調査観察データの統計科学ゼミ

第二回

Adachi Lab. M1 伊藤真道

大阪大学大学院人間科学研究科

担当した章

1. 2.4　欠測モデルから見た調査観察データと因果効果の定義
2. 2.5 共変量調整による因果効果推定のための条件

2.4　欠測モデルから見た調査観察データと因果効果の定義

2.4 欠測モデルから見た調査観察データと因果効果の定義

	早期教育あり			早期教育なし		
所属群	1	1	1	0	0	0
対象者番号	1	2	...	...	N-1	N
y_1	y_{11}	y_{21}	$\cdots$	$\cdots$	y_{N-11}	y_{N1}
y_0	y_{10}	y_{20}	$\cdots$	$\cdots$	y_{N-10}	y_{N0}

y_1：早期教育をした場合の子供の中学校での成績
y_0：早期教育をしない場合の子供の中学校での成績

2.4 欠測モデルから見た調査観察データと因果効果の定義

- 潜在的な結果変数
  - 独立変数が取りうる値の数だけ存在する．仮想的な従属変数
- ルービンの因果モデル
  - あえて独立変数の条件の数だけ「本来あり得た結果」を考える
  - 反実仮想モデル/アプローチ (counterfactual model / approach) とも呼ばれる
- 実際に観測される y は，2 つの潜在的な結果変数 y_1, y_0 と独立変数 z を用いて

$$y = zy_1 + (1 - z)y_0$$

と表される

因果効果の定義

- 対象者 i に対する因果効果
  - 対象者 i での 2 つの潜在的な結果変数の差 $y_1 - y_0$
  - 条件の割り当て z 以外の対象者の要因が除去されている量
  - 実際には結果変数のうち必ず一方は観測できないため，この推定量は観測されたデータから計算できない (<-因果公開における根本問題)

- ルービンの因果効果 (平均処置効果)

$$E[y_1 - y_0] = E[y_1] - E[y_0]$$

  - 母集団の対象者全員が「処置群に割り当てられた際の結果」と「対照群に割り当てられた際の結果」の差の平均
  - 因果効果の素直な推定量は，

$$\hat{E}(y_1 - y_0) = \frac{1}{N}\sum_{i=1}^{N}(y_{i1}) - y_{i0}$$

  - 結果変数の片方は欠測しているため，観測データからこれを計算することはできない．

無作為割り当て

- 無作為割り当て
  - 結果変数への割り当てが無作為割り当てなら，割り当て z と従属変数 y_1, y_0 は独立となる．つまり，

$$p(y_1|z) = \frac{p(y_1, z)}{p(z)} = \frac{p(y_1)p(z)}{p(z)} = p(y_1)$$
$$p(y_0|z) = \frac{p(y_0, z)}{p(z)} = \frac{p(y_0)p(z)}{p(z)} = p(y_0)$$
$$E[y_1|z] = E[y_1],\ E[y_0|z] = E[y_0], (z = 0, 1) \tag{2.10}$$

となる．因果効果は，

$$E[y_1 - y_0] = E[y_1] - E[y_0] = E[y_1|z = 1] - E[y_0|z = 0] \tag{2.11}$$

が成立する．

無作為割り当て続き

- (無作為割り当て続き)
  - さらに $y = zy_1 + (1 - z)y_0$ から,

$$E[y_1|z=1] = E[y|z=1],\ E[y_0|z=0] = E[y|z=0]$$

が成立する.
  - →無作為割り当てが成立しているなら欠測値の存在を無視できる！
  - 観測された各群の平均値の差

$$\frac{1}{N_1}\sum_{i:z_i=1}^{N} y_i - \sum_{i:z_i=0}^{N} y_i$$

を用いることで，因果効果を普遍推定することが可能.
  - ここで，$E[y_1], E[y_0]$ を各群での周辺期待値と呼ぶ

処置群での介入効果/対照群での介入効果

- 処置群での介入効果 (average treatment effect on the treated: TET)
  - 処置群における潜在的な結果変数の差の期待値

$$TET = E[y_1 - y_0|z = 1]$$

  - e.g. 失業者に対する失業給付の効果，教育ニーズのある子どもに対する教育プログラムの効果 etc....

