応用統計学会チュートリアルセミナー
「観察データからの因果分析－共変量調整の立場から」
日程：平成18年5月26日（金）
於：国立保健医療科学院

SEMによる因果分析入門

--パス解析から傾向スコアまで--

大阪大学 大学院基礎工学研究科
狩野　裕

内　容

1. 構造方程式モデリング（SEM）とは

What is SEM?

SEMとは

- 直接観測できない潜在変数を導入し，その潜在変数と観測変数との間の因果関係を同定することにより社会現象や自然現象を理解するための統計的アプローチ
- 基本的には非実験データ（観察データ）の多変量解析で，因子分析とパス解析を統合したモデルを提供

5

SEMの特徴

- 理論に基づくモデルの検証
  - 探索的なモデリングではない
- 潜在変数
  - 誤差の分離
  - (構成)概念の測定
- 因果分析
  - パス解析

2. 回帰分析と第三変数の制御

Regression Analysis and Controlling
Third Variables

6

7

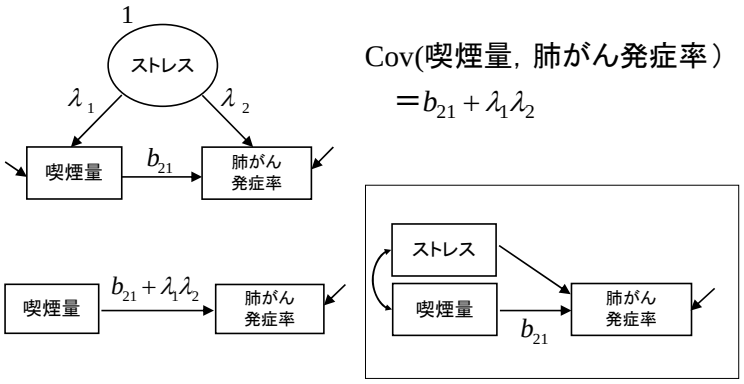
回帰分析の目的

- 予測
  - 因果とは無関係？
- 因果分析
  - 因果構造の解明
    - 変数選択
  - 因果効果の推定
    - 交絡変数のコントロール
    - 偏回帰係数：他の原因変数が一定であるときに，当該変数の変化がyへ影響する割合

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

8

交絡変数とその制御



Cov(喫煙量，肺がん発症率)
$= b_{21} + \lambda_1 \lambda_2$

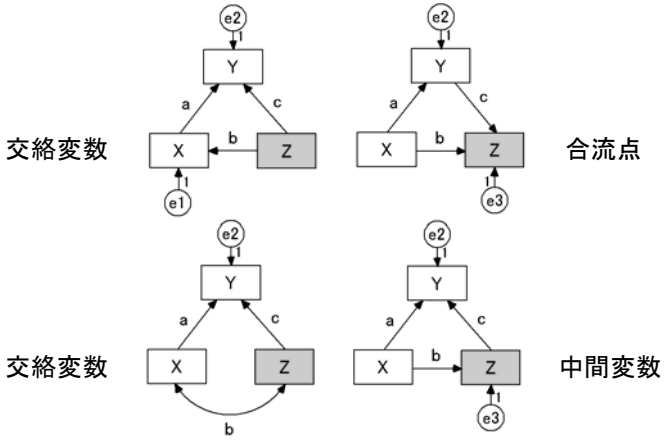
9

交絡変数と回帰分析

- 交絡変数(confounder)
  - 分野によって呼称が違う
  - 第三変数，剰余変数，二次変数，媒介変数，共変量
- 回帰分析は交絡変数の制御に利用可能
  - 交絡変数を説明変数に加える
- 回帰分析は未分析交絡変数の影響を受ける
  - 観察研究の場合（無作為割付でない場合）

