

いまさら聞けない基礎統計学3 ～初めての多変量解析

ダイバーシティ推進室

衛生学・予防医学講座

各務竹康

今回の目的

- ▶ 基本的な多変量解析の仕組みについて学ぶ
- ▶ 多変量解析の結果を解釈できる

多変量解析とは

- ▶ 単変量解析：要因Aと要因Bの関連を分析する
それぞれの裏側にあるものはすべて無視！
- ▶ 特に「人の集団を対象とした研究」では、他の要因を無視できない
年齢、性別、身長、体重、その他・・・・・・・・
- ▶ 全く「同一」の対照集団を設定することは無理（双子でも性格の違い）
- ▶ “無視できない”要因も考慮した上で比較を行う

多変量解析

ざっくりとした概念図

A~Dは要因と結果の関係に何

A~Dは要因と結果の関係に何らかの影響を与える(結果を捻じ曲げる)→交絡因子

要因

結果

A

B



D

A~Dの関与を排除
→多変量解析

たまにある勘違い

- ▶ 多変量解析≠多重比較
- ▶ 多重比較：一つのアウトカムに対する変数同士の差を調べる
 - A群とB群の年齢差を調べる→t検定
 - A群、B群、C群の年齢差を調べる→多重比較(Bonferroni検定)
 - 3群間に差があるか調べる検定(ANOVAなど)は多重比較ではない
- ▶ 多変量解析：一つのアウトカム変数に対する複数の因子の相互作用を取り除いた影響を調べる

よく用いる多変量解析の手法

- ▶ 重回帰分析
- ▶ ロジスティック回帰分析
- ▶ Cox回帰分析

大切なこと

- ▶ 求めるのは「関連」ではなく「因果関係」
- ▶ $A \rightarrow B$ の式が成り立つことが前提
意外と“見逃された発表”が散見される
- ▶ まずは「因果関係が成り立つか」を吟味する！！！！

“関連”と“因果”

- ▶ “関連”：時間的な関係は問わない(むしろ同時期)

飲酒習慣のある人は喫煙者が多い

喫煙→飲酒、飲酒→喫煙のどちらもありえる

- ▶ “因果”：“原因”と“結果”

必ず結果は原因の後にくる(逆転はあり得ない)

歩行習慣のある人は死亡率が低い

因果の逆転に注意

- ▶ “原因”と“結果”の逆転は起こりうる(特に横断研究)

