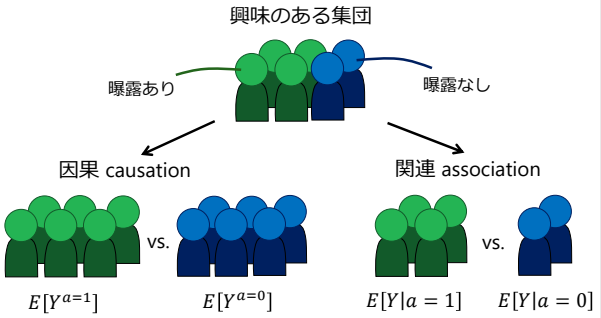
8

- 喫煙本数、体重変化量、バイオマーカーの変化量・・・
- $E[Y^a] = \beta_0 + \beta_1 a$
  - a ：Pack Years（タバコの箱数×年数）
  - 20PYに対する60PYでの発生割合の差は？
    - $40\beta_1$

8

Association is not causation

9



Hernán MA, Robins JM (2020). Causal Inference. What If. Chapman & Hall/CRC. を基に作成

9

因果効果指標と関連効果指標

10

- 因果効果指標は定義、概念的なもの
  - 反事実アウトカムを用いて定義されるため
- 関連効果指標は観察データから求まる

- 関連効果指標をもって因果効果指標を求めるには？
  - どのような条件が成立すれば？
  - どのような解析を行えば？

10

交絡 confounding

11

- 実際の曝露群での結果と集団全体が曝露した場合が違う
  - $E[Y^{a=1}|A=1] \neq E[Y^{a=1}]$

and / or

- 実際の非曝露群での結果と集団全体が曝露しなかった場合が違う
  - $E[Y^{a=0}|A=0] \neq E[Y^{a=0}]$

11

ランダム化による交換可能性の成立

exchangeability

12

- 曝露群での結果と非曝露群が、仮に曝露を受けた場合の結果が一致（その逆も）
  - $\Pr[Y^{a=1}|A=1] = \Pr[Y^{a=1}|A=0]$
    $\Pr[Y^{a=0}|A=0] = \Pr[Y^{a=0}|A=1]$
  - $Y^a \perp\!\!\!\perp A$ for all a
- 片方の集団と全体集団での結果と一致
  - $\Pr[Y^{a=1}|A=1] = \Pr[Y^{a=1}|A=0] = \Pr[Y^{a=1}]$
    $\Pr[Y^{a=0}|A=0] = \Pr[Y^{a=0}|A=1] = \Pr[Y^{a=0}]$

12

条件付き交換可能性 13

- 予後因子Lが同じ値を持つ集団（層内）では交換可能性が成立
  - $\Pr[Y^{a=1}|A=1,L=1]=\Pr[Y^{a=1}|A=0,L=1]$
    $\Pr[Y^{a=0}|A=0,L=1]=\Pr[Y^{a=0}|A=1,L=1]$
  - $\Pr[Y^{a=1}|A=1,L=0]=\Pr[Y^{a=1}|A=0,L=0]$
    $\Pr[Y^{a=0}|A=0,L=0]=\Pr[Y^{a=0}|A=1,L=0]$
  - $Y^a \perp\!\!\!\perp A|L$ for all a
- No unmeasured confounding
  - 残差交絡 residual confounding がない

13

擬似集団 pseudo-population 14

- 因果リスクを知る上で必要な、全員が曝露／非曝露である場合の仮想集団
- 実際に曝露を受けた人は9/12
  - 曝露を受けた割合の逆数をかけてみよう
  - $\frac{1}{\frac{9}{12}}=\frac{12}{9}$ 倍
- 曝露を受けなかった人は3/12
  - 逆数をかけよう；$\frac{1}{\frac{3}{12}}=4$ 倍

