2022年3月27日

第24回春の合宿セミナー（日本行動計量学会）

（統計的因果推論入門）

講義6a
傾向スコア
と
重回帰モデル

長崎大学 情報データ科学部 准教授

高橋 将宜

博士（理工学）

m-takahashi@nagasaki-u.ac.jp

観察研究におけるデータ

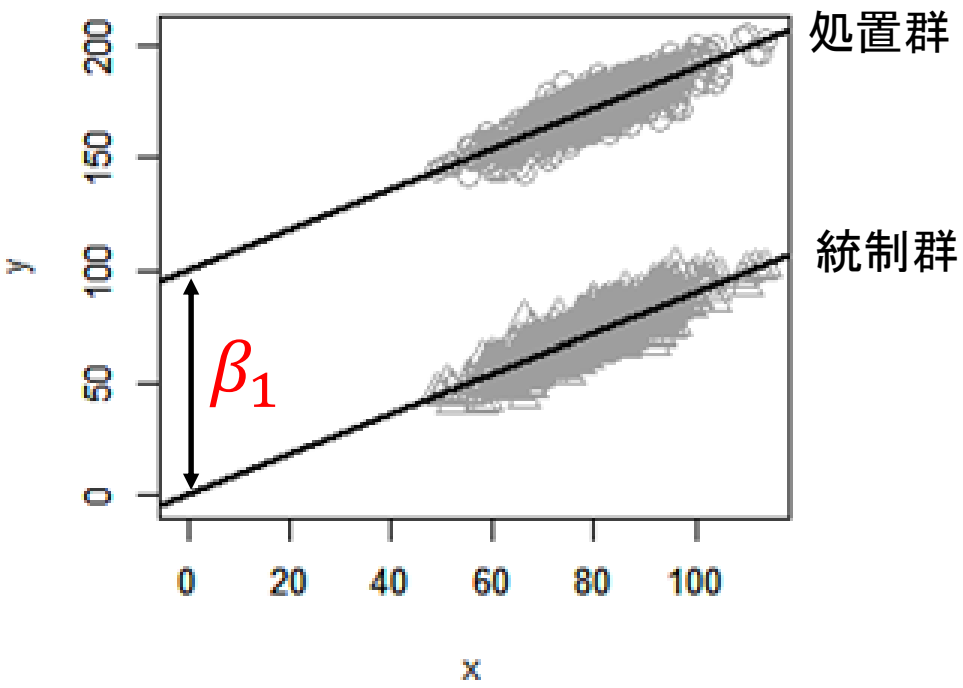
id	結果	処置	X1	X2	…	Xp
1	42	0	100	300	…	220
2	91	0	200	250	…	280
3	85	1	100	350	…	390
4	27	1	150	400	…	410
⋮	⋮	⋮	⋮	⋮	⋱	⋮
$n-1$	74	1	150	300	…	190
n	64	0	200	400	…	350

交絡を取り除くために，多数の共変量を統制する必要がある.

重回帰モデル（共分散分析）

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \varepsilon_i$$

- Y_i：結果変数
- T_i：処置の割付け
  - 0 ＝ 統制群
  - 1 ＝ 処置群
- X_i：共変量



統計的因果推論では，なぜ重回帰モデルではなく傾向スコアを使うのか？

傾向スコア（propensity score）

$$e(\boldsymbol{X}) = Pr(T_i = 1|\boldsymbol{X})$$

- 共変量$\boldsymbol{X}$が与えられたとき，処置に割付けられる確率
- 共変量$\boldsymbol{X}$を条件としたときに，$T_i = 1$となる確率