＊横断研究：一時点での情報で行う研究

- ▶ 糖尿病の人は運動習慣が多い
- ▶ 高血圧の人は摂取塩分が少ない

どう捉えますか？

“経路”を考える

要因



結果

医学研究における多変量解析の2つの目的

- ▶ 交絡因子の調整

立証したい因果関係に対して様々な交絡因子の影響を除外する
デザイン段階でどうしても調整できない因子を投入

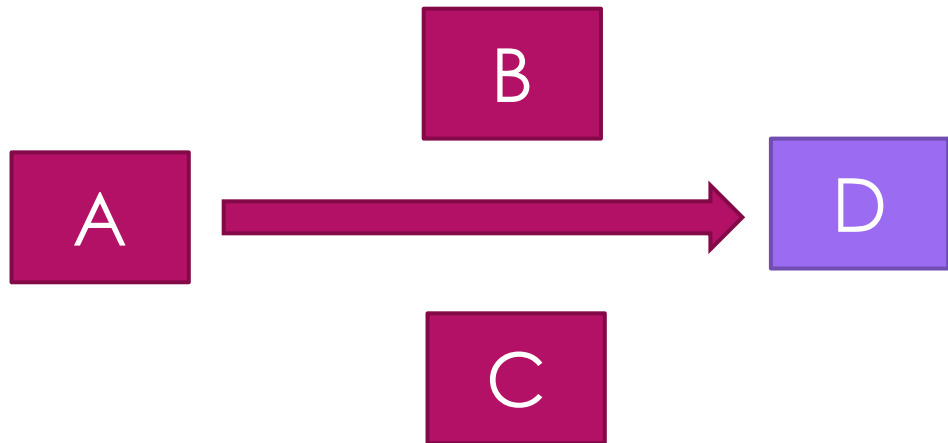
- ▶ 変数の予測

様々な要因の中から“真に”影響を与える要因を探索する
どの因子が有意な関連であるのかを調べる

その他、目的変数の予測モデルを作成することも

多変量解析の考え方1

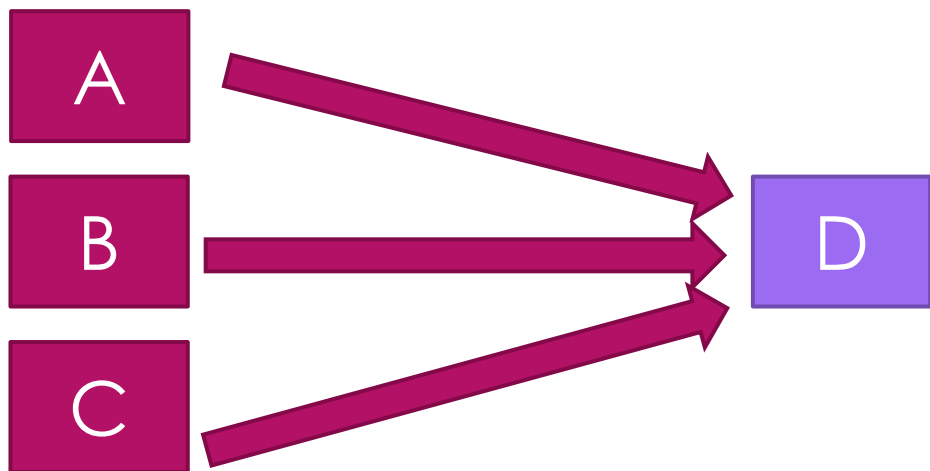
- ▶ $A \rightarrow D$ の関係が真に意味のあるものなのか、他に影響を与える要因を統計モデルに投入してその結果を求める



- ▶ B、C → 交絡因子

多変量解析の考え方2

- ▶ ABCのどの因子が真にDに影響を与えるのか明らかにする



交絡が成り立つのか？

- ▶ 交絡が成り立つ3つの条件
- ▶ 交絡因子が疾病(結果)発生の危険因子である
- ▶ 交絡因子が曝露と関係ある
- ▶ 交絡因子が曝露と疾病(結果)の中間因子でない

性・年齢は特に理由がなければ交絡因子として扱う

多国籍国家、多国間研究では人種も

交絡因子が疾病(結果)発生の危険因子である

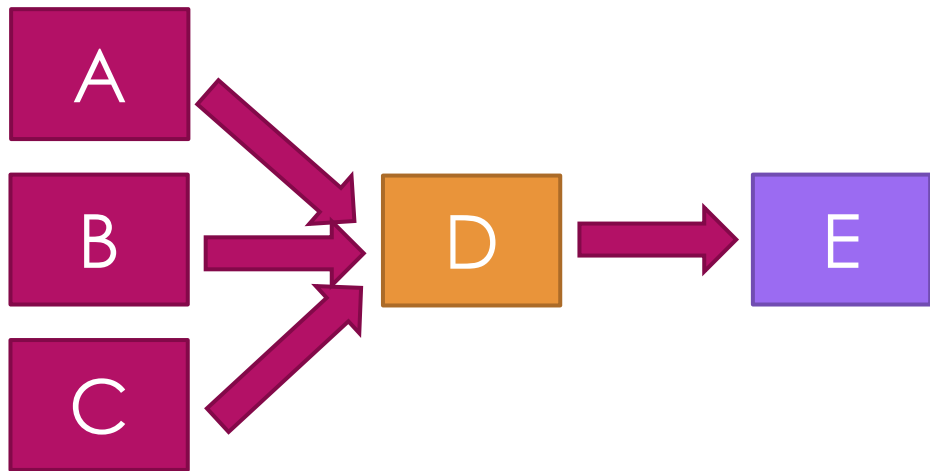
- ▶ 喫煙と癌の関係を調べる場合
- ▶ 癌の発症に関係ある(と思われる)要因を含める
 - 先行研究
 - 経験則
 - その他