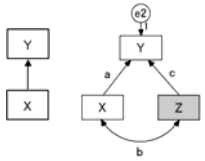
10

第三変数とは



11

回帰分析による因果推論

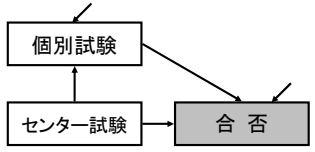


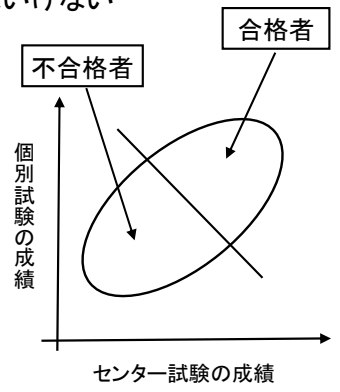
	中間変数	交絡変数	合流点
直接効果	a	a	a
総合効果	a+bc	a	a
単回帰分析	a+bc	a+bc	a
重回帰分析	a	a	≠a

12

回帰分析の御法度

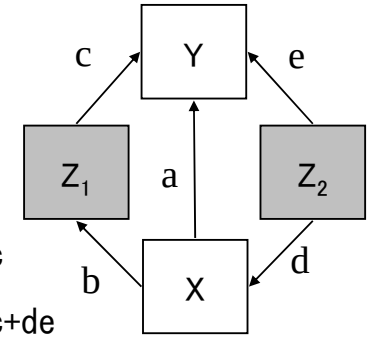
- yの結果変数で調整してはいけない
  - 予測の場合はよい
- 例
  - X: センター試験
  - Y: 個別試験
  - Z: 合否





13

複数個の第三変数



直接効果	a
総合効果	a+bc
単回帰分析X	a+bc+de
重回帰分析X, Z_1, Z_2	a
重回帰分析X, Z_2	a+bc

14

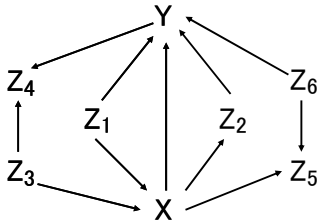
back-door criterion _1

- XからYへの総合効果を求めたいとき，コントロールすべき第三変数を同定するための条件
  - その第三変数**z**を観測
  - **z**とXとを併せて重回帰分析
- back-door criterion
  - [B1] Xから**z**への有向道がない
  - [B2] Xから出る矢線を全て除いたグラフにおいて，zがXとYを有向分離する
  - 文献　　宮川-黒木(1999, 応用統計学, p.153)

15

back-door criterion _2

- [B1] Xから**z**への有向道がない
  - 間接効果を殺さない
  - 合流点を調整しない
- [B2] Xから出る矢線を全て除いたグラフにおいて，**z**がXとYを有向分離する
  - 合流点を調整しない
  - 交絡変数を調整
- 有向分離
  - XとYを結ぶ各道において，以下のどちらかが成立
    - [D1] 合流点があるとき，zは合流点とその子孫を含まない
    - [D2] 非合流点があるとき，zは少なくとも1つの非合流点を含む



16

回帰分析からパス解析へ

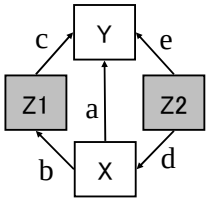
- 単回帰分析と重回帰分析を組み合せると，直接効果と総合効果を同定することが可能
  - 交絡変数の調整ができる
  - 必要な変数を観測できるという仮定
  - パス図が真の因果関係を表すという仮定
- そのためには第三変数Zの役割を正確に掴むことが必要
  - 説明変数間の関係も知る必要がある
- 従来の回帰分析よりも(SEMによる)パス解析が望ましい