傾向スコア定理と仮定

- 傾向スコア定理

$$\{Y(1), Y(0)\} \perp T | e(\boldsymbol{X})$$

  - 処置割付けに影響を与えるのは傾向スコア$e(\boldsymbol{X})$のみ

- 無交絡性の仮定

$$\{Y(1), Y(0)\} \perp T | \boldsymbol{X}$$

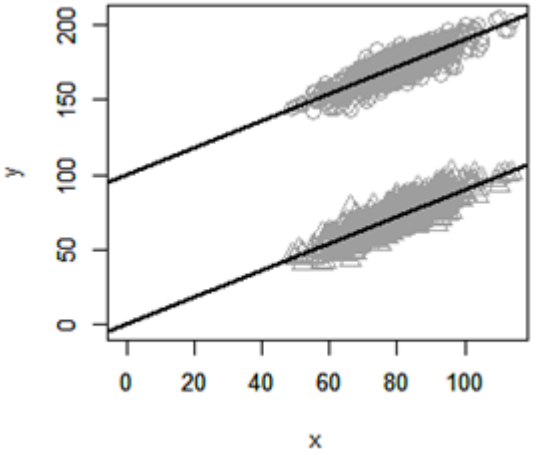
  - 処置割付けに影響を与えるのは観測された共変量のみ

無交絡性の仮定は，重回帰モデルにも必要

共分散分析の仮定？

- 共分散分析では，ガウス・マルコフの仮定に加えて，以下の仮定も満たす必要があるとされる.
  - （Jamieson, 2004; 星野, 2009; Huitema, 2011）
  - $Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \varepsilon_i$

回帰の傾きβ_2が処置群と統制群の間で共通（傾きが平行）である.



- しかし，これは半分正しく，半分正しくない.

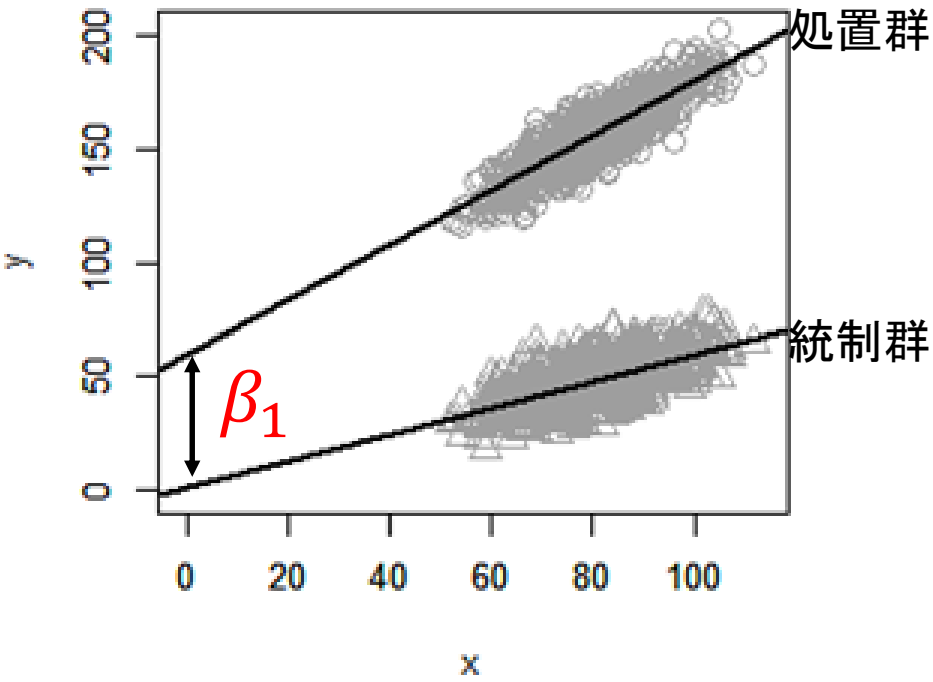
交差項（交互作用項：Interaction Term）

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 X T_i + \varepsilon_i$$

- 平均因果効果

  $\beta_1 + \beta_3 E[X_i]$

ただし，標準誤差を計算するには，β_1とβ_3の共分散を考慮に入れる必要があり，少し複雑になる．
var(A+B)
=var(A)+var(B)+2cov(A,B)