- 対照群での介入効果 (average treatment effect on the untreated: TEU)
  - 対照群における潜在的な結果変数の差の期待値

$$TEU = E[y_1 - y_0|z = 0]$$

- 一方を正しく推定できれば他方も正しく推定可能

- 但し，どちらも観測できない値を含むため，単純な解析では推定不可能

処置群での介入効果/対照群での介入効果　続き

- TET と TEU を用いて，因果効果を表すと，

$$E[y_1 - y_0] = TET \times p(z=1); TEU \times p(z=0)$$

- TET,TEU が推定でき，母集団での構成比 $p(z=1), p(z=0)$ がわかるなら，因果効果も推定可能である．
- 因果効果の推定には構成比必要 → 因果効果は処置群と対照群がどのような母集団から抽出されたかに依存する．

分位点での因果効果と周辺構造モデル

- $Q_\alpha(a)$ を変数 a の $100 \times (1-\alpha)\%$ 分位点とすると，分位点での因果効果は
$$Q_\alpha(y_1) - Q_\alpha(y_0)$$
で与えられる．
  - 因果効果と同様に，$p(y_1|z=1), p(y_0|z=0)$ ではなく，$p(y_1), p(y_0)$ での分位点の計算を行なっていることに注意．
  - 一般に
$$Q_\alpha(y_1) - Q_\alpha(y_0) \neq Q_\alpha(y_1 - y_0)$$
である．

- 周辺構造モデル (marginal structural model)
  - 今日変量の影響を除去した「潜在的な結果変数の周辺期待値構造」
  - e.g. 潜在的な結果変数に対する，割り当てと共変量の効果を見たいなら，
$$y_1 = \beta_1^t \mathbf{v} + \epsilon_1,\ y_0 = \beta_0^t \mathbf{v} + \epsilon_0$$
とモデリングすれば良い (詳しくは 3.4 節で !!)

2.5 共変量調整による因果効果推定のための条件

2.5 共変量調整による因果効果推定のための条件

- 共変量調整 (今日変量についての周辺化)
  - 共変量の値に依存しない量を得るために共変量の分布について期待値をとること，
  - 選択モデル $p(y, m|\theta, \phi) = p(y|\theta)p(m|y, \phi)$ の考え方に基づくと，

$$\begin{aligned} p(y_1, y_0, \mathbf{x}, z) &= p(z|y_1, y_0, \mathbf{x})p(y_1, y_0, \mathbf{x}) \\ &= p(z|y_1, y_0, \mathbf{x})p(y_1, y_0|\mathbf{x})p(\mathbf{x}) \end{aligned} \tag{2.14}$$

のように分解可能．この時，例えば，$p(y_1)$ は，

$$p(y_1) = \int p(z|y_1, y_0, \mathbf{x})p(y_1, y_0|\mathbf{x})p(\mathbf{x})dy_0dy_1d\mathbf{x}$$

となり，$z, y_0, \mathbf{x}$ の値には依存しない．

- 疑似相関
  - 本来は，独立変数単独での結果変数への因果効果がないにも関わらず，見かけ上両者の関係が生じている状況

強く無視できる割り当て

- 強く無視できる割り当て (strongly ignorable treatment assignment) 条件
  - 「割り当てはあくまでも共変量のみに依存し，結果変数には依存しない」という仮定
  - 共変量 $\mathbf{x}$ の値を条件付けると，処置群 $z=1$ の時の y_1 と対照群 $z=0$ の時の y_0 の同時分布が z と独立である，つまり

$$(y_1, y_0) \perp\!\!\!\perp z|\mathbf{x} \tag{2.15}$$

    という条件
  - これを同時分布で書くと，

$$p(y_1, y_0, \mathbf{x}, z) = p(y_1, y_0|\mathbf{x})p(z|\mathbf{x})p(\mathbf{x})$$