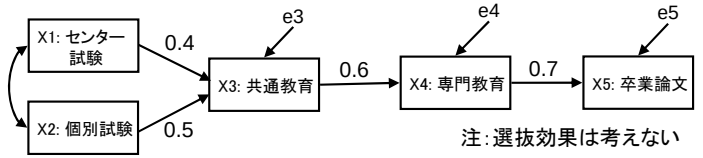
3. パス解析

Path Analysis

パス解析モデル

- (観測)変数間の因果モデル
  - 複数個の(線型)回帰モデル
- 推測
  - 適合度の吟味
  - パス係数の推定
  - 効果の分解

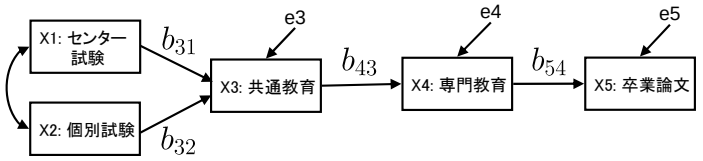




注:選抜効果は考えない

構造方程式

$$X_3 = b_{31}X_1 + b_{32}X_2 + e_3$$
$$X_4 = b_{43}X_3 + e_4$$
$$X_5 = b_{54}X_4 + e_5$$



誘導形

$$\boldsymbol{y} = B\boldsymbol{y} + \Gamma\boldsymbol{x} + \boldsymbol{e}$$

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix} = \begin{bmatrix}B & \Gamma\\ O & O\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix} + \begin{bmatrix}\boldsymbol{e}\\ \boldsymbol{x}\end{bmatrix}$$

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix} = \begin{bmatrix}I-B & -\Gamma\\ O & I\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{e}\\ \boldsymbol{x}\end{bmatrix}$$

共分散構造とパラメータ

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} I-B & -\Gamma \\ O & I \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{e} \\ \boldsymbol{x} \end{bmatrix}$$

$$V\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} I-B & -\Gamma \\ O & I \end{bmatrix}^{-1} \begin{bmatrix} V(\boldsymbol{e}) & O \\ O & V(\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} I-B' & O \\ -\Gamma' & I \end{bmatrix}^{-1} = \Sigma(\theta)$$

- 推定すべきパラメタ θ
  - パス係数
  - 独立変数の分散・共分散

統計的推測

- 尤度

$$L(\mu, \Sigma(\theta)) := \prod_{\alpha=1}^{n} N_p\left(\begin{bmatrix} \boldsymbol{y}_\alpha \\ \boldsymbol{x}_\alpha \end{bmatrix} \middle| \mu, \Sigma(\theta)\right)$$

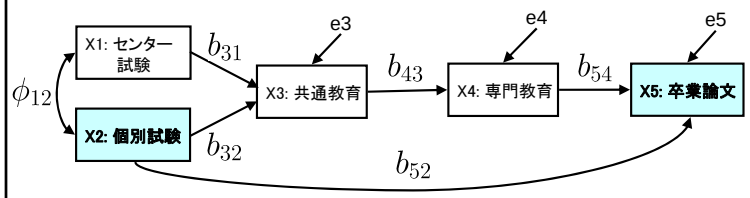
- 最尤推定

$$(\hat{\mu}, \hat{\theta}) := \underset{(\mu,\theta)}{\operatorname{argmax}}\ L(\mu, \Sigma(\theta))$$

- 適合度検定

$$H_0 : V\begin{bmatrix} \boldsymbol{y}_\alpha \\ \boldsymbol{x}_\alpha \end{bmatrix} = \Sigma(\theta) \quad \text{vs.} \quad H_1 : V\begin{bmatrix} \boldsymbol{y}_\alpha \\ \boldsymbol{x}_\alpha \end{bmatrix} = \Sigma$$

効果の分解(標準解)

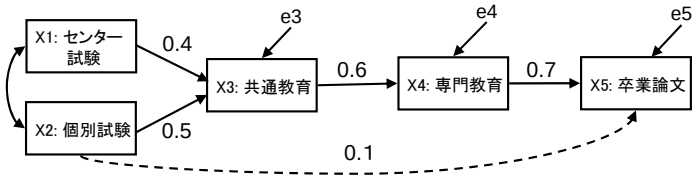


$$\mathrm{Cor}(X_2, X_5) = \underbrace{b_{52}}_{\text{直接効果}} + \underbrace{b_{32}b_{43}b_{54}}_{\text{間接効果}} \Big\} \text{総合効果} + \underbrace{\phi_{12}b_{31}b_{43}b_{54}}_{\text{擬似相関}}$$

