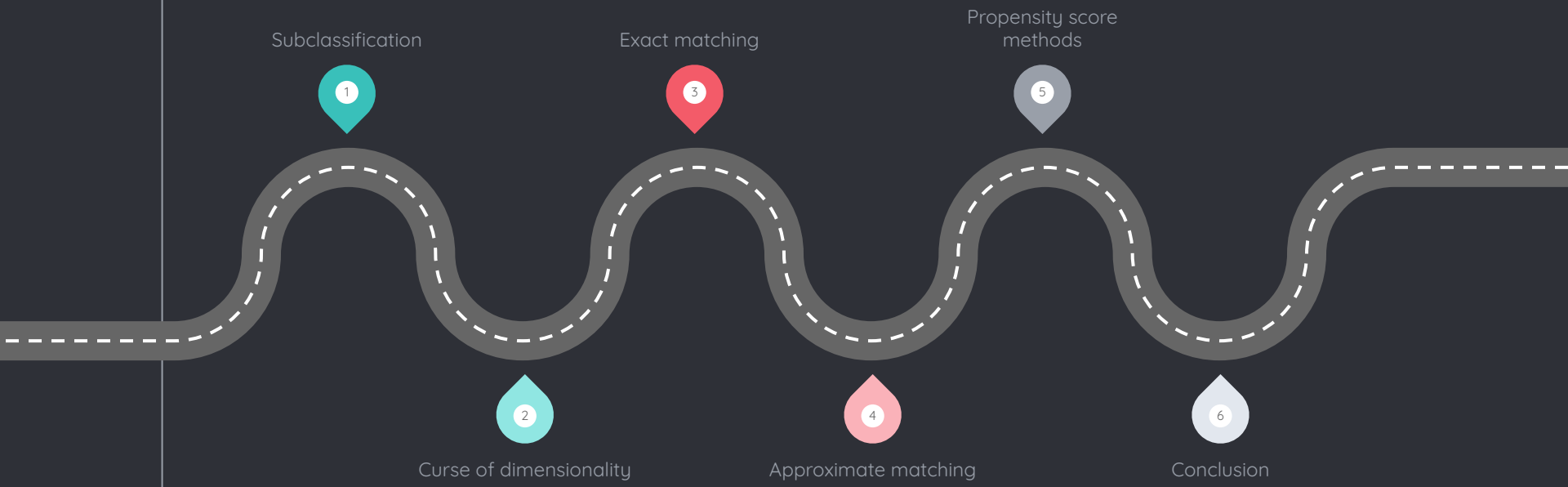
Causal Inference: *The Mixtape*
Ch5 Matching and Subclassification

三澤崇治　倉橋 優亜

Roadmap


Subclassification
1
Exact matching
3
Propensity score methods
5
2
Curse of dimensionality
4
Approximate matching
6
Conclusion

