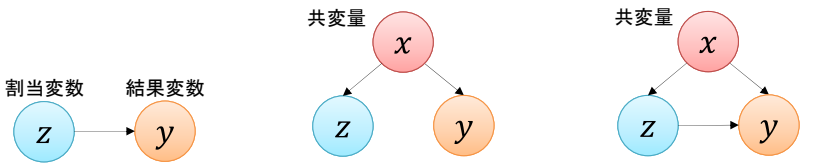
数理統計 補助資料
～統計的因果推論と共変量調整～

2023年度2学期：　月曜1限, 水曜3限
担当教員：　石垣　司

共変量

割当変数 z と結果変数 y の両者に影響を与える変数 x



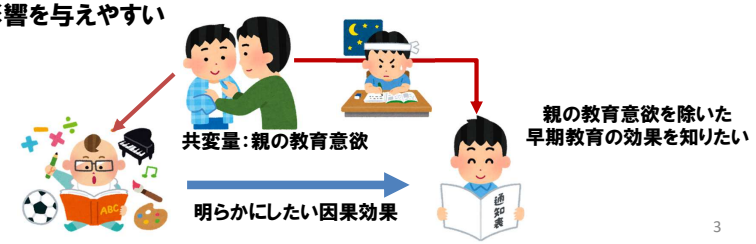
- 統計的因果推論において明らかにしたいのは，共変量の影響を取り除いた処置(割当変数)と結果の因果効果
  - 結果変数 y は処置/非処置の割り当てに値が依存する確率変数
  - ランダム割り当てならば平均処置効果の推定量 $\hat{E}[y_1 - y_0]$ は共変量の影響を受けず不偏推定量(共変量についてもランダム化されている)
  - ランダム割り当てではない観察データでは共変量の影響で因果効果を正確に測定できない(疑似相関の例)

社会科学の問題と共変量の例

職業訓練プログラム受講者とその後の賃金の因果関係
- 職業訓練プログラムは自主的に参加するためプログラム参加者は非参加者と比べて就労意欲が高い。意欲の高低はその後の賃金に影響を与えやすい

ベトナム戦争での軍隊経験とその後の賃金の因果関係
- 軍隊は危険かつ訓練が厳しいため多くの人は忌避。望む職に就労できなかった元々賃金の低くなる傾向にある人が軍隊に入りやすい

早期教育と中学校での成績の因果関係
- 親の教育意欲が高いほど早期教育を受ける傾向。親の教育意欲は中学校での成績にも影響を与えやすい



共変量調整 #1

共変量の影響を推定結果から取り除くための対処法
- 同じ共変量を持つグループ間で平均処置効果を算出すれば，共変量の影響は取り除くことができる(と仮定する)

例：人間ドックを受診した人ほどコレステロール値が高い?
(疑似的なストーリー)
- 母集団：健康不安のない20～60歳の日本人男性
- 割当変数：人間ドックを受診/非受診の観察データ
- 結果変数：受診日の血清LDLコレステロールの値
- 共変量：年齢
  - 母集団内の年齢とコレステロール値はほぼ比例関係
  - 母集団内の人間ドック受診者の約9割は40歳～60歳

共変量調整 #2

例：人間ドックを受診した人ほどコレステロール値が高い?

共変量調整無し

受診した群（処置群）

コレステロール値
平均130mg/dl

非受診群（対照群）

コレステロール値
平均120mg/dl

因果効果 +10 mg/dl ?

共変量調整あり

おじさんグループ
（40, 50歳代）

受診した群（処置群）

平均130mg/dl

非受診群（対照群）

平均130mg/dl

若者グループ
（20, 30歳代）

受診した群（処置群）

平均110mg/dl

非受診群（対照群）

平均110mg/dl

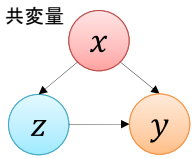
おじさんグループ間の差と若者グループ間の差の
年齢の分布に関する期待値を計算

因果効果ゼロ

共変量調整の数理的記述 #1

推定したいのは平均処置効果 $E[y_1 - y_0]$

- **共変量ベクトル $\boldsymbol{x} = [x_1, \ldots, x_P]^T$**
- **割当がランダムのとき，z と y_1, z と y_0 は独立なので，**
  $E[y_1] = E[y_1|z = 1]$ と $E[y_0] = E[y_0|z = 0]$ が成立した
  処置群（対照群）に割当てられたグループの結果変数の期待値は，母集団全体の処置（非処置）をしたグループの結果変数の期待値と一致



同時分布

$$P(y, z, \boldsymbol{x}) = P(y|z, \boldsymbol{x})P(z|\boldsymbol{x})P(\boldsymbol{x})$$