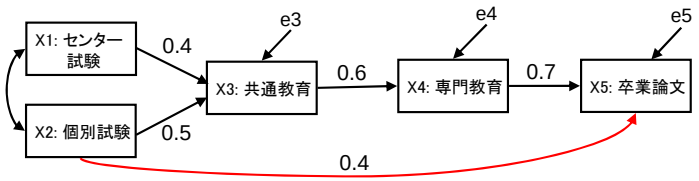
注:モデル適合が良いことが必要

多母集団同時分析 _1

A学科



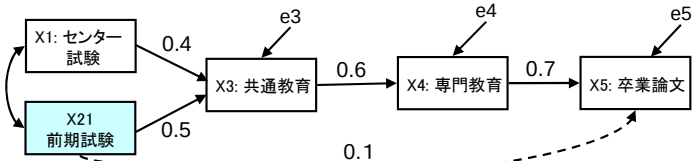
B学科



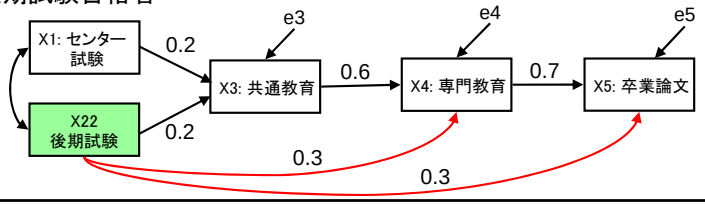
25

多母集団同時分析 _2

前期試験合格者



後期試験合格者



4. 傾向スコア

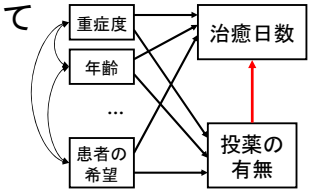
Propensity Score

26

27

セットアップ

- 調査(or 実験)研究において
  - X:二値の原因変数
  - Y:結果変数
  - $\mathbf{z}=[Z_1,Z_2,\ldots,Z_m]'$:交絡変数

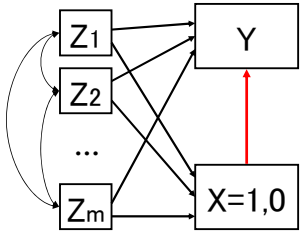


Y	Y_1	Y_2	$\cdots$	$\cdots$	Y_{n-1}	Y_n
X	0	$\cdots$	0	1	$\cdots$	1
$\mathbf{z}$	$\mathbf{z}_1$	$\mathbf{z}_2$	$\cdots$	$\cdots$	$\mathbf{z}_{n-1}$	$\mathbf{z}_n$

28

SEMで分析するとすれば

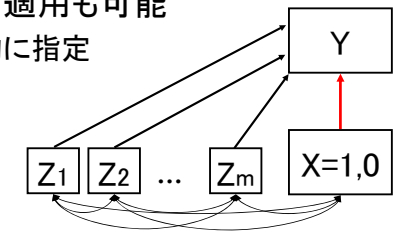
- パス解析
  - 従属二値変数をプロビット法によってモデリング
  - Yへの影響もモデリング(線型)できている



29

共分散分析である

- やや制約的なモデル
  - 「z→Y」の関係は線型
  - Xとzの交互作用はないという仮定
- 非線形モデルの適用も可能
  - モデルを明示的に指定



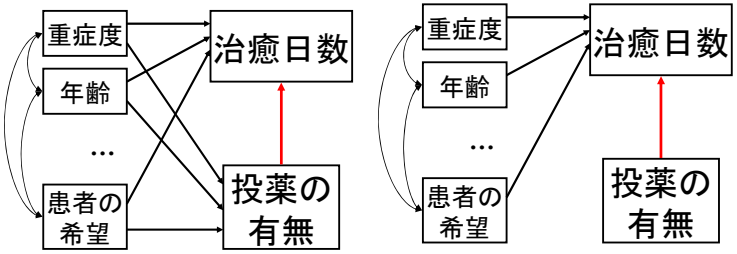
30

傾向スコアの定義

- 傾向スコア(propensity score)
  - by Rosenbaum-Rubin (Biometrika, 1983)
  - e(**z**)=P[X=1|**z**]
    - X=1を割付ける条件付確率
- 重要な性質
  - $X \perp\!\!\!\perp \mathbf{z} \mid e(\mathbf{z})$
  - e(**z**)は1次元