0 復習と導入

5.0 復習と導入

【全体観】

因果関係を推論したい

⇓そのために

バックドア基準を満たす

⇓そのために

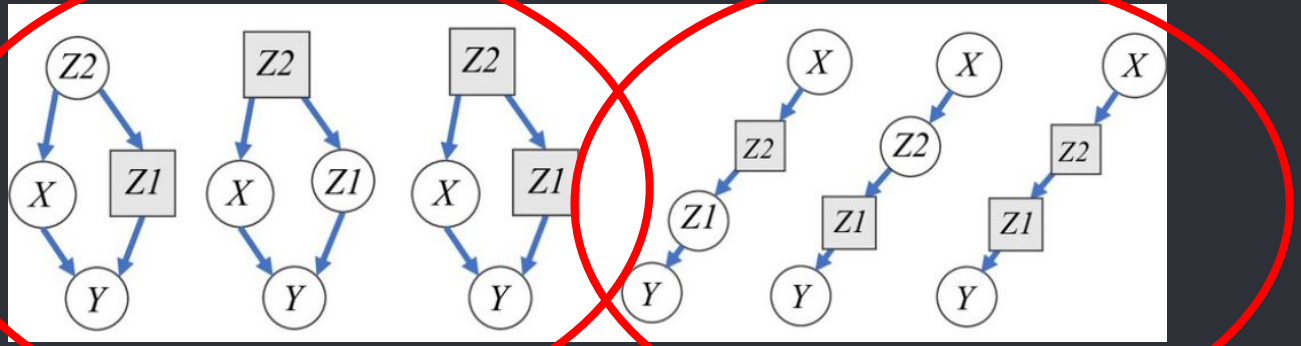
背景情報（交絡変数）を揃える

5.0.1 復習

例:「Zがバックドア基準を満たす」

≒「開いているバックドアパスがない

＋処理(X)→結果(Y)の道がブロックされていない」

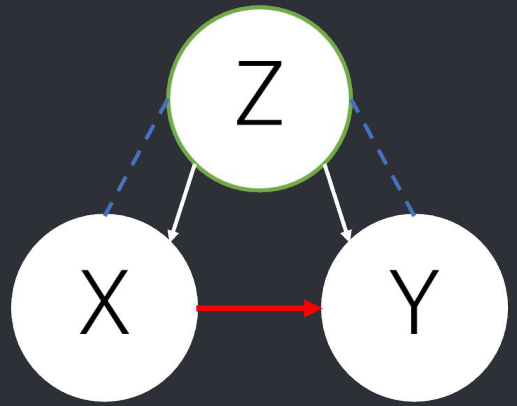


バックドアパスをブロック○

バックドアパスをブロック×

5.0.1 復習

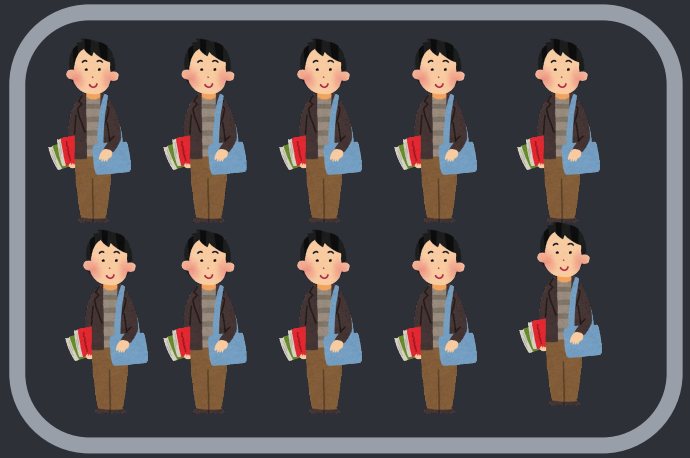
例:「Zがバックドア基準を満たす」
条件:
①XからZに有向道がない
②ZがXからYへの矢印を含む
すべての交絡変数間の経路を
塞いでいること



効果:⇒「X→Yの介入効果をバイアスなく推定できる」

【対象のあり方】

仮想：10人の20歳男性がいます



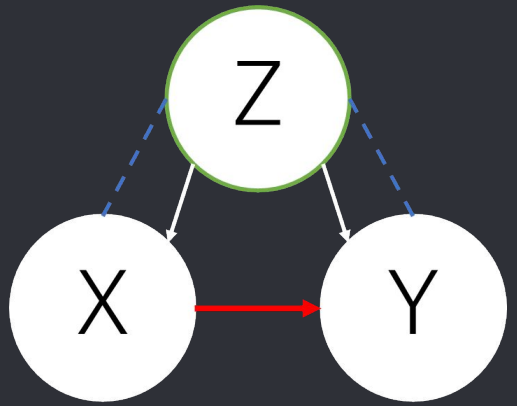
現実：不揃い

CIA（条件付独立性の仮定）

結果に影響を与える**共変量（Z）**を所与（条件）として、**結果変数（Y）**と**介入効果（X）**が独立であることを支える

$$(Y^1, Y^0) \perp\!\!\!\perp D \mid X$$

$$E[Y^1 \mid D = 1, X] = E[Y^1 \mid D = 0, X]$$
$$E[Y^0 \mid D = 1, X] = E[Y^0 \mid D = 0, X]$$