交絡因子が曝露と関係ある

- ▶ 結果の危険因子であっても曝露と関係がない(独立)の場合は交絡因子でない
独立の要因は結果に影響を与えない

交絡因子が曝露と疾病(結果)の中間因子でない

- ▶ この見極めが重要(かつ難しい)

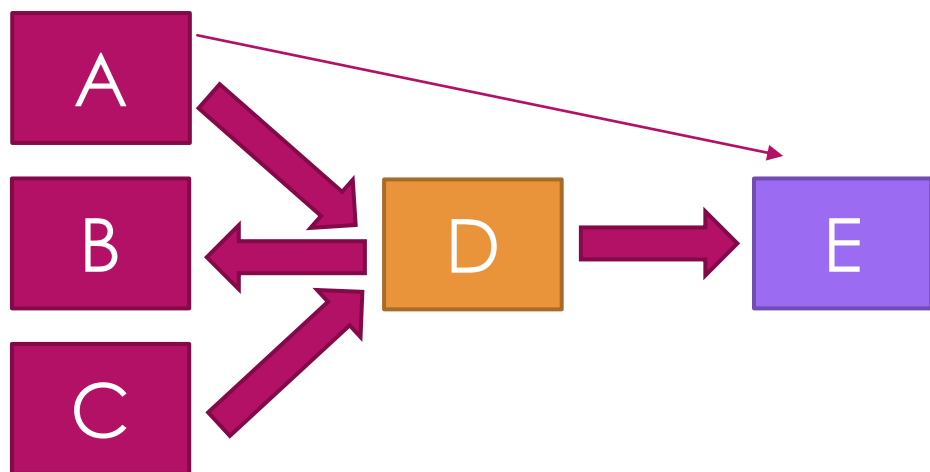


ABCはそれぞれDの原因となる(曝露)
ABCはそれぞれ関連がある(交絡)
ABCはDを介してEの原因となる(Dが中間因子)

中間因子をどうやって見極めるか？

- ▶ 過去の文献などにより、交絡因子なのか、中間因子なのか見極める
- ▶ 生理学的、生化学的機序から推測することも可能

パス解析



- ▶ 様々な因子の関連を経路として示す
- ▶ 経路を「解析者が」描いていることが前提(統計ソフトはその経路の確からしさ、影響の程度を計算する。

多変量解析 実践の際に

用語解説

- ▶ 目的変数(従属変数)：アウトカムとなる因子。この因子を決定する他の因子の関与について調べる。
- ▶ 説明変数(独立変数)：アウトカムに影響を与える因子。他の因子の影響を受けることなく、互いに独立(連動しない)関係にある。
- ▶ 信頼区間：求めた影響の度合いがどの程度の幅で変化しうるか。一般的には95%信頼区間を用いる。この信頼区間の範囲で結果が有意であるか決まる。

最初に確認すること

- ▶ 多変量解析に耐えられるデータであるのか？
- ▶ サンプルサイズ
- ▶ アウトカム変数
- ▶ その他の変数

サンプルサイズの方

▶ 交絡因子1因子につきアウトカムイベント数10例 (線形回帰モデルでは15例)