- **今，z と y_1, z と y_0 は，x から影響を受けるので，**
  一般には $E[y_1] \neq E[y_1|z = 1]$ と $E[y_0] \neq E[y_0|z = 0]$

共変量調整の数理的記述 #2

平均での独立性の仮定の下での平均処置効果

- **平均での独立性の仮定：共変量 x が同じ（おじさんグループと若者グループに分けて考えた場合）ならば，次式を満たすという仮定**

$$E[y_1|\boldsymbol{x}] = E[y_1|z = 1, \boldsymbol{x}] \;\&\; E[y_0|\boldsymbol{x}] = E[y_0|z = 0, \boldsymbol{x}]$$

- **平均での独立性の仮定を満たす場合，**

$$E[y_1 - y_0|\boldsymbol{x}] = E[y_1|z = 1, \boldsymbol{x}] - E[y_0|z = 0, \boldsymbol{x}]$$

- **x の分布に対する期待値 E_x を取り平均処置効果を推定**

$$E[y_1 - y_0] = E_x\big[E[y_1|z = 1, \boldsymbol{x}] - E[y_0|z = 0, \boldsymbol{x}]\big]$$

おじさんの確率 = 14/20, **若者の確率** = 6/20
おじさんグループの処置群と対照群の因果効果 = 130-130 = 0
若者グループの処置群と対照群の因果効果 = 110-110=0

$$\hat{E}[y_1 - y_0] = (130 - 130) \times 14/20 + (110 - 110) \times 6/20 = 0$$

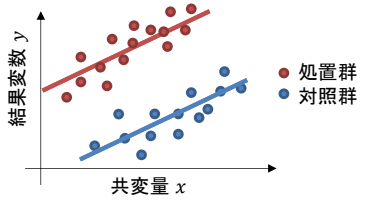
共変量調整の方法

マッチング，層別解析

- **処置群と対照群で同じような共変量のペアを作りマッチングしたり、同じようなグループを作成し層別（おじさんグループ/若者グループ）に分析する**
  弱点：共変量が複数のとき同じようなペアや層を作成するのが本質的に困難になる。“同じような”の区分けが恣意的になる

回帰モデルの利用

- **ーとーの差の x に関する分布の期待値を因果効果とする（共分散分析）**
  弱点：この差を因果効果と見なすためには強い仮定が必要。その仮定が崩れると推定結果に大きなバイアスがかかる



傾向スコアによる共変量調整

傾向スコア(Rosenbaum & Rubin 1983)

- 観測された共変量ベクトルが与えられた時に処置群に割当てられる条件付き確率 $e \equiv P(z=1|\boldsymbol{x})$

傾向スコアによる共変量調整の性質

- 傾向スコアが同じならば z と y_1, z と y_0 は独立と仮定すると

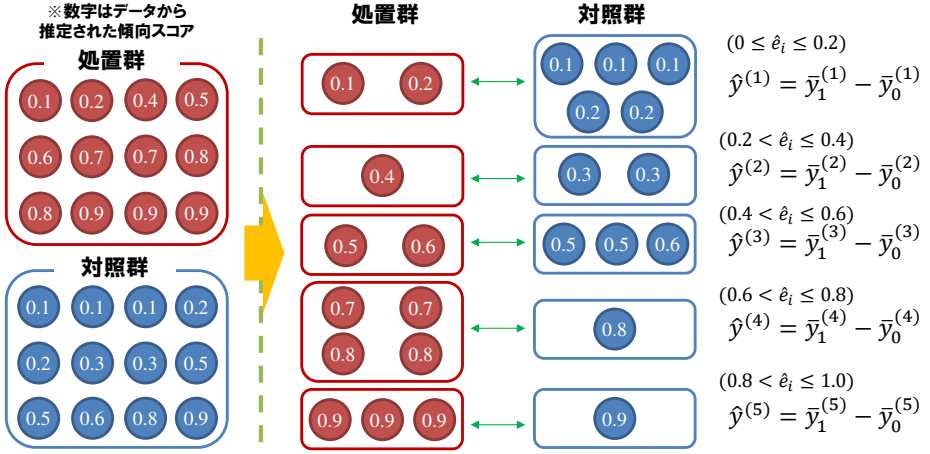
$$E[y_1 - y_0] = E_e\big[E[y_1|z=1,e] - E[y_0|z=0,e]\big]$$ (理屈はp.7と同様)

- 各主体 i の傾向スコア $e_i = P(z=1|\boldsymbol{x}_i)$ は観測データ $\{z_i, \boldsymbol{x}_i\}_{i=1,\dots,N}$ からロジスティック回帰モデルで推定可能
- 複数の共変量の情報を1次元の変数に縮約