⇒Xの値を決めたときの、Y1とY0の期待値は治療群と対照群とで同じ

5.0.2 導入

「バックドア基準を満たすためには？」
⇒条件付けをする＝背景情報を同じにする

条件付けの種類:

1. Subclassification
2. Matching
    a. Exact Matching
    b. Approximate Matching

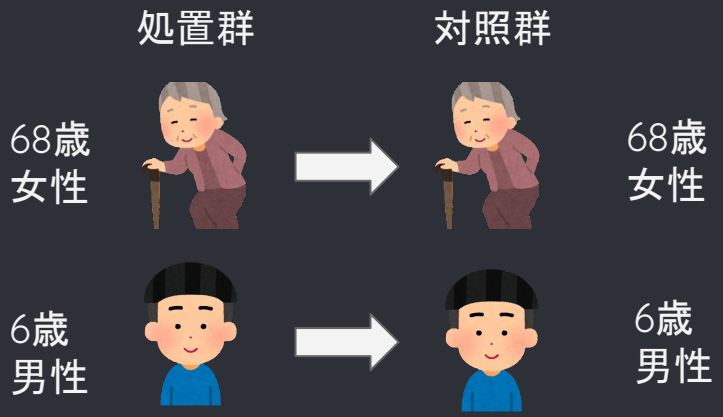
5.0.2 導入

Subclassification

	処置群 （薬飲んだ）	対照群 （薬飲でない）
20代	70	10
30代	20	20
40代	10	70
死亡率	10%	20%

交絡変数：年齢

Matching



交絡変数：年齢と性別

1 Subclassification

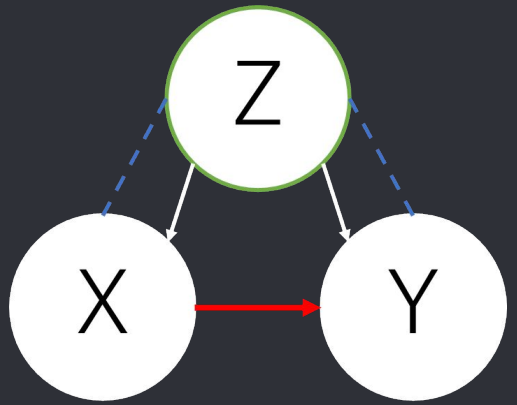
Subclassificationとは？

目的：背景情報を揃えて交絡バイアスを小さくする

方法：交絡変数の値によって層に分ける