31

傾向スコアの性質



Conditioned on e(z)

- 「**z**→Y」の関係は線型に限らない
- 「X→Y」の関係は傾向スコアの値に依存してもよい

32

傾向スコアの利用 _1

- 交絡変数**z**が多い場合はe(**z**)の利用が有効
  - サブグループ化
    - e(**z**)の値の近い被験者をグループ化してX=0,1を比較
  - マッチング
    - e(**z**)の分布が両群で等質になるようにする
    - e(**z**)の値の近い被験者でX=0とX=1を割付けられたものを組にし，対応のあるデータの分析を行う(ペアマッチ)
  - e(**z**)を共変量とした共分散分析
  - データの重み付け

33

傾向スコアの利用 _2

- e(**z**) =P[X=1|z]の推定
  - ロジスティック回帰分析の利用
- 重要な仮定
  - Strongly ignorable
  - **z**を与えた下で，バランスがとれた割付けがなされている
  - **z**がすべての交絡要因を含んでいる

因果効果の推定と傾向スコア

34

35

データの構造と欠測

X＝0を選択した被験者　X＝1を選択した被験者

	1	⋯	m	$m+1$	⋯	n
Y_0	Y_{01}	⋯	Y_{0m}	欠　測		
Y_1	欠　測			$Y_{1,m+1}$	⋯	Y_{1n}
X	0	⋯	0	1	⋯	1
$\mathbf{z}$	$\mathbf{z}_1$	⋯	$\mathbf{z}_m$	$\mathbf{z}_{m+1}$	⋯	$\mathbf{z}_n$

z：共変量

36

データの構造と欠測

比較に意味無し

X＝0を選択した被験者　X＝1を選択した被験者

	1	⋯	m	$m+1$	⋯	n
Y_0	Y_{01}	⋯	Y_{0m}	$Y_{0,m+1}$	⋯	Y_{0n}
Y_1	Y_{11}	⋯	Y_{1m}	$Y_{1,m+1}$	⋯	Y_{1n}
X	0	⋯	0	1	⋯	1
$\mathbf{z}$	$\mathbf{z}_1$	⋯	$\mathbf{z}_m$	$\mathbf{z}_{m+1}$	⋯	$\mathbf{z}_n$

z：共変量

性質が異なる

$\begin{pmatrix} Y_{0,i} \\ Y_{1,i} \end{pmatrix}$ と X は独立に分布しない

因果効果: $E(Y_1)-E(Y_0)$

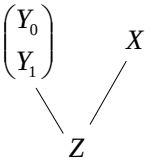
$$\left[E(Y_1 \mid X=0)-E(Y_0 \mid X=0)\right]P(X=0)$$
$$+$$
$$\left[E(Y_1 \mid X=1)-E(Y_0 \mid X=1)\right]P(X=1)$$

	X=0を選択した被験者			X=1を選択した被験者		
	1	⋯	m	$m+1$	⋯	n
Y_0	$E(Y_0 \mid X=0)$			$E(Y_0 \mid X=1)$		
Y_1	$E(Y_1 \mid X=0)$			$E(Y_1 \mid X=1)$		
X	0			1		
$\mathbf{z}$	$\mathbf{z}_0$			$\mathbf{z}_1$		

Strongly Ignorable and Estimable

- Strongly Ignorable
  - 因果効果を推定できる仮定

$$\begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix} \perp\!\!\!\perp X \mid Z$$



  - X は **z** からのみ直接的な影響を受ける
  - **z** を与えた下でバランスのとれた割付け
  - MAR
    - Xの割付を欠測と見たとき

Strongly Ignorable and Estimable

- Strongly Ignorableの下で因果効果が推定可能

$$E\left[\frac{Y_1 X}{e(\mathbf{z})}\middle|\mathbf{z}\right] = E\left[Y_1|\mathbf{z}\right]\times E\left[\frac{X}{e(\mathbf{z})}\middle|\mathbf{z}\right] = E\left[Y_1|\mathbf{z}\right]$$