14

逆確率重み付け法 15

- Inverse probability weighting (IPW) 法
  - Horvitz-Thompson推定量(1952, *J Am Stat Assoc*)

$$\hat{E}\left[\frac{I(A=1)Y}{\Pr(A=1|L)}\right]$$

- 生成した擬似集団での関連効果指標は
  - 擬似集団での因果効果指標
  - 元の集団での因果効果指標　に同じ
    - $Y^a \perp\!\!\!\perp A|L$ for all a　ゆえ

15

擬似集団にてモデル化 16

- $E[Y^a]=E_{ps}[Y|a]$
  - 擬似集団において実際に曝露aを受けた人のアウトカムの期待値
- 周辺構造モデルをおける因果パラメータを擬似集団の関連指標として推定
  - $E[Y^a]=\beta_0+\beta_1 a$
  - $E_{ps}[Y|a]=\theta_0+\theta_1 a$

16

Lがたくさんあった場合 17

- 性別、重症度、高血圧の既往、糖尿病の既往、心不全の既往・・・・
  - 2 x 2 x 2 x 2 x ・・・
- 年齢、eGFR、LDL、HDL、TG・・・
  - 連続量データで厳密に等しい人は他にいない
- 統計モデルの利用
  - 個人ごとの曝露確率を予測する
    ロジスティック回帰

17

信頼区間の構成 18

- IPW法を用いたことの考慮が必要
- 漸近ロバスト分散
  - Proc CAUSALTRT (SAS/STAT14.2以降)
- ブートストラップ法
  - Proc CAUSALTRT (SAS/STAT14.2以降)
  - Proc SURVEYSELECTを利用
- 通常のロバスト分散
  - Proc GENMOD, Proc GEE, Proc GLIMMIX等

18

練習①　IPW解析をやる

19

年齢	性別	曝露	イベント	非発生	合計
高齢	男	あり	360	240	600
高齢	男	なし	40	60	100
高齢	女	あり	150	150	300
高齢	女	なし	30	70	100
若年	男	あり	80	120	200
若年	男	なし	20	80	100
若年	女	あり	30	70	100
若年	女	なし	10	90	100

19

練習①　IPW解析をやる

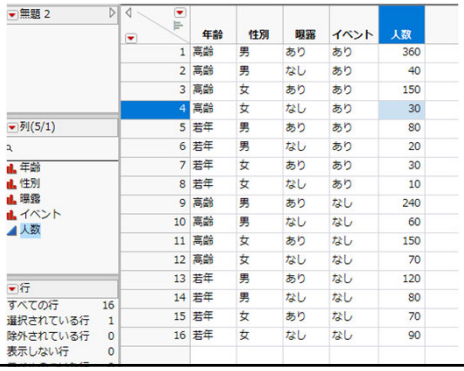
20

- 年齢による曝露オッズ比は3倍
  性別による曝露オッズ比は2倍
  - ロジスティックモデルが正しい数値例
- ロジスティック回帰モデルから
  (非)曝露確率を予測し、
  擬似集団でのリスク差を計算
  - リスク差回帰モデルにIPWを重みとして指定

20

データの作成

21

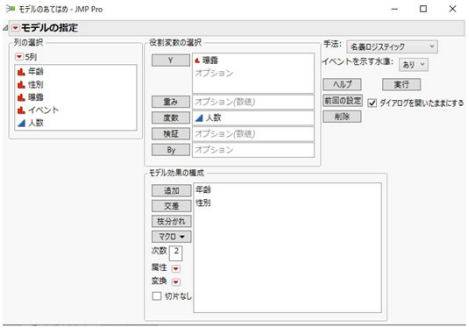


	年齢	性別	曝露	イベント	人数
1	高齢	男	あり	あり	360
2	高齢	男	なし	あり	40
3	高齢	女	あり	あり	150
4	高齢	女	なし	あり	30
5	若年	男	あり	あり	80
6	若年	男	なし	あり	20
7	若年	女	あり	あり	30
8	若年	女	なし	あり	10
9	高齢	男	あり	なし	240
10	高齢	男	なし	なし	60
11	高齢	女	あり	なし	150
12	高齢	女	なし	なし	70
13	若年	男	あり	なし	120
14	若年	男	なし	なし	80
15	若年	女	あり	なし	70
16	若年	女	なし	なし	90