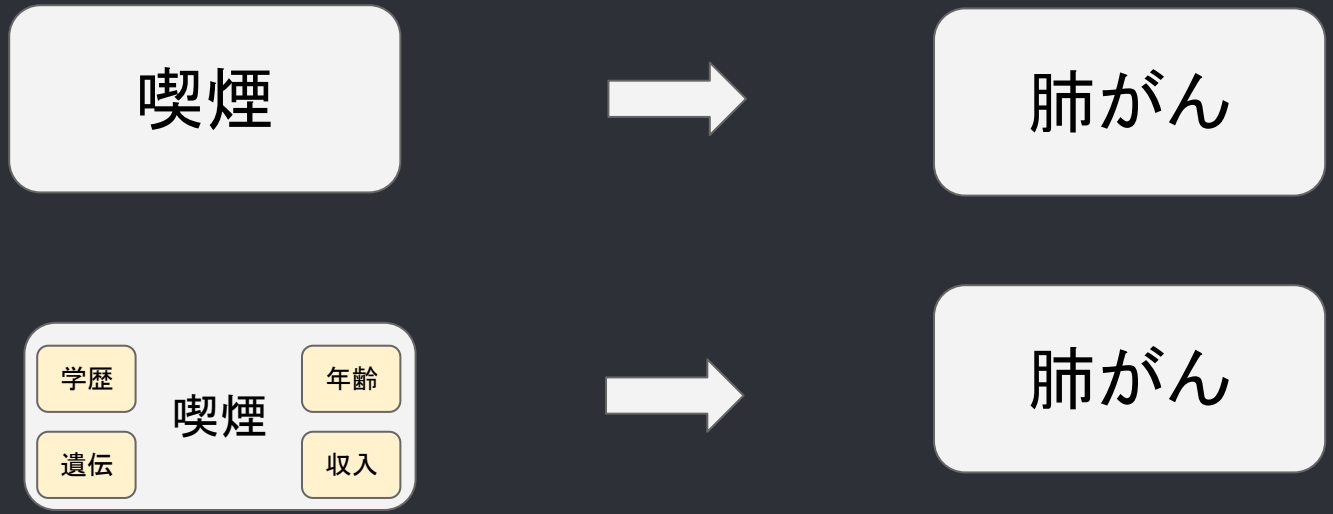
⇒層ごとに重みづけ

	処置群	対照群
20代	70	10
30代	20	20
40代	10	70



5.1.1 背景

20世紀半ば～後半：喫煙⇒肺がんの研究



例：喫煙タイプと死亡率の関係

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Table 5.1: Death rates per 1,000 person-years (Cochran 1968)

死亡率

5.1.1 背景：Cochran(1968)の研究

単純な平均値では比較できない

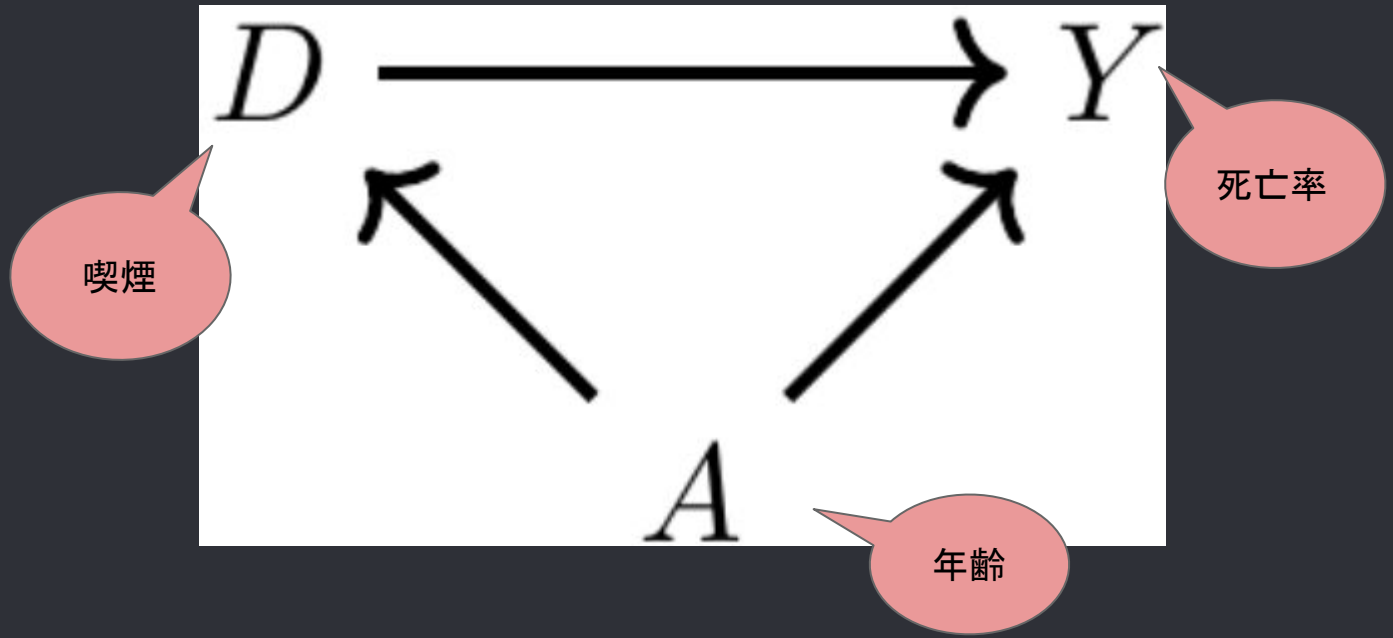
Smoking group	Canada	British	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Table 5.2: Mean ages, years (Cochran 1968).