$\therefore$

$$E\left[Y_1\right] = E\left[\frac{Y_1 X}{e(\mathbf{z})}\right]$$

同様にして

$$E\left[Y_0\right] = E\left[\frac{Y_0(1-X)}{1-e(\mathbf{z})}\right]$$

推定_1

	1	⋯	m	$m+1$	⋯	n
Y_0	Y_{01}	⋯	Y_{0m}	欠 測		
Y_1	欠 測			$Y_{1,m+1}$	⋯	Y_{1n}
X	0	⋯	0	1	⋯	1
$\mathbf{z}$	$\mathbf{z}_1$	⋯	$\mathbf{z}_m$	$\mathbf{z}_{m+1}$	⋯	$\mathbf{z}_n$

$$E\left[Y_1\right] = E\left[\frac{Y_1 X}{e(\mathbf{z})}\right]$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_{1i} \approx \frac{1}{n}\sum_{i=1}^{n}\frac{Y_{1i}x_i}{e(\mathbf{z}_i)}$$

$$= \frac{1}{n}\sum_{i=m+1}^{n}\frac{Y_{1i}}{e(\mathbf{z}_i)} \approx \frac{1}{\sum_{i=m+1}^{n}\frac{1}{e(\mathbf{z}_i)}}\sum_{i=m+1}^{n}\frac{Y_{1i}}{e(\mathbf{z}_i)}$$

$$\because \quad E\left[\sum_{i=m+1}^{n}\frac{1}{e(\mathbf{z}_i)}\right] = E\left[\sum_{i=1}^{n}\frac{X_i}{e(\mathbf{z}_i)}\right] = nE\left[\frac{X}{e(\mathbf{z})}\right] = n$$

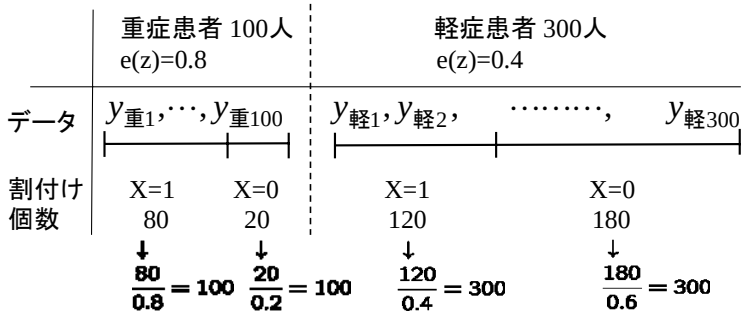
推定_2

$$\widehat{E[Y_1 - Y_0]} = E\left[\frac{Y_1 X}{e(\mathbf{z})} - \frac{Y_0(1-X)}{1-e(\mathbf{z})}\right]$$

$$= \left(\frac{1}{n}\sum_{i=m+1}^{n}\frac{Y_{1i}}{e(\mathbf{z}_i)}\right) - \left(\frac{1}{n}\sum_{i=1}^{m}\frac{Y_{0i}}{1-e(\mathbf{z}_i)}\right)$$

- Propensity score weighting
  - 傾向スコアを用いて各observationに重み付けすることで，**z**の影響を殺す
- IPTW推定
  - Inverse Probability of Treatment Weighted Estimation
  - 被験者が処置を受ける確率の逆数で重みづける