▶ アウトカムイベントの発生率が10%

▶ 交絡因子は最低5因子含めたい

▶ 必要アウトカムイベント数 $5 \times 10 = 50$ 例

▶ 必要サンプル数 $50 \times 10 = 500$ 例

年齢、性別(多国籍研究では人種も)
はよほどの理由がない限り交絡因子として扱う

よく用いる多変量解析の手法

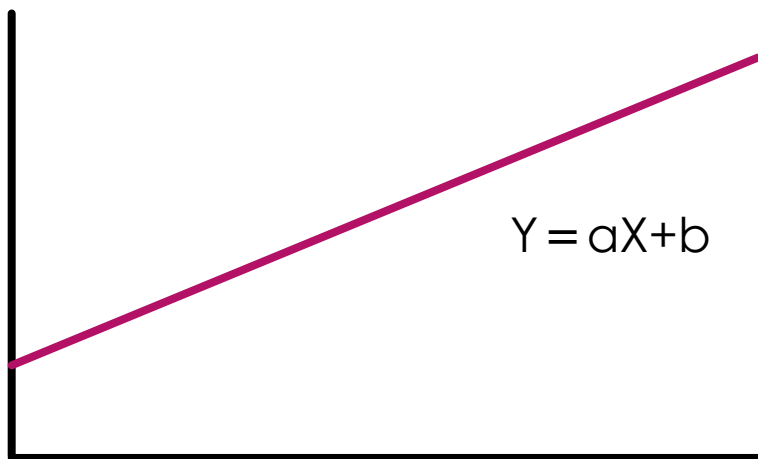
- ▶ 重回帰分析
- ▶ ロジスティック回帰分析
- ▶ Cox回帰分析

統計ソフトでは

- ▶ 目的変数、説明変数をそれぞれ入力
- ▶ Cox回帰は説明変数は2つ(イベントの有無、観察期間)、その他は共変量(交絡因子)として扱う
- ▶ 説明変数(交絡因子)の選択が悩みどころ

重回帰分析1

- ▶ 目的変数が連続変数(比率尺度、もしくは間隔尺度)
順序変数は対象にならない
- ▶ 説明変数は連続変数、カテゴリ変数どちらでも可
順序変数はカテゴリ化することが多い



重回帰分析2

- ▶ 求めるものは“偏回帰係数”(coefficient もしくはBと表現される)

$y=ax+b$ のaの部分

xが1増加した場合、yがどの程度変化するのか

- ▶ 偏回帰係数は、正、負どちらの関係もある
- ▶ 偏回帰係数の95%信頼区間が0をまたがない場合、その結果を有意と解釈できる
- ▶ 偏回帰係数の大きさ(どの要因が最も影響を与える)を比較することは出来ない
因子同士の比較は標準化偏回帰係数(β)を用いる

ロジスティック回帰分析1

- ▶ 目的変数が連続変数(比率尺度、もしくは間隔尺度)

順序変数は対象にならない

- ▶ 説明変数は連続変数、カテゴリ変数どちらでも可

順序変数はカテゴリ化することが多い

連続変数(年齢など)をグループ化するかどうかで結果が大きく変わる

		結果	
		あり	なし
要因	あり	a	b
	なし	c	d

$$\text{オッズ比} = \frac{a \times d}{b \times c}$$

ロジスティック回帰分析2

- ▶ 求めるものは“オッズ比”
- ▶ 説明変数が“あり”となった場合目的変数が“あり”となるオッズが○倍変化する
 - オッズ比が1より大きい→説明変数がアウトカムイベント出現の促進因子である
 - オッズ比が1より小さい→説明変数がアウトカムイベント出現の抑制因子である
- ▶ オッズ比の95%信頼区間が1をまたがない場合、その結果が有意であると解釈できる

オッズ比について1

- ▶ オッズ比は、相対危険の近似値として用いる
相対危険：要因がある場合、ない場合と比べてどれくらいリスクが上昇するか
- ▶ 本当に求めたい相対危険 = リスク比(コホート研究のみで適応可能)
- ▶ オッズ比10：要因がある場合イベント発生の可能性が10倍上昇

オッズ比について2

- ▶ 説明変数が連続変数の場合解釈が難しい
- ▶ オッズ比10→変数が1変化した場合オッズが10倍上昇する
2変化した場合オッズ比が 10×2 で20倍上昇するとの解釈にはならない
- ▶ オッズ比10のときの係数 $= \ln(10) \doteq 2.302$
- ▶ 変数が10変化した場合 $OR_{10} = \exp(10 \times 2.302) \doteq 10^{10}$
連続変数の場合、オッズ比は低くなる傾向にある

連続変数を説明変数に投入する場合の注意点

- ▶ 直線的にオッズが変化する変数のみを投入する
- ▶ 例：年齢はある地点で発症率が急上昇
 - 実は連続変数のまま投入することは不自然なことが多い
- ▶ 解決策1：リスクが上昇する年齢で区切って2値変数とする
- ▶ 解決策2：年齢を5歳、もしくは10歳刻みで年齢群としてダミー変数化