    とかける．これと (2.14) から，強く無視できる割り当て条件は，

$$p(z|y_1, y_0, \mathbf{x}) = p(z|\mathbf{x}) \tag{2.16}$$

    である．

強く無視できる割り当て条件続き

- 「どちらの群に割り当てられるかは共変量の値に依存」
- マッチングや層別解析でもこの仮定を利用

- ランダムな欠測

$$p(z=j|y_1,y_0,\mathbf{x}) = p(z=j|y_j,\mathbf{x}) \ (j=1,0) \tag{2.17}$$

- つまり $z=1$ に割り当てられる確率は y_1 に依存し，$z=0$ に割り当てられる確率は y_0 に依存する．
- (2.16), (2.17) を比較すると，
  強く無視できる割り当て $\Rightarrow$ ランダムな欠測

強く無視できる割り当て条件とランダムな欠測続き

- (2.16) 式 $p(z|y_1, y_0, \mathrm{x}) = p(z|\mathrm{x})$ を書き換えると，

$$p(z|y_1, y_0, \mathrm{x}) = \frac{p(y_1, y_0|z, \mathrm{x})p(z|\mathrm{x})}{p(y_1, y_0|\mathrm{x})} = p(z|\mathrm{x})$$

  から，

$$p(y_1, y_0|z, \mathrm{x}) = p(y_1, y_0|\mathrm{x}) \tag{2.18}$$

  とかける．
- → 共変量を条件づければ，y_1, y_0 の同時分布の形は，どちらの群に割り当てられたか z に依存しないという仮定と同等!!

強く無視できる割り当て条件続き

- 平均での独立性 (mean independence)
  - 強く無視できる割り当て条件が成立しているとして，(2.18) 式を周辺化したものの期待値

$$\left.\begin{array}{l} E(y|Z,\mathbf{x}) = E(y_1|z,\mathbf{x}) = E(y_1|\mathbf{x}) \\ E(y|Z,\mathbf{x}) = E(y_0|z,\mathbf{x}) = E(y_0|\mathbf{x}) \end{array}\right. \quad (z = 1, 0) \tag{2.19}$$

    のこと．
  - 平均の独立性の仮定が成立しているとする．この時，共変量を条件付けた因果効果 $E(y_1 - y_0|\mathbf{x})$ は

$$E(y_1 - y_0|\mathbf{x}) = E(y_1|z = 1, \mathbf{x}) - E(y_0|z = 0, \mathbf{x}) \tag{2.20}$$

    となる．
  - →観測された結果変数についての回帰関数

$$E(y_1|z = 1, \mathbf{x}), E(y_0|z = 0, \mathbf{x})$$

    で表現可能！

強く無視できる割り当て条件続き

- さらに共変量に関して期待値をとると，

$$
\begin{aligned}
E(y_1 - y_0) = E_{\mathrm{x}}[E(y_1 - y_0|\mathrm{x})] &= E_{\mathrm{x}}[E(y_1|z=1,\mathrm{x}) - E(y_0|z=0,\mathrm{x})] \\
&= E_{\mathrm{x}}[E(y|z=1,\mathrm{x}) - E(y|z=0,\mathrm{x})]
\end{aligned}
$$

  となり，観測されているデータのみを用いて因果効果を推定することが可能
- 因果効果を推定するためだけなら，平均での独立性のみが成立していればよい．
- 強く無視できる割り当て条件は，潜在的な結果変数 y の期待値以外の母数を推定したい場合や，潜在的な結果変数 y のモデリングが必要な場合に必要となる．

周辺効果と条件付き効果

周辺効果 (marginal effect)

共変量の影響を除去した後の結果変数の期待値の差 (= 因果効果)

$$E(y_1 - y_0)$$

条件付き効果 (conditional effect)

共変量の値を所与とした結果変数の条件付き期待値の差

$$E(y_1 - y_0 | \mathbf{x})$$

(ブロック使ってみたかったんや…)

メモに使ってね！

メモに使ってね！

Questions?