平均年齢

5.1.1 背景：Cochran(1968)の研究

例：喫煙タイプと死亡率の関係



5.1.1 背景：Cochran(1968)の研究

Subclassification：全体の流れ

①年齢を層に分ける：20～40歳、41～70歳、71歳～
②処置群（タバコを吸う人）の層（ここでは年齢）別の死亡率を計算する
③処置群の死亡率に、対照群に対応する層別（年齢別）重みづけをする⇒処置群の**年齢調整死亡率**が計算できる

⇒バックドア基準が満たされる
⇒CIAが達成される

5.1.1 背景：Cochran(1968)の研究

Subclassification：①年齢を層別に分類

②タバコを吸う人の死亡率（/10万人）

若い人が多い群

高齢者が多い群

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

Table 5.3: Subclassification example.

年齢層の割合

5.1.1 背景：Cochran(1968)の研究

③Cigarette smokerの平均死亡率を計算

タバコ（ 若い人が多い群

年齢層別重みづけ＝

$$20 \times \frac{65}{100} + 40 \times \frac{25}{100} + 60 \times \frac{10}{100} = 29.$$

パイプ、葉巻（ 高齢者が多い群

年齢層別重みづけ＝

$$20 \times \frac{10}{100} + 40 \times \frac{25}{100} + 60 \times \frac{65}{100} = 51$$

20~40歳の
喫煙者の死亡率
（10万人当たり）

20~40歳の割合

結果：年齢を層別に分類（調整済）

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	29.5	14.8	21.2
Cigars/pipes	19.8	11.0	13.7

Table 5.4: Adjusted mortality rates using 3 age groups (Cochran 1968).

Subclassification：今までのまとめ

因果関係を推論したい

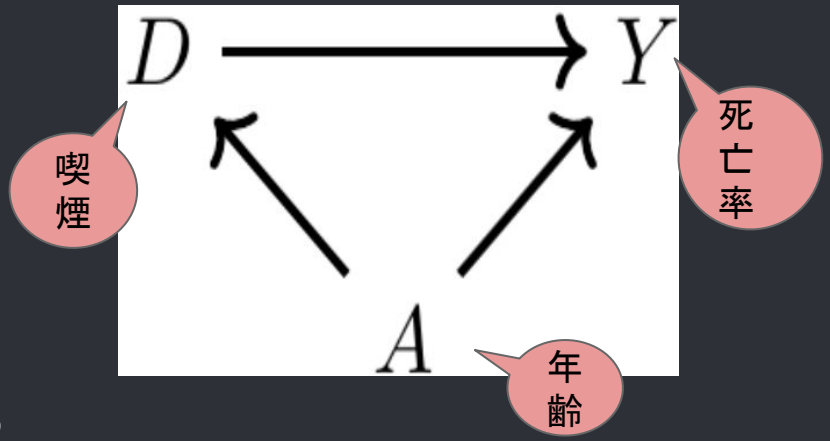
⇓そのために

バックドア基準を満たす

⇓そのために

背景情報（交絡変数）を揃える

→ Subclassificationを使う



5.1.2 前提条件の確認

- Subclassificationに必要な仮定

仮定1：CIA(条件つき独立の仮定)

$$(Y^1, Y^0) \perp\!\!\!\perp D \mid X$$

仮定2：Common Support(コモン・サポート)

$$0 < P_r(D = 1|X) < 1$$

ーデータを重みづけするため。

ATEの推定値を算出する

仮定1：CIAが成り立つとき、ある階層における平均処置効果は、

$$E[Y^1 - Y^0 \mid X] = E[Y^1 - Y^0 \mid X, D = 1]$$
$$= E[Y^1 \mid X, D = 1] - E[Y^0 \mid X, D = 0]$$
$$= E[Y \mid X, D = 1] - E[Y \mid X, D = 0]$$

さらに、仮定2：Common Supportより全ての階層において

[処置群のyの平均]ー[対照群のyの平均](前章だとτとされていた)が識別できるため、

$$\widehat{\delta_{ATE}} = \int (E[Y \mid X, D = 1] - E[Y \mid X, D = 0])\, d\mathrm{P_r}(\mathrm{X})$$