21

曝露オッズをモデル化

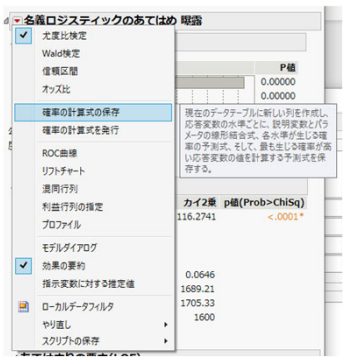
22



22

曝露の予測確率をデータ化

23



23

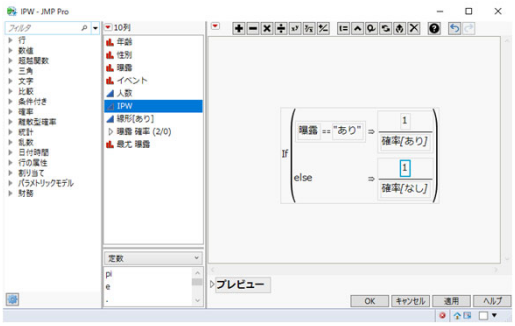
曝露確率が保存

24

24

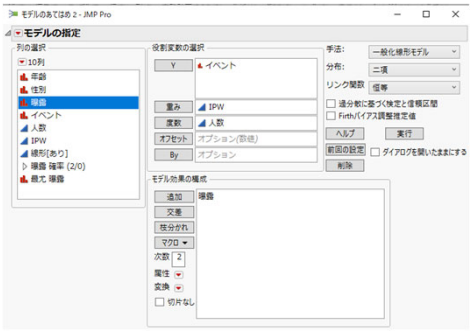
IPWの作成 25

- 列の新規作成＞計算式



25

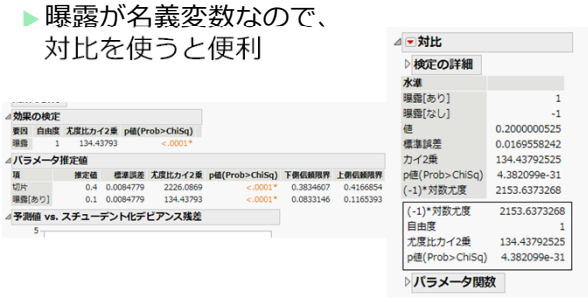
IPWを重みにしたリスク差回帰 26



26

リスク差の推定 27

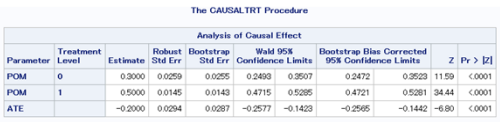
- 曝露が名義変数なので、対比を使うと便利



27

分散推定だけはご注意 28

- IPWを考慮しない分散：0.0170（Biased）
- 漸近ロバスト分散：0.0294
- ブートストラップ分散：0.0287



The CAUSALTRT Procedure

Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Bootstrap Std Err	Wald 95% Confidence Limits		Bootstrap Bias Corrected 95% Confidence Limits		Z	Pr > |Z|
POM	0	0.3000	0.0259	0.0255	0.2493	0.3507	0.2472	0.3523	11.59	<.0001
POM	1	0.5000	0.0145	0.0143	0.4715	0.5285	0.4721	0.5281	34.44	<.0001
ATE		−0.2000	0.0294	0.0287	−0.2577	−0.1423	−0.2565	−0.1442	−6.80	<.0001

- 通常のロバスト分散：0.0301

GEE パラメータ推定値の分析

経験的標準誤差推定

パラメータ	推定値	標準誤差	95% 信頼限界		Z	Pr > |Z|
Intercept	0.3000	0.0263	0.2485	0.3515	11.42	<.0001
a	0.2000	0.0301	0.1410	0.2590	6.65	<.0001