  1変数であるので，マッチングや層別のような複数の共変量に関する問題は起こらない

傾向スコアを用いた層別分析

例：5段階の層別化による平均処置効果の推定量



$N = 24$

$$\hat{E}[y_1 - y_0] = \sum_{k=1}^{5} \frac{N_k}{N}\hat{y}^{(k)} = \frac{7}{24}\hat{y}^{(1)} + \frac{3}{24}\hat{y}^{(2)} + \frac{5}{24}\hat{y}^{(3)} + \frac{5}{24}\hat{y}^{(4)} + \frac{4}{24}\hat{y}^{(5)}$$

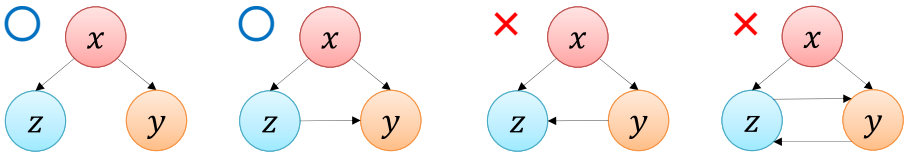
共変量調整による因果効果推定のための条件 #1

強く無視できる割り当て条件

- 共変量の値を固定したとき，結果変数 y から割当変数 z への直接の影響は無いという条件

$$P(z|y,\boldsymbol{x}) = P(z|\boldsymbol{x})$$

- 本条件が成り立つ/成り立たないパターン



-  「先日の職場の健康診断でコレステロール値が異常値だったので人間ドックを受けに来た」という人たちがいると，強く無視できる割り当て条件を満たさない

共変量調整による因果効果推定のための条件 #2

強く無視できる割り当て条件と平均での独立性の仮定

$P(z|y,\boldsymbol{x}) = P(z|\boldsymbol{x})$ であるならば，ベイズの定理より，

$$P(y|z,\boldsymbol{x}) = \frac{P(z|y,\boldsymbol{x})P(y|\boldsymbol{x})}{P(z|\boldsymbol{x})} = P(y|\boldsymbol{x})$$

$$\Rightarrow P(y_1|z=1,\boldsymbol{x}) = P(y_1|\boldsymbol{x}) \text{ and } P(y_0|z=0,\boldsymbol{x}) = P(y_0|\boldsymbol{x})$$

$$\Rightarrow E[y_1|z=1,\boldsymbol{x}] = E[y_1|\boldsymbol{x}] \text{ and } E[y_0|z=0,\boldsymbol{x}] = E[y_0|\boldsymbol{x}]$$

2行目の意味：強く無視できる割り当て条件を満たすならば，共変量の値を固定すると結果変数は割当変数には依存しない

全体の意味：強く無視できる割り当て条件を満たすならば，平均での独立性の仮定も満たされる

- 平均での独立性の仮定を満たさない場合，本日紹介した共変量調整による因果効果推定は成立しない

演習問題 #1

ある職業訓練プログラムが受講者のスキルに対して与える影響を受講者（処置群8人）と非受講者（対照群8人）のスキルテストの点数を用いて因果効果を推定したい

ただし，プログラム参加自体とスキルテストの点数の両方が，元々のその受講者の能力・意欲・年齢・学歴・経歴などから影響を受けると考えられるため，各受講者 i について推定された傾向スコア $\hat{e}_i$ を用いて共変量調整を行うこととする

また，強く無視できる割り当て条件やその他の平均処置効果推定に必要な仮定は満たされている

加えて，共変量の影響は推定された傾向スコアにより十分に調整できているとする

※実用的には各群8名のサンプルサイズ少ないが，演習問題のため気にしない

演習問題 #2

以上の条件で，傾向スコアによる共変量調整を用いて職業訓練プログラムへの参加がスキルテストの点数に与える平均処置効果を求めなさい

ただし，ここでは下記の分類による2段階の層別化を利用しなさい

第1層： $0 \leq \hat{e}_i < 0.5$，　**第2層：** $0.5 \leq \hat{e}_i \leq 1.0$

	処置群			対照群	
受講者	点数	傾向スコア	非受講者	点数	傾向スコア
1	60	0.2	9	60	0.1
2	60	0.3	10	50	0.1
3	60	0.4	11	70	0.2
4	80	0.6	12	50	0.3
5	70	0.8	13	70	0.4
6	90	0.8	14	60	0.4
7	95	0.9	15	80	0.8
8	90	0.9	16	80	0.8
平均点	75.6			65.0	

単純な平均点の差は 10.6