5.1.3 Subclassification exercise: Titanic data set

目的：富と規範が乗客の生存率に与えた影響を知りたい

仮説：ファーストクラスに座っていたことで生存確率は上昇していたのではないか

ー様々な席があり、富裕層は上層デッキに集中していた

ー女性や子どもは？→**問題！**

5.1.3 Subclassification exercise: Titanic data set

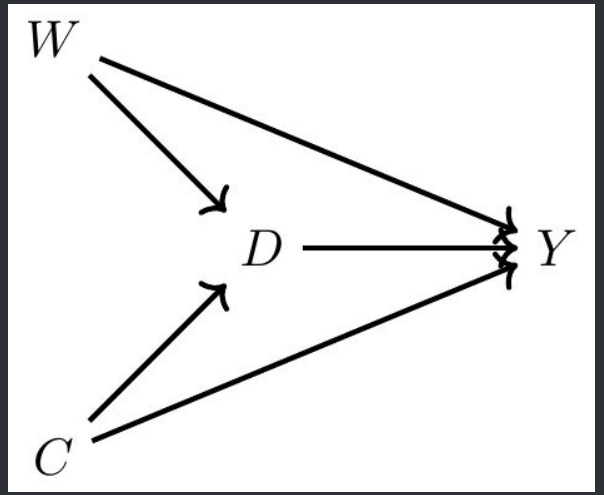
- 問題：

  女性や子供は救命ボートに優先的に乗船できた

  ↪女性や子どもがファーストクラスに座る確率が高かったら、ファーストクラスに座っていたことによる生存率の差は
  単にその社会規範の影響を拾い上げているだけなのかもしれない

5.1.3 Subclassification exercise: Titanic data set

- DAGを使用して因果関係を特定する
- 各有向辺の意味
  - W/C→D：女性/子どもであれば
    ファーストクラスに
    座る可能性が高い
  - W/C→Y：女性/子どもであれば
    救命ボートが優先的に
    割り当てられるので、
    生き残りやすい

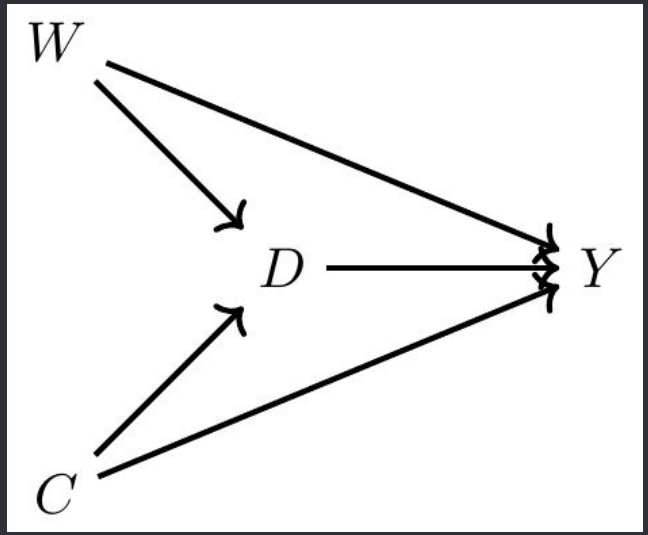


※観察・非観察を問わず、ほかにconfounderはないとする

5.1.3 Subclassification exercise: Titanic data set

- DとYの間の直接的なパス(因果関係)は1つ
- バックドアパスは2つ
  - D←C→Y
  - D←W→Y

→subclassificationを用いる



5.1.3 Subclassification exercise: Titanic data set

- Subclassificationによる統制の手順
  1. 若い男性、若い女性、年配の男性、年配の女性の4つのグループにデータを層別する。
  2. 各グループにおいて処置群と対照群の生存率の差を計算する。
  3. 各グループのファーストクラスでなかった人数を計算し、ファーストクラスでなかった総人数で割る。これが層別の重みとなる。
  4. 層別の重みを用いて加重平均生存率を算出する