少し具体的な例



・割付けのアンバランスを交絡変数によって調整

欠測の母数を推定する

	X=0を選択した被験者			X=1を選択した被験者		
	1	⋯	m	m+1	⋯	n
Y_0	$E(Y_0 \mid X=0)$			$E(Y_0 \mid X=1)$		
Y_1	$E(Y_1 \mid X=0)$			$E(Y_1 \mid X=1)$		
X	0			1		
$\mathbf{z}$	$\mathbf{z}_0$			$\mathbf{z}_1$		

$$E[Y_1] = E\left[\frac{Y_1 X}{e(\boldsymbol{z})}\right]$$

$$\begin{aligned} E[Y_1|X=0]P(X=0) &= E[Y_1(1-X)] \\ &= E[Y_1] - E[Y_1 X] \\ &= E\left[\frac{Y_1 X}{e(\boldsymbol{z})}\right] - E[Y_1 X] \\ &= E\left[\frac{1-e(\boldsymbol{z})}{e(\boldsymbol{z})} Y_1 X\right] \end{aligned}$$

$$E[Y_1|X=0] = E\left[\frac{(1-e(\boldsymbol{z}))Y_1 X}{e(\boldsymbol{z})P(X=0)}\right]$$

欠測の母数を推定する

$$\Longrightarrow \quad E[Y_1|X=0] = E\left[\frac{(1-e(\boldsymbol{z}))Y_1 X}{e(\boldsymbol{z})P(X=0)}\right]$$

$$\widehat{E[Y_1|X=0]} = \frac{1}{n}\sum_{i=m+1}^{n}\frac{(1-e(\boldsymbol{z}_i))Y_{1i}}{e(\boldsymbol{z}_i)P(X=0)}$$

$$\approx \frac{\sum_{i=m+1}^{n}\frac{(1-e(\boldsymbol{z}_i))Y_{1i}}{e(\boldsymbol{z}_i)}}{\sum_{i=m+1}^{n}\frac{1-e(\boldsymbol{z}_i)}{e(\boldsymbol{z}_i)}}$$

45

一般の確率モデルへ（星野2005）

X＝0を選択した被験者　X＝1を選択した被験者

	1	⋯	m	$m+1$	⋯	n
Y_0	$f(y_0 \mid X=0,\theta_{00})$			$f(y_0 \mid X=1,\theta_{10})$		
Y_1	$f(y_1 \mid X=0,\theta_{01})$			$f(y_1 \mid X=1,\theta_{11})$		
X	0			1		
$\mathbf{z}$	$\mathbf{z}_0$			$\mathbf{z}_1$		

$\mathbf{z}$：共変量

性質が異なる

$\begin{pmatrix} Y_{0i} \\ Y_{1i} \end{pmatrix}$ と X は独立に分布しない

46

一般の確率モデルへ

$$E[Y_1|X=0] = E\left[\frac{(1-e(\boldsymbol{z}))Y_1X}{e(\boldsymbol{z})P(X=0)}\right]$$

$$\Longrightarrow$$

$$E[h(Y_1)|X=0] = E\left[\frac{(1-e(\boldsymbol{z}))h(Y_1)X}{e(\boldsymbol{z})P(X=0)}\right]$$

$$\Longrightarrow$$

$$E[\partial \log f(Y_1|\theta_{01})|X=0] = E\left[\frac{(1-e(\boldsymbol{z}))\partial \log f(Y_1|\theta_{01})X}{e(\boldsymbol{z})P(X=0)}\right]$$

$$\Longrightarrow$$

$$\sum_{i=m+1}^{n}\left[\frac{(1-e(\boldsymbol{z}_i))\partial \log f(Y_{1i}|\theta_{01})}{e(\boldsymbol{z}_i)P(X=0)}\right] = 0$$

θ_{01} についての推定方程式

47

正規母集団の場合

推定方程式

$$\sum_{i=m+1}^{n}\left[\frac{(1-e(\boldsymbol{z}_i))\partial \log f(Y_{1i}|\boldsymbol{\theta}_{01})}{e(\boldsymbol{z}_i)P(X=0)}\right] = 0$$

正規母集団

$$f(y_1|\boldsymbol{\theta}_{01}) = N(y_1|\theta_{01},\sigma^2)$$

推定量

$$\hat{\theta}_{01} = \sum_{i=m+1}^{n}\frac{(1-e(\boldsymbol{z}_i))Y_{1i}}{e(\boldsymbol{z}_i)} \Big/ \sum_{i=m+1}^{n}\frac{1-e(\boldsymbol{z}_i)}{e(\boldsymbol{z}_i)}$$