28

標準化 standardization 29

- 層ごとの条件付き期待値の重み付き平均

$$\mathrm{E}[Y^a] = \sum_l \mathrm{E}[Y^a|L=l]\Pr[L=l]$$

- Lが連続量である場合、確率密度fで置換

$$\int_l \mathrm{E}[Y^a|L=l]f[l]\,dl$$

29

$\mathrm{E}[Y^a|L=l]$ のモデル化 30

- Lのすべてのとりうる値ごとに
  - 高次元、連続量の場合、困難
- 回帰モデル等を利用

30

$\Pr[L = l]$ での重み付き平均

31

- 個人まで遡れば、単純平均でOK

$$\hat{E}[Y^a] = \frac{1}{n}\sum_{i}^{n} \hat{E}[Y^a | L = l_i]$$

- 特定の集団、外挿する場合は
  Lのprevalenceが必要

31

データの複製による標準化

32

- 共変量部分をそのまま、
  アウトカムを欠測
  {全員の曝露をなしと加工
  {全員の曝露をありと加工
  したデータを複製
  - このデータはアウトカムが欠測のため、
    条件付き期待値の計算には寄与しない
  - 共変量・曝露の情報があるので、
    予測された期待値をデータとしてはきだせる

32

練習②

33

- 練習①のデータに対して
  標準化を行ってみよう

33

データを複製

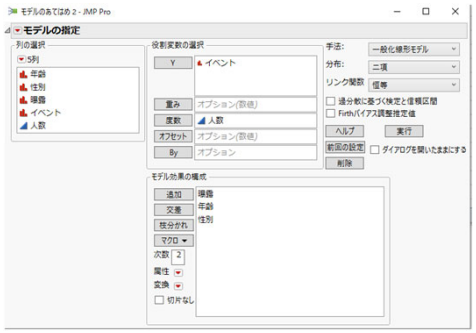
34

	年齢	性別	曝露	イベント	人数		年齢	性別	曝露	イベント	人数		年齢	性別	曝露	イベント	人数
1	高齢	男	あり	あり	360	17	高齢	男	あり		360	33	高齢	男	なし		360
2	高齢	男	なし	あり	40	18	高齢	男	あり		40	34	高齢	男	なし		40
3	高齢	女	あり	あり	150	19	高齢	女	あり		150	35	高齢	女	なし		150
4	高齢	女	なし	あり	30	20	高齢	女	あり		30	36	高齢	女	なし		30
5	若年	男	あり	あり	80	21	若年	男	あり		80	37	若年	男	なし		80
6	若年	男	なし	あり	20	22	若年	男	あり		20	38	若年	男	なし		20
7	若年	女	あり	あり	30	23	若年	女	あり		30	39	若年	女	なし		30
8	若年	女	なし	あり	10	24	若年	女	あり		10	40	若年	女	なし		10
9	高齢	男	あり	なし	240	25	高齢	男	あり		240	41	高齢	男	なし		240
10	高齢	男	なし	なし	60	26	高齢	男	あり		60	42	高齢	男	なし		60
11	高齢	女	あり	なし	150	27	高齢	女	あり		150	43	高齢	女	なし		150
12	高齢	女	なし	なし	70	28	高齢	女	あり		70	44	高齢	女	なし		70
13	若年	男	あり	なし	120	29	若年	男	あり		120	45	若年	男	なし		120
14	若年	男	なし	なし	80	30	若年	男	あり		80	46	若年	男	なし		80
15	若年	女	あり	なし	70	31	若年	女	あり		70	47	若年	女	なし		70
16	若年	女	なし	なし	90	32	若年	女	あり		90	48	若年	女	なし		90