5.1.3 Subclassification exercise: Titanic data set

ファーストクラスに座ることで生存確率が

SDO : 35.4%

⇩subclassification

The weighted ATE : 18.9%

上昇する

2 Curse of dimensionality ～次元の呪い～

5.1.4 Curse of dimensionality ～次元の呪い～

- Curse of dimensionality ～次元の呪い～
  - Titanicのケース：{2共変量, 2値}
    ={(性別,年齢), (男性/女性, 子ども/大人)}

↪もし年齢の取り得る値が複数あったら？

↪層内の差を計算するために必要な情報が
得られず層別の重みを計算できない可能性

5.1.4 Curse of dimensionality ～次元の呪い～

Age and Gender	Survival Prob. 1st Class	Controls	Diff	# of 1st Class	# of Controls
Male 11-yo	1.0	0	1	1	2
Male 12-yo	—	1	—	0	1
Male 13-yo	1.0	0	1	1	2
Male 14-yo	—	0.25	—	0	4

年齢の詳細なデータを持っていると仮定

↪Common Supportが成立していない

5.1.4 Curse of dimensionality ～次元の呪い～

年齢と性別の全ての組み合わせについて考えると、Common Supportの不成立はかなり一般的

⇩

Subclassificationを用いてATEを推定することが出来ない

つまり「次元の呪い」とは. . .

層別に使用した変数が多次元になりすぎて、その結果、サンプルが小さすぎるが故に、いくつかのセルでデータが欠損してしまっていること。

5.1.4 Curse of dimensionality ～次元の呪い～

Age and Gender	Survival Prob. 1st Class	Controls	Diff	# of 1st Class	# of Controls
Male 11-yo	1.0	0	1	1	2
Male 12-yo	–	1	–	0	1
Male 13-yo	1.0	0	1	1	2
Male 14-yo	–	0.25	–	0	4

この問題が処置群のみに生じる場合(実際に対照群よりも処置群の方がデータが少ないことが多い)

⇩

ATTは計算できる。

$$\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \times \left(\frac{N_T^k}{N_T}\right)$$

5.1.4 Curse of dimensionality ～次元の呪い～

つまり. . .

有限標本の場合、共変量の数が増えるにつれ、

Subclassificationの実現性は低くなる

∵多くのセルでいずれか片方、もしくはその両方を含まない

可能性が高まり、Common Supportを満たさないため

☞別の方法は. . . ？

3

Exact Matching

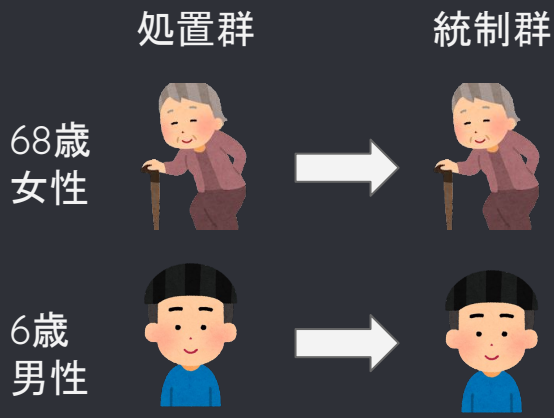
5.2 Exact Matching

Subclassification

	処置群	対照群
20代	70	10
30代	20	20
40代	10	70

交絡変数：年齢

Matching



交絡変数：年齢と性別

❖ マッチングの種類

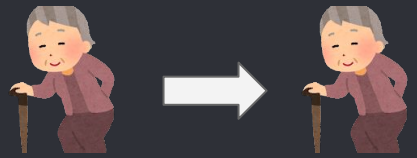
- ➢ Exact Matching（厳格なマッチング）
- ➢ Approximate Matching（近似一致）