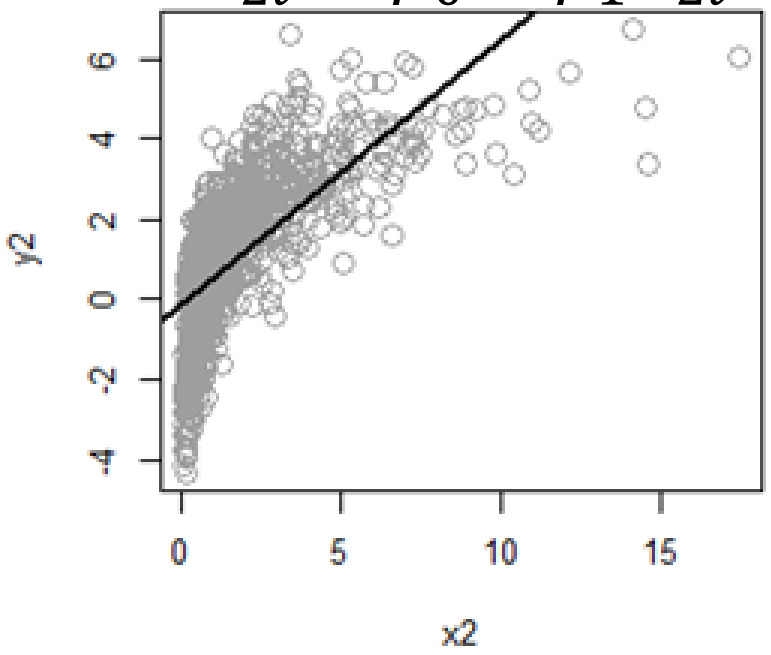
- 適切なモデリングさえできれば，共分散分析から平均因果効果を推定できる．

重回帰モデルの重要な仮定

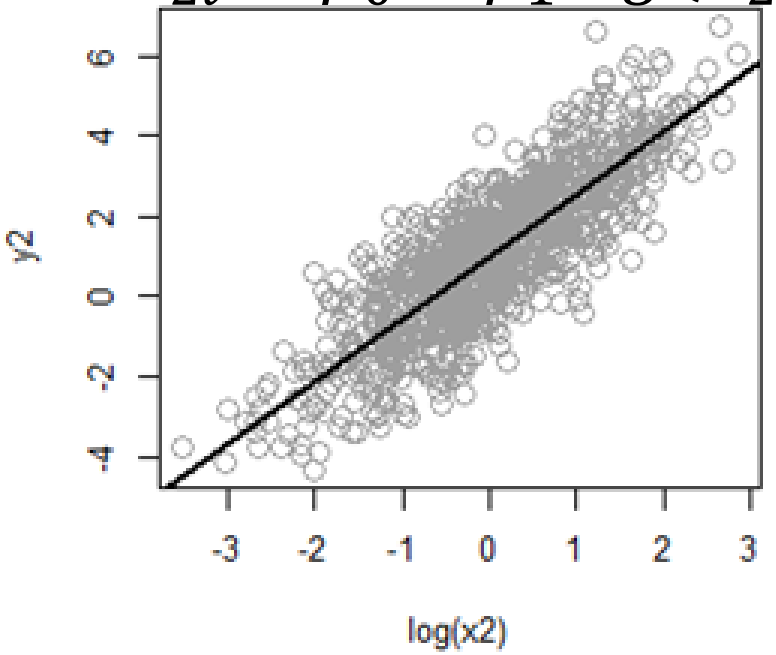
- パラメータにおける線形性

$$Y_{2i} = \beta_0 + \beta_1 log(X_{2i}) + \varepsilon_i$$

$$\hat{Y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 X_{2i}$$



$$\hat{Y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 log(X_{2i})$$



適切な関数の形を特定する必要がある.

重回帰モデルの限界

- 無交絡性の仮定

$$\{Y(1), Y(0)\} \perp T|\boldsymbol{X}$$

  - 処置割付けに影響を与えるのは観測された共変量のみ
  - この$\boldsymbol{X}$は多変量である.
  - 共変量が2個（X_1とX_2）の場合を考えよう.

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_{1i} + \beta_3 X T_{1i} + \beta_4 X_{2i} + \beta_5 X T_{2i} + \beta_6 X X_{12i} + \varepsilon_i$$

  - すべての交差項の組み合わせを考えなければならない.
  - X_1とX_2について，対数変換，平方根変換，二乗項，三乗項など多くの関数形を考慮しなければならない.

傾向スコアの利点1

- 傾向スコア

$\boldsymbol{X}$は多変量

$$e(\boldsymbol{X}) = Pr(T_i = 1|\boldsymbol{X})$$

- ロジスティック回帰モデル

$$Pr(T_i = 1|\boldsymbol{X})$$
$$= \frac{exp\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots + \beta_p X_{pi}\right)}{1 + exp\left(\beta_0 + \beta_1 X_i + \beta_2 X_{2i} \ldots + \beta_p X_{pi}\right)}$$

β_jを正しく推定することには興味がない．確率を予測できればよい．したがって，真のモデルと異なるモデルで推定したとしても，誤設定の影響が小さい．

ただし，共変量のバランシングが取れていない場合，モデルの設定を見直す必要はある．

傾向スコアの利点2

id	結果	処置	PS
1	42	0	0.4
2	91	0	0.3
3	85	1	0.8
4	27	1	0.7
⋮	⋮	⋮	⋮
$n-1$	74	1	0.6
n	64	0	0.1

交絡を取り除くために，多数の共変量を統制する必要があったが，1次元の傾向スコアに縮約できる．