5. まとめ

Summary

48

SEMについて

- パス解析を用いて検証的因果推論
  - 適合度，パス係数の有意性検定
  - 効果の分解
  - 回帰モデルよりも，因果関係を素直に表現できるパス解析に優位性
- 基本的には線形モデル
  - 非線型項を扱うSEMもある
  - 非線型項を明示的にモデリング

傾向スコア

- 傾向スコア
  - 傾向スコアは交絡変数zと割付変数Xの関係を切る
  - 高次元の交絡変数zを1次元に落とす
    - マッチングやサブグループ化を容易にする
  - zからYへのモデリングが不要
    - 適切にモデリングできるなら，した方が良い
  - 傾向スコアによって重み付けする方法も有効
    - 広く適用できる可能性（星野他）
- SEM
  - zを調整する基本モデルを提供
    - 共分散分析

参考文献

- Bollen, K. A. (1989). Structural Equations with Latent Variables. Wiley.
- Bullock, H. E., Harlow, L. L. & Mulaik, S. A. (1994). Causal issues in structural equation modeling research. Structural Equation Modeling, 1, 253–267.
- Holland, P. W. (1986). Statistics and causal inference (with discussion). Journal of the American Statistical Association, 81, 945–970.
- Hirano, K., Imbens, G. & Ridder, G. (2003). Efficient estimation of average treatment effect using the estimated propensity score. Econometrica, 71, 1161–1189.
- Mulaik, S. A. & James, L. R. (1995). Objectivity and reasoning in science and structural equation modeling. In Structural Equation Modeling: Concepts, Issues, and Applications, (Hoyle, H., Ed.), pp.118–137. Sage Publications: CA.
- Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.
- Rosenbaum, P. R. (2002). Observational Studies. 2nd ed. Springer.

- 岩崎 学（2002）. 不完全データの統計解析. エコノミスト社
- 狩野 裕（2002）.「構造方程式モデリング，因果推論，そして非正規性」竹内啓（編著）多変量解析の展開 Part II. 岩波書店
- 佐藤俊哉・松山 裕（2002）.「疫学・臨床研究における因果推論」竹内啓（編著）多変量解析の展開 Part III. 岩波書店
- 竹内啓(1986). 因果関係と統計的方法. 行動計量学, 14, 85–90.
- 豊田秀樹(1998). 共分散構造分析[入門編]. 朝倉書店
- 星野・繁桝（2004）. 傾向スコア解析法による因果効果の推定と調査データの調整について. 行動計量学, 31, 43–61.
- 星野崇宏(2005). 欠測群の周辺分布の母数に対する傾向スコアを用いた重み付きM推定量の提案と介入効果研究への応用. 行動計量学, 32, 121–132.
- 宮川雅巳（1997）. グラフィカルモデリング. 朝倉書店
- 宮川雅巳（2004）. 統計的因果推論. 朝倉書店
- 宮川・黒木(1999). 因果ダイアグラムにおける介入効果推定のための共変量選択. 応用統計学, 28, 151–162.

後註

欠測の母数を推定する

$$E[Y_1|X=0] = E\left[\frac{(1-e(z))Y_1X}{e(z)P(X=0)}\right]$$

$$\Longrightarrow$$

$$\widehat{E[Y_1|X=0]} = \frac{1}{n}\sum_{i=m+1}^{n}\frac{(1-e(z_i))Y_{1i}}{e(z_i)P(X=0)}$$

$$E\left[\sum_{i=m+1}^{n}\frac{1-e(z_i)}{e(z_i)}\right]$$
$$=E\left[\sum_{i=1}^{n}\frac{1-e(z_i)}{e(z_i)}X_i\right]$$
$$=nE\left[\frac{1-e(z)}{e(z)}X\right]=nE\left[\frac{X}{e(z)}-X\right]$$
$$=n\Big(1-P(X=1)\Big)=nP(X=0)$$

$$\approx \frac{\sum_{i=m+1}^{n}\frac{(1-e(z_i))Y_{1i}}{e(z_i)}}{\sum_{i=m+1}^{n}\frac{1-e(z_i)}{e(z_i)}}$$