- ➢ Propensity score Matching（傾向スコアマッチング）

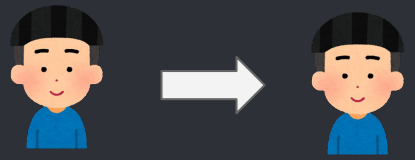
Exact Matching

処置群　　対照群

68歳
女性



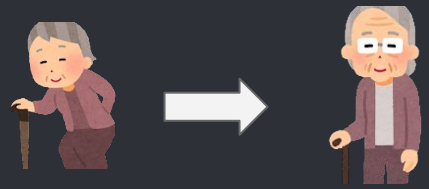
6歳
男性



Approximate Matching

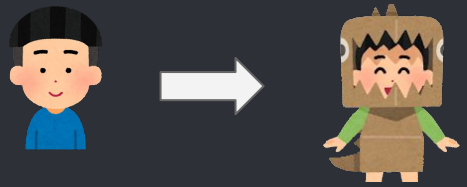
処置群　　対照群

68歳
女性



69歳
女性

6歳
男性



7歳
男性

交絡変数：年齢と性別

単純なマッチング推定量

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

5.2 Exact Matching

例：
研修への参加と収入の関係

Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	$11,075		31.95	$11,101.25

Table 5.6: Training example with exact matching

5.2 Exact Matching

例：
研修への参加と収入の関係

Trainees			Non-Trainees			Matched Sample		
Unit	Age	Earnings	Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500	14	18	8050
2	29	12250	2	27	10075	6	29	10525
3	24	11000	3	21	8725	9	24	9400
4	27	11750	4	39	12775	8	27	10075
5	33	13250	5	38	12550	11	33	11425
6	22	10500	6	29	10525	13	22	8950
7	19	9750	7	39	12775	17	19	8275
8	20	10000	8	33	11425	1	20	8500
9	21	10250	9	24	9400	3	21	8725
10	30	12500	10	30	10750	10,18	30	9875
			11	33	11425			
			12	36	12100			
			13	22	8950			
			14	18	8050			
			15	43	13675			
			16	39	12775			
			17	19	8275			
			18	30	9000			
			19	51	15475			
			20	48	14800			
Mean	24.3	$11,075		31.95	$11,101.25		24.3	$9,380

Table 5.7: Training example with exact matching (including matched sample)

4 Approximate Matching

Exact Matching

⇩

Approximate Matching

- 共変量が連続変数、多次元の場合、
  「完全に一致」するケースは無い場合がほとんど

⇒「一致」ではなく、「最も似ている」ケース同士と比較

Nearest-neighbor Matching

近さの基準

- The Euclidean distance
- The normalized Euclidean distance
- The Mahalanobis distance

5.3.1 Nearest-neighbor Matching ～最近傍マッチング～

The Euclidean distance

$$\left\|X_i - X_j\right\| = \sqrt{(X_i - X_j)'(X_i - X_j)} = \sqrt{\sum_{n=1}^{k}(X_{ni} - X_{nj})^2}$$

- 問題点：距離尺度自体が変数自体のスケールに依存する

5.3.1 Nearest-neighbor Matching ～最近傍マッチング～

The normalized Euclidean distance

$$\left\|X_i - X_j\right\| = \sqrt{(X_i - X_j)'\hat{V}^{-1}(X_i - X_j)} = \sqrt{\sum_{n=1}^{k} \frac{(X_{ni} - X_{nj})}{\widehat{\sigma_n}^2}}$$

Where

$$\hat{V}^{-1} = \begin{pmatrix} \widehat{\sigma_1}^2 & 0 & \cdots & 0 \\ 0 & \widehat{\sigma_2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\sigma_k}^2 \end{pmatrix}$$

- Xのスケールに変化があっても、その変化は分散にも影響するので、the normalized Euclidean distanceは変化しない。

5.3.1 Nearest-neighbor Matching ～最近傍マッチング～

- The Mahalanobis distance

$$\left\|X_i - X_j\right\| = \sqrt{(X_i - X_j)' \widehat{\sum}_X^{-1} (X_i - X_j)}$$

Where

$\sum_X$ はXの標本分散・共分散行列

5.3.1 Nearest-neighbor Matching ～最近傍マッチング～

マッチングの不一致はサンプルサイズが大きくなるほど0に収束する。

⇩

次元が大きければ大きいほど、マッチングの不一致度は高くなり、より多くのデータが必要になる。

⇩

マッチング問題では大きなデータセットが必要になる！！