34

通常のリスク差回帰

35



35

条件付け期待値をはきだし

36

36

複製データに適当に対比を設定

37

	年齢	性別	曝露	イベント	人数	P(イベント) 計算式	対比
13	若年	男	あり	なし	120	0.4	.
14	若年	男	なし	なし	80	0.2	.
15	若年	女	あり	なし	70	0.3	.
16	若年	女	なし	なし	90	0.1	.
17	高齢	男	あり		360	0.6	1
18	高齢	男	あり		40	0.6	1
19	高齢	女	あり		150	0.5	1
20	高齢	女	あり		30	0.5	1
21	若年	男	あり		80	0.4	1
22	若年	男	あり		20	0.4	1
23	若年	女	あり		30	0.3	1
24	若年	女	あり		10	0.3	1
25	高齢	男	あり		240	0.6	1
26	高齢	男	あり		60	0.6	1
27	高齢	女	あり		150	0.5	1
28	高齢	女	あり		70	0.5	1
29	高齢	男	あり		120	0.4	1
30	若年	男	あり		80	0.4	1
31	若年	女	あり		70	0.3	1
32	若年	女	あり		90	0.3	1
33	高齢	男	なし		360	0.4	0
34	高齢	男	なし		40	0.4	0
35	高齢	女	なし		150	0.3	0
36	高齢	女	なし		30	0.3	0

37

群間差を計算

38

信頼区間はBiased

38

IPW法と標準化

39

- IPW法

$$\hat{E}\left[\frac{I(A=1)Y}{\Pr(A=1|L)}\right]$$

  - 曝露確率をモデル化

- 標準化

$$\sum_l \mathrm{E}[Y^a|L=l]\Pr[L=l]$$

  - アウトカムをモデル化

パラメトリックなモデルを仮定しない（単純な層別ですむ）ならば両者は一致

39

"All models are wrong; but some are useful"

Box GEP. 1978

40

- IPW推定値と標準化推定値は大して変わらないはず
  - 大きく乖離するなら、モデルの誤特定を疑う
- モデルが(そこそこ)正しく特定できたかに注意

40

二重ロバスト推定量

Bang and Robins. 2005 *Biometrics*

41

- 曝露確率モデル、アウトカムモデルのいずれか一方でも正しければ、漸近的にバイアスをゼロにできる

$$\hat{E}[Y^{a=1}]$$
$$=\frac{1}{n}\sum_i^n\left[\hat{E}(Y|A=1,L_i)+\frac{A_i}{\Pr(A=1|L_i)}\{Y_i-\hat{E}(Y|A=1,L_i)\}\right]$$

アウトカムモデルが正しければゼロ

$$=\frac{1}{n}\sum_i^n\left[\frac{A_iY_i}{\Pr(A=1|L_i)}-\left\{\frac{A_i}{\Pr(A=1|L_i)}-1\right\}\hat{E}(Y|A=1,L_i)\right]$$

曝露確率モデルが正しければゼロ

41

Augmented IPW 推定量

42

- 二重ロバスト推定量は両方ともモデルが正しければセミパラメトリック有効
  - IPW法よりも漸近分散が小さい（効率がよい）
    - Robins, Rotnitzky, and Zhao. 1994 *JASA*
    - Rotnitzky, Robins, and Scharfstein. 1998 *JASA*
    - Scharfstein, Rotnitzky, and Robins. 1999 *JASA*

42

Target trial emulation

43

- 疫学研究での因果推論は、
  仮想的なランダム化比較試験を模倣
- 妥当な推測のために必要な性質
  - Exchangeability
  - Positivity
    - (非)曝露確率が0より大きい
  - Consistency
    - 受けた曝露通りの潜在アウトカムが観察される
  - No measurement error
  - No model misspecification

43

今回の内容

44

- 周辺構造モデル
  - 平均因果効果をモデル化
  - IPW法による擬似集団でそのままあてはめ
- 標準化
  - 平均因果効果を求めるもう一つのアプローチ
- IPW法か標準化か
  - 曝露確率のモデルかアウトカムのモデルか
  - 二重ロバスト推定量

44