ダミー変数

- ▶ 3種類の変数(A, B, C)がある場合
2種類の変数(X1, X2)を作成

- ▶ X1 : Aに対するBのオッズ比
- ▶ X2 : Aに対するCのオッズ比

	X1	X2
A	0	0
B	1	0
C	0	1

- ▶ どの変数をリファレンスとして設定するのか、事前に十分に吟味すること

Cox回帰分析2

- ▶ 目的変数がイベント発生日(追跡期間)
- ▶ 主要な説明変数はカテゴリ変数
- ▶ 単変量解析：カプランマイヤー法

Cox回帰分析

- ▶ 求めるものは“ハザード比”
- ▶ 要因のあった場合のハザードがどの程度上昇するかを計算
Cox比例ハザードモデルとも呼ばれる
- ▶ オッズ比同様、ハザード比の95%信頼区間を計算
- ▶ ハザード比の95%信頼区間が1をまたがない場合、その結果が有意であると解釈できる

交絡因子の選びかた

- ▶ 性、年齢はよほどの理由がない限り交絡因子として扱う(査読で必ずツッコまれる。単変量解析で有意でない、は理由にならない)
- ▶ 先行研究で多く投入されている変数も含める
- ▶ 変数をソフトに委ねることは近年推奨されない

因子の選択方法

- ▶ 事前に先行文献等で当たりをつける

統計ソフトのアルゴリズム

▶ ステップワイズ法を例に

	AIC
None	100
-A	90
-B	80
-C	110
-D	120
-E	150



-B

	AIC
None	80
-A	100
+B	100
-C	120
-D	70
-E	130



-D

	AIC
None	70
-A	110
+B	120
-C	130
+D	80
-E	140



投入する変数は
A, C, Eに決定

ソフトに委ねる危険性

- ▶ 必ず投入すべき要因を排除してしまう可能性
SPSSだと強制投入変数を指定できる
- ▶ ダミー変数にした因子の一部を排除する可能性
- ▶ 一般化の問題点

多重共線性(交互作用)

- ▶ 複数の因子が説明変数に対して同じ方向に影響を与える
- ▶ 多重共線性があると
 - 一定の方向に結果を歪める力が増強
 - 結果、他の因子の影響も歪んでしまう

多重共線性の確認

- ▶ 現在はVIF (Variance Inflation Factor) を用いることが多い
VIFの値が大きい→他の因子との作用が大きい
- ▶ VIFの基準は10を超えないこと、とされているものが多いが、それより低い値でも問題となることもある
- ▶ VIFが大きい変数があった場合
2つの変数が同じくらいの値→どちらかを削除
3つ以上の変数がそれぞれ大きな値→大きい変数から順に削除

結果の記載例

	Crude OR (95%CI)	Adjusted OR (95%CI)
因子A	1.503 (1.401-2.345)	1.326 (1.208-1.895)
因子B	2.254 (1.278-3.554)	2.044 (1.195-3.334)
因子C	0.980 (0.501-1.124)	0.889 (0.408-1.081)
因子D	1.214 (1.084-3.681)	1.058 (0.576-2.567)

ロジスティック回帰分析
Cox回帰分析ではORがHRに変化
雑誌により右にp-valueを記載することもある

	coefficient (95%CI)	β
因子A	1.156 (0.985, 1.256)	2.554
因子B	0.851 (-0.095, 1.165)	0.985
因子C	-1.258 (-3.419, -0.594)	-1.654
因子D	2.597 (1.168, 4.661)	1.996

重回帰分析
雑誌により右にp-valueを記載することもある

まとめ

- ▶ 最も大切→多変量解析に耐えられるデータか
- ▶ デザイン段階で多変量解析をイメージして情報収集を行う
- ▶ 変数の選択は過去の文献をひたすら辿ること
 - 機械的な決定は近年推奨されない傾向
- ▶ 最初に実施する際は、必ず誰かの指導を仰ぐ！
- ▶ いつでも質問受け付けます